

**Mass spectrometry-based proteomics to study the
development of *D. melanogaster* and telomere-
associated proteins with a focus on human ZBTB10**

Dissertation

zur Erlangung des Grades

„Doktor der Naturwissenschaften“

am Fachbereich Biologie

der Johannes Gutenberg-Universität in Mainz

Alina Bluhm

Mainz, Januar 2019

Abstract

Mass spectrometry-based proteomics is a versatile tool that can be used to analyze proteins in a global and unbiased fashion. In this thesis, state-of-the-art LC-MS/MS technology has been used to 1) investigate protein expression changes during development by quantitative proteomics, 2) identify novel telomere-associated proteins and 3) determine protein-protein interactions by interactomics.

Article I presents the first protein-based study covering the complete life cycle of the commonly used model organism *Drosophila melanogaster*. Using label-free quantitative mass spectrometry, we generated two large datasets: the lifecycle proteome and a highly time-resolved embryogenesis proteome. We identified stage- and gender-specific proteins, as well as maternally-provided proteins that might play a role in oogenesis or early development. Moreover, we observed only moderate correlation at all time points between mRNA and protein, which is important to consider in developmental studies. By integrating tissue-specific expression data, we further inferred possible roles for uncharacterized proteins in the development of certain tissues. These comprehensive proteomic datasets serve as powerful resources for future investigations.

Article II expands our current knowledge about telomere-associated proteins. We identified TTAGGG-repeat binders in 16 vertebrate cell lines by affinity purification coupled to quantitative mass spectrometry. We detected an unexpected loss of telomeric binding by the shelterin complex member TRF1 in non-therian species. In addition, we identified several novel proteins associated with the telomeric sequence presenting interesting candidates for future functional studies at telomeres.

In the most recent publication (*Article III*), we focused on one of these candidates: ZBTB10. We analyzed the binding behavior of ZBTB10 to different telomeric repeats and determined direct binding via its two zinc fingers. While we confirmed binding to the TTAGGG motif, binding affinity for the TTGGGG telomere variant repeat was two-times higher. We observed localization of ZBTB10 at a subset of telomeres in ALT-maintaining U2OS cells. To gain insights into the possible molecular function of ZBTB10, we analyzed

the interaction network in two human cell lines by quantitative mass spectrometry. Of note, the shelterin complex members TRF2 and RAP1 were reproducibly identified as ZBTB10 interactors in both cell lines. The N-terminus of TRF2 was defined to be necessary and sufficient for binding ZBTB10. Therefore, ZBTB10 presents a novel telomere-binding protein at ALT telomeres.

In conclusion, I investigated a wide range of research questions that can be addressed by MS-based proteomics and interactomics. I then further characterized a novel telomere interactor derived from our phylointeractomics screen.

Zusammenfassung

In dieser Doktorarbeit wurde die hochmoderne LC-MS/MS Technologie verwendet um 1) Proteinexpressionsänderungen während der Entwicklung zu untersuchen, 2) Proteine zu identifizieren, die spezifisch an die Telomersequenz binden und 3) Protein-Protein Interaktionen zu bestimmen.

Artikel I präsentiert die bisher erste proteinbasierte Studie, die den gesamten Lebenszyklus des genetischen Modellorganismus *Drosophila melanogaster* abdeckt. Mittels quantitativer Massenspektrometrie haben wir zwei Datensätze erstellt: ein Proteom des gesamten Lebenszyklus und ein zeitlich hochaufgelöstes Proteom während der Embryogenese. Wir haben Proteine analysiert, die spezifisch für jedes Lebensstadium und für das Geschlecht sind. Außerdem berichten wir erstmals von Proteinen, die von der Mutter weitergegeben wurden und somit wichtig für Oogenese oder die frühe embryonale Entwicklung sein können. Zu allen Zeitpunkten haben wir lediglich eine moderate Korrelation zwischen mRNA und Proteinexpression festgestellt; ein wichtiger Aspekt, der bei Studien während der Entwicklung zu beachten ist. Die Einarbeitung von gewebespezifischen Expressionsdaten ermöglichte uns die Funktion von bisher nicht charakterisierten Proteinen in der Entwicklung eines bestimmten Gewebes abzuleiten. Die umfangreichen Proteomikdatensätze stehen auf unserer interaktiven Webseite zur Verfügung.

Artikel II verfolgt das Ziel unser momentanes Wissen über Telomer-bindende Proteine zu erweitern. Mittels Affinitätsreinigung und quantitativer Massenspektrometrie haben wir TTAGGG-bindende Proteine aus Zelllinien von 16 Wirbeltieren identifiziert. Dabei wurde ein Unterschied in der Bindung des Shelterin Komplexmitglieds TRF1 in Vögeln Amphibien und Fischen an die Telomersequenz festgestellt. Zusätzlich haben wir neue Proteine identifiziert, die spezifisch mit TTAGGG assoziieren. Diese stellen interessante Kandidaten für zukünftige funktionelle Studien an Telomeren dar.

In unserer neuesten Publikation (*Artikel III*) haben wir einen dieser Kandidaten, ZBTB10, näher untersucht. ZBTB10 bindet mit seinen beiden Zinkfingern direkt an Telomersequenzen. Interessanterweise haben wir eine doppelt so hohe Bindeaffinität an die variable Telomersequenz TTGGGG im Vergleich zu TTAGGG gemessen. ZBTB10 lokalisiert an einen Teil der Telomere *in vivo*. Um Erkenntnisse über die Funktion von ZBTB10 zu bekommen, haben wir das Interaktionsnetzwerk in zwei humanen Zelllinien mittels quantitativer Massenspektrometrie bestimmt. Dabei wurden die beiden Shelterin Komplexmitglieder TRF2 und RAP1 reproduzierbar als spezifische Interaktionspartner von ZBTB10 identifiziert. Der N-Terminus von TRF2 ist notwendig und ausreichend, um die Interaktion mit ZBTB10 zu ermöglichen. Damit ist ZBTB10 ein neues Telomer-bindendes Protein.

Zusammenfassend habe ich ein breites Spektrum an Fragen mittels MS-basierter Proteomik und Interaktomik untersucht und ein neues Telomer-bindendes Protein unseres Phylointerakomikscreens charakterisiert.

List of publications

Article I:

The developmental proteome of *Drosophila melanogaster*

Article II:

Phylointeractomics reconstructs functional evolution of protein binding

Article III:

ZBTB10 binds the telomeric variant repeat TTGGGG and interacts with TRF2

Additional publications

Quantifying post-transcriptional regulation in the development of *D. melanogaster*

ZBTB48 is both a vertebrate telomere-binding protein and a transcriptional activator

Reader interactome of epigenetic histone marks in birds

Abbreviations

53BP1	p53-binding protein 1
4-nqo	4-Nitroquinoline 1-oxide
aa	amino acid
ALT	alternative lengthening of telomeres
AP	affinity purification
APB	ALT-associated PML bodies
AQUA	absolute quantification
Arg, R	arginine
Asp, D	aspartic acid
ATM	Ataxia telangiectasia mutated
ATR	Ataxia telangiectasia and Rad3 related
BAC	bacterial artificial chromosome
BIR	break-induced replication
bp	base pairs
BTB	Broad-complex, Tramtrack and Bric a brac
C18	octadecyl carbon chain
CAA	2-chloroacetamide
ChIP	chromatin immunoprecipitation
CpG	Cytosine-phosphate-Guanine
DC	direct current
DDA	data-dependent acquisition
DDR	DNA damage response
DIA	data-independent acquisition
ds	double-strand
DSB	DNA double-strand break
DTT	dithiothreitol
EMSA	electro mobility shift assay
enChIP	engineered DNA-binding molecule-mediated
ESI	electrospray ionization
FDR	false discovery rate
FISH	fluorescence in situ hybridization

FT-ICR	Fourier-transform ion cyclotron resonance
Glu, E	glutamic acid
H ₂ O ₂	water peroxide
HCD	higher-energy collisional dissociation
HDR	homology directed repair
HPLC	high performance liquid chromatography
HR	homologous recombination
IAA	2-iodoacetamide
iBAQ	intensity-based absolute quantification
ICAT	isotope-coded affinity tag
iTRAQ	isobaric tag for relative and absolute quantification
ITS	interstitial telomeric sequences
kD	kilo Dalton
KO	knockout
LC	liquid chromatography
LFQ	label-free quantification
Lys, K	lysine
m/z	mass-to-charge ratio
Mb	mega bases
miR	microRNA
MMS	methyl methanesulfonate
MRM	multi reaction monitoring
MS	mass spectrometry
MZT	maternal-to-zygotic transition
NaBH ₃ CN	sodium cyanoborohydride
NF-κB	nuclear factor kappa-light-chain-enhancer of activated B cells
NHEJ	non-homologous end joining
nt	nucleotides
PAGE	polyacrylamide gel electrophoresis
PEP	posterior error probability
PiCh	proteomics of isolated chromatin segments
PML	promyelocytic leukaemia
Pol	polymerase
POT1	protection of telomeres 1

ppm	parts per million
pre-RC	pre-replication complex
PrEST	Protein Epitope Signature Tag
PTM	post-translational modification
QconCAT	quantitative concatamer
QTIP	quantitative telomeric chromatin isolation protocol
RAP1	repressor activator protein 1
Rb	Retinoblastoma
RF	radio frequency
RNP	ribonucleoprotein
ROS	reactive oxygen species
SDS	sodium dodecyl sulfate
SILAC	stable isotope labeling of amino acids in cell culture
SP1,3,4	specificity protein 1,3,4
SP3	Single-Pot Solid-Phase-enhanced Sample Preparation
SRM	single reaction monitoring
ss	single-strand
SWATH	sequential windowed acquisition of all theoretical fragment ion mass spectra
TAL	transcription activator-like
TBD	telomere biology disorders
TERRA	telomeric repeat containing RNA
TERT	telomerase reverse transcriptase
TIF	telomere induced foci
TIN2	TRF1 interacting nuclear factor 2
TiO ₂	titanium dioxide
TMT	tandem mass tag
TOF	time of flight
TPE	telomere position effect
TPP1	TINF2-interacting protein 2
TR, TERC	telomerase RNA
TRAP	telomere repeat amplification protocol
TRF1, 2	telomere repeat binding factor 1, 2
T-SCE	telomere sister-chromatid exchange
WT	wildtype

XIC	extracted ion chromatogram
ZBTB10	Zinc finger and BTB domain-containing protein 10
Znf	Zinc finger

Table of Contents

Abstract	i
Zusammenfassung.....	ii
List of publications	iv
Additional publications.....	iv
Abbreviations.....	v
Introduction	1
Mass spectrometry-based proteomics.....	1
MS sample preparation	3
From the biological source to peptides	3
Peptide separation by chromatography	4
The mass spectrometer	5
Peptide ionization.....	5
Mass analyzer	6
Ion fragmentation.....	8
Data analysis.....	9
Quantification strategies	11
Stable isotope labeling	12
Label-free quantification	14
Applications of MS-based proteomics	15
Analysis of post-translational modifications	16
Interaction proteomics	16
Studying expression dynamics.....	18
Proteomics to study organismal development.....	18
Development of <i>Drosophila melanogaster</i>	18
Proteomics studies of <i>Drosophila</i> development	19
Telomere biology	21

End protection	22
End replication.....	23
Telomere maintenance by telomerase	25
Telomere maintenance by alternative lengthening of telomeres.....	27
Chromatin organization.....	28
Transcription of telomeres	29
The shelterin complex	30
Telomeres in disease and aging.....	31
Functions of shelterin proteins outside of telomeres	32
Zinc finger and BTB proteins	34
Aims of the thesis.....	36
Publications	37
The developmental proteome of <i>Drosophila melanogaster</i>	38
Summary.....	38
Zusammenfassung.....	39
Statement of Contribution	40
Phylointeractomics reconstructs functional evolution of protein binding	79
Summary.....	79
Zusammenfassung.....	79
Statement of Contribution	80
ZBTB10 binds the telomeric variant repeat TTGGGG and interacts with TRF2.....	97
Summary.....	97
Zusammenfassung.....	97
Statement of Contribution	98
Conclusions and Future Perspectives.....	122
MS-based proteomics to study development	122
MS-based proteomics to identify telomere-associated proteins	124
The TTGGGG variant repeat-binding protein ZBTB10	126

Acknowledgements 130

References 131

Introduction

Cellular processes are constantly active, well organized and controlled. Proteins are involved in nearly all of these processes as the active functional molecules. Thus, the proteome determines the phenotype of a cell at all times by describing all expressed proteins. Proteomes of most cellular systems are similar and only few proteins are unique per tissue demonstrating that protein abundance is of similar importance as presence or absence of any protein (Lundberg et al., 2010). Upon inner or outer perturbations, the proteome dynamically adapts through regulation, changes in protein expression and post-translational modifications (PTMs). While there are about 19,000-20,000 protein-coding genes in human (Ezkurdia et al., 2014), the actual number of possible proteins is much higher considering gene expression regulation. Alternative splicing, for example, increases protein diversity by generating alternative mRNA variants from a single gene that can code for different protein isoforms with distinct functions (Pan et al., 2008). Moreover, cellular localization, PTMs or protein-protein interactions and the formation of specific complexes further increase proteome complexity and influence the functionality of individual cells (Khoury et al., 2011).

Mass spectrometry-based proteomics

Mass spectrometry (MS) is an analytical method that measures the mass-to-charge (m/z) ratio of ionized analytes. MS-based proteomics has become a popular method to precisely analyze proteins from complex samples in an unbiased and systematic fashion (Aebersold and Mann, 2003). The availability of genome datasets for many species contributed to the successful application of MS-based proteomics (Nesvizhskii, 2014; Ruggles et al., 2017). It is a fast evolving field with constant improvements in instruments, quantification approaches and computational analysis (Aebersold and Mann, 2016; Leitner, 2018). Thus, quantitative proteomics is a powerful tool to identify organelle-, cell type- or tissue-specific proteins and to compare different conditions with high sensitivity and accuracy. While proteomics aims at identifying and analyzing all expressed proteins, there are two basic concepts: either measurement of intact proteins (top-down) or their digested peptides

(bottom-up) (Chait, 2006). In top-down proteomics, the entire protein with all its modifications is analyzed in high quality. This is advantageous for analyzing purified proteins, specific isoforms and PTMs because high sequence coverage can be achieved (Catherman et al., 2014). In contrast, bottom-up proteomics gains more quantification and therefore is usually used for analyzing complex protein mixtures. Proteolytic enzymes are used to digest proteins into specific peptides (Hildonen et al., 2014), which are reassembled to a protein after identification by computational analysis.

One distinguishes between data-dependent or data-independent acquisition (DDA or DIA) (Aebersold and Mann, 2016). While both modes allow an unbiased investigation free of any hypothesis, DDA is more commonly used in discovery proteomics. In contrast, the DIA mode is currently more often used for post-acquisition targeted analysis to sequence all peptides in a sample (Vidova and Spacil, 2017). Typically, a selected mass window of precursor ions is used for fragmentation, which results in complex spectra and requires a peptide fragmentation library for analysis (Bilbao et al., 2015; Gillet et al., 2012). A third acquisition mode is the single/multi reaction monitoring (SRM/MRM), which is a hypothesis-driven approach (Domon and Aebersold, 2010). Here, selected precursor ions are fragmented to monitor specific ions with high sensitivity.

In the following pages, a typical bottom up MS experiment using DDA is explained including sample preparation steps, the instrument itself as well as how data is analyzed. A first overview of the complete workflow can be found in Figure 1.

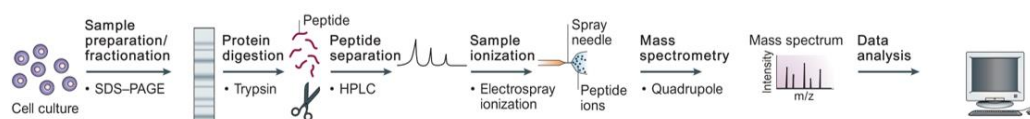


Figure 1: Standard MS workflow in bottom-up proteomics. Proteins are extracted from a biological source and can be fractionated by SDS-PAGE. Proteins are digested into peptides using sequence-specific enzymes such as trypsin. The complex peptide mixture is separated by high performance liquid chromatography (HPLC) and peptides are ionized using electrospray ionization (ESI). Ions are analyzed in the mass spectrometer, for example, in a quadrupole. A mass spectrum is generated and each ion with its m/z ratio (x-axis) has an intensity (y-axis). In tandem mass spectrometry, ions of a specific m/z are fragmented and detected in the mass analyzer (MS/MS). The peptide sequence is identified using a database search algorithm and corresponding proteins are inferred. Modified from (Steen and Mann, 2004).

MS sample preparation

During sample preparation it is important to maintain the state of the proteome, to avoid reagents that would interfere with the quantification method to be used or chemicals that are incompatible with MS and to minimize bias towards a specific class of proteins (e.g. membrane proteins).

From the biological source to peptides

A typical MS workflow of bottom-up proteomics starts with protein extraction from a biological source (Figure 2). Depending on the biological material (cell, tissue or whole organism) an appropriate lysis approach is used. The aim of lysis is the efficient extraction of proteins. Protein extraction can be performed using buffers with mild detergents under physiological conditions and in the presence of protease inhibitors to avoid proteolysis and preserve the native protein state. Thereby, non-denatured protein lysates can still be used for affinity enrichment experiments. In contrast, lysis conditions can also be harsh using buffers with high detergent concentrations or strong chaotropes such as guanidinium chloride or urea resulting in protein denaturation. Lysis is often accompanied by an additional mechanical disruption step, for instance, sonication or bead-milling. After the extraction of proteins, they can be fractionated by e.g. SDS-PAGE to gain deeper proteome coverage (Mostovenko et al., 2013). Subsequent steps will then be performed using chopped gel pieces (“in-gel”) (Shevchenko et al., 2006). Alternatively, when protein fractionation is spared, all following steps can be performed “in-solution”. In both approaches, cysteine residues of denatured proteins are reduced using dithiothreitol (DTT) and alkylated with 2-iodoacetamide (IAA) or 2-chloroacetamide (CAA) (Shevchenko et al., 2006). These modifications are critical to prevent the occurrence of disulfide bonds and oxidation of cysteines, which would result in various peptide forms complicating MS analysis. Afterwards, proteins are digested into peptides using proteolytic enzymes such as trypsin, Lys-C, Asp-N or Glu-C. The most common protease used in bottom-up proteomics is trypsin, which specifically cleaves at the carboxyl-termini of lysine (K) and arginine (R) residues (if not followed by a proline residue) (Olsen et al., 2004).

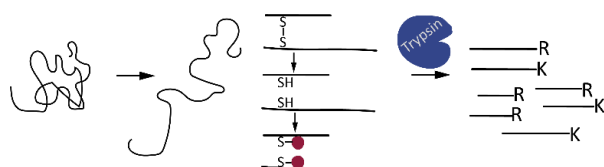


Figure 2: Protein extraction and processing. Proteins are extracted from a biological sample and denatured by heat or chaotropic reagents. Disulfide bonds are reduced, alkylated and proteins are digested into peptides using specific proteases, e.g. trypsin, which cleaves after arginine (R) and lysine (K) if not followed by a proline.

At low pH, each peptide has a minimum of two positive charges (N-terminus and side chain of the terminal residue). When performing in-gel digest, peptides need to be extracted from gel pieces, a step that is more time consuming with possibly lower recovery of peptides compared to in-solution digest. Prior to separation by liquid chromatography (LC), peptides are desalted with a StageTip using C18 material as solid phase (Rappsilber et al., 2007).

Peptide separation by chromatography

Efficient peptide separation is a critical part of bottom-up proteomics as it influences data quality of the MS measurement. Chromatography-based separation relies on the affinity of an analyte (peptides) to the stationary phase. Common chromatography principles are reverse-phase, ion exchange or isoelectric focusing, which are based on hydrophobic or Coulomb interactions. With the introduction of high performance LC (HPLC) in the 1970s, chromatography has significantly improved mainly because very high pressures can be handled (up to 1,000 bar) (Cristobal et al., 2012; Köcher et al., 2011; Thakur et al., 2011). In our setup, we use HPLC columns with a reverse-phase made of long hydrophobic alkyl chains (C18) as stationary phase. Under acidic conditions, positively charged peptides exhibit high affinity to the C18 material. To elute peptides according to their hydrophobicity, a gradient with increasing organic content (acetonitrile) is used. Depending on column length, inner diameter, column filling and gradient length, the resolving power can be influenced, which has an effect on how many peaks can be separated from each other in the gradient (Jorgenson, 2010; Köcher et al., 2011; Smith et al., 2004). Once HPLC is directly coupled to the mass spectrometer it is referred to be “online” (LC-MS).

The mass spectrometer

The main elements of a mass spectrometer are the ionization source and the mass analyzer. Peptides separated by HPLC are ionized, sprayed into the mass spectrometer and measured in the mass analyzer resulting in mass spectra with m/z values and a respective intensity for each analyzed ion. For all studies presented in this thesis, the Q Exactive Plus instrument was used and its components will be explained in more detail (Figure 3).

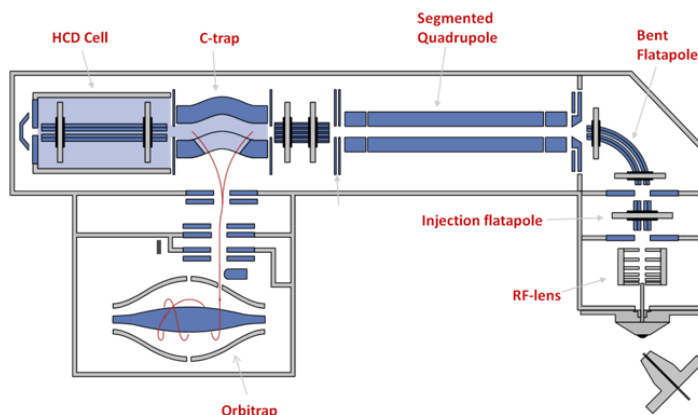


Figure 3: Construction details of the Q Exactive Plus mass spectrometer (Thermo Scientific). Gas-phase ions enter the radio frequency (RF)-lens and pass through flatapoles guided by an RF voltage and a direct current (DC) voltage. Neutral species or solvent droplets are removed by using the bent flatapole. Ions pass through the quadrupole, are stored and compressed in the C-trap and analyzed in the Orbitrap. A full scan MS spectrum is acquired. For MS/MS analysis, precursor ions of specific m/z are selected in the quadrupole with high resolution, fragmented in presence of an inert gas in the HCD cell and analyzed in the Orbitrap.
From: <http://planetorbitrap.com/q-exactive-plus>.

Peptide ionization

The development of soft ionization techniques has revolutionized the field of MS. The introduction of electrospray ionization (ESI) has been acknowledged a Nobel Prize in Chemistry in 2002 (Fenn et al., 1989). ESI-MS systems are preferred for complex sample analysis and ESI is mostly coupled to liquid-based separation methods (Aebersold and Mann, 2003). In our setup, a microscale capillary HPLC column is coupled online to the mass spectrometer. When peptides are eluted according to their hydrophobicity, they are ionized by ESI (Figure 4). By using a nanospray ionization source, flow rates are reduced to

the nanoliter range (200 nL/min). Ions are generated by passing the sample through a needle under high electric potential (2.4 kV), thereby generating charged droplets consisting of both solvent and analyte. When the solvent evaporates, the charge density increases and the droplet breaks down. Positively charged ions enter the vacuum system of the mass spectrometer and are guided by different RF and DC voltages to the mass analyzer.

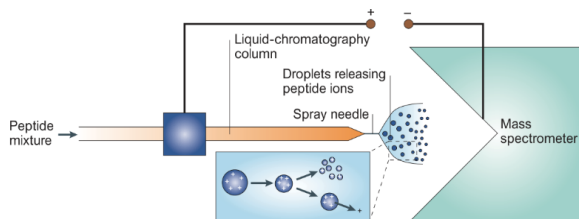


Figure 4: HPLC coupled to ESI-MS. Peptides are separated by HPLC and eluted according to their hydrophobicity. The tip of the HPLC column is under high electrical pressure resulting in highly positively charged droplets. Upon evaporation of the solvent, desolvated ions are formed and enter the mass spectrometer. From (Steen and Mann, 2004).

Mass analyzer

The mass analyzer is the key part of a mass spectrometer. The core principle of mass analyzers is based on ion movement, which is dependent on the mass-to-charge ratio and is influenced by magnetic and/or electric fields. Mass analyzers can be grouped into beam type (quadrupole and time of flight (TOF)) and trap type (ion trap, Fourier-transform ion cyclotron resonance (FT-ICR) and orbitrap) analyzers (Aebersold and Mann, 2003). The mass analyzer is highly important for data quality as it determines mass resolution, accuracy, sensitivity, dynamic range and scan speed (Scigelova and Makarov, 2009). The parameter mass resolution describes the ability to differentiate between two ion peaks with similar m/z values. A high mass resolution is usually classified as $>10,000$ (Steen and Mann, 2004; Xian et al., 2012). Discrepancy between measured and theoretical mass is defined by mass accuracy as the mass error in parts per million (ppm) (Olsen et al., 2010). In general, accuracy depends on the resolution and thus, orbitrap analyzers with very high mass resolution ($>100,000$) also achieve sub-ppm mass accuracy (Hu et al., 2005; Makarov and Denisov, 2009). Another important measure is the dynamic range, which determines the capability of the mass analyzer to detect low abundant ions. Thus, the actual limit of

detection is not only dependent on the detection principle but also on sample complexity because peptide abundances can differ by several orders of magnitude in complex biological samples. The parameter scan speed expresses the time needed for a certain m/z range. Beam type analyzers with lower resolution are the fastest.

Modern instruments combine elements of each group (trap and beam type analyzers) to achieve best results. The Q Exactive Plus instrument is a typical hybrid instrument coupling the quadrupole for mass selection with an orbitrap for mass analysis. Thereby, high speed and good sensitivity combined with very high accuracy, resolution and a high dynamic range are achieved (Michalski et al., 2011).

The quadrupole – a beam type mass analyzer

The quadrupole contains four metal rods that are electrically connected with opposing voltages. DC voltages and alternating voltages are applied to control ion movement (Schwartz and Senko, 2002). Ions oscillate within the electric field and only ions with the desired m/z maintain the trajectory and pass through the analyzer. At each m/z value, the intensity of an ion is recorded and a mass spectrum is generated. Quadrupoles are affordable mass analyzers with a good speed but low resolution. In the Q Exactive Plus instrument, the quadrupole is used as a mass filter to select precursor ions. The advanced quadrupole technology features a hyperbolic segmented quadrupole (see Figure 3), hereby optimizing dynamic range and precursor selection as well as ion transmission efficiency.

The orbitrap – a trap type mass analyzer

Trap analyzers are able to store and isolate ions. The orbitrap is the most recent developed analyzer, which requires a nitrogen-filled C-trap to inject a population of ions as a package upon high voltages (Makarov, 2000). Ions are then trapped in the orbitrap. They move in orbitals around the central spindle-like electrode and additionally, forth and back along the electrode axis resulting in a complex spiral (Makarov et al., 2006). Importantly, the frequency of ion oscillation depends on the m/z ratio and is independent of the initial energy. The fast Fourier transform algorithm is used to deconvolute the m/z information to

generate a mass spectrum (Zubarev and Makarov, 2013). Orbitraps are known for their high resolution, mass accuracy and large dynamic range (Hu et al., 2005; Makarov and Denisov, 2009; Olsen et al., 2010). To conduct tandem MS analysis (MS/MS), ions are passed through the C-trap into the HCD cell for fragmentation. Fragmented ions are then transferred back to be injected into the orbitrap analyzer for detection.

Ion fragmentation

First, all peptide ions are analyzed in a full scan mass spectrum (MS^1) in which m/z values are plotted against measured intensities (Figure 5A). To obtain the primary sequence information, a peptide ion of a specific m/z is selected in the quadrupole (precursor ion) and fragmented (product ions) in a separate cell to be analyzed in the orbitrap resulting in an MS/MS spectrum (MS^2) (Figure 5B) (Steen and Mann, 2004). In tandem MS, typically the ten most abundant peptide peaks are isolated for fragmentation (Top10 method). A common fragmentation technique is higher-energy collisional dissociation (HCD), which uses an inert gas such as helium, nitrogen or argon in a separate collision cell (Olsen et al., 2007). By HCD, peptide ions are fragmented at amide bonds. A series of b- and y-ions is generated if the charge is retained by the amino- and carboxyl-terminal part, respectively (Figure 5C) (Cañas Montalvo et al., 2006).

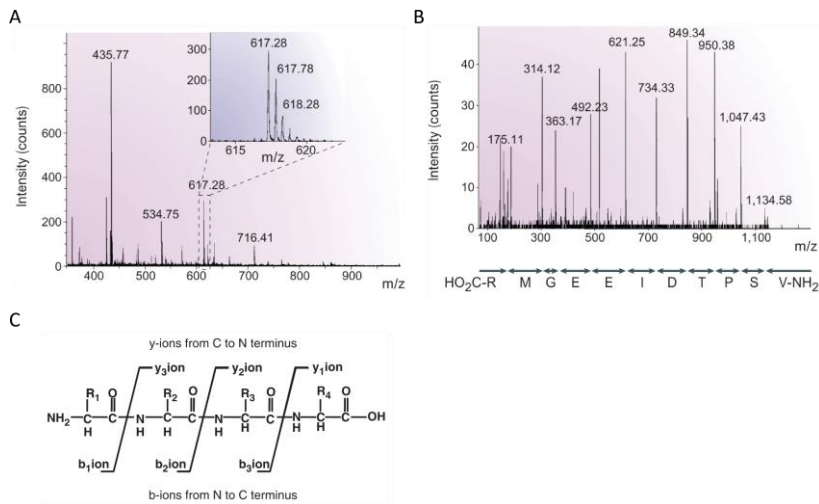


Figure 5: Representative mass spectra and nomenclature of peptide fragmentation. (A) Mass spectrum of all peptide ions with respective intensities at a specific elution time. m/z values are indicated at different peptide peaks. (B) MS/MS spectrum of the precursor ion (dashed box in A) that was fragmented. The primary sequence is indicated below the spectrum starting from the C-terminus (y-ion series). (C) Peptide fragmentation by HCD occurs at amide bonds and results in b-ions when the charge is retained by the N-terminal fragment or y-ions when the charge is retained by the C-terminal fragment. Ions are consecutively labeled. Adapted from (Cañas Montalvo et al., 2006; Steen and Mann, 2004).

Because a precursor ion can be fragmented at different amide bonds, the MS² spectrum is a collection of peptides that differ by at least one amino acid. In beam type instruments, HCD often results in further fragmentation of these product ions, preferably of the less stable b-ions resulting in a y-ion dominated spectrum (Olsen et al., 2007). Combination of liquid chromatography and tandem mass spectrometry (LC-MS/MS) is the most popular format in MS-based proteomics because it enables the analysis of complex samples within a limited amount of time.

Data analysis

The MaxQuant software can be used for data analysis (Cox and Mann, 2008). It was introduced in 2008 and has continuously improved. To determine the peptide sequence of each fragmentation spectrum, an implemented Andromeda search engine is used for automated identification of peptides (Cox et al., 2011). It is desirable to know the protein sequences so that all possible peptides are known. Hence, it is important to decide for an

appropriate database. Using MaxQuant search algorithms, each MS² spectrum is searched against the database and in addition, a second peptide search of each MS² spectrum is performed to also identify co-fragmentation signals of additional precursor ions. MaxQuant uses the target-decoy search strategy to provide statistical evidence for the quality of a peptide spectrum match and to limit the number of incorrect identifications (Elias and Gygi, 2007). For instance, the Andromeda score or probability score is a measure of how well the acquired spectrum matches the theoretical fragmentation spectrum. The posterior error probability (PEP) defines the probability of a peptide to be a false hit by taking the Andromeda score, peptide length, charge and the number of miscleavages into account. To control the false discovery rate (FDR), reverse peptide sequences of the database are used and searched against the tandem mass spectra. As soon as 1% of the sequences are matched to the reverse database, the search is stopped and peptides are accepted with a 1% FDR. After identification of peptide sequences, they are reassembled to protein identifications and peptide scores are added up resulting in protein scores. As an output, tables of all identified peptides and protein groups (a protein group contains all proteins that cannot be distinguished from their identified peptides) are generated and can be used for downstream analysis. In general, proteins identified based on a single peptide hit should be treated with caution and the respective spectra should be manually inspected. For downstream data analysis, short peptides with less than seven amino acids and all proteins identified by the reverse database are filtered.

MaxQuant uses a multilevel scoring system to optimize peptide identification and to have a low percentage of false positive hits. Importantly, the absence of a protein in a large-scale experiment does not necessarily mean the protein is indeed not present. Possibly its peptides co-eluted with more abundant peptides of other proteins, hence these peptides were not selected for sequencing by the instrument. Other important aspects to consider are protein size, accessibility to the used protease, peptide solubility, as well as ionization efficiency.

Quantification strategies

Peptide LC-MS/MS experiments need to be quantitative in order to analyze complex proteomes and to detect changes between different conditions. Quantification can be either absolute or relative.

Using absolute quantification, protein copy numbers can be determined by spiking-in isotopically labeled standard peptides. The absolute quantification (AQUA) approach determines the absolute amount of a specific protein in a sample by introducing stable isotope-labeled internal peptides with a known concentration during proteolytic protein digestion (Gerber et al., 2003; Kirkpatrick et al., 2005). Similarly, fully-labeled proteins can be added in a known concentration and used for label-based absolute quantification (Singh et al., 2009). Moreover, the QconCAT technology relies on stable isotope-labeled Qpeptides that are generated by recombinant expression of synthetic genes. Typically, each gene encodes an ORF that is a concatenation of tryptic peptides and additionally carries a quantification standard as well as a purification tag. These artificial proteins of about 50-70 kD are purified via a His-tag and mixed with the sample prior to the digestion step. All Qpeptides can be used as internal standards allowing absolute quantification of proteins (Brownridge et al., 2011). The Protein Epitope Signature Tag (PrEST) consists of a short unique sequence of the protein of interest, which is fused to solubility and quantification tags. Because of the Human Protein Atlas Project (www.proteinatlas.org), in which antibodies have been generated against parts of all human proteins, different PrESTs are available. PrESTs can be SILAC-labeled and spiked in cell lysates. Thus, SILAC ratios of the PrEST and the endogenous peptide are compared and absolute quantification is possible (Zeiler et al., 2012). An alternative label-free approach that can estimate absolute amounts of protein is: intensity-based absolute quantification (iBAQ). iBAQ adds up all peptide intensities matching a protein and divides it by the number of theoretical peptides in order to describe protein abundance (Schwannhäuser et al., 2011).

Relative quantification describes the relative fold change of a protein between different samples. It can be obtained by stable isotope labeling as well as in a label-free manner (Figure 6).

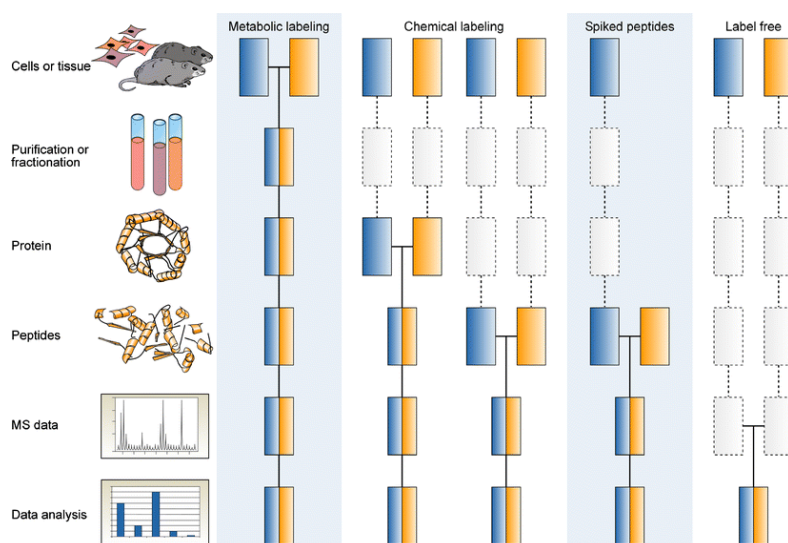


Figure 6: Comparison of different approaches for quantitative proteomics. In metabolic labeling, samples of two experimental conditions (blue and yellow boxes) are combined in the beginning of the workflow. Thereby, accurate quantification is achieved as possible experimental changes or errors equally affect both samples. Chemical labeling can be carried out on peptide or protein level. Samples are processed for a while in parallel possibly allowing experimental variation (dashed lines), which might lead to quantification errors. Isotope labels can be introduced by spike-in of synthetic peptides serving as an external standard. In label-free quantification, samples are processed in parallel. Therefore, the complete workflow needs to be reproducible to obtain reliable quantification results. From (Bantscheff et al., 2007).

Stable isotope labeling

Commonly used heavy isotopes for labeling are hydrogen (^2H), carbon (^{13}C), nitrogen (^{15}N) and oxygen (^{18}O), which can be introduced either chemically or metabolically. These isotopes do not significantly change chemical properties and can be differentiated from the unlabeled variant in the mass spectrometer (Bantscheff et al., 2012; Steen and Mann, 2004).

Chemical labeling is inexpensive and fast. It is introduced either on protein or on peptide level allowing for variability and systematic errors possibly resulting in less robustness. Reactive side chains of amino acids or protein termini are typically used for labeling. One popular technique is dimethyl labeling, which was introduced in 2003 (Hsu et al., 2003). All primary amines (N-termini and ϵ -amino group of Lys) are converted to dimethylamines through reductive amination by isotopically labeled formaldehyde (and

NaBH₃CN). Isotope-coded affinity tag (ICAT) labeling is restricted to peptides that contain cysteines. They are covalently modified at protein level and a biotin within the ICAT allows affinity enrichment (Gygi et al., 1999). Isobaric labeling strategies such as tandem mass tag (TMT) (Thompson et al., 2003) and isobaric tag for relative and absolute quantification (iTRAQ) (Ross et al., 2004) quantify different peptides on the MS² level. Both labeling techniques introduce isobaric tags via an amine-specific reactive group. In the MS¹, peptides have an identical mass but upon fragmentation, distinct reporter ion masses can be detected in the MS². Multiplexing (up to 11-plex) and application for any sample type is possible. Possible drawbacks are co-eluting peptides affecting quantification accuracy and undesired side reactions complicating data analysis (Bantscheff et al., 2007).

An alternative strategy is metabolic labeling. It describes the *in vivo* integration of non-radioactive isotopes into newly synthesized proteins of an organism resulting in a defined mass shift. To detect the introduced mass difference by MS, the label should ideally differ at least by four Da and overlapping isotopic distributions should be avoided. The greatest advantage of metabolic labeling is the early combination of two experimental conditions (Figure 6). Hence, variations in the workflow equally affect all peptides. The easiest approach is ¹⁵N labeling via nutrients, hereby labeling whole organisms. This has been done in yeast, *Caenorhabditis elegans*, *Drosophila melanogaster* and cultured cells (Conrads et al., 2001; Krijgsvelde et al., 2003). However, only two conditions can be compared and the incorporation of heavy nitrogen atoms varies between proteolytic peptides complicating data analysis. The introduction of stable isotope labeling by amino acids in cell culture (SILAC) addressed these challenges (Ong et al., 2003). SILAC is an efficient, robust and easy way to label newly synthesized proteins resulting in a specific mass shift per peptide pair. Despite the introduction of a mass difference, physical and chemical properties of the labeled peptide do not change. SILAC is considered the most accurate way to obtain quantitative information (Ong and Mann, 2006). In a typical SILAC experiment, two conditions are compared and cells are grown in “light” or “heavy” medium containing the normal or isotopically labeled analogue (¹³C and/or ¹⁵N) of lysine and arginine. It is also feasible to compare three biological samples by using an additional SILAC

state. If peptides are digested by trypsin at least one amino acid per peptide is labeled. Hence, a defined mass shift between labeled and unlabeled peptide is achieved. In the precursor spectrum, light and heavy peptides generate characteristic mass pairs. Direct comparison of signal intensities of isotope clusters determines the relative amount and results in accurate protein quantification. Although SILAC was initially introduced for mammalian cell culture, it has become a great approach for quantitative MS in model organisms such as yeast (de Godoy et al., 2006), *D. melanogaster* (Sury et al., 2010), *C. elegans* (Fredens et al., 2011), zebrafish (Westman-Brinkmalm et al., 2011) and mouse (Krüger et al., 2011). The introduction of super SILAC enabled SILAC quantification in biological systems that cannot be metabolically labeled, for instance, human tissue or certain cell lines. Super SILAC uses a spike-in standard of SILAC labeled cells (Geiger et al., 2010). Furthermore, SILAC can also be used to study the rate of protein synthesis and degradation of the protein of interest (Pratt et al., 2002), protein turnover by pulsed SILAC and translation rates (Doherty and Beynon, 2006; Mcshane et al., 2016; Schwanhäusser et al., 2009), protein-protein interactions (Hubner and Mann, 2010) and the methylation status of proteins (Ong et al., 2004).

Label-free quantification

Label-free quantification (LFQ) includes all strategies of quantifying proteins without introducing stable isotopes (Al Shweiki et al., 2017). The biggest advantages are that LFQ can be applied to any kind of sample, is independent of sample number, inexpensive and technically straightforward (Bantscheff et al., 2007). LFQ constitutes the simplest proteomics approach.

The plainest form of LFQ relies on spectral counting, which compares the number of identified MS² spectra matching one protein to estimate protein abundance (Lasonder et al., 2002). Concomitantly, detection of low abundant proteins is unreliable as only few sequenced peptides are accurately quantified. The extracted ion chromatogram (XIC) estimates protein abundance by comparing MS¹ intensities of peptides, which are directly proportional to peptide concentration (Bilbao et al., 2015). Peptide counting is another

early quantification method that relies on the number of identified peptides as an approximation for protein abundance. This strategy was modified by defining a protein abundance index (PAI), which divides the number of measured peptides by the number of theoretically obtained peptides to account for differences between proteins (Ishihama et al., 2005; Rappsilber et al., 2002). A machine learning-based method is the absolute protein expression profiling (APEX) approach that corrects for the probability to observe a peptide (Lu et al., 2007). Four years ago, an advanced label-free algorithm has been introduced in the MaxQuant software that improved quantification accuracy: MaxLFQ (Cox and Mann, 2008; Cox et al., 2014).

All MS experiments presented in this thesis used MaxLFQ, which will be explained in more detail. Retention times of all runs (typically eight runs with four replicates per condition) are aligned to correct for small differences in chromatography and to compare detected peptide features. Activation of the match between runs option in MaxQuant allows peptide identification between runs. In case a peptide is present in one run (MS^1) but not selected for fragmentation (MS^2), it can still be identified if it is sequenced in another run. Hereby, the maximum amount of information is extracted on peptide level. To control for experimental variability between samples, peptide intensities are normalized and protein intensities are calculated by using multiple pairwise peptide ratios. As a result, LFQ intensities are generated for each identified protein representing an approximation for protein abundance in each sample. Comparison of large sample numbers as well as quantification of a three-fold change became possible with MaxLFQ and high quality datasets have been generated.

Applications of MS-based proteomics

MS-based proteomics enables a broad range of studies. It can be used for studying a proteome under specific conditions (expression proteomics), proteins that bind to a specific bait such as protein, DNA, RNA or a peptide (interaction proteomics) and the occurrence of PTMs, which are important for many cellular functions. A general challenge of proteomics

is the dynamic range, which complicates detection of low abundant proteins or transient protein interactions.

Analysis of post-translational modifications

The importance of PTMs is illustrated by a two to three orders of magnitude more complex proteome compared to the predicted encoding genome (Walsh et al., 2005). Studying PTMs in a global and unbiased manner is of high importance to understand the complex signaling networks for cellular control. Most commonly studied PTMs are phosphorylation, acetylation and methylation, which all lead to changes in mass shifts and can be matched to a single amino acid (Aebersold and Mann, 2016; Doll and Burlingame, 2015). Distinct strategies have been developed to investigate ubiquitinylation and glycosylation (Kim et al., 2011; Zielinska et al., 2010).

Although mass spectrometry has contributed to the identification of more than 50,000 phosphorylated peptides (Sharma et al., 2014) and over 20,000 ubiquitinylation sites (Wagner et al., 2012), investigating PTMs remains challenging. General problems are more complex fragmentation spectra of modified peptides and less abundance in comparison to the unmodified counterpart (Choudhary and Mann, 2010). To identify PTMs with high confidence, an enrichment step for peptides carrying the modification of interest is usually added to the standard sample preparation workflow by using specific antibodies, ionic interactions or affinity ligands (Zhao et al., 2009; Olsen and Mann, 2013). For example, phosphopeptides are specifically enriched using TiO_2 strong cation exchange or Ti^{4+} -immobilized metal ion affinity chromatography (Zhou et al., 2011). A diglycine-specific antibody allows capturing the remnant modification of ubiquitinylated peptides after trypsin digestion (Udeshi et al., 2013).

Interaction proteomics

The investigation of protein interactions with DNA, RNA or proteins is essential to gain a better understanding of cellular processes and individual protein function. By using a specific bait for affinity purification (AP), MS can be used to study protein-protein, DNA-

protein, RNA-protein or peptide-protein (e.g. a peptide carrying specific histone modifications or PTMs) interactions (Dunham et al., 2012; El Kennani et al., 2018; Makowski et al., 2018). To this end, the bait of interest is immobilized on a matrix to fish interacting proteins present in the cell lysate. In most cases, an antibody against the protein of interest or against the tag, to which the protein of interest is fused to, is used for immobilization (Rothbauer et al., 2008; Smits and Vermeulen, 2016). Nucleic acids or peptides can be biotinylated *in vitro* allowing immobilization on streptavidin-coated beads via the strong streptavidin-biotin bond. Various washing steps remove unspecific binders and bound proteins are analyzed in comparison to a control by quantitative MS. In all AP studies it is crucial to carry out respective control pulldowns by using an IgG antibody, a scrambled nucleic acids sequence or an unmodified peptide sequence as bait. While background proteins are found in both pulldowns, specific interactors are determined by their significant enrichment in target pulldowns (Keilhauer et al., 2015; Vermeulen et al., 2008). In case of protein-protein interaction studies, it is generally best to perform pulldowns under endogenous conditions to ensure correct expression and localization of the bait protein. In 2011, nearly 1,800 primary antibodies were used in a large-scale study to investigate interactions of proteins mostly involved in transcription and signaling (Malovannaya et al., 2011). A lot of other studies used GFP-, TAP- or FLAG-tagged protein variants. In human cells, the use of bacterial artificial chromosomes (BAC) coupled to quantitative interactomics enables the investigation of protein-protein interactions in a nearly endogenous situation. In addition to a GFP tag, the BAC vector contains the complete human gene of interest including all regulatory elements ensuring expression levels very close to the endogenous level (Poser et al., 2008). In a large-scale study with more than 1,000 HeLa BAC cell lines, Hein et al. (2015) revealed that the protein network is dominated by weak interactions. Other studies used FLAG-HA-tagged proteins as bait for AP-MS in HEK293T cells and reported more than 1,000 tightly connected communities of proteins that correspond to larger complexes or proteins involved in the same pathway (Huttlin et al., 2015, 2017). However, the human interactome is still incomplete and remains challenging due to large differences in protein abundance as well as dynamic interactions.

Studying expression dynamics

Proteomics enables a global analysis of protein composition in a specific cell line, tissue or across development of an organism, as well as comparisons between a treated and an untreated situation. Alternative splicing events, polymorphisms and PTMs further increase proteome complexity. They result in different protein variants with possibly distinct roles in cellular processes. Thus, the need for efficient and accurate identification is high. Although MS is a powerful method, it is still challenging to quantify low abundant proteins expressed under certain conditions or in distinct cell types. Detection of protein isoforms is challenging due to their high percentage of sequence homology. In addition, incomplete sequence coverage is limiting as it is difficult to distinguish isoforms if not the whole protein sequence is covered.

Proteomics to study organismal development

*Development of *Drosophila melanogaster**

The fruit fly *Drosophila melanogaster* is a widely used model organism to study genetics and behavior in developmental biology. *Drosophila* undergoes complete metamorphosis by progressing through four distinct developmental stages: embryo, larva, pupa and the adult fly. During embryogenesis, the fertilized egg develops into a complete organism and already within the first two hours, the diploid cell undergoes multiple nuclear divisions forming the blastoderm comprising thousands of cells. A key event during embryogenesis is the maternal-to-zygotic transition (MZT), in which maternally provided mRNAs are cleared and new zygotic transcripts are synthesized at high levels (Tadros and Lipshitz, 2009). Other key events defining early development are axis specification, gastrulation, germ layer formation and organogenesis before the egg hatches after 22-24 hours at 25°C (Tyler et al., 2003). The first instar larva emerges, feeds and molts into the next larval stage. In total, there are three larval stages each lasting for about one day, which are mainly characterized by an increase in body size. The L3 larva ages into a crawling larva. When climbing out of its food, the larva starts pupation in a dry surrounding. While the pupa progressively darkens, it develops into

the adult fly within four days. Epithelial cells differentiate into wings, eyes, legs, mouthparts or genital ducts, whereas histoblasts form internal organs and the abdominal epidermis. Controlled by the circadian rhythm, the fly emerges in a process called eclosion. Within the first 4-8 hours, flies are virgins before they become sexually active. When mating, the female fly stores the sperm in specific body compartments that allows fertilization to take place internally. A peak of egg production with about 80 eggs per day is reached in four to seven day old female flies. On average, flies live for around eight weeks (Tyler et al., 2003). All metamorphological changes require a tight regulation by transcriptional, translational and epigenetic control mechanisms.

Proteomics studies of Drosophila development

The development of *Drosophila* has been widely studied on RNA level. Various genetic screens (St Johnston, 2002) and RNAi screens (Dietzl et al., 2007) focused on the identification of genes that are involved in a specific biological process and revealed certain phenotypes. While there are about 19,800 genes in *D. melanogaster* (Adams et al., 2000), around 7,000 proteins listed in FlyBase start with the identifier “CG” (computed gene) probably presenting uncharacterized proteins. Until now, proteomics studies mainly focused on specific developmental stages or on body compartments using different quantification methods. In 2003, *D. melanogaster* together with *C. elegans* were the first model organisms analyzed after *in vivo* ^{15}N metabolic labeling (Krijgsveld et al., 2003). In a two-step approach, yeast was first ^{15}N -labeled and used as food source throughout *Drosophila*'s development. As a result, second generation embryos were successfully labeled and analyzed by quantitative MS. However, the frequent occurrence of N atoms is a disadvantage of ^{15}N labeling as it leads to partial labeling of peptide sequences resulting in complex isotope clusters and peptide-specific mass shifts, which complicate data analysis. Two other proteomics studies used the same labeling approach. On the one hand, the MZT transition was investigated and 500 proteins were specifically upregulated. On the other hand, a study about seminal fluid proteins revealed their transfer from the male to the female fly during mating (Findlay et al., 2008; Gouw et al., 2009). By feeding larvae and

F1 flies with SILAC-labeled yeast, a systematic study for female and male-specific proteins was conducted and around 1,900 proteins were quantified with high confidence (Sury et al., 2010). Using label-free quantification as well as iTRAQ and combining separately analyzed proteins extracted from the fly's head and the body, Xing and co-workers quantified around 5,600 proteins in two biological replicates of five day old adult male flies (Xing et al., 2014). Another proteomics study quantified around 1,700 proteins across the lifetime of adult flies at nine time points from eclosion to day 60 and reported a downregulation of proteins related to metabolism, reproduction and the defense response with age (Sowell et al., 2007). Additional proteomics studies in the adult focused on the male reproductive system (Takemori and Yamamoto, 2009) and composition of the sperm proteome (Dorus et al., 2006; Karr, 2007; Wasbrough et al., 2010). To identify proteins regulating oogenesis, egg activation and early embryogenesis, dimethyl labeling was used to study the oocyte-to-embryo transition in *Drosophila* (Kronja et al., 2014a, 2014b). Comparisons during oocyte maturation and egg activation revealed substantial proteome remodeling with about 500 proteins being each up- and downregulated accompanied with the transition from meiosis to mitosis (proteins involved in cell cycle, spindle formation, DNA replication or chromosome architecture). Embryogenesis has also been studied by MS using two different approaches: the common DDA mode and a recently developed label-free algorithm based on the DIA mode without normalization (SWATH). Precursor ions are selected by a specific m/z window size. Here, it is necessary to first generate a peptide library to know about peptide spectra and retention times. Both methods resulted in similar protein profiles of about 400 proteins that were measured at five time points during 0-22.5 hours of embryogenesis (Fabre et al., 2016). The largest proteomics study so far catalogued 9,124 proteins that correspond to around 63% of *Drosophila* proteins and 8,672 genes. In this study, two different *Drosophila* cell lines (embryonic S2 and larval Kc167) and samples from all developmental stages were combined for analysis to increase sample diversity. Moreover, the authors used different fractionation techniques to reduce sample complexity. An analysis-driven feedback strategy was used to optimize experimental conditions focusing on under- or overrepresented proteins (Brunner et al., 2007). More

defined MS studies have been conducted in *Drosophila* cell lines to determine complex components in a phosphorylation dependent manner using fourplex iTRAQ isobaric labeling (Pflieger et al., 2008), to identify over 10,000 phosphorylation sites from nearly 4,600 phosphoproteins (Bodenmiller et al., 2007) and to monitor expression kinetics after viral infections using SILAC-MS (Go et al., 2008). Although all of these studies cover various stages during the fly's lifecycle, a direct comparison is not possible due to technical differences such as the used quantification method and different instrument types. Thus, a comprehensive analysis throughout development is desirable.

Telomere biology

Telomeres are nucleoprotein structures at the end of linear chromosomes that protect the ends from DNA repair activities, fusion and degradation. While telomeres are conserved across species, their sequence and length differs between species and individuals. In vertebrates, telomeres consist of double-stranded (ds) TTAGGG repeats ending in a 50-300 nucleotide long 3' single-stranded (ss) DNA overhang (Wu et al., 2012) and telomere-associated proteins (Figure 7) (Palm and de Lange, 2008). The shelterin complex is a telomere-specific complex consisting of six proteins. While TRF1 and TRF2 directly bind the dsDNA, POT1 binds the ssDNA 3' overhang. Additional interactors are TPP1, RAP1 and the bridging protein TIN2 (de Lange, 2005). The shelterin complex is crucial for telomere maintenance and end protection, thereby preserving genome integrity.

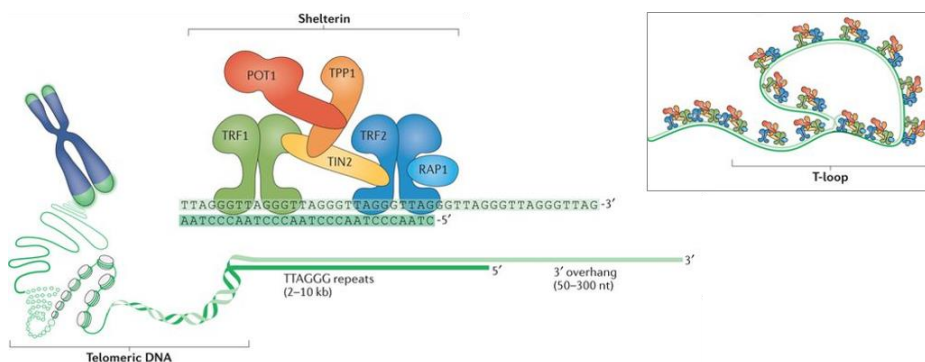


Figure 7: Structure of human telomeres. The ends of linear chromosomes consist of long TTAGGG repeat arrays bound by the six-subunit shelterin complex (TRF1, TRF2, RAP1, POT1, TPP1 and TIN2). The 3' overhang invades the duplex repeat sequence forming a t-loop (right window). Modified from (Maciejowski and de Lange, 2017).

End protection

Due to the structural similarity of chromosomal ends and DNA double-strand breaks (DSB), telomeres are indispensable capping structures to prevent misidentification by the DNA repair machinery. About 20 years ago, electron microscopy studies revealed the ability of human telomeres to fold back forming the so-called t-loop (Figure 7) (Griffith et al., 1999). The ssDNA 3' overhang invades the duplex DNA repeat array resulting in a t-loop that generates a small displacement loop. The t-loop structure hereby contributes to the protection of telomere integrity and safeguards chromosomal ends from nucleolytic attacks, DNA repair activities and checkpoint activation (De Lange, 2010; O'Sullivan and Karlseder, 2010; van Steensel et al., 1998). Among the six shelterin components, TRF2 is essential for t-loop maintenance (Doksani et al., 2013). The ability of the TRFH domain of TRF2 to wrap ca. 90 bp DNA around itself is thought to induce t-loop formation (Benarroch-Popivker et al., 2016). Upon TRF2 depletion, chromosome ends are fused resulting in long ligated or circular chromosomes (van Steensel et al., 1998). In addition to the role of TRF2 in regulating t-loop formation, it also suppresses the DNA damage response (DDR) machinery at telomeres. Therefore, TRF2 has been attributed a major role in telomere end protection (Feuerhahn et al., 2015).

DSBs can be repaired either by non-homologous end joining (NHEJ), an error-prone mechanism that directly joins two ends (Lieber, 2010) or by homologous recombination (HR). HR requires DSB end resection by nucleases and helicases, consequently allowing RAD51 recombinase to assemble to the ssDNA overhang for repair synthesis (San Filippo et al., 2008). NHEJ is active throughout the cell cycle, but predominates in G1 cells, whereas HR is favored in S/G2 phases after DNA replication (Feng et al., 2015). Both mechanisms are inhibited at telomeres to avoid disastrous end-to-end fusions by NHEJ or telomeric recombination events by HR. While TRF2 inhibits NHEJ (Celli and de Lange, 2005), the TRF2-RAP1 heterodimer represses homology directed repair (HDR) by preventing telomeric localization of PARP1 and SLX4 (Rai et al., 2016). Moreover, TRF2 prevents MRN-dependent ATM kinase activation at telomeres (Stracker and Petrini, 2011) and blocks propagation of damage signaling via its iDDR domain (Okamoto et al., 2013). If the DNA-binding ability of

TRF2 is impaired, ATM activity results in a p53-dependent cell cycle arrest or apoptosis (Karlseder et al., 1999). POT1 inhibits ATR kinase activity at telomeres by preventing RPA binding to the ssDNA overhang (Denchi and de Lange, 2007). Although the DDR is suppressed at telomeres, numerous checkpoint proteins, repair factors and DNA damage sensors associate as they play major roles in telomere maintenance (D'Adda Di Fagagna et al., 2004).

End replication

DNA replication is a semi-conservative mechanism that ensures proper duplication of genetic material, preserves epigenetic information and cell identity. Thus, efficient maintenance mechanisms are required. Accurate DNA replication relies on the temporal separation of loading the MCM2-7 helicase complex onto chromatin in late M/G1 phases and activating it for DNA synthesis in S phase. Formation of the pre-replication complex (pre-RC) is dependent on ORC, CDC6 and the replication licensing factor CDT1. TRF2 directly interacts with ORC1. Thereby, ORC and the pre-RC are recruited to telomeric chromatin suggesting an implication of TRF2 in activating telomeric origins (Higa et al., 2017; Tatsumi et al., 2008). In S phase, additional proteins remodel the MCM helicase complex into an active state to initiate DNA synthesis. While DNA polymerase α and the primase complex initiate DNA synthesis, DNA Pol δ and Pol ϵ proceed with synthesis of the lagging and leading strand, respectively, throughout S phase in human (Sfeir and de Lange, 2012; Wright et al., 1999).

Because of the semi-conservative replication mode, telomeres shorten during each round of cell division, a phenomenon known as the end replication problem (Gilson and Géli, 2007). In budding yeast it was demonstrated that the end replication problem arises from the leading strand (Soudet et al., 2014). Leading strand replication results in the formation of a blunt end. End resection generates the ssDNA G-rich overhang and subsequently, the C-strand is filled-in. The last Okazaki fragment at 5-10 nt from the end of the leading strand hereby results in the observed telomere shortening phenomenon each

replication round. G-strand overhang generation is tightly regulated and requires a set of proteins (Figure 8) (Wu et al., 2012).

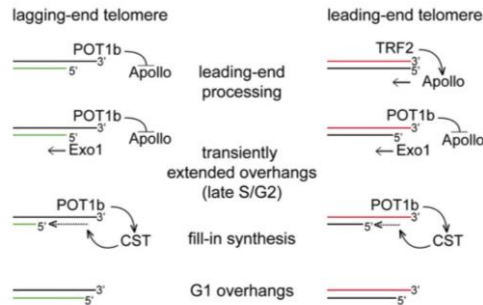


Figure 8: Model for the generation of the 3' ssDNA overhang in mouse cells. TRF2 recruits Apollo nuclease (also known as SMN1B) at leading end telomeres to generate an overhang. POT1b binds the ssDNA 3' overhang and inhibits resection by Apollo. On both ends, end resection is fulfilled by Exo1 endonuclease and transiently generated overhangs are filled-in by the CST complex. From (Wu et al., 2012).

Telomeric replication challenges the replication machinery due to the occurrence of secondary structures that need to be resolved to preserve genome integrity (Figure 9) (Higa et al., 2017). For instance, G-quadruplexes are highly stable structures that are preferably formed at the 3' overhang (Tang et al., 2008). They can be removed by helicases BLM, WRN or by RTEL1 interacting with PCNA, whereas t-loops are resolved by RTEL1 interacting with TRF2 during S phase (Crabbe et al., 2004; Law et al., 2010; Vannier et al., 2012; Sarek et al., 2015). While WRN localizes to telomeres in late S phase to facilitate lagging strand replication, BLM localizes to telomeres in G2/M to process both telomere strands (Martínez and Blasco, 2015).

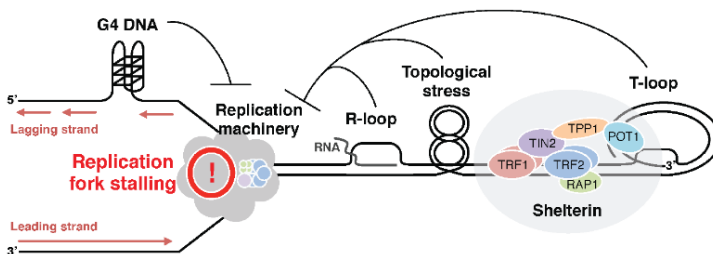


Figure 9: Secondary structures present obstacles during replication of telomeres. Telomeric DNA is duplicated by lagging and leading strand synthesis. Causes for replication fork stalling can be secondary structures such as G-quadruplexes (G4), R-loops (DNA-RNA hybrids) or topological stress. From (Higa et al., 2017).

The shelterin protein TRF1 is necessary for efficient replication and prevention of replication fork stalling (Sfeir et al., 2009). Other important factors are Timeless and Topoisomerase II α , which interact with TRF1 and are essential for stabilizing replication forks and resolving DNA replication intermediates, respectively. A combined action of TRF2, Apollo and Topoisomerase II α ensures the relieve of occurring topological stress during replication (Ye et al., 2010). Thereby, telomere fragility is repressed (D'Alcontres et al., 2014; Leman et al., 2010, 2012).

Telomere maintenance by telomerase

About 30 years ago the first terminal transferase activity was detected in the ciliated protozoan *Tetrahymena thermophila* (Greider and Blackburn, 1985). Subsequently, the reverse transcriptase telomerase (TERT) was found in vertebrates. Telomerase uses its own RNA moiety (TR) to add telomeric repeats to the 3' end of telomeres, hereby overcoming the end replication problem (Figure 10A) (Feng et al., 1995; Jiang et al., 2018; Lingner, 1997). While TR is widely expressed, TERT expression is downregulated in somatic cells, which results in progressive shortening of telomeres. In stem and germ cells as well as in 85-90% of cancer cells expression of TERT counteracts telomere shortening (Kim et al., 1994). Due to the low abundance of telomerase (ca. 240 active telomerase monomers in HEK293 cells (Xi and Cech, 2014)), it was proposed that telomerase only elongates a certain amount of telomeres per round of replication. Indeed, telomerase was preferentially recruited to short telomeres for their elongation in budding yeast (Bianchi and Shore, 2007; Marcand et al., 2010; Teixeira et al., 2004). Critically short telomeres were elongated first in murine embryonic stem cells heterozygous for *mTert* (Liu et al., 2002).

The TR component of telomerase consists of three major structural and functional domains: 1) a pseudoknot containing the complementary template sequence for annealing to telomeric DNA, 2) a stem terminus element required for TERT binding and activity and 3) the biogenesis domain containing an H/ACA motif, which interacts with several proteins (e.g. dyskerin) constituting the core ribonucleoprotein (RNP) complex of about 650-670 kD

(Robinson and Reddel, 2007). Physiological active telomerase requires additional proteins for recruitment, activity and C-strand synthesis (Figure 10B) (Chan et al., 2017).

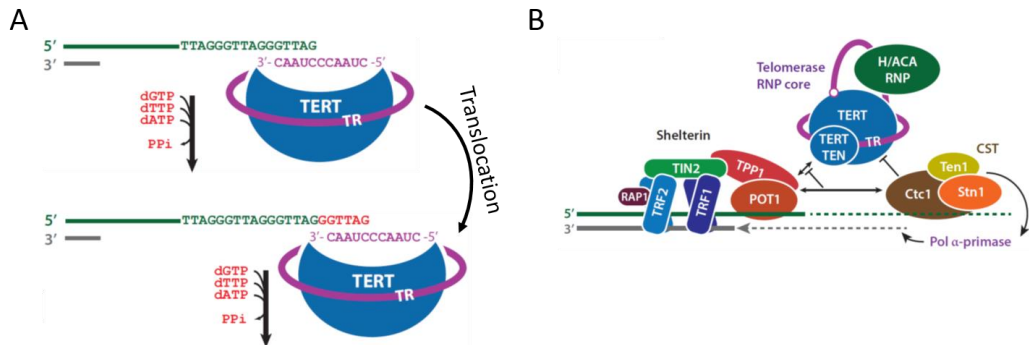


Figure 10: Elongation of telomeres by telomerase. (A) The TR component of telomerase contains the complementary RNA template (purple) to synthesize a telomeric repeat at the 3' end of human telomeres. After translocation, the next repeat is added. (B) The shelterin complex recruits the RNP core complex (TERT, TR, H/ACA RNP) to telomeres. Upon interaction of the TPP1-POT1 heterodimer with the CST complex, telomerase activity is inhibited. CST further recruits the DNA polymerase α -primase complex for C-strand synthesis. Modified from (Robinson and Reddel, 2007).

Telomerase levels are controlled by transcriptional regulation of *TERT* through epigenetic modifications or by DNA-binding proteins. A direct interaction of TERT and the shelterin protein TPP1 results in telomerase recruitment to telomeres and its association with the ssDNA overhang is supported by a bridging function of TIN2 and POT1 (Nandakumar et al., 2012; Xin et al., 2007; Zhong et al., 2012). The TPP1-POT1 heterodimer activates telomerase and increases its processivity (Nandakumar and Cech, 2013; Wang et al., 2007). However, when TPP1-POT1 interacts with the CST complex (CTC1-STN1-TEN1), association of telomerase and shelterin is disrupted and its telomere elongation activity is inhibited. In addition, the DNA polymerase α -primase complex is recruited for C-strand synthesis in late S/G2 phase (Figure 10B) (Chen et al., 2012). In human HCT116 cells, CTC1-STN1 terminates DNA synthesis by telomerase, whereas TEN1 is required for stabilizing CTC1-STN1 binding to ssDNA and enabling C-strand fill-in synthesis (Feng et al., 2018). Phosphorylation of TRF1 by ATM/ATR kinases increases telomerase recruitment and deletion of both kinases results in telomerase inactivation (Lee et al., 2015; Tong et al., 2015). While TRF1 overexpression increases telomere shortening, TRF1 deletion results in telomere elongation (van Steensel

and de Lange, 1997). Hence, a quantitative effect on telomerase regulation has been suggested by the amount of bound TRF1 or TRF2 at telomeres. Only recently it was demonstrated that the telomere-binding protein TZAP, also known as ZBTB48, regulates a maximal telomere length. TZAP binds long telomeres and promotes trimming by t-loop excision (Li et al., 2017).

Telomere maintenance by alternative lengthening of telomeres

In the absence of telomerase, replicative immortality is achieved by alternative lengthening of telomeres (ALT) in about 15% of cancers (Bryan et al., 1995). ALT is based on homologous recombination and is mediated by enhanced DNA damage within ALT telomeres that triggers break-induced replication (BIR) (Constantino et al., 2014; Sobinoff and Pickett, 2017). In a study using a TRF1-Fok1 fusion protein to specifically induce DSBs at telomeres of ALT cells, telomeres showed enhanced mobility and clustering. Concomitantly, telomeric movement was homology-directed and RAD51-dependent (Cho et al., 2014). A triple fluorescence *in situ* hybridization (FISH) experiment showed conservatively replicated telomeres in U2OS cells (Roumelioti et al., 2016). Deletion experiments further support the importance of DNA Pol δ components (POLD3,4) in break-induced telomere synthesis (Dilley et al., 2016; Roumelioti et al., 2016).

As a result of enhanced recombination activities, ALT cells exhibit long heterogeneous telomeres that are prone to replication stress. Typical ALT characteristics include clustering of telomeres in promyelocytic leukaemia (PML) bodies, which form ALT-associated PML bodies (APBs), the accumulation of extra-chromosomal telomeric DNA (circular and linear), elevated levels of telomere sister-chromatid exchange (T-SCE) and a decreased or no expression of the chromatin remodeling helicase ATRX (Heaphy et al., 2011; Henson and Reddel, 2010; Lovejoy et al., 2012). Another consequence of enhanced recombination is the interspersion of telomere variant sequences throughout telomeres (Figure 11) (Cesare and Reddel, 2010; Conomos et al., 2012; Lee et al., 2014). In contrast, these repeats are restricted to subtelomeres (i.e. the proximal 2 kb of telomeres) in telomerase-positive cells. Sequencing of genomic DNA identified TCAGGG, TGAGGG and

TTGGGG variant repeats as most common in the ALT cell line WI38-VA13/2RA (Conomos et al., 2012). Variant repeats present additional binding sites at telomeres. The orphan nuclear receptors NR2C/F preferably bind TCAGGG and TGAGGG motifs (Déjardin and Kingston, 2009). The transcription factor NR2C2 binds with highest affinity TCAGGG and hereby contributes to ALT phenotypes (Conomos et al., 2012). Nuclear receptors additionally promote targeted telomere insertion throughout the genome. Thereby, fragile sites are generated possibly leading to genome instability (Marzec et al., 2015).

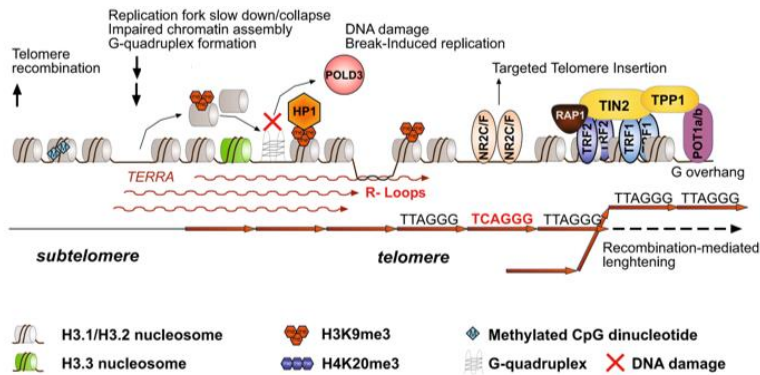


Figure 11: Chromatin organization at ALT telomeres. Telomeres are elongated by HR events leading to the insertion of telomeric variant repeats such as TCAGGG. NR2C/F transcription factors are able to bind telomeres, which can result in genomic rearrangements and genomic instability. DNA damage induces POLD3-dependent break-induced replication. Transcription of the long non-coding RNA TERRA is increased in ALT cells and leads to an enhanced occurrence of R-loops. Modified from (Tardat and Déjardin, 2017).

Chromatin organization

The organization of human subtelomeres and telomeres resembles non-telomeric chromatin with DNA being wrapped around nucleosomes. However, telomeric nucleosomes exhibit a 20-40 bp shorter repeat size and higher mobility (O'Sullivan and Almouzni, 2014). Nucleosomes carry PTMs that can regulate telomere length, which is controlled by the histone methyltransferases SUV39H1/H2 (García-Cao et al., 2004). The structure of subtelomeric chromatin is regulated by direct DNA methylation of CpG dinucleotides through DNMT3a/3b and further maintenance by DNMT1 preventing inter-telomeric HR (Gonzalo et al., 2006). In contrast, telomeres are free of DNA methylation as they are poor in cytidines. For many years it has been assumed that telomeres consist of

heterochromatic material with histone PTMs such as H3K9me3, H4K20me3, hypoacetylation of histone H3 and H4 and binding of the heterochromatin protein HP1 (Schoeftner and Blasco, 2010). However, the epigenetic nature of telomeres remains controversial. Only recently it was shown that most human cancer cell lines do not carry the heterochromatic H3K9me3 mark, but instead reveal features of euchromatin with H4K20ac1 and H3K27ac marks (Cubiles et al., 2018).

A typical hallmark of telomeres is the repression of genes in the subtelomeric region known as telomere position effect (TPE), which is mainly studied in yeast (Gottschling et al., 1990). Also in human cells TPE exists. The Wright laboratory identified 15 subtelomeric genes in fibroblasts that are silenced over distances of up to 10 Mb because of their proximity to the telomere by formation of a loop structure. Upon telomere shortening, loops diminish and genes are increasingly expressed as they are further separated from the telomere (Robin et al., 2014). Thus, telomere length influences the strength of gene silencing (Robin et al., 2014; Baur et al., 2001).

Transcription of telomeres

About ten years ago transcription of human subtelomeres was detected (Azzalin et al., 2007). The long non-coding telomeric repeat containing RNA (TERRA) is mainly transcribed by RNA Pol II (Schoeftner and Blasco, 2008) from the C-rich strand (Luke and Lingner, 2009). At about 25% of telomeres, CpG islands present specific transcription start sites and result in TERRA transcripts heterogeneous in length (Maicher et al., 2014). TERRA transcripts can range from 100 bp up to 9 kb (Azzalin et al., 2007). They carry a 7-methylguanosine cap at the 5' end (Porro et al., 2010). In HeLa cells, polyadenylation occurs at around 7% of TERRA contributing to transcript stability and nucleoplasmic localization (Azzalin and Lingner, 2008). In contrast, non-polyadenylated TERRA is mostly chromatin-associated (Porro et al., 2010). Expression of human TERRA is cell cycle regulated with a peak in early G1 phase and decreasing levels until late S phase (Porro et al., 2010). TERRA levels are tightly regulated, which mainly occurs on transcriptional level or by specific chromatin marks (Azzalin and Lingner, 2008). For instance, CTCF and cohesin are positive regulators of TERRA

transcription by maintaining an open chromatin region (Deng et al., 2012). On the contrary, the DNA methyltransferases DNMT1/3b, HP1 α or SUV39H1 repress transcription after they are recruited by TERRA, generating a negative feedback loop (Arnoult et al., 2012; Porro et al., 2014a). Moreover, TRF2 forms a ternary complex with TERRA and ORC1. Complex formation enables an indirect interaction with HP1 and H3K9me3 resulting in heterochromatin expansion and protection of genome integrity (Deng et al., 2009; Takahama et al., 2013). Interestingly, TRF2 depletion results in TERRA upregulation at all transcribed telomeres (Porro et al., 2014a, 2014b). In a recent report, the interaction network as well as chromatin binding sites of TERRA were identified in mouse embryonic stem cells (Chu et al., 2017). An increase in telomere induced foci upon TERRA depletion revealed a protective role for TERRA at chromosome ends. While TERRA directly interacts with ATRX, both proteins have antagonistic roles. In particular, TERRA may promote ALT in cancer cells by repressing ATRX function and telomerase expression (Chu et al., 2017). A study in yeast showed that TERRA accumulates at critically short telomeres and forms R-loops. Hence, it triggers the DDR and promotes HDR as a protective mechanism against senescence (Graf et al., 2017). In human ALT cells, R-loops are key mediators of telomere maintenance, which are controlled by the enzyme RNaseH (Arora et al., 2014). The amount of hybrids needs to be balanced to not risk telomere integrity, but still allow HR and new DNA synthesis.

The shelterin complex

As already mentioned, the shelterin complex consists of six proteins (see Figure 7). The two telomere repeat-binding factors TRF1 and TRF2 directly bind the dsDNA via their C-terminal DNA-binding domain (Court et al., 2005), while RAP1 is recruited through interaction with TRF2 (Li et al., 2000). POT1 binds the ssDNA overhang (Baumann and Cech, 2001) and forms a heterodimer with TPP1 (Liu et al., 2004a; Ye et al., 2004a). TIN2 acts as a bridging protein for complex assembly (O'Connor et al., 2006). In human, about 100-1,000 copies of the shelterin complex are present per chromosome end. With an estimated density of one complex per 100 bp, a recent study supports the “beads on a string” view of shelterin

complexes binding to telomeric DNA (Erdel et al., 2017). Even in the absence of telomeric DNA, shelterin is able to form a stable complex and subcomplexes without TRF1 or TRF2-RAP1 have been described as well (Liu et al., 2004b; Ye et al., 2004b).

The presence of shelterin at each telomere is crucial for genome integrity and cell viability. A three state model classifies length and functionality of telomeres: closed, intermediate and uncapped (Cesare et al., 2009). While the closed state describes fully protected telomeres with t-loop structures, the intermediate state can either be found on short telomeres, telomeres that fail to form the protective structure or upon POT1 or partial TRF2 inhibition. Thus, the DDR is triggered resulting in aberrant telomere recombination, whereas fusion events are still inhibited (Cesare et al., 2013). Only the uncapped state of telomeres results in spontaneous end-to-end telomere fusions supposedly because of insufficient shelterin loading at shortened telomeres. When telomeres are critically short or when shelterin is disrupted, the DNA damage checkpoint is activated (D'Adda di Fagagna et al., 2003). Four to five critically short dysfunctional telomeres are enough to trigger p53-dependent DDR, which induces replicative senescence, a permanent cell cycle arrest, eventually leading to apoptosis (Kaul et al., 2012; Vettorelli and Passos, 2017). Telomere length thereby contributes to the lifetime of a cell and presents a putative tumor suppressor mechanism (Maciejowski and de Lange, 2017). Almost 60 years ago, the limited capacity of human cells to divide was described and is known as the Hayflick limit (Hayflick and Moorhead, 1961).

Telomeres in disease and aging

Continuous cell divisions in human somatic cells result in programmed telomere shortening, which leads to insufficient end protection and activation of the DDR. Critically short telomeres trigger replicative senescence, a safeguard mechanism against cancer (Shay, 2016). While senescent cells are metabolically active, they stop proliferating and exhibit a persistent DDR. Senescence is an essential cellular mechanism to protect the organism from cancerogenesis. On the downside, senescent cells contribute to age-related diseases through tissue degeneration, the release of reactive oxygen species and pro-inflammatory

cytokines (Correia-Melo et al., 2014). A study in mouse tissue demonstrated a correlation between the accumulation of senescent cells and organismal aging (Wang et al., 2009). Strikingly, pharmacological elimination of senescent cells alleviates aging phenotypes in mice (Ovadya et al., 2018). Telomerase-deficient mouse models show premature aging phenotypes (Blasco et al., 1997; Lee et al., 1998) and an increased expression of TERT in cancer-resistant mice extended mouse longevity by 40% (Tomas-Loba et al., 2008).

Several clinical phenotypes are associated with mutations in genes encoding telomere maintenance and repair proteins such as the telomerase complex (TERT, TR, DKC1), the shelterin complex (TIN2, TPP1, POT1, RAP1), capping factors (CTC1 or STN1) and proteins involved in these processes (e.g. RTEL1) (Bertuch, 2016). For this reason, associated diseases have been named telomeropathies or telomere biology disorders (TBDs). Dyskeratosis congenital, aplastic anemia, pulmonary fibrosis and liver disease are inherited TBDs that exhibit telomere shortening. Patients generally suffer from a weak immune system, tissue impairment, premature aging and susceptibility to a subset of cancers (Armanios and Blackburn, 2012).

By inactivation of both the tumor suppressor p53 and Retinoblastoma (Rb), mammalian cells bypass senescence and regain their proliferative capacity (Shay and Wright, 2005; Shay, 2016). Continuous cell divisions result in further telomere shortening eventually leading to telomere crisis classified as the uncapped state (Cesare et al., 2009). The spindle activation checkpoint induces mitotic arrest that serves as another proliferative barrier. As a consequence, cells undergo mitotic catastrophe, which describes cell death during mitosis. Reactivation of telomerase or ALT results in the escape from crisis and hereby, transformation to a cancerous cell. Genomic rearrangements through breakage-fusion-bridge cycles, aneuploidy and translocations promote cancer progression (Martínez and Blasco, 2017).

Functions of shelterin proteins outside of telomeres

In the last years, more and more studies revealed extratelomeric roles for some shelterin complex members. A role for TRF1 as tumor suppressor has been proposed as its

expression is controlled by miRNA-155, which is involved in leukemia, breast, colon, and lung cancer progression. Higher levels of miR-155 repress TRF1 expression in the absence of p53. Hence, not only telomere protection and replication are affected, but also tumor formation is promoted (Dinami et al., 2014). Increased TRF2 levels activate the angiogenic PDGFR β promoter and lead to enhanced expression of angiogenic factors in endothelial cells of tumor vessels. Thus, early cell differentiation in endothelial cells as well as migration, proliferation and tube formation is facilitated (El Maï et al., 2014). Furthermore, TRF2 binds non-coding regions that are enriched for the TTAGGG motif. Lamin proteins stabilize TRF2 binding at interstitial telomeric sequences (ITS) and a function in promoting long-range telomere loops to these sequences has been proposed (Wood et al., 2014). Additional extratelomeric binding sites of TRF2 that account for gene expression regulation are the CDKN1A promoter and the PCGF3 promoter, the latter being bound in a G-quadruplex dependent manner (Simonet et al., 2011; Yang et al., 2011).

A study using ChIP-sequencing revealed that RAP1 binds non-telomeric regions with a (TTAGGG)₂ binding motif. Changes in subtelomeric gene expression and in several genes involved in metabolism, cancer or cell adhesion were observed in *Rap1*-null mouse embryonic fibroblasts (Martinez et al., 2010; Yeung et al., 2013). Electro mobility shift assays (EMSA) revealed altered TRF2 DNA-binding affinity in the presence of RAP1, thereby suggesting a role for RAP1 in modulating binding at telomeres (Janouskova et al., 2015). Among telomeric proteins, RAP1 is the first described protein that stably localizes in the cytoplasm, whereas its interaction partner TRF2 is restricted to the nucleus (Martínez and Blasco, 2011). In various cancer cells as well as in macrophages, RAP1 associates with inhibitor kinases to ensure nuclear factor (NF)- κ B activation by phosphorylation to recruit chromatin remodeling proteins for transcriptional activation. Thus, RAP1 has been attributed an unexpected role as a modulator of the NF- κ B-mediated inflammation pathway (Teo et al., 2010). Upon telomere shortening, less RAP1 is bound within the shelterin complex and can accumulate in the cytoplasm to associate with inhibitor kinases. As a consequence, NF- κ B activity increases driving age-associated inflammation. In general,

tumor samples exhibit higher levels of TRF2 and RAP1 suggesting facilitation of oncogenesis and their requirement for proliferation and migration (Biroccio et al., 2013).

In addition to the protective role of TRF2 at telomeres, it also inhibits ATM kinase activity at pericentromeres. TRF2 preferentially localizes to the satellite III repeats during S phase and recruits the helicase RTEL1 assisting replication fork progression (Mendez-Bermudez et al., 2018). In another study, an interaction between TRF2 and NOLC1 was reported, which regulates retention of TRF2 in the nucleolus (Yuan et al., 2017). Furthermore, TRF2 plays a role in neuronal differentiation by interaction with the REST transcription factor regulating neuronal genes (Zhang et al., 2008).

It will be exciting to characterize extratelomeric functions of telomere-associated proteins to fully understand their regulation of cell integrity.

Zinc finger and BTB proteins

The family of Broad-Complex, Tramtrack and Bric a brac (BTB) proteins consists of about 200 proteins, which differ by additional domains such as Kelch or zinc finger (Znf) domains (Chen and Chen, 2016). The shape of a BTB domain is compact with five α -helices ending in a β -sheet and a core domain of mainly hydrophobic amino acids. In contrast, the exposed surface varies a lot between different BTB proteins. BTB-Znf proteins offer a large surface for protein-protein interactions and are able to form homo- and heterodimers (Perez-Torrado et al., 2006). Thus, they have been suggested to play a role in the regulation of gene expression by acting as transcription factors or by recruiting regulatory proteins (Perez-Torrado et al., 2006). Zinc fingers are one of the most common types of DNA-binding domains. They are about 25-30 amino acids long with the zinc atom being typically coordinated by two cysteine and two histidine residue pairs (C_2H_2). In human, there are about 80 zinc finger and BTB domain-containing proteins with several being classified as transcriptional regulators although experimental validation is still lacking.

In 2017, we identified three Znf-BTB proteins as telomere-associated in vertebrates by quantitative mass spectrometry (*Article II* (Kappei et al., 2017)). One of these proteins is ZBTB48/TZAP, which directly binds telomeres via its zinc finger 11 and controls telomere

length (Jahn et al., 2017; Li et al., 2017; Zhao et al., 2018). In contrast, little is known about ZBTB10. Almost 20 years ago it was identified as a binder of the cis-regulatory CACC element within the gastrin promoter and inhibitor of specificity protein SP1 (Tillotson, 1999). In a miRNA array analysis, ZBTB10 transcript levels were upregulated when miR-27a expression was decreased upon treatment with an histone deacetylase inhibitor in a breast cancer cell line (Scott et al., 2006). Upregulated ZBTB10 was shown to transcriptionally repress SP1, SP3 and SP4 (Mertens-Talcott et al., 2007) and concomitantly SP-regulated genes (Lai et al., 2013; Li et al., 2010). The SP transcription factors play a role in many cellular processes such as growth, differentiation, apoptosis, immune and DNA damage response as well as chromatin remodeling. Up until now, a possible implication of ZBTB10 in telomere biology remains unknown.

Aims of the thesis

I explored two fields of research using mass spectrometry-based proteomics and interactomics: the development of *Drosophila melanogaster* and telomere-associated proteins.

Drosophila is a widely used model organism that presents a powerful tool to study genetics and biochemical mechanisms of development and aging. The award of six Nobel prizes emphasizes the impact of research performed in fruit flies. Until now, proteomic studies were primarily restricted to specific developmental stages. The aim of this project was to create a proteomic dataset covering the complete life cycle, and allowing analysis of protein expression changes throughout development. Because embryogenesis is of particular interest, we aimed at providing a second proteomic dataset with high temporal resolution. Due to the large number of uncharacterized proteins, we utilized data integration approaches to infer possible biological functions for these proteins.

Telomeres protect the ends of eukaryotic chromosomes from fusions, degradation and illegitimate DNA repair activities. Thus, telomeres are crucial to preserve genome stability and integrity. In a large-scale phylointeractomics approach, we aimed to identify proteins associated with the vertebrate telomere sequence TTAGGG. Functional follow-up characterization of novel candidate proteins constitute future objectives. In my thesis, the goal was to characterize one of these candidates: ZBTB10. This zinc finger protein has not been associated with telomeres before, and I determined its binding ability to telomeric repeat sequences. Moreover, I globally identified ZBTB10 interactors by interactomics to gain insights into the role of ZBTB10 in cellular processes.

Publications

The developmental proteome of *Drosophila melanogaster*

Summary

In this project, we used label-free quantitative mass spectrometry to explore the proteome of *Drosophila melanogaster* at 15 distinct developmental time points. We robustly quantified 7,952 proteins with more than half of them being detected at all stages. In addition to the analysis of stage-specific proteins, for the first time we report maternally-loaded proteins suggesting a putative role in oogenesis and early embryogenesis. Knockdown of one such protein, CG17018, drastically reduced the hatching rate and affected embryonic development.

Moreover, we provide a highly time-resolved proteome during embryogenesis at 14 time points and quantified 5,458 proteins. Differentially expressed proteins were grouped together using an automated clustering approach, thereby proposing functions for uncharacterized proteins. Integration of tissue-specific RNA expression data (Lécuyer et al., 2007) allowed us to infer a specific function of similarly expressed proteins in the development of certain tissues. The availability of transcriptome data (Graveley et al., 2011) further enabled us to compare mRNA and protein expression levels revealing moderate correlation at any time point. This is especially important when studying proteins and mRNA during development.

In summary, this study presents a powerful resource for future studies and is applicable at: <https://butterlab.imb-mainz.de/flydev/>.

Zusammenfassung

An 15 Zeitpunkten während der vier Lebensstadien von *Drosophila melanogaster* (Embryo, Larve, Puppe und adulte Fliege) wurden mittels quantitativer Massenspektrometrie 7952 Proteine robust quantifiziert. Damit liefern wir die erste proteinbasierte Studie, die alle Entwicklungsstadien abdeckt. Während die Hälfte der Proteine in allen Lebensstadien vorkommt, wurden Proteine analysiert, die spezifisch für jeweils eine der vier Lebensstadien sind. Des Weiteren erfassen wir erstmalig Proteine, die mütterlicherseits weitergegeben wurden und somit wichtig für die Oogenese oder die frühe embryonale Entwicklung sein können. Anhand eines Beispiels des bisher nicht charakterisierten Proteins CG17018 wurde gezeigt, dass eine reduzierte Proteinexpression zu einem deutlichen Verlust der Schlupfrate und Fehlern in der embryonalen Entwicklung führt.

Da die Embryogenese ein wichtiges streng reguliertes Entwicklungsstadium ist, haben wir zusätzlich das Proteom in zeitlich engen Abständen über 20 Stunden gemessen und insgesamt 5458 Proteine quantifiziert. Durch automatisiertes Clustering ist es gelungen differentiell exprimierte Proteine zu gruppieren, um somit auf eine mögliche Funktion von nicht charakterisierten Proteinen schließen zu können. Durch die Einarbeitung von gewebespezifischen RNA Expressionsdaten lässt sich außerdem eine mögliche Funktion für Proteine in der Entwicklung eines bestimmten Gewebes ableiten. Der Vergleich von Proteom- und Transkriptionsdaten zeigt, dass mRNA und Proteinlevel zu allen Zeitpunkten lediglich moderat korrelieren. Dieser wesentliche Aspekt sollte bei der Entwicklungsforschung beachtet werden.

Insgesamt haben wir eine wertvolle Ressource erstellt, die auf unserer interaktiven Homepage genutzt werden kann: <https://butterlab.imb-mainz.de/flydev/>.

Statement of Contribution

This is a shared project among Nuria Casas-Vila, Sergi Sayols and myself. While Sergi did all data analysis, Nuria and I planned and conducted all sample collections, we processed and measured samples to determine the embryonic and whole life cycle proteomes. Moreover, we performed Western blot analysis and the Engrailed antibody staining of embryos. For about one year, I discussed data analysis and how to structure and present our data together with Sergi, Nuria and Falk in a weekly meeting. Nuria and I assembled and finalized all figures for the manuscript, which we wrote together with Falk.

Supervisor confirmation

The developmental proteome of *Drosophila melanogaster*

Nuria Casas-Vila,^{1,8} Alina Bluhm,^{1,8} Sergi Sayols,^{2,8} Nadja Dinges,³ Mario Dejung,⁴ Tina Altenhein,⁵ Dennis Kappei,⁶ Benjamin Altenhein,^{5,7} Jean-Yves Roignant,³ and Falk Butter¹

¹Quantitative Proteomics, Institute of Molecular Biology, 55128 Mainz, Germany; ²Bioinformatics Core Facility, Institute of Molecular Biology, 55128 Mainz, Germany; ³RNA Epigenetics, Institute of Molecular Biology, 55128 Mainz, Germany; ⁴Proteomics Core Facility, Institute of Molecular Biology, 55128 Mainz, Germany; ⁵Institute of Genetics, Johannes Gutenberg University, 55128 Mainz, Germany; ⁶Cancer Science Institute of Singapore, National University of Singapore, Singapore 117599; ⁷Institute of Zoology, University of Cologne, 50674 Cologne, Germany

Drosophila melanogaster is a widely used genetic model organism in developmental biology. While this model organism has been intensively studied at the RNA level, a comprehensive proteomic study covering the complete life cycle is still missing. Here, we apply label-free quantitative proteomics to explore proteome remodeling across *Drosophila*'s life cycle, resulting in 7952 proteins, and provide a high temporal-resolved embryogenesis proteome of 5458 proteins. Our proteome data enabled us to monitor isoform-specific expression of 34 genes during development, to identify the pseudogene *Cyp9f3Ψ* as a protein-coding gene, and to obtain evidence of 268 small proteins. Moreover, the comparison with available transcriptomic data uncovered examples of poor correlation between mRNA and protein, underscoring the importance of proteomics to study developmental progression. Data integration of our embryogenesis proteome with tissue-specific data revealed spatial and temporal information for further functional studies of yet uncharacterized proteins. Overall, our high resolution proteomes provide a powerful resource and can be explored in detail in our interactive web interface.

[Supplemental material is available for this article.]

Drosophila melanogaster is among the best-described model organisms for development and aging. During its life cycle, it progresses through well-defined stages including embryo, larva, pupa, and adult, undergoing a complete phenotypic metamorphosis (Lawrence 1992). These transitions are based on tightly regulated gene expression at the transcriptional, epigenetic, and translational level. Currently, most developmental gene expression studies in *Drosophila* rely on in situ hybridization of RNA (Lécuyer et al. 2007; Tomancak et al. 2007), transcriptome analysis using large-scale microarray/RNA-seq data sets (Chintapalli et al. 2007; Kalinka et al. 2010; Graveley et al. 2011; Brown et al. 2014), or a combination of both (Jambor et al. 2015). However, mRNAs are further translated into proteins, which perform the actual cellular functions. It has been shown in multiple species such as *Saccharomyces cerevisiae* (Griffin et al. 2002), *Trypanosoma brucei* (Butter et al. 2013), *Caenorhabditis elegans* (Grün et al. 2014), and human (Schwanhäusser et al. 2011), as well as in *Drosophila melanogaster* (Bonaldi et al. 2008), that transcript levels are only a moderate predictor for protein expression as they do not account for post-transcriptional processes such as translational regulation or protein stability (Vogel and Marcotte 2012; Liu et al. 2016). Recently, this has also been addressed with a developmental perspective in *Caenorhabditis elegans* (Grün et al. 2014), *Xenopus laevis* (Peshkin et al. 2015), and *Trypanosoma brucei* (Dejung et al. 2016), but not yet in *Drosophila*.

The number of fly proteins with available antibodies increased in the last decade from around 450 (Adams et al. 2000) to 1586 (listed in FlyBase version 6.01) but still covers only a small fraction of expressed genes. To accelerate protein studies in *Drosophila*, several

tagging strategies were devised. Around 100 genes have been fused with GFP using piggyBac transposition (Morin et al. 2001), and 400 GFP-tagged fly lines have been established using MiMICs (Minos Mediated Integration Cassettes) to permit systematic protein investigations (Nagarkar-Jaiswal et al. 2015). In an alternative approach, BAC TransgeneOmics allowed the creation of 880 lines and the systematic study of 207 GFP-tagged fly proteins (Sarov et al. 2016). In principle, all protein-coding genes can be investigated, but this requires the establishment of a line for each protein. Additionally, a putative caveat of tagging strategies is altered protein behavior like mislocalization, changes in protein stability, or a dominant negative regulatory effect (Margolin 2012).

The fruit fly is one of the model species investigated by modENCODE (The modENCODE Consortium et al. 2010), and thus several large data sets are available for mapping histone modifications (Kharchenko et al. 2011), global RNA levels during development (Graveley et al. 2011), and tissue-specific splicing (Brown et al. 2014). In contrast, proteomic studies in *Drosophila* have been restricted to certain developmental stages. For example, changes in the proteome during aging from eclosion to 60-d-old flies (Sowell et al. 2007), the adult itself (Sury et al. 2010; Xing et al. 2014), larva and pupa (Chang et al. 2013), the embryo (Fabre et al. 2016), and the oocyte-to-embryo transition (Kronja et al. 2014) have been investigated. However, these studies have relatively low proteome coverage (around 2000 proteins), do not cover the complete developmental process, and are not directly comparable because of technical differences.

*These authors contributed equally to this work.

Corresponding author: f.butter@imb.de

Article published online before print. Article, supplemental material, and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.213694.116>.

© 2017 Casas-Vila et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <http://genome.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

Applying label-free quantitative proteomics (Cox et al. 2014), we here measured protein expression throughout the *Drosophila* life cycle with a coverage of 7952 proteins to provide insight into proteome remodeling. With embryogenesis being a focus in *Drosophila* developmental studies, we amended the life cycle proteome with an embryogenesis proteome of 5458 proteins with high temporal resolution. Finally, data integration with tissue-specific (Lécuyer et al. 2007) and developmental transcriptomic studies (Graveley et al. 2011) allows investigation of the importance of spatial and translational regulation.

Results

Proteomics screen of the life cycle

We collected whole-animal samples at 15 representative time points during the *Drosophila* life cycle (Fig. 1A). The embryonic time points were chosen according to major stages of embryonic development: prior to zygotic gene activation (0–2 h, E02), gastrulation (4–6 h, E06), organogenesis (10–12 h, E10), and the late stages of embryogenesis (18–20 h, E20). For larva, the three different instar larvae (L1, L2, and early L3) and a late stage (L3 crawling larva

va) were examined. Pupae were collected daily starting with the white pupa, and, for adults, the virgin males and females (up to 4 h after eclosure) as well as 1-wk-old animals of each sex were chosen. All samples were collected as biological quadruplicates and processed by mechanical disruption with a universal protein extraction protocol. For each replicate, a 5-h mass spectrometry (MS) run was used, resulting in 340 h of measurement (68 MS runs). We searched the resulting eight million MS/MS spectra against a *Saccharomyces cerevisiae* and *Drosophila melanogaster* database using the MaxQuant software suite (Cox and Mann 2008). Overall, we identified 9627 protein groups (a protein group contains proteins indistinguishable by the peptides that were identified) with 144,067 unique peptide sequences at a FDR < 0.01. This number includes 1078 yeast and 8549 *Drosophila* protein groups (Supplemental Fig. S1A). The identification of yeast proteins is nearly exclusively restricted to the larval stages, where it is a food source (Supplemental Fig. S1B). The number of 8549 identified fly proteins is comparable to a previous in-depth measurement of multiple sources of *Drosophila* material reaching 9124 proteins (Brunner et al. 2007). After filtering for robust detection in at least two replicates of any time point, we performed our subsequent analysis on a set of 7952 protein groups (Fig. 1B; Supplemental Table S1).

Developmental processes are tightly regulated and thus highly reproducible in each organism. Nevertheless, to visualize biological variability of this process, we performed correlation and principal component analysis (PCA). To increase quantitation reliability, all label-free quantitation (LFQ) values were solely based on unique peptide intensities for each protein group. Despite the fact that our replicates are originating from different egg-laying events, are being processed independently, and are measured several days apart on the mass spectrometer, we find a very high correlation within the time points ($R = 0.84$ – 0.98) (Supplemental Fig. S1C) and clear formation of clusters in PCA (Fig. 1C). These findings demonstrate a very high reproducibility of our experimental conditions from the biological system to the mass spectrometry measurement.

Core proteome and protein expression dynamics

To identify a core proteome, i.e., proteins detected at all stages of development, we grouped the proteins according to their presence in the four major stages of the life cycle (Fig. 2A). We found 4627 protein groups, more than half of our proteome, to be detectable in all stages. To obtain an overview of the functionality of these continuously expressed proteins, we performed gene ontology (GO) annotation enrichment analysis and reduced the GO term complexity to uncover major descriptors (Fig. 2B). As

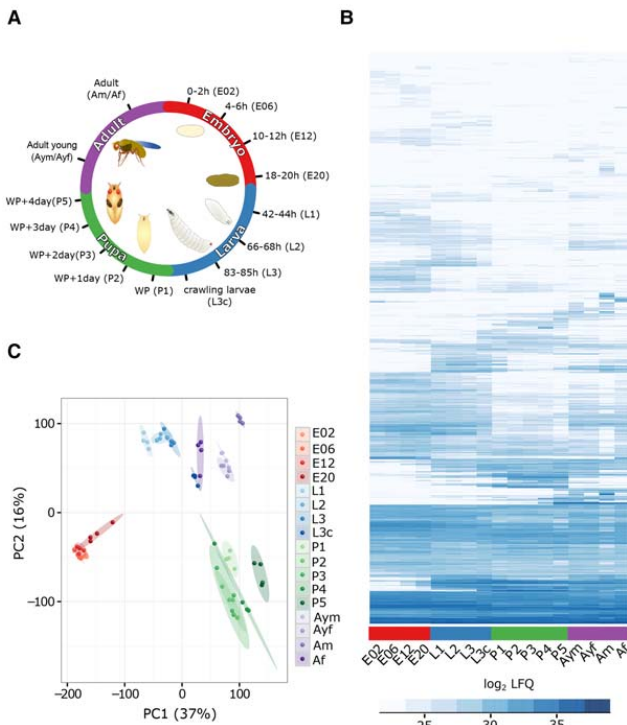


Figure 1. *Drosophila* developmental life cycle proteome. (A) Scheme depicting the collected time points throughout the four major metamorphic stages of *Drosophila* (embryo [red], larva [blue], pupa [green], and adult [violet]). (WP) White pupa, (L3c) crawling third instar larva. (B) Heat map of log₂ LFQ values of the 7952 protein groups quantified during fly development. (C) Visualization of the first two principal components separating samples according to their developmental stage. The biological replicates are indicated in the same color, with elliptic areas representing the standard error of the two depicted components.

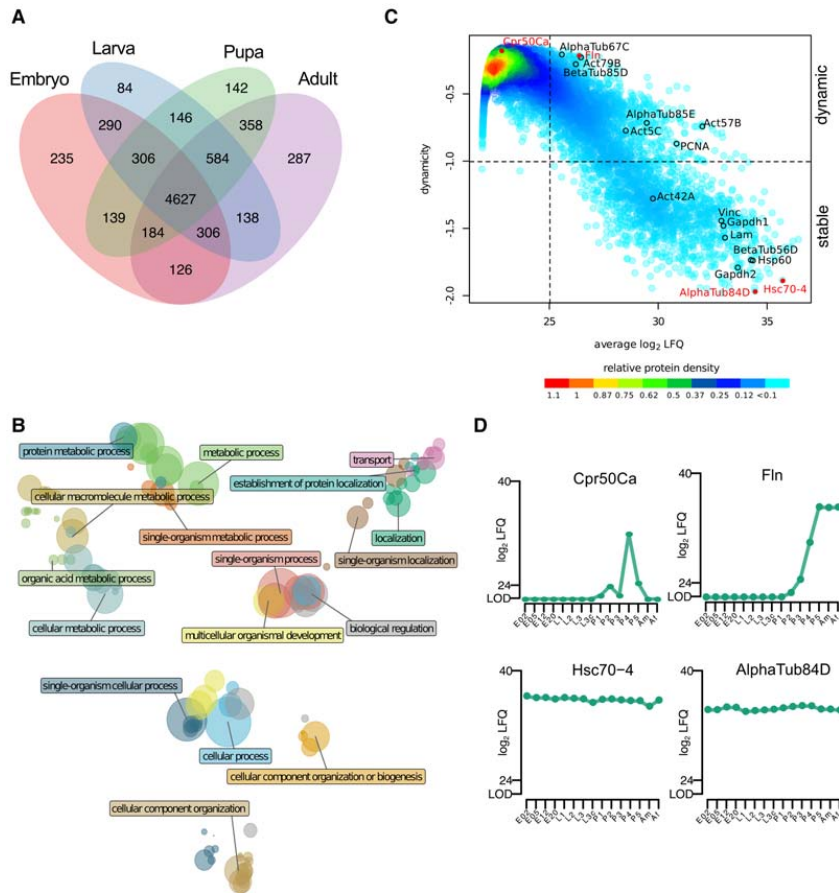


Figure 2. Characteristics of the developmental proteome. (A) Overlap of quantified protein groups between developmental stages results in a core proteome of 4627 proteins. (B) Clusters of enriched GO terms obtained from the core proteome are plotted in a coordinate system defined by the first two dimensions of a multidimensional scaling according to their similarity scores. The color of the circle represents the GO cluster with a representative term highlighted. The diameter of the circle is proportional to the size of the GO category. (C) The density plot relates protein abundance with a dynamicity score during developmental protein expression (log₁₀ transformed Gini index). In the lower-right quadrant, highly stable proteins are represented, while the upper-right quadrant contains proteins with changing expression levels during development. (D) Expression profiles for two highly dynamic (upper panel) and two stably expressed (lower panel) proteins highlighted in red in the dynamicity plot.

expected, our core proteome is enriched for metabolic and cellular processes describing the basic activities of any cellular system, exemplified by covering all known proteins for such essential processes as tRNA aminoacylation, endosome transport via a multivesicular body sorting pathway, cell junction maintenance, nuclear pore organization, and ribosome assembly (Fig. 2B; Supplemental Fig. S2A; Supplemental Table S2). We also analyzed developmental expression dynamics for all proteins with an averaged abundance above the detection limit, log₂ LFQ intensity > 25 (Fig. 2C; Supplemental Table S3). We additionally applied a Gini coefficient filter of 0.1, which divided our proteome into 1386 stably expressed proteins throughout life cycle and 1978 differentially expressed proteins. Consistent with a previous developmental study in *Xenopus*, we see that the dynamicity decreases with protein abundance (Fig. 2C; Peshkin et al. 2015). We show exam-

ples of highly dynamic and stably expressed proteins (Fig. 2D; Supplemental Fig. S2D). The stable proteins include the widely accepted loading controls: tubulins, actins, heat-shock proteins, Gapdh1, Gapdh2, and Vinculin.

Developmental expression profiles of highly abundant proteins

We first characterized the 100 most abundant proteins per stage, comprising around 10% of the total protein mass (Supplemental Table S1). Among proteins with the highest LFQ values, we find ribosomal proteins, being especially prevalent in the top 100 list during embryogenesis, a phase of rapid cell proliferation. The fly uses different storage proteins at specific developmental stages: yolk proteins (Yp1, Yp2, and Yp3) in embryogenesis and Lsp proteins, whose protein abundance rises drastically in L3. Among

these highly abundant proteins, there are several preliminary annotated genes that are not further characterized. CG1850, representing the most abundant protein in the pupal stage, shares a small stretch of similarity to the cuticular protein Cpr72Eb (BLAST *E*-value: 0.019) (Supplemental Fig. S2B). Interestingly, some other highly expressed computed genes (CG) also show similar protein expression patterns to well-studied cuticular proteins like Cpr72Ea (CG1850 and CG13023), Cpr64Aa and Cpr64Ac (CG34461 and CG42323), and Cpr66D (CG16886 and CG30101). While thus far we have looked at the most highly expressed 100 proteins, our proteome can be interrogated to reveal the temporal expression pattern of any quantified protein.

Proteome remodeling throughout the life cycle

Our proteome covers a dynamic range of more than six orders of magnitude, showing expression changes of individual proteins of more than 100,000-fold (Supplemental Fig. S2C). We interrogated our data set for stage-specific proteins by applying ANOVA ($FDR < 0.01$) on the $\log_2$ LFQ values (Fig. 3A). The majority of these 1535 differentially regulated protein groups are found in adult flies (556), followed by embryos (473), pupae (317), and larvae (189). To connect the proteome differences to stage-specific biological functions, we performed GO enrichment analysis on clustered protein expression profiles (Fig. 3A; Supplemental Tables S4, S5). The most enriched GO terms during embryogenesis include mitotic cell cycle regulation and nuclear division represented by cyclins (CycE, CycA, CycB) and developmental kinases, such as Loki (Lok), Greatwall (Gwl), and Grapes (Grp). By this clustering, we were able to separate an early and late embryogenesis phase (Fig. 3B). The early phase (0–6 h) is characterized by high expression of proteins involved in cytoskeleton organization (Dgt4, AlphaTub67C, and GammaTub37C), microtubule binding proteins (Mars and Wee Augmin [Wac]), as well as the classical examples Bicardal C (BicC) and Cup, important in translational regulation of the *oskar* mRNA. In contrast, proteins involved in tissue morphogenesis, such as Bazooka (Baz), Fat (Ft), Ribbon (Rib), and Tramtrack (Ttk), are up-regulated in later phases (12–20 h). Stage-specific proteins in larvae and pupae include expected structural constituents of the chitin-based cuticle: Lcp, Tweedle (Twd), and cuticular proteins. Intriguingly, several proteins that are highly up-regulated only at a single pupal stage, like CG13376, CG13082, and CG42449, are poorly characterized (Fig. 3B; Supplemental Fig. S3A). In the adult, odorant-binding proteins (Obp83b and Obp57a), proteins involved in light perception and phototransduction (Arr1 and Arr2), and the retinal degeneration protein A (RdgA) show strong expression, consistent with the adult fly having a fully developed light sensory system. Also, proteins involved in muscle contraction, like flightin (Fln) and Eaet1, increase their expression 100-fold in adult stages (Fig. 3B).

Overall, our data are in agreement with previously published studies and connects protein expression with well-described morphological changes during *Drosophila* development. Therefore, our screen defines the developmental stage to study molecular or phenotypic effects of yet uncharacterized proteins. All protein profiles can be interrogated using the interactive web interface (<http://www.butterlab.org/flydev>).

Developmentally regulated functions: ecdysone-induced proteins and cuticle formation

The regulation of molting by endogenous 20-hydroxyecdysone (20E) is a prototype example of hormonal gene regulation path-

ways in insects (Yamanaka et al. 2013). Previous microarray studies focused on 20E-induced gene regulation of mRNA transcripts between the L3 larval stage and 12 h after puparium formation (Beckstead et al. 2005; Gonsalves et al. 2011). However, for the ecdysone-induced gene family 71E (*Eig71E*), we find intriguing differences between the expression profiles of mRNA and protein in pupae. Messenger RNA expression is detectable in three different waves: *Eig71Ee* spikes at L3c, another group represented by *Eig71Ed* at P1, and a later group represented by *Eig71Ek* at P2 (Fig. 3D; Supplemental Fig. S3B; Graveley et al. 2011). While the mRNA is detectable only in early pupal stages, the corresponding *Eig71E* proteins show prolonged high expression levels until P5 (Fig. 3C,D). Likewise, second puff genes display a similar transcriptome versus proteome pattern. A 1000-fold up-regulation of glue proteins (Sgs5, Sgs7, and Sgs8) at late L3 concordant with the detection of their mRNA in a narrow window of ~24 h between crawling L3 and P1 (Beckstead et al. 2005) is followed by the presence of the protein in all pupal stages (Fig. 3D; Supplemental Fig. S3C). Our data show that for selected puff proteins, protein stability is the major determinant of their expression patterns during development. In contrast, in a high number of cases, we detect the protein at a single time point, while the RNA is detectable at multiple time points (Fig. 3E). In the aforementioned cases, protein levels cannot be directly predicted by transcriptomics, which demonstrates the necessity of proteome data for studying fly development.

Comparison of sex-specific protein patterns in adult flies

Sex-specific proteins are of high interest and have already been investigated by several proteomics studies (Dorus et al. 2006; Takemori and Yamamoto 2009; Sury et al. 2010; Wasbrough et al. 2010). To benchmark our label-free quantitative approach, we compared our adult time point to the published SILAC data set (Sury et al. 2010) and found a high overlap of sex-specific proteins ($R = 0.84$) (Supplemental Fig. S4A), showing that our developmental proteome recapitulates previous studies that are more specialized. To identify sex-specific proteins, we defined a fourfold expression difference with a *P*-value > 0.01 between male and female flies (1 wk old) and found 308 male- and 374 female-specific proteins (Fig. 4A; Supplemental Table S6). The 308 male proteins include Tektin-A and Tektin-C as sperm-specific flagellar proteins (Amos 2008), several less characterized genes known to be expressed in fly testes and seminal vesicles (Dorus et al. 2006; Takemori and Yamamoto 2009), and some proteins functioning in male development, like Lectin-46Ca, Lectin-46Cb, and Lectin-30A. For some proteins, like Hsp60B, Hsp60C, and the male fertility factor Kl-5, as well as Aquarius (Aqus) and Antares (Antr), an essential role in sperm development or sperm storage has already been demonstrated. The list of 374 female-specific proteins include vitelline membrane (Vm32E) and chorion proteins (Cp15, Cp18 and Cp36), which are important for eggshell assembly, the vitellogenins (Yp1, Yp2, and Yp3), and the fatty acid desaturase Fad2.

Additionally, our developmental proteome allows the investigation of young flies, which were collected as virgins within 4 h after eclosure (Supplemental Table S7). The majority of proteins are equally expressed in both sexes (Supplemental Fig. S4B). While we detect only 21 female-specific proteins in young flies, there are 155 proteins with higher expression in its male counterpart (Fig. 4B; Supplemental Fig. S4C,D). In agreement with this observation, a previous transcriptomic study showed up-regulation of genes in

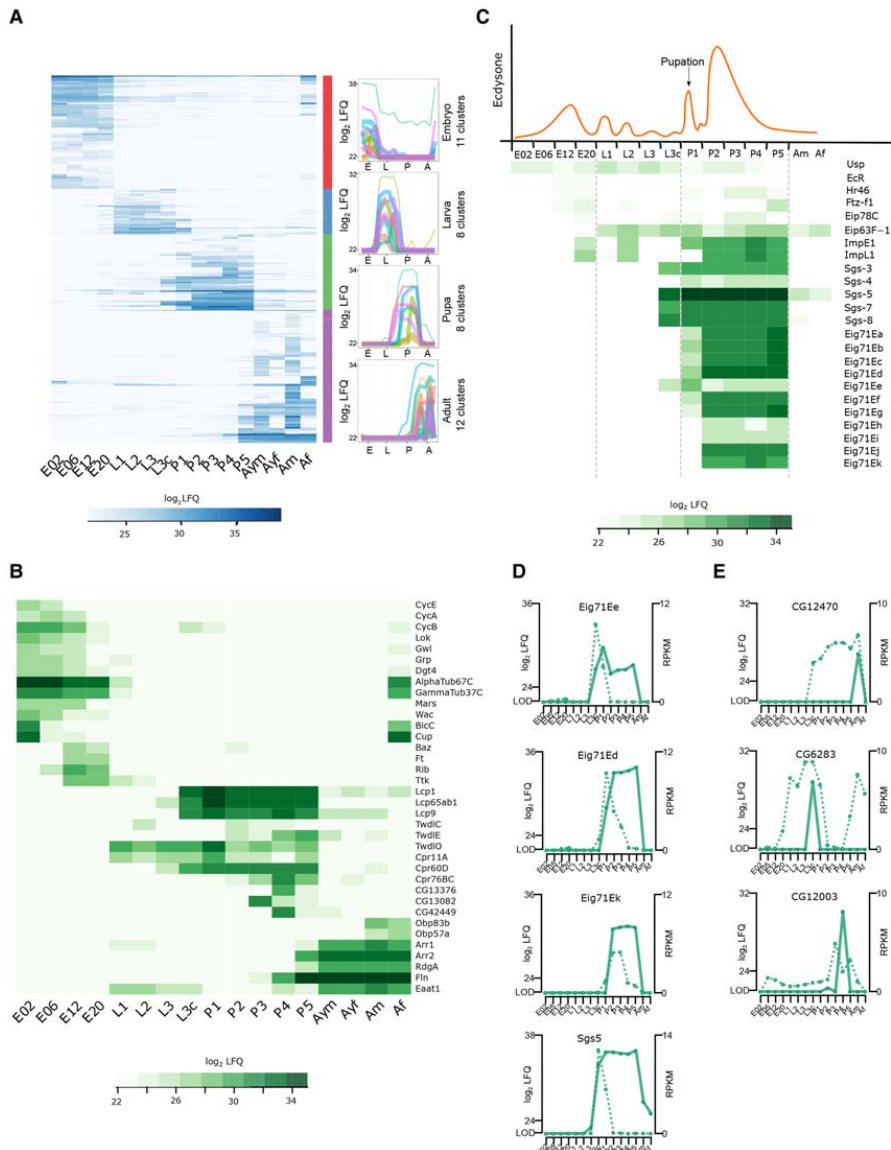


Figure 3. Stage-specific proteins and ecdysone-induced developmental regulation. (A) Heat map showing 1535 protein groups found to be differentially (ANOVA, FDR < 0.01) regulated during the life cycle. These protein groups were clustered into up to 12 stage-specific profiles. Average profiles of the individual clusters for each developmental stage are shown. (B) Heat map showing log₂ LFQ abundance of proteins with stage-specific expression profiles discussed in the text. (C) Schematic representation of ecdysone pulses during fly development (upper panel) and heat map of log₂ LFQ expression levels of selected proteins of 20-hydroxyecdysone regulated genes (lower panel). (D) For the *Eig71E* and *Sgs* gene family, RNA expression profiles (dotted line) differ from protein levels (solid line) during the pupal phase, demonstrating prolonged protein stability. (E) Three examples showing single protein expression burst, but more broadly detectable RNA indicating more tightly controlled protein expression.

female flies after mating, suggestively triggered by sperm and seminal fluid proteins (McGraw et al. 2008). In contrast, the majority of male-specific proteins are already present in young male flies prior

to mating (Fig. 4B; Supplemental Fig. S4D). Interestingly, the only two proteins with more than 30-fold up-regulation in virgin females compared to males are not characterized: CG13862 and

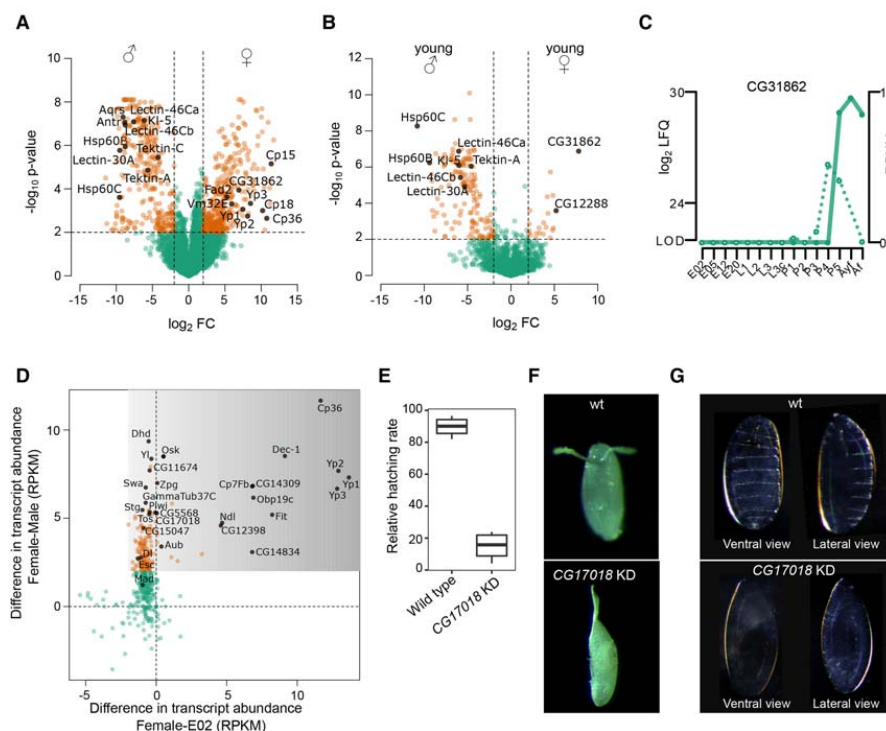


Figure 4. Sex-specific proteome and maternally loaded proteins. (A) Volcano plot comparing protein expression levels between 1-wk-old male and female flies. Candidates discussed in the text are highlighted (filled black circles). Dashed lines indicate a fourfold expression difference with $P < 0.01$. (B) Volcano plot comparing protein expression levels between young male and female flies (<4 h old after eclosion) shows very few female-specific proteins. Candidates discussed in the text are highlighted (filled black circles). Dashed lines indicate a fourfold expression difference with $P < 0.01$. (C) Developmental expression profile of the female-specific protein CG31862 shows detection of mRNA (dotted line) in late pupal stage, while the protein (solid line) is also found in female flies. (D) Integration of mRNA levels with embryo-specific proteins allows identifying maternally loaded proteins. The mRNA levels of the adult female flies compared to embryos (x-axis) and males (y-axis) distinguishes cases in which either both the mRNA and protein ($x = 0, y > 2$), or only the protein (darker shaded area) is maternally loaded. (E) Relative embryonic hatching rate (four biological replicates) of *CG17018* knockdown embryos compared to wild type. (F) Image of representative wild type and the *CG17018* knockdown embryo with fused dorsal appendages. (G) Cuticle preparation of embryos revealed absence of denticle belts patterning in the *CG17018* knockdown line.

CG12288. Noteworthy, CG31862 is found in P5 and shows a continuously high protein level, while its RNA expression is restricted to the late pupal phase (Fig. 4C).

Maternally loaded proteins

While there is ample knowledge about maternally loaded RNA in *Drosophila* embryos (Tadros and Lipshitz 2005), no systematic analysis for maternally loaded proteins has been conducted yet. We interrogated our data for proteins enriched during embryogenesis whose RNA levels were higher in adult females compared to adult males. Among this subset of likely maternally loaded material should be candidates that have a functional importance during early development. In most cases, protein and mRNA are present in 2-h-old embryos, suggesting that both are maternally loaded (Fig. 4D; Supplemental Table S8). These include well-known examples such as Oskar (Osk), String (Stg), Piwi, Aubergine (Aub), Extra sexcombs (Esc), Dorsal (Dl), Mothers against dpp (Mad), and Swallow (Swa) (Chao et al. 1991; Edgar and Datar 1996; Luschnig et al. 2004;

Simmons et al. 2010; Mani et al. 2014). However, also yet undescribed candidates like CG11674, CG5568, CG17018, CG15047, Zpg, GammaTub37C, and Tosca found in this set represent interesting candidates with a putative role in oogenesis and early embryogenesis. In order to investigate potential germline-specific functions of these candidates, we performed RNAi-mediated knockdown using the driver *nanos*-GAL4 and specific transgenic lines expressing double-stranded RNA from inverted repeats (shRNAs). Germline-specific expression of two independent shRNAs targeting *CG17018* RNA revealed drastic effects on the embryonic hatching rate. While the number of laid eggs was unaffected (Supplemental Fig. S4E), hatching was reduced by almost 80% (Fig. 4E). In addition, ~30% of unhatched eggs displayed defective dorsal appendages that are fused (Fig. 4F). Cuticle preparations showed that *CG17018* knockdown embryos miss the denticle belts, revealing an absence of patterning at early stages (Fig. 4G). Also of note, *CG17018* knockdown ovaries were indistinguishable from wild-type ones, as we could not detect any obvious morphological or differentiation defects (Supplemental Fig. S4F). Taken together,

our findings imply a critical role of CG17018 during early embryogenesis.

Furthermore, our proteomic data set allows a comprehensive classification of maternally loaded proteins when the RNA is not present. The most prominent proteins include the major egg yolk vitellogenins (Yp1, Yp2, and Yp3), Dec-1, Cp36, and Cp7Fb as part of the chorion, the oxidoreductase family member CG12398 for which a role in vitelline membrane formation has been previously suggested (Fakhouri et al. 2006), the serine protease Nudel (Ndl), the sensor protein Obp19c, and the female-specific protein Fit, as well as two uncharacterized candidates, CG14309 and CG14834 (Fig. 4D; Supplemental Table S8).

Small proteins in the developmental proteome

Recently, there has been an increased interest in small proteins and translated small ORFs (smORFs) with up to 100 amino acids (aa) (Ramamurthi and Storz 2014), as their protein-coding potential is difficult to assess bioinformatically (Ladoukakis et al. 2011). These small proteins localize to specific subcellular compartments and perform cellular functions as any other protein (Magny et al. 2013). Our data set detects 268 small proteins (Fig. 5A), of which 84% have two or more unique peptides and temporal expression information (Supplemental Fig. S5A; Supplemental Table S1). This number is similar to a previous investigation using ribosome profiling (Aspden et al. 2014), demonstrating that mass spectrometry-based proteomics is on par with next generation sequencing approaches to detect translation of small proteins.

Peptides originating from noncoding regions of the genome

Peptides originating from putative noncoding regions have been reported in diverse organisms. Therefore, we re-analyzed our data including ncRNA sequences from FlyBase, which we in silico

translated for open reading frames of at least 20 aa. Overall, we identified 29 putative proteins that unambiguously map to non-translated transcripts at a FDR < 0.01 (Supplemental Table S9). Due to short open reading frames of these small proteins, we usually detect a single peptide per transcript. However, only two of these ncRNA-derived peptides showed a good MS2 fragmentation pattern and were independently identified with more than 10 different MS/MS spectra in several replicates and time points. One of these, FBtr0340701, has also been found in a control experiment using human cell lysate (data not shown), classifying it as a false positive identification originating from a contaminant. The only remaining peptide with strong evidence of identification matches to CR43476 (Fig. 5B).

Other genes classified as nonexpressed are pseudogenes. These genes have mutations in their promoter regions or other functional elements that make their expression unlikely (Harrison et al. 2003). We checked for protein evidence of the 2902 reported pseudogenes (FlyBase 6.01) and found nine protein groups in our data set to include peptides unambiguously mapping to pseudogenes. Whereas most of these proteins are represented by a single peptide (Supplemental Table S10), the most prominent hit, FBtr0082602, encoding Cyp9f3Ψ, is supported by 23 peptides including five unique sequences. The measured peptides match to the N-terminal and C-terminal regions, demonstrating that the complete pseudogene is most likely translated (Fig. 5C). Furthermore, Cyp9f3Ψ and Cyp9f2 present distinct expression patterns, further indicating that, despite their close genomic vicinity, they are differently regulated during development (Supplemental Fig. S5B).

Despite an extremely low expression of peptides originating from ncRNA transcripts, only very few detected peptides map to noncoding regions of the genome, illustrating a low false discovery rate in our screen and a carefully curated gene

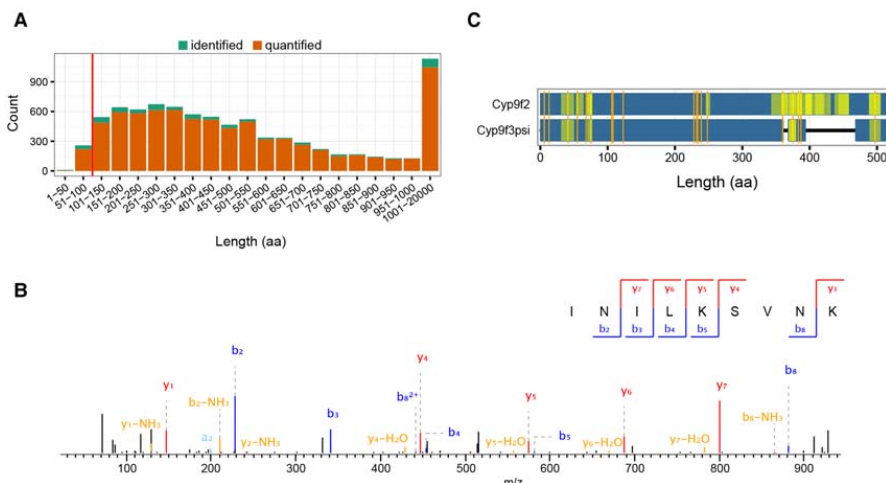


Figure 5. Small proteins and peptides from noncoding regions of the genome. (A) Protein length distribution of identified (green, not enough quantitation values) and quantified (orange) protein groups of the life cycle proteome. Most proteins have quantitation values (>90%), and this fraction only marginally depends on protein length. The red line demarcates the fraction of 268 small proteins (<100 aa). (B) Representative MS/MS spectrum with annotated b- and y-ions of the peptide INILKSVNK⁽²⁺⁾ from the putative noncoding gene CR43476. (C) Sequence comparison of Cyp9f2 and the “pseudogene” product Cyp9f3Ψ with amino acid substitution between both proteins marked in orange. Coverage of peptides for either protein is shown (yellow, more intense regions have overlapping peptides).

annotation of the *Drosophila melanogaster* genome (Matthews et al. 2015).

Highly temporal-resolved embryogenesis proteome

Because they are intensely studied, we were particularly interested in proteome changes during embryogenesis. To investigate the process in a high time-resolved and systematic manner, we collected whole embryos at narrow intervals: every hour after egg laying for up to 6 h and then every 2 h until 20 h (Fig. 6A). These 14 time points were also measured in four independent biological replicates to account for technical, biological and environmental variation. To control for our collections, we staged embryos of selected time points by morphology and Engrailed antibody staining (Supplemental Fig. S6A; Campos-Ortega and Hartenstein 1997). Protein expression levels were determined using label-free quantitation based on unique peptides provided by MaxLFQ (Cox et al. 2014). We detected 6487 expressed protein groups, of which 5458 were quantified in at least two replicates of any time point (Supplemental Table S11). PCA revealed that embryo stages correlate well with our collected time points ($R=0.93$), showing a developmental progression through embryogenesis (Fig. 6A; Supplemental Fig. S6C). Noteworthy, all four independent biological replicates show very high reproducibility ($R=0.92-0.96$) (Supplemental Fig. S6B,C). We also validated the expression profiles of seven proteins by immunostaining with antibodies against endogenous proteins (Fig. 6C; Supplemental Fig. S6D).

Expression profiles during embryogenesis

We analyzed the time course data using a multivariate empirical Bayes approach and identified 1644 protein groups with differential expression during embryogenesis (Fig. 6B). To obtain a functional overview on the embryogenesis process, we performed GO enrichment analysis on this set of differentially expressed proteins. Based on this analysis, we observed enrichment of terms related to very early embryogenesis cellular processes (0–1 h), such as zygotic determination of anterior/posterior axis and syncytial blastoderm mitotic cell cycle (Fig. 6D). Additionally, proteins involved in ribosome biogenesis up-regulate at 2 h to initiate active translation concomitant with zygotic gene activation (ZGA) starting at 2 h. We also noted high enrichment of proteins involved in cell cycle and cytoskeleton organization during early phases of embryogenesis (2–3 h). While proteins involved in nervous system development are highly present at 3 h, muscle structure development proteins are more prominent later in embryogenesis, at 14 h.

As an alternative approach to analyze the data, we automatically clustered the differentially expressed proteins with similar temporal profiles, resulting in 70 distinct clusters, and performed GO enrichment analysis on these clusters (Fig. 6E; Supplemental Tables S12, S13). As a result, the known embryonic developmental program can be followed by temporal alignment of individual clusters (Fig. 6E), possibly hinting at putative functions of not yet characterized proteins.

Integrating the developmental proteome and spatial expression

To integrate spatial information, we fused our proteome profiles with tissue-specific RNA expression data from fluorescence in situ hybridizations (Lécuyer et al. 2007). We chose muscle development to highlight the value provided by the merged data. In the muscle-specific clusters (Fig. 6F; Supplemental Tables S14, S15), we noted up-regulation of proteins involved in muscle develop-

ment such as Mlc2, Mp20, and Mlp60A (Sandmann et al. 2006) at 14 h. Later in embryogenesis (20 h), we found high expression of Eaat1 and EcR, which control muscle contraction at larval stages. Furthermore, this data integration allowed us to identify similarly expressed, not yet characterized proteins (CG1674, CG6040, and CG15022), shown to localize in muscle tissue, suggesting a role in muscle development. In order to test this hypothesis, we performed RNAi-mediated knockdown of two candidates. Remarkably, mesodermal knockdown of either *CG1674* or *CG6040* severely affects locomotion behavior of adult flies (Fig. 6G; Supplemental Fig. S6E). Likewise, the complete *CG6040* loss-of-function produces viable flies that display similar climbing defects, confirming the specificity of the RNAi phenotype. Importantly, neuronal knockdown of both genes did not impair locomotion performance, supporting their muscle-specific functions. We next performed in situ hybridization on embryos and observed a strong enrichment of *CG1674* mRNA in muscle tissue, more specifically in somatic and pharyngeal muscles, whereas *CG6040* exhibits moderate ubiquitous expression (Supplemental Fig. S6F). Altogether, our findings strongly suggest a muscular function for *CG1674* and *CG6040*. However, further investigation will be required to understand their specific role in muscle development.

Alternatively, other tissue data can be inspected for biological insights. The analysis of the central nervous system (CNS) revealed an up-regulation of proteins involved in neural development (Roughest [Rst], Smooth [Sm], and Erect wing [Ewg]) at 8–12 h and in synapse organization and axon ensheathment (Ank2 and Wrapper) at 14 h (Fig. 6F; Supplemental Tables S14, S15). Likewise, all 21 tissue clusters can be examined in our web interface.

Comparing transcriptome and proteome to study translational delay

We compared our embryogenesis proteome with the transcriptome generated as part of the modENCODE Project (Graveley et al. 2011). In agreement with the transcriptome analysis, we found that the general protein complexity is increased during embryogenesis (Fig. 7A; Supplemental Fig. S7A). Similar to a previous study in yeast (Fournier et al. 2010), we found only a moderate correlation (maximum $R=0.5$) between transcriptome and proteome and noted that the best correlation is nonsynchronous, showing a 4- to 5-h proteome delay (Fig. 7B).

By multidimensional scaling followed by clustering, we subgrouped the RNA/protein expression profiles into six clusters (Fig. 7C; Supplemental Table S16). In the majority of cases, the mRNA is more abundant at early time points, while the protein expression peaks at later stages. Except for cluster 1, the remaining clusters illustrate different behavior of RNA and protein during embryogenesis. We observed a temporal proteome delay in clusters 5 and 6: while the RNA expression peaks around 7 h, proteins steadily up-regulate later in embryogenesis, putatively due to translational control mechanisms (Fig. 7C).

Quantification of protein isoforms during embryogenesis

As distinct protein isoforms may show differential developmental regulation, we mined our proteomic data for protein isoforms. We found 34 genes with various quantified isoforms, some of them showing differential expression, such as *lola*, *mod(mdg)4*, and *Rtn1* (Fig. 7D; Supplemental Fig. S7B). We further validated our isoform quantitation by immunoblotting following the expression of Lola- RAA/RI (also known as Lola-K) (Giniger et al. 1994).

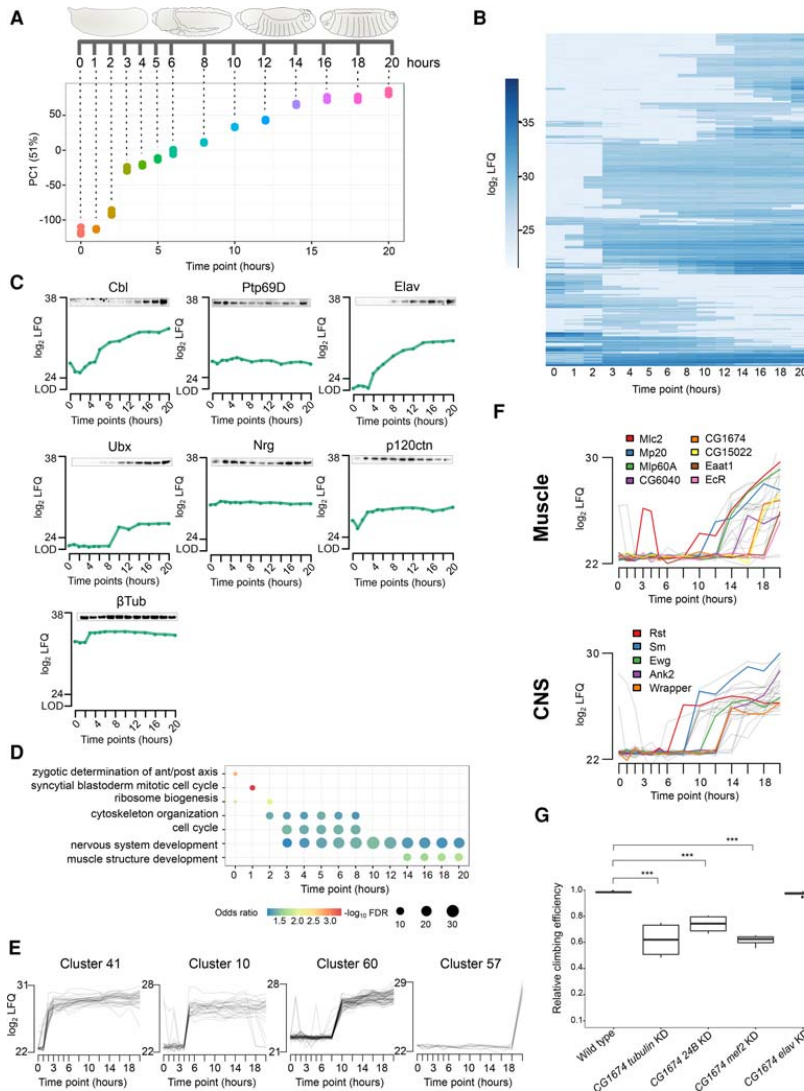


Figure 6. The embryogenesis proteome time course. (A) Scheme indicating the collected time points. PCA shows high reproducibility of replicates, and the first component shows high correlation with developmental progression ($R = 0.93$). (B) Heat map of $\log_2$ LFQ expression values for 1644 developmentally regulated protein groups in embryogenesis. (C) Western blots of seven selected proteins validate their temporal expression profile from the proteomics screen. (D) Dot plot connecting the selected enriched GO terms with developmental progression. The circle size indicates the odds ratio of each GO term category. (E) The regulated protein groups were assigned automatically to 70 clusters based on expression profiles, of which four representative clusters with an up-regulation at 2–3 h (cluster 41), 5 h (cluster 10), 10 h (cluster 60), and 20 h (cluster 57) are shown. (F) Profiles of tissue-specific protein expression created by integrating RNA fluorescence in situ hybridization data. Muscle and central nervous system (CNS) clusters were chosen as examples. (G) Ubiquitous (*tubulin-GAL4*) and mesodermal (*24B-* and *mel2-GAL4*) but not neuronal (*elav-GAL4*) knockdown of *CG1674* results in reduced locomotion activity (Dunnett's test; [***] P -value < 0.001).

While *Lola-RAA/RI* is highly expressed at 20 h (Fig. 7E), its mRNA shows an expression peak at 14 h shown by in situ hybridization (Fig. 7F). This underscores again the importance of a developmental proteome as an addition to transcriptomic studies.

Discussion

We generated high-quality proteome data sets for embryogenesis and the full life cycle of *Drosophila melanogaster* that close the

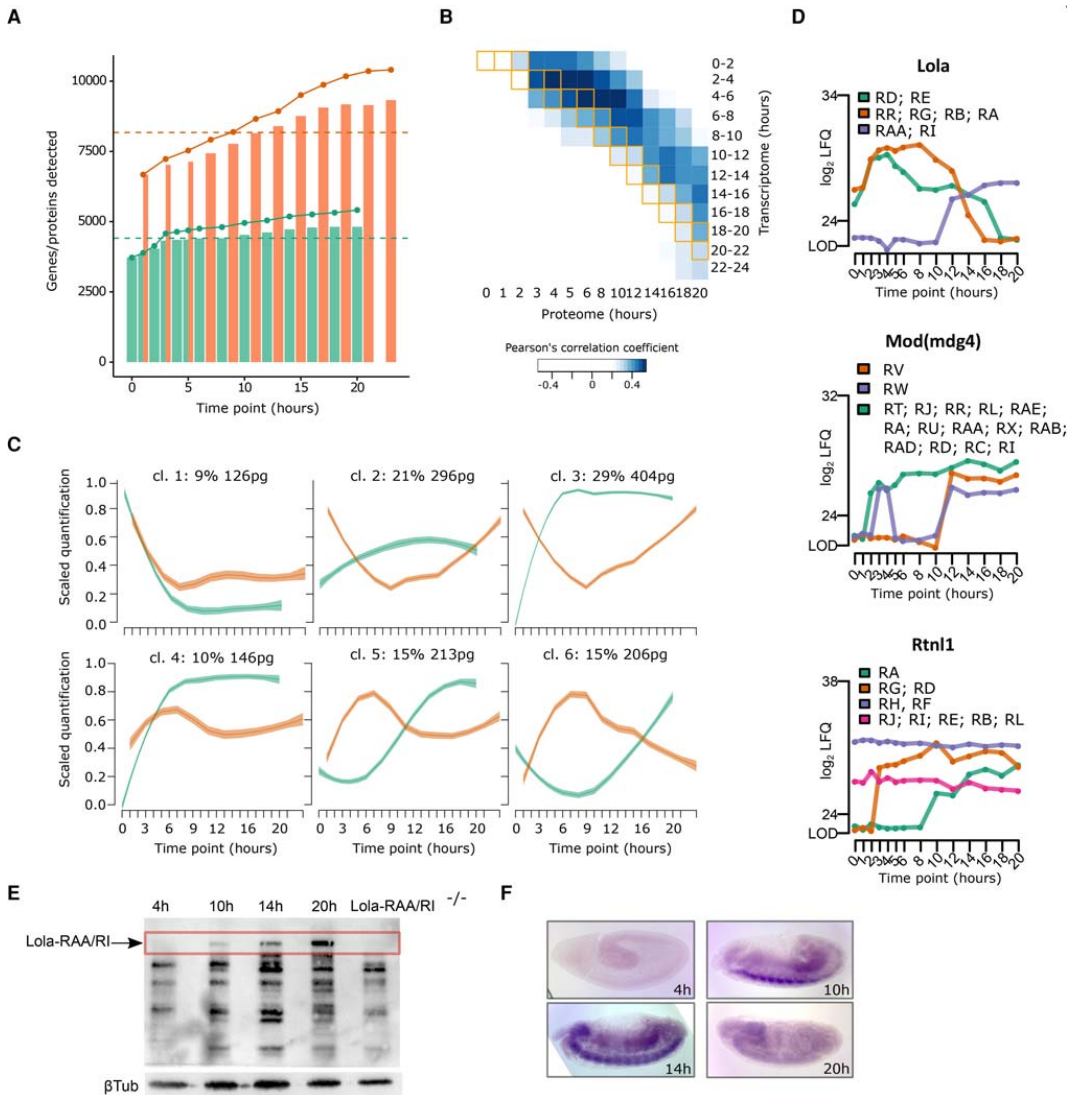


Figure 7. Temporal transcriptome/proteome dynamics and isoform quantitation. (A) Plot showing the number (bars) of detected transcripts (orange) and proteins (green) at each time point. The solid line depicts the cumulative sum of unique transcripts (orange) and proteins (green). The dashed line represents the median across all time points. (B) Heat map displaying the Pearson correlation between transcript and protein expression levels. Matching time points between the two data sets are indicated by orange boxes. (C) Median scaled quantification plotted after clustering of the first PCA component of RNA (orange) and protein (green) expression into six different categories. Shaded regions display the standard error of the fitted line. (D) Expression profiles with isoform-specific information of three proteins: Lola, Mod(mdg4), and Rtnl1. Isoforms are colored according to the legend. (E) Validation of Lola-RAA/Lola-RI isoform quantitation by immunoblotting against Lola at four selected time points. Protein lysate of *lola-RAA/lola-RI* mutant embryos at 20 h were used to identify the isoform-specific band (arrow) corresponding to Lola-RAA/Lola-RI. Beta-tubulin was used as a loading control. (F) RNA levels were determined by in situ hybridization at the selected time points with a specific probe for *lola-RAA/lola-RI*.

gap for systematic developmental investigation of protein expression. Both proteomes cover nearly 8000 and 5500 protein groups during the life cycle and embryogenesis, respectively, accounting for at least one-third of annotated *Drosophila* genes. However,

while these two data sets are larger than previous ones, they are not complete. Especially low abundant proteins or proteins that are highly expressed in a restricted number of small tissues will likely not be present in our proteomes. Thus, a not-quantified

protein can either be absent in this stage or expressed below our limit of detection (LOD) enforced by the mass spectrometry measurement. Nevertheless, these large-scale data sets allow us to assess the developmental expression of proteins and protein isoforms, report maternally provided proteins, validate small proteins (≤ 100 aa), identify Cyp9f3P as an expressed protein-coding gene, and describe peptides originating from noncoding regions.

We scored significant developmentally regulated protein groups: 1535 for the whole life cycle and 1644 for embryogenesis. Nearly half of them are not characterized in depth, suggesting a large area of developmental gene regulation still to be discovered.

We used our data to follow the well-characterized regulation by the hormone ecdysone at a protein level. This revealed intriguing differences to previously reported transcriptome analysis (Beckstead et al. 2005; Gonsalves et al. 2011). For several ecdysone-induced genes, the protein abundance relies on protein stability rather than the presence of RNA transcripts. Overall, transcript abundance and protein levels correlate only modestly. The same observation holds true even when considering the temporal delay between transcript and protein expression. The temporal difference in RNA and protein expression needs to be taken into account when studying phenotypic differences of protein-coding genes using mRNA as a proxy.

As previous transcriptomic studies reported maternally loaded RNAs, our proteomic data enable systematic identification of maternally provided proteins. Here, we catalog not yet reported maternally loaded proteins such as CG14309, CG14834, and CG12398, whose functions in early development need further investigation. For instance, the knockdown of the maternally loaded protein CG17018 results in a severe defect in embryo development.

To gain further insight, we complemented our data sets with other available published data. For example, to de-convolute tissue-specific expression information, we merged our embryogenesis proteome with RNA in situ hybridization data (Lécuyer et al. 2007). This allowed us to pinpoint individual proteins showing tissue-specific developmental regulation, as exemplified with the impaired muscular phenotypes of CG1674 and CG1640 knock-down lines. Additionally, this analysis can be extended to other tissues to uncover currently unknown proteins that might play an essential role in the development of a specific tissue. This underscores the power to combine available high-quality *Drosophila* data sets to achieve a more holistic model for developmental gene regulation.

Methods

Collection of embryos, larvae, pupae, and adult flies

Population cages of wild-type Oregon R flies containing only fertilized females were maintained at 25°C. For the whole life cycle comparative analysis, embryos were collected from cages on agar apple juice plates in 2-h laying time windows and processed immediately (0–2 h) or aged at 25°C for the required time (4–6, 10–12, and 18–20 h). Early larval collections were performed from embryo plates, whereas crawling larvae and pupae stages were collected directly from flasks at indicated time points. Virgin young flies within 4 h after eclosion were collected separately for males and females, as well as 1-wk-old flies (adult flies). For the time course analysis, embryos were collected on apple juice agar plates in 30-min laying time windows, processed immediately (0 h time point) or aged at 25°C for the required time. All samples were mechanically lysed prior to mass spectrometry sample preparation (see Supplemental Material for detailed descriptions).

Mass spectrometry measurement and label-free analysis

Peptides were separated by nanoflow liquid chromatography on an EASY-nLC 1000 system (Thermo) coupled to a Q Exactive Plus mass spectrometer (Thermo). Separation was achieved by a 25-cm capillary (New Objective) packed in-house with ReproSil-Pur C18-AQ 1.9- μ m resin (Dr. Maisch). Peptides were separated chromatographically by a 280-min gradient from 2% to 40% acetonitrile in 0.5% formic acid with a flow rate of 200 nL/min. Spray voltage was set between 2.4 and 2.6 kV. The instrument was operated in the data-dependent mode (DDA) performing a top15 MS/MS per MS full scan. Isotope patterns with unassigned and charge state 1 were excluded. MS scans were conducted with 70,000 and MS/MS scans with 17,500 resolution. The raw measurement files were analyzed with MaxQuant 1.5.2.8 standard settings except LFQ (Cox et al. 2014) and match between run options were activated, as well as quantitation was performed on unique peptides only. The raw data were searched against the translated Ensembl transcript databases (release 79) of *D. melanogaster* (30,362 translated entries) and the *S. cerevisiae* protein database (6692 entries). Known contaminants, protein groups only identified by site, and reverse hits of the MaxQuant results were removed. In the life cycle data set, the imputation was performed in two distinct ways for proteins with a measured intensity (raw) missing an LFQ intensity or proteins with no intensity value. In the first case, values were imputed from a normal distribution with a mean value shifted by -0.6 from the mean value of all measured LFQ intensities and half of the standard deviation. In contrast, proteins with no intensity value were replaced with the smallest measured value in the set. For the embryo time course, missing values were drawn from a distribution calculated with the logspline R package (<https://cran.r-project.org/package=logspline>). For cases where three or more replicates were measured, the mean of the measured replicates was used as the mean parameter of the distribution. Otherwise, the average of the two neighboring time points was used. In cases of no measured values in neighboring time points or for proteins measured only in one replicate with no surrounding values, a fixed value of 22.5 close to the LOD was used.

Bioinformatics analysis

Significant changes during the life cycle were calculated by analysis of variance (ANOVA), flagging as stage-specific proteins those with $FDR < 0.01$ (Benjamini–Hochberg procedure) and present in either one unique stage or differing in one stage compared to the rest ($\log_2$ LFQ FC > 4 in all stages, allowing only one not fulfilling the condition). The effect of the differences was assessed calculating Cohen's effect size and the Tukey HSD post-hoc test. The Gini ratio was used to measure the stability of protein abundance throughout the time course. Automatic clustering of genes and samples was performed using Affinity Propagation (Frey and Dueck 2007) on the significant proteins, taking negative squared Euclidean distances as a measure of similarity. The goodness of the clusters was assessed from the Silhouette information according to the given clustering. Gene set enrichment analysis (GSEA) was done in R (R Core Team 2017), followed by a strategy of scoring similar (redundant) terms calculating the information content (IC) between two terms. Results were presented as a treemap or a scatterplot of terms clustered based on the first two components of a PCA of the IC similarity scores. For the embryo development, significant changes of protein abundance along the time course were assessed. FPKM levels for FlyBase 5.12 Transcripts from short poly(A)⁺ RNA-seq (Graveley et al. 2011) and localization data from <http://fly-fish.ccbcr.utoronto.ca> were integrated with our proteome data.

Data access

The mass spectrometry raw data from this study have been submitted to the ProteomeXchange (<http://www.proteomexchange.org>) under the data set identifiers PXD005691 (life cycle) and PXD005713 (embryogenesis).

Acknowledgments

We thank Junaid Akhtar for training, as well as Marion Scheibe, Anja Freiwald, and Christian Berger for critical reading of the manuscript. Antibodies were obtained from the Developmental Studies Hybridoma Bank, created by the NICHD of the NIH or kindly provided by Christian Berger. We acknowledge support of the Zentrum für Datenverarbeitung (ZDV) at the University of Mainz in hosting the web application. The study was partly funded by the Rhineland Palatinate Forschungsschwerpunkt GeneRED (Gene Regulation in Evolution and Development). D.K. is supported by the National Research Foundation Singapore and the Singapore Ministry of Education under its Research Centres of Excellence initiative.

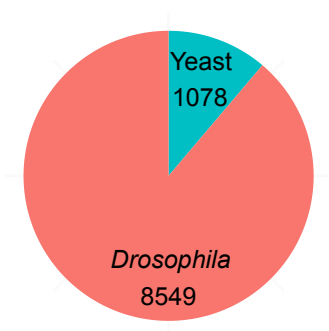
References

- Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287**: 2185–2195.
- Amos LA. 2008. The tektin family of microtubule-stabilizing proteins. *Genome Biol* **9**: 229.
- Aspden JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MAS, Brocard M, Couso J-P. 2014. Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *eLife* **3**: e03528.
- Beckstead RB, Lam G, Thummel CS. 2005. The genomic response to 20-hydroxyecdysone at the onset of *Drosophila* metamorphosis. *Genome Biol* **6**: R99.
- Bonaldi T, Straub T, Cox J, Kumar C, Becker PB, Mann M. 2008. Combined use of RNAi and quantitative proteomics to study gene function in *Drosophila*. *Mol Cell* **31**: 762–772.
- Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM, et al. 2014. Diversity and dynamics of the *Drosophila* transcriptome. *Nature* **512**: 393–399.
- Brunner E, Ahrens CH, Mohanty S, Baetschmann H, Loevenich S, Potthast F, Deutsch EW, Panse C, de Lichtenberg U, Rinner O, et al. 2007. A high-quality catalog of the *Drosophila melanogaster* proteome. *Nat Biotechnol* **25**: 576–583.
- Butter F, Bucerius F, Michel M, Cicova Z, Mann M, Janzen CJ. 2013. Comparative proteomics of two life cycle stages of stable isotope-labeled *Trypanosoma brucei* reveals novel components of the parasite's host adaptation machinery. *Mol Cell Proteomics* **12**: 172–179.
- Campos-Ortega JA, Hartenstein V. 1997. *The embryonic development of Drosophila melanogaster*, 2nd ed. Springer, Berlin.
- Chang YC, Tang HW, Liang SY, Pu TH, Meng TC, Khoo KH, Chen GC. 2013. Evaluation of *Drosophila* metabolic labeling strategies for in vivo quantitative proteomic analyses with applications to early pupa formation and amino acid starvation. *J Proteome Res* **3**: 2138–2150.
- Chao YC, Donahue KM, Pokrywka NJ, Stephenson EC. 1991. Sequence of swallow, a gene required for the localization of bicoid message in *Drosophila* eggs. *Dev Genet* **12**: 333–341.
- Chintapalli VR, Wang J, Dow JAT. 2007. Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. *Nat Genet* **39**: 715–720.
- Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**: 1367–1372.
- Cox J, Hein MY, Lubner CA, Paron I, Nagaraj N, 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.
- Dejung M, Subota I, Bucerius F, Dindar G, Freiwald A, Engstler M, Boshart M, Butter F, Janzen CJ. 2016. Quantitative proteomics uncovers novel factors involved in developmental differentiation of *Trypanosoma brucei*. *PLoS Pathog* **12**: e1005439.
- Dorus S, Busby SA, Gerike U, Shabanowitz J, Hunt DF, Karr TL. 2006. Genomic and functional evolution of the *Drosophila melanogaster* sperm proteome. *Nat Genet* **38**: 1440–1445.
- Edgar BA, Datar SA. 1996. Zygotic degradation of two maternal Cdc25 mRNAs terminates *Drosophila*'s early cell cycle program. *Genes Dev* **10**: 1966–1977.
- Fabre B, Korona D, Groen A, Vowinckel J, Gatto L, Deery MJ, Ralsler M, Russell S, Lilley KS. 2016. Analysis of *Drosophila melanogaster* proteome dynamics during embryonic development by a combination of label-free proteomics approaches. *Proteomics* **16**: 2068–2080.
- Fakhouri M, Elalayli M, Sherling D, Hall JD, Miller E, Sun X, Wells L, LeMosy EK. 2006. Minor proteins and enzymes of the *Drosophila* eggshell matrix. *Dev Biol* **293**: 127–141.
- Fournier ML, Paulson A, Pavelka N, Mosley AL, Gaudenz K, Bradford WD, Glynn E, Li H, Sardi ME, Fleharty B, et al. 2010. Delayed correlation of mRNA and protein expression in rapamycin-treated cells and a role for Ggc1 in cellular sensitivity to rapamycin. *Mol Cell Proteomics* **9**: 271–284.
- Frey BJ, Dueck D. 2007. Clustering by passing messages between data points. *Science* **315**: 972–976.
- Giniger E, Tietje K, Jan LY, Jan YN. 1994. Iola encodes a putative transcription factor required for axon growth and guidance in *Drosophila*. *Development* **120**: 1385–1398.
- Gonsalves SE, Neal SJ, Kehoe AS, Westwood JT. 2011. Genome-wide examination of the transcriptional response to ecdysteroids 20-hydroxyecdysone and ponasterone A in *Drosophila melanogaster*. *BMC Genomics* **12**: 475.
- Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, et al. 2011. The developmental transcriptome of *Drosophila melanogaster*. *Nature* **471**: 473–479.
- Griffin TJ, Gygi SP, Ideker T, Rist B, Eng J, Hood L, Aebersold R. 2002. Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. *Mol Cell Proteomics* **1**: 323–333.
- Grün D, Kirchner M, Thierfelder N, Stoeckius M, Selbach M, Rajewsky N. 2014. Conservation of mRNA and protein expression during development of *C. elegans*. *Cell Rep* **6**: 565–577.
- Harrison PM, Milburn D, Zhang Z, Bertone P, Gerstein M. 2003. Identification of pseudogenes in the *Drosophila melanogaster* genome. *Nucleic Acids Res* **31**: 1033–1037.
- Jambor H, Surendranath V, Kalinka AT, Meijstrik P, Saalfeld S, Tomancak P. 2015. Systematic imaging reveals features and changing localization of mRNAs in *Drosophila* development. *eLife* **4**: e05003.
- Kalinka AT, Varga KM, Gerrard DT, Preibisch S, Corcoran DL, Jarrells J, Ohler U, Bergman CM, Tomancak P. 2010. Gene expression divergence recapitulates the developmental hourglass model. *Nature* **468**: 811–814.
- Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, et al. 2011. Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* **471**: 480–485.
- Kronja I, Whitfield ZJ, Yuan B, Dzyek K, Kirkpatrick J, Krijgsvel J, Orr-Weaver TL. 2014. Quantitative proteomics reveals the dynamics of protein changes during *Drosophila* oocyte maturation and the oocyte-to-embryo transition. *Proc Natl Acad Sci* **111**: 16023–16028.
- Ladoukakis E, Pereira V, Magny EG, Eyre-Walker A, Couso JP. 2011. Hundreds of putatively functional small open reading frames in *Drosophila*. *Genome Biol* **12**: R118.
- Lawrence PA. 1992. *The making of a fly. The genetics of animal design*. Blackwell Scientific, Oxford.
- Lécuyer E, Yoshida H, Parthasarathy N, Alm C, Babak T, Cerovina T, Hughes TR, Tomancak P, Krause HM. 2007. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell* **131**: 174–187.
- Liu Y, Beyer A, Aebersold R. 2016. On the dependency of cellular protein levels on mRNA abundance. *Cell* **165**: 535–550.
- Luschnig S, Moussian B, Krauss J, Desjeux I, Perkovic J, Nüsslein-Volhard C. 2004. An F1 genetic screen for maternal-effect mutations affecting embryonic pattern formation in *Drosophila melanogaster*. *Genetics* **167**: 325–342.
- Magny EG, Pueyo JI, Pearl FMG, Cespedes MA, Niven JE, Bishop SA, Couso JP. 2013. Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science* **341**: 1116–1120.
- Mani SR, Megosh H, Lin H. 2014. PIWI proteins are essential for early *Drosophila* embryogenesis. *Dev Biol* **385**: 340–349.
- Margolin W. 2012. The price of tags in protein localization studies. *J Bacteriol* **194**: 6369–6371.
- Matthews BB, Dos Santos G, Crosby MA, Emmert DB, St Pierre SE, Gramates LS, Zhou P, Schroeder AJ, Falls K, Strelets V, et al. 2015. Gene model annotations for *Drosophila melanogaster*: impact of high-throughput data. *G3 (Bethesda)* **5**: 1721–1736.
- McGraw LA, Clark AG, Wolfner MF. 2008. Post-mating gene expression profiles of female *Drosophila melanogaster* in response to time and to four male accessory gland proteins. *Genetics* **179**: 1395–1408.

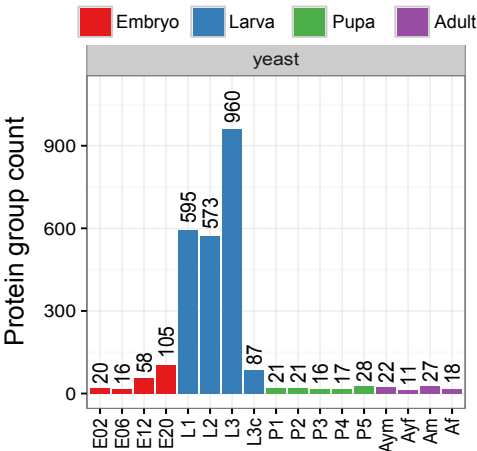
- The modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, et al. 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**: 1787–1797.
- Morin X, Daneman R, Zavortink M, Chia W. 2001. A protein trap strategy to detect GFP-tagged proteins expressed from their endogenous loci in *Drosophila*. *Proc Natl Acad Sci* **98**: 15050–15055.
- Nagarkar-Jaiswal S, Lee P-T, Campbell ME, Chen K, Anguiano-Zarate S, Gutierrez MC, Busby T, Lin W-W, He Y, Schulze KL, et al. 2015. A library of MiMICs allows tagging of genes and reversible, spatial and temporal knockdown of proteins in *Drosophila*. *eLife* **4**: e05338.
- Peshkin L, Wühr M, Pearl E, Haas W, Freeman RM, Gerhart JC, Klein AM, Horb M, Gygi SP, Kirschner MW. 2015. On the relationship of protein and mRNA dynamics in vertebrate embryonic development. *Dev Cell* **35**: 383–394.
- R Core Team. 2017. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Ramamurthi KS, Storz G. 2014. The small protein floodgates are opening; now the functional analysis begins. *BMC Biol* **12**: 96.
- Sandmann T, Jensen LJ, Jakobsen JS, Karzynski MM, Eichenlaub MP, Bork P, Furlong EEM. 2006. A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. *Dev Cell* **10**: 797–807.
- Sarov M, Barz C, Jambor H, Hein MY, Schmied C, Suchold D, Stender B, Janosch S, Kij VV, Krishnan RT, et al. 2016. A genome-wide resource for the analysis of protein localisation in *Drosophila*. *eLife* **5**: e12068.
- Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. 2011. Global quantification of mammalian gene expression control. *Nature* **473**: 337–342.
- Simmons MJ, Thorp MW, Buschette JT, Peterson K, Cross EW, Bjorklund EL. 2010. Maternal impairment of transposon regulation in *Drosophila melanogaster* by mutations in the genes aubergine, piwi and Suppressor of variegation 205. *Genet Res (Camb)* **92**: 261–272.
- Sowell RA, Hersberger KE, Kaufman TC, Clemmer DE. 2007. Examining the proteome of *Drosophila* across organism lifespan. *J Proteome Res* **6**: 3637–3647.
- Sury MD, Chen J-X, Selbach M. 2010. The SILAC fly allows for accurate protein quantification in vivo. *Mol Cell Proteomics* **9**: 2173–2183.
- Tadros W, Lipshitz HD. 2005. Setting the stage for development: mRNA translation and stability during oocyte maturation and egg activation in *Drosophila*. *Dev Dyn* **232**: 593–608.
- Takemori N, Yamamoto M-T. 2009. Proteome mapping of the *Drosophila melanogaster* male reproductive system. *Proteomics* **9**: 2484–2493.
- Tomančák P, Berman BP, Beaton A, Weizmann R, Kwan E, Hartenstein V, Celniker SE, Rubin GM. 2007. Global analysis of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol* **8**: R145.
- Vogel C, Marcotte EM. 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* **13**: 227–232.
- Wasbrough ER, Dorus S, Hester S, Howard-Murkin J, Lilley K, Wilkin E, Polpitiya A, Petritis K, Karr TL. 2010. The *Drosophila melanogaster* sperm proteome-II (DmSP-II). *J Proteomics* **73**: 2171–2185.
- Xing X, Zhang C, Li N, Zhai L, Zhu Y, Yang X, Xu P. 2014. Qualitative and quantitative analysis of the adult *Drosophila melanogaster* proteome. *Proteomics* **14**: 286–290.
- Yamanaka N, Rewitz KF, O'Connor MB. 2013. Ecdysone control of developmental transitions: lessons from *Drosophila* research. *Annu Rev Entomol* **58**: 497–516.

Received July 29, 2016; accepted in revised form March 30, 2017.

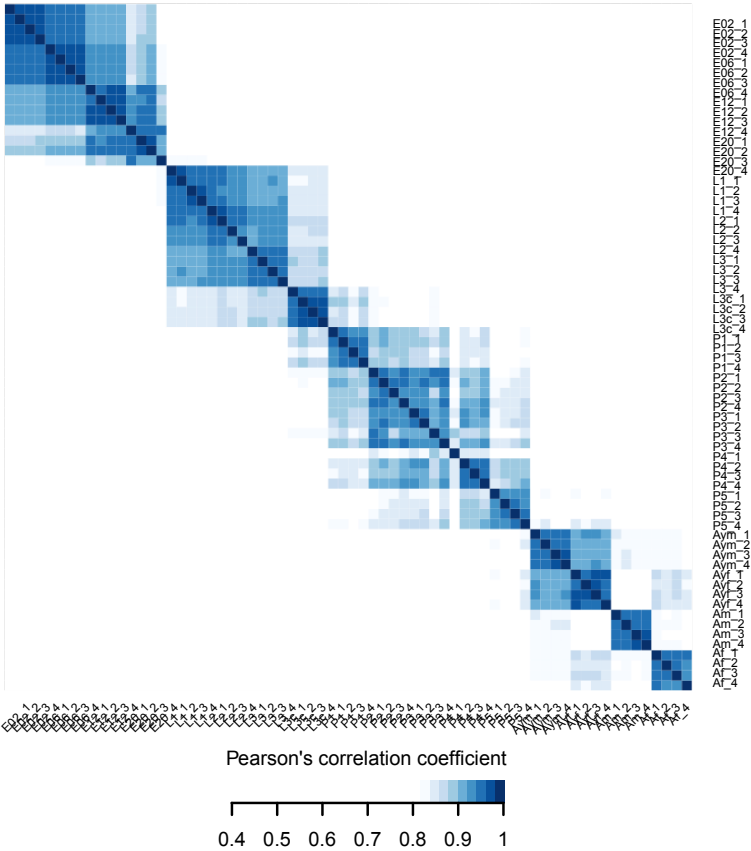
A



B

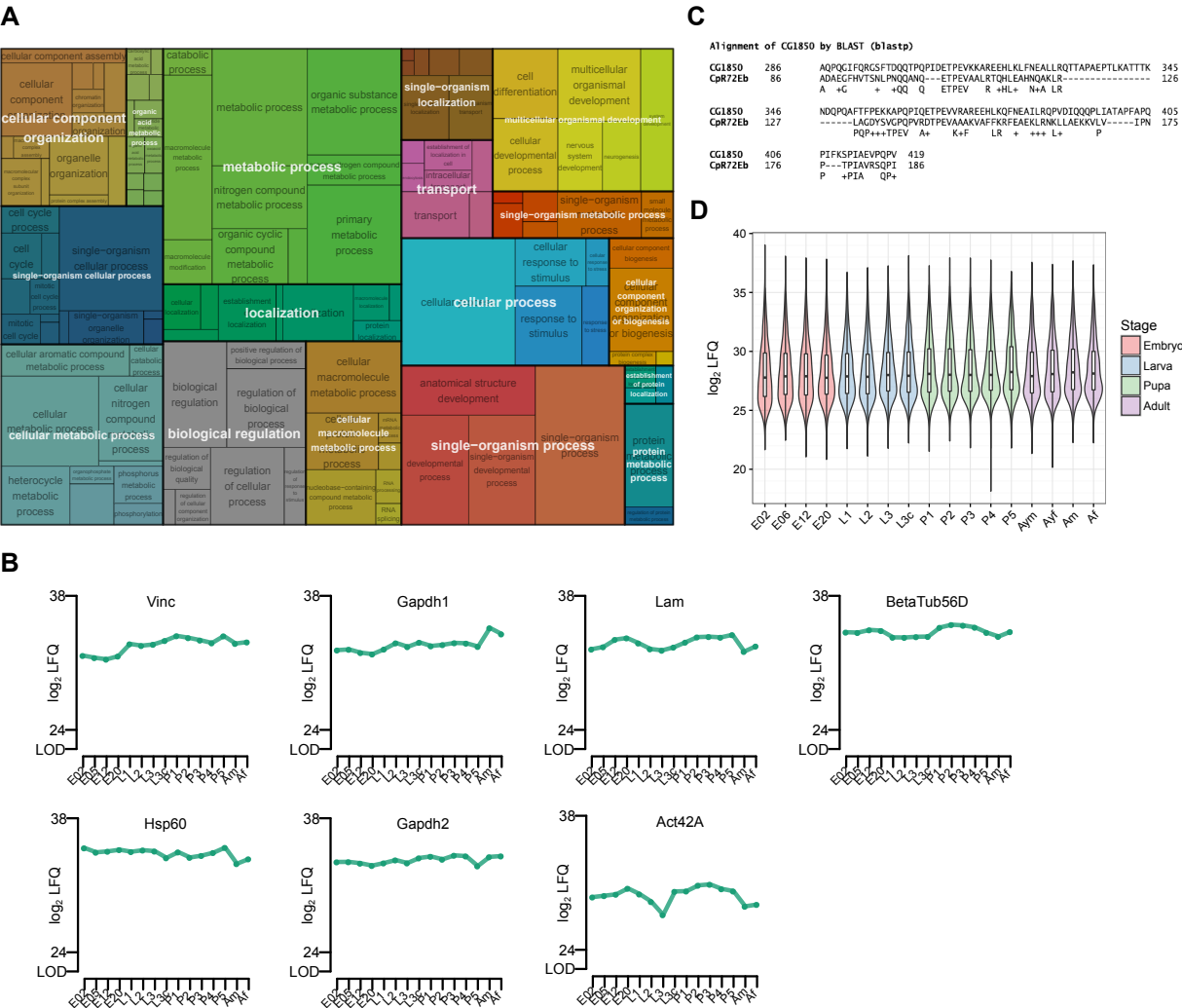


C



Supplemental Fig. S1:

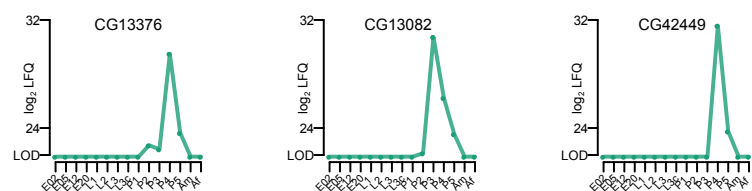
(A) Pie chart of detected yeast and *Drosophila* protein groups. Yeast proteins represent 12 percent (1078) of the identified proteins in the screen (9627). (B) Nearly all yeast proteins are detected in feeding L1, L2 and early L3 larval stages. (C) Correlation plot for the 17 samples of the life cycle experiment with all four replicates.



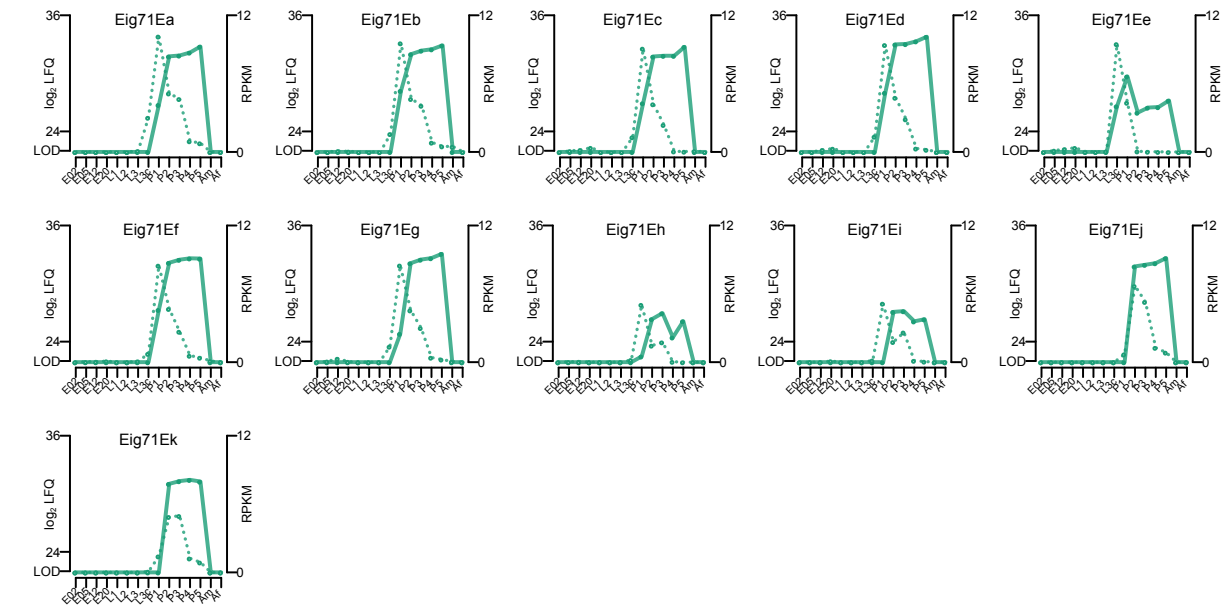
Supplemental Fig. S2:

(A) Treemap with overrepresented GO terms. The area is proportional to the size of the GO terms. (B) Alignment using BLAST between CpR72Eb and CG1850 shows a stretch of moderate sequence similarity between both proteins. (C) Overall protein expression values based on LFQ quantitation shows similar abundance levels and distribution across all conditions of the experiment. (D) Expression profiles of stably and dynamically expressed proteins annotated in Figure 2C and not displayed in Figure 2D.

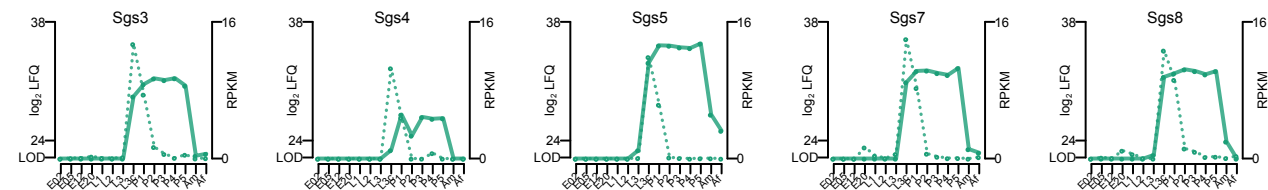
A



B

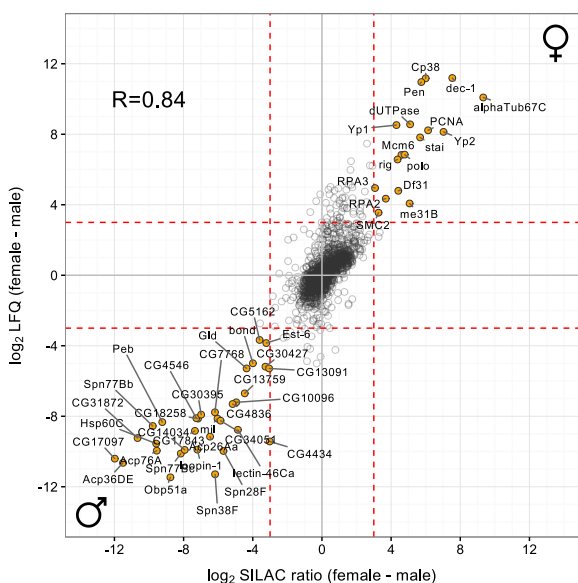
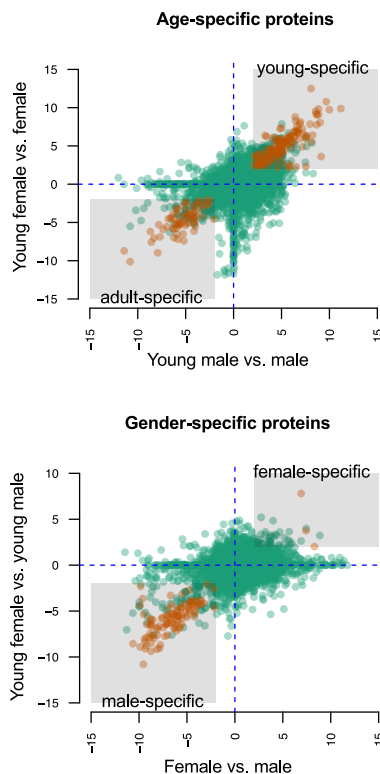
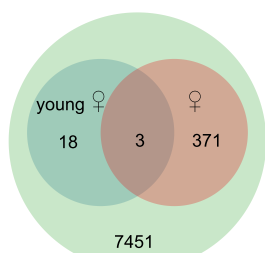
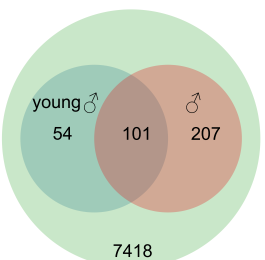
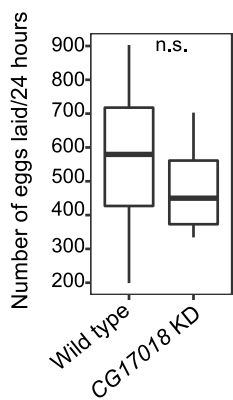
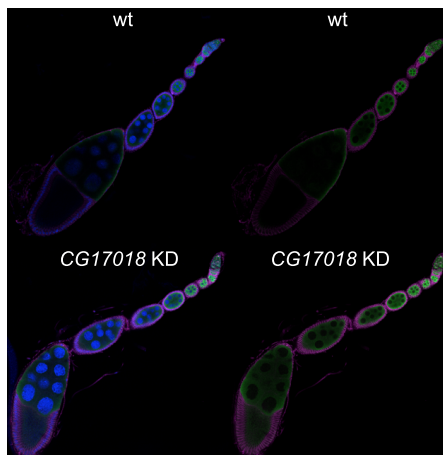


C



Supplemental Fig. S3:

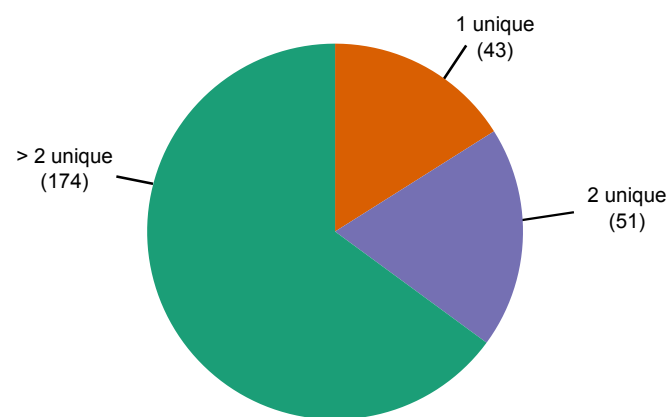
Expression profiles of uncharacterized proteins upregulating at a single pupal stage (A). RNA (dotted line) and protein (solid line) expression profiles of all quantified Eig72E family proteins (B) and Sgs proteins (C).

A**B****C****D****E****F**

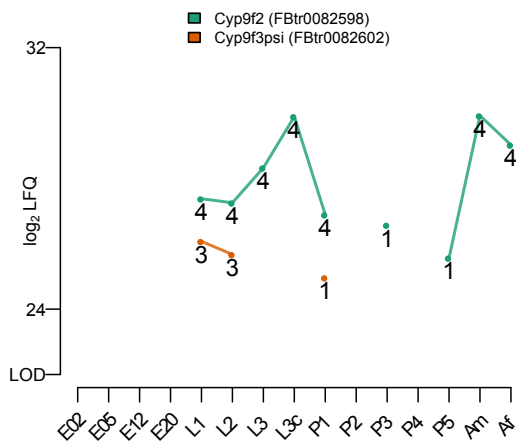
Supplemental Fig. S4:

(A) Comparison of w1118 SILAC data (Sury et al., 2010) to our Oregon R strain LFQ data. More than 8-fold enriched protein groups are highlighted (orange fill) and annotated. (B) Scatter plot showing age-specific and gender-specific proteins. (C) Venn diagram to compare female-specific proteins already present in the young fly. (D) Venn diagram to compare male-specific proteins already present in the young fly. (E) Number of laid eggs from wild type and the *CG17018* knockdown line. (F) Dissected ovaries from 5 day old female flies stained for DAPI (blue), Vasa (green) and 1B1 (purple).

A

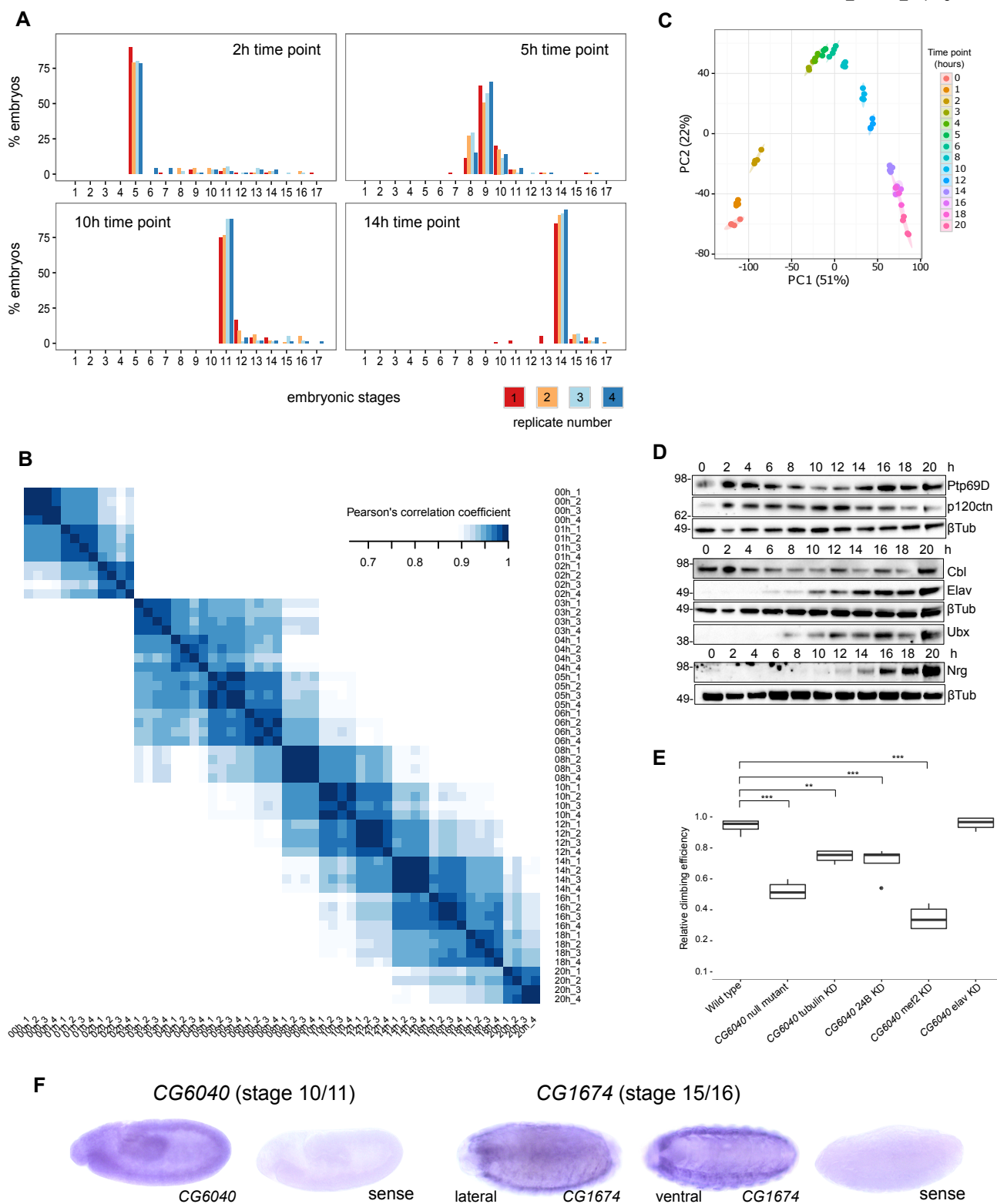


B



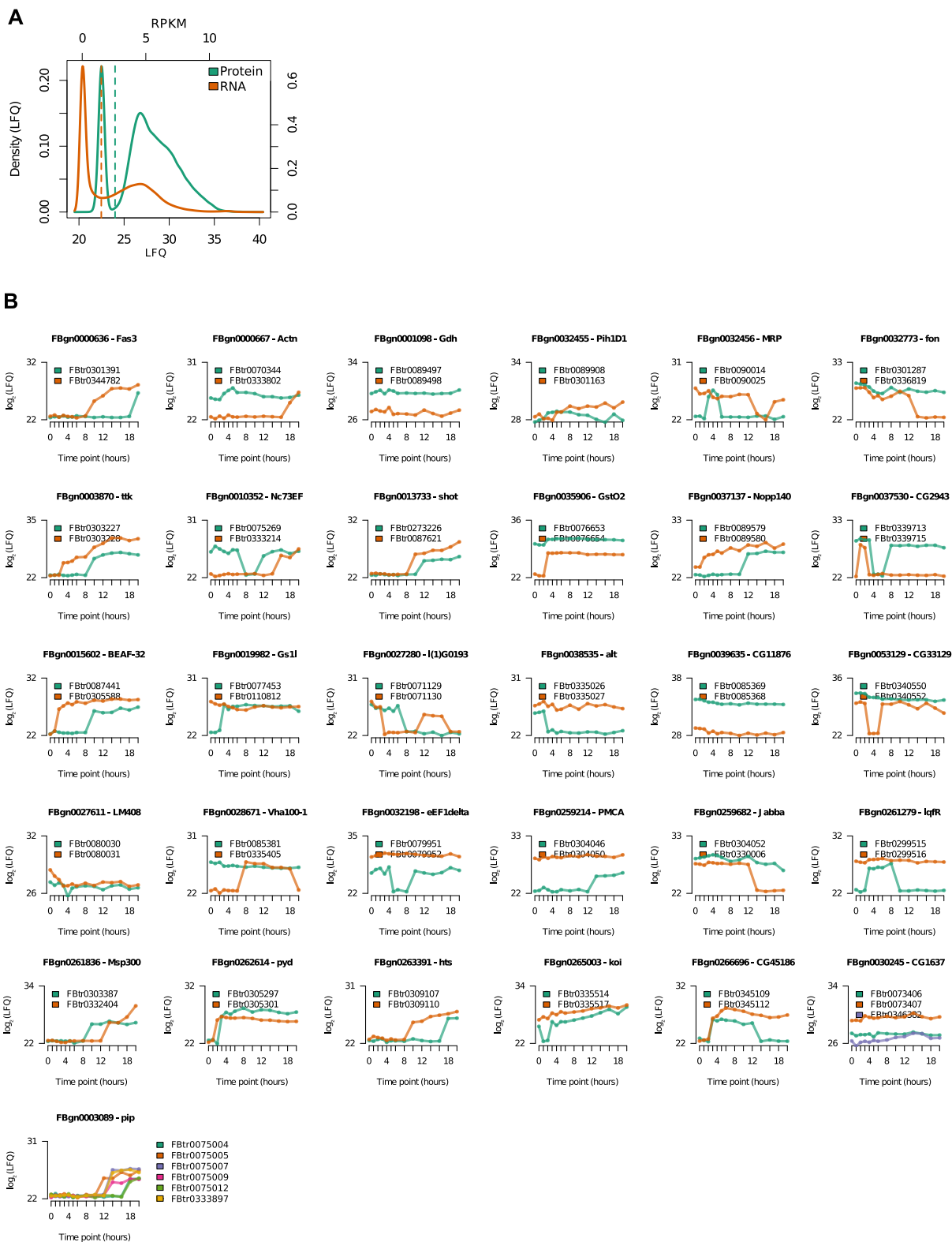
Supplemental Fig. S5:

(A) Pie chart showing the distribution of small proteins identified by the number of unique peptides. For a large majority (>80 percent) of the small proteome more than a single unique peptide was measured. (B) Developmental expression profile for the expressed previously annotated pseudogene Cyp9f3Psi in comparison to its paralog Cyp9f2.



Supplemental Fig. S6:

(A) Histogram showing the distribution of embryo stages in the four replicates of selected staged time points with a representative picture. (B) Correlation analysis of the 500 most variant proteins show high correlation between individual biological replicates. (C) PCA analysis with the first two components. Replicates are shown in identical colors and the standard error is represented by elliptic areas. (D) Western blot validation of chosen protein profiles in embryogenesis shown in Fig. 6. The respective tubulin loading controls are shown. (E) Ubiquitous (tubulin driver) and mesodermal (24B and mef2 driver), but not neuronal (elav) knockdown of *CG6040* results in reduced locomotion activity (Dunnett's test; ** p-value < 0.005, *** p-value < 0.001). (F) *in situ* hybridization of *CG6040* and *CG1674* showing expression in muscle tissue.



Supplemental material

The developmental proteome of *Drosophila melanogaster*

Nuria Casas-Vila, Alina Bluhm, Sergi Sayols, Nadja Dinges, Mario Dejung, Tina Altenhein, Dennis Kappei, Benjamin Altenhein, Jean-Yves Roignant and Falk Butter

Supplemental Methods

Collection of embryos, larvae, pupae and adult flies

After collection, all embryo samples were dechorionated using 7.5% hypochlorite for 2 min and rinsed with water. For each time point, approximately 20 µl embryo pellets were transferred to PBS buffer for lysis and mass spectrometry measurement. To assess the homogeneity of embryonic stages of each collection, approximately 10% of the sample was fixed and staged. Unused samples were snap-frozen in liquid nitrogen and stored at -80°C. Early larval collections were performed from agar apple juice plates in 2 hour laying time windows. Crawling larvae and pupae stages were collected from flasks at respective time points and rinsed with water.

Fixation and antibody staining for embryo staging

A small fraction (approx. 10%) of each embryo collection during the time course was fixed in fixation buffer (450 µl PBS, 600 µl heptane, 70 µl 37% formaldehyde) for 20 min while agitating. To remove the vitelline membrane, heptane was exchanged for methanol and embryos were vortexed for 2 minutes. Embryos were washed several times in methanol and finally snap-frozen in liquid nitrogen. Fixed embryos were rinsed three times in PBT (0.1% Triton-X-100 in PBS) for 10 min and incubated with the 4D9-engrailed/invec-ted-s antibody (1:7 in PBT, Developmental Studies Hybridoma Bank (DSHB)) at 4°C while slowly agitating. Embryos were washed three times in PBT for 10 min and incubated with the anti-mouse antibody conjugated with alkaline phosphatase (AP) (1:250 in PBT, Jackson ImmunoResearch) for 2 hours at RT. Three washes in PBT were followed by a 5 min wash in AP detection buffer (0.1 M NaCl, 0.05 M MgCl₂, 0.1 M Tris pH

9.5, 0.1% Tween20). The AP staining solution (150 µg/mL nitro blue tetrazolium (NBT) and 75 µg/mL 5-bromo-4-chloro-3-indolyl-phosphate (BCIP) in AP buffer) was added, embryos were transferred to a small dish and the color reaction was monitored using a binocular. To stop the AP reaction, embryos were rinsed three times in PBT and incubated for 10 min in methanol. After three PBT and three PBS washes, embryos were stored in 1 mL 70% glycerol at RT. Staging was done according to morphology and antibody staining.

Mass spectrometry sample preparation

For proteome analysis of the whole life cycle, snap-frozen samples were mechanically lysed in lysis buffer (140 mM NaCl, 10 mM Tris-HCl, pH 8.0, 0.5 mM EDTA and 1x protease inhibitor (Roche)) by bead milling using 0.5 mm diameter zircona/silica beads (Carl-Roth). Bead milling was done three times for 30 sec at 6800 rpm at 4°C in a tissue lyzer (Precellys). Homogenates were collected by centrifugation and proteins were acetone precipitated. Protein pellets were resuspended in 1x NuPAGE sample loading buffer (1x LDS) supplemented with 0.1 M DTT, boiled for 10 min at 95°C, sonicated for 10 min and proteins were separated on a 4-12% NuPAGE Bis/Tris gel for 10 min at 180 V in MOPS buffer.

For the embryonic time course analysis, embryos in PBS were homogenized with a microtube pestle, cells were pelleted at 1000 x g for 5 min at 4°C and resuspended in 1x LDS buffer complemented with 0.1 M DTT. Samples were boiled for 10 min at 80°C and proteins were separated on a 4-12% NuPAGE Bis/Tris gel for 10 min at 180 V in MOPS buffer. In-gel digestion and MS analysis was done as essentially described (Kappei et al. 2013).

Western blotting

100 embryos were homogenized in 50 µl of lysis buffer (140 mM NaCl, 10 mM Tris-HCl pH 8, 1 mM EDTA pH 8, 0.5% Triton X-100) using a microtube pestle (8-10 strokes). After 30 min incubation at 4°C on a rotation wheel, the lysate was centrifuged at maximum speed for 20 min and the supernatant transferred to a new tube. The previous two steps were repeated four times until the lysate was clear. Total embryonic protein extract was separated by SDS-PAGE. Western blot

analysis with affinity purified anti-Lola antibody (1:500, kindly provided by Edward Giniger) was performed by standard methods and visualized using ultra-sensitive enhanced chemiluminescent reagent (Thermo). Anti- β -Tubulin antibody (Covance catalog #MMS-410P, BioLegend) was used at a 1:2000 dilution. Additionally, Lola-RAA/RI depleted flies were generated as previously described (Kondo and Ueda 2013) using a gRNA (GTGTTGCACGTAAAGAAGCT) in exon 21 leading to a 2-bp deletion (Chr2R:10510060-10510061).

Embryo samples of selected time points (0h, 2h, 4h, 6h, 8h, 10h, 12h, 14h, 16h, 18h, 20h) prepared for mass spectrometry were used for western blot validation of protein expression profiles during embryogenesis. Proteins were separated by SDS-PAGE and blotted onto a nitrocellulose membrane, and subsequently blocked with 5% milk powder in TBST (0.1% Tween20 in 1xTBS). The following antibodies from DSHB were used in a 1:50 dilution (5% milk powder TBST): anti-p120ctn (p1B2), anti-Nrg (BP 102), anti-Ubx (kindly provided by Christian Berger), anti-PTP69D (3F11), anti-Cbl (8C4). Anti-Elav (7E8A10, DSHB) raised in rat was used 1:200 in 5% milk powder TBST.

***in situ* hybridization**

Primers were designed to amplify a unique region within respective coding sequences using a reverse primer containing the SP6 sequence. The PCR was performed on embryonic cDNA using Phusion DNA polymerase (NEB). Amplicon size comprised 861 bp for *lola*-RAA/RI, 1095 bp for *CG6040* and 1070 bp for *CG1674*. 250 ng of template PCR product was used to perform *in vitro* transcription using the SP6 RNA polymerase (Roche) and DIG labeled UTP (Roche). The reaction was incubated over night at 37°C and probes were carbonated to approximately 500 bp using carbonate buffer. The probes were then ethanol precipitated and resuspended in DEPC treated water to obtain a concentration of 100 ng/ μ l. Probes were diluted 1:50 in hybridization buffer for *in situ* hybridization. Embryos were fixed for 25 min in fixation solution (400 μ l PBS, 500 μ l n-heptane, 100 μ l 37% formaldehyde) while shaking at RT. After washing in methanol several times the embryos were snap-frozen in liquid nitrogen and

stored at -80°C. Embryos were gradually transferred into PBT (0.1% Tween20 in PBS), followed by three washes for 15 min, and finally into HB4 hybridization buffer (50 ml formamide, 25 ml 20xSSC buffer (3 M NaCl and 0.3 mM trisodium citrate-HCl pH 7.0), 200 µl Heparin (50 mg/ml), 100 µl Tween20, 500 mg Torula Yeast RNA extract). After equilibration at RT, embryos were pre-hybridized in HB4 at 56°C for several hours. Upon denaturation of the diluted RNA probe at 80°C for 10 min, embryos were hybridized over night at 65°C. Embryos were subsequently incubated in washing buffer (formamide:2xSSC (1:1) and 0.1% Tween20) for 30 min at 65°C and transferred into PBT at RT before they were incubated with anti-DIG-AP antibody (1:1000, Roche) for 2h at RT. Upon several washes in PBT and one rinse with AP buffer probes were visualized using the NBT/BCIP solution in AP buffer (1:100). Primers used in 5'-3' orientation:

<i>lola</i> -RAA/RI_fwd	AACCACAACAATTGCCACACATCATC
<i>lola</i> -RAA/RI_rev	GAGAATGGTGTAGCTCTTGCTC
<i>CG6040</i> _fwd	CCTTTGCCGCCTTAAACTGG
<i>CG6040</i> _rev	CGCTACCCAAGCTAATGCCG
<i>CG1674</i> _fwd	CACTAAAGCAGACCTTGTTTCG
<i>CG1674</i> _rev	TTTCGCACTGCTGTGAAG

Locomotion assay

Freshly hatched male and female flies of the respective genotype were separated and directly placed into measuring cylinders. The locomotion was assessed using the climbing assay described previously (Bahadorani and Hilliker 2008). Flies were tapped to the bottom and flies passing 8 cm in 10 or 5 seconds, respectively, were counted. Measurements were repeated four times in two independent biological replicates for each of the shRNA expressing-lines.

Fertility assay

20 young female flies were separated upon hatching and mated with 5 males for three days. Upon fertilization, flies were transferred onto fresh agar plates every 24 hours, eggs were counted and allowed to develop for further 36 hours. The hatching rate was determined by counting the number of unhatched eggs for each

lay. Counting was done on four subsequent days. Experiments were performed in three biological replicates with two different RNAi knockdown-lines.

Cuticle preparation

Cuticle preparations were prepared as previously described (Liu and Lasko 2015). Embryos were dechorionated for 2 min with 50% bleach, washed with water, transferred into an Eppendorf tube and washed with PBT (0.1% Tween20 in PBS). The supernatant was removed and Hoyer's medium (30 g gum Arabic, 200 g chloral hydrate, 20 g glycerol ad 50 ml water) was added to cover the embryos. The tube was incubated over night at 65°C, embryos were mounted onto a glass slide and examined under dark field illumination.

Immunohistochemistry

Ovaries were dissected from 5 days-old virgin females in cold PBS and fixed in 5% formaldehyde for 20 min. Samples were washed three times for 10 min with PBT (0.3% Triton X-100 in PBS) and blocked for 20 min in PBT + 5% donkey serum at RT. Samples were incubated with primary antibodies (rabbit anti-Vasa, 1:500; mouse anti-1B1, 1:100) over night at 4°C, washed three times for 10 min with PBT and incubated with secondary antibody for two hours at RT. Samples were mounted in Vectashield and examined using a confocal Leica SP5.

Extended bioinformatics analysis

Dynamicity of protein profiles

We used the Gini ratio (Damgaard and Weiner 2000; Gini 1912) to measure the stability of protein abundance throughout time. The Gini ratio calculates a score ranging from 0 to 1, which depicts the normalized mean difference of LFQ values between every possible combination of two stages for each protein:

$$G = \frac{\sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|}{2n^2\mu}$$

where “n” is the number of stages, “x_i” the protein quantification (LFQ) at stage “i”, and “μ” the average protein quantification throughout time. The minimum score refers to proteins that are stably expressed regardless of their average quantitation, while proteins having high abundance in only one stage present scores close to 1.

Significantly changing proteins throughout the life cycle

We used analysis of variance (ANOVA) in order to detect significant changes in protein expression in the life cycle data set. The resulting p-values were adjusted by Benjamini and Hochberg to control for the false discovery rate (FDR). We defined a protein as differentially expressed with a cut-off of 1% FDR. To identify in which stages proteins significantly change, we used the Tukey HSD post-hoc test. The test defines the Honest Significant Difference as the minimal distance between two groups to be considered statistically significant. To have a measure of the strength of the changes, we also calculated the effect size as suggested for one-way ANOVA analysis (Cohen 1988):

$$f = \sqrt{\frac{\sum_{i=1}^k p_i * (\mu_i - \mu)}{\sigma^2}}$$

Stage-specific proteins of the life cycle

To classify proteins into stages, we required that they are differentially expressed at 1% FDR and either detected in only one stage (embryo, larva, pupa or adult) or showing high differences in abundance (LFQ fold change > 4) in at least two other stages.

Getting the significant changes during the embryo development

We used the *timecourse* (Tai 2007) and *qvalue* (Storey et al. 2015) packages to assess significant changes in protein expression during time. The *timecourse* package implements a multivariate empirical Bayes method to calculate

moderated T^2 -statistics from longitudinal data, taking into account replicate information and correlation among gene expression along time points.

The significance of the T^2 -statistics was empirically estimated using the *qvalue* package, by comparing the obtained T^2 -statistics with the ones obtained from bootstrapping the original data. We performed a 1000-times bootstrap of each protein, permuting with replacement the values and calculating the statistic again. Then, we calculated the empirical p-values comparing the statistics from the original data and the pool of null statistics (bootstrapped data). To control the false discovery rate we use the *qvalue* function of the same package. Based on the distribution of our q-values, we set the significance cut-off at an FDR adjusted p-value of 0.0001.

Clustering strategies

We use Affinity Propagation (AP) (Bodenhofer et al. 2011; Frey and Dueck 2007) to cluster the differentially expressed proteins with similar expression profiles into an optimal number of clusters. AP is a well-established method to automatically calculate the best number of fitting clusters to the data. This method takes the data points as potential “exemplars” of clusters and further passing messages between points to decide which are [the best] exemplars and to which exemplar the rest belong to.

We calculated a similarity matrix between protein profiles using the negative Euclidean distance, defined as the negative squared distance between two points:

$$s = -d^r$$

Then we called AP with this similarity matrix, without setting any preference for any protein to be the exemplar of a cluster (default settings are the median of non-infinite values in the similarity matrix), and allowing up to 1000 iterations (i.e. rounds of messages passed between data points).

To reduce the number of clusters of the life cycle data set, we run AP with the preference parameter set at the 10% quantile of the distribution of similarities. For the embryo data set, the default parameters were used instead.

The goodness of the clustering was assessed using silhouette plots.

GO analysis

Using the R packages *GSEABase* (Morgan et al. 2016), *GOstats* (Falcon and Gentleman 2007) and *org.Dm.eg.db* (Carlson 2016, date stamp from the source of: 2015-Sep27) we performed gene set enrichment analyses of GO terms (biological processes only).

Due to the hierarchical nature of the GO annotation, usually many terms appear from the same set of genes. To remove redundancy of terms we scored the similarity between terms using the *GOSemSim* package (Yu et al. 2010). This package implements several methods to calculate the functional similarity of different terms. We used the Relevance method (Schlicker et al. 2006), based on the Information Content of two terms and their closest common ancestor.

Embryogenesis data set: For each cluster of the embryo data set we performed a hypergeometric test to find GO enriched categories. To assign GO terms from the clusters back to the different time points, we require for each time point that at least $\frac{3}{4}$ of the proteins comprised in each cluster were at least 2-fold enriched relative to their minimum LFQ value. To calculate the time point-specific GO terms, only terms not ubiquitously enriched in all time points were kept. Terms of each cluster were sorted by FDR and similarity scores were calculated. Eventually, only the term with the lowest p-value among similar terms (similarity score higher than 0.7 in a range between 0 and 1) was kept.

Life cycle data set: For each stage of the life cycle, we performed a hypergeometric test with the stage-specific proteins.

We additionally performed GO terms analysis on the core proteome (defined as the fraction of proteins detected in all time points) and also scored the similarity between terms. The terms were summarized in a scatter plot and colored based on their similarity score: 1) we performed hierarchical clustering of the distances between terms, defined as 1-score; 2) we cut the tree into subtrees, grouping together terms with similarity higher than 0.7 (range 0-1, same threshold used

before); 3) the dimensionality of the similarity matrix was reduced to 2 dimensions using classical multidimensional scaling, which were used as xy coordinates to distribute the terms in a scatter plot colored based on the cluster assignment; 4) as the cluster representative GO term, the most enriched term (FDR) of each cluster was selected, with the size of the circle representing the number of genes a term contains.

Alternatively, we summarized the GO terms of the core proteome as a Treemap, which is a way of displaying hierarchical data using nested proportional rectangles. In our case, we colored (grouped) the terms based on the subtrees of the previously calculated hierarchical clustering of the similarity matrix. The cluster representative and the size of the rectangle were assigned as described above for the scatter plot.

Integration of RNA-seq data

RNA quantification comes from the Supplementary table S10 of Graveley et al., 2011, titled *FPKM levels for FlyBase 5.12 Transcripts from short poly(A)+ RNA-Seq*. We used this table to estimate the gene expression naively summing the FPKM values of the different transcripts quantified. In order to control the variance, the FPKM values were log-transformed.

We obtained from Ensembl 84 (Yates et al. 2016) the corresponding FlybaseName Gene ids to the Flybase Transcript ids, which were later used to merge the RNA and protein quantitation.

The plot comparing the similarity between RNA and protein expression at each time point is a false color image matrix with the pairwise Pearson correlation coefficients of all time points between the RNA and protein quantitation.

Translational delay and identification of RNA to protein translation patterns

As part of the integration of RNA and protein data, we grouped together genes that have similar RNA and protein profiles. To do this, we calculated a PCA for the protein and RNA data sets in order to accumulate the maximal variability in one dimension. The first component of the two reduced data sets was then used to cluster the genes by a k-means clustering (max. 1000 iterations, 100 starting

random centers). Based on Silhouette plots we chose 6 clusters that explained best our data.

Integration of the Fly-FISH data

We downloaded the localization data for embryos from <http://fly-fish.ccb.utoronto.ca/terms/> and complemented them with further annotation from the Ensembl release 84 (ensembl_gene_id, flybasename_gene) using Biomart. Proteins were classified into tissues based on FISH data at not only embryo stages but also taking into account RNA expression in later developmental stages. Genes were automatically clustered based on the LFQ data available, using the Affinity Propagation method described before with negative squared Euclidean distances. The average expression trend of each cluster was then calculated by lowess smoothing (Cleveland 1979, 1981). Eventually, we calculated the GO terms enriched per cluster using the R packages *GSEABase*, *GOstats* and *org.Dm.eg.db* as previously described.

Supplemental Tables

Supplemental Table 1. Filtered MaxQuant output table with calculated and imputed LFQ values.

Supplemental Table 2. GO enrichment information on the life cycle proteome.

Supplemental Table 3. Calculated dynamicity score table.

Supplemental Table 4. Clusters obtained for the complete life cycle experiment.

Supplemental Table 5. GO enrichment information on the life cycle clusters.

Supplemental Table 6. One week old male and female protein comparison.

Supplemental Table 7. Young male and female protein comparison.

Supplemental Table 8. Overview of maternally loaded RNA and proteins.

Supplemental Table 9. Filtered MaxQuant output table obtained by including ncRNA sequences.

Supplemental Table 10. Filtered MaxQuant output table obtained by including pseudogene sequences.

Supplemental Table 11. Filtered MaxQuant output table with calculated and imputed LFQ values for the embryogenesis time course.

Supplemental Table 12. Automatically generated 70 clusters for the embryogenesis proteome.

Supplemental Table 13. GO enrichment information on the embryogenesis time-course analysis clusters.

Supplemental Table 14. Tissue-specific cluster information.

Supplemental Table 15. GO enrichment information on the tissue-specific clusters.

Supplemental Table 16. Six RNA and protein clusters based on the first PCA component.

Supplemental References

Badoahrani S, Hilliker AJ. 2008. Antioxidants cannot suppress the lethal phenotype of a *Drosophila melanogaster* model of Huntington's disease. *Genome* **51**: 392–395.

Bodenhofer U, Kothmeier A, Hochreiter S. 2011. APCluster: an R package for affinity propagation clustering. *Bioinformatics* **27**: 2463–2464.

Carlson M. 2016. *org.Dm.eg.db: Genome wide annotation for Fly*.

- Cleveland WS. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**: 829–836.
- Cleveland WS. 1981. Lowess - A program for smoothing scatterplots by robust locally weighted regression. *American Statistician* **35**: 54.
- Cohen J. 1988. *Statistical power analysis for the behavioral sciences* (2nd ed.) L. Erlbaum Associates, Hillsdale, New Jersey
- Damgaard C, Weiner J. 2000. Describing inequality in plant size or fecundity. *Ecology* **81**: 1139–1142.
- Falcon S, Gentleman R. 2007. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**: 257–258.
- Frey BJ, Dueck D. 2007. Clustering by passing messages between data points, *Science* **315**: 972–976.
- Gini C. 1912. Variabilità e mutabilità, contribuito allo studio delle distribuzioni e delle relazioni statistiche. Fascicolo Ier: Introduzione - Indici di variabilità - Indici di mutabilità. Bologna.
- Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW et al. 2011. The developmental transcriptome of *Drosophila melanogaster*. *Nature* **471**: 473–479.
- Kappei D, Butter F, Benda C, Scheibe M, Draskovic I, Stevense M, Novo CL, Basquin C, Araki M, Araki K, et al. 2013. HOT1 is a mammalian direct telomere repeat-binding protein contributing to telomerase recruitment. *EMBO J* **32**: 1681–1701.
- Kondo S, Ueda R. 2013. Highly improved gene targeting by germline-specific Cas9 expression in *Drosophila*. *Genetics* **195**: 715–721.
- Liu N, Lasko P. 2015. Analysis of RNA interference lines identifies new functions of maternally-expressed genes involved in embryonic patterning in *Drosophila melanogaster*. *G3 (Bethesda)* **5**: 1025–1034.
- Morgan M, Falcon S, Gentleman R. 2016. *GSEABase: Gene set enrichment data structures and methods*.

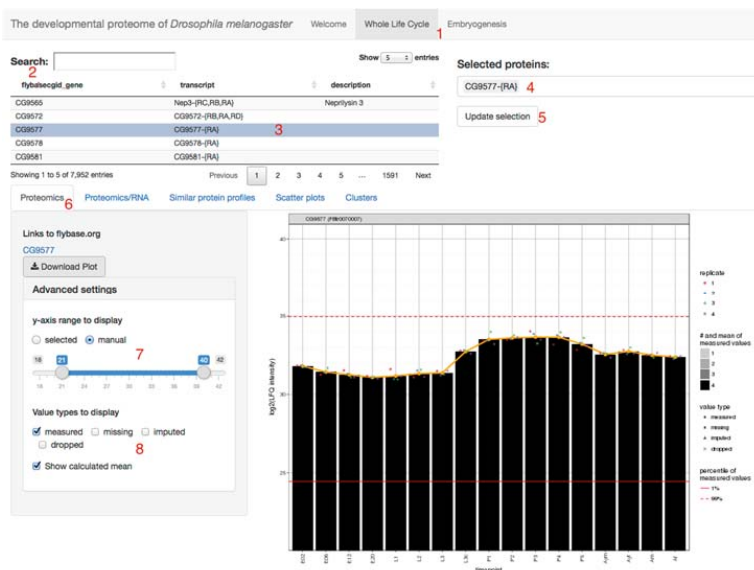
- Schlicker A, Domingues FS, Rahnenfuehrer J, Lengauer T. 2006. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* **7**: 302.
- Storey D, Bass AJ, Dabney A, Robinson D. 2015. *Q-value estimation for false discovery rate control*.
- Sury MD, Chen JX, Selbach M. 2010. The SILAC fly allows for accurate protein quantification in vivo. *Mol Cell Proteomics* **9**: 2173-2183.
- Tai YC. 2007. *Timecourse: Statistical Analysis for Developmental Microarray Time Course Data*.
- Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L et al. 2016. Ensembl 2016. *Nucleic Acids Res.* **44**: D710-D716.
- Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. 2010. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products, *Bioinformatics* **26**: 976–978.

How-To for the web interface

This is a short manual on how to navigate the web interface accompanied with the manuscript "The developmental proteome of *Drosophila melanogaster*".

Contents

- 1 Data sets
 - 1.1 Whole life cycle
 - 1.2 Embryogenesis
- 2 Web interface structure
 - 2.1 Search
 - 2.2 Selected proteins
 - 2.3 Data analysis
 - 2.3.1 Proteomics
 - 2.3.2 Proteomics/RNA
 - 2.3.3 Similar protein profiles
 - 2.3.4 Scatter plots
 - 2.3.5 Clusters
- 3 Advanced settings



1 Data sets

In the top menu [1], you can select the data set to work with: “Whole Life Cycle” or “Embryogenesis”.

1.1 Whole life cycle

The data set contains the following 15 time points (measured in quadruplicates):

- E02: 0-2 hours old embryos
- E06: 4-6 hours old embryos
- E12: 10-12 hours old embryos
- E20: 18-20 hours old embryos
- L1: 40-44 hours after egg laying
- L2: 66-68 hours after egg laying
- eL3: 83-85 hours after egg laying
- L3c: crawling larvae
- P1: white pupae
- P2-P5: every 24 hours after pupation (white pupae)
- Aym/Ayf: adult young male/female flies 4 hours after hatching
- Am/Af: 1 week old adult male/female flies

1.2 Embryogenesis

Embryos were collected in quadruplicates with a 30 min laying window for the following time points: 0h, 1h, 2h, 3h, 4h, 5h, 6h, 8h, 10h, 12h, 14h, 16h, 18h, 20h.

2 Web interface structure

2.1 Search

After selecting one of the data sets, a table with all detected proteins will appear. The protein of interest can be queried in the “Search” bar [2] and selected from the

table by clicking [3]. After selecting one entry, the protein will be saved in the “update selection” window on the upper right site.

2.2 Selected proteins

This feature allows keeping information about previous searches. For a completely new search, please delete the entry in the upper right window [4] and click the update selection button [5].

2.3 Data analysis

After selecting the protein of interest, the data can be chosen from a second menu [6] with the following tabs:

- Proteomics
- Proteomics/RNA
- Similar protein profiles
- Scatter plots
- Clusters

All the graphical representations are navigable by hovering over the dots/lines. Additionally, it is possible to download any graphical representation via the “download” button.

2.3.1 Proteomics

Bar plots representing protein abundance (y-axis) throughout the collected time points of the selected data set (x-axis). Colored dots indicate the different replicates. The grey scale of the bars represents the number of replicates in which the protein was measured (grey=1, black=4). The value type legend specifies the origin of the data (for detailed information, please check supplemental methods). In each plot two thresholds are presented: 1% (continuous line) and 99% (dashed line) of the LFQ intensities found in the complete data set. The orange line

indicates the tendency of protein expression based on measured and imputed values.

2.3.2 Proteomics/RNA

Protein and RNA profiles of the selected protein are shown. The RNA data is retrieved from the modENCODE project (Graveley et al., 2011). All transcripts for each protein are color-coded and by hovering over the lines, their gene names are displayed.

2.3.3 Similar protein profiles

This feature allows searching for proteins with similar expression profiles to the protein of interest. The 100 most similar proteins are listed in the table on the left (scored by similarity) and can be searched in the search tab. By selecting proteins, expression profiles of protein and RNA will be added to the graphical output in a different color.

2.3.4 Scatter plots

Different scatter plots from the manuscript are available for interactive searches (see left menu for selection). The selected proteins from the search (if available in the scatter plot) are highlighted.

2.3.5 Clusters

Significantly changing whole life cycle protein profiles were clustered and are shown by stage (embryo, larva, pupa, adult). The cluster of interest can be selected and respective protein profiles will be visualized. The clusters that contain information about the selected proteins will be displayed (lower right).

Embryogenesis data was complemented with tissue-specific expression data. Therefore, a second tab “tissue” allows the selection of tissue and cluster number showing the respective protein expression profiles.

3 Advanced settings

This tab allows modifying different aspects of the graphical output:

- y-axis: the ordinate range can be defined [7]
- Value types [8]:
 - **measured:** calculated LFQ intensities
 - **missing:** no LFQ intensity calculated, but measured intensity
 - **imputed:** calculated by imputation (see supplemental methods)
 - **dropped:** a single measured replicate value without intensities in neighboring time points was dropped and replaced by an imputed value
- Show/hide calculated mean (orange line)

Phylointeractomics reconstructs functional evolution of protein binding

Summary

Here, we combined a MS-based interactomics approach with phylogenomics to study proteins binding to the vertebrate telomeric TTAGGG sequence. Biotinylated oligonucleotides were immobilized on streptavidin beads and incubated with nuclear extracts from 16 vertebrate cell lines covering 450 million years of evolution. Bound proteins were quantified by label-free MS. In non-therian species, we observed loss of binding of the shelterin complex member TRF1 (=TERF1). DNA pulldowns using the recombinantly expressed DNA-binding domain of TRF1 from different species validated the altered binding behavior. Moreover, we report previously unknown telomere-associated proteins that were specifically enriched at telomeric repeats in at least five species. These proteins present interesting candidates for future functional studies at telomeres.

Overall, phylointeractomics is a powerful tool that can be used to study evolutionary changes in protein binding to any bait of interest.

Zusammenfassung

Mittels Kombination aus quantitativer Interaktomik und Phylogenomik haben wir in diesem Projekt Telomer-bindende Proteine identifiziert. Biotinylierte Oligonukleotide wurden an Streptavidin-Beads immobilisiert und mit nuklearem Extrakt von 16 verschiedenen Vertebraten-Zelllinien, die 450 Jahre Evolution abdecken, inkubiert. Spezifisch gebundene Proteine wurden mittels quantitativer Massenspektrometrie identifiziert. Dabei haben wir herausgefunden, dass das Shelterin Komplexmitglied TRF1 (=TERF1) von Vögeln, Amphibien und Fischen nicht mehr an die Telomersequenz bindet. Dieser Unterschied im Bindeverhalten wurde validiert, indem die TRF1 DNA-Bindedomäne verschiedener Spezies in DNA Pulldowns auf Binden getestet wurde. Zusätzlich dazu haben wir Proteine, die bisher nicht mit Telomeren in Zusammenhang gebracht wurden, in mindestens fünf der 16

analysierten Spezies als spezifische TTAGGG-Binder identifiziert. Diese Proteine stellen interessante Kandidaten für weitere funktionelle Studien an Telomeren dar.

Das Prinzip der Phylointeraktomik kann genutzt werden, um evolutionäre Unterschiede im Proteinbindeverhalten an eine beliebige Matrix zu untersuchen.

Statement of Contribution

I cultured cells and prepared nuclear extracts of several species: opossum, guinea pig, zebra finch, zebrafish, clawed frog, medaka, duck, and chicken. In addition, I performed telomere pulldown assays for opossum, zebra finch, rabbit, guinea pig, clawed frog, and duck and prepared samples for MS measurements. As part of the revision process, I conducted telomere pulldowns using HeLa S3 and chicken nuclear extracts. Moreover, I determined telomerase activity in the 16 vertebrate cell lines by telomere repeat amplification protocol (TRAP) assays. I critically read and commented the manuscript.

Supervisor confirmation

ARTICLE

Received 8 Jan 2016 | Accepted 16 Dec 2016 | Published 8 Feb 2017

DOI: 10.1038/ncomms14334

OPEN

Phylointeractomics reconstructs functional evolution of protein binding

Dennis Kappel^{1,2,*}, Marion Scheibe^{3,4,*}, Maciej Paszkowski-Rogacz², Alina Bluhm³, Toni Ingolf Gossmann⁵, Sabrina Dietz³, Mario Dejung³, Holger Herlyn⁶, Frank Buchholz^{2,7,8,9,10}, Matthias Mann⁴ & Falk Butter³

Molecular phylogenomics investigates evolutionary relationships based on genomic data. However, despite genomic sequence conservation, changes in protein interactions can occur relatively rapidly and may cause strong functional diversification. To investigate such functional evolution, we here combine phylogenomics with interaction proteomics. We develop this concept by investigating the molecular evolution of the shelterin complex, which protects telomeres, across 16 vertebrate species from zebrafish to humans covering 450 million years of evolution. Our phylointeractomics screen discovers previously unknown telomere-associated proteins and reveals how homologous proteins undergo functional evolution. For instance, we show that TERF1 evolved as a telomere-binding protein in the common stem lineage of marsupial and placental mammals. Phylointeractomics is a versatile and scalable approach to investigate evolutionary changes in protein function and thus can provide experimental evidence for phylogenomic relationships.

¹Cancer Science Institute of Singapore, National University of Singapore, 14 Medical Drive, Singapore 117599, Singapore. ²Medical Systems Biology, UCC, University Hospital and Medical Faculty Carl Gustav Carus, TU Dresden, Fetscherstraße 74, Dresden D-01307, Germany. ³Institute of Molecular Biology (IMB) gGmbH, Ackermannweg 4, Mainz D-55128, Germany. ⁴Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Am Klopferspitz 18, Martinsried D-82152, Germany. ⁵Department of Animal & Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, UK. ⁶Institute of Anthropology, University of Mainz, Anselm-Franz-von-Bentzel-Weg 7, Mainz D-55099, Germany. ⁷Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstraße 108, Dresden D-01307, Germany. ⁸German Cancer Research Center (DKFZ), Neuenheimer Feld 280, 69120 Heidelberg, Germany. ⁹German Cancer Consortium (DKTK) partner site, Fetscherstr. 74, 01307 Dresden Germany. ¹⁰National Center for Tumor Diseases (NCT), University Hospital Carl Gustav Carus, TU Dresden, Dresden D-01307, Germany. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to F.Buchholz (email: frank.buchholz@tu-dresden.de) or to M.M. (email: mmann@biochem.mpg.de) or to F.Butter (email: f.butter@imb-mainz.de).

The analysis of evolutionary relationships of gene sequences was strongly advanced in recent years because of the advent of high-throughput DNA sequencing technologies^{1,2}. However, while overall gene sequence might be conserved, even single amino-acid exchanges can change the functionality of the corresponding proteins and thus drives evolution in regulatory networks^{3,4}. Mass spectrometry-based interaction proteomics has been the technique of choice to identify protein interactions in a systematic manner⁵.

To systematically assess functional evolutionary changes in protein binding, we developed the concept of phylointeractomics combining phylogenomics with interaction proteomics. We demonstrate the power of our approach by investigation of telomere-binding proteins across 16 vertebrate species, sharing the common telomeric repeat motif TTAGGG⁶. These repeats are bound by the shelterin complex, which protects the linear chromosome ends from recognition as DNA double-strand breaks and is composed of six subunits (TERF1, TERF2, TIN2, TPP1, RAP1 and POT1) in human. It is generally assumed that throughout 450 million years of vertebrate evolution⁷ this complex is conserved^{8,9}. However, experimental validation of this phylogenomic assumption is lacking.

Here we show that TERF1 actually evolved its intrinsic telomere-binding ability in the common stem lineage of marsupial and placental mammals. This observation exemplifies that the assumption to equate phylogenomic homology and functional conservation has restrictions, and with phylointeractomics, we provide a versatile and scalable approach to uncover these functional differences.

Results

Interaction proteomics recapitulates the core telosome. In our phylointeractomics screen to identify telomere-binding proteins, we used a DNA-protein interaction approach combined with quantitative mass spectrometry. Polymerized biotinylated DNA of either telomeric sequence (TTAGGG) or a scrambled control sequence (GTGAGT) was immobilized on paramagnetic streptavidin beads. Both sequences were separately incubated with nuclear protein extracts from each of the 16 species (Fig. 1a), and bound proteins were analysed by label-free quantitative mass spectrometry (Fig. 1b). We discovered significant interactors (identical cutoff values of $S_0 = 0.6$ and $P < 0.05$, see Methods section) with the telomeric TTAGGG sequence by quadruplicate pull-down experiments for all 16 vertebrates. Each pull-down experiment quantified on average 1,300 proteins. This allowed us to determine significant interactors by their enrichment rather than the presence or absence on target and control sequence, making the analysis much more robust¹⁰. To further focus on high confidence interactors in the core telosome that potentially have a role in a larger number of organisms, we focussed on those candidates that were identified in at least five species (Figs 1c and 2a, Supplementary Data 1 and 2).

The shelterin complex (TERF1, TERF2, TIN2, TPP1, RAP1 and POT1) involved in telomere end protection and telomere homeostasis clustered tightly together with high enrichment scores in mammals, providing a positive control (Figs 1c and 2a, Supplementary Data 1 and 2). Furthermore, we observed the recent gene duplication of POT1 into POT1A and POT1B in the rodent lineage¹¹ when using extracts from mouse and rat and we identified both TERF2 paralogues in medaka. This demonstrates the required comprehensiveness and sensitivity to detect specific molecular evolution events (Fig. 1c).

Phylointeractomics identifies putative novel telomere binders. Our screen recapitulated the TERF2-interaction partner

DCLRE1B, a nuclease implicated in proper end processing^{12,13}, and the nuclear receptors NR2C2 and NR2C1, previously described as subtelomere-binding proteins^{14–16}. So far, only the homeobox-domain-containing proteins TERF1, TERF2 and HOT1 have been demonstrated to directly bind to double-stranded telomeric DNA while both POT1 and the CST complex are single-stranded telomere binders^{17–19}. In addition to the established double-strand binders, we identified eight zinc finger proteins (ZBTB7A, ZBTB10, ZBTB48, ZNF276, ZNF524, ZNF827, VEZF1 and KLF12) enriched at the TTAGGG repeat sequence in at least five species. We speculate that some of these candidates bind double-stranded TTAGGG repeats directly and have functional roles at the telomere. For instance, human ZNF827 has, in the meantime, been reported to localize to telomeres, to induce telomere remodeling and to promote telomere–telomere recombination²⁰. Consistently, with our observed enrichment of RECQL1 to TTAGGG repeats in 8 species, this helicase is involved in telomere maintenance, actively resolving telomeric D-loops and Holliday junction substrates, regulated via an interaction with TERF2 in human cells²¹. The recently discovered direct telomere-binding protein HOT1 was also enriched on telomeric DNA; consistent with a role of this protein in active telomere elongation, HOT1 is identified predominantly in those species with detectable telomerase activity (Fig. 2a, Supplementary Fig. 2) even in normal somatic tissue²² as used in this screen, which is in agreement to its previously described differential binding behaviour in human cells, in which it associates with telomeres in cellular contexts with active telomere elongation¹⁷. Indeed, when performing the same analysis with nuclear protein extracts from telomerase-positive HeLa cells we readily detect HOT1 among the specific telomere binders, but it is not identified from the telomerase-negative IMR90 human fibroblasts used in our initial screen (Fig. 2a, Supplementary Fig. 1a, Supplementary Fig. 2). RUNX1, RUNX2, CBFB and, in fewer cases, also RUNX3 (Fig. 2a, Supplementary Data 1 and 2) as well as three poly r(C) binding proteins PCBP1, PCBP2 and PCBP3 were consistently enriched at telomeric DNA in various vertebrates, including humans. CBFB had been previously detected by proteomics of isolated chromatin segments (PICH)¹⁵ and PCBP1 as shelterin-associated by immunoprecipitation/mass spectrometry (IP/MS)²³. RUNX proteins are transcription factors and RUNX1 regulates the differentiation of hematopoietic stem cells, whereas the PCBP proteins are generally thought to be RNA-binding proteins regulating several cancer relevant transcripts^{24,25}. Of note, two other candidates, FSBP and NAIF1, contain a homeobox/myb domain, typical for double-strand telomere binders^{17,26}, and SRBD1 features an OB-fold as found in POT1 (ref. 18), suggesting that these candidates may bind directly to TTAGGG and regulate and maintain telomeres. Thus our screen resulted in numerous proteins already implicated in telomere regulation and provides evidence for several novel candidates.

The affinity purification assay²⁷ used in this study may have putative limitations that need to be carefully considered. The assay primarily identifies direct binding proteins to a particular DNA sequence, together with their tight interaction partners. This is highlighted by the fact that we have identified the entire shelterin complex together with DCLRE1B but none of the more transient shelterin interactors¹⁸ (Fig. 2a, Supplementary Fig. 1, Supplementary Data 1 and 2). Thus, to address research questions that involve temporal and spatial resolution at telomeres other interactomics approaches such as PICH¹⁵ or a quantitative telomeric chromatin isolation protocol (QTIP)²⁸, an adaptation of the more general concept of combining chromatin immunoprecipitation with MS^{5,29,30}, have been developed. However, quantitative telomeric chromatin isolation

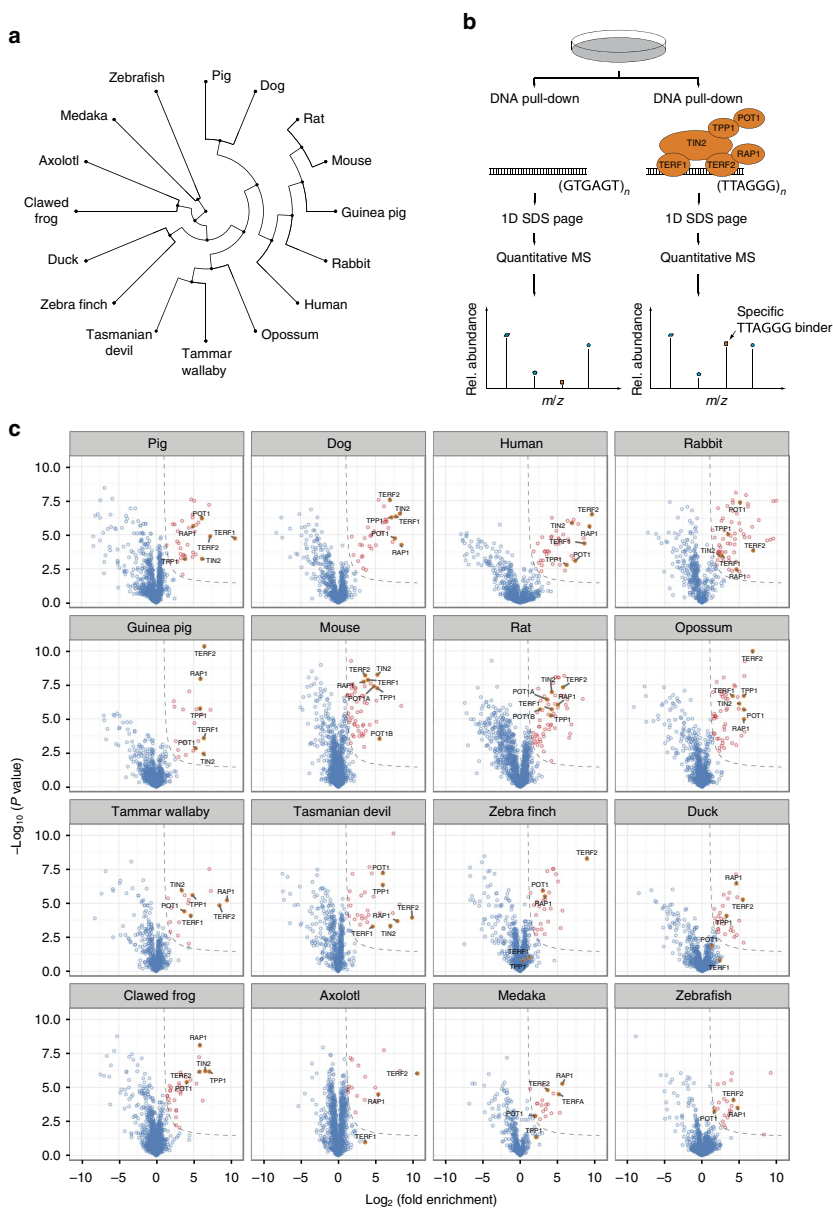


Figure 1 | Phylointeractomics screen for telomere binders in 16 vertebrate species. (a) Phylogenetic tree of vertebrate species analysed in this study—all higher ranked vertebrate taxa apart from monotremata are represented. (b) Quantitative label-free DNA interaction screen with DNA oligonucleotides containing either the telomeric repeat sequence (TTAGGG) or a control sequence (GTGAGT). Specific interaction partners are differentiated from background binders by a ratio significantly different from 1:1. All pull-downs were performed in biological replicates ($n = 4$). (c) Volcano plots for all tested vertebrate species. Specifically enriched proteins (red circles) are distinguished from background binders (blue circles) by a two-dimensional cutoff with $S_0 = 0.6$ and $P < 0.05$ (Welch's t -test). Detected members of the shelterin complex (TERF1, TERF2, TIN2, TPP1, RAP1 and POT1) are highlighted (filled orange dots).

protocol is specifically targeted to telomeres and dependent on the behaviour of TERF1 and TERF2, while PICH is currently limited to repetitive elements that provide multiple binding sites for sufficient enrichment of endogenous chromatin. These technical biases are likely the reason why other studies aiming

at identifying novel telomeric factors are showing surprisingly little overlap between each other beyond the shelterin proteins (Supplementary Fig. 3). Interestingly, when compared with other screens, our approach shows the strongest overlap with PICH, which uses telomeric DNA as the bait as well. Affinity purification

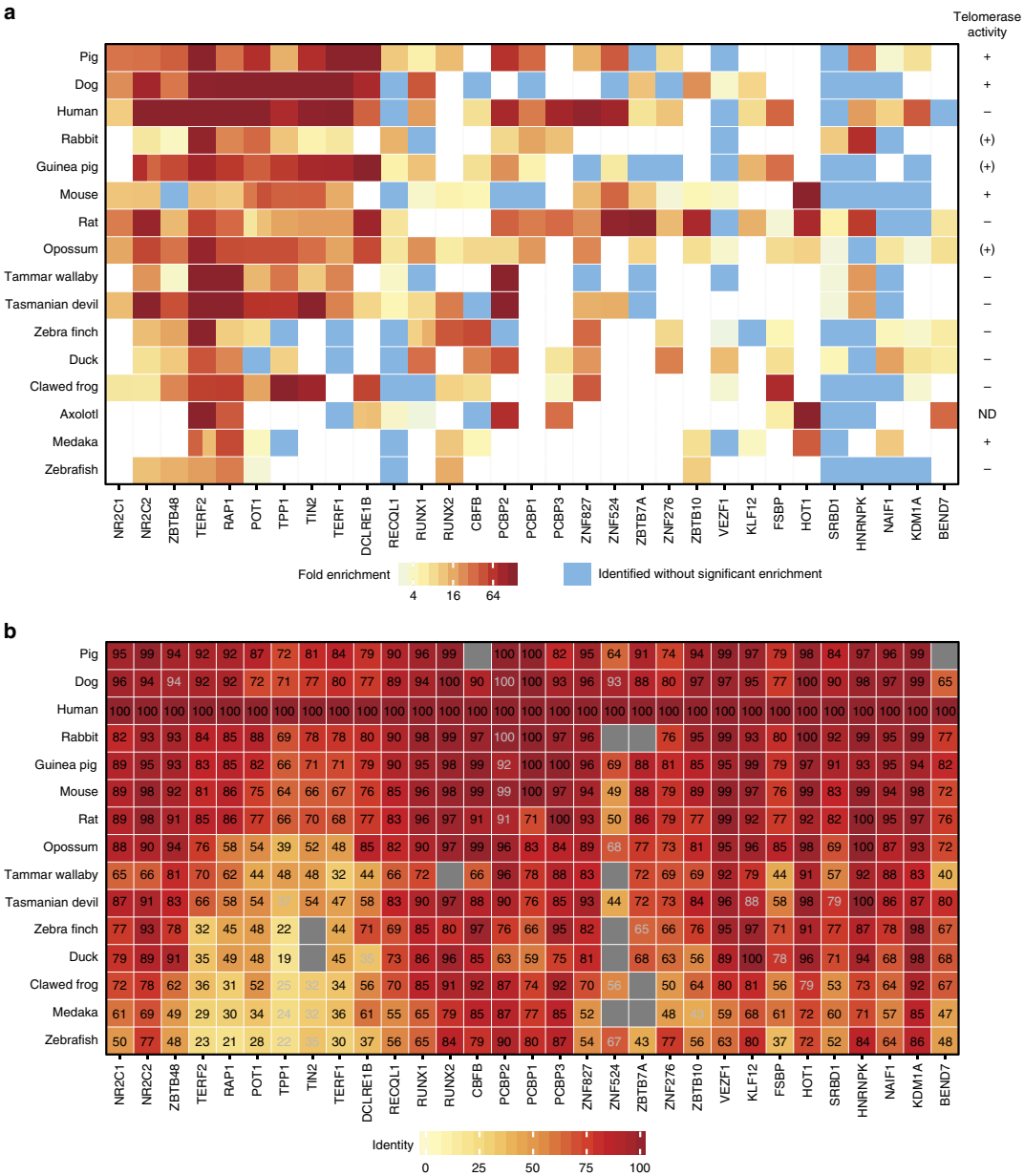


Figure 2 | Telomere-binding proteins are identified by cross-species validation. (a) Heat map of identified proteins with specific enrichment at TTAGGG repeats in at least 5 out of the 16 vertebrate species. Species (rows) are arranged according to their phylogenetic relations while proteins (columns) are clustered based on their binding pattern across all pull-downs. Colour gradient represents relative enrichment from TTAGGG binding (red) to equal enrichment on the telomeric and control sequence (blue). Only events that were identified as hits according to the criteria in Fig. 1c are shown. Gene names are based on the human versions, and in the occurrence of paralogues, both enrichment values are displayed side by side in the species concerned (for example, POT1 in mouse). The presence of telomerase activity is indicated on the right. (+) indicates cell lines with activity at the detection limit. IMR90 (human)⁴⁷ and NIH3T3 cells (mouse)⁴⁸ are known telomerase-negative and -positive cells, respectively. Quantifications can be found in Supplementary Fig. 2. (b) Heat map of protein sequence identity for all candidate telomere binders from **a** across 15 vertebrate species. Axolotl was excluded from this analysis as there is currently no published whole-genome annotation available. The protein sequence identity is displayed as a colour gradient relative to the human sequences with the percentage value displayed in each square. Grey squares represent absent genes based on ENSEMBL genome assemblies. Grey values represent non-annotated homologues based on ENSEMBL genome assemblies for which we could identify homologues by reciprocal BLAST search.

procedures such as the widely used immunoprecipitations and peptide pull-downs are prone to reveal all putative interaction candidates. Therefore, a careful validation of localization, binding and functional relevance in an endogenous context is required for any novel factor. Nevertheless, the presence of several already established telomeric factors in our screen suggests that our candidates may be important for telomere biology pending further in-depth characterization for each protein.

Constitutive TERF1 telomere binding is a therian invention.

Importantly, our phylointeractomics approach also provides information about the evolutionary history of binding patterns of individual proteins, which here led to unexpected findings. Although the shelterin complex is thought to be universal to vertebrates^{8,9} and indeed TERF2 and RAP1 are strongly enriched on TTAGGG repeats in all species analysed, TERF1 was absent or not specifically enriched in non-therian vertebrates despite the presence of a TERF1 orthologue in all 16 species (Figs 1c and 2a,b). Our data therefore raised the possibility that only therian (placental and marsupial) TERF1 can bind to telomeric TTAGGG repeats. To investigate this hypothesis, we recombinantly expressed the TERF1 DNA-binding domain (DBD) of 13 vertebrate species spanning from ray-finned fish to human and performed DNA-binding assays. Although all 8 recombinantly expressed TERF1 homeobox domains of therian mammals clearly showed binding to TTAGGG repeats, no binding was detectable for any TERF1 DBD outside this group (Fig. 3b). Our data therefore places a gain of direct TTAGGG repeat binding of TERF1 in the therian lineage after separation from monotremata around 225 million years ago.

Specific TERF1 residues were positively selected in therians.

We next performed domain-specific multiple sequence alignment and conducted substitution rate analysis to infer selective pressures between therians and other species sampled. Using a branch-site model, selective pressures across sites were inferred from posterior distributions using a Naive Empirical Bayes (NEB) approach (Fig. 3a,c). We pinpointed seven amino-acid positions (A371, W374, Q378, K379, S382, Q393, S397 based on the opossum sequence) in the TERF1 DBD that are positively selected in therians but that evolve neutrally across non-therian branches. W374 and S382 are identical between platypus and therians and therefore cannot explain the observed binding pattern. Of the remaining five residues, only the Q393 residue confers direct contact with DNA based on the co-crystal structure of the human TERF1 homeobox domain with telomeric DNA²⁶. Two additional residues, T400 and E420, differ between platypus and theria but are not under evolutionary constraints based on our branch-site model (Fig. 3a).

To test whether the identified amino-acid residues are important for telomeric binding, we mutated all residues that differ between platypus and therians individually and performed DNA-binding assays with the recombinant expressed opossum TERF1 DBD variants. In this screen, A371M, T400Q and E420D did not show any obvious change in their binding behaviour while Q378M, K379Q and Q393T reduced TERF1 binding. Exchange of S397 against the bulkier phenylalanine residue completely abrogated binding to telomeric DNA (Fig. 3d). Based on this information, we attempted to recapitulate a gain-of-function-binding switch for the platypus TERF1 DBD. To further quantify the change in binding activity, we used purified recombinant TERF1 domains and tested multiple platypus variants along with the platypus and opossum wild-type DBDs. Although the exchange of the bulky phenylalanine in F313S was not sufficient to induce binding, the combination with other

substitutions (T309Q and M294Q/Q295K) transferred TTAGGG binding to the TERF1 DBD from platypus (Fig. 3e,f, Supplementary Fig. 4), demonstrating that these are indeed the key residues for the evolutionary switch of TERF1 telomere binding.

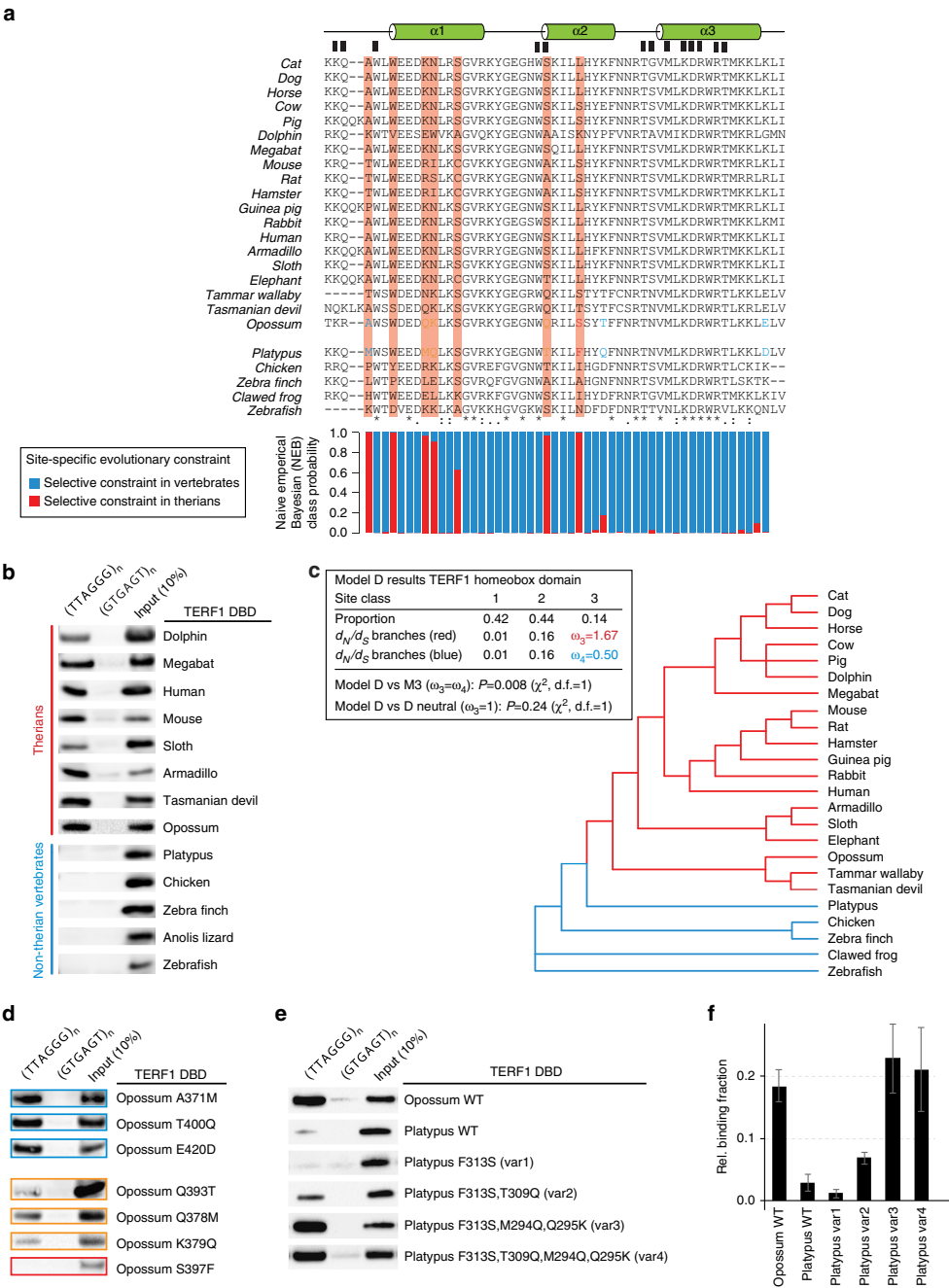
Discussion

Although TERF1 has not been extensively studied in non-therian vertebrates, available data from clawed frog and chicken suggests that TERF1 binding to telomeres and its telomeric function might indeed not be conserved. Using *in vitro* transcribed/translated *Xenopus laevis* TERF1, binding to chromatin or a plasmid containing TTAGGG repeats could only be observed when added to mitotic but not to interphase extracts with the mitotic phenotype appearing rather faint compared with that of TERF2 (ref. 31). Although direct binding to TTAGGG was not assessed, the fact that a specific extract has to be present to achieve any enrichment suggests that the interaction is likely indirect. For chicken TERF1 (cTERF1), Okamoto *et al.*³² have reported as data not shown that cTERF1 can localize to telomeres in chicken cells. However, FLAG-cTERF1 does not localize to telomeres in mouse cells but telomeric localization can be enforced by fusing the chicken TRFH domain of TERF1 to the mouse TERF1 DBD³². Similarly, Cooley *et al.*³³ reported the localization of myc-cTERF1 to telomeres in DT40 *Terf1*^{-/-} cells³³. As the result by Okamoto *et al.*³² could suggest indirect recruitment of cTERF1 to telomeres in an endogenous context, we performed a TTAGGG pull-down using nuclear protein extracts from 6C2 chicken cells (Supplementary Fig. 1b). In contrast to the results obtained with zebra finch and duck, cTERF1 is specifically enriched on telomeric DNA in our proteomics analysis (Supplementary Fig. 1b). Given that the cTERF1-DBD does not directly bind to telomeric DNA *in vitro* (Fig. 3b), non-therian TERF1 seems to be passively recruited to telomeres in some species and/or specific cellular contexts (for example, chicken, mitotic cells in *xenopus*). Importantly, our findings suggest that among the investigated species only therian TERF1 DBDs have an intrinsic TTAGGG-binding ability. A difference in cellular function is also supported by the observation that DT40 *Terf1*^{-/-} cells do not display any defective cell viability contrary to mouse *Terf1* knockout cells³⁴ or major telomere dysfunctions³³. These data functionally underscore our finding and indeed Cooley *et al.*³³ suggest as one possible explanation that the shelterin complex composition in chicken may differ from mammals.

Our results raise interesting questions about the evolution of the shelterin complex in vertebrates. Although TERF1 directly binds to telomeric dsDNA in therians, it is also connected to the shelterin complex via direct interaction with TIN2, mediated by the TRFH domain of TERF1 and the FxLxP motif in the C-terminus of TIN2 (ref. 35). Furthermore, TIN2 connects the remaining shelterin members through direct interactions with both TERF2-RAP1 and TPP1-POT1. Similar to TERF1, TIN2 and TPP1 are absent or not enriched in our pull-downs from several non-mammalian vertebrates (Figs 1c and 2a). Absence of TIN2 in interaction screens with bird proteomes turned out to be due to a lack of an annotated *TIN2* gene in this lineage in current genome assemblies (Fig. 2b). Without the central TIN2 hub, the six protein complex does not exist in birds. In the remaining non-mammalian vertebrates, the complex stability seems also impaired by the lack of TERF1 or at least by the lack of direct TERF1-TTAGGG binding (Fig. 1c). In addition, branch-site modeling of the TERF1 TRFH domain, which is important for dimerization³⁶, reveals the opposite conservation pattern compared with the homeobox domain as several residues are under purifying selection in non-therian vertebrates and evolve

neutrally otherwise (Supplementary Fig. 5). For both the homeobox and the TRFH domain in TERF2, which together with its interactor RAP1 is consistently bound to TTAGGG in all the 16 vertebrates in our screen, we could not detect such evolutionary differences (Supplementary Table 1 and 2). Together our data suggest that, after gene duplication of the ancestral TERF gene, TERF2 retained telomere-binding activity, whereas TERF1

evolved otherwise. Although it is possible that convergent evolution of TERF1 telomere binding is yet to be discovered in specific lineages of non-therian vertebrates, our data clearly illustrate that TERF1 regained the ability to directly bind TTAGGG repeats in therians where the six protein shelterin complex is found in its previously described composition and function¹⁸. Beyond the specific example of telomeres, these data



also highlight that the reasonable assumption to equate genetic and functional homology requires careful experimental testing and that proteins and protein complexes evolve dynamically in function and composition.

In conclusion, our approach combines the proliferation of sequenced genomes with the increasing power of interactomics screens to investigate functional and evolutionary important protein binding. Current synthesis and expression technologies allow for an easy production of baits, such as DNA sequences, RNA structures, (modified) peptides, protein domains or full-length proteins, while streamlined interaction screens can be performed in a large number of systems, given access to their proteomes. Thus phylointeractomics is capable to investigate the molecular evolution of domain-specific binding across any species and could serve as a blueprint for a future analysis of how full-length proteins evolve. It can provide experimental evidence for phylogenomics relationships and helps to extrapolate the results obtained in model organisms to a broader group of species.

Methods

Cell culture. IMR90 (human), HeLa (human), NIH3T3 (mouse), PC12 (rat), Vx-2 (rabbit), LLC-PK1 (pig), MDCK (dog), CCL-141 (duck) and tasmanian devil cells were cultivated in 4.5 g l^{-1} glucose, 4 mM glutamine, 1 mM sodium pyruvate Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% fetal bovine serum (FBS; Sigma), 100 U ml^{-1} penicillin and $100 \mu\text{g ml}^{-1}$ streptomycin (Gibco), which is referred to as full DMEM hereafter. 104C1 cells (guinea pig) were cultured in RPMI-1640 medium supplemented with 10% FBS, 2 mM glutamine, 100 U ml^{-1} penicillin and $100 \mu\text{g ml}^{-1}$ streptomycin. Medium for the cultivation of OK (opossum) cells consisted of full DMEM complemented with 1% non-essential amino acids (PAA). Medium for culturing ZFTMA (zebra finch) cells was composed of full DMEM supplemented with 2% chicken serum. Chicken 6C2 cells were cultivated in DMEM with 10% FBS, 100 U ml^{-1} penicillin and $100 \mu\text{g ml}^{-1}$ streptomycin (Gibco), 1 mM sodium pyruvate, 1% non-essential amino acids and 0.4% chicken serum. Culture conditions of all above listed cell lines were 37°C and 5% CO_2 . Leibovitz's (L-15) medium (67%) supplemented with 10% FBS, 100 U ml^{-1} penicillin and $100 \mu\text{g ml}^{-1}$ streptomycin defined medium composition for the cultivation of speedy (frog) cells at 28°C and 0% CO_2 . BRFA1 (zebrafish) cells were grown in L-15 medium, including 15% FBS, 100 U ml^{-1} penicillin and $100 \mu\text{g ml}^{-1}$ streptomycin at 33°C and 0% CO_2 . Medium for growing OLF-136 (medaka) cells was composed of DMEM supplemented with 15% FBS, 100 U ml^{-1} penicillin and $100 \mu\text{g ml}^{-1}$ streptomycin. Medaka cells were cultivated at 28°C and 5% CO_2 .

Nuclear protein extraction. Cells were harvested and nuclear extracts were prepared as previously described³⁷. For extraction of nuclear extracts, cells were harvested and incubated in hypotonic buffer (10 mM Hepes, pH 7.9, 1.5 mM MgCl_2 , 10 mM KCl) on ice for 10 min. Cells were transferred to a dounce homogenizer in hypotonic buffer supplemented with 0.1% Igepal CA630 (Sigma) and 0.5 mM DTT and lysed by 40 strokes. Nuclei were washed once in $1 \times$ PBS and extracted in hypertonic buffer (420 mM NaCl, 20 mM Hepes, pH 7.9, 20% glycerol, 2 mM MgCl_2 , 0.2 mM EDTA, 0.1% Igepal CA630 (Sigma), 0.5 mM DTT) for 2 h at 4°C on a rotating wheel.

Telomere pull-down. Forward and reverse sequence oligonucleotides ($25 \mu\text{g}$) (see Supplementary Table 3) were diluted in annealing buffer (20 mM Tris-HCl, pH 7.5, 10 mM MgCl_2 , 100 mM KCl), denatured at 95°C and annealed by cooling. Annealed double-stranded oligonucleotides were incubated with 100 units T4 kinase (Life Technologies) for 2 h at 37°C followed by incubation with 20 units T4 ligase overnight. Concatenated DNA strands were purified using phenol-chloroform extraction. Following biotinylation with desthiobiotin-dATP (Jena Bioscience) and 60 units DNA polymerase (Thermo), the biotinylated probes were purified using microspin G-50 columns (GE Healthcare). Telomeric or control DNA was immobilized on $500 \mu\text{g}$ paramagnetic streptavidin beads (Dynabeads MyOne C1, Life Technologies) on a rotation wheel for 30 min at room temperature. Subsequently, baits were incubated with 400 or $800 \mu\text{g}$ (frog) of nuclear extract in PBB buffer (150 mM NaCl, 50 mM Tris-HCl pH 7.5, 5 mM MgCl_2 , 0.5% Igepal CA-630 (Sigma)) while rotating for 1.5 h at 4°C . Sheared salmon sperm DNA ($10 \mu\text{g}$; Ambion) was added as a competitor for DNA binding. After three washes with PBB buffer, bound proteins were eluted in $1 \times$ LDS sample buffer supplemented with 0.1 M DTT, boiled for 10 min at 70°C and separated on a 10% NuPAGE Novex Bis-Tris precast gel (Life Technologies).

MS data acquisition. For in-gel digestion, samples were reduced in 10 mM DTT for 1 h at 56°C followed by alkylation with 55 mM iodoacetamide (Sigma) for 45 min in the dark. Tryptic digest was performed in 50 mM ammonium bicarbonate buffer with $1 \mu\text{g}$ trypsin (Promega) at 37°C overnight. Peptides were desalted on StageTips and analysed by nanoflow liquid chromatography on an EASY-nLC 1000 system (Thermo) coupled online to a Q Exactive Plus mass spectrometer (Thermo). Peptides were separated on a C18-reversed phase capillary (25 cm long, $75 \mu\text{m}$ inner diameter, packed in-house with ReproSil-Pur C18-AQ $1.9 \mu\text{m}$ resin (Dr Maisch)) directly mounted on the electrospray ion source. We used a 90 min gradient from 2% to 60% acetonitrile in 0.5% formic acid at a flow of 200 nl min^{-1} . The Q Exactive Plus was operated with a Top10 MS/MS spectra acquisition method per MS full scan.

MS data analysis. The raw files were processed with MaxQuant³⁸ (version 1.4.0.8) against the ENSEMBL annotated protein and the genescan databases of the respective species (duck: BGI1.0.74; dog: CanFam3.1.71; guinea pig: cavPor3.75; zebrafish: Zv9.71; human: GRCh37.71; wallaby: Meug.1.0.74; opossum: BROADO5.71; mouse: GRCh38.71; rabbit: oryCun2.71; medaka: MEDAKA1.7; rat: Rnor_5.0.71; Tasmanian devil: DEVIL7.0.74; pig: Sscrofa10.2.71; zebra finch: taegu3.2.4.71; xenopus: JGI.4.2.71; rabbit: oryCun 2.71) with the exception of axolotl where the Am2.0 database was used. Carbamidomethylation was set as fixed modification while methionine oxidation and protein N-acetylation were considered as variable modifications. The search was performed with an initial mass tolerance of 7 p.p.m. mass accuracy for the precursor ion and 20 p.p.m. for the MS/MS spectra in the HCD fragmentation mode. Search results were processed with MaxQuant and filtered with a false discovery rate of 0.01. The match between run option and the LFQ quantitation were activated.

Bioinformatic analysis. After peak detection and label-free quantitation was performed in MaxQuant, the files were further analysed using self-developed R and Python scripts. In detail, protein groups marked as reverse, contaminants or only identified by site were removed. A further filtering step removed protein groups that were not identified in at least three out of the four replicates at either bait. The missing values were imputed for each sample individually. The values were calculated with a beta distribution using the fitdistr function (MASS R package) fitted to 5% of the smallest values. Afterwards, a BLAST search was performed to map the human homologues to the different species. For creating the volcano plots, a two-tailed Welch's *t*-test to calculate the *P* value for each protein was used.

Figure 3 | TERF1 acquired ability to bind telomeric DNA in the therian stem lineage. (a) Sequence alignment of the TERF1 DBD of several species. Residues involved in DNA binding are marked by a black rectangle. Asterisks (*) indicate positions with fully conserved residues; colons (:) indicate exchanges with biochemical similar and periods (.) with related amino acid. Below each residue, a quantitative representation of the Naive Empirical Bayesian class probability derived from the branch-site modeling in **c** for selective constraints in therians (red; also in the background of these residues in the sequence alignment) and constraints in vertebrates (blue). Coloured residues refer to results of binding tests of amino-acid exchange variants from **d** with unchanged binding (blue), reduced binding (orange) or no binding (red). (b) Pull-down of the TERF1 DBD of various different vertebrate species with either telomeric repeats or a control oligonucleotide. (c) Phylogenetic relationship of the species used for substitution rate analysis. The colours denote the branch classifications, with red representing selective constraints across therians and blue representing selective constraints across vertebrates. Substitution rates were calculated using PAML⁴⁵ to obtain the non-synonymous to synonymous substitution rate ratio ($d_N/d_S = \omega$). ω values <1 , $=1$ and >1 indicate purifying selection, neutral evolution and diversifying (positive) selection, respectively. A branch-site model (model D) was applied and compared with a homogeneous site model (discrete Model M3) and with a Model D that assumes neutral evolution for a predefined set of branches, representing our null hypothesis ($\omega = 1$). (d) Sequence-specific pull-down of single amino-acid exchange variants of the opossum TERF1 DBD. All seven residues exclusively found in platypus but not in therians were tested. (e) Sequence-specific pull-down of platypus TERF1 DBD variants using purified TERF1 DBDs. Combinations of the four identified residues from **d** were mutated to the opossum sequence to test whether their substitution can attribute TERF1 the capacity of directly binding to telomeric dsDNA. (f) Quantification of western blotting intensities from **e** ($n = 3$) shows similar binding affinity for platypus variants 3 and 4 compared with wild-type opossum TERF1-DBD. Mean with s.d. error bars.

Hit selection was based on P values obtained from an analogous t -statistic with variance increased by a constant factor S_0 equal 0.6. The P value cutoff indicated by dashed lines was set to 0.05. For creating the heat map, only proteins reported as enriched in at least five different species were considered. All plots were created using the `ggplot2` and `ggrepel` R package. Data formatting and filtering was performed with the `plyr` and `reshape2` R packages as well as base R commands.

Telomeric repeat amplification protocol. For all vertebrate species, 3.6 million cells were lysed in 100 μ l lysis buffer (50 mM Tris-HCl (pH 8.0), 150 mM NaCl and 1% Igepal CA-630 (Sigma) supplemented with protease inhibitor). The quantitative telomeric repeat amplification protocol assay was carried out using GoTaq qPCR Master Mix (Promega) with both TS (5'-AATCCGTCGAGCAGAGTT-3') and ACX primer (5'-GCGGGCTTACCCCTTACCCCTACCCCTAACCC-3') at 200 nM. Reactions were run on a ViiA 7 real-time PCR system (Thermo Fisher Scientific) with the following protocol: 25 °C for 20 min, 95 °C for 10 min and 40 cycles with 95 °C for 30 s, 60 °C for 30 s, and 72 °C for 1 min.

Recombinant expression of TERF1 variants and binding test. TERF1 DBDs were ordered as gene synthesis constructs (Genescript). The sequence was sub-cloned into the SLIC-compatible pETM44 vector via SLIC cloning³⁹ and expressed in *Escherichia coli* Rosetta at 25 °C. Amino-acid exchanges were introduced using site-directed mutagenesis⁴⁰ and validated by sequencing (GATC Biotech). Autoinduction was performed according to the published protocol³⁹ and bacteria were lysed in PBB buffer (150 mM NaCl, 50 mM Tris-HCl pH 7.5, 5 mM MgCl₂, 0.5% Igepal CA-630 (Sigma)) using a precooled Fastprep 24 system (Peglab) with silica beads. Soluble supernatant of *E. coli* extracts with overexpressed recombinant proteins were used for telomere pull-downs. For purified protein domains, 100 ml cultures were autoinduced, harvested and treated with 5 mg lysozyme (Sigma) prior to sonication on ice with a Branson Sonifier 450 for 10 times 15 s with 1 min breaks in between. Protein purification was carried out as previously described⁴¹. In short, samples were centrifuged for 30 min at 3,500 g and supernatant was filtrated with a 45 μ m (Fisher Scientific) syringe filter. The supernatant was loaded on an equilibrated 1 ml HisTrap HP column (GE Healthcare). The column was washed with buffer containing 50 mM imidazole (Sigma) and the bound proteins were eluted in 250 μ l fractions by buffer containing 500 mM imidazole (Sigma). The fractions were dialyzed overnight in storage buffer (0.5 M NaCl, 20 mM Tris-HCl pH 7.5, 5 mM MgCl₂, 10% glycerol and 1 mM DTT). The protein concentration was measured by Bradford assay (Bio-rad) and the purity of the elution fraction was assessed by polyacrylamide gel electrophoresis and Coomassie blue staining. Purified protein (10 μ g) was incubated with 500 μ g of paramagnetic streptavidin beads (Dynabeads, Thermo) coated with 600 nmol of biotinylated TTAGGG oligonucleotides for 2 h at 4 °C with slight agitation in PBB buffer (supplemented with 1 mM DTT and protease inhibitor). After three washes with PBB buffer, beads were transferred to a new Eppendorf tube and boiled in 1 $\times$ LDS buffer (Thermo) containing 100 mM DTT for 10 min at 70 °C. Samples were loaded on a NuPage 4–12% Bis-Tris polyacrylamide gel (Thermo), which was run with 1 $\times$ MES buffer (Thermo) for 45 min at 180 V.

Western blotting. Gels were blotted to nitrocellulose membranes (Protran83; Schleicher & Schuell) for 1 h constantly at 300 mA, blocked for 1 h at room temperature and incubated for 1 h with the anti-His₆ horseradish peroxidase-conjugated antibody following the manufacturer's instructions (Penta-His HRP Conjugate Kit, Qiagen). Membranes were washed twice in TBS-Tween-Triton buffer and once in TBS buffer for 10 min each. Detection was followed by incubating with enhanced chemiluminescence Prime Western Blotting Detection Reagent (GE-Healthcare). As a molecular weight standard, Seelue 2 (Invitrogen) was used. For the purified domains, western blotting intensities were analysed using the integrated density function of ImageJ (<https://imagej.nih.gov/ij/index.html>).

Multiple alignments and PAML statistical analysis. DNA and protein sequences of human TERF1 and TERF2 orthologues from up to 24 vertebrate species were obtained from the ENSEMBL database⁴² (release 75), including all species from the MS screen except axolotl for which there is currently no published genome available. To obtain multiple DNA sequence alignments, the corresponding protein sequences were aligned using MUSCLE⁴³ (version 3.8.31) and files were prepared with PAL2NAL⁴⁴ (version 14) to set up codon alignments and to remove gaps. Because whole-protein alignments for highly divergent species are difficult to obtain, we restricted the analyses to domain-specific alignments based on the human domain annotation. Here sequences were manually inspected and domains were separately analysed for the homeobox as well as TRFH domains of TERF1 and TERF2. Species for which the respective domain was not fully sequenced were excluded from further analysis. The exact species used for the analysis of the four different domains are depicted in the corresponding figure elements (Fig. 3c, Supplementary Fig. 4, Supplementary Table 1 and 2). Substitution rates were calculated using PAML⁴⁵ (version 4.7) to obtain the non-synonymous to synonymous substitution rate ratio ($d_N/d_S = \omega$). ω values < 1 , $= 1$ and > 1 indicate purifying selection, neutral evolution and diversifying (positive) selection, respectively. A branch-site model (model D) was applied and compared with a homogeneous site model (discrete model M3) and to a model D that assumes

neutral evolution for a predefined set of branches (for example, for the therian clade). In particular, we used a three-site class model, because we found a highly significant difference when compared with a discrete two-site class model, indicating heterogeneous levels of purifying selection within the protein domains. Significant differences between models were assessed by likelihood-ratio tests, which assume that the $2\Delta\ln L$ is approximately χ^2 distributed with degrees of freedom being the number of free parameters.

Code availability. Scripts are available from the authors upon request.

Data availability. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE⁴⁶ partner repository with the data set identifier PXD005517.

References

- Jarvis, E. D. *et al.* Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–1331 (2014).
- Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
- Stefflova, K. *et al.* Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell* **154**, 530–540 (2013).
- Villar, D., Flicek, P. & Odom, D. T. Evolution of transcription factor binding in metazoans—mechanisms and functional implications. *Nat. Rev. Genet.* **15**, 221–233 (2014).
- Wierer, M. & Mann, M. Proteomics to study DNA-bound and chromatin-associated gene regulatory complexes. *Hum. Mol. Genet.* **25**, R106–R114 (2016).
- Meyne, J., Ratliff, R. L. & Moyzis, R. K. Conservation of the human telomere sequence (TTAGGG)_n among vertebrates. *Proc. Natl Acad. Sci. USA* **86**, 7049–7053 (1989).
- Kumar, S. & Hedges, S. B. A molecular timescale for vertebrate evolution. *Nature* **392**, 917–920 (1998).
- Price, C. M. *et al.* Evolution of CST function in telomere maintenance. *Cell Cycle* **9**, 3157–3165 (2010).
- Ye, J., Renault, V. M., Jamet, K. & Gilson, E. Transcriptional outcome of telomere signalling. *Nat. Rev. Genet.* **15**, 491–503 (2014).
- Keilhauer, E. C., Hein, M. Y. & Mann, M. Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS). *Mol. Cell. Proteomics* **14**, 120–135 (2015).
- Hockemeyer, D., Daniels, J. P., Takai, H. & de Lange, T. Recent expansion of the telomeric complex in rodents. Two distinct POT1 proteins protect mouse telomeres. *Cell* **126**, 63–77 (2006).
- Lenain, C. *et al.* The Apollo 5' exonuclease functions together with TRF2 to protect telomeres from DNA repair. *Curr. Biol.* **16**, 1303–1310 (2006).
- van Overbeek, M. & de Lange, T. Apollo, an Artemis-related nuclease, interacts with TRF2 and protects human telomeres in S phase. *Curr. Biol.* **16**, 1295–1302 (2006).
- Conomos, D. *et al.* Variant repeats are interspersed throughout the telomeres and recruit nuclear receptors in ALT cells. *J. Cell Biol.* **199**, 893–906 (2012).
- Dejardin, J. & Kingston, R. E. Purification of proteins associated with specific genomic Loci. *Cell* **136**, 175–186 (2009).
- Marzec, P. *et al.* Nuclear-receptor-mediated telomere insertion leads to genome instability in ALT cancers. *Cell* **160**, 913–927 (2015).
- Kappei, D. *et al.* HOT1 is a mammalian direct telomere repeat-binding protein contributing to telomerase recruitment. *EMBO J.* **32**, 1681–1701 (2013).
- Palm, W. & de Lange, T. How shelterin protects mammalian telomeres. *Annu. Rev. Genet.* **42**, 301–334 (2008).
- Miyake, Y. *et al.* RPA-like mammalian Ctc1-Stn1-Ten1 complex binds to single-stranded DNA and protects telomeres independently of the Pot1 pathway. *Mol. Cell* **36**, 193–206 (2009).
- Conomos, D., Reddel, R. R. & Pickett, H. A. NuRD-ZNF827 recruitment to telomeres creates a molecular scaffold for homologous recombination. *Nat. Struct. Mol. Biol.* **21**, 760–770 (2014).
- Popuri, V. *et al.* Human RECQL1 participates in telomere maintenance. *Nucleic Acids Res.* **42**, 5671–5688 (2014).
- Gomes, N. M. *et al.* Comparative biology of mammalian telomeres. hypotheses on ancestral states and the roles of telomeres in longevity determination. *Aging Cell* **10**, 761–768 (2011).
- Giannone, R. J. *et al.* The protein network surrounding the human telomere repeat binding factors TRF1, TRF2, and POT1. *PLoS ONE* **5**, e12407 (2010).
- Czyzyk-Krzeska, M. F. & Bendixen, A. C. Identification of the poly(C) binding protein in the complex associated with the 3' untranslated region of erythropoietin messenger RNA. *Blood* **93**, 2111–2120 (1999).

25. Ito, Y., Bae, S. C. & Chuang, L. S. The RUNX family: developmental regulators in cancer. *Nat. Rev. Cancer* **15**, 81–95 (2015).
26. Court, R., Chapman, L., Fairall, L. & Rhodes, D. How the human telomeric proteins TRF1 and TRF2 recognize telomeric DNA: a view from high-resolution crystal structures. *EMBO Rep.* **6**, 39–45 (2005).
27. Kadonaga, J. T. & Tjian, R. Affinity purification of sequence-specific DNA binding proteins. *Proc. Natl Acad. Sci. USA* **83**, 5889–5893 (1986).
28. Grolimund, L. *et al.* A quantitative telomeric chromatin isolation protocol identifies different telomeric states. *Nat. Commun.* **4**, 2848 (2013).
29. Lambert, J.-P., Mitchell, L., Rudner, A., Baetz, K. & Figey, D. A novel proteomics approach for the discovery of chromatin-associated protein networks. *Mol. Cell. Proteomics* **8**, 870–882 (2009).
30. Wang, C. I. *et al.* Chromatin proteins captured by ChIP-mass spectrometry are linked to dosage compensation in *Drosophila*. *Nat. Struct. Mol. Biol.* **20**, 202–209 (2013).
31. Nishiyama, A. *et al.* Cell-cycle-dependent *Xenopus* TRF1 recruitment to telomere chromatin regulated by Polo-like kinase. *EMBO J.* **25**, 575–584 (2006).
32. Okamoto, K., Iwano, T., Tachibana, M. & Shinkai, Y. Distinct roles of TRF1 in the regulation of telomere structure and lengthening. *J. Biol. Chem.* **283**, 23981–23988 (2008).
33. Cooley, C. *et al.* Trf1 is not required for proliferation or functional telomere maintenance in chicken DT40 cells. *Mol. Biol. Cell* **20**, 2563–2571 (2009).
34. Karlseder, J. *et al.* Targeted deletion reveals an essential function for the telomere length regulator Trf1. *Mol. Cell. Biol.* **23**, 6533–6541 (2003).
35. Chen, Y. *et al.* A shared docking motif in TRF1 and TRF2 used for differential recruitment of telomeric proteins. *Science* **319**, 1092–1096 (2008).
36. Fairall, L., Chapman, L., Moss, H., de Lange, T. & Rhodes, D. Structure of the TRFH dimerization domain of the human telomeric proteins TRF1 and TRF2. *Mol. Cell* **8**, 351–361 (2001).
37. Butter, F., Kappei, D., Buchholz, F., Vermeulen, M. & Mann, M. A domesticated transposon mediates the effects of a single-nucleotide polymorphism responsible for enhanced muscle growth. *EMBO Rep.* **11**, 305–311 (2010).
38. 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).
39. Scholz, J., Besir, H., Strasser, C. & Suppmann, S. A new method to customize protein expression vectors for fast, efficient and background free parallel cloning. *BMC Biotechnol.* **13**, 12 (2013).
40. Zheng, L., Baumann, U. & Raymond, J.-L. An efficient one-step site-directed and site-saturation mutagenesis protocol. *Nucleic Acids Res.* **32**, e115 (2004).
41. Scheibe, M., Bonin, S., Hajnsdorf, E., Betat, H. & Morl, M. Hfq stimulates the activity of the CCA-adding enzyme. *BMC Mol. Biol.* **8**, 92 (2007).
42. Flicke, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42**, 55 (2014).
43. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
44. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, 12 (2006).
45. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
46. Vizcaino, J. A. *et al.* 2016 update of the PRIDE database and related tools. *Nucleic Acids Res.* **44**, D447–D456 (2016).
47. Ouellette, M. M., Aisner, D. L., Savre-Train, I., Wright, W. E. & Shay, J. W. Telomerase activity does not always imply telomere maintenance. *Biochem. Biophys. Res. Commun.* **254**, 795–803 (1999).
48. Holt, S. E., Wright, W. E. & Shay, J. W. Regulation of telomerase activity in immortal cell lines. *Mol. Cell. Biol.* **16**, 2932–2939 (1996).

Acknowledgements

We thank the following people for providing cell lines: Kai Simmons (dog), Janine Deakin (tammar wallaby and tasmanian devil), Nicolas Pollet (clawed frog), and Art Arnold (zebra finch), as well as Elly Tanaka for axolotl cells. Funding was provided by the Rhineland Pfalz Forschungsschwerpunkt GeneRED and the Deutsche Forschungsgemeinschaft (BU 2996/1) to F. Butter. D.K. is supported by the National Research Foundation Singapore and the Singapore Ministry of Education under its Research Centres of Excellence initiative. Work in the Buchholz laboratory was supported by the Excellence Initiative of the German Federal and State Governments (Institutional Strategy, measure 'support the best ZUK 64'). The research in the Mann laboratory received funding from the European Community's Health Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 259867, the Center for Integrative Protein Science Munich and the Max Planck Society for the Advancement of Science. T.I.G. is supported by a Leverhulme Early Career Fellowship (Grant ECF-2015-453) and NERC grant N013832/1.

Author contributions

D.K. and F. Butter initiated the research; D.K., M.S., A.B. and S.D. performed experiments; M.P.-R. and M.D. performed large-scale data analysis; T.I.G. analysed the evolutionary conservation of TERC domains; H.H. advised on vertebrate phylogeny and evolutionary analysis; F. Buchholz, M.M. and F. Butter supervised the research and contributed to the planning of the experiments; D.K., M.M. and F. Butter wrote the manuscript with input from all authors.

Additional information

Supplementary Information accompanies this paper at <http://www.nature.com/naturecommunications>

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

How to cite this article: Kappei, D. *et al.* Phylointeractomics reconstructs functional evolution of protein binding. *Nat. Commun.* **8**, 14334 doi: 10.1038/ncomms14334 (2017).

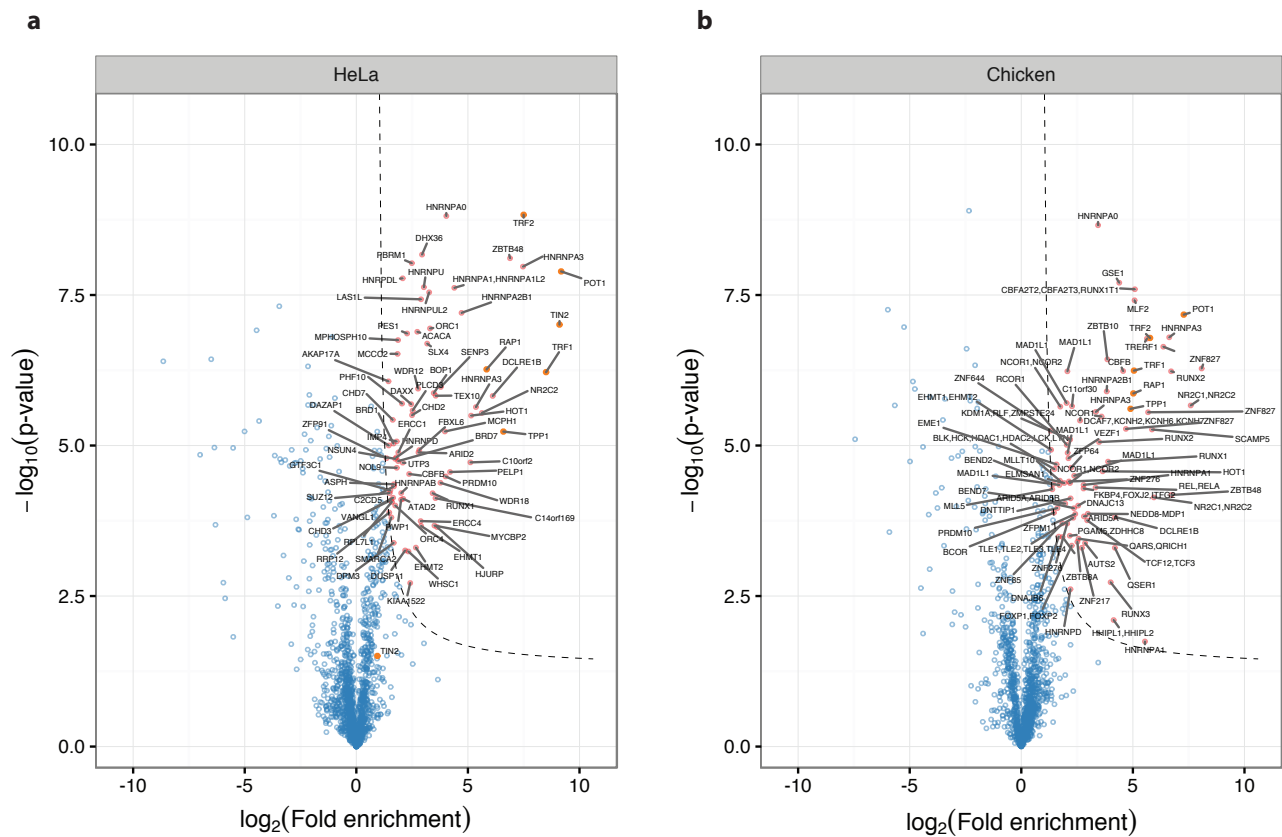
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

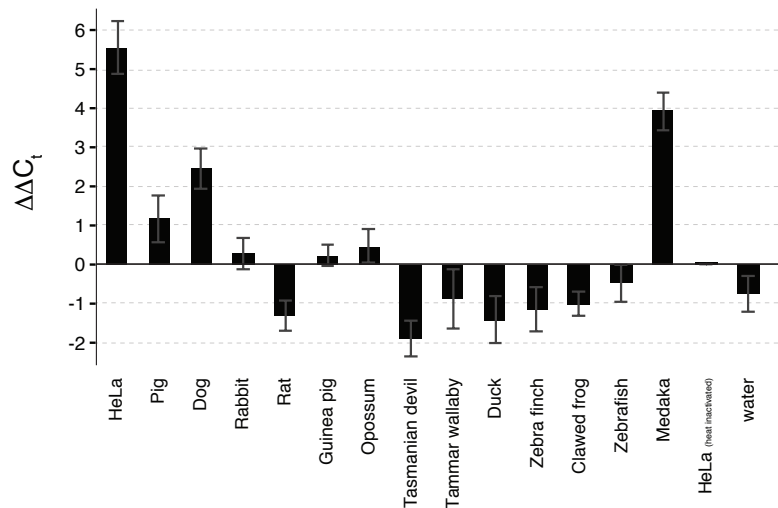
© The Author(s) 2017

Supplementary Figure 1



Supplementary Figure 1: Quantitative label-free DNA interaction screen with human HeLa and chicken 6C2 cells

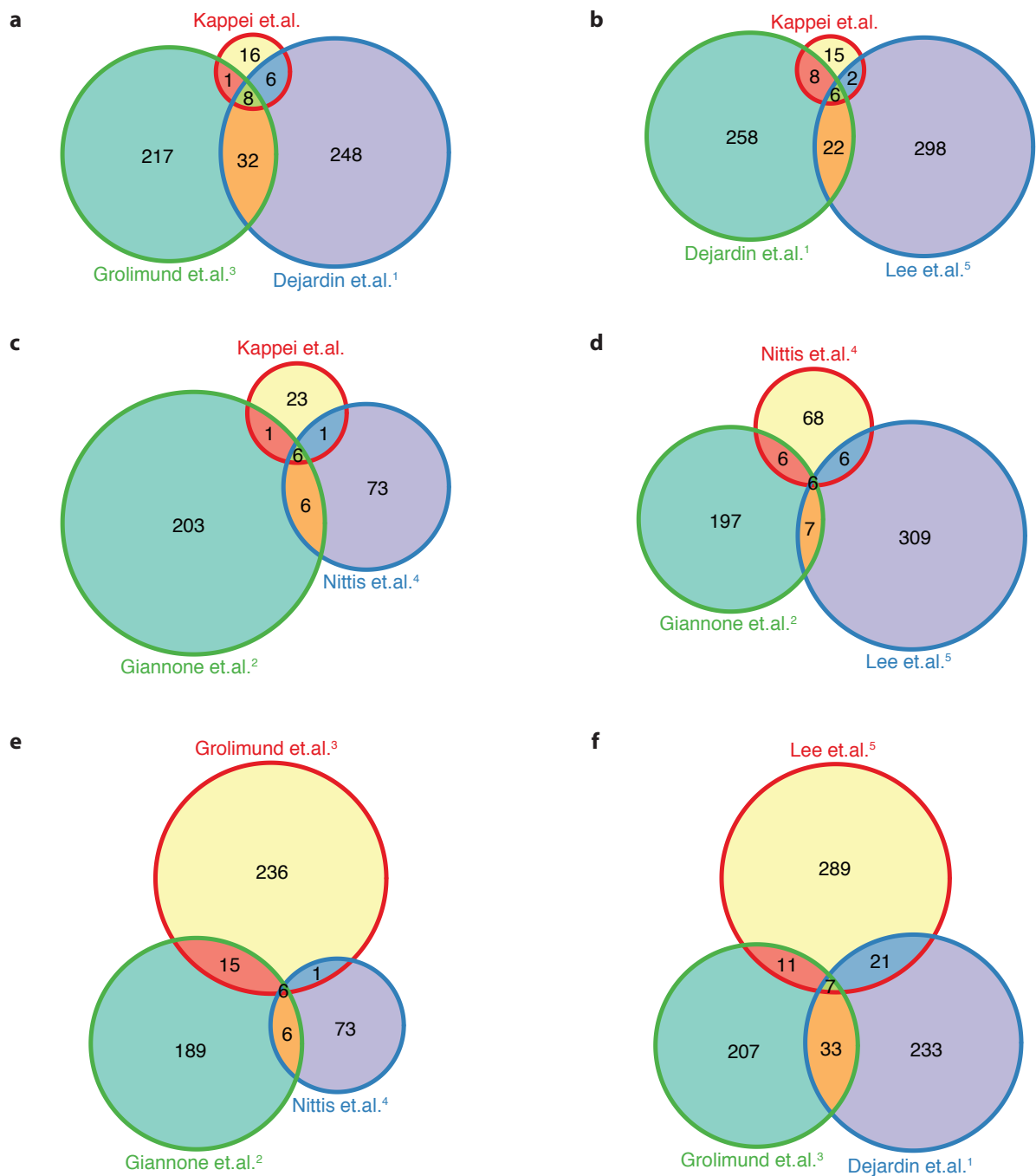
DNA pull-downs were performed as for the evolutionary screen shown in Fig. 1. Volcano plots for (a) HeLa and (b) 6C2 cells. Specifically enriched proteins (red circles) are distinguished from background binders (blue circles) by a two-dimensional cut-off with $S0=0.6$ and $p<0.05$ identical to the one used in the initial screen. Detected members of the shelterin complex (TERF1, TERF2, TIN2, TPP1, RAP1 and POT1) are highlighted (filled orange dots) and all specifically enriched proteins are annotated.



Supplementary Figure 2: Quantitative telomerase activity detection

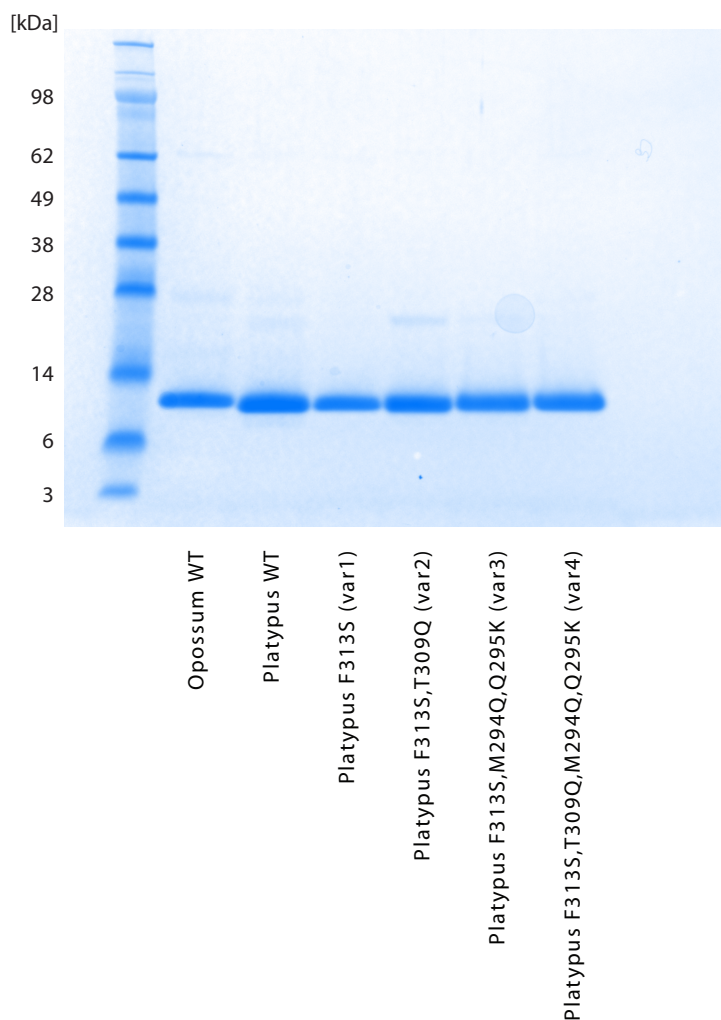
The presence of telomerase activity in each cell line as listed in Fig. 2a was determined based on a quantitative TRAP assay. HeLa cells served as a positive control for a telomerase-positive cell line and heat-inactivated HeLa extracts were used as a minimal threshold to determine telomerase-positive cells. Differences in Ct values from the quantitative PCR measurements are displayed. Rabbit, guinea pig and opossum are considered putatively positive due to a minor deviation ($<0.5\Delta\Delta C_t$) whereas pig, dog and medaka show clear telomerase activity. Error bars represent standard deviations (n=4).

Supplementary Figure 3



Supplementary Figure 3: Venn diagram comparison of telosome screens

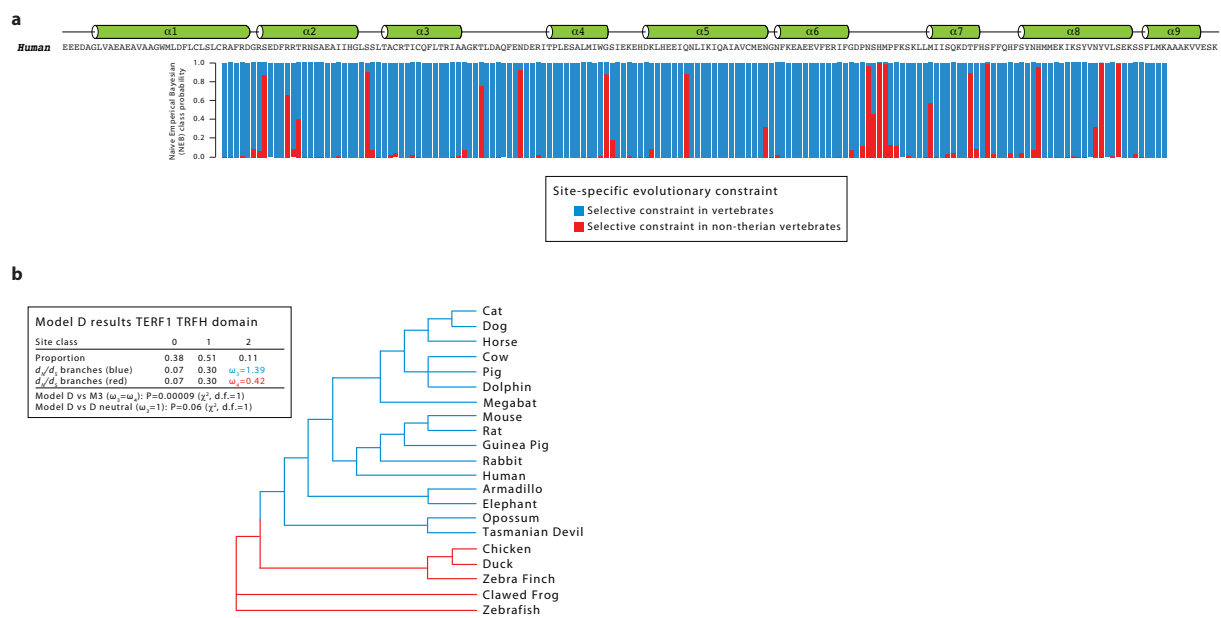
The list of telomeric factors from various screens¹⁻⁵ was obtained from the TeloPIN database⁶ and the overlap was calculated based on NCBI accession numbers. Numbers in the Venn diagrams (a-f) represent number of proteins that are unique or overlapping between the corresponding studies. Please note that all studies share the six shelterin proteins as a common set of factors.



Supplementary Figure 4: Coomassie blue gel of purified TERF1 DBDs

Representative Coomassie blue gel picture of the purified TERF1 DBDs used in Fig. 3e and Fig. 3f. 5 µg of each domain were loaded on the gel. All domains show high purity and migrate at the expected molecular weight.

Supplementary Figure 5



Supplementary Figure 5: PAML statistical analysis for the TERF1 TRFH domain

(a) Sequence of the human TERF1 TRFH domain. A schematic representation of the domain structure with nine α -helices (green) is shown. Below each residue is a quantitative representation of the Naive Empirical Bayesian class probability used for the branch-site modeling. Red represents selective constraints in non-therians and blue selective constraints in vertebrates. (b) Substitution rates were calculated using PAML7 to obtain the non-synonymous to synonymous substitution rate ratio ($dN/dS=\omega$). ω values <1 , $=1$, and >1 indicate purifying selection, neutral evolution, and diversifying (positive) selection, respectively. A branch-site model (model D) was applied and compared to a homogeneous site model (discrete Model M3) and to a Model D that assumes neutral evolution for a predefined set of branches. The phylogenetic tree represents 21 vertebrate species with available full TERF1 TRFH domain sequences that were included in this analysis.

Supplementary Table 1: PAML statistical analysis showing no significant difference in the model comparison between TERF2 homeobox domain sites. Complete TERF2 DBD domain sequences were retrieved for the following 24 vertebrate species and used for this analysis: cat, dog, horse, cow, pig, dolphin, megabat, mouse, rat, hamster, guinea pig, rabbit, human, armadillo, sloth, opossum, tammar wallaby, tasmanian devil, platypus, chicken, duck, zebra finch, clawed frog and zebrafish.

Model D results TERF2 myb			
Site class	0	1	2
Proportion	0.31	0.48	0.21
d_N/d_S branches (blue)	0	0.1	$\omega_3=0.65$
d_N/d_S branches (red)	0	0.1	$\omega_4=0.9$
Model D vs M3 ($\omega_3=\omega_4$): $P=0.41$ (χ^2 , d.f.=1)			
Model D vs D neutral ($\omega_3=1$): $P=0.75$ (χ^2 , d.f.=1)			

Supplementary Table 2: PAML statistical analysis showing no significant difference in the model comparison between TERF2 TRFH domain sites. Complete TERF2 TRFH domain sequences were retrieved for the following 18 vertebrate species and used for this analysis: dog, cow, pig, dolphin, megabat, mouse, rat, hamster, guinea pig, rabbit, human, armadillo, opossum, tammar wallaby, chicken, zebra finch, clawed frog and zebrafish.

Model D results TERF2 dim			
Site class	0	1	2
Proportion	0.63	0.18	0.19
d_N/d_S branches (blue)	0.01	0.003	$\omega_3=0.7$
d_N/d_S branches (red)	0.01	0.003	$\omega_4=0.32$
Model D vs M3 ($\omega_3=\omega_4$): $P=0.07$ (χ^2 , d.f.=1)			
Model D vs D neutral ($\omega_3=1$): $P=0.05$ (χ^2 , d.f.=1)			

Supplementary Table 3: Oligonucleotides used in this study

Sequence motif	primer sequence (5'-->3')
TTAGGG for	TTAGGG TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
TTAGGG rev	AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCT
GTGAGT for	GTGAGT GTGAGTGTGAGTGTGAGTGTGAGTGTGAGTGTGAGTGTGAGTGTGAGTGTGAGT
GTGAGT rev	ACACTCACTCACTCACTCACTCACTCACTCACTCACTCACTCACTCACTCACTCACTC

Supplementary References

1. Dejardin, J. & Kingston, R. E. Purification of proteins associated with specific genomic Loci. *Cell* **136**, 175–186 (2009).
2. Giannone, R. J. *et al.* The protein network surrounding the human telomere repeat binding factors TRF1, TRF2, and POT1. *PLoS One* **5**, e12407 (2010).
3. Grolimund, L. *et al.* A quantitative telomeric chromatin isolation protocol identifies different telomeric states. *Nat Commun* **4**, 2848 (2013).
4. Nittis, T. *et al.* Revealing novel telomere proteins using in vivo cross-linking, tandem affinity purification, and label-free quantitative LC-FTICR-MS. *Mol Cell Proteomics* **9**, 1144–1156 (2010).
5. Lee, O.-H. *et al.* Genome-wide YFP fluorescence complementation screen identifies new regulators for telomere signaling in human cells. *Mol Cell Proteomics* **10**, M110.001628 (2011).
6. Luo, Z. *et al.* TeloPIN: a database of telomeric proteins interaction network in mammalian cells. *Database (Oxford)* **2015** (2015).
7. Yang, Z. PAML 4. phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586–1591 (2007).

ZBTB10 binds the telomeric variant repeat TTGGGG and interacts with TRF2

Summary

The project focused on the zinc finger and BTB domain-containing protein 10 (ZBTB10). In our previous phylointeractomics screen, ZBTB10 was identified as a binder of telomeric TTAGGG repeats in six species (*Article II*). While we confirmed binding to TTAGGG, we further showed that both zinc fingers are necessary and sufficient for direct binding. Notably, we determined highest binding affinity of ZBTB10 for the telomeric variant repeat TTGGGG. Telomeric variant repeats are usually found in subtelomeres, but in cells that use alternative lengthening of telomeres to maintain their telomeres (ALT cells), variant repeats are also interspersed throughout telomeres. We visualized co-localization of ZBTB10 and TRF2 at a subset of telomeres in the ALT cell line U2OS.

To gain insights into the molecular function of ZBTB10, we analyzed its interaction network in telomerase-positive HEK293 and ALT-maintaining U2OS cells. Of note, the two shelterin complex members TRF2 and RAP1 were identified as specific interaction partners of ZBTB10 in both cell lines, and the N-terminus of TRF2 is specifically required for the interaction.

ZBTB10 is a novel binder of the telomere variant TTGGGG at telomeres in ALT cells.

Zusammenfassung

In diesem Projekt wurde das Zinkfingerprotein ZBTB10 untersucht, welches zuvor als Telomer-bindender Kandidat in sechs Vertebraten unseres Phylointeraktomikscreens identifiziert wurde (*Artikel II*). Hier zeigen wir nun, dass ZBTB10 mit den beiden Zinkfingern direkt an die TTAGGG Sequenz bindet. Außerdem wurde die stärkste Bindeaffinität von ZBTB10 für die alternative Telomersequenz TTGGGG im niedrigen nanomolaren Bereich bestimmt. Das Vorkommen alternativer Telomersequenzen beschränkt sich normalerweise auf Subtelomere. In Zellen, die ihre Telomere über homologe Rekombination verlängern (ALT Zellen), kommen diese Sequenzen allerdings auch in Telomeren vor. Einhergehend mit

der *in vitro* Bindefähigkeit von ZBTB10, wurde eine co-Lokalisierung von ZBTB10 und TRF2 an einem Teil der Telomere in U2OS Zellen quantifiziert.

Um weitere Hinweise bezüglich der molekularen Funktion von ZBTB10 zu erhalten, wurde das Interaktionsnetzwerk in Telomerase-positiven HEK293 Zellen und in ALT-erhaltenden U2OS Zellen mittels quantitativer Massenspektrometrie bestimmt. Unter den Interaktionspartnern wurden interessanterweise die beiden Shelterin-Komplexproteine TRF2 und RAP1 in beiden Zelllinien identifiziert. Weiterhin konnte gezeigt werden, dass der N-Terminus von TRF2 notwendig und ausreichend ist, um die Interaktion mit ZBTB10 zu ermöglichen.

ZBTB10 ist ein neues Protein, das die variable Telomersequenz TTGGGG in ALT Zellen bindet.

Statement of Contribution

I was in charge of that project, planned and conducted experiments and talked to collaborators for their help with specific experiments. I wrote the manuscript and prepared all figures. Together with Dennis and Falk we finalized the manuscript.

Supervisor confirmation

ZBTB10 binds the telomeric variant repeat TTGGGG and interacts with TRF2

Alina Bluhm¹, Nikenza Viceconte¹, Fudong Li², Grishma Rane³, Sandra Ritz⁴, Suman Wang², Michal Levin¹, Yunyu Shi², Dennis Kappei^{1,3,*} and Falk Butter^{1,*}

¹Quantitative Proteomics, Institute of Molecular Biology, 55128 Mainz, Germany, ²Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science and Technology of China, 230027 Hefei, Anhui, China, ³Cancer Science Institute of Singapore, National University of Singapore, Singapore 117599 and ⁴Microscopy Core Facility, Institute of Molecular Biology, 55128 Mainz, Germany

Received August 07, 2018; Revised November 27, 2018; Editorial Decision December 13, 2018; Accepted December 14, 2018

ABSTRACT

Telomeres are nucleoprotein structures at the ends of linear chromosomes and present an essential feature for genome integrity. Vertebrate telomeres usually consist of hexameric TTAGGG repeats, however, in cells that use the alternative lengthening of telomeres (ALT) mechanism, variant repeat sequences are interspersed throughout telomeres. Previously, it was shown that NR2C/F transcription factors bind to TCAGGG variant repeats and contribute to telomere maintenance in ALT cells. While specific binders to other variant repeat sequences have been lacking to date, we here identify ZBTB10 as the first TTGGGG-binding protein and demonstrate direct binding via the two zinc fingers with affinity in the nanomolar range. Concomitantly, ZBTB10 co-localizes with a subset of telomeres in ALT-positive U2OS cells and interacts with TRF2/RAP1 via the N-terminal region of TRF2. Our data establishes ZBTB10 as a novel variant repeat binding protein at ALT telomeres.

INTRODUCTION

Telomeres are nucleoprotein structures at the end of linear chromosomes that in vertebrates consist of double-stranded (ds) TTAGGG repeats. They are constitutively bound by the shelterin complex, which consists of six proteins: the telomere repeat binding factors TRF1 and TRF2, the TRF2-interacting protein RAP1, the telomeric ssDNA-binding protein POT1 with its direct interactor TPP1 as well as the bridging factor TIN2 (1). The shelterin complex is important for telomere end protection by preventing recognition of the chromosome end as a DNA double strand break (DSB) and protecting it from illegitimate DNA repair activities (2). Within the complex, TRF2 plays

an important role in end protection by inducing and stabilizing t-loop structures (3–5). Thereby, chromosomal ends are safeguarded from nucleolytic degradation and protected from the DNA damage response and repair activities such as non-homologous end joining. TRF2 and RAP1 together repress homology directed repair by preventing telomeric localization of PARP1 and SLX4 (6). TRF2 has furthermore been shown to repress ATM kinase activity at telomeres (7,8). Moreover, TRF2 recruits various factors such as RTEL1 during S-phase to enable t-loop unwinding (9) or ORC to stimulate pre-replication complex assembly at telomeres (10).

During each cell division telomeres progressively shorten, which is counteracted by the ribonucleoprotein telomerase in the majority of cancers (11). In the absence of telomerase, replicative immortality can be achieved by alternative lengthening of telomeres (ALT) in 10–15% of cancers (12). ALT is based on homology directed repair (13,14), leading to long heterogeneous telomeres and the interspersal of variant repeats such as TCAGGG, TGAGGG and TTGGGG that are otherwise confined to subtelomeres (15,16). As a result, NR2C/F nuclear receptors, which have been shown to preferentially bind TCAGGG repeats, are recruited along the interspersed telomeric sequence and contribute to the ALT phenotype (i.e. ALT-associated PML bodies, C-circles, telomere sister chromatid exchanges) (16). Moreover, NR2C/F receptors have also been described as bridging proteins leading to targeted telomeric insertion throughout chromosomes in ALT cells (17). However, proteins specifically binding to other telomeric variant repeat sequences have not been described yet.

Zinc finger and BTB domain containing proteins offer a large surface for protein-protein interactions, are able to form homo- and heterodimers and have been suggested to play a role in regulating gene expression (18). Recently the family member ZBTB48, also known as TZAP, has been described as a telomeric protein involved in telomere length

*To whom correspondence should be addressed. Tel: +49 6131 39 21570; Fax: +49 6131 39 21521; Email: f.butter@imb.de
Correspondence may also be addressed to Dennis Kappei. Tel: +65 6516 1333; Fax: +656873 9664; Email: dennis.kappei@nus.edu.sg

regulation (19,20). In a phylointeractomics screen across 16 vertebrate species, we found multiple ZBTB proteins including ZBTB10 as a binder of TTAGGG repeats (21). Here, we show that ZBTB10 preferably binds to the TTGGGG variant repeat, partially localizes to telomeres in the ALT cancer cell line U2OS and interacts with the TRF2/RAP1 heterodimeric complex irrespective of the telomere maintenance mechanism.

MATERIALS AND METHODS

Cell culture and plasmids

All cell lines (HEK293, HeLa E1 (22), U2OS, GM847 and SAOS2) were cultivated in 4.5 g/l glucose, 1 mM sodium pyruvate Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum (Gibco), 2 mM glutamine (Life Technologies), 100 U/ml penicillin and 100 µg/ml streptomycin (Gibco) in a humidified incubator at 5% CO₂ and 37°C.

The ZBTB10 clone was obtained from Bioscience (Q96DT7-isoform 3; ENST00000610895.2) and full-length TRF2 (Q15554; ENST00000254942.7) was amplified from HEK293 cDNA (First strand synthesis Kit; Thermo Scientific). Deletion constructs of both proteins were obtained by PCR amplification using specific primers (Supplementary Table S1). Constructs were confirmed by Sanger sequencing (GATC) and sequences were LR recombined into Gateway-compatible expression vectors: pDest-pcDNA3.1 with N-terminal FLAG tag or pcDNA-Dest47 with C-terminal GFP tag.

Transfections

Plasmid transfection in HEK293 cells was performed using linear polyethylenimine (PEI, MW 25 000; Polysciences). 450 000 cells/ml were seeded one day before in 10 cm dishes and transfected with 12 µg DNA and 48 µl PEI diluted in DMEM. U2OS cells were transfected using linear PEI HCl MAX (MW 4000; Polysciences) or with the Amaxa Cell Line Nucleofactor Kit V (Lonza) according to manufacturer's instructions. SAOS2, GM847 and HeLa E1 cells were transfected using Lipofectamine 2000 (Thermo) diluted in Opti-MEM (Life Technologies). Stable cell lines were generated by selection in 0.8 mg/ml G418 (Sigma) for one week, enriched by sorting GFP-positive cells (Becton Dickinson FACSaria III SORP) and further cultured in 0.4 mg/ml G418.

DNA pulldown

Twenty five microgram of forward and reverse oligonucleotides of telomeric, variant repeat or control sequences (Supplementary Table S1) were denatured at 80°C and annealed by cooling. Double-stranded oligonucleotides were then polymerized using 50 U T4 polynucleotide kinase (Thermo Scientific) and 80 U T4 DNA ligase (NEB), biotinylated with desthiobiotin-dATP (Jena Bioscience) by Klenow fragment (Thermo Scientific) and purified using G-50 columns (GE Healthcare). Chemically synthesized DNA was immobilized on 0.5 mg streptavidin-coated magnetic beads (Dynabeads MyOne Streptavidin C1, Invitrogen) for

15 min at RT and incubated with 400–800 µg protein lysates diluted in PBB buffer (150 mM NaCl, 50 mM Tris-HCl pH 7.5, 10 mM MgCl₂, 0.5% IGEPAL CA-630, 1 µM Pepstatin, 1 µg/ml Leupeptin and 0.5 mM PMSF) for 1.5 h on a rotation wheel at 4°C. Sheared salmon sperm DNA (16.7 µg; Ambion) was added as a competitor. After three washes with PBB buffer, bound proteins were eluted in LDS sample buffer supplemented with 0.1 M DTT, boiled for 10 min at 70°C and separated on a 4–12% NuPAGE Novex Bis-Tris precast gel (Life Technologies) for 50 min at 180 V in 1× MES buffer.

Protein purification

ZBTB10-2Znf (aa 540–606) and ZBTB10-2Znf-C (aa 540–682) were cloned into a modified pET28a vector with a SUMO protein fused at the N-terminus after a His₆ tag. Proteins were expressed in *Escherichia coli* BL21 (DE3) cells (Stratagene). After induction for 24 h with 0.1 mM IPTG at 16°C, the cells were harvested and then lysed by sonication. Proteins were purified by Ni-NTA agarose beads (Qiagen), followed by Ulp1 cleavage to remove the His-SUMO tag and size-exclusion chromatography on a Hiload 16/60 Superdex 75 column (GE Healthcare) in a buffer containing 20 mM Tris-HCl (pH 7.5) and 1 M NaCl. Purified proteins were dialyzed with Buffer A (20 mM Tris-HCl pH 7.5, 150 mM NaCl) and concentrated for subsequent analysis.

Fluorescence polarization assay

The protein stock was diluted in 1/2 series in Buffer A to the lowest desired concentration. The protein dilution series was then incubated with 50 nM FAM-labeled probe in Buffer A (200 µl final volume) for 30 min at RT. Fluorescence polarization was measured using a Spectra-Max M5 plate reader (Molecular Devices). Curves were fit individually using the equation $[mP] = \frac{[maximum\ mP] \times [C]}{K_D + [C]}$ + [baseline mP], where mP is millipolarization, and [C] is protein concentration. K_D values and the fitting figures were derived by fitting the experimental data (two experimental replicates) to the equation using a fitting script written in python.

Nuclear and chromatin extract preparation

Cells were harvested, washed in PBS and incubated for 10 min on ice in five pellet volumes of cold buffer A (10 mM HEPES-KOH pH 7.9, 1.5 mM MgCl₂, 10 mM KCl). Cells were pelleted, resuspended in two volumes of cold buffer A supplemented with 0.2% IGEPAL CA-630 and complete protease inhibitor mix (Roche) and homogenized with 30 strokes (type B pestle) in a dounce homogenizer. The cytoplasmic fraction was collected after centrifugation for 15 min at 3900 rpm at 4°C. After a PBS wash, nuclei were resuspended in two volumes of buffer C (420 mM NaCl, 20 mM HEPES-KOH pH 7.9, 20% glycerol, 2 mM MgCl₂, 0.2 mM EDTA pH 8, 0.1% IGEPAL CA-630, 0.5 mM DTT and complete protease inhibitor mix (Roche)), incubated for 1 h on a rotating wheel at 4°C and centrifuged at 14 800 rpm for 1 h at 4°C to collect the nuclear extract. The chromatin pellet was incubated for 30 min on a rotating wheel at 4°C

in buffer C supplemented with 0.05% SDS, sonicated for 15 cycles (30 s on and 45 s off) and further incubated for 1 h on a rotating wheel at 4°C. The extract was cleared by centrifugation at 14 500 rpm for 1 h at 4°C.

GFP pulldown

Ten microliters slurry of GFP-M Nanotrap (Chromotek) magnetic beads were used per immunoprecipitation (IP) and equilibrated in wash buffer (150 mM NaCl, 10 mM Tris-HCl pH 7.5, 0.5 mM EDTA). Cells were lysed in wash buffer supplemented with 0.5% IGEPAL CA-630 and 0.1% sodium deoxycholate, 1 μ M Pepstatin, 1 μ g/ml Leupeptin and 0.5 mM PMSF for 30 min on ice. Cleared supernatants were either first incubated with 500 U Sm-nuclease (in-house) for 30 min on ice or directly incubated with equilibrated GFP beads for 2 h on a rotation wheel at 4°C. After three washes on ice, proteins were eluted in LDS sample buffer supplemented with 0.1 M DTT, boiled for 10 min at 95°C and separated on a 4–12% NuPAGE Novex Bis-Tris precast gel (Life Technologies) for 10 min at 180 V in 1× MOPS buffer.

Western blot

Separated proteins were transferred from a SDS gel to a nitrocellulose membrane (Amersham) at 300 mA for 1 h. The membrane was blocked in PBS containing 5% (w/v) non-fat milk (Sigma) and 0.1% Tween-20 (PBST) for 1 h at RT and incubated with primary antibodies for 1–2 h at RT or overnight at 4°C. Detailed information about antibodies can be found in Supplementary Table S1. Antibodies were diluted in PBST containing 5% BSA and 0.02% sodium azide. After three washes in PBST, the membrane was incubated for 1 h at RT with secondary antibody. Chemiluminescence detection was performed using SuperSignal West Pico solution (Pierce). The SeeBlue Plus2 Pre-stained Protein Standard (Thermo Scientific) was used as a marker.

MS sample preparation, data acquisition and data analysis

In-gel digestion was performed as previously described (23). Tryptic peptides were desalted on StageTips and loaded on an in-house packed C18 column (25 cm long, 75 μ m inner diameter) for reverse-phase chromatography. The EASY-nLC 1000 system (Thermo Scientific) was mounted to a Q Exactive Plus mass spectrometer (Thermo Scientific) and peptides were eluted from the column in an optimized 90 min (pulldown) or 5 h (proteome) gradient from 2 to 40% MS grade acetonitrile/0.1% formic acid solution at a flow rate of 200 nl/min. The mass spectrometer was used in a data-dependent acquisition mode with one MS full scan and up to 10 MS/MS scans using HCD fragmentation. All raw files were processed with MaxQuant (version 1.5.2.8) and searched against the human Uniprot database (81 194 entries). Carbamidomethylation (Cys) was set as fixed modification, while oxidation (Met) and protein *N*-acetylation were considered as variable modifications. For enzyme specificity, trypsin was selected with a maximum of two miscleavages. LFQ quantification (without fast LFQ)

using at least two LFQ ratio counts and the match between run option were activated. Data analysis was performed in R using existing libraries (knitr, reshape2, dplyr, ggplot2, ggrepel) and in-house scripts. Protein groups reported by MaxQuant were filtered removing known contaminants, protein groups only identified by site and those marked as reverse hits. Missing values were imputed at the lower end of LFQ values using random values from a beta distribution fitted at 0.1–1.5%. For statistical analysis, *P*-values were calculated using the Welch's *t*-test. Enrichment values in the volcano plots represent the median difference of log₂ transformed LFQ intensities between ZBTB10-GFP and GFP. For the definition of enriched protein groups in all GFP pulldowns we applied an identical cut-off ($S_0 = 1.6$, $c = 0.8$ and $P = 0.05$).

Immunofluorescence (IF) staining

One day before, cells were seeded on a 96-well imaging plate (Perkin Elmer) or on glass slides (#1216691, Marienfeld). For synchronization at the G2/M border, cells were incubated with 10 μ M RO-3306 (SML0569#; Sigma) inhibitor for 20 h and after two washes in medium released into cell cycle progression for 4.5 h to reach G1 phase. Cells were briefly washed with PBS and fixed in 10% formalin solution (Sigma) for 10 min at RT followed by two washes with 30 mM glycine-PBS. Permeabilization was performed for 5 min at 4°C in 0.5% Triton X-100 in PBS. After two more washes in 30 mM glycine-PBS, cells were blocked for 15 min at RT in blocking solution (0.2% fish skin gelatin in PBS). Primary antibodies were diluted in blocking solution and incubated for 1 h at RT or overnight at 4°C. Detailed information regarding antibodies can be found in Supplementary Table S1. After three washes in blocking buffer for 3 min each, cells were incubated for about 45 min with secondary antibodies in the dark at RT. After three more washes, slides were briefly rinsed in distilled water and mounted using DAPI Prolong Diamond Antifade Reagent (Invitrogen). In case of imaging plates, cells were incubated for 10 min in 10 μ g/ml Hoechst 33342 (Life Technologies) followed by PBS washes and stored in PBS at 4°C. Imaging was performed using the 63× water magnification and the spinning disk confocal mode of the high-content screening microscope Opera Phenix (Perkin Elmer) or the 63× oil magnification of the inverted TCS SP5 confocal microscope (Leica). Raw images were used for quantification of co-localization events using the maximum projection of acquired z-stacks. Images were deconvolved using the Huygens Essentials software.

Single molecule localization microscopy (SMLM)

2D SMLM was performed on a SR GSD microscope set up (Leica Microsystems CMS GmbH) based on an inverse widefield microscope DMI AF6000 equipped with an oil immersion objective (160×/1.43 plan apochromatic, HCX PL APO, TIRF). The temporal isolation of single molecules was obtained by fluorescence activation and deactivation (on and off switching) of organic dyes by laser excitation in an appropriate buffer system, which supports reversible photoswitching. The method is known

as ground state depletion and individual molecule return (GSDIM) (24) or direct stochastically optical reconstruction microscopy (dSTORM) (25). U2OS cells were seeded on imaging dishes (#81156, ibidi) overnight and IF staining was performed as described using the anti-GFP Atto488 nanobody and anti-TRF2 antibody coupled to rabbit Alexa-647 (F(ab)2). Detailed information regarding antibodies can be found in Supplementary Table S1. Before imaging, fixed cells were submersed with freshly prepared photoswitching buffer containing the oxygen scavenging system Pyranose-Oxidase-Catalase in combination with a thiol as reducer to provide stable pH conditions during measurement time. The buffer was adapted from (26): pyranose-oxidase (3 U/ml), catalase (90 U/ml), glucose 100 mg/ml (10% w/v), 5 mM cysteamine/ β -mercaptoethylamine hydrochloride, Tris buffer (20 mM Tris-HCl pH 7.5, 50 mM NaCl), 2 mM cyclooctatetraene and 20% glycerol. Lasers (488 nm/300 mW, 642 nm/500 mW) were used for excitation and photoactivation, and a 405 nm laser (30 mW) was used for backpumping. TRF2-Alexa647 was imaged before ZBTB10-Atto488 to prevent photobleaching of the far-red dye by the 488 nm laser. Bleed through was controlled by single fluorophore stained samples. Fluorescent emission was filtered (488 HP-T, dichroic 405/488 nm, em 505–595 nm; 647 HP-T, dichroic 405/642nm, em 660–760 nm) and imaged on an EMCCD camera with a pixel size of 100 nm (iXon3 ultra 897, Andor Technologies). Sequential two-color SMLM imaging was performed in TIRF mode ca. 200 nm above the surface of the imaging dish. First, corresponding widefield images of ZBTB10-GFP positive cells and controls were created in the 488 nm and the 647 nm channel. On average 15 000 frames of a 180×180 pixel ($18 \times 18 \mu\text{m}$) camera sub-region were recorded with a camera frame rate of 50 Hz (20 ms). The laser intensity for activation was progressively increased during acquisition to keep the amount of emitting fluorophores constant ($0.5\text{--}2 \text{ kW}/\text{cm}^2$) and ca. 5% of the 405 nm laser was used for backpumping. Chromatic aberrations were controlled by a ‘sandwich method’, imaging TetraSpeck microspheres (0.1 μm , 1000 $\times$ dilution in PBS, pre-adsorbed on the surface of the imaging dish; Thermo Scientific) before and after the biological sample. The offset (shift between channels 647 and 488) was obtained by autocorrelation (chromatic shift correction, Huygens software, SVI), and the correction matrix was transferred to the respective sample. Fluorescence image acquisition and processing was performed with the LAS AF 4.0 software (Leica microsystems), Fiji, and ThunderSTORM (27). The post-processing procedure was done in ThunderSTORM: uncertainty 2–25, merging maximum distance of 20 nm in x,y , drift correction. The shift between the 642 and 488 nm channel was corrected by cross correlation of the TetraSpeck beads (chromatic aberration corrector, Huygens Essential, SVI). Distances between TRF1/TRF2 and ZBTB10/TRF2 protein pairs were measured between the intensity maxima (peak-to-peak distance) in the reconstructed two-color SMLM images.

Generation of knockout cells

ZBTB10 knockouts (KOs) were generated using three different guide RNAs targeting the second exon of the

ZBTB10 gene (Supplementary Table S1). Guide oligonucleotides carrying the PAM sequence were cloned into the pX459 V2 vector containing Cas9 and the sgRNA scaffold and validated by Sanger sequencing (GATC). One day after transfection, HEK293 and U2OS cells were selected in 750 ng/ml puromycin. After successful selection gDNA was extracted using the QIAamp DNA Mini Blood Kit (Qiagen) according to manufacturer’s instructions and a ~ 1100 bp long target region around the cut locus was amplified by PCR. Cas9 activity was tested in a T7 endonuclease assay. In a first annealing step mismatched DNA was allowed to form, which was visualized by agarose gel electrophoresis after T7 endonuclease digest for 30 min at 37°C . After confirming Cas9 activity, single cell sorting was performed (Becton Dickinson FACSARIA III SORP) and cells were expanded to screen for ZBTB10 KO clones by Western blot using an anti-ZBTB10 antibody (Supplementary Table S1).

RESULTS

ZBTB10 binds telomeric and telomeric variant repeats directly via its zinc fingers

In our previous phylointeractomics screen for telomere-associated proteins across 16 vertebrate species, we identified the zinc finger and BTB domain-containing protein 10 (ZBTB10) as a binder to the telomeric TTAGGG sequence in six different species ranging from fish to mammals (21). To verify the ability of human ZBTB10 to bind to telomeric repeats and to compare specificity with variant repeat sequences, we performed DNA pulldown assays with whole cell lysates from telomerase-positive HEK293 and telomerase-negative U2OS cell lines stably expressing a C-terminal GFP-tagged ZBTB10 (Figure 1A, Supplementary Table S1). Indeed, while human ZBTB10 was able to bind to TTAGGG repeats, we also observed binding to TTCGGG and TTGGGG variant repeats. In fact, ZBTB10 showed strongest enrichment at TTGGGG variant repeats. To investigate whether ZBTB10’s binding ability is due to a direct interaction with DNA, we tested a recombinantly expressed His-tagged ZBTB10 in DNA pulldowns. The recombinant His₆-ZBTB10 showed the same binding behavior (Figure 1B), demonstrating that recognition of the DNA sequence is a function of the protein and not mediated by auxiliary factors. To identify the DNA-binding domain, we generated four deletion constructs: ΔBTB (aa $\Delta 164\text{--}297$), Δmid (aa $\Delta 306\text{--}529$), ΔZnf (aa $\Delta 540\text{--}595$; two classical Cys₂–His₂ zinc fingers) and ΔC (aa $\Delta 602\text{--}698$) (Figure 1C, Supplementary Table S1). Cells stably expressing ZBTB10 fusion proteins did not show any growth defect and no change in cell cycle distribution (Supplementary Figure S1A). While ZBTB10 Δmid and ZBTB10 ΔC were still able to bind telomeric variant repeats, the loss of either the BTB domain or the zinc finger region abolished binding to any tested sequence (Figure 1D). To further substantiate the importance of both zinc fingers for DNA binding, we introduced a single point mutation in either one of the two zinc fingers (C551A or C579A) or in both. DNA pulldown assays revealed that both zinc fingers need to be intact for binding to telomeric DNA (Figure 1E). As we observed different enrichment of ZBTB10 to the distinct repeat sequences, we next quantified binding affinity

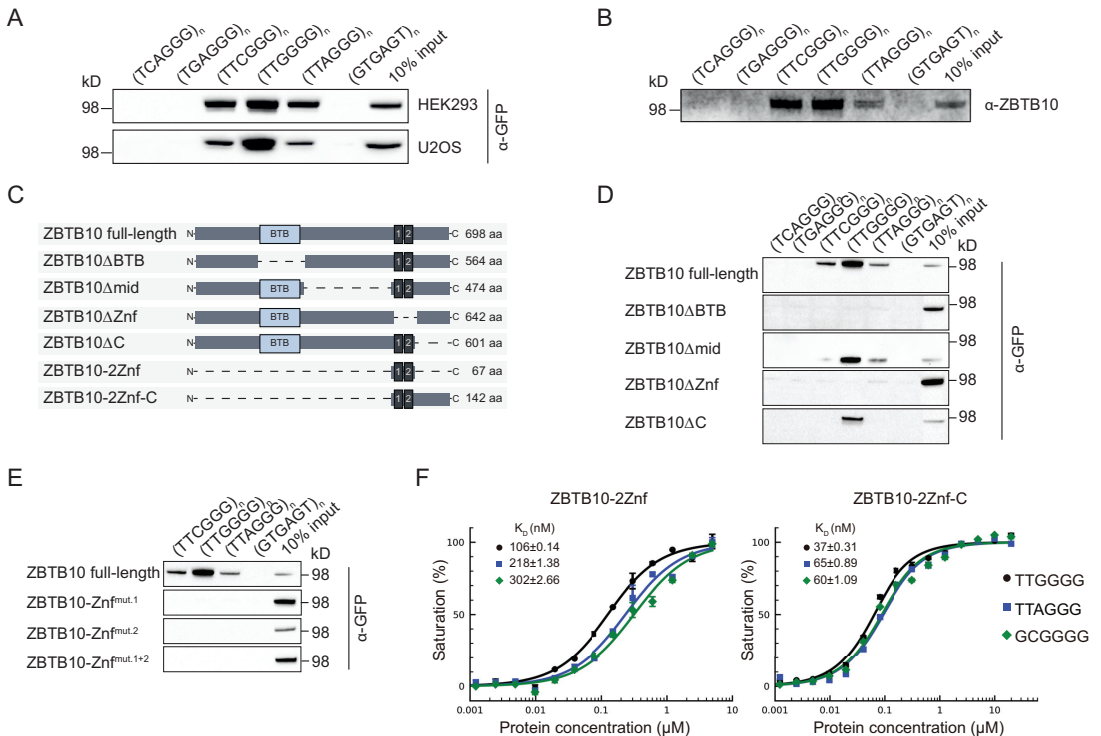


Figure 1. ZBTB10 preferably binds the TTGGGG variant repeat sequence. (A) Binding of stably expressed ZBTB10-GFP from HEK293 and U2OS cells to telomeric and telomeric variant repeats. Biotinylated concatenated DNA oligonucleotides were immobilized on streptavidin magnetic beads and bound proteins were analyzed by Western blot. A scrambled GTGAGT repeat sequence was used as control bait. (B) DNA pulldowns with extracts containing recombinantly expressed His₆-ZBTB10 to telomeric and telomeric variant repeats. (C) Schematic representation of constructed ZBTB10 deletion variants together with their respective length in amino acids (aa). (D) Binding of ZBTB10-GFP variants in DNA pulldown experiments stably expressed in U2OS cells. (E) DNA pulldowns of ZBTB10 zinc finger mutants zinc finger 1 (C551A) and zinc finger 2 (C579A) stably expressed in U2OS cells. (F) Fluorescence polarization assays of the minimal ZBTB10 construct consisting of both zinc fingers (ZBTB10-2Znf) and the longer construct containing both C2H2 fingers and the C-terminus (ZBTB10-2Znf-C). Binding affinities to the telomeric variant repeat TTGGGG (black circle), the telomeric repeat TTAGGG (blue square) and the predicted binding motif GCGGGG (green diamond) were determined.

ties by fluorescence polarization. Here, we determined the affinity of ZBTB10 to its predicted DNA-binding motif GCGGGG, the telomeric sequence TTAGGG and the variant repeat TTGGGG. To this end, we purified a minimal construct consisting of both zinc fingers (ZBTB10-2Znf: aa 540–606, Figure 1C) and a longer construct containing the C-terminus (ZBTB10-2Znf-C: aa 540–682, Figure 1C). We decided to include the longer construct as we found a putative atypical C2HR finger C-terminally adjacent to the two canonical C2H2 fingers, which might also be important for DNA binding. In agreement with the DNA pulldown data, ZBTB10-2Znf showed strongest binding to the TTGGGG probe with 106 nM, a two-fold higher affinity in comparison to 218 nM for the TTAGGG probe and to 302 nM for the GCGGGG motif (Figure 1F). Noteworthy, the longer ZBTB10-2Znf-C construct showed increased specificity for all tested motifs exhibiting strongest affinity for TTGGGG with a K_D value of 37 nM, again twice as high as for the other tested motifs (Figure 1F). We thereby quantitatively confirmed the preferred binding of ZBTB10 to the

TTGGGG variant repeat sequence and established that the two classical zinc fingers combined are necessary and sufficient for DNA binding, while the C-terminal region further increases binding affinity.

ZBTB10 localizes to a subset of telomeres in ALT cells

To investigate whether ZBTB10 localizes to telomeres *in vivo*, we analyzed U2OS cells stably expressing ZBTB10-GFP. First, we assessed whether the expression level of the ZBTB10-GFP fusion is similar to the endogenous level and compared expression of ZBTB10 (WT), stably expressed (SE) and transiently overexpressed (OE) ZBTB10-GFP in U2OS cells by label-free quantitative mass spectrometry. In contrast to a >100-fold overexpression when transiently transfected, the expression level of the stably expressed protein were within 10-fold of the endogenous protein level (Supplementary Figure S1B). Using these stable ZBTB10-GFP cells, we observed a punctuated pattern of ZBTB10 with an average of 17 foci per nucleus (Figure 2A, Sup-

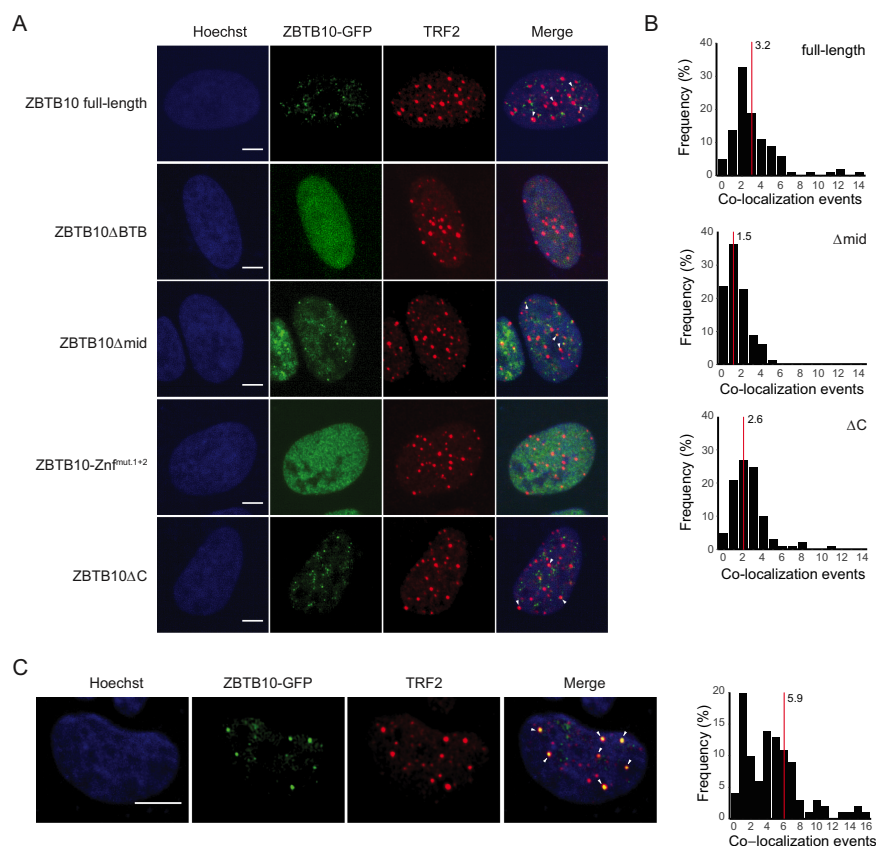


Figure 2. Cellular localization of ZBTB10 in U2OS cells. (A) Immunofluorescence staining of TRF2 in U2OS cells stably expressing the different GFP-tagged ZBTB10 deletion variants. DNA was stained with Hoechst. Scale bars represent 5 μ m. (B) Histogram of co-localization events between TRF2 and ZBTB10 full-length ($n = 103$ cells), ZBTB10 Δ mid ($n = 101$ cells) or ZBTB10 Δ C ($n = 100$ cells). The average number of co-localization events per cell is indicated with a red line. Due to the diffuse distribution of ZBTB10 Δ BTB and ZBTB10 Δ Zn^{mut.1+2} co-localization events could not be determined. (C) Immunofluorescence staining of TRF2 in G1-synchronized U2OS cells stably expressing full-length ZBTB10-GFP. DNA was stained with Hoechst. Scale bars represent 5 μ m. Quantification of co-localization events between TRF2 and ZBTB10 ($n = 101$ cells). The average number of co-localization events is marked with a red line.

plementary Figure S1C). To examine whether ZBTB10 foci co-localize with telomeres, we performed immunofluorescence staining of endogenous TRF2 as a bona fide marker for telomeres in unsynchronized U2OS ZBTB10-GFP cells. For full-length ZBTB10 we quantified on average three foci to co-localize with TRF2 showing that 19% of ZBTB10 foci are overlapping with telomeres (Figure 2A and B). To test whether we observe a difference in co-localization of ZBTB10 and TRF2 between ALT and telomerase-positive cells, we analyzed U2OS, GM847 and SAOS2 cells overexpressing ZBTB10-GFP as well as HeLa E1 and HEK293 cells. On average, co-localization events are slightly higher in the three ALT cell lines compared to the two telomerase-positive cell lines (Supplementary Figure S2A and B). Moreover, it is interesting to note that we quantified up to 20 co-localization events in ALT cells, whereas we found a maximum of five events per nucleus in HeLa E1

or HEK293 cells (Supplementary Figure S2A and B). When we analyzed co-localization of ZBTB10 deletion constructs and TRF2, we find that distinct foci are completely lost upon removal of the dimerization domain (ZBTB10 Δ BTB) or mutation of both zinc fingers (ZBTB10 Δ Zn^{mut.1+2}) and thus, co-localization with TRF2 cannot be determined (Figure 2A). In accordance with our *in vitro* DNA-binding data, deletion of the middle region (ZBTB10 Δ mid) and the C-terminus (ZBTB10 Δ C) only moderately affected the number of ZBTB10 and TRF2 co-localization events (Figure 2A and B). The co-localization of full-length ZBTB10-GFP with TRF2 was increased to on average six overlapping foci in G1 cells that were synchronized at the G2/M border and released into cell cycle for 4.5 h (Figure 2C). To further study ZBTB10 nuclear localization during cell cycle progression, we performed live-cell imaging for 40 h in the stable U2OS cell line (Supplementary Figure S2C and Supple-

mentary Video). We observed dynamic changes in foci number, intensity and size, which did not correlate with a specific cell cycle phase. However, we monitored absence of foci during mitosis and foci formation in both daughter cells after cell division.

As conventional microscopy has limited capacity in further resolving the localization, we performed super-resolution microscopy based on single molecule localization. For the TRF1/TRF2 complex, serving as a positive control, we measured distances ranging from 1 to 200 nm with an average of 41 nm (Figure 3A–C). In comparison, ZBTB10/TRF2 pairs revealed a higher diversity in measured distances ranging from 14 to 760 nm with an average of 267 nm (Figure 3A, C and D). The observed distances between ZBTB10/TRF2 in comparison to TRF1/TRF2 pairs indicate that a fraction of ZBTB10 (measured distances < 200 nm) co-localizes with telomeres.

Depletion of ZBTB10 showed only few changes on transcript level

To study the effect of the putative transcription factor ZBTB10 on gene expression, we generated CRISPR/Cas9 knockout (KO) cells using three different guide RNAs targeting the second exon of ZBTB10 in two cell lines: HEK293 and U2OS (Supplementary Figure S3A). KO clones were validated by Western blot (Supplementary Figure S3B) and sequencing of genomic DNA (Supplementary Figure S3C). To compare transcript levels between WT and KO, we performed RNA-seq analysis using at least three cell clones per cell line (Supplementary Figure S3D, Supplementary Tables S2 and S3). Differential expression analysis was performed for 16,182 genes in HEK293 and 17,936 genes in U2OS cells. When employing a threshold of $\log_2(\text{fold change}) > 1$ and $-\log_{10}(\text{ad. } P\text{-value}) > 2$, we detected only a small number of differentially regulated genes: 39 up- and 17 downregulated genes in HEK293 cells as well as five up- and two downregulated in U2OS cells. Our data therefore shows very few cell line-specific transcriptional changes upon ZBTB10 knockout and further bioinformatics analysis did not reveal any functionally enriched category.

ZBTB10 is not involved in telomere homeostasis

To probe a role of ZBTB10 in telomere homeostasis, we performed telomerase activity measurements using the quantitative telomeric repeat amplification protocol in telomerase-positive HEK293 cells, C-circle assay in telomerase-negative U2OS cells and telomere restriction fragment analysis in both cell lines. Using lysates of the five biological HEK293 WT and KO clones we did not detect differences in telomerase activity in three independent experiments (Supplementary Figure S4A). Additionally, for HEK293 WT clones we measured an average telomere restriction fragment length of 3.5 kb in comparison to a slightly shorter telomere fragment length of 2.8 kb in HEK293 KO clones that did not reach statistical significance (Supplementary Figure S4B). Similarly, no difference in telomere restriction fragment length was observed in U2OS WT and KO cells after culturing them for five months

(Supplementary Figure S4C). Concomitantly, quantitative telomeric fluorescence *in situ* hybridization experiments to assess telomere length in WT and KO clones of HEK293 and U2OS cells did not reveal changes (Supplementary Figure S4D and E). To analyze possible ALT phenotypes, we quantified extrachromosomal telomeric DNA by C-circle assay (Supplementary Figure S4F) and counted ALT-associated promyelocytic leukaemia bodies (APBs) (Supplementary Figure S4G). For both, we did not observe any significant changes between WT and KO U2OS cells. These data suggest that ZBTB10 is not involved in the assayed telomeric functions.

ZBTB10 interacts with TRF2 and RAP1

We then performed GFP pulldowns coupled to quantitative label-free proteomics in HEK293 and U2OS stable ZBTB10-GFP cells to gain insights into the molecular function of ZBTB10 (Figure 4A, Supplementary Figure S5A and B, Supplementary Table S4). We reproducibly detected (in at least two of three experiments per cell line) 42 proteins enriched in HEK293, 34 proteins enriched in U2OS and 38 commonly enriched proteins delineating a set of cell line-specific and common interaction partners (Figure 4B and C). Performing functional annotation clustering on the 37 proteins putatively interacting with ZBTB10, 27 of them formed a network with five high confidence clusters (EASE score > 3) that share either annotated protein domains, specific gene ontology descriptors or KEGG pathway terms (Supplementary Figure S5C, Supplementary Table S5). Four of these clusters are associated with RNA-related terms, originating from several proteins involved in cleavage and polyadenylation of pre-mRNA 3' ends, splicing regulation and RNA binding. In addition to RNA-associated proteins, we found the core transcription factor II H complex (GTF2H), which plays a role in nucleotide excision repair of DNA and when bound to the CAK complex, it functions in RNA polymerase II-dependent transcription (28). Moreover, we identified several subunits of the PTW/PP1 phosphatase complex, which plays a role in the control of chromatin structure, RNA processing, and cell cycle progression during the transition from mitosis to interphase (29). Our interaction data indicates that ZBTB10 functions in several cellular processes and might be involved in post-transcriptional activities via interaction with a diverse set of proteins. In addition to these factors, we also co-purified the shelterin complex members TRF2 and RAP1 with very high fold enrichments (Figure 4C). To note, none of the other shelterin subunits were detected in any of our experiments implying that ZBTB10 only interacts with the TRF2/RAP1 heterodimer (30). As observed in telomerase-positive HEK293 and telomerase-negative U2OS cells, the interaction is independent of the telomere maintenance mechanism. To investigate whether loss of ZBTB10 influences TRF2 levels, we compared protein abundance by Western blot and IF staining but could not detect differences in HEK293 or U2OS cells (Supplementary Figure S6A and B). To nevertheless exclude putative effects on telomere end protection, we analyzed chromosomal rearrangements on metaphase chromosomes, changes in the occurrence of telomere-induced foci (TIF) and ATM kinase

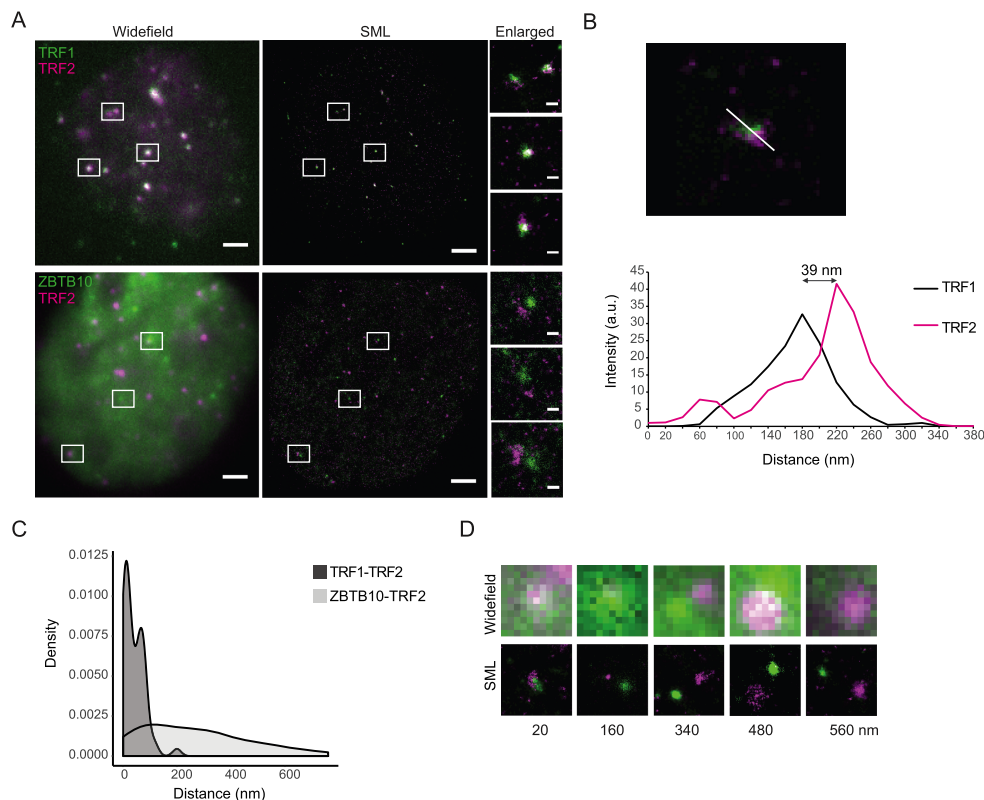


Figure 3. Super-resolution microscopy (GSDIM/dSTORM) and nanometer resolved intermolecular two-color distance analysis of ZBTB10-GFP stably expressed in U2OS cells in relation to TRF2. (A) Dual-color immunostaining of TRF1 (green)/TRF2 (magenta) or ZBTB10-GFP (green)/TRF2 (magenta) demonstrate the different interrelation between the respective protein pairs. Widefield images (left) and corresponding single molecule localizations (SML) rendered as Normalized Gaussian distribution with a localization precision of $15\text{--}20 \pm 6$ nm. Scale bars represent $2 \mu\text{m}$ and 200 nm in enlarged SMLM images. (B) Intermolecular distances were measured in the reconstructed two-color images by drawing an intensity line profile (in arbitrary units) crossing corresponding protein clusters in the respective channels and recording the distances between the intensity maxima of Atto488 and Alexa647. In this example, the measured distance is 39 nm. (C) Density plot distribution of peak-to-peak distances measured between TRF1/TRF2 pairs ($n = 58$) with on average 41 nm and ZBTB10/TRF2 ($n = 67$) pairs with on average 267 nm in reconstructed SMLM images. Image reconstruction and shift correction with a localization precision of $30\text{--}40$ nm. (D) Examples of selected ZBTB10 (green)/TRF2 (magenta) pairs with measured peak-to-peak distances demonstrate the capabilities of SML (lower panel) to further differentiate between co-localization and co-distribution events compared to conventional light microscopy (upper panel). Images are $1 \mu\text{m} \times 1 \mu\text{m}$.

activation (Supplementary Figure S6C–G). In none of these assays, we could detect a difference between ZBTB10 KO and WT U2OS cells. Overall, we show that ZBTB10 does not influence TRF2 protein stability or its role in end protection.

ZBTB10 interacts with TRF2 through its N-terminal domain

Prior to mapping the interaction between ZBTB10 and the TRF2/RAP1 complex, we validated TRF2 and RAP1 enrichment in the ZBTB10-GFP IP by immunostaining and show that the interaction is not mediated by RNA or DNA binding as Sm-nuclease treatment did not abrogate TRF2/RAP1 enrichment (Figure 5A, Supplementary Figure S7A). To test if the ZBTB10-GFP/TRF2 interaction occurs in the nucleoplasm or in the context of chromatin, we

prepared nuclear and chromatin extracts from HEK293 and U2OS cells stably expressing ZBTB10-GFP. While TRF2 is equally abundant in both subcellular fractions, the amount of chromatin-bound ZBTB10 is $6\times$ less in HEK293 cells and even $34\times$ less in U2OS cells (Figure 5B and C). GFP pulldowns of nuclear and chromatin extracts showed similar stoichiometry of TRF2 and ZBTB10 (Figure 5B and C) revealing that their interaction is also stable in nucleoplasmic extracts and does not require telomeric chromatin. To determine the interaction site of TRF2 required for binding to ZBTB10, we generated various deletion constructs of TRF2: Δ Basic (aa $\Delta 1\text{--}81$; N-terminal domain), Δ TRFH (aa $\Delta 83\text{--}290$; the dimerization domain), Δ Hinge (aa $\Delta 293\text{--}482$) and Δ Myb (aa $\Delta 483\text{--}542$; the DNA-binding domain) (Figure 5D) (31). We also constructed a TRF2 Δ RAP1 site (aa $\Delta 328\text{--}341$) (32) construct to evaluate whether RAP1

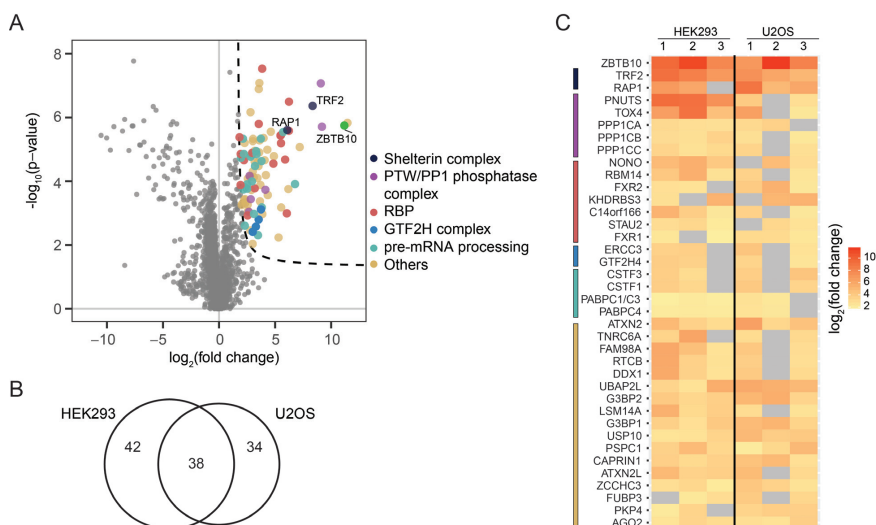


Figure 4. Identification of ZBTB10 interaction partners. (A) Volcano plot of a representative GFP pulldown from HEK293 cells stably expressing ZBTB10-GFP using quantitative mass spectrometry. GFP transfected cells served as control and each experiment was performed in quadruplicates. Proteins co-purifying with ZBTB10 are determined by $S_0 > 1.6$, $c = 0.8$ and $p < 0.05$ (dotted line) and colored according to functional description or complex membership (see legend). (B) Venn diagram to show overlap of proteins found enriched in at least two out of three independent experiments from U2OS and HEK293 cells. (C) Heat map illustrating enrichment of the 37 common interactors and ZBTB10. Interaction partners are grouped according to their functional description or complex membership: shelterin complex (dark blue), PTW/PP1 phosphatase complex (purple), RBPs (red), GTF2H complex (blue), pre-mRNA processing (green) and others (beige). Grey squares represent no enrichment in the respective pulldown.

binding to TRF2 was required for the interaction with ZBTB10. Except for TRF2 Δ Basic and TRF2 Δ TRFH, all TRF2 variants co-immunoprecipitated with ZBTB10-GFP (Figure 5E) revealing that the interaction between ZBTB10 and TRF2 is mediated via the Basic and TRFH domains of TRF2. Indeed, a minimal construct of the N-terminal region of TRF2 missing the Hinge and Myb domain (TRF2 Δ Hinge Δ Myb; aa 1–292) was sufficient to be co-IPed by ZBTB10-GFP (Figure 5F). These results establish that the interaction of ZBTB10 with TRF2 does not require the DNA-binding domain of TRF2 and is independent of RAP1.

DISCUSSION

Our previous phylointeractomics screen suggested that ZBTB10 can associate with the telomeric repeat sequence. Indeed, we here show that the dsTTAGGG motif is recognized by the ZBTB10 minimum DNA-binding fragment with similar affinity ($K_D = 218$ nM) as the recently described telomeric protein ZBTB48/TZAP ($K_D = 170$ nM) (33). Unlike ZBTB48/TZAP that surprisingly uses a single zinc finger (Znf11) aided by an adjacent C-terminal arm to bind telomeric DNA (19,33), we demonstrate that the two classical Cys₂His₂ zinc fingers of ZBTB10 combined are necessary and sufficient for binding. Moreover, when we investigated the binding ability of ZBTB10 at telomeric variant repeats, we found that ZBTB10 exhibited a two-fold stronger binding to the TTGGGG motif ($K_D = 106$ nM) than to the TTAGGG motif, differentiating it from ZBTB48/TZAP, which has less affinity for this telomeric

variant repeat ($K_D = 380$ nM) (33). We further reveal that the additional atypical zinc finger in the C-terminal region strongly enhances ZBTB10 binding affinity to telomeric DNA with a K_D value of 37 nM for the TTGGGG motif. In the same way, a cooperative binding mode for Zif268 has been reported by addition of a third zinc finger that strongly increased binding affinity (34). We thus identified ZBTB10 as a new telomere-binding protein with a preference for the TTGGGG motif. Similarly to ZBTB10, the nuclear receptor NR2C2 has been characterized to preferably bind the TCAGGG sequence with a two-fold higher binding affinity ($K_D = 28$ nM) compared to the telomeric motif ($K_D = 58$ nM) (16). While variant repeat sequences are restricted to the proximal 2 kb of telomeres in telomerase-positive cells, they are more frequently occurring in the longer heterogeneous telomeres of ALT cells due to homology-directed telomere maintenance (16). Concomitantly with the *in vitro* affinity of ZBTB10 to telomeric repeats, we observe colocalization of ZBTB10 and telomeres. There seems to be a subset of ALT cells that have more than five co-localization events compared to the telomerase-positive cell lines.

Our additional experiments using HEK293 and U2OS ZBTB10 knockout cells demonstrated that absence of ZBTB10 has no effect on telomere homeostasis in telomerase-positive or ALT cells for multiple hallmarks such as telomere length, C-circle abundance and APB frequency. In our phylointeractomics screen, we identified several TTAGGG-associated proteins and we cannot exclude a redundant function of ZBTB10 with any of them. However, using quantitative interactomics, we identify TRF2

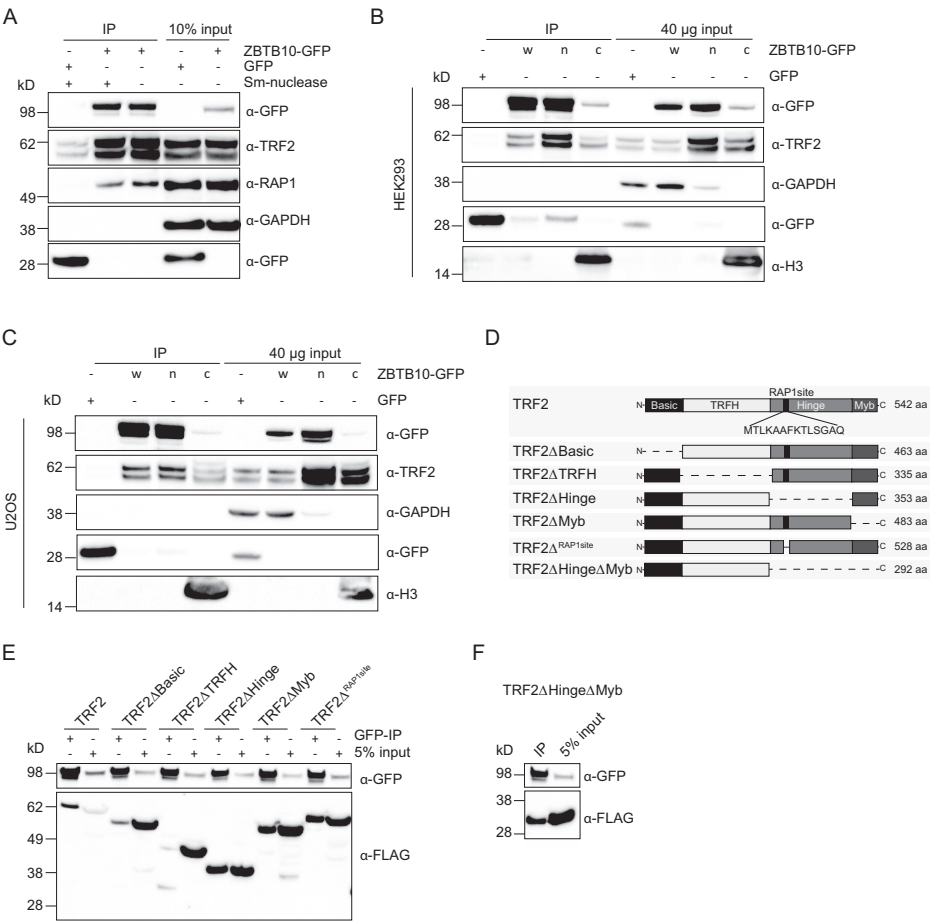


Figure 5. Mapping of the interaction site between ZBTB10 and TRF2. (A) Western blot validation of endogenous TRF2 and RAP1 co-purifying with ZBTB10 from stable HEK293 cells with and without Sm-nuclease treatment. GFP transfected cells served as control pulldown and GAPDH was used as a loading control. (B) GFP pulldowns using 1 mg whole lysate (w), 400 µg nuclear (n) and 400 µg chromatin (c) extracts of HEK293 cells stably expressing ZBTB10-GFP. GFP transfected cells served as control pulldown. Histone H3 served as a marker protein for the chromatin fraction. (C) GFP pulldowns using 1 mg whole lysate (w), 400 µg nuclear (n) and 400 µg chromatin (c) extracts of U2OS cells stably expressing ZBTB10-GFP. GFP transfected cells served as control pulldown. Histone H3 served as a marker protein for the chromatin fraction. (D) Schematic overview of constructed TRF2 deletions with protein length in amino acids (aa). (E) GFP pulldowns from HEK293 cells stably expressing ZBTB10-GFP in combination with different transiently overexpressed FLAG-TRF2 variants pinpoint the N-terminal region of TRF2 as the interaction site. (F) Co-immunoprecipitation of ZBTB10-GFP with the minimal FLAG-TRF2ΔHingeΔMyb construct from stable HEK293 cells.

and RAP1 as interaction partners of ZBTB10 and we show that ZBTB10 interacts with the N-terminal region of TRF2 in a DNA and RNA-independent manner. The interactome further revealed an interaction with the GTF2H complex and the PTW/PP1 phosphatase complex. Interestingly, the regulatory subunit of this complex, PNUTS, was highly enriched in our pulldowns and has previously been reported as an interactor of TRF2 and shown to co-localize with a subset of telomeres (35). In a recent quantitative proximity ligation assay to decipher the DNA repair network, ZBTB10 showed between 1.50- and 1.74-fold enrichment

with all three APEX2-tagged DNA repair proteins: 53BP1, MDC1 and BRCA1 (36). Previously, ZBTB10 has also been found in a TAP-MS screen to possibly interact with the BRCT-domain of 53BP1 (37). Together with the interaction of PNUTS that influences 53BP1-mediated DNA repair (38), it is tempting to speculate that ZBTB10 in cooperation with the TRF2/RAP1 complex is involved in DNA damage repair at telomeres, especially as ZBTB10 has been shown to localize to sites of damaged chromatin after UV laser microirradiation (39).

DATA AVAILABILITY

The mass spectrometry proteomics data has been deposited to the ProteomeXchange Consortium via the PRIDE (40) partner repository with the dataset identifier PXD010672. The sequencing data has been deposited in the NCBI Gene Expression Omnibus database with the accession number GSE117214.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Assistance by the following IMB facilities is gratefully acknowledged: Microscopy Core Facility and the use of its Opera Phenix (INST 247/845-1 FUGG), Cytometry Core Facility, Proteomics Core Facility, Genomics Core Facility and Media Lab. We further thank Joachim Lingner for providing the HeLa E1 cell line.

FUNDING

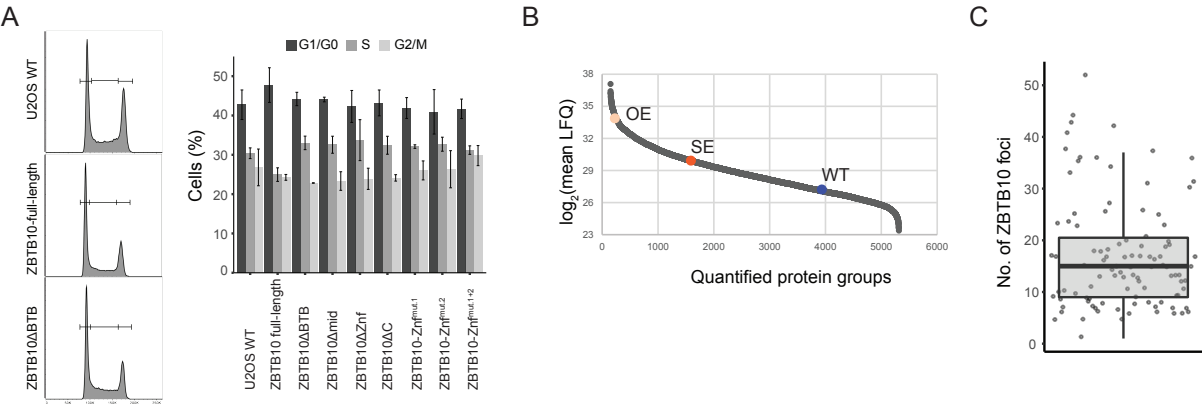
Deutsche Forschungsgemeinschaft [Bu2996/1-2]; Research in the Kappei laboratory was supported by the National Research Foundation Singapore and the Singapore Ministry of Education under its Research Centres of Excellence initiative and by the RNA Biology Center at the Cancer Science Institute of Singapore, NUS, as part of funding under the Singapore Ministry of Education's AcRF Tier 3 grants [MOE2014-T3-1-006]; Research in the Shi laboratory was supported by The Strategic Priority Research Program (Pilot study) 'Biological basis of aging and therapeutic strategies' of the Chinese Academy of Sciences [XDPB10]. Funding for open access charge: Deutsche Forschungsgemeinschaft [Bu2996/1-2].

Conflict of interest statement. None declared.

REFERENCES

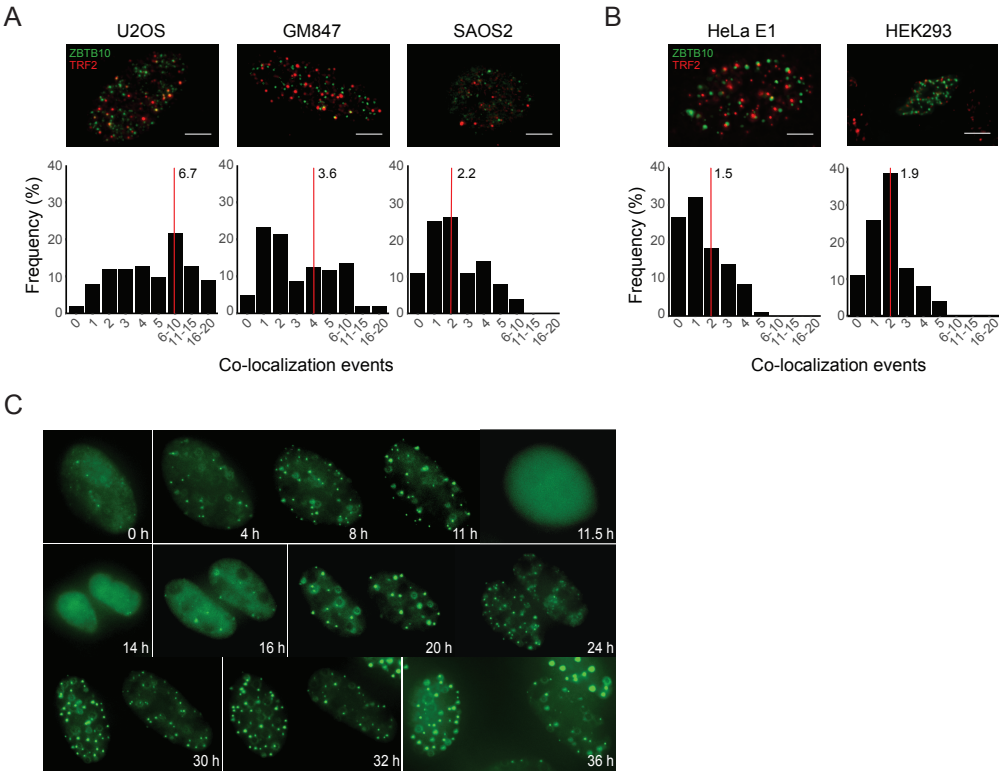
- Palm, W. and de Lange, T. (2008) How shelterin protects mammalian telomeres. *Annu. Rev. Genet.*, **42**, 301–334.
- Maciejowski, J. and de Lange, T. (2017) Telomeres in cancer: tumour suppression and genome instability. *Nat. Rev. Mol. Cell Biol.*, **18**, 175–186.
- Doksani, Y., Wu, J.Y., de Lange, T. and Zhuang, X. (2013) Super-resolution fluorescence imaging of telomeres reveals TRF2-dependent T-loop formation. *Cell*, **155**, 345–356.
- Griffith, J.D., Comeau, L., Rosenfield, S., Stansel, R.M., Bianchi, A., Moss, H. and de Lange, T. (1999) Mammalian telomeres end in a large duplex loop. *Cell*, **97**, 503–514.
- van Steensel, B., Smogorzewska, A. and de Lange, T. (1998) TRF2 protects human telomeres from end-to-end fusions. *Cell*, **92**, 401–413.
- Rai, R., Chen, Y., Lei, M. and Chang, S. (2016) TRF2-RAP1 is required to protect telomeres from engaging in homologous recombination-mediated deletions and fusions. *Nat. Commun.*, **7**, 1–13.
- Karlseder, J., Hoke, K., Mirzoeva, O.K., Bakkenist, C., Kastan, M.B., Petrini, J.H.J. and de Lange, T. (2004) The telomeric protein TRF2 binds the ATM Kinase and Can Inhibit the ATM-dependent DNA damage response. *PLoS Biol.*, **2**, 1150–1156.
- Denchi, E.L. and de Lange, T. (2007) Protection of telomeres through independent control of ATM and ATR by TRF2 and POT1. *Nature*, **448**, 1068–1071.
- Sarek, G., Vannier, J., Panier, S., Petrini, J.H.J. and Boulton, S.J. (2015) TRF2 recruits RTEL1 to telomeres in S phase to promote T-Loop unwinding. *Mol. Cell*, **57**, 622–635.
- Higa, M., Kushiya, T., Kurashige, S., Kohmon, D., Enokitani, K., Iwahori, S., Sugimoto, N., Yoshida, K. and Fujita, M. (2017) TRF2 recruits ORC through TRFH domain dimerization. *BBA - Mol. Cell Res.*, **1864**, 191–201.
- Kim, N.W., Piatyszek, K.R., Harley, C.B., West, M.D., Ho, P.L., Coviello, G.M., Wright, W.E., Weinrich, S.L., Shay, J.W. and Prowse, M.A. (1994) Specific association of human telomerase activity with immortal cells and cancer. *Science*, **226**, 2011–2015.
- Bryan, T.M., Englezou, A., Gupta, J., Bacchetti, S. and Reddel, R.R. (1995) Telomere elongation in immortal human cells without detectable telomerase activity. *Embo. J.*, **14**, 4240–4248.
- Dilley, R.L., Verma, P., Cho, N.W., Winters, H.D., Wondisford, A.R. and Greenberg, R.A. (2016) Break-induced telomere synthesis underlies alternative telomere maintenance. *Nature*, **539**, 54–58.
- Sobinoff, A.P. and Pickett, H.A. (2017) Alternative lengthening of Telomeres: DNA repair pathways converge. *Trends Genet.*, **33**, 921–932.
- Lee, M., Hills, M., Conomos, D., Stutz, M.D., Dagg, R.A., Lau, L.M.S., Reddel, R.R. and Pickett, H.A. (2014) Telomere extension by telomerase and ALT generates variant repeats by mechanistically distinct processes. *Nucleic Acids Res.*, **42**, 1733–1746.
- Conomos, D., Stutz, M.D., Hills, M., Neumann, A.A., Bryan, T.M., Reddel, R.R. and Pickett, H.A. (2012) Variant repeats are interspersed throughout the telomeres and recruit nuclear receptors in ALT cells. *J. Cell Biol.*, **199**, 893–906.
- Marzec, P., Armenise, C., Pérot, G., Roumelioti, F., Basyuk, E., Gagos, S., Chibon, F. and Déjardin, J. (2015) Nuclear-Receptor-Mediated telomere insertion leads to genome instability in ALT cancers. *Cell*, **160**, 913–927.
- Perez-Torrado, R., Yamada, D. and Defossez, P.A. (2006) Born to bind: The BTB protein-protein interaction domain. *BioEssays*, **28**, 1194–1202.
- Jahn, A., Rane, G., Paszkowski-Rogacz, M., Sayols, S., Bluhm, A., Han, C., Drašković, I., Londoño-Vallejo, J.A., Kumar, A.P., Buchholz, F. et al. (2017) ZBTB48 is both a vertebrate telomere-binding protein and a transcriptional activator. *EMBO Rep.*, **18**, 929–946.
- Li, J.S.Z., Fusté, J.M., Simavorian, T., Bartocci, C., Tsai, J., Karlseder, J. and Denchi, E.L. (2017) TZAP: A telomere-associated protein involved in telomere length control. *Science*, **355**, 638–641.
- Kappei, D., Scheibe, M., Paszkowski-Rogacz, M., Bluhm, A., Gossmann, T.I., Dietz, S., Dejung, M., Herlyn, H., Buchholz, F., Mann, M. et al. (2017) Phylotranscriptomics reconstructs functional evolution of protein binding. *Nat. Commun.*, **8**, 14334.
- Cristofari, G. and Lingner, J. (2006) Telomere length homeostasis requires that telomerase levels are limiting. *EMBO J.*, **25**, 565–574.
- Shevchenko, A., Tomas, H., Havliš, J., Olsen, J. V. and Mann, M. (2007) In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.*, **1**, 2856–2860.
- Fölling, J., Bossi, M., Bock, H., Medda, R., Wurm, C.A., Hein, B., Jakobs, S., Eggeling, C. and Hell, S.W. (2008) Fluorescence nanoscopy by ground-state depletion and single-molecule return. *Nat. Methods*, **5**, 943–945.
- Van De Linde, S., Löschberger, A., Klein, T., Heidbreder, M., Wolter, S., Heilemann, M. and Sauer, M. (2011) Direct stochastic optical reconstruction microscopy with standard fluorescent probes. *Nat. Protoc.*, **6**, 991–1009.
- Swoboda, M., Henig, J., Cheng, H.M., Brugger, D., Haltrich, D., Plumeré, N. and Schlierf, M. (2012) Enzymatic oxygen scavenging for photostability without pH drop in single-molecule experiments. *ACS Nano*, **6**, 6364–6369.
- Ovesný, M., Křížek, P., Borkovec, J., Švindrych, Z. and Hagen, G.M. (2014) ThunderSTORM: a comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging. *Bioinformatics*, **30**, 2389–2390.
- Compe, E. and Egly, J.-M. (2016) Nucleotide excision repair and transcriptional regulation: TFIIH and beyond. *Annu. Rev. Biochem.*, **85**, 265–290.
- Lee, J.H., You, J., Dobrota, E. and Skalknik, D.G. (2010) Identification and characterization of a novel human PP1 phosphatase complex. *J. Biol. Chem.*, **285**, 24466–24476.

30. Bae,N.S. and Baumann,P. (2007) A RAP1/TRF2 complex inhibits nonhomologous end-joining at human telomeric DNA ends. *Mol. Cell*, **26**, 323–334.
31. Arnoult,N. and Karlseder,J. (2015) Complex interactions between the DNA-damage response and mammalian telomeres. *Nat. Struct. Mol. Biol.*, **22**, 859–866.
32. Okamoto,K., Bartocci,C., Ouzounov,I., Diedrich,J.K., Iii,J.R.Y. and Denchi,E.L. (2013) A two-step mechanism for TRF2-mediated chromosome-end protection. *Nature*, **494**, 502–505.
33. Zhao,Y., Zhang,G., He,C., Mei,Y., Shi,Y. and Li,F. (2018) The 11th C2H2 zinc finger and an adjacent C-terminal arm are responsible for TZAP recognition of telomeric DNA. *Cell Res.*, **28**, 130–134.
34. Lee,J., Kim,J. and Seok,C. (2010) Cooperativity and specificity of Cys 2 His 2 zinc finger protein–DNA interactions: a molecular dynamics simulation study.
35. Kim,H., Lee,O.-H., Xin,H., Chen,L.-Y., Qin,J., Chae,H.K., Lin,S.-Y., Safari,A., Liu,D. and Songyang,Z. (2009) TRF2 functions as a protein hub and regulates telomere maintenance by recognizing specific peptide motifs. *Nat. Struct. Mol. Biol.*, **16**, 372–379.
36. Gupta,R., Somyajit,K., Narita,T., Maskey,E., Stanlie,A., Kremer,M., Typas,D., Lammers,M., Mailand,N., Nussenzweig,A. *et al.* (2018) DNA repair network analysis reveals shieldin as a key regulator of NHEJ and PARP inhibitor sensitivity. *Cell*, **173**, 972–974.
37. Woods,N.T., Mesquita,R.D., Sweet,M., Carvalho,M.A., Li,X., Liu,Y., Nguyen,H., Thomas,C.E., Iversen,E.S. Jr, Marsillac,S. *et al.* (2012) Charting the landscape of tandem BRCT domain-mediated protein interactions. *Sci. Signal*, **5**, rs6.
38. Landsverk,H.B., Mora-Bermúdez,F., Landsverk,O.J.B., Hasvold,G., Naderi,S., Bakke,O., Ellenberg,J., Collas,P., Syljuåsen,R.G. and Küntziger,T. (2010) The protein phosphatase 1 regulator PNUTS is a new component of the DNA damage response. *EMBO Rep.*, **11**, 868–875.
39. Izhar,L., Adamson,B., Ciccio,A., Lewis,J., Pontano-vaite,L., Liang,A.C., Westbrook,T.F., Harper,J.W. and Stephen,J. (2015) A Systematic Analysis of Factors Localized to Damaged Chromatin Reveals PARP-Dependent Recruitment of Transcription Factors. *Cell Rep.* **11**, 1486–1500.
40. Vizcaino,J.A., Csordas,A., del-Toro,N., Dianes,J.A., Griss,J., Lavidas,I., Mayer,G., Perez-Riverol,Y., Reisinger,F., Ternent,T. *et al.* (2016) 2016 update of the PRIDE database and related tools. *Nucleic Acids Res.*, **44**, D447–D456.



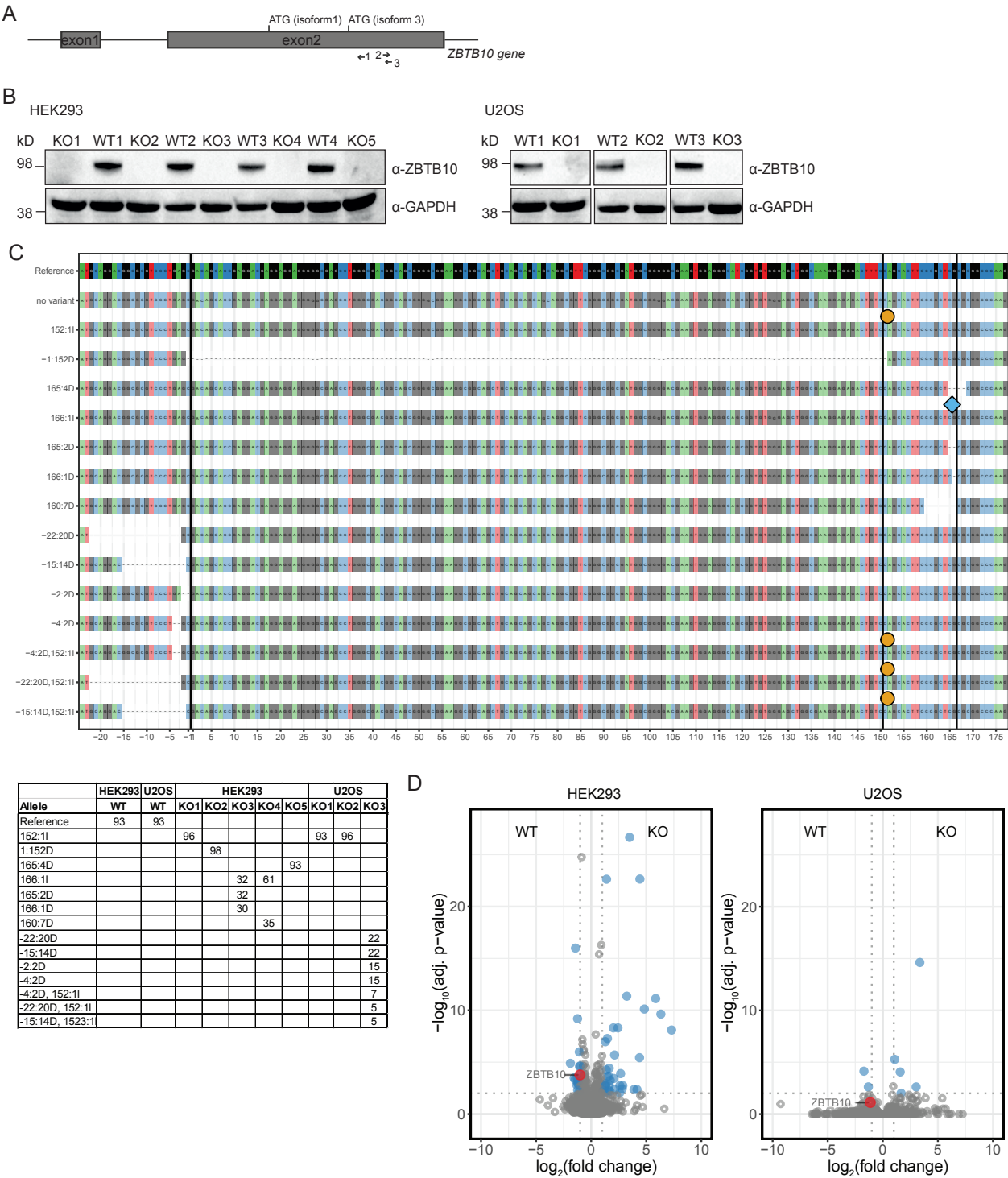
Supplementary Figure 1

Comparison of ZBTB10 constructs and nuclear localization of ZBTB10. **(A)** Cell cycle distribution of U2OS WT cells in comparison to U2OS cells stably expressing ZBTB10-GFP fusion proteins. Representative histograms are shown on the left. The experiment was performed in three independent replicates. Error bars represent S.E.M. **(B)** Abundance rank of quantified proteins determined in U2OS cells by label-free proteomics. Endogenous ZBTB10 (WT), stably expressed (SE) and transiently overexpressed (OE) ZBTB10-GFP are highlighted with $\log_2$ LFQ intensity values of 27 (WT), 30.3 (SE) and 33.8 (OE). **(C)** Number of ZBTB10-GFP foci per nucleus in stable U2OS cells, which were used to determine co-localization with TRF2 in Figure 2A (n=101).



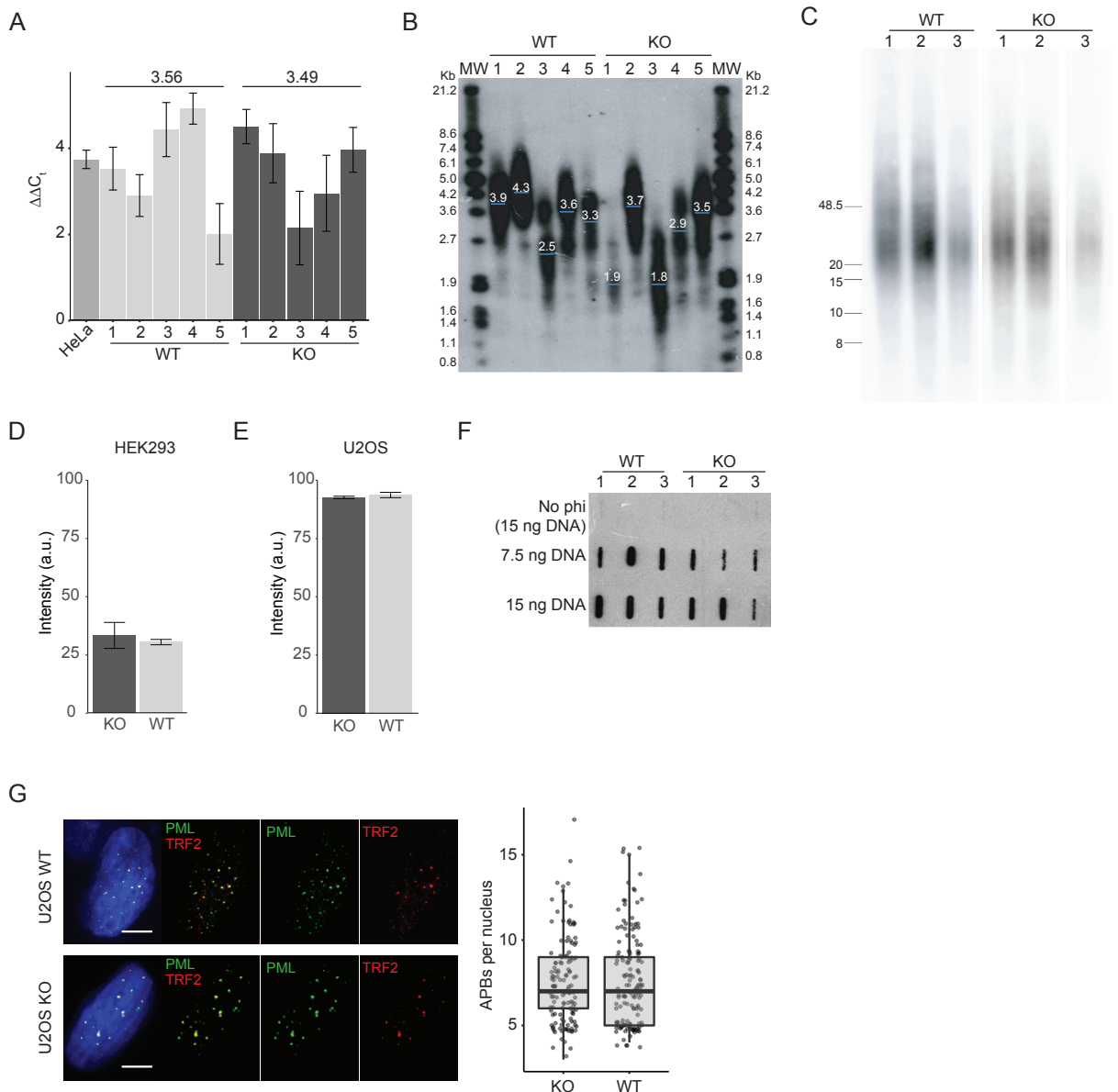
Supplementary Figure 2

Localization of ZBTB10 in different cell lines. **(A)** Representative images are shown for the three ALT cell lines U2OS, GM847 and SAOS2 with transiently overexpressed ZBTB10-GFP (green) and TRF2 IF (red). Scale bars represent 5 μ m. Quantification of ZBTB10-GFP foci overlapping with TRF2 for each cell line ($n \geq 100$ cells). The average of co-localization events is indicated by a red line. **(B)** Representative images are shown for the two telomerase-positive cell lines HeLa E1 and HEK293. ZBTB10-GFP (green) was transiently overexpressed and TRF2 (red) was stained by IF. Scale bars represent 5 μ m. Quantification of ZBTB10-GFP foci overlapping with TRF2 for each cell line ($n \geq 100$ cells). The average of co-localization events is indicated by a red line. **(C)** Images of a representative U2OS cell stably expressing ZBTB10-GFP at indicated time points during 40 hours live-cell imaging.



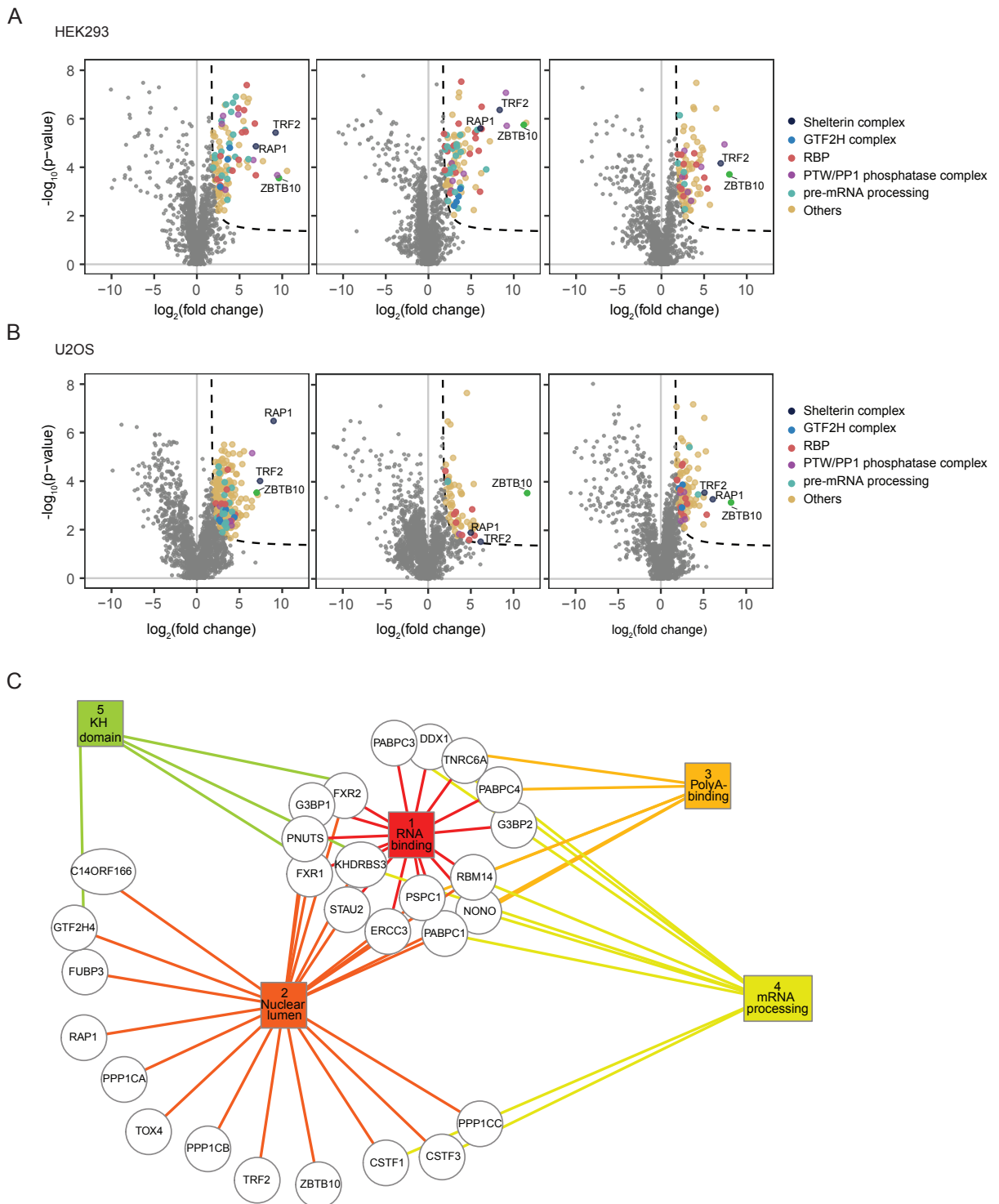
Supplementary Figure 3

Characterization of ZBTB10 knockout cells. **(A)** Three different guide RNAs were used to induce a cut in the second exon of the ZBTB10 gene. Arrows indicate guide RNAs position and direction. **(B)** Western blot validation for ZBTB10 KO clones in HEK293 and U2OS cells. GAPDH served as a loading control. **(C)** Analysis of the genomic modifications in ZBTB10 clones by next generation sequencing. The sequences of KO clones compared to the reference genes are plotted: A (green), C (blue), G (grey) and T (red). Deletions are indicated by a dotted line, whereas nucleotide insertions are indicated by an orange circle for adenine or a blue diamond for guanine. Calculated allele frequencies (in %) for each KO clone are noted in the table below. **(D)** Differential gene expression analysis of HEK293 and U2OS WT and KO cells (adjusted p-value 0.01). Significantly enriched genes with $\log_2(\text{fold change}) > 1$ and a $-\log_{10}(\text{adj. p-value}) > 2$ are plotted in blue.



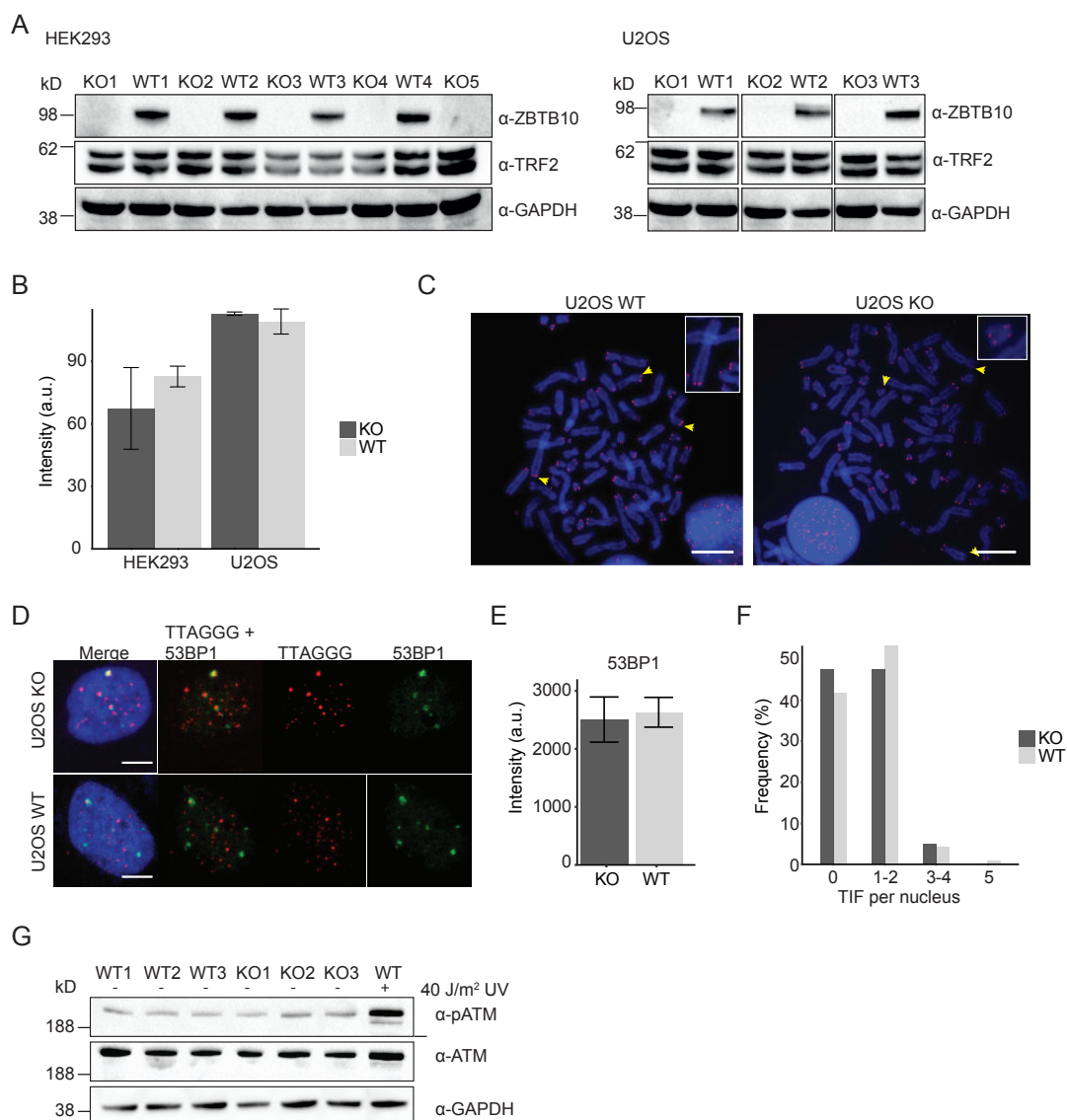
Supplementary Figure 4

Analysis of telomere maintenance mechanisms in HEK293 and U2OS cells. **(A)** Detection of telomerase activity in HEK293 clones using a quantitative TRAP assay. HeLa cells served as a positive control. Differences after normalization of Ct values against IMR90 cells are shown. Three independent experiments were performed and error bars represent standard deviation. WT and KO clones do not show significant changes (p-value 0.92). **(B)** Telomere restriction fragment analysis of HEK293 cell clones with an average telomere length of 3.5 kb for WT and 2.8 kb for KO clones (p-value 0.16). Fragment lengths in kb are indicated for each cell clone. **(C)** Telomere restriction fragment analysis of U2OS cell clones after five months culturing. **(D)** Quantitative fluorescence *in situ* hybridization for telomeres in WT and KO HEK293 cell clones. Bar plot representing the calculated average of telomeric signal intensities (in arbitrary units) for HEK293 WT 30.4 ± 1.2 a.u. (n=38,700 foci) and KO 33.2 ± 5.6 a.u. (n=29,047 foci). Error bars represent standard deviation. **(E)** Quantitative fluorescence *in situ* hybridization for telomeres in U2OS WT and ZBTB10 KO clones. Bar plot representing the calculated average of telomeric signal intensities (in arbitrary units) for WT 93.7 ± 1.2 a.u. (n=4,177 foci) and KO 92.7 ± 0.5 a.u. (n=3,734 foci). Error bars represent standard deviation. **(F)** Detection of extrachromosomal DNA in WT and KO U2OS clones by C-circle assay showed no significant difference (p-value 0.09 for 7.5 ng and 0.35 for 15 ng DNA). **(G)** PML (green) and TRF2 (red) immunofluorescence staining to quantify ALT-associated promyelocytic leukaemia bodies (APBs) in U2OS WT and KO cells. DNA was stained with DAPI. Scale bar represent 5 μ m. Quantification of APBs in U2OS WT (n=140) and ZBTB10 KO (n=129) cells with on average eight APBs per nucleus in both WT and KO cells.



Supplementary Figure 5

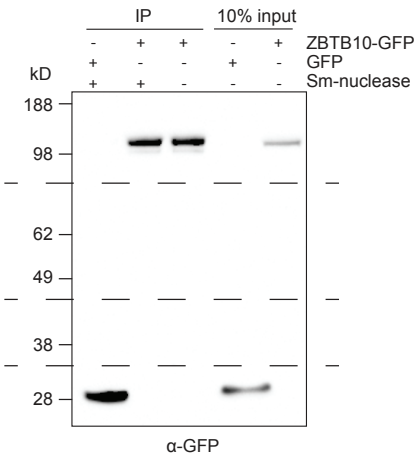
Interactomics data for ZBTB10. **(A)** Volcano plots of three independent GFP pulldown experiments using lysate of HEK293 cells stably expressing ZBTB10-GFP and GFP transfected control cells. Each experiment was performed in quadruplicates. Highlighted proteins are colored according to their described functions. The enrichment threshold ($S_0=1.6$, $c=0.8$, $p=0.05$) is indicated by a dotted line. **(B)** Volcano plots of three independent GFP pulldown experiments performed with lysate of U2OS cells stably expressing ZBTB10-GFP. GFP transfected U2OS cells served as a control. Each experiment was performed in quadruplicates. Highlighted proteins are colored according to their described functions and the enrichment threshold ($S_0=1.6$, $c=0.8$, $p=0.05$) is indicated by a dotted line. **(C)** 27 of the common ZBTB10 interactors (and ZBTB10 itself) were grouped into five high confidence clusters: RNA binding (red), nuclear lumen (orange), poly(A) binding (yellow), mRNA processing (light green) and KH domain (green).



Supplementary Figure 6

TRF2 levels are not affected upon ZBTB10 knockout. **(A)** Western blot analysis of TRF2 levels in HEK293 and U2OS cells. GAPDH served as a loading control. **(B)** Barplots of quantified TRF2 intensities in KO and WT clones of HEK293 and U2OS cells after immunofluorescence staining: for HEK293 KO 67.42 ± 19.61 ($n=2,016$ foci) and WT 82.73 ± 4.96 ($n=4,070$ foci), for U2OS KO 112.92 ± 0.79 ($n=32,027$ foci) and WT 109.14 ± 6.02 ($n=22,956$ foci). Error bars represent standard deviation. **(C)** Representative metaphase spreads of U2OS cells. Telomeres were stained using a TAMRA-labeled C-rich telomere probe and DNA was stained with DAPI. Yellow arrow heads indicate telomeric duplets that were scored in 50 metaphase spreads of each U2OS WT and KO cells. On average three telomeric duplets were scored per metaphase spread in both WT and KO cells. Scale bar represents 10 μ m. **(D)** 53BP1 immunofluorescence staining (green) coupled to telomeric FISH using a C-rich telomere PNA probe (red) in U2OS WT and ZBTB10 KO clones. DNA was stained with DAPI. Scale bar represents 5 μ m. **(E)** Barplot of quantified 53BP1 intensities in KO and WT clones of U2OS cells from (D). Error bars represent standard deviation. **(F)** Occurrence of telomere induced foci (TIF) in WT and ZBTB10 KO U2OS clones from (D). **(G)** Western blot analysis of ATM and phospho-ATM levels in U2OS WT and ZBTB10 KO cells. GAPDH served as a loading control.

A



Supplementary Figure 7

ZBTB10 interacts with TRF2. (A) Full-sized Western blot from Figure 5A. All membrane parts (indicated by the dotted line) were re-incubated with anti-GFP antibody and detected simultaneously.

Supplementary Methods

Cell cycle analysis

80% confluent cells were harvested, fixed with ice-cold 70% ethanol and stored at 4°C. After two PBS washes, cells were treated with RNaseA (40 µg/µL) and DNA was stained with propidium iodide (80 µg/µL) for 30 min at RT. DNA content measurements of at least 30,000 cells was done by flow cytometry using a LSRFortessa (Becton Dickinson). Cell cycle analysis was done using FlowJo V10 (TreeStar).

Genomic next generation sequencing

Genomic DNA of cell clones was extracted as described and a 1175 bp target region around the cut locus was amplified by PCR using specific primers, followed by a second PCR using P5 and P7 overhang primers (Supplementary Table 1). PCR products were purified using AMPure XP beads (Beckman Coulter) and DNA concentration was determined using the Qubit dsDNA HS Assay Kit (Invitrogen). In a third PCR index primers specific for each sample were used and DNA was purified with AMPure XP beads. All samples were pooled in equimolar ratio and sequenced on a MiSeq Nano Flowcell, paired-end for 2x159 cycles plus 7 cycles for the index read.

DNA-Seq measurements of HEK293 and U2OS WT and ZBTB10 KO samples yielded on average 68 K reads of 159 nt length per sample. We assessed the quality of the sequenced reads with fastqc (Andrews S. (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>). Adapter sequences were removed from both ends of both reads using cutadapt version 1.14 (1). Paired reads were merged using pandaseq version 2.11 (2) with the following parameters: -d BFSrk -A pear (on average 97% of paired reads were merged successfully). Merged reads (average size 288 nt) were mapped to chromosome 8 of the homo sapiens GRCh38 reference genome using gmap version 2017-02-15 (3) with the following parameters: --min-intron length=200 -f samse --nofails. For localization and visualization of the mutations we summarized mapped sequences using R version 3.4.3 (R Core Team (2017) R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. Available online at: <http://www.R-project.org/>) and CrispRVariants bioconductor package version 1.6.0 (4). In short, variants within the region of interest were localized and mutation rates of all alleles were calculated for each sample.

Transcriptome sequencing

Total RNA was extracted from cells using the RNeasy Mini Kit (Qiagen) according to manufacturer's instructions including on-column DNA digestion. Next generation sequencing library preparation was performed with Illumina's TruSeq stranded mRNA LT Sample Prep Kit following Illumina's standard protocol. Libraries were prepared with a starting amount of 250 ng and amplified in 11 PCR cycles. Libraries were profiled using the High Sensitivity DNA Kit on a 2100 Bioanalyzer (Agilent technologies) and quantified using the Qubit dsDNA HS Assay Kit in a Qubit 2.0 Fluorometer (Life Technologies). Libraries were pooled in equimolar ratio and sequenced on a NextSeq 500 instrument (Illumina) in a Highoutput Flowcell, single reads for 1x75 cycles plus 2x 8 cycles for dual index read. The RNA-seq measurement of WT and ZBTB10 KO HEK293 and U2OS cells yielded on average 25.5 M reads per sample. Reads were mapped to the homo sapiens GRCh38 reference genome using STAR version 2.5.2b (5), allowing up to two mismatches, a minimum intron length of 21 and keeping only uniquely aligned reads (on average 90% of all reads). We assessed the quality of the sequenced reads with fastqc (Andrews S. (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>) and dupRadar (6). We then counted reads per gene using featureCounts (7) from the subread package version 1.5.1 with the default parameters and using the gene model provided by Ensembl for the same assembly version used for mapping (homo sapiens GRCh38 version 84). For the differential expression analysis, we used R version 3.4.3 and DESeq2 version 1.18.1 (8) to normalize, transform, and model the data. The counts were

fitted to a Negative Binomial generalized linear model and the Wald significance test was used to determine the differentially expressed genes between WT and KO samples. Finally, RPKM values were calculated per gene per sample using the library size-normalized FPM (robust counts per million mapped fragments) values from DESeq2. We applied automatic independent filtering to avoid testing genes, which were poor candidates of being differentially expressed (maximizes the number of adjusted p-values less than $\alpha = 0.1$). Differentially expressed mRNAs were identified using a threshold of BH corrected p-values < 0.01 and a fold-change cut-off of 100%. Gene-ontology, KEGG pathway, Reactome and Gene Set enrichment analyses were performed using clusterProfiler (version 3.6.0)(9), ReactomePA (version 1.22.0)(10) and GSEASBase (version 1.40.1)(Morgan M, Falcon S, Gentleman R (2018). GSEASBase: Gene set enrichment data structures and methods.) bioconductor (11) packages in R.

Quantitative Telomere Repeat Amplification Protocol (qTRAP)

HEK293 WT and KO cells were lysed in RIPA lysis buffer (150 mM NaCl, 50 mM Tris-HCl pH 7.5, 1% IGEPAL CA-630) for 30 min on ice and protein concentration was determined via Bradford assay. Equal amounts of protein extract were diluted in TRAPeze Chaps buffer (Millipore), 200 nM of the TS (5'-AATCCGTCGAGCAGAGTT-3') and the ACX (5'-GCGCGGCTTACCCTTACCCTTACCCTAACCT-3') primer as well as the GoTaq qPCR Master Mix (Promega) were used. The quantitative TRAP assay was performed on a ViiA7 real-time PCR system (Thermo Scientific) for 20 min at 25°C and 10 min at 95°C and 40 cycles with 95°C for 30 sec, 60°C for 30 sec and 72°C for 1 min.

Telomere restriction fragment analysis (TRF)

TRF analysis for HEK293 cell clones was performed using the TeloTAGG telomere length assay kit (Sigma-Aldrich) according to manufacturer's instructions with slight modifications (12). 10 U each of HinfI and RsaI restriction enzymes were used to digest 2-5 μ g of DNA at 37°C overnight, and the digested DNA was resolved on a 0.8% agarose gel at 120 V for 4 h in 1x TAE buffer. The gel was submerged in 0.25 M HCl for 30 min for depurination, rinsed twice with distilled water, and washed twice with denaturation solution (0.5 M NaOH, 1.5 M NaCl) for 20 min each. Followed by two rinses in distilled water, the gel was washed twice with neutralizing solution (0.5 M Tris-HCl pH 7.5, 3 M NaCl) for 20 min each. Capillary osmosis using 20x SSC (3M NaCl, 0.3M sodium citrate tribasic dehydrate, pH 7) was performed overnight for transfer of the digested DNA onto a nylon membrane (Hybond, N⁺, Amersham). The DNA was fixed onto the membrane by UV-crosslinking at 120 mJ using a Stratalinker® UV Crosslinker (Stratagene). The membrane was rinsed twice with 2x SSC and incubated in pre-warmed DIG Easy Hyb Granules for 1 h at 42°C for pre-hybridization. 1 μ l DIG-labelled telomere probe/5 ml of Hyb Granules was added for hybridization and incubated for 3 h, followed by two washes with 100 ml stringent buffer 1 (2x SSC, 0.1% SDS), 5 min each at RT. The membrane was washed twice with 100 ml pre-warmed stringent buffer 2 (0.2x SSC, 0.1% SDS), 20 min each at 50°C, rinsed once with 1x wash buffer for 5 min, and blocked with 100 ml 1x blocking solution provided by the kit for 30 min at RT. Subsequently, the membrane was incubated with Anti-DIG-AP antibody (1:10,000) diluted in blocking solution for 30 min at RT, then washed twice with 100 ml of 1x washing buffer, 15 min each at RT, followed by 100 ml 1x detection for 5 min at RT. Digoxigenin luminescent detection (CDP star) system was used for detection of the TRF smears. The substrate solution was added dropwise to the membrane, incubated for 5 mins at RT, and the membrane was exposed to X-ray films. Average telomere length was calculated by comparison to the 1 kb plus DNA ladder provided by the kit.

For U2OS cells, 15 μ g gDNA were digested with 40 U HinfI and RsaI overnight at 37°C and separated by pulsed field gel electrophoresis on a 1% agarose gel (Agarose Megabase, Biorad) for 14 h at 6 V/cm. DNA was broken in a crosslinker for 1 min and transferred onto Byodine B membrane (Pall) overnight at RT. After fixing in 0.4 M NaOH for 15 min, the membrane was neutralized with two washes of 2x SSC for 5 min each and hybridized in hybridization buffer (6x SSC buffer, 0.1% SDS and 1% w/v non-fat milk powder) for at least 1 h at 42°C. The telomere oligonucleotide (TTAGGG_rev, Supplementary Table 1) was labeled using 3 μ l γ ³²P-ATP (3000 Ci/mmol; Hartmann Analytic) and T4 polynucleotidekinase (Thermo Scientific) for 1h at 37°C and purified using G50-columns (GE Healthcare). Labelled

oligonucleotides were denatured for 10 min at 95°C and incubated with the membrane in hybridization buffer for up to three days at 42°C. The membrane was rinsed in wash buffer (0.1% SDS in 2x SSC buffer), washed twice for 5 min each and incubated for 2 min in 37°C pre-warmed wash buffer 2 (0.1% SDS in 0.2x SSC buffer). The blot was dried on Whatman paper, exposed to a storage phosphor screen (GE Healthcare) overnight and scanned using the Typhoon FLA 9500 scanner (GE Healthcare).

Fluorescence in situ hybridization (FISH)

HEK293 and U2OS WT and KO clones were seeded overnight on glass slides (#1216691, Marienfeld). In brief, cells were washed in PBS, fixed for 10 min in 4% paraformaldehyde at RT and incubated in permeabilization buffer (20 mM Tris-HCl pH 8.0, 3 mM MgCl₂, 300 mM sucrose, 0.5% Triton X-100, 50 mM NaCl) for 1h at 37°C, followed by an incubation with RNaseA (0.1 mg/ml) for 1h at 37°C. Cells were dehydrated in successive baths of 70%, 85% and 100% ethanol for 2 min each, air-dried and hybridized with the TAMRA-labeled C-rich telomere probe (#507207, Eurogentec) for 3 min at 85°C and incubated overnight at 37°C. The next day, cells were washed in 2x SSC buffer in 50% formamide solution for 20 min at 37°C, followed by a wash in Tris-NaCl buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.05% Tween20). Samples were imaged with the confocal microscope TCS SP5 (Leica) using a 63x oil objective. Quantification of FISH signal was performed with ImageJ using the 3D Objects Counter plugin.

C-Circle assay

Genomic DNA concentration was estimated using Qubit dsDNA HS assay kit (Thermo Scientific) and 300 ng of DNA was digested using 10 U each of HinfI and RsaI restriction enzymes at 37°C for 2h. 15 ng and 7.5 ng of DNA was amplified with 7.5 U Φ DNA polymerase (NEB) in 1x Φ 29 buffer supplemented with 2 mM each of dATP, dGTP, dTTP (NEB) and 0.1 mg/ml BSA for 6h at 30°C followed by heat inactivation at 70°C for 20 min. 5 μ l of the reaction was diluted with 2X SSC (300 mM NaCl, 30 mM sodium citrate tribasic dehydrate, pH 7) and slot blotted on a nylon membrane (Hybond, N⁺, Amersham). DNA cross linking, probe hybridization and detection was performed as described for telomere restriction fragment analysis.

Immunofluorescence (IF) staining

Cells were washed with PBS and fixed for 10 min in 4% Formaldehyde at RT. After two PBS washes, cells were permeabilized for 30 min at 37°C in permeabilization buffer (20 mM Tris-HCl pH 8, 50 mM NaCl, 3 mM MgCl₂, 300 mM sucrose and 0.5% Triton X-100). For TRF1 and PML IF, cells were washed in PBS, blocked for 30 min in 20% FBS in PBS-0.1% Tween20 and incubated with primary antibody (Supplementary Table 1) for 2h at RT. After PBS washes, cells were incubated with secondary antibody for 1h at RT and mounted using the ProLong Diamond Antifade Mountant with DAPI (Life Technologies). In case of 53BP1 IF, permeabilized cells were washed in PBS-0.5% Triton X-100 and blocked in 5% BSA-PBS-0.5% Triton X-100 for 30 min. Primary antibody incubation followed overnight at 4°C and secondary incubation for 30 min at RT the next day.

Meta-FISH

Cells were incubated for 6h with 200 nM Nocodazole (Sigma-Aldrich) for metaphase spreads preparation. A hypotonic shock was performed with 75 mM KCl at 37°C for 30 min. After fixation in ethanol/acetic acid solution (3:1, v/v), suspensions were dropped onto clean superfrost slides to obtain metaphase spreads. FISH was performed as described. Metaphases were imaged with the confocal microscope TCS SP5 (Leica) using a 63x oil objective. Telomeric duplications were scored by manual inspection of deconvolved images.

Supplementary References

1. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10.
2. Masella,A.P., Bartram,A.K., Trzaskowski,J.M., Brown,D.G. and Neufeld,J.D. (2012) PANDAsseq: Paired-end assembler for illumina sequences. *BMC Bioinformatics*, **13**, 31.
3. Wu,T.D. and Watanabe,C.K. (2005) GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.
4. Lindsay,H., Burger,A., Biyong,B., Felker,A., Hess,C., Zaugg,J., Chiavacci,E., Anders,C., Jinek,M., Mosimann,C., *et al.* (2016) CrispRVariants charts the mutation spectrum of genome engineering experiments. *Nat. Biotechnol.*, **34**, 701–702.
5. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
6. Sayols,S., Scherzinger,D. and Klein,H. (2016) dupRadar: A Bioconductor package for the assessment of PCR artifacts in RNA-Seq data. *BMC Bioinformatics*, **17**, 1–5.
7. Liao,Y., Smyth,G.K. and Shi,W. (2014) FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
8. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 1–21.
9. Yu,G., Wang,L.-G., Han,Y. and He,Q.-Y. (2012) clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *Omi. A J. Integr. Biol.*, **16**, 284–287.
10. Yu,G. and He,Q.Y. (2016) ReactomePA: An R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.*, **12**, 477–479.
11. Huber,W., Carey,V.J., Gentleman,R., Anders,S., Carlson,M., Carvalho,B.S., Bravo,H.C., Davis,S., Gatto,L., Girke,T., *et al.* (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*, **12**, 115–121.
12. Kimura,M., Stone,R.C., Hunt,S.C., Skurnick,J., Lu,X., Cao,X., Harley,C.B. and Aviv,A. (2010) Measurement of telomere length by the southern blot analysis of terminal restriction fragment lengths. *Nat. Protoc.*, **5**, 1596–1607.

Conclusions and Future Perspectives

MS-based proteomics to study development

Proteomic studies in *D. melanogaster* are mostly restricted to a specific protein of interest as either antibodies or GFP-tagged fly lines are used (Morin et al., 2001; Nagarkar-Jaiswal et al., 2015; Sarov et al., 2016). In contrast, we used label-free quantitative mass spectrometry to investigate the development of *D. melanogaster* in a global fashion and generated two large datasets (*Article I* (Casas-Vila et al., 2017)). While the first dataset included 15 representative time points covering all stages of *Drosophila*'s life cycle, the second dataset focused on embryogenesis to study protein expression changes in a highly temporal-resolved manner. Almost half of the 1,535 and 1,644 differentially expressed proteins during life cycle and embryogenesis, respectively, are not characterized. Therefore, they present interesting candidates for future functional studies. By integrating tissue-specific RNA expression data, we were able to infer possible roles for similarly expressed proteins in the development of specific tissues. Our study presents a valuable and powerful resource that can be used for future projects.

Up until now, the majority of studies analyzing gene expression in development was based on *in situ* hybridization of RNA or microarray and RNA-sequencing datasets (Tomancak et al., 2007; Lecuyer et al., 2007; Kalinka et al., 2010; Graveley et al., 2011; Brown et al., 2014; Jambor et al., 2015). Although proteins are translated from mRNA, mRNA cannot be used as a precise predictor for protein concentrations. Several studies in different model organisms such as *Saccharomyces cerevisiae* (Griffin et al., 2002), *C. elegans* (Gruen et al., 2014), *Xenopus laevis* (Peshkin et al., 2015) or *Drosophila* SL2 cells (Bonaldi et al., 2008) and mammalian cells (Schwannhäuser et al., 2011) reported a rather poor correlation between mRNA and protein levels at any time point (for review see also (Maier et al., 2009; Liu et al., 2016)). Although it has been shown that ribosome occupancy is a better predictor for protein expression than the mRNA concentration itself (Ingolia et al., 2009), it still does not match up to proteomics. It is important to consider differences in

mRNA and protein levels at a specific time, when studying phenotypic differences during development due to tightly controlled gene expression.

Most mRNA-protein studies focused on a single time point of development. In contrast, we combined the embryogenesis proteome including 14 time points (Casas-Vila et al., 2017) with its paired transcriptome in a follow-up study to analyze post-transcriptional gene regulation (Becker et al., 2018). Using four mathematical models to describe translation and protein degradation, the majority of 3,761 mRNA-protein pairs was explained. 16% of mRNA-protein pairs were rejected and hence, might be controlled post-transcriptionally. It is well established that maternally-deposited mRNAs such as *oskar*, *hunchback* or *nanos* are regulated post-transcriptionally in the early development of *Drosophila* (Murata and Wharton, 1995; Kalifa et al., 2006; Andrews et al., 2011) because the transcription machinery of the embryo is not switched on until two-three hours of embryogenesis when the MZT has taken place. Post-transcriptional regulation is dynamically regulated by RNA-binding proteins to ensure clearing of maternal transcripts and controlling translation (Tadros and Lipshitz, 2009).

It is still a matter of debate whether protein abundance is predominantly regulated by transcription or rather controlled at translational level and/or by protein stability. A study in early development of *Xenopus* showed that mainly low abundant proteins change expression levels dynamically. This is controlled by fast changes in transcription and protein degradation rather than translational control (Peshkin et al., 2015). Although it is still challenging to fully understand the relation between mRNA and protein, transcript levels predominantly influence protein abundance (Liu et al., 2016).

While we investigated changes in protein abundance during *Drosophila* development, it would be exciting to additionally study post-translational changes of proteins. PTMs are important as they influence the activity, function, structure, as well as the steady state concentration of proteins. Changes in phosphorylation patterns can provide insights into regulation and deregulation of phosphatases and kinases. Furthermore, an increase in ubiquitinated proteins would hint towards perturbations in the process of protein degradation. It will be beneficial to measure both protein expression

and PTMs to expand our knowledge about cellular signaling networks during development. To gain a deep coverage of PTMs, different enrichment methods for modified peptides would need to be applied prior to MS measurements. This was exemplified by analysis of phosphorylation and acetylation in the previously mentioned *Xenopus* study (Peshkin et al., 2015).

Although we cover at least one third of all annotated *Drosophila* genes, our proteomics datasets are incomplete. Low abundant proteins or proteins that are restricted to a certain tissue are probably missing. Thus, tissue-specific MS-based proteomics is desirable, as it would enable a more comprehensive and precise analysis. The main limitation is the amount of required material and the necessity for advances in workflows and quantification of low-input samples. During the last years, two methods have been published to reduce contact with plastics and to shorten sample preparation time. The Mann laboratory proposed an in-StageTip method. The complete workflow from lysis to peptide purification is performed in a single pipette tip (Kulak et al., 2014). Single-Pot Solid-Phase-enhanced Sample Preparation (SP3) was introduced by Krijgsveld and his colleagues. SP3 uses paramagnetic beads in a single tube to trap proteins and peptides (Hughes et al., 2014). Both methods present valuable improvements but still need to be tested for their successful use in different applications. Current MS workflows and instruments are not sensitive enough to achieve complete sequence coverage. Moreover, limited material as well as the complexity of biological samples are still challenging. Nevertheless, MS-based proteomics underwent massive technical improvements during the last years and can provide important insights into global protein expression changes.

MS-based proteomics to identify telomere-associated proteins

To shed light into the understanding of genome biology, it is essential to identify proteins associated with distinct genomic loci. In the past years, technologies have advanced and identification of DNA- and chromatin-bound proteins by MS-based proteomics enabled global analysis of an entire set of associated proteins in an unbiased and quantitative fashion (Lambert et al., 2009; Makowski et al., 2018; Wang et al., 2013; Wierer and Mann,

2016). The first large-scale study to identify telomere-associated proteins employed proteomics of isolated chromatin segments (PICH) by using locus-specific pulldowns (Déjardin and Kingston, 2009). In this way, genomic DNA with its associated proteins is isolated by affinity purification using antisense hybridization of a telomeric probe. A few years later, this method was optimized to determine the protein composition at dysfunctional telomeres upon TRF2 depletion leading to DNA damage (Bartocci et al., 2014). Another approach to compare protein levels at telomeres between different states (e.g. long versus short telomeres) uses a quantitative telomeric chromatin isolation protocol (QTIP) (Grolimund et al., 2013). While both approaches are conducted *in vivo*, PICH requires a vast amount of material with more than 5×10^8 cells per experiment and includes an RNase digestion step. Thereby, TERRA-dependent factors are lost. Another advantage of QTIP is the use of SILAC quantification. Despite the identification of the six shelterin members by both approaches, overlap of additional factors is relatively low. BioID-proteomics presents another method to capture all transient and tight associations in close proximity of telomeres, which does not require direct binding to telomeric DNA. BioID was applied in ALT and telomerase-positive cells and identified most of the previously described proteins (Garcia-Exposito et al., 2016). Using stable cell lines expressing tagged TRF1, TRF2 or POT1 fusion proteins, an *in vitro* affinity purification method (IP-MS) defined the interaction network at telomeres (Giannone et al., 2010). In contrast to immobilizing the protein, in our study we used the telomeric sequence as a bait to identify associated proteins in 16 vertebrate species (*Article II* (Kappei et al., 2017)). Only proteins that specifically bound the telomeric sequence in at least five species were defined as telomere-associated. While we detected all shelterin members and known accessory factors (Conomos et al., 2012, 2014; Kappei et al., 2013; Lenain et al., 2006; Marzec et al., 2015; Miyake et al., 2009; Overbeek and De Lange, 2006; Popuri et al., 2014; Ye et al., 2010), we also identified novel candidate proteins. In general, it is important to consider that *in vivo* protein binding to chromatin is highly dependent on different factors, for instance, histone and DNA modifications, the presence of other DNA-binding proteins and the three-dimensional structure of chromatin (Bannister and Kouzarides, 2011). Some differences

between the described screens might be explained by the *in vivo* versus *in vitro* situation or by cell type-specific binders. Candidates that overlap between these studies probably have a higher affinity for telomeric DNA or are present at higher concentrations. Transiently associating factors might be missing in most MS-based approaches. Overall it is advisable to take various screens as complementary information. In any case, novel candidate proteins have to be validated by different means. This requires validating the *in vivo* localization and determining a functional relevance of the telomere-binding activity. The TeloPIN database provides an overview of a telomere regulatory network by including protein interactions with DNA, RNA and proteins that were determined in various large-scale studies (Luo et al., 2015).

All previously mentioned screens were conducted in unsynchronized cell populations and therefore detect proteins of different cell cycle phases. Telomeres bind individual proteins at a specific time resulting in different compositions of telomere-associated proteins throughout the cell cycle. It would be fascinating to conduct large-scale studies in synchronized cells to expand and define our current knowledge at specific cell cycle phases. It is still one of the biggest challenges to obtain a comprehensive overview of molecular changes at telomeres in certain conditions.

MS-based studies require further technical improvements and experimental optimizations as the major drawback is the little amount of bound protein that can be extracted per cell. Multiple copies of a specific locus have to be present to successfully perform ChIP-MS, which until now only works for repeat sequences. A strategy that evolved in the last years makes use of the CRISPR/Cas9 system. An inactivated Cas9 fusion protein is co-expressed with a specific guide RNA to target a genomic region of interest (enChIP-MS) (Fujita and Fujii, 2013). Similarly, genomic loci can also be enriched using transcription activator-like (TAL) effector proteins (Bogdanove and Voytas, 2011; Fujita et al., 2013).

The TTGGGG variant repeat-binding protein ZBTB10

In our follow-up project (*Article III* (Bluhm et al., 2019)), we investigated the candidate protein ZBTB10, which was initially identified as a TTAGGG binder in six vertebrate species

(Article II (Kappei et al., 2017)). While binding to the TTAGGG sequence was defined to be direct via the two zinc fingers of ZBTB10, a stronger affinity for the variant motif TTGGGG was determined. Telomeric variant repeats are present in the proximal 2 kb region of telomeres and in telomeres of ALT cells. Due to the HR-based maintenance mechanism in ALT cells, it is difficult to compare absolute numbers of variant repeat sequences. They change stochastically and over time. Until now, orphan nuclear receptors NR2C/F are the only proteins identified as binders of TCAGGG repeats exhibiting specific functions. They contribute to ALT phenotypes (Conomos et al., 2012) and bridge their bound genomic loci to ALT telomeres. Thereby, a platform for targeted telomere insertion at regions distant from gene promoters is provided (Marzec et al., 2015). Our data establishes ZBTB10 as a novel variant repeat-binding protein that localizes to telomeres in U2OS cells. It would be fascinating to gain a comprehensive overview of variant repeat-binders (especially in ALT cells) and to investigate functional consequences.

Our interactomics data in telomerase-positive HEK293 and in ALT-maintaining U2OS cells stably expressing ZBTB10-GFP, reproducibly identified TRF2 and RAP1 as interaction partners of ZBTB10. Further analysis revealed that TRF2 and ZBTB10 co-IP in the nucleoplasm and the chromatin fraction suggesting that telomeric chromatin is not required. TRF2 is essential for the inhibition of NHEJ repair at telomeres (Bae and Baumann, 2007) and together with RAP1, they protect telomeres from processing by HDR (Rai et al., 2016). Repair by NHEJ dominates in G1 phase due to ATM-dependent phosphorylation of 53BP1, the key regulator. In S and G2 phases, BRCA1 counteracts 53BP1 signaling assuring that HDR predominates (Feng et al., 2015). In a TAP-MS screen, ZBTB10 was found as a putative interactor of 53BP1 (Woods et al., 2014). In addition, a large-scale DNA repair network analysis by the Choudhary laboratory detected ca. 1.6-fold enriched ZBTB10 in close proximity to 53BP1, BRCA1 and the DNA damage checkpoint protein MDC1 (Gupta et al., 2018). We found a remarkable overlap of proteins in proximity to 53BP1, BRCA1 or MDC1 and ZBTB10 interactors identified in our GFP pulldowns such as PNUTS, TOX4, NONO, members of the GTF2H complex, TRF2 and RAP1. Taking all data into account, it is tempting to speculate that ZBTB10 plays a role in DNA damage repair and is increasingly recruited to

telomeres upon damage. So far, we can exclude an increase of ZBTB10 at telomeres after three days knockdown of TRF1 or TRF2 in U2OS cells. An effect on the localization of ZBTB10 after inducing DNA damage by different means needs further investigation.

Functional annotation clustering of the additional 35 interactors revealed strong association with RNA-related terms. In the past years, RNA-binding proteins as well as RNA-processing factors have been described to regulate the expression of genes involved in DDR (e.g. RBM14) or to act as a DNA damage sensor (e.g. hnRNPUL1) (Kai, 2016; Lackner et al., 2011; Paulsen et al., 2009). Cells depleted for the RNA helicase DDX1 show decreased cell survival and delayed DSB repair in G2 phase after treatment with ionizing radiation. A combination of experiments further revealed that DDX1 is important for maintaining the ssDNA during repair and resolving RNA-DNA hybrids at DSBs (Li et al., 2016). The strongly enriched subunit PNUTS is a DDR component and involved in RNA processing and setting the phosphorylation state of cell cycle proteins (Landsverk et al., 2010; Lee et al., 2010). Of note, the QTIP study detected PNUTS and other subunits of the PTW/PP1 phosphatase complex at telomeres in HeLa cells (Grolimund et al., 2013).

Several of the significantly enriched interaction partners of ZBTB10 (NONO, SFPQ, DAZAP1, RBM14 and FUS) are required for the formation of paraspeckles. These are nuclear organelles involved in gene regulation by sequestration of proteins and RNAs (Hirose et al., 2014; Imamura et al., 2014; Fox et al., 2017). It would be exciting to see whether ZBTB10 localizes to paraspeckles and to investigate a functional link.

Due to the diverse roles of interactors, it would be interesting to decipher a role for ZBTB10 in several cellular processes and whether they relate to telomeric ZBTB10 or whether they point to non-telomeric functions. We focused on telomere assays testing telomere length changes, telomere maintenance or ALT phenotypes and did not observe differences between WT and ZBTB10 KO cells. At the moment, we cannot exclude a possible redundant function of ZBTB10. Sequence alignments showed highest overlap with ZBTB8A and ZBTB46. However, we neither observe expression level changes on RNA nor on protein level in ZBTB10 KO cells. Moreover, none of the two proteins have a described telomeric function and they were not detected in any large-scale telomere-related study. Redundancy

with other zinc finger or BTB domain-containing proteins, which have been identified as telomere-associated, might also be possible.

As all interaction studies have been performed using a GFP-tagged protein version, we cannot exclude a tag effect (Stadler et al., 2013). However, we used GFP as a control to only identify ZBTB10-specific interactors. We worked with cell lines stably expressing ZBTB10-GFP and showing similar expression levels as the wildtype protein. We further ensured nuclear localization of ZBTB10.

Overall, telomere regulation is a complex cellular mechanism exemplified by the interplay of telomere-binding proteins, DNA repair and damage proteins. Further studies are necessary to decipher individual and combined functions throughout the cell cycle.

Acknowledgements

References

- Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. (2002). The Genome Sequence of *Drosophila melanogaster*. *Genetics* 161, 1507–1516.
- Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature* 422, 198–207.
- Aebersold, R., and Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. *Nature* 537, 347–355.
- Andrews, S., Snowflack, D.R., Clark, I.R.A.E., and Gavis, E.R. (2011). Multiple mechanisms collaborate to repress nanos translation in the *Drosophila* ovary and embryo. *RNA* 17, 967–977.
- Arnoult, N., Van Beneden, A., and Decottignies, A. (2012). Telomere length regulates TERRA levels through increased trimethylation of telomeric H3K9 and HP1 α . *Nat. Struct. Mol. Biol.* 19, 948–956.
- Arora, R., Lee, Y., Wischnewski, H., Brun, C.M., Schwarz, T., and Azzalin, C.M. (2014). RNaseH1 regulates TERRA-telomeric DNA hybrids and telomere maintenance in ALT tumour cells. *Nat. Commun.* 5, 1–11.
- Azzalin, C.M., and Lingner, J. (2008). Telomeres: The silence is broken. *Cell Cycle* 7, 1161–1165.
- Azzalin, C.M., Reichenbach, P., Khoraiuli, L., Giulotto, E., and Lingner, J. (2007). Telomeric Repeat-Containing RNA and RNA Surveillance Factors at Mammalian Chromosome Ends. *Science* 318, 798–801.
- Bae, N.S., and Baumann, P. (2007). A RAP1/TRF2 Complex Inhibits Nonhomologous End-Joining at Human Telomeric DNA Ends. *Mol. Cell* 26, 323–334.
- Bannister, A.J., and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Res.* 21, 381–395.
- Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., and Kuster, B. (2007). Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* 389, 1017–1031.
- Bantscheff, M., Lemeer, S., Savitski, M.M., and Kuster, B. (2012). Quantitative mass spectrometry in proteomics: Critical review update from 2007 to the present. *Anal. Bioanal. Chem.* 404, 939–965.

- Bartocci, C., Diedrich, J.K., Ouzounov, I., Li, J., Piunti, A., Pasini, D., Yates, J.R., and Lazzerini Denchi, E. (2014). Isolation of chromatin from dysfunctional telomeres reveals an important role for Ring1b in NHEJ-mediated chromosome fusions. *Cell Rep.* 7, 1320–1332.
- Baumann, P. (2001). Pot1, the Putative Telomere End-Binding Protein in Fission Yeast and Humans. *Science.* 292, 1171–1175.
- Baur, J.A., Zou, Y., Shay, J.W., and Wright, W.E. (2016). Telomere Position Effect in Human Cells. *Science.* 292, 2075–2077.
- Becker, K., Bluhm, A., Casas-Vila, N., Dinges, N., Dejung, M., Sayols, S., Kreutz, C., Roignant, J., Butter, F., and Legewie, S. (2018). Quantifying post-transcriptional regulation in the development of *Drosophila melanogaster*. *Nat. Commun.* 9, 1–14.
- Benarroch-Popivker, D., Pisano, S., Mendez-Bermudez, A., Lototska, L., Kaur, P., Bauwens, S., Djerbi, N., Latrick, C.M., Fraiser, V., Pei, B., et al. (2016). TRF2-Mediated Control of Telomere DNA Topology as a Mechanism for Chromosome-End Protection. *Mol. Cell* 61, 274–286.
- Bianchi, A., and Shore, D. (2007). Increased association of telomerase with short telomeres in yeast. *Genes Dev.* 21, 1726–1730.
- Bilbao, A., Varesio, E., Luban, J., Strambio-De-Castillia, C., Hopfgartner, G., Müller, M., and Lisacek, F. (2015). Processing strategies and software solutions for data-independent acquisition in mass spectrometry. *Proteomics* 15, 964–980.
- Biroccio, A., Cherfils-Vicini, J., Augereau, A., Pinte, S., Bauwens, S., Ye, J., Simonet, T., Horard, B., Jamet, K., Cervera, L., et al. (2013). TRF2 inhibits a cell-extrinsic pathway through which natural killer cells eliminate cancer cells. *Nat. Cell Biol.* 15, 818–828.
- Bodenmiller, B., Malmstrom, J., Gerrits, B., Campbell, D., Lam, H., Schmidt, A., Rinner, O., Mueller, L.N., Shannon, P.T., Pedrioli, P.G., et al. (2007). PhosphoPep-A phosphoproteome resource for systems biology research in *Drosophila* Kc167 cells. *Mol. Syst. Biol.* 3, 1–11.
- Bogdanove, A.J., and Voytas, D.F. (2011). TAL effectors: customizable proteins for DNA targeting. *Science* 333, 1843–1846.
- Bonaldi, T., Straub, T., Cox, J., Kumar, C., Becker, P.B., and Mann, M. (2008). Combined Use of RNAi and Quantitative Proteomics to Study Gene Function in *Drosophila*. *Mol. Cell* 31, 762–772.

- Brown, J.B., Boley, N., Eisman, R., May, G.E., Stoiber, M.H., Duff, M.O., Booth, B.W., Wen, J., Park, S., Suzuki, A.M., et al. (2014). Diversity and dynamics of the *Drosophila* transcriptome. *Nature* *512*, 1–7.
- Brownridge, P., Holman, S.W., Gaskell, S.J., Grant, C.M., Harman, V.M., Hubbard, S.J., Lanthaler, K., Lawless, C., O’cualain, R., Sims, P., et al. (2011). Global absolute quantification of a proteome: Challenges in the deployment of a QconCAT strategy. *Proteomics* *11*, 2957–2970.
- Brunner, E., Ahrens, C.H., Mohanty, S., Baetschmann, H., Loevenich, S., Potthast, F., Deutsch, E.W., Panse, C., de Lichtenberg, U., Rinner, O., et al. (2007). A high-quality catalog of the *Drosophila melanogaster* proteome. *Nat. Biotechnol.* *25*, 576–583.
- Bryan, T.M., Englezou, A., Gupta, J., Bacchetti, S., and Reddel, R.R. (1995). Telomere elongation in immortal human cells without detectable telomerase activity. *Embo J* *14*, 4240–4248.
- Cañas Montalvo, B., López-Ferrer, D., Ramos-Fernández, A., Camafeita, E., and Calvo, E. (2006). Mass spectrometry technologies for proteomics. *Briefings Funct. Genomics Proteomics* *4*, 295–320.
- Casas-Vila, N., Bluhm, A., Sayols, S., Dinges, N., Dejung, M., Altenhein, T., Kappei, D., Altenhein, B., Roignant, J., and Butter, F. (2017). The developmental proteome of *Drosophila melanogaster*. *Genome Res.* *27*, 1273–1285.
- Catherman, A.D., Skinner, O.S., and Kelleher, N.L. (2014). Top Down proteomics: Facts and perspectives. *Biochem. Biophys. Res. Commun.* *445*, 683–693.
- Celli, G.B., and de Lange, T. (2005). DNA processing is not required for ATM-mediated telomere damage response after TRF2 deletion. *Nat. Cell Biol.* *7*, 712–718.
- Cesare, A.J., and Reddel, R.R. (2010). Alternative lengthening of telomeres: Models, mechanisms and implications. *Nat. Rev. Genet.* *11*, 319–330.
- Cesare, A., Hayashi, M., Crabbe, L., and Karlseder, J. (2013). The Telomere deprotection response is functionally distinct from the Genomic DNA damage response. *Mol. Cell* *51*, 141–155.
- Cesare, A.J., Kaul, Z., Cohen, S.B., Napier, C.E., Pickett, H.A., Neumann, A.A., and Reddel, R.R. (2009). Spontaneous occurrence of telomeric DNA damage response in the absence of chromosome fusions. *Nat. Struct. Mol. Biol.* *16*, 1244–1251.
- Chait, B.T. (2006). Mass Spectrometry: Bottom-Up or Top-Down? *Science*. *314*, 65–66.

- Chan, H., Wang, Y., and Feigon, J. (2017). Progress in Human and *Tetrahymena* Telomerase Structure Determination. *Annu. Rev. Biophys.* **46**, 199–225.
- Chen, H.-Y., and Chen, R.-H. (2016). Cullin 3 Ubiquitin Ligases in Cancer Biology: Functions and Therapeutic Implications. *Front. Oncol.* **6**, 1–10.
- Chen, L.Y., Redon, S., and Lingner, J. (2012). The human CST complex is a terminator of telomerase activity. *Nature* **488**, 540–544.
- Cho, N.W., Dilley, R.L., Lampson, M.A., and Greenberg, R.A. (2014). Interchromosomal Homology Searches Drive Directional ALT Telomere Movement and Synapsis. *Cell* **159**, 108–121.
- Choudhary, C., and Mann, M. (2010). Decoding signalling networks by mass spectrometry-based proteomics. *Nat. Rev. Mol. Cell Biol.* **11**, 427–439.
- Chu, H., Cifuentes-Rojas, C., Kesner, B., Aeby, E., Lee, H., Wei, C., Oh, H.J., Boukhali, M., Haas, W., and Lee, J.T. (2017). TERRA RNA Antagonizes ATRX and Protects Telomeres. *Cell* **170**, 86–101.
- Conomos, D., Stutz, M.D., Hills, M., Neumann, A. a., Bryan, T.M., Reddel, R.R., and Pickett, H.A. (2012). Variant repeats are interspersed throughout the telomeres and recruit nuclear receptors in ALT cells. *J. Cell Biol.* **199**, 893–906.
- Conomos, D., Reddel, R.R., and Pickett, H. a (2014). NuRD–ZNF827 recruitment to telomeres creates a molecular scaffold for homologous recombination. *Nat. Struct. Mol. Biol.* **21**, 760–770.
- Conrads, T.P., Alving, K., Veenstra, T.D., Belov, M.E., Anderson, G.A., Anderson, D.J., Lipton, M.S., Paša-Tolić, L., Udseth, H.R., Chrisler, W.B., et al. (2001). Quantitative Analysis of Bacterial and Mammalian Proteomes Using a Combination of Cysteine Affinity Tags and ¹⁵N-Metabolic Labeling. *Anal. Chem.* **73**, 2132–2139.
- Costantino, L., Sotiriou, S.K., Rantala, J.K., Magin, S., Mladenov, E., Helleday, T., Haber, J.E., Iliakis, G., Kallioniemi, O.P., and Halazonetis, T.D. (2014). Break-Induced Replication Repair of Damaged Forks Induces Genomic Duplications in Human Cells. *Science*. **343**, 88–92.
- Court, R., Chapman, L., Fairall, L., and Rhodes, D. (2005). How the human telomeric proteins TRF1 and TRF2 recognize telomeric DNA: a view from high-resolution crystal structures. *EMBO* **6**, 39–45.

- 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.
- Cox, J., Hein, M.Y., Lubner, 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.
- Cristobal, A., Hennrich, M.L., Giansanti, P., Goerdal, S.S., Heck, A.J.R., and Mohammed, S. (2012). In-house construction of a UHPLC system enabling the identification of over 4000 protein groups in a single analysis. *Analyst* **137**, 3541–3548.
- Cubiles, M.D., Barroso, S., Vaquero-Sedas, M.I., Enguix, A., Aguilera, A., and Vega-Palas, M.A. (2018). Epigenetic features of human telomeres. *Nucleic Acids Res.* 1–9.
- D’Adda di Fagagna, F., Reaper, P.M., Clay-Farrace, L., Fiegler, H., Carr, P., von Zglinicki, T., Saretzki, G., Carter, N.P., and Jackson, S.P. (2003). A DNA damage checkpoint response in telomere-associated senescence. *Nature* **426**, 194–198.
- D’Adda Di Fagagna, F., Teo, S.H., and Jackson, S.P. (2004). Functional links between telomeres and proteins of the DNA-damage response. *Genes Dev.* **18**, 1781–1799.
- D’Alcontres, M.S., Palacios, J.A., Mejias, D., and Blasco, M.A. (2014). Topoll α prevents telomere fragility and formation of ultra thin DNA bridges during mitosis through TRF1-dependent binding to telomeres. *Cell Cycle* **13**, 1463–1481.
- Déjardin, J., and Kingston, R.E. (2009). Purification of Proteins Associated with Specific Genomic Loci. *Cell* **136**, 175–186.
- Denchi, E.L., and de Lange, T. (2007). Protection of telomeres through independent control of ATM and ATR by TRF2 and POT1. *Nature* **448**, 1068–1071.
- Deng, Z., Norseen, J., Wiedmer, A., Riethman, H., and Lieberman, P.M. (2009). TERRA RNA Binding to TRF2 Facilitates Heterochromatin Formation and ORC Recruitment at Telomeres. *Mol. Cell* **35**, 403–413.

- Deng, Z., Wang, Z., Stong, N., Plasschaert, R., Moczan, A., Chen, H.S., Hu, S., Wikramasinghe, P., Davuluri, R. V., Bartolomei, M.S., et al. (2012). A role for CTCF and cohesin in subtelomere chromatin organization, TERRA transcription, and telomere end protection. *EMBO J.* *31*, 4165–4178.
- Dietzl, G., Chen, D., Schnorrer, F., Su, K.C., Barinova, Y., Fellner, M., Gasser, B., Kinsey, K., Oppel, S., Scheiblaue, S., et al. (2007). A genome-wide transgenic RNAi library for conditional gene inactivation in *Drosophila*. *Nature* *448*, 151–156.
- Dilley, R.L., Verma, P., Cho, N.W., Winters, H.D., Wondisford, A.R., and Greenberg, R.A. (2016). Break-induced telomere synthesis underlies alternative telomere maintenance. *Nature* *539*, 54–58.
- Dinami, R., Petti, E., Sestito, R., Benetti, R., and Schoeftner, S. (2014). microRNAs control the function of telomeres in cancer. *Rna Dis.* 12–14.
- Doherty, M.K., and Beynon, R.J. (2006). Protein turnover on the scale of the proteome. *Expert Rev. Proteomics* *3*, 97–110.
- Doksani, Y., Wu, J.Y., de Lange, T., and Zhuang, X. (2013). Super-resolution fluorescence imaging of telomeres reveals TRF2-dependent T-loop formation. *Cell* *155*, 345–356.
- Doll, S., and Burlingame, A.L. (2015). Mass Spectrometry-Based Detection and Assignment of Protein Posttranslational Modifications. *ACS Chem. Biol.* *10*, 63–71.
- Domon, B., and Aebersold, R. (2010). Options and considerations when selecting a quantitative proteomics strategy. *Nat. Biotechnol.* *28*, 710–721.
- Dorus, S., Busby, S.A., Gerike, U., Shabanowitz, J., Hunt, D.F., and Karr, T.L. (2006). Genomic and functional evolution of the *Drosophila melanogaster* sperm proteome. *Nat. Genet.* *38*, 1440–1445.
- Dunham, W.H., Mullin, M., and Gingras, A. (2012). Affinity-purification coupled to mass spectrometry: Basic principles and strategies. *Proteomics* *12*, 1576–1590.
- Elias, J.E., and Gygi, S.P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* *4*, 207–214.
- Erdel, F., Kratz, K., Willcox, S., Griffith, J.D., Greene, E.C., and de Lange, T. (2017). Telomere Recognition and Assembly Mechanism of Mammalian Shelterin. *Cell Rep.* *18*, 41–53.

- Ezkurdia, I., Juan, D., Rodriguez, J.M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., and Tress, M.L. (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Hum. Mol. Genet.* **23**, 5866–5878.
- Fabre, B., Korona, D., Groen, A., Vowinckel, J., Gatto, L., Deery, M.J., Ralser, M., Russell, S., and Lilley, K.S. (2016). Analysis of *Drosophila melanogaster* proteome dynamics during embryonic development by a combination of label-free proteomics approaches. *Proteomes* **16**, 2068–2080.
- Feng, J., Funk, W.D., Wang, S.S., Weinrich, S.L., Avilion, A.A., Chiu, C.P., Adams, R.R., Chang, E., Allsopp, R.C., and Yu, J. (1995). The RNA component of human telomerase. *Science* **269**, 1236–1241.
- Feng, L., Li, N., Li, Y., Wang, J., Gao, M., Wang, W., and Chen, J. (2015). Cell cycle-dependent inhibition of 53BP1 signaling by BRCA1. *Cell Discov.* **1**, 1–11.
- Feng, X., Hsu, S.J., Bhattacharjee, A., Wang, Y., Diao, J., and Price, C.M. (2018). CTC1-STN1 terminates telomerase while STN1-TEN1 enables C-strand synthesis during telomere replication in colon cancer cells. *Nat. Commun.* **9**.
- Fenn, J.B., Mann, M., Meng, C.K.A.I., Wong, S.F., and Whitehouse, C.M. (1989). Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science*. **246**, 64–71.
- Feuerhahn, S., Chen, L., Luke, B., and Porro, A. (2015). No DDRama at chromosome ends: TRF2 takes centre stage. *Trends Biochem. Sci.* **40**, 275–285.
- Fredens, J., Engholm-Keller, K., Giessing, A., Pultz, D., Larsen, M.R., Højrup, P., Møller-Jensen, J., and Førgeman, N.J. (2011). Quantitative proteomics by amino acid labeling in *C. elegans*. *Nat. Methods* **8**, 845–847.
- Fujita, T., and Fujii, H. (2015). Isolation of specific genomic regions and identification of associated molecules by engineered DNA-binding molecule-mediated chromatin immunoprecipitation (enChIP) using CRISPR. *Biochem. Biophys. Res. Commun.* **439**, 43–52.
- Fujita, T., Asano, Y., Ohtsuka, J., Takada, Y., Saito, K., Ohki, R., and Fujii, H. (2013). Identification of telomere-associated molecules by engineered DNA-binding molecule-mediated chromatin immunoprecipitation (enChIP). *Sci. Rep.* **3**, 1–7.
- García-Cao, M., O’Sullivan, R., Peters, A.H.F.M., Jenuwein, T., and Blasco, M.A. (2004). Epigenetic regulation of telomere length in mammalian cells by the Suv39h1 and Suv39h2 histone methyltransferases. *Nat. Genet.* **36**, 94–99.

- Garcia-Exposito, L., Bournique, E., Bergoglio, V., Bose, A., Barroso-Gonzalez, J., Zhang, S., Roncaioli, J.L., Lee, M., Wallace, C.T., Watkins, S.C., et al. (2016). Proteomic Profiling Reveals a Specific Role for Translesion DNA Polymerase η in the Alternative Lengthening of Telomeres. *Cell Rep.* 17, 1858–1871.
- Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J.R., and Mann, M. (2010). Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat. Methods* 7, 383–385.
- Gerber, S.A., Rush, J., Stemman, O., Kirschner, M.W., and Gygi, S.P. (2003). Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci.* 100, 6940–6945.
- Giannone, R.J., McDonald, H.W., Hurst, G.B., Shen, R.F., Wang, Y., and Liu, Y. (2010). The protein network surrounding the human telomere repeat binding factors TRF1, TRF2, and POT1. *PLoS One* 5, 34–39.
- Gillet, L.C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (2012). Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Mol. Cell. Proteomics* 11.
- Gilson, E., and Géli, V. (2007). How telomeres are replicated. *Nat. Rev. Mol. Cell Biol.* 8, 825–838.
- de Godoy, L.M.F., Olsen, J. V., de Souza, G.A., Li, G., Mortensen, P., and Mann, M. (2006). Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. *Genome Biol.* 7, 1–15.
- Gonzalo, S., Jaco, I., Fraga, M.F., Chen, T., Li, E., Esteller, M., and Blasco, M.A. (2006). DNA methyltransferases control telomere length and telomere recombination in mammalian cells. *Nat. Cell Biol.* 8, 416–424.
- Gottschling, D.E., Aparicio, O.M., Billington, B.L., and Zakian, V.A. (1990). Position effect at *S. cerevisiae* telomeres: Reversible repression of Pol II transcription. *Cell* 63, 751–762.
- Gouw, J.W., Pinkse, M.W.H., Vos, H.R., Moshkin, Y., Verrijzer, C.P., Heck, A.J.R., and Krijgsveld, J. (2009). *In Vivo* Stable Isotope Labeling of Fruit Flies Reveals Post-transcriptional Regulation in the Maternal-to-zygotic Transition. *Mol. Cell. Proteomics* 8, 1566–1578.

- Graf, M., Bonetti, D., Lockhart, A., Serhal, K., Kellner, V., Maicher, A., Jolivet, P., Teixeira, M.T., and Luke, B. (2017). Telomere Length Determines TERRA and R-Loop Regulation through the Cell Cycle. *Cell* 170, 72–85.
- Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., Van Baren, M.J., Boley, N., Booth, B.W., et al. (2011). The developmental transcriptome of *Drosophila melanogaster*. *Nature* 471, 473–479.
- Greider, C.W., and Blackburn, E.H. (1985). Identification of a specific telomere terminal transferase activity in tetrahymena extracts. *Cell* 43, 405–413.
- Griffin, T.J., Gygi, S.P., Ideker, T., Rist, B., Eng, J., Hood, L., and Aebersold, R. (2002). Complementary Profiling of Gene Expression at the Transcriptome and Proteome Levels in *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* 1.4, 323–333.
- Griffith, J.D., Comeau, L., Rosenfield, S., Stansel, R.M., Bianchi, a, Moss, H., and de Lange, T. (1999). Mammalian telomeres end in a large duplex loop. *Cell* 97, 503–514.
- Grolimund, L., Aeby, E., Hamelin, R., Armand, F., Chiappe, D., Moniatte, M., and Lingner, J. (2013). protocol identifies different telomeric states. *Nat. Commun.* 4, 1–12.
- Gruen, D., Kirchner, M., Thierfelder, N., Stoeckius, M., Selbach, M., and Rajewsky, N. (2014). Conservation of mRNA and Protein Expression during Development of *C. elegans*. *Cell Rep.* 6, 565–577.
- Gupta, R., Somyajit, K., Narita, T., Maskey, E., Stanlie, A., Kremer, M., Typas, D., Lammers, M., Mailand, N., Nussenzweig, A., et al. (2018). DNA Repair Network Analysis Reveals Shieldin as a Key Regulator of NHEJ and PARP Inhibitor Sensitivity. *Cell* 173, 972–974.
- Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., and Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* 17, 1–6.
- Hayflick, L., and Moorhead, P.S. (1961). The serial cultivation of human diploid cell strains. *Exp. Cell Res.* 25, 585–621.
- Heaphy, C.M., Wilde, R.F. De, Jiao, Y., Klein, A.P., Edil, B.H., Shi, C., Bettegowda, C., Rodriguez, F.J., Eberhart, C.G., Hebbar, S., et al. (2011). Altered Telomeres in Tumors with ATRX and DAXX Mutations. *Science.* 333, 425.

- 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.
- Henson, J.D., and Reddel, R.R. (2010). Assaying and investigating Alternative Lengthening of Telomeres activity in human cells and cancers. *FEBS Lett.* 584, 3800–3811.
- Higa, M., Kushiya, T., Kurashige, S., Kohmon, D., Enokitani, K., Iwahori, S., Sugimoto, N., Yoshida, K., and Fujita, M. (2017). Biochimica et Biophysica Acta TRF2 recruits ORC through TRFH domain dimerization. *BBA - Mol. Cell Res.* 1864, 191–201.
- Hildonen, S., Halvorsen, T.G., and Reubsaet, L. (2014). Why less is more when generating tryptic peptides in bottom-up proteomics. *Proteomics* 14, 2031–2041.
- Hirose, T., Virnicchi, G., Tanigawa, A., Naganuma, T., Li, R., Kimura, H., Yokoi, T., Nakagawa, S., Benard, M., Fox, A.H., et al. (2014). NEAT1 long noncoding RNA regulates transcription via protein sequestration within subnuclear bodies. *Mol. Biol. Cell* 25, 169–183.
- Hsu, J.L., Huang, S.Y., Chow, N.H., and Chen, S.H. (2003). Stable-Isotope Dimethyl Labeling for Quantitative Proteomics. *Anal. Chem.* 75, 6843–6852.
- Hu, Q., Noll, R.J., Li, H., Makarov, A., Hardman, M., and Cooks, R.G. (2005). The Orbitrap: A new mass spectrometer. *J. Mass Spectrom.* 40, 430–443.
- Hubner, N.C., and Mann, M. (2011). Extracting gene function from protein-protein interactions using Quantitative BAC Interactomics (QUBIC). *Methods* 53, 453–459.
- Hughes, C.S., Foehr, S., Garfield, D.A., Furlong, E.E., Steinmetz, L.M., and Krijgsveld, J. (2014). Ultrasensitive proteome analysis using paramagnetic bead technology. *Mol. Syst. Biol.* 10, 757–757.
- Huttlin, E.L., Ting, L., Bruckner, R.J., Gebreab, F., Gygi, M.P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K., et al. (2015). The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* 162, 425–440.
- Huttlin, E.L., Bruckner, R.J., Paulo, J.A., Cannon, J.R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M.P., Parzen, H., et al. (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature* 545, 505–509.

- Imamura, K., Imamachi, N., Akizuki, G., Kumakura, M., Kawaguchi, A., Nagata, K., Kato, A., Kawaguchi, Y., Sato, H., Yoneda, M., et al. (2014). Long Noncoding RNA NEAT1-Dependent SFPQ Relocation from Promoter Region to Paraspeckle Mediates IL8 Expression upon Immune Stimuli. *Mol. Cell* 53, 393–406.
- Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide in vivo analysis of translation with sub-codon resolution by ribosome profiling. *Science* 324, 218–224.
- Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., and Mann, M. (2005). Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein. *Mol. Cell. Proteomics* 4, 1265–1272.
- Jahn, A., Rane, G., Paszkowski-Rogacz, M., Sayols, S., Bluhm, A., Han, C., Draškovič, I., Londoño-Vallejo, J.A., Kumar, A.P., Buchholz, F., et al. (2017). ZBTB48 is both a vertebrate telomere-binding protein and a transcriptional activator. *EMBO Rep.* 1–18.
- Jambor, H., Surendranath, V., Kalinka, A.T., Mejsstrik, P., Saalfeld, S., and Tomancak, P. (2015). Systematic imaging reveals features and changing localization of mRNAs in *Drosophila* development. *Elife* 2015, 1–18.
- Janouskova, E., Necasova, I., Pavlouskova, J., Zimmermann, M., Hluchy, M., Marini, V., Novakova, M., and Hofr, C. (2015). Human Rap1 modulates TRF2 attraction to telomeric DNA. *Nucleic Acids Res.* 43, 2691–2700.
- Jiang, J., Wang, Y., Sušac, L., Chan, H., Basu, R., Zhou, Z.H., and Feigon, J. (2018). Structure of Telomerase with Telomeric DNA. *Cell* 173, 1179–1190.
- Jorgenson, J.W. (2010). Capillary Liquid Chromatography at Ultrahigh Pressures. *Annu. Rev. Anal. Chem.* 3, 129–150.
- Kai, M. (2016). Roles of RNA-binding proteins in DNA damage response. *Int. J. Mol. Sci.* 17, 1–9.
- Kalifa, Y., Huang, T., Rosen, L.N., Chatterjee, S., and Gavis, E.R. (2006). Glorund, a *Drosophila* hnRNP F/H Homolog, Is an Ovarian Repressor of nanos Translation. *Dev. Cell* 10, 291–301.
- Kalinka, A.T., Varga, K.M., Gerrard, D.T., Preibisch, S., Corcoran, D.L., Jarrells, J., Ohler, U., Bergman, C.M., and Tomancak, P. (2010). Gene expression divergence recapitulates the developmental hourglass model. *Nature* 468, 811–816.

- Kappei, D., Butter, F., Benda, C., Scheibe, M., Draškovič, I., Stevense, M., Novo, C.L., Basquin, C., Araki, M., Araki, K., et al. (2013). HOT1 is a mammalian direct telomere repeat-binding protein contributing to telomerase recruitment. *EMBO J.* 32, 1681–1701.
- Kappei, D., Scheibe, M., Paszkowski-Rogacz, M., Bluhm, A., Gossmann, T.I., Dietz, S., Dejung, M., Herlyn, H., Buchholz, F., Mann, M., et al. (2017). Phylointeractomics reconstructs functional evolution of protein binding. *Nat. Commun.* 8, 14334.
- Karlseder, J., Broccoli, D., Dai, Y., Hardy, S., and de Lange, T. (1999). p53- and ATM-dependent apoptosis induced by telomeres lacking TRF2. *Science* 283, 1321–1325.
- Karr, T.L. (2007). Fruit flies and the sperm proteome. *Hum. Mol. Genet.* 16, 124–133.
- Kaul, Z., Cesare, A.J., Huschtscha, L.I., Neumann, A.A., and Reddel, R.R. (2012). Five dysfunctional telomeres predict onset of senescence in human cells. *EMBO Rep.* 13, 52–59.
- 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.
- El Kennani, S., Crespo, M., Govin, J., and Pflieger, D. (2018). Proteomic Analysis of Histone Variants and Their PTMs: Strategies and Pitfalls. *Proteomes* 6, 29.
- Khoury, G.A., Baliban, R.C., and Floudas, C.A. (2014). Proteome-wide post-translational modification statistics: Frequency analysis and curation of the swiss-prot database. *Sci. Rep.* 1, 1–5.
- Kim, N.W., Piatyszek, K.R., Harley, C.B., West, M.D., Ho, P.L., Coviello, G.M., Wright, W.E., Weinrich, S.L., Shay, J.W., and Prowse, M.A. (1994). Specific association of human telomerase activity with immortal cells and cancer. *Science*. 226, 2011–2015.
- Kim, W., Bennett, E.J., Huttlin, E.L., Guo, A., Li, J., Possemato, A., Sowa, M.E., Rad, R., Rush, J., Comb, M.J., et al. (2011). Resource Systematic and Quantitative Assessment of the Ubiquitin-Modified Proteome. *Mol. Cell* 44, 325–340.
- Kirkpatrick, D.S., Gerber, S.A., and Gygi, S.P. (2005). The absolute quantification strategy: A general procedure for the quantification of proteins and post-translational modifications. *Methods* 35, 265–273.
- Köcher, T., Swart, R., and Mechtler, K. (2011). Ultra-high-pressure RPLC hyphenated to an LTQ-orbitrap velos reveals a linear relation between peak capacity and number of identified peptides. *Anal. Chem.* 83, 2699–2704.

- Krijgsveld, J., Ketting, R.F., Mahmoudi, T., Johansen, J., Artal-Sanz, M., Verrijzer, C.P., Plasterk, R.H.A., and Heck, A.J.R. (2003). Metabolic labeling of *C. elegans* and *D. melanogaster* for quantitative proteomics. *Nat. Biotechnol.* *21*, 927–931.
- Kronja, I., Whitfield, Z.J., Yuan, B., Dzek, K., Kirkpatrick, J., Krijgsveld, J., and Orr-Weaver, T.L. (2014b). Quantitative proteomics reveals the dynamics of protein changes during *Drosophila* oocyte maturation and the oocyte-to-embryo transition. *Proc. Natl. Acad. Sci. U. S. A.* *111*, 16023–16028.
- Kronja, I., Yuan, B., Eichhorn, S.W., Dzek, K., Krijgsveld, J., Bartel, D.P., and Orr-Weaver, T.L. (2014a). Widespread changes in the posttranscriptional landscape at the *Drosophila* oocyte-to-embryo transition. *Cell Rep.* *7*, 1495–1508.
- Krueger, M., Moser, M., Ussar, S., Thievensen, I., Lubner, C.A., Forner, F., Schmidt, S., Zanivan, S., Faessler, R., and Mann, M. (2008). SILAC Mouse for Quantitative Proteomics Uncover Kindlin-3 as an Essential Factor for Red Blood Cell Function. *Cell* *134*, 353–364.
- 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.
- Lackner, D.H., Durocher, D., and Karlseder, J. (2011). A siRNA-based screen for genes involved in chromosome end protection. *PLoS One* *6*, 1–8.
- Lai, Y., Zhang, X., Zhang, Z., Shu, Y., Luo, X., Yang, Y., Wang, X., Yang, G., Li, L., and Feng, Y. (2013). The microRNA-27a: ZBTB10-specificity protein pathway is involved in follicle stimulating hormone-induced VEGF, Cox2 and survivin expression in ovarian epithelial cancer cells. *Int. J. Oncol.* *42*, 776–784.
- Lambert, J., Mitchell, L., Rudner, A., Baetz, K., and Figgeys, D. (2009). A Novel Proteomics Approach for the Discovery of Chromatin-associated Protein Networks. *Mol. Cell. Proteomics* *8.4*, 870–882.
- Landsverk, H.B., Mora-Bermúdez, F., Landsverk, O.J.B., Hasvold, G., Naderi, S., Bakke, O., Ellenberg, J., Collas, P., Syljuåsen, R.G., and Küntziger, T. (2010). The protein phosphatase 1 regulator PNUTS is a new component of the DNA damage response. *EMBO Rep.* *11*, 868–875.
- De Lange, T. (2005). Shelterin: the protein complex that shapes and safeguards human telomeres. *Genes Dev.* *19*, 2100–2110.

- De Lange, T. (2010). How shelterin solves the telomere end-protection problem. *Cold Spring Harb. Symp. Quant. Biol.* 75, 167–177.
- Lasonder, E., Ishihama, Y., Andersen, J.S., Vermunt, A.M.W., Pain, A., Sauerwein, R.W., Eling, W.M.C., Hall, N., Waters, A.P., Stunnenberg, H.G., et al. (2002). Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* 419, 537–542.
- Law, M.J., Lower, K.M., Voon, H.P.J., Hughes, J.R., Garrick, D., Viprakasit, V., Mitson, M., Gobbi, M. De, Marra, M., Morris, A., et al. (2010). ATR-X Syndrome Protein Targets Tandem Repeats and Influences Allele-Specific Expression in a Size-Dependent Manner. *Cell* 143, 367–378.
- Lecuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., Babak, T., Cerovina, T., Hughes, T.R., Tomancak, P., and Krause, H.M. (2007). Global Analysis of mRNA Localization Reveals a Prominent Role in Organizing Cellular Architecture and Function. *Cell* 131, 174–187.
- Lee, J.H., You, J., Dobrota, E., and Skalnik, D.G. (2010). Identification and characterization of a novel human PP1 phosphatase complex. *J. Biol. Chem.* 285, 24466–24476.
- Lee, M., Hills, M., Conomos, D., Stutz, M.D., Dagg, R. a., Lau, L.M.S., Reddel, R.R., and Pickett, H. a. (2014). Telomere extension by telomerase and ALT generates variant repeats by mechanistically distinct processes. *Nucleic Acids Res.* 42, 1733–1746.
- Lee, S.S., Bohrson, C., Pike, A.M., Wheelan, S.J., and Greider, C.W. (2015). ATM Kinase Is Required for Telomere Elongation in Mouse and Human Cells. *Cell Rep.* 13, 1623–1632.
- Leitner, A. (2018). A review of the role of chemical modification methods in contemporary mass spectrometry-based proteomics research. *Anal. Chim. Acta* 1000, 2–19.
- Leman, A.R., Noguchi, C., Lee, C.Y., and Noguchi, E. (2010). Human Timeless and Tipin stabilize replication forks and facilitate sister-chromatid cohesion. *J. Cell Sci.* 123, 660–670.
- Leman, A.R., Dheekollu, J., Deng, Z., Lee, S.W., Das, M.M., Lieberman, P.M., and Noguchi, E. (2012). Timeless preserves telomere length by promoting efficient DNA replication through human telomeres. *Cell Cycle* 11, 2337–2347.
- Lenain, C., Bauwens, S., Amiard, S., Brunori, M., Giraud-Panis, M.J., and Gilson, E. (2006). The Apollo 5' Exonuclease Functions Together with TRF2 to Protect Telomeres from DNA Repair. *Curr. Biol.* 16, 1303–1310.

- Li, B., Oestreich, S., and Lange, T. De (2000). Identification of Human Rap1: Implications for Telomere Evolution. *Cell* 101, 471–483.
- Li, J.S.Z., Fuste, J.M., Simavorian, T., Bartocci, C., Tsai, J., Karlseder, J., and Denchi, E.L. (2017). TZAP: A telomere-associated protein involved in telomere length control. *Science*. 1–5.
- Li, L., Germain, D.R., Poon, H.-Y., Hildebrandt, M.R., Monckton, E.A., McDonald, D., Hendzel, M.J., and Godbout, R. (2016). DEAD Box 1 Facilitates Removal of RNA and Homologous Recombination at DNA Double Strand Breaks. *Mol. Cell. Biol.* 36, MCB.00415-16.
- Li, X., Mertens-Talcott, S.U., Zhang, S., Kim, K.H., Ball, J., and Safe, S. (2010). MicroRNA-27a indirectly regulates estrogen receptor α expression and hormone responsiveness in MCF-7 breast cancer cells. *Endocrinology* 151, 2462–2473.
- Lieber, M.R. (2010). The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway. *Annu. Rev. Biochem.* 79, 181–211.
- Lingner, J. (1997). Reverse transcriptase motifs in the catalytic subunit of telomerase. *Science*. 276, 561–566.
- Liu, D., Safari, A., O'Connor, M.S., Chan, D.W., Laegeler, A., Qin, J., and Songyang, Z. (2004a). PTOP interacts with POT1 and regulates its localization to telomeres. *Nat. Cell Biol.* 6, 673–680.
- Liu, D., O'Connor, M.S., Qin, J., and Songyang, Z. (2004b). Telosome, a mammalian telomere-associated complex formed by multiple telomeric proteins. *J. Biol. Chem.* 279, 51338–51342.
- Liu, Y., Kha, H., Ungrin, M., Robinson, M.O., and Harrington, L. (2002). Preferential maintenance of critically short telomeres in mammalian cells heterozygous for *mTert*. *Proc. Natl. Acad. Sci.* 99, 3597–3602.
- Liu, Y., Beyer, A., and Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* 165, 535–550.
- Lovejoy, C.A., Li, W., Reisenweber, S., Thongthip, S., Bruno, J., de Lange, T., De, S., Petrini, J.H.J., Sung, P.A., Jasin, M., et al. (2012). Loss of ATRX, Genome Instability, and an Altered DNA Damage Response Are Hallmarks of the Alternative Lengthening of Telomeres Pathway. *PLOS Genet.* 8, 12–15.

- Lu, P., Vogel, C., Wang, R., Yao, X., and Marcotte, E.M. (2007). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* 25, 117–124.
- Luke, B., and Lingner, J. (2009). TERRA: telomeric repeat-containing RNA. *EMBO J.* 28, 2503–2510.
- Lundberg, E., Fagerberg, L., Klevebring, D., Matic, I., Geiger, T., Cox, J., Älgenäs, C., Lundberg, J., Mann, M., and Uhlen, M. (2010). Defining the transcriptome and proteome in three functionally different human cell lines. *Mol. Syst. Biol.* 6, 1–10.
- Luo, Z., Dai, Z., Xie, X., Feng, X., Liu, D., Songyang, Z., and Xiong, Y. (2015). TeloPIN: a database of telomeric proteins interaction network in mammalian cells. *Database* 2015, 1–9.
- Maciejowski, J., and de Lange, T. (2017). Telomeres in cancer: tumour suppression and genome instability. *Nat. Rev. Mol. Cell Biol.* 1–12.
- El Maï, M., Wagner, K.D., Michiels, J.F., Ambrosetti, D., Borderie, A., Destree, S., Renault, V., Djerbi, N., Giraud-Panis, M.J., Gilson, E., et al. (2014). The telomeric protein TRF2 regulates angiogenesis by binding and activating the PDGFR β promoter. *Cell Rep.* 9, 1047–1060.
- Maicher, A., Lockhart, A., and Luke, B. (2014). Breaking new ground: digging into TERRA function. *Biochim. Biophys. Acta* 1839, 387–394.
- Maier, T., Güell, M., and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. *FEBS Lett.* 583, 3966–3973.
- Makarov, A. (2000). Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Anal. Chem.* 72, 1156–1162.
- Makarov, A., and Denisov, E. (2009). Dynamics of Ions of Intact Proteins in the Orbitrap Mass Analyzer. *JAM* 20, 1486–1495.
- Makarov, A., Denisov, E., Kholomeev, A., Balschun, W., Lange, O., Strupat, K., and Horning, S. (2006). Performance Evaluation of a Hybrid Linear Ion Trap / Orbitrap Mass Spectrometer. *Anal. Chem.* 78, 2113–2120.
- Makowski, M.M., Gräwe, C., Foster, B.M., Nguyen, N. V., Bartke, T., and Vermeulen, M. (2018). Global profiling of protein-DNA and protein-nucleosome binding affinities using quantitative mass spectrometry. *Nat. Commun.* 9, 1–10.

- Malovannaya, A., Lanz, R.B., Jung, S.Y., Bulynko, Y., Le, N.T., Chan, D.W., Ding, C., Shi, Y., Yucer, N., Krenciute, G., et al. (2011). Analysis of the human endogenous coregulator complexome. *Cell* **145**, 787–799.
- Marcand, S., Gilson, E., and Shore, D. (2010). A Protein-Counting Mechanism for Telomere Length Regulation in Yeast. *Science*. **986**.
- Martinez, P., Thanasoula, M., Carlos, A.R., Gómez-López, G., Tejera, A.M., Schoeftner, S., Dominguez, O., Pisano, D.G., Tarsounas, M., and Blasco, M.A. (2010). Mammalian Rap1 controls telomere function and gene expression through binding to telomeric and extratelomeric sites. *Nat. Cell Biol.* **12**, 768–782.
- Martínez, P., and Blasco, M.A. (2011). Telomeric and extra-telomeric roles for telomerase and the telomere-binding proteins. *Nat. Rev. Cancer* **11**, 161–176.
- Martínez, P., and Blasco, M.A. (2015). Replicating through telomeres: A means to an end. *Trends Biochem. Sci.* **40**, 504–515.
- Martínez, P., and Blasco, M.A. (2017). Telomere-driven diseases and telomere-targeting therapies. *J. Cell Biol.* **216**, 875–887.
- Marzec, P., Armenise, C., Perot, G., Roumelioti, F., Basyuk, E., Gagos, S., Chibon, F., and Dejardin, J. (2015). Nuclear-Receptor-Mediated Telomere Insertion Leads to Genome Instability in ALT Cancers. *Cell* **160**, 913–927.
- Mcshane, E., Sin, C., Zauber, H., Wells, J.N., Donnelly, N., Wang, X., Hou, J., Chen, W., Storchova, Z., Marsh, J.A., et al. (2016). Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation. *Cell* **167**, 803–806.
- Mendez-Bermudez, A., Lototska, L., Bauwens, S., Giraud-Panis, M.J., Croce, O., Jamet, K., Irizar, A., Mowinckel, M., Koundrioukoff, S., Nottet, N., et al. (2018). Genome-wide Control of Heterochromatin Replication by the Telomere Capping Protein TRF2. *Mol. Cell* **70**, 449–461.
- Mertens-Talcott, S.U., Chintharlapalli, S., Li, X., and Safe, S. (2007). The oncogenic microRNA-27a targets genes that regulate specificity protein transcription factors and the G2-M checkpoint in MDA-MB-231 breast cancer cells. *Cancer Res.* **67**, 11001–11011.

- Michalski, A., Damoc, E., Hauschild, J.-P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M., and Horning, S. (2011). Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer. *Mol. Cell. Proteomics* *10*.
- Miyake, Y., Nakamura, M., Nabetani, A., Shimamura, S., Tamura, M., Yonehara, S., Saito, M., and Ishikawa, F. (2009). RPA-like Mammalian Ctc1-Stn1-Ten1 Complex Binds to Single-Stranded DNA and Protects Telomeres Independently of the Pot1 Pathway. *Mol. Cell* *36*, 193–206.
- Morin, X., Daneman, R., Zavortink, M., and Chia, W. (2001). A protein trap strategy to detect GFP-tagged proteins expressed from their endogenous loci in *Drosophila*. *Proc. Natl. Acad. Sci.* *98*, 15050–15055.
- Mostovenko, E., Hassan, C., Rattke, J., Deelder, A.M., van Veelen, P.A., and Palmblad, M. (2013). Comparison of peptide and protein fractionation methods in proteomics. *Eur. Proteomics Assoc.* *1*, 30–37.
- Murata, Y., and Wharton, R.P. (1995). Binding of pumilio to maternal hunchback mRNA is required for posterior patterning in *drosophila* embryos. *Cell* *80*, 747–756.
- Nagarkar-Jaiswal, S., DeLuca, S.Z., Lee, P., Lin, W., Pan, H., Zuo, Z., Lv, J., Spradling, A.C., and Bellen, H.J. (2015). A genetic toolkit for tagging intronic MiMIC containing genes. *Elife* *4*, 1–8.
- Nandakumar, J., and Cech, T.R. (2013). Finding the end: Recruitment of telomerase to telomeres. *Nat. Rev. Mol. Cell Biol.* *14*, 69–82.
- Nandakumar, J., Bell, C.F., Weidenfeld, I., Zaug, A.J., Leinwand, L.A., and Cech, T.R. (2012). The TEL patch of telomere protein TPP1 mediates telomerase recruitment and processivity. *Nature* *492*, 285–289.
- Nesvizhskii, A.I. (2014). Proteogenomics: Concepts, applications and computational strategies. *Nat. Methods* *11*, 1114–1125.
- O'Connor, M.S., Safari, A., Xin, H., Liu, D., and Songyang, Z. (2006). A critical role for TPP1 and TIN2 interaction in high-order telomeric complex assembly. *Proc. Natl. Acad. Sci.* *103*, 11874–11879.
- O'Sullivan, R.J., and Almouzni, G. (2014). Assembly of telomeric chromatin to create ALternative endings. *Trends Cell Biol.* *24*, 675–685.

- O'Sullivan, R.J., and Karlseder, J. (2010). Telomeres: protecting chromosomes against genome instability. *Nat. Rev. Mol. Cell Biol.* **11**, 171–181.
- Okamoto, K., Bartocci, C., Ouzounov, I., Diedrich, J.K., Iii, J.R.Y., and Denchi, E.L. (2013). A two-step mechanism for TRF2-mediated chromosome-end protection. *Nature* **494**, 502–505.
- Olsen, J. V., and Mann, M. (2013). Status of Large-scale Analysis of Post-translational Modifications by Mass Spectrometry. *Mol. Cell. Proteomics* **12**, 3444–3452.
- Olsen, J. V., Ong, S.-E., and Mann, M. (2004). Trypsin Cleaves Exclusively C-terminal to Arginine and Lysine Residues. *Mol. Cell. Proteomics* **3**, 608–614.
- Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S., and Mann, M. (2007). Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **4**, 709–712.
- Olsen, J. V., Godoy, L.M.F. De, Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S., and Mann, M. (2010). Parts per Million Mass Accuracy on an Orbitrap Mass Spectrometer via Lock Mass Injection into a C-trap. *Mol. Cell. Proteomics* **4**, 2010–2021.
- Ong, S.-E., and Mann, M. (2006). A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat. Protoc.* **1**, 2650–2660.
- Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., and Mann, M. (2002). Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Mol. Cell. Proteomics* **1**, 376–386.
- Ong, S., Mittler, G., and Mann, M. (2004). Identifying and quantifying in vivo methylation sites by heavy methyl SILAC. *Nat. Methods* **1**, 1–8.
- Ovadya, Y., Landsberger, T., Leins, H., Vadai, E., Gal, H., Biran, A., Yosef, R., Sagiv, A., Agrawal, A., Shapira, A., et al. (2018). Impaired immune surveillance accelerates accumulation of senescent cells and aging. *Nat. Commun.* **9**, 1–15.
- Overbeek, M. Van, and De Lange, T. (2006). Apollo, an Artemis-Related Nuclease, Interacts with TRF2 and Protects Human Telomeres in S Phase. *Curr. Biol.* **16**, 1295–1302.
- Palm, W., and de Lange, T. (2008). How shelterin protects mammalian telomeres. *Annu. Rev. Genet.* **42**, 301–334.
- Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415.

- Paulsen, R.D., Soni, D. V., Wollman, R., Hahn, A.T., Yee, M.C., Guan, A., Hesley, J.A., Miller, S.C., Cromwell, E.F., Solow-Cordero, D.E., et al. (2009). A Genome-wide siRNA Screen Reveals Diverse Cellular Processes and Pathways that Mediate Genome Stability. *Mol. Cell* 35, 228–239.
- Perez-Torrado, R., Yamada, D., and Defossez, P.A. (2006). Born to bind: The BTB protein-protein interaction domain. *BioEssays* 28, 1194–1202.
- Peshkin, L., Wuehr, M., Pearl, E., Haas, W., Freeman, R.M., Gerhart, J.C., Klein, A.M., Horb, M., Gygi, S.P., and Kirschner, M.W. (2015). On the Relationship of Protein and mRNA Dynamics in Vertebrate Embryonic Development. *Dev. Cell* 35, 383–394.
- Pflieger, D., Jünger, M.A., Müller, M., Rinner, O., Lee, H., Gehrig, P.M., Gstaiger, M., and Aebersold, R. (2008). Quantitative Proteomic Analysis of Protein Complexes. *Mol. Cell. Proteomics* 7, 326–346.
- Popuri, V., Hsu, J., Khadka, P., Horvath, K., Liu, Y., Croteau, D.L., and Bohr, V.A. (2014). Human RECQL1 participates in telomere maintenance. *Nucleic Acids Res.* 42, 5671–5688.
- Porro, A., Feuerhahn, S., Reichenbach, P., and Lingner, J. (2010). Molecular Dissection of Telomeric Repeat-Containing RNA Biogenesis Unveils the Presence of Distinct and Multiple Regulatory Pathways. *Mol. Cell. Biol.* 30, 4808–4817.
- Porro, A., Feuerhahn, S., Delafontaine, J., Riethman, H., Rougemont, J., and Lingner, J. (2014a). Functional characterization of the TERRA transcriptome at damaged telomeres. *Nat. Commun.* 5, 1–13.
- Porro, A., Feuerhahn, S., and Lingner, J. (2014b). TERRA-Reinforced Association of LSD1 with MRE11 Promotes Processing of Uncapped Telomeres. *Cell Rep.* 6, 765–776.
- Poser, I., Sarov, M., Hutchins, J.R.A., Heriche, J., Toyoda, Y., Pozniakovsky, A., Weigl, D., Nitzsche, A., Hegemann, B., Bird, A.W., et al. (2008). BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nat. Methods* 5, 409–415.
- Pratt, J.M., Petty, J., Riba-Garcia, I., Robertson, D.H.L., Gaskell, S.J., Oliver, S.G., and Beynon, R.J. (2002). Dynamics of Protein Turnover, a Missing Dimension in Proteomics. *Mol. Cell. Proteomics* 1, 579–591.
- Rai, R., Chen, Y., Lei, M., and Chang, S. (2016). TRF2-RAP1 is required to protect telomeres from engaging in homologous recombination-mediated deletions and fusions. *Nat. Commun.* 7, 1–13.

- Rappsilber, J., Ryder, U., Lamond, A.I., and Mann, M. (2002). Large-Scale Proteomic Analysis of the Human Spliceosome. *Genome Res.* 12, 1231–1245.
- 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.
- Robin, J.D., Ludlow, A.T., Batten, K., Magdinier, F., Stadler, G., Wagner, K.R., Shay, J.W., and Wright, W.E. (2014). Telomere position effect: Regulation of gene expression with progressive telomere shortening over long distances. *Genes Dev.* 28, 2464–2476.
- Robinson, P.J., and Reddel, R.R. (2007). Protein Composition of Catalytically Active Human Telomerase from Immortal Cells. *Science.* 315, 1850–1853.
- Ross, P.L., Huang, Y.N., Marchese, J.N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., et al. (2004). Multiplexed Protein Quantitation in *Saccharomyces cerevisiae* Using Amine-reactive Isobaric Tagging Reagents. *Mol. Cell. Proteomics* 3, 1154–1169.
- Rothbauer, U., Zolghadr, K., Muyldermans, S., Schepers, A., Cardoso, M.C., and Leonhardt, H. (2008). A Versatile Nanotrap for Biochemical and Functional Studies with Fluorescent Fusion Proteins. *Mol. Cell. Proteomics* 7, 282–289.
- Roumelioti, F., Sotiriou, S.K., Katsini, V., Chiourea, M., Halazonetis, T.D., and Gagos, S. (2016). Alternative lengthening of human telomeres is a conservative DNA replication process with features of break-induced replication. *EMBO Rep.* e201643169.
- Ruggles, K. V., Krug, K., Wang, X., Clauser, K.R., Wang, J., Payne, S.H., Fenyő, D., Zhang, B., and Mani, D.R. (2017). Methods, Tools and Current Perspectives in Proteogenomics. *Mol. Cell. Proteomics* 16, 959–981.
- San Filippo, J., Sung, P., and Klein, H. (2008). Mechanism of Eukaryotic Homologous Recombination. *Annu. Rev. Biochem.* 77, 229–257.
- Sarek, G., Vannier, J., Panier, S., Petrini, J.H.J., and Boulton, S.J. (2015). TRF2 Recruits RTEL1 to Telomeres in S Phase to Promote T-Loop Unwinding. *Mol. Cell* 57, 622–635.
- Sarov, M., Barz, C., Jambor, H., Hein, M.Y., Schmied, C., Suchold, D., Stender, B., Janosch, S., Vinay Vikas, K.J., Krishnan, R.T., et al. (2016). A genome-wide resource for the analysis of protein localisation in *Drosophila*. *Elife* 5, 1–38.

- Schoeftner, S., and Blasco, M.A. (2008). Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II. *Nat. Cell Biol.* *10*, 228–236.
- Schoeftner, S., and Blasco, M.A. (2010). Chromatin regulation and non-coding RNAs at mammalian telomeres. *Semin. Cell Dev. Biol.* *21*, 186–193.
- Schwanhäusser, B., Gossen, M., Dittmar, G., and Selbach, M. (2009). Global analysis of cellular protein translation by pulsed SILAC. *Proteomics* *9*, 205–209.
- Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature* *473*, 337–342.
- Schwartz, J.C., and Senko, M.W. (2002). A Two-Dimensional Quadrupole Ion Trap Mass Spectrometer. *Am. Soc. Mass Spectrom.* *13*, 659–669.
- Scigelova, M., and Makarov, A. (2009). Advances in bioanalytical LC – MS using the Orbitrap™ mass analyzer. *Bioanalysis* *1*, 741–754.
- Scott, G.K., Mattie, M.D., Berger, C.E., Benz, S.C., and Benz, C.C. (2006). Rapid alteration of microRNA levels by histone deacetylase inhibition. *Cancer Res.* *66*, 1277–1281.
- Sfeir, Agnel J, de Lange, T. (2012). Removal of Shelterin Reveals the Telomere End-Replication Problem. *Science*. *336*, 593–597.
- Sfeir, A., Kosiyatrakul, S.T., Hockemeyer, D., MacRae, S.L., Karlseder, J., Schildkraut, C.L., and de Lange, T. (2009). Mammalian Telomeres Resemble Fragile Sites and Require TRF1 for Efficient Replication. *Cell* *138*, 90–103.
- Sharma, K., D’Souza, R.C.J., Tyanova, S., Schaab, C., Wiśniewski, J.R., Cox, J., and Mann, M. (2014). Ultradeep Human Phosphoproteome Reveals a Distinct Regulatory Nature of Tyr and Ser/Thr-Based Signaling. *Cell Rep.* *8*, 1583–1594.
- Shevchenko, A., Tomas, H., Havliš, J., Olsen, J. V., and Mann, M. (2006). In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* *1*, 2856–2860.
- Al Shweiki, M.H.D.R., Mönchgesang, S., Majovsky, P., Thieme, D., Trutschel, D., and Hoehenwarter, W. (2017). Assessment of Label-Free Quantification in Discovery Proteomics and Impact of Technological Factors and Natural Variability of Protein Abundance. *J. Proteome Res.* *16*, 1410–1424.

- Simonet, T., Zaragosi, L.-E., Philippe, C., Lebrigand, K., Schouteden, C., Augereau, A., Bauwens, S., Ye, J., Santagostino, M., Giulotto, E., et al. (2011). The human TTAGGG repeat factors 1 and 2 bind to a subset of interstitial telomeric sequences and satellite repeats. *Cell Res.* *21*, 1028–1038.
- Singh, S., Springer, M., Steen, J., Kirschner, M.W., and Steen, H. (2009). FLEXIQuant: A novel tool for the absolute quantification of proteins, and the simultaneous identification and quantification of potentially modified peptides. *J. Proteome Res.* *8*, 2201–2210.
- Smith, R.D., Shen, Y., and Tang, K. (2004). Ultrasensitive and Quantitative Analyses from Combined Separations - Mass Spectrometry for the Characterization of Proteomes. *Acc. Chem. Res.* *37*, 269–278.
- Smits, A.H., and Vermeulen, M. (2016). Characterizing Protein–Protein Interactions Using Mass Spectrometry: Challenges and Opportunities. *Trends Biotechnol.* *34*, 825–834.
- Sobinoff, A.P., and Pickett, H.A. (2017). Alternative Lengthening of Telomeres : DNA Repair Pathways Converge. *Trends Genet.* *33*, 921–932.
- Sonenberg, N., and Hinnebusch, A.G. (2009). Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. *Cell* *136*, 731–745.
- Soudet, J., Jolivet, P., and Teixeira, M.T. (2014). Elucidation of the DNA end-replication problem in *saccharomyces cerevisiae*. *Mol. Cell* *53*, 954–964.
- Sowell, R.A., Hersberger, K.E., Kaufman, T.C., and Clemmer, D.E. (2007). Examining the Proteome of *Drosophila* Across Organism Lifespan. *J. Proteome Res.* *6*, 3637–3647.
- St Johnston, D. (2002). The art and design of genetic screens: *Drosophila melanogaster*. *Nat. Rev. Genet.* *3*, 176–188.
- Stadler, C., Rexhepaj, E., Singan, V.R., Murphy, R.F., Pepperkok, R., Uhlén, M., Simpson, J.C., and Lundberg, E. (2013). Immunofluorescence and fluorescent-protein tagging show high correlation for protein localization in mammalian cells. *Nat. Methods* *10*, 315–323.
- Steen, H., and Mann, M. (2004). THE ABC' S (AND XYZ' S) OF PEPTIDE SEQUENCING. *Nat. Rev. Mol. Cell Biol.* *5*, 699–711.
- van Steensel, B., and de Lange, T. (1997). Control of telomere length by the human telomeric protein TRF1. *Nature* *385*, 740–743.
- van Steensel, B., Smogorzewska, a, and de Lange, T. (1998). TRF2 protects human telomeres from end-to-end fusions. *Cell* *92*, 401–413.

- Stracker, T.H., and Petrini, J.H.J. (2011). The MRE11 complex: starting from the ends. *Nat. Rev. Mol. Cell Biol.* *12*, 90–103.
- Sury, M.D., Chen, J.-X., and Selbach, M. (2010). The SILAC fly allows for accurate protein quantification in vivo. *Mol. Cell. Proteomics* *9*, 2173–2183.
- Tadros, W., and Lipshitz, H.D. (2009). The maternal-to-zygotic transition: a play in two acts. *Development* *136*, 3033–3042.
- Takahama, K., Takada, A., Tada, S., Shimizu, M., Sayama, K., Kurokawa, R., and Oyoshi, T. (2013). Regulation of telomere length by G-quadruplex telomere DNA- and TERRA-binding protein TLS/FUS. *Chem. Biol.* *20*, 341–350.
- Takemori, N., and Yamamoto, M. (2009). Proteome mapping of the *Drosophila melanogaster* male reproductive system. *Proteomics* *9*, 2484–2493.
- Tang, J., Kan, Z.Y., Yao, Y., Wang, Q., Hao, Y.H., and Tan, Z. (2008). G-quadruplex preferentially forms at the very 3' end of vertebrate telomeric DNA. *Nucleic Acids Res.* *36*, 1200–1208.
- Tardat, M., and Déjardin, J. (2017). Telomere chromatin establishment and its maintenance during mammalian development. *Chromosoma* *1–16*.
- Tatsumi, Y., Ezura, K., Yoshida, K., Yugawa, T., Narisawa-Saito, M., Kiyono, T., Ohta, S., Obuse, C., and Fujita, M. (2008). Involvement of human ORC and TRF2 in pre-replication complex assembly at telomeres. *Genes to Cells* *13*, 1045–1059.
- Teixeira, M.T., Arneric, M., Sperisen, P., and Lingner, J. (2004). Telomere length homeostasis is achieved via a switch between telomerase-extendible and-nonextendible states. *Cell* *117*, 323–335.
- Teo, H., Ghosh, S., Luesch, H., Ghosh, A., Wong, E.T., Malik, N., Orth, A., de Jesus, P., Perry, A.S., Oliver, J.D., et al. (2010). Telomere-independent Rap1 is an IKK adaptor and regulates NF- κ B-dependent gene expression. *Nat. Cell Biol.* *12*, 758–767.
- Thakur, S.S., Geiger, T., Chatterjee, B., Bandilla, P., Fröhlich, F., Cox, J., and Mann, M. (2011). Deep and Highly Sensitive Proteome Coverage by LC-MS/MS Without Prefractionation. *Mol. Cell. Proteomics* *10*.
- Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., and Hamon, C. (2003). Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* *75*, 1895–1904.

- Tillotson, L.G. (1999). RIN ZF, a Novel Zinc Finger Gene, Encodes Proteins That Bind to the CACC Element of the Gastrin Promoter. *J. Biol. Chem.* **274**, 8123–8128.
- Tomancak, P., Berman, B.P., Beaton, A., Weiszmamm, R., Kwan, E., Hartenstein, V., Celniker, S.E., and Rubin, G.M. (2007). Global analysis of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.* **8**, 1–24.
- Tong, A.S., Stern, J.L., Sfeir, A., Kartawinata, M., de Lange, T., Zhu, X.D., and Bryan, T.M. (2015). ATM and ATR Signaling Regulate the Recruitment of Human Telomerase to Telomeres. *Cell Rep.* **13**, 1633–1646.
- Tyler, F. (2003). Development of the Fruit Fly *Drosophila melanogaster*. p.
- Udeshi, N.D., Mertins, P., Svinkina, T., and Carr, S.A. (2013). Large-scale identification of ubiquitination sites by mass spectrometry. *Nat. Protoc.* **8**, 1950–1960.
- Vannier, J., Pavicic-Kaltenbrunner, V., Petalcorin, M.I.R., Ding, H., and Boulton, S.J. (2012). RTEL1 Dismantles T Loops and Counteracts Telomeric G4-DNA to Maintain Telomere Integrity. *Cell* **149**, 795–806.
- Vermeulen, M., Hubner, N.C., and Mann, M. (2008). High confidence determination of specific protein-protein interactions using quantitative mass spectrometry. *Curr. Opin. Biotechnol.* **19**, 331–337.
- Victorelli, S., and Passos, J.F. (2017). Telomeres and Cell Senescence - Size Matters Not. *EBioMedicine* **21**, 14–20.
- Vidova, V., and Spacil, Z. (2017). A review on mass spectrometry-based quantitative proteomics: Targeted and data independent acquisition. *Anal. Chim. Acta* **964**, 7–23.
- Wagner, S.A., Beli, P., Weinert, B.T., Schölz, C., Kelstrup, C.D., Young, C., Nielsen, M.L., Olsen, J. V., Brakebusch, C., and Choudhary, C. (2012). Proteomic Analyses Reveal Divergent Ubiquitylation Site Patterns in Murine Tissues. *Mol. Cell. Proteomics* **11**, 1578–1585.
- Walsh, C.T., Garneau-tsodikova, S., and Gatto, G.J. (2005). Protein Posttranslational Modifications: The Chemistry of Proteome Diversifications. *Angew. Chemie* **44**, 7342–7372.
- Wang, C.I., Alekseyenko, A.A., Leroy, G., Elia, A.E.H., Gorchakov, A.A., Britton, L.M.P., 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, F., Podell, E.R., Zaug, A.J., Yang, Y., Baciú, P., Cech, T.R., and Lei, M. (2007). The POT1-TPP1 telomere complex is a telomerase processivity factor. *Nature* **445**, 506–510.
- Wasbrough, E.R., Dorus, S., Hester, S., Howard-murkin, J., Lilley, K., Wilkin, E., Polpitiya, A., Petritis, K., and Karr, T.L. (2010). The *Drosophila melanogaster* sperm proteome-II (DmSP-II). *J. Proteomics* **73**, 2171–2185.
- Westman-Brinkmalm, A., Abramsson, A., Pannee, J., Gang, C., Gustavsson, M.K., von Otter, M., Blennow, K., Brinkmalm, G., Heumann, H., and Zetterberg, H. (2011). SILAC zebrafish for quantitative analysis of protein turnover and tissue regeneration. *J. Proteomics* **75**, 425–434.
- Wood, A.M., Danielsen, J.M.R., Lucas, C.A., Rice, E.L., Scalzo, D., Shimi, T., Goldman, R.D., Smith, E.D., Beau, M.M. Le, and Kosak, S.T. (2014). TRF2 and lamin A/C interact to facilitate the functional organization of chromosome ends. *Nat. Commun.* **5**, 1–9.
- Woods, N.T., Mesquita, R.D., Sweet, M., Carvalho, M.A., Li, X., Liu, Y., Nguyen, H., Thomas, C.E., Iversen Jr, E.S., Marsillac, S., et al. (2014). Charting the Landscape of Tandem BRCT Domain-Mediated Protein Interactions. *Sci. Signal.* **5**, 1–32.
- Wright, W.E., Tesmer, V.M., Liao, M.L., and Shay, J.W. (1999). Normal Human Telomeres Are Not Late Replicating. *Exp. Cell Res.* **251**, 492–499.
- Wu, P., Takai, H., and De Lange, T. (2012). Telomeric 3' overhangs derive from resection by Exo1 and Apollo and fill-in by POT1b-associated CST. *Cell* **150**, 39–52.
- Xi, L., and Cech, T.R. (2014). Inventory of telomerase components in human cells reveals multiple subpopulations of hTR and hTERT. *Nucleic Acids Res.* **42**, 8565–8577.
- Xian, F., Hendrickson, C.L., and Marshall, A.G. (2012). High Resolution Mass Spectrometry. *Anal. Chem.* **84**, 708–719.
- Xin, H., Liu, D., Wan, M., Safari, A., Kim, H., Sun, W., O'Connor, M.S., and Songyang, Z. (2007). TPP1 is a homologue of ciliate TEBP- β and interacts with POT1 to recruit telomerase. *Nature* **445**, 559–562.
- Xing, X., Zhang, C., Li, N., Zhai, L., and Zhu, Y. (2014). Qualitative and quantitative analysis of the adult *Drosophila melanogaster* proteome. *Proteomics* **14**, 286–290.
- Yang, D., Xiong, Y., Kim, H., He, Q., Li, Y., Chen, R., and Songyang, Z. (2011). Human telomeric proteins occupy selective interstitial sites. *Cell Res.* **21**, 1013–1027.

- Ye, J., Lenain, C., Bauwens, S., Rizzo, A., Saint-Léger, A., Poulet, A., Benarroch, D., Magdinier, F., Morere, J., Amiard, S., et al. (2010). TRF2 and Apollo Cooperate with Topoisomerase 2 α to Protect Human Telomeres from Replicative Damage. *Cell* **142**, 230–242.
- Ye, J.Z.S., Hockemeyer, D., Krutchinsky, A.N., Loayza, D., Hooper, S.M., Chait, B.T., and De Lange, T. (2004a). POT1-interaction protein PIP1: A telomere length regulator that recruits POT1 to the TIN2/TRF1 complex. *Genes Dev.* **18**, 1649–1654.
- Ye, J.Z.S., Donigian, J.R., Van Overbeek, M., Loayza, D., Luo, Y., Krutchinsky, A.N., Chait, B.T., and De Lange, T. (2004b). TIN2 binds TRF1 and TRF2 simultaneously and stabilizes the TRF2 complex on telomeres. *J. Biol. Chem.* **279**, 47264–47271.
- Yeung, F., Ramírez, C.M., Mateos-Gomez, P.A., Pinzaru, A., Ceccarini, G., Kabir, S., Fernández-Hernando, C., and Sfeir, A. (2013). Nontelomeric Role for Rap1 in Regulating Metabolism and Protecting against Obesity. *Cell Rep.* **3**, 1847–1856.
- Yuan, F., Li, G., and Tong, T. (2017). Nucleolar and coiled-body phosphoprotein 1 (NOLC1) regulates the nucleolar retention of TRF2. *Cell Death Discov.* **3**, 1–9.
- Zeiler, M., Straube, W.L., Lundberg, E., Uhlen, M., and Mann, M. (2012). A Protein Epitope Signature Tag (PreST) Library Allows SILAC-based Absolute Quantification and Multiplexed Determination of Protein Copy Numbers in Cell Lines. *Mol. Cell. Proteomics* **11**.
- Zhang, P., Pazin, M.J., Schwartz, C.M., Becker, K.G., Wersto, R.P., Dilley, C.M., and Mattson, M.P. (2008). Nontelomeric TRF2-REST Interaction Modulates Neuronal Gene Silencing and Fate of Tumor and Stem Cells. *Curr. Biol.* **18**, 1489–1494.
- Zhao, Y., and Jensen, O.N. (2009). Modification-specific proteomics: Strategies for characterization of post-translational modifications using enrichment techniques. *Proteomics* **9**, 4632–4641.
- Zhao, Y., Zhang, G., He, C., Mei, Y., Shi, Y., and Li, F. (2018). The 11th C2H2 zinc finger and an adjacent C-terminal arm are responsible for TZAP recognition of telomeric DNA. *Cell Res.* **28**, 130–134.
- Zhong, F.L., Batista, L.F.Z., Freund, A., Pech, M.F., Venteicher, A.S., and Artandi, S.E. (2012). TPP1 OB-fold domain controls telomere maintenance by recruiting telomerase to chromosome ends. *Cell* **150**, 481–494.

- Zhou, H., Low, T.Y., Hennrich, M.L., van der Toorn, H., Schwend, T., Zou, H., Mohammed, S., and Heck, A.J.R. (2011). Enhancing the Identification of Phosphopeptides from Putative Basophilic Kinase Substrates Using Ti (IV) Based IMAC Enrichment. *Mol. Cell. Proteomics* 10.
- Zielinska, D.F., Gnad, F., Wisniewski, J.R., and Mann, M. (2010). Precision Mapping of an In Vivo N-Glycoproteome Reveals Rigid Topological and Sequence Constraints. *Cell* 141, 897–907.
- Zubarev, R.A., and Makarov, A. (2013). Orbitrap mass spectrometry. *Anal. Chem.* 85, 5288–5296.