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

Novel proteomic approaches to study gene regulatory interactions

Alexander Reim

aus

Erlangen, Deutschland

2021

Erklärung

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

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

Grasbrunn, 21.04.2021

Alexander Reim

Dissertation eingereicht am:	22.02.2021
1. Gutachterin/Gutachter:	Prof. Dr. Matthias Mann
2. Gutachterin/Gutachter:	Prof. Dr. Elena Conti
Mündliche Prüfung am:	14.04.2021

Table of contents

Abstract

The genome encodes the building blocks of every cell. Proper transcription and translation of the genome into its functional entities, i.e. proteins, is key to every organism's survival and proliferation. Therefore, sophisticated mechanisms have evolved to ensure proper transcriptional and translational regulation. Protein-protein, protein-DNA and protein-RNA interactions are crucial parts of these regulatory networks.

Given their importance, it has been a longstanding interest in the field to characterize these interactions. Researchers have continuously developed and optimized methods to improve analyses of transcriptional and post-transcriptional regulatory interactions. In this regard, mass spectrometry (MS) has emerged to become a powerful tool to investigate almost every aspect in protein science including protein-protein and protein-nucleic acids interactions.

In this dissertation, I aimed at using mass spectrometry to study different features of transcriptional regulatory protein interactions. I developed a mass spectrometry UV laser crosslinking pipeline to localize regions in proteins binding to DNA. In a collaborative effort, I further employed UV crosslinking and mass spectrometry to define the RNA-binding proteome in T helper cells. Moreover, I set up a formaldehyde-crosslinking chromatin immunoprecipitation mass spectrometry (ChIP-MS) workflow to map the interactomes of 104 yeast transcription factors resulting in the identification of novel chromatin-associated proteins and regulatory interactions. I also successfully transferred this workflow to brain tissue of mice in a collaborative endeavour to characterize the transcription factor Tbx3 and its role in body weight regulation. Finally, I used my knowledge on mass-spectrometry based protein interactomics to dissect the dual functionalities of two different chromatin-associated *D. melanogaster* Domino isoforms defined by their distinctive interactomes.

1 Introduction

1.1 Mass spectrometry-based proteomics

Understanding the global state of the genome, the transcriptome and the proteome is pivotal for researchers to describe cellular development, maintenance and disease mechanisms. The advancements of technologies over the past decades have allowed researchers to sequence the human genome [1] and analyze large parts of the transcriptome at any given point in time [2]. The complexity drastically increases from studying genomes to the entirety of expressed proteins, i.e. proteomes. Therefore, the characterization of whole proteomes long remained just an aspiration for researchers. This began to change with the introduction of mass spectrometry to the field of protein science.

What prevented the use of MS in proteomics in the early days was the difficulty of ionizing relatively labile molecules like proteins and peptides while keeping them intact. This obstacle was overcome by the development of matrix-assisted laser desorption ionization (MALDI) [3] and electrospray ionization (ESI) [4]. In ESI-MS a high voltage is applied between the tip of the chromatographic column and the inlet of the mass spectrometer to attract the analyte and disperse the solvent droplets. The tremendous impact that electrospray ionization had on protein analysis by mass spectrometry was later awarded with a share of the nobel prize to John B. Fenn in 2002.

Ever since, the advancements in mass spectrometry-based proteomics were substantial. Various types of mass spectrometers, different chromatography instruments and methodological approaches are available today. The projects in this thesis used bottom-up proteomics on Thermo Fisher Scientific Orbitrap mass spectrometers for protein analysis. For this reason, I will mainly focus on introducing these methods and instruments. Other techniques will also be briefly described for comparison.

1.1.1 Protein analysis methods in mass spectrometry

In MS-based proteomics a distinction has to be made between bottom-up, top-down and middle-down approaches. The fundamental difference lies in the digestion of proteins prior to injection into the MS (bottom-up and middle-down proteomics) or the direct analysis of intact proteins (top-down proteomics, Figure 1.1).

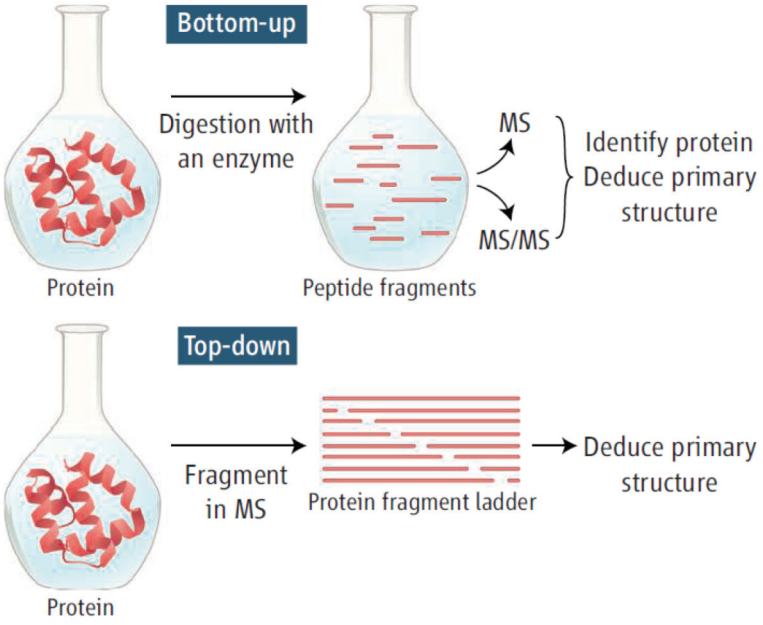


Figure 1.1: Differences between bottom-up and top-down mass spectrometry experiments. Proteins are digested to short peptides before analysis in bottom-up proteomics. They are identified based on the MS and MS/MS data from precursor and product ions. In top-down approaches proteins are directly injected to the mass spectrometer and identified from the fragmentation patterns. Adapted from [5].

In bottom-up proteomics, also known as shotgun-proteomics, proteins are subjected to enzymatic digestion at specific cleavage sites prior to analysis. Typically trypsin is used which cuts C-terminally of every lysine or arginine residue. Subsequent MS analysis acquires mass spectra of the peptides. Comparison of the experimental data to protein sequence databases allows identification of the proteins present in the sample [5]. Of course, not the entirety of peptides will be captured, which limits the applicability of shotgun approaches in the

identification of specific protein isoforms.
In these cases, the top-down approach can be advantageous. Top-down describes the injection of intact proteins which are fragmented in the mass spectrometer. Information on the analyte is obtained from both the molecular mass of the protein and the fragmented ions [5]. The missing digestion step creates a key advantage over shotgun proteomics as it solves the 'protein inference problem' known from peptide-centric approaches. This describes the issue that proteins often exist in different isoforms. Bottom-up approaches cannot unambiguously pinpoint the specific splice form unless peptides are detected that are unique to the individual proteoforms [6]. The same may hold true for peptide modifications [5].
Considering these issues it would be appealing to use top-down proteomics, which distinguishes between proteoforms and may detect all modifications on the protein. However, the solubility of intact proteins is much more variable than of peptides. This impedes the prefractionation of samples compared to bottom-up proteomics [7]. Moreover, particularly when it comes to global analyses of proteomes, top-down proteomics has many practical difficulties [7]. Hence, bottom-up proteomics is still the method of choice even in the analysis of protein modifications, mainly because of its ease of use and the ability to study thousands of proteins in one go.
More recently, a compromise between top-down and bottom-up proteomics emerged, which was termed middle-down proteomics. It also encompasses a protein digestion step, but typically uses the proteases Asp-N and Glu-C which produce longer peptides than trypsin. Middle-down proteomics is still a niche application, as it is mostly used to study proteins with co-existing modifications [8]. This is predominantly the case for chromatin-associated proteins, and particularly for histones. The basic N-terminal tail carries most of the modifications on histones. Modifications of lysines like propionylation can occur, which create missed cleavages in conventional bottom-up proteomics [9]. In contrast, Glu-C, which cleaves C-terminally of glutamic acids and Asp-N, which cleaves N-terminally of aspartic acids, will keep the histone tail intact [8] and is not affected by lysine propionylation. Middle-down proteomics was successfully used in chromatin biology. It characterized 233 histone H4 proteoforms in breast cancer cell lines [10]. H4 modifications were monitored during

the cell cycle and several intriguing changes in phosphorylation, methylation and acetylation of the N-terminal tail were discovered. For instance, an increase of H4 Serine-1 phosphorylation in S to G2/M phase transition indicated a crucial role of this modification in mitotic chromatin condensation [10].
Despite its use in studying combinatorial modifications on histones [8], middle-down proteomics still remains a niche technology. This is due to the much more complicated data analysis compared to shotgun proteomics as combinatorial PTMs on peptides exponentially increase data complexity [8]. Yet, in answering specialized biological question, e.g. the crosstalk of PTMs on histones, middle-down proteomics has become a valuable tool and needs to be considered in the experimental design.

1.1.2 Sample preparation in shotgun proteomics

Sample preparation is a central part of any mass-spectrometry based proteomics workflow. The approaches towards efficient protein digestion and purification of peptides before MS analysis can essentially be divided into 'in-gel' and 'in-solution' digestion. Both approaches require disruption of disulfide bridges in order to avoid missed cleavages. To this end, cysteines are reduced by the addition of dithiothreitol (DTT), tris(2-carboxyethyl)phosphine (TCEP) or other reducing agents. Free thiols are alkylated with iodoacetamide (IAA) or chloroacetamide (CAA) to prevent reassembly of disulfide bridges. TCEP may be added together with CAA, while DTT needs to be incubated with the sample before addition of the alkylation reagent to prevent cross-reactions [11]. Proteolytic digestion is usually carried out using trypsin. Lys-C may be added for improved digestion of tightly folded proteins [12]. Some biological questions, e.g. the study of specific modifications, need different enzymes than trypsin and were discussed in section 1.1.1.
'In-gel' protocols require solubilization of proteins with detergents and separate them by SDS–polyacrylamide gel electrophoresis (SDS-PAGE). Excision of individual gel bands and immediate digestion with trypsin allows MS analysis of separated proteins [13]. Measurement of individual gel slices has the advantage of improving the dynamic range, which is defined by the ratio of the lowest to the highest abundant proteins in MS experiments. In addition, gel separation removes low molecular weight contaminants [14]. The major con-

cerns with in-gel digestion are incomplete peptide recovery and the difficulty of implementing it in automated, high-throughput workflows.
In my projects I relied on 'in-solution' digestion given its high peptide recovery and compatibility with high-throughput workflows. Chaotropic reagents like urea or guanidinium chloride are used for denaturation of proteins. However, parts of the proteome (e.g. membrane proteins) may not be recovered with these reagents. Sodium deoxycholate (SDC) emerged as a MS-compatible buffer component, which effectively solubilized also hydrophobic proteins and even enhanced the activity of trypsin [15]. The development of workflows like 'Filter-Aided Sample Preparation' (FASP) or 'STop And Go Extraction Tips' (StageTips) streamlined 'in-solution' digestion by combining sample retention, removal of detergents, exchanging of buffers, reduction of disulfide bonds, alkylation and protein digestion in a single reaction chamber [16], [17]. Particularly, the invention of StageTips allowed one-step sample preparation workflows as outlined in Figure 1.2, which strongly reduced the risk of loss or contamination of material, allowed easier automation of the whole process and is a routinely used method in proteomics [17], [11].

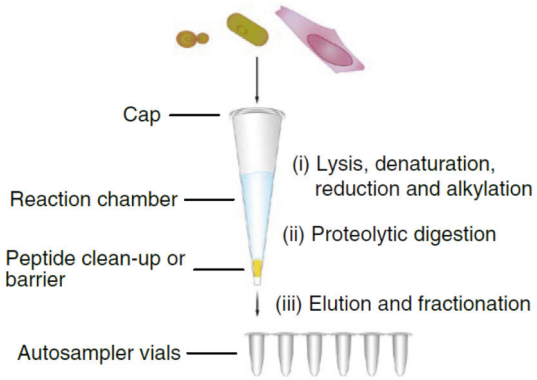


Figure 1.2: Scheme of an optimized in-StageTip sample preparation workflow. Combination of lysis, denaturation, reduction and alkylation into a single step reduces risk of sample loss and contamination. Proteolysis and elution of peptides is carried out in the same reaction chamber. Adapted from [11].

1.1.3 Liquid-chromatography coupled to Orbitrap mass spectrometers

Several generic workflows are available to researchers for the analysis of proteomic samples including different chromatographic methods, mass analyzers and mass spectrometry settings. In this PhD thesis I exclusively used Orbitrap mass spectrometers in the data-dependent acquisition mode, which is the

focus of this section. Yet, other techniques will also briefly be introduced.

Separation of peptides via reversed-phase chromatography

Before ions are injected into the mass spectrometer, peptides need to be separated. The choice of chromatography greatly impacts throughput and quality of the analysis. Typically, separation of peptides in proteomics experiments is achieved by nano-high performance liquid chromatograpy (nano-HPLC), which refers to the low flows of 250-350 nl/min compared to standard high performance liquid chromatography (HPLC). Reversed-phase chromatography in proteomics is typically based on C_{18} silica material as stationary phase, which achieves good separation of peptides. A gradient starting with an aqueous buffer (0.1% formic acid in water) with increasing proportions of a hydrophobic buffer (80% acetonitrile, 0.1% formic acid in water) gradually elutes peptides depending on their hydrophobicity.
For the projects in this thesis I used the EASY-nLC system of Thermo Fisher. Columns were packed with C_{18} silica resin of 1.9 μm particle size and with an inner column diameter of 75 μm. This setup with its low flow rates allows better ionization of peptides in ESI-based proteomics experiments and reaches higher sensitivities [18]. Larger column diameters promise better separation efficiency, however they also lead to higher flow rates and lower sensitivity. Yet, it was shown that also micro flow columns reach a proteome depth comparable to nano LC setups (for a recent example see [19]). Hence, micro flow chromatography setups may become more widely used in proteomics applications in the future.
Another important consideration is the throughput of experiments, which is defined by the LC gradient and sample loading onto the column. The improvements of mass spectrometers have also enabled higher sample throughput with ever shorter gradients. This development caused loading and washing steps to take up a larger proportion of the overall gradient length. Moreover, the high pressure required for low flow rates in the EASY-nLC system has a negative impact on the robustness of the entire system [20].
While the EASY-nLC system provides a well established setup for peptide separation in proteomics experiments, there are still advancements to be expected in the future. Alternative LC systems like Evosep reduce the time between

samples by preparing injections in parallel and improve robustness by relying on a single high-pressure pump [20]. This will further increase the throughput and reproducibility of retention times between samples, which is particularly important in clinical proteomics [20].

Orbitrap mass analyzers

The choice of mass analyzers is crucial in proteomics analysis. The major benchmarks of mass spectrometers, which are sensitivity, mass accuracy, scan rate, mass resolution and dynamic range, vary between instruments. There are five types of mass analyzers that are typically used in proteomics experiments. These include orbitrap analyzers, quadrupoles (Q), ion traps (including quadrupole ion traps: QIT and linear ion traps: LIT or LTQ), time-of-flight (TOF) and, less frequently, Fourier-transform ion cyclotron resonance (FTICR) mass analyzers [21]. These instruments are combined in so called 'hybrid' mass spectrometers, which allow selection and fragmentation of peptides in separate parts.
An example of a linear quadrupole instrument is the 'Q Exactive' instrument, which contains a quadrupole, a C-trap and an Orbitrap [22]. Compared to previous instruments, it reached double the resolution due to an 'enhanced Fourier Transformation' algorithm, had faster cycle times and allowed multiplexing at the MS and MS2 levels compared to previous instruments [22].
The 'Q Exactive HF' is a further development of the Q Exactive instrument of which the schematic structure is depicted in Figure 1.3. A low-resolution prefilter was integrated in the injection flatapole of the Q Exactive HF. Moreover, a novel quadrupole guaranteed a higher fidelity over a broad range of isolation widths. Together, this improves peptide and protein detection by 40% or 20%, respectively, compared to its predecessors. Identification of peptide phosphorylation sites increased by 60% [23]. Ions enter the S lens through the capillary, which focuses and propels them forward. The injection flatapole guides the ions into the bent flatapole and further focuses them into a compact beam. The bent flatapole directs the ions in a 90° turn into the quadrupole, removing solvent and neutral gas molecules. As in all quadrupoles, a particular ratio of radio frequency (RF) to direct current (DC) voltages applied to the quadrupole filters for ions with specific mass-to-charge (m/z) properties.

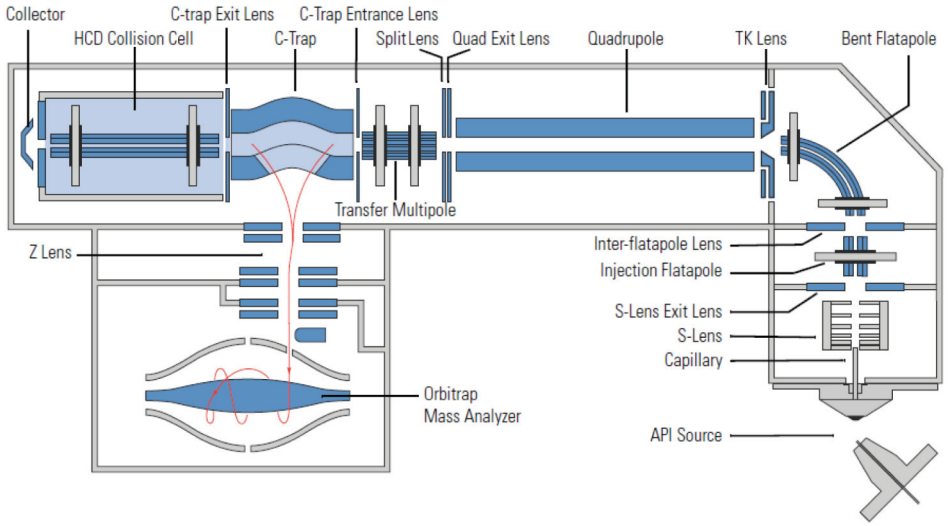


Figure 1.3: Construction of the Q Exactive benchtop mass spectrometer series. The Q Exactive HF is a hybrid instrument combining a quadrupole with an orbitrap mass analyzer. The HCD collision cell is directly connected to the C-trap. Detailed description of the individual parts is provided in the text. Adapted from [24].

Continuously adjusting the RF-to-DC settings enables m/z scan ranges. The transfer multipole guides the ions into the curved linear trap (C-trap). The C-trap stores ions in an RF-only quadrupole, as ions lose their kinetic energy by collision with nitrogen gas [24]. In full scans, the C-trap transmits ions into the Orbitrap by decreasing the radio frequency and applying a high voltage [25]. In MS/MS scans, ions are passed through the C-trap into the Higher-energy C-trap dissociation (HCD) cell. Energetic collision with nitrogen molecules creates fragment ions. These product ions are transferred back into the C-trap and injected into the Orbitrap for analysis [24].

At the center of the Orbitrap (Figure 1.4a) is a solid metal electrode. It is surrounded by two outer electrodes bearing the same shape, which are the receiver plates for image current detection [25]. When ions enter the Orbitrap, they will orbit both around (r-axis) and along (z-axis) the central electrode, which is schematically shown in figure 1.4b [26]. The oscillation along the z-axis is only dependent on the ion's mass-to-charge ratio, whereas rotation around the central electrode depends on various parameters like the entry velocity of the ions [26]. The image current caused by the movement along the z-axis is measured on the outer electrodes and amplified (Figure 1.4a). Fourier-

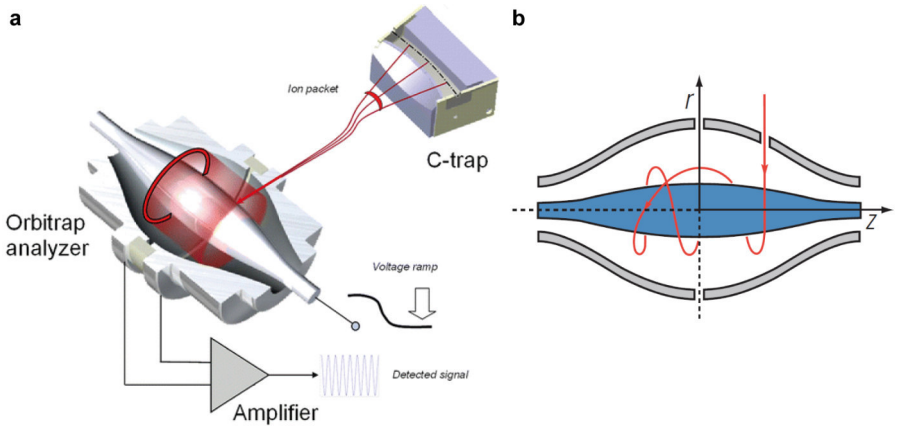


Figure 1.4: Cross-section and schematic structure of the Orbitrap mass analyzer. a, Ions are stored in the RF-only quadrupole of the C-trap. One of all m/z is ejected into the Orbitrap as a short packet. Ions oscillate around the electrode and induce a current on the outer electrodes, which is amplified and monitored. **b,** Schematic of the inner and outer Orbitrap electrodes. Ions circulate around and along the inner electrode, the designated z-axis. Adapted from [24] and [25].

transformation of the signal allows analysis of the m/z ratio of the detected ions. The first generation of Orbitrap mass analyzers already provided a high resolution up to 150,000, mass accuracy with a root-mean-square deviation of 5 ppm, higher dynamic range and a higher upper mass limit [27]. This had a strong impact on the advancement of mass spectrometers. The first Orbitrap-based mass spectrometer combined the accurate mass detection and high resolution of the Orbitrap with the fragmentation capabilities, sensitivity and exact precursor isolation of linear ion traps in a hybrid LTQ Orbitrap mass spectrometer [28]. The Orbitrap was optimized by adapting the geometry and voltages of the electrodes in the 'high-field Orbitrap' mass analyzer [29] and 'ultra high-field' Orbitrap [30], which improved the resolution of mass spectrometers. The invention of the Orbitrap mass analyzer, its combination with a quadrupole mass filter in the bench-top Q Exactive mass spectrometer platform allowed researchers to perform qualitative and quantitative analysis of proteomic samples with high mass accuracy, resolution and sensitivity.

Peptide fragmentation methods and identification

An essential part of bottom-up proteomics is the fragmentation of peptides and associated acquisition modes of the mass spectrometer. Peptides are subject

to bond breakages in MS/MS scans, which is crucial for their computational identification. In theory, peptides can fragment into so-called a, b, c- and x, y, z-ions. The fragmentation method determines which ions are actually produced in mass spectrometry MS/MS scans. The most prominent fragmentation methods in tandem mass spectrometry are collision-induced dissociation (CID), and its variant higher-energy C-trap dissociation (HCD), electron transfer dissociation (ETD) and electron capture dissociation (ECD). The different techniques cause distinctive bond breakages in peptides, which is important for database searches and accurate identification of the analyte.

CID causes peptide fragmentation by collision with an inert gas like helium or nitrogen [31]. HCD is similar to CID, but applies higher energy dissociations [32]. This changes the characteristic fragmentation patterns compared to CID MS/MS spectra. In ECD [33], which is almost exclusively used in FTICR mass spectrometers [34], and in the closely related ETD [35], fragmentation is caused by transferring electrons to the multiply protonated peptides. This results in the formation of an unstable radical ion, which will subsequently be cleaved at the N-C_α bond [35].

ETD, CID and HCD generate distinctive fragmentation patterns in the MS/MS scans. ETD predominantly results in the formation of c- and z-ions and to a much lesser extent also y-ions. CID and HCD almost exclusively create b- and y-ions [36]. They also produce a-ions, but much less frequently than b- or y-ions and predominantly the a2 ion after loss of carbon monoxide from the b2-ion [37]. Finally, they also create peaks in the low m/z range including immonium ions and internal fragments (i.e. combination of b- and y-type cleavage) [36]. The difference between CID and HCD mainly exists in the lower proportion of b-ions in HCD fragmentation.

CID has been shown to identify more proteins in a direct comparison with ETD, although the fragment coverage was higher in ETD fragmentation [38]. However, in studies of some protein modifications, ETD may outperform CID and HCD. This is due to the lability of many post-translational modifications (PTM) like phosphorylation, O-glycosylation and sulfonation. In CID-type fragmentation these modifications can be lost as they may be preferentially fragmented compared to backbone cleavage. In contrast, ETD would fragment modified peptides predominantly along the peptide backbone retaining

the PTM localization [39]. Hence, the choice of fragmentation method to some extent depends on the experimental aim. Nevertheless, HCD is the most widely used technique as it provides convincing results in almost any application and is compatible with most mass spectrometers.

The fragmentation of a peptide into sequence-specific ions allows its identification and ultimately the assignment to a protein (Figure 1.5). Several search en-

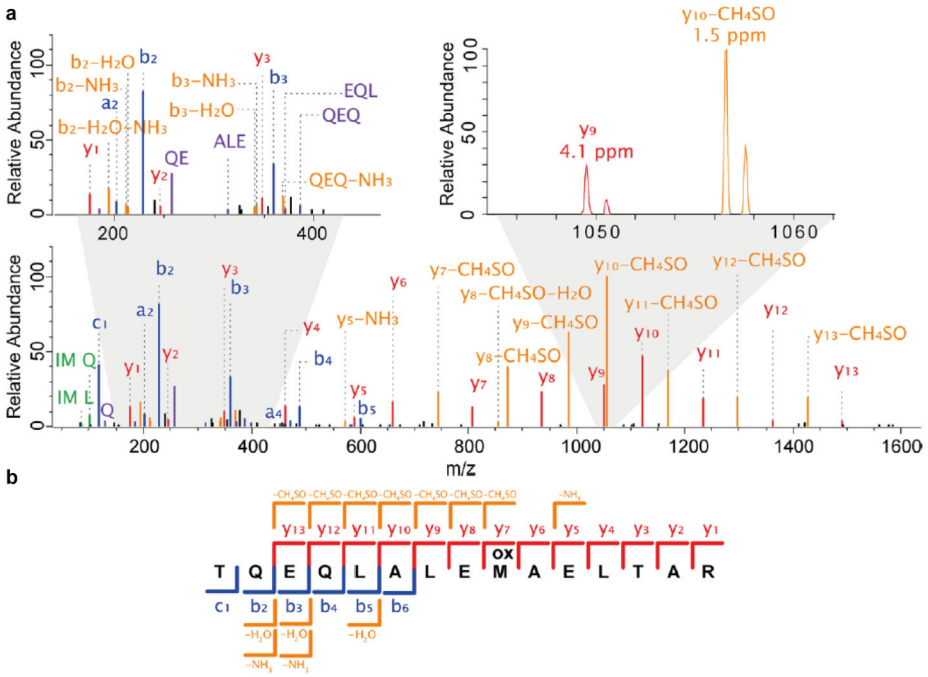


Figure 1.5: Example of an HCD-type MS/MS product ion spectrum used for peptide identification. a, Full tandem mass spectrum (bottom) shows predominant generation of y- and b-type ions in HCD-type fragmentation. Neutral loss of CH_4SO from oxidized methionine is frequently observed. A closer view of the 200-400 m/z mass range (top left) reveals internal fragments, which may also be used for identification of the peptide. Mass fragments are detected with a high accuracy allowing unambiguous assignment of the ion (top right). **b,** Conventional MS software matches the experimentally assigned fragment ions to theoretical fragmentation spectra from a database. Typical search engines incorporate the overlap between experimental and theoretical spectra in a scoring system to identify the correct peptide within a certain false discovery rate. Adapted from [40].

gines like SEQUEST [41], Mascot [42] or Andromeda [43] have been developed to accurately assign peptide sequences to the experimental data. Andromeda, which is the search engine implemented in MaxQuant, was used for most of the analyses reported in this thesis. It calculates the score based on the number of

matching peaks to a theoretical spectrum. An experimental peak is considered as a match to the theoretical database only if the difference does not exceed a defined mass window specified in ppm for Orbitrap instruments. Moreover, only the top q peaks within a range of 100 m/z are considered. In order to control for false positive assignments and calculate a false discovery rate, the reverse sequences of all proteins in the database are introduced as decoys [43]. Identified peptide sequences are subsequently combined into protein IDs.

Data acquisition modes

For the results shown in section 3.1 to 3.5, I relied on data-dependent acquisition (DDA). However, data-independent acquisition (DIA) is becoming more and more popular in MS-based proteomics.
An outline of DDA and DIA acquisition and their differences are shown in Figure 1.6. The main one is that DIA aims at capturing all peptides by applying broader isolation windows across the entire m/z range of the full scans. In contrast, DDA selects the top most abundant precursor ions from the full scan for fragmentation in the MS2 scan. For example, a 'top10' method would select the 10 most abundant ions of the MS1 scan for MS2 fragmentation. Most mass spectrometers can perform at least one 'top10' scan cycle in 2 seconds [45]. Other settings in DDA scans also strongly affect the identification rates. For instance apart from MS/MS events per cycle, the maximum injection time (IT) has a significant impact on peptide identifications [46]. Although some parameters should be optimized for DDA acquisition, it remains the easiest mode in MS-based proteomics regarding both settings and data analysis. On the downside DDA performs worse in terms of data reproducibility and consistency than DIA due to stochastic sampling of the most abundant peptide species [47].
The development of novel DIA methods is rapidly becoming more popular. It is especially useful when only small sample amounts are available. Moreover, DIA allows much shorter gradients compared to DDA, which enables an overall higher throughput of samples [48]. The comparably more difficult data analysis is a negative aspect, yet this will likely change in the near future as DIA becomes more and more popular in the field [49]. One major development was the development of the BoxCar DIA method. In conventional DDA

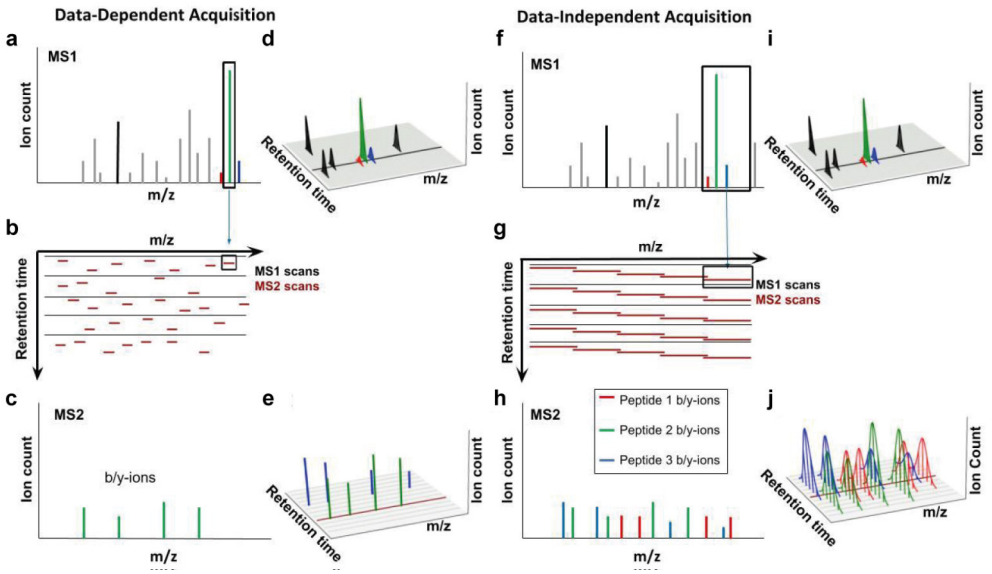


Figure 1.6: Comparison of data-dependent and data-independent acquisition. **a,** In DDA the most abundant ions are selected with a defined isolation window for MS/MS fragmentation. **b,** Each MS1 spectrum (black lines) are followed by several MS2 spectra (red lines). **c,** Fragmentation of an isolated ion after the full MS1 scan will result in the characteristic MS2 spectrum. **d,** Ion intensities of peptides eluting over different retention times are monitored in each full scan (black line) and used for quantification. **e,** MS2 scans of the blue and green peptides from figure a are shown. MS2 scans are not acquired for lower abundant peptides observed in the MS1. The MS2 scan of the green peptide is indicated by a red line. **f,** The major goal of data-independent acquisition is to fragment all eluting peptides. Much wider isolation windows are chosen and all ions in this window are fragmented simultaneously. **g,** MS2 scans (red lines) in DIA cover the entire m/z range of each full scan (black lines) in contrast to DDA. **h,** Co-fragmentation of multiple peptides contained in the broad isolation window results in more complex MS2 spectra with signals from the red, green and blue precursor ions. **i,** Ion intensities can be used for quantification similar to DDA scans. **j,** MS2 ion intensities are monitored for each precursor ion throughout its entire elution profile. Thus, MS2 intensities can be used for quantification in contrast to DDA modes. All precursor ions from the full scan are fragmented in MS2 scans and also low abundant peptides are identified. Adapted from [44].

studies only a tiny fraction of ions (less than 1%) is used for MS1 scans. BoxCar increases this fraction by an order of magnitude by segmenting the mass range of full scans into multiple windows and identified more than 90% of the proteome of a human cancer cell line [50]. Another well-known DIA workflow is 'Sequential Window Acquisition of all Theoretical Mass Spectra' [51]. In a typical SWATH-MS workflow, precursor ions in the m/z range of 400-1200

are selected in 32 steps with an isolation window of 25 Da. DIA has emerged to become particularly powerful in studies where accurate and reproducible quantification of large parts of a proteome is desired. In general, the advantages attributed to DIA workflows are its data completeness especially in samples with a high dynamic range, its data reproducibility and its ability to retrospectively query the data as MS2 scans are recorded for all precursors [49].

1.1.4 Protein quantification methods

The quantification of proteomics data in mass spectrometry is one of the most important things to consider in MS-based proteomics. The ultimate aim is to detect and quantify changes between biological or clinical samples. The major distinction is made between label-free and label-based quantification experiments. Label-based methods require isotope-labeled peptides, such that a quantitative comparison of peptide intensities in a mixture of samples is feasible. Several distinctive workflows have emerged which differ in their techniques of how to label the analytes, namely metabolic labeling, chemical labeling or spiking in heavy peptides. Stable-isotope labeling by amino acids in cell culture (SILAC) is a metabolic labeling strategy where cell culture media either contain 'heavy' (H^2, C^{13} or N^{15} isotope-labeled) or 'light' (normal) amino acids [52]. Quantitation is performed at the MS1 or MS2 level [53]. SILAC maps quantitative ratios extremely accurately [53], which makes it particularly useful when small biological differences are expected. Moreover, SILAC combines samples after cell culture, which minimizes quantitative errors due to inconsistent sample processing (compare figure 1.7). Chemical labeling is achieved by incorporation of an isotope mass-tag to an amino acid side chain of proteins or peptides, which was described in several methods like Isotope-Coded Affinity Tag (ICAT) [55], Tandem Mass Tags (TMT) [56], Easily Abstractable Sulfoxide-based Isobaric-tag (EASI-tag) [57] or isobaric Tags for Relative and Absolute Quantification (iTRAQ) [58]. Isobaric labeling methods like TMT, EASI-tag or iTRAQ utilize chemical reagents consisting of a reporter region, a balance region and an amine-specific reactive group. The reactive group allows covalent linking of the tag typically to the N-terminus of the polypeptide or lysine. Different samples are tagged with

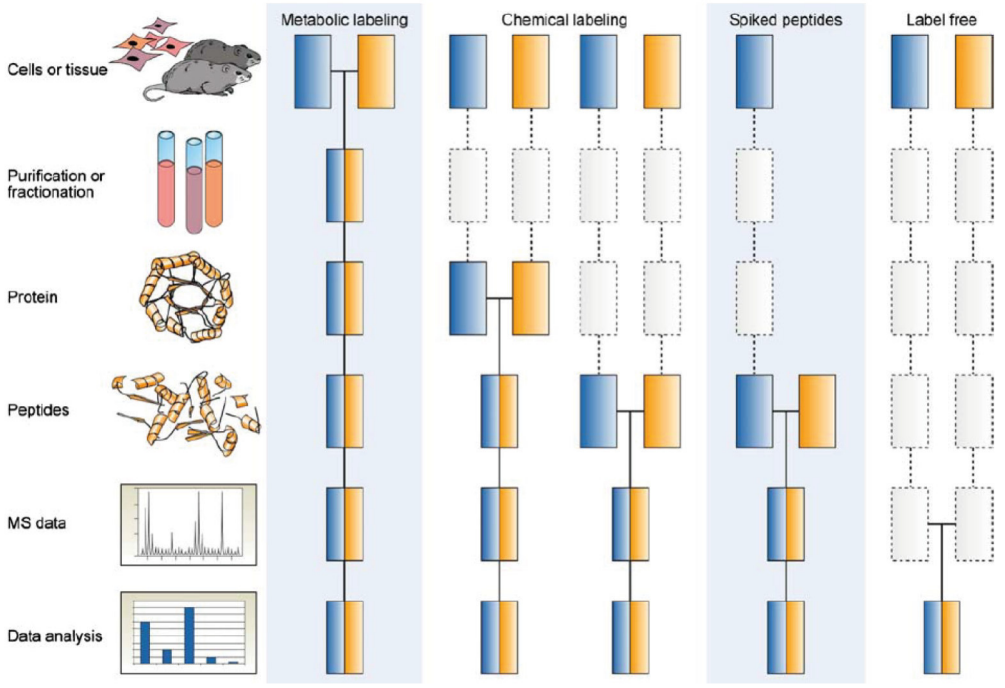


Figure 1.7: Overview of label-based and label-free approaches to data quantification in MS-based proteomics. Label-based methods comprise metabolic, chemical labeling or the use of spiked-in peptides. The earlier samples are combined, the lower is the variation between analyses caused by inconsistent sample handling. Label-free workflows separate different samples completely until comparison during data analysis. Adapted from [54].

labeling reagents distinctive in the molecular weight of their reporter regions. However, the balancing region equalizes the overall mass of the tag such that the same polypeptides from different samples cannot be compared in the full scan. MS/MS fragmentation leads to the cleavage of reporter regions, which allows quantitative comparison of product ion intensities of the peptides from different samples. Samples in chemical labeling are combined either at the protein level or at the peptide level (Figure 1.7).

Spiking in isotope-labeled peptides profoundly differs from other quantitative methods. It is used when absolute quantification of a subset of proteins is desired, which also coined its alternative term 'Absolute Quantification of Proteins' (AQUA) [59]. The basic principle is to add known quantities of a labeled peptide to the samples which enables the absolute quantification of that specific peptide and its respective protein. Usually, the labeled peptides are added to the sample peptides (Figure 1.7). Yet, other methods also used

labeled synthetic proteins, which can be added earlier in the process [54].
A completely different approach is label-free quantification (LFQ). Although LFQ provides less accurate quantification results than label-based methods, it has become attractive due to the experimental simplicity, cost-effectiveness and unlimited number of samples that can be compared [60]. One of the earliest approaches was spectral counting, which is based on the correlation of protein quantity with sequence coverage and number of identified peptides. Various computational approaches have emerged using spectral counting to quantify proteomics data [61]. However, the quantitative accuracy of spectral counting is lower compared to other approaches.
Alternatively, and much more accurately, label-free quantification is performed by peak-intensity based approaches, which quantifies proteins by summing up the extracted ion current peak areas of their peptide precursor ions. The increasing mass accuracy of modern mass spectrometers has consistently improved the matching and quantification of peaks across all samples. [60].
In all LFQ workflows samples are processed separately and are only combined during data analysis (Figure 1.7). Thus, LC-MS data of individual samples need to be aligned and data normalized in order to be able to compare biological samples [62]. Therefore, sample processing needs to be particularly reproducible and technical variation caused during sample preparation has to be minimized. However, label-free quantification has greatly advanced over the past years and, to a certain extent, differences caused during sample processing are adressed for during data normalization. Even copy number estimation has become feasible by taking into account the label-free quantification intensities, sequence length, molecular weight and the total protein amount of a cell [63]. The developments of MS instruments and data analysis software has widely expanded the use of LFQ workflows.
In summary, the choice of data quantification method depends on the experimental aims. In general, if small quantitative ratios between samples are expected and if quantification needs to be particularly accurate, one should opt for metabolic or chemical labeling. Spike-in of isotope-labeled peptides is useful, if absolute copy numbers of a few proteins have to be determined. Label-free quantification workflows are well suited in large-scale experiments and if experimental costs are an issue.

1.2 Mechanisms of gene regulation - an overview

One key to life is the storage of genetic information and the accurate translation into its functional units. Not without any reason it is referred to as the central dogma of molecular biology and was described by Francis Crick in as early as 1958 and refined in 1970 [64]. It essentially described that biological information is transferred from DNA to RNA to proteins.
Ever since the first publication of this concept, the underlying mechanisms have been studied in great detail, which today allows us to precisely understand how genetic information is transcribed and regulated. While the main idea of the central dogma has not changed over the past decades, technological advances enabled researchers to add much more information to it like transcription rates and half-lives (see Figure 1.8). This revealed that the process of gene expression is far more complex than initially described by the central dogma (reviewed in [65]). Given the importance and complexity of the gene

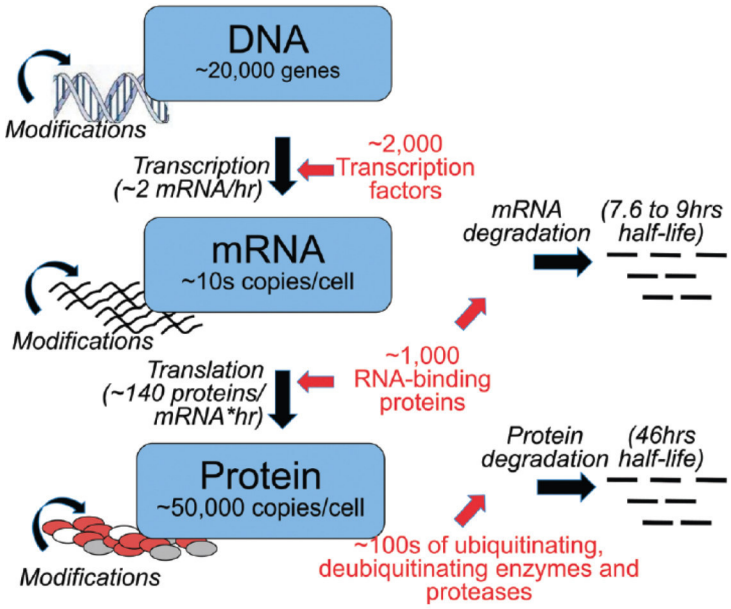


Figure 1.8: Overview of the central dogma of molecular biology shown for mammalian cells. The biological flow of information is determined by the transcription of DNA into RNA and the translation of RNA into proteins. Several regulatory processes like DNA or RNA modifications fine-tune gene expression. Technological advances have allowed the estimation of transcription and translation rates. Likewise, the half-lives of transcripts and proteins have been determined. Adapted from [66].

expression system, these processes have to be tightly controlled in order to prevent aberrant expression of genes. For this reason various mechanisms have evolved to adapt the expression and turnover of proteins to intracellular and environmental signals. DNA methylation [67], histone (de)acetylation [68] or alternative splicing [69] are just a few prominent examples of gene regulatory mechanisms. While many of these pathways have been described, it is very likely that key regulatory interactions or even novel signaling cascades remain to be discovered and novel technologies may be crucial in that process. Considering the pathological effects of aberrant gene expression it is crucial to fill the gaps of knowledge and to develop more sophisticated methods helping in that endeavour.

1.2.1 Regulatory mechanisms in gene transcription

The activation of the vast majority of genes begins with a transcription factor binding to promoters or enhancers. TFs recruit other proteins which control chromatin accessibility and initiate transcription. This leads to the opening of the DNA duplex by RNA polymerases and synthesis of RNA (Figure 1.9a-c). Most TFs bind to naked DNA, but some pioneering factors have the ability to target nucleosomal DNA and recruit chromatin remodeling proteins to provide access to these genomic regions [70]. Regulation of gene expression at the transcriptional level can essentially be exerted at two intertwined stages involving a) transcription factors and the transcriptional machinery and b) the chromatin architecture modulated by specific proteins [71].

Gene regulation involving the transcription machinery

Transcriptional control by transcription factors (TF) is one of the most important mechanisms of how organisms ensure dynamic expression of genes.
Regulation involving TFs and the entire transcription apparatus occurs both before and after Pol II recruitment [71,72]. TFs binding to their specific target DNA sequences in promoter or enhancer regions can be either gene activating or repressing. One key regulatory function is cooperative DNA binding by TFs. Large enhancer complementation assays showed that a TF's function heavily depends on the enhancer context, that TFs can be substituted by another TF

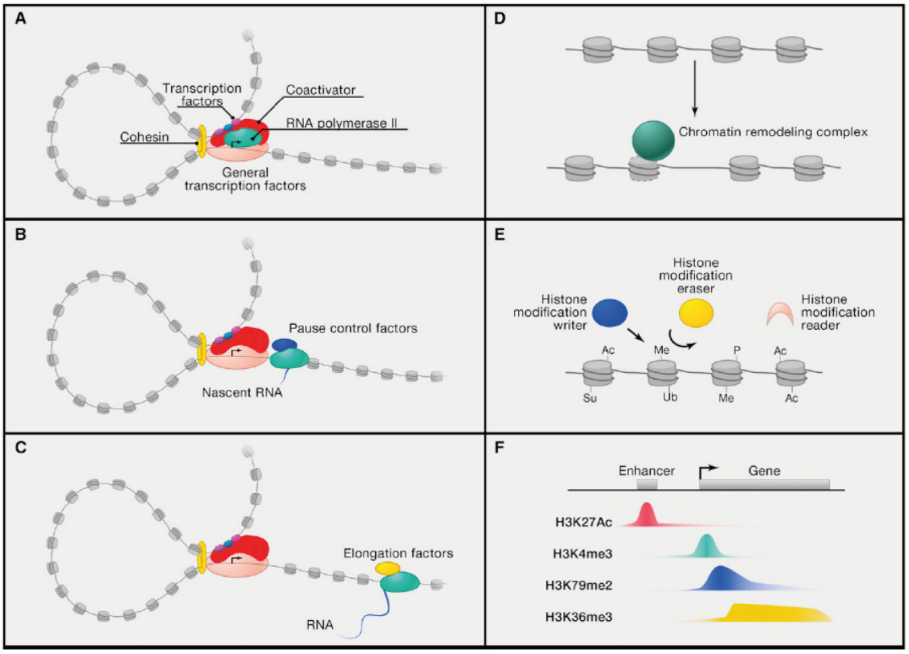


Figure 1.9: Regulatory mechanisms in gene expression. **A,** Transcription factors bind to sequence-specific DNA regions and recruit co-activators. The RNA Pol II associates with these factors and general transcription factors (GTF). The Mediator and cohesin complexes stabilize the DNA loop formed between enhancer and transcription start site (TSS). **B,** The RNA polymerase II initiates transcription, but can be halted by specific pause control factors. **C,** Co-factors recruit elongation factors, which induce phosphorylation of polymerase and pause release factors to continue elongation. **D,** ATP-dependent chromatin remodelers shift nucleosomes to facilitate access of specific regions for other transcriptional proteins. **E,** Histones can be methylated (Me), acetylated (Ac), phosphorylated (P), sumoylated (Su) or ubiquitinylated (Ub). These modifications are introduced by writers, removed by erasers and detected and acted upon by readers. **F,** Example of the characteristic modification patterns of histone acetylations and methylations within a gene locus, which affect transcription rates. Adapted from [71].

and that specific functions of TFs can be explained by cooperativity with different co-factors [73]. Direct physical interactions can lead to cooperativity. However, it was also reported that the majority of cooperative TF-TF interactions are mediated by DNA [74] indicating that protein-DNA interactions play a crucial role in the recruitment of regulatory TF-TF pairs. Many prominent examples of cooperative TF-TF regulation are known from tissue development and depend on the cell type. For instance, Smad2/3 associates with Oct4 in ES cells, Myod1 in Myotubes and PU.1 in Pro-B-cells. Interactions of Smad2/3

with the other cell-type specific TFs is regulated by their abundance [75].
TFs also recruit other protein complexes, which are capable of directly binding to the RNA Polymerase machinery or which block the engagement of the transcription apparatus. TFs can additionally associate with remodeller proteins, which alter chromatin structure to regulate gene expression [72].
While it was long believed that transcriptional control occurs exclusively before initiation of the Polymerase complex [76], later studies described that regulation also happens by promoter-proximal pausing and efficiency of elongation (Figure 1.9b-c) [77]. Pause control factors like negative elongation factor (NELF) and DRB-sensitivity-inducing factor (DSIF) can halt recruited polymerases early in the elongation process. Paused polymerases can be induced to restart.

Gene regulation by alteration of the chromatin structure

Important gene-regulatory features of chromatin structure comprise nucleosome organization and histone or DNA modifications. Active promoters show a characteristic pattern of lower mean nucleosome occupancy, referred to as nucleosome-depleted regions (NDR) [78, 79]. Transcription factors may recruit nucleosome remodellers to achieve a more accessible chromatin architecture (Figure 1.9d) [80, 81]. However, it was recently shown that these remodelling complexes themselves can cause dissociation of TFs by sliding nucleosomes past them. In this way nucleosome remodellers regulate TFs and influence gene activity [82].
Histone and DNA modifications are other hallmarks of epigenetic gene regulation. Most prominent examples are histone acetylation and methylation of histones or DNA. These modifications are executed by so-called 'writers', detected by 'reader' domains of effector proteins and removed by 'erasers' (Figure 1.9e).
DNA methylation describes the conversion of cytosine into 5-methylcytosine (5-mC). In the context of gene regulation, 5-mC is usually found in dinucleotides of cytosine and guanosine (CpG) [83]. As methylcytosine shows an increased mutagenic potential, genomes of vertebrates are low in CpG content [83]. However, there exist CpG islands (CGI) with a relatively higher amount of CpG dinucleotides [83]. More than two thirds of gene promoters

contain CGIs [83]. Transcriptional regulation by DNA methylation is executed mainly by two different mechanisms (reviewed in [84]). First, TFs with GC-rich recognition motifs are not capable of recognizing their target DNA sequence if cytosines are methylated. Second, methyl-CpG binding proteins can associate with methyl-CpGs. These proteins were shown to recruit other co-repressor complexes and thus promote repression of gene transcription [84].
Histone acetylation is also tightly controlled across the entire genome [85, 86] by multisubunit histone acetylase and deacetylase complexes, many of which are highly conserved across various eukaryotes [87, 88]. It affects gene expression in two ways. Histone acetylation causes a change in the local chromatin structure, which makes it more accessible for other transcriptional activating proteins [89]. Secondly, bromodomain (BRD)-containing proteins detect and bind acetylated lysines. They can act as a scaffold for other co-regulatory complexes, induce transcription themselves or exert a catalytic reaction like transferring methyl groups [90].
Histones are also methylated on lysines and arginines. Effects vary depending on the position of the modification and the amount of methylations (i.e. mono-, di- or tri-methylations). A wide variety of histone methylases and demethylases have emerged, which are often part of large complexes. As for acetylations, transcriptional co-regulators containing chromo- or bromodomains recognize methylations and subsequently impact gene expression [91].

1.2.2 Protein-RNA interactions in post-transcriptional regulation of gene expression

Post-transcriptional control of transcribed RNAs is a central feature in the regulation of gene expression. RNA-binding proteins execute the vast majority of control mechanisms. From the transcription of DNA to RNA and the translation of RNA into proteins, RNA is altered and relocalized in many ways (Figure 1.10). Alternative splicing or polyadenylation affect expression levels of different proteoforms. Translation of the transcript strongly depends on its stability, which plays a key role in post-transcriptional regulation [93]. Spatial and temporal control of gene expression is ensured by accurate localization of mRNAs [94]. Finally, ribosomes translate the mRNA into polypeptides.

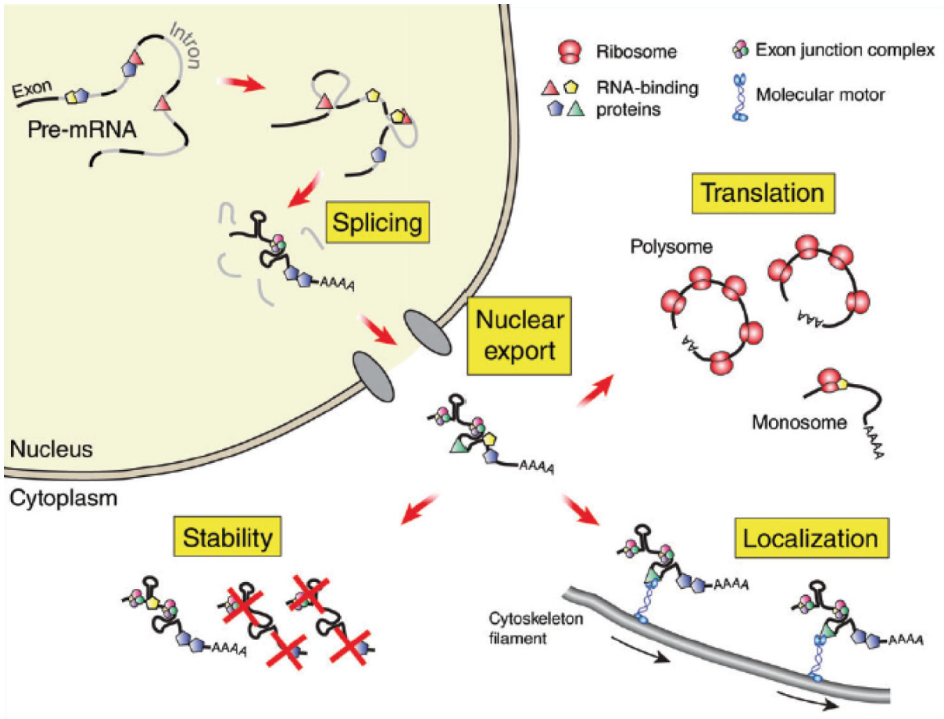


Figure 1.10: Different stages of mRNAs from transcription to translation. Interactions with proteins determine the entire lifecycle of RNA including regulatory steps and the translation of the RNA into a protein. Adapted from [92].

Protein-RNA interactions are crucial to ensure proper execution of these regulatory pathways.

Alternative splicing

Alternative splicing allows the generation of more than one mRNA from a single gene. Three properties impact the product of alternative splicing, which are the splice site strength, the existence of cis-regulatory elements in the pre-mRNA and the expression of specific RBPs [95]. Splicing is executed by the spliceosome complex. In mammals, recognition of the 5' splice site by snRNP U1 is followed by the binding of the splicing factor 1. Recruitment of the U2 auxiliary factor completes formation of the E complex. Replacement of splicing factor 1 by U2 snRNP and binding of U4/U6-U5 tri-snRNP forms the ATP-dependent B complex. Conformational changes in the B complex (e.g. loss of U1 and U4 snRNPs) leads to the catalytically active spliceosome (C complex) [96]. The decision to splice exons depends on various signals by

exonic and intronic splicing enhancers (ESE or ISE) or silencers (ESS or ISS). SR proteins recognize these elements and recruit the splicing machinery and other regulatory protein complexes. They may also switch between repressing and activating splicing [96, 97]. Splicing inhibitor proteins bind to specific elements of the RNA (e.g. polypyrimidine tract or ISS) and inhibit recruitment of crucial splicing factors or prevent E complex formation [96, 98]. Alternative splicing involves the interplay of various multi-protein complexes to fine-tune expression of specific proteoforms from the same transcript. Thus, it is a crucial process in lineage commitment and tissue development [99–101].

Alternative polyadenylation

Similar to alternative splicing, cells use alternative polyadenylation as a means to obtain mutliple mRNAs from the same gene. Virtually all 3' ends of eukaryotic pre-mRNAs undergo endonucleolytic cleavage and subsequent polyadenylation. Four protein complexes execute this process, which are CPSF (Cleavage and Polyadenylation Specificity Factor), CstF (Cleavage stimulation Factor), CFI and CFII (Cleavage Factor I and II). Other auxiliary proteins like poly(A) polymerase (PAP) and RNA polymerase II help in establishing the poly(A) tail [102]. Polyadenylation impacts the expression of proteins both quantitatively and qualitatively. If alternative poly(A)-sites locate between internal introns and exons, this will lead to different protein isoforms (coding region-alternative polyadenylation). If poly(A)-tails are placed differently in the 3'- untranslated region (3'-UTR), the outcome will be distinctive expression levels of the same protein (UTR-alternative polyadenylation) [102]. Shorter 3'-UTRs have been shown to increase protein abundance compared to longer 3'-UTRs [103]. The regulation of coding region- and UTR-alternative polyadenylation largely depends on protein-RNA interaction. For instance, the overexpression of the polyadenylation factor CstF64 increases proximal poly(A)-tails on the IgM mRNA, which leads to a switch from membrane-bound to secreted IgM [104]. In mouse male germ cell maturation, CstF64t is more abundant in all germ cells, whereas levels of CstF64 decreased. This was attributed to different 3'-UTR alternative polyadenylation regulation during spermatogenesis [102, 105]. Remarkably, about 54% of human genes harbor multiple poly(A) sites indicating a widespread role of polyadenylation in post-

transcriptional regulation [106].

Regulation of mRNA stability

Transcription and degradation of mRNA is closely intertwined as it defines the expression levels of proteins. Various proteins control the decay of transcripts. The degradation of most eukaryotic mRNAs begins with deadenylation of the poly(A)-tail by deadenylases like Ccr4, Pop2 or the Pan2-Pan3 complex. Next, the Dcp1-Dcp2 enzyme removes the 5' cap. Subsequently, the 5'->3' exonuclease Xrn1 can digest the mRNA. Transcripts can also be degraded from 3'->5' by the exosome, a complex of several exonucleases, which also digests pre-mRNAs failing to successfully complete mRNA processing [107, 108]. Eukaryotic cells identify premature translational stop codons (nonsense-mediated decay, NMD) or missing translation termination codons (nonstop decay, NSD) and degrade these nonsense transcripts [107, 109]. AU-rich elements (ARE) are a more specialized means of mRNA stablity control. About 5-8% of the human transcriptome contain ARE sites. They are enriched in the 3' UTR of mRNAs and were shown to have an impact on transcript half-lives. Various proteins have been identified that bind to these elements. The majority of these proteins induces transcript destabilization by associating with deadenylases and exonucleases [110].
Stability of mRNA is additionally controlled by other molecular mechanisms. One example is the regulation by microRNAs (miRNA), which affect expression levels of about 30% of mammalian genes [111]. The proteins Drosha and Dicer process pre-miRNAs to obtain the approximately 21-bp long miRNA. Together with proteins, most importantly from the Argonaute family AGO1-4 in mammals, they form micro-ribonucleoproteins (miRNP) [111]. If the miRNA has near-perfect complementarity to the target mRNA, Ago2 endonucleolytically cleaves the mRNA, which becomes fully digested by the normal degradation machinery outlined above. Alternatively, the miRNP promotes the common pathway of deadenylation, decapping and degradation of the mRNA body without initial endonucleolytic cleavage [112]. Moreover, miRNAs affect the translation of mRNAs, although it is not yet clear whether that happens during or after initiation of translation [111].
A plethora of regulatory mechanisms like alternative splicing, polyadenylation

or mRNA decay fine-tune the expression levels of coding genes at the post-transcriptional level. Considering the underlying molecular mechanisms, it has become apparent that protein-RNA interactions play a fundamental role in them. Hence, characterizing and mapping these interactions is a central part in understanding how cells adapt and regulate protein abundance post-transcriptionally.

1.3 Mass spectrometry-based investigation of gene-regulatory protein interactions

Cells are incredibly complex structures which have to fulfill a very large number of tasks typically carried out by proteins. To this end, they need to interact with a variety of other compounds, most prominently other proteins, but also nucleic acids, lipids or small molecules. This also holds true for the regulation of gene expression, which requires interactions of proteins with other proteins, DNA and RNA. Given their importance in cell development and maintenance, but also in pathology, researchers have continuously improved methods to study these interaction events.
While the toolbox for protein interactome studies comprises various methods from Yeast Two-Hybrid to affinity-enrichment western blotting, mass spectrometry has emerged as the most versatile readout of protein interactomics. Soon after mass spectrometry was introduced to the world of protein science, global maps of protein complexes in yeast were created by mass spectrometry-based proteomics [113,114]. Ever since, methods have been developed to study also interactions of proteins with other biomolecules like DNA and RNA. In this section, I would like to introduce important MS-based methods to study protein interactions, particularly in the context of transcriptional regulation.

1.3.1 General considerations in protein interactomics

Owing to the importance of protein interactions in cell biology there is a great interest in shaping system-wide protein interactome maps. After the first groundbreaking publications in mass spectrometry-based protein interactomics were published in 2002 [113,114], other large-scale interaction maps quickly fol-

lowed [115–117]. Except for Ho et al., all of these studies used TAP-tagged bait proteins and a two-step purification procedure to separate non-specific interactors from true-positive ones. Affinity purification-mass spectrometry (AP-MS) was a popular strategy in interaction proteomics when quantitative mass spectrometry was still in its infancy. It required stringent purification as every identified protein was considered a true-positive prey in non-quantitative MS. In this context, AP-MS on TAP-tagged baits entails a dual purification method of the bait-prey complexes. The original TAP-tag is attached C-terminally to the protein of interest and consists of Protein A, a TEV protease cleavage and the calmodulin binding peptide (CBP) [118]. These early non-quantitative studies resulted in binary matrices of interacting proteins (see example in Figure 1.11). They needed rather sophisticated approaches to extract protein

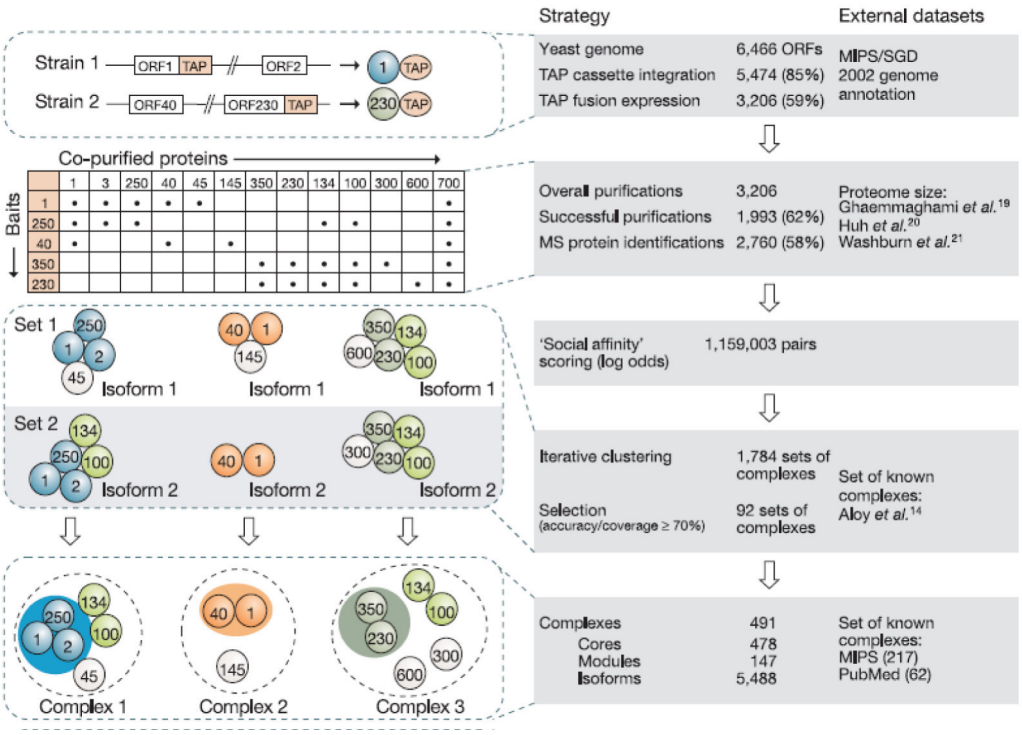


Figure 1.11: Outline of a common strategy to identify protein complexes from nonquantitative AP-MS experiments. System-wide AP-MS of TAP-tagged bait proteins produce a binary dataset of protein-protein interactions. Socio-affinity index calculation helps removing spurious interactions involving promiscuously identified proteins. Repeated cluster analysis allows identification of protein complexes. Adapted from [116].

complexes from this type of data [116]. 'Socio-affinity indices' were calculated,

which assess the odds of observing two proteins together in relation to their frequency of identification in the dataset. It helps to distinguish between true and false interactions of promiscuously identified proteins. Iterative cluster analysis of the resulting matrix led to the characterization of protein complexes, many of which had not been discovered before [116]. While the information obtained by early high-throughput AP-MS studies provided highly important insights into the organization of protein complexes in eukaryotic organisms, two main problems arise in non-quantitative AP-MS.

First, the stringent purification may prevent the identification of weak or transient interactors. Second, despite the dual-step purification, spurious interactors could never be completely removed. Considerable efforts were made to create lists of non-specific binders and remove these proteins from the dataset. Nonspecific interactors were usually defined as such if they were identified in the control or in a high number of bait pulldowns [113, 114]. This prevents a fully unbiased analysis of protein interactomes. The advancements in quantitative MS allowed novel approaches towards identifying interactors. Various studies applied quantitative mass spectrometry in the context of protein interactomics (reviewed in [119]). Nevertheless, these developments by no means substituted non-quantitative mass spectrometry. Likely, the higher costs of reagents and lower throughput made it less appealing to use quantitative labeling strategies.

This paradigm shifted with the development of label-free quantification and higher resolution mass spectrometers. The strength of label-based quantitative MS compared to label-free MS is to accurately identify small, but biologically meaningful expression changes (see section 1.1.4). In contrast, AP-MS workflows massively enrich true-positive interactors by large ratios, which makes label-free quantitative MS perfectly suited for these studies. Not surprisingly, it was quickly adopted in the study of protein-protein interactions (e.g. [120–122]). One key characteristic in modern label-free quantitative AP-MS experiments is the use of a large non-specific background proteome for data analysis [122]. This provides a consistent background for data normalization in quantification algorithms like MaxLFQ. It is also a quality control feature and allows global correlation analysis, which can add valuable information to the confidence in true-positive interactors [122]. The background

proteome forms the 'base' of the volcano plots in the statistical data analysis (compare figure 1.12). This characteristic also heavily affected workflows. Single affinity-enrichment protocols now became much more attractive, which opened up completely new avenues in streamlining global studies and enabling novel analyses. It rendered the development of protocols feasible, which identify also weak interactors at a high throughput of up to 96 interactomes in 24 hours [123]. Moreover, high-throughput studies acquiring quantitative information now allowed the introduction of interaction stoichiometries and the distinction between weak and strong interactions (see figure 1.12) on a system-wide scale [124].

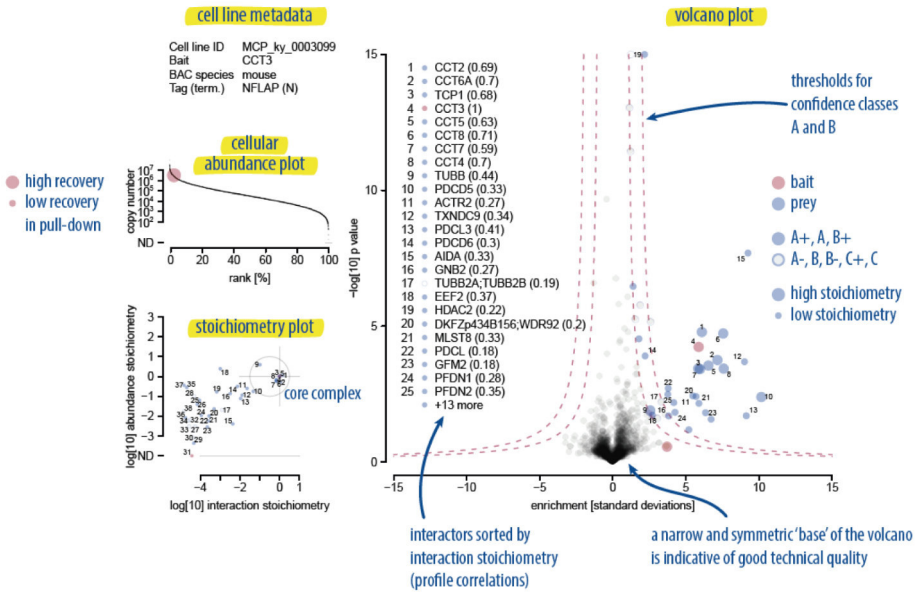


Figure 1.12: Data analysis in label-free quantitative AP-MS. Background binders locate at the 'base' of the volcano plot. True-positive interactors are identified based on their enrichment over the control group and can be subdivided into different classes of confidence based on a FDR rate and correlation of the interacting proteins across several bait pulldowns. Additional information like interaction stoichiometries can also be acquired in quantitative MS-based protein interactomics. Adapted from [124].

Furthermore, the quantitative information enable downstream analyses far beyond what is possible with binary interaction matrices in non-quantitative AP-MS. Importantly, members of protein complexes are similarly enriched with all of their interacting baits. Thus, LFQ intensities of proteins can be correlated for each protein leading to a correlation map where protein complexes are

strongly correlated (see figure 1.13). This allows a much more unbiased way

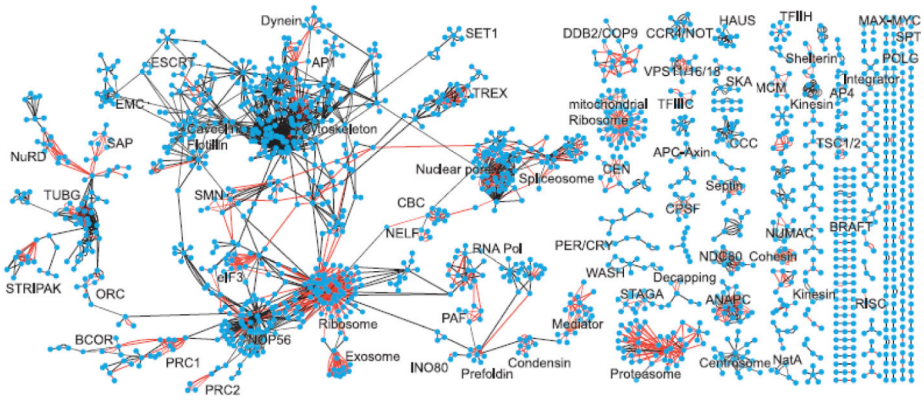


Figure 1.13: Correlation map obtained from global label-free quantitative AP-MS data. Nodes represent individual proteins and edges depict the correlation between them. Here, only correlations matching the core stoichiometry signature defined in the publication are shown. Protein complexes cluster together as all of their members are strongly correlating with each other. Adapted from [124].

of detecting (novel) protein complexes. Strong and consistent interactions can be directly read out from the correlation map with no need of filtering out any proteins beforehand.

Ultimately, these developments have made label-free single-step affinity enrichment mass spectrometry the easiest and most streamlined method of choice to map even weak interactions with a powerful distinction between true- and false-positive interactors. In general, many of the methods in label-free quantitative interactomics can be transferred to study gene-regulatory protein complexes. However, specific aspects both in sample preparation and statistical analysis need to be carefully considered to optimize the characterization of trans-regulatory protein interactions.

1.3.2 A case of its own: Chromatin-associated protein-protein interactions

The developments in MS-based protein interactomics over the past two decades has contributed to a large increase in knowledge about protein-protein interactions in various organisms. How trans-regulatory proteins associate on chromatin is a principal question to ask in order to better understand gene regulation. Ample evidence exists that chromatin-related cooperativity of proteins

is highly dynamic and depends for instance on mediation by DNA [74], the epigenetic context [75] and the developmental status of the cell [84].
Hence, chromatin-associated interactions of trans-regulatory proteins may change depending on various factors and depend on the chromatin environment. It is therefore important to discuss whether the advances in protein interaction studies may be directly transferred to the investigation of chromatin-associated interactions. While this certainly holds true with regards to general aspects of proteomic sample preparation (e.g. enrichment and StageTip purification) and data acquisition, the study of transcriptional regulatory interactomes is still more complicated. The main issue is the lower solubility and tight integrity of chromatin. Therefore, the chromatin needs to be disrupted without affecting the stability of associated protein complexes. In conventional AP-MS workflows the DNA is simply degraded. While this can also be used to study the interactomes of trans-regulatory proteins (e.g. [123]), it loses any information contained in the chromatin environment of the bait proteins. Thus, other methods have been developed to capture also chromatin-associated interactions. Many of them are based on combining chromatin immunoprecipitation strategies with mass spectrometry (see figure 1.14). Here, I will focus on typi-

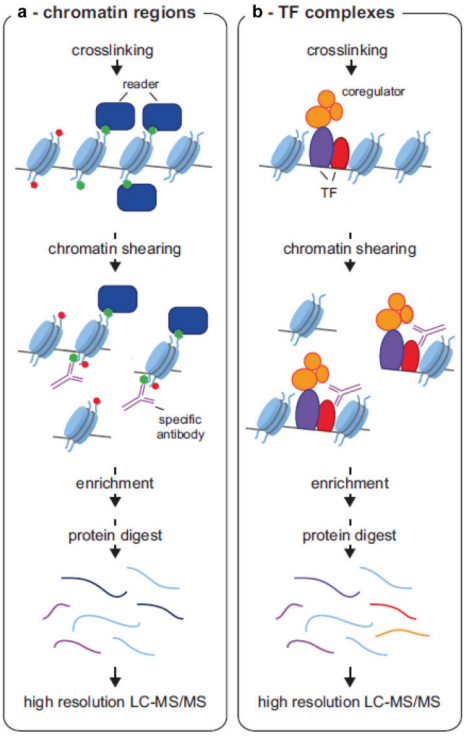


Figure 1.14: Commonly used ChIP-MS methods applied to study chromatin associated protein-protein interactions. a, Immunocapture of modifications on histones allows the identification of proteins localized on specific genomic regions. **b,** ChIP-MS workflow on transcription factors to analyze their co-regulatory interactions on the chromatin. Unlike other ChIP-MS methods, mChIP [125] does not contain a formaldehyde crosslinking step. Adapted from [126].

cal ChIP-MS workflows, which investigate chromatin-associated interactomes of trans-regulatory proteins in a near physiological state (see Figure 1.14a and b). I use ChIP-MS as the umbrella term for similar workflows like Chromatin Proteomics (chroP) [127], modified ChIP (mChIP) [125] and rapid immunoprecipitation mass spectrometry of endogenous proteins (RIME) [128]. ChIP-MS has been applied to various aspects of chromatin proteomics like studying interactions of TFs or associations of proteins with histone modifications (reviewed in [126]). Common ChIP-based methods contain a formaldehyde crosslinking step followed by chromatin shearing (e.g. sonication).

Only few native ChIP-MS protocols have been employed to study chromatin-associated protein interactions [125, 129, 130]. Native pulldowns on reconstituted nucleosomes demonstrated how protein binding depends on the methylation state of histones [129]. In another native ChIP-MS study termed mChIP, the authors used sonication to shear chromatin to about 2000 bp and applied mild washing buffers to preserve protein interactions on the chromatin. The risk of crosslinking indirect protein-protein interactions mediated by RNA or DNA does not exist in native immunoprecipitations, however, it may identify many non-specific interactions due to mild extraction and washing buffers. Specific mChIP abundance factors had to be introduced to remove proteins identified across various baits [130].

In contrast, formaldehyde crosslinking in ChIP-MS allows stringent washing and specific capture of more transient interactions. It is still the predominantly used technique in ChIP-Seq, but has also been applied in various ChIP-MS studies. Researchers have created interaction maps of genomic regions (compare Figure 1.14a) conducting pulldowns on specific histone acetylation and methylation marks (e.g. in mouse embryonic stem cells (ESC) [131,132]). Another central application of ChIP-MS is the identification of TF co-regulatory interactions. This can be done both by endogenously tagging the TF or by using a specific antibody. The latter was used to identify GREB1 as a specific co-regulator of the oestrogen receptor (ER) [128]. The same protocol was later also utilized to show that the progesterone receptor (PR) directly associates with estrogen receptor-α (ER-α) on chromatin [133]. Wang et al. applied ChIP-MS on two subunits of the Drosophila male-specific lethal (MSL) complex, MSL2 and MSL3, tagged with the HTB tag. Mass spectrometric analysis

revealed specific histone modification in MSL target regions including H4 Lys16 acetylation and H3 Lys36 methylation, as well as a novel H3K36-me3 binding protein [134].

The following aspects need to be particularly considered in ChIP-MS in comparison to conventional AP-MS. Firstly, the duration and concentration of formaldehyde crosslinking is crucial. Formaldehyde is toxic for cells and triggers a multitude of repair mechanisms [135] and alters gene expression [136]. Therefore, a balance between effective fixing of complexes and minimization of cellular damage is crucial. Moreover, the crosslinking step has to be highly reproducible and consistent. Significant variations between samples have an imminent impact on data analysis and increases the risk of identifying false-positive interactors. I found crosslinking with 1% formaldehyde to be sufficient and that durations of 10-15 minutes need not to be exceeded. Yet, formaldehyde concentrations of up to 3% have been reported in ChIP-MS studies [134]. Secondly, shearing of chromatin is a trade-off between effective fragmentation and keeping the complexes intact. Experiments on yeast are more simple in that regard as it requires mechanical lysis of the cell wall prior to sonication of the chromatin. This enables highly reproducible fragmentation of the chromatin by sonication. For mammalian cells, sonication is typically used both for lysis and shearing. In my experience this may sometimes lead to variations between individual experiments and thus settings need to be optimized and controlled more tightly. Alternatively to sonication MNase may be used for chromatin fragmentation.

Lastly, buffers are different in ChIP-MS experiments than in AP-MS. Formaldehyde crosslinked complexes withstand low detergent concentrations, but are reversible at high salt concentrations (e.g. 5M NaCl) and high temperatures. Small amounts of detergents like sodium dodecyl sulfate (SDS) are important for effective cell disruption, chromatin shearing and removal of spurious interactors. SDS concentrations should be diluted to concentrations around 0.1% (v/v) in order not to interfere with the affinity enrichment for most tags. Integrating a high salt washing step with 500 mM sodium chloride (NaCl) also helps to wash off non-specific binders. A mild detergent like Triton-X100 should also be included in washing buffers.

Similarly to conventional AP-MS experiments, data analysis is an essential

part. It was reported that establishing bait-specific control groups helps in minimizing the identification of false-positive interactors [122]. Finding unrelated baits allows the assembly of control groups in large scale studies [124]. Chromatin-associated baits are much more likely to have overlapping set of interactors, which makes it more difficult to define specific control groups. Nevertheless, I found the use of bait-specific control groups to be absolutely essential for accurate discrimination between true- and false-positive interactors. Therefore, I developed a concept to find unrelated baits based on the correlation of enriched chromatin-associated and non-chromatin proteins (see Results section 3.3).
ChIP-MS has certainly emerged as the method of choice in studying chromatin-associated interactomes given the wealth of information contained in the chromatin environment.

1.3.3 Mass spectrometry-based analysis of gene-regulatory associations of proteins with chromatin

DNA-centric workflows provide an alternative approach towards the analysis of chromatin-associated protein complexes. Unlike protein-centric methods they focus on enriching specific genomic regions and analyzing the bound proteins (see Figure 1.15). Long-non coding RNAs (lncRNA) have sparked much interest over the past decade as they were identified as a surprisingly crucial regulator of gene expression [137]. Chromatin Isolation by RNA Purification (ChIRP) is an approach to identify proteins associating with these lncRNAs [138]. After formaldehyde crosslinking of lncRNA, DNA and associated proteins, biotin labeled oligonucleotides hybridize to the target lncRNA. Enrichment of the oligonucleotide probes by streptavidin binding allows the analysis of lncRNA DNA targets and the bound protein complexes.
Typically, ChIP-MS and related methods capture the interactions of the bait across the entire genome. Naturally, these associations may vary between different loci. Therefore, it is appealing to study protein complexes at genomic sites individually. Purification of protein-interactions on a single locus is particularly challenging due to the small sample amounts. Considering the detection limit of current standard mass spectrometers (e.g. Q-Exactive HF-X) and

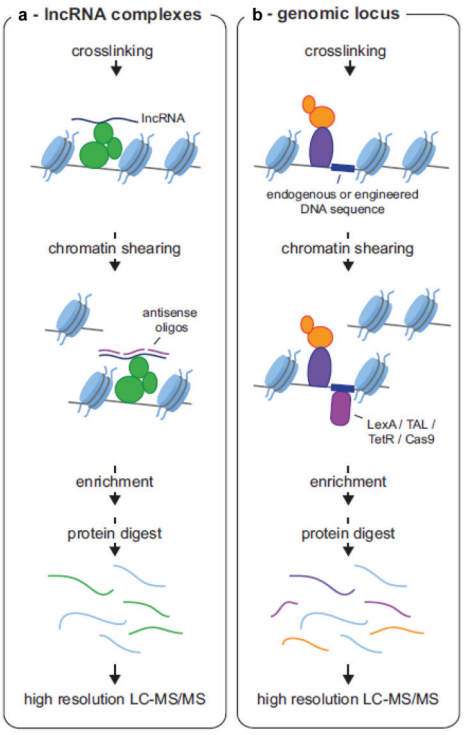


Figure 1.15: Commonly used MS methods to study gene-regulatory associations of proteins with chromatin. **a,** Chromatin-associated lncRNAs are captured by hybridization with biotinylated antisense oligonucleotides ('ChIRP-MS'). **b,** Specific genomic loci are precipitated by binding to a complementary DNA probe or an artifically introduced target site of a DNA-binding protein (e.g. LexA or TetR). Other methods use proteins directly binding to a recognition site like TAL or Cas9-sgRNA. Publications, that employed these methods are referred to in the text. Adapted from [126].

assuming zero loss of material, it would still require at least 500 million cells to identify the most abundant protein on a locus [139]. The first locus-specific enrichment coupled to MS analysis was performed using hybridization-capture with synthetic oligonucleotides termed proteomics of isolated chromatin segments (PICh). In the original development of the method it was used to identify proteins associated with telomere regions. As expected the most abundant proteins were shelterins, nevertheless it also led to the discovery of novel telomere binders [140]. Alternatively, specific recognition sites of a DNA-binding region (e.g. LexA) can be inserted into the DNA. This strategy was successfully employed to show that protein binding and histone modifications strongly deferred in *GAL1* promoter regions between transcriptionally active and repressive states. Gal3, Spt16, Rpb1, Rpb2 and acetylated histones were enriched under active conditions, while H3K36me3 was identified under repressive ones [141]. Specific modification of DNA regions by escorting the DNA endonuclease Cas9 using a guide RNA (CRISPR-Cas9 technique) to genomic regions of interest provided a breakthrough in gene editing. Similarly, it can also be used for MS analysis of proteins associated with single loci. CRISPR affinity purification in situ of regulatory elements (CAPTURE)

combines the expression of a specific guide RNA with a N-terminal FLAG and biotin-acceptor-site (FB)-tagged deactivated Cas9 (dCas9). After biotinylation of dCas9 by the biotin ligase BirA the targeted chromatin locus is purified by binding to a streptavidin matrix [142]. Applied to human telomeres the method effectively captured known telomere maintenance proteins like TERF2, Terf2IP, APEX1, POT1 and 8 novel telomere-associated proteins. Yet, it also missed three major shelterin proteins indicating that improvements are still necessary. Various adaptations and modifications of existing protocols have been made and published (reviewed in [139]) reflecting the interest in single locus proteomics in the field. Nevertheless, many limitations still exist. This is mainly with regards to missing out known interactors or identifying far more non-canonical binders than can be expected [139]. Hence, single locus chromatin proteomics has not been 'solved' yet in a way that it can be applied to a broad variety of biological questions. Considering the advances in MS sensitivity that are currently being made in single cell proteomics will likely have a positive impact on single locus studies, too.

1.4 UV crosslinking mass spectrometry to identify protein-nucleic acid interactions

1.4.1 Characteristics of UV crosslinking

The effect of UV irradiation on the stability and integrity of deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) has sparked the interest of researchers for a long time. This is particularly true for the various cellular effects of UV irradiation like DNA strand breaks, gene mutations and molecular repair mechanisms of UV-induced DNA damages. Here, I focus on a different aspect of UV light, which is its ability to covalently crosslink amino acids to RNA and DNA nucleotides. The concept of photo-crosslinking is appealing in various fields of biological research, as it promises the formation of 'zero-length' crosslinks, which reflect actual contact sites between protein and DNA or RNA.
The first direct evidence of UV-induced formation of a covalent bonds between an amino acid and a nucleobase dates back to 1966, when a cysteine-uracil het-

erodimer was characterized after irradiation with a UV lamp [143]. Four years after this observation, the same phenomenon was reported for the DNA nucleobase thymine and cysteine [144]. These findings accelerated the use of *in vitro* photo-crosslinking to analyze protein-RNA and protein-DNA binding. It was especially helpful to identify protein-RNA interactions as it led to the discovery of proteins binding to mRNA [145], ribosomal RNA [146] and tRNA [147]. UV crosslinking in the characterization of protein-DNA contacts was also described in the 1970s [148, 149], nevertheless not nearly to the extent as in the RNA field. This is likely due to the fact that crosslinking to thymine is less efficient than to uracil, making the downstream analysis much more complicated. This was also established soon after the first publications on the subject by directly comparing uracil and thymine crosslinking efficiencies to several amino acids [150]. Additionally, it was later established that crosslinking to single-stranded DNA is about an order of magnitude higher than to double-stranded DNA [151], which also contributes to differences between RNA and DNA crosslinking. The goal of maximizing crosslinking rates and the idea to prevent non-specific crosslinks by delivering photons on a timescale lower than that of macromolecular rearrangements prompted the use of UV lasers [152]. They achieve crosslinking efficiencies which are orders of magnitude higher than those of conventional UV lamps [153]. Using femtosecond pulses is 30 times more efficient than nanosecond pulses [154]. UV lasers subsequently strongly enhanced the applicability of UV crosslinking to solve biological questions regarding protein-DNA interactions. This is in stark contrast to the field of protein-RNA studies, where conventional UV lamps have been widely used to effectively generate crosslinked protein-RNA complexes.

1.4.2 Characterization of RNA-binding proteins by UV crosslinking mass spectrometry

Two different types of downstream analyses are most often applied to characterize the UV captured protein-RNA interactions (Figure 1.16). Arguably the most prominent protein-centric method to study protein-RNA binding is UV

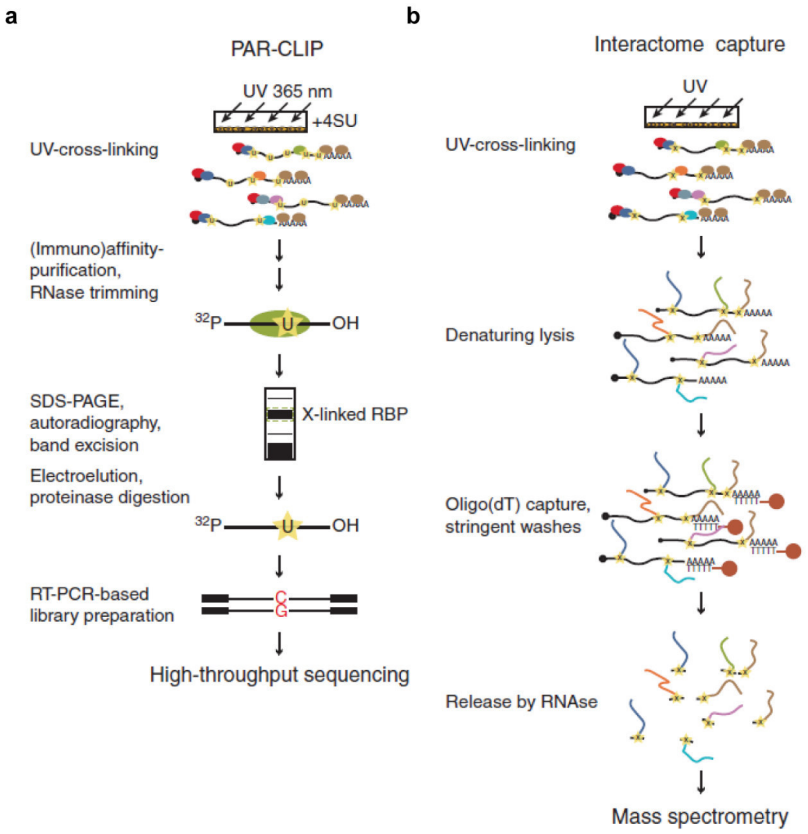


Figure 1.16: Standard protein-centric and RNA-centric analyses after UV protein-RNA crosslinking. a, Outline of the PAR-CLIP method. After UV irradiation, the protein of interest is immunoprecipitated, RNA-protein crosslinks subjected to limited RNA digestion, radioactive end-labeling and SDS-PAGE separation. Proteins are digested and RNA fragments sequenced. **b,** RNA interactome capture. After lysis of crosslinked cells, protein-RNA complexes are purified by oligo(dT) capture. RNA-bound proteins are released by RNAse digestion and analyzed via mass spectrometry. Adapted from [155].

crosslinking and immunoprecipitation (CLIP) or modifications thereof (e.g. PAR-CLIP, HITS-CLIP or iCLIP) as outlined in figure 1.16a. The original CLIP method was published in 2003 and allowed researchers to identify the RNA targets of specific RNA-binding proteins [156]. UV crosslinked RNA-protein complexes are extracted from cells and the RNA is partially digested to a length of 60 to 100 nucleotides. Following immunoprecipitation of the protein of interest, the RNA-protein complex is separated by SDS-PAGE and transferred to a nitrocellulose membrane. Protein digestion, 3'- and 5'-ligation

allow PCR amplification and sequencing of the RNA targets. Further developments of this technique used 4-thiouridine (4-SU) labeling to increase crosslinking efficiency with a wavelength of 365 nm instead of 254 nm (PAR-CLIP), next-generation sequencing (HITS-CLIP) or double-affinity purification (iCLIP) [157].

In contrast, RNA-centric approaches identify the proteins binding to specific RNAs. Early methods characterized RNA-binding proteins using monoclonal antibodies [158]. However, it took further advances in mass spectrometry to establish UV crosslinking as the method of choice in the study of RNA-binding proteomes (RBPome). Castello et al. described the first global mRNA interactome combining UV crosslinking and mass spectrometry [159]. Crosslinked cells were lysed and RNA captured by binding to an oligo(dT) resin (Figure 1.16b), which specifically pulls on the polyadenylated mRNA. Purification, elution and MS analysis of the mRNA-crosslinked proteins allowed the authors to define a HeLa mRNA interactome of 860 proteins [159]. Baltz et al. published a similar purification of the crosslinked mRNA interactome back-to-back. They used a SILAC approach for MS analysis resulting in a mRNA interactome of 797 protein, which is comparable to the label-free approach by Castello [160]. These two seminal papers sparked the interest in combining UV crosslinking and mass spectrometry to characterize entire RBPomes. Another study later investigated the conservation of mRNA-binding proteins from yeast to man using a similar mRNA enrichment approach [161]. The pitfall of oligo(dT) capture strategies is the strong enrichment of mRNA over non-polyadenylated RNA. Therefore, it is not suitable for identifying the entire RBPome. Other approaches are necessary for an unbiased isolation of all RNA-protein complexes. The recent developments of protein-Xlinked RNA eXtraction (XRNAX) [162] and orthogonal organic phase separation (OOPS) [163] were a huge leap forward in mapping entire RBPomes. Both methods rely on the phenomenon that in acid guanidinium thiocyanate-phenol-chloroform (TRIZOL) extraction RNA ends up in the aqueous phase, proteins in the organic phase and RNA-protein crosslinks in the insoluble interphase.

The interphase of XRNAX extracts is intensively washed and DNA digested to yield the purified RNA-protein complexes. Crosslinked species contained RNA from all major biotypes. Different downstream analyses of the crosslinked

RBPs are possible [162]. First, individual RBPs can be immunoprecipitated and RNA transcripts sequenced after protein digestion similar to CLIP approaches (compare Figure 1.16a). Second, partial digestion of RNA-protein complexes with trypsin followed by silica column purification enables mapping of entire RBPomes. Finally, complete proteolysis and silica column purification isolates crosslinked peptides. Subsequent RNA digestion and MS analysis precisely identifies the crosslinked amino acid and nucleotide [162].
OOPS follows the same strategy of enriching crosslinked RNA-protein complexes in the interphase of a trizol extract. The authors showed that glycoproteins and other proteins bearing similar physicochemical properties also accumulate in the interphase. However, RNAse digestion of proteins led to the specific migration of RNA-crosslinked proteins into the organic phase [163]. Thus, OOPS circumvents silica column purification in comparison to the XRNAX protocol. In summary, UV crosslinking has become a versatile tool to create specific and irreversible crosslinks between protein and RNA, which allow stringent purification and subsequent analysis of both the crosslinked proteins and RNA.

1.4.3 Novel concepts for the analysis of protein-DNA interactions

In contrast to protein-RNA interactions, UV crosslinking has not had nearly the same impact on studying protein-DNA interactions owing to the aforementioned low crosslinking efficiencies. Nevertheless, some efforts have been made to use UV crosslinking to study protein-DNA interactions. Especially studies employing UV lasers provided insights into the structural features of histone-DNA association [164, 165] and shed light onto the binding kinetics of TBP to its Ad-2 E4 gene promoter [166]. The ultrafast delivery of a high density of photons provides snapshots of dynamic physiological processes, which unraveled the periodic promoter binding of the glucocorticoid receptor and the SWI/SNF complex during chromatin remodeling [167]. In some studies conventional UV lamps enabled analysis of promoter binding of sequence-specific TFs in living cells [168,169]. Yet, it is doubtful if conventional UV light sources

are efficient enough to apply them for global studies in cells. Moreover, the long irradiation times are causing DNA and protein damage which may impede the identification of physiological protein-DNA association studies [170]. The first proof-of-concept study that described a broader use of UV laser crosslinking in cell biology was only published recently. Steube et al. combined UV nanosecond laser crosslinking with ChIP-Seq as a readout (UV-ChIP-Seq). This novel method led to the discovery of previously missed direct binding sites of the B-cell lymphoma 6 (BCL6) TF particularly in heterochromatin areas [170].

In this thesis I combined femtosecond UV laser crosslinking with state-of-the-art mass spectrometers (see Results section 3.1). These data provide the first showcase example for the robust identification of DNA-binding domains in proteins by UV crosslinking and mass spectrometry analysis. This established that the technique can be applied to living cells. Two aspects of UV crosslinking MS workflows are particularly challenging. First, even UV femtosecond lasers often only reach protein-DNA crosslinking efficiencies of about 15% in recombinant complexes (see Results 3.1) and likely much smaller rates in cells. While this is good enough to map DNA-binding domains *in vitro*, it still makes it difficult to obtain enough crosslinked species for system-wide studies on the DNA-binding proteome. This is especially true as purification of DNA-crosslinked proteins needs to be stringent to enable unambiguous identification by MS. Secondly, computational analysis of crosslinked protein-DNA complexes is difficult. The identification of peptides depends for the most part on the MS/MS fragmentation pattern. However, irreversible crosslinks shift the m/z signal of product ions and even reversible crosslinks can alter fragmentation pathways. Conventional analysis tools use the unmodified ion series to identify peptides. Nucleotide shifted fragment ions will not be taken into account for assessment of the fragmented precursor peptide. Moreover, the nucleotide modification may fragment by different mechanisms. Loss of the entire nucleotide is possible, but also sequential neutral losses of water, ammonia, phosphate or the deoxyribose. This creates a multitude of possible fragment ions shifted by different masses. Dedicated software like Rnp(xl), which was created to study protein-RNA crosslinks, prefilters the MS data to peptides carrying any combination of up to 4 nucleotides. Any neutral

loss can be defined Rnp(xl), which makes it well suited for identifying DNA-binding domains up to single amino acid and nucleotide resolution. However, the software does not entail a false discovery rate (FDR) filter for the identified crosslinked peptides and may introduce false-positive IDs. Therefore, individual crosslinks still need to be manually inspected to identify and remove potential false discoveries. This complicates the use of Rnp(xl) in global analyses of DNA-binding domains by UV crosslinking.

While there are still some limitations in the applicability of protein-DNA UV crosslinking for routine laboratories, recent developments like UV-ChIP-Seq and novel approaches in combining femtosecond UV laser crosslinking with state-of-the-art mass spectrometers (see Results section 3.1) will likely spark new interest in the technology. With further improvements of existing purification methods and bioinformatic analyses, it may become feasible to map a DNA-binding proteome, which makes it possible to discriminate between indirect and direct protein-DNA interaction or discovers novel DNA-binding domains in proteins.

2 Aims of the thesis

In my PhD thesis I aimed at developing and applying mass spectrometry-based methods to the diverse field of gene regulation in different organisms. Tight control of transcription is crucial for any living organism to maintain cell functions, proliferate and to react to environmental cues. Various mechanisms have evolved to ensure proper regulation of gene expression. These pathways function at all three stages of gene expression, i.e. the transcriptional, post-transcriptional and post-translational level. They involve a multitude of molecular reactions and signaling pathways, e.g. protein-protein, protein-RNA or protein-DNA interactions, as well as modification or degradation of key regulators.
In order to better understand these interactions, I set out to use mass spectrometry and transfer or improve existing protein interactomics methods to the field of (post-)transcriptional regulation. In five diverse projects, I studied different aspects of gene expression regulation including protein-protein associations, protein-RNA and protein-DNA interactions. This encompassed establishing a UV femtosecond laser crosslinking pipeline to identify protein-DNA interactions by mass spectrometry analysis, identifying interactomes of specific transcriptional regulators in collaborative projects, global analyses of the trans-regulatory network in yeast and the RNA-binding proteome in immune cells. I will briefly outline the aims for each of projects and refer to the results sections for the detailed description of the findings.

Protein-DNA interactions are crucial in the regulation of gene expression. The event of a transcription factor recognizing its target DNA sequence is a pivotal step to initiate or repress transcription of the respective gene. Consequently, methods to study these protein-DNA interactions are of utmost interest to researchers. Typically, chromatin immunoprecipitation combined with deep

sequencing (ChIP-Seq) is applied to analyze genome binding sites of DNA-binding proteins. Recently, UV laser crosslinking has improved the detection of specific DNA-binding compared to the conventionally used formaldehyde, which also causes indirect non-specific protein-DNA crosslinks [170]. I set out to investigate whether mass spectrometry could be used to analyze proteins crosslinked to DNA and localize the peptide or amino acid involved in binding to the DNA. In order to achieve high crosslinking rates, I used a femtosecond UV laser at the Institute of Applied Physics, University of Jena, with the help of Roland Ackermann and Christoph Russmann. Together we optimized laser settings to maximize crosslinking efficiency. I set up a protocol to enrich protein-DNA crosslinks for mass spectrometry analysis and successfully identified peptides shifted by the mass of single or multiple nucleotides. Furthermore, these were highly specific for the DNA-binding regions of the studied transcription factors. Finally, I optimized an existing protocol for chromatin purification to extract crosslinks from UV laser irradiated cells (see Results section 3.1). These findings may pave the way for future global UV crosslinking studies, which would enable the unbiased identification of DNA-binding proteins without any *a priori* knowledge.

UV crosslinking is already widely used in the field of protein-RNA interaction studies. This is due to the higher susceptibility of the single-stranded RNA to be crosslinked to proteins in comparison to double-stranded DNA. For this reason conventional UV lamps cause crosslinks specifically at sites where proteins contact RNA, which allows use of the technique in non-specialized laboratories. A variety of protocols like RNA-IC, XRNAX or OOPS have been developed to analyze proteins crosslinked to RNA by mass spectrometry [159, 162, 163]. In a collaboration with Kai Höfig in the lab of Vigo Heissmeyer at the Helmoltz Center Munich, we intended to define the RNA-binding proteome in T cells as a resource to study post-transcriptional gene regulation. This allowed me to apply my knowledge of protein-nucleic acid UV crosslinking and MS-based proteomics to another field of gene regulatory interactions. Post-transcriptional gene regulation is highly dynamic and crucial for immune cell function. Yet studies of RNA-protein associations in the context of post-transcriptional gene expression control have been limited to specific proteins. We performed label-

free RNA-IC and OOPS experiments in mouse and human T cells. I conducted extensive data analysis which resulted in a core RBPome defined for both organisms, which serves as a database to obtain insights into novel RNA-binding proteins and post-transcriptional regulation in immune cells (Results section 3.2).

Apart from protein-nucleic acid interactions, co-regulatory protein complexes play a fundamental role in gene expression control. One key application of mass spectrometry is the global investigation of protein interactomes. Many studies have mapped soluble protein complexes in yeast and other organisms, however, there have been only very few publications characterizing also the chromatin-associated interactomes of transcription factors. These are crucial as TFs are rarely organized in stable soluble complexes and the chromatin environment strongly impacts their interaction profile. As our group had access to a library of C-terminally GFP-tagged *S. cerevisiae* strains [171], I set out to investigate the interactome of a large proportion of yeast transcription factors in order to obtain insights into the global trans-regulatory network. I first started testing ChIP-MS workflows to establish a robust, straightforward protocol and applied it to 104 GFP-tagged transcription factors, which account for more than half of all confirmed yeast TFs [172]. I realized that the amount of data required a versatile and streamlined analysis tool. Therefore, I began to learn how to use and implement proteomic analyses in Python. In the end, I scripted an extensive, unbiased data analysis pipeline in Python, that allows definition of control groups, correction of skewed enrichment profiles of promiscuous factors, identification of significantly enriched outliers, global LFQ profile correlation, calculation of overlaps with known protein complexes and other analyses. The resulting dataset adds important information to the interactome profile of various TFs compared to previous conventional methods and led to the discovery of proteins of unknown function to be involved in transcriptional regulation (see Results section 3.3).

In a collaborative effort with Alessandro Scacchetti and Peter Becker from the Biomedical Centre, Ludwig-Maximilians-Universität München, our goal was to disentangle the interactomes of the two *D. melanogaster* isoforms Domino

A (DOM-A) and Domino B (DOM-B). I contributed with my knowledge in mapping protein interactions by mass spectrometry. We initially optimized pulldown protocols as we found Domino to be present in the background proteome at high intensities, which hampered enrichment. In the end we opted for anti-FLAG pulldowns of endogenously tagged Domino isoforms and stringent washes with Radioimmunoprecipitation assay (RIPA) buffer. This allowed optimal enrichment of both proteoforms. We recovered the major complex that is shared between both isoforms and has been described before. Moreover, we found interactors that were unique to either DOM-A or DOM-B. These specific interactors hinted at distinctive functions of DOM-A and DOM-B in transcriptional regulation in Drosophila. Indeed, further follow-up experiments by Alessandro Scacchetti on the unique interactors revealed the mechanics behind the trans-regulatory function of DOM-A and DOM-B (see Results section 3.4).

In another collaboration with Carmelo Quarta and Alexandre Fisette in the group of Matthias Tschöp at the Helmholtz Diabetes Center in Munich we aimed at characterizing the interactome of the transcription factor Tbx3 to better understand its role in neural development and body weight regulation. I transferred the ChIP-MS protocol established for yeast to tissues. Tbx3 is only expressed in specific cells of the hypothalamus in adult mice, which initially made it difficult to sufficiently enrich it, but was solved after several adjustments of the protocol. Among known interactors, I discovered novel associated proteins, which are involved in neuronal development and inter- and intracellular signaling. Remarkably, we were thus able to retrieve the interactome of a transcription factor expressed only in sub-regions of the mouse hypothalamus, which severely limited input material. Together with genomic data this led to the hypothesis that Tbx3 controls the cellular fate and differentiation stage in the arcuate nucleus and affect their peptidergic profile. The proteomic results paved the way for a detailed characterization of Tbx3 and its crucial role in the terminal specification of neurons, for maintaining their peptidergic identity and its importance in body weight regulation (see Results section 3.5).

3 Results

3.1 Article 1: Atomic-resolution mapping of transcription factor-DNA interactions by femtosecond laser crosslinking and mass spectrometry

Alexander Reim, Roland Ackermann, Jofre Font-Mateu, Robert Kammel, Miguel Beato, Stefan Nolte, Matthias Mann, Christoph Russmann & Michael Wierer. **Atomic-resolution mapping of transcription factor-DNA interactions by femtosecond laser crosslinking and mass spectrometry**. *Nat Commun* 11, 3019 (2020). https://doi.org/10.1038/s41467-020-16837-x

Mass spectrometry analysis of UV crosslinked protein-RNA complexes is a widely used method. However, it has not entered the field of protein-DNA interactions yet owing to the severely decreased UV crosslinking efficiencies. We used a UV femtosecond laser to boost crosslinking rates and combined it with MS analysis to investigate its application in characterizing DNA-binding proteins. Crosslinking rates strongly depend on the pulse energy and total energy. Using optimal laser settings I analyzed the nature of crosslinks generated in recombinant nucleosomes and TF-DNA complexes by mass spectrometry. I showed that UV laser-induced crosslinks are highly specific for DNA-binding domains of proteins. Finally, we irradiated embryonic stem cells to scrutinize whether we can expand the use of the UV femtosecond laser pipeline. Mass spectrometric analysis revealed various peptides shifted by the mass of mono- or di-nucleotides, that belonged to TFs and were part of the DNA-binding domains. Collectively, this publication showed the specificity of UV laser crosslinking and the use of combining it with MS analysis to characterize DNA-binding of proteins both in recombinant complexes and in cells.

nature COMMUNICATIONS

ARTICLE

Check for updates

https://doi.org/10.1038/s41467-020-16837-x OPEN

Atomic-resolution mapping of transcription factor-DNA interactions by femtosecond laser crosslinking and mass spectrometry

Alexander Reim[1], Roland Ackermann[2], Jofre Font-Mateu[3], Robert Kammel[2], Miguel Beato[3,4], Stefan Nolte[2,5], Matthias Mann[1], Christoph Russmann[6,7] & Michael Wierer[1]

Transcription factors (TFs) regulate target genes by specific interactions with DNA sequences. Detecting and understanding these interactions at the molecular level is of fundamental importance in biological and clinical contexts. Crosslinking mass spectrometry is a powerful tool to assist the structure prediction of protein complexes but has been limited to the study of protein-protein and protein-RNA interactions. Here, we present a femtosecond laser-induced crosslinking mass spectrometry (fliX-MS) work flow, which allows the mapping of protein-DNA contacts at single nucleotide and up to single amino acid resolution. Applied to recombinant histone octamers, NF1, and TBP in complex with DNA, our method is highly specific for the mapping of DNA binding domains. Identified crosslinks are in close agreement with previous biochemical data on DNA binding and mostly fit known complex structures. Applying fliX-MS to cells identifies several bona fide crosslinks on DNA binding domains, paving the way for future large scale ex vivo experiments.

[1] Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany. [2] Institute of Applied Physics, Abbe Center of Photonics, Friedrich-Schiller-Universität Jena, Albert-Einstein-Straße 15, 07745 Jena, Germany. [3] Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, 08003 Barcelona, Spain. [4] University Pompeu Fabra (UPF), 08002 Barcelona, Spain. [5] Fraunhofer Institute for Applied Optics and Engineering (IOF), Albert-Einstein-Straße 7, 07745 Jena, Germany. [6] University of Applied Sciences and Arts Hildesheim/Holzminden/Goettingen (HAWK), Von-Ossietzky-Straße 99, 37085 Göttingen, Germany. [7] Brigham and Women's Hospital, Harvard Medical School, 75 Francis Street, Boston, MA 02115, USA. email: christoph.russmann@hawk.de; wierer@biochem.mpg.de

Transcription factors (TFs) are key players in the regulation of gene expression and control a multitude of cellular functions, including differentiation, maintenance of cellular identity, cell homeostasis, as well as highly cell specific functions such as immune response[1]. Due to their pivotal role in cellular signaling, mutations of TFs are often linked to human diseases[2–4].

TFs exert their gene regulatory function through the recognition of specific DNA-binding elements in spatial vicinity of target genes and by the recruitment of coregulators, which may have transcriptional activating or repressing functions. DNA binding is mediated by specific DNA-binding domains (DBDs). Evolution gave rise to various different classes, including zinc finger, HMG-box, leucine zipper, helix-turn-helix, and helix-loop-helix domains[1]. Most DBDs of known and putative TFs are identified and classified by sequence homology to a previously characterized DBD[5] and large-scale studies verified the DNA-binding specificity of several hundred individual domains[6,7]. Nevertheless, for several DNA-binding proteins the DBD is unknown, due to the lack of homology with classical domains. Even for domains that have been proven to bind DNA in a stand-alone context, it is not certain that the domain will have the same functionality in the full-length protein.

The molecular mechanism by which TFs bind to DNA can be elucidated by cocrystallization of protein–DNA complexes, which provides insight into the amino acids that are in closest vicinity to the DNA and therefore most likely involved in DNA binding[8,9]. NMR spectroscopy has been used to gain similar information[10]. Furthermore, the composition and stoichiometry of large protein–DNA complexes can be disentangled using high-resolution electron microscopy (EM)[11]. While all those methods allow to study protein–DNA complexes in great detail, for many TFs they are very time consuming or not feasible at all. In addition, especially for crystallization, they reflect a frozen state, which can be different from the dynamic binding behavior of TFs to DNA in solution.

With the advances in mass spectrometry (MS) over the past decade[12], cross-linking MS (XL-MS) has become a viable complementary method to study the structure of protein complexes. The use of chemical crosslinkers allowed the analysis of stoichiometry and spatial arrangement of proteins organized into large complexes (reviewed in ref. [13]). More recently, XL-MS has also entered the field of protein–RNA interactions. Here, ultraviolet (UV) irradiation can create "zero-length" cross-links in the native state of a protein–RNA complex, meaning the direct covalent attachment of an amino acid to a nucleotide. Pioneering studies applied UV irradiation and MS analysis to identify RNA-binding proteins on a system-wide scale in yeast and mammalian cells[14–16]. Improvements in bioinformatic tools further allowed the localization of RNA-protein cross-links at the level of single amino acids[17], providing complementary information about RNA-binding domains.

Despite these developments in applying UV XL-MS to study protein interaction with RNA, the technology has not been applicable for protein–DNA interactions so far. This is largely due to the fact that double-stranded oligonucleotides are about an order of magnitude less efficiently cross-linked by UV than single-stranded oligonucleotides[18]. Yet, over the last three decades a small number of studies have shown that the efficiency of protein–DNA cross-linking can be increased by using UV lasers[19–24]. For a given total energy, the efficiency of protein–DNA cross-linking was shown to largely depend on the length of the laser pulses. Highest cross-linking efficiency can be reached with an ultrafast femtosecond laser, providing 30 times higher efficiency than a nanosecond laser[20].

To map protein–DNA interaction in a highly specific manner, we here present a pipeline for femtosecond UV-laser-induced cross-linking combined with high-resolution MS (fliX-MS). Our workflow is capable of mapping protein–DNA interactions of in vitro assembled nucleosomes as well as in vitro and ex vivo TF–DNA interactions. Our method successfully confirms protein–DNA binding sites predicted by structural studies, and provides insights into the extent of flexibility within DBDs.

Results

A fliX-MS pipeline to map protein–DNA interactions. UV-laser cross-linking with ultrafast pulses can cross-link TFs and DNA with high efficiency[20]. Here, we developed a pipeline, which combines that technology with a high-resolution MS methodology in order to map DNA–protein interactions on amino acid level (Fig. 1). To this end, we used a femtosecond fiber laser at 515 nm, and further doubled its wavelength to 258 nm with a beta barium borate (BBO) crystal (Fig. 1a). Its frequency was 0.5 MHz and pulse duration about 500 fs. The laser beam was adjusted to 2.5 mm (e^{-2}), in order to match the inner diameter of a 1.5 ml Eppendorf tube containing the sample. Following UV irradiation, we denatured protein–DNA complexes, cut the DNA to mono or short oligonucleotide size using a mix of three different nucleases, and digested proteins to peptides with trypsin and Lys-C (Fig. 1b). We then separated peptides from free DNA with StageTips loaded with C18 material[25], enriched peptide–DNA cross-links using titanium dioxide (TiO_2) coated beads, and analyzed them by high-resolution MS (see "Methods"). Peptide–DNA cross-links were searched in MS data using the RNP(xl) software, which was originally developed for the identification of peptide–RNA cross-links[17] (Fig. 1c). Processing nonirradiated control samples in parallel allowed us to subtract any spectra that were not UV cross-linking specific, massively reducing the search space. To improve detection of true DNA cross-links, we further manually validated and annotated all cross-linked peptide fragmentation spectra, considering y-, b-, and a-ion series, as well as internal fragment ions.

Optimization of cross-linking conditions. To maximize the cross-linking rate and therefore the identification of protein–DNA cross-links, we first optimized the femtosecond UV-laser parameters. UV-dependent DNA cross-linking is a two-photon process and depends on both intensity and pulse length[20]. As the pulse length is determined by the laser setup, we tested different pulse energies, as well as increasing amounts of total energy.

We used a recombinant TF—porcine nuclear factor 1/C (NF1)—and let it bind to a biotinylated oligonucleotide containing its specific DNA-consensus binding site or a mutated version of it (Fig. 2a). As the binding was much stronger for the wild-type binding site, compared with its mutant counterpart, we concluded that the protein–DNA interaction was functional. The minor binding to the mutant consensus site can be explained by the ability of NF1 to bind DNA also in unspecific manner[26]. Next, we UV-irradiated the NF1–DNA complex with a pulse energy of 7 nJ and increasing amounts of total energy followed by western blotting and detection of protein–DNA cross-links using a streptavidin–HRP conjugate (Fig. 2b). There was a direct relationship between total energy and cross-linking yield at the beginning of the curve and only a minor increase of cross-linked species from 350 mJ onward. With higher total energy, we also observed protein–protein cross-links bound to biotin-DNA, reflected in an increasing signal in the higher molecular weight range (Supplementary Fig. 1a).

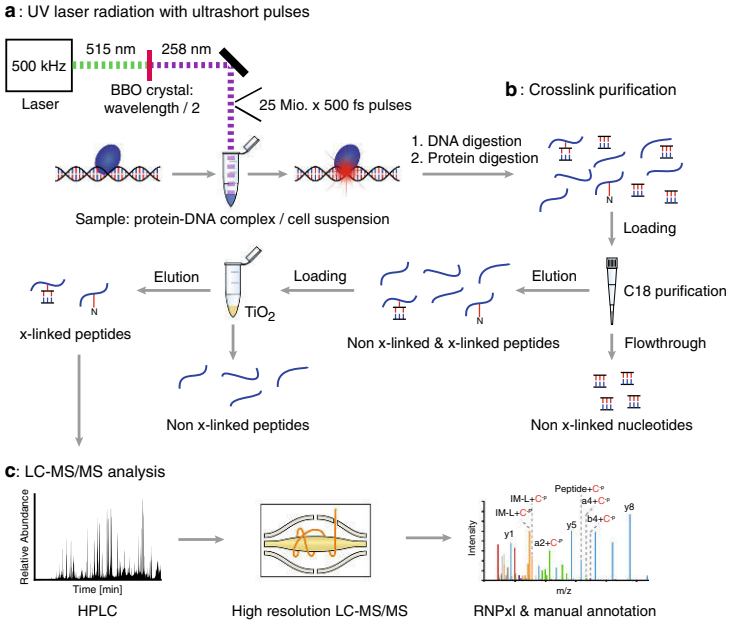


Fig. 1 Schematic workflow of the fliX-MS pipeline. a A pulsed laser beam was generated using a femtosecond fiber laser with 515 nm wavelength, repetition rate of 0.5 MHz, and pulse duration of 500 fs. The wavelength was doubled to 258 nm by second harmonic generation (SHG) over a beta barium borate (BBO) crystal and the laser beam adjusted to fit the inner diameter of a regular 1.5 ml Eppendorf tube. **b** Protein–DNA complexes were irradiated or left untreated as control. Samples were denatured, DNA digested to mono/short oligonucleotides by a mix of Mnase, DNase I, and Benzonase, and proteins digested to peptides by trypsin and Lys-C. Peptides and peptide–nucleotide cross-links were separated from free DNA on C18 StageTips[25], and cross-links subsequently enriched with TiO_2 beads. **c** Peptides were measured by LC–MS/MS and data analyzed with the RNP(xl) software package implemented in the proteome discoverer software[50] followed by manual annotation of candidate spectra.

To determine the optimal pulse energy, we next irradiated the TF–DNA complex with increasing pulse energies, keeping the total energy at 1 J (Fig. 2c, Supplementary Fig. 1b). Maximum cross-linking efficiency occurred at about 40 nJ pulse energy, whereas it strongly decreased at both lower and higher pulse energies. While the lower cross-linking efficiency with less pulse energy can be explained by a minimum energy requirement for the two-photon processes to take place, the reduction at higher pulse rates is either due to saturation effects or DNA damage. We conclude that a maximal energy of 50 nJ per pulse is sufficient to cross-link protein–DNA complexes, and an increase of pulse energy does not enhance the process.

To investigate whether the formed protein–DNA cross-links reflected functional TF–DNA interactions, we repeated the titration of the total pulse energy with the optimal pulse energy of 50 nJ for NF1 bound to a DNA oligo containing either its wild-type consensus binding sequence or a mutated form of it (Fig. 2d, Supplementary Fig. 1c). Western Blot analysis of the biotin-DNA complex revealed that protein–DNA cross-linking was specific for the wild-type sequence. Notably, this was also the case for the higher molecular weight fraction, indicating that protein–protein cross-linking does not affect DNA-binding specificity, even at a total energy of 1.25 J.

To quantify the cross-link efficiency, we irradiated NF1–DNA complex (pulse energy of 7 nJ and a total energy of 350 mJ) and probed the western blot with an antibody directed against the His-tag of NF1 (Fig. 2e, Supplementary Fig. 1d). We observed a shifted double band at 60–65 kDa, which disappeared when digesting the sample with either DNase I or proteinase K suggesting that the signal is derived from the NF1 bound to single- and double-stranded DNA. Reblotting the stripped membrane with the streptavidin–HRP conjugate recognizing biotinylated DNA confirmed this observation. Quantification of the mono-NF1–DNA cross-links revealed a cross-linking efficiency of 7.5%. Taking into account also the high-molecular weight population and extrapolating from the cross-linking efficiency of mono-NF1–DNA and the intensities of the 65, 130, and 185 kDa bands in the DNA-biotin blot, we estimate a cross-linking efficiency of 14% under these energy conditions (Supplementary Fig. 1d).

To validate the observations with another TF–DNA complex, we UV-irradiated recombinant TATA-box binding protein (TBP) bound to an oligo containing either the wild-type TATAA sequence or a single point mutant of it (TGTAA), known to decrease TBP binding by 49%[27] (Fig. 2f). As expected, we observed a stronger signal for the TBP–TATAA complex compared with the TBP–TGTAA, which disappeared with either DNase I or proteinase K treatment indicating that fliX-MS works effectively also for TBP. Of note, the difference in the cross-link efficiency for the two sequences was also visible in the high-molecular weight fraction, corresponding to multiple copies of TBP bound to DNA (Supplementary Fig. 1e).

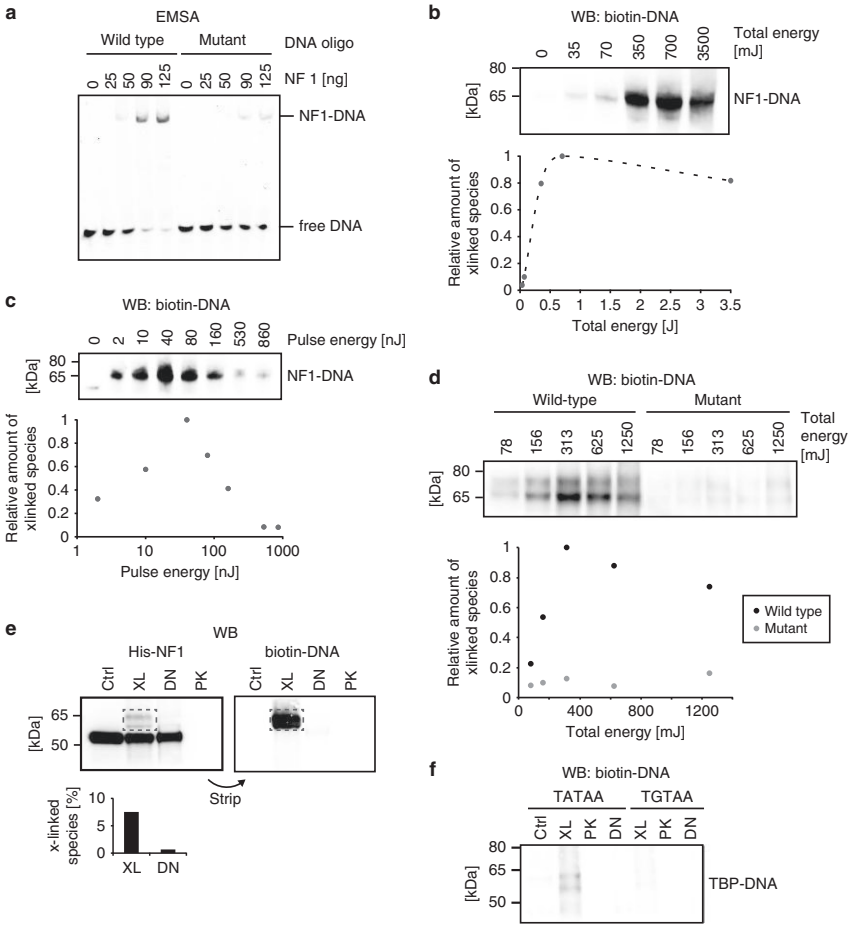


Fig. 2 Assessment of cross-linking efficiencies. a Electrophoretic mobility shift assay (EMSA) with increasing amount of NF1 bound to a DNA oligomer harboring its consensus site, or a mutated version of it. The molar ratios of protein to DNA were 2.1:1 (25 ng), 4.4:1 (50 ng), 8:1 (90 ng), and 11:1 (125 ng). The NF1–DNA complex was separated from free DNA by nondenaturing gel electrophoresis and visualized by SYBR Green staining. **b** NF1–DNA (5′-biotinylated) complex was irradiated with increasing total energy and constant pulse energy of 7 nJ. Samples were separated by denaturing gel electrophoresis, protein–DNA complexes transferred to a nitrocellulose membrane, and biotinylated DNA visualized by probing with an HRP-coupled streptavidin conjugate. Intensities of the cross-linked protein–DNA bands (x-linked species) were quantified and plotted relative to the most intense band at 700 mJ. **c** NF1–DNA (5′-biotinylated) complex was cross-linked applying increasing pulse energies, and a constant total energy of 1 J. Cross-linked protein–DNA complexes were detected as in **b**. Band intensities were plotted relative to the most intense band at a pulse energy of 40 nJ. **d** NF1 bound to a DNA oligo harboring its consensus site or a mutated version of it was irradiated with increasing total energy and constant pulse energy of 50 nJ: cross-linking depended on a functional protein–DNA interaction. **e** NF1–DNA complex was cross-linked with a pulse energy of 7 nJ and 350 mJ total energy (XL) or left untreated (Ctrl). Cross-linked samples were further optionally treated with DNase I (DN) or proteinase K (PK) and loaded on a SDS-PAGE followed by western blotting. After detection of His-NF1 using an anti-His antibody, the membrane was stripped and reprobed with an HRP-coupled streptavidin conjugate to detect biotin-labeled DNA. The percentage of cross-linked protein–DNA complexes (x-linked species) was calculated as the intensity of the cross-linked band (dashed rectangle) divided by the sum of intensities of all bands observed in the cross-linked sample. **f** TBP bound to DNA oligos containing either a wild-type (TATAA) or point-mutated (TGTAA) consensus motif were UV irradiated (pulse energy 50 nJ, total energy 1.25 J) and biotin-DNA detected by western blot. Full-scale versions of all blots are depicted in Supplementary Fig. 1.

a

DCP id	Histone	Peptide sequence	Crosslinked nucleotide	Charge	m/z	Δm [ppm]
1	H4	^{21}KVLRDNIQGITKPAIR^{36}	ACCG-H_3PO_4	3	1002.4468	6.2
2	H4	^{22}VLRDNIQGITKPAIR^{36}	CT-H_2O	3	777.3659	9.5
3	H3.1	^{29}SAPATGGVK^{37}	C-HPO_3	2	507.7652	1.6
4	H3.1	^{74}EIAQDFK^{80}	AT-HPO_3	2	703.2883	6.5
5	H3.1	^{118}VTIMPKDIQLAR^{129}	C-HPO_3	3	543.3030	6.7
6	H2B-1K	^{45}VLKQVHPDTGISSK^{58}	AAGT+HPO_3	2	1433.5139	4.6
7	H2B-1K	^{45}VLKQVHPDTGISSK^{58}	AGGT+HPO_3	2	1441.5155	1.7
8	H2B-1K	^{110}HAVSEGTKAVTK^{121}	GT	2	961.3941	1.7
9	H2A-1B	^{79}IIPRHLQLAIRNDEELNK^{96}	AT	4	702.5960	8.3
10	H2A-1B	^{83}HLQLAIRNDEELNK^{96}	T-HPO_3	3	660.0110	10.0
11	H2A-1B	^{83}HLQLAIR^{89}	AAT-HPO_3	3	573.5804	2.6
12	H2A-1B	^{90}NDEELNK^{96}	AC	2	741.2575	1.1

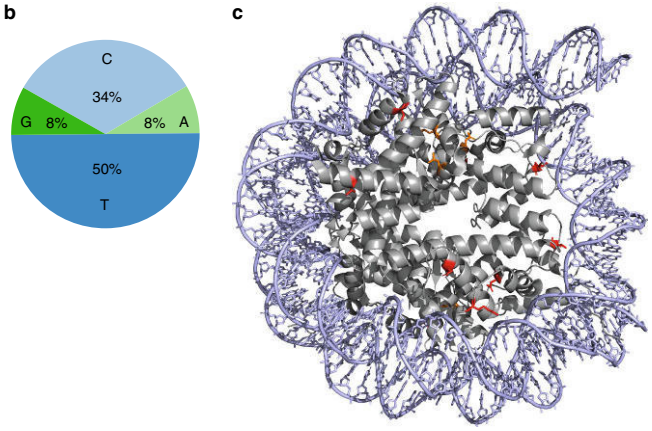


Fig. 3 fliX-MS of in vitro assembled human nucleosomes. a fliX-MS revealed 12 unique peptide–nucleotide cross-links. The sequences of the cross-linked peptides are shown. For cross-links that could be located on one or several amino acids, the location within the peptide is marked in red letters. The cross-linked nucleotide sequence derived from precursor mass differences (A: deoxyadenosine, C: deoxycytidine, G: deoxyguanosine, T: thymidine), charge state, mass-to-charge ratio (*m/z*), and mass error (Δ*m*) are shown. The cross-linked base is marked in red letters. **b** Base distribution among cross-links. **c** Crystal structure of the human recombinant nucleosome (PDB ID: 2CV5[8]), with cross-linked amino acids marked in red (close to DNA) and orange (distant to DNA). For cross-links with more than one potential cross-linked amino acid, the residue closest to the DNA is marked.

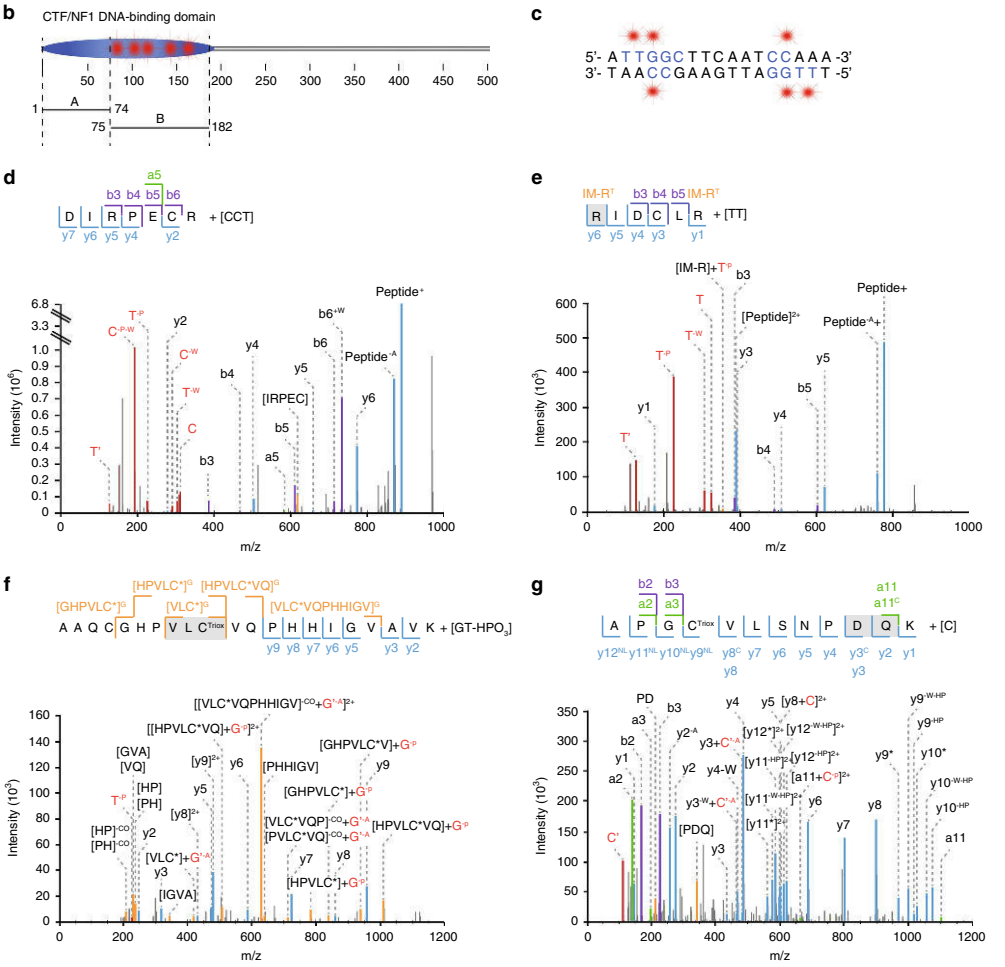
Protein–DNA cross-linking of recombinant human nucleosomes. We next applied the fliX-MS workflow to in vitro assembled human nucleosomes, as this structure involves a large number of protein–DNA contacts. This identified 12 unique peptide–nucleotide cross-links, located on seven different peptides (Fig. 3a, Supplementary Data 1). The cross-linked peptides had MS1 mass shifts corresponding to one to four nucleotides. Considering the base specific MS2 mass shifts, we were able to unambiguously call the nucleotide that was cross-linked in all of the DNA-modified peptides. Cross-links to nucleotides of pyrimidine bases represented the large majority, with six and four cross-links on thymidine and deoxycytidine, respectively. However, fliX-MS also revealed cross-links to nucleotides with purine bases, with one cross-link to deoxyadenosine and one to deoxyguanosine (Fig. 3b). This imbalance between the different base classes is likely due to their different susceptibility to the two-photon processes[28]. In any case, our results show that ultrashort laser UV pulses are capable to cross-link nucleotides of all four bases.

Cross-link-derived mass shifts in MS2 spectra also allowed the localization of the cross-link within the DNA-modified peptides. In seven cases we could pinpoint the cross-link to a single amino acid and in five other cases, we could narrow down the cross-link localization to stretches of two to six amino acids (Fig. 3a).

Comparing our results with the crystal structure of the human nucleosome[8], 8 of the 12 cross-links were in close vicinity to the DNA, with side chains of the respective amino acids pointing toward the DNA double helix (Fig. 3c). Yet, for four DNA-cross-linked peptides (DCPs 9–12, Fig. 3a, c), the distance of the closest possible cross-linked amino acid to the DNA was between 16.5 and 22.1 Å and therefore too large to be explained by a direct protein–DNA contact. As nucleosomes are known to undergo structural changes due to transient unwrapping of DNA[29,30], we hypothesized that the distant cross-links were derived from different conformational states that are not reflected in the crystal structure. In support of this notion, all cross-links that were unexpectedly far away from the DNA in the nucleosome structure, were located on the α3-helix of H2A, which is

a

DCP id	XL-peptide sequence	XL-nucleotide	XL-base	Charge	m/z	Δm [ppm]
13	^{83}DIRPECR^{89}	CCT	C	2	894.7907	6.0
14	^{101}APGCVLSNPDQK^{112}	C	C	2	792.3290	6.4
15	^{117}RIDCLR^{122}	TT	T	2	701.2614	8.2
16	^{140}GIPLESTDGER^{150}	AC-H_3PO_4	C	2	848.3651	6.3
17	^{154}AAQCGHPVLCVQPHHIGVAVK^{174}	GT-HPO_3	G	3	928.4280	5.5



particularly rearranged during partial unwrapping of DNA from the nucleosome[29,30]. We therefore conclude that fliX-MS is able to detect different conformational states of a protein–DNA complex in solution.

fliX-MS applied to the NF1–DNA complex. Next, we enriched peptide–DNA cross-links from the NF1–DNA complex following femtosecond laser irradiation. Subjecting the cross-linked peptides to high-resolution MS, we identified five unique peptides shifted by a mass corresponding to mono-, di-, or trinucleotides in the precursor ions (Fig. 4a). All cross-linked peptides were part of the DBD of the porcine nuclear factor 1/C (amino acids 2–195), demonstrating the structural specificity of fliX-MS (Fig. 4b). In addition, all cross-links were located on peptides between amino acids 83 and 174 indicating a specific binding region in this part of the protein. NF1 and especially its CTF/NF1–DBD are highly conserved across species (Supplementary Fig. 2). Previous experiments using truncated versions of rat NF1 showed that amino acids 75–182 were responsible for

Fig. 4 Mapping protein–DNA interactions in the transcription factor nuclear factor 1/C. a Overview of the identified peptide–nucleotide cross-links. The possible cross-link locations are indicated by red letters in the peptide sequence. Cross-linked (XL)-nucleotide and XL-base information derived from specific MS1 and MS2 mass shifts are specified. **b** Location of the annotated DNA-binding domain of nuclear factor 1/C (NF1) and location of the detected cross-links (red stars). A represents the unspecific DNA-binding subdomain and B the sequence-specific DNA-binding subdomain according to Dekker et al.[26]. **c** Location of the cross-links (red stars) on the palindromic consensus DNA-binding sequence of NF1. Blue letters indicate nucleotides, which fit to the NF1 consensus sequence TTGGC(N)6CC[32]. **d–g** MS2 ion series and spectra of four NF1–DNA cross-links. In the MS2 spectra, nucleotides are annotated in red, amino acids in regular letters. N′ denotes the nucleobase, and N the deoxynucleotide monophosphate (with N being one of the four bases A/T/G/C). The following abbreviations describe neutral losses after MS2 fragmentation: Asterisk: neutral loss of H_2SO_3, −CO: neutral loss of carbon monoxide, −A: neutral loss of ammonia, −/+W: neutral loss or adduct of water, −HP: neutral loss of hydrogen peroxide, −p: neutral loss of HPO_3, −P: neutral loss of H_3PO_4. In the MS2 ion series, cross-linked fragments are depicted with the cross-linked nucleotide (A/T/G or C) in superscript. M^{Ox} represents oxidated methionine and $C^{Trio x}$ trioxidated cysteine (cysteic acid). The prefix IM before the respective amino acid indicates an immonium ion. The superscripted NL represents the neutral loss of sulfurous acid or hydrogen peroxide. All other symbols represent the same neutral losses as in the MS2 spectra.

sequence-specific DNA binding, while amino acids 1–78 had only nonspecific DNA-binding affinity[26]. Notably, all our cross-linked peptides located in the region responsible for sequence-specific DNA binding, highlighting the capability of fliX-MS to detect specific protein–DNA contacts (Fig. 4b, c).

For all cross-linked peptides, we defined the nucleotides that were cross-linked to the peptides making use of characteristic differences in the precursor mass. In addition, specific product ion mass shifts in the MS/MS spectra allowed us to define the exact bases that formed the cross-links (Fig. 4a, d–g). In addition to three cytosine cross-links and one thymine cross-link, one cross-link occurred to guanine, once more underscoring the potential of fliX-MS to cross-link purine bases.

The DNA contact sites of NF1 are known from DNA modification studies[31,32]. To a large extent, DNA binding is mediated by contact to the TTGG motif in the forward strand, as well as additional nucleotides in the reverse strand, which point in the same direction of the double helix (Fig. 4c). Our cross-link data covered interactions of the TTGG motif with two unique peptides (DCPs 15 and 17). In addition, we identified three cytosine cross-links, two of which were specific for the reverse strand (DCPs 13 and 16). While cytosine interactions have not been investigated previously, our data strongly suggest binding to the cytosines opposite of the TTGG sequence. Taken together, all identified cross-links fit to the defined NF1 consensus motif TTGGC(N)6CC[32].

In four out of the five DCPs, mass shifts in the MS2 spectra allowed us to locate the interactions to one, two, or three amino acids. For instance, the peptide RIDCLR cross-linked to a thymidine dinucleotide (DCP15), revealed a specific marker ion of the mass of an arginine immonium ion shifted by the mass of thymidine (Fig. 4e). As the presence of a DNA cross-link on the C-terminal arginine is unlikely due to steric interference during trypsin digest, we allocated the cross-link to R117. This residue is in close vicinity to L121/R122, which in a previous mutation study conferred DNA-binding activity of NF1[33]. On the same line, the seven amino acid long DCP13, which did not reveal a specific cross-linked amino acid (Fig. 4d), overlaps with the C88/R89 mutation site, which also significantly reduced DNA-binding affinity in the previous study.

Analysis of fragment spectra of the other cross-linked NF1 peptides provided additional technical characterization of fliX-MS. Both C104 and C163 were trioxidated to cysteic acid, likely as a result of sample preparation under nonreducing conditions[34–36] (Fig. 4f–g). In the MS2 fragmentation, the trioxidized cysteine underwent neutral loss of sulfurous acid H_2SO_3 (Fig. 4f–g), as has been reported previously[37]. Yet, in case of ^{101}APGCVLSNPDQK^{112}, we also observed an alternative neutral loss of 34.005 Da, which corresponds to the molecular weight of hydrogen peroxide H_2O_2 (Fig. 4g). Moreover, we observed multiple fragments with neutral losses of ammonia on the guanine (Fig. 4f) and cytosine base (Fig. 4g). Such neutral losses have been reported previously for the measurement of free guanine, cytosine, and adenine per MS[38–40]. Including neutral loss of ammonia in the search for MS2 fragment ions that are characteristic for these base adducts strongly enhanced the capability of localizing DNA modifications on individual amino acids. In case of DCP14, the loss of the mononucleotide indicates a cross-link between the amino group of cytosine and the aspartate side chain, which dissociated during higher-energy collisional dissociation (HCD) fragmentation.

Cross-linking of the TATA-box binding TF TBP. We next applied the fliX-MS workflow to human TBP bound to the adenovirus major late promoter containing a TATA box. MS analysis of the cross-linked protein identified four cross-links on three unique peptides (Fig. 5a). As in the case of NF1, all of the TBP peptides with DNA modifications were exclusively located on the DBD of TBP (Fig. 5b).

The precursor of the peptide ^{255}IQNMVGSCDVK^{265} was shifted by the mass of a TT-HPO_3 dinucleotide. Detailed analysis of the MS2 spectrum narrowed down the cross-link to either N257 or M258 (Supplementary Data 1). In the crystal structure of TBP bound to the Adml promoter[41,42], N257 is in close contact to the DNA and located between the two thymines and the two adenines of the complementary strand (Fig. 5c). The distance to either of the thymines is very short with 6.1 or 6.3 Å, respectively, thus both thymines are likely to be cross-linked to the contacting aspartic acid.

In addition, we observed an adenine cross-link to one of the amino acids G217–V220 (Fig. 5d). Based on information from the crystal structure, V220 has been mapped to interact with an adenine in the TATA box[9,42], given an extremely short distance of 3.5 Å (Fig. 5c). Hence, also this cross-link fits to the published structure with high probability. Notably, the same peptide, which contains the V220-A modification, has a second cross-link to a cytosine on A211, which in the crystal structure is located on the fourth strand of the beta sheet (Fig. 5c, d). The closest cytosine is the first nucleotide downstream of the TATAAAA sequence, on the opposite strand, with a distance of 13.4 Å. The coexistence of both cross-links on the same peptide indicates that A211 infers additional DNA binding of TBP, reaching toward a nucleotide adjacent to the TATA box.

The third TBP DCP (^{178}LDLKTIALR^{186}) reflected a cross-link of a cytosine to L178 (Supplementary Fig. 3a). This leucine is located between the four adenine bases and the following guanine stretch downstream of the TATA box. The closest cytosine is the same nucleotide, which we found cross-linked to A211. However, compared with the other TBP cross-links, the distance in the crystal structure to the cytosine is comparably large (17.3 Å, Supplementary Fig. 3b). One explanation to this discrepancy

a

	fLiX-MS					Structure Integration	
DCP id	Peptide sequence	Crosslinked nucleotide	Charge	m/z	Δm [ppm]	Crosslink location	Distance to nucleotide [Å]
18	^{209}TTALIFSSGKMVCTGAK^{225}	(A+C)-HPO$_3$	3	773.6829	8.5	A211(C) & V220(A)	3.5/13.4
19	^{255}IQNMVGSCDVK^{265}	TT-HPO$_3$	2	902.3474	3.4	N257(T)	6.1
20	^{178}LDLKTIALR^{186}	C-HPO$_3$	2	635.3821	3.4	L178(C)	17.2

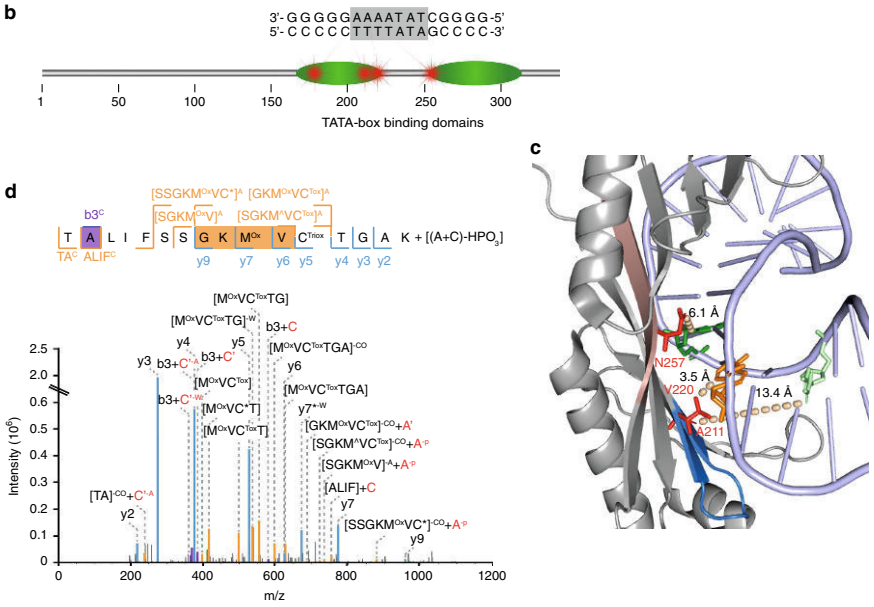


Fig. 5 Cross-linking of the TATA-box binding transcription factor TBP. a Overview of the four identified cross-links on three unique peptides. Blue background indicates information obtained from cross-linking experiments and gray background information obtained from the crystal structure (PDB ID: 1C9B[41]). Red letters indicate possible cross-linked amino acids or cross-linked nucleotides, respectively. **b** Schematic view of the domain structure of TBP, the TATA box (gray shading), and surrounding nucleotides, as well as the cross-links. **c** Crystal structure of TBP bound to an extended AdmI promoter (PDB ID: 1C9B[41]). Location of the cross-link of N257 to one of the two thymidines (green) and of the cross-links of amino acid V220 to deoxyadenosine (orange) and A211 to deoxycytidine (light green). Cross-linked amino acids are depicted in red, the peptide ^{255}IQNMVGSCDVK^{265} in light red, and the peptide ^{209}TTALIFSSGKMVCTGAK^{225} in blue. Dashed lines represent the distance of amino acid to nucleotide and distance is shown above the line. **d** MS2 ion series and spectra of the cross-linked TTALIFSSGKMVCTGAK peptide. Abbreviations in the MS2 spectra and MS2 ion series: caret: neutral loss of CH_4SO, Tox: trioxidated cysteine (cysteic acid). Other abbreviations as in Fig. 4.

could be a higher flexibility of the TBP–DNA complex in solution, compared with the "frozen" picture of the crystal structure.

An interesting observation in the MS2 spectrum of the ^{178}LDLKTIALR^{186} peptide is that its fragment ions y6, y7, y8, and y9 are exclusively observed with a mass shift of +27.995 Da, corresponding to the addition of carbon monoxide (CO) (Supplementary Fig. 3a). Searching for the source of this adduct, we analyzed all peaks in the lower *m/z* range and identified a prominent peak at $m/z = 89.06$ that equaled deoxyribose after loss of CO. Together with a strong marker ion of [deoxycytidine −CO], this provides evidence that the CO adduct is derived from the deoxyribose part of the deoxycytidine, which is additionally cross-linked to the central lysine of the peptide and cut off during HCD fragmentation (Supplementary Fig. 3c). Therefore, we hypothesize that both L178 and K181 were cross-linked to deoxycytidine at the same time and to different parts of the nucleotide.

Ex vivo fliX-MS in mouse embryonic stem cells (ESCs). Having established that fliX-MS is highly specific for cross-linking protein–DNA interactions in in vitro assembled protein–DNA complexes, we next asked whether the method could be also applied to cells. To investigate this, we resuspended mouse ESCs (mESCs) in phosphate-buffered saline (PBS) and subjected them to femtosecond UV-laser radiation. We isolated chromatin from the cross-linked cells, following a DNA biotinylation protocol[43], and enriched peptide–DNA cross-links as in the standard fliX-MS workflow (Fig. 6a). Comparison with a nonirradiated control allowed the identification of specific peptide–DNA cross-links.

Analyzing the data with the RNP(xl) software identified several high-confidence cross-links on TFs. Among those, we manually annotated and validated six bona fide cross-links (Fig. 6b, d, e, Supplementary Fig. 4c–e). All cross-links were exclusively present on the DBDs, which once more highlights the specificity of fliX-MS. In addition, fliX-MS was capable to cover different types of protein–DNA interactions, as cross-linked DBDs represented

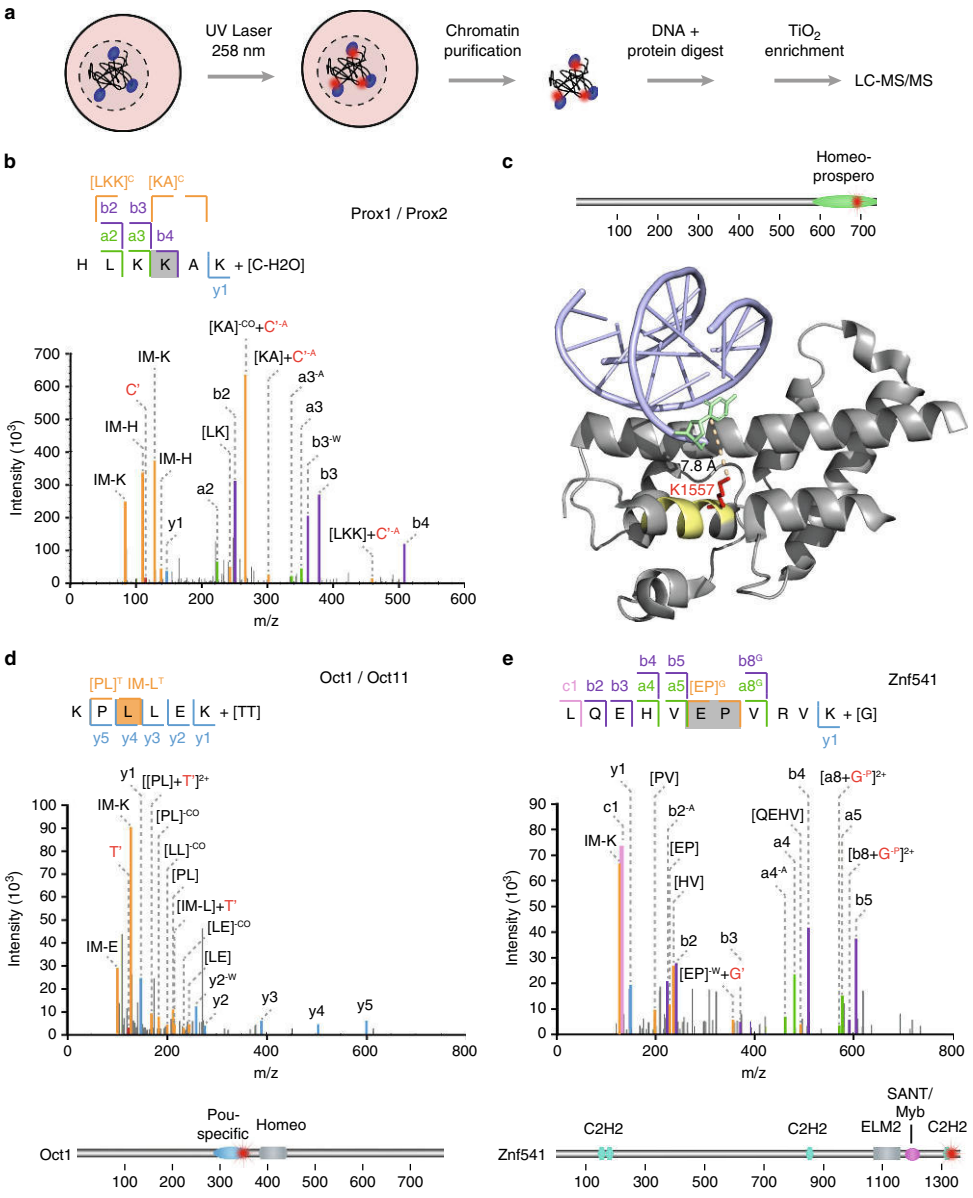


Fig. 6 fliX-MS applied to mouse embryonic stem cells. a Schematic overview of the chromatin purification and enrichment of peptide–DNA adducts from laser UV-cross-linked embryonic stem cells. **b** MS2 ion series and spectra of the transcription factor Prox1/2 peptide ^{584}HLKKAK^{589} cross-linked to a deoxycytidine monophosphate. Abbreviations as in Fig. 4. **c** Crystal structure of the *D. melanogaster* prospero domain (PDB ID: 1XPX[64]) and location of the ^{1552}HLRKAK^{1557} peptide, which is homologous to the cross-linked ^{584}HLKKAK^{589} peptide. The peptide ^{1552}HLRKAK^{1557} is highlighted in yellow, deoxycytidine in green, and K1557 in red. The distance of K1557 to the cytosine is indicated. **d**, **e** MS2 ion series of two other high-confidence cross-links of the transcription factor Oct1/Oct11 and the zinc finger protein Znf541. Abbreviations as in Fig. 4. Schematic representation of the cross-link location and domain structure is shown below the respective MS2 spectrum.

four different classes, including homeo–prospero, bHLH, ZNF, and SANT/Myb domains.

The Prox1/2 peptide ^{584}HLKKAK^{589} was cross-linked to a deoxycytidine monophosphate via K587 and is part of the DNA-binding homeo–prospero domain (Fig. 6b, c). Since this domain is fairly large, we wondered whether the interaction would agree with known structural data. Locating the ^{584}HLKKAK^{589} cross-link in the crystal structure of the highly conserved prospero protein in *D. melanogaster* (Supplementary Fig. 4a), we observed that the *Drosophila* counterpart peptide (^{1552}HLRKAK^{1557}) is in an alpha helix in close vicinity to the DNA, where K1557 points toward the deoxycytidine with a distance of 7.8 Å (Fig. 6c). This demonstrates that the ex vivo generated cross-link specifically reflected a TF–DNA binding event.

The peptide KPLLEK was cross-linked to a dithymidine and could be mapped to several different TFs, namely Oct1, Oct2, Oct11, and Hes2, as well as to the mitotic spindle assembly checkpoint protein Mad2l2 (Fig. 6d). Analyzing the proteome of the same murine ES cell line to a depth of >9700 proteins (Supplementary Fig. 4b) revealed exclusive expression of Oct1 and Oct11 in this dataset, suggesting that the cross-linked peptide is derived from one of the two proteins. In both cases the peptide forms part of the conserved Pou-specific DBD, again underlining the feasibility of fliX-MS to identify functional ex vivo protein–DNA contacts.

The high-confidence cross-linked peptides of Znf541, Smarca1, Zfp91, and Znf354c supported this further (Fig. 6e, Supplementary Fig. 4c–e). As for the other two ex vivo cross-links, our data defined both the exact cross-linked nucleotide, as well as the amino acid position with a precision of maximum two adjacent amino acids. Of technical note, the spectrum of Znf541 contained a rare C1 ion, which can be formed during HCD fragmentation of peptides with an asparagine or glutamine in second position[44].

Discussion

Although interaction of TFs with DNA is a hallmark of gene transcription, it has remained an understudied area of biology due to several technical limitations: (i) Current methodologies such as chromatin immunoprecipitation followed by next-generation sequencing (ChIP-Seq) or proteomics (ChIP-MS) cannot differentiate between direct DNA binding and co-recruitment via other DNA-binding proteins. (ii) Direct TF–DNA binding assays depend on the availability of recombinant proteins and do not necessarily reflect DNA binding in living cells. (iii) Cocrystallization or NMR of protein–DNA complexes are highly laborious and not even possible for many TFs. Hence, a tool to directly assign protein–DNA interactions with amino acid and nucleotide resolution would have a strong impact on biological research.

High-intensity femtosecond lasers provide a plethora of applications reaching from ultrafine material processing[45], high-precision medical surgery[46], to the detection of biomolecular processes[47]. In the search for effective cross-linking methods of proteins and DNA, we and others have previously shown that femtosecond lasers are promising for this purpose because they provide high cross-linking yields while minimizing DNA damage[20,24,48,49]. With recent advances in XL-MS in the sample preparation, MS instrumentation, and bioinformatics side[17,50], we here combined this highly effective cross-linking strategy with an optimized purification protocol for cross-linked peptides, and MS-based read out of protein–DNA cross-links. Our method can map protein–DNA interactions both in vitro as well as in cells, making it a powerful tool for many different research topics.

As a proof of principle, we applied our femtosecond laser-induced cross-linking followed by high-resolution MS (fliX-MS) pipeline to in vitro assembled nucleosomes, as well as to recombinant TFs. Notably, we were able to detect cross-links to all four DNA bases. For recombinant TFs, all cross-links mapped exclusively to annotated DBDs, providing confidence for future applications of fliX-MS for the de novo identification of protein–DNA interactions. Although UV cross-linking in addition to DNA–protein cross-links also produced protein–protein cross-links, the observed DNA–protein cross-links strongly depended on a specific DNA-consensus site, suggesting that femtosecond UV-laser irradiation does not interfere with the protein conformation.

One technical limitation of the current fliX-MS workflow is the dependency on enzymatic protein digestion for MS analysis. In case of the nucleosome, many of the annotated amino acid–DNA contacts locate in regions, which are enriched in lysine and arginine residues and the resulting peptides are often too short to be measurable by LC–MS/MS. For instance, histone H2A has seven annotated DNA-binding sites (R30, R33, R36, K37, R43, K75, and R78, Interpro: P04908) in regions where tryptic digestion would produce peptides less than seven amino acids in length, which are difficult to observe by MS analysis. This limitation could be overcome by the use of enzymes with different specificity such as Arg-C or chymotrypsin, or by chemically modifying all lysine residues in the protein complex, which is commonly applied for the analysis of histone posttranslational modifications by MS[51].

Apart from localizing protein regions, our method revealed detailed structural information of DNA–protein interactions, especially where no crystal structure was available. Despite being one of the first studied DNA-binding proteins[52], mechanistic information on DNA interaction of nuclear factor 1/C (NF1) has been limited to mutation[33] and truncation[26] studies, as well as DNA-binding analyses in combination with modified bases[31]. Notably, our fliX-MS data on the NF1–DNA complex were in close agreement with the previous biochemical data. All cross-links were in the subregion of the CTF/NF1 binding domain that was reported to confer sequence-specific DNA-binding activity[26], while no cross-link was found in the remaining part of the CTF/NF1 binding domain that mediates only unspecific DNA binding. Furthermore, two cross-linked amino acids were in close vicinity to mutation sites that had been shown to reduce or eliminate DNA binding[33]. Taking advantage of the sequence information provided by the cross-linked di- and trinucleotides, we explicitly localized the cross-links on the NF1 consensus sequence in four out of five cases, confirming the interaction with both DNA strands originally proposed of early NF1–DNA contact site analyses[31]. In addition, we revealed interactions of NF1 with the cytosines on the TTGGA reverse strand, which have not been observed before. Given the detailed information of binding contacts from our experiments, molecular modeling of the NF1–DNA complex might now be feasible. In fact, the CTF/NF1 domain shares structural homology with the structurally resolved SMAD DBD[53]. With the additional information gained by fliX-MS, we envision that the structure of the NF1–DBD in complex with DNA can be finally resolved.

Comparing our data on recombinant nucleosomes and TBP bound to its target DNA with the respective crystal structures showed that the peptide–DNA cross-links were largely in agreement with the intramolecular distances in the electron density maps. However, three cross-links of the nucleosome, and two of TBP revealed distances between amino acid side chains and nucleotides that were too large (>16 Å) to support a direct contact according to the crystal structure. The most likely explanation is that our method is capable to detect different conformational states of protein–DNA complexes in solution, while crystal structures reflect only a single discrete structural conformation. In

support, cryo-EM studies on nucleosomes[29,30] revealed a large degree of structural dynamics, based on partial unwrapping of the DNA, also known as DNA breathing. Notably, all distant cross-links lie on an H2A helix, which was described to be especially susceptible to conformational rearrangements in the nucleosome[29,30]. In case of TBP, the two distant cross-links all pointed to the same nucleotide, namely the first cytosine downstream of the TATA box on the reverse strand. TBP binding to the DNA requires significant DNA deformation, including opening of the minor groove and a reduction of the helical twist[9,54]. To generate the cross-links identified here, the DNA must be able to take up a much stronger deformed conformation than the crystal structure would suggest. Taken together, this demonstrates that fliX-MS is capable to add additional information to crystal structure data, by providing evidence for structural flexibility of certain subregions.

Having established the potential of fliX-MS to accurately map DNA-binding contacts in vitro, we were encouraged to also extend our cross-linking strategy to cells. Despite a potential for optimizing both chromatin enrichment efficiency and MS sensitivity much further, we were able to identify several bona fide examples of TF–DNA cross-links. Reassuringly, these cross-links were all located on DBDs, suggesting that fliX-MS can indeed identify specific protein–DNA interactions in cells. In addition, our method might be also applicable to DNA-pulldown experiments[55], after laser irradiation of the eluted protein–DNA complexes. This would be especially useful for analysis of selected TFs, which cannot be expressed recombinantly.

In conclusion, we have developed a workflow to map protein–DNA contacts in both in vitro and cellular contexts. Given the scientific importance of such contacts, we believe that fliX-MS will have major impacts in many fields of biology and even clinical research. Current developments on both MS technology and data analysis side may even allow the mapping of global DNA interactomes in near future.

Methods

Protein expression and purification. For the reconstitution of recombinant human nucleosomes, histone proteins (H3.1, H4, H2A, H2B) were expressed in *E. Coli* BL21 (DE3) cells and purified via inclusion body preparation, denaturing gel filtration, and ion exchange chromatography[56–58] (see Supplementary Fig. 5). The pUC19-16×601 plasmid was amplified in *E. Coli*, the 601 strong positioning DNA sequence excised by digestion with EcoRV and purified by PEG precipitation[58]. Purified DNA was further digested with EcoRI and biotinylated with biotin-11-dUTP (Jena Biosciences) using Klenow fragment (3′ → 5′ exo-) polymerase (NEB). Finally, histones were refolded into octamers and nucleosomes reconstituted by salt gradient dialysis[56].

6xHis-tagged recombinant NF1 was cloned into a baculovirus vector, expressed in Sf9 cells, and purified by nickel column chromatography[59].

Recombinant TBP was purchased from Active Motif (81114).

Assembly of protein–DNA complexes. Sequences of the DNA oligonucleotides used for in vitro experiments were: NF1: 5′-AAT TCC TTT TTT TGG ATT GAA GCC AAT CGG ATA ATG AGG-3′ (sense, wild type), AAT TCC TTT TTT TGC GCT AAA GCG TAG TGG ATA ATG AGG (sense, mutant) for all experiments except Fig. 2d, 5′-AAG TCC TTT TTT AGG ATT GAA GCC AAT CGG CTG ATG AGG-3′ (sense, wild type), 5′-AAG TCC TTT TTT AGC GCT AAA GCG TAG TGG CTG ATG AGG-3′ (sense, mutant) for Fig. 2d; TBP: 5′-CCT GAA GGG GGG CTA TAA AAG GGG GTG GGG GCG CG-3′ (sense, wild type), 5′-CCT GAA GGG GGG CTG TAA AAG GGG GTG GGG GCG CG-3′ (sense, mutant). For each sequence both sense and antisense oligonucleotides were synthesized with a biotin covalently linked to the 5′-end. Double-stranded DNA probes were generated by incubating 100 pmol of each sense and antisense oligo in 25 µl of annealing buffer (10 mM Tris-Cl pH 7.5, 50 mM NaCl, 1 mM EDTA) at 95 °C for 5 min followed by cooling down to room temperature for 60 min. For all western blots and fliX-MS experiments other than Fig. 2d, 13 µg NF1 protein (234 pmol) was incubated with 30 pmol of annealed DNA for 25 min at room temperature in 50 µl NF1 binding buffer (90 mM NaCl, 0.5 mM EDTA, 5% glycerol, 0.55 mM β-mercaptoethanol, 5 mM Tris-Cl (pH 8.0), 1 µg BSA). For the western blot in Fig. 2d, 4.2 µg NF1 (74 pmol) was incubated with 14 pmol of annealed DNA for 25 min at room temperature in 50 µl NF1 binding buffer containing 200 mM NaCl. For all experiments with TBP 15 µg (380 pmol) of recombinant protein was incubated with 30 pmol of DNA for 25 min at RT in 200 µl TBP binding buffer (NF1 binding buffer + 2 mM $MgCl_2$). For electrophoretic mobility shift assays (EMSA), 0, 25, 50, 90, or 125 ng NF1 (corresponding to 0, 442, 885, 1592, and 2212 fmol) were incubated for 25 min with 200 fmol DNA in 20 µl NF1 binding buffer containing 200 mM NaCl for 25 min at room temperature.

Electrophoretic mobility shift assay. 5× TBE Hi-Density buffer (15% Ficoll (w/v), 5% glycerol (v/v), 1× TBE Buffer (Invitrogen, 15581044)) was added in a 1:4 ratio to assembled protein–DNA complexes and samples separated on a 6% DNA retardation gel (Invitrogen) at 100 V in 0.5× TBE buffer for 45 min. The gel was incubated with 3 µl SYBR Green I (Sigma-Aldrich) diluted in 30 ml of 1× TBE buffer and rocked for 20 min at room temperature. Excess SYBR Green I dye was washed off by rinsing the gel three times with MilliQ water. DNA was visualized on a LAS4000 Image Quant (GE Healthcare).

Western blot. For DNase I and proteinase K experiments from Fig. 2e, f, samples were diluted fourfold to final concentrations of 10 mM Tris-Cl (pH 8.0), 2.5 mM $MgCl_2$, and 0.5 mM $CaCl_2$. One microliter of DNase I or one microliter of proteinase K was added for digestion experiments and left at 37 °C (DNase I and untreated) or 56 °C (proteinase K) for 1.5 h before denaturation. All other cross-linked samples were denatured directly before denaturing gel electrophoresis. To this end, 4× LDS sample buffer (Invitrogen, NP0007) was added to the cross-linked samples in a 1:3 ratio and samples boiled for 5 min at 95 °C. Proteins were separated on a NuPAGE 4–12% Bis-Tris Gel (Invitrogen) at 150 V for 45 min in 1× MOPS Running Buffer (Invitrogen) and transferred onto a nitrocellulose membrane (GE Healthcare) at 75 V for 90 min at 4 °C in 1× blotting buffer (25 mM Tris-Cl, 192 mM glycine (pH 8.3), 20% methanol). For detection of biotin-labeled DNA, blots were blocked with 15 ml blocking buffer (Active Motif EMSA kit, 37341) for 15 min and incubated with 50 µl streptavidin–HRP conjugate (Active Motif EMSA kit, 37341) in 15 ml blocking buffer for 15 min. Blots were washed three times with 10 ml TBS-T buffer (150 mM NaCl, 50 mM Tris-Cl (pH 7.6), 0.1% Tween-20). Fifteen milliliters Substrate Equilibration Buffer (Active Motif EMSA kit, 37341) was added and blots incubated at RT for 5 min. DNA was visualized by incubation with chemiluminescent reagent (WESTAR ηC 2.0, Cyanagen) for 1 min and imaged on the LAS4000 Image Quant. For the detection of 6xHis-tagged NF1 protein, blots were blocked with western blocking buffer (5% skim milk powder in 1× TBS-T buffer) for 45 min followed by incubation with anti-6xHis antibody (MA1-21315, Invitrogen) diluted 1:2000 in western blocking buffer overnight at 4 °C. Blots were washed three times for 10 min with 1× TBS-T buffer and incubated with HRP-coupled anti-mouse IgG (GE Healthcare, NA931) antibody, diluted 1:4000 in 0.5× western blocking buffer (2.5% skim milk powder in 1× TBS-T buffer), for 1 h at room temperature. Blots were washed three times for 5 min with 1× TBS-T, incubated with chemiluminescent reagent for 1 min, and visualized on the LAS4000 Image Quant. For membrane stripping, blots were washed twice in TBS and incubated 10 min with 10 ml of Restore PLUS Western Blot Stripping Buffer (Thermo) at room temperature and washed four times with TBS. Band intensities were quantified using ImageJ version 1.52a. All blots are shown in Supplementary Figs. 1 and 6.

Femtosecond laser-induced cross-linking. For UV irradiation, a femtosecond fiber laser (active fiber systems GmbH, Jena, Germany) with a wavelength of 1030 nm, doubled to 515 nm, was used with a pulse duration of about 500 fs. The wavelength was further doubled to 258 nm using a BBO crystal (Laser components GmbH, Olching, Germany). The laser average power is limited for repetition rates of 0.5–20 MHz, which was adjusted to 0.5 MHz to provide sufficiently high pulse energy for frequency conversion. The laser beam diameter was adjusted to 2.5 mm (e^{-2}), in order to fit the inner diameter of a 1.5 ml Eppendorf tube. For a typical pulse energy of 50 nJ (or 25 mW average power), this diameter results in a peak intensity of 4 MW cm^{-2} on the beam axis.

For fliX-MS experiments, 100 µl (45.9 µg) of recombinant human nucleosomes, 100 µl assembled NF1–DNA complex, or 200 µl assembled TBP–DNA complex (see above) were irradiated with 1.25 J total energy and a pulse energy of 50 nJ (25 Mio. pulses with 500 fs pulse length), or left untreated as control.

For UV radiation titration experiments, 25 µl of NF1–DNA complex was irradiated with varying settings as mentioned in the text. For ex vivo cross-linking experiments, 20 Mio. mESCs (E14TG2a) were resuspended in 100 µl PBS and irradiated with 2.1 J total energy and 42 nJ pulse energy.

Digestion of in vitro samples. Individual enrichments were performed with 75 pmol of cross-linked NF1–DNA complex (molar amount of DNA), 150 pmol of cross-linked TBP–DNA complex, or 200 µg of cross-linked recombinant mononucleosomes (containing 91.8 µg of histone octamer and 941 pmol of DNA), each pooled from multiple UV radiation samples. Urea and Tris-Cl (pH 8.0) were added to the cross-linked samples to final concentrations of 4 M and 50 mM, respectively. After 5 min incubation, urea was diluted to 1 M with 50 mM Tris-Cl (pH 8.0). $CaCl_2$ was added to 5 mM and $MgCl_2$ to a final concentration of 2 mM. One microliter of MNase (New England Biolabs, M0247S), one microliter of DNase I (New England Biolabs, M0303S), and three microliters of Benzonase (Merck Millipore, 70746) were added to every 150 pmol of DNA. DNA digestion was

carried out for 90 min at 37 °C. Trypsin and Lys-C were added at a ratio of 1:40 (w/w) compared with the protein amount and incubated for 30 min at 37°C, followed by overnight incubation at 25 °C. The next day, formic acid (FA) was added to 0.1% final concentration.

Purification of chromatin associated proteins. Chromatin extraction and purification from cross-linked mESCs was performed by adapting a published chromatin biotinylation protocol[43]. Three cross-linked or three non-cross-linked control cell pellets, respectively, were resuspended in 300 µl cell lysis buffer (20 mM Tris-Cl (pH 8.0), 85 mM KCl, 0.5% NP-40, 1× PIC (Roche cOmplete)) and immediately centrifuged for 5 min at 2000 × *g* and 4 °C. Pellets were resuspended in 300 µl SPC-NEB buffer (1 M sorbitol, 50 mM Tris-Cl (pH 8.0), 5 mM $CaCl_2$, 1× PIC) and incubated at 37 °C for 1 min. Six microliter of MNase (New England Biolabs) was added and samples incubated for 13 min at 37 °C. EDTA was added to a final concentration of 50 mM. Nuclei were pelleted by centrifugation at 2000 × *g* and 4 °C for 5 min. After resuspension with 300 µl 0.2% SDS buffer (0.2% SDS, 20 mM Tris-Cl (pH 8.0), 1 mM EDTA pH 8.0, 1× PIC), samples were sonicated in a Bioruptor (Diagenode) at a low setting for three cycles, 30 s on/30 s off. After centrifugation at 8600 × *g* and 4 °C for 5 min, the supernatant was removed and dialyzed twice (once overnight and once for 6 h) using a 10,000 MWCO membrane (Slide-A-Lyzer, Thermo Fisher) against 3 l of NEB buffer 2 (50 mM NaCl, 10 mM Tris-Cl (pH 7.9), 10 mM $MgCl_2$, 1 mM DTT). BSA was added to 100 µg ml^{-1} and chromatin diluted with NEB buffer 2 (+100 µg ml^{-1} BSA) to a concentration of about 1 µg $µl^{-1}$. Forty-five microliters of T4 PNK (New England Biolabs, M0201S) and five microliters NEBuffer 2 (+100 µg ml^{-1} BSA) were added to 45 µg of cross-linked or non-cross-linked chromatin, respectively. Samples were incubated for 15 min at 37 °C. For the biotin-replacement synthesis, the following reagents were added to 45 µg of cross-linked or 45 µg non-cross-linked chromatin at a concentration of 0.5 µg $µl^{-1}$, respectively: 3.1 µl of 10 mg ml^{-1} BSA (New England Biolabs), 21 µl of 10× NEB buffer 2 (New England Biolabs), 76.3 µl of 0.4 mM Biotin-dATP (Jena Biosciences), 76.3 µl of 0.4 mM Biotin-dCTP (Jena Biosciences), 3.1 µl of 10 mM dTTP/dGTP (New England Biolabs), and 30 µl of T4 Polymerase at a concentration of 3 U $µl^{-1}$ (New England Biolabs, M0203S). After incubation at 12 °C for 15 min, EDTA was added to a final concentration of 50 mM. Next, the chromatin was dialyzed overnight against 3 l of dialysis buffer (50 mM Tris-Cl (pH 7.5), 1 mM EDTA, 150 mM NaCl, 1× PIC) at 4 °C. GdmCl denaturation buffer (8 M guanidine hydrochloride (GdmCl), 13.33 mM TCEP, 133.33 mM Tris-Cl (pH 8)) was added in a ratio of 3:1 and samples were boiled for 10 min at 99 °C. After allowing samples to cool down to room temperature, chloroacetamide was added to 40 mM and incubated for 20 min. 1.4 mg of T1 streptavidin beads (Thermo Scientific) were equilibrated by washing once with 1 ml of 1× B&W buffer (5 mM Tris-Cl (pH 7.5), 0.5 mM EDTA, 1 M NaCl) and once with 1 ml of 1× GdmCl wash buffer (0.6 M GdmCl, 1 mM TCEP, 10 mM Tris-Cl (pH 8)). Chromatin samples were diluted tenfold with 25 mM Tris-Cl (pH 8) and added to the beads. After incubation for 90 min at room temperature, beads were washed by incubating the beads 15 min with 1 ml of GdmCl wash buffer (0.6 M GdmCl, 10 mM Tris-Cl (pH 8)) for three times. After two washes with 1 ml of BW2× buffer (10 mM Tris-Cl (pH 8.0), 1 mM EDTA, 0.1% Tritone-X100, 2 M NaCl), two washes with 1 ml of SDS wash buffer (25 mM Tris-Cl (pH 8.0), 1 mM EDTA, 1% SDS, 200 mM NaCl), and two washes with TBS buffer (150 mM NaCl, 50 mM Tris-Cl (pH 7.6)), beads were resuspended in 50 µl MNase/Benzonase digestion buffer (2 mM $MgCl_2$, 5 mM $CaCl_2$). DNA was digested by the addition of 1 µl of MNase (New England Biolabs), 1 µl of DNase I (New England Biolabs) and 3 µl of Benzonase (Merck Millipore) and incubation for 90 min at 37 °C. GdmCl was added to 0.6 M and Tris-Cl (pH 8.0) to 10 mM. Two microliters of trypsin (0.5 µg $µl^{-1}$) and Lys-C (0.5 µg $µl^{-1}$) were added and proteins digested overnight.

Enrichment of DNA-cross-linked peptides. Peptides were desalted on StageTips containing C18 material (3× C18 disks) (Empore)[60]. StageTips were equilibrated sequentially with 100 µl methanol, 100 µl buffer B3 (95% acetonitrile (ACN)/0.1% FA), 100 µl buffer B2 (80% ACN/0.1% FA), 100 µl buffer B1 (50% ACN/0.1% FA), and 100 µl buffer A (0.1% FA). Samples were loaded and washed twice with buffer A. Peptides were eluted twice with 50 µl buffer B1 and once with 50 µl buffer B2. Eluates were combined and dried on a centrifugal evaporator. Three hundred microliters of TiO_2 blocking buffer (60% ACN, 0.1% trifluoroacetic acid (TFA), 300 mM lactic acid) was added and samples resuspended at 25 °C and 2000 rpm for 5 min. Fifteen milligrams of TiO_2 beads (GL Sciences) were resuspended in 25 µl of buffer B2 and added to the sample. After 5 min incubation at 25 °C and 2000 rpm, beads were pelleted by centrifugation at 2000 × *g* for 1 min. Beads were washed once with TiO_2 blocking buffer (centrifugation at 2000 × *g*, 1 min) and three times with buffer B2 (centrifugation at 2000 × *g*, 1 min). Beads were resuspended with 100 µl of buffer B2 and loaded onto C8 StageTips (3× C8 disks, Empore). Beads remaining in the tube after the first transfer were resuspended once more with 100 µl of buffer B2 and loaded onto the same C8 StageTip. Peptide–nucleotide cross-links were eluted twice with 40 µl TiO_2 elution buffer (60% ACN/5% NH_4OH) and samples dried on a centrifugal evaporator. Samples were dissolved in 5 µl buffer A* (2% ACN/0.1% TFA) for MS analysis.

LC–MS/MS analysis. Online chromatography was performed with a Thermo EASY-nLC 1200 UHPLC system (Thermo Fisher Scientific, Bremen, Germany) coupled online to a Q Exactive HF-X mass spectrometer with a nano-electrospray ion source (Thermo Fisher Scientific). Analytical columns (50 cm long, 75 µm inner diameter) were packed in-house with ReproSil-Pur C18 AQ 1.9 µm reversed phase resin (Dr. Maisch GmbH, Ammerbuch, Germany) in buffer A (0.1% FA). During online analysis the analytical columns were placed in a column heater (Sonation GmbH, Biberach, Germany) regulated to a temperature of 60 °C. Peptide mixtures were loaded onto the analytical column in buffer A and separated with a linear gradient of 5–20% buffer B (80% ACN and 0.1% FA) for 50 min, and 20–30% buffer B for 10 min, at a flow rate of 250 nl min^{-1}. MS data were acquired with a Q Exactive HF-X instrument programmed with a data-dependent top 12 method in positive mode using Tune 2.9 and Xcalibur 4.1. The S-lens RF level was 40.0 and capillary temperature was 250 °C. Full scans were acquired at 60,000 resolution with a maximum ion injection time of 20 ms and an AGC target value of 3E6. Selected precursor ions were isolated in a window of 2.0 *m/z*, fragmented by HCD with normalized collision energies of 30 for in vitro complexes and 35 for samples derived from cell cross-linking), and measured at 15,000 resolution with maximum injection time of 60 ms and AGC target of 1E5 ions. Precursor ions with unassigned or single states were excluded from fragmentation selection and repeated sequencing minimized by a dynamic exclusion window of 20 s.

Data analysis. Raw MS data were analyzed using the RNP(xl) workflow from the OpenMS Nodes v2.0.3 package implemented in the Proteome Discoverer software (v. 2.1.1.21)[17,50]. Control and UV-irradiated files were aligned by the retention time and precursors present in both conditions removed[17]. For in vitro fliX-MS experiments, searches were performed with modified Uniprot databases for human (nucleosomes and TBP) or pig (NF1), in which isoforms of the recombinant proteins were removed. Ex vivo fliX-MS data were searched against the Uniprot database for mouse (mESCs) in combination with contaminant sequences from the MaxQuant software package[61]. FDR control was performed by searching against a target-decoy version of the respective database. For in vitro flix-MS data, oxidation of methionine, trioxidation of cysteine (cysteic acid), and carbamylation of lysine and N-termini were allowed as variable modifications. For ex vivo flix-MS data, oxidation of methionine was defined as variable modification and carbamidomethylation of cysteines as static modification. The maximum allowed number of missed cleavages was 2 in all cases. Precursor DNA modifications were searched against all possible combinations of up to four connected nucleotides and possible modifications of $-H_2O$, $-HPO_3$, $-H_3PO_4$, and $+HPO_3$. Precursor mass tolerance was set to 10 ppm and fragment mass tolerance to 20 ppm. Incremental masses of shifted ions were set in the following order: nucleotide, nucleotide $-H_3PO_4$, $-HPO_3$, $-H_2O$, nucleobase, nucleobase $-NH_3$, and nucleobase $-CO$ (only thymine).

Manual curation of spectra proposed by RNP(xl) was performed as follows: (i) Precursor ions were evaluated for the correct assignment of the charge state and monoisotopic peak. (ii) The corresponding MS2 spectra were evaluated for >40% amino acid coverage combining a, b, and y ions. (iii) If the mass shift on the precursor ion reflected more than one nucleotide, nucleotides were required to be observed as marker ions in the low mass range. (iv) High-intensity fragment ions, which did not represent the unmodified peptide sequence, needed to be explainable by the DNA cross-link. RNP(xl) automatically annotates a, b, and y ions and immonium ions shifted by a nucleotide or nucleobase. In addition, all spectra were further analyzed for shifted and nonshifted internal ions using ProteinProspector (v 5.24.0). Supplementary Data 2 lists the identified a, b, and y ions and mass-shifted ions with additional information for all spectra.

Analysis of crystal structures and validation of the cross-links in crystal structures was performed using PyMol (the PyMOL Molecular Graphics System, version 1.2r3pre, Schrödinger, LLC).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The MS proteomics data have been deposited to the ProteomeXchange Consortium[62] (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository[63] with the dataset identifier PXD014898. All other data are available from the corresponding authors on reasonable request.

Received: 30 August 2019; Accepted: 27 May 2020;
Published online: 15 June 2020

References

1. Lambert, S. A. et al. The human transcription factors. *Cell* **172**, 650–665 (2018).

2. Barrera, L. A. et al. Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science* **351**, 1450–1454 (2016).
3. Kohler, S. et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* **42**, D966–D974 (2014).
4. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251 (2013).
5. Weirauch, M. T. & Hughes, T. R. A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. in *A Handbook of Transcription Factors* (ed. Hughes, T. R.) 25–73 (Springer Netherlands, Dordrecht, 2011).
6. Jolma, A. et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384–388 (2015).
7. Jolma, A. et al. DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).
8. Tsunaka, Y., Kajimura, N., Tate, S. & Morikawa, K. Alteration of the nucleosomal DNA path in the crystal structure of a human nucleosome core particle. *Nucleic Acids Res.* **33**, 3424–3434 (2005).
9. Nikolov, D. B. et al. Crystal structure of a human TATA box-binding protein/TATA element complex. *Proc. Natl Acad. Sci. USA* **93**, 4862–4867 (1996).
10. Kalodimos, C. G., Boelens, R. & Kaptein, R. A residue-specific view of the association and dissociation pathway in protein–DNA recognition. *Nat. Struct. Biol.* **9**, 193–197 (2002).
11. Grob, P. et al. Electron microscopy visualization of DNA-protein complexes formed by Ku and DNA ligase IV. *DNA Repair* **11**, 74–81 (2012).
12. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
13. Leitner, A., Faini, M., Stengel, F. & Aebersold, R. Crosslinking and mass spectrometry: an integrated technology to understand the structure and function of molecular machines. *Trends Biochem. Sci.* **41**, 20–32 (2016).
14. Baltz, A. G. et al. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell* **46**, 674–690 (2012).
15. Castello, A. et al. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393–1406 (2012).
16. Mitchell, S. F., Jain, S., She, M. & Parker, R. Global analysis of yeast mRNPs. *Nat. Struct. Mol. Biol.* **20**, 127–133 (2013).
17. Kramer, K. et al. Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nat. Methods* **11**, 1064–1070 (2014).
18. Angelov, D. et al. Protein DNA crosslinking in reconstituted nucleohistone, nuclei and whole cells by picosecond Uv laser irradiation. *Nucleic Acids Res.* **16**, 4525–4538 (1988).
19. Walter, J. & Biggin, M. D. Measurement of in vivo DNA binding by sequence-specific transcription factors using UV cross-linking. *Methods* **11**, 215–224 (1997).
20. Russmann, C. et al. Crosslinking of progesterone receptor to DNA using tuneable nanosecond, picosecond and femtosecond UV laser pulses. *Nucleic Acids Res.* **25**, 2478–2484 (1997).
21. Nagaich, A. K. & Hager, G. L. UV laser cross-linking: a real-time assay to study dynamic protein/DNA interactions during chromatin remodeling. *Sci. STKE* **2004**, pl13 (2004).
22. Steube, A., Schenk, T., Tretyakov, A. & Saluz, H. P. High-intensity UV laser ChIP-seq for the study of protein-DNA interactions in living cells. *Nat. Commun.* **8**, 1303 (2017).
23. Moss, T., Dimitrov, S. I. & Houde, D. UV-laser crosslinking of proteins to DNA. *Methods* **11**, 225–234 (1997).
24. Nebbioso, A. et al. Time-resolved analysis of DNA-protein interactions in living cells by UV laser pulses. *Sci. Rep.* **7**, 11725 (2017).
25. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
26. Dekker, J., van Oosterhout, J. A. & van der Vliet, P. C. Two regions within the DNA binding domain of nuclear factor I interact with DNA and stimulate adenovirus DNA replication independently. *Mol. Cell. Biol.* **16**, 4073–4080 (1996).
27. Gilfillan, S., Stelzer, G., Piaia, E., Hofmann, M. G. & Meisterernst, M. Efficient binding of NC2.TATA-binding protein to DNA in the absence of TATA. *J. Biol. Chem.* **280**, 6222–6230 (2005).
28. Hockensmith, J. W., Kubasek, W. L., Vorachek, W. R. & von Hippel, P. H. Laser cross-linking of nucleic acids to proteins. Methodology and first applications to the phage T4 DNA replication system. *J. Biol. Chem.* **261**, 3512–3518 (1986).
29. Bilokapic, S., Strauss, M. & Halic, M. Histone octamer rearranges to adapt to DNA unwrapping. *Nat. Struct. Mol. Biol.* **25**, 101–108 (2018).
30. Bilokapic, S., Strauss, M. & Halic, M. Structural rearrangements of the histone octamer translocate DNA. *Nat. Commun.* **9**, 1330 (2018).
31. de Vries, E., van Driel, W., van den Heuvel, S. J. & van der Vliet, P. C. Contactpoint analysis of the HeLa nuclear factor I recognition site reveals symmetrical binding at one side of the DNA helix. *EMBO J.* **6**, 161–168 (1987).
32. Misawa, H. & Yamaguchi, M. Involvement of nuclear factor-1 (NF1) binding motif in the regucalcin gene expression of rat kidney cortex: the expression is suppressed by cisplatin administration. *Mol. Cell. Biochem.* **219**, 29–37 (2001).
33. Armentero, M. T., Horwitz, M. & Mermod, N. Targeting of DNA polymerase to the adenovirus origin of DNA replication by interaction with nuclear factor I. *Proc. Natl Acad. Sci. USA* **91**, 11537–11541 (1994).
34. Claiborne, A. et al. Protein-sulfenic acids: diverse roles for an unlikely player in enzyme catalysis and redox regulation. *Biochemistry* **38**, 15407–15416 (1999).
35. Williams, B. J., Barlow, C. K., Kmiec, K. L., Russell, W. K. & Russell, D. H. Negative ion fragmentation of cysteic acid containing peptides: cysteic acid as a fixed negative charge. *J. Am. Soc. Mass Spectrom.* **22**, 1622–1630 (2011).
36. Bayer, M. & Konig, S. Abundant cysteine side reactions in traditional buffers interfere with the analysis of posttranslational modifications and protein quantification—how to compromise. *Rapid Commun. Mass Spectrom.* **30**, 1823–1828 (2016).
37. Wang, Y., Vivekananda, S., Men, L. & Zhang, Q. Fragmentation of protonated ions of peptides containing cysteine, cysteine sulfinic acid, and cysteine sulfonic acid. *J. Am. Soc. Mass Spectrom.* **15**, 697–702 (2004).
38. Gregson, J. M. & McCloskey, J. A. Collision-induced dissociation of protonated guanine. *Int. J. Mass Spectrom.* **165**, 475–485 (1997).
39. Jensen, S. S., Ariza, X., Nielsen, P., Vilarrasa, J. & Kirpekar, F. Collision-induced dissociation of cytidine and its derivatives. *J. Mass Spectrom.* **42**, 49–57 (2007).
40. Qian, M. et al. Ammonia elimination from protonated nucleobases and related synthetic substrates. *J. Am. Soc. Mass Spectrom.* **18**, 2040–2057 (2007).
41. Tsai, F. T. F. & Sigler, P. B. Structural basis of preinitiation complex assembly on human Pol II promoters. *Embo J.* **19**, 25–36 (2000).
42. Bleichenbacher, M., Tan, S. & Richmond, T. J. Novel interactions between the components of human and yeast TFIIA/TBP/DNA complexes. *J. Mol. Biol.* **332**, 783–793 (2003).
43. Hsieh, T. H. S. et al. Mapping nucleosome resolution chromosome folding in yeast by micro-C. *Cell* **162**, 108–119 (2015).
44. Winter, D., Seidler, J., Hahn, B. & Lehmann, W. D. Structural and mechanistic Information on c(1) ion formation in collision-induced fragmentation of peptides. *J. Am. Soc. Mass Spectrom.* **21**, 1814–1820 (2010).
45. Malinauskas, M. et al. Ultrafast laser processing of materials: from science to industry. *Light Sci. Appl.* **5**, e16133 (2016).
46. Kim, T. I., Del Barrio, J. L. A., Wilkins, M., Cochener, B. & Ang, M. Refractive surgery. *Lancet* **393**, 2085–2098 (2019).
47. Choudhary, D., Mossa, A., Jadhav, M. & Cecconi, C. Bio-molecular applications of recent developments in optical tweezers. *Biomolecules* **9**, 23 (2019).
48. Russmann, C., Beigang, R. & Beato, M. High DNA-protein crosslinking yield with two-wavelength femtosecond laser irradiation. *Methods Mol. Biol.* **148**, 611–620 (2001).
49. Russmann, C., Stollhof, J., Weiss, C., Beigang, R. & Beato, M. Two wavelength femtosecond laser induced DNA-protein crosslinking. *Nucleic Acids Res.* **26**, 3967–3970 (1998).
50. Veit, J. et al. LFQProfiler and RNP(xl): open-source tools for label-free quantification and protein-RNA cross-linking integrated into proteome discoverer. *J. Proteome Res.* **15**, 3441–3448 (2016).
51. Volker-Albert, M. C., Schmidt, A., Forne, I. & Imhof, A. Analysis of histone modifications by mass spectrometry. *Curr. Protoc. Protein Sci.* **92**, e54 (2018).
52. Nagata, K., Guggenheimer, R. A. & Hurwitz, J. Specific binding of a cellular DNA replication protein to the origin of replication of adenovirus DNA. *Proc. Natl Acad. Sci. USA* **80**, 6177–6181 (1983).
53. Stefancsik, R. & Sarkar, S. Relationship between the DNA binding domains of SMAD and NFI/CTF transcription factors defines a new superfamily of genes. *DNA Seq.* **14**, 233–239 (2003).
54. Etheve, L., Martin, J. & Lavery, R. Protein-DNA interfaces: a molecular dynamics analysis of time-dependent recognition processes for three transcription factors. *Nucleic Acids Res.* **44**, 9990–10002 (2016).
55. Butter, F. et al. Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding. *PLoS Genet.* **8**, e1002982 (2012).
56. Luger, K., Rechsteiner, T. J. & Richmond, T. J. Expression and purification of recombinant histones and nucleosome reconstitution. in *Chromatin Protocols* (ed. Becker, P. B.) 1–16 (Humana Press, Totowa, NJ, 1999).
57. Dyer, P. N. et al. Reconstitution of nucleosome core particles from recombinant histones and DNA. *Methods Enzymol.* **375**, 23–44 (2004).
58. Bartke, T. et al. Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell* **143**, 470–484 (2010).
59. Di Croce, L. et al. Two-step synergism between the progesterone receptor and the DNA-binding domain of nuclear factor 1 on MMTV minichromosomes. *Mol. Cell* **4**, 45–54 (1999).

60. Sharma, K. et al. Analysis of protein-RNA interactions in CRISPR proteins and effector complexes by UV-induced cross-linking and mass spectrometry. *Methods* **89**, 138–148 (2015).
61. 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).
62. Vizcaino, J. A. et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32**, 223–226 (2014).
63. Vizcaino, J. A. et al. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **41**, D1063–D1069 (2013).
64. Yousef, M. S. & Matthews, B. W. Structural basis of Prospero-DNA interaction: implications for transcription regulation in developing cells. *Structure* **13**, 601–607 (2005).

Acknowledgements
We thank Henning Urlaub and Alexandra Stützer from the Max Planck Institute for Biophysical Chemistry for advice on data analysis. We thank Till Bartke and Benjamin Foster from the Institute of Functional Epigenetics (Helmholtz Zentrum München) for providing expression plasmids and assistance in generating recombinant nucleosomes. We thank our colleagues at the Department of Proteomics and Signal Transduction for help and fruitful discussions. We thank Alexander Strasser, Igor Paron, and Christian Deiml for excellent technical assistance.

Author contributions
M.W., C.R., and A.R. designed the study; A.R. performed all experiments and data analysis; R.A. and R.K. conducted laser irradiation; J.F.-M. prepared recombinant NF1 protein; A.R. and M.W. wrote the manuscript; M.W., C.R., M.M., S.N., and M.B. supervised the study.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-16837-x.

Correspondence and requests for materials should be addressed to C.R. or M.W.

Peer review information *Nature Communications* thanks Yamini Dalal, and other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

3.2 Article 2: Defining the RBPome of T helper cells to study higher order post-transcriptional gene regulation

Kai P. Hoefig*, **Alexander Reim***, Christian Gallus*, Elaine H. Wong, Gesine Behrens, Christine Conrad, Meng Xu, Taku Ito-Kureha, Kyra Defourny, Arie Geerlof, Josef Mautner, Stefanie M. Hauck, Dirk Baumjohann, Regina Feederle, Matthias Mann, Michael Wierer, Elke Glasmacher & Vigo Heissmeyer. **Defining the RBPome of T helper cells to study higher order post-transcriptional gene regulation.** *bioRxiv* 2020.08.20.259234; doi:https://doi.org/10.1101/2020.08.20.259234

* Equal contributions

In revision at Nature Communications.

In this project I collaborated with Kai Höfig in the group of Vigo Heissmeyer at the Helmholtz Center Munich. They are interested in RNA-binding proteins in the context of post-transcriptional gene regulation in immune cells. We defined a RNA-binding proteome (RBPome) in mouse and human primary T cells. To this end we performed RNA interactome capture (RNA-IC) and orthogonal organic phase separation (OOPS) experiments and defined a core RBPome. I was involved in the OOPS experiments and performed extensive data analysis to identify and characterize the proteins identified in RNA-IC and OOPS experiments. The RBPome contained 798 protein in mouse and 801 proteins in human T helper cells. We identified several novel unexpected RNA-binding proteins like the signaling proteins Vav1, Stat1 and Stat4. We further characterized several of these proteins in follow-up experiments for their regulatory function on mRNA. This proteomic dataset serves as an important resource for RNA-binding proteins in immune cells.

Defining the RBPome of T helper cells to study higher order post-transcriptional gene regulation

Kai P. Hoefig[1*], Alexander Reim[2*], Christian Gallus[3*], Elaine H. Wong[4], Gesine Behrens[1], Christine Conrad[4], Meng Xu[1], Taku Ito-Kureha[4], Kyra Defourny[4,10], Arie Geerlof[5], Josef Mautner[6], Stefanie M. Hauck[7], Dirk Baumjohann[4,11], Regina Feederle[8], Matthias Mann[2], Michael Wierer[2,12#], Elke Glasmacher[3,9#], Vigo Heissmeyer[1,4#]

1 Research Unit Molecular Immune Regulation, Helmholtz Center Munich, Munich, Germany

2 Department of Proteomics and Signal Transduction, Max-Planck-Institute of Biochemistry, Munich, Germany

3 Institute of Diabetes and Obesity, Helmholtz Center Munich, Munich, Germany

4 Institute for Immunology at the Biomedical Center, Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany

5 Institute of Structural Biology, Helmholtz Center Munich, Neuherberg, Germany

6 Research Unit Gene Vectors, Helmholtz Center Munich & Children's Hospital, TU Munich, Germany

7 Research Unit Protein Science, Helmholtz Center Munich, Munich, Germany

8 Monoclonal Antibody Core Facility and Research Group, Institute for Diabetes and Obesity, Helmholtz Center Munich, Neuherberg, Germany.

9 present address: Roche Pharma Research and Early Development, Large Molecule Research, Roche Innovation Center Munich, Penzberg, Germany

10 present address: Department of Biomolecular Health Sciences, Utrecht University, Utrecht, the Netherlands

11 present address: Medical Clinic III for Oncology, Immuno-Oncology and Rheumatology University Hospital Bonn, University of Bonn, Bonn, Germany

12 present address: Proteomics Research Infrastructure, University of Copenhagen, Copenhagen, Denmark

* These authors contributed equally

Corresponding Authors: vigo.heissmeyer@med.uni-muenchen.de

elke.glasmacher@roche.com

wierer@biochem.mpg.de

One sentence statement:

We provide an atlas of RNA-binding proteins in human and mouse T helper cells as a resource for studying higher order post-transcriptional gene regulation.

Abstract

Post-transcriptional gene regulation is complex, dynamic and ensures proper T cell function. The targeted transcripts can simultaneously respond to various factors as evident for *Icos*, an mRNA regulated by several RNA binding proteins (RBPs), including Roquin. However, fundamental information about the entire RBPome involved in post-transcriptional gene regulation in T cells is lacking. Here, we applied global RNA interactome capture (RNA-IC) and orthogonal organic phase separation (OOPS) to human and mouse primary T cells and identified the core T cell RBPome. This defined 798 mouse and 801 human proteins as RBPs, unexpectedly containing signaling proteins like Stat1, Stat4 and Vav1. Based on the vicinity to Roquin-1 in proximity labeling experiments, we selected ~50 RBPs for testing coregulation of Roquin targets. Induced expression of these candidate RBPs in wildtype and Roquin-deficient T cells unraveled several Roquin-independent contributions, but also revealed Celf1 as a new Roquin-1-dependent and target-specific coregulator of *Icos*.

T lymphocytes as central entities of the adaptive immune system must be able to make critical cell fate decisions fast. To exit quiescence, commit to proliferation and differentiation, exert effector functions or form memory they strongly depend on programs of gene regulation. Accordingly, they employ extensive post-transcriptional regulation through RBPs and miRNAs or 3' end oligo-uridylation and m6A RNA modifications. These RBPs, or RBPs that recognize the modifications, directly affect the expression of genes by controlling mRNA stability or translation efficiency [1, 2, 3, 4, 5]. The studies in T helper cells have focused on a small number of RNA-binding proteins, including HuR and TTP/Zfp36l1/Zfp36l2 [6, 7, 8, 9], Roquin-1/2 [10, 11] and Regnase-1/4 [12, 13] as well as some miRNAs like miR-17~92, miR-155, miR-181, miR-125 or miR-146a [14]. Moreover, first evidence for m6A RNA methylation in this cell type has been provided [4]. Underscoring the relevance for the immune system, loss-of-function of these factors has often been associated with profound alterations in T cell development or functions which caused immune-related diseases [14, 15, 16, 17]. Intriguingly, many key factors of the immune system have acquired long 3'-UTRs enabling their regulation by multiple, and often overlapping sets of post-transcriptional regulators [18]. Some RBPs recruit additional co-factors as for example Roquin binding with Nufip2 to RNA [19] and some of them have antagonistic RBPs like HuR and TTP [20] or Regnase-1 and Arid5a [21]. Such functional or physical interactions together with interdependent binding to the transcriptome create enormous regulatory potential. The challenge is therefore to integrate our current knowledge about individual RBPs into concepts of higher order gene regulation that reflect the interplay of different, and ideally of all cellular RBPs.

A prerequisite for studying higher order post-transcriptional networks is to know cell type specific RBPomes to account for differential expression and RBP plasticity. To this end several global methods have been developed over the last decade, revealing a growing number of RBPs that may even exceed recent estimates of ~7.5% of the human proteome [22]. RNA interactome capture (RNA-IC) [23] is one widely used, unbiased technique, however, it is constricted by design, exclusively identifying proteins binding to polyadenylated RNAs. In

contrast, orthogonal organic phase separation (OOPS) analyzes all UV-crosslinked protein-RNA adducts from interphases after organic phase separation.

The interactions of RBPs with RNA typically involve charge-, sequence- or structure-dependent interactions and to date over 600 structurally different RNA-binding domains (RBD) have been identified in canonical RBPs of the human proteome [22]. However, global methods also identified hundreds of non-canonical RBPs, which oftentimes contained intrinsically disordered regions (IDRs). Surprisingly, as many as 71 human proteins with well-defined metabolic functions were found to interact with RNA [24] introducing the concept of "moonlighting". Depending on availability from their "day job" in metabolism such proteins also bear the potential to impact on RNA regulation. Recent large-scale approaches have increased the number of EuRBPDB-listed human RBPs to currently 2949 [25] , suggesting that numerous RNA/RBPs interactions and cell-type specific gene regulations have gone unnoticed so far.

As a first step towards a global understanding of post-transcriptional gene regulation we experimentally defined all proteins that can be crosslinked to RNA in T helper cells. RNA-IC and OOPS identified 310 or 1200 proteins in primary $CD4^+$ T cells interacting with polyadenylated transcripts or all RNA species, respectively. Importantly, this dataset now enables the study of higher order gene regulation by determining for example how the cellular RBPs participate or intervene with post-transcriptional control of targets by individual RBPs.

Results

Icos exhibits simultaneous and temporal regulation through several RBPs

A prominent example for complex post-transcriptional gene regulation is the mRNA encoding inducible T-cell costimulator (Icos). It harbors a long 3´-UTR, which responds in a redundant manner to Roquin-1 and Roquin-2 proteins, is repressed by Regnase-1 and microRNAs, but also contains TTP binding sites [7, 10, 11, 13, 26, 27, 28]. Moreover, the Icos 3´-UTR was proposed to be modified by m6A methylation [29], which could either attract m6A-specific RBPs with YTH

domains [30], recruit or repel other RBPs [31], or interfere with base-pairing and secondary structure- or miRNA/mRNA-duplex formation [32]. Because of transcriptional and post-transcriptional regulation, Icos expression exhibits a hundred-fold upregulation on the protein level during T cell activation, which then declines after removal of the TCR stimulus (**Fig. 1**). To determine the temporal impact on Icos expression by Roquin, Regnase-1, m6A and miRNA regulation we analyzed inducible, CD4-specific inactivation of Roquin-1 together with Roquin-2 (*Rc3h1-2*), Regnase-1 (*Zc3h12a*) or Wtap, an essential component of the m6A methyltransferase complex [33], as well as Dgcr8, which is required for pre-miRNA biogenesis [34]. To this end, we performed tamoxifen gavage on mice expressing a Cre-ERT2 knockin allele in the CD4 locus [35] together with the floxed, Roquin paralogs encoding, *Rc3h1* and *Rc3h2* alleles (**Fig. 1a-c**), Regnase-1 encoding *Zc3h12a* alleles (**Fig. 1d-f**) as well as *Wtap*, (**Fig. 1g-i**) and *Dgcr8* alleles, (**Fig. 1j-l**). We isolated CD4$^+$ T cells from these mice and expanded them for five days. Confirming target deletion on the protein level (**Fig. 1c, f, i, l**) we determined a strong negative effect on Icos expression by Roquin and Regnase-1 on days 2-5 (**Fig. 1a-b** and **d-e**), a moderate positive effect of Wtap on days 2-5 (**Fig. 1g-h**), and only a small effect of Dgcr8, with an initial tendency of negative (day 1) and later on positive effects (days 4-5) (**Fig. 1j-k**). We next asked whether T cell activation affects the expression levels of known regulators of Icos, as well as other RBPs, to install temporal compartmentalization. To do so, we monitored the expression of a panel of RBPs in mouse CD4$^+$ T cells over the same time course (**Fig. 1m-q**). Indeed, we observed fast upregulation of RBPs as determined with pan-Roquin, Nufip2 [19], Fmrp, Fxr1, Fxr2, pan-TTP/Zfp36 [7], pan-Ythdf (**Supplementary Fig. 1a**), or Celf1 specific antibodies, but also slower accumulation as determined with Regnase-1 or Rbms1 specific antibodies (**Fig. 1m, o**). There was also downregulation of RBPs as shown with pan-AGO [36] and Cpeb4 specific antibodies (**Fig. 1m-n and p**). Of note, we also observed signs of post-translational regulation showing incomplete or full cleavage of Roquin or Regnase-1 proteins [10, 13], respectively (**Fig. 1m**), and the induction of a slower migrating band for Celf1, likely phosphorylation [37] (**Fig. 1q**). Factors with the potential to cooperate, as involved for Roquin and Nufip2[19] or Roquin and

Regnase-1[10], showed overlapping temporal regulation (**Fig. 1m**), suggesting reinforcing effects on the Icos target. Together these data indicate that mRNA targets respond to simultaneous inputs from several RBPs, which are aligned by dynamic expression and post-translational regulation to orchestrate redundant, cooperative and antagonistic effects into a higher order regulation.

The polyadenylated RNA-bound proteome of mouse and human T helper cells

To decipher this post-transcriptional network we set out to determine the RNA-binding protein signature in primary mouse and human $CD4^+$ T cells. To identify mRNA-binding proteins we first performed mRNA capture experiments on $CD4^+$ T cells expanded under T_H0 culture conditions (**Fig. 2a**). Pull-down with oligo-dT beads enabled the enrichment of mRNA-bound proteins as determined by silver staining, which were increased in response to preceding UV irradiation of the cells (**Supplementary Fig. 1b**). Reverse transcription quantitative PCR (RT-qPCR) confirmed the recovery of specific mRNAs, such as for the housekeeping genes *Hprt* and *β-actin*. Both mRNAs were enriched at least 2-3 fold after UV crosslink, but there was no detection of non-polyadenylated 18S rRNA (**Supplementary Fig. 1c**). Focusing on protein recovery, we determined greatly enriched polypyrimidine tract-binding protein 1 (Ptbp1) RBP compared to the negative control β-tubulin in mRNA capture experiments using the EL-4 thymoma cell line (**Supplementary Fig. 1d**). Next we performed mass spectrometry (MS) on captured proteins from murine and human T cells (**Fig. 2b-c**). Quantifying proteins bound to mRNA in crosslinked (CL) versus non-crosslinked (nCL) samples we defined a total of 312 mouse (**Fig. 2b**) and 308 human mRNA-binding proteins (mRBPs) (**Fig. 2c**) with an overlap of ~70% (**Supplementary Table 1**), which is in concordance with the overlap of all listed RBPs for these two species in the eukaryotic RBP database (http://EuRBPDB.syshospital.org). Gene ontology (GO) analysis identified the term 'mRNA binding' as most significantly enriched (**Fig. 2d**). The top ten GO terms in mouse were also strongly enriched in the human dataset with comparable numbers of proteins assigned to the individual GO terms in both species (**Fig. 2d**). RBPs not only bind RNA by

classical RBDs, but RNA-interactions can also map to intrinsically disordered regions (IDRs) [38]. Furthermore, low complexity regions (LCRs) have also been reported to be overrepresented in RBPs [23]. Indeed, IDRs (**Fig. 2e**) and LCRs (**Supplementary Fig. 1e**) were strongly enriched protein characteristics of mouse and human RNA-IC-identified RBPomes. We next wondered how much variation exists in the composition of RBPs between different T helper cell subsets. Repeating RNA-IC experiments in mouse and human iTreg cells (**Supplementary Fig. 2a**), we find an overlap of 96% or 90% with the respective mouse or human iTreg with Teff RBPomes, suggesting that the same RBPs bind to the transcriptome in different T helper cell subsets (Table 1). Nevertheless, 47 or 48 proteins were exclusively identified in mouse or human effector T cells, respectively, and 10 or 28 proteins were only found in mouse or human Treg cells, respectively (**Supplementary Fig. 2b**). Although these differences could indicate the existence of subset-specific RBPs, they could also be related to an incomplete assessment of RBPs by the RNA-IC technology.

The global RNA-bound proteome of mouse and human T helper cells

We therefore employed a second RNA-centric method of orthogonal organic phase separation (OOPS) to extend our definition of the RBPome and cross-validate our results. Similar to interactome capture, OOPS preserves cellular protein/RNA interactions by UV crosslinking of intact cells. The physicochemical properties of the resulting aducts direct them towards the interphase in the organic and aqueous phase partitioning procedure (**Fig. 3a**). Following several cycles of interphase transfer and phase partitioning, RNase treatment releases RNA-bound proteins into the organic phase, making them amenable to mass spectrometry [39, 40]. Evaluating the method, we investigated RNA and proteins from purified interphases derived from CL and nCL MEF cell samples. RNAs with crosslinked proteins purified from interphases hardly migrated into agarose gels, but regained normal migration behavior after protease digest, as judged from the typical 18S and 28S rRNA pattern (**Supplementary Fig. 3a**). Conversely, known RBPs like Roquin-1 and Gapdh appeared after crosslinking in the interphase and could be recovered after RNase treatment from the

organic phase (**Supplementary Fig. 3b**). Utilizing the same cell numbers and culture conditions of T cells, this method identified in total 1255 and 1159 significantly enriched RBPs for mouse or human T cells, respectively, when comparing CL and nCL samples (**Supplementary Table 1**). The overlap between both organisms was 55% (**Fig. 3b**) and 60% (**Fig. 3c**) in relation to the individual mouse and human RBPomes. Although glycosylated proteins are known to also accumulate in the interphase [39], we experimentally verified that they did not migrate into the organic phase after RNase treatment (**Supplementary Fig. 3c-d**). Analyzing OOPS-derived RBPomes for gene ontology enrichment using the same approach as for RNA-IC the GO term 'mRNA binding' was again most significantly enriched in mouse and man (**Fig. 3d**). The top 10 GO terms were RNA related and six of them overlapped with those identified for RNA-IC-derived RBPomes. High similarity between mouse and human RBPomes becomes apparent by the similarity in all GO categories, including 'molecular function' (**Fig. 3d**), 'biological process' and 'cellular component' (**Supplementary Fig. 4**). Although our OOPS approach exceeded by far the quantity of RNA-IC identified RBPs, the number of ~1200 RBPs well matched published RBPomes of HEK293 (1410 RBPs), U2OS (1267 RBPs) and MCF10A (1165 RBPs) cell lines [39].

Defining the core T helper cell RBPome

To define a T helper cell RBPome we first made sure that RNA-IC and OOPS did not preferentially identify high abundance proteins (**Fig. 4a-b**). In comparison to single shot total proteomes OOPS-identified RBPs spanned the whole range of protein expression without apparent bias. In general, this was also true for RNA-IC, with a slight tendency to more abundantly expressed proteins. This however might be a true effect since messenger RNA binding RBPs have been reported to be higher in expression even in comparison to other RBPs [22]. We used the recently established comprehensive eukaryotic RBP database as reference to compare OOPS and RNA-IC-identified canonical and non-canonical RBPs from mouse and human $CD4^+$ T cells. The numbers of proteins in the mouse T helper cell

RBPomes created by RNA-IC and OOPS ranging from 312 to 1255 made up 10% to 40% of all listed EuRBPDB proteins, respectively (**Fig. 4c**). OOPS-identified T cell RBPs outnumbered those from RNA-IC experiments by a factor of four, which was predominantly due to the eight times higher number of non-canonical RBPs. Interestingly though, there were also twice as many canonical RBPs significantly enriched by OOPS (**Fig. 4c**). Analyzing the ten most abundantly annotated mouse RBDs (comprising 26 to 224 members) showed that RNA-IC and OOPS often identified the same canonical RBPs (**Supplementary Fig. 5**), however at least equal, higher or much higher numbers were detected in OOPS samples depending on the specific RBD (**Fig. 4d**). These findings suggested that the OOPS method recovered RBPs from additional, non-polyadenylated RNAs and that RNA-IC-derived RBPomes are specific but incomplete. These conclusions are also supported by highly similar results obtained for the human $CD4^+$ T cell RBPome (**Fig. 4e-f**). In a 4-way comparison of mouse and human RBPs identified by OOPS and RNA-IC (**Fig. 4g**) we conservatively defined all proteins that were identified by at least two datasets as 'core $CD4^+$ T cell RBPomes' discovering 798 mouse and 801 human RBPs in this category (**Supplementary Table 1**). A sizable number of 519 mouse and 424 human proteins were exclusively enriched by the OOPS method. While we cannot rule out false-positives, more than 55% of the proteins of both subsets matched to EuRBPDB-listed annotations (not shown). These findings suggested that genuine RBPs are found even outside of the intersecting set of OOPs and RNA-IC identified proteins and that the definition of RBPomes profits from employing different biochemical approaches.

T cell signaling proteins with unexpected RNA-binding function

Some of the identified RBPs of the core proteome including Stat1, Stat4 and Crip1 are not expected to be associated with mRNA in cells. We therefore established an assay to confirm RNA-binding of these candidates. To do so, GFP-tagged candidate proteins were overexpressed in HEK293T cells, which were UV crosslinked, and extracts were used for immunoprecipitations with GFP specific antibodies. By Western or North-Western blot

analyses with oligo(dT) probes we could verify pull-down of GFP-tagged proteins and the association with mRNA for the RBPs Roquin-1 and Rbms1, but also for the lactate dehydrogenase (Ldha) protein, a metabolic enzyme with known ability to also bind RNA [23] (**Fig. 5a**). Via this approach, the determined RNA association of Stat1, Stat4 and Crip1 by global methods was indeed confirmed (**Fig. 5b**), but appeared less pronounced as compared to prototypic RPBs, and similar with regard to Ldha (**Fig 5a**). This finding supports a potential moonlighting function of these signaling proteins. As the identity of the associated RNA species for these RBPs is unknown, we established a dual luciferase assay to determine the impact of the different proteins on the expression of the renilla luciferase transcript when they were tethered to an artificial 3'-UTR (**Fig. 5c**). We utilized the λN/5xboxB system [41] and confirmed the expression of fusion proteins with a newly established λN-specific antibody (**Fig. 5d-e**). Importantly, Stat1 and Stat4 repressed luciferase function almost to the same extend as the known negative regulators Pat1b and Roquin-1, or other known RBPs, such as Celf1, Rbms1 and Cpeb4 (**Fig. 5f**). λN-Crip1 and λN-Vav1 expression did not exert a positive or negative influence, since their relative luciferase expression appeared unchanged compared to cells transfected to only express the λN polypeptide (**Fig. 5f**). These data suggest that the signaling proteins Stat1 and Stat4 that we defined here as part of the T helper cell RBPome not only have the capacity to bind mRNA but can also exert RNA regulatory functions.

Analyzing higher order post-transcriptional regulation

We next devised an experimental strategy to uncover RBPs with the potential to antagonize or cooperate with the Roquin-1 RBP in the repression of its target mRNAs by performing 'BioID' experiments (**Fig. 6a**). In this protein-centric, proximity-based labelling method we expressed a Roquin-1 BirA* fusion protein to identify proteins that reside within a short distance of approximately ~10 nm [42] in T cells (**Fig. 6a**). In this dataset we sought for matches with the T cell RBPome (**Fig. 4g**) to identify proteins that shared the features, 'RNA-binding' and 'Roquin-1 proximity'. We first verified that the mutated version of the biotin

ligase derived from *E. coli* (BirA*) which was N-terminally fused to Roquin-1 or GFP was able to biotinylate residues in Roquin or other cellular proteins (**Fig. 6b**) but did not interfere with the ability of Roquin-1 to downregulate Icos (**Fig. 6c**). Doxycycline-induced BirA*-Roquin-1 compared to BirA*-GFP expression in $CD4^+$ T cells significantly enriched biotin labelling of 64 proteins (**Supplementary Table 2**), including Roquin-1 (Rc3h1) itself or Roquin-2 (Rc3h2) (**Fig. 6d**) as well as previously identified Roquin-1 interactors and downstream effectors, such as Ddx6 and Edc4 [26], components of the Ccr4/Not complex [43, 44] and Nufip2 [19] (**Fig. 6d**). More than half of all proteins in proximity to Roquin-1 were also part of the defined RBPome (**Fig. 6e-f**). Increasing this BioID list with additional proteins that we determined in proximity to Roquin-1 when establishing and validating the BioID method in fibroblasts (**Supplementary Fig. 6a-d**), we arrived at 143 proteins (**Supplementary Table 2**) of which 96 (67%) were part of the RBPome (**Supplementary Fig. 6e** and **Supplementary Fig. 7a**). Cloning of 46 candidate genes in the context of N-terminal GFP fusions (**Supplementary Fig. 8a**) and generating retroviruses to overexpress candidate proteins we analyzed their effect on endogenous Roquin-1 targets (**Supplementary Fig. 8b**). $CD4^+$ T cells were used from mice with *Rc3h1*$^{fl/fl}$;*Rc3h2*$^{fl/fl}$;rtTA alleles in combination with (iDKO) or without the CD4Cre-ERT2 allele (WT) allowing induced inactivation of Roquin-1 and -2 by 4'-OH-tamoxifen treatment. Reflecting deletion, the Roquin-1 targets Icos, Ox40, Ctla4, $I\kappa B_{NS}$ and Regnase-1 became strongly derepressed in induced double-knockout (iDKO) T cells (**Supplementary Fig. 8c**). This elevated expression was corrected to wildtype levels in iDKO T cells that were retrovirally transduced and doxycycline-induced to express GFP-Roquin-1 (**Supplementary Fig. 8c**). The target expression in WT T cells was only moderately reduced through overexpression of GFP-Roquin-1 (**Supplementary Fig. 8c**). For the majority of the 46 candidate genes induced expression in WT and iDKO $CD4^+$ T cells did not alter expression of the five analyzed Roquin-1 targets, exemplified here by the results obtained for Vav1 (**Fig. 7a-b**). Interestingly, we identified a new function for Rbms1 (transcript variant 2), specifically upregulating Ctla-4 (**Fig. 7c-d**). Furthermore, we demonstrated that Cpeb4 strongly upregulates Ox40 and, most strikingly, in the same cells

Cpeb4 repressed Ctla-4 levels (**Fig. 7e-f**). While these findings are new and noteworthy, they occurred in a Roquin-1 independent manner. In contrast to these effects, we determined a higher order regulation of Icos by Celf1. Induced Celf1 expression in wildtype T cells clearly upregulated Icos, Ctla-4 and Ox40 expression and this function was obliterated in Roquin-1-deficient iDKO cells (**Fig. 7g-h**). Importantly, this effect could not be explained by Celf1-mediated repression of Roquin-1 on the protein or mRNA level (**Supplementary Fig. 8d-e**). Together these findings demonstrate responsiveness of Roquin targets to many other RBPs and a higher order, Roquin-dependent regulation of transcripts encoding for the costimulatory receptors Icos, Ox40 and Ctla-4 by Celf1.

Discussion

The work on post-transcriptional gene regulation in T helper cells has focused on some miRNAs and several RNA-binding proteins, and few reports described m6A RNA methylation in this cell type. Although arriving at a more or less detailed understanding of individual molecular relationships and regulatory circuits, this isolated knowledge assembles into a very incomplete picture. Creating an atlas of the human and mouse T helper cell RBPomes has now opened the stage, allowing to work towards understanding connections, complexity and principles of post-transcriptional regulatory networks in these cells.

RNA-IC and OOPS are two complementary methods to define RBPs on a global scale. While RNA-IC specifically queries for proteins bound to polyadenylated RNAs, OOPS captures the RNA bound proteome in its whole. Applying both methods to T helper cells of two different organisms allowed us to cross-validate the results from both methods and solidified our description of the core mouse and human T helper cell RBPome. While the vast majority of RNA-IC-identified $CD4^+$ T cell RBPs were previously known RNA binders, OOPS typically repeated and profoundly expanded these results (**Supplementary Fig. 5**), with the possible caveat of identifying false-positive proteins. Contradicting this possibility at least in part, more than half of OOPS identified proteins that were exclusively found in mouse or man were EuRBPDB-listed.

Strikingly, the signaling proteins Stat1 and Stat4 were identified by mouse RNA-IC and human OOPS and were just below the cut-off in the mouse OOPS dataset, and we could support their RBP function by additional RNA-binding assays. Undoubtedly, the defined human and mouse T cell RBPomes contain many more unusual RBPs that would warrant further investigations. We assume that even interactions of proteins without prototypic RBDs, like the Stat proteins, with RNA will have consequences for both binding partners. As the identity of the interacting mRNA(s) is unknown, we could only speculate about the post-transcriptional impact. Nevertheless, Stat1 and Stat4 showed regulatory capacity in our tethering assays. Intriguingly, RNA binding may also impact the function of Stat proteins as transcription factors. Supporting this notion, early results found Stat1 bound to the non-coding, polyadenylated RNA 'TSU', derived from a trophoblast cDNA library, and translocation of Stat1 into the nucleus was reduced after TSU RNA microinjection into HeLa cells [45, 46] .

Many 3'-UTRs, which effectively instruct post-transcriptional control, exhibit little sequence conservation between species, and the exact modules which determine regulation are not known. This for example is true for the *Icos* mRNA [19, 26]. On the side of the *trans*-acting factors, we find high similarity between the RBPomes of T helper cells of mouse and human origin, actually reflecting the general overlap of so far determined RBPomes from many cell lines of these species. We interpret this evolutionary conservation as holding ready similar sets of RBPs in T helper cells across species, which are then able to work together on composite *cis*-elements to reach comparable regulation of 3'-UTRs in the different organisms, despite sequence variability and differences in the composition of elements.

We not only define the first RBPomes of human and mouse T helper cells. We also provide potential avenues of how to make use of this information. Screening a set of candidates from the T cell RBPome for effects on Roquin targets, our findings support a concept in which post-transcriptional targets are separated into "regulons" [47]. These regulons comprise coregulated mRNA subsets responding to the same inputs and often functioning in the same

biological process. Therefore, complex and differential binding of targets by RBPs of the cellular RBPome specifies the possible regulons and differential functions of the cell. Roquin cooperated, coregulated, or antagonized in the regulation of Icos with Regnase-1, m6A methylation and miRNA functions. Moreover, our screening approach added Rbms1 as Roquin-independent and Celf1 as Roquin-dependent coregulator in the Icos containing regulon. It also revealed that different targets responded very differently to the expression of specific coregulators, as for example Ctla4 and Ox40 were inhibited or induced by the same RBP, Cpeb4, respectively. Together these data indicated an unexpected wealth of possible inputs from the T helper cell RBPome and suggested highly variable and combinatorial mRNP compositions in higher order post-transcriptional gene regulation of individual targets.

To solve the seemingly simple question which RBPs regulate which mRNA in T helper cells, we will require further knowledge about individual contributions, binding sites and composite *cis*-elements, temporal and interdependent occupancies, interactions among RBPs and with downstream effector molecules. In this endeavor global protein and RNA centric approaches make fundamental contributions.

Acknowledgements

We thank Hemalatha Mutiah (Ludwig-Maximilians-Universität Munich) for screening hybridoma supernatants and also Claudia Keplinger (Helmholtz Center Munich) for excellent technical support. For the provision of mouse lines we would like to thank Marc Schmidt-Supprian (Rc3h1), Wolfgang Wurst (Rc3h2), Robert Blelloch (Dgcr8flox), Mingui Fu (Zc3h12aflox) and Thorsten Buch (Cd4-Cre-ERT2). The work was supported by the German Research Foundation grants SPP-1935 (to VH), SFB-1054 projects A03 and Z02 as well as HE3359/7-1 and HE3359/8-1 to V.H. as well as grants from the Wilhelm Sander, Fritz Thyssen, Else Kro□ner-Fresenius and Deusche Krebshilfe foundations to V.H.. D.B. was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)

under Emmy Noether Programme BA 5132/1-1 and BA 5132/1-2 (252623821), as well as SFB 1054 project B12 (210592381). D.B. is a member of Germany's Excellence Strategy EXC2151 577 (390873048).

Author contribution

EG and VH conceived the idea for the project and together with MW and MM supervised the experimental work. KPH and VH wrote the manuscript with contributions from EG, MW and AR. KPH performed OOPS, BioID and T cell transductions with help from GB, KD, SMH, JM and CC. CG conducted RNA-IC experiments and RNA-binding assays. AR and SMH performed mass spectrometry and AR analyzed RBPome data. SMH and KPH analyzed the BioID data. MX contributed the tethering assays and EW analyzed Icos regulation for which AG, RF, TI-K and DB provided unpublished reagents.

Data availability statement

The data sets generated and/or analyzed during the current study are available from the corresponding authors on reasonable request.

Methods

Isolation, *in vitro* cultivation and transduction of primary CD4$^+$ T cells

For *in vitro* cultivation of primary murine CD4$^+$ T cells, mice were sacrificed and spleen as well as cervical, axillary, brachial, inguinal and mesenteric lymph nodes were dissected and pooled. Organs were squished and passed through a 100 µm filter under rinsing with T cell isolation buffer (PBS supplemented with 2% FCS and 1 mM EDTA). Erythrocytes were eliminated by incubating cells with TAC-lysis buffer (13 mM Tris, 140 mM NH_4Cl, pH 7.2) for 5 min at room temperature. CD4$^+$ T cells were isolated by negative selection using EasySep™ Mouse CD4$^+$ T cell isolation Kit (STEMCELL) according to manufacturer's

protocol. Purified $CD4^+$ T cells were cultured in DMEM (Invitrogen) T cell culture medium supplemented with 10% FCS (Gibco), 1% Pen-Strep (Thermo Fisher), 1% HEPES-Buffer (Invitrogen), 1% non-essential amino acids (NEAA; Invitrogen) and 1 mM β-mercaptoethanol (Invitrogen). For activation and differentiation under T_H1 conditions the T cells were stimulated with α-CD3 (0.5 µg/mL; cl. 2C11, in house production), α-CD28 (2.5 µg/mL; cl. 37.5N, in house production), 10 µg/mL α-IL-4 (cl. 11B11, in house production) and 10 ng/mL IL-12 (BD Pharmingen) and cultured on goat α-hamster IgG (MP Biochemicals) pre-coated 6-or 12-well culture plates for 40-48h at an initial cell density of 5 or 1.5×10^6 cells/mL, respectively. The cells were then resuspended and cultured in medium supplemented with 200 IE/mL recombinant hIL-2 (Novartis) in a 10% CO_2 incubator and expanded for 2-4 days, as indicated. Subsequently cells were fed with fresh IL-2-containing medium every 24h and cultured at a density of $0.5\text{-}1\times10^6$ cells/mL. For *in vitro* deletion of floxed alleles of $Rc3h1^{fl/fl}$;$Rc3h2^{fl/fl}$;CD4Cre-ERT2;rtTA3 (but with no effect on the Cre-deficient WT control $Rc3h1^{fl/fl}$;$Rc3h2^{fl/fl}$;rtTA3) $CD4^+$ T cells were treated with 1 µM 4´OH-tamoxifen (Sigma) for 24h prior to activation at a cell density of $0,5\text{-}1\times10^6$ cells/mL. We performed retroviral transduction 40h after the start of anti-CD3/CD28 activation, T cells were transduced with retroviral particles using spin-inoculation (1h, 18 °C, 850 g), and after 4-6h co-incubation of T cells and virus, virus particles were removed and T cells resuspended in T cell medium supplemented with IL-2 as described before. To induce expression of pRetro-Xtight-GFP constructs in rtTA expressing T cells, the transduced cells were cultured for 16h in the presence of doxycycline (1 µg/mL) prior to flow cytometry analysis of expression of targets in transduced GFP+ cells.

***In vivo* deletion of loxP-flanked alleles and *in vitro* culture of $CD4^+$ T cells**

For *in vitro* culture analysis, deletion of Roquin ($Rc3h1^{fl/fl}$;$Rc3h2^{fl/fl}$;CD4-Cre-ERT2) [48], Regnase-1 ($Zc3h12a^{fl/fl}$;CD4-Cre-ERT2) [49], Wtap ($Wtap^{fl/fl}$;CD4-Cre-ERT2) [33], and Dgcr8 ($Dgcr8^{fl/fl}$;CD4-Cre-ERT2) [34] encoding alleles in Cd4-cre-ERT2 mice was induced *in vivo* by oral transfer of 5 mg tamoxifen (Sigma) in corn oil. Two doses of tamoxifen each day were given on two consecutive days (total of 20 mg tamoxifen per mouse). Mice with the genotype

CD4-Cre-ERT2 (without floxed alleles) were used for wild-type controls. Mice were scarified three days after the last gavage and total $CD4^+$ T cells were isolated using the EasySep™ Mouse T Cell Isolation Kit (Stem Cell) and activated under T_H1 conditions as described above.

Flow cytometry

Following *in vivo* deletion and $CD4^+$ T cell isolation (above) cells were activated and cultured under T_H1 conditions. Cells were obtained daily for FACS analysis. The single cell suspensions were stained with fixable violet dead cell stain (Thermo Fisher) for 20 min at 4°C. For the detection of surface proteins, cells were stained with the appropriate antibodies in FACS buffer for 20 min at 4°C. After staining, cells were acquired on a FACS Canto II (3-laser). The data were further processed with the software FlowJo 10 software (BD Bioscience). The following antibodies were used: anti-CD4 (cl. GK1.5), anti-CD44 (cl. IM7), anti-CD62L (cl. MEL-14), anti-Icos (cl. C398.4A), anti-Ox40 (cl. OX-86), all from eBioscience, anti-CD25 (cl. PC61, Biolegend).

Effects of doxycycline-induced expression of 46 GFP-GOI fusion protein in $2x10^6$ wild-type and Roquin iDKO $CD4^+$ T cells were analyzed on day 6 after isolation (compare **Suppl. Fig. 8b**). First, proteins were treated with a fixable blue dead cell stain (Invitrogen) and after washing, stained in three panels to interrogate the surface expression of Icos and Ox40 (Icos-PE, clone 7E.17G9/Ox40-APC, clone OX-86; both eBioscience) and to intracellularly measure Ctla4, $I\kappa B_{NS}$ (Ctla4-PE, UC10-4B9; eBioscience/cl. 4C1 rat monoclonal; in house production) as well as Regnase-1 (cl. 15D11 rat monoclonal; in house production). For intracellular staining cells were fixed in 2% formaldehyde for 15 min at RT, permeabilized by washing in Saponin buffer and stained with the appropriate antibodies for 1h at 4 °C. After washing, anti-rat-AF647 antibody (cl. poly4054; Biolegend) was added for 30 min. After additional rounds of Saponin- and FACS buffer washing, acquisition was performed using a LSR Fortessa.

Isolation and differentiation of effector and regulatory T cells for RNA-IC

Naïve $CD4^+$ T cells were isolated by using Dyna- and Detachabeads (11445D and 12406D, Invitrogen) from spleens and mesenteric lymph nodes of 8-12 week old C57BL/6J mice. For iTreg culture, cells were additionally selected for $CD62L^+$ with anti-CD62L (clone: Mel14) coated beads. All cells were then activated with plate-bound anti-CD3 (using first anti-hamster, 55397, Novartis, then anti-CD3 in solution: clone: 2C11H: 0.1 µg/mL) and soluble anti-CD28 (clone: 37N: 1 µg/mL) and cultured in RPMI medium (supplemented with 10% (vol/vol) FCS, β-mercaptoethanol (0.05 mM, Gibco), penicillin-streptomycin (100 U/ml, Gibco), Sodium Pyruvate (1 mM, Lonza), Non-Essential Amino Acids (1x, Gibco), MEM Vitamin Solution (1x, Gibco), Glutamax (1x, Gibco) and HEPES pH 7.2 (10 mM, Gibco)). For iTreg cell differentiation we additionally added the following cytokines and blocking antibodies: rmIL-2 and rmTGF-β (both: R&D Systems, 5 ng/ml), anti-IL-4 (clone: 11B11, 10 µg/ml) and anti-IFNγ (clone: Xmg-121, 10 µg/ml). All antibodies were obtained in collaboration with and from Regina Feederle (Helmholtz Center Munich). After differentiation for 36-48 h cells were expanded for 2-3 days. iTreg cells were cultured in RPMI and 2000 units Proleukin S (02238131, MP Biomedicals) and T_{eff} cells with 200 U Proleukin S. We only used iTreg cells for experiments if samples achieved at least 80% Foxp3 positive cells (00552300, Foxp3 Staining Kit, BD Bioscience).

EL-4 T cells were cultured in the same medium as primary T cells. HEK293T cells were cultured in DMEM (supplemented with 10% (vol/vol) FCS, penicillin-streptomycin (100 U/ml, Gibco) and Hepes, pH 7.2 (10 mM, Gibco)).

RBPome capture

For RBPome capture for mass spectrometry, 20 x 10^6 T_{eff} or iTreg cells were either lysed directly (nonirradiated, control) in 1 ml lysis buffer from the µMACS mRNA Isolation Kit (130-075-201, Miltenyi) or suspended in 1 ml PBS, dispensed on a 10 cm dish and UV irradiated at 0.2 J/cm^2 at 254 nm for 1 min, washed with PBS, pelleted and subsequently lysed (UV

irradiated) and mRNA was isolated from both samples with the µMACS mRNA Isolation Kit according to the manufacturer's instructions. RNAs and crosslinked proteins were eluted with 70° C RNase-free H_2O. For RBPome capture for Western blot analysis, 400 x 10^6 EL-4 T cells were either lysed directly in 8 ml lysis buffer (nonirradiated, control) or suspended in 16 ml PBS, dispensed on sixteen 10 cm dishes and UV irradiated at 0.2 J/cm^2 at 254 nm for 1 min, washed with PBS, pelleted and subsequently lysed in 8 ml lysis buffer (UV irradiated) and then mRNA was isolated from both samples with the µMACS mRNA Isolation Kit using 500 µl oligo(dT) beads per sample. Each sample was split and run over two M columns (130-042-801, Miltenyi) and each column was eluted with two times 100 µl RNase-free H_2O. The eluate was concentrated in Amicon centrifugal filter units (UFC501008, Merck) to a final volume of ~25 µl. 8 µl Lämmli buffer (4x) with 10% (vol/vol) β-mercaptoethanol was added and the sample boiled for 5 min at 95° C. For protein analysis, samples were either flash-frozen in liquid nitrogen for MS analysis or Lämmli loading dye was added for subsequent analysis by Western blotting or silver staining.

OOPS

20-30 x 10^6 mouse $CD4^+$ T cells were isolated as described above and activated without bias. Cells were washed in PBS once and resuspended in 1 ml of cold PBS and transferred into one well of a six well dish. Floating on ice, the cells in the open 6 well plate were UV irradiated once at 0.4 J/cm^2 and twice at 0.2 J/cm^2 at 254 nm with shaking in-between sessions. After irradiation the 1ml of cell suspension was transferred into a FACS tube and the well was washed with another 1 ml of cold PBS which was also added to the FACS tube. After centrifugation the cell pellet was completely dissolved in 1 ml of Trizol (Sigmal). The remainder of the procedure was performed strictly according to [39] with the exception that we broke up and regenerated interphases in five successive rounds of phase partitioning, rather than three.

Culture preparation of human $CD4^+$ T cell blasts

Blood (120 ml) was collected by venepuncture from two times four donors for RNA-IC and OOPS. Peripheral blood mononuclear cells (PBMCs) were isolated by standard Ficoll gradient (Pancoll^{R}) centrifugation and CD4^{+} T cells were isolated from 1x10^{8} cells using CD4^{+} Microbeads (Miltenyi) to arrive at 2-3x10^{7} cells. These were resuspended at 2x10^{6}/ml in T cell medium ((AIM-V (Invitrogen), 10% heat-inactivated human serum, 2mM L-glutamine, 10 mM HEPES and 1,25 µg/ml Fungizone), supplemented with 500 ng/ml PHA (Murex) and 100 IU IL-2/ml (Novartis). Cells were dispensed in 24-well plates at 2 ml per well and on day 3 old medium was replaced by fresh T cell medium and cell were harvested and counted at day 4,5. For generating iTreg, naïve CD4^{+} T cells were isolated from PBMCs using Microbeads (Naive CD4+ T Cell Isolation Kit II, Miltenyi) and resuspended at 1x10^{6} cells/ml in T cell medium supplemented with 500 ng/ml PHA (Murex), 500 IU IL-2/ml (Novartis), 500 nM Rapamycin (Selleckchem), and 5 ng/ml TGF-ß1 (Miltenyi). Cells were cultured in 24-well plates at 2 ml per well together with 1x10^{6} irradiated (40 Gy) PBMCs derived from three donors. On day 3, old medium was replaced by fresh T cell medium including supplements and the cells harvested and counted at day 5.

SDS-PAGE, Western blotting and silver staining

All protein visualization procedures were performed according to standard protocols. For silver staining we used the SilverQuest kit (LC6070, Invitrogen). Antibodies used were anti-GFP, (1:10, clone: 3E5-111, in house), anti-Ptbp1, (1:1000, 8776, Cell Signaling), anti-β-tubulin, (1:1000, 86298, Cell Signaling). Proteins were visualized by staining with anti-rat (1:3000, 7077, Cell Signaling) or anti-mouse (1:3000, 7076, Cell Signaling) secondary antibodies conjugated to HRP.

Sample preparation for mass spectrometry

For RNA-capture, eluates were incubated with 10 µg/ml RNase A in 100 mM Tris, 50 mM NaCl, 1 mM EDTA at 37°C for 30 min. RNase-treated eluates were acetone precipitated and

resuspended in denaturation buffer (6 M urea, 2 M thiourea, 10 mM Hepes, pH 8), reduced with 1 mM DTT and alkylated with 5.5 mM IAA. Samples were diluted 1:5 with 62.5 mM Tris, pH 8.1 and proteins digested with 0.5 µg Lys-C and 0.5 µg Trypsin at room temperature overnight. The resulting peptides were desalted using stage-tips containing three layers of C18 material (Empore).

For OOPS experiments, 100 µl of lysis buffer (Preomics, iST kit) were added and samples incubated at 100°C for 10 min at 1,400 rpm. Samples were sonicated for 15 cycles (30s on/30s off) on a bioruptor (Diagenode). Protein concentration was determined using the BCA assay and about 30 µg of proteins were digested. To this end, trypsin and Lys-C were added in a 1:100 ratio, samples diluted with lysis buffer to contain at least 50 µl of volume and incubated overnight at 37°C. To 50 µl of sample, 250 ul Isopropanol/1% TFA were added and samples vortexed for 15s. Samples were transferred on SDB-RPS (Empore) stagetips (3 layers), washed twice with 100 µl Isopropanol/1% TFA and twice with 100 µl 0.2% TFA. Peptides were eluted with 80 µl of 2% ammonia/80% acetonitrile, evaporated on a centrifugal evaporator and resuspended with 10 µl of buffer A* (2% ACN, 0.1% TFA).

LC-MS/MS analysis

Peptides were separated on a reverse phase column (50 cm length, 75 µm inner diameter) packed in-house with ReproSil-Pur C18-AQ 1.9 µm resin (Dr. Maisch GmbH). Reverse-phase chromatography was performed with an EASY-nLC 1000 ultra-high pressure system, coupled to a Q-Exactive HF Mass Spectrometer (Thermo Scientific) for mouse RNA-capture experiments or a Q-Exactive HF-X Mass Spectrometer (Thermo Scientific) for human RNA-capture, OOPS experiments and single-shot proteomes. Peptides were loaded with buffer A (0.1% (v/v) formic acid) and eluted with a nonlinear 120-min (100-min gradient for human RNA-capture and OOPS experiments) gradient of 5–60% buffer B (0.1% (v/v) formic acid, 80% (v/v) acetonitrile) at a flow rate of 250 nl/min (300 nl/min for human RNA-capture and OOPS). After each gradient, the column was washed with 95% buffer B and re-equilibrated with buffer A. Column temperature was kept at 60° C by an in-house designed oven with a

Peltier element and operational parameters were monitored in real time by the SprayQc software. MS data were acquired using a data-dependent top 15 (top 12 for human RNA-capture and OOPS experiments) method in positive mode. Target value for the full scan MS spectra was 3×10^6 charges in the 300–1,650 m/z range with a maximum injection time of 20 ms and a resolution of 60,000. The precursor isolation windows was set to 1.4 m/z and capillary temperature was 250°C. Precursors were fragmented by higher-energy collisional dissociation (HCD) with a normalized collision energy (NCE) of 27. MS/MS scans were acquired at a resolution of 15,000 with an ion target value of 1×10^5, a maximum injection time of 120 ms (60 ms for human RNA-capture and OOPS experiments). Repeated sequencing of peptides was minimized by a dynamic exclusion time of 20 s (30 ms for human RNA-capture and OOPS).

Raw data processing

MS raw files were analyzed by the MaxQuant software[50] (version 1.5.1.6 for RNA-capture files and version 1.5.6.7 for OOPS files) and peak lists were searched against the mouse or human Uniprot FASTA database, respectively, and a common contaminants database (247 entries) by the Andromeda search engine[51]. Cysteine carbamidomethylation was set as fixed modification, methionine oxidation and N-terminal protein acetylation as variable modifications. False discovery rate was 1% for both proteins and peptides (minimum length of 7 amino acids). The maximum number of missed cleavages allowed was 2. Maximal allowed precursor mass deviation for peptide identification was 4.5 ppm after time-dependent mass calibration and maximal fragment mass deviation was 20 ppm. Relative quantification was performed using the MaxLFQ algorithm[52]. “Match between runs” was activated with a retention time alignment window of 20 min and a match time window of 0.5 min for RNA-capture experiments, while matching between runs was disabled for OOPS experiments. The minimum ratio count was set to 2 for label-free quantification.

Data analysis

Statistical analysis of MS data was performed using Perseus (version 1.6.0.28). Human RNA-capture, mouse RNA-capture, human OOPS and mouse OOPS data was processed separately. For all experiments, MaxQuant output tables were filtered to remove protein groups matching the reverse database, contaminants or proteins only observed with modified peptides. Next, protein groups were filtered to have at least two valid values in either the crosslinked or control triplicate. LFQ intensities were logarithmized (base 2) and missing values were imputed from a normal distribution with a downshift of 1.8 standard deviations and a width of 0.2 (0.25 for OOPS data). For RNA-capture experiments, a Student's T-test was performed to find proteins significantly enriched in the crosslinked sample over the non-crosslinked control (false-discovery rate (FDR) < 0.05). As many proteins were identified specifically in the crosslinked sample at intensities too low to find significant differences compared to the imputed values, we additionally considered proteins only identified in two or three replicates of the crosslinked sample, but never in the non-crosslinked control, as RNA-binding proteins. For OOPS experiments, proteins significantly enriched in a Student's T-tests of the organic phase after RNase-treatment over the same sample of the non-crosslinked control (FDR < 0.05) were considered RBPs.

GO term enrichment analysis was performed using the clusterProfiler package in R (version 4.1.0) as described in the original publication[53]. The mouse or human proteomes served as background for the respective enrichment analysis. Relative abundance of proteins in the single-shot proteome (**Fig. 4a**-**b**) was determined by calculating the logarithm (base 2) of the ratio of the LFQ intensity and the number of theoretical peptides. Relative abundance of proteins significant in the mouse OOPS and RNA-IC is shown in the single-shot proteome of mouse $CD4^+$ T cells measured with the OOPS samples. Relative abundance of proteins significant in the human OOPS and RNA-IC experiments is shown in the single-shot proteome of human $CD4^+$ T cells measured with the OOPS samples or RNA-IC samples, respectively. For the 4-way Venn comparison to find RBPs identified in more than one RNA-IC or OOPS experiment, human gene names were converted into their homologous mouse

counterparts. Protein group entries containing more than one isoform were expanded for the comparison and subsequently collapsed into one entry again for calculation of the Venn diagram. Multiple gene name entries for different unambiguously identified isoforms were collapsed into the major isoform. This affected six protein groups in the human RNA-IC dataset and three protein groups in the mouse RNA-IC dataset.

Intrinsically disordered regions were retrieved from the Disorder Atlas[54]. We referred to the LCR-eXXXplorer[55] to obtain low complexity regions in proteins.

Plasmid construction

To generate a vector that expresses N-terminally GFP-tagged proteins, we amplified the respective genes from cDNA of T_{eff} or T_{Foxp3+} cells, added *Hind*III and *Kpn*I restriction sites in front of the start codon and cloned them into the pCR™8/GW/TOPO® vector. We then used *Hind*III and *Kpn*I to insert a GFP sequence where we removed the bases for the stop codon. The respective sequences were subsequently transferred to the expression vector pMSCV via the gateway cloning technology. Only GFP-Roquin-1 (a kind gift of Vigo Heissmeyer) was expressed from the vector pDEST14. For oligonucleotide sequences see **Supplementary Table 3**.

Validation of RNA binding ability

HEK293T cells were transfected by calcium phosphate transfection with plasmids expressing the respective proteins with an N-terminal GFP-tag or GFP alone. After three days, cells were washed with PBS on plates, UV crosslinked (CL) as before or directly scraped from the plates (nCL). Cell lysates were generated by flash-freezing pellets in liquid nitrogen and incubation in NP-40 lysis buffer (150 mM NaCl, 1% NP-40, 50 mM Tris-HCl, pH 7.4, 5 mM EDTA, 1 mM DTT, 1 mM PMSF and protease inhibitor mixture (Complete, Roche)). After lysis, extracts were cleared by centrifugation at 17000 g for 15 min at 4° C. We then determined protein concentration via the BCA method and used 2-10 mg of protein

for the subsequent GFP immunoprecipitation, depending on transfection efficiency and expected RNA-binding capacity. We pre-coupled 200 µl Protein-G beads (10004D, Dynabeads Protein G, Invitrogen) with 20 µg antibody (anti-GFP, clone: 3E5-111, in house) in PBS (1 h, RT), washed beads in lysis buffer, added them to cell lysates and incubated with rotation for 4 h at 4 °C. Beads were then washed three times with IP wash buffer (50 mM Tris-HCl, pH 7.5) with decreasing salt (500 mM, 350 mM, 150 mM, 50 mM NaCl) and SDS (0.05%, 0.035%, 0.015%, 0.005%) concentrations. Proteins and crosslinked RNAs were eluted with 50 mM glycine, pH 2.2 at 70° C for 5 min. Lämmli buffer (4x) was added and samples were divided for mRNA and protein detection and separated via SDS gel electrophoresis (6% SDS gels for detection of mRNA samples and 9% gels to verify immunoprecipitation efficiency). For RNA detection we blotted onto Nitrocellulose membranes and for protein detection on PVDF membranes. After transfer, the Nitrocellulose membrane was prehybridized with Church buffer (0.36 M Na_2HPO_4, 0.14 M NaH_2PO_4, 1 mM EDTA, 7% SDS) for 30 min and then incubated for 4 h with Church buffer containing 40 nM 3´-and 5´-Biotin labeled oligo$(dT)_{20}$ probe to anneal to the poly-A tail of the bound mRNA. The membrane was washed twice with 1 x SSC, 0.5% SDS and twice with 0.5 x SSC, 0.5% SDS. Bound mRNA was detected with the Chemiluminescent Nucleic Acid Detection Kit Module (89880, Thermo Fisher) according to the manufacturer´s instructions.

Real-Time PCR

Total RNA of input was purified from lysates with Agencourt RNAClean XP Beads (A63987, Beckman Coulter) according to the manufacturer´s instructions and eluted in nuclease-free H_2O. cDNA was synthesized from total input RNA and oligo(dT)-isolated RNA with the QuantiTect Reverse Transcription Kit (205311, Qiagen). All qRT-PCRs were performed with the SYBR green method. For Primer sequences see **Supplementary Table 3**.

RNA isolation, reverse transcription and quantitative RT PCR as shown in Figure 8d was performed as published [56] using the universal probes systems (Roche). Primers for Rc3h1 (F: gagacagcaccttaccagca; R: gacaaagcgggacacacat; probe 22) and Hprt (F:

Tgatagatccattcctatgactgtaga; R: aagacattctttccagttaaagttgag; probe 95) were efficiency tested (both E=1.99).

Tethering assay

Hela cells were seeded in 24-well plates using 5×10^4 cells per well. Transfection was performed the following day using Lipofectamine2000 (Invitrogen) and 300 ng of total constructs. Each transfection consisted of 75 ng of luciferase reporter plasmid psiCHECK2 (Promega) or luciferase-5boxB plasmid psiCHECK2 -5boxB, 225 ng of pDEST12.2-λN fused constructs. After 24 h, cells were harvested for luciferase activity assays using a Dual-Luciferase Reporter Assay System (Promega). Renilla luciferase activity was normalized to Firefly luciferase activity in each well to control for variation in transfection efficiency. psiCHECK2 lacking boxB sites served as a negative control, and each transfection was analysed in triplicates.

BioID

The proximity-dependent biotin identification assay was performed according to Roux (Roux et al., 2012) with modifications. For each sample $2x10^7$ MEF cells were grown on ten 15-cm cell culture dishes for 24h before BirA*-Roquin-1 or BirA* expression was induced by doxycycline treatment. For T cells, transduction with the same BirA*-fusions cloned into the plasmid pRetroXtight was performed as described above and the same number of cells was used for the experiment. Six hours after addition of doxyclycline, biotin was added for 16h to arrive at an end concentration of 50 µM. Approximately $8x10^7$ cells per sample were trypsinized, washed twice with PBS and lysed in 5 ml lysis buffer (50 mM Tris-HCL, ph7,4; 500 mM NaCl, 0,2% SDS; 1x protease inhibitors (Roche), 20 mM DTT, 25 U/ml Benzonase) for 30 min at 4 °C using an end-over-end mixer. After adding 500 µl of 20% Triton X-100 the samples were sonicated for two sessions of 30 pulses at 30% duty cycle and output level 2, using a Branson Sonifier 450 device. Keep on ice for two minutes in between sessions.

Pipetting of 4,5 ml prechilled 50 mM Tris-HCL, pH 7,4 was followed by an additional round of sonication. During centrifugation at 16500 x g for 10 min at 4 °C 500 µl streptavidin beads (Invitrogen) or each sample was equilibrated in a 1:1 mixture of lysis buffer and 50 mM Tris-HCL pH 7,4. After overnight binding on a rotator at 4 °C Streptavidin beads were stringently washed using wash buffers 1, 2 and 3 (Roux et al, 2012) and prepared for mass spectrometry by three additional washes with buffer 4 (1 mM EDTA, 20 mM NaCl, 50 mM Tris-HCL pH 7,4). Proteins were eluted from streptavidin beads with 50 µl of biotin-saturated 1x sample buffer (50 mM Tris-HCL pH 6,8, 12% sucrose, 2% SDS, 20 mM DTT, 0,004% Bromphenol blue, 3 mM Biotin) by incubation for 7 min at 98 °C. For identification and quantification of proteins, samples were proteolysed by a modified filter aided sample preparation [57] and eluted peptides were analysed by LC-MSMS on a QExactive HF mass spectrometer (ThermoFisher Scientific) coupled directly to a Ultimate 3000 RSLC nano-HPLC (Dionex). Label-free quantification was based on peptide intensities from extracted ion chromatograms and performed with the Progenesis QI software (Nonlinear Dynamics, Waters). Raw files were imported and after alignment, filtering and normalization, all MSMS spectra were exported and searched against the Swissprot mouse database (16772 sequences, Release 2016_02) using the Mascot search engine with 10 ppm peptide mass tolerance and 0.02 Da fragment mass tolerance, one missed cleavage allowed, and carbamidomethylation set as fixed modification, methionine oxidation and asparagine or glutamine deamidation allowed as variable modifications. A Mascot-integrated decoy database search calculated an average false discovery of < 1% when searches were performed with a mascot percolator score cut-off of 13 and significance threshold of 0.05. Peptide assignments were re-imported into the Progenesis QI software. For quantification, only unique peptides of an identified protein were included, and the total cumulative normalized abundance was calculated by summing the abundances of all peptides allocated to the respective protein. A t-test implemented in the Progenesis QI software comparing the normalized abundances of the individual proteins between groups was calculated and corrected for multiple testing resulting in q values (FDR adjusted p values) given in

Supplemental Table 2. Only proteins identified by two or more peptides were included in the list of Roquin-1 proximal proteins.

Antibodies

To generate monoclonal antibodies against pan-Ythdf proteins or λN-peptide, Wistar rats were immunized with purified GST-tagged full-length mouse Ythdf3 protein or ovalbumin-coupled peptide against λN (MNARTRRRERRAEKQWKAAN) using standard procedures as described [58]. The hybridoma cells of Ythdf- or λN-reactive supernatants were cloned at least twice by limiting dilution. Experiments in this study were performed with anti-pan-Ythdf clone DF3 17F2 (rat IgG2a/**κ**) and anti-λN clone LAN 4F10 (rat IgG2b/**κ**).

References

1. Hoefig, K.P. & Heissmeyer, V. Degradation of oligouridylated histone mRNAs: see UUUUU and goodbye. *Wiley interdisciplinary reviews. RNA* **5**, 577-589 (2014).

2. Hoefig, K.P. & Heissmeyer, V. Posttranscriptional regulation of T helper cell fate decisions. *The Journal of cell biology* (2018).

3. Salerno, F., Turner, M. & Wolkers, M.C. Dynamic Post-Transcriptional Events Governing CD8(+) T Cell Homeostasis and Effector Function. *Trends in immunology* **41**, 240-254 (2020).

4. Shulman, Z. & Stern-Ginossar, N. The RNA modification N(6)-methyladenosine as a novel regulator of the immune system. *Nature immunology* **21**, 501-512 (2020).

5. Turner, M. & Hodson, D.J. An emerging role of RNA-binding proteins as multifunctional regulators of lymphocyte development and function. *Adv Immunol* **115**, 161-185 (2012).

6. Chen, J. *et al.* Posttranscriptional gene regulation of IL-17 by the RNA-binding protein HuR is required for initiation of experimental autoimmune encephalomyelitis. *Journal of immunology* **191**, 5441-5450 (2013).

7. Moore, M.J. *et al.* ZFP36 RNA-binding proteins restrain T cell activation and anti-viral immunity. *eLife* **7** (2018).

8. Stellato, C. *et al.* Coordinate regulation of GATA-3 and Th2 cytokine gene expression by the RNA-binding protein HuR. *Journal of immunology* **187**, 441-449 (2011).

9. Vogel, K.U., Bell, L.S., Galloway, A., Ahlfors, H. & Turner, M. The RNA-Binding Proteins Zfp36l1 and Zfp36l2 Enforce the Thymic beta-Selection Checkpoint by Limiting DNA Damage Response Signaling and Cell Cycle Progression. *Journal of immunology* **197**, 2673-2685 (2016).

10. Jeltsch, K.M. *et al.* Cleavage of roquin and regnase-1 by the paracaspase MALT1 releases their cooperatively repressed targets to promote T(H)17 differentiation. *Nature immunology* **15**, 1079-1089 (2014).

11. Vogel, K.U. *et al.* Roquin paralogs 1 and 2 redundantly repress the Icos and Ox40 costimulator mRNAs and control follicular helper T cell differentiation. *Immunity* **38**, 655-668 (2013).

12. Minagawa, K. *et al.* Posttranscriptional modulation of cytokine production in T cells for the regulation of excessive inflammation by TFL. *Journal of immunology* **192**, 1512-1524 (2014).

13. Uehata, T. *et al.* Malt1-induced cleavage of regnase-1 in CD4(+) helper T cells regulates immune activation. *Cell* **153**, 1036-1049 (2013).

14. Baumjohann, D. & Ansel, K.M. MicroRNA-mediated regulation of T helper cell differentiation and plasticity. *Nature reviews. Immunology* **13**, 666-678 (2013).

15. Akira, S. Regnase-1, a ribonuclease involved in the regulation of immune responses. *Cold Spring Harb Symp Quant Biol* **78**, 51-60 (2013).

16. Heissmeyer, V. & Vogel, K.U. Molecular control of Tfh-cell differentiation by Roquin family proteins. *Immunological reviews* **253**, 273-289 (2013).

17. Hodson, D.J., Screen, M. & Turner, M. RNA-binding proteins in hematopoiesis and hematological malignancy. *Blood* **133**, 2365-2373 (2019).

18. Turner, M., Galloway, A. & Vigorito, E. Noncoding RNA and its associated proteins as regulatory elements of the immune system. *Nature immunology* **15**, 484-491 (2014).

19. Rehage, N. *et al.* Binding of NUFIP2 to Roquin promotes recognition and regulation of ICOS mRNA. *Nature communications* **9**, 299 (2018).

20. Mukherjee, N. *et al.* Global target mRNA specification and regulation by the RNA-binding protein ZFP36. *Genome biology* **15**, R12 (2014).

21. Masuda, K. *et al.* Arid5a controls IL-6 mRNA stability, which contributes to elevation of IL-6 level in vivo. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 9409-9414 (2013).

22. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nature reviews. Genetics* **15**, 829-845 (2014).

23. Castello, A. *et al.* Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393-1406 (2012).

24. Hentze, M.W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of RNA-binding proteins. *Nat Rev Mol Cell Biol* **19**, 327-341 (2018).

25. Liao, J.Y. *et al.* EuRBPDB: a comprehensive resource for annotation, functional and oncological investigation of eukaryotic RNA binding proteins (RBPs). *Nucleic acids research* **48**, D307-D313 (2020).

26. Glasmacher, E. *et al.* Roquin binds inducible costimulator mRNA and effectors of mRNA decay to induce microRNA-independent post-transcriptional repression. *Nature immunology* **11**, 725-733 (2010).

27. Pratama, A. *et al.* MicroRNA-146a regulates ICOS-ICOSL signalling to limit accumulation of T follicular helper cells and germinal centres. *Nature communications* **6**, 6436 (2015).

28. Vinuesa, C.G. *et al.* A RING-type ubiquitin ligase family member required to repress follicular helper T cells and autoimmunity. *Nature* **435**, 452-458 (2005).

29. Zhu, Y. *et al.* The E3 ligase VHL promotes follicular helper T cell differentiation via glycolytic-epigenetic control. *The Journal of experimental medicine* **216**, 1664-1681 (2019).

30. Zaccara, S., Ries, R.J. & Jaffrey, S.R. Reading, writing and erasing mRNA methylation. *Nat Rev Mol Cell Biol* **20**, 608-624 (2019).

31. Edupuganti, R.R. *et al.* N(6)-methyladenosine (m(6)A) recruits and repels proteins to regulate mRNA homeostasis. *Nature structural & molecular biology* **24**, 870-878 (2017).

32. Liu, N. *et al.* N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. *Nature* **518**, 560-564 (2015).

33. Weichmann, F. *et al.* Validation strategies for antibodies targeting modified ribonucleotides. *RNA* (2020).

34. Rao, P.K. *et al.* Loss of cardiac microRNA-mediated regulation leads to dilated cardiomyopathy and heart failure. *Circ Res* **105**, 585-594 (2009).

35. Sledzinska, A. *et al.* TGF-beta signalling is required for CD4(+) T cell homeostasis but dispensable for regulatory T cell function. *PLoS Biol* **11**, e1001674 (2013).

36. Bronevetsky, Y. *et al.* T cell activation induces proteasomal degradation of Argonaute and rapid remodeling of the microRNA repertoire. *The Journal of experimental medicine* **210**, 417-432 (2013).

37. Bohjanen, P.R., Moua, M.L., Guo, L., Taye, A. & Vlasova-St Louis, I.A. Altered CELF1 binding to target transcripts in malignant T cells. *RNA* **21**, 1757-1769 (2015).

38. Castello, A. *et al.* Comprehensive Identification of RNA-Binding Domains in Human Cells. *Mol Cell* **63**, 696-710 (2016).

39. Queiroz, R.M.L. *et al.* Comprehensive identification of RNA-protein interactions in any organism using orthogonal organic phase separation (OOPS). *Nat Biotechnol* **37**, 169-178 (2019).

40. Villanueva, E. *et al.* Efficient recovery of the RNA-bound proteome and protein-bound transcriptome using phase separation (OOPS). *Nat Protoc* (2020).

41. Amaya Ramirez, C.C., Hubbe, P., Mandel, N. & Bethune, J. 4EHP-independent repression of endogenous mRNAs by the RNA-binding protein GIGYF2. *Nucleic acids research* **46**, 5792-5808 (2018).

42. Roux, K.J., Kim, D.I., Raida, M. & Burke, B. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *The Journal of cell biology* **196**, 801-810 (2012).

43. Leppek, K. *et al.* Roquin promotes constitutive mRNA decay via a conserved class of stem-loop recognition motifs. *Cell* **153**, 869-881 (2013).

44. Sgromo, A. *et al.* A CAF40-binding motif facilitates recruitment of the CCR4-NOT complex to mRNAs targeted by Drosophila Roquin. *Nature communications* **8**, 14307 (2017).

45. Peyman, J.A. Repression of major histocompatibility complex genes by a human trophoblast ribonucleic acid. *Biol Reprod* **60**, 23-31 (1999).

46. Peyman, J.A. Mammalian expression cloning of two human trophoblast suppressors of major histocompatibility complex genes. *Am J Reprod Immunol* **45**, 382-392 (2001).

47. Keene, J.D. RNA regulons: coordination of post-transcriptional events. *Nature reviews. Genetics* **8**, 533-543 (2007).

48. Tavernier, S.J. *et al.* A human immune dysregulation syndrome characterized by severe hyperinflammation with a homozygous nonsense Roquin-1 mutation. *Nature communications* **10**, 4779 (2019).

49. Li, Y. *et al.* Central role of myeloid MCPIP1 in protecting against LPS-induced inflammation and lung injury. *Signal Transduct Target Ther* **2**, 17066 (2017).

50. 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).

51. Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **10**, 1794-1805 (2011).

52. 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).

53. Yu, G., Wang, L.G., Han, Y. & He, Q.Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-287 (2012).

54. Vincent, M. & Schnell, S. Disorder Atlas: Web-based software for the proteome-based interpretation of intrinsic disorder predictions. *Comput Biol Chem* **83**, 107090 (2019).

55. Kirmitzoglou, I. & Promponas, V.J. LCR-eXXXplorer: a web platform to search, visualize and share data for low complexity regions in protein sequences. *Bioinformatics* **31**, 2208-2210 (2015).

56. Hoefig, K.P. *et al.* Eri1 degrades the stem-loop of oligouridylated histone mRNAs to induce replication-dependent decay. *Nature structural & molecular biology* **20**, 73-81 (2013).

57. Grosche, A. *et al.* The Proteome of Native Adult Muller Glial Cells From Murine Retina. *Mol Cell Proteomics* **15**, 462-480 (2016).

58. Feederle, R. *et al.* Generation of Pax1/PAX1-Specific Monoclonal Antibodies. *Monoclon Antib Immunodiagn Immunother* (2016).

59. Sysoev, V.O. *et al.* Global changes of the RNA-bound proteome during the maternal-to-zygotic transition in Drosophila. *Nat Commun* **7**, 12128 (2016).

Figure legends

Figure 1: Icos responds to inputs from several post-transcriptional regulators. a, d, g, j, Bar diagrams of Icos mean fluorescence intensities (MFI) over a five-day period of T cell

activation illustrating changes induced by the $CD4^+$-specific, induced depletion of the respective genes. Significance was calculated using the unpaired t-test (two-tailed) with n=3. **b, e, h, k,** Representative histograms of Icos expression at the specified days. **c, f, i, l,** Histograms demonstrating successful depletion of the respective target proteins. **o-r,** Western blots showing patterns of dynamic RBP regulation after CD3 and CD28 $CD4^+$ T cell activation. ns = not significant; * significant (p value = 0.01 – 0.05); ** very significant (p value = 0.001 – 0.01); *** very significant (p value = < 0.001).

Figure 2: The $CD4^+$ T cell RBPome of polyadenylated RNAs. **a,** Schematic illustration of the RNA-interactome capture (RNA-IC) method that was carried out to identify RBPs from mouse and human $CD4^+$ T cells. **b, c,** Volcano plot showing the -log10 p-value plotted against the log2 fold-change comparing the RNA-capture from crosslinked (CL) mouse CD4 T cells (b) or human $CD4^+$ T cells (c) versus the non-crosslinked (nCL) control. Red dots represent proteins significant at a 5% FDR cut-off level in both mouse and human RNA-capture experiments and blue dot proteins were significant only in mouse or human, respectively. **d,** Enrichment analysis of GO Molecular Function terms of significant proteins in mouse or human RNA capture data. The 10 most enriched terms in mouse (dark blue) and the respective terms in human (light blue) are shown. The y-axis represents the number of proteins matching the respective GO term. Numbers above each term depict the adjusted p-value after Benjamini-Hochberg multiple testing correction. **e,** Distribution of IDRs in all Uniprot reviewed protein sequences (black line), in proteins of the mouse EuRBPDB database (green line) and in proteins significant in the mouse RNA-IC experiment (red line). The same plot is shown for human data at the bottom. According to Kolmogorov-Smirnov testing the IDR distribution differences between RNA-IC (red lines) and all proteins (black lines) are highly significant in mouse and man and reach the smallest possible p-value ($p<2.2x10^{-16}$).

Figure 3: The global RNA-bound proteome of $CD4^+$ T cells. **a,** Schematic overview of the OOPS method [39] with phase partitioning cycles increased to five. **b-c,** Volcano plots showing

the -log10 p-value plotted against the log2 fold-change comparing the organic phase after RNase treatment of the interphase of OOPS experiments of crosslinked mouse $CD4^+$ T cells (b) or human $CD4^+$ T cells (c) versus the same non-crosslinked sample. Red dots represent proteins significant at a 5% FDR cutoff level in both mouse and human OOPS experiments and blue dots proteins significant only in mouse or human, respectively. **d,** Enrichment analysis of GO Molecular Function terms of significant proteins in mouse or human OOPS data. Enriched terms are depicted exactly as described for RNA-capture data in Figure 2d. **e,** Distribution of intrinsically disordered regions in all Uniprot reviewed protein sequences (black line), in proteins in the mouse EuRBPDB database (green) and in proteins significant in the mouse OOPS data (red line). The same plot is shown for human data at the bottom. According to Kolmogorov-Smirnov testing the IDR distribution differences between RNA-IC (red lines) and all proteins (black lines) are highly significant in mouse and man and reach the smallest possible p-value ($p<2.2x10^{-16}$).

Figure 4: Defining the mouse and human $CD4^+$ T cell RBPomes.

a, Relative abundance (Log2) of proteins identified in a single-shot mouse proteome (orange dots) plotted by their rank from highest to lowest abundant protein. RNA-binding proteins detected by RNA capture (top plots) or by OOPS (bottom plots) are highlighted as blue diamonds. **b,** Same plots as shown in (a) for human data. **c, e,** The recently established database EuRBPDB was used as a reference for eukaryotic RBPs to determine the numbers of canonical and non-canonical RBPs identified by RNA-IC or OOPS on mouse (c) or human (e) $CD4^+$ T cells. **d, f,** The occurrence of RBDs in RNA-IC and OOPS-identified RBPs was analyzed in comparison with the ten most abundant motifs described for the mouse (d) or human (f) proteome. **g,** Venn diagram using four datasets for RBPs in $CD4^+$ T cells as determined by RNA-IC and OOPS in mouse and human cells. *The core RBPomes contain proteins present in at least two datasets.

Figure 5: Stat1 and Stat4 are RBPs with regulatory potential. a, b, Semi-quantitative identification method for RBPs as in [59]. In short, GFP-fused proteins were transfected into

HEK293T cells, crosslinked or left untreated and subsequently immunoprecipitated using an anti-GFP antibody. The obtained samples were divided for protein and RNA detection by Western or Northern blotting, respectively. **c,** Schematic representation of the tethering assay that was used to investigate a possible influence of the genes of interest (GOI) on renilla luciferase expression. The affinity of the λN peptide targets the respective fusion protein to boxB stem-loop structures (5x) in the 3'-UTR of a Renilla luciferase gene where it can exert its function, if exciting. **d,** FACS plots using the new monoclonal antibody 4F10 demonstrate specific λN detection of a λN-GFP fusion protein expressed in 293T cells. **e,** Western blot showing the expression of the indicated λN-GOI proteins in 293T cells after transfection. **f,** Tethering assay results performed in HeLa cells as explained in (c). Two negative controls were implemented, using constructs without boxB sites or λN expression without fusion to a GOI. Each measurement was performed in triplicate and was independently repeated at least twice (n=2).

Figure 6: Identification of proteins in proximity to Roquin-1 in CD4$^+$ T cells. a, Schematic overview of the BioID method showing how addition of biotin to the medium leads to the activation of biotin, diffusion of biotinoyl-5'AMP and the biotinylation of the bait (Roquin-1) and all preys in the circumference. **b,** Equimolar amounts of protein were loaded onto a PAGE gel for Western blotting applying an anti-biotin antibody. Efficient biotinylation of both baits BirA*-Roquin-1 and BirA*-GFP (control) could be demonstrated. **c,** Histogram showing that transduction of CD4$^+$ T cells with retrovirus to inducibly express BirA*-Roquin-1 lead to the efficient downregulation of endogenous Icos. **d,** Identified preys from Roquin-1 BioIDs (n=5) in CD4$^+$ T cells. Depicted are all significantly enriched proteins with the exception of highly abundant ribosomal and histone proteins. Dot sizes equal p-values and positioning towards the center implies increased x-fold enrichment over BirA*-GFP BioID results. **e,** Venn diagram showing the overlap of RBPs from the CD4$^+$ T cell RBPome with the proteins identified by Roquin-1 BioID in T cells. **f,** Listed are all 38 proteins (59%) that

are Roquin-1 preys and RBPs. Colored fields indicate proteins that were chosen for further analysis.

Figure 7: Higher order Icos regulation by Roquin-1 and Celf1. Treatment with 4'-OH-tamoxifen of $CD4^+$ T cell with the genotypes $Rc3h1^{fl/fl}$;$Rc3h2^{fl/fl}$;rtTA3 without (WT) or with the CD4Cre-ERT2 allele (iDKO) were used for transduction with retroviruses. Expression levels of Icos and four additional Roquin-1 targets in WT and iDKO cells were analyzed 16h after individual, doxycycline-induced overexpression of 46 GFP-GOI fusion genes. **a, c, e, g,** The geometrical mean of each GOI-GFP divided by GFP for each Roquin-1 target was calculated and the summarized results are shown as bar diagrams for (a) Vav1, (b) Rbms1, (c) Cpeb4 and (d) Celf1. **b, d, f, h,** Representative, original FACS data are depicted as histograms or contour plots. Experiments for the negative Vav1 result were repeated twice, those for Rbms1, Cpeb4 and Celf1 at least three times.

Supplementary Fig. 1: RNA-IC supporting results.

a, Western blots demonstrating specificity of the newly established pan-Ythdf monoclonal antibody 17F2 for N-terminal GFP fusions of Ythdf1, Ythdf2 and Ythdf3, which are absent from non-transfected 293T cells. **b,** Silver staining analysis of oligo(dT) captured samples with and without UV irradiation. **c,** Quantitative RT-PCR to determine RNA pull-down efficiency with (CL) and without crosslink (nCL). Error bars show the standard deviation around the means of three independent experiments. **d,** Western blotting of UV irradiated and nonirradiated samples of EL-4 T cells. Membranes were probed with antibodies for the known mRNA-binding protein polypyrimidine tract binding protein 1 (Ptbp1) and β-tubulin. **e,** Distribution of low complexity regions in all Uniprot reviewed protein sequences (black line), in proteins in the EuRBPDB database (green) and in proteins significant in the RNA-IC data (red line). The left plot shows mouse data and the right plot human data. According to Kolmogorov-Smirnov testing the LCR distribution differences between RNA-IC (red lines) and all proteins (black lines) are highly significant in mouse ($p<2.2x10^{-16}$) and man ($p=7.8x10^{-16}$).

Supplementary Fig. 2: RNA-IC on mouse and human iTreg cells. a, Volcano plot showing the -log10 p-value in relation to the log2 fold-change comparing the RNA-capture from crosslinked mouse regulatory T cells (left plot) or human regulatory T cells (right plot) versus the non-crosslinked control. Red dots represent proteins significant at a 5% FDR cutoff level in both mouse and human RNA-capture experiments and blue dots proteins significant only in mouse or human, respectively. **b,** Venn diagram using four datasets to compare RNA-IC derived RBPomes of effector T and induced Treg cells from mouse and man.

Supplementary Fig. 3: OOPS supporting results. a, Agarose gel demonstrating disappearance of the typical 18S/28S rRNA bands after crosslinking and appearance of upshifted protein-RNA adducts (black arrows) in MEF cells with and without doxycycline-induce Roquin-1 expression. rRNA bands reappear after proteinase K (PK) treatment (grey arrows) and after RNase treatment the protein-RNA adducts in the wells disappear. **b,** Western blots showing that known RBPs, such as Roquin-1 and Gapdh can be detected in interphases, in the case of Roquin-1 only after induced expression. * cleavage product. **c, d,** Volcano plot showing the -log10 p-value plotted against the log2 fold-change comparing the interphase of OOPS experiments of crosslinked mouse $CD4^+$ T cells versus the organic phase after RNase treatment of the interphase. Glycoproteins are highlighted in red. Enrichment analysis of GO Molecular Function terms was performed for proteins with a log2 fold-change larger than 2. The 20 most enriched terms are depicted.

Supplementary Figure 4: Gene ontology enrichment analysis. Enrichment analysis of GO Biological Process and GO Cellular Component terms of significant proteins in mouse or human RNA-IC data (top row) or OOPS data (bottom row). The ten most enriched terms in mouse (dark blue) and the respective terms in human (light blue) are shown. The y-axis represents the number of proteins matching the respective GO term. Numbers above each term depict the adjusted p-value after multiple testing correction (Benjamini-Hochberg).

Supplementary Fig. 5: All currently annotated CCCH and KH domain containing RBPs and their detection in the RNA-IC and OOPS-identified RBPomes. All RBPs with the indicated RBDs are listed according to EuRBPDB. Colored boxes indicate significant enrichment by RNA-IC or OOPS in mouse or human $CD4^+$ T cells.

Supplementary Fig. 6: Identification of Roquin-1 preys by BioID in MEF cells. a, Western blots showing doxycycline-induced expression of Myc-BirA*-Roquin-1 or Myc-BirA* in MEF cell clones. **b,** FACS blot demonstrating that as in T cells (Fig. 6c) N-terminal fusion of Myc-BirA* to Roquin-1 does not affect its function and **c,** the fusion protein can biotinylate Roquin-1 and **d,** its cofactor Nufip2. **e,** Identified preys from Roquin-1 BioID in MEF cell clones. Depicted are 55 of 143 significantly enriched proteins (n=4) for better comparison with Fig. 6d. Yellow dot color indicates identification of the Roquin-1 prey in MEF and T cells. Dot sizes equal p-values and positioning towards the center implies increased x-fold enrichment.

Supplementary Fig. 7: Overlap between Roquin-1 preys in MEF cells and the $CD4^+$ T cell RBPome. a, Venn diagram showing that 96 proteins (67%) are Roquin-1 preys and also RBPs in T cells. **b,** Table listing these 96 proteins. Colors indicate which genes were clone for downstream experiments and if they were identified by MEF BioID only (red) or additionally in the T cell BioID.(blue).

Supplementary Fig. 8: Supporting results for higher order Icos regulation by Roquin-1 and Celf1. a, Microscopical images showing different localizations of the GFP signal as a result of GFP-GOI subcellular targeting in transfected 293T cells. **b,** Schematic representation of the experiment performed in Fig. 7. **c,** Treatment with 4'-OH-tamoxifen of $CD4^+$ T cell with the genotypes $Rc3h1^{fl/fl}$;$Rc3h2^{fl/fl}$;rtTA3 without (WT) or with the CD4Cre-ERT2 allele (iDKO) were used for transduction with a retrovirus expressing GFP-Roquin-1. Expression levels of the Roquin-1 targets Icos, Ox40, Ctla4 and Ikb_{NS} were analyzed and work as a positive control for the experiment. **d, e,** Cells taken from an experiment as performed (b). qPCR (d) and Western blot (e) performed on cDNA or protein lysates,

respectively, derived from $CD4^+$ T cells (WT) after doxycycline-induced expression of the indicated fusion proteins.

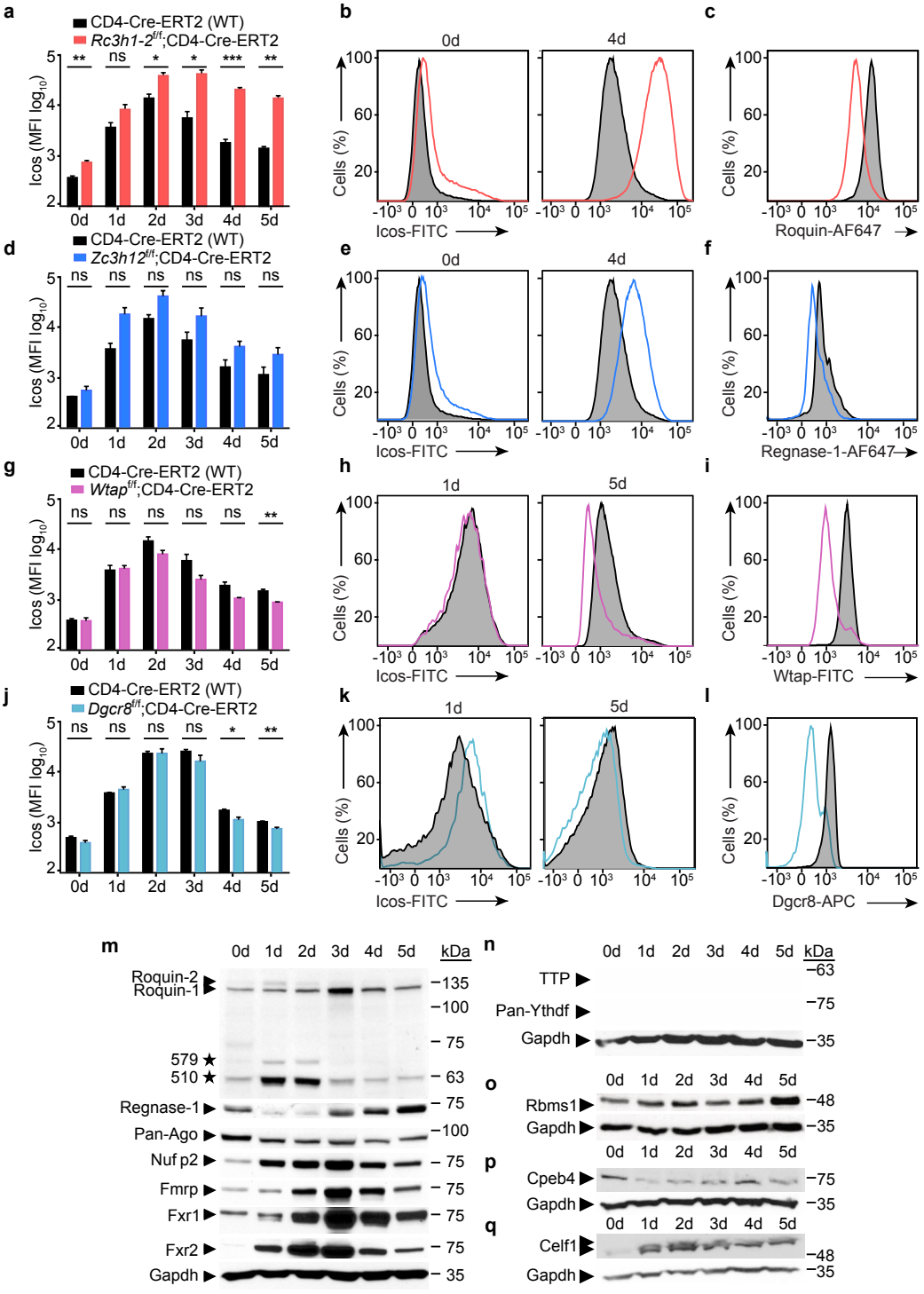

a
CD4-Cre-ERT2 (WT)
Rc3h1-2f/f;CD4-Cre-ERT2
Icos (MFI log10)
0d 1d 2d 3d 4d 5d
b
0d
4d
Cells (%)
Icos-FITC
c
Roquin-AF647
d
CD4-Cre-ERT2 (WT)
Zc3h12f/f;CD4-Cre-ERT2
e
f
Regnase-1-AF647
g
CD4-Cre-ERT2 (WT)
Wtapf/f;CD4-Cre-ERT2
h
1d
5d
i
Wtap-FITC
j
CD4-Cre-ERT2 (WT)
Dgcr8f/f;CD4-Cre-ERT2
k
l
Dgcr8-APC
m
kDa
Roquin-2
Roquin-1
579 ★
510 ★
Regnase-1
Pan-Ago
Nuf p2
Fmrp
Fxr1
Fxr2
Gapdh
n
TTP
Pan-Ythdf
o
Rbms1
p
Cpeb4
q
Celf1

Figure 1

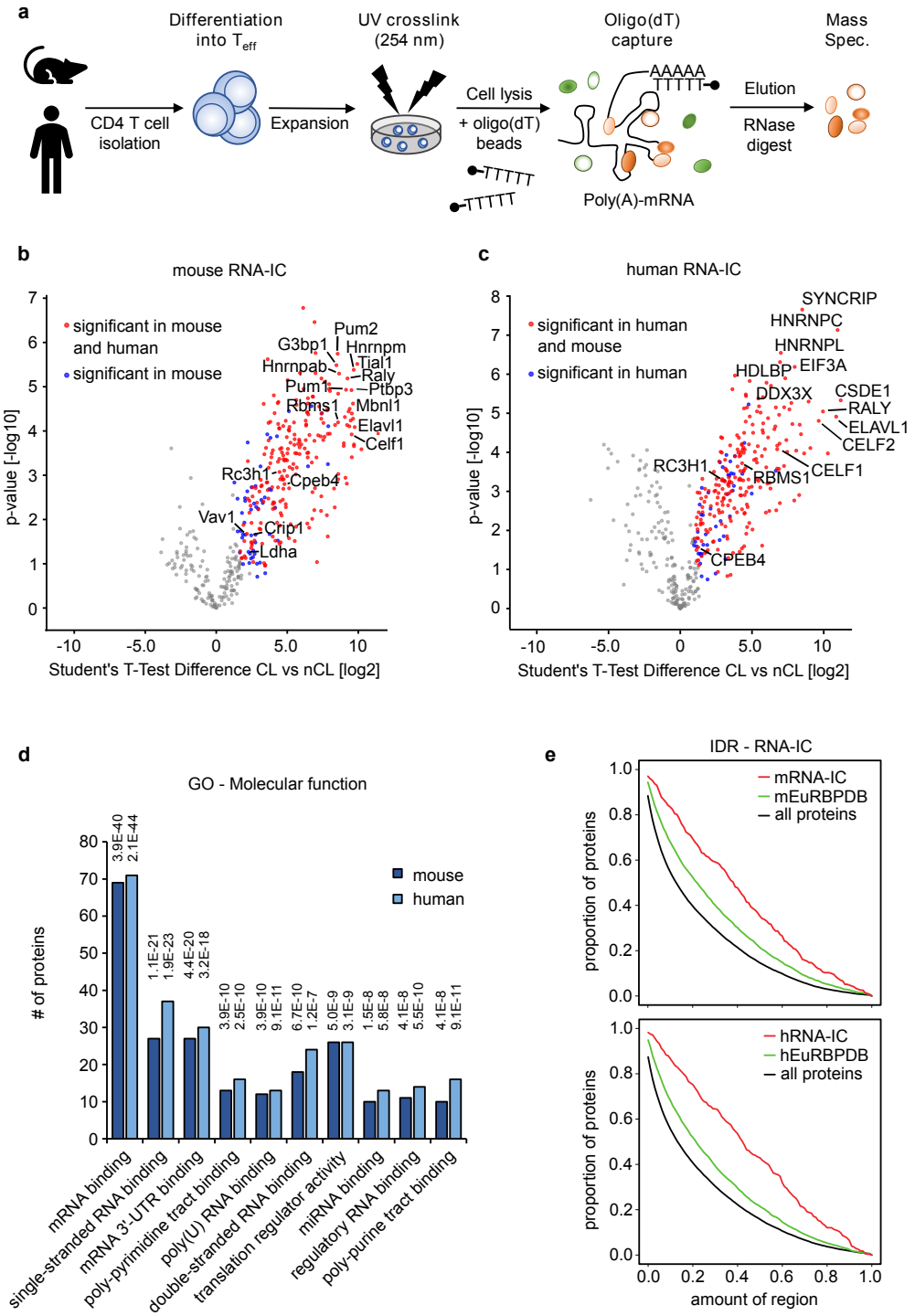

a
Differentiation into T_eff
UV crosslink (254 nm)
Oligo(dT) capture
Mass Spec.
CD4 T cell isolation
Expansion
Cell lysis + oligo(dT) beads
Elution
RNase digest
AAAAA
TTTTT
Poly(A)-mRNA
b
mouse RNA-IC
significant in mouse and human
significant in mouse
Pum2
G3bp1
Hnrnpm
Hnrnpab
Tial1
Raly
Pum1
Ptbp3
Rbms1
Mbnl1
Elavl1
Celf1
Rc3h1
Cpeb4
Vav1
Cnp1
Ldha
p-value [-log10]
Student's T-Test Difference CL vs nCL [log2]
c
human RNA-IC
significant in human and mouse
significant in human
SYNCRIP
HNRNPC
HNRNPL
HDLBP
EIF3A
DDX3X
CSDE1
RALY
ELAVL1
CELF2
RC3H1
RBMS1
CELF1
CPEB4
p-value [-log10]
Student's T-Test Difference CL vs nCL [log2]
d
GO - Molecular function
mouse
human
# of proteins
3.9E-40
2.1E-44
1.1E-21
1.9E-23
4.4E-20
3.2E-18
3.9E-10
2.5E-10
3.9E-10
9.1E-11
6.7E-10
1.2E-7
5.0E-9
3.1E-9
1.5E-8
5.8E-8
4.1E-8
5.5E-10
4.1E-8
9.1E-11
mRNA binding
single-stranded RNA binding
mRNA 3'-UTR binding
poly-pyrimidine tract binding
poly(U) RNA binding
double-stranded RNA binding
translation regulator activity
miRNA binding
regulatory RNA binding
poly-purine tract binding
e
IDR - RNA-IC
mRNA-IC
mEuRBPDB
all proteins
hRNA-IC
hEuRBPDB
all proteins
proportion of proteins
amount of region

Figure 2

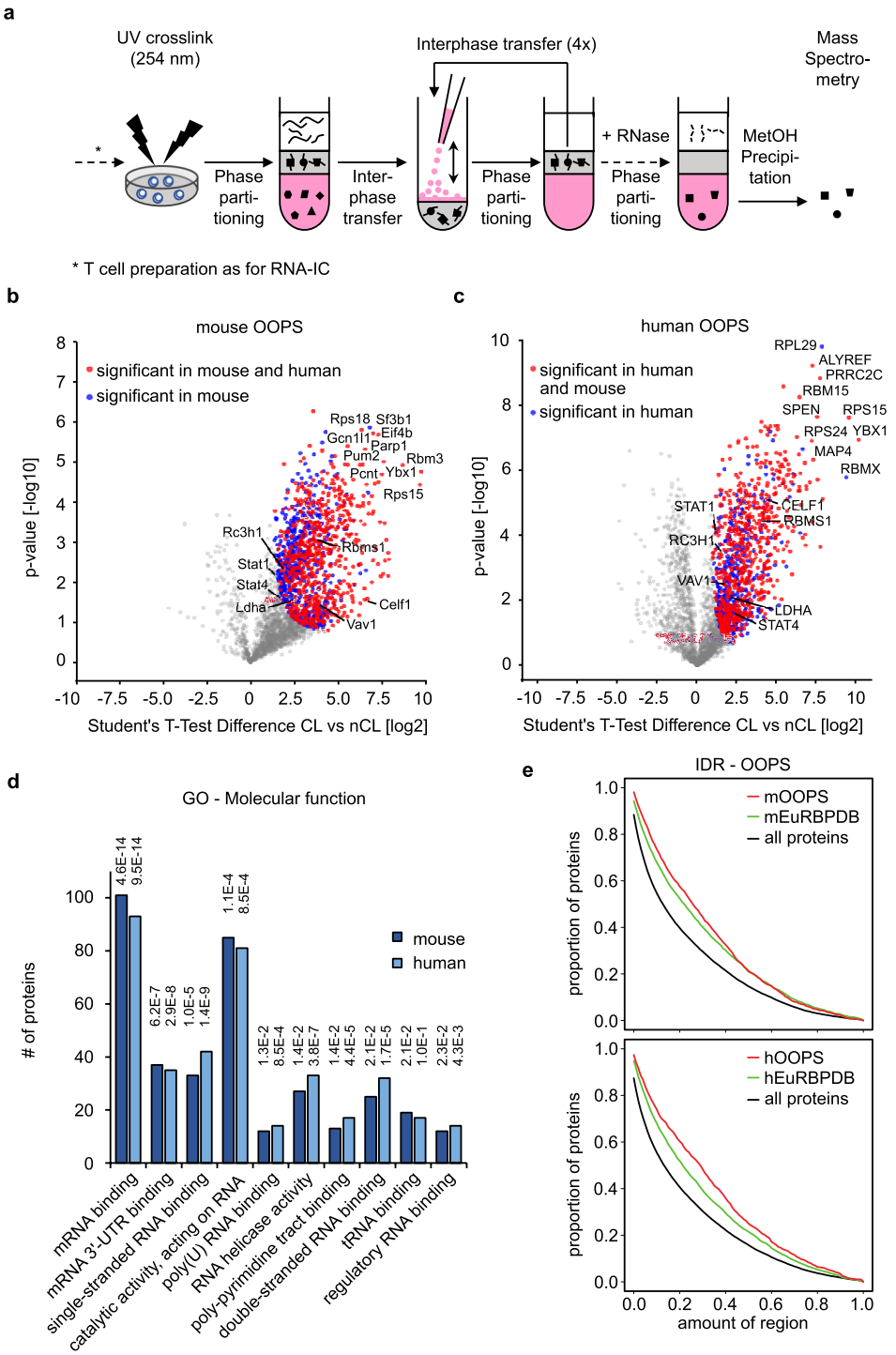

a
UV crosslink (254 nm)
Phase parti-tioning
Inter-phase transfer
Interphase transfer (4x)
Phase parti-tioning
+ RNase
Phase parti-tioning
MetOH Precipi-tation
Mass Spectro-metry
* T cell preparation as for RNA-IC
b
mouse OOPS
significant in mouse and human
significant in mouse
p-value [-log10]
Student's T-Test Difference CL vs nCL [log2]
Rps18
Sf3b1
Gcn1l1
Eif4b
Parp1
Pum2
Rbm3
Pcnt
Ybx1
Rps15
Rc3h1
Rbms1
Stat1
Stat4
Ldha
Celf1
Vav1
c
human OOPS
significant in human and mouse
significant in human
RPL29
ALYREF
PRRC2C
RBM15
SPEN
RPS15
RPS24
YBX1
MAP4
RBMX
STAT1
CELF1
RBMS1
RC3H1
VAV1
LDHA
STAT4
d
GO - Molecular function
# of proteins
mouse
human
4.6E-14
9.5E-14
6.2E-7
2.9E-8
1.0E-5
1.4E-9
1.1E-4
8.5E-4
1.3E-2
8.5E-4
1.4E-2
3.8E-7
1.4E-2
4.4E-5
2.1E-2
1.7E-5
2.1E-2
1.0E-1
2.3E-2
4.3E-3
mRNA binding
mRNA 3'-UTR binding
single-stranded RNA binding
catalytic activity, acting on RNA
poly(U) RNA binding
RNA helicase activity
poly-pyrimidine tract binding
double-stranded RNA binding
tRNA binding
regulatory RNA binding
e
IDR - OOPS
mOOPS
mEuRBPDB
all proteins
hOOPS
hEuRBPDB
all proteins
proportion of proteins
amount of region

Figure 3

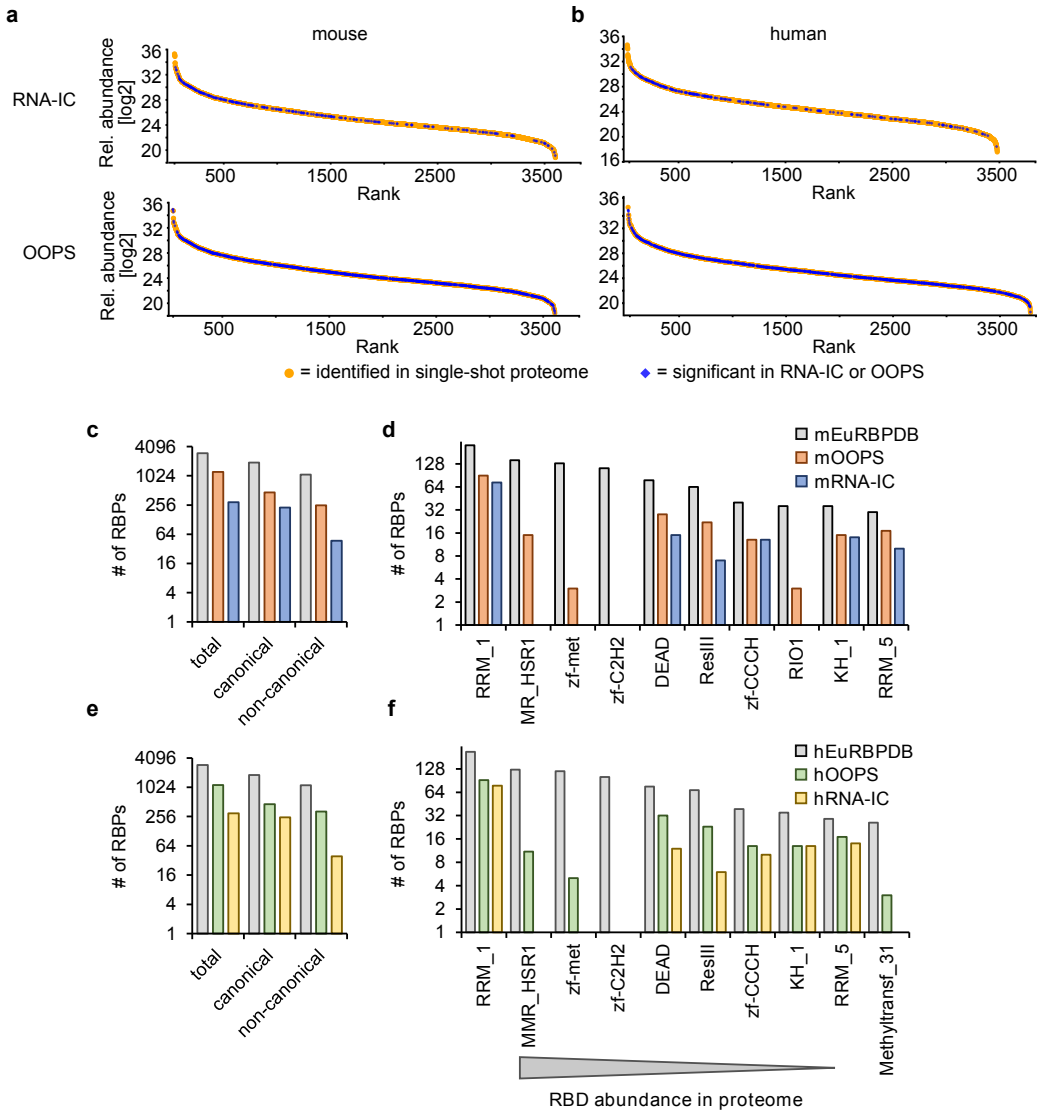

a
mouse
b
human
RNA-IC
OOPS
Rel. abundance [log2]
Rank
= identified in single-shot proteome
= significant in RNA-IC or OOPS
c
# of RBPs
total
canonical
non-canonical
d
mEuRBPDB
mOOPS
mRNA-IC
RRM_1
MR_HSR1
zf-met
zf-C2H2
DEAD
ResIII
zf-CCCH
RIO1
KH_1
RRM_5
e
f
hEuRBPDB
hOOPS
hRNA-IC
MMR_HSR1
KH_1
Methyltransf_31
RBD abundance in proteome
g
mOOPS 1255
hOOPS 1159
mRNA-IC 312
hRNA-IC 308
519
424
11
486
14
35
23
27
24
8
161
8
16
20
38
Mouse RBPome: 1352
*Core RBPome: 798
Human RBPome: 1250
*Core RBPome: 801

Figure 4

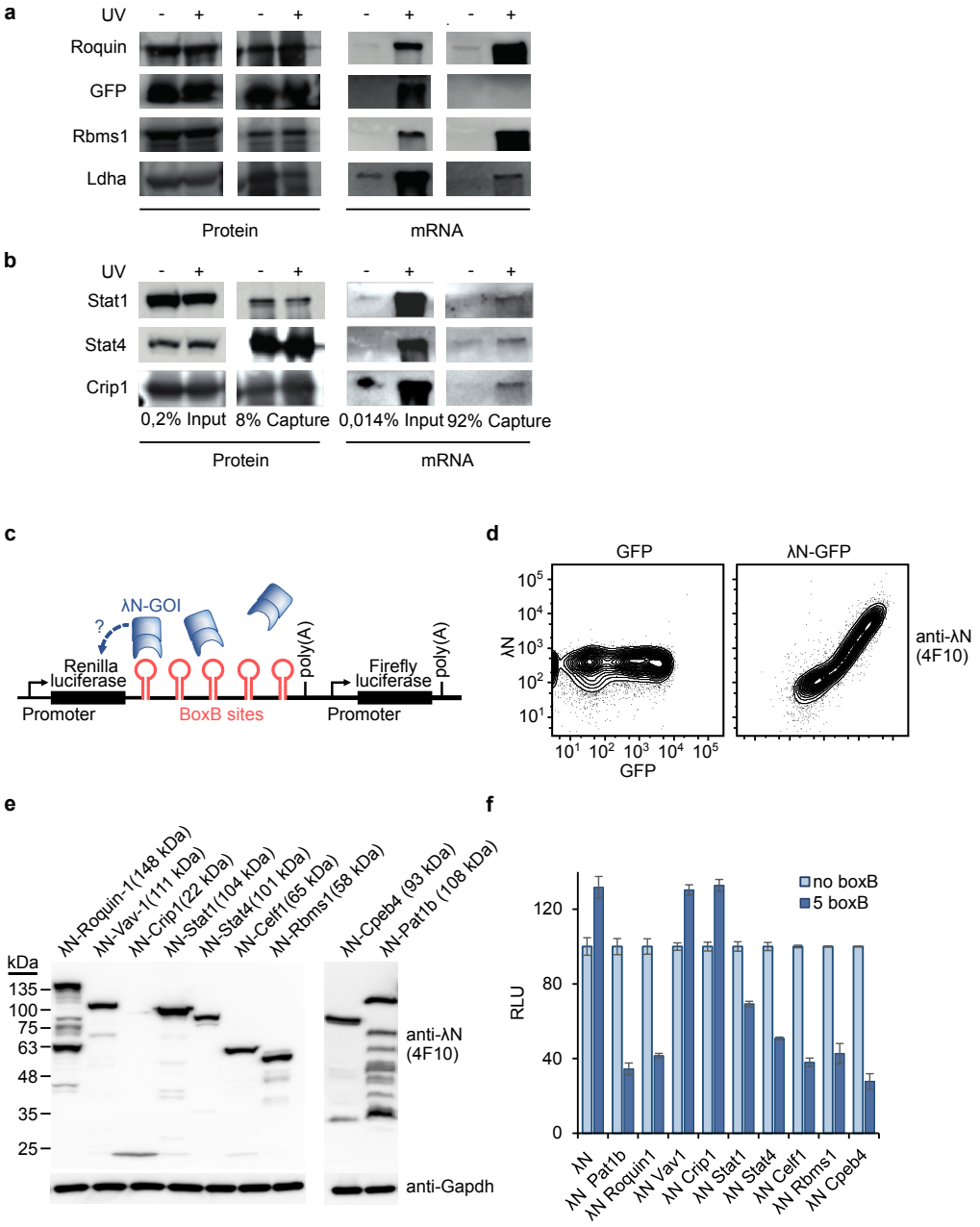
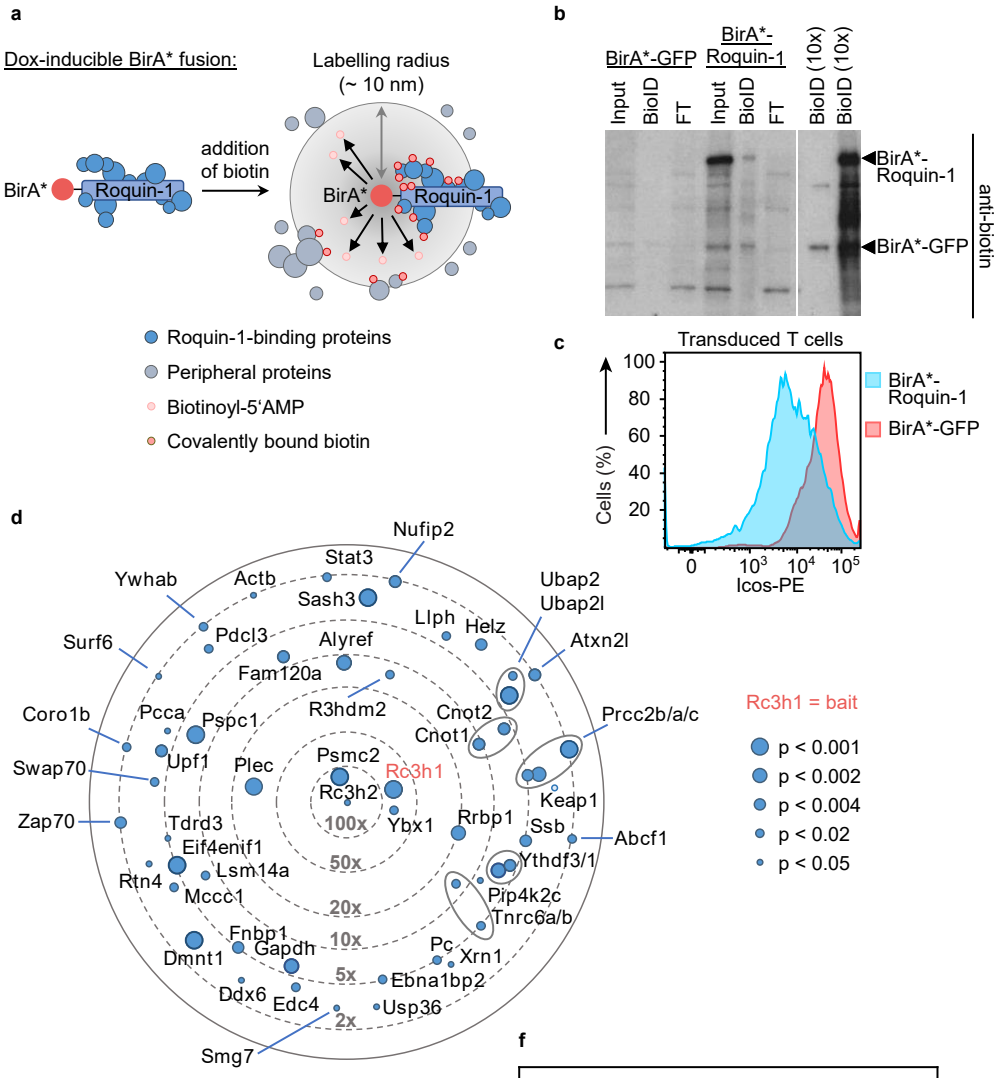

a
UV
Roquin
GFP
Rbms1
Ldha
Protein
mRNA
b
UV
Stat1
Stat4
Crip1
0,2% Input
8% Capture
0,014% Input
92% Capture
Protein
mRNA
c
λN-GOI
Renilla luciferase
poly(A)
Firefly luciferase
poly(A)
Promoter
BoxB sites
Promoter
d
GFP
λN-GFP
λN
GFP
anti-λN (4F10)
e
λN-Roquin-1(148 kDa)
λN-Vav-1(111 kDa)
λN-Crip1(22 kDa)
λN-Stat1(104 kDa)
λN-Stat4(101 kDa)
λN-Celf1(65 kDa)
λN-Rbms1(58 kDa)
λN-Cpeb4 (93 kDa)
λN-Pat1b (108 kDa)
kDa
anti-λN (4F10)
anti-Gapdh
f
no boxB
5 boxB
RLU
λN
λN_Pat1b
λN Roquin1
λN Vav1
λN Crip1
λN Stat1
λN Stat4
λN Celf1
λN Rbms1
λN Cpeb4

Figure 5



Roquin-1 T cell BioID overlap with RBPome			
Abcf1	Fnbp1	Prrc2c	Tnrc6b
Alyref	Helz	Pspc1	Ubap2
Atxn2l	Hist1h1c	R3hdm2	Ubap2l
Cnot1	Hist1h3b	Rc3h1	Upf1
Cnot2	Lsm14a	Rpl23a	Xrn1
Ddx6	Nufip2	Smg7	Ybx1
Ebna1bp2	Pc	Ssb	Ythdf1
Edc4	Plec	Stat3	Ythdf3
Eif4enif1	Prrc2a	Tdrd3	
Fam120a	Prrc2b	Tnrc6a	

Figure 6

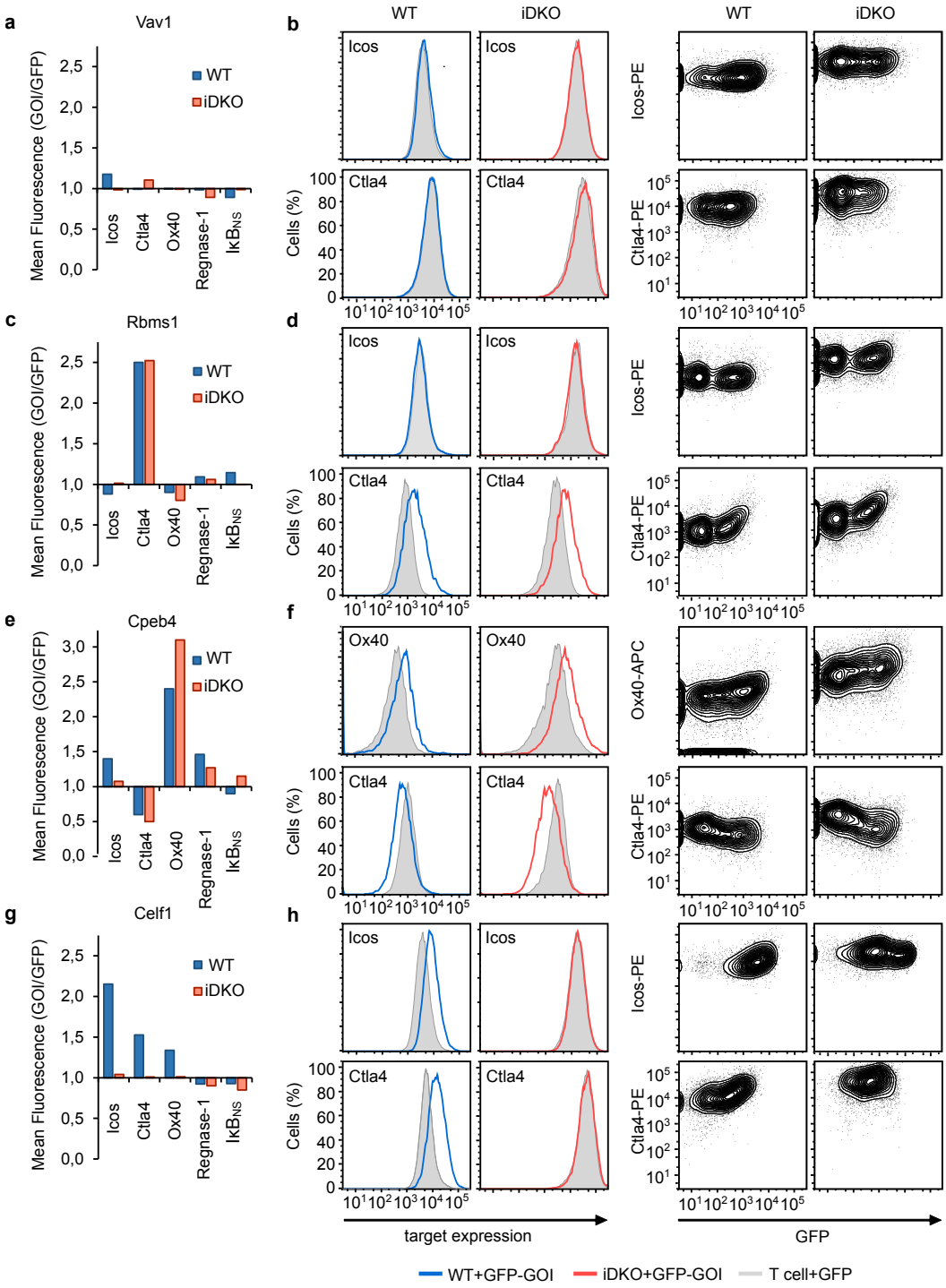

a
Vav1
Mean Fluorescence (GOI/GFP)
WT
iDKO
Icos
Ctla4
Ox40
Regnase-1
IκBNS
b
WT
iDKO
Icos-PE
Ctla4-PE
Cells (%)
c
Rbms1
d
e
Cpeb4
f
Ox40-APC
g
Celf1
h
target expression
GFP
WT+GFP-GOI
iDKO+GFP-GOI
T cell+GFP

Figure 7

3.3 Article 3: Mapping the trans-coregulatory network in yeast by ChIP-MS

Alexander Reim, Matthias Mann & Michael Wierer. **Mapping the trans-coregulatory network in yeast by ChIP-MS.**
In preparation for submission to Cell Systems.

The aim of conventional MS-based interactomics studies is to map the soluble interactions of proteins. Transcription factors are different from other proteins in such that chromatin-associated interactions differ from soluble ones. I established a straightforward ChIP-MS workflow to capture both the soluble and chromatin-related interactome of yeast TFs. Using a chromatin-immunoprecipitation based workflow, I greatly expanded the trans-regulatory network in yeast. Moreover, I identified co-regulatory transcription factors and enrichment correlation provided clues towards functional overlaps between TFs. Remarkably, the LFQ profile correlation of interactors detected proteins of unknown function associated with other trans-regulatory complexes or proteins. This led to the identification of a novel strong interactor of the RNA polymerase I machinery. The data demonstrates how system-wide ChIP-MS can vastly expand the knowledge about trans-regulatory TF networks.

Reim et al. 2020

Mapping the trans-coregulatory network in yeast by ChIP-MS

Alexander Reim[1], Matthias Mann[1]*, and Michael Wierer[1,2]*

1 Department of Proteomics and Signal Transduction, Max-Planck-Institute of Biochemistry, Munich, Germany

2 Proteomics Research Infrastructure, University of Copenhagen, Copenhagen, Denmark

* Corresponding Authors:

mmann@biochem.mpg.de

wierer@biochem.mpg.de

Abstract

Precise transcriptional regulation of gene expression plays a fundamental role for survival and proliferation in every living organism. Refined networks of protein-protein and protein-DNA interactions have evolved to ensure accurate transcriptional regulation. Mass spectrometry analysis of these interactions allows a system-wide view on the physical relationship of co-regulatory proteins. We obtained chromatin-associated protein-protein interaction data from 104 transcription factors in yeast by performing straightforward chromatin immunoprecipitation mass spectrometry (ChIP-MS) analysis followed by streamlined, unbiased data analysis. We observe 775 interactions of which 649 have not been mapped by previous immunocapture experiments. We show that this workflow specifically identifies interactions of transcription factors with transcription-related proteins. Strikingly, we observed differences in the number of interactors, which depended on the TF's molecular function, but were independent of the expression level. The dataset lays out co-regulatory relationships by uncovering heterodimeric transcription factor

complexes and previously unknown associations with gene-regulatory protein complexes. We further perform global correlation analysis to reveal the association of proteins of unknown function with chromatin-related proteins. Using this approach we provide first insights into the interactions of YDR249C with the RNA polymerase I core factor and upstream activation factor complexes. Collectively, we created a resource of the trans-regulatory network of transcription factors in yeast including various novel co-regulatory interactions and implication of proteins of unknown function in gene regulation.

Introduction

Precise regulation of gene transcription is one of the key mechanisms of cell homeostasis. The main regulators of transcription are transcription factors (TFs) that are sequence-specific DNA binding proteins which directly influence the expression levels of genes. Each cell depends on a plethora of different TFs ensuring correct temporal regulation of transcription and maintenance of cell identity (D'Alessio et al., 2015; Dufourt et al., 2018; Hahn et al., 2004; Kratsios et al., 2012; Meng et al., 2020; Natoli, 2010; Panman et al., 2011; Uyehara et al., 2017). The expression levels of TFs themselves need to be tightly controlled, which is partly achieved by feedback loops within transcriptional networks (Bornstein et al., 2014; Lee et al., 2001; Ngondo and Carbon, 2014).

Upon binding to promoter or enhancer elements on the DNA, the vast majority of TFs exert their function by recruiting other cofactors (Reiter et al., 2017). Coactivator complexes possess diverse biological functions to reorganize and modify the chromatin barrier in order to facilitate or impede the formation of the transcriptional pre-initiation complex (PIC) (Fuda et al., 2009; Li et al., 2007; Malik and Roeder, 2010).

Given their essential role in cells, many TF-recruited coregulator complexes are highly conserved from yeast to mammalian cells (Doyon et al., 2004; Gregoretti et al., 2004; Lardenois et al., 2015; Srivastava et al., 2015). One example are histone acetylation or deacetylation complexes, which

affect transcription by altering the accessibility of DNA or by introducing binding surfaces for other regulators on the acetylated lysines (Dhalluin et al., 1999; Lee et al., 1993; Owen et al., 2000). Given their importance in transcriptional regulation, histone acetylation marks are tightly controlled across the entire genome (Kurdistani et al., 2004; Wang et al., 2009).

Other examples comprise histone methylase and demethylase complexes, which place or remove histone methylation marks that form the basis of the epigenetic code (Sardina et al., 2018; Thambyrajah et al., 2016). Nucleosome remodeling complexes allow access of other regulatory proteins to the DNA by opening the chromatin structure (Mivelaz et al., 2020; Reddy et al., 2010; Vierbuchen et al., 2017). Dedicated bridging proteins link sequence specific TFs to the basal transcription machinery such as the mediator complex (Allen and Taatjes, 2015; Kagey et al., 2010; Soutourina, 2018).

TFs also frequently interact with other TFs to coregulate genes in a synergistic manner (Jolma et al., 2015; Lee et al., 2002; Stampfel et al., 2015; Xie et al., 2013). In fact, more than one third of TF-targeted genes are bound by more than one TF in yeast and three or more TFs consistently co-occupy more promoter regions than expected by random distribution (Lee et al., 2002).

Considering the diversity of interactions of TFs with other proteins in the chromatin environment, it becomes apparent that understanding the trans-regulatory interactome is important to unravel how the function of a TF is executed on a molecular level.

In yeast, several studies have investigated the protein-protein interactome both by yeast-two hybrid techniques (Fields and Song, 1989; Ito et al., 2001; Uetz et al., 2000) and affinity purification mass spectrometry (AP-MS) (Gavin et al., 2006; Gavin et al., 2002; Ho et al., 2002; Krogan et al., 2006). With the advancement of quantitative mass spectrometry, true-positive interactors can be analyzed by their enrichment over background proteins (Blagoev et al., 2003; Hein et al., 2015; Paul et al., 2011; Ranish et al., 2003). The population of non-specific background binders needs to be carefully controlled, which can be achieved by comparison of unrelated baits to improve the distinction between true interactors and artifacts (Keilhauer et al.,

2015). One major advantage of AP-MS experiments compared to other interactomic methods is their ability to identify transient or low-affinity interactors (Hein et al., 2015; Hosp et al., 2015; Keilhauer et al., 2015).

Chromatin-associated protein complexes are particularly challenging to study due to their high affinity to non-soluble chromatin. In conventional AP-MS protocols, DNA is usually fully degraded by the addition of nucleases like benzonase, which solubilize stable chromatin complexes (Gavin et al., 2006; Gavin et al., 2002; Ho et al., 2002; Hosp et al., 2015; Krogan et al., 2006). However, as the DNA is often an integral part of chromatin-associated protein complexes, linking together different TFs and cofactors, full enzymatic digestion of DNA is prone to lose information on local protein arrangements (Kim et al., 2013; Narasimhan et al., 2015; Panne, 2008; Slattery et al., 2011). In the epigenetic field this limitation is overcome by crosslinking the chromatin environment with formaldehyde and mechanic or enzymatic shearing of chromatin into 200 to 500 bp-long pieces. We and other have previously shown that such strategy can also be applied to AP-MS (Engelen et al., 2015; Hemmer et al., 2019; Rafiee et al., 2016; Tardiff et al., 2007; Wang et al., 2013).

Here, we used ChIP-MS and novel data analyses approaches to build up a global network of the transcriptional co-regulatome in yeast.

Results

A robust ChIP-MS workflow in yeast

To study chromatin-associated transcription factor complexes, we developed a fast and robust chromatin immunoprecipitation workflow followed by mass spectrometry (ChIP-MS) analysis in yeast. We used a library of yeast strains with endogenously C-terminally GFP-tagged proteins covering about 63% of all yeast open-reading frames (Huh et al., 2003). To select bona-fide transcription factors, we searched the Yeast Transcription Factor Specificity Compendium

database (de Boer and Hughes, 2012) for all yeast transcription factors with a published DNA-binding motif. Of 249 transcription factors in the database, 134 were present in the library, and 104 strains survived histidine dropout growth and expressed the GFP-tagged protein (Supplemental Table 1). The selected baits ranged from very high abundant TFs like Wtm1 with 73,381 molecules per cell to very low abundant TFs with an average of only four copy numbers per cell (Zap1) (Supplemental Figure 1a).

To perform ChIP-MS in these strains, we grew yeast cells to log phase (OD600 = 0.8) and performed formaldehyde cross-linking to freeze protein-DNA and protein-protein interactions (Figure 1). Following mechanical lysis by bead beating, we sonicated the chromatin to 100 to 500 bp, and enriched GFP-tagged TFs using a GFP nanobody coupled to agarose beads. After stringent washing steps we eluted precipitated proteins by on-bead trypsin digestion and analyzed the peptides in single shot LC-MS/MS runs on a Q-Exactive HF Orbitrap instrument. This regularly resulted in the quantification of roughly 50-60% of the yeast proteome in every single sample, of which most of the identifications were part of a consistently identified background proteome that served as accurate means of normalization.

Global statistics identifies a large number of novel interactions for individual TFs

Following label free protein quantification in MaxQuant, we set up a streamlined data analysis pipeline in Python (see Material and Methods). Thereby, we compared each bait to a bait specific control group of unrelated transcription factors and performed individual t-tests. Following a correction strategy for global interactome datasets (Hein et al., 2015), we adjusted the fold enrichment values by a bait-specific penalty factor to account for broadly enriched chromatin proteins by high abundant general TFs. The resulting normalized enrichment values of all pulldowns served as the basis for the definition of a global false discovery rate (FDR). This was calculated by the ratio of left (false positive) to right (true positive)-sided outliers at a given cutoff line, which was defined as -log10(p-value) = c/(x-x0), where x0 is the minimal enrichment and c

the curvature parameter (Hein et al., 2015; Keilhauer et al., 2015). Using an optimization function of the Scipy python package, we automatically adjusted both minimal enrichment and curvature parameter to maximize the number of interactors while staying below a defined false discovery rate (FDR). This analysis resulted in a total of 775 interactors at a FDR of 5% (Figure 2a, Supplemental data 1).

Notably, mapping known interactions from the BioGRID database to this dataset identified 80% of interactions as novel. This can be likely explained by our chromatin focused interaction approach, in contrast to most interaction datasets focusing on soluble complexes. In fact, a GO term enrichment analysis on the set of novel interactors revealed that almost all of these proteins were related to DNA binding, transcription factor activity and RNA polymerase activity (Supplemental Figure 1b).

TFs with a general function in transcriptional regulation interact with a much higher number of other regulators than very specific TFs. A principal component analysis shows that general TFs form a cluster separated from other baits (Figure 2b).

Next, we analyzed, whether our dataset is biased for the abundance of the bait in the sample. Plotting absolute intensity values against the number of interactors for each bait, shows a group of high abundant TFs that cluster in the top right corner, namely Spt15, Sua7, Nhp6b, Reb1, Cbf1 and Abf1 (Figure 2c). These TFs possess a general function in transcription or chromatin remodelling, which are present troughout the majority of promoter sites in the genome and therefore, a high number of interactors is expected. They also recruit various complexes involved in transcription or chromatin remodeling. For instance, Sua7 associates with various general transcription factor complexes, the RNA polymerase II and the Core mediator complex (Figure 2d).

Excluding this set of general TFs from the analysis, there was no apparent correlation of TF abundance and number of interactors (Figure 2c). We map the co-regulatory interactomes of both low and high abundant TFs. For instance, the TFs Hms2 and Hcm1 are very low abundant with

14 and 17 copies on average, respectively. In the case of Hms2 we observe novel transcriptional coregulators like Skn7, Hsf1 and Tbs1 and in the case of Hcm1 several members of the NuA4 acetyltransferase complex (Figure 2e and f). Among the most abundant TFs of this study is Hmo1 (25,398 copies), which interacts with TFIID members, the SMC5-6 complex and other coregulators (Figure 2g). Thus, we obtain relevant interactomes for both very low and high abundant TFs.

Analysis of TF-TF interactions reveal novel co-regulatory relationships

TFs often bind to and regulate gene enhancers in cooperation with other TFs (Jolma et al., 2015; Stampfel et al., 2015; Xie et al., 2013). To identify TF-TF interactions in an unbiased way, we generated a binary matrix showing bait proteins significantly enriched with another bait protein (Figure 3a). We identified 117 significant TF-TF co-enrichments. Notably, general transcription factors such as Sua7, Spt15 and Reb1 had the highest numbers of copurified TFs with 17, 11 and 14 TF interactions, respectively, reflecting their omnipresence at promoter sites (Figure 3b).

The majority of TFs with cooperative binding had a very selective set of TF cobinders. Those included well known TF-TF associations such as Dal81 with Stp1 and Stp2 (Boban and Ljungdahl, 2007), Tec1-Ste12 (Chou et al., 2006) and Rtg1-Rtg3 (Jia et al., 1997) (Supplemental Figure 2a-c). Strikingly, our dataset also revealed several novel TF associations.

One example is the association of the paralogous TFs Hms2 with Skn7 (Figure 3c). While Skn7 has been described to interact with Hsf1 to regulate stress-response genes (Raitt et al., 2000), not much is known about the function of Hms2 other than its role in pseudohyphal differentiation determined by overexpression screens (Lorenz and Heitman, 1998). Notably, our dataset confirmed the interaction of Skn7 with Hsf1, and also identified the same interaction for Hms2, suggesting that all three TFs cooperate in the transcriptional regulation of stress-response.

Another interesting connection is the interaction of Hal9 and Tbs1 (Figure 3d). Also, Hal9 and Tbs1 are paralogs and both proteins strongly enrich the other TF. Little is known about both TFs, however, overexpression of Hal9 in yeast has been linked to an increased salt tolerance (Mendizabal et al., 1998). Performing an alignment of both sequences revealed a high conservation of the DNA-binding domains (Supplemental Figure 2d). This indicates that likely both TFs consistently bind the same targets as a heterodimeric TF complex.

In addition, we identified six more cases for cooperative binding of paralogous TFs (Supplemental Figure 3), out of which 5 were previously unknown.

We also identified novel relationships among non-paralogous TFs. One example is the interaction between the two TFs Sfl1 and Sok2 (Figure 2d). Given the strong reciprocal enrichment, our dataset indicates a high degree of common promoter binding of Sfl1 and Sok2. This hypothesis is backed by a recent study, identifying simultaneous binding of Sfl1 and Sok2 at the promoter of the meiosis-specific *IME1* gene (Tam and van Werven, 2020).

Yeast transcription factors interact with a broad variety of DNA-modulatory complexes

Transcription factors interact with other chromatin modulatory or transcriptional complexes to execute their regulatory function. To study these interactions on a global level we created a matrix of all TFs, which interacted at least once with 20% or more of the members of a yeast protein complex (Figure 4a). As expected, general factors like Spt15, Sua7, Reb1, Nhp6b, Cbf1 or Abf1 bind to a large number of general transcriptional complexes and histone modifying complexes. The promiscuous binding to other complexes can be explained by their molecular functions. Reb1 displaces nucleosomes (Koerber et al., 2009) and is required for RNA polymerase termination (Colin et al., 2014). Indeed, we can observe that Reb1 interacts mostly with nucleosome remodeling complexes (e.g. INO80, ISW chromatin remodeling complexes, RSC complex, FUN30 complex), but also with the general TF complex TFIID, TFIIE, TFIIF complexes. Nhp6b and Abf1 have also been shown to be involved in chromatin reorganization and gene regulation (Lascaris

et al., 2000; Moreira and Holmberg, 2000) and similarly to Reb1 they interact with chromatin remodeling complexes and general transcription factor complexes.

Apart from binding to centromere DNA, Cbf1 may bind to the regulatory DNA of up to 10% of genes in yeast where it has an impact on chromatin structure (Kent et al., 2004). Accordingly, we observe Cbf1 to interact with centromere complexes like the CBF3 complex, the COMA complex and the central kinetochore CTF19 complex, but additionally associates with general transcription factor complexes and several chromatin remodeling complexes. This underscores the dual function of Cbf1 that has been proposed before (Kent et al., 2004).

The interaction matrix of transcription factors with other gene regulatory complexes also reveals unknown associations of TFs with transcriptional complexes. One example is Hcm1, which we found to interact with the NuA4 histone acetyltransferase complex (Figure 4a, 2f). The specificity and intensity with which we identify the NuA4 components co-enriched with Hcm1 in the ChIP-MS pulldowns provides strong evidence that Hcm1 associates with the NuA4 complex. Given the essentiality of NuA4 for cell proliferation, we searched for genetic relationship in double knock out data sets (Collins et al., 2007; Costanzo et al., 2010; Costanzo et al., 2016). Strikingly in all three datasets Hcm1 was genetically linked to members of the NuA4 complex, suggesting a strong functional association.

We also looked into the previously unknown interactors of the TFs and which complexes they are connected to (Figure 4b). Most of the novel interactions are with members of large transcriptional complexes (TFIIC, TFIID, TFIIE, TFIIF, TFIIH), RNA polymerase complexes or chromatin modulatory complexes (Rpd3L, RSC). Moreover, also the remaining interactions (less than 5 novel interactions) are almost exclusively with transcriptionally relevant complexes (Supplemental figure 4b).

Enrichment correlation sheds light on novel co-regulatory mechanisms of TFs

Next, we asked, whether our dataset can reveal novel functional relationships of TFs, based on similar sets of chromatin-associated protein interactions. To this end, we filtered for proteins, which were significantly enriched in at least one pulldown, and correlated their profiles of fold enrichment values across all baits (Figure 5a).

The TFs Abf1, Cbf1, Nhp6b and Reb1 with a broad role in transcriptional regulation and chromatin remodeling form a specific cluster (cluster 1, Figure 5a). The LFQ profile across general TF complexes (e.g. TFIIF, TFIIK) and chromatin remodeling complexes (e.g. INO80, ISW2 and RSC) is highly consistent (Figure 5b). The general TFs Spt15 and Sua7 (cluster 2, Figure 5a) correlate strongly with cluster 1. They enrich chromatin remodelers like INO80, ISW2 and RSC at similar levels (Figure 5b). Yet, they separate from Abf1, Cbf1, Nhp6b and Reb1 in the hierarchical clustering as they enrich general transcriptional complexes stronger (TFIIF, TFIIK) or even exclusively (TFIIA).

Cluster 3 reflects TFs interacting with the histone deacetylase complex Rpd3L (Figure 5a and c). This cluster consists of Ash1, Stb3, Rph1, Gis1 and Sfl1. Ash1 is a known subunit of the Rpd3L complex itself. However, we provide first physical evidence that Stb3, Rph1 and Gis1 associate with the entire Rpd3L complex, while Sfl1 enriches specific subunits. These data also back previous genetic or biochemical data linking these TFs to the Rpd3L complex.

Rpd3 regulates 58 of 158 Rph1-repressed genes, which indicates a cooperative relationship (Liang et al., 2013). Rpd3 was also shown to associate with Sfl1 at *FLO11* promoters (Bumgarner et al., 2009). There is no published data linking Gis1 to the Rpd3L complex.

The correlation matrix also revealed several new relationships of TFs based on common interactor sets (cluster 4, Figure 5a). For instance, Sfl1 and Sok2, which interact with each other (Figure 3d), show a strikingly similar enrichment of a set of three TFs, Sko1, Swi4, Phd1, and the general transcriptional corepressor Cyc8-Tup1 (Figure 5d). A physical interaction of any of these TFs has not been reported yet, however biochemical evidence and sequence similarity in the

DNA-binding domains indicate a cooperative relationship. Tam et al. recently showed Sok2, Phd1, Sko1 and Sfl1 to be enriched at the promoter of the meiosis-specific *IME1* gene (Tam and van Werven, 2020). Furthermore, the same study showed these TFs to recruit the corepressor complex Cyc8-Tup1 to the promoter.

We also observe a cluster of the TFs Ste12, Tec1 and Kar4 (cluster 5, Figure 5a). Ste12 interacts with Tec1 to regulate invasive growth (Figure 5e, left panel). Dig1 and Dig2 inhibit this complex and unphosphorylated Kss1 binds directly to Ste12. Phosphorylation of Kss1 ultimately leads to the dissociation of Dig1/2 and derepression of the Ste12/Tec1 transcriptional complex (Bardwell et al., 1998). Interestingly, Kar4 also interacts with Dig1, Kss1 and Ste12, but not with Tec1 and Dig2 (Figure 5e, right panel). Kar4 associates with Ste12 independently of Tec1 on genes required for karyogamy (Lahav et al., 2007). Based on our dataset we can postulate a transcriptional complex of Ste12 and Kar4 cooperating with Dig1 and Kss1, but not with Dig2.

Collectively, correlation of ChIP-MS data across transcription factors allows the assessment of similarities between TFs leading to novel insights into their regulatory interactions.

Global correlation network of interactors reveals novel coregulators of transcriptional complexes

Correlation of prey proteins across high throughput quantitative mass spectrometry experiments can yield information on the organization of protein complexes beyond the information obtained by bait individual interactomes (Hein et al., 2015). We correlated the LFQ intensities of all significantly enriched proteins in each pulldown resulting in a large heatmap of protein-protein LFQ correlation values (Supplemental Figure 5). This heatmap can be converted into a correlation network of transcription regulating proteins (Figure 6a). Reassuringly, we observe that known chromatin remodeling and histone modulatory complexes (e.g. RSC, Rpd3L, NuA4) as well as transcriptional complexes (e.g. RNA polymerase complex, Core Mediator complex) show strong correlations between their subunits and thus form subclusters. The most intriguing question to

answer with these types of network is whether we identify proteins of unknown function strongly correlating with known transcriptional complexes or other chromatin binding proteins suggesting a role in transcriptional regulation. One prominent example is the protein of unknown function YDR249C. We observed a pearson correlation larger than 0.65 with the RNA polymerase I upstream activating factor complex (UAF) subunits Uaf30, Rrn5, Rrn9, the RNA polymerase I core factor complex (CF) subunits Rrn7, Rrn11 and the RNA polymerase I transcription initiation factor Rrn3. Notably, none of these proteins had been used as baits in the initial ChIP-MS dataset. YDR249C is a largely uncharacterized protein with no mapped physical interactions. The only connection to the UAF/CF complexes are a negative genetic relationship with Rrn3 and Rrn9 as observed in a high throughput global genetic interaction network (Costanzo et al., 2016). To validate an interaction between YDR249C and the UAF/CF complexes, we performed parallel ChIP-MS pulldowns of Rrn3, Rrn5, Rrn6, Rrn7, Rrn10 and Rrn11. Indeed, we enriched YDR249C with every member of the UAF/CF complexes, but not with the bridging factor Rrn3 (Supplemental Figure 6). Rrn3 recruits RNA polymerase I to rDNA promoters by linking it to the CF complex and is not a stable subunit of neither CF nor UAF. Accordingly, we also did not find Rrn3 enriched in the pulldowns of other RRN subunits. Including the UAF and CF pulldowns in the correlation, we observe a pearson correlation of 0.93 for YDR249C with the UAF/CF member Rrn9, which is comparable to the correlation of other UAF/CF members among each other (Figure 6b). We therefore hypothesize that YDR249C may be a cofactor of the UAF/CF complexes and also binding at RNA Pol I promoters. To verify that the interaction is specific for the chromatin context, we compared a regular affinity-enrichment pulldown of Rrn7 with a ChIP-MS pulldown (Figure 6c). The conventional pulldown of Rrn7 resulted in the significant enrichment of the CF complex (Rrn6-Rrn7-Rrn11), of which the structure has been determined before (Engel et al., 2017). However, we did not see any enrichment of the other UAF members or Rrn3. Performing a ChIP-MS pulldown of Rrn7 recovers the CF complex, the UAF complex, Rrn3, Net1, Spt15 and YDR249C, as well as several subunits of the DNA-directed RNA polymerase I complex. This

underscores the power of using ChIP-MS pulldowns for studying transcriptional complexes, as cross-linking and preservation of the chromatin environment largely expands the biological insights obtained by the observed interactions.

Discussion

Yeast is a key model organism that has been used to unravel molecular and cellular pathways, signaling cascades and transcriptional networks. It has also been extensively used to map protein complexes in high throughput studies. Even though being a well-studied system, we expected to expand the transcriptional interactome by using a chromatin immunoprecipitation workflow combined with mass spectrometry analysis.

In DNA-binding studies (e.g. ChIP-Seq or ChIP-ChIP) the information obtained from the resulting DNA-binding information is often utilized to identify co-regulatory relationships of TFs. This is however limited to the studied baits. ChIP-mass spectrometry also captures TF-TF co-occupancies if they cooperate on a large set of genes or even form soluble heterodimers.

While various publications used affinity enrichment workflows with mass spectrometry or western blotting as a readout in both a low- and high throughput, substantially less studies investigated interactions in the chromatin environment. With a direct comparison of a conventional pulldown of the RNA Polymerase I CF complex subunit Rrn7 to a ChIP-MS pulldown, we show how preserving the chromatin environment severely affects the interactome. Combining formaldehyde cross-linking with a stringent GFP-affinity enrichment of GFP-tagged TFs analyzed by quantitative MS, we could identify a core interactome of yeast TFs and expand the trans-regulatory network. Yet, we also observe that most TFs only show less than 5 interactors. This likely reflects the fact that our method captures both soluble and chromatin-associated protein interactions. Therefore, if a TF does not possess many soluble interactors and only a minor fraction of it is bound to the chromatin, we will likely do not observe a high number of interactors. On the other hand, functions

of a TF not involving DNA-binding and the chromatin environment will also be identified. For instance, although having a DNA-binding motif, Tho2 is also part of the THO complex which links transcription to the export of mRNA to the cytoplasm. Accordingly, we predominantly identify THO complex members and other mRNA-binding proteins as members, but we also capture the histone acetyltransferase Hat2 as an interactor. Applying our ChIP-MS workflow, we capture heterodimeric coregulations of TFs which share a homologous DNA-binding domain or which regulate overlapping sets of target genes. Moreover, we find novel interactions with regulatory complexes or corroborate previous links obtained by genetic studies. Association of Stb3 with the Rpd3L complex or Hcm1 with the NuA4 complex are just two of several examples. Finally, this can also lead to mechanistic insights as in the case of Kar4, where the strong correlation with Ste12 and Tec1 suggests a similar regulatory mechanism as described in the literature for the heterodimeric complex of Ste12 and Tec1. All of this data demonstrates that ChIP-MS workflows outperform conventional methods in the context of TF interactome studies.

Another great advantage of a large ChIP-based quantitative mass spectrometry dataset is the opportunity of creating global correlation networks. While general transcriptional protein complexes like RNA polymerases are difficult to resolve in this network, more specific complexes are clustering together with only a few edges to other proteins not belonging to the complex. This allows the identification of novel interactors of gene regulatory complexes beyond the studied baits. Even though yeast is one of the most studied model organisms and its gene regulatory complexes have been studied in great detail, we observe proteins of unknown function strongly correlating with other DNA-binding or trans-regulatory proteins. One remarkable case is the protein YDR249C. The correlation network led us to identify YDR249C as an interactor of the entire RNA Pol I UAF/CF complex. As this interaction does not exist with the soluble RNA Pol I CF complex, it likely depends on the chromatin environment. We cannot yet describe the precise role of YDR249C within the UAF/CF complexes or its role in regulation of genes transcribed by RNA polymerase I. Still, we provide evidence that the molecular function of YDR249C is linked to

the UAF and CF complex and transcription of RNA polymerase I. Collectively, we show that chromatin immunoprecipitation mass spectrometry can serve as a valuable tool to study transcription factor interactomes allowing the identification of novel transcriptional coregulators of TFs even in a well characterized organism like yeast.

Methods

Yeast strains and culture

Yeast strains were taken from the Yeast-GFP clone collection, which is a library of endogenously GFP-tagged proteins covering about 63% of the S. cerevisiae open reading frames (Huh et al., 2003). We searched for bona fide transcription factors with a DNA-binding motif in the YetFaSco database (de Boer and Hughes, 2012). Of this dataset, we cultured 104 GFP-tagged TFs available in our library. The parental strain of the library is BY4741 (ATCC 201388). The histidine synthesizing strain pHis3-GFP-HIS3kMX6 served as an alternative control strain and was identical to the strained designed in a previous study (Keilhauer et al., 2015). Yeast strains were streaked on fresh YPD plates (BY4741 strain) or SC-His plates (tagged strains). Twenty-five milliliters of YPD medium were inoculated with the tagged strains, the parental BY4741 control strain or the pHis3-GFP control and grown overnight at 30°C. The next day, the cultures were diluted to an OD600nm of 0.2 in 400 ml of YPD medium. OD600nm was regularly checked and cultures were harvested when an OD of about 0.8 was reached. For ChIP-MS experiments, cultures were crosslinked with 1% of formaldehyde for 15 min at room temperature (RT). Subsequently, formaldehyde was quenched with 300 mM of glycine for 15 min at RT. Next, cultures were centrifuged at 4°C and 500g for 5 minutes. For GFP pulldown experiments, cultures were directly centrifuged at an OD600nm of 0.8. Pellets were washed once with 25 ml of ice-cold PBS and centrifuged again. Finally, pellets were resuspended in 1 ml of ice-cold PBS and centrifuged. PBS was aspirated and pellets flash-frozen and stored at -80°C.

Chromatin immunoprecipitation

Formaldehyde crosslinked cell pellets were dissolved in 1 ml of ChIP lysis buffer (50 mM HEPES, pH 8), 150 mM NaCl, 1 mM EDTA (pH 8.0), 1% Triton-X-100, 0.1% Sodium Deoxycholate (SDC), 0.1% Sodium Dodecyl Sulfate (SDS), complete protease inhibitors (Roche), phosphatase inhibitors (Roche)) and transferred into FastPrep tubes (MP Biomedicals) with 1 mm silica spheres (lysing matrix C, MP Biomedicals). Dissolved pellets were lysed in a FastPrep instrument (MP Biomedicals) for 60 s per cycle at maximum speed with 6 cycles at 4°C and breaks of 5 min between cycles. For comparison with non-crosslinked pulldowns, mechanical lysis was performed as described for the non-crosslinked yeast cell lysis. Lysates were pelleted at 16,100 x g for 15 min at 4°C. Pellets were again dissolved in 1 ml of ChIP lysis buffer and transferred into conical centrifuge tubes (Falcon). Samples were then sonicated with a metallic rod in a bioruptor (Diagenode) for 20 cycles at 4°C. One cycle lasted 1 min with 30 s of sonication and 30 s break. After 10 cycles, samples were put on ice for 5 min. After sonication, cell debris was separated from the lysate by centrifugation for 5 min at 4°C and 6200 x g.

Checking chromatin length after sonication

Successful sonication of chromatin to a length of between 100-500 bp was checked by taking 25 µl of lysate and reverse the crosslinking by adding 1 µl of 5 M NaCl and boiling at 99°C for 15 min. The sample was cooled down and 1 µl of Rnase A (Thermo, EN0531) was added. RNA was digested by incubation for 20 min at 37°C. Four microliters of Proteinase K (Thermo, AM2546), 2 µl of Tris, pH 7.5 and 1 µl of EDTA, pH 8.0, were added and proteins digested at 56°C for 20 min. DNA was purified with the Qiagen PCR purification kit (Qiagen, 28104) according to the manufacturer's protocol. DNA was separated on a 1% Agarose gel for 45 min at 110 V and DNA was checked for a length of between 100-500 bp.

GFP-pulldowns of crosslinked cultures

Protein concentration was determined using the BCA Protein-Assay-Kit (Thermo, 23227) according to the manufacturer's protocol.

For each GFP pulldown, 20 µl of anti-GFP agarose beads were washed twice with TBS (150 mM NaCl, 50 mM NaCl, pH 7.5) and distributed to a 96-deep well plate. Each GFP-tagged bait was split into triplicates and 2.5 mg of protein lysate were added to the beads. 1.3 mg of protein lysate were used for characterization of the RNA polymerase I activator complex and the comparison to non-crosslinked pulldowns. Each well was diluted to 1 ml with ChIP lysis buffer. Plates were sealed and lysates incubated with the beads for 2 h at 4°C and 1,200 rpm. Beads were pelleted by centrifugation at 2,000 rpm, 3 min at 4°C. Beads were washed three times with 1 ml Low Salt wash buffer (50 mM HEPES, pH 7.5, 140 mM NaCl, 1% Triton-X-100), once with 1 ml High Salt wash buffer (50 mM HEPES, pH 7.5, 500 mM NaCl, 1% Triton-X-100) and twice with 1 ml TBS (50 mM Tris-HCl, pH 7.5, 150 mM NaCl). Fifty microliters of elution buffer (2 M Urea, 50 mM Tris-HCl, pH 7.5, 2 mM DTT and 20 µg μl^{-1} Trypsin) were added and beads incubated for 30 min at 37°C and 1400 rpm. Eluates were transferred to tubes and beads incubated for 5 min with 50 µl of alkylation buffer (2 M Urea, 50 mM Tris-HCl, pH 7.5, 10 mM Chloroacetamide) at 37°C and 1,400 rpm. Eluates were again removed and combined with the first eluate. Samples were further digested overnight at 25°C and 800 rpm.

Cell Lysis and GFP pulldown of non-crosslinked cultures

Yeast cell pellets were dissolved in 1 ml of IP lysis buffer (150 mm NaCl, 50 mm Tris-HCl (pH 8.0), 1 mm $MgCl_2$, 5% glycerol, 1% IGEPAL CA-630, complete protease inhibitors (Roche), phosphatase inhibitors (Roche), 1% benzonase (Novagen, 70746) and transferred into a 96-deep well plate containing Zirconia beads. Plates were sealed and pellets were lysed in a Geno/Grinder (Horiba) for 90 s per cycle at 1750 rpm with 5 cycles at 4°C and breaks of 5 min between cycles. Subsequently, lysates were centrifuged at 16,100 x g for 15 min at 4°C. The supernatant was

taken and the protein concentration determined using the BCA Protein-Assay-Kit (Thermo, 23227) according to the manufacturer's protocol.

For each GFP pulldown, 20 µl of anti-GFP agarose beads were washed twice with TBS (150 mM NaCl, 50 mM NaCl, pH 7.5) and distributed to a 96-deep well plate. The lysate of GFP-tagged bait and control strain was split into triplicates and 1.3 mg of protein lysate was added to the beads. Each well was diluted to 1 ml with IP lysis buffer. Plates were sealed and lysates incubated with the beads for 2 h at 4°C and 1,200 rpm. Beads were pelleted by centrifugation at 2,000 rpm, 3 min at 4°C. Beads were washed twice with 1 ml of IP wash buffer 1 (150 mm NaCl, 50 mm Tris-Cl (pH 8.0), 0.25% IGEPAL CA-630) and four times with 1 ml IP wash buffer 2 (150 mm NaCl, 50 mm Tris-Cl, pH 8.0). Elution of proteins was performed exactly as for crosslinked samples.

Stage-tip purification of eluted peptides

After elution of peptides from the beads by overnight digestion with trypsin, peptides were acidified by the addition of 1 µl of Trifluoroacetic acid (TFA). Peptides were purified following previously described standard protocols (Rappsilber et al., 2007). To this end, stage tips were prepared by filling regular pipette tips with 3 layers of C18 material (Empore). Stage tips were equilibrated with 100 µl of methanol, 100 µl of Buffer B (0.5% Acetic Acid, 80% Acetonitrile (ACN)) and 100 µl of Buffer A (0.5% Acetic Acid). Peptides were loaded onto the stage tip. After one wash with Buffer A, peptides were eluted with 60 µl of Buffer B. Samples were dried in a SpeedVac concentrator for 60 min. Samples were resuspended with 10 µl of Buffer A* (2% ACN, 0.1% TFA) and incubated at 2000 rpm and RT for 10 min to ensure complete resuspension of the sample. 4 µl of sample were used for mass spectrometry analysis.

LC-MS/MS measurements

A Thermo EASY-nLC 1200 UHPLC system (Thermo Fisher Scientific, Bremen, Germany) was used for online chromatography which was coupled online to a Q Exactive HF mass spectrometer with a nano-electrospray ion source (Thermo Fisher Scientific). Comparison of crosslinked and

non-crosslinked samples and pulldowns of the RNA Polymerase I UAF/CF complex were measured on a Q Exactive HF-X mass spectrometer using the same gradient and settings. 50 cm analytical columns (75 µm inner diameter) were packed in-house with ReproSil-Pur C18 AQ 1.9 µm reversed phase resin (Dr. Maisch GmbH, Ammerbuch, Germany) in 0.1% formic acid. The analytical columns were placed in a column heater (Sonation GmbH, Biberach, Germany) during online analysis, which was regulated to a temperature of 60˚C. Peptide mixtures were loaded onto the analytical column in 0.1% formic acid and separated with a linear gradient of 5-32% buffer B+ (80% ACN and 0.1% formic acid) for 100 min at a flow rate of 300 nl min^{-1}. A washout with up to 95% of ACN followed the gradient to prepare the column for the next sample. The overall gradient length was 115 min. For MS data acquisition we used a data-dependent top 10 method in positive mode using Tune 2.9 and Xcalibur 4.1. The capillary temperature was 250°C and S-lens RF level was set to 40.0. MS full scan data acquisition was set to a resolution of 60,000 and a maximum ion injection time of 20 ms and an AGC target value of 3E6. The isolation window for selection of precursor ions was 1.4 m/z and the fragmentation by HCD occurred at a normalized collision energy of 27. Product ions were measured at a resolution of 15.000 with a maximum injection time of 60 ms and an AGC target of 1E5 ions. Precursor ions with unassigned, single charge states or charge states larger than 6 were excluded from fragmentation selection and repeated sequencing minimized by a dynamic exclusion window of 30 s.

Raw data processing

All raw files except the comparison of non-crosslinked versus crosslinked samples and the pulldowns of RNA polymerase I activating complex factors were processed together using the MaxQuant software (Cox and Mann, 2008) (version 1.5.6.7). Peak lists were searched against the yeast Uniprot FASTA database combined with 262 common contaminants by the integrated Andromeda search engine (Cox et al., 2011). The database was searched for peptides with trypsin-specific C-terminal cleavages after lysine or arginine. Two missed cleavages were allowed and the minimum length of peptides was set to 7. Oxidation of methionine and acetylation of

protein N-termini were considered as variable modifications, while carbamidomethylation was set as a fixed modification on cysteine. PSM and protein identifications were filtered using a target-decoy calculation at a false discovery rate (FDR) of 1%. Matching between runs was carried out with a window of 0.7 min and an alignment time window of 20 min. Label-free quantification was performed using the MaxLFQ algorithm (Cox et al., 2014). The FastLFQ option was enabled and unique and razor peptides were used for quantification. The minimal ratio count for LFQ calculations was set to 2.

Data analysis

First, the MaxQuant output file was loaded into Perseus (version 1.6.0.7) and contaminants, proteins from the reverse database and proteins never identified without a modified peptide were removed. The logarithm (base 2) of the LFQ intensities was calculated and proteins that did not have at least 2 valid values in one set of triplicates were eliminated. Next, missing values were imputed from a normal distribution around the detection limit of the mass spectrometer. The normal distribution had a downshift of 1.8 standard deviations and a width of 0.25 standard deviations. This table was exported and used for further analysis in Python (version 3.7.3).

Since sonication of chromatin was a rate-limiting step, we could only process subsets of baits together. In order to minimize batch effects, we grouped baits if they were processed together and controls and enrichments were calculated within these groups (group A through H). First, we created a bait-specific control group (BSCG) for each bait. It has been shown that a BSCG improves the distinction between true- and false-positive interactors (Hein et al., 2015; Keilhauer et al., 2015). In addition to the control pulldown, we used other TFs as controls to create a BSCG for every TF. However, a large set of transcription factor pulldowns is by default more likely to be related and therefore selection of control baits cannot be performed as previously described (Hein et al., 2015). Therefore, we came up with a different approach. Usually, we require the majority of proteins (background proteome) to behave most similar in the control and bait pulldown in order to allow a good separation of background and specific interactors. However, TFs with a similar

interactome will behave even more similar in the entirety of their enrichment profile than unrelated TFs. For this reason, we divided the preys into a transcriptional and into a non-transcriptional group by filtering for the GO Molecular Function terms “DNA binding”, “chromatin binding”, “regulation of transcription”, “DNA-dependent RNA polymerase” and “chromatin modification”. Next, we calculated the correlation between the baits for both the transcriptional and non-transcriptional prey group. We sorted the non-transcriptional correlations by a descending correlation value. Starting from the top correlation, we required the correlation of the transcriptional prey group to be below the median of all correlations plus one interquartile range. The first 4 transcription factors meeting these criteria and the control pulldown were chosen as the BSCG for each TF. This allowed an automatic assignment of control baits in most cases. Nevertheless, we applied some red flags if baits were subunits of complexes or had an overlap in interactions with another TF. Using these control sets we performed Student’s T-Tests for every bait with the “scipy.stats.ttest_ind” package in Python. Resulting differences of LFQ intensities and p-values can be used to create volcano plots and define true positive outliers. We observed that some baits showed comparably wide distributions in the volcano plots. This has been observed before with baits organized in large complexes (e.g. ribosome) (Hein et al., 2015). In order to still apply the same cutoff criteria to an entire group of baits, a penalty factor for each bait was calculated to account for these deviations in the background proteome. This factor was calclulated as described before (Hein et al., 2015) and multiplied with the t-test differences of the respective bait.

Identification of significant interactors was performed modifying a previously reported strategy (Hein et al., 2015; Keilhauer et al., 2015):

Cutoff calculation was based on the formula: $-\log_{10}(p) \geq c/(x-x_0)$ with p: p-value, c: curvature parameter, x: fold enrichment of a protein and x_0: fixed minimum fold enrichment.

By the nature of an affinity-enrichment of a bait, left-sided outliers are impossible and can therefore be defined as false-positives. With that definition, we can calculate a FDR based on the

ratio of left-sided (false positive) to right-sided (true positive) outliers. To this end, we used the scipy.optimize.minimize package in Python to evaluate the optimal c and x_0 parameters that return the maximum number of right-sided outliers while keeping the ratio of left-sided outliers to right-sided outliers below 0.01 or 0.05 (i.e. a FDR of 1% or 5%, respectively). We defined the optimal c and x_0 parameters for every set of experiments individually (i.e set A through H).

For the identification of cooperative TF-TF pairs (Figure 3a), we expanded our cut-off calculation to a false discovery rate of 10% as we consider unspecific enrichment of TFs less likely than other proteins and created a square matrix where all TFs represent the bait on the x-axis and the prey on the y-axis.

Yeast complexes were downloaded from the Complex Portal (Meldal et al., 2019). The interaction map of TFs with yeast complexes in Figure 4a was filtered for complexes where at least one TF interacted with at least 20% of the complex members.

For enrichment correlation between baits to identify functional overlaps (Figure 5a), we calculated the pearson correlation between baits based on the enrichment of proteins significantly enriched by at least one TF.

The global correlation network was created by calculating the pearson correlation for each protein of the mean LFQ intensities across all baits in Python. Correlation values were filtered to remove those below 0.65. This correlation network was imported into Cytoscape (version 3.4.0). Edges represent the correlation value between two proteins.

To determine the GO enrichment of molecular functions and biological processes of novel interactors, we used the clusterProfiler package implemented in R (version 3.5.0). Uniprotkb identifiers were converted to EntrezGene identifiers and GO enrichment was performed as described for the clusterProfiler package (Yu et al., 2012).

For comparison with previous yeast protein interactomics, we downloaded the interactions annotated on Biogrid (version 3.5.186). These were filtered for physical interactions identified by affinity enrichment mass spectrometry or affinity enrichment western blotting.

Acknowledgements

We thank Alexander Strasser for help with yeast culturing. We also thank Igor Paron for his assistance with mass spectrometry maintenance and measurements.

Author contributions

A.R. conducted experiments and performed data analysis. A.R-.and M.W. designed experiments and wrote the manuscript. M.W. and M.M. supervised the study.

Conflict of interest

The authors declare that they have no conflict of interest.

Figure legends

Figure 1: Chromatin immunoprecipitation mass spectrometry workflow. a, GFP-tagged strains were cultured in YPD medium to an OD_{600nm} of about 0.8. Cells were formaldehyde crosslinked and stored at -80˚C until further processing. After mechanical lysis, chromatin was sheared by sonication to a length of between 100-500 bp. GFP-tagged TFs were enriched with anti-GFP nanobeads, stringently washed and proteins digested with trypsin. After mass spectrometry analysis, data analysis was performed using Python. Statistically significant interactors were identified by a student's T-test and applying FDR control as described before (Hein et al., 2015; Keilhauer et al., 2015).

Figure 2: Comparison of interactors to earlier datasets and differences between TF interactomes. a, Overlay of significant interactors (FDR = 5%) from all volcano plots showing the

enrichment of baits, known interactors and novel interactors. The x-axis represents the fold-enrichment (log2) of proteins and y-axis depicts the p-value (log10) resulting from t-test analysis. Baits are highlighted in red, novel interactors in orange and known interactors in blue. **b,** 2D-Principal component analysis of LFQ intensities. Baits distinctive from other baits by the second principal component are labeled. **c,** The relationship between the number of interactors (log2, y-axis) with the average copy number (log10) of the baits. **d,** Volcano plot of the general transcription factor Sua7 shows interactions with a broad set of co-regulatory protein complexes like general TFs (light pink), the RNA Polymerase II apparatus (green) or the Core Mediator complex (orange). The dashed line depicts the cutoff of the 5% FDR (inner line) or 1% FDR (outer), respectively. Axis labeling as in Figure 2a. **e,** Volcano plot of the low abundant transcription factor Hms2. Hms2 interacts with a specific set of other transcription factors like Skn7, Hsf1 and Tbs1. Axis labeling as in Figure 2d. **f,** Volcano plot of the low abundant transcription factor Hcm1. ChIP-MS data for Hcm1 reveals a novel interaction with the NuA4 histone acetyltransferase complex. Axis labeling as in Figure 2d. **g,** Volcano plot of the one of the most abundant baits in this study, Hmo1. Axis labeling as in Figure 2d.

Figure 3: Analysis of co-regulatory transcription factors. a, Binary matrix of TFs significantly enriched by another TF with a 10% FDR cutoff. Blue squares indicate a positive event where the bait TF (x-axis) enriched the prey TF (y-axis). **b,** Table listing all interactions of a bait with another TF at a 10% FDR cutoff studied in this dataset**. c,** Volcano plots showing the ChIP-MS data of Skn7 and Hms2. Dashed lines depict the FDR cutoffs of 10% (inner line), 5% (middle), 1% (outer). The bait is shown in red and the prey TF in orange. d. Volcano plots of the largely uncharacterized TFs Tbs1 and Hal9. Proteins are highlighted as in d. **e,** Scatter plot of the mutual enrichment of the TFs Sfl1 and Sok2.

Reim et al. 2020

Figure 4: Interactions of transcription factors with yeast protein complexes. a, For each bait TF its interactors were compared to known yeast protein complexes. The overlap with each protein complex was calculated and the resulting matrix filtered for TFs (y-axis) to have at least one complex they interact with at an overlap of more than 20%. Size of the dots represent the overlap with all proteins from the complex. Red dots show previously known interactions with the complex and blue novel interactions. **b,** Proteins identified as novel interactors were compared to yeast protein complexes. If the novel interactor is part of a complex, this was counted as an interaction with the respective complex (# of interaction with complex). All interactions to a given complex were summed and the number of interactions plotted on the y-axis against the respective complex (x-axis).

Figure 5: Correlation matrix of enrichments of all TFs. a, Pearson correlation matrix of enrichments of all significant interactors across all baits. Correlation values are colored from 0 (white) to 1 (dark red). **b,** LFQ intensity profile (log2) of general TF complexes and chromatin remodeling complexes across the broad chromatin binding TFs Abf1, Cbf1, Nhp6b, Reb1, Spt15, Sua7 and the control strain. **c,** LFQ intensity profile (log2) of Rpd3L complex members across the TFs Ash1, Gis1, Stb3, Rph1, Sfl1 and the control strain. **d,** Volcano plots of Sok2 and Sfl1 showing the similar interactome of the TFs Sok2 (left) and Sfl1 (right). Dashed lines represent the FDR cutoffs of 10% (inner line), 5% (middle line) and 1% (outer line). The bait is shown in red and the interacting TF in orange. **e.** Scatterplot comparing enrichments in the pulldown of Ste12 and Tec1 (left) and Ste12 and Kar4 (right). The baits are shown in red and blue, respectively. The t-test enrichment (log2) of proteins by Tec1 or Kar4 is shown on the y-axis, respectively, and by Ste12 on the x-axis.

Figure 6: Global correlation network of proteins and comparison of ChIP-MS and conventional affinity-enrichment MS. a, Pearson correlation network of LFQ intensities of all

25

proteins in all bait pulldowns. Correlations of less than 0.65 were removed. Edge thickness represents the strength of the correlation. Proteins of unknown function are highlighted in orange. **b,** Pearson correlation of LFQ intensities of RNA Pol I Core factor complex members (Rrn6, Rrn7, Rrn11), RNA Pol I upstream activating factor complex members (Rrn5, Rrn9, Rrn10, Uaf30) and proteins associated with both complexes (Net1, Spt15, Ydr249C) in ChIP-MS pulldowns of the respective proteins or control strains. Pearson correlation values are shown below the preys on the right. **c,** Comparison of a ChIP-MS pulldown of the RNA Pol I Core factor complex member Rrn7 and a conventional affinity-enrichment pulldown of Rrn7. The bait is highlighted in red and the protein of unknown function Ydr249c in orange.

Supplemental Figure 1: Characterization of baits and interactors. a, Yeast proteins ranked by their average copy number as calculated in (Kulak et al., 2014). Baits used in this study are highlighted in red. **b,** Venn diagram showing the overlap of interactors found in this study and affinity capture data mapped on Biogrid. GO enrichment analysis data of GO-Molecular Function and Biological Processes was performed for previously unknown interactors. Dot size represents the number of proteins and the x-axis shows the percentage of all novel interactors made up from the respective GO term. p-values were adjusted by Benjamini-Hochberg multiple t-test correction.

Supplemental Figure 2: TF-TF interactions observed by ChIP-MS. a-c, Volcano plots of previously described TF-TF interactions Stp1-Dal81, Stp2-Dal81, Tec1-Ste12 and Rtg1-Rtg3. The bait is highlighted in red and the interacting paralogous protein in orange. **d,** Sequence alignment of the paralogous TFs Hal9 and Tbs1.

Supplemental Figure 3: Volcano plots of co-regulatory interactions between paralogous transcription factors. The interacting factors are Spt23-Mga2, Sok2-Phd1, Rsc3-Rsc30, Msn2-

Msn4, Gis1-Rph1 and Cst6-Aca1. The bait is highlighted in red and the interacting paralogous protein in orange.

Supplemental Figure 4: Novel interactions with yeast protein complexes observed by ChIP-MS. a, Number of previously unknown interactions with proteins from specified yeast protein complexes. **b,** Volcano plot of Stb3 (red). Members of the Rpd3 histone deacetylase complex are highlighted in orange. **c,** Volcano plot of Gis1 (red). Members of the Rpd3 histone deacetylase complex are highlighted in orange.

Supplemental Figure 5: Heatmap of correlations between all proteins significantly enriched with at least one bait. Correlation values are colored from dark blue (correlation -1) to dark red (correlation 1).

Supplemental Figure 6: Volcano plots of the RNA Pol I activator complex members Rrn5, Rrn6, Rrn10, Rrn11 and Rrn3. The bait is highlighted in red and the protein of unknown function Ydr249C is highlighted in orange.

References

Allen, B.L., and Taatjes, D.J. (2015). The Mediator complex: a central integrator of transcription. Nat Rev Mol Cell Bio *16*, 155-166.

Bardwell, L., Cook, J.G., Voora, D., Baggott, D.M., Martinez, A.R., and Thorner, J. (1998). Repression of yeast Ste12 transcription factor by direct binding of unphosphorylated Kss1 MAPK and its regulation by the Ste7 MEK. Genes Dev *12*, 2887-2898.

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.

Boban, M., and Ljungdahl, P.O. (2007). Dal81 enhances Stp1- and Stp2-dependent transcription necessitating negative modulation by inner nuclear membrane protein Asi1 in Saccharomyces cerevisiae. Genetics *176*, 2087-2097.

Bornstein, C., Winter, D., Barnett-Itzhaki, Z., David, E., Kadri, S., Garber, M., and Amit, I. (2014). A Negative Feedback Loop of Transcription Factors Specifies Alternative Dendritic Cell Chromatin States. Molecular Cell *56*, 749-762.

Bumgarner, S.L., Dowell, R.D., Grisafi, P., Gifford, D.K., and Fink, G.R. (2009). Toggle involving cis-interfering noncoding RNAs controls variegated gene expression in yeast. P Natl Acad Sci USA *106*, 18321-18326.

Chou, S., Lane, S., and Liu, H. (2006). Regulation of mating and filamentation genes by two distinct Ste12 complexes in Saccharomyces cerevisiae. Mol Cell Biol *26*, 4794-4805.

Colin, J., Candelli, T., Porrua, O., Boulay, J., Zhu, C., Lacroute, F., Steinmetz, L.M., and Libri, D. (2014). Roadblock termination by reb1p restricts cryptic and readthrough transcription. Mol Cell *56*, 667-680.

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.

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.

Costanzo, M., VanderSluis, B., Koch, E.N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S.D.*, et al.* (2016). A global genetic interaction network maps a wiring diagram of cellular function. Science *353*.

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.

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.

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.

D'Alessio, A.C., Fan, Z.P., Wert, K.J., Baranov, P., Cohen, M.A., Saini, J.S., Cohick, E., Charniga, C., Dadon, D., Hannett, N.M.*, et al.* (2015). A Systematic Approach to Identify Candidate Transcription Factors that Control Cell Identity. Stem Cell Rep *5*, 763-775.

de Boer, C.G., and Hughes, T.R. (2012). YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. Nucleic Acids Res *40*, D169-179.

Dhalluin, C., Carlson, J.E., Zeng, L., He, C., Aggarwal, A.K., and Zhou, M.M. (1999). Structure and ligand of a histone acetyltransferase bromodomain. Nature *399*, 491-496.

Doyon, Y., Selleck, W., Lane, W.S., Tan, S., and Cote, J. (2004). Structural and functional conservation of the NuA4 histone acetyltransferase complex from yeast to humans. Molecular and Cellular Biology *24*, 1884-1896.

Dufourt, J., Trullo, A., Hunter, J., Fernandez, C., Lazaro, J., Dejean, M., Morales, L., Nait-Amer, S., Schulz, K.N., Harrison, M.M.*, et al.* (2018). Temporal control of gene expression by the pioneer factor Zelda through transient interactions in hubs. Nature Communications *9*.

Engel, C., Gubbey, T., Neyer, S., Sainsbury, S., Oberthuer, C., Baejen, C., Bernecky, C., and Cramer, P. (2017). Structural Basis of RNA Polymerase I Transcription Initiation. Cell *169*, 120-131 e122.

Engelen, E., Brandsma, J.H., Moen, M.J., Signorile, L., Dekkers, D.H., Demmers, J., Kockx, C.E., Ozgur, Z., van, I.W.F., van den Berg, D.L.*, et al.* (2015). Proteins that bind regulatory regions identified by histone modification chromatin immunoprecipitations and mass spectrometry. Nat Commun *6*, 7155.

Fields, S., and Song, O. (1989). A novel genetic system to detect protein-protein interactions. Nature *340*, 245-246.

Fuda, N.J., Ardehali, M.B., and Lis, J.T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. Nature *461*, 186-192.

Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B.*, et al.* (2006). Proteome survey reveals modularity of the yeast cell machinery. Nature *440*, 631-636.

Gavin, A.C., Bosche, 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.

Gregoretti, I.V., Lee, Y.M., and Goodson, H.V. (2004). Molecular evolution of the histone deacetylase family: Functional implications of phylogenetic analysis. Journal of Molecular Biology *338*, 17-31.

Hahn, J.S., Hu, Z.Z., Thiele, D.J., and Iyer, V.R. (2004). Genome-wide analysis of the biology of stress responses through heat shock transcription factor. Molecular and Cellular Biology *24*, 5249-5256.

Hein, M.Y., Hubner, N.C., Poser, I., Cox, J., Nagaraj, N., Toyoda, Y., Gak, I.A., Weisswange, I., Mansfeld, J., Buchholz, F.*, et al.* (2015). A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell *163*, 712-723.

Hemmer, M.C., Wierer, M., Schachtrup, K., Downes, M., Hubner, N., Evans, R.M., and Uhlenhaut, N.H. (2019). E47 modulates hepatic glucocorticoid action. Nat Commun *10*, 306.

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

Hosp, F., Scheltema, R.A., Eberl, H.C., Kulak, N.A., Keilhauer, E.C., Mayr, K., and Mann, M. (2015). A Double-Barrel Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS) System to Quantify 96 Interactomes per Day. Mol Cell Proteomics *14*, 2030-2041.

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.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A *98*, 4569-4574.

Jia, Y.K., Rothermel, B., Thornton, J., and Butow, R.A. (1997). A basic helix-loop-helix-leucine zipper transcription complex in yeast functions in a signaling pathway from mitochondria to the nucleus. Molecular and Cellular Biology *17*, 1110-1117.

Jolma, A., Yin, Y.M., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature *527*, 384-+.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S.*, et al.* (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430-435.

Keilhauer, E.C., Hein, M.Y., and Mann, M. (2015). Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS). Mol Cell Proteomics *14*, 120-135.

Kent, N.A., Eibert, S.M., and Mellor, J. (2004). Cbf1p is required for chromatin remodeling at promoter-proximal CACGTG motifs in yeast. J Biol Chem *279*, 27116-27123.

Kim, S., Brostromer, E., Xing, D., Jin, J., Chong, S., Ge, H., Wang, S., Gu, C., Yang, L., Gao, Y.Q.*, et al.* (2013). Probing allostery through DNA. Science *339*, 816-819.

Koerber, R.T., Rhee, H.S., Jiang, C., and Pugh, B.F. (2009). Interaction of transcriptional regulators with specific nucleosomes across the Saccharomyces genome. Mol Cell *35*, 889-902.

Kratsios, P., Stolfi, A., Levine, M., and Hobert, O. (2012). Coordinated regulation of cholinergic motor neuron traits through a conserved terminal selector gene. Nat Neurosci *15*, 205-214.

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

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.

Kurdistani, S.K., Tavazoie, S., and Grunstein, M. (2004). Mapping global histone acetylation patterns to gene expression. Cell *117*, 721-733.

Lahav, R., Gammie, A., Tavazoie, S., and Rose, M.D. (2007). Role of transcription factor Kar4 in regulating downstream events in the Saccharomyces cerevisiae pheromone response pathway. Mol Cell Biol *27*, 818-829.

Lardenois, A., Stuparevic, I., Liu, Y.C., Law, M.J., Becker, E., Smagulova, F., Waern, K., Guilleux, M.H., Horecka, J., Chu, A.*, et al.* (2015). The conserved histone deacetylase Rpd3 and its DNA binding subunit Ume6 control dynamic transcript architecture during mitotic growth and meiotic development. Nucleic acids research *43*, 115-128.

Lascaris, R.F., Groot, E., Hoen, P.B., Mager, W.H., and Planta, R.J. (2000). Different roles for abf1p and a T-rich promoter element in nucleosome organization of the yeast RPS28A gene. Nucleic Acids Res *28*, 1390-1396.

Lee, C., Etchegaray, J.P., Cagampang, F.R.A., Loudon, A.S.I., and Reppert, S.M. (2001). Posttranslational mechanisms regulate the mammalian circadian clock. Cell *107*, 855-867.

Lee, D.Y., Hayes, J.J., Pruss, D., and Wolffe, A.P. (1993). A Positive Role for Histone Acetylation in Transcription Factor Access to Nucleosomal DNA. Cell *72*, 73-84.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I.*, et al.* (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science *298*, 799-804.

Li, B., Carey, M., and Workman, J.L. (2007). The role of chromatin during transcription. Cell *128*, 707-719.

Liang, C.Y., Wang, L.C., and Lo, W.S. (2013). Dissociation of the H3K36 demethylase Rph1 from chromatin mediates derepression of environmental stress-response genes under genotoxic stress in Saccharomyces cerevisiae. Mol Biol Cell *24*, 3251-3262.

Lorenz, M.C., and Heitman, J. (1998). Regulators of pseudohyphal differentiation in Saccharomyces cerevisiae identified through multicopy suppressor analysis in ammonium permease mutant strains. Genetics *150*, 1443-1457.

Malik, S., and Roeder, R.G. (2010). The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation. Nature Reviews Genetics *11*, 761-772.

Meldal, B.H.M., Bye, A.J.H., Gajdos, L., Hammerova, Z., Horackova, A., Melicher, F., Perfetto, L., Pokorny, D., Lopez, M.R., Turkova, A.*, et al.* (2019). Complex Portal 2018: extended content and enhanced visualization tools for macromolecular complexes. Nucleic Acids Res *47*, D550-D558.

Mendizabal, I., Rios, G., Mulet, J.M., Serrano, R., and de Larrinoa, I.F. (1998). Yeast putative transcription factors involved in salt tolerance. Febs Lett *425*, 323-328.

Meng, J.L., Wang, Y.P., Carrillo, R.A., and Heckscher, E.S. (2020). Temporal transcription factors determine circuit membership by permanently altering motor neuron-to-muscle synaptic partnerships. Elife *9*.

Mivelaz, M., Cao, A.M., Kubik, S., Zencir, S., Hovius, R., Boichenko, I., Stachowicz, A.M., Kurat, C.F., Shore, D., and Fierz, B. (2020). Chromatin Fiber Invasion and Nucleosome Displacement by the Rap1 Transcription Factor. Molecular Cell *77*, 488-+.

Moreira, J.M., and Holmberg, S. (2000). Chromatin-mediated transcriptional regulation by the yeast architectural factors NHP6A and NHP6B. EMBO J *19*, 6804-6813.

Narasimhan, K., Pillay, S., Huang, Y.H., Jayabal, S., Udayasuryan, B., Veerapandian, V., Kolatkar, P., Cojocaru, V., Pervushin, K., and Jauch, R. (2015). DNA-mediated cooperativity facilitates the co-selection of cryptic enhancer sequences by SOX2 and PAX6 transcription factors. Nucleic Acids Res *43*, 1513-1528.

Natoli, G. (2010). Maintaining Cell Identity through Global Control of Genomic Organization. Immunity *33*, 12-24.

Ngondo, R.P., and Carbon, P. (2014). Transcription factor abundance controlled by an auto-regulatory mechanism involving a transcription start site switch. Nucleic acids research *42*, 2171-2184.

Owen, D.J., Ornaghi, P., Yang, J.C., Lowe, N., Evans, P.R., Ballario, P., Neuhaus, D., Filetici, P., and Travers, A.A. (2000). The structural basis for the recognition of acetylated histone H4 by the bromodomain of histone acetyltransferase Gcn5p. Embo Journal *19*, 6141-6149.

Panman, L., Andersson, E., Alekseenko, Z., Hedlund, E., Kee, N., Mong, J., Uhde, C.W., Deng, Q.L., Sandberg, R., Stanton, L.W.*, et al.* (2011). Transcription Factor-Induced Lineage Selection of Stem-Cell-Derived Neural Progenitor Cells. Cell Stem Cell *8*, 663-675.

Panne, D. (2008). The enhanceosome. Curr Opin Struct Biol *18*, 236-242.

Paul, F.E., Hosp, F., and Selbach, M. (2011). Analyzing protein-protein interactions by quantitative mass spectrometry. Methods *54*, 387-395.

Rafiee, M.R., Girardot, C., Sigismondo, G., and Krijgsveld, J. (2016). Expanding the Circuitry of Pluripotency by Selective Isolation of Chromatin-Associated Proteins. Mol Cell *64*, 624-635.

Raitt, D.C., Johnson, A.L., Erkine, A.M., Makino, K., Morgan, B., Gross, D.S., and Johnston, L.H. (2000). The Skn7 response regulator of Saccharomyces cerevisiae interacts with Hsf1 in vivo and is required for the induction of heat shock genes by oxidative stress. Mol Biol Cell *11*, 2335-2347.

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.

Rappsilber, J., Mann, M., and Ishihama, Y. (2007). Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat Protoc *2*, 1896-1906.

Reddy, B.A., Bajpe, P.K., Bassett, A., Moshkin, Y.M., Kozhevnikova, E., Bezstarosti, K., Demmers, J.A.A., Travers, A.A., and Verrijzer, C.P. (2010). Drosophila Transcription Factor Tramtrack69 Binds MEP1 To Recruit the Chromatin Remodeler NuRD. Molecular and Cellular Biology *30*, 5234-5244.

Reiter, F., Wienerroither, S., and Stark, A. (2017). Combinatorial function of transcription factors and cofactors. Curr Opin Genet Dev *43*, 73-81.

Sardina, J.L., Collombet, S., Tian, T.V., Gomez, A., Di Stefano, B., Berenguer, C., Brumbaugh, J., Stadhouders, R., Segura-Morales, C., Gut, M.*, et al.* (2018). Transcription Factors Drive Tet2-Mediated Enhancer Demethylation to Reprogram Cell Fate. Cell Stem Cell *23*, 727-+.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H.J.*, et al.* (2011). Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell *147*, 1270-1282.

Soutourina, J. (2018). Transcription regulation by the Mediator complex. Nat Rev Mol Cell Biol *19*, 262-274.

Srivastava, R., Rai, K.M., Pandey, B., Singh, S.P., and Sawant, S.V. (2015). Spt-Ada-Gcn5-Acetyltransferase (SAGA) Complex in Plants: Genome Wide Identification, Evolutionary Conservation and Functional Determination. Plos One *10*.

Stampfel, G., Kazmar, T., Frank, O., Wienerroither, S., Reiter, F., and Stark, A. (2015). Transcriptional regulators form diverse groups with context-dependent regulatory functions. Nature *528*, 147-+.

Tam, J., and van Werven, F.J. (2020). Regulated repression governs the cell fate promoter controlling yeast meiosis. Nat Commun *11*, 2271.

Tardiff, D.F., Abruzzi, K.C., and Rosbash, M. (2007). Protein characterization of Saccharomyces cerevisiae RNA polymerase II after in vivo cross-linking. Proc Natl Acad Sci U S A *104*, 19948-19953.

Thambyrajah, R., Mazan, M., Patel, R., Moignard, V., Stefanska, M., Marinopoulou, E., Li, Y.Y., Lancrin, C., Clapes, T., Moroy, T.*, et al.* (2016). GFI1 proteins orchestrate the emergence of haematopoietic stem cells through recruitment of LSD1. Nat Cell Biol *18*, 21-+.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P.*, et al.* (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature *403*, 623-627.

Uyehara, C.M., Nystrom, S.L., Niederhuber, M.J., Leatham-Jensen, M., Ma, Y.Q., Buttitta, L.A., and McKay, D.J. (2017). Hormone-dependent control of developmental timing through regulation of chromatin accessibility. Gene Dev *31*, 862-875.

Vierbuchen, T., Ling, E., Cowley, C.J., Couch, C.H., Wang, X.F., Harmin, D.A., Roberts, C.W.M., and Greenberg, M.E. (2017). AP-1 Transcription Factors and the BAF Complex Mediate Signal-Dependent Enhancer Selection. Molecular Cell *68*, 1067-+.

Wang, C.I., Alekseyenko, A.A., LeRoy, G., Elia, A.E., Gorchakov, A.A., Britton, L.M., Elledge, S.J., Kharchenko, P.V., Garcia, B.A., and Kuroda, M.I. (2013). Chromatin proteins captured by ChIP-mass spectrometry are linked to dosage compensation in Drosophila. Nat Struct Mol Biol *20*, 202-209.

Wang, Z.B., Zang, C.Z., Cui, K.R., Schones, D.E., Barski, A., Peng, W.Q., and Zhao, K.J. (2009). Genome-wide Mapping of HATs and HDACs Reveals Distinct Functions in Active and Inactive Genes. Cell *138*, 1019-1031.

Xie, D., Boyle, A.P., Wu, L.F., Zhai, J., Kawli, T., and Snyder, M. (2013). Dynamic trans-Acting Factor Colocalization in Human Cells. Cell *155*, 713-724.

Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284-287.

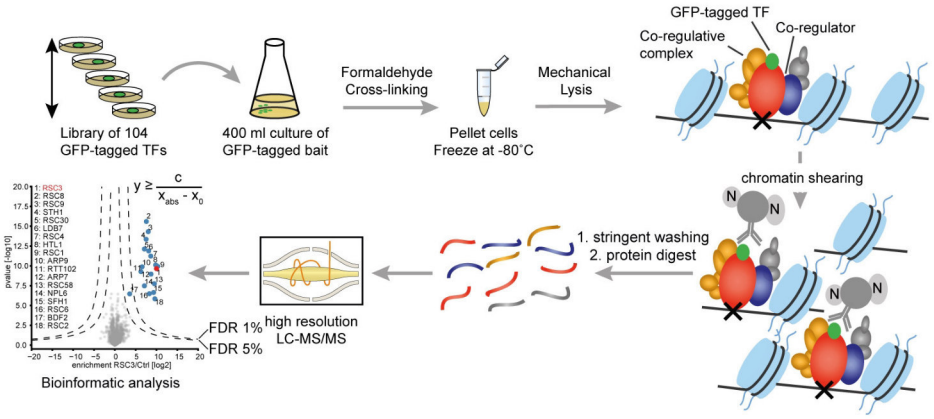

Library of 104
GFP-tagged TFs
400 ml culture of
GFP-tagged bait
Formaldehyde
Cross-linking
Pellet cells
Freeze at -80°C
Mechanical
Lysis
GFP-tagged TF
Co-regulative
complex
Co-regulator
chromatin shearing
1. stringent washing
2. protein digest
high resolution
LC-MS/MS
FDR 1%
FDR 5%
Bioinformatic analysis

Figure 1

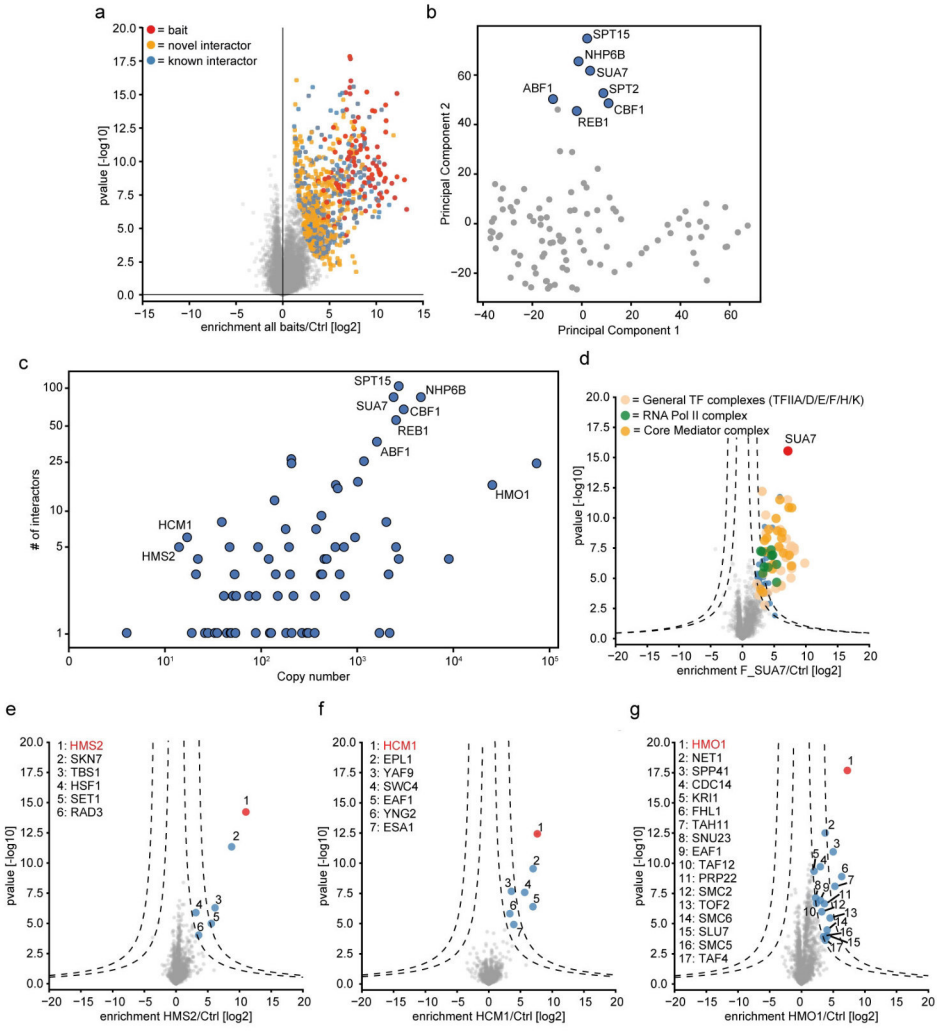
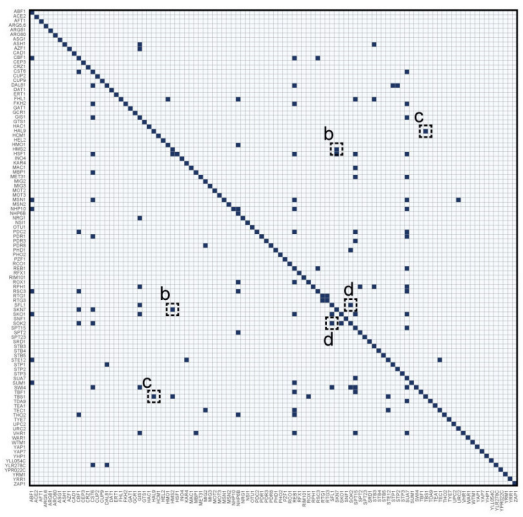

a
= bait
= novel interactor
= known interactor
pvalue [-log10]
enrichment all baits/Ctrl [log2]
b
SPT15
NHP6B
SUA7
SPT2
CBF1
REB1
ABF1
Principal Component 2
Principal Component 1
c
SPT15
NHP6B
SUA7
CBF1
REB1
ABF1
HMO1
HCM1
HMS2
# of interactors
Copy number
d
= General TF complexes (TFIIA/D/E/F/H/K)
= RNA Pol II complex
= Core Mediator complex
SUA7
pvalue [-log10]
enrichment F_SUA7/Ctrl [log2]
e
1: HMS2
2: SKN7
3: TBS1
4: HSF1
5: SET1
6: RAD3
pvalue [-log10]
enrichment HMS2/Ctrl [log2]
f
1: HCM1
2: EPL1
3: YAF9
4: SWC4
5: EAF1
6: YNG2
7: ESA1
pvalue [-log10]
enrichment HCM1/Ctrl [log2]
g
1: HMO1
2: NET1
3: SPP41
4: CDC14
5: KRI1
6: FHL1
7: TAH11
8: SNU23
9: EAF1
10: TAF12
11: PRP22
12: SMC2
13: TOF2
14: SMC6
15: SLU7
16: SMC5
17: TAF4
pvalue [-log10]
enrichment HMO1/Ctrl [log2]

Figure 2

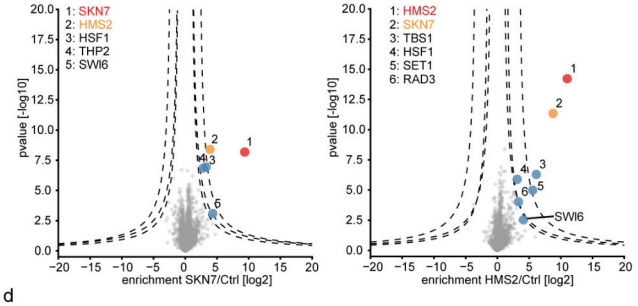
a



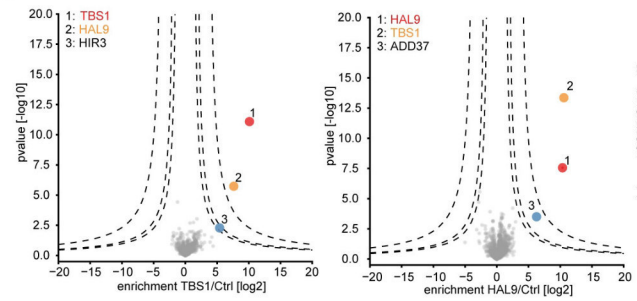
b

Bait TF	Interacting TFs
Abf1	Cbf1, Msn1, Nhp10, Rsc3, Sko1, Ste12, Sum1
Cbf1	Cst6, Pdc2, Rsc3, Skn7, Sok2, Tho2
Cst6	Dal81, Fkh2, Gis1, Hsf1, Mbp1, Msn1, Pdr1, Skn7, Sok2, Swi4, YLR278C
Dal81	Stp1, YLR278C
Gis1	Ash1, Azf1, Nrg1, Rph1, Sfl1, Sko1, Swi4, Vhr1
Hal9	Tbs1
Hmo1	Fhl1
Hms2	Hsf1, Skn7, Tbs1
Kar4	Ste12
Mbp1	Vhr1
Mig2	Pdr8, Tec1
Nhp6b	Fhl1, Hmo1, Nhp10, Rox1, Rsc3, Spt2, Tho2
Reb1	Cbf1, Dal81, Fkh2, Hsf1, Msn1, Nhp10, Pdc2, Rox1, Sko1, Sum1, Tbf1, Tec1, Tho2, Vhr1
Rox1	Tbs1
Rsc3	Cbf1, Reb1
Rtg1	Rtg3
Rtg3	Rtg1
Sfl1	Sko1, Sok2, Swi4
Skn7	Hms2, Hsf1
Sko1	Sok2
Sok2	Phd1, Sfl1, Swi4
Spt15	Fhl1, Hsf1, Mac1, Met31, Pdc2, Pdr3, Sko1, Sua7, Swi4, Tbf1, Vhr1
Spt2	Rph1
Stb3	Ash1, Fhl1, Rph1
Ste12	Tbs1, Tec1
Stp1	Dal81
Stp2	Dal81
Sua7	Cst6, Fkh2, Gis1, Hsf1, Met31, Msn1, Nhp10, Pdc2, Pdr1, Reb1, Rph1, Rsc3, Sko1, Spt15, Swi4, Tda9, Vhr1
Tbs1	Hal9
Tec1	Ste12
Urc2	Msn1

c



d



e

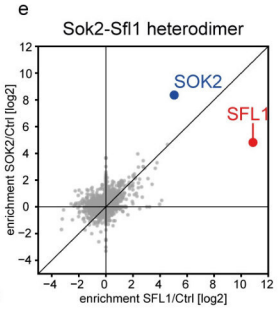


Figure 3

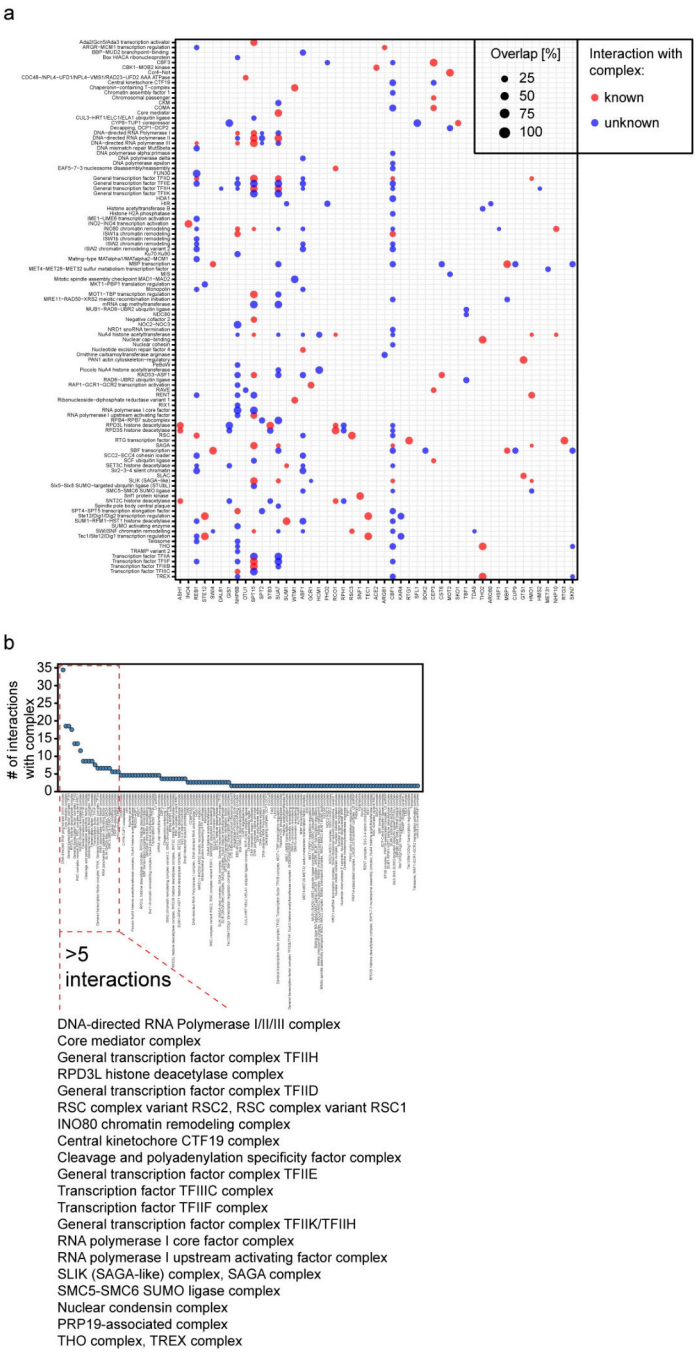

a
Overlap [%]
25
50
75
100
Interaction with complex:
known
unknown
b
# of interactions with complex
35
30
25
20
15
10
5
0
>5 interactions
DNA-directed RNA Polymerase I/II/III complex
Core mediator complex
General transcription factor complex TFIIH
RPD3L histone deacetylase complex
General transcription factor complex TFIID
RSC complex variant RSC2, RSC complex variant RSC1
INO80 chromatin remodeling complex
Central kinetochore CTF19 complex
Cleavage and polyadenylation specificity factor complex
General transcription factor complex TFIIE
Transcription factor TFIIIC complex
Transcription factor TFIIF complex
General transcription factor complex TFIIK/TFIIH
RNA polymerase I core factor complex
RNA polymerase I upstream activating factor complex
SLIK (SAGA-like) complex, SAGA complex
SMC5-SMC6 SUMO ligase complex
Nuclear condensin complex
PRP19-associated complex
THO complex, TREX complex

Figure 4

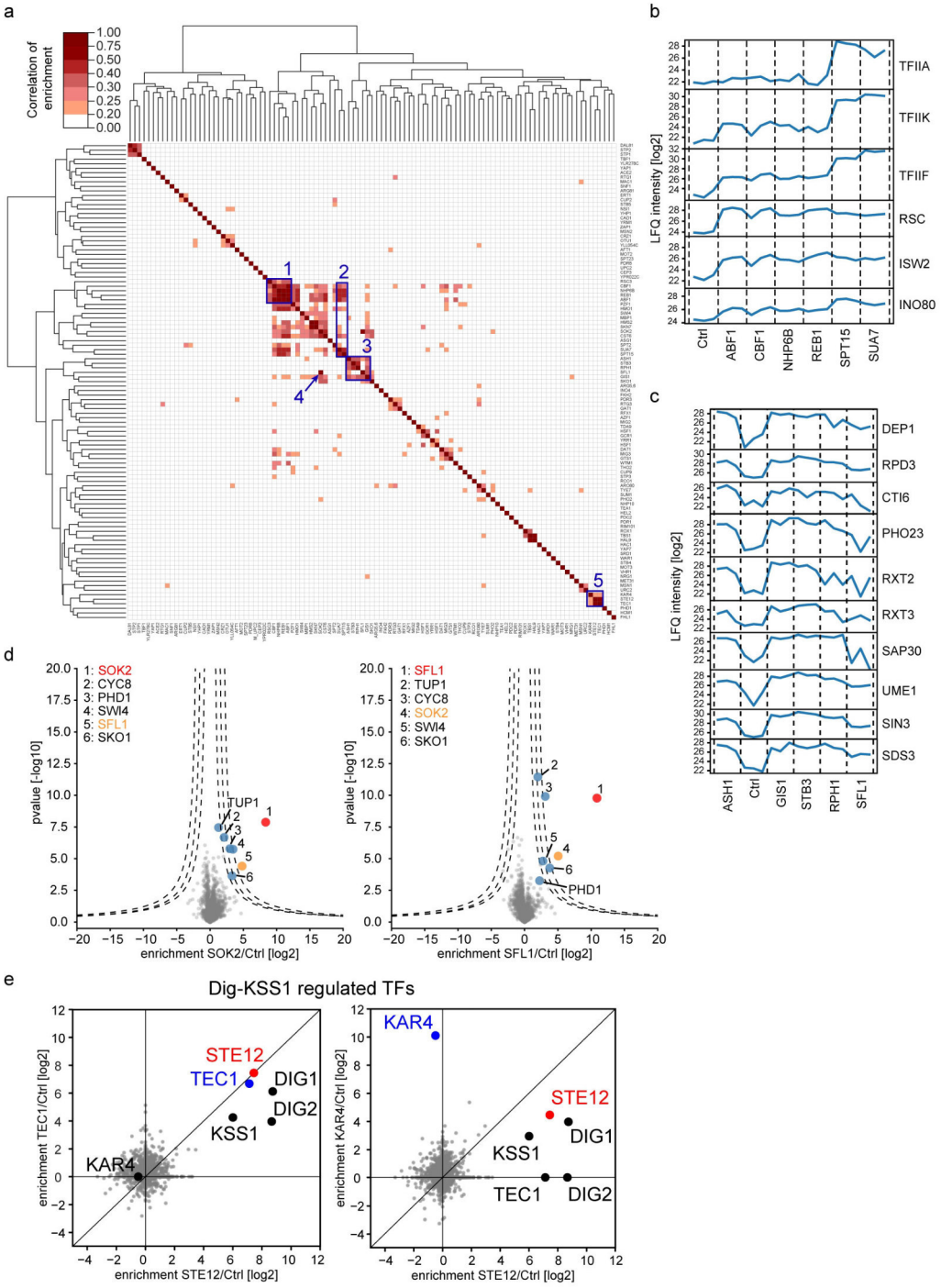

a
Correlation of enrichment
1.00
0.75
0.50
0.40
0.30
0.25
0.20
0.00
1
2
3
4
5
b
LFQ intensity [log2]
TFIIA
TFIIK
TFIIF
RSC
ISW2
INO80
Ctrl
ABF1
CBF1
NHP6B
REB1
SPT15
SUA7
c
LFQ intensity [log2]
DEP1
RPD3
CTI6
PHO23
RXT2
RXT3
SAP30
UME1
SIN3
SDS3
ASH1
Ctrl
GIS1
STB3
RPH1
SFL1
d
1: SOK2
2: CYC8
3: PHD1
4: SWI4
5: SFL1
6: SKO1
pvalue [-log10]
enrichment SOK2/Ctrl [log2]
TUP1
1: SFL1
2: TUP1
3: CYC8
4: SOK2
5: SWI4
6: SKO1
enrichment SFL1/Ctrl [log2]
PHD1
e
Dig-KSS1 regulated TFs
enrichment TEC1/Ctrl [log2]
enrichment STE12/Ctrl [log2]
STE12
TEC1
DIG1
DIG2
KSS1
KAR4
enrichment KAR4/Ctrl [log2]
enrichment STE12/Ctrl [log2]
KAR4
STE12
DIG1
KSS1
TEC1
DIG2

Figure 5

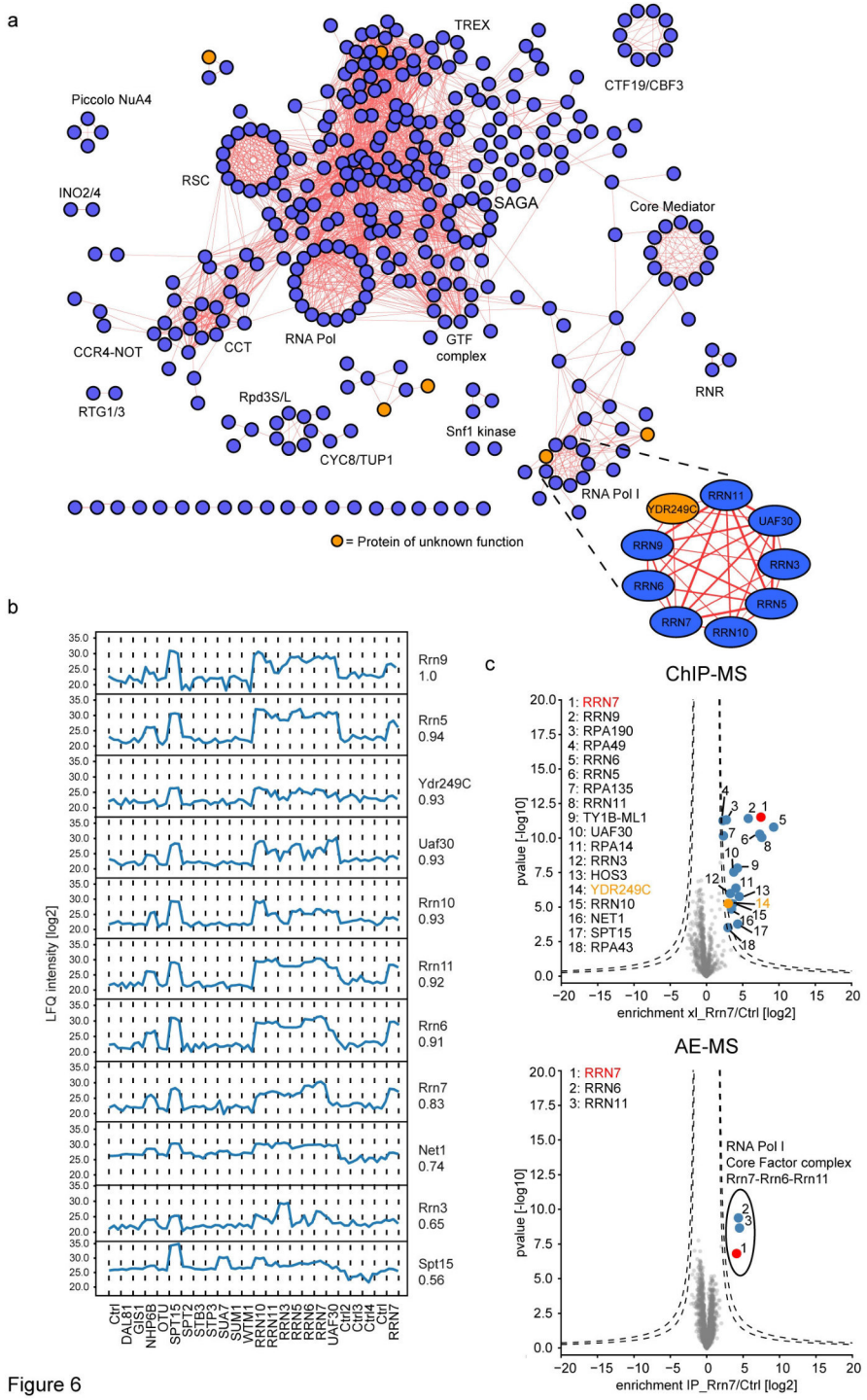
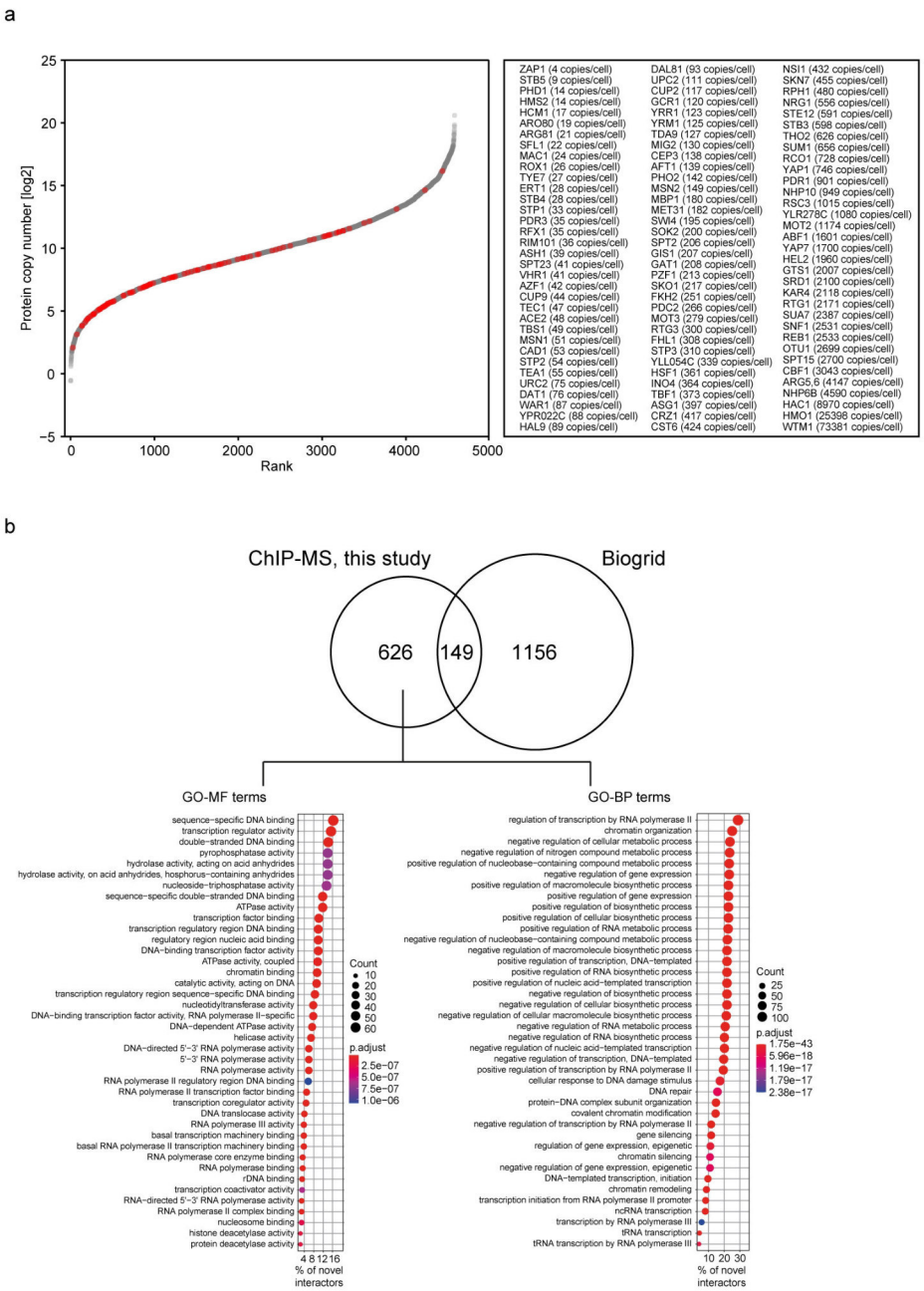
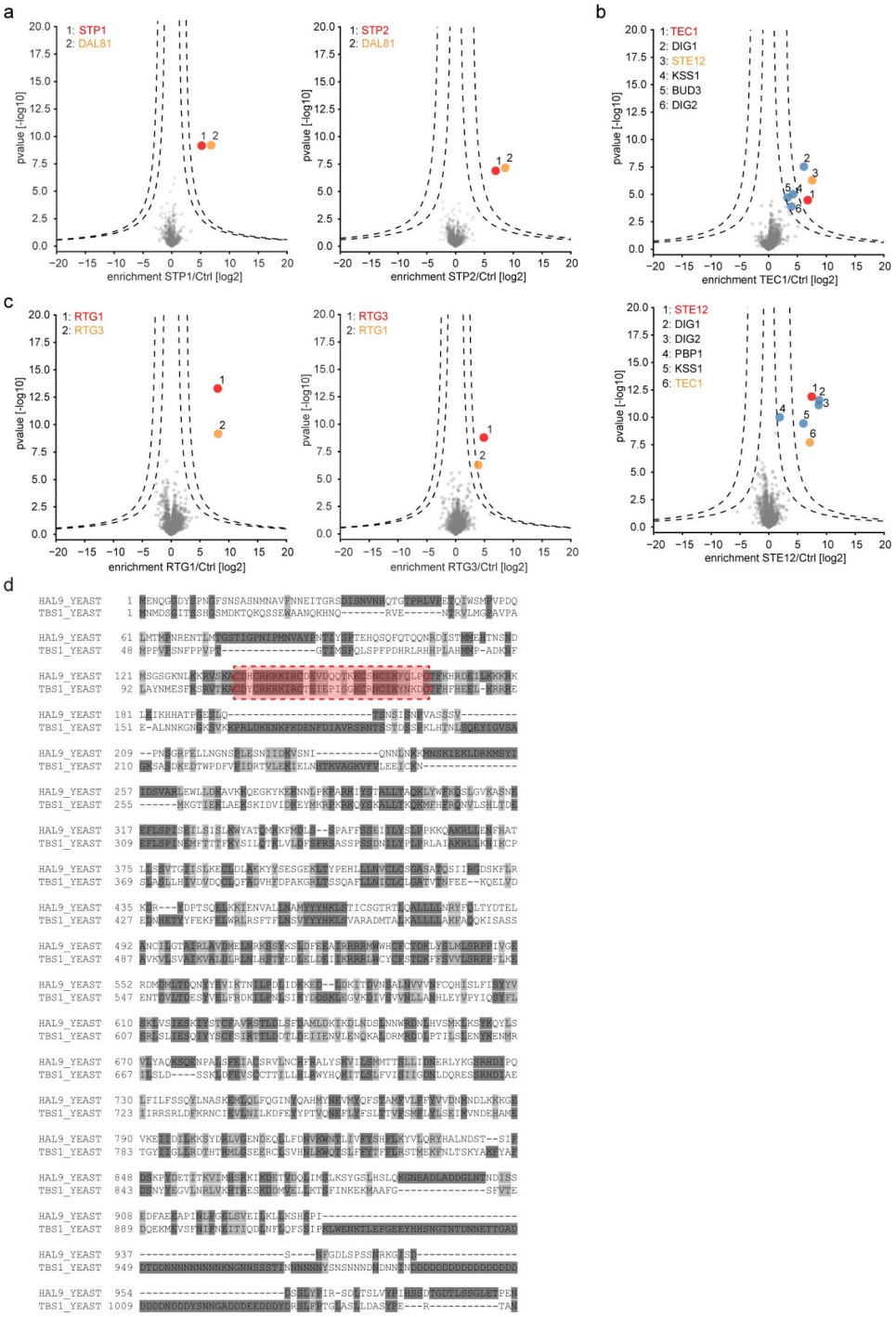

a
TREX
CTF19/CBF3
Piccolo NuA4
RSC
INO2/4
SAGA
Core Mediator
CCR4-NOT
CCT
RNA Pol
GTF complex
RNR
RTG1/3
Rpd3S/L
Snf1 kinase
CYC8/TUP1
RNA Pol I
= Protein of unknown function
RRN11
YDR249C
UAF30
RRN9
RRN3
RRN6
RRN5
RRN7
RRN10
b
LFQ intensity [log2]
Rrn9 1.0
Rrn5 0.94
Ydr249C 0.93
Uaf30 0.93
Rrn10 0.93
Rrn11 0.92
Rrn6 0.91
Rrn7 0.83
Net1 0.74
Rrn3 0.65
Spt15 0.56
Ctrl
DAL81
GIS1
NHP6B
OTU
SPT15
STB2
STB3
STP3
SUA7
SUM1
WTM1
RRN10
RRN11
RRN3
RRN5
RRN6
RRN7
UAF30
Ctrl2
Ctrl3
Ctrl4
Ctrl
RRN7
c
ChIP-MS
1: RRN7
2: RRN9
3: RPA190
4: RPA49
5: RRN6
6: RRN5
7: RPA135
8: RRN11
9: TY1B-ML1
10: UAF30
11: RPA14
12: RRN3
13: HOS3
14: YDR249C
15: RRN10
16: NET1
17: SPT15
18: RPA43
pvalue [-log10]
enrichment xl_Rrn7/Ctrl [log2]
AE-MS
1: RRN7
2: RRN6
3: RRN11
RNA Pol I
Core Factor complex
Rrn7-Rrn6-Rrn11
pvalue [-log10]
enrichment IP_Rrn7/Ctrl [log2]

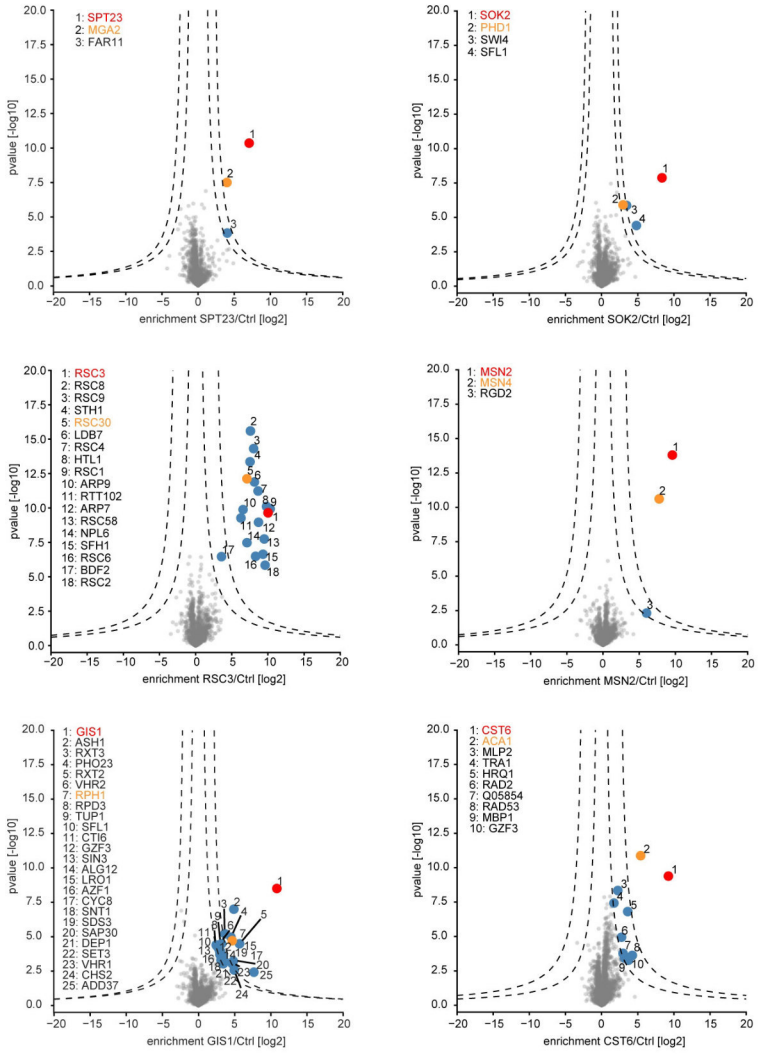
Figure 6



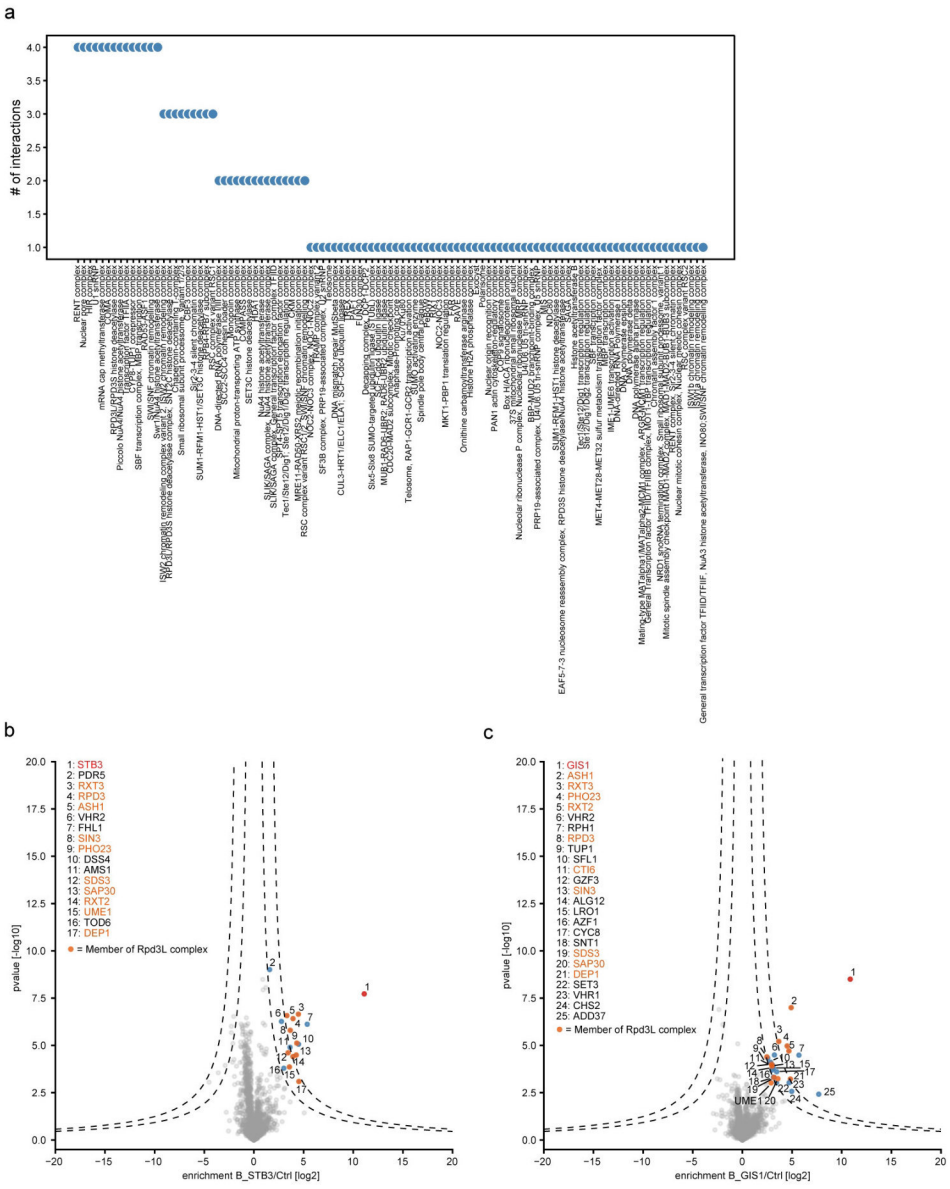
Supplemental Figure 1



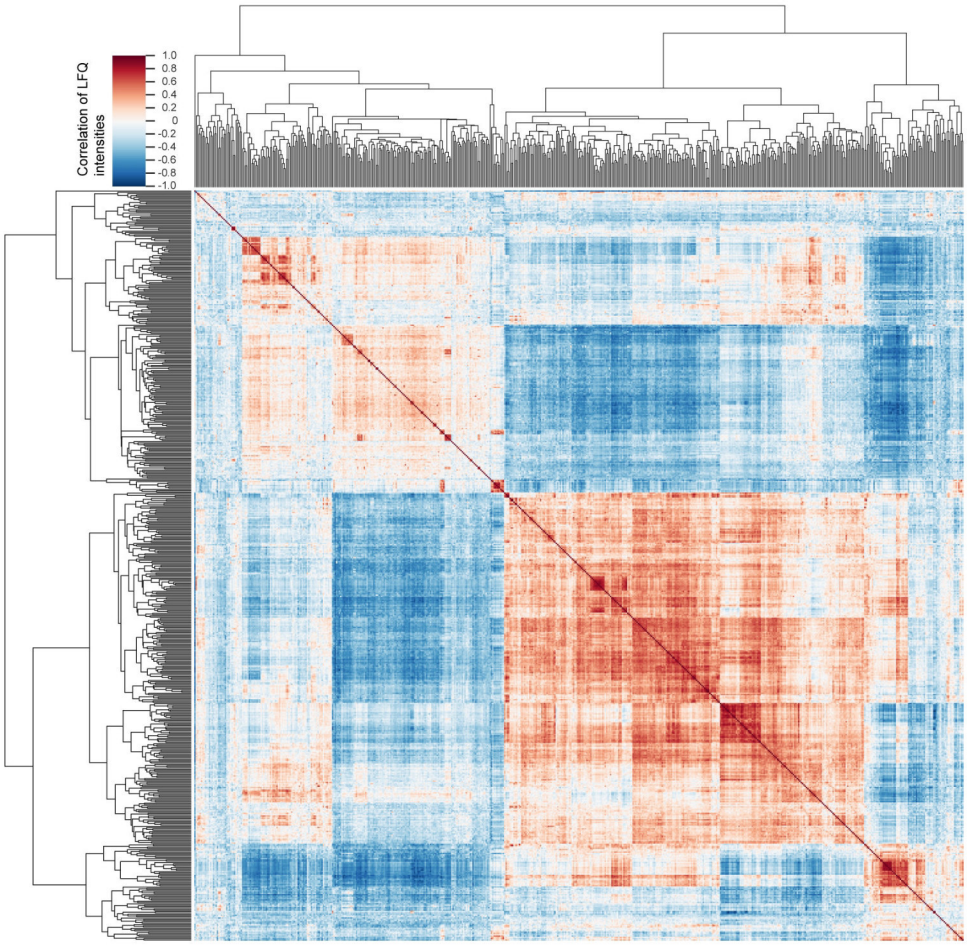
Supplemental Figure 2



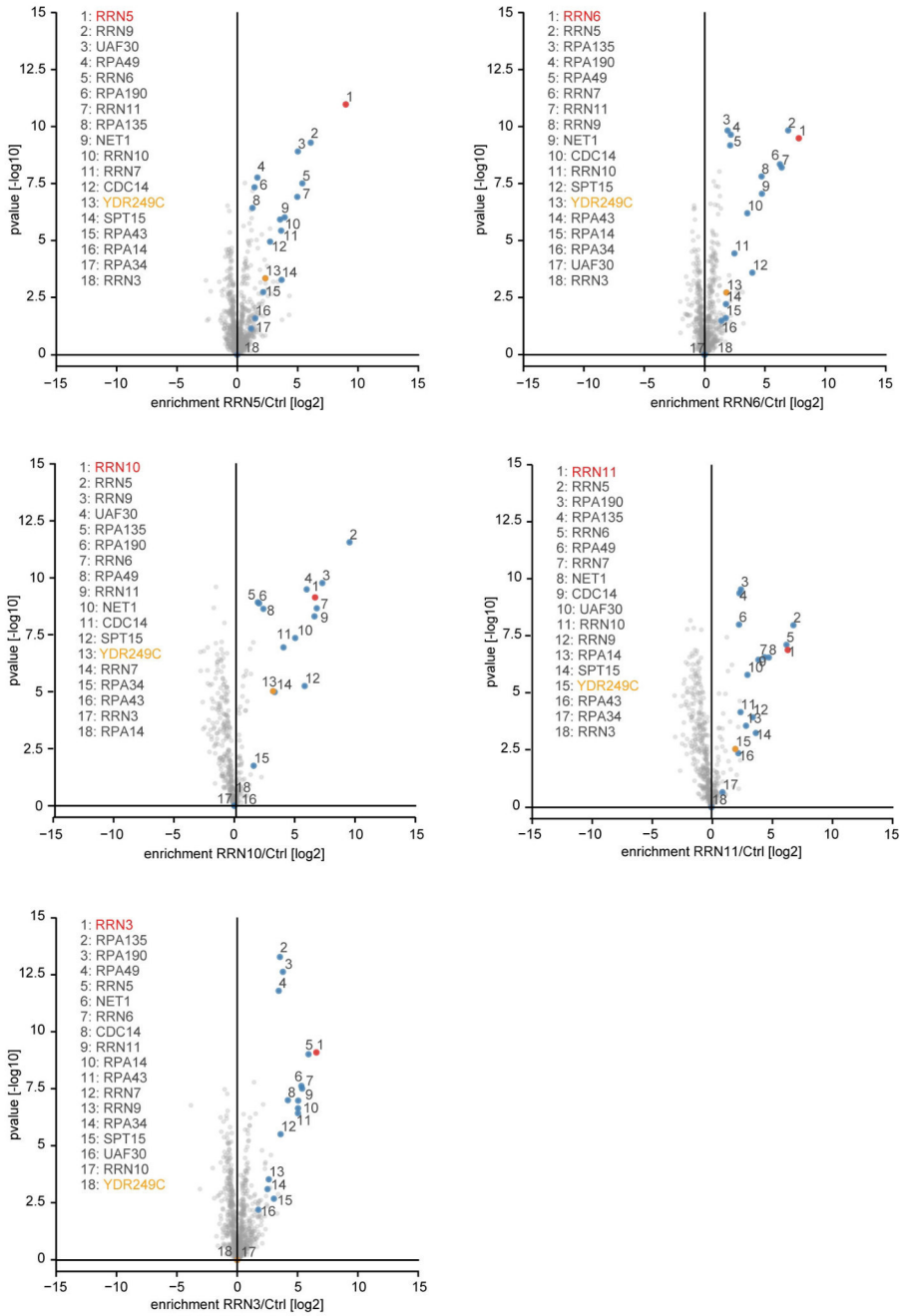
Supplemental Figure 3



Supplemental Figure 4



Supplemental Figure 5



Supplemental Figure 6

3.4 Article 4: *Drosophila* SWR1 and NuA4 complexes are defined by DOMINO isoforms

Alessandro Scacchetti, Tamas Schauer, **Alexander Reim**, Zivkos Apostolou, Aline Campos Sparr, Silke Krause, Patrick Heun, Michael Wierer & Peter B Becker. ***Drosophila* SWR1 and NuA4 complexes are defined by DOMINO isoforms.** *eLife* 2020;9:e56325 doi: 10.7554/eLife.56325

In this project I collaborated with Alessandro Scacchetti and Peter Becker from the Biomedical Center, Ludiwg-Maximilians University Munich, to understand whether the two Domino isoforms Domino A (DOM-A) and Domino B (DOM-B) exert distinctive functions in Drosophila. Initially, we asked whether we can detect differences in the interactome of both proteoforms, which would point towards individual molecular functions. Alessandro Scacchetti generated 3xFLAG-tagged Domino A or Domino B S2 cells. Together, we set up a stringent AP-MS protocol for endogenously FLAG-tagged DOM-A and DOM-B proteins in Drosophila S2 cells. We identified a complex of 7 proteins shared between both isoforms. However, we further identified 6 novel interactors specific for DOM-A and 5 specific interactors of DOM-B. Four of the interactors unique to DOM-A shared remarkable homology with the yeast NuA4 histone acetyltransferase complex. The other two interactors were transcription factors. In the case of DOM-B the 5 specific interactors were ARP6, PPS, HCF, PONT and REPT. In yeast ARP6 is crucial for H2A.Z remodeling. These findings led to the hypothesis that DOM-A and DOM-B are both involved in transcriptional regulation, but through different pathways. Further follow-up experiments on the individual interactions by Alessandro Scacchetti indeed showed that the DOM-A complex functions as an ATP-independent histone acetyltransferase complex, comparable to the yeast NuA4 complex. Conversely, the DOM-B complex incorporates H2A-V in the genome similarly to the yeast SWR1 complex.

eLife RESEARCH ARTICLE

Drosophila SWR1 and NuA4 complexes are defined by DOMINO isoforms

Alessandro Scacchetti[1], Tamas Schauer[2], Alexander Reim[3], Zivkos Apostolou[1], Aline Campos Sparr[1], Silke Krause[1], Patrick Heun[4], Michael Wierer[3], Peter B Becker[1]*

[1]Molecular Biology Division, Biomedical Center, Ludwig-Maximilians-University, Munich, Germany; [2]Bioinformatics Unit, Biomedical Center, Ludwig-Maximilians-University, Munich, Germany; [3]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Munich, Germany; [4]Wellcome Trust Centre for Cell Biology and Institute of Cell Biology, School of Biological Sciences, The University of Edinburgh, Edinburgh, United Kingdom

Abstract Histone acetylation and deposition of H2A.Z variant are integral aspects of active transcription. In *Drosophila*, the single DOMINO chromatin regulator complex is thought to combine both activities *via* an unknown mechanism. Here we show that alternative isoforms of the DOMINO nucleosome remodeling ATPase, DOM-A and DOM-B, directly specify two distinct multi-subunit complexes. Both complexes are necessary for transcriptional regulation but through different mechanisms. The DOM-B complex incorporates H2A.V (the fly ortholog of H2A.Z) genome-wide in an ATP-dependent manner, like the yeast SWR1 complex. The DOM-A complex, instead, functions as an ATP-independent histone acetyltransferase complex similar to the yeast NuA4, targeting lysine 12 of histone H4. Our work provides an instructive example of how different evolutionary strategies lead to similar functional separation. In yeast and humans, nucleosome remodeling and histone acetyltransferase complexes originate from gene duplication and paralog specification. *Drosophila* generates the same diversity by alternative splicing of a single gene.

***For correspondence:** pbecker@bmc.med.lmu.de

Competing interests: The authors declare that no competing interests exist.

Funding: See page 19

Received: 24 February 2020
Accepted: 23 April 2020
Published: 20 May 2020

Reviewing editor: Jerry L Workman, Stowers Institute for Medical Research, United States

Introduction

Nucleosomes, the fundamental units of chromatin, are inherently stable and organized in polymeric fibers of variable compactness (**Baldi et al., 2018**; **Erdel and Rippe, 2018**). The dynamic properties of the fiber required for gene regulation are implemented by several broad principles. ATP-dependent nucleosome remodeling factors slide or evict nucleosomes (**Clapier et al., 2017**), chemical modifications of histones create new interaction surfaces (**Bowman and Poirier, 2015**; **Zhao and Garcia, 2015**) and histone variants furnish nucleosomes with special features (**Talbert and Henikoff, 2017**).

The very conserved H2A variant H2A.Z accounts for ~5–10% of the total H2A-type histone pool in vertebrates (**Redon et al., 2002**; **Thatcher and Gorovsky, 1994**) and flies (**Bonnet et al., 2019**). H2A.Z is primarily found at active promoters and enhancers, where it is thought to be important to regulate transcription initiation and early elongation (**Adam et al., 2001**; **Weber et al., 2010**; **Weber et al., 2014**). In *Saccharomyces cerevisiae*, H2A.Z is introduced into chromatin by the SWR1 complex (SWR1.C), a multi-subunit ATP-dependent chromatin remodeler with the INO80-type ATPase SWR1 at its core (**Mizuguchi et al., 2004**; **Ranjan et al., 2013**; **Wang et al., 2018a**; **Willhoft et al., 2018**; **Wu et al., 2005**). In humans, the two SWR1 orthologs EP400 and SRCAP may also be involved in H2A.Z incorporation (**Greenberg et al., 2019**; **Pradhan et al., 2016**).

In *Drosophila melanogaster*, where H2A.Z is named H2A.V (**Baldi and Becker, 2013**; **van Daal and Elgin, 1992**), only one gene codes for a SWR1 ortholog: *domino* (*dom*) (**Ruhf et al., 2001**). The

eLife digest Cells contain a large number of proteins that control the activity of genes in response to various signals and changes in their environment. Often these proteins work together in groups called complexes. In the fruit fly *Drosophila melanogaster*, one of these complexes is called DOMINO. The DOMINO complex alters gene activity by interacting with other proteins called histones which influence how the genes are packaged and accessed within the cell.

DOMINO works in two separate ways. First, it can replace certain histones with other variants that regulate genes differently. Second, it can modify histones by adding a chemical marker to them, which alters how they interact with genes. It was not clear how DOMINO can do both of these things and how that is controlled; but it was known that cells can make two different forms of the central component of the complex, called DOM-A and DOM-B, which are both encoded by the same gene.

Scacchetti et al. have now studied fruit flies to understand the activities of these forms. This revealed that they do have different roles and that gene activity in cells changes if either one is lost. The two forms operate as part complexes with different compositions and only DOM-A includes the TIP60 enzyme that is needed to modify histones. As such, it seems that DOM-B primarily replaces histones with variant forms, while DOM-A modifies existing histones. This means that each form has a unique role associated with each of the two known behaviors of this complex.

The presence of two different DOMINO complexes is common to flies and, probably, other insects. Yet, in other living things, such as mammals and yeast, their two roles are carried out by protein complexes originating from two distinct genes. This illustrates a concept called convergent evolution, where different organisms find different solutions for the same problem. As such, these findings provide an insight into the challenges encountered through evolution and the diverse solutions that have developed. They will also help us to understand the ways in which protein activities can adapt to different needs over evolutionary time.

first biochemical characterization revealed the presence of a multi-subunit complex composed of 15 proteins associated with the DOM ATPase (**Kusch et al., 2004**). While many of the interactors identified are orthologous to the yeast SWR1.C subunits, additional interactors were found. Surprisingly, they showed similarity to components of a distinct yeast complex, the Nucleosome Acetyltransferase of H4 (NuA4.C) (**Kusch et al., 2004**). NuA4.C is a histone acetyltransferase (HAT) complex with the histone H4 N-terminal domain as a primary target (**Allard et al., 1999**; **Doyon et al., 2004**; **Wang et al., 2018a**; **Xu et al., 2016**). The DOM complex (DOM.C) appeared then to be a chimera, a fusion between two complexes with different biochemical activities. It has been proposed that both enzymatic activities of DOM.C, histone acetylation and histone variant exchange, are required for H2A.V turnover during DNA damage response (**Kusch et al., 2004**). It is unclear, however, if this model of DOM.C action could be generalized to other processes, such as transcription regulation. Furthermore, it is still not known how H2A.V is incorporated globally into chromosomes while, at the same time, enriched at promoters.

It is long known that the *dom* transcripts are alternatively spliced to generate two major isoforms, DOM-A and DOM-B (**Ruhf et al., 2001**). We and others previously found that the two splice variants play non-redundant, essential roles during development with interesting phenotypic differences (**Börner and Becker, 2016**; **Liu et al., 2019**).

In this work, we systematically characterized the molecular context and function of each DOM splice variant in *D. melanogaster* cell lines and assessed their contribution to the activity of the DOM.C in the context of transcription. We discovered the existence of two separate, isoform-specific complexes with characteristic composition. Both are involved in transcription regulation, but through different mechanisms. On the one hand, we found that the DOM-B.C is the main ATP-dependent remodeler for H2A.V, responsible for its deposition across the genome and specifically at active promoters. On the other hand, we discovered that DOM-A.C is not involved in bulk H2A.V incorporation, despite the presence of an ATPase domain and many shared subunits with DOM-B.C. Rather, we realized that DOM-A.C might be the 'missing' acetyltransferase NuA4.C of *D. melanogaster*, which specifically targets lysine 12 of histone H4 (H4K12), the most abundant and yet

uncharacterized H4 acetylation in flies (**Feller et al., 2015**). Surprisingly, our data also suggest that the ATPase activity of DOM-A is dispensable for H4K12 acetylation by the DOM-A.C, a principle that might be conserved across metazoans. Our work illustrates how alternative splicing generates functional diversity amongst chromatin regulators.

Results

The splice variants of DOMINO, DOM-A and DOM-B, define two distinct complexes

The isoforms of the DOMINO ATPase, DOM-A and DOM-B, are identical for the first 2008 amino acids, but alternative splicing diversifies their C-termini (**Figure 1A**). Both proteins share an N-terminal HSA domain and a central, INO80-like ATPase domain. DOM-A has a longer C-terminus characterized by a SANT domain and a region rich in poly-glutamine stretches (Q-rich). The shorter C-terminus of DOM-B, instead, folds in no predictable manner. Given these differences, we wondered if the interaction partners of the two isoforms might differ. To avoid artefactual association of DOM isoforms with proteins upon overexpression, we inserted a 3XFLAG tag within the endogenous *dom* gene in *D. melanogaster* embryonic cell lines using CRISPR/Cas9. The sites were chosen such that either DOM-A (DOM-RA) or DOM-B (DOM-RE) would be tagged at their C-termini. Of note, the editing of DOM-A C-terminus results in the additional tagging of a longer, DOM-A-like isoform (DOM-RG, which compared to DOM-RA has an insertion of 35 residues at its N-terminus starting from residue 401), but leaves a second DOM-A-like isoform untagged (DOM-RD, 16 residues shorter than DOM-RA at the very C-terminus). We obtained three different clonal cell lines for each isoform (3 homozygous clones for DOM-A, 2 homozygous and 1 heterozygous clone for DOM-B) (**Figure 1—figure supplement 1A,B**). The *dom* gene editing resulted in the expression of 3XFLAG-tagged proteins of the correct size and with similar expression levels across clones (**Figure 1B**).

To identify the strongest and most stable interactors, we enriched the isoforms and associated proteins from nuclear extracts by FLAG-affinity chromatography under very stringent conditions. Mass-spectrometry analysis revealed 13 and 12 strongly enriched interactors (FDR < 0.05 and log2 fold-change >0) for DOM-A and DOM-B, respectively (**Figure 1C**, **Supplementary file 1**). Of those, 7 are common between the two isoforms and were previously characterized as DOM interactors (**Kusch et al., 2004**). Two of the expected subunits, PONT and REPT, associated more strongly with DOM-B than with DOM-A under these conditions (log2 DOM-A IP/CTRL = 1.19 and 1.16, FDR = 0.373 and 0.338). A newly identified DOM-B interactor, HCF, also interacts less strongly DOM-A (log2 DOM-A IP/CTRL = 0.70, FDR = 0.466). The unique interactors revealed interesting differences between DOM-A and DOM-B (**Figure 1D**). Three of the proteins that specifically associate with DOM-A [ING3, E(Pc) and TIP60] share extensive homology with the acetyltransferase module of the yeast NuA4 complex. Another component of the yeast NuA4.C, NIPPED-A, specifically associates with DOM-A. Additionally, we found two transcription factors, XBP1 and CG12054, amongst the DOM-A specific interactors. On the DOM-B side, only ARP6 and PPS appear to be specific interactors of this isoform. While ARP6 was not described before in *Drosophila*, its yeast homolog is essential for H2A.Z remodeling by the SWR1.C (**Wu et al., 2005**). To validate the DOM-A/TIP60 interaction, we raised monoclonal antibodies against TIP60. Co-immunoprecipitation confirmed that TIP60 interacts with DOM-A and not with DOM-B (**Figure 1E** and **Figure 1—figure supplement 1C**). The immunoprecipitation of DOM-A appears to be more efficient when probing with the anti-FLAG antibody compared to the anti-DOM-A polyclonal antibody. This difference might be explained by the presence of one of the DOM-A-like isoforms (DOM-RD), which was left untagged. This isoform is therefore not immunoprecipitated by the anti-FLAG antibody, but it is recognized by the DOM-A specific antibody. Importantly, the same co-immunoprecipitation showed that DOM-A and DOM-B do not interact with each other under these conditions. Taken together, these findings document the existence of two distinct DOM complexes: DOM-A.C and DOM-B.C.

Specific effects of DOM isoforms on transcription

Previous observation in flies suggested that DOM-A and DOM-B have different, non-redundant functions during *Drosophila* development (**Börner and Becker, 2016**; **Ruhf et al., 2001**). Isoform-specific depletion by RNA interference (RNAi) of either DOM variant in a *Drosophila* embryonic cell line did

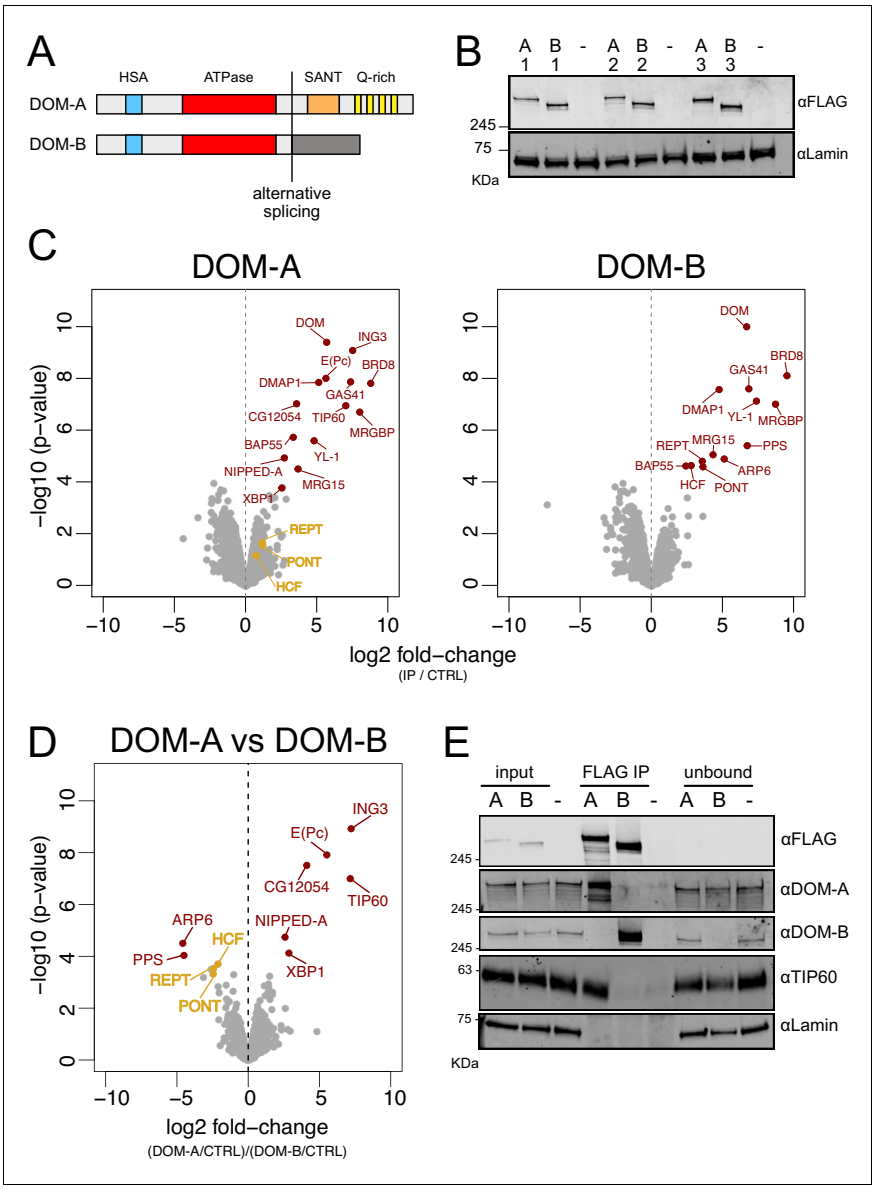


Figure 1. DOMINO isoform-specific affinity enrichment reveals distinct DOM-A and DOM-B complexes. (**A**) Schematic representation of the DOM-A (*dom-RA*) and DOM-B (*dom-RE*) isoforms. The two proteins are derived through alternative splicing and differ in their C-termini. (**B**) Western blot showing the expression of 3XFLAG-tagged DOM-A and DOM-B in nuclear fractions derived from three different clonal S2 cell lines (A = DOM A, B = DOM B). Endogenously tagged proteins were detected using αFLAG antibodies. Nuclear extract from S2 cells lacking the tag (-) serves as negative

Figure 1 continued on next page

Figure 1 continued

control. Lamin: loading control. (**C**) Volcano plot showing -log10 p-value in relation to average log2 fold-change (n = 3 biological replicates) comparing FLAG AP-MS from 3XFLAG DOM-A or DOM-B cell lines (IP) versus 'mock' purifications from untagged S2 cells (CTRL). Red dots represent enriched proteins with FDR < 0.05 and log2 fold-change >0. Orange dots represent proteins significantly enriched in DOM-B AP-MS but not considered as DOM-B specific interactors. (**D**) Volcano plot as described in (C) comparing DOM-A FLAG AP-MS (DOM-A/CTRL) and DOM-B FLAG AP-MS (DOM-B/CTRL). Positive log2 fold-change indicate enrichment in DOM-A pulldown, while negative log2 fold-change indicate enrichment in DOM-B pulldown. Red dots represent isoform-specific enriched proteins with FDR < 0.05. Orange dots represent proteins enriched in DOM-B AP-MS but not considered as DOM-B specific interactors, due to lower statistical significance (FDR > 0.05). (**E**) Western blot validating the mass-spectrometry results. DOM-A clone #2 and DOM-B clone #1 cells were used. Untagged cells (-) serve as negative control. 2% of input and unbound fractions was loaded. Proteins were detected using the antibodies described in the panel.
The online version of this article includes the following figure supplement(s) for figure 1:

Figure supplement 1. DOMINO isoform-specific affinity enrichment reveals distinct DOM-A and DOM-B complexes.

not lead to depletion of the other isoform (***Figure 2A***). Interestingly, knock-down of DOM-A (but not DOM-B) led to a strong reduction of TIP60 protein levels. *tip60* mRNA levels were unchanged (***Figure 2—figure supplement 1A***), indicating that TIP60 requires DOM-A for stability (***Figure 2A***). This suggests that most of TIP60 resides in the DOM-A complex in *Drosophila* cells.

The yeast SWR-1 and NuA4 complexes are both implicated in transcription (***Morillo-Huesca et al., 2010***; ***Searle et al., 2017***). We therefore explored the functional differences of the two DOM isoforms on transcription by RNAseq. In our analysis, we also included knock-downs of H2A.V and TIP60. Knock-down of either DOM-A, DOM-B, TIP60 or H2A.V individually resulted in significant perturbation of transcription, with notable differences (***Figure 2—figure supplement 1B***, ***Supplementary file 2***). Principal Components Analysis (PCA) revealed clearly different transcriptional responses upon loss of DOM-A or DOM-B (***Figure 2B***), which can be visualized by comparing their log2 fold-changes relative to control (***Figure 2C***). The correlation value of 0.45 indicates that many genes are regulated similarly by both ATPases, but a significant number of genes are also differentially affected upon specific depletion of either DOM-A or DOM-B (***Figure 2—figure supplement 1B***). As expected, the transcriptional effects of DOM-A knock-down, but not of DOM-B, resemble the ones caused by knock-down of TIP60 (***Figure 2B,D***). Depletion of H2A.V led to a global reduction of transcription, only observable by normalization to spiked-in *D.virilis* RNA (see methods) (***Figure 2—figure supplement 1C***). The effects of H2A.V depletion were better correlated to DOM-B (r = 0.51) than to DOM-A knock-down (r = 0.25) (***Figure 2B,D***). Many effects of DOM-B depletion may be explained by its H2A.V deposition function, but the ATPase also affects transcription through different routes.

In summary, we found that the depletion of the two DOM isoforms in cells caused specific transcriptional perturbations. The partially overlapping responses upon DOM-B and H2A.V depletions motivated a more in-depth analysis of the relationship between DOM-B and H2A.V levels in the genome.

The DOM-B complex is the main ATP-dependent remodeler for H2A.V

Both SWR1-type ATPases, DOM-A and DOM-B may contribute to H2A.V incorporation and turnover. We explored global changes in H2A.V levels upon isoform-specific RNAi in nuclear extracts containing chromatin and soluble nuclear proteins. We found a strong H2A.V reduction upon DOM-B depletion (***Figure 3A***, ***Figure 3—figure supplement 1A***), while H2A.V mRNA level was unchanged (***Figure 3—figure supplement 1B***). Among the interactors found in our mass-spectrometry analysis, only RNAi against the DOM-B.C-specific subunit ARP6 reduced H2A.V levels to a similar extent (***Figure 3—figure supplement 1C,D***). H2A.V was not affected by the knock-down of DOM-A, TIP60 or other DOM-A.C-specific subunits (***Figure 3A***, ***Figure 3—figure supplement 1A,C,D***).

While western blots reveal global changes, they are not sufficiently sensitive to detect changes of H2A.V occupancy at specific sites in chromatin. We therefore employed a more sensitive chromatin immunoprecipitation (ChIP-seq) approach, in which we included *D. virilis* spike-in cells to quantify global changes in H2A.V levels. As expected, we scored dramatic effects on H2A.V levels along the entire genome upon depletion of DOM-B, including promoters and transcriptional termination sites (***Figure 3B,C***, ***Figure 3—figure supplement 1E***). Depletion of DOM-A did not affect chromosomal

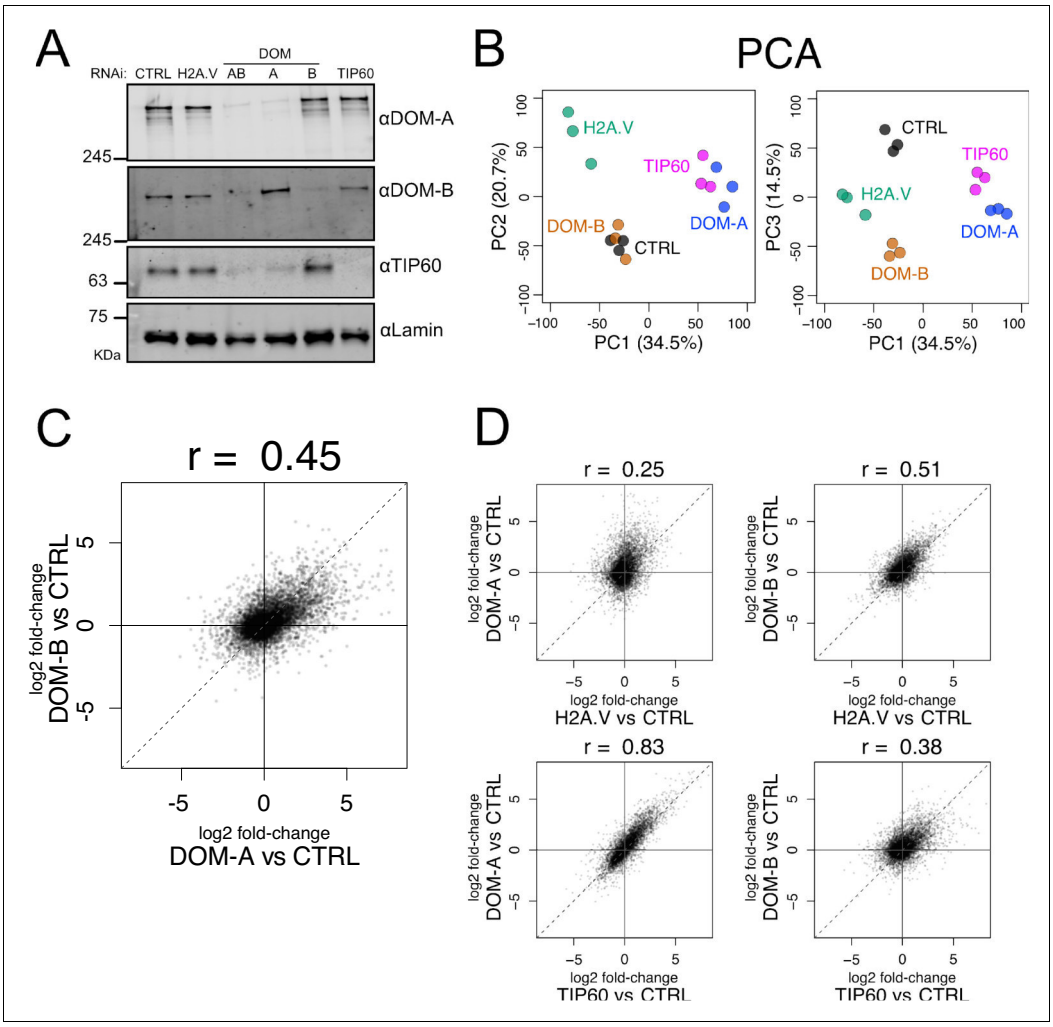


Figure 2. Isoform-specific depletion of DOM-A and DOM-B causes distinct transcriptional effects. (**A**) Western blot showing the expression of DOM-A, DOM-B and TIP60 in nuclear extracts of Kc167 cells treated with dsRNA against GST (CTRL), H2A.V, both DOM isoforms (AB), DOM-A (A), DOM-B (B) and TIP60. Proteins were detected with specific antibodies. Lamin: loading control. (**B**) Principal Component Analysis (PCA) comparing transcriptome profiles derived from Kc167 cells treated with dsRNA against GST or GFP (CTRL), H2A.V, DOM-A, DOM-B and TIP60 (n = 3 biological replicates). Three components (PC1, PC2 and PC3) are shown. Percentage of variance is indicated in parenthesis. (**C**) Scatter plot comparing log2 fold-changes in expression of DOM-A against CTRL RNAi and log2 fold-changes in expression of DOM-B against CTRL RNAi for every gene analyzed (N = 10250). Spearman's correlation coefficient (r) is shown above the plot. (**D**) Same as (**C**) but depicting the comparison between DOM-A or DOM-B RNAi and H2A.V or TIP60 RNAi.

The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. Isoform-specific depletion of DOM-A and DOM-B causes distinct transcriptional effects.

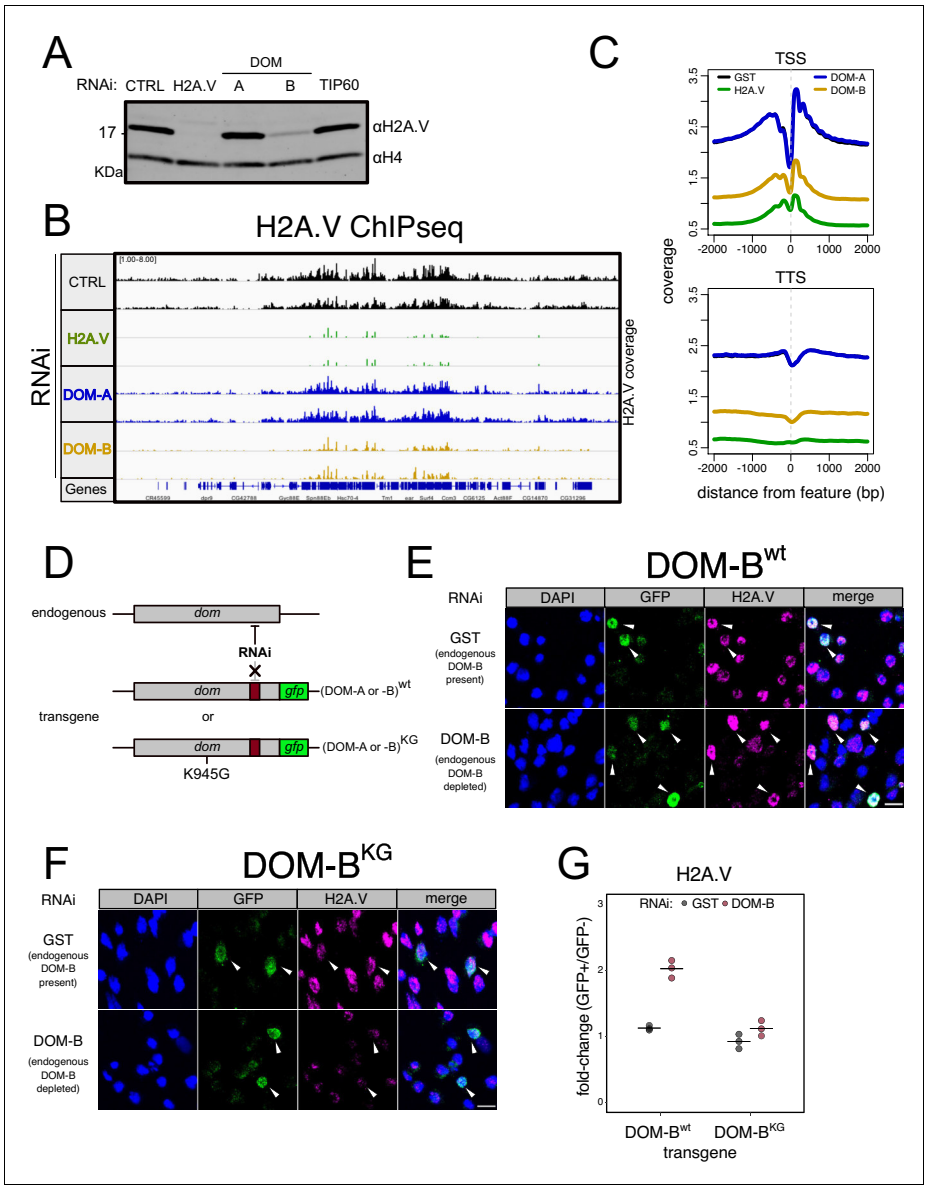


Figure 3. DOM-B is responsible for H2A.V incorporation into chromatin in an ATP-dependent manner. (**A**) Western blot showing the expression of H2A.V in nuclear extracts derived from Kc167 cells treated with dsRNA against GST (CTRL), H2A.V, DOM-A (A), DOM-B (B) and TIP60. Histone H4 (H4): loading control. (**B**) Screenshot of genome browser illustrating a region on Chromosome 3R. Each track shows the spike-in and input normalized H2A.V ChIPseq signal derived from Kc167 cells treated with dsRNA as in (**A**). Two biological replicates are shown. (**C**) Composite plot showing spike-in and

Figure 3 continued on next page

Figure 3 continued

input normalized H2A.V coverage around Transcription Start Sites (TSS) and Transcription Termination Sites (TSS) (N = 10139). Each represent the average coverage (n = 2 biological replicates) of H2A.V in Kc167 cells treated with dsRNA as described in (**A**). (**D**) Schematic representation of the experimental setup to test the requirement for ATPase activity of DOM (A or B) for functionality. A transgene encoding a GFP-tagged wild type or mutant (K945G) DOM is codon-optimized to be resistant to specific dsRNA targeting. The transgene is transfected into Kc167 cells while the endogenous DOM (A or B) are depleted by RNAi. (**E**) Representative immunofluorescence pictures for the DOM-B complementation assay. Cells were treated either with control (GST) dsRNA (endogenous DOM-B present) or with a dsRNA targeting only the endogenous DOM-B (endogenous DOM-B depleted). Cells were transfected with a wild-type transgene encoding RNAi-resistant DOM-B. Cells were stained with DAPI and with GFP and H2A.V antibodies. Arrows indicate the cells where the transgene is expressed and nuclear. Scale bar: 10 μm **F** Same as (**E**) but cells were transfected with a mutant DOM-B (K945G) (**G**) Dot plot showing the quantification of the immunofluorescence-based complementation assay. Each dot represents the fold-change of mean H2A.V signal between GFP-positive (in which the transgene is expressed) and GFP-negative cells in one biological replicate (>100 total cells/replicate). Cells were treated with dsRNAs as in (**E**) Wild-type or mutant DOM-B transgenes are compared.
The online version of this article includes the following figure supplement(s) for figure 3:

Figure supplement 1. DOM-B is responsible for H2A.V incorporation into chromatin in an ATP-dependent manner.

H2A.V at any of these sites (***Figure 3B,C***). These observations support the notion that the DOM-B.C, and not the DOM-A.C, is the remodeler dedicated to H2A.V incorporation.

Since SWR1-type remodelers bind and hydrolyze ATP to incorporate H2A.Z variants (***Hong et al., 2014***; ***Willhoft et al., 2018***), we wanted to confirm the ATP-requirement for in vivo incorporation of H2A.V by DOM-B. We devised an RNAi-based complementation strategy in which we rescued the effects of depleting endogenous *dom-B* mRNA by expression of RNAi-resistant *dom-B* transgenes. The functional complementation involved wild-type DOM-B or a mutant predicted to be deficient in ATP-binding (K945G) (***Hong et al., 2014***; ***Mizuguchi et al., 2004***; ***Figure 3D***). GFP-tagging of the DOM-B proteins allowed to selectively monitor the H2A.V levels by immunofluorescence microscopy in the cells in which the transgenes are expressed. We detected higher levels of H2A.V in cells complemented with a wild-type DOM-B transgene, indicating the expected rescue (***Figure 3E***). Conversely, the remodeling-defective mutant transgene did not increase the residual H2A.V levels (***Figure 3F***). Comparing the mean H2A.V signal between cells that express the transgene (GFP+) and cells that don't (GFP-), revealed once more that only the wild-type could restore H2A.V levels (***Figure 3G***, ***Figure 3—figure supplement 1F***). The data suggest that the DOM-B.C is responsible for the incorporation of H2A.V in an ATP-dependent manner.

The DOM-A complex is related to the yeast NuA4 complex and catalyzes H4K12 acetylation

Despite the presence of an ATPase domain identical to DOM-B, DOM-A does not seem to be responsible for H2A.V incorporation in steady state. Therefore, we considered other functions for DOM-A.C. The striking correlation between transcriptional responses upon TIP60 and DOM-A depletion suggests a unique association with functional relevance. Our mass-spectrometry analysis had identified several proteins that are homologous to corresponding subunits of the yeast NuA4 HAT complex. The core NuA4.C subunit EAF1 is a small protein with prominent N-terminal HSA and C-terminal SANT domains. DOM-A also features similarly arranged domains, but they are separated by the long ATPase domain (***Figure 4—figure supplement 1A***). This raises the question whether DOM-A might serve as the central subunit of a NuA4-type complex in *Drosophila*. The existence of such a complex with functional and structural similarity to the well-studied yeast complex has not been reported so far. Since NuA4.C is responsible for histone acetylation, we looked at H3 and H4 acetylation changes upon DOM isoform-specific knock-down by targeted mass-spectrometry.

The hypothesis of a *Drosophila* NuA4.C poses the acetyltransferase TIP60 as the main effector of DOM-A.C. This is supported by the earlier finding that TIP60 is unstable in the absence of DOM-A (***Figure 2A***). We therefore included TIP60 knock-down for our targeted mass-spectrometry. Our analysis showed that RNAi against DOM-A, but not against DOM-B, specifically reduces H4K12ac (average 28.9% reduction) and, to a lesser extent, H4K5ac (average 23.1% reduction) (***Figure 4A***, ***Figure 4—figure supplement 1B***). Importantly, unsupervised clustering shows that depleting DOM-A or TIP60 lead to very similar changes in histone acetylation patterns: depletion of TIP60 also reduces H4K12ac, by on average 36.3% and H4K5ac by 16.4% (***Figure 4A***, ***Figure 4—figure supplement 1B***). Interestingly, we detected a decrease in monomethylation and increase of trimethylation

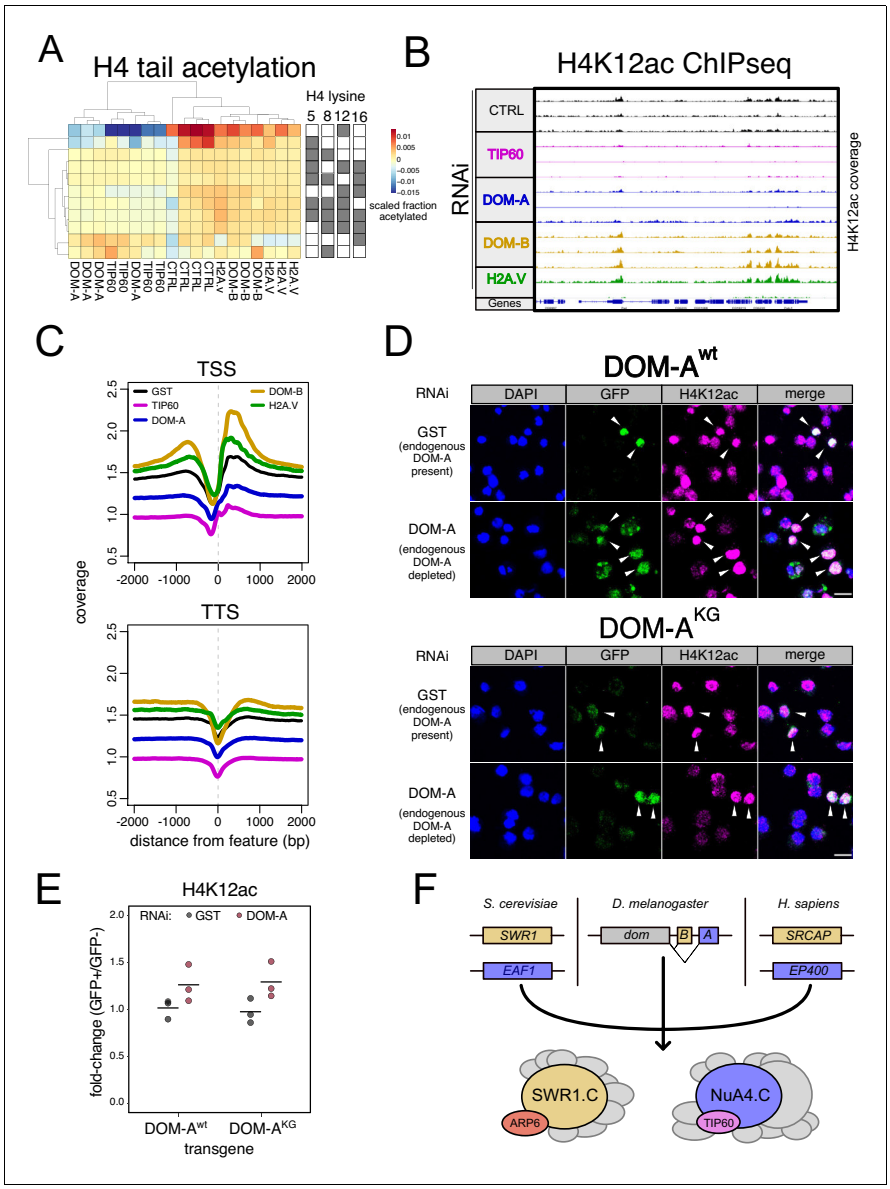


Figure 4. The DOM-A.C acetylates H4K12. (**A**) Heatmap shows scaled acetylation levels for various histone H4 residues (measured by mass-spectrometry) in Kc167 cells treated with dsRNA against GST or GFP (CTRL), H2A.V, DOM-A (A), DOM-B (B) and TIP60. Individual biological replicates are shown. Rows and columns are clustered based on Euclidean distance. (**B**) Screenshot of genome browser illustrating a region on Chromosome 3R. Each track shows spike-in and input normalized H4K12ac ChIPseq signal derived from Kc167 cells treated with dsRNA against GST or GFP (CTRL),

Figure 4 continued on next page

Figure 4 continued

TIP60, DOM-A, DOM-B and H2A.V. 3 biological replicates are shown for all RNAi except H2A.V (2 biological replicates). (**C**) Composite plot showing spike-in and input normalized H4K12ac coverage around Transcription Start Sites (TSS) and Transcription Termination Sites (TSS) (N = 10139). Each represent the average coverage (n = 2 biological replicates for H2A.V RNAi, n = 3 biological replicates for all the other knock-down) of H4K12ac in Kc167 cells treated with dsRNA as described in (**A**). (**D**) Representative immunofluorescence pictures for the DOM-A complementation assay. Cells were treated either with control (GST) dsRNA (endogenous DOM-A present) or with a dsRNA targeting the endogenous DOM-A (endogenous DOM-A depleted). In the top panel, cells were transfected with a wild-type transgene encoding for DOM-A. In the bottom panel, cells were transfected with a mutant (K945G) DOM-A transgene. Cells were stained with DAPI, and with GFP and H4K12ac antibodies. Arrows indicate the cells where the transgene is expressed and nuclear. Scale bar: 10 μm (**E**) Dot plot showing the quantification of the immunofluorescence-based complementation assay. Each dot represents the fold-change of mean H4K12ac signal between GFP-positive (in which the transgene is expressed) and GFP-negative cells in one biological replicate (>100 total cells/replicate). Cells were treated with dsRNAs as in (**E**) Wild-type or mutant DOM-A transgenes are compared. (**F**) Model for SWR1.C and NuA4.C specification in *S. cerevisiae*, *D. melanogaster* and *H. sapiens*.
The online version of this article includes the following figure supplement(s) for figure 4:

Figure supplement 1. The DOM-A.C acetylates H4K12.

at H3K27 by DOM depletion, a bit stronger for DOM-A or TIP60 (***Figure 4—figure supplement 1C, D***). H4 methylation was unchanged (***Figure 4—figure supplement 1E***).

The H4K12 seems to be the most prominent chromatin target of the DOM-A/TIP60 complex. We sought to confirm the mass spectrometric result by an orthogonal ChIP-seq experiment. We found the H4K12ac signal reduced in many regions of the genome, including promoters and transcriptional termination sites, upon TIP60 RNAi and to a lesser extent upon DOM-A RNAi (***Figure 4B,C***, ***Figure 4—figure supplement 1F,H,I***), but the results suffer from variability, probably due to a low ChIP efficiency of the H4K12ac antibody (***Figure 4—figure supplement 1F***). Comparison between H4K12ac and transcription showed that genes downregulated in DOM-A or TIP60 knock-down tend to have higher basal levels of H4K12ac at promoters (***Figure 4—figure supplement 1G***). The reduction of H4K12ac by DOM-A or TIP60 knock-down, however, is global and does not affect some genes specifically (***Figure 4—figure supplement 1H***). Remarkably, depletion of DOM-B caused an unexpected increase in H4K12 acetylation at many chromosomal regions that lose the mark upon DOM-A ablation (***Figure 4B,C***). Depletion of H2A.V also causes a small H4K12ac increase similarly to DOM-B knock-down, as if the absence of this remodeler and/or its substrate allowed more DOM-A activity at promoters.

The yeast NuA4.C does not contain a functional ATPase at its core. To explore whether the acetylation of H4K12 catalyzed by the DOM-A.C depends on ATP-dependent nucleosome remodeling activity, we employed the same RNAi-based complementation strategy we had used for DOM-B (***Figure 3D***). DOM-A was depleted and RNAi-resistant DOM-A wild-type or ATPase mutant derivatives were tested for their ability to rescue the loss of H4K12ac. As expected, the wild-type DOM-A transgene restored H4K12 acetylation (***Figure 4D***). Remarkably, this acetylation did not depend on a functional DOM-A ATPase. Comparison of the mean H4K12ac signals in cells that do or do not express the transgene confirmed that both, the wild-type and the mutant transgenes, could restore H4K12ac levels (***Figure 4E***, ***Figure 4—figure supplement 1J***).

Discussion

Our mass-spectrometry analysis of endogenously expressed DOM isoforms purified under stringent conditions revealed two separate complexes. A DOM.C was previously reported after overexpression and affinity-purification of tagged PONTIN (***Kusch et al., 2004***), which yielded a mixture of DOM-A and DOM-B complexes and may be contaminated with the dINO80 complex, which also contains PONTIN (***Klymenko et al., 2006***). In light of our results, we think the model for H2A.V exchange during DNA damage response proposed in this early work (***Kusch et al., 2004***) should be re-visited accounting for the contribution of both DOM-A.C, DOM-B.C and possibly dINO80.C. It will be interesting to define the role of each complex on the recognition and restoration of damaged chromatin, especially at the level of H2A.V remodeling and acetylation-based signaling.

We previously showed that DOM-B, and not DOM-A or TIP60, affects H2A.V levels during fly oogenesis (***Börner and Becker, 2016***). In addition to confirming this finding in a different system and with complementary experimental approaches, we now also showed that the DOM-B.C is

responsible for H2A.V incorporation into chromatin. The reaction requires ATP, like SWR1.C-mediated H2A.Z incorporation. We also discovered a previously unidentified subunit, ARP6, which is necessary for the maintenance of H2A.V global levels, just like DOM-B. In yeast SWR1.C and human SRCAP.C, the ARP6 orthologs are indispensable for nucleosome remodeling since they couple the ATPase motor to productive nucleosome sliding (**Matsuda et al., 2010**; **Willhoft et al., 2018**; **Willhoft and Wigley, 2020**; **Wu et al., 2005**). The *Drosophila* DOM-B.C is likely to employ a similar remodeling mechanism. Knock-down of DOM-B affects transcription, but the effects overlap only partially with those that follow H2A.V depletion. This discrepancy could be explained in several ways. First, the reduction of H2A.V levels upon DOM-B knock-down is not as extensive as the one caused by direct depletion of H2A.V. The residual levels of H2A.V upon DOM-B depletion may suffice to regulate transcription at many promoters. Second, we cannot exclude that DOM-B.C also impacts transcription independently of H2A.V incorporation. Third, the global increase of H4K12ac at promoters upon DOM-B knock-down might indirectly compensate for the loss of H2A.V at some specific genes.

The DOM-A.C, surprisingly, did not affect H2A.V incorporation under physiological conditions in any of our assays, in agreement with what has been observed for the DOM-A isoform during oogenesis (**Börner and Becker, 2016**). DOM-A.C lacks the ARP6 subunit that is a mechanistic requirement for nucleosome remodeling by INO80-type remodelers (**Willhoft and Wigley, 2020**). Because the ATPase domain of DOM-A is identical to the one in DOM-B, it is possible that DOM-A utilizes ATP under circumstances that we did not monitor in our study. It is also possible that DOM-A.C-specific subunits have an inhibitory effect on DOM-A ATPase activity through allosteric regulation. Of note, the recombinant human ortholog of DOM-A, EP400, can incorporate H2A.Z (**Park et al., 2010**), but H2A.Z levels are unaffected if EP400 is depleted in vivo, where it resides in a multi-subunit complex (**Pradhan et al., 2016**). Regulation of nucleosome remodeling through autoinhibitory domains or associated subunits is a widespread mechanism (**Clapier et al., 2017**).

Our data suggest that the DOM-A.C is the functional equivalent of the yeast NuA4.C, which acetylates the H4 N-terminus (**Kuo et al., 2015**) and possibly other proteins. Depletion of DOM-A.C causes a significant reduction of H4K12ac at a global level. Some genomic regions that still retain a high H4K12ac ChIP signal in the absence of DOM-A may be explained by the presence of additional acetyltransferases targeting H4K12, such as CHAMEAU (**Feller et al., 2015**; **Peleg et al., 2016**). The function of H4K12ac is still largely unknown in *Drosophila*, although it has been implicated in aging (**Peleg et al., 2016**). We speculate that H4K12ac may participate in transcriptional regulation since knock-down of DOM-A or TIP60 perturb the transcriptional program in a very similar manner. Genes down-regulated upon DOM-A and TIP60 RNAi show high H4K12ac around their TSS, but the H4K12ac is not specifically reduced at their promoters. We speculate that these genes might rely more on H4K12ac for their expression or be more sensitive to changes in acetylation. Given that H4K12ac is the most abundant H4 acetylation (**Feller et al., 2015**; **Peleg et al., 2016**), we also cannot exclude that some of these effects are due to global and aberrant chromosomal condensation. The increase in H4K12ac at promoter observed upon DOM-B RNAi appears to be partially phenocopied by the knock-down of H2A.V. It is possible that H2A-containing nucleosomes are a better substrate for the DOM-A.C compared to the ones containing only H2A.V. The loss of the variant might therefore result in higher H4K12ac catalyzed by TIP60.

In the absence of DOM-A, TIP60 is unstable suggesting that the DOM-A.C is the major form of TIP60, at least in *D. melanogaster* cells. It also suggests that during evolution a HAT module and some components of the SWR1.C became stably associated in a new functional complex, the dNuA4.C, as it has been proposed for the human EP400 complex (**Auger et al., 2008**). An intermediate case is found in *C. albicans*, where acetylation of EAF1 by TIP60 mediates a reversible association between the NuA4.C and SWR1.C (**Wang et al., 2018b**). We found that the DOM-A.C H4K12 HAT activity does not need the DOM-A ATPase activity. In yeast, inserting the ATPase domain of the *Drosophila* DOM between the HSA and SANT domains of EAF1, the central subunit of the yeast NuA4, does not affect its function (**Auger et al., 2008**). Such a situation could also apply to DOM-A. Lastly, our mass-spectrometry analysis revealed new, uncharacterized interactors for DOM-A. Of those, the transcription factor CG12054 has been found as a potential DOM partner in a previous screen (**Rhee et al., 2014**). Its human ortholog, JAZF1, appears to be involved in transcriptional repression (**Nakajima et al., 2004**) and has been associated to endometrial stromal tumors (**Koontz et al., 2001**). Its function in flies is unknown.

Division of labor between chromatin modifying enzymes is key to ensure efficient regulation of nuclear processes. During evolution, genome duplications and genetic divergence expand and diversify activities. The case of DOM illustrates beautifully how evolution can take different routes to achieve what must be assumed as an important functional specification (***Figure 4F***, ***Supplementary file 3***). In yeast, the SWR1 and NuA4 complexes are entirely separate entities. In humans, whose genomes underwent duplication events, the paralogous SRCAP and EP400 protein ATPases each organize different complexes that may serve distinct, conserved functions. In *Drosophila* a similar specialization was achieved by alternative splicing. Surprisingly, the gene orthologs of *dom* in honeybee (*A. mellifera*, LOC413341), jewel wasp (*N. vitripennis*, LOC100115939), Jerdon's jumpin ant (*H. saltator*, LOC105183375), red flour beetle (*T. castaneum*, LOC656538) and even in the common house spider (*P. tepidariorum*, LOC107448208) undergo alternative splicing to generate at least two isoforms with different C-termini. The mode of specification of SWR1 and NuA4 through splice variants might therefore not be limited to *Drosophila*, but more wide-spread throughout the Arthropoda phylum.

Materials and methods

Key resources table

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Cell line (*D. melanogaster*)	Kc167	DGRC	FLYB; FBtc0000001	
Cell line (*D. melanogaster*)	S2 (Clone L2-4)	***Villa et al., 2016***		Gift from P Heun lab
Cell line (*D. virilis*)	79f7Dv3	***Albig et al., 2019***		Gift from BV Adrianov
Antibody	DOM-A (17F4) (rat monoclonal)	***Börner and Becker, 2016*** and this publication		1:5 (WB)
Antibody	DOM-A (SA-8977) (rabbit polyclonal)	This publication		1:1000 (WB)
Antibody	DOM-B (SA-8979) (rabbit polyclonal)	This publication		1:1000 (WB)
Antibody	TIP60 (11B10) (rat monoclonal)	This publication		1:20 (WB)
Antibody	H2A.V (Rb-H2Av) (rabbit polyclonal)	***Börner and Becker, 2016***		1:1000 (WB) 1:2500 (IF) 25 µl/IP (ChIP)
Antibody	H4 (rabbit polyclonal)	abcam	ab10158	1:5000 (WB)
Antibody	H4K12ac (rabbit polyclonal)	Merck-Millipore	07–595	1:2500 (IF) 2 µl/IP (ChIP)
Antibody	FLAG-m2 (mouse monoclonal)	Sigma-Aldrich	F3165	1:1000 (WB)
Antibody	GFP (mouse monoclonal)	Roche	11814460001	1:500 (IF)
Antibody	Lamin (mouse monoclonal)	Gift from H Saumweber		1:1000 (WB)
Recombinant DNA reagent	pENTR3C (plasmid)	Thermo Fischer Scientific	A10464	

Continued on next page

Continued

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Recombinant DNA reagent	pHWG (plasmid)	DGRC		Kind gift from P Korber

Plasmids, primers, cell lines and antibodies from this study are available upon request from Peter B Becker (pbecker@bmc.med.lmu.de).

Cell lines and culture conditions

D. melanogaster embryonic Kc167 cell line was obtained from the *Drosophila* Genomic Resource Center (https://dgrc.bio.indiana.edu/Home). *D. melanogaster* S2 (subclone L2-4) cell line was a kind gift of P Heun (***Villa et al., 2016***). *D. virilis* 79f7Dv3 cell line was a kind gift of BV Adrianov (***Albig et al., 2019***). The identity of cell lines was verified by high-throughput sequencing. Cells were subjected to mycoplasma testing. Cells were maintained at 26˚C in Schneider's *Drosophila* Medium (Thermo-Fischer, Cat. No 21720024) supplemented with 10% FBS (Kc167 and S2) or 5% FBS (79f7Dv3) and 1% Penicillin-Streptomycin solution (Sigma-Aldrich, Cat No P-4333).

CRISPR/Cas9 tagging

gRNAs targeting exon 14 (DOM-A and DOM-G, Flybase transcripts *dom-RA* and *dom-RG*) or exon 11 (DOM-B, Flybase transcript *dom-RE*) were initially designed using GPP sgRNA designer (https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design; ***Doench et al., 2016***). gRNA candidates were checked for off-targets using flyCRISPR Target Finder (https://flycrispr.org/target-finder/, guide length = 20, Stringency = high, PAM = NGG only) (***Gratz et al., 2014***). Two gRNAs each for DOM-A and DOM-B were selected (***Supplementary file 4***). The 20 bp gRNAs were fused to a tracrRNA backbone during synthesis and cloned downstream of the *Drosophila* U6 promoter. These constructs were synthesized as gBlocks (Integrated DNA Technologies) and PCR-amplified before transfection using Q5 polymerase (New England Biolabs, Cat No. M0491S). To generate the repair template, the sequence encoding for 3XFLAG tag, including a stop codon, was inserted between two homology arms of 200 bp each by gene synthesis (Integrated DNA Technologies) (***Supplementary file 4***). The repair templates were cloned in pUC19 to generate repair plasmids. For CRISPR editing, one million S2 cells (subclone L2-4) in 500 µL medium were seeded in each well of a 24-well plate. After 4 hr, cells were transfected with 110 ng of gRNAs (55 ng each), 200 ng of repair plasmid and 190 ng of pIB_Cas9_Blast (encoding SpCas9 and carrying Blasticidin resistance, kind gift of P Heun) using X-tremeGENE HP DNA Transfection Reagent (Roche, Cat. No 6366236001). 24 hr after transfection, medium was replaced with 500 µL of fresh medium containing 25 µg/ml Blasticidin (Gibco, Cat. No A1113903). Three days after selection the cells were collected and seeded into 6 cm tissue culture dishes at three different concentrations (1000, 2000 and 5000 cells/well) and allowed to attach for 1–2 hr. Medium was then removed and cells carefully overlaid with 2.5 mL of a 1:1 mix of 2X Schneider's Medium (prepared from powder, Serva, Cat. No 35500) + 20% FBS + 2% Penicillin-Streptomycin and 0.4% low-melting agarose equilibrated to 37˚C. The dishes were sealed with parafilm, inserted into 15 cm dishes together with a piece of damp paper and sealed once more with parafilm. After 2 to 3 weeks, individual cell colonies were picked using a pipet, suspended in 100 µL of Schneider's *Drosophila* Medium + 10% FBS + 1% Penicillin-Streptomycin and plated into 96-well plates. Clones were expanded for 1–2 weeks and further expanded into 48-well plates. For PCR-testing of clones, 50 µL of cells were collected, 50 µL water was added and DNA was purified using Nucleospin Gel and PCR Cleanup (Macherey-Nagel, Cat. No 740609.250). Extracted DNA was PCR-amplified to check the insertion of the 3XFLAG tag. The PCR product of DOM-A clone #2 results larger due to the presence of an insertion 29 bp downstream of the stop codon (***Figure 1—figure supplement 1B***).

Selected clones were further expanded and stored in liquid nitrogen in 90% FBS + 10% DMSO.

Nuclear extraction and FLAG affinity enrichment

For nuclear extraction from 3XFLAG-tagged cells lines, 0.5–1 billion cells were collected by centrifugation at 500 g for 5 min. Cells were washed with 10 ml of PBS, resuspended in 10 ml of ice-cold NBT-10 buffer [15 mM HEPES pH 7.5, 15 mM NaCl, 60 mM KCl, 0.5 mM EGTA pH 8, 10% Sucrose, 0.15% Triton-X-100, 1 mM PMSF, 0.1 mM DTT, 1X cOmplete EDTA-free Protease Inhibitor (Roche, Cat. No 5056489001)] and rotated for 10 min at 4˚C. Lysed cells were gently overlaid on 20 ml of ice-cold NB-1.2 buffer (15 mM HEPES pH 7.5, 15 mM NaCl, 60 mM KCl, 0.5 mM EGTA pH 8, 1.2 M Sucrose, 1 mM PMSF, 0.1 mM DTT, 1X cOmplete EDTA-free Protease Inhibitor) and spun at 4000 g for 15 min. Pelleted nuclei were washed once with 10 ml of ice-cold NB-10 buffer (15 mM HEPES pH 7.5, 15 mM NaCl, 60 mM KCl, 0.5 mM EGTA pH 8, 10% Sucrose, 1 mM PMSF, 0.1 mM DTT, 1X cOmplete EDTA-free Protease Inhibitor) and resuspended in ice-cold Protein-RIPA buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM EDTA, 0.5% IGEPAL CA-630, 0.5% Na-Deoxycholate, 0.1% SDS, 1 mM PMSF, 0.1 mM DTT, 1X cOmplete EDTA-free Protease Inhibitor). Nuclei were sonicated in 15 mL Falcons tubes using Diagenode Bioruptor (20 cycles, 30 s ON/30 s OFF). Extract was spun at 16000 g for 15 min at 4˚C in a table-top centrifuge. Soluble extract was collected and total protein concentration determined using Protein Assay Dye Reagent Concentrate (BIO-RAD, Cat No 5000006) with BSA as standard. 2 mg aliquots were flash-frozen. For FLAG-immunoprecipitation, 2 mg of nuclear protein were thawed, spun at 160,000 g for 15 min at 4˚C to remove aggregates. Extracts were diluted 1:1 with Benzonase dilution buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 4 mM $MgCl_2$, 0.5% NP-40, 1 mM PMSF, 0.1 mM DTT, 1X cOmplete EDTA-free Protease Inhibitor). 60 µL (50% slurry) of Protein-RIPA-equilibrated FLAG-m2 beads (Sigma-Aldrich, Cat. No A2220) were added together with 1 µL of Benzonase (Merck-Millipore, Cat. No 1.01654.0001). After 3 hr of incubation at 4˚C on a rotating wheel, the beads were washed 3 times with ice-cold Protein-RIPA buffer and thrice with ice-cold TBS (5 mM Tris-HCl pH 7.5, 150 mM NaCl) (5 min rotation each, 4˚C). For western blots, beads were then resuspended in 50 µL of 5X Laemmli Sample buffer (250 mM Tris-HCl pH 6.8, 10% w/v SDS, 50% v/v glycerol, 0.1% w/v bromophenol blue, 10% β-mercaptoethanol) and boiled for 5 min at 95˚C. For mass-spectrometry, beads were incubated with 50 µL of elution buffer (2 M urea, 50 mM Tris-HCl, pH 7.5, 2 mM DTT and 10 µg ml^{-1} trypsin) for 30 min at 37˚C. The eluate was removed and beads were incubated in 50 µL of alkylation buffer (2 M urea, 50 mM Tris-HCl, pH 7.5 and 10 mM chloroacetamide) at 37˚C for 5 min. Combined eluates were further incubated overnight at room temperature. Tryptic-peptide mixtures were acidified with 1% Trifluoroacetic acid (TFA) and desalted with Stage Tips containing three layers of SDB-RPS (Polystyrene-divinylbenzene copolymer partially modified with sulfonic acid) material. To this end, samples were mixed 1:1 with 1% TFA in isopropanol and loaded onto the stagetip. After two washes with 100 µL 1%TFA in Isopropanol and two washes with 100 µl 0.2%TFA in water, samples were eluted with 80 µl of 2% (v/v) ammonium hydroxide, 80% (v/v) acetonitrile (ACN) and dried on a centrifugal evaporator. Samples were dissolved in 10 µL Buffer A* (2% ACN/0.1% TFA) for mass spectrometry. Peptides were separated on 50-cm columns packed in house with ReproSil-Pur C18-AQ 1.9 µm resin (Dr Maisch). Liquid chromatography was performed on an EASY-nLC 1200 ultra-high-pressure system coupled through a nanoelectrospray source to a Q-Exactive HF-X mass spectrometer (Thermo Fisher). Peptides were loaded in buffer A (0.1% formic acid) and separated by application of a non-linear gradient of 5–30% buffer B (0.1% formic acid, 80% ACN) at a flow rate of 300 nl min^{-1} over 70 min. Data acquisition switched between a full scan and 10 data-dependent MS/MS scans. Full scans were acquired with target values of 3×10^6 charges in the 300–1,650 *m/z* range. The resolution for full-scan MS spectra was set to 60,000 with a maximum injection time of 20 ms. The 10 most abundant ions were sequentially isolated with an ion target value of 1×10^5 and an isolation window of 1.4 *m/z*. Fragmentation of precursor ions was performed by higher energy C-trap dissociation with a normalized collision energy of 27 eV. Resolution for HCD spectra was set to 15,000 with a maximum ion-injection time of 60 ms. Multiple sequencing of peptides was minimized by excluding the selected peptide candidates for 30 s. In total, 3 technical replicates (parallel immunoprecipitations) for each of the 3 biological replicates (1 clone = 1 replicate, extract prepared on different days) were analyzed. Raw mass spectrometry data were analyzed with MaxQuant (version 1.5.6.7) (***Cox and Mann, 2008***) and Perseus (version 1.5.4.2) software packages. Peak lists were searched against the *Drosophila melanogaster* UniProt FASTA database combined with 262 common contaminants by the integrated Andromeda search engine (***Cox et al., 2011***). The false

discovery rate (FDR) was set to 1% for both peptides (minimum length of 7 amino acids) and proteins. 'Match between runs' (MBR) with a maximum time difference of 0.7 min was enabled. Relative protein amounts were determined with the MaxLFQ algorithm (**Cox et al., 2014**), with a minimum ratio count of two. Missing values were imputed from a normal distribution, by applying a width of 0.2 and a downshift of 1.8 standard deviations. Imputed LFQ values of the technical replicates for each biological replicate were averaged. Differential enrichment analysis was performed in R using the *limma* package as previously described (**Kammers et al., 2015**; **Ritchie et al., 2015**). Adjusted p-values (FDR) were calculated using the *p.adjust* function (*method = 'fdr'*) (**Source code 1**).

RNAi

Primers for dsRNA templates were either obtained from the TRiP website (https://fgr.hms.harvard.edu/fly-in-vivo-rnai), designed using SnapDragon (https://www.flyrnai.org/snapdragon) or designed using E-RNAi (https://www.dkfz.de/signaling/e-rnai3/) (**Horn and Boutros, 2010**; **Perkins et al., 2015**), except one primer pair for *Tip60* RNAi which was obtained from **Kusch et al. (2004)** (**Supplementary file 4**). Templates for in vitro transcription were generated by PCR-amplification using Q5 Polymerase (New England Biolabs, Cat No. M0491S). dsRNAs were generated by in vitro transcription using MEGAScript T7 kit (Invitrogen, Cat. No AMB 13345), followed by incubation at 85˚C for 5 min and slow cool-down to room temperature. *D. melanogaster* Kc167 cell were collected by spinning at 500 g for 5 min. Cells were washed once with PBS and resuspended in Schneider's *Drosophila* Medium without serum and Penicillin-Streptomycin at a concentration of 1.5 million/ml (for RNAi in 12-well and 6-well plates) or 3 million/ml (for RNAi in T-75 flasks). 0.75 million (12-well), 1.5 million (6-well) or 15 million (T-75 flasks) cells were plated and 5 µg (12-well), 10 ug (6-well) or 50 µg (T-75 flasks) of dsRNA was added. Cells were incubated for 1 hr with gentle rocking and 3 volumes of Schneider's *Drosophila* Medium (supplemented with 10% FBS and 1% Penicillin-Streptomycin solution) was added. After 6 days cells were collected and analyzed.

RNAseq

Two million of Kc167 cells were pelleted at 500 g for 5 min. Cells were resuspended in 1 mL of PBS and 1 million of *D. virilis* 79f7Dv3 cells were added. Cells were pelleted at 500 g for 5 min and total RNA was extracted using RNeasy Mini Kit (QIAgen, Cat No. 74104), including DNAse digestion step (QIAgen, Cat No. 79254). mRNA was purified using Poly(A) RNA selection kit (Lexogen, Cat. No M039100). Both total RNA and mRNA quality was verified on a 2100 Bioanalyzer (Agilent Technologies, Cat. No G2939BA). Libraries for sequencing were prepared using NEBnext Ultra II directional RNA library prep kit for Illumina (New England Biolabs, Cat. No E7760L). Libraries were sequenced on an Illumina HiSeq 1500 instrument at the Laboratory of Functional Genomic Analysis (LAFUGA, Gene Center Munich, LMU). For the analysis, 50 bp single reads were mapped to the *D. melanogaster* (release 6) or independently to the *D. virilis* (release 1) genome using STAR (version 2.5.3a) with the GTF annotation `dmel-all-r6.17.gtf` or `dvir-all-r1.07.gtf`, respectively. Multi-mapping reads were filtered out by the parameter `–outFilterMultimapNmax 1`. Genic reads were counted with the parameter `–quantMode GeneCounts`. Read count tables were imported to R and low count genes were removed (at least 1 read per gene in 6 of the samples). Normalization factors (`sizeFactors`) were calculated for *D. melanogaster* or *D. virilis* count tables independently using DESeq2 package (version 1.24). Normalization factors derived from *D. virilis* were applied to *D. melanogaster* counts. Statistical analysis was carried out using DESeq2 by providing replicate information as batch covariate. Estimated log2 fold-change and adjusted p-values were obtained by the results function (DESeq2) and adjusted p-value threshold was set 0.01. Batch effect was corrected by the *ComBat* function from the sva package (version 3.32) on the log2-transformed normalized read counts. Batch adjusted counts were plotted relative to control or used for principal component analysis (PCA). Plots were generated using R graphics. Scripts are available on GitHub (https://github.com/tschauer/Domino_RNAseq_2020).

RT-qPCR

cDNA was synthesized from 1 µg of total RNA (extracted as previously described but omitting the addition of *D. virilis* spike-in cells) using Superscript III First Strand Synthesis System (Invitrogen, Cat. No 18080–051, random hexamer priming) and following standard protocol. cDNA was diluted

1:100, qPCR reaction was assembled using Fast SYBR Green Mastermix (Applied Biosystem, Cat. No 4385612) and ran on a Lightcycler 480 II (Roche) instrument. Primer efficiencies were calculated via serial dilutions. Sequences as available in ***Supplementary file 4***.

Nuclear fractionation and western blot

For nuclear fractionation, 5–10 million cells were pelleted at 500 g for 5 min and washed with PBS. Cell pellets were either used directly or flash-frozen for later processing. Pellets were suspended in 300 µL of ice-cold NBT-10 buffer and rotated for 10 min at 4˚C. Lysed cells were gently overlaid on 500 µL of ice-cold NB-1.2 buffer and spun at 5000 g for 20 min. Pelleted nuclei were washed once with 500 µL of ice-cold NB-10 buffer. Nuclei were resuspended in 60 µL of 1:1 mix of Protein-RIPA buffer and 5X Laemmli Sample buffer. Samples were boiled for 5 min at 95˚C and ran on pre-cast 8% or 14% polyacrylamide gels (Serva, Cat. No 43260.01 and 43269.01) under denaturing conditions in 1X Running Buffer (25 mM Tris, 192 mM glycine, 0.1% SDS). Gels were wet-transferred onto nitrocellulose membranes (GE Healthcare Life Sciences, Cat. No 10600002) in ice-cold 1X Transfer Buffer (25 mM Tris, 192 mM glycine) with 20% methanol (for histones) or with 10% methanol + 0.1% SDS (for DOM proteins) at 400 mA for 45–60 min. Membranes were blocked with 5% BSA in TBS buffer for 1 hr and incubated with primary antibodies in TBST buffer (TBS + 0.1% Tween-20) + 5% non-fat milk at 4˚C overnight. Membrane were washed 3 times (5 min each) with TBST buffer and incubated with secondary antibodies in TBST buffer for 1 hr at room temperature. Membrane were washed 3 times with TBST buffer, 2 times with TBS buffer (5 min each), and dried before imaging. Images were taken on a LI-COR Odyssey or a LI-COR Odyssey CLx machine (LI-COR Biosciences).

ChIPseq

70 to 130 million *D. melanogaster* Kc167 cells, resuspended in 20 ml of complete Schneider's *Drosophila* Medium, were crosslinked by adding 1:10 of the volume of Fixing Solution (100 mM NaCl, 50 mM Hepes pH 8, 1 mM EDTA, 0.5 mM EGTA, 10% methanol-free formaldehyde) and rotated at room temperature for 8 min. 1.17 ml of freshly-prepared 2.5 M glycine was added to stop the fixation (final conc. 125 mM). Cells were pelleted at 500 g for 10 min (4˚C) and resuspended in 10 mL of ice-cold PBS. 3.5 million of fixed *D. virilis* cells (fixed as described for *D. melanogaster* cells) were added for every 70 million *D. melanogaster* cells. Cells were pelleted at 526 g for 10 min (4˚C) and resuspended in 1 ml of PBS + 0.5% Triton-X-100 + 1X cOmplete EDTA-free Protease Inhibitor for every 70 million *D. melanogaster* cells and rotated at 4˚C for 15 min to release nuclei. Nuclei were pelleted at 2000 g for 10 min and washed once with 10 ml of ice-cold PBS. Nuclei were suspended in 1 ml of RIPA buffer (10 mM Tris-HCl pH 8, 140 mM NaCl, 1 mM EDTA, 0.1% Na-deoxycholate, 1% Triton-X-100, 0.1% SDS, 1 mM PMSF, 1X cOmplete EDTA-free Protease Inhibitor) + 2 mM $CaCl_2$ for every 70 million *D. melanogaster* cells, divided into 1 ml aliquots and flash-frozen in liquid N_2. 1 mL of fixed nuclei was quickly thawed and 1 µL of MNase (to 0.6 units) (Sigma-Aldrich, Cat. No N5386) added. Chromatin was digested for 35 min at 37˚C. MNase digestion was stopped by transferring the samples on ice and adding 22 µL of 0.5 M EGTA. Samples were mildly sonicated using a Covaris S220 instrument with the following settings: 50 W peak power, 20% duty factor, 200 cycles/burst, 8 min total time. Insoluble chromatin was removed by centrifugation at 16,000 g for 30 min at 4˚C. Soluble chromatin was pre-cleared by incubation with 10 µL of 50% RIPA-equilibrated Protein A + G sepharose bead slurry (GE Healthcare, Cat. No 17-5280-11 and 17-0618-06) for every 100 µL of chromatin for 1 hr at 4˚C. 100 µL of pre-cleared chromatin were set aside (input) and kept overnight at 4˚C, while each primary antibody was added to 300 µL of chromatin and incubated overnight at 4˚C. 40 µL of 50% RIPA-equilibrated Protein A + G sepharose bead slurry was added for each immunoprecipitation and rotated 3 hr at 4˚C. Beads were washed 5 times with 1 ml of RIPA (5 min rotation at 4˚C, pelleted at 3000 g for 1 min between washes) and resuspended in 100 µL of TE (10 mM Tris pH 8, 1 mM EDTA). 0.5 µL of RNAseA (Sigma-Aldrich, Cat. No. R4875) was added to both input samples and resuspended beads, followed by incubation at 37˚C for 30 min. After addition of 6 µL of 10% SDS, protease digestion (250 ng/ µL Proteinase K, Genaxxon, Cat.no. M3036.0100) and crosslink reversal were performed simultaneously at 68˚C for 2 hr. DNA was purified using 1.8X Agencourt AMPure XP beads (Beckman Coulter, Cat No A63880) following the standard protocol and eluted in 30 µL of 5 mM Tris-HCl pH 8. Libraries for sequencing were prepared using NEBNext Ultra II DNA library prep kit for Illumina (New England Biolabs, E7465). Libraries

were sequenced on an Illumina HiSeq 1500 instrument at the Laboratory of Functional Genomic Analysis (LAFUGA, Gene Center Munich, LMU). 50 bp single-end reads were mapped to the *D. melanogaster* (release 6) or independently to the *D. virilis* (release 1) genome using *bowtie2* (**Langmead and Salzberg, 2012**) using standard parameters. Tag directories and input-normalized coverage files were generated using *Homer* (**Heinz et al., 2010**) with the parameters `-fragLength 150` and `-totalReads` was set to the number of reads mapped to *D. virilis* genome. Input-normalized, scaled *D. melanogaster* coverage files were visualized using the Integrative Genomics Viewer (**Robinson et al., 2011**). Scripts for *D. virilis* scaling and input normalization are available on GitHub (https://github.com/tschauer/Domino_ChIPseq_2020). Composite plots were generated using *tsTools* (https://github.com/musikutiv/tsTools) and base R graphics. Annotations are derived from *TxDb.Dmelanogaster.UCSC.dm6.ensGene_3.4.4* (http://bioconductor.org/packages/release/data/annotation/html/TxDb.Dmelanogaster.UCSC.dm6.ensGene.html). Heatmaps were generated using *pheatmap* (https://cran.r-project.org/web/packages/pheatmap/index.html). Violin-boxplots were generated using *ggplot2* (https://cran.r-project.org/web/packages/ggplot2/index.html).

Cloning of DOM constructs

DOM-A and DOM-B cDNAs were cloned into pENTR3c vector (Thermo Fischer Scientific, Cat. No A10464) by In-Fusion Cloning (Takara Bio, Cat. No 638909) using LD35056, LD03212, LD32234 plasmids *Drosophila* Genomic Resource Center) as templates for PCR. RNAi-resistant DOM-A and DOM-B were generated by substituting around 500 bp of the original cDNA sequence with manually mutagenized, synthesized DNA constructs (gBlock, Integrated DNA Technology), by restriction cloning. DOM-A and DOM-B K945G mutants were generated by site-directed mutagenesis (New England Biolabs, Cat. No E0554S). For transfection in *Drosophila* cells, the constructs in pENTR3c were recombined in pHWG (expression driven by *hsp70* promoter, C-terminal GFP tag;*Drosophila*Genomic Resource Center) by Gateway cloning (Thermo Fischer Scientific).

Complementation assays and immunofluorescence

1–2 million Kc167 cells in 2 ml complete Schneider's*Drosophila*Medium were seeded in each well of a 6-well plate. After 4 hr, cells were transfected with 500 ng of complementation plasmid (described before) + 25 ng of pCoBlast (Thermo Fischer, Cat. No K5150-01) using Effectene Transfection Reagent (QIAgen, Cat. No 301425). 48 hr after transfection, 2 ml of the cells were collected, transferred into T-25 flasks and diluted with 4 ml of complete Schneider's*Drosophila*Medium + Blasticidin at a final concentration of 50 ng/ul. 7–8 days after selection the cells were collected and treated with dsRNA as described before.

For immunofluorescence, 0.2–0.4 million cells in 200 µL of complete Schneider's *Drosophila* Medium were seeded onto a round 12 mm coverslips (Paul Marienfeld GmbH and Co., Cat No. 0117520) placed separately inside wells of 12-well plates. Cell were allowed to attach for 2–4 hr and the coverslips were gently rinsed with 500 µL of PBS. Cells were fixed in 500 µL of ice-cold PBS + 2% formaldehyde for 7.5 min. After removal of fixative, cells were permeabilized by adding 500 µL of ice-cold PBS + 0.25% Triton-X-100 + 1% formaldehyde and incubating for 7.5 min. Coverslips were washed two times with 1 ml of PBS and blocked with PBS + 3% BSA for 1 hr at room temperature. Coverslip were transferred onto a piece of parafilm, placed into a wet chamber and 40 µL of primary antibody solution was gently added onto the coverslip. After overnight incubation at 4˚C, coverslips were transferred back to 12-well plates and washed twice with 1 ml of PBS. Coverslip were transferred again onto a piece of parafilm, placed into a wet chamber and 40 µL of secondary antibody was gently added onto the coverslip. After 1 hr incubation at room temperature, coverslips were transferred back to 12-well plates and washed twice with 1 ml of PBS. Cells were incubated with 1 ml of 0.2 µg/ml DAPI (Sigma-Aldrich, Cat. No 10236276001) for 5 min at room temperature. Coverslips were washed with PBS and with deionized water, mounted on slides with 8 µL of Vectashield (Vector Laboratories, Cat. No H-1000) and sealed with nail polish. Images were taken on a Leica SP5 confocal microscope. Images were processed and analyzed using Fiji (**Source code 2**) and data plotted using R-Studio. p-values were calculated using linear regression (*lm* function in R).

Histone extraction and targeted mass-spectrometry

Kc167 cells were treated with dsRNAs in 6-well plates as described before. Cells were counted, pelleted and snap-frozen in liquid N_2. For histone acid extraction, pellets from 4 to 12 million cells were resuspended in 500 µL of ice-cold 0.2M H_2SO_4 and histone were extracted by rotating overnight at 4˚C. Cell debris were removed by centrifugation at 16,000 g for 10 min at 4˚C. Histone were precipitated by adding trichloroacetic acid to reach 26% final concentration. Tubes were mixed and incubated at 4˚C for 2 hr and spun at 16,000 g for 45 min. Pellets were washed twice with ice-cold 100% acetone (5 min rotation at 4˚C, 10 min of 16,000 g spin at 4˚C between washes), dried for 30 min at room temperature and resuspended in 10 µL of 2.5x Laemmli sample buffer for every initial cell million and boiled at 95˚C for 5 min. Samples were stored at −20˚C until further use. The histones corresponding to 10 million cells were separated onto 4–20% pre-cast polyacrylamide gels (Serva, Cat. No 43277.01). Gels were briefly stained with Coomassie (Serva, Cat. No 17524.01) and stored in water at 4˚C. For targeted mass-spectrometry analysis, histones were excised, washed once with water and de-stained twice by incubating 30 min at 37˚C with 200 µL of 50% acetonitrile (ACN) in 50 mM NH_4HCO_3. Gel pieces were then washed twice with 200 µL water and twice with 200 µL of 100% ACN to dehydrate them, followed by 5 min of speed-vac to remove residual ACN. Histones were in-gel acylated by adding 10 µL of deuterated acetic anhydride (Sigma-Aldrich, Cat. No 175641) and 20 µL of 100 mM NH_4HCO_3. After 1 min, 70 µL of 1 M NH_4HCO_3 were slowly added to the reaction. Samples were incubated at 37˚C for 45 min with vigorous shaking. Samples were washed 5 times with 200 µL of 100 mM NH_4HCO_3, 5 times with 200 µL of water and twice with 200 µL of 100% ACN, followed by 3 min of speed-vac. Gel pieces were rehydrated in 20 µL of trypsin solution (25 ng/ µL trypsin in 100 mM NH_4HCO_3) (Promega, Cat. No V5111) and incubated at 4˚C for 20 min. After the addition of 100 µL of 50 mM NH_4HCO_3, histones were in-gel digested overnight at 37˚C. Peptides were sequentially extracted by incubating 10 min at room temperature twice with 60 µL of 50% ACN 0.25% trifluoroacetic acid (TFA) and twice 40 µL of 100% ACN. The total volume (around 250 µL) was evaporated in a centrifugal evaporator and the dried peptides were stored at −20˚C until resuspension in 100 µL of 0.1% TFA. Peptides were loaded in a C18 Stagetip (prewashed with ACN and conditioned with 0.1% TFA), washed 3 times with 20 µL of 0.1% TFA and peptides were eluted 3 times with 20 µL of 80% ACN 0.25% TFA. Eluted peptides were evaporated in a centrifugal evaporator, resuspended in 15 µl of 0.1% TFA and stored at −20˚C. Desalted peptides were injected in an RSLCnano system (Thermo Fisher Scientific) and separated in a 15 cm analytical column (75 µm ID home-packed with ReproSil-Pur C18-AQ 2.4 µm from Dr. Maisch) with a 50 min gradient from 4% to 40% acetonitrile in 0.1% formic acid at 300 nl/min flowrate. The effluent from the HPLC was electrosprayed into Q Exactive HF mass spectrometer (Thermo Fisher Scientific). The MS instrument was programmed to target several ions as described before (***Feller et al., 2015***) except for the MS3 fragmentation. Survey full scan MS spectra (from m/z 270–730) were acquired with resolution R = 60,000 at m/z 400 (AGC target of 3×10^6). Targeted ions were isolated with an isolation window of 0.7 m/z to a target value of 2×10^5 and fragmented at 27% normalized collision energy. Typical mass spectrometric conditions were: spray voltage, 1.5 kV; no sheath and auxiliary gas flow; heated capillary temperature, 250˚C. Peak integration was performed using Skyline (https://skyline.ms/project/home/software/Skyline/begin.view). Quantified data was further analyzed in R according to the formulas described in ***Feller et al. (2015)*** (***Supplementary file 5***; ***Source code 3*** and ***Source code 4***).

Antibodies

DOM-A and DOM-B polyclonal antibody were generated by expression of C-terminal specific polypeptides. For DOM-A, residues 2963 to 3188 were expressed as C-terminal Glutathione-S-transferase (GST) fusion in *E. coli*, purified using Glutathione Sepharose resin (GE Healthcare, Cat. No 17075605) and eluted with glutathione. For DOM-B, residues 2395 to 2497 were expressed as C-terminal Maltose Binding Protein (MBP) fusion in *E. coli*, absorbed to amylose resin (New England Biolabs, Cat. No E8121S) and eluted with maltose. Antibody production in rabbit was done by Eurogentec (https://secure.eurogentec.com/eu-home.html). Both antibodies were validated by RNAi and western blot. For the monoclonal antibody against TIP60, N-terminal 6xHis-tagged TIP60 (full length) was expressed in *E. coli*, purified over a Ni-NTA column and eluted with imidazole.

Monoclonal antibodies were developed by Dr. Elizabeth Kremmer (BioSysM, LMU Munich). Antibodies were validated by RNAi and western blot.

Sources of other antibodies were: DOM-A monoclonal and H2A.V polyclonal (***Börner and Becker, 2016***). Histone H4 rabbit polyclonal antibody: Abcam (Cat. No ab10158). Mouse anti- FLAG monoclonal antibody: Sigma-Aldrich (Cat. No F3165). Anti H4K12ac rabbit polyclonal antibody: Merck-Millipore (Cat. No 07–595). Anti-GFP mouse monoclonal antibody: Roche (Cat. No 11814460001). Anti-lamin mouse monoclonal antibody: kind gift of Dr. Harald Saumweber.

Data and code availability

Next Generation sequencing data are available at the Gene Expression Omnibus under accession number GSE145738.

Targeted proteomics data are available at ProteomeXchange under accession number PXD017729.

Scripts for *D. virilis* scaling and input normalization for ChIP-seq are available on GitHub (https://github.com/tschauer/Domino_ChIPseq_2020; ***Schauer, 2020a***; copy archived at https://github.com/elifesciences-publications/Domino_ChIPseq_2020).

Scripts for RNA-seq analysis are available on GitHub (https://github.com/tschauer/Domino_RNA-seq_2020; ***Schauer, 2020b***; copy archived at https://github.com/elifesciences-publications/Domino_RNAseq_2020).

Immunofluorescence images used for quantification of the complementation assays are available on Dryad (https://doi.org/10.5061/dryad.1rn8pk0qt).

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Council) – Project ID 213249687 - SFB1064-A1 and -Z04. ZA was supported by the EMBO long-term fellowship (ALTF 168–2018). PH was supported by a Wellcome Trust Senior Fellowship award (103897) and the Wellcome Centre for Cell Biology is supported by core funding from the Wellcome Trust (092076). We thank K Förstemann, J Müller and the Becker laboratory for discussion; R Villa, M Müller and S Baldi for useful feedback and critical reading of the manuscript; H Blum and S Krebs for the next-generation sequencing; I Fornè, AV Venkatasubramani and A Imhof for the targeted mass-spectrometry; M Prestel for the initial generation of the TIP60 antibody and N Steffen and K Börner for initial characterization.

Additional information

Funding

Funder	Grant reference number	Author
Deutsche Forschungsgemeinschaft	SFB1064-A1	Alessandro Scacchetti Aline Campos Sparr Silke Krause Peter B Becker
Deutsche Forschungsgemeinschaft	SFB1064-Z04	Tamas Schauer
Wellcome	103897	Patrick Heun
EMBO	ALTF 168-2018	Zivkos Apostolou

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Alessandro Scacchetti, Conceptualization, Data curation, Formal analysis, Investigation, Visualization, Methodology, Writing - original draft, Writing - review and editing; Tamas Schauer, Data curation, Software, Formal analysis, Visualization; Alexander Reim, Data curation, Formal analysis, Investigation, Methodology; Zivkos Apostolou, Aline Campos Sparr, Silke Krause, Investigation; Patrick Heun,

Supervision, Methodology, Supervised the CRISPR/Cas9 tagging; Michael Wierer, Supervision; Peter B Becker, Conceptualization, Supervision, Funding acquisition, Writing - original draft, Project administration, Writing - review and editing

Author ORCIDs
Alessandro Scacchetti https://orcid.org/0000-0002-0254-3717
Patrick Heun http://orcid.org/0000-0001-8400-1892
Peter B Becker https://orcid.org/0000-0001-7186-0372

Decision letter and Author response
Decision letter https://doi.org/10.7554/eLife.56325.sa1
Author response https://doi.org/10.7554/eLife.56325.sa2

Additional files

Supplementary files

- Source code 1. R script for analysis of AP-MS data.
- Source code 2. Java script for quantification of immunofluorescence pictures.
- Source code 3. R script for quantification of acetylated peptides.
- Source code 4. R script for analysis of acetylated peptides.
- Supplementary file 1. Excel spreadsheet containing imputed LFQ values obtained from the MaxLFQ algorithm, *limma* output and DOM-A or DOM-B specific interactors.
- Supplementary file 2. Excel spreadsheet containing result tables from DEseq2 analysis.
- Supplementary file 3. Comparison of the known subunits of SWR1- and NuA4-type complexes between *D. melanogaster*, *S. cerevisiae* and *H. sapiens*. Subunit composition of the yeast SWR1 and NuA4 were obtained from the manually-curated SGD database (https://www.yeastgenome.org) (CPX-2122 and CPX3155). For the human complexes, we refer to the EP400 complex subunits described in ***Dalvai et al., 2015*** and to the SRCAP subunits described in ***Feng et al., 2018***.
- Supplementary file 4. gRNAs, repair templates and primers used in this study.
- Supplementary file 5. Excel spreadsheet containing raw output from Skyline analysis and results from quantification of acetylated peptides.
- Transparent reporting form

Data availability

Sequencing data have been deposited in GEO under accession code GSE145738. Targeted proteomics data are available at ProteomeXchange under accession number PXD017729. Immunofluoresce images are available at Dryad under accession number https://doi.org/10.5061/dryad.1rn8pk0qt.

The following datasets were generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Scacchetti A, Schauer TR, Apostolou Z, Sparr AC, Krause S, Heun P,	2020	Drosophila SWR1 and NuA4 complexes originate from DOMINO splice isoforms	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE145738	NCBI Gene Expression Omnibus, GSE145738

Wierer M, Becker PB				
Scacchetti A, Schauer TR, Apostolou Z, Sparr AC, Krause S, Heun P, Wierer M, Becker PB	2020	Drosophila SWR1 and NuA4 complexes are defined by DOMINO isoform	http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD017729	ProteomeXchange, PXD017729
Scacchetti A, Schauer TR, Apostolou Z, Sparr AC, Krause S, Heun P, Wierer M, Becker PB	2020	Drosophila SWR1 and NuA4 complexes are defined by DOMINO isoform	https://doi.org/10.5061/dryad.1rn8pk0qt	Dryad Digital Repository, 10.5061/dryad.1rn8pk0qt

References

Adam M, Robert F, Larochelle M, Gaudreau L. 2001. H2A.Z is required for global chromatin integrity and for recruitment of RNA polymerase II under specific conditions. *Molecular and Cellular Biology* **21**:6270–6279. DOI: https://doi.org/10.1128/MCB.21.18.6270-6279.2001, PMID: 11509669

Albig C, Wang C, Dann GP, Wojcik F, Schauer T, Krause S, Maenner S, Cai W, Li Y, Girton J, Muir TW, Johansen J, Johansen KM, Becker PB, Regnard C. 2019. JASPer controls interphase histone H3S10 phosphorylation by chromosomal kinase JIL-1 in *Drosophila*. *Nature Communications* **10**:5343. DOI: https://doi.org/10.1038/s41467-019-13174-6, PMID: 31767855

Allard S, Utley RT, Savard J, Clarke A, Grant P, Brandl CJ, Pillus L, Workman JL, Côté J. 1999. NuA4, an essential transcription adaptor/histone H4 acetyltransferase complex containing Esa1p and the ATM-related cofactor Tra1p. *The EMBO Journal* **18**:5108–5119. DOI: https://doi.org/10.1093/emboj/18.18.5108, PMID: 10487762

Auger A, Galarneau L, Altaf M, Nourani A, Doyon Y, Utley RT, Cronier D, Allard S, Côté J. 2008. Eaf1 is the platform for NuA4 molecular assembly that evolutionarily links chromatin acetylation to ATP-Dependent exchange of histone H2A variants. *Molecular and Cellular Biology* **28**:2257–2270. DOI: https://doi.org/10.1128/MCB.01755-07, PMID: 18212047

Baldi S, Krebs S, Blum H, Becker PB. 2018. Genome-wide measurement of local nucleosome array regularity and spacing by nanopore sequencing. *Nature Structural & Molecular Biology* **25**:894–901. DOI: https://doi.org/10.1038/s41594-018-0110-0, PMID: 30127356

Baldi S, Becker PB. 2013. The variant histone H2A.V of *Drosophila*–three roles, two guises. *Chromosoma* **122**:245–258. DOI: https://doi.org/10.1007/s00412-013-0409-x, PMID: 23553272

Bonnet J, Lindeboom RGH, Pokrovsky D, Stricker G, Çelik MH, Rupp RAW, Gagneur J, Vermeulen M, Imhof A, Müller J. 2019. Quantification of proteins and histone marks in *Drosophila* Embryos Reveals Stoichiometric Relationships Impacting Chromatin Regulation. *Developmental Cell* **51**:632–644. DOI: https://doi.org/10.1016/j.devcel.2019.09.011

Börner K, Becker PB. 2016. Splice variants of the SWR1-type nucleosome remodeling factor domino have distinct functions during *Drosophila melanogaster* oogenesis. *Development* **143**:3154–3167. DOI: https://doi.org/10.1242/dev.139634, PMID: 27578180

Bowman GD, Poirier MG. 2015. Post-translational modifications of histones that influence nucleosome dynamics. *Chemical Reviews* **115**:2274–2295. DOI: https://doi.org/10.1021/cr500350x, PMID: 25424540

Clapier CR, Iwasa J, Cairns BR, Peterson CL. 2017. Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. *Nature Reviews Molecular Cell Biology* **18**:407–422. DOI: https://doi.org/10.1038/nrm.2017.26, PMID: 28512350

Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. 2011. Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research* **10**:1794–1805. DOI: https://doi.org/10.1021/pr101065j, PMID: 21254760

Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M. 2014. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & Cellular Proteomics* **13**:2513–2526. DOI: https://doi.org/10.1074/mcp.M113.031591, PMID: 24942700

Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **26**:1367–1372. DOI: https://doi.org/10.1038/nbt.1511, PMID: 19029910

Dalvai M, Loehr J, Jacquet K, Huard CC, Roques C, Herst P, Côté J, Doyon Y. 2015. A scalable Genome-Editing-Based approach for mapping multiprotein complexes in human cells. *Cell Reports* **13**:621–633. DOI: https://doi.org/10.1016/j.celrep.2015.09.009, PMID: 26456817

Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard R, Virgin HW, Listgarten J, Root DE. 2016. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology* **34**:184–191. DOI: https://doi.org/10.1038/nbt.3437, PMID: 26780180

Doyon Y, Selleck W, Lane WS, Tan S, Côté J. 2004. Structural and functional conservation of the NuA4 histone acetyltransferase complex from yeast to humans. *Molecular and Cellular Biology* **24**:1884–1896. DOI: https://doi.org/10.1128/MCB.24.5.1884-1896.2004, PMID: 14966270

Erdel F, Rippe K. 2018. Formation of chromatin subcompartments by Phase Separation. *Biophysical Journal* **114**: 2262–2270. DOI: https://doi.org/10.1016/j.bpj.2018.03.011, PMID: 29628210

Feller C, Forné I, Imhof A, Becker PB. 2015. Global and specific responses of the histone acetylome to systematic perturbation. *Molecular Cell* **57**:559–571. DOI: https://doi.org/10.1016/j.molcel.2014.12.008, PMID: 25578876

Feng Y, Tian Y, Wu Z, Xu Y. 2018. Cryo-EM structure of human SRCAP complex. *Cell Research* **28**:1121–1123. DOI: https://doi.org/10.1038/s41422-018-0102-y, PMID: 30337683

Gratz SJ, Ukken FP, Rubinstein CD, Thiede G, Donohue LK, Cummings AM, O'Connor-Giles KM. 2014. Highly specific and efficient CRISPR/Cas9-catalyzed homology-directed repair in *Drosophila*. *Genetics* **196**:961–971. DOI: https://doi.org/10.1534/genetics.113.160713, PMID: 24478335

Greenberg RS, Long HK, Swigut T, Wysocka J. 2019. Single amino acid change underlies distinct roles of H2A.Z subtypes in human syndrome. *Cell* **178**:1421–1436. DOI: https://doi.org/10.1016/j.cell.2019.08.002

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell* **38**:576–589. DOI: https://doi.org/10.1016/j.molcel.2010.05.004, PMID: 20513432

Hong J, Feng H, Wang F, Ranjan A, Chen J, Jiang J, Ghirlando R, Xiao TS, Wu C, Bai Y. 2014. The catalytic subunit of the SWR1 remodeler is a histone chaperone for the H2A.Z-H2B dimer. *Molecular Cell* **53**:498–505. DOI: https://doi.org/10.1016/j.molcel.2014.01.010, PMID: 24507717

Horn T, Boutros M. 2010. E-RNAi: a web application for the multi-species design of RNAi reagents–2010 update. *Nucleic Acids Research* **38**:W332–W339. DOI: https://doi.org/10.1093/nar/gkq317, PMID: 20444868

Kammers K, Cole RN, Tiengwe C, Ruczinski I. 2015. Detecting significant changes in protein abundance. *EuPA Open Proteomics* **7**:11–19. DOI: https://doi.org/10.1016/j.euprot.2015.02.002, PMID: 25821719

Klymenko T, Papp B, Fischle W, Köcher T, Schelder M, Fritsch C, Wild B, Wilm M, Müller J. 2006. A polycomb group protein complex with sequence-specific DNA-binding and selective methyl-lysine-binding activities. *Genes & Development* **20**:1110–1122. DOI: https://doi.org/10.1101/gad.377406, PMID: 16618800

Koontz JI, Soreng AL, Nucci M, Kuo FC, Pauwels P, van Den Berghe H, Dal Cin P, Fletcher JA, Sklar J. 2001. Frequent fusion of the JAZF1 and JJAZ1 genes in endometrial stromal tumors. *PNAS* **98**:6348–6353. DOI: https://doi.org/10.1073/pnas.101132598, PMID: 11371647

Kuo YM, Henry RA, Tan S, Côté J, Andrews AJ. 2015. Site specificity analysis of piccolo NuA4-mediated acetylation for different histone complexes. *Biochemical Journal* **472**:239–248. DOI: https://doi.org/10.1042/BJ20150654, PMID: 26420880

Kusch T, Florens L, Macdonald WH, Swanson SK, Glaser RL, Yates JR, Abmayr SM, Washburn MP, Workman JL. 2004. Acetylation by Tip60 is required for selective histone variant exchange at DNA lesions. *Science* **306**: 2084–2087. DOI: https://doi.org/10.1126/science.1103455, PMID: 15528408

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with bowtie 2. *Nature Methods* **9**:357–359. DOI: https://doi.org/10.1038/nmeth.1923, PMID: 22388286

Liu Z, Tabuloc CA, Xue Y, Cai Y, Mcintire P, Niu Y, Lam VH, Chiu JC, Zhang Y. 2019. Splice variants of DOMINO control *Drosophila* circadian behavior and pacemaker neuron maintenance. *PLOS Genetics* **15**:e1008474. DOI: https://doi.org/10.1371/journal.pgen.1008474, PMID: 31658266

Matsuda R, Hori T, Kitamura H, Takeuchi K, Fukagawa T, Harata M. 2010. Identification and characterization of the two isoforms of the vertebrate H2A.Z histone variant. *Nucleic Acids Research* **38**:4263–4273. DOI: https://doi.org/10.1093/nar/gkq171, PMID: 20299344

Mizuguchi G, Shen X, Landry J, Wu WH, Sen S, Wu C. 2004. ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. *Science* **303**:343–348. DOI: https://doi.org/10.1126/science.1090701, PMID: 14645854

Morillo-Huesca M, Clemente-Ruiz M, Andújar E, Prado F. 2010. The SWR1 histone replacement complex causes genetic instability and genome-wide transcription misregulation in the absence of H2A.Z. *PLOS ONE* **5**:e12143. DOI: https://doi.org/10.1371/journal.pone.0012143, PMID: 20711347

Nakajima T, Fujino S, Nakanishi G, Kim YS, Jetten AM. 2004. TIP27: a novel repressor of the nuclear orphan receptor TAK1/TR4. *Nucleic Acids Research* **32**:4194–4204. DOI: https://doi.org/10.1093/nar/gkh741, PMID: 15302918

Park JH, Sun XJ, Roeder RG. 2010. The SANT domain of p400 ATPase represses acetyltransferase activity and coactivator function of TIP60 in basal p21 gene expression. *Molecular and Cellular Biology* **30**:2750–2761. DOI: https://doi.org/10.1128/MCB.00804-09, PMID: 20351180

Peleg S, Feller C, Forne I, Schiller E, Sévin DC, Schauer T, Regnard C, Straub T, Prestel M, Klima C, Schmitt Nogueira M, Becker L, Klopstock T, Sauer U, Becker PB, Imhof A, Ladurner AG. 2016. Life span extension by targeting a link between metabolism and histone acetylation in *Drosophila*. *EMBO Reports* **17**:455–469. DOI: https://doi.org/10.15252/embr.201541132, PMID: 26781291

Perkins LA, Holderbaum L, Tao R, Hu Y, Sopko R, McCall K, Yang-Zhou D, Flockhart I, Binari R, Shim HS, Miller A, Housden A, Foos M, Randkelv S, Kelley C, Namgyal P, Villalta C, Liu LP, Jiang X, Huan-Huan Q, et al. 2015. The transgenic RNAi project at Harvard medical school: resources and validation. *Genetics* **201**:843–852. DOI: https://doi.org/10.1534/genetics.115.180208, PMID: 26320097

Pradhan SK, Su T, Yen L, Jacquet K, Huang C, Côté J, Kurdistani SK, Carey MF. 2016. EP400 deposits H3.3 into Promoters and Enhancers during Gene Activation. *Molecular Cell* **61**:27–38. DOI: https://doi.org/10.1016/j.molcel.2015.10.039, PMID: 26669263

Ranjan A, Mizuguchi G, FitzGerald PC, Wei D, Wang F, Huang Y, Luk E, Woodcock CL, Wu C. 2013. Nucleosome-free region dominates histone acetylation in targeting SWR1 to promoters for H2A.Z replacement. *Cell* **154**:1232–1245. DOI: https://doi.org/10.1016/j.cell.2013.08.005, PMID: 24034247

Redon C, Pilch D, Rogakou E, Sedelnikova O, Newrock K, Bonner W. 2002. Histone H2A variants H2AX and H2AZ. *Current Opinion in Genetics & Development* **12**:162–169. DOI: https://doi.org/10.1016/S0959-437X(02)00282-4, PMID: 11893489

Rhee DY, Cho DY, Zhai B, Slattery M, Ma L, Mintseris J, Wong CY, White KP, Celniker SE, Przytycka TM, Gygi SP, Obar RA, Artavanis-Tsakonas S. 2014. Transcription factor networks in *Drosophila melanogaster*. *Cell Reports* **8**:2031–2043. DOI: https://doi.org/10.1016/j.celrep.2014.08.038, PMID: 25242320

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**:e47. DOI: https://doi.org/10.1093/nar/gkv007, PMID: 25605792

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nature Biotechnology* **29**:24–26. DOI: https://doi.org/10.1038/nbt.1754, PMID: 21221095

Ruhf ML, Braun A, Papoulas O, Tamkun JW, Randsholt N, Meister M. 2001. The domino gene of*Drosophila*encodes novel members of the SWI2/SNF2 family of DNA-dependent ATPases, which contribute to the silencing of homeotic genes. *Development* **128**:1429–1441. PMID: 11262242

Schauer T. 2020a. ChIP-seq pipeline used for "Drosophila SWR1 and NuA4 complexes originate from DOMINO splice isoforms". *GitHub*. d5eaded.https://github.com/tschauer/Domino_ChIPseq_2020

Schauer T. 2020b. Analysis code for "Drosophila SWR1 and NuA4 complexes originate from DOMINO splice isoforms". *GitHub*. fc31618.https://github.com/tschauer/Domino_RNAseq_2020

Searle NE, Torres-Machorro AL, Pillus L. 2017. Chromatin regulation by the NuA4 acetyltransferase complex is mediated by essential interactions between enhancer of polycomb (Epl1) and Esa1. *Genetics* **205**:1125–1137. DOI: https://doi.org/10.1534/genetics.116.197830, PMID: 28108589

Talbert PB, Henikoff S. 2017. Histone variants on the move: substrates for chromatin dynamics. *Nature Reviews Molecular Cell Biology* **18**:115–126. DOI: https://doi.org/10.1038/nrm.2016.148, PMID: 27924075

Thatcher TH, Gorovsky MA. 1994. Phylogenetic analysis of the core histones H2A, H2B, H3, and H4. *Nucleic Acids Research* **22**:174–179. DOI: https://doi.org/10.1093/nar/22.2.174, PMID: 8121801

van Daal A, Elgin SC. 1992. A histone variant, H2AvD, *is essential in Drosophila melanogaster*. *Molecular Biology of the Cell* **3**:593–602. DOI: https://doi.org/10.1091/mbc.3.6.593, PMID: 1498368

Villa R, Schauer T, Smialowski P, Straub T, Becker PB. 2016. PionX sites mark the X chromosome for dosage compensation. *Nature* **537**:244–248. DOI: https://doi.org/10.1038/nature19338, PMID: 27580037

Wang X, Ahmad S, Zhang Z, Côté J, Cai G. 2018a. Architecture of the *Saccharomyces cerevisiae* NuA4/TIP60 complex. *Nature Communications* **9**:1147. DOI: https://doi.org/10.1038/s41467-018-03504-5, PMID: 29559617

Wang X, Zhu W, Chang P, Wu H, Liu H, Chen J. 2018b. Merge and separation of NuA4 and SWR1 complexes control cell fate plasticity in *Candida Albicans*. *Cell Discovery* **4**:45. DOI: https://doi.org/10.1038/s41421-018-0043-0, PMID: 30109121

Weber CM, Henikoff JG, Henikoff S. 2010. H2A.Z nucleosomes enriched over active genes are homotypic. *Nature Structural & Molecular Biology* **17**:1500–1507. DOI: https://doi.org/10.1038/nsmb.1926, PMID: 21057526

Weber CM, Ramachandran S, Henikoff S. 2014. Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. *Molecular Cell* **53**:819–830. DOI: https://doi.org/10.1016/j.molcel.2014.02.014, PMID: 24606920

Willhoft O, Ghoneim M, Lin CL, Chua EYD, Wilkinson M, Chaban Y, Ayala R, McCormack EA, Ocloo L, Rueda DS, Wigley DB. 2018. Structure and dynamics of the yeast SWR1-nucleosome complex. *Science* **362**:eaat7716. DOI: https://doi.org/10.1126/science.aat7716, PMID: 30309918

Willhoft O, Wigley DB. 2020. INO80 and SWR1 complexes: the non-identical twins of chromatin remodelling. *Current Opinion in Structural Biology* **61**:50–58. DOI: https://doi.org/10.1016/j.sbi.2019.09.002, PMID: 31838293

Wu WH, Alami S, Luk E, Wu CH, Sen S, Mizuguchi G, Wei D, Wu C. 2005. Swc2 is a widely conserved H2AZ-binding module essential for ATP-dependent histone exchange. *Nature Structural & Molecular Biology* **12**:1064–1071. DOI: https://doi.org/10.1038/nsmb1023, PMID: 16299513

Xu P, Li C, Chen Z, Jiang S, Fan S, Wang J, Dai J, Zhu P, Chen Z. 2016. The NuA4 core complex acetylates nucleosomal histone H4 through a double recognition mechanism. *Molecular Cell* **63**:965–975. DOI: https://doi.org/10.1016/j.molcel.2016.07.024, PMID: 27594449

Zhao Y, Garcia BA. 2015. Comprehensive catalog of currently documented histone modifications. *Cold Spring Harbor Perspectives in Biology* **7**:a025064. DOI: https://doi.org/10.1101/cshperspect.a025064, PMID: 26330523

3.5 Article 5: Functional identity of hypothalamic melanocortin neurons depends on Tbx3

Carmelo Quarta, Alexandre Fisette, Yanjun Xu, Gustav Colldén, Beata Legutko, Yu-Ting Tseng, **Alexander Reim**, Michael Wierer, Maria Caterina De Rosa, Valentina Klaus, Rick Rausch, Vidhu V. Thaker, Elisabeth Graf, Tim M. Strom, Anne-Laure Poher, Tim Gruber, Ophélia Le Thuc, Alberto Cebrian-Serrano, Dhiraj Kabra, Luigi Bellocchio, Stephen C. Woods, Gert O. Pflugfelder, Rubén Nogueiras, Lori Zeltser, Ilona C. Grunwald Kadow, Anne Moon, Cristina García-Cáceres, Matthias Mann, Mathias Treier, Claudia A. Doege & Matthias H. Tschöp. **Functional identity of hypothalamic melanocortin neurons depends on Tbx3.** *Nat Metab* 1, 222–235 (2019). https://doi.org/10.1038/s42255-018-0028-1

In a collaboration with Carmelo Quarta and Alexandre Fisette in the group of Matthias Tschöp at the Institute for Diabetes and Obesity, Helmholtz Center Munich, we were interested in characterizing the transcription factor Tbx3 in neuronal development and body weight control. I performed ChIP-MS of Tbx3, which is expressed only in sub-regions of the mouse hypothalami, which consequently presented a very limited source of material. Remarkably, these experiments revealed known and novel interactors of Tbx3 linked to inter- and intracellular signaling and neuronal development. The ChIP-MS and genomic data gave insights into the molecular function of Tbx3 in hypothalami. Following these initial findings our collaborators further investigated the altered neuropeptide expression, namely Pomc and Agrp, after loss of Tbx3. This led to the discovery that loss of Tbx3 in Pomc but not in Agrp neurons caused a significant increase of body weight. The publication further highlights the importance of Tbx3 in differentiation of Pomc neurons and how it maintains their identity in mature neurons. Finally, Carmelo Quarta and Alexandre Fisette showed that much of the functionality of Tbx3 is also conserved in *D. melanogaster* and human neurons. Collectively, we characterize the transcription factor Tbx3 and its role in differentiation and maintenance of neuropeptidergic neurons and, consequently, its role in body weight regulation.

ARTICLES
https://doi.org/10.1038/s42255-018-0028-1
nature metabolism

Functional identity of hypothalamic melanocortin neurons depends on Tbx3

Carmelo Quarta[1,2,3,4,22], Alexandre Fisette[1,2,22], Yanjun Xu[1,2,5], Gustav Colldén[1,2], Beata Legutko[1,2], Yu-Ting Tseng[6,7], Alexander Reim[8], Michael Wierer[8], Maria Caterina De Rosa[9], Valentina Klaus[1,2,5], Rick Rausch[9], Vidhu V. Thaker[10], Elisabeth Graf[11], Tim M. Strom[11], Anne-Laure Poher[1,2], Tim Gruber[1,2], Ophélia Le Thuc[1,2], Alberto Cebrian-Serrano[1,2], Dhiraj Kabra[1,2], Luigi Bellocchio[12,13], Stephen C. Woods[14], Gert O. Pflugfelder[15], Rubén Nogueiras[16,17], Lori Zeltser[18], Ilona C. Grunwald Kadow[19], Anne Moon[20,21], Cristina García-Cáceres[1,2], Matthias Mann[8], Mathias Treier[6,7], Claudia A. Doege[18] and Matthias H. Tschöp[1,2,5]*

Heterogeneous populations of hypothalamic neurons orchestrate energy balance via the release of specific signatures of neuropeptides. However, how specific intracellular machinery controls peptidergic identities and function of individual hypothalamic neurons remains largely unknown. The transcription factor T-box 3 (Tbx3) is expressed in hypothalamic neurons sensing and governing energy status, whereas human TBX3 haploinsufficiency has been linked with obesity. Here, we demonstrate that loss of Tbx3 function in hypothalamic neurons causes weight gain and other metabolic disturbances by disrupting both the peptidergic identity and plasticity of Pomc/Cart and Agrp/Npy neurons. These alterations are observed after loss of Tbx3 in both immature hypothalamic neurons and terminally differentiated mouse neurons. We further establish the importance of Tbx3 for body weight regulation in *Drosophila melanogaster* and show that TBX3 is implicated in the differentiation of human embryonic stem cells into hypothalamic Pomc neurons. Our data indicate that Tbx3 directs the terminal specification of neurons as functional components of the melanocortin system and is required for maintaining their peptidergic identity. In summary, we report the discovery of a key mechanistic process underlying the functional heterogeneity of hypothalamic neurons governing body weight and systemic metabolism.

Energy-sensing neuronal populations of the hypothalamic arcuate nucleus (ARC), including proopiomelanocortin (Pomc)- and agouti-related protein (Agrp)-expressing neurons, release specific neuropeptides that control energy homeostasis by modulating appetite and energy expenditure. Dysregulated activity of these neurons, which constitute key components of the melanocortin system[1], is causally linked with energy imbalance and obesity[2–4]. Considering the constantly changing input into these neurons throughout development and adult life, an intricate intracellular regulatory network must be in place to accommodate plasticity adjustments (as an adequate response to energy state) as well as maintenance of cell identity. Whether extrinsic signals can induce in vivo reprogramming of neuropeptidergic identity has not been resolved, partly because of the limited knowledge of the intracellular factors involved.

To identify genes implicated in the maintenance of ARC neuronal identity and energy-sensing function, we took advantage of cell-specific transcriptomic approaches that allow profiling of subpopulations of hypothalamic neurons under basal and metabolically stimulated conditions. We cross-referenced publicly available analysed datasets from phosphorylated ribosome profiling[5], translating ribosome affinity purification (TRAP)-based sequencing of leptin-receptor-expressing neurons[6], and single-cell sequencing[7]. We determined that the transcription factor termed Tbx3 is expressed

[1]Institute for Diabetes and Obesity, Helmholtz Diabetes Center, Helmholtz Zentrum München, Neuherberg, Germany. [2]German Center for Diabetes Research (DZD), Neuherberg, Germany. [3]INSERM, Neurocentre Magendie, Physiopathologie de la Plasticité Neuronale, U1215, Bordeaux, France. [4]University of Bordeaux, Neurocentre Magendie, Physiopathologie de la Plasticité Neuronale, Bordeaux, France. [5]Division of Metabolic Diseases, Technische Universität München, Munich, Germany. [6]Cardiovascular and Metabolic Sciences, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany. [7]Charité-Universitätsmedizin Berlin, Berlin, Germany. [8]Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Martinsried, Germany. [9]Naomi Berrie Diabetes Center, Columbia Stem Cell Initiative, Department of Pediatrics, Columbia University, New York, NY, USA. [10]Naomi Berrie Diabetes Center, Division of Molecular Genetics, Department of Pediatrics, Columbia University, New York, NY, USA. [11]Institute of Human Genetics, Helmholtz Zentrum München, Neuherberg, Germany. [12]INSERM U1215, NeuroCentre Magendie, Bordeaux, France. [13]Université de Bordeaux, NeuroCentre Magendie, Bordeaux, France. [14]University of Cincinnati College of Medicine, Department of Psychiatry and Behavioral Neuroscience, Metabolic Diseases Institute, Cincinnati, OH, USA. [15]Institute of Developmental and Neurobiology. Johannes Gutenberg-University, Mainz, Germany. [16]Department of Physiology, CIMUS, University of Santiago de Compostela-Instituto de Investigación Sanitaria, Santiago de Compostela, Spain. [17]CIBER Fisiopatología de la Obesidad y Nutrición (CIBERobn), Madrid, Spain. [18]Naomi Berrie Diabetes Center, Columbia Stem Cell Initiative, Department of Pathology and Cell Biology, Columbia University, New York, NY, USA. [19]Technical University of Munich, School of Life Sciences, ZIEL - Institute for Food and Health, Freising, Germany. [20]Department of Molecular and Functional Genomics, Geisinger Clinic, Danville PA, USA. [21]Departments of Pediatrics and Human Genetics, University of Utah School of Medicine, Salt Lake City, UT, USA. [22]These authors contributed equally: Carmelo Quarta, Alexandre Fisette. *e-mail: matthias.tschoep@helmholtz-muenchen.de

with unique abundance in hypothalamic neuronal populations critically involved in energy balance regulation, including ghrelin- and leptin-responsive cells[5,6], and that its expression is regulated by scheduled feeding[5].

Although Tbx3 is known to influence proliferation[8], fate commitment and differentiation[9–11] of several non-neuronal cell types, its functional role during neuronal development or in post-mitotic neurons located in the CNS is currently uncharted. Intriguingly, this factor appears to be selectively expressed in the ARC in the adult murine hypothalamus[12]. Moreover, *TBX3* mutations in humans have been described to cause ulnar–mammary syndrome (UMS), exhibiting hallmark symptoms theoretically consistent with ARC neuron dysfunction, including impaired puberty, deficiency in growth hormone production and obesity[13,14].

Thus, we hypothesized that Tbx3 in ARC neurons may control neuronal identity and consequently be of critical relevance for systemic energy homeostasis. To test this hypothesis, we explored the functional role of Tbx3 in both mouse and human hypothalamic neurons and investigated whether loss of neuronal Tbx3 affects systemic energy homeostasis in mice and in *Drosophila melanogaster*.

We report that Tbx3 directs postnatal fate and is critical for defining the peptidergic identity of both immature and terminally differentiated mouse melanocortin neurons, a biological process essential for the regulation of energy balance.

Results

Tbx3 expression profile in the CNS and pituitary. To characterize Tbx3 expression in the central nervous system (CNS), we generated a targeted knock-in mouse model in which the Venus reporter protein is expressed under the control of the *Tbx3* locus (*Tbx3*-Venus mice) (Supplementary Fig. 1). Two areas of the brain displayed a detectable Venus signal: the ARC (Fig. 1a) and the nucleus of the solitary tract (NTS; Supplementary Fig. 1), both of which are important in the regulation of systemic metabolism[15,16]. This hypothalamic expression pattern was confirmed via qRT–PCR (Supplementary Fig. 1) and anti-Tbx3 immunohistochemistry (Supplementary Fig. 1), using an antibody validated in house with Tbx3-deficient embryos (Supplementary Fig. 1). All Venus-positive cells in the ARC and NTS of *Tbx3*-Venus mice coexpressed Tbx3, as assessed by immunohistochemistry (Supplementary Fig. 1), and the model was further validated via Southern blot analysis (Supplementary Fig. 1), thus underlining the quality of the newly developed transgenic model.

To further address the cell-specific expression profile of Tbx3, we performed bioinformatic-based reanalysis of a publicly available single-cell RNA sequencing (RNA-seq) dataset from the ARC[7]. Our analysis demonstrated overlap of *Tbx3* with neurons expressing *Pomc*, *Agrp*, kisspeptin (*Kiss*) and somatostatin (Fig. 1b and Supplementary Fig. 1), in addition to overlapping with the transcriptional profile of tanycytes, the 'gateway' cells to the metabolic hypothalamus[17] (Supplementary Fig. 1).

Neuroanatomical analysis in *Tbx3*-Venus mice demonstrated *Tbx3* (Venus) expression in almost all ARC Pomc neurons (Fig. 1c) and NTS Pomc neurons (Supplementary Fig. 1), with a comparable pattern of expression from embryonic (embryonic day (E) 18.5) to postnatal life (postnatal day (P) 0, P4 and adults) (Fig. 1d), thus indicating that Tbx3 expression in Pomc neurons is switched on embryonically and maintained throughout adult life. As suggested by the analysis of the single-cell RNA-seq data from the cells from the arc-median eminence, a considerable fraction of *Tbx3*-positive cells do not express *Pomc*. *Tbx3* transcripts have been observed within the pituitary gland[18]. We found that *Tbx3* (Venus) expression was restricted to the posterior pituitary and that no signal was observed in Pomc-expressing cells of the anterior pituitary (Supplementary Fig. 1). Moreover, no signs of *Tbx3* (Venus) expression were detected in glial fibrillary acidic protein (GFAP)-positive astrocytes (Supplementary Fig. 1) or in microglia (Iba1-positive glial cells) (Supplementary Fig. 1), whereas a substantial number of *Tbx3* (Venus)-positive cells coexpressed the tanycyte and reactive astrocyte marker vimentin (Supplementary Fig. 1), in agreement with results from single-cell sequencing analysis (Supplementary Fig. 1).

Thus, within the CNS, Tbx3 is expressed in both neuronal and non-neuronal cells known to affect energy homeostasis.

Loss of Tbx3 in hypothalamic neurons promotes obesity. ARC neurons detect changes in energy status, via both direct and indirect sensing of circulating nutrients and hormones, and accordingly modulate their activity to maintain energy balance[16]. Overnight fasting significantly decreased hypothalamic *Tbx3* mRNA levels in the ARC in C57BL/6J mice, whereas refeeding partially restored *Tbx3* expression (Fig. 1e). This finding suggests that changes in hypothalamic Tbx3 levels are likely to be involved in the control of systemic metabolism. To test this notion, we used a viral-based approach to selectively ablate *Tbx3* via Cre-LoxP recombination (adeno-associated virus (AAV)-Cre) from the mediobasal hypothalamus (MBH) of 12-week-old *Tbx3*$^{loxP/loxP}$ littermate mice (Supplementary Fig. 2).

AAV-Cre-treated mice developed pronounced obesity over the course of 7 weeks, with elevated cumulative food intake and higher fat mass relative to control mice (AAV-green fluorescent protein (GFP)-treated *Tbx3*$^{loxP/loxP}$ mice), whereas no difference was observed in lean mass (Fig. 1f–i). Indirect calorimetry did not reveal changes in hourly uncorrected energy expenditure (Fig. 1j), nor in the relationship between total uncorrected energy expenditure and body weight, as demonstrated by analysis of covariance (ANCOVA)[19] (Fig. 1k). Although the average respiratory exchange ratio (RER) was not altered in AAV-Cre-treated mice, these mice displayed metabolic inflexibility relative to controls, as indicated by a flat RER with minimal diurnal fluctuations (Fig. 1l,m).

We next asked whether loss of function of Tbx3 selectively in either Agrp or Pomc neurons would recapitulate the obesity-prone phenotype observed in the MBH loss-of-function model. No difference in body weight, food intake, glucose tolerance, fat or lean mass was observed in littermate mice bearing a conditional deletion of *Tbx3* in Agrp-expressing neurons (*Agrp*-Cre;*Tbx3*$^{loxP/loxP}$) relative to controls (Fig. 2a–e). The quality of this previously validated[20] transgenic model was confirmed by the presence of reduced *Tbx3* mRNA levels in ARC homogenates (Supplementary Fig. 2), together with a specific decrease in Tbx3 expression within *Npy*-positive neurons (Supplementary Fig. 2).

In contrast, mice bearing *Tbx3* deletion in Pomc-expressing cells[21] (*Pomc*-Cre;*Tbx3*$^{loxP/loxP}$) displayed body weight higher than that of control littermates, independently from changes in food intake (Fig. 2f,g). They also had glucose intolerance (Fig. 2h) and increased fat and lean mass (Fig. 2I,j). Indirect calorimetry demonstrated similar hourly energy expenditure, in spite of higher body weight (Fig. 2k), and further revealed lower energy expenditure with respect to body weight in *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$ mice than controls (Fig. 2l), thus suggesting that lower systemic energy dissipation may contribute to their obese phenotype. These mice also displayed a higher average RER (Fig. 2m,n), thereby indicating that lower lipid utilization might favour the increased adiposity of these mice. A significant decrease in *Tbx3* mRNA levels was observed in the hypothalamus in *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$ mice, whereas no changes in *Tbx3* mRNA levels were detected in extra-hypothalamic sites expressing Pomc, including the pituitary and adrenals (Supplementary Fig. 2). This transgenic model was further validated via costaining between Tbx3 and a Cre-dependent membrane GFP reporter, an analysis that revealed blunted Tbx3 immunoreactivity in Cre-positive neurons of *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$ mice relative to controls (Supplementary Fig. 2). Thus, the metabolic alterations observed in this model are attributable to the specific deletion of

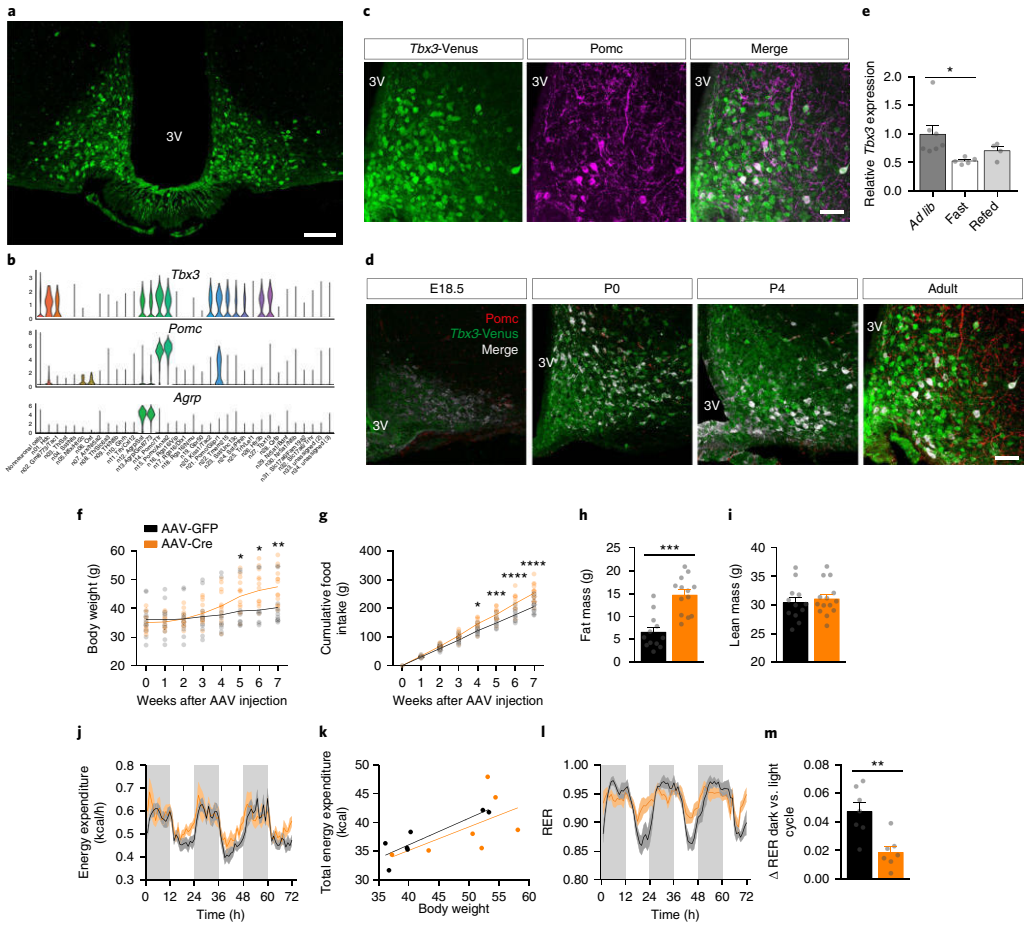


Fig. 1 | Loss of Tbx3 in hypothalamic neurons promotes obesity. a, Representative image depicting *Tbx3*-positive neurons in the ARC in *Tbx3*-Venus mice, enhanced with GFP immunohistochemistry. 3V, third ventricle. Scale bar, 100 µm. **b**, Violin plots depicting expression of *Tbx3*, *Pomc* and *Agrp* across neuronal clusters identified by Campbell et al.[7]. Of the 21,086 cells analysed, 13,079 were identified as neurons, and 8,007 were identified as non-neurons on the basis of expression of the canonical neuronal marker *Tubb3*. The width of the violin plot at different levels of the log-transformed and scaled expression levels indicates high levels of expression of Tbx3 in neuron clusters 14 (Pomc/Ttr, $n=512$), 15 (Pomc/Anxa2, $n=369$) and 21 (Pomc/Glipr1, $n=310$) compared with that of the other neuronal clusters. **c**, Colocalization between *Tbx3*-Venus and Pomc in the ARC in *Tbx3*-Venus mice, assessed by immunohistochemistry. Scale bar, 50 µm. **d**, Colocalization between *Tbx3*- and Pomc-expressing cells by immunohistochemistry in *Tbx3*-Venus mice during embryonic (E18.5), neonatal (P0, P4) and adult life (shown in **c**). **e**, Quantification of *Tbx3* mRNA levels by qRT–PCR in ARC micropunches isolated from adult (12-week-old) C57BL/6J mice after 24 h of fasting ($n=5$) or 24 h of fasting followed by 6 h of refeeding ($n=4$), relative to mice fed ad libitum ($n=7$). **f**,**g**, Body weight change (**f**) and cumulative food intake (**g**) in adult $Tbx3^{loxP/loxP}$ mice after stereotaxic injection in the MBH of AAV-Cre ($n=14$) or AAV-GFP ($n=12$) particles. **h**, Fat mass of AAV-Cre-treated ($n=13$) or AAV-GFP-treated ($n=12$) $Tbx3^{loxP/loxP}$ mice 7 weeks after surgery. **i**, Lean mass of AAV-Cre-treated ($n=14$) or AAV-GFP-treated ($n=12$) $Tbx3^{loxP/loxP}$ mice 7 weeks after surgery. **j**,**k**, Hourly energy expenditure (**j**) and total uncorrected energy expenditure correlated to body weight (**k**) in AAV-Cre-treated ($n=7$) or AAV-GFP-treated ($n=7$) $Tbx3^{loxP/loxP}$ mice 4 weeks after surgery. **l**,**m**, Hourly RER (**l**) and ΔRER averaged between night and day cycles (**m**) in AAV-Cre-treated ($n=7$) or AAV-GFP-treated ($n=7$) $Tbx3^{loxP/loxP}$ mice 4 weeks after surgery. In **k**, individual data are presented, and lines depict the fitted regression. In all other analyses, data are mean ± s.e.m. In **e**, $^{*}P=0.0476$ relative to ad libitum feeding, by analysis of variance (ANOVA) followed by Tukey's post test. In **f**, $^{*}P=0.0177$, $^{**}P=0.0095$ with ANOVA followed by Sidak's post test. In **g**, $^{**}P=0.0028$, $^{***}P=0.0001$ and $^{****}P<0.0001$ with ANOVA followed by Sidak's post test. In **h** and **m**, $^{***}P<0.0001$ and $^{**}P=0.0029$ and with a two-tailed *t* test. The experiments in **a** and **c** were repeated more than three times independently and yielded similar results. The experiments in **d** were performed once, with several samples showing similar results.

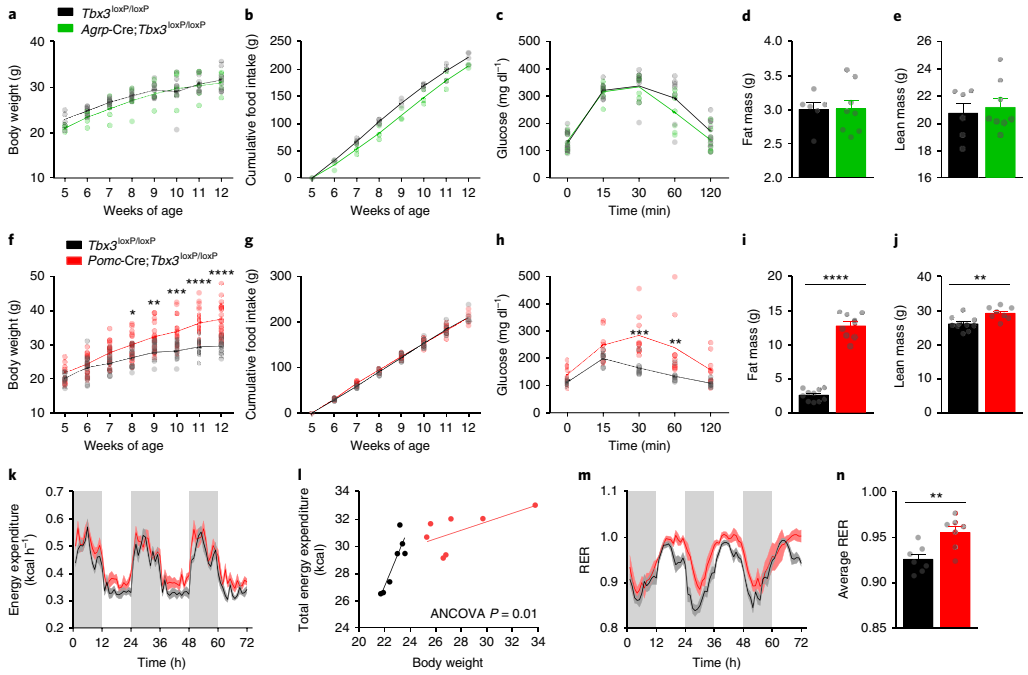


Fig. 2 | Loss of Tbx3 in Pomc but not Agrp neurons triggers obesity. a, Body weight in *Agrp*-Cre;*Tbx3*$^{loxP/loxP}$ mice ($n=5$) relative to control littermates ($n=8$). **b**, Cumulative food intake in *Agrp*-Cre;*Tbx3*$^{loxP/loxP}$ mice ($n=3$) relative to control littermates ($n=4$). **c**, Glucose tolerance test in adult *Agrp*-Cre;*Tbx3*$^{loxP/loxP}$ mice ($n=7$) relative to control littermates ($n=8$). **d**,**e**, Fat mass (**d**) and lean mass (**e**) in adult *Agrp*-Cre;*Tbx3*$^{loxP/loxP}$ mice ($n=8$) relative to control littermates ($n=6$). **f**, Body weight in *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$ mice ($n=18$) relative to control littermates ($n=11$). **g**, Cumulative food intake in *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$ mice ($n=7$) relative to control littermates ($n=7$). **h**, Glucose tolerance test in adult *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$ mice ($n=9$) relative to control littermates ($n=8$). **i**,**j**, Fat mass (**i**) and lean mass (**j**) in adult *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$ mice ($n=9$) relative to control littermates ($n=10$). **k**–**n**, Hourly energy expenditure (**k**) and energy expenditure correlated to body weight (**l**), hourly RER (**m**) and average RER values (**n**) in 7-week-old *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$ mice ($n=7$) relative to control littermates ($n=7$). Data in **a**–**k**,**m**,**n**, are mean ± s.e.m. In **f**, $^{*}P=0.02$, $^{**}P=0.003$, $^{***}P=0.0001$, $^{****}P<0.0001$ with ANOVA followed by Sidak's post test. In **h**, $^{**}P=0.001$, $^{***}P=0.0002$ with ANOVA followed by Sidak's post test. In **i**,**j**, $^{****}P<0.0001$ and $^{**}P=0.0035$ with two-tailed *t* test. In **n**, $^{**}P=0.0055$ with two-tailed *t* test.

Tbx3 in Pomc neurons located in the CNS. Collectively, these data demonstrate that ablation of *Tbx3* in ARC neurons has profound functional consequences on energy balance and that most of these metabolic alterations can be reproduced after specific deletion of this gene in Pomc-positive neurons located in the brain.

Loss of Tbx3 impairs the postnatal melanocortin system. Although Tbx3 is known to control the cell cycle and programming of highly proliferative stem cells and cancer cells[9–11,22], its functional role in neurons has remained unexplored. To investigate possible biological mechanisms underlying the metabolic phenotypes observed, we performed *Tbx3*-focused RNA sequencing and proteomic analyses in hypothalamic tissue as well as in primary hypothalamic cultures. The effect of *Tbx3* deletion on transcription in hypothalamic neurons was assessed by using primary neurons isolated from *Tbx3*$^{loxP/loxP}$ mice and infected with adenoviral (Ad) particles carrying the coding sequence for Cre recombinase (Ad-Cre) or GFP (Ad-GFP) as a control (Supplementary Fig. 3), an approach that effectively allows knockdown of *Tbx3* (Supplementary Fig. 3) in the absence of cell toxicity (Supplementary Fig. 3). Because we had found the most important in vivo metabolic effects with *Tbx3* deletion uniquely within *Pomc*-expressing cells, we performed RNA sequencing of the wild-type (WT) and *Tbx3*-knockout (*Tbx3*-KO) primary hypothalamic cultures and identified genes that were both differentially expressed in this in vitro model and known to be expressed in Pomc neurons. This analysis highlighted 449 transcripts that were differentially expressed (243 downregulated and 206 upregulated). Unbiased pathway analysis revealed that *Tbx3* deletion significantly downregulated the expression of genes controlling cellular proliferation, differentiation and determination of cellular fate (Supplementary Fig. 3). In turn, several genes linked with intracellular metabolic pathways were upregulated, albeit in a less significant way (Supplementary Fig. 3). To complement this unbiased approach, in silico analysis of the genomic loci coding for Pomc, Cart and Agrp for potential Tbx3-binding sites (T-box-binding motifs) was performed[23]. Potential Tbx3-binding sites were found in all three genes, thus suggesting that Tbx3 altered their transcription directly (Supplementary Fig. 3). To further explore the molecular machinery linked with Tbx3 in hypothalamic neurons, we performed immunoprecipitation of Tbx3 from adult C57BL/6J mouse hypothalami, then used mass spectrometry to identify Tbx3-interacting proteins. We identified 142 proteins that were

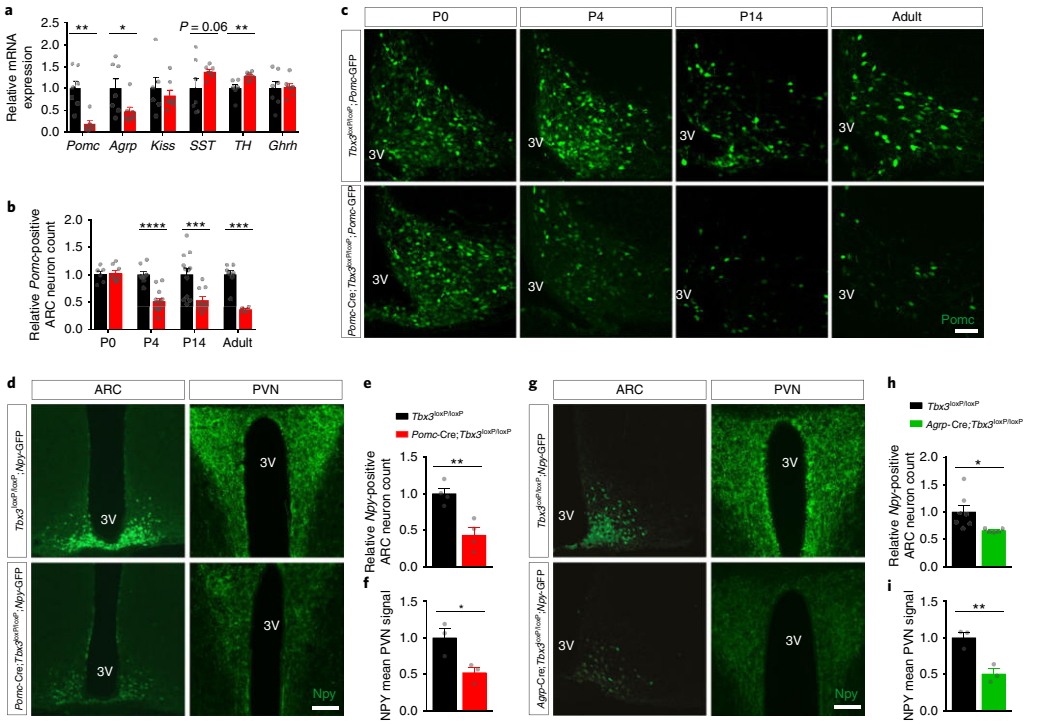


Fig. 3 | Loss of Tbx3 impairs the postnatal melanocortin system. a, Quantification of enzyme and neuropeptide mRNA levels by qRT-PCR in ARC micropunches isolated from adult (12-week-old) *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$ mice ($n=7$) and control *Tbx3*$^{loxP/loxP}$ littermates ($n=7$). *Kiss*, kisspeptin; *SST*, somatostatin; *TH*, tyrosine hydroxylase; *Ghrh*, growth-hormone-releasing hormone. **b**,**c**, Quantification (**b**) and representative images (**c**) of the relative number of *Pomc*-expressing neurons in the ARC in *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;*Pomc*-GFP mice and in control littermates (*Tbx3*$^{loxP/loxP}$;*Pomc*-GFP) at different stages of neonatal life and in adult animals. *Tbx3*$^{loxP/loxP}$;*Pomc*-GFP: $n=6$ (P0); $n=8$ (P4); $n=12$ (P14); $n=7$ (adult). *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;*Pomc*-GFP: $n=8$ (P0); $n=13$ (P4); $n=10$ (P14); $n=4$ (adult). **d**,**e**, Representative images (**d**) and relative quantification (**e**) of *Npy*-positive neurons in the ARC in adult *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;*Npy*-GFP mice ($n=4$) and control littermates (*Tbx3*$^{loxP/loxP}$;*Npy*-GFP, $n=4$). **f**, Npy-positive neuronal fibres in the PVN of adult *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;*Npy*-GFP mice ($n=3$) and control littermates (*Tbx3*$^{loxP/loxP}$;*Npy*-GFP, $n=3$). **g**,**h**, Representative images (**g**) and relative quantification (**h**) of *Npy*-positive neurons in the ARC in adult *Agrp*-Cre;*Tbx3*$^{loxP/loxP}$;*Npy*-GFP mice ($n=5$) and control littermates (*Tbx3*$^{loxP/loxP}$;*Npy*-GFP, $n=7$). **i**, Npy-positive neuronal fibres in the PVN of adult *Agrp*-Cre;*Tbx3*$^{loxP/loxP}$;*Npy*-GFP mice ($n=3$) and control littermates (*Tbx3*$^{loxP/loxP}$;*Npy*-GFP, $n=3$). 3V, third ventricle. Scale bar in **c**, 50 µm; scale bars in **d**–**g**, 100 µm. Data **a**,**b**,**e**,**f**,**h** and **i**, are mean ± s.e.m. In **a**, $^{**}P=0.0017$ (*Pomc*), $^{**}P=0.0097$ (*TH*), $^{*}P=0.0049$ with two-tailed *t* test. In **b**, $^{****}P<0.0001$ (P4), $^{***}P=0.0032$ (P14), $^{***}P=0.0002$ (adult) with two-tailed *t* test. In **e**,**f**, $^{**}P=0.0043$, $^{*}P=0.034$ with two-tailed *t* test. In **h**,**i**, $^{*}P=0.04$, $^{**}P=0.0087$ with two-tailed *t* test. The experiments in **c** were repeated more than three independent times and yielded similar results. The experiments in **d** and **g** were repeated two independent times and yielded similar results.

significantly enriched by Tbx3 precipitation (Supplementary Fig. 3 and Supplementary Table 2), including previously known Tbx3 interactors such as Kif21 (ref. [24]), AES[25] and Tollip[26]. Pathway analysis of these interacting proteins highlighted their roles in several processes, notably including inter- and intracellular signalling and neuronal development (Supplementary Fig. 3). These genomic and proteomic data led us to test the hypothesis that a lack of Tbx3 in the ARC might interfere with the cellular fate and differentiation stage of these neurons and therefore affect their peptidergic profile in addition to potentially affecting neuropeptide generation via direct transcriptional actions.

Accordingly, we measured *Pomc* and *Agrp* mRNA expression in WT and *Tbx3*-KO primary hypothalamic neurons through qRT–PCR. Both transcripts were significantly downregulated after Ad-Cre-mediated *Tbx3* deletion (Supplementary Fig. 4). These changes were reproducible in vivo, because we found significantly lower expression levels of *Pomc* and *Agrp* mRNA in the ARC in *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$ mice than in control littermates (Fig. 3a). No changes in *Kiss* or growth-hormone-releasing hormone (*Ghrh*) mRNA levels were observed in these animals, whereas the mRNA levels of tyrosine hydroxylase were elevated, and there was a trend toward elevated levels of somatostatin (Fig. 3a). To explore whether these changes in the peptidergic expression profile were caused by neurodevelopmental alterations, *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$ mice were crossed with *Pomc*-GFP reporter animals to precisely quantify *Pomc*-expressing cells during both embryonic and postnatal life, when ARC-Pomc neurons are generated and acquire their terminal peptidergic identity[27,28]. No difference in Pomc neuronal cell

number was detected in this model at E14.5, E15.5, or E18.5, thus implying normal neuronal generation in utero (Supplementary Fig. 4). No change in Pomc counts was observed at P0, whereas a substantial decrease in the number of Pomc-positive neurons was found at P4, and this relative decrement remained at P14 and 12 weeks (adult) (Fig. 3b,c). Despite progressive loss of *Pomc* expression at P2 and P4, no significant apoptotic activity was observed in this region (Supplementary Fig. 4), nor did we detect any proliferation leading to new Pomc-positive neurons between P0 and P3, as assessed with BrdU (Supplementary Fig. 4), thus confirming that most Pomc neurons are generated during embryonic life[28] and suggesting that neurogenesis and/or cellular turnover do not contribute to the Tbx3-mediated control of *Pomc* expression observed during neonatal life. Furthermore, no compensatory change was observed in the Pomc-processing enzymes of *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$ mice (Supplementary Fig. 4). Collectively, these data demonstrate that constitutive loss of Tbx3 in Pomc-expressing neurons undermines the melanocortin system, probably by interfering with the proper terminal differentiation of this neuronal population during postnatal life and possibly via direct transcriptional actions.

Loss of Tbx3 alters the peptidergic profile of Agrp neurons. In agreement with the results from the mRNA analysis documenting decreased ARC *Agrp* mRNA (Fig. 3a), *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$ mice displayed a diminished number of neuropeptide Y (Npy)-expressing neurons (co-expressed in most Agrp neurons[29]) in the ARC (Fig. 3d,e). This finding was also reflected by reduced Npy projection density in the paraventricular nucleus of the hypothalamus (PVN) (Fig. 3d–f), as demonstrated by crossing *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$ mice with *Npy*-GFP reporter mice. Because a substantial fraction of Agrp and Npy neurons are derived from Pomc-expressing cells[27], Cre-mediated ablation of *Tbx3* in these cells may interfere with *Agrp* and *Npy* expression in this animal model, thus suggesting that Tbx3's action in Agrp-expressing neurons may be similarly implicated in controlling the peptidergic profile of this specific neuronal subpopulation. To test this hypothesis, we crossed *Agrp*-Cre;*Tbx3*$^{loxP/loxP}$ mice with *Npy*-GFP reporter mice and quantified the number of *Npy*-positive neurons and their neuronal projections. Significantly fewer *Npy*-positive neurons in the ARC (Fig. 3g,h) and less Npy immunoreactivity in the PVN (Fig. 3g–i) were observed in *Agrp*-Cre;*Tbx3*$^{loxP/loxP}$;*Npy*-GFP mice than in littermate controls, as well as a significant decrease in ARC *Agrp* mRNA levels (Supplementary Fig. 4). Thus, Tbx3 action in hypothalamic ARC neurons controls the peptidergic expression profiles of different neuronal subpopulations.

Tbx3 is critical for the differentiation of Pomc neurons. To further delineate the process underlying Tbx3-mediated control of neuropeptide expression, we used a cell lineage approach and crossed *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$ mice with ROSA$^{mT/mG}$ reporter mice to genetically and permanently label cells undergoing Cre-mediated recombination (via the *Pomc*-Cre driver) as well as their neuronal projections. We then quantified Pomc expression and assessed its colocalization with GFP, which was indicative of Cre-mediated recombination. The P4 *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;ROSA$^{mT/mG}$ pups had a significantly smaller number of Pomc-positive cells than controls (Fig. 4a,b; raw counts available in Supplementary Table 3), thus reproducing our previously obtained results (Fig. 3b,c). However, no change was observed in the number of neurons or in neuronal-fibre density by analysing Cre-recombined (GFP-expressing) cells (Fig. 4a–c). These data are in agreement with the absence of apoptotic events at P2–4 (Supplementary Fig. 4) and demonstrate that loss of Tbx3 function in Pomc-expressing cells does not affect cellular survival or neuronal architecture during embryonic or early postnatal development. Instead, most Cre-recombined neurons in *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;ROSA$^{mT/mG}$ mice lacked Pomc immunoreactivity (Fig. 4a,b, arrows), thus suggesting that Tbx3 ablation in Pomc-positive cellular populations disrupts their normal peptidergic identity. Such an alteration in Pomc neuronal identity in *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;ROSA$^{mT/mG}$ was also observed in adult animals (Fig. 4d,e, arrows; raw counts available in Supplementary Table 3). Similarly, Cre-recombined cells in *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;ROSA$^{mT/mG}$ had lower expression of Cart than controls, thus indicating that the peptidergic alterations in this model are not limited to Pomc (Supplementary Fig. 5; raw counts available in Supplementary Table 3). A slight decrease in the number of Cre-recombined cells and in neuronal fibre density in the ARC and PVN was observed in adult *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;ROSA$^{mT/mG}$ mice compared with controls (Fig. 4d–i). We hypothesize that this finding is indicative of cellular loss in *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;ROSA$^{mT/mG}$ mice during adulthood, because this phenomenon occurred only after the peptidergic identity impairment observed at P4. We speculate that Tbx3 deletion in hypothalamic Pomc neurons may impair neuronal maturation during postnatal life, which might in turn provoke cell death in a subpopulation of neurons during the transition into adult life. However, these results could also be linked with decreased postnatal neurogenesis and/or impaired neuronal turnover of Pomc-positive cells in *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;ROSA$^{mT/mG}$ mice, perhaps linked with the condition of obesity observed in these animals. The concept of postnatal hypothalamic neurogenesis, however, remains controversial[30]. These data collectively demonstrate that Tbx3 has a fundamental role in maintaining the identity of ARC Pomc-expressing cells, a process that underlies changes in the neuropeptidergic profile of these neurons and consequently in systemic energy homeostasis.

Tbx3 controls the identity and plasticity of mature Pomc neurons. Because ARC *Tbx3* levels are modulated by nutritional status in mice (Fig. 1e), we asked whether Tbx3 in hypothalamic Pomc neurons might be implicated in the previously observed plastic ability of these cells to adjust Pomc expression and release in response to changes in nutritional status[31]. *Pomc*-positive cells and Pomc immunoreactivity were measured in adult *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;*Pomc*-GFP and control animals in the ad libitum–fed condition and after exposure to a fasting–refeeding paradigm. In controls, fasting reduced *Pomc*-positive cell counts (Fig. 4j,k) and Pomc immunoreactivity (Supplementary Fig. 5) relative to what occurred in ad libitum–fed mice, whereas refeeding normalized *Pomc* expression, as previously reported[31]. In contrast, changes in nutritional status did not alter *Pomc* expression in adult *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;*Pomc*-GFP mice (Fig. 4j,k and Supplementary Fig. 5), thus implicating Tbx3 in fine-tuning Pomc expression in response to energy needs. These data also suggest that Tbx3 is likely to control the peptidergic profile of fully differentiated hypothalamic neurons in adult mice. To assess this possibility, we quantified Pomc-positive neurons in our adult-onset model of viral-mediated hypothalamic Tbx3 deletion. A prominent decrease in ARC Pomc-positive cells was observed (Fig. 4l,m), with no changes in apoptotic events (Supplementary Fig. 5), thus implying that loss of Tbx3 in fully mature and specified neurons alters their peptidergic identity. To uncover whether such an alteration might underlie hyperphagia, and therefore the obese phenotype observed in AAV-Cre-treated mice, we challenged these animals with intracerebroventricular (ICV) injections of the biologically active Pomc-derived peptide alpha-melanocyte-stimulating hormone (α-MSH) at a subeffective dose. ICV injection of this dose of α-MSH had a slight, non-significant hypophagic effect in control (AAV-GFP) mice. In contrast, this approach significantly normalised food intake in AAV-Cre-treated animals to the level of control AAV-GFP mice (Fig. 4n). Together, these results indicate that *Tbx3* knockdown in fully differentiated ARC neurons impairs their peptidergic expression profile under non-stimulated

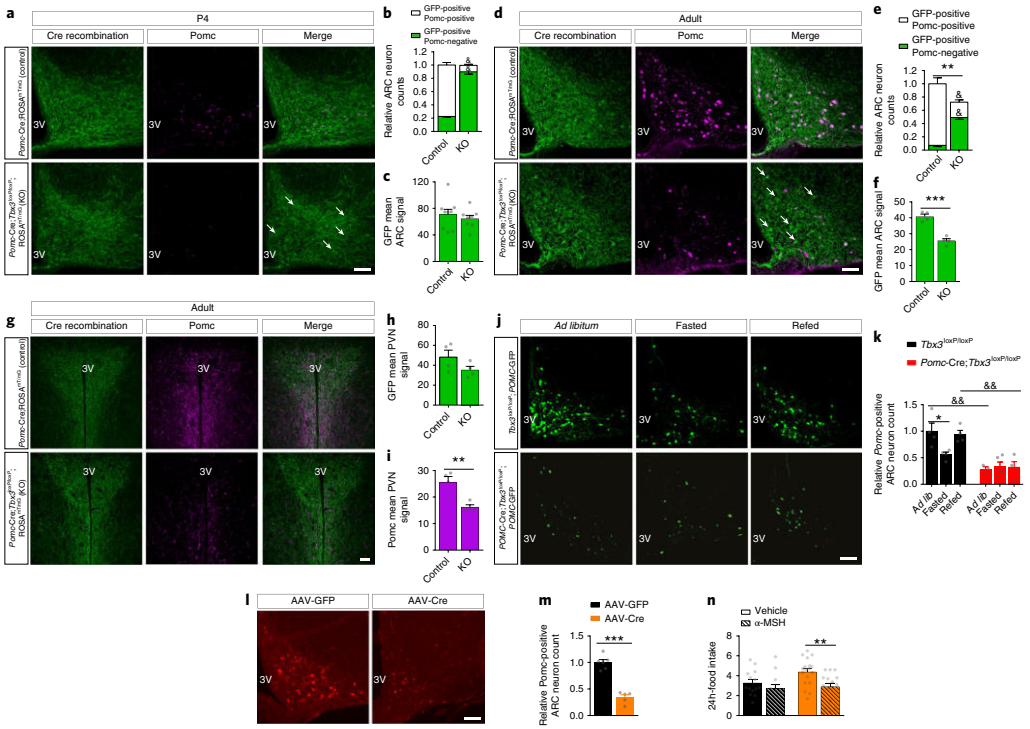


Fig. 4 | Tbx3 is critical for the differentiation of Pomc neurons. a,**b**, Representative images (**a**) and relative quantification (**b**) of GFP-expressing neurons (Cre recombination) and Pomc-positive neurons in the ARC in P4 *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;ROSA$^{mT/mG}$ mice ($n=8$) relative to controls (*Tbx3*$^{loxP/loxP}$;ROSA$^{mT/mG}$, $n=9$), assessed by immunohistochemistry. Arrows depict GFP-positive/Pomc-negative cells. **c**, Relative densitometric analysis of Cre recombination (GFP immunoreactivity) in the ARC in P4 *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;ROSA$^{mT/mG}$ mice ($n=8$) and controls ($n=9$). **d**,**e**, Representative images (**d**) and relative quantification (**e**) of GFP-expressing neurons (Cre recombination) and Pomc-positive neurons in the ARC in adult (12-week old) *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;ROSA$^{mT/mG}$ mice ($n=4$) and controls ($n=4$), assessed by immunohistochemistry. Arrows depict GFP-positive/Pomc-negative cells. **f**, Relative densitometric analysis of Cre recombination (GFP immunoreactivity) in the ARC in adult *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;ROSA$^{mT/mG}$ mice ($n=4$) and controls ($n=4$). **g**, Representative images depicting Cre recombination (GFP immunoreactivity) and Pomc-positive neuronal fibres in the PVN in adult *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;ROSA$^{mT/mG}$ mice and controls, assessed by immunohistochemistry. **h**,**i**, Relative densitometric analysis of Cre recombination (GFP immunoreactivity) (**h**) and Pomc immunoreactivity (**i**) in the PVN in adult *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;ROSA$^{mT/mG}$ mice ($n=4$) and controls ($n=4$). **j**,**k**, Representative image (**j**) and cell number quantification (**k**) of *Pomc*-positive neurons in the ARC in adult *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;*Pomc*-GFP mice or control littermates (*Tbx3*$^{loxP/loxP}$;*Pomc*-GFP) after 15 h of fasting with or without 2 h of refeeding. *Tbx3*$^{loxP/loxP}$;*Pomc*-GFP: $n=4$ for each condition. *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;*Pomc*-GFP: $n=3$ (ad libitum); $n=5$ (fasted), $n=4$ (refed). **l**,**m**, Representative images (**l**) and relative quantification (**m**) of Pomc-expressing neurons in the ARC in adult *Tbx3*$^{loxP/loxP}$ mice 7 weeks after AAV-Cre ($n=5$) or AAV-GFP ($n=6$) MBH injection. **n**, 24-h food intake measured in adult *Tbx3*$^{loxP/loxP}$ mice 7 weeks after AAV-Cre or AAV-GFP MBH injection, after intracerebroventricular administration of vehicle or αMSH. AAV-Cre: $n=18$ (vehicle); $n=15$ (αMSH). AAV-GFP: $n=14$ (vehicle); $n=12$ (αMSH). 3V, third ventricle. Scale bars in **a**,**d**,**g**,**j** and **l**, 50 µm. Data are mean ± s.e.m. In **b** and **e**, $^{\&}P<0.0001$ for comparisons of GFP-positive/Pomc-positive or GFP-positive/Pomc-negative subpopulation counts between *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;ROSA$^{mT/mG}$ mice and controls, $^{**}P=0.0025$ for comparison between total number of Cre-recombined neurons of *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;ROSA$^{mT/mG}$ mice and controls, with ANOVA followed by Sidak's post test. In **f** and **i**, $^{***}P=0.0003$ and $^{**}P=0.0071$ with two-tailed *t* test. In **k**, $^{*}P=0.04$, $^{\&\&}P=0.011$ comparing ad libitum–fed *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;ROSA$^{mT/mG}$ mice versus ad libitum fed controls; $^{\&\&}P=0.027$ comparing refed *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$;ROSA$^{mT/mG}$ mice versus refed controls, by ANOVA followed by Tukey's post test. In **m**, $^{***}P<0.0001$ with two-tailed *t* test. In **n**, $^{**}P=0.0061$ by ANOVA followed by Tukey's post test. The experiments in **a** were repeated two independent times and yielded similar results. The experiments in **d**,**g**,**j** were performed one time with several samples showing similar results. The experiments in **l** were repeated two independent times and yielded similar results.

conditions and undermines the ability of Pomc neurons to adjust Pomc expression and release in response to changes in nutritional status. These alterations in turn provoke dysregulated central melanocortin tone, a blunted neuronal response to the organism's nutritional status, and ultimately obesity.

Tbx3 functions are conserved in *Drosophila* and human neurons. The T-box family of transcription factors is remarkably conserved among species[32]. In *Drosophila melanogaster*, a Tbx3 homologue protein is encoded by the gene *omb* (or *bifid*). Omb is expressed in the CNS in adult flies, as assessed by double immunohistochemistry

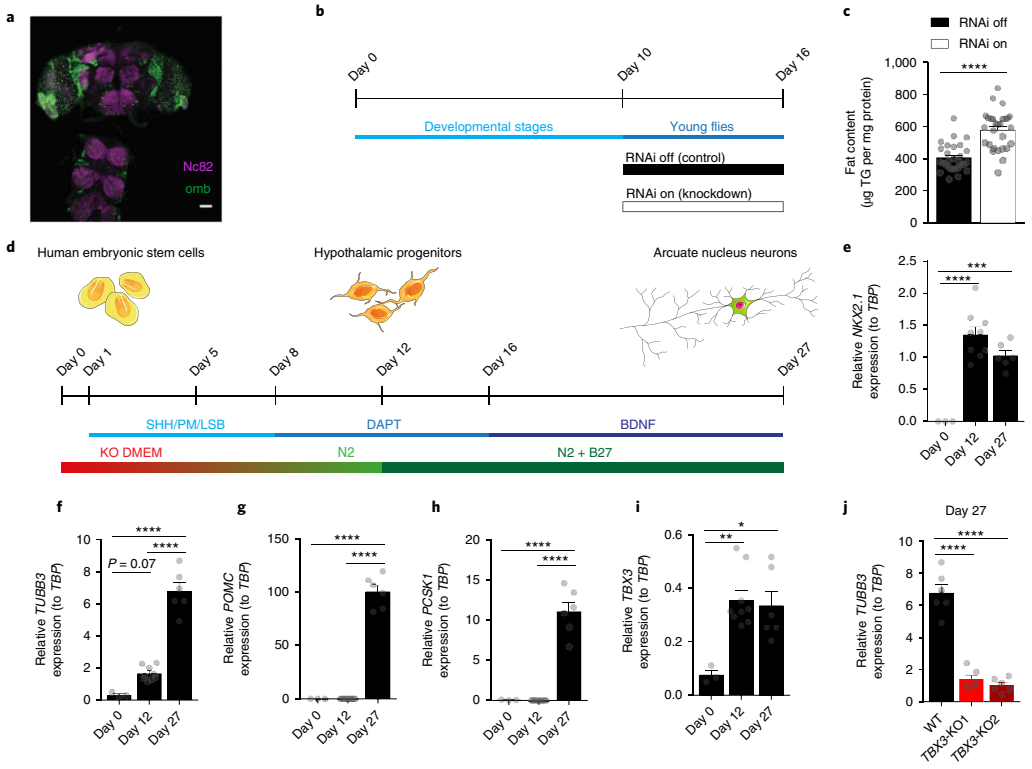


Fig. 5 | Tbx3 functions in *Drosophila* and human neurons. a, Representative image depicting expression of the *Drosophila Tbx3* orthologue omb (omb expression assessed via GFP in ombP3-Gal4>GFP flies) and Nc82 (neuronal marker) in the central nervous system of *Drosophila melanogaster*. Scale bar, 50 µm. **b**, Timeline of RNA interference (RNAi) knockdown of *omb* (RNAi on), and control flies (RNAi off). **c**, Quantification of *Drosophila* body-fat content after knockdown of omb (RNAi on, $n = 28$) compared with controls (RNAi off, $n = 28$) using the omb-RNAi line 1. **d**, Differentiation of human ESC into hypothalamic arcuate-like neurons. The combination of dual SMAD inhibition (L, LDN193189, 2.5 µM; SB, SB431542, 10 µM), early activation of sonic hedgehog (SHH) signalling (100 ng ml^{-1} SHH; SHH agonist PM, purmorphamine, 2 µM) and a step-wise switch from ESC medium (KO DMEM) to neural progenitor medium (N2) followed by inhibition of Notch signalling (DAPT, 10 µM) converts hESC into hypothalamic progenitors. For neuronal maturation, cells were cultured in neuronal medium (N2 + B27), treated with DAPT and subsequently exposed to brain-derived neurotrophic factor (BDNF, 20 ng ml^{-1}). **e**–**i**, Gene expression analyses of *NKX2.1* (**e**), *TUBB3* (**f**), *POMC* (**g**), *PCSK1* (**h**) and *TBX3* (**i**) over the time course of differentiation of ESC into hypothalamic neurons, determined by qRT-PCR. **j**, Gene expression analysis by qRT-PCR of *TUBB3* in wild-type (WT) human ESC clones and in *TBX3* knockout (*TBX3*-KO1 and *TBX3*-KO2) cell lines at ARC-like neurons (day 27) stage. Data are mean ± s.e.m. In **e**–**i**, $n = 3$ (day 0), $n = 9$ (day 12), $n = 6$ (day 27). In **j**, $n = 6$ per group. In **c**, $^{****}P < 0.0001$ with two-tailed *t* test. In **e**, $^{****}P < 0.0001$ and $^{***}P = 0.0005$ with ANOVA followed by Tukey's post test. In **f**–**i**, $^{*}P = 0.01$, $^{**}P = 0.0039$ and $^{****}P < 0.0001$ with ANOVA followed by Tukey's post test. In **j**, $^{****}P < 0.0001$ with ANOVA followed by Dunnett's post test. The experiment in **a** was repeated two independent times with similar results.

between omb and the synaptic marker bruchpilot (labelled by the Nc82 antibody) (Fig. 5a and Supplementary Fig. 6). To address whether neuronal Tbx3 action on energy homeostasis might be conserved in *Drosophila*, we generated flies bearing an inducible nervous-system-specific *omb*-knockdown system (Fig. 5b). Relative to the restults for controls (RNAi off), knockdown of *omb* (RNAi on) induced a significantly higher body-fat content (Fig. 5c). These results were reproduced in a second transgenic *Drosophila* model by using a different *omb* RNAi targeted sequence (Supplementary Fig. 6).

To determine whether Tbx3 loss-of-function phenotypes could be recapitulated in a relevant human neurocellular model system, we investigated the role of TBX3 in the control of differentiation and the peptidergic profile of human hypothalamic neurons. H9 human embryonic stem cells (hESC; WA09; WiCell) were differentiated into ARC-like neurons over the course of 27 d (Fig. 5d), as previously described[33–35]. In this in vitro human hypothalamic neuronal model, *NKX2.1* expression was observed by day 12 of differentiation, corresponding to the hypothalamic progenitor stage (Fig. 5e). Low-level expression of class III β-tubulin (*TUBB3*), a neuronal differentiation marker, occurred by day 12 and reached a maximum at day 27 (Fig. 5f). Expression of *POMC* and its processing enzyme proprotein convertase subtilisin/kexin type 1 (*PCSK1*) was detected after neuronal maturation at day 27 (Fig. 5g,h). *TBX3* expression

was observed in this model at day 12 of differentiation, corresponding to the NKX2.1-positive hypothalamic progenitor stage, and *TBX3* levels remained stable in differentiated ARC-like neurons, as obtained on day 27 (Fig. 5i).

To assess the effect of *TBX3* deletion on human hypothalamic neuronal differentiation, we generated two independent *TBX3*-KO hESC lines by using CRISPR–Cas9 (Supplementary Fig. 6). Despite efficient *TBX3* ablation (Supplementary Fig. 6), no change in the hypothalamic progenitor marker *NKX2.1* was observed at day 12 in either *TBX3*-KO line (Supplementary Fig. 6), thus suggesting normal differentiation into hypothalamic progenitors. At day 27, *NKX2.1* as well as *TUBB3*, the marker for neuronal differentiation, were greatly diminished in *TBX3*-KO cells compared with WT cells (Supplementary Fig. 6 and Fig. 5j, respectively), a result indicative of an impaired neuronal maturation state in the *TBX3*-KO condition in this in vitro human neurocellular model system. In silico analysis of the genomic loci of genes encoding human POMC, CART and AGRP for potential Tbx3-binding sites (T-box-binding motifs) revealed, as in mice, Tbx3-binding sites in all three genomic loci (Supplementary Fig. 6). However, because the strong decrease in *TUBB3* in the absence of TBX3 indicated that some hypothalamic differentiation programmes were halted, further analysis of expression levels for neuropeptides such as *POMC* was precluded.

Together, these data reveal that TBX3 is essential for the maturation of hypothalamic progenitors into ARC-like POMC-expressing neurons. Furthermore, our data suggest that Tbx3 has a conserved role in the regulation of energy homeostasis in invertebrates and mammals, including humans, although the molecular and cellular underpinnings might differ across different species.

Discussion

The heterogeneity of hypothalamic ARC neurons allows for rapid and precise physiological adaptation to changes in body energy status and is thus highly relevant for adequate maintenance of energy homeostasis. Although several transcriptional nodes are known to establish hypothalamic neuronal identity by controlling early neurogenesis and cellular fate during embryonic life[28,36–38], the molecular programme driving the terminal specification and identity maintenance of ARC neurons during postnatal life remains incompletely understood; some advances have identified Islet-1 (refs. [39,40]), Bsx[41] and microRNAs[42] as crucial regulators.

In the present experiments, we demonstrate that the transcription factor Tbx3 is required for terminal specification of hypothalamic ARC melanocortin neurons during neonatal development and is also required for the normal maintenance and plasticity of their peptidergic programme throughout adulthood.

Our work highlights a previously uncharacterized role of Tbx3 in the regulation of energy metabolism. The brain expression profile and the functional data presented reveal that Tbx3 action in hypothalamic neurons contributes to the CNS-mediated control of systemic metabolism. Loss of Tbx3 in Pomc-expressing neurons during development causes glucose intolerance and obesity secondary to decreased energy expenditure and lipid utilization in adult mice.

These metabolic alterations are accompanied by a massive decrease in the number of Pomc-expressing neurons during postnatal life, independently of changes in cell number, which probably underlies the observed obesity phenotype. In agreement, neonatal Pomc neuronal ablation promotes similar metabolic alterations[43]. Intriguingly, constitutive loss of Tbx3 specifically in Agrp/Npy-co-expressing neurons does not translate into phenotypic metabolic changes, although there is a significant decrease in *Agrp* and *Npy* expression. Such a lack of metabolic alterations in this model is probably the result of compensatory developmental mechanisms masking the ability of Agrp and Npy to modulate systemic metabolism[44], a phenomenon previously observed after neonatal Agrp/Npy neuronal ablation[45,46]. Thus, Tbx3 affects systemic energy homeostasis by controlling the peptidergic identity profile of different populations that directly modulate the activity of the melanocortin system in ARC neurons during neonatal life, when maturation of the melanocortin system occurs[28,47].

Importantly, *Tbx3* deletion in fully mature adult hypothalamic ARC neurons selectively decreases the number of Pomc-expressing neurons, a phenotype mimicking the observations in mice with *Pomc*-promoter-driven deletion of *Tbx3* from the genome at mid-term developmental stages. This translates into dysregulated central melanocortin tone that is in turn linked to hyperphagia, alterations in systemic lipid oxidation capacity and obesity. All of these findings are in agreement with the physiological role of Pomc neurons and the central melanocortin system during adulthood[48,49]. Thus, Tbx3 not only is required for establishing Pomc identity during neonatal life but also is likely to play a key role in maintaining the peptidergic identity and functional activity of fully differentiated ARC neurons.

The cellular and metabolic effects provoked by *Tbx3* ablation in hypothalamic ARC neurons are independent of neuronal survival and/or turnover, as demonstrated by our cell-lineage tracing approach. Instead, Tbx3 seems to direct intracellular programmes controlling the neuronal differentiation state, in agreement with previous studies linking Tbx3 intracellular activity with differentiation and cell fate commitment in non-neuronal cells[9–11]. Whether Tbx3 loss of function in immature and/or fully differentiated hypothalamic neurons may induce cellular reprogramming and a peptidergic identity switch is a compelling hypothesis requiring further scrutiny, but it is supported by evidence of neuronal developmental plasticity within the mammalian CNS[50–52]. In this context, our in silico–based prediction of Tbx3-binding sites suggests that the observed changes in the peptidergic identity profiles might also be explained by direct transcriptional effects in *Pomc*, *Cart* and *Agrp* genomic loci. However, a more comprehensive and unbiased analysis, such as by chromatin immunoprecipitation followed by high-throughput sequencing, will be required to directly test this hypothesis. Similarly, a detailed characterization of the molecular machinery controlled by Tbx3 in hypothalamic neurons will be necessary to elucidate the main intracellular mechanisms underlying the metabolic effects observed. Our profiling of genes and proteins linked with Tbx3 does not allow them to be causally linked with the metabolic changes observed, but this initial effort may spur future research addressing the role of such Tbx3-linked machinery in the context of obesity. It will also be of paramount importance to determine whether Tbx3 influences neuropeptidergic profiles and systemic metabolism via interactions with known metabolic signals implicated in neuronal specification, such as neurogenin 3, Mash1, OTP or Islet-1 (refs. [37–39]).

Our observations in *Drosophila melanogaster* suggest that the link between neuronal Tbx3 action and systemic energy homeostasis is probably evolutionarily conserved; however, our data do not enable understanding of the cellular and molecular mechanisms underlying the obese-like phenotype observed in flies or whether these mechanisms are conserved across different species. Because *Drosophila* does not express Pomc, Agrp or any homologue peptide, another neuronal population might link Tbx3 action with adiposity regulation in this species. Intriguingly, our data show that TBX3 is essential for the maturation of hypothalamic progenitors into ARC-like human neurons. Because human subjects with *TBX3* mutations display pathological conditions consistent with ARC neuronal dysfunction (obesity, impaired GHRH release and alterations in reproductive capacity[13,14]), we speculate that mutations affecting *TBX3* in humans might undermine ARC neuronal differentiation status and/or peptidergic profiles, changes that ultimately affect body weight regulation, reproduction and growth. Thus, our findings might have implications for human pathophysiology.

Neurons sensitive to the orexigenic hormone ghrelin express Tbx3, such that ghrelin may also directly regulate Tbx3 expression[5]. Moreover, we uncovered a clear link among nutritional status, Tbx3 action and neuropeptide expression in hypothalamic Pomc neurons. Whether hormonal factors and nutritional status in turn alter the peptidergic identity of hypothalamic neurons in physiological or pathophysiological conditions via modulation of Tbx3 remains a critical question. A detailed characterization of the role of Tbx3 in the context of nutritional and hormone-based regulation of hypothalamic neuronal activity might help in deciphering the main environmental factors controlling peptidergic identity development, maintenance and potential plasticity in mammalian CNS neurons.

We uncovered a molecular switch implicated in the terminal differentiation of body-weight-regulating ARC neurons into specific peptidergic subtypes, unravelling one of the mechanisms responsible for the neuronal heterogeneity of hypothalamic ARC neurons. Our findings represent another step toward the identification the key molecular machinery controlling the functional identity of hypothalamic neurons, particularly during postnatal life, and may consequently facilitate understanding of the fundamental neuronal mechanisms implicated in the pathogenesis of obesity and its associated metabolic perturbations.

Methods

Ethical compliance statement. All animal experiments were approved and conducted under the guidelines of Helmholtz Zentrum Munich and of the Faculty Animal Committee at the University of Santiago de Compostela.

Mice. All experiments were conducted on male mice. The mice were fed a standard chow diet and group housed under a 12 h:12 h light:dark cycle at 22 °C and given free access to food and water unless indicated otherwise. C57BL/6J mice were provided by Jackson Laboratories. *Tbx3*$^{loxP/loxP}$ mice were generated previously[53] and back-crossed on a C57BL/6J background for five generations. *Pomc*-Cre mice (Jax mice stock 5965 (ref. [21])) and *Agrp*-Cre mice (Jax mice stock 012899 (ref. [20])) were mated with *Tbx3*$^{loxP/loxP}$ mice to generate *Pomc*- and *Agrp*-specific *Tbx3*-knockout mice (*Pomc*-Cre;*Tbx3*$^{loxP/loxP}$ or *Agrp*-Cre;*Tbx3*$^{loxP/loxP}$). *Pomc*-Cre; *Tbx3*$^{loxP/loxP}$ and control (*Pomc*-Cre) mice were crossed with a ROSA$^{mT/mG}$ reporter line (Jax mice stock 007576 (ref. [54])) so that neurons expressing *Pomc* were permanently marked. *Pomc*-Cre;*Tbx3*$^{loxP/loxP}$, *Agrp*-Cre;*Tbx3*$^{loxP/loxP}$ or control mice (*Tbx3*$^{loxP/loxP}$) were crossed with mice selectively expressing GFP in Npy-expressing neurons (*Npy*-GFP, Jax mice stock 006417 (ref. [55])) or in Pomc-expressing neurons (*Pomc*-GFP Jax mice stock 009593 (ref. [56])).The *Tbx3*-Cre-Venus mouse line was created by using CRISPR–Cas9 technology. The coding sequences for 2A peptide bridges, Cre recombinase, Venus fluorescent protein and bovine growth hormone polyadenylation signal were cloned into a targeting vector between 5′ and 3′ homology arms flanking the stop codon of the *Tbx3* locus. Homologous recombination was confirmed by PCR and Southern blot analysis (using the DIG system from Roche). For the studies involving embryos, the breeders were mated 1 h before the dark phase and checked for a vaginal plug the next day. The day of conception (sperm-positive vaginal smear) was designated as E0. The day of birth was considered P0. Additional information can be found in the Nature Research Reporting Summary.

Physiological measures. To measure food consumption, we housed mice at two or three per cage. Body composition (fat and lean mass) was measured with quantitative nuclear magnetic resonance technology (EchoMRI). Energy expenditure and the respiratory exchange ratio were assessed with a combined indirect calorimetry system (TSE PhenoMaster, TSE Systems). O_2 consumption and CO_2 production were measured every 10 min for a total of up to 120 h (after a minimum of 48 h of adaptation). Energy expenditure (EE, kcal/h) values were correlated to the body weight of the animals recorded at the end of the measurement with analysis of covariance (ANCOVA)[19]. For the analysis of glucose tolerance, mice were injected intraperitoneally with 1.75 g glucose per kg of body weight (*Agrp*-Cre;*Tbx3*$^{loxP/loxP}$ mice) or 1.5 g glucose per kg of body weight (*Pomc*-Cre;*Tbx3*$^{loxP/loxP}$ mice). 20% (w/v) D-glucose (Sigma-Aldrich) in 0.9% (w/v) saline was used. Tail-blood glucose concentrations (mg/dl) were measured with a handheld glucometer (TheraSense Freestyle).

Viral-mediated deletion of Tbx3. To ablate *Tbx3* in the MBH, recombinant adeno-associated viruses (AAV) carrying the Cre recombinase and the haemagglutinin (HA)-tag (AAV-Cre) or control viruses carrying *Renilla* GFP (AAV-GFP) were generated as previously described[57] and injected bilaterally (0.5 µl per side; 1.0×10^{11} viral genomes ml^{-1}) into the MBH in *Tbx3*$^{loxP/loxP}$ mice (12 weeks old), with a motorized stereotaxic system from Neurostar. The nuclear localization signal (nls) of the simian virus 40 large T antigen and the Cre-recombinase coding region was fused downstream of the HA tag, in an rAAV plasmid backbone containing the 1.1-kb CMV immediate early enhancer/chicken β-actin hybrid promoter (CBA), the woodchuck post-transcriptional regulatory element (WPRE) and the bovine growth hormone poly(A) (bGH) to obtain rAAV-CBA-WPRE-bGH carrying Cre-recombinase (AAV-Cre). The rAAV-CBA-WPRE-bGH backbone carrying the *Renilla* GFP cDNA (Stratagene) was used as negative control. rAAV chimeric vectors (virions containing a 1:1 ratio of AAV1 and AAV2 capsid proteins with AAV2 ITRs) were generated by transfection of HEK293 cells with the AAV cis plasmid, the AAV1 and AAV2 helper plasmids, and the adenovirus helper plasmid through standard PEI transfection methods. At 60 h after transfection, cells were harvested, and the vector was purified through an OPTIPREP density gradient (Sigma). Genomic titres were determined with an ABI 7700 real-time PCR cycler (Applied Biosystems) with primers designed for WPRE. Virus was injected bilaterally (0.5 µl per side; 1.0×10^{11} viral genomes ml^{-1}) into the MBH in *Tbx3*$^{loxP/loxP}$ mice (12 weeks old), with a motorized stereotaxic system from Neurostar. Stereotaxic coordinates were −1.6 mm posterior and ±0.25 mm lateral to the bregma and −5.8 mm ventral from the dura. During the same procedure, a stainless-steel cannula (Bilaney Consultants) was implanted into the lateral cerebral ventricle. Stereotaxic coordinates for ICV injections were −0.8 mm posterior, −1.4 mm lateral from the bregma and −2.0 mm ventral from the dura. Surgeries were performed with a mixture of ketamine and xylazine (100 mg per kg and 7 mg per kg, respectively) as anaesthetic agents and Metamizol (200 mg per kg, subcutaneous), then Meloxicam (1 mg per kg, on three consecutive days, subcutaneous) for postoperative analgesia. For ICV studies, mice were infused with 1 µl of either vehicle (aCSF; Tocris Bioscience) or α-MSH (1 nmol, R&D systems, Tocris) 2 h before the onset of the dark cycle, and food intake followed immediately for 24 h.

BrdU experiments. Bromodeoxyuridine (BrdU, 50 mg per kg) in ~50 µl of sterile saline was injected daily at postnatal days 0, 1, 2 and 3 in the dorsal neck fold of pups. The pups were euthanized at P7, and the brains were processed for immunohistochemistry.

Immunohistochemistry. Adult mice were transcardially perfused with PBS, then with 4% neutral buffered paraformaldehyde (PFA) (Fisher Scientific). Brains from embryos and pups were isolated from non-PFA-perfused animals. After dissection, brains were post-fixed for 24 h with 4% PFA, equilibrated in 30% sucrose for 24 h and sectioned on a cryostat (Leica Biosystems) at 25 µm. Brain sections were incubated with the following primary antibodies: rabbit anti-Pomc precursor (Phoenix Pharmaceuticals, H-029-30), goat anti-Agrp (R&D systems, AF634), chicken anti-GFP (Acris, AP31791PU-N), goat anti-GFP (Abcam, ab6673), rabbit anti-Npy (Abcam, ab30914), rabbit anti-Cart (Phoenix Pharmaceuticals), mouse anti-BrdU (Sigma), goat anti-Tbx3 (A20, Santa Cruz Biotechnology), rabbit anti-Tbx3 (A303–098A, Bethyl Laboratories), rabbit anti-cleaved caspase-3 (5A1E, Cell Signaling), goat anti-Iba1 (Abcam ab107519), rabbit anti-GFAP (Dako, Z0334), chicken anti-vimentin (Sigma, Abcam ab24525) and rabbit anti-HA tag (C29F4, Cell Signaling). Primary antibodies were incubated at a concentration of 1:500 overnight at 4 °C in 0.1 M Tris-buffered saline (TBS) containing gelatine (0.25%) and Triton X-100 (0.5%). Sections were washed with 0.1 M TBS and incubated for 1 h at room temperature with 0.1 M TBS containing gelatine (0.25%) and Triton X-100 (0.5%), using the following secondary antibodies (1:1,000) from Jackson ImmunoResearch Laboratories: goat anti-rabbit (Alexa 647), goat anti-chicken (Alexa 488), donkey anti-goat (Alexa 488) and donkey anti-mouse (Alexa 488).

Image analysis. Images were obtained with a BZ-9000 microscope (Keyence) or a Leica SP5 confocal microscope, and automated analysis was performed in Fiji 1.0 (ImageJ) when technically feasible. Manual counts were performed blinded. When anatomically possible, neuronal cell counts were performed on several sections spanning the medial arcuate nucleus and averaged.

Gene expression analysis by qRT–PCR. Dissected tissues were immediately frozen on dry ice, and RNA was extracted with RNeasy Mini Kits (Qiagen). Whole hypothalamus was isolated and immediately frozen on dry ice. To obtain RNA from ARC micropunches, freshly dissected whole brains were immersed in RNAlater (AM7021, Thermo Fisher) for a minimum of 24 h at 4 °C. The RNAlater-immersed brains were subsequently cut coronally in 280-µm slices with a vibratome, and the ARC was dissected from each slice with a scalpel, as visually aided by binoculars. RNA was extracted with RNeasy Mini Kits (Qiagen). cDNA was generated with a reverse-transcription QuantiTech reverse transcription kit (Qiagen). Quantitative real-time RT–PCR (qRT–PCR) was performed with a ViiA 7 Real-Time PCR System (Applied Biosystems) with the following TaqMan probes (Thermo Fisher): Hprt (Mm01545399_m1), *Ppib* (Mm00478295_m1), *Npy* (Mm03048253_m1), Pomc (Mm00435874_m1), Agrp (Mm00475829_g1), Kisspeptin (Mm03058560_m1), Somatostatin (Mm0043667_m1), Tyrosine hydroxylase (Mm00447557_m1), Ghrh (Mm00439100_m1), Tbx3 (Mm01195726_m1), Pcsk1 (Mm00479023_m1), Pcsk2 (Mm00500981_m1), Pam (Mm01293044_m1) and Cpe (Mm00516341_m1). Target gene expression was normalized to expression of the reference genes *Hprt* or *Ppib*. Calculations were performed with a comparative method ($2^{-\Delta\Delta CT}$).

Primary mouse hypothalamic cell cultures. Hypothalami were extracted from *Tbx3*$^{loxP/loxP}$ mouse foetuses on E14 in ice-cold calcium- and magnesium-free HBSS (Life Technologies), digested for 10 min at 37 °C with 0.05% trypsin (Life Technologies), washed three times with serum-free MEM supplemented with L-glutamine (2 mM) and glucose (25 mM) and dispersed in the same medium. Cells were plated on 12-well plates coated with poly-L-lysine (Sigma-Aldrich) at a density of 1.5×10^6 per well in MEM supplemented with heat-inactivated 10% horse serum and 10% foetal bovine serum, 2 mM L-glutamine and glucose (25 mM) without antibiotics. On day 4, half the medium was replaced with fresh culture medium lacking foetal bovine serum and containing 10 µM of the mitotic inhibitor AraC (cytosine-1-β-D-arabinofuranoside, Sigma-Aldrich) to inhibit non-neuronal cell proliferation. On day 6, neurons were infected with a recombinant adenovirus carrying the coding sequence for the recombinase Cre (Ad5-CMV-Cre-eGFP, named Ad-Cre) to delete the loxP-flanked portion of the *Tbx3* gene, or with a control virus (Ad5-CMV-eGFP, named Ad-GFP) from Vector Development Laboratory. On day 7, after 12 h of incubation, virus-containing medium was removed and replaced with fresh growth medium. Neurons were further incubated for 48 h to ensure efficient recombination before performing experiments. Cell cytotoxicity was assessed with a Pierce LDF Cytotoxicity Assay Kit (88953, Thermo Fisher).

ChIP–MS. For ChIP experiments followed by mass spectrometry (ChIP–MS), hypothalamic samples from 34 individual mice were pooled, and five hypothalami at a time were homogenized in 9 ml of 1% formaldehyde in PBS for 10 min. After quenching for 5 min with 125 mM glycine, samples were washed twice with PBS. Pellets were resuspended in 1 ml of lysis buffer (0.3% SDS, 1.7% Triton, 5 mM EDTA, pH 8, 50 mM Tris, pH 8, and 100 mM NaCl), and the chromatin was sonicated to an average size of 200 bp. After incubation with either an antibody to Tbx3 (A303–098A, Bethyl Laboratories) or an IgG antibody (rabbit IgG 2729 S, Cell Signaling Technology), antibody–bait complexes were bound by Protein G–coupled agarose beads (Cell Signaling Technology) and washed three times with wash buffer A (50 mM HEPES, pH 7.5, 140 mM NaCl and 1% Triton), once with wash buffer B (50 mM HEPES pH 7.5, 500 mM NaCl and 1% Triton) and twice with TBS. Beads were incubated for 30 min with elution buffer 1 (2 M urea, 50 mM Tris-HCl, pH 7.5, 2 mM DTT and 20 µg ml^{-1} trypsin) followed by a second elution with elution buffer 2 (2 M urea, 50 mM Tris-HCl, pH 7.5 and 10 mM chloroacetamide) for 5 min. Both eluates were combined and further incubated overnight at room temperature. Tryptic-peptide mixtures were acidified with 1% TFA and desalted with Stage Tips containing three layers of C18 reverse-phase material and analysed by mass spectrometry. Peptides were separated on 50-cm columns packed in house with ReproSil-Pur C18-AQ 1.9 µm resin (Dr Maisch). Liquid chromatography was performed on an EASY-nLC 1000 ultra-high-pressure system coupled through a nanoelectrospray source to a Q-Exactive HF mass spectrometer (all from Thermo Fisher). Peptides were loaded in buffer A (0.1% formic acid) and separated by application of a non-linear gradient of 5–32% buffer B (0.1% formic acid, 80% acetonitrile) at a flow rate of 300 nl min^{-1} over 100 min. Data acquisition switched between a full scan and 15 data-dependent MS/MS scans. Full scans were acquired with target values of 3×10^6 charges in the 300–1,650 m/z range. The resolution for full-scan MS spectra was set to 60,000 with a maximum injection time of 20 ms. The 15 most abundant ions were sequentially isolated with an ion target value of 1×10^5 and an isolation window of 1.4 m/z. Fragmentation of precursor ions was performed by higher energy C-trap dissociation with a normalized collision energy of 27 eV. Resolution for HCD spectra was set to 15,000 with a maximum ion-injection time of 60 ms. Multiple sequencing of peptides was minimized by excluding the selected peptide candidates for 25 s. Raw mass spectrometry data were analysed with MaxQuant (version 1.5.6.7)[58] and Perseus (version 1.5.4.2) software packages. Peak lists were searched against the mouse UniProt FASTA database (2015_08 release) combined with 262 common contaminants by the integrated Andromeda search engine[59]. The false discovery rate was set to 1% for both peptides (minimum length of seven amino acids) and proteins. 'Match between runs' (MBR) with a maximum time difference of 0.7 min was enabled. For a gain in peptide identification, MS spectra were matched to a library of Tbx3 ChIP MS data derived from murine neuronal progenitor cells. Relative protein amounts were determined with the MaxLFQ algorithm[60], with a minimum ratio count of two. Missing values were imputed from a normal distribution, by applying a width of 0.2 and a downshift of 1.8 standard deviations. Significant outliers were defined by permutation-controlled Student's t test (FDR < 0.05, s0 = 1) comparing triplicate ChIP–MS samples for each antibody. Additional information is in the Nature Research Reporting Summary.

RNA sequencing. RNA-seq was performed in primary neurons isolated from *Tbx3*$^{loxP/loxP}$ mice and treated with Ad-Cre or Ad-GFP viruses. Sequencing was performed in three independent neuronal isolations totalling 9 Ad-GFP-treated and 11 Ad-Cre-treated independent samples. Before library preparation, RNA integrity was determined with an Agilent 2100 Bioanalyzer and an RNA 6000 Nano Kit. All samples had RNA integrity number (RIN) values >7. One microgram of total RNA per sample was used for library preparation. Library construction was performed as described in the low-throughput protocol of the TruSeq RNA Sample Prep Guide (Illumina) in an automated manner, by using the Bravo Automated Liquid Handling Platform (Agilent). cDNA libraries were assessed for quality and quantity with a Lab Chip GX (Perkin Elmer) and the Quant-iTPicoGreendsDNA Assay Kit (Life Technologies). cDNA libraries were multiplexed and sequenced as 100-bp paired-end runs on an Illumina HiSeq2500 platform. Approximately 8 Gb of sequence per sample were obtained. The GEM mapper[61] (v 1.7.1) with modified parameter settings (mismatches = 0.04, min-decoded-strata = 2) was used for split-read alignment against the mouse genome assembly mm9 (NCBI37) and UCSC knownGene annotation. Duplicate reads were removed. To quantify the number of reads mapping to annotated genes, we used HTseq-count[62] (v0.6.0). We normalized read counts to correct for possibly varying sequencing depths across samples using the R/Bioconducter package DESeq2 (ref. [63]) and excluded genes with low expression levels (mean read count <25) from the analysis. We combined RNA-seq data of the three independent neuronal isolations. Because the independence of the three neuronal isolations might have introduced batch effects, we applied surrogate variable analysis implemented in the R package sva[64] to remove them. Gene expression levels between the two virus treatments were compared with DESeq2. We chose 0.001 to be the P-value cutoff after FDR correction (Benjamini–Hochberg). To obtain genes selectively expressed in Pomc neurons, we used the single-cell sequencing data set previously published[7] and selected the n14 (Pomc/Ttr), n15 (Pomc/Anxa2) and n21 (Pomc/Glipr1) neuronal clusters as gene expression references. We chose all genes that had a normalized expression value above a noise level of 4.5. Additionally, we required the selected genes to be expressed in at least 10% of the 1,191 samples in our *Pomc*-neuron reference. We intersected the genes differentially expressed in our Tbx3 Ad-Cre[65] to test these genes against GO biological process terms[66]. After the overrepresentation test, we excluded GO terms whose gene list overlapped the list of another term completely. All calculations were performed in R (v3.4.3).

Drosophila. The *Drosophila melanogaster* neuronal GeneSwitch Gal4 driver line (Elav-Gal4^{GS}) was obtained from the Bloomington Drosophila Stock Center (BDSC43642). GeneSwitch drivers can be activated by progesterone steroids[67]. The RNAi transgenic lines for the *Tbx3*-homologue gene *omb* (line 1, UAS-ombRNAi-C4, line 2, UAS-ombRNAi-C1) were as described in ref. [68]. Elav-Gal4^{GS} virgin females and 15 *omb* RNAi transgenic males were crossed in big-fly food vials for 24 h, and approximately 600 F_1 embryos were seeded and kept at 25 °C for growth on standard cornmeal medium (12 h:12 h light:dark cycle; 60–70% humidity). After eclosion (24 h), 50 young adult male and virgin females were fed on small fresh drug-food vials (mifepristone, 200 µM) and control food vials (ethanol, same volume as the mifepristone dissolved volume), respectively, for 6 d. At least eight technical replicates (five flies each) were collected for body-fat content measurement (TAG value normalized to protein value), on the basis of a coupled-colorimetric assay (triglyceride, Pointe Scientific (T7532))[69] and bicinchoninic acid assay (protein, Pierce, Thermo; 23225)[70,71]; five adult male flies per technical replicate, 600 µl homogenization buffer (0.05% Tween-20 in water) and 5-mm metal beads (Qiagen 69989) were homogenized in 1.2-ml collecting tubes (Qiagen 19560; caps, 19566) with a tissue lyser II (Qiagen, 85300) and immediately incubated at 70 °C (water bath) for 5 min. The fly homogenates were spun down at 5,000 r.p.m. for 3 min, and 2 × 50 µl supernatant for each replicate, TAG standards solutions (Biomol, Cay-10010509; 0, 5.5, 11, 22, 33 and 44 µg in 50 µl homogenization buffer), BSA (bovine serum albumin) and protein-standard samples (0, 25, 125, 250, 500 and 750 mg ml^{-1}) were measured at 500 nm (for TAG) and 570 nm (for protein). The assay kit for the colorimetric assay was from Pointe Scientific (T7532). Immunostainings were carried out in 5-d-old adult male flies (omb-Gal4>UAS-GFP transgenic line[72]) through a method reported previously[73]. Brains were dissected in cold PBS and fixed in 4% PFA in PBS at room temperature (RT) for 30 min. Brain tissues were incubated with 0.25% Triton X-100 in PBS (0.25% PBST) at RT for 25 min and blocked with 1% BSA & 3% normal goat serum (NGS) in 0.25% PBST for 1 h at RT with mild rotation. The following primary antibodies were used: mouse anti-Bruchpilot (nc82, 1:50) (nc82, deposited in the DSHB by Buchner, E. (DSHB Hybridoma Product nc82)), chicken anti-GFP (1:1000) (Acris, AP31791PU-N) and rabbit anti-Omb serum (1:1,000)[74]. The following secondary antibodies (Jackson ImmunoResearch Laboratories) were used: donkey anti-mouse (Alexa 568), goat anti-rabbit (Alexa 647) and goat anti-chicken (Alexa 488). Secondary antibodies were incubated at RT for 2 h. After 5 × 10 min washing with 0.25% PBST and 1× overnight washing with PBS, tissues were mounted on gelatine-coated glass slides and coverslipped for image analysis. Images were obtained with Leica SP8 confocal system (×20 air objective) and processed with Fiji 1.0 (ImageJ).

Human embryonic stem cells. The human H9 ESC line was purchased from WiCell. Cells were maintained in a humidified incubator at 37 °C on irradiated murine embryonic fibroblasts (MEFs; CF-1 MEF 4 M IRR; GLOBALSTEM) in DMEM KO medium (10829018; Thermo Fisher) supplemented with 15% KnockOut Serum Replacement (10828028; Thermo Fisher), 0.1 mM MEM non-essential amino acids (11140050; Thermo Fisher), 2 mM GlutaMAX (35050061; Thermo Fisher), 0.06 mM 2-mercaptoethanol (21985023; Thermo Fisher), FGF-basic (AA 1–155), (20 ng ml^{-1} medium; PHG0263; Thermo Fisher), and 10 µM Rock inhibitor (S1049; Selleckchem). Cells were passaged with Accutase (00–4555–56; Thermo Fisher). For CRISPR–Cas9-mediated deletion of *Tbx3*,

pCas9_GFP was obtained from Addgene (Kiran Musunuru; 44719). As previously published, the GFP was replaced by a truncated CD4 gene from the GeneArt CRISPR Nuclease OFP Vector (Thermo Fisher) by GenScript through CloneEZ seamless cloning technology, thus resulting in vector pCas9_CD4 (ref. [73]). The full vector sequence of pCas9_CD4 is given in Supplementary Table 4. The guide RNA sequence 5′-TCATGGCGAAGTCCGGCGCC-3′ was obtained by using Optimized CRISPR Design (MIT; http://crispr.mit.edu/). Cloning of the gRNA into pGS-U6-gRNA was performed by GenScript. 800,000 human ESCs were collected and mixed in nucleofection buffer (Human Stem Cell Nucleofector Kit 2; VPH-5022) with gRNA and pCas9_CD4 plasmids (2.5 µg each). Nucleofection was performed in an Amaxa Nucleofector II (Programme A-023) with a Human Stem Cell Nucleofector Kit 2 according to the manufacturer's instructions. Cells were plated on MEFs for 2 d for recovery, and transfected cells were purified through positive selection of *CD4*-expressing cells by using human CD4 MicroBeads (130–045–101; MS Column, 130–042–20; MACS Miltenyi Biotec) and replated at clonal density in 10 cm^2 tissue culture plates on MEFs. After 7–12 d, ESC colonies were picked into 96-well plates and, 4–5 d later were split 1:2 (one well for genomic DNA extraction followed by sequence analysis as described below, and one well for amplification of clones and further analysis and freezing, if indicated). For genomic-DNA extraction and PCR analysis, genomic DNA was extracted with HotShot buffer according to a published protocol[76]. The DNA region of interest was PCR-amplified with the following primers: 5′-GAGAGCGCCGCCGCGCCGT-3′ and 5′-GCTGCGGACTTGTCCCCGGCTGGA-3′[76]. Sequences were generated by Sanger sequencing (Macrogen). Sequence analysis was performed to identify clones carrying mutations resulting in *TBX3* knockout. Positive clones were amplified, and genomic DNA was extracted with a Gentra Puregene Core Kit A (Qiagen). Topo TA Cloning Kit for sequencing (K457501; Thermo Fisher) was used to determine the zygosity of *TBX3* knockout with the following primers: 5′- CACCTTGGGGTCGTCCTCCA-3′ and 5′- CGCAAGGCACAAGGACGGTCA-3′. G-band karyotyping analysis was done by Cell Line Genetics. Chromosome analysis was performed on 20 cells per cell line.

Differentiation of human ESCs into arcuate-like neurons. Human ESCs differentiated into hypothalamic arcuate-like neurons were derived from human ESCs through a previously published protocol[33–35]. H9 cells were plated on dishes coated with Matrigel (08–774–552; Thermo Fisher) dishes at a density of 100,000 cells per cm^2) in human ESC medium, as described above, supplemented with bFGF and Rock inhibitor. Cell density was observed after 24 h. If the cells were not yet at 100% confluency, medium was aspirated and replaced with ESC medium with bFGF and Rock inhibitor for another 24 h. After cells reached 100% confluency, differentiation was initiated. 10 µM SB 431542 (S1067; Selleckchem) and 2.5 µM LDN 193189 (S2618; Selleckchem) were used from day 1 to day 8 to inhibit TGFβ and BMP signalling to promote neuronal differentiation from human ES cells[77]. 100 ng ml^{-1} SHH (248-BD; (R&D Systems) and 2 µM purmorphamine (PM; S3042; Selleckchem) were added from days 1–8 to induce ventral brain development and NKX2.1 expression. Cells were cultured on days 1–4 in ESC medium, from days 5–8, the medium was switched stepwise from ESC medium to N2 medium (3:1, 1:1, 1:3). N2 medium (500 ml) consisted of 485 ml DMEM/F12 (11322; Thermo Fisher) supplemented with 5 ml MEM non-essential amino acids (11140050; Thermo Fisher), 5 ml of a 16 % glucose solution and 5 ml N2 (1370701; Thermo Fisher). Ascorbic acid (A0278; Sigma-Aldrich) was added just before use at a final concentration of 200 nM. From Day 9 onward, cells were cultured in N2-B27 medium consisting of 475 ml DMEM/F12 (11322; Thermo Fisher) supplemented with 5 ml MEM non-essential amino acids (11140050; Thermo Fisher), 5 ml of a 16 % glucose solution, 5 ml N2 (1370701; Thermo Fisher) and 10 ml B27 (12587010; Thermo Fisher). Ascorbic acid (A0278; Sigma-Aldrich) was added just before use at a final concentration of 200 nM. Inhibition of Notch signalling by 10 µM DAPT (S2215; Selleckchem) was performed from days 9 to 12. Nkx2.1 + progenitors were collected and re-plated on extracellular matrix (poly-L-ornithine (A-004-C; Millipore) and laminin (23017015; Thermo Fisher)) to enhance the attachment and differentiation of neuron progenitors. The Notch inhibitor DAPT was used to inhibit the proliferation of progenitor cells and promote further neuronal differentiation[78,79]. The neurotrophic factor BDNF (20 ng ml^{-1}; 450–02; PeproTech) was introduced after DAPT treatment to improve the survival, differentiation and maturation of these neurons. For RT–PCR analyses, Cells at day 0, day 12 and day 27 of differentiation were homogenized in Trizol reagent (15596026; Thermo Fisher), and total RNA was extracted with an RNeasy Plus Micro Kit (74034; Qiagen) with on-column DNase I (79254; Qiagen) treatment to remove genomic DNA contamination and stored at −80 °C until further processing. A total of 500 ng of total RNA was used for reverse transcription with a Transcriptor First Strand cDNA Synthesis Kit (04897030001; Roche Diagnostic) by using a mixture of anchored $oligo(dT)_{18}$ and random-hexamer primers according to the manufacturer's instructions. Quantitative PCR was performed with a Light-Cycler 480 (Roche Diagnostics) with SYBR Green in a total volume of 10 µl with 1 µl of template, 1 µl of forward and reverse primers (10 µM) and 5 µl of SYBR Green I Master-Mix (04707516001; Roche Diagnostic). Reactions included an initial cycle at 95 °C for 10 min, followed by 40 cycles of denaturation at 95 °C for 10 s, annealing at 60 °C for 5 s and extension at 72 °C for 15 s. Crossing points were determined by Light-Cycler 480 software, by using the second-derivative maximum technique. Relative expression data were calculated with the delta-delta Ct method, with normalization of the raw data to expression of *TBP*. Quantitative PCR was performed to determine the mRNA levels of *TBX3*, *POMC*, *TUBB3*, *PCSK1*, *NKX2.1*. Primer sequences are shown in Supplementary Table 1.

Western blot analysis. Human ESC (H9) and hypothalamic arcuate-like neurons (day 27 of differentiation) were washed with DPBS and lysed in RIPA Lysis and Extraction Buffer (Thermo Fisher) with protease and phosphatase inhibitors (78442; Thermo Fisher), incubated at 4 °C for 15 min and then centrifuged at 12,000 r.p.m. for 15 min at 4 °C. Fifteen micrograms of total protein from each extract was loaded on a 4–12% gradient Bis-Tris gel (NP0335BOX; Thermo Fisher) and transferred onto nitrocellulose membrane with an iBlot 2 Dry Blotting System (Thermo Fisher). The membrane was blocked for 1 h at room temperature with SuperBlock T20 (TBS) Blocking Buffer (37536; Thermo Fisher) and then incubated with primary antibody to TBX3 (1:100; ab99302; Abcam) overnight at 4 °C, washed three times with TBS with 0.1% Tween-20 (1706531; Bio-Rad) and incubated with secondary antibody anti-rabbit HRP (1:10,000; 7074 S; Cell Signaling) for 1 h at room temperature. Specific bands were then detected through electrochemiluminescence analysis with SuperSigna West Pico PLUS Chemiluminescent Substrate (34577; Thermo Fisher). An antibody to beta-actin (1:1,000; ab8226; Abcam), with anti-mouse HRP (1:10,000; 7076 S; Cell Signaling) as a secondary antibody, was used as a loading control. Validation of the goat anti-Tbx3 antibody (A20, Santa Cruz Biotechnology) was performed by using Tbx3-deficient embryos (E13.5) kindly provided by A. Kispert[80]. Proteins were extracted with RIPA buffer containing protease- and phosphatase-inhibitor cocktails (Thermo Fisher) 1 mM phenyl-methane-sulfonyl fluoride (PMSF) and 1 mM sodium butyrate (Sigma-Aldrich). Proteins were transferred on nitrocellulose membranes by using a Trans Blot Turbo transfer apparatus (Bio-Rad), and stained with primary antibody goat anti-Tbx3 (1:500) and a secondary antibody anti-goat HRP (1:1,000). Detection was carried out on a LiCor Odyssey instrument (software Image studio 2.0), by using electrochemiluminescence (Amersham). Additional information is available in the Nature Research Reporting Summary.

Tbx3-focused single-cell RNA-sequencing analysis. Data for the scRNA-seq analysis were obtained from GEO accession codes GSE90806 and GSE93374 (ref. [7]). The data matrix comprised 21,086 cells and 22,802 genes generated from the arcuate-median eminence (Arc-ME) of the mouse hypothalamus by Campbell et al.[7]. We used Seurat software[81] to perform clustering analysis. We identified the 2,250 most variable genes across the entire dataset, controlling for the known relationship between mean expression and variance. After scaling and centring the data along each variable gene, we performed principal component analysis and identified 25 significant prinicipal components for downstream analysis that were used to identify 20 clusters. Similarly to those identified by Campbell et al.[7], a total of 13,079 neurons and 8,007 non-neuronal cells were identified in our study. We further used the neuronal identities assigned by the authors for clustering the neurons into their respective neuronal clusters. For differential expression between cell type clusters, we used the negative binomial test, a likelihood-ratio test assuming an underlying negative binomial distribution for UMI-based datasets.

Statistics. Statistical analyses were conducted in GraphPad Prism (version 5.0a). For each experiment, slides were numerically coded to obscure the treatment group. Statistical significance was determined with unpaired two-tailed Student's *t* test, one-way ANOVA or two-way ANOVA followed by an appropriate post hoc test, as indicated in figure legends, and linear regression when appropriate. $P \leq 0.05$ was considered statistically significant. Additional information is provided in the Nature Research Reporting Summary.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The authors declare that all data supporting the findings of this study are available within the paper and its supplementary information files. The RNA-seq database generated in our paper has been made publicly available through Gene Expression Omnibus (GEO accession number GSE119883).

Received: 14 June 2018; Accepted: 13 December 2018;
Published online: 28 January 2019

References

1. Cone, R. D. Anatomy and regulation of the central melanocortin system. *Nat. Neurosci.* **8**, 571–578 (2005).
2. Gautron, L., Elmquist, J. K. & Williams, K. W. Neural control of energy balance: translating circuits to therapies. *Cell* **161**, 133–145 (2015).
3. Koch, M. & Horvath, T. L. Molecular and cellular regulation of hypothalamic melanocortin neurons controlling food intake and energy metabolism. *Mol. Psychiatry* **19**, 752–761 (2014).

4. Morton, G. J., Meek, T. H. & Schwartz, M. W. Neurobiology of food intake in health and disease. *Nat. Rev. Neurosci.* **15**, 367–378 (2014).
5. Knight, Z. A. et al. Molecular profiling of activated neurons by phosphorylated ribosome capture. *Cell* **151**, 1126–1137 (2012).
6. Allison, M. B. et al. TRAP-seq defines markers for novel populations of hypothalamic and brainstem LepRb neurons. *Mol. Metab.* **4**, 299–309 (2015).
7. Campbell, J. N. et al. A molecular census of arcuate hypothalamus and median eminence cell types. *Nat. Neurosci.* **20**, 484–496 (2017).
8. Wansleben, S., Peres, J., Hare, S., Goding, C. R. & Prince, S. T-box transcription factors in cancer biology. *Biochim. Biophys. Acta* **1846**, 380–391 (2014).
9. Ang, L. T. et al. A roadmap for human liver differentiation from pluripotent stem cells. *Cell Rep.* **22**, 2190–2205 (2018).
10. Suzuki, A., Sekiya, S., Büscher, D., Izpisúa Belmonte, J. C. & Taniguchi, H. Tbx3 controls the fate of hepatic progenitor cells in liver development by suppressing p19ARF expression. *Development* **135**, 1589–1595 (2008).
11. Weidgang, C. E. et al. TBX3 Directs cell-fate decision toward mesendoderm. *Stem Cell Rep.* **1**, 248–265 (2013).
12. Eriksson, K. S. & Mignot, E. T-box 3 is expressed in the adult mouse hypothalamus and medulla. *Brain Res.* **1302**, 233–239 (2009).
13. Linden, H., Williams, R., King, J., Blair, E. & Kini, U. Ulnar mammary syndrome and TBX3: expanding the phenotype. *Am. J. Med. Genet. A* **149A**, 2809–2812 (2009).
14. Schinzel, A. The ulnar-mammary syndrome: an autosomal dominant pleiotropic gene. *Clin. Genet.* **32**, 160–168 (1987).
15. Grill, H. J. & Hayes, M. R. Hindbrain neurons as an essential hub in the neuroanatomically distributed control of energy balance. *Cell. Metab.* **16**, 296–309 (2012).
16. Joly-Amado, A. et al. The hypothalamic arcuate nucleus and the control of peripheral substrates. *Best Pract. Res. Clin. Endocrinol. Metab.* **28**, 725–737 (2014).
17. Clasadonte, J. & Prevot, V. The special relationship: glia-neuron interactions in the neuroendocrine hypothalamus. *Nat. Rev. Endocrinol.* **14**, 25–44 (2018).
18. Pontecorvi, M., Goding, C. R., Richardson, W. D. & Kessaris, N. Expression of Tbx2 and Tbx3 in the developing hypothalamic-pituitary axis. *Gene. Expr. Patterns* **8**, 411–417 (2008).
19. Tschöp, M. H. et al. A guide to analysis of mouse energy metabolism. *Nat. Methods* **9**, 57–63 (2011).
20. Tong, Q., Ye, C.-P., Jones, J. E., Elmquist, J. K. & Lowell, B. B. Synaptic release of GABA by AgRP neurons is required for normal regulation of energy balance. *Nat. Neurosci.* **11**, 998–1000 (2008).
21. Balthasar, N. et al. Leptin receptor signaling in POMC neurons is required for normal body weight homeostasis. *Neuron* **42**, 983–991 (2004).
22. Kumar, P. P. et al. Coordinated control of senescence by lncRNA and a novel T-box3 co-repressor complex. *eLife* **3**, e02805 (2014).
23. Coll, M., Seidman, J. G. & Müller, C. W. Structure of the DNA-bound T-box domain of human TBX3, a transcription factor responsible for ulnar-mammary syndrome. *Structure* **10**, 343–356 (2002).
24. Hein, M. Y. et al. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163**, 712–723 (2015).
25. Rolland, T. et al. A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).
26. Bandyopadhyay, S. et al. A human map kinase interactome. *Nat. Methods* **7**, 801–805 (2010).
27. Padilla, S. L., Carmody, J. S. & Zeltser, L. M. Pomc-expressing progenitors give rise to antagonistic neuronal populations in hypothalamic feeding circuits. *Nat. Med.* **16**, 403–405 (2010).
28. Toda, C., Santoro, A., Kim, J. D. & Diano, S. POMC neurons: from birth to death. *Annu. Rev. Physiol.* **79**, 209–236 (2017).
29. Hahn, T. M., Breininger, J. F., Baskin, D. G. & Schwartz, M. W. Coexpression of Agrp and NPY in fasting-activated hypothalamic neurons. *Nat. Neurosci.* **1**, 271–272 (1998).
30. Sousa-Ferreira, L., de Almeida, L. P. & Cavadas, C. Role of hypothalamic neurogenesis in feeding regulation. *Trends Endocrinol. Metab.* **25**, 80–88 (2014).
31. Mizuno, T. M. et al. Hypothalamic pro-opiomelanocortin mrna is reduced by fasting in ob/ob and db/db mice, but is stimulated by leptin. *Diabetes* **47**, 294–297 (1998).
32. Wilson, V. & Conlon, F. L. The T-box family. *Genome Biol.* **3**, REVIEWS3008 (2002).
33. Wang, L. et al. Differentiation of hypothalamic-like neurons from human pluripotent stem cells. *J. Clin. Invest.* **125**, 796–808 (2015).
34. Wang, L., Egli, D. & Leibel, R. L. Efficient generation of hypothalamic neurons from human pluripotent stem cells. *Curr. Protoc. Hum. Genet.* **90**, 21.5.1–21.5.14 (2016).
35. Wang, L. et al. PC1/3 deficiency impacts pro-opiomelanocortin processing in human embryonic stem cell-derived hypothalamic neurons. *Stem Cell Rep.* **8**, 264–277 (2017).
36. Coupe, B. & Bouret, S. G. Development of the hypothalamic melanocortin system. *Front Endocrinol. (Lausanne)* **4**, 38 (2013).
37. Pelling, M. et al. Differential requirements for neurogenin 3 in the development of POMC and NPY neurons in the hypothalamus. *Dev. Biol.* **349**, 406–416 (2011).
38. Lee, B. et al. Dlx1/2 and Otp coordinate the production of hypothalamic GHRH- and AgRP-neurons. *Nat. Commun.* **9**, 2026 (2018).
39. Nasif, S. et al. Islet 1 specifies the identity of hypothalamic melanocortin neurons and is critical for normal food intake and adiposity in adulthood. *Proc. Natl Acad. Sci. USA* **112**, E1861–E1870 (2015).
40. Lee, B., Lee, S., Lee, S.-K. & Lee, J. W. The LIM-homeobox transcription factor Isl1 plays crucial roles in the development of multiple arcuate nucleus neurons. *Development* **143**, 3763–3773 (2016).
41. Sakkou, M. et al. A role for brain-specific homeobox factor Bsx in the control of hyperphagia and locomotory behavior. *Cell Metab.* **5**, 450–463 (2007).
42. Messina, A. et al. A microRNA switch regulates the rise in hypothalamic GnRH production before puberty. *Nat. Neurosci.* **19**, 835–844 (2016).
43. Greenman, Y. et al. Postnatal ablation of POMC neurons induces an obese phenotype characterized by decreased food intake and enhanced anxiety-like behavior. *Mol. Endocrinol.* **27**, 1091–1102 (2013).
44. Morton, G. J. & Schwartz, M. W. The NPY/AgRP neuron and energy homeostasis. *Int. J. Obes. Relat. Metab. Disord.* **25** (Suppl. 5), S56–S62 (2001).
45. Luquet, S., Perez, F. A., Hnasko, T. S. & Palmiter, R. D. NPY/AgRP neurons are essential for feeding in adult mice but can be ablated in neonates. *Science* **310**, 683–685 (2005).
46. Tan, K., Knight, Z. A. & Friedman, J. M. Ablation of AgRP neurons impairs adaption to restricted feeding. *Mol. Metab.* **3**, 694–704 (2014).
47. Bouret, S. G. & Simerly, R. B. Minireview: leptin and development of hypothalamic feeding circuits. *Endocrinology* **145**, 2621–2626 (2004).
48. Zhan, C. et al. Acute and long-term suppression of feeding behavior by pomc neurons in the brainstem and hypothalamus, respectively. *J. Neurosci.* **33**, 3624–3632 (2013).
49. Nogueiras, R. et al. The central melanocortin system directly controls peripheral lipid metabolism. *J. Clin. Invest.* **117**, 3475–3488 (2007).
50. Burbridge, S., Stewart, I. & Placzek, M. Development of the neuroendocrine hypothalamus. *Compr. Physiol.* **6**, 623–643 (2016).
51. Dulcis, D., Jamshidi, P., Leutgeb, S. & Spitzer, N. C. Neurotransmitter switching in the adult brain regulates behavior. *Science* **340**, 449–453 (2013).
52. Gascón, S., Masserdotti, G., Russo, G. L. & Götz, M. Direct neuronal reprogramming: achievements, hurdles, and new roads to success. *Cell Stem Cell* **21**, 18–34 (2017).
53. Frank, D. U., Emechebe, U., Thomas, K. R. & Moon, A. M. Mouse Tbx3 mutants suggest novel molecular mechanisms for ulnar-mammary syndrome. *PLoS One* **8**, e67841 (2013).
54. Muzumdar, M. D., Tasic, B., Miyamichi, K., Li, L. & Luo, L. A global double-fluorescent Cre reporter mouse. *Genesis* **45**, 593–605 (2007).
55. van den Pol, A. N. et al. Neuromedin B and gastrin-releasing peptide excite arcuate nucleus neuropeptide Y neurons in a novel transgenic mouse expressing strong Renilla green fluorescent protein in NPY neurons. *J. Neurosci.* **29**, 4622–4639 (2009).
56. Cowley, M. A. et al. Leptin activates anorexigenic POMC neurons through a neural network in the arcuate nucleus. *Nature* **411**, 480–484 (2001).
57. Monory, K. et al. The endocannabinoid system controls key epileptogenic circuits in the hippocampus. *Neuron* **51**, 455–466 (2006).
58. 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).
59. Cox, J. et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
60. Cox, J. et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell Proteom.* **13**, 2513–2526 (2014).
61. Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods* **9**, 1185–1188 (2012).
62. Anders, S., Pyl, P. T. & Huber, W. HTSeq: a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
63. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
64. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, e161 (2007).
65. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
66. Ashburner, M. et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat. Genet.* **25**, 25–29 (2000).
67. Osterwalder, T., Yoon, K. S., White, B. H. & Keshishian, H. A conditional tissue-specific transgene expression system using inducible GAL4. *Proc. Natl Acad. Sci. USA* **98**, 12596–12601 (2001).

68. Shen, J., Dorner, C., Bahlo, A. & Pflugfelder, G. O. *optomotor-blind* suppresses instability at the A/P compartment boundary of the *Drosophila* wing. *Mech. Dev.* **125**, 233–246 (2008).
69. Hildebrandt, A., Bickmeyer, I. & Kühnlein, R. P. Reliable Drosophila body fat quantification by a coupled colorimetric assay. *PLoS One* **6**, e23796 (2011).
70. Gáliková, M., Klepsatel, P., Xu, Y. & Kühnlein, R. P. The obesity-related adipokinetic hormone controls feeding and expression of neuropeptide regulators of Drosophila metabolism. *Eur. J. Lipid Sci. Tech.* **119**, 1600138 (2017).
71. Klepsatel, P., Gáliková, M., Xu, Y. & Kühnlein, R. P. Thermal stress depletes energy reserves in Drosophila. *Sci. Rep.* **6**, 33667 (2016).
72. Mayer, L. R., Diegelmann, S., Abassi, Y., Eichinger, F. & Pflugfelder, G. O. Enhancer trap infidelity in Drosophila optomotor-blind. *Fly* **7**, 118–128 (2013).
73. Baumbach, J., Xu, Y., Hehlert, P. & Kühnlein, R. P. Gαq, Gγ1 and Plc21C control Drosophila body fat storage. *J. Genet. Genom.* **41**, 283–292 (2014).
74. Shen, J., Dahmann, C. & Pflugfelder, G. O. Spatial discontinuity of optomotor-blind expression in the Drosophila wing imaginal disc disrupts epithelial architecture and promotes cell sorting. *BMC Dev. Biol.* **10**, 23 (2010).
75. Stratigopoulos, G., De Rosa, M. C., LeDuc, C. A., Leibel, R. L. & Doege, C. A. DMSO increases efficiency of genome editing at two non-coding loci. *PLoS One* **13**, e0198637 (2018).
76. Santos, D. P., Kiskinis, E., Eggan, K. & Merkle, F. T. Comprehensive protocols for crispr/cas9-based gene editing in human pluripotentstem cells. *Curr. Protoc. Stem Cell Biol.* **38**, 5B.6.1–5B.6.60 (2016).
77. Chambers, S. M. et al. Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nat. Biotechnol.* **27**, 275–280 (2009).
78. Crawford, T. Q. & Roelink, H. The notch response inhibitor DAPT enhances neuronal differentiation in embryonic stem cell-derived embryoid bodies independently of sonic hedgehog signaling. *Dev. Dyn.* **236**, 886–892 (2007).
79. Nelson, B. R., Hartman, B. H., Georgi, S. A., Lan, M. S. & Reh, T. A. Transient inactivation of Notch signaling synchronizes differentiation of neural progenitor cells. *Dev. Biol.* **304**, 479–498 (2007).
80. Trowe, M.-O. et al. Inhibition of Sox2-dependent activation of Shh in the ventral diencephalon by Tbx3 is required for formation of the neurohypophysis. *Development* **140**, 2299–2309 (2013).
81. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).

Acknowledgements

We thank A. Kispert and M.-O. Trowe (Institut für Molekularbiologie, Medizinische Hochschule Hannover, Hannover, Germany) for kindly providing Tbx3-deficient embryos, J. Friedman (The Rockefeller University, Howard Hughes Medical Institute, New York, NY, USA) for scientific guidance and for graciously providing access to data shown in ref. [3], M. Guzmán (Complutense University, Madrid, Spain) for assistance with the generation of AAV-GFP and AAV-Cre viral particles, the Bloomington Drosophila Stock Center (BDSC) (NIH P40OD018537) for fly stocks, and C. Layritz, H. Hoffmann, N. Wiegert and C. L. Holleman for technical assistance and assistance with animal studies. A.F. is supported by a postoctoral fellowship from the Canadian Institutes of Health Research (Funding reference no. 152588). V.V.T. is supported by NIH-NIDDK grant 5K23DK110539 and in part by the Baylor-Hopkins Center for Mendelian Genomics through NHGRI grant 5U54HG006542. C.A.D. is supported by funding from the NIH (R01 DK52431, R01 DK110113 and P30 DK26687) and Columbia Stem Cell Initiative Seed Fund Program. We thank the Fondation Recherche Medicale (ARF20140129235, L.B.). This work was strongly supported by the Helmholtz Alliance ICEMED & the Helmholtz Initiative on Personalized Medicine iMed by Helmholtz Association. This work was supported in part by the Helmholtz cross-program topic 'Metabolic Dysfunction', the European Research Council ERC (AdG *HypoFlam* no. 695054) and in part by funding to M.H.T., Y.L., B.L. and V.K. from the Alexander von Humboldt Foundation.

Author contributions

C.Q. and A.F. designed and performed the experiments and interpreted the data. Y.X., G.C., B.L. Y.-T.T., A.R., M.W., M.C.D., V.K., R.R., V.V.T., E.G., T.M.S., A.-L.P., T.G., O.L., A.C.-S., D.K., L.B., S.C.W., G.O.P., R.N., L.Z., I.C.G.K., A.M., C.G.-C., M.M., M.T. and C.A.D. performed experiments and/or edited the manuscript. M.H.T. conceptualized the project, interpreted the data, and cowrote the manuscript together with C.Q. and A.F.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42255-018-0028-1.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.H.T.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

4 Discussion

Mass spectrometry-based proteomics has without any doubt revolutionized protein science over the past decades. Before, researchers had to characterize individual proteins with laborious experiments depending on the experimental question. Today, thousands of proteins expressed at a given time in cells or tissues can be analyzed within hours of MS measurement time. The protein complexes organized around hundreds of proteins are dissected in high-throughput experiment within days or weeks. Whole signaling cascades are disentangled by the mass spectrometry-based analysis of phosphorylated peptides. Detection of specific modifications like glycation or acetylation in dedicated MS analyses provides new information about biological functions of proteins. The list of applications in which mass spectrometry has had a major impact on the characterization of proteins appears endless. In theory, MS can be used to answer any biological question regarding the abundance, interactions, location, sequence modifications or structural features of a protein. In this thesis I applied mass spectrometry to study gene regulation. I developed and optimized UV crosslinking methods to investigate interactions of proteins with nucleic acids in a gene regulatory context. Moreover, I conducted ChIP-MS experiments to improve our understanding of trans-regulatory protein interactions.

UV crosslinking mass spectrometry - Method of choice to map protein-RNA interactions and the key to identify DNA-binding proteins?

Crosslinking mass spectrometry is an attractive approach to study interactions of proteins. It can reveal structural features and help to capture transient or weak interactions, that may otherwise escape. Ideally, crosslinking agents will specifically connect domains which are in close contact. Chemical crosslinking reagents typically comprise two reactive groups and a spacer. As outlined in

the introduction, they introduce bonds between molecules in close proximity depending on the spacer length. In contrast, UV crosslinking triggers a radical reaction between two molecules and causes 'zero-length' crosslinks. Owing to the short half-lives of the excited radical intermediates it is generally more specific than chemical reagents. Nevertheless, chemical crosslinking remains the mainstay in structural analysis of protein-protein interactions. This is in contrast to the field of protein-RNA complexes where UV crosslinking has largely replaced chemical crosslinkers. Single-stranded RNA is particularly susceptible to excitation by UV irradiation, which will readily induce the formation of radical ions within the nucleotides. These radicals will immediately react with amino acids in close vicinity. This is not only feasible for reconstituted complexes, but also for cultured cells, which opens a broad range of applications for UV crosslinking.

UV crosslinking was first adopted in a method termed CLIP (crosslinking and immunoprecipitation) for identifying the RNA-binding sites of proteins by sequencing the precipitated RNA fragments [156]. CLIP was modified and optimized for various applications and modern sequencing techniques [173]. Later the combination of UV crosslinking with mass spectrometry enabled global analyses of RNA-binding proteins. Pulldowns of specific RNA species, e.g. of messenger RNA, identified proteins binding these types of RNA [159, 160]. Recently established protocols like OOPS and XRNAX provide means to map entire RNA-binding proteomes [162, 163]. Finally, in at least 90% of cases protein-RNA crosslinking accurately detected the RNA-binding domain in proteins [174].

Disadvantages mainly concern the long irradiation times necessary with conventional UV light sources to achieve decent crosslinking effciencies. This creates the risk of causing cellular damages or conformational rearrangements in macromolecules, which do no longer reflect the true physiological state in the cell. Another potential limitation revealed in a comprehensive study is that the amino acids aspartic acid, asparagine, glutamic acid and glutamine may not be susceptible to crosslinking. Moreover, almost exclusively uracil appears to be involved in crosslinking by UV lamps [174]. This limits the applicability of the described methods in structural investigations.

Despite these potential limitations the insights obtained into dynamics and

structural features of RNA-binding by UV crosslinking MS underscores its applicability in protein-RNA studies. The fact that conventional UV light sources are highly effective enables non-specialized laboratories to use this technique for RNA-binding investigations. In this thesis I made use of this technique and shaped the RNA-binding proteome in mouse and human immune cells by adapting and modifying the OOPS protocol. For the publication we used both OOPS and RNA-IC method to perform straightforward, label-free quantification analysis of the RBPome. Indeed, we showed that both methods accurately identify proteins interacting with RNA. We further describe how RNA-IC detects a specific set of canonical RNA-binders, whereas OOPS recovers both more canonical, but also more non-canonical RNA-binding proteins. Particularly with OOPS we recovered various novel unexpected RNA-binders. For several of these we demonstrate an immunological mRNA-regulatory function by follow-up biochemical experiments.

I expect that UV crosslinking will continue to be the method of choice to characterize protein-RNA interactions. It is highly specific and effective, comparably easy to use and can be performed in routine laboratories. With our study we add to previous publications in describing entire RBPomes by UV crosslinking mass spectrometry in specialized cell types.

As mentioned before, photo-crosslinking has hardly entered the field of protein-DNA interactions owing to much lower crosslinking efficiencies. The identification of true DNA-binders has been of crucial interest to the community as they control various biological processes by regulating protein expression. Typical approaches to study DNA-binding of proteins are electrophoretic mobility shift assays and ChIP-Seq experiments. However, these methods require *a priori* knowledge of the proteins and ChIP-Seq relies on chemicals like formaldehyde, which crosslinks proteins both to DNA and other proteins. This leads to the identification of indirect DNA-binders, that associate with the DNA through another protein. Currently, there is no method that allows the unbiased, system-wide detection of protein-DNA interactions. This may change with the introduction of photo-crosslinking similarly as it has revolutionized the study of RNA-binders. This will predominantly depend on the increase in crosslinking efficiency of proteins to DNA. Decades ago UV lasers were shown to be orders of magnitude more effective than UV lamps. They de-

liver photons on a timescale that is shorter than that in which rearrangement of macromolecules occurs [152]. Among nano-, pico- and femtosecond lasers the latter achieves the highest crosslinking rates [154]. Despite the early research on effective photo-crosslinking of proteins to DNA by UV lasers, the application to biological research remained limited over the following decades. Recently, a nanosecond UV laser was used to crosslink the transcription factor BCL6 in human lymphoma cells and analyze BCL6 binding sites by ChIP-Seq [170]. The study identified novel BCL6 binding sites and outperformed formaldehyde-based ChIP-Seq in specificity and resolution. It highlighted the strong benefits of UV laser crosslinking compared to the use of formaldehyde. This raised the question whether the crosslinked proteins may also be analyzed via mass spectrometry, which I was able to answer in this thesis. ChIP-Seq has the advantage of having intrinsic amplification steps, which reduces the amount of crosslinked species needed for detection. MS analysis lacks such a step and limited amount of crosslinked species may cause detection problems. Moreover, the global structural analysis of proteins crosslinked to DNA nucleotides comprises an enormous search space including all protein sequences, different possible modifications and one or multiple DNA nucleotides. Additionally, protein-RNA complexes are likely the dominating crosslinked species. Thus, sample preparation requires very stringent purification of protein-DNA crosslinks to reduce search space in computational analysis and separate them from protein-RNA complexes. In this thesis I conducted a showcase study to obtain insights into the nature and specificity of photo-crosslinking protein-DNA complexes using a UV femtosecond laser. Likewise, I established that extracting DNA-protein crosslinks from cells is possible, although I did not yet perform a global analysis of DNA-binders. Nevertheless, I expect this study to prime further investigation into photo-crosslinking, which will improve existing protocols to render a global mapping of DNA-binders feasible in the future.
In conclusion, I show in this thesis that UV crosslinking combined with mass spectrometry is the most powerful tool to investigate nucleic acid binding of proteins. In line with previous studies I used UV crosslinking for a system-wide analysis of RNA-binders and, together with our collaborators, shaped the first RBPome in immune cells. In addition, I expanded the applicability of UV crosslinking to the MS analysis of DNA-binding proteins, which will hopefully

pave the way for its wider use in DNA interaction studies similar to the field of RNA.

Expanding transcription factor interactomes by ChIP-MS

Soon after mass spectrometry became established in protein science, it was employed to study protein-protein interactions. This quickly yielded large interactome maps in different organisms by AP-MS workflows [113–117]. These early publications relied on non-quantitative MS and interactors were identified based on whether or not they were present in the control pulldowns. This requires stringent purification of bait-prey complexes and thus weak interactions escaped detection. Moreover, lists of contaminants needed to be compiled to remove non-specific background binders. The introduction of quantitative proteomics was clearly a breakthrough in protein interactomics. Now, interactors were simply detected by being signficantly enriched over control samples [175, 176]. Quantitative AP-MS captures also weak interactors as it allows for a large background proteome due to less stringent washing buffers. The quantitative dimension also expands the scope of information obtained from the data. Correlation of intensities across multiple pulldowns adds an additional layer of confidence to the identification of true positive interactors. It can also be used to determine the stoichiometry of protein complexes [124]. However, all these publications analyzed the soluble interactions, whereas the interactome of transcription factors additionally depends on the chromatin environment. This challenge has been adressed in several studies [125, 132, 134, 177]. In contrast to conventional workflows, the DNA is not digested, but rather fragmented into stretches of 100-500 bp length. This allows the identification of co-regulatory TF interactomes by preserving the protein complexes on the chromatin. Despite the efforts made in optimizing protocols to characterize the entire trans-regulatory protein interactome, large scale investigation of mammalian gene regulatomes have not been published yet. Even in yeast only one study used native ChIP-MS on about one hundred chromatin-associated proteins [130]. All of the above greatly extended the interactome of these chromatin-regulatory proteins, although these studies used non-quantitative MS, which limited the scope of information obtained from the data. In this thesis I used a crosslinking-based ChIP-MS approach on more

than one hundred *bona fide* transcription factors and set up a streamlined, unbiased data analysis pipeline in Python. The data showed notable differences between broad chromatin-binding TFs and regulators of specific pathways in terms of their interaction profile. Enrichment correlation unraveled functional overlaps between various TFs defined by their interactors. Strikingly, global correlation analysis suggests a role in gene expression regulation for proteins of unknown function by showing association with other transcriptional regulators.

Considering the novel insights that are obtained even in well-studied systems like yeast (see Results section 3.3), I think that a chromatin-centric investigation of transcription factors in mammalian cells would tremendously expand our knowledge on gene regulation. I have also shown in this thesis how important it is to analyze the interactome of transcriptional regulators in order to understand their molecular function as showcased in the collaborative projects (see Results 3.4 and 3.5).

Recently, our group has performed ChIP-MS with improved MNase-based fragmentation of chromatin and streamlined pulldowns on 100 TFs in HeLa cells. First data analysis promises exciting findings, which will lead to a comprehensive human trans-regulatory network. It will expand the knowledge on gene regulation in mammals and will once more highlight the benefits of ChIP-MS similar to what I have shown in yeast (Results section 3.3)

In closing, I am convinced that my results underscore the importance of viewing proteomics of gene regulators as an individual field, which promises a better characterization and novel insights into how cells execute transcriptional control.

Bibliography

[1] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[2] Z. Wang, M. Gerstein, and M. Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, 2009.

[3] M. Karas and F. Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem*, 60(20):2299–301, 1988.

[4] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse.

Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, 1989.

[5] B. T. Chait. Chemistry. mass spectrometry: bottom-up or top-down? *Science*, 314(5796):65–6, 2006.

[6] A. I. Nesvizhskii and R. Aebersold. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics*, 4(10):1419–40, 2005.

[7] T. K. Toby, L. Fornelli, and N. L. Kelleher. Progress in top-down proteomics and the analysis of proteoforms. *Annu Rev Anal Chem (Palo Alto Calif)*, 9(1):499–519, 2016.

[8] S. Sidoli and B. A. Garcia. Middle-down proteomics: a still unexploited resource for chromatin biology. *Expert Rev Proteomics*, 14(7):617–626, 2017.

[9] A. Moradian, A. Kalli, M. J. Sweredoski, and S. Hess. The top-down, middle-down, and bottom-up mass spectrometry approaches for characterization of histone variants and their post-translational modifications. *Proteomics*, 14(4-5):489–97, 2014.

[10] T. Jiang, M. E. Hoover, M. V. Holt, M. A. Freitas, A. G. Marshall, and N. L. Young. Middle-down characterization of the cell cycle dependence of histone h4 posttranslational modifications and proteoforms. *Proteomics*, 18(11):e1700442, 2018.

[11] N. A. Kulak, G. Pichler, I. Paron, N. Nagaraj, and M. Mann. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat Methods*, 11(3):319–24, 2014.

[12] W. H. McDonald, R. Ohi, D. T. Miyamoto, T. J. Mitchison, and J. R. Yates. Comparison of three directly coupled hplc ms/ms strategies for identification of proteins from complex mixtures: single-dimension lc-ms/ms, 2-phase mudpit, and 3-phase mudpit. *International Journal of Mass Spectrometry*, 219(1):245–251, 2002.

[13] A. Shevchenko, M. Wilm, O. Vorm, and M. Mann. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal Chem*, 68(5):850–8, 1996.

[14] A. Shevchenko, H. Tomas, J. Havlis, J. V. Olsen, and M. Mann. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat Protoc*, 1(6):2856–60, 2006.

[15] T. Masuda, M. Tomita, and Y. Ishihama. Phase transfer surfactant-aided trypsin digestion for membrane proteome analysis. *J Proteome Res*, 7(2):731–40, 2008.

[16] J. R. Wisniewski, A. Zougman, N. Nagaraj, and M. Mann. Universal sample preparation method for proteome analysis. *Nat Methods*, 6(5):359–62, 2009.

[17] J. Rappsilber, M. Mann, and Y. Ishihama. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using stagetips. *Nat Protoc*, 2(8):1896–906, 2007.

[18] M. Wilm and M. Mann. Analytical properties of the nanoelectrospray ion source. *Anal Chem*, 68(1):1–8, 1996.

[19] Y. Bian, R. Zheng, F. P. Bayer, C. Wong, Y. C. Chang, C. Meng, D. P. Zolg, M. Reinecke, J. Zecha, S. Wiechmann, S. Heinzlmeir, J. Scherr, B. Hemmer, M. Baynham, A. C. Gingras, O. Boychenko, and B. Kuster. Robust, reproducible and quantitative analysis of thousands of proteomes by micro-flow lc-ms/ms. *Nat Commun*, 11(1):157, 2020.

[20] N. Bache, P. E. Geyer, D. B. Bekker-Jensen, O. Hoerning, L. Falkenby, P. V. Treit, S. Doll, I. Paron, J. B. Muller, F. Meier, J. V. Olsen, O. Vorm, and M. Mann. A novel lc system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics. *Mol Cell Proteomics*, 17(11):2284–2296, 2018.

[21] X. M. Han, A. Aslanian, and J. R. Yates. Mass spectrometry for proteomics. *Current Opinion in Chemical Biology*, 12(5):483–490, 2008.

[22] A. Michalski, E. Damoc, J. P. Hauschild, O. Lange, A. Wieghaus, A. Makarov, N. Nagaraj, J. Cox, M. Mann, and S. Horning. Mass spectrometry-based proteomics using q exactive, a high-performance benchtop quadrupole orbitrap mass spectrometer. *Molecular & Cellular Proteomics*, 10(9), 2011.

[23] R. A. Scheltema, J. P. Hauschild, O. Lange, D. Hornburg, E. Denisov, E. Damoc, A. Kuehn, A. Makarov, and M. Mann. The q exactive hf, a benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field orbitrap analyzer. *Molecular & Cellular Proteomics*, 13(12):3698–3708, 2014.

[24] Thermo Fisher Scientific Inc. Operating manual: Exactive series, 2017.

[25] R. A. Zubarev and A. Makarov. Orbitrap mass spectrometry. *Analytical Chemistry*, 85(11):5288–5296, 2013.

[26] Q. Z. Hu, R. J. Noll, H. Y. Li, A. Makarov, M. Hardman, and R. G. Cooks. The orbitrap: a new mass spectrometer. *Journal of Mass Spectrometry*, 40(4):430–443, 2005.

[27] A. Makarov. Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Analytical Chemistry*, 72(6):1156–1162, 2000.

[28] S. Eliuk and A. Makarov. Evolution of orbitrap mass spectrometry instrumentation. *Annual Review of Analytical Chemistry, Vol 8*, 8:61–80, 2015.

[29] A. Makarov, E. Denisov, and O. Lange. Performance evaluation of a high-field orbitrap mass analyzer. *Journal of the American Society for Mass Spectrometry*, 20(8):1391–1396, 2009.

[30] C. D. Kelstrup, R. R. Jersie-Christensen, T. S. Batth, T. N. Arrey, A. Kuehn, M. Kellmann, and J. V. Olsen. Rapid and deep proteomes by faster sequencing on a benchtop quadrupole ultra-high-field orbitrap mass spectrometer. *Journal of Proteome Research*, 13(12):6187–6195, 2014.

[31] J. Yinon. Tandem mass-spectrometry (ms/ms) and collision-induced dissociation (cid) - an introduction. *Chemistry and Physics of Energetic Materials*, 309:685–693, 1990.

[32] J. V. Olsen, B. Macek, O. Lange, A. Makarov, S. Horning, and M. Mann. Higher-energy c-trap dissociation for peptide modification analysis. *Nature Methods*, 4(9):709–712, 2007.

[33] R. A. Zubarev, N. L. Kelleher, and F. W. McLafferty. Electron capture dissociation of multiply charged protein cations. a nonergodic process. *Journal of the American Chemical Society*, 120(13):3265–3266, 1998.

[34] L. M. Mikesh, B. Ueberheide, A. Chi, J. J. Coon, J. E. P. Syka, J. Shabanowitz, and D. F. Hunt. The utility of etd mass spectrometry in proteomic analysis. *Biochimica Et Biophysica Acta-Proteins and Proteomics*, 1764(12):1811–1822, 2006.

[35] J. E. P. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz, and D. F. Hunt. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, 101(26):9528–9533, 2004.

[36] A. Guthals and N. Bandeira. Peptide identification by tandem mass spectrometry with alternate fragmentation modes. *Molecular & Cellular Proteomics*, 11(9):550–557, 2012.

[37] H. Steen and M. Mann. The abc's (and xyz's) of peptide sequencing. *Nature Reviews Molecular Cell Biology*, 5(9):699–711, 2004.

[38] H. Molina, R. Matthiesen, K. Kandasamy, and A. Pandey. Comprehensive comparison of collision induced dissociation and electron transfer dissociation. *Analytical Chemistry*, 80(13):4825–4835, 2008.

[39] L. M. Mikesh, B. Ueberheide, A. Chi, J. J. Coon, J. E. P. Syka, J. Shabanowitz, and D. F. Hunt. The utility of etd mass spectrometry in proteomic analysis. *Biochimica Et Biophysica Acta-Proteins and Proteomics*, 1764(12):1811–1822, 2006.

[40] A. Michalski, N. Neuhauser, J. Cox, and M. Mann. A systematic investigation into the nature of tryptic hcd spectra. *Journal of Proteome Research*, 11(11):5479–5491, 2012.

[41] J. K. Eng, A. L. Mccormack, and J. R. Yates. An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, 1994.

[42] D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.

[43] J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, and M. Mann. Andromeda: A peptide search engine integrated into the maxquant environment. *Journal of Proteome Research*, 10(4):1794–1805, 2011.

[44] N. Pappireddi, L. Martin, and M. Wuhr. A review on quantitative multiplexed proteomics. *Chembiochem*, 20(10):1210–1224, 2019.

[45] A Hu, WS Noble, and A Wolf-Yadlin. Technical advances in proteomics: new developments in data-independent acquisition [version 1; peer review: 3 approved]. *F1000Research*, 5(419), 2016.

[46] G. L. Andrews, R. A. Dean, A. M. Hawkridge, and D. C. Muddiman. Improving proteome coverage on a ltq-orbitrap using design of experiments. *Journal of the American Society for Mass Spectrometry*, 22(4):773–783, 2011.

[47] A. Michalski, J. Cox, and M. Mann. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent lc-ms/ms. *Journal of Proteome Research*, 10(4):1785–1793, 2011.

[48] F. F. Zhang, W. G. Ge, G. Ruan, X. Cai, and T. N. Guo. Data-independent acquisition mass spectrometry-based proteomics and software tools: A glimpse in 2020. *Proteomics*, 2020.

[49] C. Ludwig, L. Gillet, G. Rosenberger, S. Amon, B. Collins, and R. Aebersold. Data-independent acquisition-based swath-ms for quantitative proteomics: a tutorial. *Molecular Systems Biology*, 14(8), 2018.

[50] F. Meier, P. E. Geyer, S. V. Winter, J. Cox, and M. Mann. Boxcar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nature Methods*, 15(6):440–+, 2018.

[51] L. C. Gillet, P. Navarro, S. Tate, H. Rost, N. Selevsek, L. Reiter, R. Bonner, and R. Aebersold. Targeted data extraction of the ms/ms spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis. *Molecular & Cellular Proteomics*, 11(6), 2012.

[52] M. Mann. Functional and quantitative proteomics using silac. *Nature Reviews Molecular Cell Biology*, 7(12):952–958, 2006.

[53] S. E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and M. Mann. Stable isotope labeling by amino acids in cell culture, silac, as a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics*, 1(5):376–386, 2002.

[54] M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, and B. Kuster. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry*, 389(4):1017–1031, 2007.

[55] S. P. Gygi, B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb, and R. Aebersold. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, 17(10):994–999, 1999.

[56] L. Dayon, A. Hainard, V. Licker, N. Turck, K. Kuhn, D. F. Hochstrasser, P. R. Burkhard, and J. C. Sanchez. Relative quantification of proteins in human cerebrospinal fluids by ms/ms using 6-plex isobaric tags. *Analytical Chemistry*, 80(8):2921–2931, 2008.

[57] S. V. Winter, F. Meier, C. Wichmann, J. Cox, M. Mann, and F. Meissner. Easi-tag enables accurate multiplexed and interference-free ms2-based proteome quantification. *Nature Methods*, 15(7):527–+, 2018.

[58] P. L. Ross, Y. L. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, and D. J. Pappin. Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & Cellular Proteomics*, 3(12):1154–1169, 2004.

[59] S. A. Gerber, J. Rush, O. Stemman, M. W. Kirschner, and S. P. Gygi. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem ms. *Proceedings of the National Academy of Sciences of the United States of America*, 100(12):6940–6945, 2003.

[60] C. J. Tu, J. Li, Q. H. Sheng, M. Zhang, and J. Qu. Systematic assessment of survey scan and ms2-based abundance strategies for label-free quantitative proteomics using high-resolution ms data. *Journal of Proteome Research*, 13(4):2069–2079, 2014.

[61] M. Bantscheff, S. Lemeer, M. M. Savitski, and B. Kuster. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Analytical and Bioanalytical Chemistry*, 404(4):939–965, 2012.

[62] Thibaut Léger, Camille Garcia, Mathieu Videlier, and Jean-Michel Camadro. *Label-Free Quantitative Proteomics in Yeast*, pages 289–307. Springer New York, New York, NY, 2016.

[63] J. R. Wisniewski, P. Ostasiewicz, K. Dus, D. F. Zielinska, F. Gnad, and M. Mann. Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. *Molecular Systems Biology*, 8, 2012.

[64] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[65] J. A. Shapiro. Revisiting the central dogma in the 21st century. *Ann N Y Acad Sci*, 1178:6–28, 2009.

[66] J. McManus, Z. Cheng, and C. Vogel. Next-generation analysis of gene expression regulation–comparing the roles of synthesis and degradation. *Mol Biosyst*, 11(10):2680–9, 2015.

[67] J. D. McGhee and G. D. Ginder. Specific dna methylation sites in the vicinity of the chicken beta-globin genes. *Nature*, 280(5721):419–20, 1979.

[68] V. G. Allfrey, R. Faulkner, and A. E. Mirsky. Acetylation and methylation of histones and their possible role in the regulation of rna synthesis. *Proc Natl Acad Sci U S A*, 51:786–94, 1964.

[69] R. F. Luco, M. Allo, I. E. Schor, A. R. Kornblihtt, and T. Misteli. Epigenetics in alternative pre-mrna splicing. *Cell*, 144(1):16–26, 2011.

[70] P. Cramer. Organization and regulation of gene transcription. *Nature*, 573(7772):45–54, 2019.

[71] T. I. Lee and R. A. Young. Transcriptional regulation and its misregulation in disease. *Cell*, 152(6):1237–51, 2013.

[72] N. J. Fuda, M. B. Ardehali, and J. T. Lis. Defining mechanisms that regulate rna polymerase ii transcription in vivo. *Nature*, 461(7261):186–192, 2009.

[73] G. Stampfel, T. Kazmar, O. Frank, S. Wienerroither, F. Reiter, and A. Stark. Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature*, 528(7580):147–51, 2015.

[74] A. Jolma, Y. Yin, K. R. Nitta, K. Dave, A. Popov, M. Taipale, M. Enge, T. Kivioja, E. Morgunova, and J. Taipale. Dna-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, 527(7578):384–8, 2015.

[75] A. C. Mullen, D. A. Orlando, J. J. Newman, J. Loven, R. M. Kumar, S. Bilodeau, J. Reddy, M. G. Guenther, R. P. DeKoter, and R. A. Young. Master transcription factors determine cell-type-specific responses to tgf-beta signaling. *Cell*, 147(3):565–76, 2011.

[76] M. Ptashne and A. Gann. Transcriptional activation by recruitment. *Nature*, 386(6625):569–577, 1997.

[77] K. Adelman and J. T. Lis. Promoter-proximal pausing of rna polymerase ii: emerging roles in metazoans. *Nature Reviews Genetics*, 13(10):720–731, 2012.

[78] V. Haberle and A. Stark. Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology*, 19(10):621–637, 2018.

[79] A. Oruba, S. Saccani, and D. van Essen. Role of cell-type specific nucleosome positioning in inducible activation of mammalian promoters. *Nature Communications*, 11(1), 2020.

[80] T. C. Voss, R. L. Schiltz, M. H. Sung, P. M. Yen, J. A. Stamatoyannopoulos, S. C. Biddie, T. A. Johnson, T. B. Miranda, S. John, and G. L. Hager. Dynamic exchange at regulatory elements during chromatin remodeling underlies assisted loading mechanism. *Cell*, 146(4):544–554, 2011.

[81] M. L. Dechassa, A. Sabri, S. Pondugula, S. R. Kassabov, N. Chatterjee, M. P. Kladde, and B. Bartholomew. Swi/snf has intrinsic nucleosome disassembly activity that is dependent on adjacent nucleosomes. *Molecular Cell*, 38(4):590–602, 2010.

[82] M. Li, A. Hada, P. Sen, L. Olufemi, M. A. Hall, B. Y. Smith, S. Forth, J. N. McKnight, A. Patel, G. D. Bowman, B. Bartholomew, and M. D. Wang. Dynamic regulation of transcription factors by nucleosome remodeling. *Elife*, 4, 2015.

[83] A. M. Deaton and A. Bird. Cpg islands and the regulation of transcription. *Genes Dev*, 25(10):1010–22, 2011.

[84] E. Li and Y. Zhang. Dna methylation in mammals. *Cold Spring Harb Perspect Biol*, 6(5):a019133, 2014.

[85] S. K. Kurdistani, S. Tavazoie, and M. Grunstein. Mapping global histone acetylation patterns to gene expression. *Cell*, 117(6):721–733, 2004.

[86] Z. B. Wang, C. Z. Zang, K. R. Cui, D. E. Schones, A. Barski, W. Q. Peng, and K. J. Zhao. Genome-wide mapping of hats and hdacs reveals

distinct functions in active and inactive genes. *Cell*, 138(5):1019–1031, 2009.

[87] A. Lardenois, I. Stuparevic, Y. C. Liu, M. J. Law, E. Becker, F. Smagulova, K. Waern, M. H. Guilleux, J. Horecka, A. Chu, C. Kervarrec, R. Strich, M. Snyder, R. W. Davis, L. M. Steinmetz, and M. Primig. The conserved histone deacetylase rpd3 and its dna binding subunit ume6 control dynamic transcript architecture during mitotic growth and meiotic development. *Nucleic Acids Research*, 43(1):115–128, 2015.

[88] R. Srivastava, K. M. Rai, B. Pandey, S. P. Singh, and S. V. Sawant. Spt-ada-gcn5-acetyltransferase (saga) complex in plants: Genome wide identification, evolutionary conservation and functional determination. *Plos One*, 10(8), 2015.

[89] P. J. Horn and C. L. Peterson. Molecular biology: Chromatin higher order folding: Wrapping up transcription. *Science*, 297(5588):1824–1827, 2002.

[90] T. Fujisawa and P. Filippakopoulos. Functions of bromodomain-containing proteins and their roles in homeostasis and cancer. *Nature Reviews Molecular Cell Biology*, 18(4):246–262, 2017.

[91] A. Jambhekar, A. Dhall, and Y. Shi. Roles and regulation of histone methylation in animal development. *Nature Reviews Molecular Cell Biology*, 20(10):625–641, 2019.

[92] L. J. Pilaz and D. L. Silver. Post-transcriptional regulation in corticogenesis: how rna-binding proteins help build the brain. *Wiley Interdisciplinary Reviews-Rna*, 6(5):501–515, 2015.

[93] S. Carpenter, E. P. Ricci, B. C. Mercier, M. J. Moore, and K. A. Fitzgerald. Post-transcriptional regulation of gene expression in innate immunity. *Nature Reviews Immunology*, 14(6):361–376, 2014.

[94] K. C. Martin and A. Ephrussi. mrna localization: Gene expression in the spatial dimension. *Cell*, 136(4):719–730, 2009.

[95] F. E. Baralle and J. Giudice. Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology*, 18(7):437–451, 2017.

[96] M. Chen and J. L. Manley. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nature Reviews Molecular Cell Biology*, 10(11):741–754, 2009.

[97] Y. Feng, M. Chen, and J. L. Manley. Phosphorylation switches the general splicing repressor srp38 to a sequence-specific activator. *Nature Structural & Molecular Biology*, 15(10):1040–1048, 2008.

[98] T. O. Tange, C. K. Damgaard, S. Guth, J. Valcarcel, and J. Kjems. The hnrnp a1 protein regulates hiv-1 tat splicing via a novel intron silencer element. *Embo Journal*, 20(20):5748–5758, 2001.

[99] C. K. Vuong, D. L. Black, and S. K. Zheng. The neurogenetics of alternative splicing. *Nature Reviews Neuroscience*, 17(5):265–281, 2016.

[100] D. D. Licatalosi. Roles of rna-binding proteins and post-transcriptional regulation in driving male germ cell development in the mouse. *Rna Processing: Disease and Genome-Wide Probing*, 907:123–151, 2016.

[101] J. G. Conboy. Rna splicing during terminal erythropoiesis. *Current Opinion in Hematology*, 24(3):215–221, 2017.

[102] D. C. Di Giammartino, K. Nishida, and J. L. Manley. Mechanisms and consequences of alternative polyadenylation. *Molecular Cell*, 43(6):853–866, 2011.

[103] R. Sandberg, J. R. Neilson, A. Sarma, P. A. Sharp, and C. B. Burge. Proliferating cells express mrnas with shortened 3 ' untranslated regions and fewer microrna target sites. *Science*, 320(5883):1643–1647, 2008.

[104] Y. Takagaki and J. L. Manley. Levels of polyadenylation factor cstf-64 control igm heavy chain mrna accumulation and other events associated with b cell differentiation. *Molecular Cell*, 2(6):761–771, 1998.

[105] D. L. Liu, J. M. Brockman, B. Dass, L. N. Hutchins, P. Singh, J. R. McCarrey, C. C. MacDonald, and J. H. Graber. Systematic variation in mrna 3 '-processing signals during mouse spermatogenesis. *Nucleic Acids Research*, 35(1):234–246, 2007.

[106] B. Tian, J. Hu, H. B. Zhang, and C. S. Lutz. A large-scale analysis of mrna polyadenylation of human and mouse genes. *Nucleic Acids Research*, 33(1):201–212, 2005.

[107] R. Parker and H. W. Song. The enzymes and control of eukaryotic mrna turnover. *Nature Structural & Molecular Biology*, 11(2):121–127, 2004.

[108] D. L. Makino, B. Schuch, E. Stegmann, M. Baumgartner, C. Basquin, and E. Conti. Rna degradation paths in a 12-subunit nuclear exosome complex. *Nature*, 524(7563):54–U89, 2015.

[109] R. E. Halbeisen, A. Galgano, T. Scherrer, and A. P. Gerber. Post-transcriptional gene regulation: From genome-wide studies to principles. *Cellular and Molecular Life Sciences*, 65(5):798–813, 2008.

[110] D. Beisang and P. R. Bohjanen. Perspectives on the are as it turns 25 years old. *Wiley Interdisciplinary Reviews-Rna*, 3(5):719–731, 2012.

[111] W. Filipowicz, S. N. Bhattacharyya, and N. Sonenberg. Mechanisms of post-transcriptional regulation by microrna: are the answers in sight? *Nature Reviews Genetics*, 9(2):102–114, 2008.

[112] N. Standart and R. J. Jackson. Microrna repress translation of m(7) gppp-capped target mrnas in vitro by inhibiting initiation and promoting deadenylation. *Genes & Development*, 21(16):1975–1982, 2007.

[113] A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. Functional

organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–7, 2002.

[114] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sorensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. Hogue, D. Figeys, and M. Tyers. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868):180–3, 2002.

[115] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrin-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. St Onge, S. Ghanny, M. H. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, and J. F. Greenblatt. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, 440(7084):637–43, 2006.

[116] A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M. A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A. M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, and G. Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–6, 2006.

[117] T. Bouwmeester, A. Bauch, H. Ruffner, P. O. Angrand, G. Bergamini, K. Croughton, C. Cruciat, D. Eberhard, J. Gagneur, S. Ghidelli, C. Hopf, B. Huhse, R. Mangano, A. M. Michon, M. Schirle, J. Schlegl, M. Schwab, M. A. Stein, A. Bauer, G. Casari, G. Drewes, A. C. Gavin, D. B. Jackson, G. Joberty, G. Neubauer, J. Rick, B. Kuster, and G. Superti-Furga. A physical and functional map of the human tnf-alpha/nf-kappa b signal transduction pathway. *Nat Cell Biol*, 6(2):97–105, 2004.

[118] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Seraphin. A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*, 17(10):1030–2, 1999.

[119] M. Vermeulen, N. C. Hubner, and M. Mann. High confidence determination of specific protein-protein interactions using quantitative mass spectrometry. *Curr Opin Biotechnol*, 19(4):331–7, 2008.

[120] M. E. Sowa, E. J. Bennett, S. P. Gygi, and J. W. Harper. Defining the human deubiquitinating enzyme interaction landscape. *Cell*, 138(2):389–403, 2009.

[121] H. Choi, T. Glatter, M. Gstaiger, and A. I. Nesvizhskii. Saint-ms1: protein-protein interaction scoring using label-free intensity data in affinity purification-mass spectrometry experiments. *J Proteome Res*, 11(4):2619–24, 2012.

[122] E. C. Keilhauer, M. Y. Hein, and M. Mann. Accurate protein complex retrieval by affinity enrichment mass spectrometry (ae-ms) rather than affinity purification mass spectrometry (ap-ms). *Mol Cell Proteomics*, 14(1):120–35, 2015.

[123] F. Hosp, R. A. Scheltema, H. C. Eberl, N. A. Kulak, E. C. Keilhauer, K. Mayr, and M. Mann. A double-barrel liquid chromatography-tandem mass spectrometry (lc-ms/ms) system to quantify 96 interactomes per day. *Mol Cell Proteomics*, 14(7):2030–41, 2015.

[124] M. Y. Hein, N. C. Hubner, I. Poser, J. Cox, N. Nagaraj, Y. Toyoda, I. A. Gak, I. Weisswange, J. Mansfeld, F. Buchholz, A. A. Hyman, and

M. Mann. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell*, 163(3):712–23, 2015.

[125] J. P. Lambert, L. Mitchell, A. Rudner, K. Baetz, and D. Figeys. A novel proteomics approach for the discovery of chromatin-associated protein networks. *Mol Cell Proteomics*, 8(4):870–82, 2009.

[126] M. Wierer and M. Mann. Proteomics to study dna-bound and chromatin-associated gene regulatory complexes. *Hum Mol Genet*, 25(R2):R106–R114, 2016.

[127] M. Soldi and T. Bonaldi. The chrop approach combines chip and mass spectrometry to dissect locus-specific proteomic landscapes of chromatin. *J Vis Exp*, (86), 2014.

[128] H. Mohammed, C. D'Santos, A. A. Serandour, H. R. Ali, G. D. Brown, A. Atkins, O. M. Rueda, K. A. Holmes, V. Theodorou, J. L. Robinson, W. Zwart, A. Saadi, C. S. Ross-Innes, S. F. Chin, S. Menon, J. Stingl, C. Palmieri, C. Caldas, and J. S. Carroll. Endogenous purification reveals greb1 as a key estrogen receptor regulatory factor. *Cell Rep*, 3(2):342–9, 2013.

[129] T. Bartke, M. Vermeulen, B. Xhemalce, S. C. Robson, M. Mann, and T. Kouzarides. Nucleosome-interacting proteins regulated by dna and histone methylation. *Cell*, 143(3):470–84, 2010.

[130] J. P. Lambert, J. Fillingham, M. Siahbazi, J. Greenblatt, K. Baetz, and D. Figeys. Defining the budding yeast chromatin-associated interactome. *Mol Syst Biol*, 6:448, 2010.

[131] X. Ji, D. B. Dadon, B. J. Abraham, T. I. Lee, R. Jaenisch, J. E. Bradner, and R. A. Young. Chromatin proteomic profiling reveals novel proteins associated with histone-marked genomic regions. *Proc Natl Acad Sci U S A*, 112(12):3841–6, 2015.

[132] E. Engelen, J. H. Brandsma, M. J. Moen, L. Signorile, D. H. Dekkers, J. Demmers, C. E. Kockx, Z. Ozgur, IJcken W. F. van, D. L. van den Berg, and R. A. Poot. Proteins that bind regulatory regions identified

by histone modification chromatin immunoprecipitations and mass spectrometry. *Nat Commun*, 6:7155, 2015.

[133] H. Mohammed, I. A. Russell, R. Stark, O. M. Rueda, T. E. Hickey, G. A. Tarulli, A. A. Serandour, S. N. Birrell, A. Bruna, A. Saadi, S. Menon, J. Hadfield, M. Pugh, G. V. Raj, G. D. Brown, C. D'Santos, J. L. Robinson, G. Silva, R. Launchbury, C. M. Perou, J. Stingl, C. Caldas, W. D. Tilley, and J. S. Carroll. Progesterone receptor modulates eralpha action in breast cancer. *Nature*, 523(7560):313–7, 2015.

[134] C. I. Wang, A. A. Alekseyenko, G. LeRoy, A. E. Elia, A. A. Gorchakov, L. M. Britton, S. J. Elledge, P. V. Kharchenko, B. A. Garcia, and M. I. Kuroda. Chromatin proteins captured by chip-mass spectrometry are linked to dosage compensation in drosophila. *Nat Struct Mol Biol*, 20(2):202–9, 2013.

[135] M. North, B. D. Gaytan, C. Romero, V. Y. De la Rosa, A. Loguinov, M. T. Smith, L. P. Zhang, and C. D. Vulpe. Functional toxicogenomic profiling expands insight into modulators of formaldehyde toxicity in yeast. *Frontiers in Genetics*, 7, 2016.

[136] D. Yasokawa, S. Murata, Y. Iwahashi, E. Kitagawa, R. Nakagawa, T. Hashido, and H. Iwahashi. Toxicity of methanol and formaldehyde towards saccharomyces cerevisiae as assessed by dna microarray analysis. *Applied Biochemistry and Biotechnology*, 160(6):1685–1698, 2010.

[137] R. W. Yao, Y. Wang, and L. L. Chen. Cellular functions of long noncoding rnas. *Nat Cell Biol*, 21(5):542–551, 2019.

[138] C. Chu, K. Qu, F. L. Zhong, S. E. Artandi, and H. Y. Chang. Genomic maps of long noncoding rna occupancy reveal principles of rna-chromatin interactions. *Mol Cell*, 44(4):667–78, 2011.

[139] M. Gauchier, G. van Mierlo, M. Vermeulen, and J. Dejardin. Purification and enrichment of specific chromatin loci. *Nat Methods*, 17(4):380–389, 2020.

[140] J. Dejardin and R. E. Kingston. Purification of proteins associated with specific genomic loci. *Cell*, 136(1):175–86, 2009.

[141] S. D. Byrum, A. Raman, S. D. Taverna, and A. J. Tackett. Chapms: a method for identification of proteins and histone posttranslational modifications at a single genomic locus. *Cell Rep*, 2(1):198–205, 2012.

[142] X. Liu, Y. Zhang, Y. Chen, M. Li, F. Zhou, K. Li, H. Cao, M. Ni, Y. Liu, Z. Gu, K. E. Dickerson, S. Xie, G. C. Hon, Z. Xuan, M. Q. Zhang, Z. Shao, and J. Xu. In situ capture of chromatin interactions by biotinylated dcas9. *Cell*, 170(5):1028–1043 e19, 2017.

[143] K. C. Smith and R. T. Aplin. A mixed photoproduct of uracil and cysteine (5-s-cysteine-6-hydrouracil). a possible model for the in vivo cross-linking of deoxyribonucleic acid and protein by ultraviolet light. *Biochemistry*, 5(6):2125–30, 1966.

[144] K. C. Smith. A mixed photoproduct of thymine and cysteine: 5-s-cysteine, 6-hydrothymine. *Biochem Biophys Res Commun*, 39(6):1011–6, 1970.

[145] J. R. Greenberg. Ultraviolet light-induced crosslinking of mrna to proteins. *Nucleic Acids Res*, 6(2):715–32, 1979.

[146] K. Moller and R. Brimacombe. Specific cross-linking of proteins s7 and l4 to ribosomal rna, by uv irradiation of escherichia coli ribosomal subunits. *Mol Gen Genet*, 141(4):343–55, 1975.

[147] H. J. Schoemaker and P. R. Schimmel. Photo-induced joining of a transfer rna with its cognate aminoacyl-transfer rna synthetase. *J Mol Biol*, 84(4):503–13, 1974.

[148] A. Markovitz. Ultraviolet light-induced stable complexes of dna and dna polymerase. *Biochim Biophys Acta*, 281(4):522–34, 1972.

[149] Z. Hillel and C. W. Wu. Photochemical cross-linking studies on the interaction of escherichia coli rna polymerase with t7 dna. *Biochemistry*, 17(15):2954–61, 1978.

[150] H. N. Schott and M. D. Shetlar. Photochemical addition of amino acids to thymine. *Biochem Biophys Res Commun*, 59(3):1112–6, 1974.

[151] D. Angelov, VYu Stefanovsky, S. I. Dimitrov, V. R. Russanova, E. Keskinova, and I. G. Pashev. Protein-dna crosslinking in reconstituted nucleohistone, nuclei and whole cells by picosecond uv laser irradiation. *Nucleic Acids Res*, 16(10):4525–38, 1988.

[152] I. G. Pashev, S. I. Dimitrov, and D. Angelov. Crosslinking proteins to nucleic acids by ultraviolet laser irradiation. *Trends Biochem Sci*, 16(9):323–6, 1991.

[153] E. I. Budowsky, M. S. Axentyeva, G. G. Abdurashidova, N. A. Simukova, and L. B. Rubin. Induction of polynucleotide-protein cross-linkages by ultraviolet irradiation. peculiarities of the high-intensity laser pulse irradiation. *Eur J Biochem*, 159(1):95–101, 1986.

[154] C. Russmann, M. Truss, A. Fix, C. Naumer, T. Herrmann, J. Schmitt, J. Stollhof, R. Beigang, and M. Beato. Crosslinking of progesterone receptor to dna using tuneable nanosecond, picosecond and femtosecond uv laser pulses. *Nucleic Acids Res*, 25(12):2478–84, 1997.

[155] F. Gebauer, T. Preiss, and M. W. Hentze. From cis-regulatory elements to complex rnps and back. *Cold Spring Harb Perspect Biol*, 4(7):a012245, 2012.

[156] J. Ule, K. B. Jensen, M. Ruggiu, A. Mele, A. Ule, and R. B. Darnell. Clip identifies nova-regulated rna networks in the brain. *Science*, 302(5648):1212–5, 2003.

[157] X. Li, J. Song, and C. Yi. Genome-wide mapping of cellular protein-rna interactions enabled by chemical crosslinking. *Genomics Proteomics Bioinformatics*, 12(2):72–8, 2014.

[158] G. Dreyfuss, Y. D. Choi, and S. A. Adam. Characterization of heterogeneous nuclear rna-protein complexes in vivo with monoclonal antibodies. *Mol Cell Biol*, 4(6):1104–14, 1984.

[159] A. Castello, B. Fischer, K. Eichelbaum, R. Horos, B. M. Beckmann, C. Strein, N. E. Davey, D. T. Humphreys, T. Preiss, L. M. Steinmetz,

J. Krijgsveld, and M. W. Hentze. Insights into rna biology from an atlas of mammalian mrna-binding proteins. *Cell*, 149(6):1393–406, 2012.

[160] A. G. Baltz, M. Munschauer, B. Schwanhausser, A. Vasile, Y. Murakawa, M. Schueler, N. Youngs, D. Penfold-Brown, K. Drew, M. Milek, E. Wyler, R. Bonneau, M. Selbach, C. Dieterich, and M. Landthaler. The mrna-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol Cell*, 46(5):674–90, 2012.

[161] B. M. Beckmann, R. Horos, B. Fischer, A. Castello, K. Eichelbaum, A. M. Alleaume, T. Schwarzl, T. Curk, S. Foehr, W. Huber, J. Krijgsveld, and M. W. Hentze. The rna-binding proteomes from yeast to man harbour conserved enigmrbps. *Nat Commun*, 6:10127, 2015.

[162] J. Trendel, T. Schwarzl, R. Horos, A. Prakash, A. Bateman, M. W. Hentze, and J. Krijgsveld. The human rna-binding proteome and its dynamics during translational arrest. *Cell*, 176(1-2):391–403 e19, 2019.

[163] R. M. L. Queiroz, T. Smith, E. Villanueva, M. Marti-Solano, M. Monti, M. Pizzinga, D. M. Mirea, M. Ramakrishna, R. F. Harvey, V. Dezi, G. H. Thomas, A. E. Willis, and K. S. Lilley. Comprehensive identification of rna-protein interactions in any organism using orthogonal organic phase separation (oops). *Nat Biotechnol*, 37(2):169–178, 2019.

[164] VYu Stefanovsky, S. I. Dimitrov, D. Angelov, and I. G. Pashev. Interactions of acetylated histones with dna as revealed by uv laser induced histone-dna crosslinking. *Biochem Biophys Res Commun*, 164(1):304–10, 1989.

[165] S. I. Dimitrov, VYu Stefanovsky, L. Karagyozov, D. Angelov, and I. G. Pashev. The enhancers and promoters of the xenopus laevis ribosomal spacer are associated with histones upon active transcription of the ribosomal genes. *Nucleic Acids Res*, 18(21):6393–7, 1990.

[166] T. Moss, S. I. Dimitrov, and D. Houde. Uv-laser crosslinking of proteins to dna. *Methods*, 11(2):225–34, 1997.

[167] A. K. Nagaich and G. L. Hager. Uv laser cross-linking: a real-time assay to study dynamic protein/dna interactions during chromatin remodeling. *Sci STKE*, 2004(256):pl13, 2004.

[168] J. Walter and M. D. Biggin. Measurement of in vivo dna binding by sequence-specific transcription factors using uv cross-linking. *Methods*, 11(2):215–24, 1997.

[169] K. E. Boyd and P. J. Farnham. Myc versus usf: discrimination at the cad gene is determined by core promoter elements. *Mol Cell Biol*, 17(5):2529–37, 1997.

[170] A. Steube, T. Schenk, A. Tretyakov, and H. P. Saluz. High-intensity uv laser chip-seq for the study of protein-dna interactions in living cells. *Nat Commun*, 8(1):1303, 2017.

[171] W. K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, and E. K. O'Shea. Global analysis of protein localization in budding yeast. *Nature*, 425(6959):686–91, 2003.

[172] C. G. de Boer and T. R. Hughes. Yetfasco: a database of evaluated yeast transcription factor sequence specificities. *Nucleic Acids Research*, 40(D1):D169–D179, 2012.

[173] F. C. Y. Lee and J. Ule. Advances in clip technologies for studies of protein-rna interactions. *Molecular Cell*, 69(3):354–369, 2018.

[174] K. Kramer, T. Sachsenberg, B. M. Beckmann, S. Qamar, K. L. Boon, M. W. Hentze, O. Kohlbacher, and H. Urlaub. Photo-cross-linking and high-resolution mass spectrometry for assignment of rna-binding sites in rna-binding proteins. *Nat Methods*, 11(10):1064–70, 2014.

[175] B. Blagoev, I. Kratchmarova, S. E. Ong, M. Nielsen, L. J. Foster, and M. Mann. A proteomics strategy to elucidate functional protein-protein interactions applied to egf signaling. *Nat Biotechnol*, 21(3):315–8, 2003.

[176] J. A. Ranish, E. C. Yi, D. M. Leslie, S. O. Purvine, D. R. Goodlett, J. Eng, and R. Aebersold. The study of macromolecular complexes by quantitative proteomics. *Nat Genet*, 33(3):349–55, 2003.

[177] M. R. Rafiee, C. Girardot, G. Sigismondo, and J. Krijgsveld. Expanding the circuitry of pluripotency by selective isolation of chromatin-associated proteins. *Mol Cell*, 64(3):624–635, 2016.

Acknowledgements

First of all, I would like to say thank you to my supervisor Michael Wierer who provided great guidance during my PhD studies. Thank you for teaching me how to approach scientific questions and communicate results. Also for helping with your experience on MS protocols and data analysis.

A special thank you also goes to Matthias Mann for being my doctoral advisor, for being so supportive along the way and for hosting such a great lab.

I'd like to thank Alessandro Scacchetti and Prof. Dr. Peter Becker for the great collaboration. Thank you also to Kai Höfig for the fruitful discussions during our collaboration.

I am also grateful to Christoph Russmann and Roland Ackermann for providing so much help with the laser.
I'd also like to thank Igor for always helping me to solve mass spectromety issues and being in such a good mood no matter what.

My appreciation goes to all Mann group members for being so helpful and creating such a nice atmosphere.

Many thanks goes to my office, especially Paola, for the great atmosphere and always helping with any problems in the lab.

A huge thanks also goes to my friends for all the fun activities and being the best distraction from work. Special thanks to Matze for always being so positive and being the best best man possible.

A huge thank you goes to my family and especially my siblings for being an inspiration and spending great holidays together.
I am also deeply grateful to my parents for always being so encouraging and helpful and being the greatest support I could think of.
Finally, I would like to thank Lisa for all her support and love throughout the years, thank you!