# Dissertation

submitted to the

Combined Faculty of Natural Sciences and Mathematics

of the Ruperto-Carola University of Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Presented by

M.Sc. Matilde Bertolini

born in Parma, Italy

Oral examination: 02/03/2021

# Profiling interactions of proximal nascent chains reveals a general co-translational mechanism of protein complex assembly

Referees:

Prof. Dr. Bernd Bukau

Prof. Dr. Caludio Joazeiro

## Contributions

The experiments and analyses presented in this Thesis were conducted by myself unless otherwise indicated, under the supervision of Dr. Günter Kramer and Prof. Dr. Bernd Bukau. The Disome Selective Profiling (DiSP) technology was developed in collaboration with my colleague Kai Fenzl, with whom I also generated initial datasets of human HEK293-T and U2OS cells. Dr. Ilia Kats developed important bioinformatics tools for the analysis of DiSP data, including RiboSeqTools and the sigmoid fitting algorithm. He also offered great input on statistical analyses. Dr. Frank Tippmann performed analysis of crystal structures.

# TABLE OF CONTENTS

# List of Figures

# List of Tables

# List of Equations

# Abbreviations

Å = Ångström
PTC = peptidyltransferase centre
AP = arrest peptides
RAC = ribosome associated complex
NAC = nascent polypeptide associated complex
ER = endoplasmic reticulum
ORF = open reading frame
TF = trigger factor / transcription factor
RIp-chip = Ribonucleoprotein Immunoprecipitation analysed with DNA chips
SeRP = Selective Ribosome Profiling
RHD = Rel Homology Domain
TMDs = transmembrane domains
UTR = untranslated region
RBP = RNA-binding protein
mRNP = messenger ribonucleoproteins
nt = nucleotide
aa = amino acid
RNase = ribonuclease
DiSP = Disome Selective Profiling
UMI = unique molecular identifier
RPKM = reads per kilobase per million
CDS = coding sequence
GFP = green fluorescent protein
RFP = red fluorescent protein
N-terminus and C-terminus = amino- and carboxy-terminus
PK = Proteinase K
Puro = puromycin
PP-seq = Polysome profiling and RNA sequencing
smFISH = single molecule Fluorescence In Situ Hybridization

# ABSTRACT

The association of proteins into functional oligomeric complexes is crucial for nearly all cellular processes. Despite rapid progress in characterizing the structure of native assemblies, the underlying mechanisms that guide faithful complex formation in the crowded cellular environment are understood only superficially. To secure efficient complex biogenesis and limit the exposure of aggregation-prone intermediates, many proteins assemble co-translationally, via interaction of a fully synthetized and a nascent protein subunit (co-post assembly).

Here, we explore the prevalence and the mechanistic principles of a putative co-translational assembly mechanism, which involves the direct interaction of nascent subunits emerging from proximal ribosomes (co-co assembly). To obtain direct evidence of this putative assembly mode, we apply a newly developed method based on Ribosome Profiling, named Disome Selective Profiling (DiSP), which allows to monitor the conversion of single ribosomes to nascent chain connected ribosome pairs across the proteome with high resolution. We use this approach to analyse co-co assembly in two human cell lines and demonstrate that it constitutes a general mechanism inside cells that is employed by hundreds of high confidence and thousands of low confidence candidates, comprising 11 to 32% of all complex subunits. Analysing the features of the co-co assembly proteome, we reveal that this mechanism guides formation of mostly homomeric complexes and typically relies on interaction of N-terminal nascent chain segments. We further identify five dimerization domains mediating the majority of co-co interactions, which are either partially or completely exposed at the onset of nascent chain dimerization, implying different folding and assembly mechanisms. The detectable fraction of each candidate's nascent chains that co-co assemble is in median 40% and in some cases exceeds 90%, suggesting that this co-translational assembly path may be employed as the main route for complex formation.

To gain deeper insights into the mechanistic basis of co-co assembly, we took a series of experimental approaches that distinguish between interactions of nascent chains emerging from the same or different polysomes (termed assembly *in cis* and *in trans*, respectively). These experiments could not support a model of assembly *in trans*. Conversely, we find indications supporting a *cis* assembly model for nuclear lamin C, one of our high confidence candidates. This mechanism provides a simple explanation for the remarkable specificity of lamin homodimer formation *in vivo*, where splice variants with largely overlapping sequences do not mix. We propose that assembly *in cis* more generally secures specific homomer formation of isoforms and structurally-related proteins which are highly prone to promiscuous interactions inside cells.

In conclusion, this study provides a global annotation of nascent chain interactions across the human proteome and elucidates the basic principles of this widespread assembly pathway. Our findings raise a number of fundamental questions concerning the mechanisms ensuring high-fidelity protein biogenesis, including the implications of co-co assembly on polysome structure, the possible consequences of co-co assembly failure, the inter-dependence with co-translational folding and the synchronization and coordination with translation kinetics.

# ZUSAMMENFASSUNG

Die Assemblierung funktionaler Proteinkomplexe ist für nahezu alle zellulären Prozesse von fundamentaler Bedeutung. Trotz rascher Fortschritte bei der Charakterisierung der Funktion von Proteinkomplexen sind die Mechanismen der Komplexbildung im dicht gepackten Zytoplasma der Zelle noch unvollständig verstanden. Um die Zeitdauer der Exposition aggregationsanfälliger Translationsintermediate zu limitieren und damit eine effiziente Biogenese sicherzustellen, werden viele Komplexe durch die Interaktion einer vollständig synthetisierten und einer entstehenden (naszierenden) Proteinuntereinheit (Co-Post-Assemblierung) bereits co-translational assembliert.
Diese Arbeit untersucht die Prävalenz und die mechanistischen Prinzipien eines mutmaßlichen co-translationalen Assemblierungsmechanismus, der die direkte Wechselwirkung entste¬hender Untereinheiten beinhaltet, die von proximalen Ribosomen synthetisiert werden (Co-Co-Assemblierung). Zur Detektion dieses mutmaßlichen Assemblierungsmodus wurde eine neue Methode etabliert, die auf „ribosome profiling“ basiert und „Disome Selective Profiling“ (DiSP) genannt wird. DiSP detektiert mRNAs, bei denen im Verlauf der Proteinsynthese durch die Interaktion der naszierenden Ketten aus Monosomen Disomen entstehen. DiSP von zwei menschlichen Zelllinien zeigt, dass 11 % aller Proteinkomplexe mit hoher und bis zu 32 % mit geringerer Wahrscheinlicheit co-co-assemblieren. Die Analyse des Co-Co-Assemblierungs-proteoms zeigt, dass dieser Mechanismus besonders die Bildung homomerer Komplexe steuert und typischerweise auf der Wechselwirkung N-terminaler Proteinsegmente beruht. Co-Co Assemblierung wird besonders häufig durch die Interaktion von fünf verschiedenen N-terminalen Dimerisierungsdomänen vermittelt, die zum Beginn der Assemblierung entweder vollständig oder teilweise exponiert sind. Dies impliziert die Existenz unterschiedlicher Faltungs- und Assemblierungs¬mechanismen. Der nachweisbare Anteil co-co assemblierender Untereinheiten liegt im Median bei 40% und übersteigt in einigen Fällen 90%. Diese Zahlen verdeutlichen, dass co-translationale Assemblierung für viele Untereinheiten die wichtigste Route zur Herstellung eines funktionalen Komplexes darstellt.
Um tiefere Einblicke in die mechanistischen Grundlagen der Co-Co-Assemblierung zu erhalten, haben wir eine Reihe experimenteller Ansätze verfolgt, die zwischen Wechselwirkungen gleicher oder verschiedener Polysomen unterscheiden (als Assemblierung in cis bzw. in trans bezeichnet). Diese Experimente konnten keinen Fall einer Assemblierung in trans unterstützen. Umgekehrt finden wir Hinweise, dass Lamin C durch in cis-Assemblierung gebildet wird. Dieser Mechanismus liefert eine einfache Erklärung für die bemerkenswerte Spezifität der Lamin-Homodimer-Bildung in vivo, die die Assemblierung verschiedener Lamin-Isoformen mit identischen Dimerisierungs¬domänen ausschließt. Auf der Basis dieser Ergebnisse schlagen wir vor, dass die Assemblierung in cis die spezifische Homomerbildung von Isoformen und strukturell verwandten Proteinen sicherstellt, und damit promiskuitive, nicht produktive Wechselwirkungen innerhalb von Zellen effizient unterbindet.
Zusammenfassend liefert diese Studie eine globale Annotation der Wechselwirkungen des naszierenden menschlichen Proteoms und liefert fundamentale Erkenntnisse über die Grundprinzipien dieses weit verbreiteten Assemblierungsweges von Proteinkomplexen. Unsere Ergebnisse werfen eine Reihe weiterer grundlegender mechanistischer Fragen auf, darunter die Frage, welche Auswirkungen die Co-Co-Assemblierung auf die

Polysomenstruktur hat, welche Folgen eine ineffiziente Co-Co-Assemblierung für das Proteom darstellt, und wie die Co-Co Assemblierung mit der co-translationalen Faltung und der Kinetik der Translation koordiniert ist.

# 1. INTRODUCTION

## 1.1. Co-translational folding of individual polypeptides

To support all cellular activities and react to changes in the environment, cells continuously renew and adapt their proteome, by producing thousands of functionally diverse proteins. The majority of newly-synthetized proteins must fold into defined three-dimensional structures to exert their biological function. This process is extremely challenging in the crowded cellular environment, where unspecific interactions constantly compete with the specific intra- and inter-molecular contacts required for native folding, posing risks of misfolding and aggregation. To limit the exposure of endangered folding intermediates in time and space, folding of individual polypeptides – i.e. the formation of secondary and tertiary structures – often starts at the polysome (co-translational folding), a concept that is well supported by both computational and experimental work.

### 1.1.1. Benefits of co-translational folding

Co-translational folding provides multiple inherent opportunities to facilitate and regulate efficient protein biogenesis. During synthesis, the polypeptide folds in a vectorial fashion (from N- to C-terminus) as it linearly emerges from the ribosome exit tunnel. This reduces the possible routes through which a protein can fold, favouring local against long-range interactions and productive conformational intermediates against kinetic traps (*1*). Protein synthesis occurs at an average rate of ~6 (eukaryotes) to ~20 (bacteria) amino acids per second, while folding of most proteins occurs in milliseconds. Therefore, translation is rate-limiting over folding for many proteins, allowing co-translational folding to occur at quasi-equilibrium (*2*).

For single domain proteins, the ribosome can induce early formation of folding intermediates in N-terminal parts of nascent chains. On the other hand, the ribosome can sometimes keep the nascent chain unfolded until all required residues have emerged; this is the case of folds that are dominated by long-range interactions which can only be established once the full sequence is synthetized and include examples of small globular proteins that are able to rapidly refold in a test tube.

Folding of multi-domain proteins is more challenging in solution and often results in misfolding and aggregation, highlighting the requirement of the ribosome folding environment (*3*). Indeed, sequential structuring of the nascent chain segments available at different times during translation enhances folding efficiency in a domain-wise fashion (*2*, *4*, *5*). This has likely enabled the evolution of multi-domain proteins, which are especially prevalent in eukaryotes (up to 75% of the proteome) (*6*).

Importantly, the ribosome itself additionally impacts co-translational folding, by providing a constrained environment for initial folding within the ribosome exit tunnel and by actively tuning folding rates to avoid establishing premature or incorrect contacts (*3*, *7*). In some cases, the folding intermediates formed on the ribosome play an independent functional role. An example is the bacterial SRP receptor FtsY whose N-terminus folds into a distinct intermediate which does not exist in the mature protein structure and is required for its targeting to the membrane (*8*).

In addition, the ribosome provides the main platform for the organization and coordination of a number of nascent chain interaction partners, including chaperones, modifying enzymes and targeting factors which all affect the folding and maturation of the nascent polypeptide (*7*).
The intimate coupling of folding with protein synthesis also allows fine-tuning of folding with local variations in translation speed, thereby enhancing both efficient compaction of emerged folding units and timely interaction with ribosome associated biogenesis factors (*9–12*).
Finally, co-translational folding occurs in the context of the polysome structure. Recent cryo-electron tomography studies of densely packed polysomes in prokaryotic (*13*) and eukaryotic systems (*14–17*) suggested that the arrangement of ribosomes on one mRNA may have evolved to maximise the distance of ribosome exit tunnels and hence avoid entanglement of neighbouring nascent chains during co-translational folding.
Collectively, these features allow cells to fold the nascent proteome with high efficiency, as indicated by the relatively low fraction of newly-synthetized proteins that fail to properly fold and are co-translationally ubiquitinated for subsequent degradation (~1–6% in yeast and 12-15% in human cells) (*18*, *19*).

### 1.1.2. Folding on the ribosome

During translation, the ribosome catalyses the polymerization of amino acids in the peptidyltransferase centre (PTC) and the growing peptide is gradually extruded through the ~100 Å-long exit tunnel which traverses the large ribosomal subunit (Fig. 1, left). In eukaryotic ribosomes, the tunnel is mainly composed of 28S rRNA and has a variable width (15 Å in average). Two ribosomal proteins (uL4 and uL22) protrude into the tunnel about 30 Å away from the PTC, generating a first constriction point which delimits the upper tunnel region (Fig. 1, right). An additional extension of uL4 generates a second constriction point within the central tunnel region. Finally, the tunnel becomes wider in the lower and especially the vestibule region, which is formed about 80 Å away from the PTC (Fig. 1, right).
The nascent chain residues included in the exit tunnel (~25 to 40 depending on the degree of nascent chain compaction) can undergo initial co-translational folding: α-helices can form

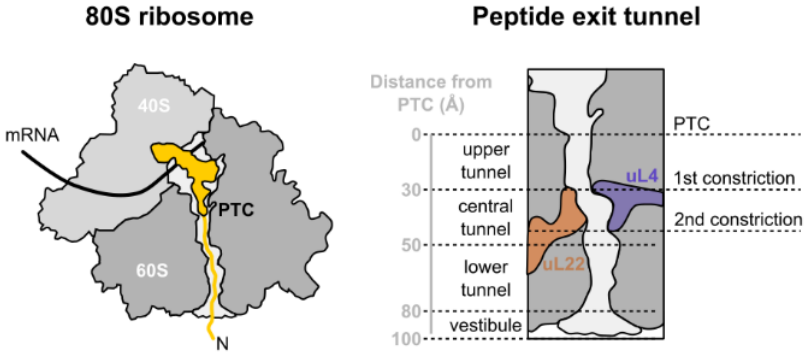



***Figure 1.*** *The peptide exit tunnel folding environment of eukaryotic ribosomes*

*(Left) Cross-section through the eukaryotic ribosome with schematic depictions of the peptidyl-tRNA (yellow). PTC = peptidyltransferase centre. Adapted from S. Bhushan et al, Mol. Cell (2010).*
*(Right) Enlargement of the peptide exit tunnel. Adapted from M. Liutkute et al, Biomolecules (2020).*

both in the upper and lower regions of the tunnel, but not at the constriction sites. Not all nascent peptides however start folding inside the tunnel: longer helices with higher helix propensity and hydrophobicity (such as transmembrane helices) are more likely to form (*20*). Minimalistic tertiary structures (including both α-helices and β-sheet topologies) can also form at the vestibule of eukaryotic ribosomes (*21*). In some cases, the co-translationally formed structures are the same as in the mature protein; in other cases, co-translational folds represent unique intermediates that are resolved at later stages.

These early folding events are influenced by the physical properties of the tunnel environment. First, the geometry of the tunnel restricts the folding space of the nascent chain. Some peptides can adopt an α-helical conformation in the upper tunnel even if they cannot form a stable helix in solution, suggesting an entropic stabilization role by the ribosome (*22*, *23*). Molecular dynamics simulations indicate that water molecules confined inside the ribosomal tunnel assume a semi-structured state, which in turn restricts nascent chain diffusion and favours compaction (*24*). In agreement with experimental data (*20*), this effect is expected to be stronger for hydrophobic nascent chains, as the solvent confined between the hydrophilic tunnel walls and the hydrophobic chain would generate repulsive forces resulting in nascent chain compaction. Second, the high net negative charge conferred by the conserved rRNA and ribosomal protein segments lining the tunnel also modulate co-translational folding. The electrostatic potential varies along the tunnel, with the lowest values coinciding with the first constriction site. Experiments employing ribosomal protein S6 as model nascent chain showed that the onset of co-translational folding is affected by the net charge of the nascent protein: positively charged proteins fold deeper in the exit tunnel, while negatively charged ones are rapidly extruded by electrostatic repulsion before folding can start (*25*). Not only folding, but also translation can be affected by the electrostatic interactions between the ribosome exit tunnel and the growing peptide: positively charged nascent chains can slow down or even stall translation by interacting with the interior of the tunnel. This occurs during translation of arrest peptides (AP), ~20 amino acids long sequences whose interactions with the exit tunnel induce a distortion of the PTC geometry required for elongation (*26*, *27*).

Ribosome proximity influences the folding landscape of a nascent polypeptide even after it has emerged from the ribosome exit tunnel. The ribosome surface can destabilize nascent chains to delay compaction until all residues required for correct folding of a domain are accessible. Examples of this holdase activity by the ribosome are nascent GFP and RFP, which are kept in a non-native folding intermediate until the last β-strand is emerged into the cytosol and can be integrated into the barrel structure (*28*). Electrostatic interactions between the ribosome and the nascent protein and the steric effect imposed by the large ribosome particle both likely contribute to this destabilizing effect. As the polypeptide emerges from the exit tunnel, it can substantially interact with the highly negatively charged ribosome surface. Such interactions can stabilize nascent chains (especially positively charged ones) into a low dynamic state, which may allow protection from unspecific interactions or co-translational degradation (*29*). Conversely, a higher net negative charge of nascent chains correlates with a larger fraction of highly dynamic states on the ribosome, which may allow the nascent peptide to sample a wider conformational space and avoid kinetic traps. Moreover, the folded state of short nascent chains is destabilized sterically by the ribosome (*30*). Accordingly, the rate of a domain native folding increases linearly with its distance from the exit tunnel as it increasingly behaves like a polypeptide in solution (*30*, *31*).

The ribosome activity in modulating the timing of co-translational folding is of particular importance for multi-domain proteins. The best understood example is the folding pathway of

bacterial elongation factor G (EF-G), a protein composed of five domains (termed G-domain, II, III, IV, and V). EF-G fails to fold efficiently in solution due to interactions among unfolded domains (*32*). The folding rate of the G domain is modulated by the ribosome based on the portion of polypeptide already synthetized: the maximum folding rate is achieved when the entire G domain is exposed out of the ribosome exit tunnel, while folding of longer or shorter nascent chain portions is slowed down on the ribosome compared to folding in solution (*32*, *33*). Hence, the ribosome decelerates folding when too little sequence information is available for correct formation of the G-domain or at longer nascent chains to avoid non-productive interactions with domain II. Hence, native folding of the two N-terminal domains (which constitute the first super-domain of EF-G) occurs in a co-translational manner. Conversely, EF-G shifts to a post-translational folding route for formation of its second super-domain, as domain III requires interactions with the C-terminal domains IV and V for stabilization. This example highlights the important role played by the ribosome in facilitating timely formation of folding units. Additional support to EF-G co-translational folding is provided by the coordination with the ribosome-associated chaperone trigger factor (*33*), another benefit offered by the ribosome folding environment.

### 1.1.3. Support by ribosome-associated chaperones

Ribosomes directly influence *de-novo* protein folding not only by imposing specific physical constrains on the nascent peptide, but also by coordinating the recruitment of ribosome-associated chaperones. Co-translational chaperone activity is particularly important to avoid misfolding of long proteins with complex topologies, often by delaying nascent chain compaction until all residues required for productive folding are available.

Among eukaryotes, the yeast ribosome-associated chaperone machinery is the best studied (Fig. 2, left). It includes the Hsp70 chaperone Ssb, the ribosome associated complex (RAC) and the nascent polypeptide associated complex (NAC) (*7*).

Recently, the Selective Ribosome Profiling technology provided detailed information on Ssb-nascent chain interactions on a proteome-wide scale (*12*). Ssb engages nascent chains close to the ribosome exit tunnel (*34*) and secures folding of the majority of the yeast nascent proteome, including proteins from all cellular compartments, with a preference for complex, aggregation-prone proteins. It typically binds via multiple binding-release cycles and recognizes nascent chain segments enriched with hydrophobic and aromatic residues, which typically become buried in protein folds, and also positively charged amino acids, which are more likely to be surface-exposed (*12*). By binding segments that will form the surface or the hydrophobic core of the complete protein, Ssb delays co-translational folding of nascent domains.

Ssb functions together with the co-chaperone RAC, composed of the non-canonical Hsp70 member Ssz and the Hsp40 Zuotin (Zuo1, Fig. 2, left). RAC binds the ribosome at the exit tunnel where it promotes substrate binding by positioning Ssb in an optimal orientation for nascent chain engagement and by stimulating ATP hydrolysis (*35*, *36*). In the absence of RAC, the timing and specificity of co-translational Ssb engagement to nascent chains are altered (*12*). Ribosome Profiling further revealed that Ssb-RAC binding correlates with a speedup of translation, mainly caused by intrinsic features of the mRNA and nascent chain. Ssb-mediated delay of co-translational folding locally uncouples folding from translation kinetics and may have allowed evolution of faster elongation during Ssb engagement (*12*). Interestingly, Zuo1 additionally interacts with regions of the conserved decoding centre in the

small ribosomal subunit, suggesting a possible role in coupling translation fidelity and kinetics with co-translational folding by Ssb.

The human mRAC complex is formed by the Zuo1 homologue MPP11 and the Ssz homologue Hsp70L1 (Fig. 2, right). Differently from yeast however, mammalian cells lack a ribosome-associated Hsp70; instead mRAC recruits the cytosolic Hsp70A1A/B to ribosomes for co-translational folding.

The second co-translationally acting chaperone, conserved in eukaryotes from yeast to humans, is NAC, a heterodimeric complex composed of α- and β-subunits (*7*). Inside cells, NAC is present at stoichiometric levels to ribosomes and binds to both translating and idle ribosomes. Like Ssb, NAC sits on the surface of translating ribosomes, close to the polypeptide exit site and engages a wide range of nascent proteins from all cellular compartments at early translation stages. Therefore, it was proposed that NAC has a chaperone function. Accordingly, NAC deletion causes enhanced co-translational ubiquitination, especially of longer proteins with high aggregation propensity, suggesting that NAC normally secures folding of these nascent chains at the ribosome (*18*). Supporting NAC's chaperone function, deletion of both NAC and Ssb in yeast causes massive protein aggregation, especially of ribosomal proteins and ribosome biogenesis factors, suggesting that both proteins play a dual role in supporting folding and regulating ribosome production (*37*). In addition, NAC was shown to re-locate from ribosomes to cytosolic aggregates upon proteostasis imbalance (including heat stress, ageing and expression of the aggregation-prone Aβ peptide), which in turn results in reduced translation activity (*38*). Together, these findings implicate NAC as a proteostasis sensor which balances translation levels with the chaperone folding capacity.

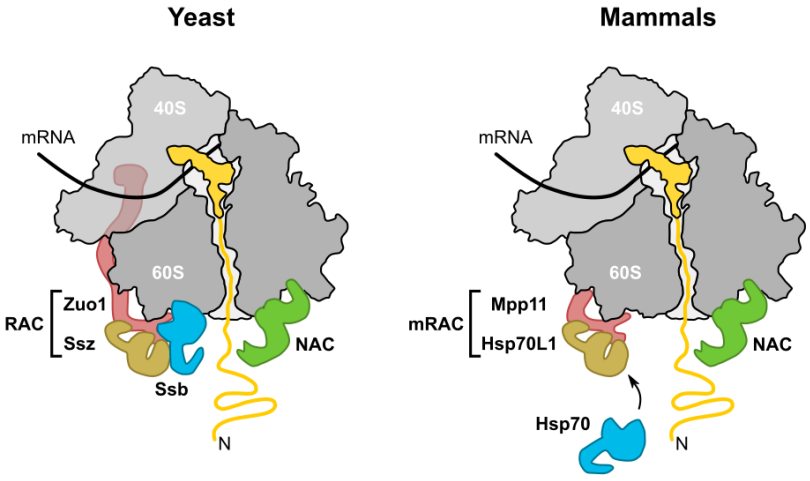



***Figure 2.*** *Eukaryotic ribosome-associated chaperones*

*Adapted from G. Kramer et al, Annu. Rev. Biochem. (2019).*

NAC is also involved in the regulation of co-translational protein sorting to different subcellular compartments (*39*). Its binding position on the ribosome surface interferes with SRP binding, preventing mistargeting of nascent chains that lack a signal sequence to the endoplasmic reticulum (ER). Moreover, NAC positively regulates co-translational targeting of mitochondrial precursors translated by cytosolic ribosomes.

In conclusion, ribosome-associated chaperones transiently engage the nascent polypeptide in coordination with other chaperones, maturation factors and with variations in translation speed to ensure efficient biogenesis of natively folded proteins. Folding of a more restricted

spectrum of proteins is further supported by post-translationally acting chaperones that do not interact directly with ribosomes, including the classical Hsp70/Hsp40 system, the TRiC/CCT chaperonin and the Hsp90 system.

### 1.1.4. Interdependence between co-translational folding and translation kinetics

Co-translational folding is inherently linked to the kinetics of protein synthesis through the ribosome. The first aspect of this mutual relationship is that translation kinetics modulates folding of nascent proteins. Indeed, translation elongation, which proceeds at an average speed of 4-6 amino acids / second in eukaryotes, imposes an upper limit on the rate of co-translational folding. Furthermore, the rate of translation elongation is not uniform along open reading frames (ORFs) and these variations, which can span over an order of magnitude, can have a profound impact on co-translational folding (*40*). Identical protein sequences encoded by different nucleotide sequences through the use of synonymous codons can result in modified outcomes, such as misfolding or altered conformation or stability of the protein (*10*, *41*, *42*). Indeed, the codon optimality (which scales with the abundance of cognate tRNAs), codon context and mRNA secondary structures all have the potential to affect the optimal pattern of local translation speed. While the negative impact of altered translation kinetics on protein biogenesis was demonstrated for a number of substrates, these experiments often required alteration of entire codon clusters, suggesting that folding is resilient to a certain amount of variation (*43*). Indeed, some proteins can be successfully expressed in heterologous systems despite the differences in organism-specific tRNA abundance. On the other hand, harmonization of codon signatures in the expression host can often enhance efficiency of protein production.

Stretches of rare codons, which are generally associated with translation slowdowns, are found at specific positions along coding sequences, in particular at subdomain boundaries, separating folding of small structural modules (*44*, *45*). These patterns of codon optimality are conserved across prokaryotic and eukaryotic genomes, indicating evolutionary pressure to optimize the timing of protein synthesis in favour of co-translational folding (*44–46*). Hence, codon degeneracy is increasingly recognized as a second genetic code defining protein structure.

A recent evolution of the Ribosome Profiling technology allows to identify sites of ribosome collisions caused by translation pausing across the transcriptome by sequencing 60 nucleotide-long RNA footprints that are protected by collided ribosome pairs (Disome Profiling) (*47–50*). Using this approach on mouse liver, Arpat and colleagues identified a distinct class of pause sites correlating with structural features of the final protein (*50*). Stalling of the leading ribosome occurred most frequently during translation of unstructured protein segments, which were immediately preceded and followed by structured regions (including α-helices and β-sheets). Stalling sites were particularly enriched towards the 5’ portion of the sequence encoding for the unstructured protein region, close to the 3’ boundary of the upstream sequence encoding for the structured parts. These findings further support the concept that ribosome pausing is important to kinetically separate folding of consecutive (sub)domains.

On the other hand, slower translation can also negatively impact translation fidelity. For example, slow or inefficient decoding of rare codons can cause ribosome frameshifting (*51*) or incorporation of wrong amino acids which may lead to misfolding (*52*). In all organisms,

from bacteria to humans, optimal codons, which are generally translated with higher accuracy, are associated with conserved protein regions, suggesting evolutionary pressure to ensure translation accuracy at sites where it is most important for proper folding (*52*). The resulting trade-off between slow and fast translation finely balances the local needs of extended folding time of protein domains and translation fidelity.

Transient ribosome slowdown can also indirectly affect folding and maturation of the nascent protein by facilitating engagement of ribosome-associated factors required to support folding, enzymatic modification and, notably, membrane targeting of the nascent peptide (*40*).

The second emerging aspect of the mutual relationship between folding and translation kinetics is that folding near the ribosome exit tunnel exerts a mechanical force on the nascent chain which is transmitted to the ribosome catalytic core and can affect translation elongation. The most rigorously studied example is stalling during translation of the arrest peptide (AP) of bacterial SecM (*53*) which can be released by external forces exerted on the nascent chain (*54*, *55*). The nascent AP of SecM interacts with the interior of the ribosomal exit tunnel, causing allosteric repression of the PTC and hindering incorporation of a proline residue (*53*). *In vivo*, stalling regulates translation of the downstream SecA open reading frame (which is part of the same polycistronic mRNA) and is rescued by interactions of nascent SecM with the translocon. In artificial experimental setups, a speedup of translation elongation during SecM translation was linked to different co-translational events, including subdomain folding and membrane insertion, making the AP of SecM a force sensor, generally employed to study co-translational processes (*54–58*). A pulling force on the nascent chain can also enhance the efficiency of programmed ribosomal frameshifting, as recently shown for the Sindbis virus structural polyprotein (*59*). The translocon-mediated insertion of a transmembrane domain in this protein generates a tension that tunes the frequency of a frameshifting event, which in turn allows generation of a secondary form of the polyprotein. Arrest peptides have been found in the proteomes of many living organisms (including a number of bacterial species, yeast, plants and humans) and of different viruses. How widespread these sequences really are within coding sequences and to what extent they are employed as general modulators of translation speed with different biological outcomes is poorly understood (*27*).

In summary, protein synthesis and folding are synchronized processes which can influence each other in several ways. The intriguing concept is emerging that this intimate relationship constitutes a feedback mechanism by which a protein can regulate its own synthesis based on co-translational structure acquisition (*54*).

## 1.2. Co-translational protein complex assembly

A large fraction of both prokaryotic and eukaryotic proteomes form long-lived protein complexes composed of identical (homomers) or different subunits (heteromers). Similar to folding of individual polypeptides, complex assembly is a demanding task in the crowded cytoplasm: since isolated subunits expose unpaired protein-protein interaction surfaces which are often hydrophobic, they are highly prone to non-productive interactions with other proteins or components of the protein quality control machinery, which may eventually lead to aggregation or degradation. To minimize the time in which the newly-synthetized subunit is accessible for unspecific interactions, the translating machinery can additionally coordinate the ultimate folding step, namely formation of the native quaternary structure (co-translational assembly) (*60*).

### 1.2.1. Emerging features of translation-coupled assembly

Remarkably, the first indications of co-translational assembly date back to the 1960s, when researchers found that homo-tetrameric β-galactosidase was enzymatically active while still bound to polysomes, indicating formation of natively folded and assembled enzymes (*61*). The concept received additional support during the 1980s, when some cytoskeletal components were found to directly associate with the cytoskeleton structure during translation (*62*). A number of studies followed, proposing, for individual protein complexes, different mechanisms of co-translational assembly (see chapter 1.2.2). Recent evidence now suggests that co-translational assembly may be a widespread phenomenon both in prokaryotic and eukaryotic cells (*60*, *63*–*66*). Given the well-established prevalence of co-translational folding of secondary and tertiary structures, the pre-requisite for assembly largely exists already on ribosomes, hence one may argue that co-translational assembly should be allowed for a similarly large number of proteins.

The features and mechanistic details of co-translational assembly have just started to emerge and, unsurprisingly, they partially overlap with the ones already known for co-translational folding of single polypeptides. Early, translation-coupled assembly can counteract the nascent protein's aggregation propensity (*66*), it provides a solution for proteins that are toxic or unstable in isolation (*67*, *68*), it ensures precisely ordered assembly pathways (*60*) and allows coordination with ribosome-associated chaperones that can shield the nascent domain to avoid premature assembly interactions (*63*, *66*). In addition, the ribosome itself may directly regulate assembly. An example is that of two intrinsically disordered proteins of opposite charge, ACTR and NCBD, that fold into a high affinity complex upon association (*69*). Completely synthetized NCBD engages nascent ACTR, but not vice versa. This unidirectionality of co-translational assembly (which is consistent with most described examples so far (*60*, *63*, *66*, *70*)), is imposed by electrostatic interactions between the nascent protein and the ribosome: nascent positively charged NCBD strongly interacts with the ribosome surface and hence is inaccessible to its binding partner; conversely, repulsion of the negatively charged nascent ACTR from the ribosome surface favours its interaction with diffusing NCBD.

Finally, similarly to folding, co-translational assembly may benefit from the fine tuning of local translation speed to create the optimal time-window for assembly (*71*, *72*), although experimental evidence for this mechanism is still scarce. Contrary to folding of individual polypeptides however, the kinetics of co-translational assembly are not only governed by the proteins' intrinsic features (i.e. affinity for the partner protein) and by translation rate, but are

also highly dependent on the concentration of interacting subunits around the polysome at specific time-points during translation (*73*). Cells regulate this important parameter through different strategies which fundamentally depend on the particular mechanism employed for co-translational assembly.

### 1.2.2. Mechanisms mediating co-translational complex assembly

#### 1.2.2.1. Assembly involving one nascent chain (co-post assembly)

Most available evidence on co-translational assembly describes the interaction of a fully-synthetized, diffusing protein with its nascent, polysome-bound partner subunit (co-post assembly, Fig. 3, left). A co-post assembly mechanism can mediate formation of both homomeric and heteromeric complexes via interactions *in cis* (if the nascent and diffusing subunits are both synthetized from the same mRNA) or *in trans* (if the two subunits are translated from separate mRNAs). A *cis* assembly mode of heteromeric complexes however requires the organization of subunits-encoding genes in operons, a feature that characterizes bacterial proteomes but was largely abandoned by eukaryotes. Shieh and co-workers recently demonstrated that interactions between the operon-encoded subunits of heteromeric bacterial luciferase LuxAB initiate co-translationally (*63*). Assembly involves the fully-synthetized product of the upstream gene (*LuxA*) interacting with the nascent chain of the downstream gene (*LuxB*) and initiates upon ribosome exposure of the complete LuxB interaction domain, suggesting that folding is a prerequisite for assembly. The ribosome-associated chaperone trigger factor (TF) delays the onset of co-translational assembly until the complete LuxB interface has emerged into the cytosol, thereby avoiding premature interactions. Importantly, efficient assembly requires that both subunits are synthesized from a single bicistronic mRNA (*63*). In prokaryotes, the subunits of a protein complex are often encoded within the same operon, allowing their translation from a single mRNA molecule, and gene order is optimized to match that of complex assembly (*64*). Thus, the evolution of bacterial gene organization into operons may reflect a general translation-coupled mechanism of complex assembly (a notion supported by additional unpublished work from the Bukau laboratory).

In eukaryotes, assembly of several heteromeric complexes was shown to start co-translationally, suggesting that co-post assembly *in trans* may be a widespread mechanism for complex formation. In a first systematic study, the authors employed RIp-chip (Ribonucleoprotein Immunoprecipitation analysed with DNA chips) to identify the mRNAs that interact with 31 protein subunits of *S. pombe* lacking RNA-binding domains (*70*). They found that 38% co-purify with the mRNA of the interacting complex subunits. This phenomenon is dependent on the presence of the encoded nascent chain and on the integrity of polysomes. A second systematic study employed the Selective Ribosome Profiling technology (SeRP), which allows to identify the sub-fraction of translating ribosomes that interacts with a subunit of choice for co-translational assembly and provides codon-resolved information on the timing of nascent chain engagement during translation (*66*). The authors provided direct evidence for co-translational assembly of 9 out of 12 pre-selected protein complexes in *S. cerevisiae*, with the remaining three complexes employing dedicated assembly chaperones. Co-translational interactions of the studied complexes are mostly unidirectional and rely on full exposure, and presumably folding, of the nascent subunits' dimerization domains. Similar to bacteria, the ribosome-associated chaperone Ssb shields

the nascent interaction domain until it is completely ribosome-exposed and can be handed over to the interaction partner (*66*).
Additional studies described the co-post assembly of a number of hetero-oligomeric complexes (*70*, *74*, *75*). A recent review highlighted that several of these complexes share a common structural feature, namely an extended conformation of the nascent subunit which allows to sequentially recruit multiple co-translationally interacting partners as their interaction sites become progressively available on the nascent chain (*60*). As mentioned above, co-post assembly mostly occurs in a unidirectional fashion, suggesting that the identity of the nascent subunit is important to build the complex. Thus, the nascent protein of co-post assembling complexes may function as coordinating subunit, which specifies the correct order of assembly through its own primary sequence.

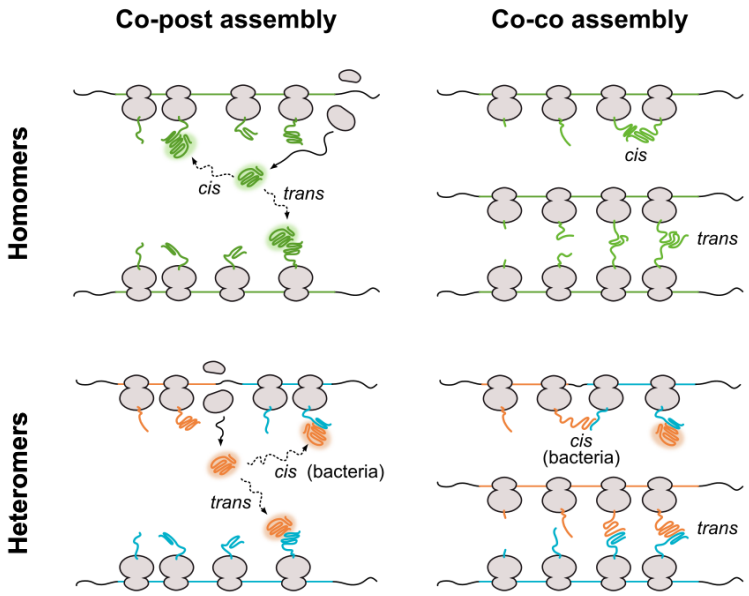



***Figure 3.*** *Co-translational complex formation involves at least one nascent subunit*

*Possible mechanisms mediating co-translational assembly are schematized.*
*Co-post assembly involves one fully-synthetized and one nascent subunit (left); co-co assembly involves two nascent subunits (right). Polysomes translating homomeric (top) or heteromeric (bottom) subunits are shown; open reading frames (ORFs) on mRNAs are coloured based on the encoded proteins; fully-synthetized proteins are highlighted by shadowed circles in the background; for co-post assembly, diffusion of a fully-synthetized protein towards its nascent partner subunit (in cis or in trans) is indicated by dashed arrows.*

#### 1.2.2.2. Assembly involving multiple nascent chains (co-co assembly)

In theory, co-translational assembly could also involve two (or more) nascent subunits (co-co assembly, Fig. 3, right), including interactions of subunits that are concurrently translated from the same (*cis*) or different mRNAs (*trans*). In bacteria, assembly of not only homomeric but also heteromeric complexes could theoretically occur in *cis*, if the interacting subunits are translated from polycistronic mRNAs, although to date no evidence for such mechanism exists. In eukaryotes on the other hand, heteromer formation by co-co assembly relies solely on interactions *in trans* (Fig. 3, right).

Due to the lack of methods to directly and selectively identify interactions between two nascent proteins, mechanistic and systematic studies on co-co assembly are still missing. Therefore, current knowledge relies on sporadic examples and does not allow to draw any conclusion on the prevalence of this mechanism relative to co-post assembly.

Most available evidence suggests co-co assembly of homomeric protein complexes via interactions *in cis*. The first described examples are the trimeric Reovirus cell attachment protein σ1 (*76*) and the hexameric human tenascin complex (*77*). The monomers of both complexes feature long N-terminal heptad repeats which associate in triple coiled coils; two tenascin trimers are further joined at a central knob forming the hexameric complex. In the first study, the authors found σ1 trimers associated to polysome of *in vitro* translation reactions; they further observed that co-translation of two σ1 variants results preferentially in homo-trimers and this bias is more pronounced at lower transcript concentrations, suggesting that assembly of σ1 trimers is transcript-templated. The second study mostly relies on the rather indirect evidence that tenascin hexamers are formed rapidly inside human cells, with no assembly intermediates detectable by pulse-chase experiments. The authors of both studies come to the same conclusion, that interactions of N-terminal domains emerging from the same polysome mediate simultaneous assembly of three or six nascent subunits (co-co assembly *in cis*).

A more recent study suggested that also assembly of a much larger complex, the vault particle, is orchestrated by a single polysome (*78*). The authors proposed a model whereby neighbouring nascent chains dimerize during translation and adjacent dimers are progressively stacked at the 3' end of the polysome, until 78 vault subunits are joined into a complete particle. This model is based on a set of observations similar to the ones reported for the σ1 and tenascin complexes, including: (i) efficient assembly in diluted solutions, (ii) absence of (detectable) assembly intermediates, (iii) visualization of vault particles in proximity of polysomes by electron microscopy (although only at low frequency) and (iv) segregation of two protein variants bearing different N-terminal tags into separate particles. It should be noted however that one of the variants employed in the latter experiment assembles into unstable vault particles, suggesting an aberrant conformation of this protein conferred by the N-terminal tag, which may not allow mixing with the wild type counterpart in solution.

Notably, formation of p53 and NF-κB1 homodimers by co-co assembly *in cis* were proposed to have important functional implications (*67*, *79*). Biochemical experiments analysing the final products of *in vitro* translation reactions suggested that p53 dimers form via interaction of neighbouring nascent subunits on one polysome, while tetramers are generated from dimers in a post-translational fashion (*79*). The mechanism and timing by which p53 tetramers are built become particularly relevant in the context of cells expressing a wild type and a mutant allele. Indeed, only 1/16 of total p53 would form wild type tetramers through a completely post-translational assembly pathway, while the proposed co-translational dimerization followed by post-translational dimerization of dimers would result in 1/4 wild type tetramers. Therefore, co-co assembly *in cis* may have partially evolved to counteract the deleterious effects of dominant negative mutations.

A series of experiments analysing the products of both *in vitro* and cellular translation indicated that the nascent subunits of NF-κB1 assemble on the polysome (*67*). The full length p105 product of *NFKB1* is composed of an N-terminal RHD (Rel Homology Domain), which mediates dimerization, and C-terminal ankyrin repeats which are a hallmark of IκB inhibitors. A shorter protein, p50, comprising only the N-terminal RHD is generated by proteasomal cleavage during translation (*80*). Thus, in this particular case, assembly *in cis* of

nascent chains ensures rapid association of the NF-κB transcription factor (p50) with its IκB inhibitor (p105), thereby avoiding undesired activation of the inflammatory pathway during NF-κB1 biogenesis. Furthermore, co-translational dimerization of the nascent RHDs appears to shield the interaction domain from the proteasome, and hence is crucial for preventing complete degradation of the p50 subunit (*67*).

The potential phenomenon of co-co assembly *in trans* has received comparably less attention, with the exception of three recent studies which proposed co-co assembly of individual heterodimeric complexes: the human voltage-gated K channel hERG 1a/1b (*81*), proteasome subunits RPT1 and RPT2 (*71*) and the TFIID transcription factor subunits TAF6 and TAF9 (*75*). These studies support the intriguing possibility that the biogenesis of two different polypeptides can be physically linked, a process that would require a remarkably high level of spatiotemporal coordination of translation, folding and assembly of the two subunits. Interestingly, assembly of the proteasomal subunits RPT1 and RPT2 was proposed to be facilitated by ribosome stalling on both mRNAs, coinciding with ribosome-exposure of the N-terminal helices involved in dimerization (*71*). On the other hand, the evidence in support of the *trans* co-co assembly mechanism is so far limited to indirect observations, including the co-purification of each monomer with its partner subunit's mRNA and the finding that the subunits encoding mRNAs are physically linked or co-localized inside cells (*71*, *75*, *81*). These findings are fully compatible with a co-post assembly mechanism that is bi-directional. Such mechanism was identified for three protein complexes in yeast using Selective Ribosome Profiling with C-terminally-tagged subunits (*66*) and these proteins did not co-co assemble according to our unpublished proteome-wide data on co-co assembly in yeast. Therefore, clear evidence for a co-co assembly mechanism *in trans* is still missing.

In conclusion, scattered evidence indicates that co-co assembly is employed to assemble protein complexes in eukaryotic cells with different proposed functional implications for the cellular physiology. However, we are still missing information on its existence and prevalence inside cells and we lack any knowledge on the regulatory mechanisms of this process.

### 1.2.3. Translocation-coupled assembly

In eukaryotes, most membrane, luminal and secreted proteins are co-translationally translocated into or across the ER membrane. Transmembrane proteins often fold co-translationally: the transmembrane domains (TMDs) of multi-spanning membrane proteins sequentially partition from the translocon into the lipid bilayer and rapidly interact with each other in order to stabilize the polar residues which mediate the tertiary fold of the protein (*55*, *82*).

This concept can be extended to the formation of quaternary structure, which is often driven by less hydrophobic TMDs that need to interact with partner TMDs to be efficiently inserted into the membrane. Indeed, assembly of transmembrane proteins can occur simultaneously to co-translational translocation, as described for homomeric (*83*) as well as heteromeric protein complexes (*72*, *81*, *84*). Furthermore, translocated proteins could associate via interactions of the proteins N-termini after they have emerged into the ER lumen, as proposed for the secreted tenascin complex (*77*).

The ER membrane itself can facilitate the co-translational association of translocated proteins: (i) it can regulate the local concentration of the interacting domains (*83*), (ii) it can facilitate the co-localization of the interacting subunits-encoding mRNAs (*81*) and (iii) it provides a common scaffold that forces the same orientation of translocated nascent subunits, thereby facilitating their interaction (*77*).

### 1.2.4. Evolutionary aspects of co-translational complex assembly

For some proteins, co-translational assembly may not be possible, as exemplified by organellar subunits whose interaction partner is translated in a separate compartment. In other cases, diffusion-driven post-translational assembly may be sufficient, as evidenced by a number of proteins that can assemble into native structures *in vitro*. Cells have evolved multiple strategies to ensure efficient complex assembly, including the tuning of subunits expression to match complex stoichiometries (*85*), the employment of dedicated assembly chaperones (*86*) and the spatial confinement of subunit translation to distinct subcellular localizations (*87–89*) or even to dedicated organelles (e.g. ribosome biogenesis in the nucleolus). Co-localized translation facilitates association of interacting subunits shortly after they have been released by ribosomes (termed peri-translational assembly) as well as in a co-translational fashion. Both in bacteria and eukaryotes, co-translational assembly is emerging as a widespread phenomenon (*63*, *66*, *70*). Indeed, several functional benefits have been described which may explain the selective pressure to couple assembly with translation. These include an increased efficiency of complex biogenesis (*63*), the protection of aggregation-prone nascent subunits from misfolding or degradation (*66*, *74*, *75*, *90*), the possibility of rapidly assembling proteins with their regulators (*65*, *67*), the coordination with additional co-translational maturation steps required for biogenesis of the protein (*67*, *80*) and the reduced impact of dominant negative mutations (*79*). Yet, the existence of alternative assembly strategies indicates that co-translational assembly is not always the preferred route and in some cases there may be evolutionary pressure to avoid it (*91*). For example, a possible risk of coordinating translation on one polysome with assembly of nascent chains *in cis* is that the high local concentration of partially folded nascent chains increases the likelihood of premature or unspecific interactions and eventually of misfolding. Recent work suggested that the polysome structure may have evolved to optimize co-translational folding by preventing unwanted interactions between neighbouring nascent chains (*13–17*). The organization of densely packed polysomes was observed by electron tomograms in intact human cells, revealing they can acquire conformations such as a three-dimensional helix, a less compacted spiral, or a planar zig-zag topology (*14*). In all cases, the small ribosomal subunits face the interior of the polysome structure, while the large subunits are oriented towards the cytosol. This general configuration maximises the distance of nascent subunits emerging from the exit tunnels and may have evolved to counteract the propensity for nascent chain entanglement *in cis*.

A recent study indicated that this may be particularly important for proteins with N-terminal homo-dimerization domains, as the prolonged exposure of the homomer interfaces during translation increases the likelihood for assembly *in cis* to occur before the C-terminal parts of the protein are properly folded. This in turn would force the partially or completely unfolded segments of nascent proteins in close proximity, increasing the chances of misfolding (*73*). The authors showed that, in fact, the inter-subunit interfaces of homomeric but not heteromeric complexes are evolutionarily selected to be located in the C-terminal halves of proteins in organisms from bacteria to humans. In order to identify sequence-specific features that affect the outcome of assembly, the authors further monitored the aggregation propensity of a library of artificial constructs containing the p53 homo-oligomerization domain (tet), a linker of varying length and a fluorescence reporter. Their results showed that successful assembly is facilitated by (i) the C-terminal positioning of the oligomerization domain, (ii) increased linker length separating the oligomerization domain from the reporter and (iii) increased folding rate of the C-terminal fluorescence reporter. The authors concluded that co-translational assembly between neighbouring nascent chains (induced by

the N-terminal positioning of the tet oligomerization domain) is often detrimental (i), unless a long enough linker allows to decrease the local concentration of the partially unfolded C-terminal parts of the nascent protein (ii), or folding of these C-terminal parts is fast enough to occur before assembly (iii).
Although for most proteins interactions between nascent chain N-termini *in cis* was likely selected against during evolution, a large number of homomeric subunits did maintain an N-terminal interaction interface. How native folding and assembly of these subunits is achieved and what polysome structure is formed during their translation is currently unknown.

## 1.3. Spatial organization of complex assembly

The importance of regulating translation in space is evidenced, in bacteria, by the widespread organization of functionally related genes in polycistronic mRNAs.
It is reasonable to assume that the spatial control of protein production is even more important for eukaryotic cells, whose volume is several orders of magnitude larger than that of bacterial cells. The proposed benefits of locally confined translation are manifold. First, protein production directly on the site where they carry out their specific task is highly energetically favourable compared to the transport of the several protein copies that are generated from a single mRNA. Moreover, rapid responses to environmental stimuli can be more easily achieved if proteins are synthetized on-site, instead of being transported from a distal cellular location when they are needed. This is particularly relevant for neurons, due to the rapid nature of synaptic transmission and the extreme distal locations of axonal or dendritic compartments where protein activity is required (*92*). Second, the confinement of mRNAs to specific cellular regions can serve to prevent the toxic effect of the protein products in other subcellular compartments. Third, multiple mRNAs can be co-localized inside cells in so-called “local transcriptomes” allowing for nearby synthesis of functionally related proteins (*93*). This can be achieved through the formation of messenger ribonucleoproteins (mRNPs) composed of RNA-binding proteins and a specific subset of mRNAs whose localization, translation and degradation can be co-regulated. It was shown for example, that mRNAs encoding for enzymes of the glycolytic pathway are co-localized and actively translated within granules in the yeast cytosol and that, upon stress, these granules recruit the mRNA decay machinery, suggesting concerted regulation (*94*). Similarly, mathematical modelling studies predict that co-localized translation of complex subunits would substantially enhance the efficiency of complex formation (*95*), by increasing the local concentration of interacting subunits and excluding aberrant interactors. Supporting this prediction, the subunits-encoding transcripts of a number of protein complexes have been shown to display a similar pattern of subcellular localization, including the co-localization of all seven subunits of the Arp2/3 actin-polymerization nucleator complex (*87*), the Ccr4–Not complex-mediated tethering of two SAGA histone acetyltransferase transcripts (*90*) and of mRNAs encoding for proteasomal subunits RPT1 and RPT2 (*71*), the localized translation and assembly of several cytoskeleton components (*62*, *96*), nuclear pore complex subunits (*97*) and others (*75*, *81*, *98*, *99*). For some of these examples, it is unknown whether assembly occurs via peri-translational or co-translational interactions, for others, the nearby synthesis of interacting subunits is known to facilitate one of the co-translational assembly mechanisms involving interactions *in trans* (Fig. 3). Given the high abundance of redundant dimerization domains and of structurally related subunits which bear a high potential for promiscuous interactions, formation of translation hotspots of protein complexes in combination with co-translational assembly may allow eukaryotic cells to control complex composition in an energy-efficient manner.

### 1.3.1. Mechanisms mediating localized translation

A classic mechanism allowing the partitioning of a selected population of mRNAs for translation is the targeting to membrane organelles. In addition to its established function for the biogenesis of translocated proteins, the ER membrane organizes translation of various cytosolic and nuclear proteins, suggesting a global role of the ER in regulating transcriptome localization and expression (*100*). Ribosomes synthesizing cytosolic proteins can associate

with the ER in a translocon-independent manner, allowing to confine their translation to a separate compartment. Moreover, proximity labelling approaches revealed a non-uniform distribution of ER-localized mRNAs: cohorts of transcripts encoding for interacting proteins are enriched at distinct neighbourhoods within the ER (termed translation centres), presumably to facilitate local translation and assembly of protein complexes (*101*). This notion was recently supported by the discovery of TIGER domains, ER-associated granules which enable the selective recruitment certain transcripts thereby enabling protein-protein interactions which would not occur outside of this compartment (*102*).

Additional mechanisms allow cells to further compartmentalize translation within the cytosolic environment. In many of the known examples, the information required for the spatial organization of transcripts is encoded in the untranslated regions (UTRs) of mRNAs in the form of linear sequences or secondary structures (so-called “zip-codes”) (*103*). These targeting elements are recognized by specific RNA-binding proteins (RBPs) that direct the transcripts targeting. In this mechanism, the resulting messenger ribonucleoproteins (mRNPs) are actively transported by motor proteins along the cytoskeleton structure. Multiple mRNAs can be transported at once and, likewise, a single mRNA can interact with multiple RBPs at the same time, creating a heterogeneous constellation of mRNPs whose composition can influence the speed and direction of transport.

To achieve a stable localization of transcripts, active transport is followed by anchoring of the mRNPs to actin or microtubules, a process usually mediated by proteins that simultaneously bind the mRNAs and the cytoskeleton component (*103*).

Although less frequent, selective mRNA degradation or protection can also mediate the formation of local transcriptomes, a mechanism that was observed in Drosophila embryos (*103*, *104*). In this mechanism, an mRNA is unstable or actively degraded unless it is found in a specific cellular location. All these mechanisms of RNA trafficking may also act in combination to modulate the final localization of a single transcript.

For the purpose of protein complex biogenesis, the spatial organization of translation likely plays an important role to facilitate those assembly mechanisms that rely on interactions *in trans* (Fig. 3). Importantly, a high local concentration of interacting subunits can be achieved independent of any mechanism of RNA subcellular localization by assembly *in cis*, as the polysome itself provides the platform for localized translation and assembly of complex subunits.

### 1.3.2. The RNA operon hypothesis

The observation that functionally related transcripts can be physically linked in the cytosol through RNA binding proteins raised the intriguing idea that mRNPs are surrogates of bacterial operons, allowing synchronized and co-localized translation and regulation of the included transcripts (*93*, *94*, *105*). Compared to bacterial operons however, the so-called “post-transcriptional RNA operons” allow a higher flexibility of regulation, thanks to the combinatorial nature of mRNPs and the possibility to dynamically change their composition in response to environmental stimuli.

Interestingly, some RBPs associate to transcripts already during transcription and remain associated while transcripts are processed and exported to the cytoplasm. For example, RIp-chip analyses revealed that the splicing factors U2AF and PTB interact both with unspliced and mature mRNAs and that, surprisingly, a distinct subset of functionally-related mRNAs is targeted by each factor (*106*). This finding may explain why these RBPs shuttle between the

nucleus and the cytoplasm despite their canonical splicing function is limited to the nucleus, suggesting that they interact with a subset of newly-transcribed mRNAs and potentially drive their coordinated export and localization in the cytosol (*93*). An independent RIp-chip study found that also the yeast RNA nuclear-export proteins YRA1 and MEX67 associate with specific subsets of newly-transcribed mRNAs encoding functionally-related proteins and mediate their export to the cytoplasm (*107*). These nuclear-cytoplasmic shuttling proteins provide a physical link between the transcription and translation processes, possibly allowing the coordination of gene expression at both levels similar to the direct coupling existing in bacteria.

The concept of post-transcriptional RNA operons is supported and further extended by a new study, which suggests that the functional regulation of gene expression starts prior to transcription, at the level of nuclear organization of genes (*108*). Employing the single-cell RNA-sequencing technology, the authors identified gene pairs whose expression co-varies across individual mouse embryonic stem cells. They found that genes encoding for proteins with related functions are more likely to co-vary. Importantly however, this effect is only present for proteins that associate in heteromeric complexes (including about 3500 co-varying gene pairs encoding for interacting proteins), while it is absent for proteins that function in the same pathway but do not physically interact, implying that co-variations are most important to secure correct complex stoichiometries. They further show that transcription of co-varying genes is often regulated by the same transcription factors or miRNAs and, most importantly, it is strongly affected by nuclear gene proximity. Specifically, co-variation is most strongly linked to the linear vicinity of genes belonging to the same chromosome, and, although to lesser extent, to the proximity of genes residing on different chromosomes. Importantly however, the latter effect may be higher than reported, as the method employed to measure gene proximity (Hi-C) is known to under-estimate inter-chromosomal contacts.

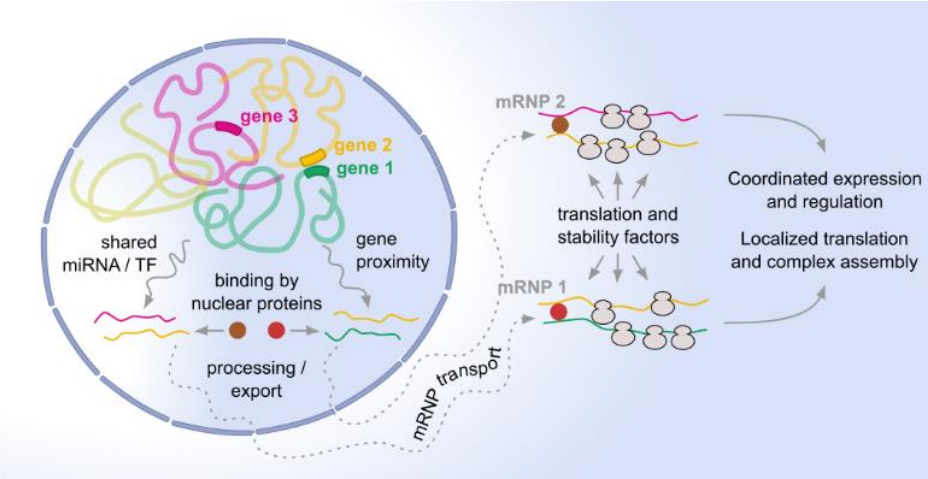



***Figure 4.*** *Coordination of eukaryotic gene expression from transcription to translation*

*Transcription of different sets of genes can be co-activated inside the nucleus by a shared miRNA or transcription factor (TF) and / or by their nuclear proximity. Nuclear proteins (for example splicing or export factors) recognize the co-expressed transcripts and remain bound during processing and export to the cytosol. Each mRNP is transported to a specific localization, where the transcripts' translation and turnover can be coordinated. Association of transcripts encoding complex subunits and their localized translation can facilitate complex assembly through peri- or co-translational interactions.*

Together, these findings support the emerging idea that eukaryotic expression of complex subunits is synchronized and co-regulated starting at the level of nuclear gene organization, all the way to subcellular targeting, translation and assembly (Fig. 4) (*93*).

# 2. AIMS OF THIS STUDY

The overall objective of this Thesis was to provide a comprehensive and mechanistic investigation of a postulated process, termed co-co assembly, where nascent chains emerging from two adjacent ribosomes fold and assemble during their concurrent translation.

Until now, the study of this assembly mechanism has been limited by the absence of suitable experimental approaches. Thus, the first aim of this study was to define a method to selectively identify and characterize co-co assembly. Specifically, we aimed at establishing a broadly applicable technology which would allow, in a single experiment, to determine (i) which proteins undergo co-co assembly on a proteome-wide scale, (ii) when, during translation, nascent chains start to interact and (iii) the efficiency of each co-co assembly event. To this end, we developed and validated an unbiased method, named Disome Selective Profiling (DiSP), which detects the formation of ribosome pairs (disomes) connected via nascent chains.

The second aim of my Thesis was to reveal the prevalence and features of co-co assembly in human cells. By applying DiSP on two human cell lines, in combination with a newly developed bioinformatics framework, we sought to identify the human co-co assembly proteome. Once established, this allows to answer the important question if co-co assembly is a general mechanism for the biogenesis of protein complexes inside cells. Moreover, exploring the features of the co-co assembly proteome can provide key insights into this assembly mode, by revealing the subcellular localization of co-co assembly candidates and indicating whether this it is mostly employed to assemble homomeric or heteromeric protein complexes. It further enables to study, for each candidate gene, the fraction of nascent chains that engages in co-co assembly, the nascent chain length required for initiation of assembly and the relative position and identity of the domains mediating this assembly mode.

The third aim of this study was to gain deeper insights into the molecular mechanisms mediating co-co assembly. We set the focus on the investigation of two possible topologies of co-co assembly interactions, namely association of proximal nascent chains emerging from the same or different polysomes (termed assembly *in cis* and *in trans*, respectively). Discriminating between these interaction modes can substantially advance our understanding of the functional implications and regulation of this assembly process. In particular, we explored the intriguing possibility that co-co assembly *in cis* enables transcript-templated homomer assembly. We further asked whether the concentration of nascent chains determined by the overall ribosome density on mRNAs or by local variations in translation speed are correlated to co-co assembly and thus may be employed to regulate this process inside cells.

# 3. RESULTS

## 3.1. Proteome-wide analysis of disome formation by Disome Selective Profiling

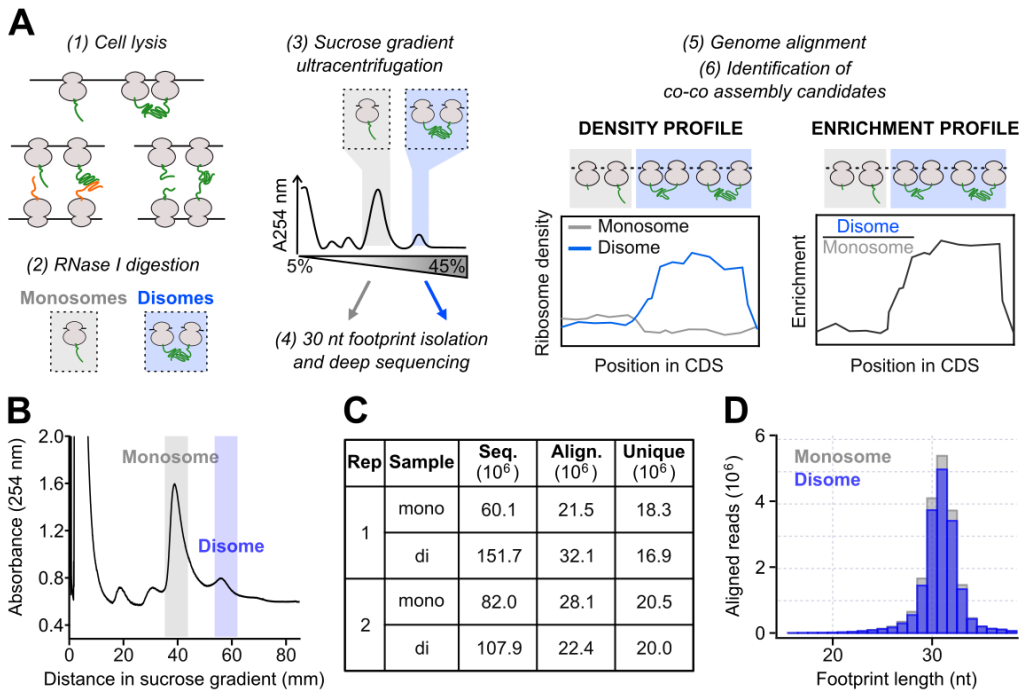
### 3.1.1. Disome formation is widespread in human cells

Aiming to establish a method for direct and proteome-wide detection of co-co assembly, we reasoned that this type of co-translational interactions will produce ribosome pairs (disomes) that are connected by their exposed nascent chains (Fig. 5A, step 1). These disomes remain connected upon RNA digestion and can be separated from single ribosomes (monosomes) by differential sedimentation in sucrose gradients (Fig. 5A, steps 2,3). Hence, disome formation across the proteome can be identified by independent sequencing of 30 nt-long footprints protected by monosomes and disomes and calculating the relative ribosome density along the coding sequence of individual genes (steps 4-6). In this method, named Disome Selective Profiling (DiSP), co-co assembly candidates are revealed by a shift of the footprint density from the monosome to the disome fraction, which can be visualized by density or enrichment profiles (step 6). Disomes can also result from inefficient RNA digestion of collided ribosomes (*47–50*). This disome sub-population protects double-length (60 nt) RNA footprints, which are excluded during library preparation of DiSP samples, but may be analysed in parallel (see chapter 3.6.2).

Importantly, compared to other existing methods employed to study co-translational assembly (usually based on purification of a bait protein and identification of the co-purified mRNAs, such as in RNA immunoprecipitation – RIP (*65*, *75*) – or in Selective Ribosome Profiling – SeRP (*66*)), DiSP provides proteome-wide information in a single experiment. Contrarily to the aforementioned approaches however, DiSP does not inform on the specificity of the pairwise interactions, which can only be deducted from prior knowledge on the final complex composition in the absence of further experiments.

I initially performed DiSP of human embryonic kidney cells (HEK293-T), in collaboration with my colleague Kai Fenzl. DiSP samples were collected from 10-25% sucrose gradients to minimize cross-contamination between the monosome and disome fractions (Fig. 5B). Monosome and disome libraries were deep-sequenced and aligned to the protein coding genome; quantitative biases introduced by the amplification step during library preparation were corrected employing unique molecular identifiers, UMIs (Fig. 5C). The resulting uniquely mapped reads of both monosome and disome samples showed typical footprint length distributions indicative of high quality ribosome profiling libraries (Fig. 5D).

In order to estimate the prevalence of disome formation on a genome-wide scale, we compared the average footprint density per gene (expressed in Reads Per Kilobase per Million – RPKM) in the disome versus the monosome fraction (Fig. 6A). This analysis revealed that over 1300 out of about 10500 expressed genes had a more than two-fold higher density in the disome compared to the monosome fraction, suggesting that disome formation is widespread in human cells. Note that the size effect of disome over monosome enrichment varies substantially depending on a number of factors, including the extent of RNA digestion, the efficiency of the monosome-disome separation in sucrose gradients and the salt concentration in buffers. All these parameters were optimized in series of DiSP




| Rep | Sample | Seq. ($10^6$) | Align. ($10^6$) | Unique ($10^6$) |
|---|---|---|---|---|
| 1 | mono | 60.1 | 21.5 | 18.3 |
| | di | 151.7 | 32.1 | 16.9 |
| 2 | mono | 82.0 | 28.1 | 20.5 |
| | di | 107.9 | 22.4 | 20.0 |

***Figure 5.*** *Disome Selective Profiling (DiSP) of human HEK293-T cells*

*A) Schematic representation of the DiSP workflow. Cells are lysed under conditions routinely employed for ribosome profiling to preserve the in vivo translational status (1); lysates are treated with RNase I producing monosomes and disomes (2) which are separated by differential sedimentation in sucrose gradients (3); 30 nt-long RNA footprints from the monosome and the disome fractions are isolated and deep-sequenced (4); after genome alignment (5), co-co assembly candidates can be identified by analysing the monosome and disome density distributions separately or by calculating disome over monosome enrichment profiles (6).*

*B) Representative 10-25% sucrose gradient absorbance profile of RNase-digested HEK293-T lysates. Collected monosome and disome fractions for DiSP are highlighted in grey and blue, respectively.*

*C) Sequencing depth of DiSP libraries of HEK293-T cells. Rep = biological replicate; Seq. = number of sequenced (raw) reads; Align. = number of reads aligning to the protein-coding genome; Unique = number of uniquely mapped reads after discarding duplicated unique molecular identifiers (umi), reflecting unique input footprints.*

*D) Footprint length distributions of uniquely mapped reads in monosome and disome libraries. One replicate shown.*

experiments (described in more detail in the PhD thesis of Kai Fenzl). Comparing the results of these experiments showed that while the intensity of the disome shift varied, the identity of the genes and the onset position of the disome shift remained the same. The analysis of an exemplary dataset shown in Fig. 6A reveals the widespread occurrence of disome formation. An unbiased method to identify co-co assembly candidates and determine the prevalence of this phenomenon is presented in chapter 3.3.

We went on to analyse footprint density distributions on individual genes. In many cases, both monosome and disome fractions followed a correlated distribution along the gene's coding sequence (CDS), indicating no disome formation. This was the case for *JUN*, encoding for c-Jun, which forms a heteromeric transcription factor with c-Fos through its C-terminal leucine zipper domain (Fig. 6B, left). A number of other genes however showed a clear shift of the footprint density from the monosome to the disome fraction. One example

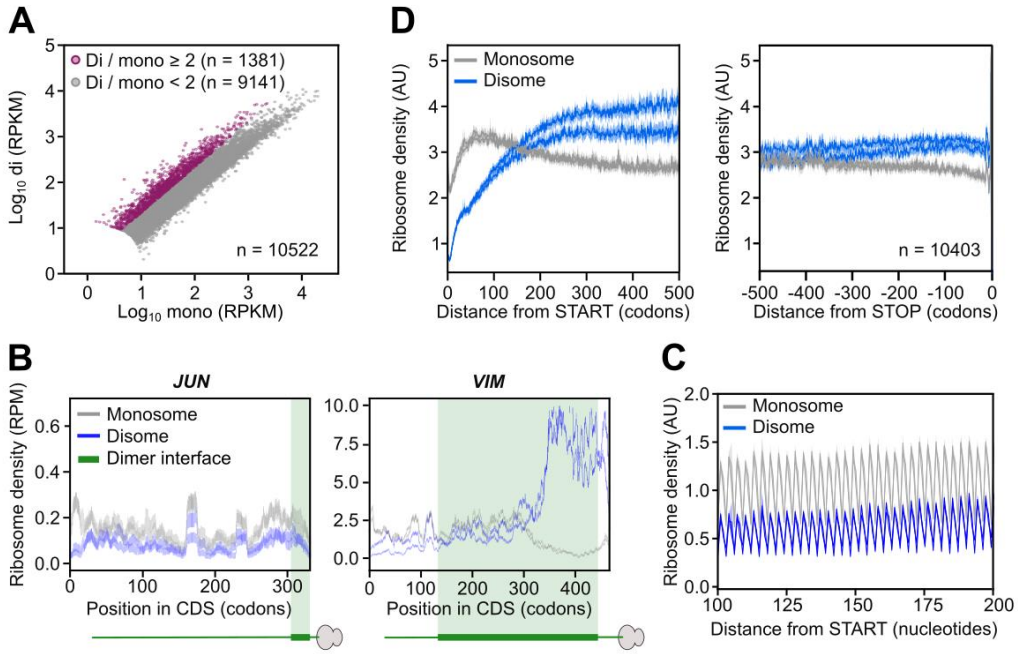



***Figure 6.*** *Disome formation in human HEK293-T cells detected by DiSP*

*A) Comparison of the gene-wise monosome and disome fooprint densities normalized by the length of the gene's coding sequence (RPKM). One replicate shown.*

*B) DiSP data of two representative genes. Monosome and disome densities are plotted as vertical bars, indicating the 95% confidence intervals at each codon position of the genes' coding sequences. Cartoons indicate the exposed nascent chain segments during translation (shifted by 30 codons compared to ribosome positions to account for the ~30 C-terminal residues that are included in the ribosome exit tunnel); green boxes represent the dimer interfaces. Two biological replicates shown.*

*C) Footprint periodicity in monosome and disome libraries is visualized by metagene profiles with nucleotide resolution. All reads were assigned to ribosome P-sites and average densities across all reliably expressed genes aligned to the +1 nt of the start codon were calculated. A region encompassing nucleotides 100 – 200 from the start is shown. Solid lines indicate average densities, shadows indicate the 95% bootstrapping confidence interval. Two biological replicates shown.*

*D) Metagene analysis of the monosome and disome footprint densities across all expressed genes aligned to the start (left) or the stop codon (right). The contribution of each gene was normalized to its expression (derived from total translatome samples). Solid lines indicate average densities, shadows indicate the 95% bootstrapping confidence interval. Two biological replicates shown.*

was *VIM*, encoding for vimentin, a type III intermediate filament which homo-dimerizes via a long coiled coil dimerization domain (named rod, Fig. 6B, right). Ribosomes translating *VIM* converted from monosomes to disomes when more than half of the rod domain was exposed on the ribosomal surface, suggesting that disome formation was mediated by co-translational homo-dimerization. Importantly, the increase in disome density near codon 320 of *VIM* correlated with a depletion of footprints in the monosome, suggesting that a large fraction of the ribosomes was engaged in disomes after this onset position.

We next performed metagene analyses, by calculating the average monosome and disome footprint density along the coding sequences of all expressed genes aligned to translation start. Calculating metagene profiles with nucleotide resolution revealed the characteristic 3-nucleotide periodicity of monosome and disome footprints, again supporting the overall

quality of the presented data (Fig. 6C). To reveal the nascent chain length dependence of disome shifts, genes were aligned to the start or stop codon and metagene profiles were computed with codon resolution. This analysis revealed that ribosomes migrated mostly as monosomes at early translation stages and converted to disomes after 100 to 200 codons (Fig. 6D, left); disome enrichment was in average stable until translation termination (Fig. 6D, right). These results are in line with a model of co-co assembly, where a nascent chain portion of sufficient length has to be synthetized and exposed on the ribosome surface to establish stable co-translational interactions.

### 3.1.2. Disome formation is robustly correlated in two human cell lines

To demonstrate that disome formation is not a phenomenon limited to HEK293-T cells, I repeated DiSP in osteosarcoma U2OS cells, in collaboration with my colleague Kai Fenzl.
DiSP can be performed using virtually any cell type without the need for prior manipulations (e.g. cell line generation), however the protocol has to be adapted based on the specific features of each cell type. Differences between the protocols employed for HEK293-T and U2OS cells are summarized in Table 1.

***Table 1.*** *Comparison of protocols employed for DiSP of HEK293-T and U2OS cells*

| Step (Fig. 5A) | DiSP of HEK293-T cells | DiSP of U2OS cells |
|---|---|---|
| 1 | Cells were first detached and pelleted. Pellets were resuspended in lysis buffer. | Cells were lysed on-dish by scraping. |
| 2 | RNA digestion was performed after lysis, on the cleared lysate. | RNA digestion was performed during lysis (RNase 1 directly supplemented in the lysis buffer) |
| 1 to 3 | All buffers contained 500 mM KCl to allow direct comparison with control experiments which required high salt concentrations (see chapter 3.2.2) | All buffers contained 150 mM KCl (similar to the physiological salt concentration). |

DiSP of U2OS cells revealed that over 400 out of ~10500 expressed genes had a more than two-fold higher footprint density in the disome compared to the monosome sample (Fig. 7A). This fraction of disome enriched genes is substantially smaller compared to the one obtained in HEK293-T cells (Fig. 6A). A direct comparison of the disome over monosome enrichment per gene revealed that, while enrichment values of the U2OS dataset were generally lower, they were robustly correlated to the ones of the HEK293-T dataset (Fig. 7B). The generally lower intensity of the disome shift can be explained by the lower salt concentration employed for the U2OS dataset (Table 1). Indeed, direct comparison of DiSP experiments performed under physiological (150 mM KCl) or high salt concentrations (500 mM KCl) in HEK293-T cells, revealed that high salt reproducibly generated more pronounced disome shifts, but did not affect the identity of disome-enriched genes, nor the shape of enrichment profiles (data shown in the PhD thesis of Kai Fenzl). This difference can be explained by a higher resolution of the monosome-disome separation in sucrose gradients containing high salt concentrations, which reduces the cross-contamination between the two fractions. Besides these apparent discrepancies, DiSP of HEK293-T and U2OS cells revealed highly

overlapping results. First, metagene analysis including all expressed genes in U2OS cells revealed a similar shift of the footprint density from monosomes to disomes after 100 to 200 codons were translated (compare Fig. 7C and 6D, left).

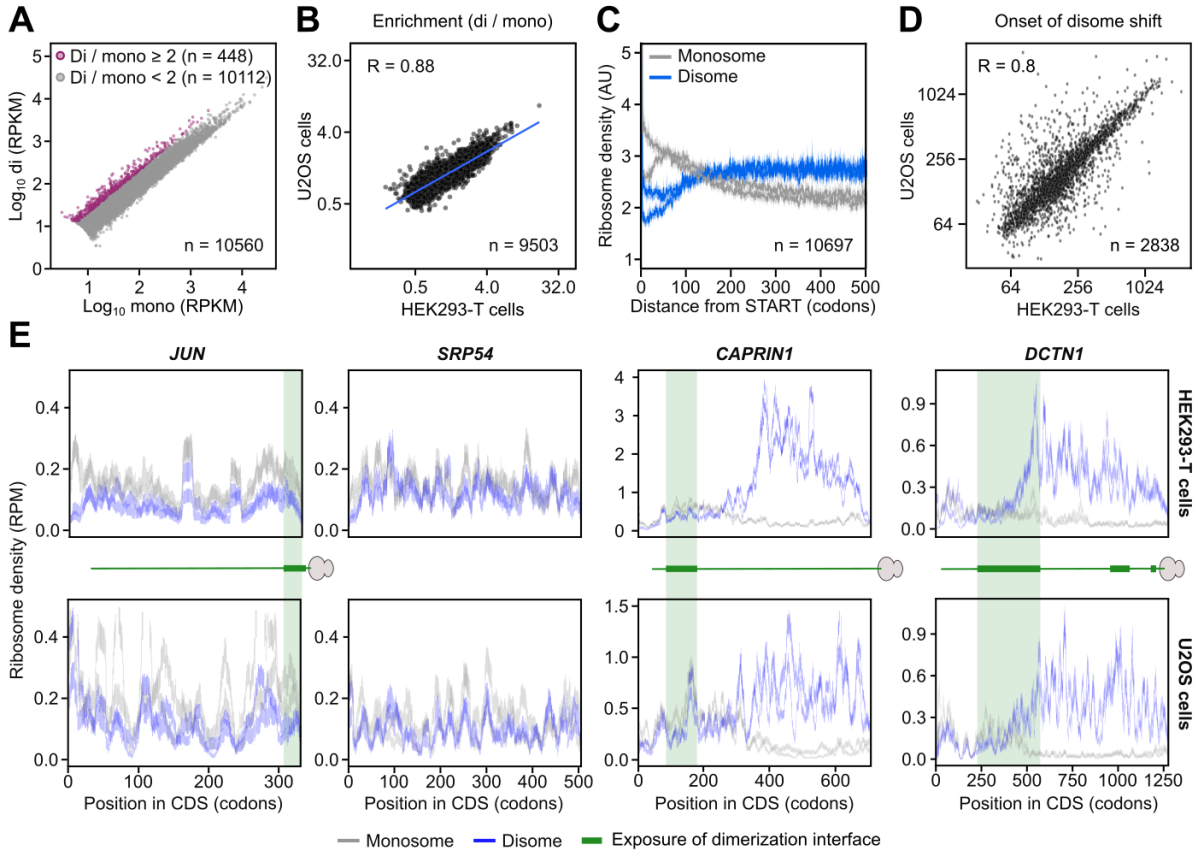



***Figure 7.*** *DiSP of HEK293-T and U2OS cells reveal largely overlapping results*

*A) Comparison of the gene-wise monosome and disome fooprint densities of U2OS cells normalized by the length of the gene's coding sequence (RPKM). One replicate shown.*

*B) Comparison of the gene-wise disome over monosome enrichment obtained by DiSP of HEK293-T and U2OS cells. Enrichment values are averages of two biological replicates. Only genes with sufficient coverage in all four datasets were included in the analysis.*

*C) Metagene analysis of the monosome and disome footprint densities across all expressed genes in U2OS cells aligned to the start codon. The contribution of each gene was normalized to its expression (derived from total translatome samples). Solid lines indicate average densities, shadows indicate the 95% bootstrapping confidence interval. Two biological replicates shown.*

*D) Comparison of the onset of disome formation, defined as the inflection point of a sigmoidal shape fitted to the gene's disome over monosome enrichment profile (see chapters 3.3.1 and 5.8.3 for details). Onsets are averages of two biological replicates for each cell line. Only genes with sufficient coverage and that fit best to a sigmoidal rather than a flat enrichment profile in all four datasets were included in the analysis.*

*E) Direct comparison of single gene density profiles obtained by DiSP of HEK293-T and U2OS cells. Monosome and disome densities are plotted as vertical bars, indicating the 95% confidence intervals at each codon position of the genes' coding sequences. Cartoons indicate the exposed nascent chain segments during translation (shifted by 30 codons compared to ribosome positions to account for the ~30 C-terminal residues that are included in the ribosome exit tunnel); green boxes represent the dimer interfaces. Two biological replicates shown.*

Second, we observed a robust correlation of the onsets of disome formation (Fig. 7D) and highly similar single gene enrichment profiles (Fig. 7E).
Taken together, these data indicate that disome formation is a general feature of a specific subset of human proteins across different cell types.

### 3.1.3. Analysis of known examples of co-translational assembly

We next asked if our proteome-wide DiSP data could recapitulate previously described examples of co-translational assembly.
Evidence dating back to the 1980s indicate that a number of cytoskeletal proteins associate to the cytoskeleton already during translation (*62*). Reported examples include myosin heavy chain (*109*), titin (*110*), vimentin (*111*), tropomyosin alpha and beta (*112*) and the more recently studied peripherin (*96*). Of these proteins, two members of the myosin family (*MYH9* and *MYH10*), vimentin (*VIM*) and the two subunits of tropomyosin (*TPM1* and *TPM2*) were expressed in at least one of the tested cell lines (HEK293-T and U2OS), and all five showed a disome enrichment correlating with the exposure of major parts of their coiled coil dimerization domains (Fig. 8, top). Note that *TPM2* contains a region with lower sequence coverage (codons 200-300), as indicated by the larger confidence intervals of enrichment in this region. This may be caused by homology of this sequence to other regions in the human genome, which results in a loss of reads that cannot be mapped unambiguously. Hence, while disome enrichment is clearly apparent towards the 3' end of *TPM1*, it is less clear whether it also occurs during translation of its partner subunit *TPM2*.
A systematic analysis of other known components of the cytoskeleton revealed that all intermediate filaments (IFs) expressed in HEK293-T or U2OS cells showed a shift of footprints from monosome to disome, coinciding with partial exposure of their common coiled coil dimerization domain (named rod, Fig. 8, bottom). Among them were type II, III, IV and V IFs, including keratins (*KRT75* and *KRT80*), vimentin (*VIM*), neurofilament polypeptides light and medium (*NEFL* and *NEFM*), α-internexin (*INA*) and all nuclear lamins (*LMNA*, *LMNB1* and *LMNB2*). This result suggests a conservation of the mechanism mediating formation of the homodimeric building block of all intermediate filaments, driven by co-translational assembly of the rod domain. In addition, we found that several cytoskeleton-associated motor proteins showed a prominent disome shift upon partial emergence of their coiled coil dimerization domains, including kinesins and subunits of the dynein and dynactin motor complexes (data not shown).
Additional works pointed to a more specific mechanism of co-translational assembly, involving neighbouring nascent chains on one mRNA molecule (named co-co assembly *in cis* in this Thesis). In two of these studies, the polysome was proposed to act as a scaffold that dictates ordered interactions of neighbouring nascent chains for co-translational assembly of large homomeric protein complexes. The first study described assembly of the homo-hexameric tenascin, a constituent of the extracellular matrix, which was proposed to assemble by simultaneous association of six nascent chains emerging from the same polysome during co-translational translocation (*77*). The second study proposed that formation of the mega-Dalton sized vault particle initiates through dimerization of neighbouring nascent chains, followed by stacking of dimers emerging from consecutive ribosome pairs on one polysome (*78*). This would allow gradual and ordered growing of the vault particle structure close to the 3' end of the polysome structure, until completion of the 78-member particle. Both tenascin and vault proteins (encoded by *TNC* and *MVP*, respectively) are expressed in U2OS (but not in HEK293-T cells), however no disome shift

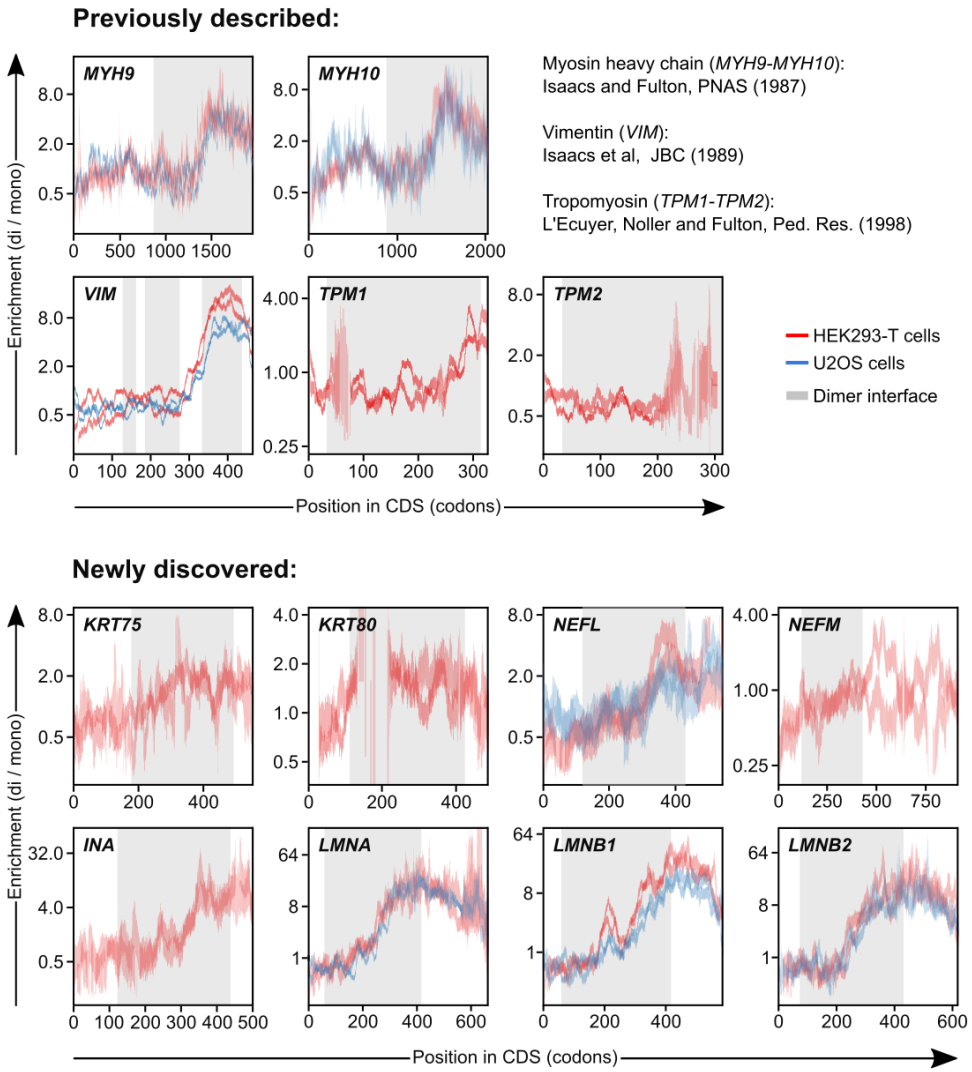



***Figure 8.*** *DiSP confirms and extends current knowledge on co-translational assembly of cytoskeletal proteins*

*Analysis of the disome over monosome enrichment distribution of individual genes encoding for cytoskeletal proteins that were previously described (top) or discovered in this study (bottom) to assemble co-translationally. Two biological replicates of HEK293-T (red) and U2OS cells (blue) shown. Grey boxes indicate exposure of coiled coil dimerization interfaces during translation.*

was observed in DiSP data of these genes (Fig. 9). On one hand, this result may agree with the described rapid assembly of these multimeric complexes, in which dimeric intermediates may not exist or be short-lived due to the high local concentration of nascent subunits. On the other hand, a shift of translating ribosomes to higher polysomes in sucrose gradients should result in a substantial loss of both monosome and disome footprint densities in DiSP data after the onset of assembly and until translation end. However, we do not find evidence for such effect in our data, speaking against the proposed mechanism of co-translational oligomerization *in cis* of the vault particle and tenascin complex.

Among the most notable examples of proteins reported to assemble via interactions of two nascent subunits are the NF-κB and p53 transcription factors (*67*, *79*). Rapid dimerization of NF-κB1 nascent chains on one polysome was indicated as critical for formation of the mature complex by shielding the N-terminal interaction domain (named Rel Homology Domain,

RHD) from co-translational cleavage by proteasome. DiSP confirmed that ribosomes translating *NFKB1* undergo disome formation and further indicated a correlation with the complete exposure of the RHD domain (grey box, Fig. 9), suggesting that folding of this globular interaction motif is required for the nascent chain interaction.
A *cis* assembly modus was also proposed to ensure transcript-templated homodimerization of p53, while tetramerization was proposed to occur by post-translational dimerization of dimers. Importantly, this combination of co- and post-translational pathways would constrain the possible composition of the final complex and impact the penetrance of dominant negative mutations, hence explaining the observed lower frequency of defective complexes. However, our analysis of DiSP data did not support disome formation during translation of *TP53* (Fig. 9). On one hand, this result may suggest possible artefacts of the findings by Nicholls and colleagues, which are solely based on *in vitro* translation systems. On the other

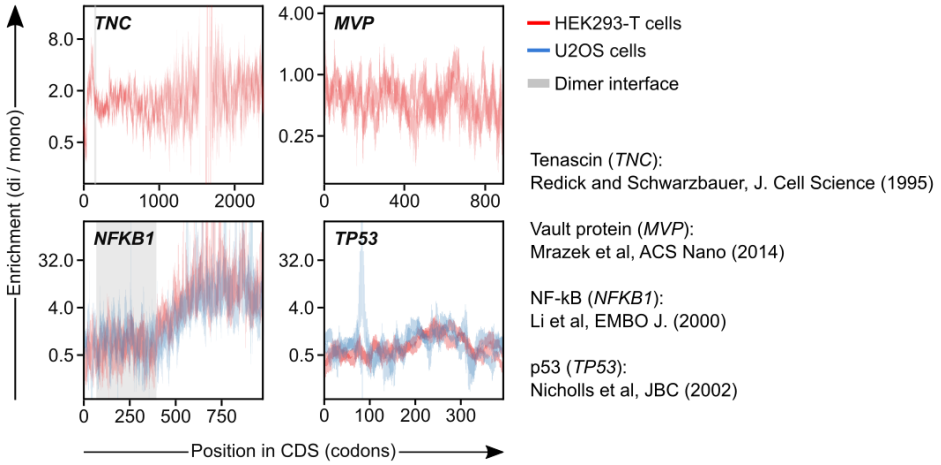



***Figure 9.*** *DiSP analysis of four previously reported examples of co-translational assembly on one polysome*

*Analysis of the disome over monosome enrichment distribution of individual genes encoding for tenascin (TNC), major vault protein (MVP), NF-κB (NFKB1) and p53 (TP53), that were previously described to assemble co-translationally. Two biological replicates of HEK293-T (red) and U2OS cells (blue) shown. Grey boxes indicate exposure of annotated dimerization interfaces during translation.*

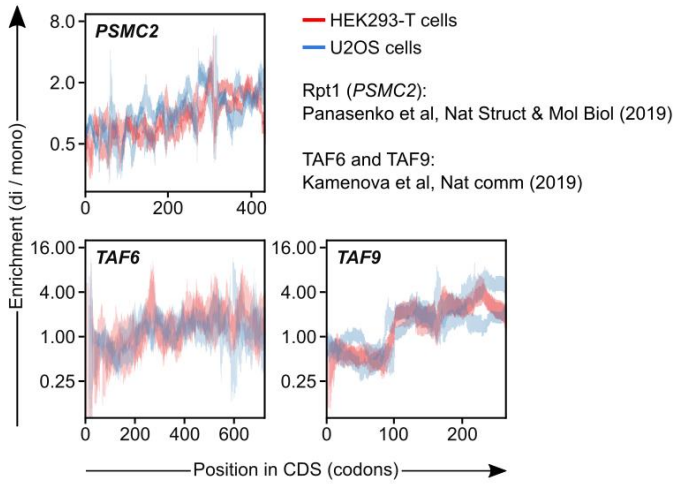



***Figure 10.*** *DiSP analysis of previously reported examples of co-translational heterodimer assembly*

*Analysis of the disome over monosome enrichment distribution of individual genes encoding for the proteasome subunit RPT1 (PMSC2) and the transcription factor TFIID subunits TAF6 and TAF9. Two biological replicates of HEK293-T (red) and U2OS cells (blue) shown.*

hand, even though both U2OS and HEK293-T cells express wild type p53, it is possible that biogenesis of this oncoprotein is differently regulated in different cell types (e.g. tumorigenic / immortalized / progenitor / differentiated cells) or under different conditions (e.g. exposure to inducing stimuli). Finally, it is also possible that co-translational dimerization of p53 is missed by our approach if the hydrophobic contacts between the nascent chain N-termini are not strong enough to endure the DiSP sample preparation protocol.
Two recent studies suggested co-translational heterodimer formation via interaction of two nascent subunits *in trans*. These studies focused on proteasome subunits RPT1 and RPT2 and the transcription initiation factor TFIID subunits TAF6 and TAF9 (*71*, *75*). Of the two proteasome subunits, RPT1 (encoded by *PSMC2*) showed a rather weak increase of the disome over monosome enrichment along the coding sequence (Fig. 10). On the other hand, its partner subunit RPT2 (encoded by *PSMC1*) is paralog to at least other five proteasomal subunits, a feature which does not allow unambiguous reads mapping and hinders inspection of this gene's profile. Thus, our data does not allow to verify the previously described co-co assembly of this heterodimer *in trans* (*71*).
Conversely, *TAF6* and *TAF9* both revealed some degree of disome formation. However, the enrichment values and the length of the disome shift differed between the two proteins (Fig. 10). Assuming that TAF6 and TAF9 heterodimerize *in trans*, the duration of the disome shift should be similar between both mRNAs (assuming comparable translation speed).
Hence, our DiSP results do not provide clear evidence for a model of co-co assembly *in trans* of these protein pairs.

### 3.1.4. DiSP with chemical crosslinking

DiSP detects disomes that remain connected throughout the ribosome purification procedure by sucrose gradient centrifugation. Disomes connected by nascent chains with intermediate or low affinity may dissociate and be missed in our analysis, even if they exist inside cells.
To test this possibility, we performed DiSP of U2OS cells that included a lysis-coupled chemical crosslinking step. Specifically, we employed a mix containing an amine-amine (2.5 mM BS3) and a carboxyl-amine crosslinker (20 mM EDC), a combination previously shown to stabilize interactions of the chaperone trigger factor with nascent chains in bacterial lysates (*113*). We first asked if crosslinking would allow capturing less stable dimerization intermediates and hence cause a shift of disome detection to earlier translation stages. Analysing the position along coding sequences where the shift from monosomes to disomes occurs (termed onset) revealed a robust correlation between the two datasets (Fig. 11A, left, Pearson R = 0.77), which was highly comparable to the correlation obtained for biological replicates of the same DiSP experiment (R = 0.74 for the replicates of U2OS without crosslinking). Analysis of single gene profiles revealed that the candidates with varying onsets (outliers in Fig. 11A, left) were characterized by less pronounced / shallower disome shifts, which causes a higher variation range in the onset determination. Importantly, we did not observe a general shift of onsets towards earlier CDS positions in cross-linked samples (Fig. 11A, right). Furthermore, crosslinking did not cause appearance of new disome shifts in genes that did not show any in the non-crosslinked dataset. While gene-wise disome over monosome enrichment values were highly correlated in the two datasets (Fig. 11B, left), we found that disome enrichment of some of the candidates showing the highest disome shifts in the non-crosslinked dataset was further increased after crosslinking (Fig. 11B, right). This was not a general feature of all top disome enriched candidates, as revealed by metagene profiles of disome over monosome enrichment which were completely unaffected by

crosslinking (Fig. 11C). Instead, only few genes (such as *NOLC1* and *CAPRIN1*) showed increased disome detection, without any detectable change in onset position (Fig. 11D).
Crosslinking of DiSP samples would have been expected to either (i) cause a shift of onsets earlier positions in some or all candidates due to stabilization of dimerization intermediates, (ii) allow detection of a larger number of candidates, thereby reducing the enrichment on individual candidates (because of the relative nature of disome quantification performed by DiSP), or (iii) show no difference compared to non-crosslinked data (if nascent chain interactions are generally stable enough to survive the centrifugation during sample preparation or if crosslinking did not work). The results presented in this chapter are therefore rather surprising. Additional experiments employing different types of crosslinkers or *in vivo* crosslinking may be an interesting future direction.

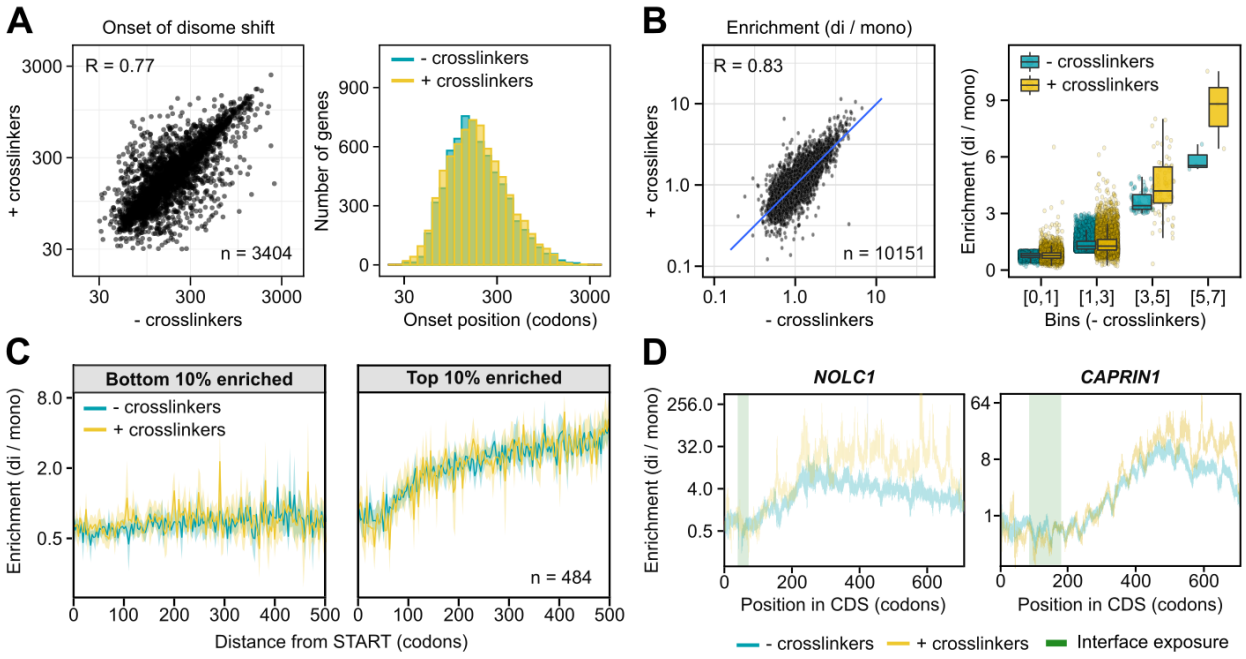



***Figure 11.*** *DiSP of U2OS cells with a lysis-coupled crosslinking step*

*A) Comparison of the onset of disome formation between non-crosslinked and crosslinked datasets. Onsets of the non-crosslinked dataset are averages of two biological replicates; one replicate of DiSP with crosslinking was generated and analysed. Only genes with sufficient coverage and that fit best to a sigmoidal rather than a flat enrichment profile in all three datasets were included in the analysis.*

*B) Enrichment values (disome / monosome) in non-crosslinked and crosslinked datasets are directly compared for each gene (left) or plotted as a function of binned enrichment values in the non-crosslinked dataset (right). Only genes with sufficient coverage in both datasets were included. One biological replicate shown.*

*C) Metagene enrichment profiles (disome / monosome) aligned to translation start of non-crosslinked and crosslinked DiSP samples of U2OS cells, including genes with the bottom (left) or the top (right) 10% disome over monosome enrichments. One biological replicate shown.*

*D) Enrichment profiles (disome / monosome) of two single gene examples showing a higher disome enrichment in the crosslinked than in the non-crosslinked dataset. Green bars indicate exposure of the nascent chain segments containing the dimerization interface during translation. One biological replicate shown.*

## 3.2. Disome formation is nascent chain-dependent

To challenge our model that disome formation is mediated by nascent chain interactions, we performed a series of controls of the DiSP approach which all relied on monitoring the reduction of disome shifts upon suppression of nascent chain connections.

### 3.2.1. DiSP with targeted nascent chain cleavage

We first tested targeted cleavage of nascent chains by TEV protease to validate co-co assembly of *LMNA*, encoding for the A and C isoforms of nuclear lamins. These alternatively spliced isoforms share most of their coding sequence, including the N-terminal head domain and the ~ 350 amino acids-long coiled-coil dimerization motif (named rod), and differ only for the length of their C-terminal tail domain (Fig. 12A). Importantly, lamins homodimerize in the cytosol by means of the rod domain and are then transported into the nucleus where they polymerize and mix with other lamin types to form the nuclear lamina. Our DiSP data indicated that the first step of lamin assembly (homodimer formation) starts co-translationally, when approximately half of the rod domain is exposed on the ribosome surface (Fig. 12A).
To test if disomes translating *LMNA* are connected by nascent chains, we generated one synthetic construct encoding for, from N- to C-terminus, the head and rod domains of lamin A/C, followed by a TEV cleavage site and eGFP (Fig. 12B, top). Upon transfection of HEK293-T cells and TEV treatment of lysates, lamin's rod domain should be removed from nascent chains (Fig. 12B, bottom). Hence, DiSP of cells transfected with the *LMNA-TEV-TS-*

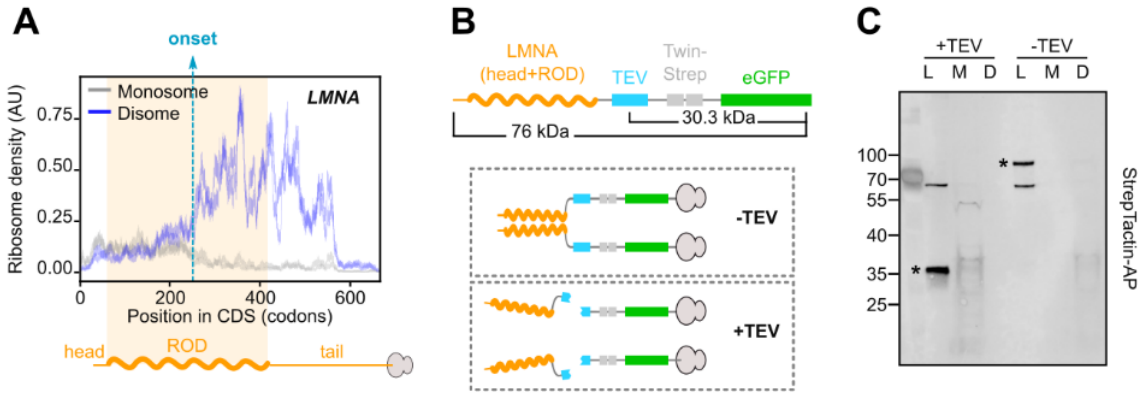



***Figure 12.*** *Cleavage of nascent lamin C by TEV protease*

*A) Monosome and disome footprint density distribution on LMNA coding sequence in HEK293-T cells. Exposed nascent chain segments during translation and domain annotations are indicated; a yellow box represents exposure of the dimerization domain (rod); the onset of disome formation is indicated by a dashed vertical arrow. Two biological replicates shown.*

*B) Schematic representation of the synthetic construct employed in this experiment (LMNA-TEV-TS-eGFP). The N-terminal LMNA segment should allow dimerization of nascent chains (-TEV); this segment is removed upon TEV digetion, resulting in splitting of disomes to monosomes (+TEV).*

*C) Western blot analysis of total lysates (L), monosome (M) and disome (D) samples derived from HEK293-T cells expressing plasmid-encoded LMNA-TEV-TS-eGFP. Samples of control (-TEV) and TEV-treated lysates (+TEV) are shown. Bands corresponding to the fusion protein are highlighted (*). An unspecific band of about 60 kDa was repeatedly observed in StrepTactin immunoblots of human cell lysates.*

*eGFP* construct should reveal a TEV-dependent reduction of the disome footprints mapping to the downstream eGFP encoding sequence.

We first showed by western blotting that TEV effectively cleaved the fully synthetized protein in cell lysates of transfected cells (L, Fig. 12C). The TwinStrep tag employed for immune-detection was positioned C-terminal to the TEV cleavage site. This should allow to monitor the nascent chain segments that remain attached to ribosomes after TEV digestion in the monosome and disome fraction (M and D, Fig. 12C). A smeared signal, indicative of nascent chains of varying lengths, was present in the disome fraction of untreated lysates (-TEV) and shifted to the monosome fraction of treated lysates (+TEV). While the TEV-dependent shift of the nascent protein from disomes to monosomes was in agreement with our model of co-co assembly, the size range of the detected nascent chains did not fit expectations (46 to 76 kDa for untreated and up to 30 kDa for TEV-treated samples).

To achieve better resolution, we next transfected cells with the plasmid-encoded construct and performed DiSP of untreated or TEV-treated lysates. The separation of monosome and disome peaks on sucrose gradients was very poor for this experiment (Fig. 13A). We later discovered that this was a consequence of RNA over-digestion, which likely affected the ribosome structure, as indicated by the apparent degradation of 40S subunits (Fig. 13A) and by the production of shorter RNA footprints (Fig. 13B). This resulted in high cross-contamination of monosome and disome samples and a generally low magnitude of disome shifts. Analysis of the disome over monosome enrichment profile of plasmid-encoded *LMNA-TEV-TS-eGFP* still revealed a mild disome shift on both untreated and TEV-treated samples (Fig. 13C). The onset of disome formation correlated with emergence of the same nascent chain length that mediates disome formation of endogenous lamin C in untreated samples (onset, Fig. 13C), but not in TEV-treated samples, where disome enrichment apparently initiated earlier. This result is unexpected, as disome formation initiates before the TEV cleavage site is synthetized and hence should not be affected by TEV treatment. On the other hand, we did observe a TEV-dependent disome reduction which correlated with

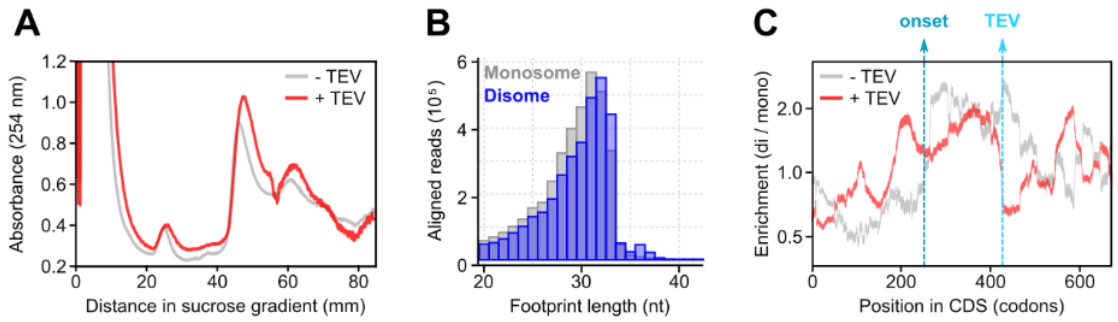



***Figure 13.*** *DiSP combined with targeted nascent chain cleavage*

*A) Absorbance profile at 254 nm of RNase-digested lysates along 10-25% sucrose gradients. Monosome and disome fractions were collected from untreated or TEV-treated lysates for DiSP.*

*B) Footprint length distributions of uniquely mapped reads in monosome and disome libraries of the untreated (-TEV) dataset (one replicate) reveal a high frequency of short (20-30 nt-long) footprints.*

*C) DiSP enrichment profile (disome / monosome) along the plasmid-encoded LMNA-TEV-TwinStrep-eGFP gene. The expected onset of disome formation (derived from DiSP data of endogenous LMNA) and the expected position of TEV-dependent suppression of disomes (coinciding with exposure of the TEV cleavage site on the ribosome surface) are indicated by dashed arrows.*

exposure of the TEV cleavage site on nascent chains (TEV, Fig. 13C).
In summary, these results can only partially support a model of nascent chain-dependent disome formation. The noise in the data is most likely due to over-digestion which caused rRNA degradation and production of ribosome particles with varying densities that sediment differently in sucrose gradients. Solving this technical problem should however allow to successfully employ this experimental approach in the future to validate and test the specificity of nascent chain interactions.

### 3.2.2. DiSP with puromycin treatment

We next asked if disome shifts were sensitive to nascent chain release by puromycin on a proteome-wide scale.
Puromycin causes premature nascent chain termination. Its structure mimics the aminoacyl moiety of an aa-tRNA, allowing it to enter the A-site of ribosomes and be incorporated to the C-terminus of the prematurely terminated peptide. This ribosome-catalysed reaction produces a deacylated P-site tRNA and a puromycilated nascent chain which is released (Fig. 14, left).

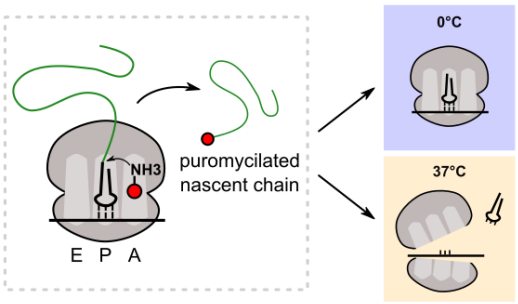



***Figure 14.*** *Puromycin mechanism of action*

*Puromycin (red) enters the ribosome A site and employs its free amino group to accept the growing peptide (green) from the P-site peptidyl tRNA in a reaction catalysed by the peptidyltransferase centre (PTC). Release of the puromycilated nascent chain leaves intact ribosomes that still contain a P-site tRNA (at 0°C) or causes release of the P-site tRNA and ribosome disassembly (at 37°C).*

Performing DiSP on lysates treated with puromycin bears two significant advantages over alternative control experiments: (i) it allows to test the whole proteome in a single experiment and (ii) it specifically probes the nascent chain nature of disome connections.
On the other hand, combining a ribosome profiling-based method with puromycin-treatment is challenging. The first issue is that puromycin cannot be used in conjunction with cycloheximide, which is employed to stall translation in all ribosome profiling-based approaches. The second, more important issue, is that puromycin-induced nascent chain release results in ribosome disassembly under most conditions, which is incompatible with the footprinting procedure of ribosome profiling. The protocol employed for our puromycin control of DiSP is based on a detailed study by Blobel and Sabatini describing the effects of puromycin on polysome integrity (*114*). Importantly, they showed that at 0°C nascent chain release leaves polysomes intact, as the deacylated tRNA remains attached to the ribosome P-site. Heating to 37°C leads to P-site tRNA release from ribosomes lacking the nascent protein and hence to ribosome disassembly (Fig. 14). Under these conditions, polysome disassembly and re-association of subunits into monosomes occurs during the long ultracentrifugation step required for polysome profiling experiments, but can be prevented by fixation of lysates with formaldehyde.
Based on these findings, we designed a protocol suitable for HEK293-T cells which includes the preparation of both undigested samples employed for polysome analysis and RNase-

digested samples employed for DiSP (Fig. 15A). Most important features of this protocol, provided in chronological order, include the following:

(1) Cell lysis is performed in the absence of cycloheximide to avoid downstream inhibition of

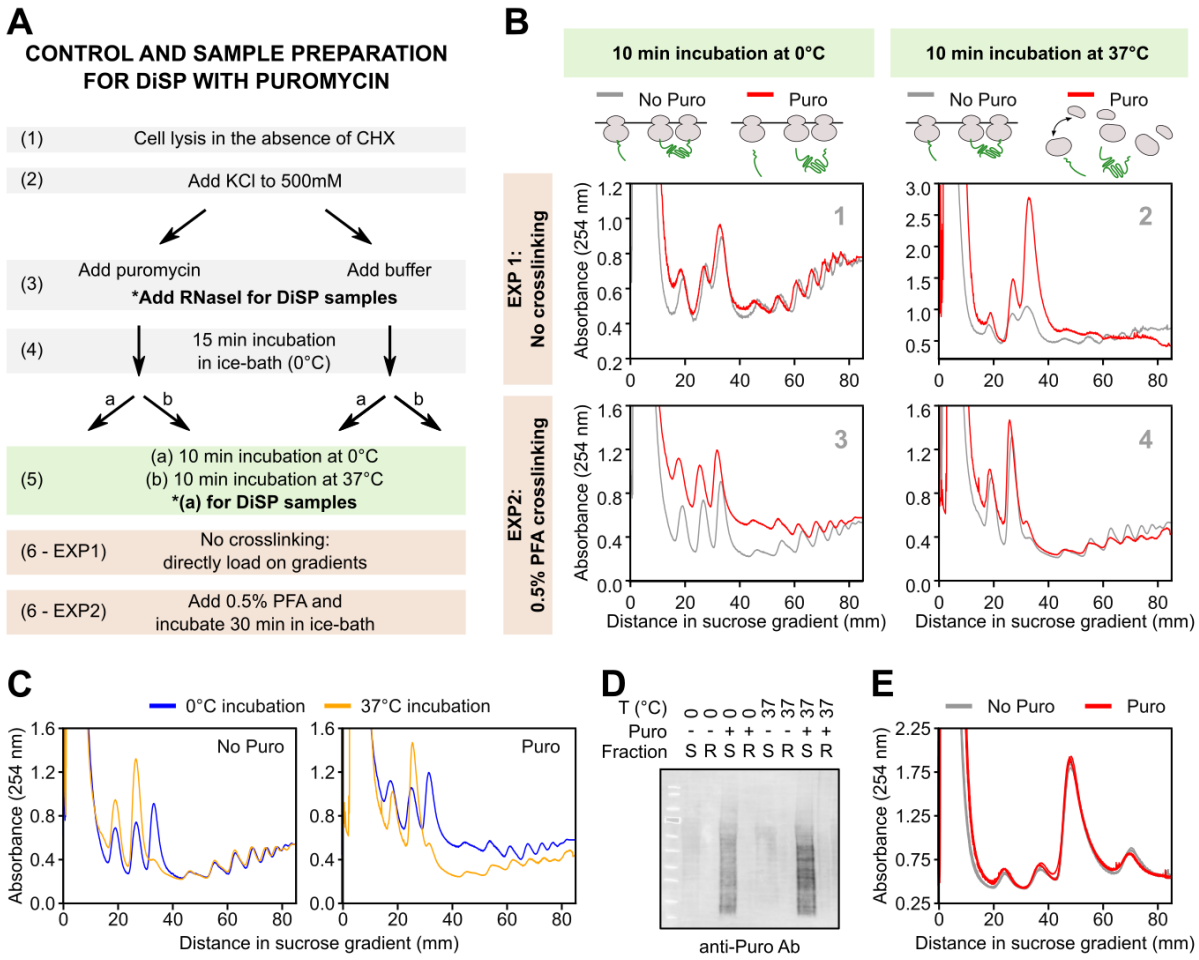



***Figure 15.*** *Establishing a protocol for DiSP with puromycin*

*A) Flow chart illustrating the protocol employed for preparation of control (undigested lysates employed for the polysome analyses shown in panel B of this figure) and DiSP samples (RNase-digested lysates). Steps specific to preparation of DiSP samples are highlighted (*).*

*B) Polysome profiles of control lysates prepared in the presence or absence of puromycin (shown in red and grey, respectively). Lysates were either incubated at 0°C or at 37°C (step 5 of the workflow) and were either crosslinked or not before loading on gradients (step 6 of the workflow).*

*C) Polysome analysis reveals the effect of temperature on polysome integrity: incubation at 37°C only affected polysome integrity in the presence (Puro) but not in the absence of puromycin treatment (No puro).*

*D) Immunostaining of puromycilated nascent proteins detected with anti-puromycin antibody in supernatant (S, including the top fractions of gradients before elution of the small ribosomal subunit) and ribosomal fractions (R, including monosome and polysome fractions).*

*E) Sucrose gradient absorbance profile of DiSP samples prepared in the presence or absence of puromycin treatment (two biological replicates shown).*

puromycin.
(2) Lysates are supplemented with KCl to a final concentration of 0.5 M. Under these conditions puromycin releases about 70% of all nascent chains (*114*).
(3-4) Lysates are either treated with 2 mM puromycin or left untreated for 15 minutes in an ice bath (~ 0°C). RNase 1 is added at this step for preparation of DiSP samples, and is omitted from polysome profiling samples.
(5) Next, lysates are split into two aliquots, one of which is kept at 0°C (a) and the other is moved to a 37°C water bath for 10 minutes (b).
(6) Finally, all samples are either directly loaded on sucrose gradients (EXP1) or are first crosslinked with formaldehyde and then loaded (EXP2).

Monitoring the polysome integrity of undigested samples revealed a high reproducibility of the results reported by Blobel and Sabatini. Indeed, polysome profiles were hardly affected by nascent chain release at 0°C, confirming that ribosomes did not dissociate upon puromycin treatment (Fig. 15B, panels 1 and 3). Heating at 37°C on the other hand caused disassembly of ribosomes lacking a nascent chain (panel 2). As expected, fixation of heated lysates partially hindered polysome disassembly and strikingly inhibited re-association of ribosome subunits into monosomes (panel 4). Importantly, heating to 37°C alone did not affect polysome integrity (Fig. 15C, left), but only in combination with puromycin treatment (Fig. 15C, right).
The degree of polysome disassembly observed in panel 2 showed that the large majority of nascent chains were released upon puromycin treatment. Indicating a similar efficiency of the puromycin treatment, immuno-detection of puromycilated nascent proteins in the supernatant (S) and ribosome-bound (R) fractions of sucrose gradients confirmed that nascent proteins were released from polysomes to a similar extent, regardless of the incubation temperature (Fig. 15D).
Having established a suitable protocol that allows efficient nascent chain release while keeping polysomes intact, we went on to prepare RNase-digested samples from both non-crosslinked (EXP1) and crosslinked lysates (EXP2) of HEK293-T cells. The sucrose gradient profiles of digested lysates revealed that puromycin caused only a negligible reduction of the disome and increase of the monosome peak (Fig. 15E). This surprising result will be discussed in detail in chapter 3.2.4.
Sequencing of monosomes and disomes isolated from both crosslinked and non-crosslinked lysates revealed that crosslinking reduced the noise of DiSP data caused by the absence of cycloheximide. We therefore selected crosslinked samples (EXP2) for further analysis.

### 3.2.3. Disomes detected by DiSP are nascent chain-connected

To test if disome formation monitored by DiSP is a direct measure of co-co assembly, we performed DiSP of lysates that were either untreated or subjected to nascent chain release by puromycin treatment, and fixed with formaldehyde (EXP2, Fig.15).
As expected, puromycin treatment efficiently suppressed the footprint density shift from monosome to disome observed by DiSP. This was apparent by an almost complete loss of disome-enriched genes (Fig. 16A and B) and by the flattening of both metagene (Fig. 16C) and single gene enrichment profiles (Fig. 16D). Note that only the subpopulation of genes that showed a disome shift in the original DiSP datasets were affected by puromycin treatment (DiSP candidates in Fig. 16B and D), while genes that did not show any disome shift in our original dataset remained unaffected.

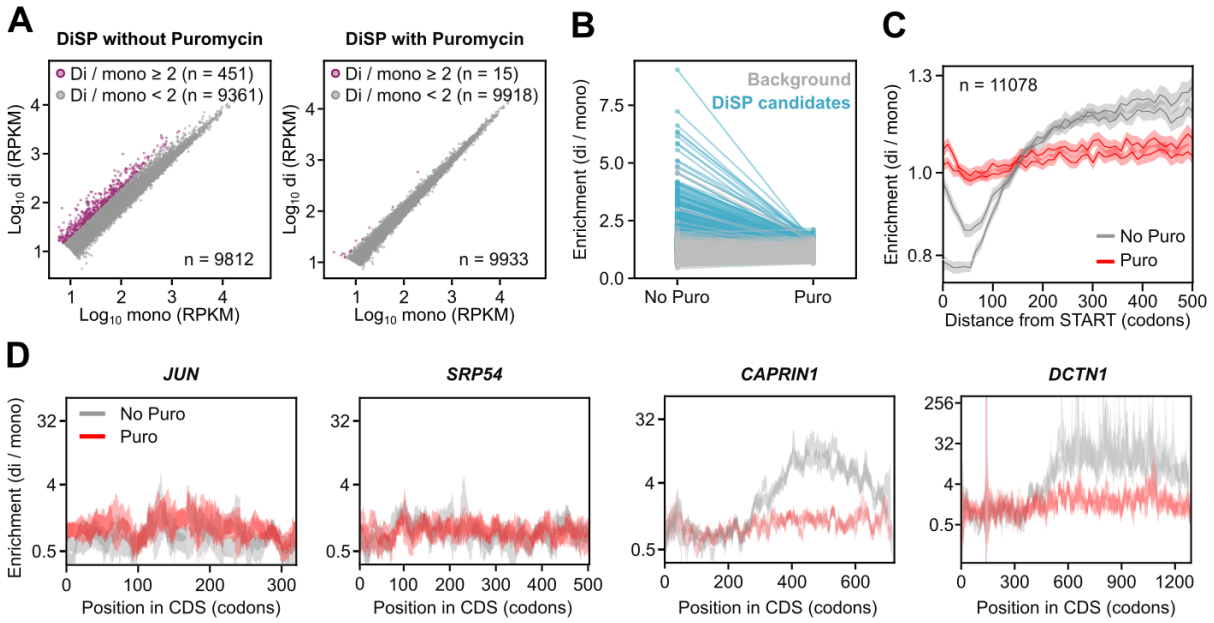



***Figure 17.*** *Nascent chain release by puromycin suppresses disome shifts*

*A) Comparing the monosome and disome densities on each gene (in RPKM) indicates an almost complete loss of disome-enriched genes in puromycin-treated samples.*

*B) Enrichments (disome / monosome) were computed for each gene in DiSP experiments with or without puromycin treatment. Lines connect the same gene across experiments. DiSP candidates defined as genes whose enrichment profile fit best to a sigmoidal shape are highlighted (see chapter 3.3.1).*

*C) Metagene profiles of disome over monosome enrichment aligned to translation start comparing DiSP experiments with or without puromycin treatment. Two biological replicates shown.*

*D) Enrichment profiles (disome / monosome) of DiSP experiments with or without puromycin treatment for four representative genes. Two biological replicates shown.*

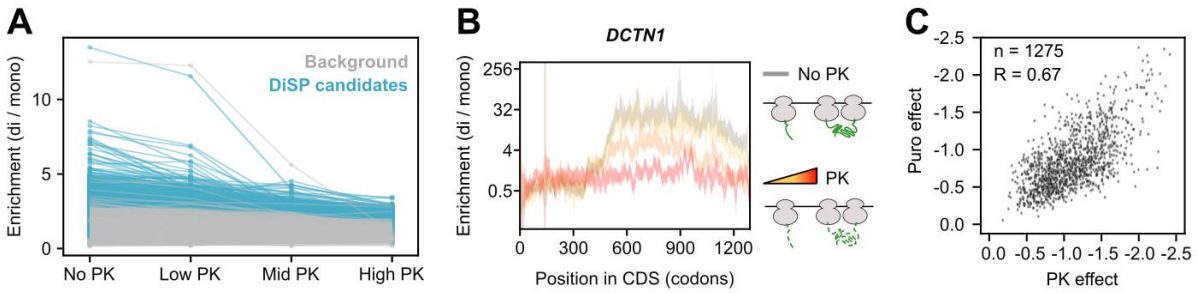



***Figure 16.*** *Limited proteolysis of nascent chains by Proteinase K (PK) treatment suppresses disome shifts*

*A) DiSP was performed on lysates treated with increasing concentrations of PK. Enrichments (disome / monosome) were computed for each gene reliably expressed in all experiments. Lines connect the same gene across experiments. DiSP candidates defined in chapter 3.3.1 are highlighted.*

*B) Enrichment profile (disome / monosome) of DCTN1 in samples treated with increasing PK concentrations. One replicate shown.*

*C) The effect of puromycin and PK treatments, expressed as a coefficient measuring the loss of sigmoidal shape of disome enrichment profiles compared to untreated samples, is correlated (analysis includes only candidates significantly affected by both controls).*

To further support our model of co-co assembly, we performed an additional DiSP experiment relying on nascent chain degradation by limited proteolysis. Specifically, we treated human cell lysates with increasing concentrations of Proteinase K (PK) and monitored the effects on the monosome and disome distributions by DiSP (data shown in detail in the PhD Thesis of Kai Fenzl).

Briefly, this experiment revealed a stepwise decrease of the disome enrichment values across the proteome with increasing PK concentrations (Fig. 17A), exemplified by the gradual flattening of the enrichment profile of *DCTN1*, encoding for the p150glued subunit of the dynactin motor complex (Fig. 17B). Importantly, the loss of disome shifts in PK and puromycin-treated lysates (termed effect in Fig. 17C) is correlated, indicating that nascent chain degradation and release both result in a similar splitting of disomes into monosomes.

Therefore, our proteome-wide controls support the notion that disomes detected by DiSP are connected by nascent chains, in agreement with the model of co-co assembly.

### 3.2.4. How many ribosomes are engaged in co-co assembly?

The results presented in the previous chapter show that disome formation can be effectively suppressed by eliminating the nascent chain connections. Therefore, it seems surprising that the absolute amounts of disomes detected in absorbance profiles of sucrose gradients were hardly affected by puromycin treatment (Fig. 15E).
One explanation for this apparent contradiction is that the disome peak contains at least one additional ribosome population, namely collided disomes caused by pausing and ribosome queuing, in agreement with published data (*47–50*). The close proximity of collided ribosomes hinders RNA cleavage by nucleases, leading to formation of double length (60 nt) footprints, which are not analysed by DiSP. This ribosome population should not be affected by nascent chain degradation or release and several studies suggest that ribosome collisions are frequently happening *in vivo* (*47*, *48*, *50*). While the contribution of collided ribosomes to the disome peak height is difficult to estimate and currently unknown, it should be possible to approximately quantify the fraction of ribosomes that are engaged in co-co assembly.
DiSP provides insights into the efficiency of disome formation (i.e. the fraction of all ribosomes downstream of assembly onsets which are engaged into disomes, see chapter 3.4). Furthermore, standard ribosome profiling, which analyses all ribosomes purified through sucrose cushions (total translatome) allows to quantify the translation levels of co-co assembly candidates. Hence, the fraction of all ribosomes engaged in co-co assembly can be calculated as the sum of ribosome densities in the total translatome downstream of assembly onset corrected for the efficiency of co-co assembly of individual candidates (Equation 1).

$$\text{D)} \qquad P_g = \sum_{i=onset(g)}^{i=offset(g)} TT_{(i)} \ * \ E$$

$$\text{E)} \qquad P = \sum P_g$$

**Equation 1.** *Calculation of the proportion of ribosomes engaged in co-co assembly*

*(i) First, the proportion of co-co assembling disomes is calculated for each gene separately ($P_g$) as the sum of the total translatome read densities ($TT_{(i)}$) at each codon position (i) ranging from the onset to the offset of co-co assembly, multiplied by the efficiency of co-co assembly (E). Note that onset(g) is determined as the inflection point of a sigmoidal shape fitted to the gene's disome enrichment profile and offset(g) is either the length of the gene's coding sequence if the disome enrichment is stable until the end of translation, or the end of co-co assembly if the disome enrichment declines before translation end (see chapter 3.3.1). TT are read densities per codon expressed in reads per million (RPM) multiplied by $10^4$ to obtain percentages, and E is a measure of efficiency ranging from 0 to 1 (see chapter 3.4).*
*(ii) Next, the proportion values of all co-co assembly candidates ($P_g$) are summed to reveal the overall fraction of ribosomes engaged in co-co assembly (P).*

We performed this calculation including all co-co assembly candidates of HEK293-T cells, defined as genes fulfilling the following criteria: (i) the gene's DiSP profile revealed a shift of footprints from monosomes to disomes and (ii) the disome shift was suppressed by both PK and puromycin treatments. According to this calculation, about 3% of all elongating ribosomes are engaged in co-co assembling disomes in human cells.
While this is of course only a rough estimate, it does suggest that puromycin treatment is not expected to substantially reduce the disome peak in sucrose gradients.

## 3.3. High and low confidence classes of co-co assembly proteins

### 3.3.1. A classification pipeline to identify co-co assembly candidates

Aiming to identify the human co-co assembly proteome, we established a bioinformatics pipeline and employed it on DiSP data of human HEK293-T cells. This pipeline employs three filters and results in the identification of a high confidence and a low confidence class of co-co assembly proteins (Fig. 18). The first filter analyses the disome over monosome enrichment distribution of each gene using a sigmoid fitting algorithm (developed by former postdoc Dr. Ilia Kats). This approach allows to categorize proteins into one of three possible types of enrichment, each agreeing with different models of translation-coupled assembly (Fig. 18, (i)). Contrarily to threshold-based analyses that are often employed in genome-wide analyses, our approach relies solely on detection of the shape of DiSP profiles and is not biased by the length or size effect of the disome over monosome enrichment. Accordingly, a single sigmoid shape agrees with models of co-co assembly where the interacting ribosomes remain connected until translation end; this may occur if ribosomes translating two distinct mRNAs terminate translation simultaneously (*in trans*), or if neighbouring ribosomes on one mRNA run in close proximity to each other (*in cis*). An example of single sigmoidal enrichment is *NFKB1*, encoding for the Nuclear factor NF-κ-B p105/p50 subunit (Fig.18, left, top). A double sigmoid shape agrees with models of co-co assembly where the interacting ribosomes disconnect before translation end; this may occur *in trans* if one ribosome in the pair terminates before the other (i.e. during heteromer assembly involving mRNAs of different lengths), or *in cis* if the leading ribosome is distant from the trailing one when it reaches translation end. An example of double sigmoidal enrichment is *DCTN1*, encoding for the $p150^{glued}$ subunit of the dynactin motor complex (Fig.18, left, middle). Finally, a flat line indicates no co-co assembly, yet it doesn't exclude other assembly pathways which are not investigated by DiSP (e.g. post-translational or co-post assembly). An example of the latter is *SOD2*, encoding for mitochondrial Superoxide dismutase (Fig. 18, left, bottom). Of almost 16000 expressed genes, 5250 showed enrichment profiles that fit better to one of the sigmoid shapes compared to the flat line and were classified as DiSP candidates.

The second filter tests whether the shape of a candidate's enrichment profile becomes significantly less sigmoidal upon PK and puromycin treatment (Fig. 18, (ii)).

The third filter aims at distinguishing between translation in the cytosol and across membranes, by examining the mature protein's subcellular localization (Fig. 18, (iii)). Indeed, we cannot formally exclude that nascent proteins in the process of translocation might migrate to the disome fraction in sucrose gradients because of secondary interactions with membrane components of the translocation machinery. Moreover, our additional validation experiments (presented here and in (*115*)) focused on cytosolic or nuclear candidates. Finally, the majority of translocated DiSP candidates are membrane proteins, a category which suffers from poor structural characterization. This complicates the downstream bioinformatics analysis aimed at supporting the authenticity of DiSP candidates. Considering these limitations, we decided to include translocated proteins in the low confidence class. This, however, does not exclude the possibility that inner-membrane proteins frequently assemble via nascent chain interactions as indicated by published work (*72*, *77*, *81*, *83*, *84*) (see also chapters 3.3.4 and 4.1.2).

829 DiSP candidates were sensitive to both controls (ii) and encoded cytoplasmic or nuclear proteins (iii), hence they were classified in the high confidence class. A large number of DiSP candidates (3301) were significantly affected by only one of our controls (ii) or did not encode for cytoplasmic or nuclear proteins (iii) and were classified in the low confidence class.

Importantly, the sigmoid fitting analysis not only allows to identify candidates, but also to mathematically compute the onset of assembly as the inflection point of the sigmoidal shift. This allows to explore the features of nascent chain segments that are ribosome exposed when co-co assembly starts.

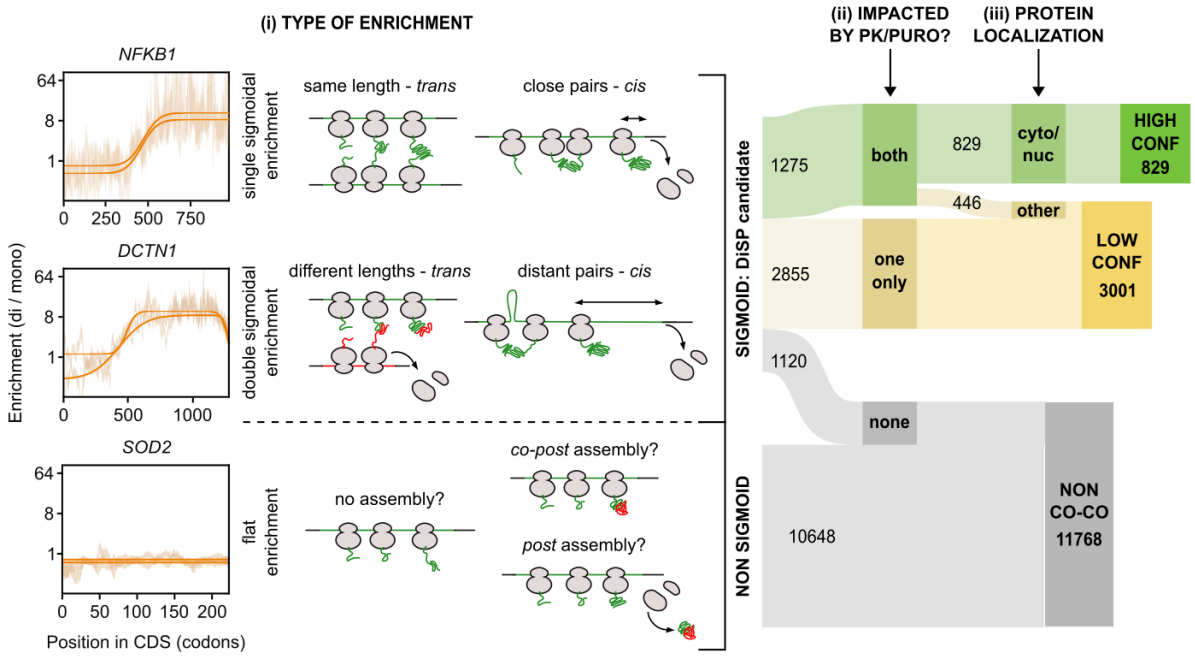



***Figure 18.*** *A bioinformatics pipeline to identify co-co assembly candidates*

*(i) A sigmoid fitting algorithm is employed to identify the shape that fits best to the genes' enrichment profiles (disome / monosome). Different co-co assembly models agree with single and double sigmoidal enrichment profiles, while absence of co-co assembly agrees with a flat enrichment profile. Examples of each type of enrichment (NFKB1, DCTN1 and SOD2) show the disome over monosome enrichment profiles (DiSP data in the background) and the corresponding fitting (solid lines). Two biological replicates shown.*

*(ii) DiSP candidates (showing a sigmoidal shape of disome enrichment) are next filtered based on their response to the PK and puromycin controls: genes impacted by none or only one of the controls are included in the non co-co assembly or low confidence class, respectively; genes impacted by both controls proceed to filter (iii).*

*(iii) Only genes encoding for cytosolic or nuclear proteins are included in the high confidence list.*

### 3.3.2. Co-co assembly is predominantly employed for homomer formation

To explore the features of proteins enclosed in the high and low confidence co-co assembly lists, we first analysed the content of annotated subunits of protein complexes.

The "frequency enrichment" in this analysis was calculated as the ratio between the frequency of complex subunits in each assembly class and their frequency in the background proteome (see absolute and relative numbers in Table 2). Note that since the high confidence class only includes cytosolic and nuclear proteins, the frequency of annotated complex subunits in this class was compared to their frequency in the human "cyto-nuclear" proteome. Conversely, the low confidence class includes proteins regardless of their subcellular localization and was therefore compared to the entire human proteome.

Our analysis revealed that while monomers were disenriched in both co-co assembly classes, subunits of protein complexes were enriched, most strongly among high confidence proteins (Fig. 19A). Importantly, we found that only enrichment of homomeric complex subunits was statistically significant in both assembly classes, while enrichment of heteromers was not, despite a similar size effect of enrichment.

***Table 2.*** *Annotation of subunits of protein complexes in the human co-co assembly proteome*

*Absolute and relative amounts of complex subunits (oligomerization state, as annotated in UniprotKB database) included in different assembly classes. The frequency enrichment (rightmost column of this table and Fig. 19A) is calculated by dividing relative amounts in each assembly class by the relative amounts in their respective human proteome background. NA = not annotated.*

| Assembly class | Oligomerization state | Absolute number | Fraction of assembly class | Frequency Enrichment (as in Fig. 19A) |
|---|---|---|---|---|
| High confidence (cyto / nuc) | Homomer | 254 | 0.296 | 1.453 |
| | Heteromer | 267 | 0.322 | 1.315 |
| | Monomer | 246 | 0.297 | 0.678 |
| | NA | 71 | 0.086 | 0.751 |
| Human proteome (cyto / nuc) for normalization of high confidence class | Homomer | 2060 | 0.203 | |
| | Heteromer | 2480 | 0.245 | |
| | Monomer | 4431 | 0.438 | |
| | NA | 1155 | 0.114 | |
| Low confidence | Homomer | 796 | 0.241 | 1.172 |
| | Heteromer | 819 | 0.248 | 1.186 |
| | Monomer | 1291 | 0.391 | 0.884 |
| | NA | 395 | 0.120 | 0.838 |
| Human proteome for normalization of low confidence class | Homomer | 3270 | 0.206 | |
| | Heteromer | 3326 | 0.209 | |
| | Monomer | 7033 | 0.442 | |
| | NA | 2269 | 0.143 | |

The discrepancy between enrichment values and statistical significance is the result of employing an enrichment test adjusted for the expression bias of our DiSP data. Because the statistical power of our co-co assembly detection method increases as a function of gene expression levels, highly expressed genes have a higher chance of being identified and included in the high or low confidence proteomes (Fig. 19B, compare “Non co-co” with “High conf” and “Low conf”). This bias is also inherently present in all annotation databases, because poorly expressed genes are comparably less likely to be experimentally characterized.

Hence, a significant enrichment of complex subunits may simply reveal the higher likelihood of these proteins to be included in both the annotation databases and our co-co assembly classes. The expression bias is particularly strong for low confidence proteins that are annotated as heteromer subunits (Fig. 19B, compare “Low conf hetero” with “Low conf” and

"Low conf homo"). This explains why the enrichment of heteromers in this class is higher than the homomer enrichment, but still not significant. We next combined our position-resolved assembly information with the analysis of available crystal structures to explore the position of residues constituting the inter-subunit interface of assembled complexes when co-co assembly starts (analysis performed by my co-worker Dr. Frank Tippmann). Our analysis of subunits enclosed in the high confidence list revealed that the onset of co-co assembly often coincided with the synthesis and ribosome exposure of homomeric but not heteromeric interfaces (Fig. 19C).

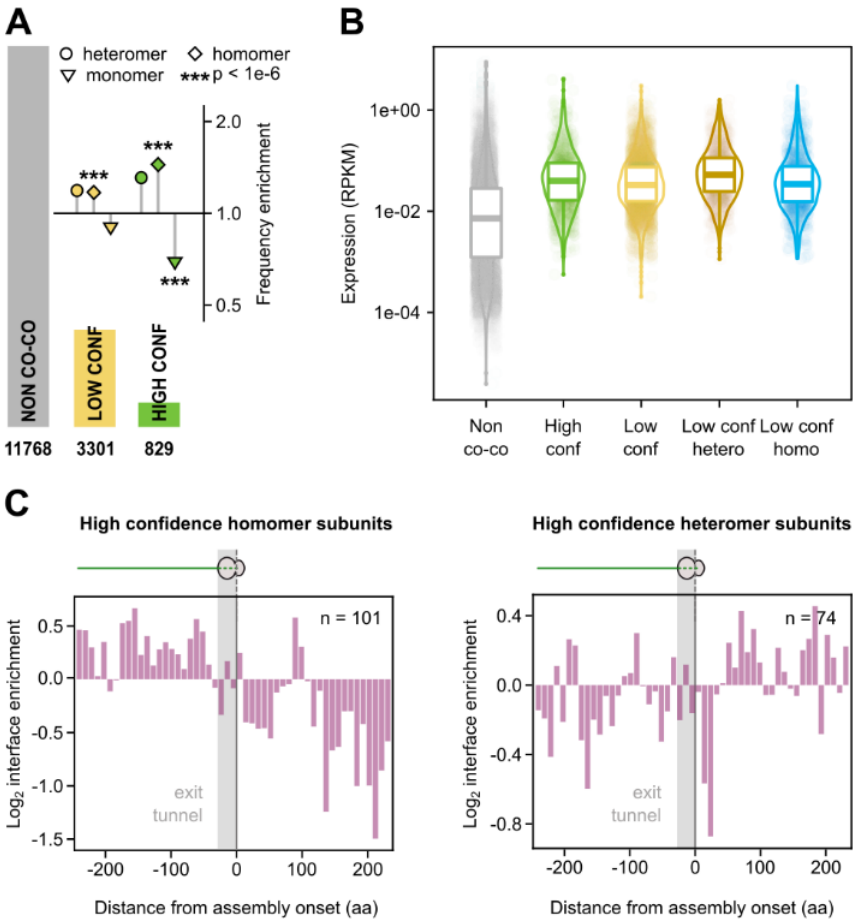



***Figure 19.*** *Homomer formation is a predominant feature of co-co assembly*

*A) The enrichment of annotated complex subunits in the high and low confidence classes is normalized by their enrichment in the respective background proteome (frequency enrichment, see absolute and relative numbers in Table 2). Statistical significance was calculated with an enrichment test adjusted for expression bias (for details see chapter 5.8.5).The bar plot indicates the number of genes in each assembly class.*

*B) Co-co assembly classes are biased towards highly expressed genes (indicated by higher RPKM values in total translatome samples).*

*C) Residues forming the inter-subunit interface of protein complexes were identified by analysis of available crystal structures of high confidence candidates. Metagene enrichment of interface residues aligned to assembly onset is plotted separately for homomeric (left) and heteromeric candidates (right).*

While we do not exclude the possibility that heteromers also co-co assemble, as suggested by previous studies (*71*, *75*, *81*), our enrichment and structural analyses both indicate that co-co assembly is a general mechanism for the formation of mainly homomeric protein complexes.

### 3.3.3. Features of the nascent chain segments mediating co-co assembly

#### 3.3.3.1. Co-co assembly largely employs N-terminal interaction interfaces

Indicating that co-co assembly interactions are most often established early during translation, 576 of 829 assembly onsets were detected in the 5' halves of transcripts encoding for high confidence proteins (Fig. 20A).
Consistently, analysis of available crystal structures revealed that homomer interfaces were enriched in N-terminal halves of high confidence candidates (Fig. 20B, left), while the opposite was true for the rest of the human proteome (Fig. 20B, right). These results are in agreement with a previous study and suggest the existence of evolutionary pressure to co-co assemble (for discussion, see chapter 4.2.3).

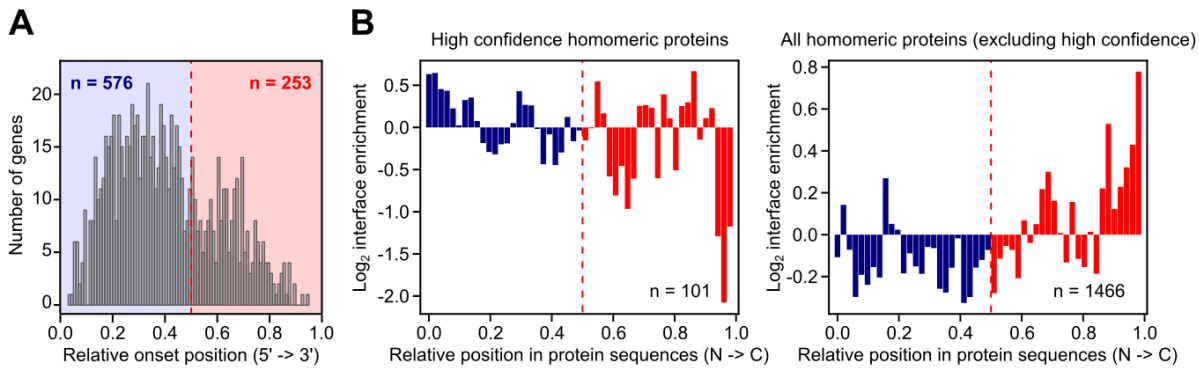



***Figure 20.*** *Initiation of co-co assembly correlates with exposure of N-terminal dimerization interfaces.*

*A) Relative distribution of co-co assembly onset positions along the coding sequence of high confidence genes. The red dashed line separates the 5' and 3' halves of the genes' coding sequences.*
*B) Residues forming the inter-subunit interface of protein complexes were identified by analysis of available crystal structures of high confidence candidates. The relative enrichment distribution of interface residues is plotted separately for homomeric (left) and heteromeric candidates (right). The red dashed line separates the N-terminal and C-terminal halves of proteins.*

#### 3.3.3.2. Five major dimerization domains mediate co-co assembly

We next asked what structural motives evolved to mediate co-co assembly. Specifically, we aimed to identify protein domains that were significantly more prevalent within the ribosome-exposed nascent chain segments at the assembly onset of high confidence candidates, compared to their general frequency in the human proteome. Because co-co assembly mostly starts at early translation stages (Fig. 20A), a simple comparison to the frequency of

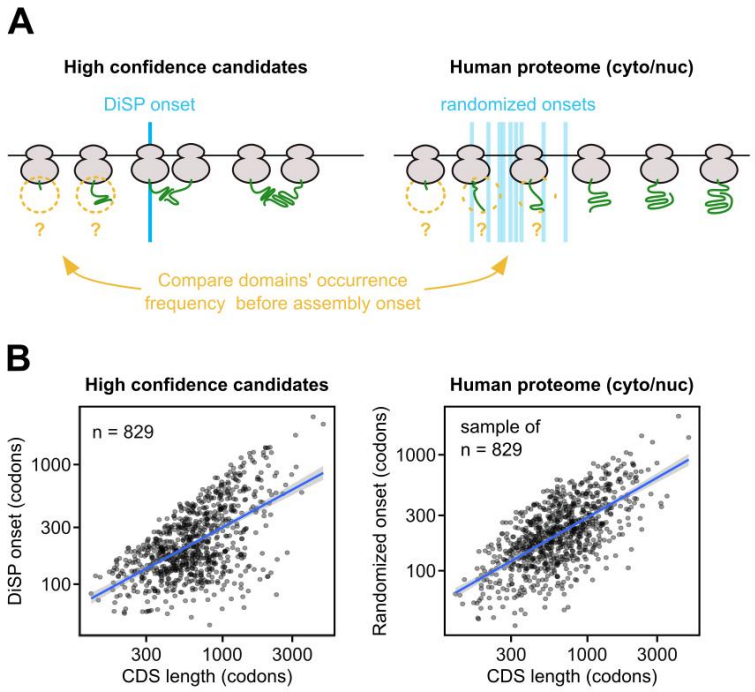



***Figure 21.*** *Enrichment analysis of protein domains exposed at assembly onset*

*A) The frequency of annotated domains in nascent chain segments exposed at the onset of assembly are compared to their frequency in similar N-terminal segments of the background proteome, by assigning randomized onsets to all human cyto-nuclear proteins.*

*B) A log-log linear relationship exists between CDS length and onset positions of high confidence candidates (left, n = 829). Randomized onsets are drawn for a sample of 829 cyto-nuclear genes based on this relationship.*

domains across all full-length human proteins would be biased towards detection of domains that are generally found at the N-terminus of proteins. To allow a fair comparison between N-terminal portions of high confidence candidates and similar N-terminal regions of the background proteome, we bioinformatically computed “randomized assembly onsets” for all cyto-nuclear human proteins (Fig. 21A). Conceptually, our calculation of randomized onsets aimed at mirroring the correlation existing in the high confidence class between protein lengths and onset positions (Fig. 21B, left): a new onset for a protein of length x was drawn from the distribution of onsets of high confidence proteins with similar lengths (Fig. 21B, right).

Analysis of protein domains annotated in the UniprotKB (knowledge-based) database identified seven domain clusters that were significantly enriched within the exposed nascent chain segments of high confidence proteins at the onset of co-co assembly (color-coded in Fig. 22A), among which five were highly conserved, well established homodimerization domains. Coiled coils were the most prevalent co-co assembly interface, exposed by 193 out of 829 high confidence proteins according to UniprotKB (Fig. 22A and B, left). Our analysis identified seven additional domains which are generally found N-terminal to the conserved coiled coils of kinesins, myosins and AGC-kinases and whose function was unrelated to protein dimerization. Since co-co assembly onset generally coincided with partial or complete exposure of the coiled coil rather than the N-terminal domains, we considered these as part of the coiled coil cluster (color-coded in orange, Fig. 22A). Partial synthesis of coiled coil interfaces was often sufficient to induce co-co assembly (Fig. 22B, left). Given the highly repetitive nature of both primary and secondary structures of coiled coils, and the possibility for α-helices to start folding inside the ribosomal exit tunnel, it is not surprising that partial exposure of coiled coils would often be sufficient to establish co-co assembly interactions.

Similar to coiled coils, BAR domains (named after Bin, Amphiphysin and Rvs) were also mostly partially exposed at assembly onset (Fig. 22B, right). BAR are conserved dimerization units consisting of three (classical BAR) to five (F-BAR) bent antiparallel α-helices.

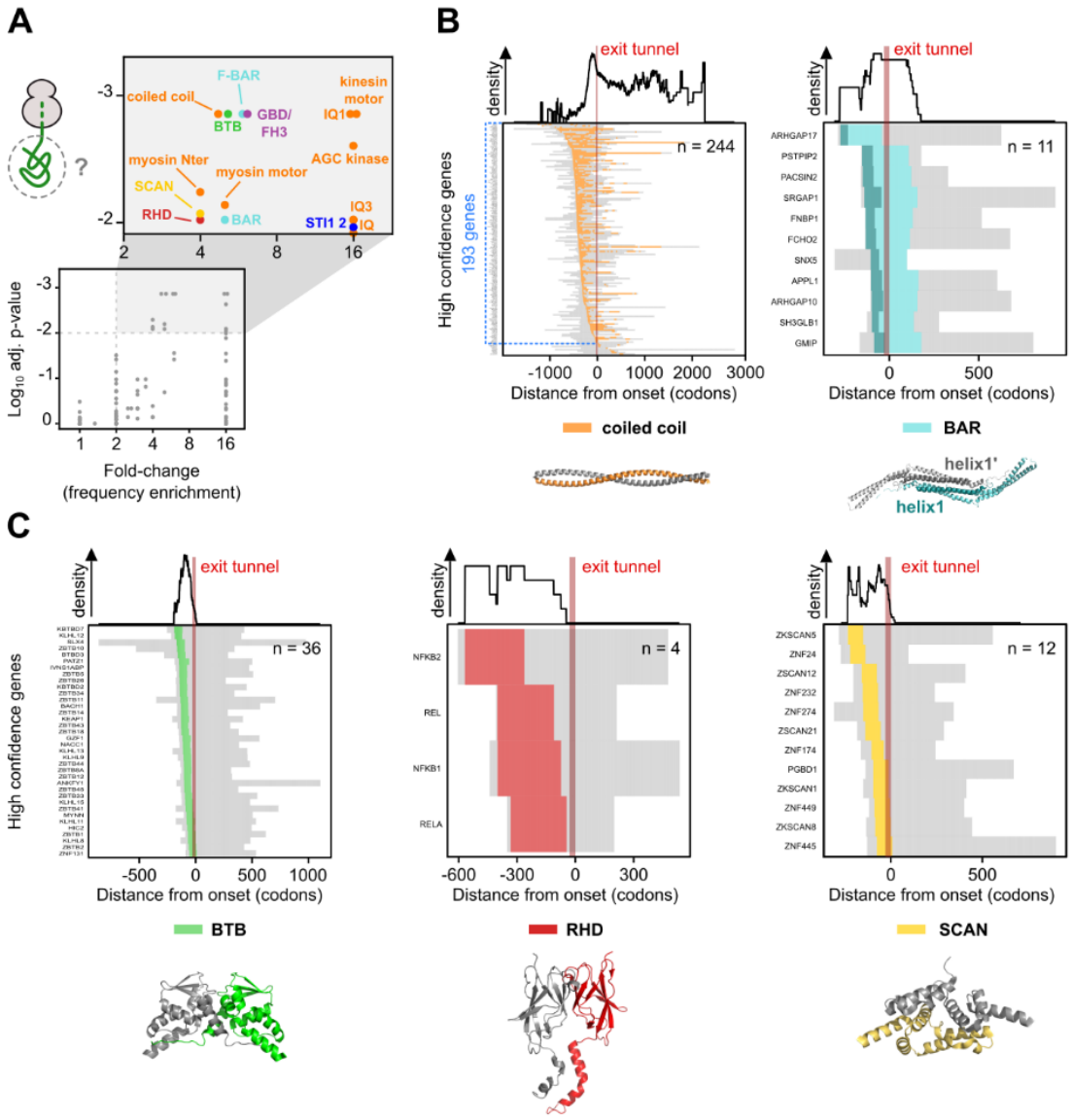



***Figure 22.*** *Onset of co-co assembly correlates with exposure of conserved dimerization folds*

*A) The frequency of annotated protein domains on nascent chain segments exposed at the onset of assembly by high confidence candidates was normalized by their frequency in N-terminal regions of the background proteome (fold-change, see Fig. 21). Significance of enrichment was calculated using a Monte-Carlo simulation and p-values adjusted for multiple testing were plotted against the fold-change (frequency enrichment) for each domain. Domains with p-adj. ≥ 0.01 and fold-change ≥ 2 are highlighted and were analysed in detail.*

*B) Analysis of domains that were only partially exposed at assembly onset. Heatmaps and domain density profiles of high confidence candidates containing coiled coil (left) and BAR domains (right) aligned to assembly onsets. The subset of genes highlighted in blue exposes a coiled coil domain upstream of assembly onset. Dark green residues are involved in formation of helix1 of BAR domains (highlighted in the heatmap and representative structure). Representative structures are PDB: 1D7M, 3Q0K.*

*C) Analysis of domains that were completely exposed at assembly onset. Heatmaps and domain density profiles of high confidence candidates containing BTB (left), RHD (middle) and SCAN domains (right) aligned to assembly onset. Representative structures are PDB: 1BUO, 1K3Z, 3LHR.*

BAR domains play a fundamental role in a variety of cellular processes by binding membranes and inducing their curvature (*116*). A closer analysis revealed that assembly minimally depended on exposure of the most N-terminal α-helix (helix1, Fig. 22B, right), which needs to be fully translated to contact helix1' on its partner subunit in an antiparallel fashion. The only exception was *ARHGAP17*, encoding for the Rho GTPase-activating protein 17, for which assembly onset closely followed the emergence of the complete BAR domain. While the structure of this protein is not resolved, prediction of the homodimeric structure with Swissmodel (*117*) indicated a "head-to-head" interaction mode, characterized by scattered contacts involving the whole domain's primary sequence, suggesting that folding of the complete BAR motif was required to create the assembly interface.

The remaining dimerization motives included BTB (Broad-Complex, Tramtrack and Bric a brac), RHD (Rel Homology Domain) and SCAN (SRE-ZBP, CTfin51, AW-1 and Number 18 cDNA). These domains are all globular and were completely ribosome-exposed at assembly onset (Fig. 22C), suggesting that they were presumably folded prior to assembly onset, a feature previously reported for co-post assembly (*66*). BTB are conserved and recurrent dimerization domains located at the N-termini of many transcription factors and actin-associated proteins and mostly employed for homomer formation. The BTB dimer forms a highly intertwined structure which in most cases involves the formation of a two-stranded β-sheet between each monomer's N-terminus and the other chain's C-terminus. The BTB domain was exposed by 36 high confidence proteins at the time point of assembly (Fig. 22C, left). Visual inspection of the enrichment profiles of the few candidates that appeared to not fully expose the BTB domain at assembly onset showed that this was very likely due to imprecise onset assignment, strongly implying that this dimerization domain was always fully exposed at the onset of assembly.

RHDs are found in a family of eukaryotic transcription factors, including nuclear factor kappa-B (NF-kB) complexes. Previous studies indicated that proteins encoded by *NFKB1* interact cotranslationally on a single polysome (*in cis*) and that early assembly interactions are required for native biogenesis of the p50 transcription factor (*67*, *80*). All genes encoding for NF-kB subunits were included in the high confidence list, except for *RELB*, which was however only excluded because of its poor expression levels in HEK293-T cells (Fig. 22C, middle). In all cases, disome formation occurred upon full exposure of the immunoglobulin-like fold positioned at the C-terminus of the RHD (Fig. 22C, middle), which creates the interface for both homo- and hetero-dimerization of RHD-containing proteins.

Among high confidence candidates were also 12 transcription factors that exposed the SCAN domain at assembly onset (Fig. 22C, right). SCANs are leucine rich dimerization folds found at the N-termini of several C2H2 zinc finger proteins. A SCAN monomer is composed of five packed α-helices, and can mediate both homo- and heteromeric interactions (*118*). In agreement with the requirement to fully expose this dimerization fold prior to co-co assembly, the SCAN dimer interface is formed by intercalating helix 2 of one chain between helices 3 and 5 of the opposing chain.

Lastly, our analysis revealed two significantly enriched, yet less characterized domains (Fig. 23). The first were STI1 repeats, found in ubiquilin proteins. Two out of four human ubiquilins were high confidence candidates (*UBQLN1* and *UBQLN2*), exposing the complete STI1 repeat 2 at the onset of assembly (Fig. 23, top). In agreement with this finding, STI1 repeats have been shown to mediate both homo- and hetero-dimerization of ubiquilin 1 and 2 (Ford et al). The second were GBD/FH3, N-terminal regulatory elements conserved across eukaryotes and found in formins, a class of proteins involved in the organization of the actin cytoskeleton. Previous studies revealed that the mouse homologue of human *DIAPH1* forms

stable homodimers through its FH3 domain segment (*119*). Our DiSP data revealed that full exposure of this domain correlated with disome formation in six high confidence candidates (Fig. 23, bottom), suggesting that human formins may homodimerize through co-co assembly of the FH3 domain.

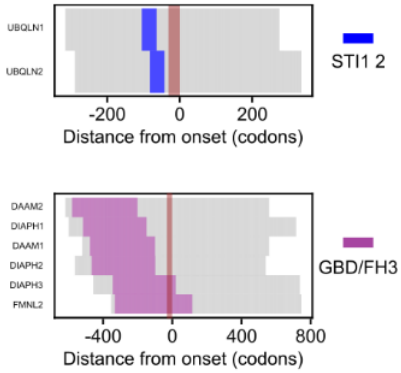



***Figure 23.*** *Correlation of co-co assembly with exposure of less characterized protein folds*

*Heatmaps of STI1 (top) and GBD/FH3 (bottom) domains aligned to assembly onset of high confidence candidates.*

#### 3.3.3.3. Coiled coils: the most widely employed co-co assembly domains

Among annotated protein domains mediating co-co assembly, coiled coils were the most prevalent in the high confidence class (Fig. 22A and B, left). To determine the relative enrichment of coiled coils among co-co assembly candidates, we analysed and compared the occurrence and position of this dimerization motif in cyto-nuclear proteins of the non co-co assembly class with randomized onset positions (Fig. 24A). While coiled coils were annotated in almost 30% of high confidence proteins and exposed by the majority of them (23%) at the assembly onset, this domain was only present in about 13% of non co-co assembly proteins, and exposed by less than 6% (Fig. 24B). In particular, the frequency of coiled coils was highest immediately upstream of assembly onset in high confidence proteins (density profile in Fig. 22B, left), while this correlation did not exist for non co-co assembly proteins (density profile in Fig. 24A), indicating that coiled coil exposure is a specific feature mediating co-co assembly.

The number of exposed coiled coil residues required for co-co assembly was variable (111 in median in the high confidence class), suggesting different propensities to form stable dimers (Fig. 24C). This number was much lower among the randomized onset positions of non co-co assembly proteins (65 residues in median, Fig. 24C), suggesting that the enrichment of exposed coiled coil residues in co-co assembling nascent chains is not observed by chance.

We next employed the DeepCoil prediction tool to reveal additional protein segments which may fold into coiled coils (*120*). This analysis extended the number of high confidence proteins containing this dimerization motif to 609, of which 408 exposed a predicted coiled coil at the assembly onset, suggesting that almost 50% of high confidence candidates could employ this fold for co-co assembly (Fig. 24D, left and E, left). As a comparison, only 23% of non co-co assembly proteins exposed a predicted coiled coil segment upstream of randomized assembly onsets (Fig. 24E, right). Again, analysis of the domain density distribution confirmed that exposure of predicted coiled coils correlated with assembly onsets of high confidence proteins, but was unrelated to randomized onsets of non co-co assembly proteins (Fig. 24D).

Together, our database and prediction analyses indicate that coiled coils are highly prevalent dimerization interfaces employed for co-co assembly.

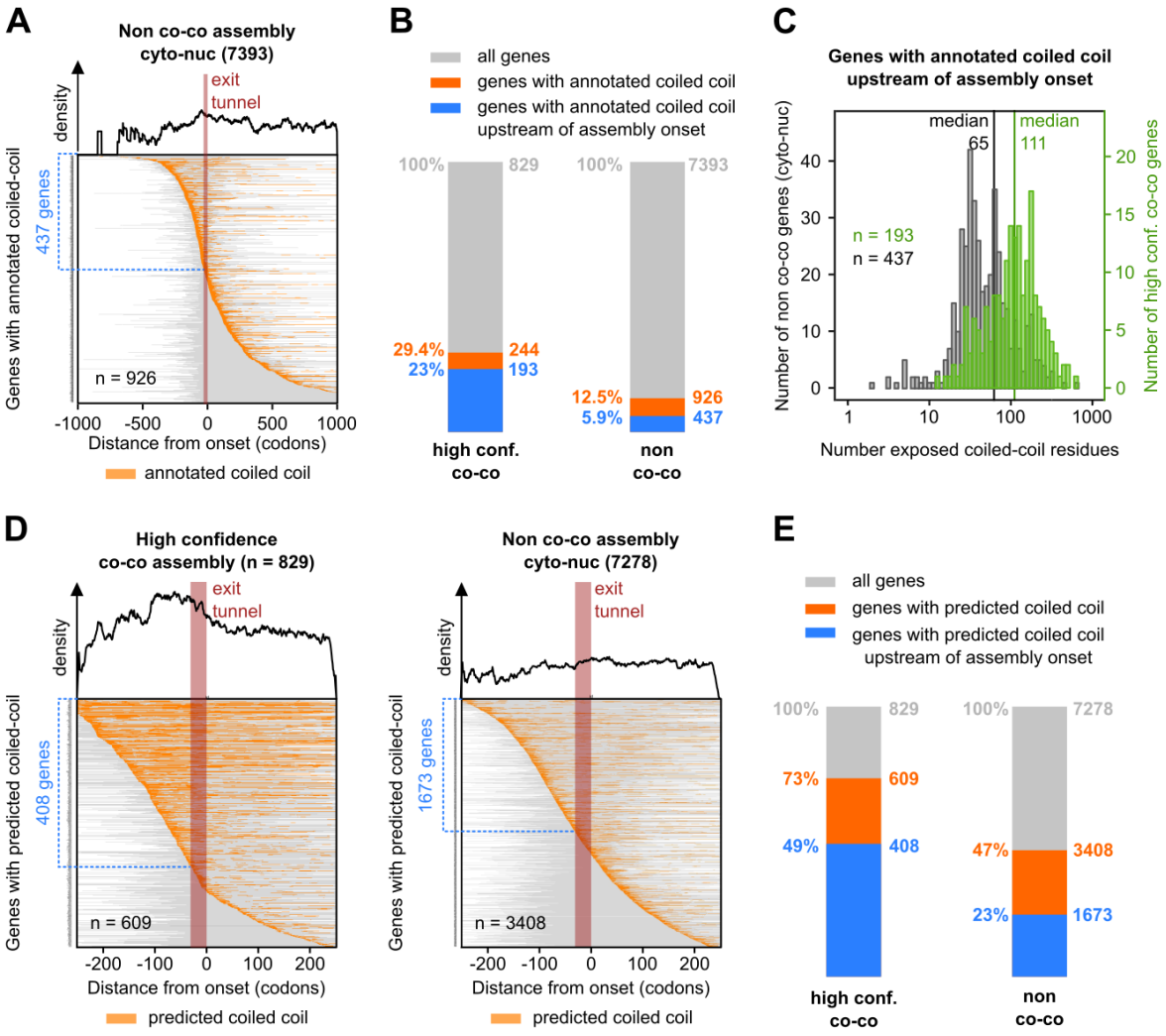



***Figure 24.*** *Coiled coils are a specific and predominant feature of co-co assembly nascent chains*

*A) Heatmap and domain density profile of cyto-nuclear proteins of the non co-co assembly class containing annotated coiled coils and aligned to randomized assembly onsets (as illustrated in Fig. 21). The subset of genes highlighted in blue exposes a coiled coil domain upstream of assembly onset.*

*B) Absolute and relative quantification of genes with annotated coiled coils (orange) and genes exposing a coiled coil segment before assembly onset (blue) in the high confidence and non co-co assembly cyto-nuclear classes.*

*C) Distribution of the number of coiled coil residues on the ribosome-exposed nascent chains at assembly onset. High confidence proteins with annotated coiled coils upstream of assembly onset are included in the analysis.*

*D) Heatmaps and domain density profiles of proteins containing predicted coiled coils (according to the DeepCoil algorithm). High confidence proteins aligned to DiSP onsets and non co-co assembly cyto-nuclear proteins aligned to randomized onsets are analysed. The subset of genes highlighted in blue exposes a coiled coil domain upstream of assembly onset.*

*E) Absolute and relative quantification of genes with predicted coiled coils (orange) and genes exposing a predicted coiled coil segment before assembly onset (blue) in the high confidence and non co-co assembly cyto-nuclear classes.*

### 3.3.4. Features of low confidence co-co assembly proteins

For 3301 DiSP candidates it was less clear whether migration into the disome fraction was caused by co-co assembly interactions. These proteins were not excluded from analysis, but were assigned to a low confidence class for separate investigation. Importantly, also this class was enriched with subunits of protein complexes (Fig. 19A) and very likely includes a number of authentic co-co assembly proteins.

The large majority of low confidence candidates (2855) did not classify as high confidence because they were at least partially resistant to one of our DiSP control experiments, e.g. their disome over monosome enrichment profile did not become considerably less sigmoidal upon PK or puromycin treatment. Among them, 2757 were sensitive only to PK and 98 were sensitive only to puromycin (Fig. 25A).

We first aimed to investigate why so many proteins were sensitive to PK but resistant to puromycin, by comparing the general features of this subgroup (puro resistant, Fig. 25B, left) to the puromycin sensitive group (puro sensitive). No clear differences could be found by analysing the annotation of complex subunits and subcellular localization (Fig. 25B, right). Conversely, metagene analyses of the disome over monosome enrichment aligned to assembly onset revealed that puromycin resistant genes had in average less pronounced disome shifts compared to puromycin sensitive genes already in the untreated samples of the puromycin datasets (grey profiles, Fig. 25C). Importantly, puromycin resistant genes had generally poorer scores of disome shifts (low CI) also in the main DiSP dataset of HEK293-T cells (Fig. 25D) and were on average less expressed (Fig. 25E).

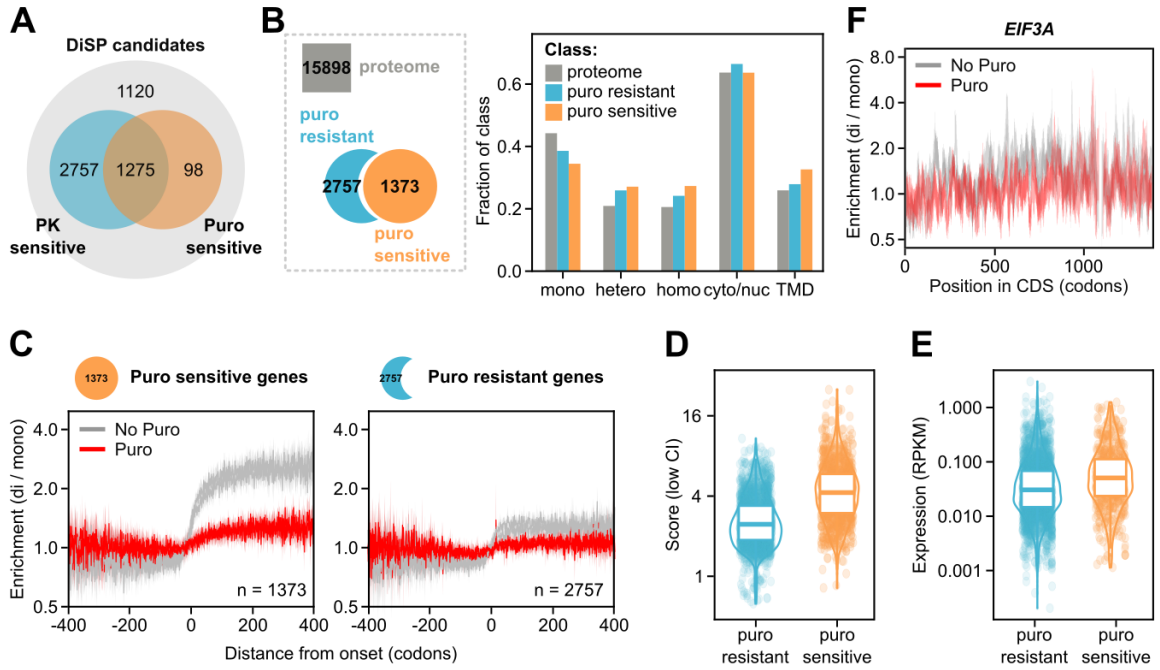



***Figure 25.*** *Features of puromycin-resistant low confidence candidates*

*A) DiSP candidates (defined as genes with a sigmoidal enrichment profile) were defined as PK or puromycin sensitive if their enrichment became substantially less sigmoidal in the respective control experiments. DiSP candidates that were sensitive to only one of the controls were classified as low confidence.*

*B) Proteins that were sensitive to PK but not to puromycin (puro resistant) were compared to proteins that were sensitive to puromycin (puro sensitive) or to the general proteome. The frequency of general features (annotation of complex composition and subcellular localization) within each class was calculated.*

*C) Metagene profiles of disome over monosome enrichment aligned to assembly onset comparing DiSP experiments with or without puromycin treatment. Puromycin sensitive (left) and resistant (right) genes are analysed separately. Two biological replicates shown.*

*D) The magnitude of disome shifts for each gene can be measured by analysing the lower bounds of codon-resolved enrichment confidence intervals (low CI), i.e. the lower limits of vertical bars visualised in single gene enrichment profiles, such as the one in panel F of this figure, and selecting the highest of them. The scores obtained in this way are a reliable approximation for the size effect of disome enrichment which account for sequencing noise. Low CI values of genes included in the puro resistant and puro sensitive groups are compared.*

*E) Expression is measured as ribosome densities per gene derived from total translatome experiments in HEK293-T cells (expressed in RPKM). Genes included in the puro resistant and puro sensitive groups are compared.*

*F) Representative example of puromycin resistant gene (EIF3A). Enrichment profiles (disome / monosome) of DiSP experiments with or without puromycin treatment. Two biological replicates shown.*

While the metagene analyses suggested that puromycin treatment did suppress the averaged disome shift of resistant candidates (Fig. 25C, right), this effect was much less clear for individual gene profiles, exemplified by *EIF3A* (Fig. 25F). Hence, the significance of the disome shift reduction was limited for individual genes with less clear sigmoidal profiles and lower sequence coverage, which ultimately hindered their identification as high confidence co-co assembly candidates.

We next analysed the 98 proteins that were sensitive to the puromycin but not to the PK control (Fig. 26A, left). Among them, 46 contained transmembrane domains (TMD), indicating that the frequency of integral membrane proteins in this class was about double compared to their frequency in the human proteome (Fig. 26A, right). A closer analysis revealed that most of them were short, multi-spanning transmembrane proteins, suggesting that nascent chain accessibility to the protease was likely hindered by ribosome docking to

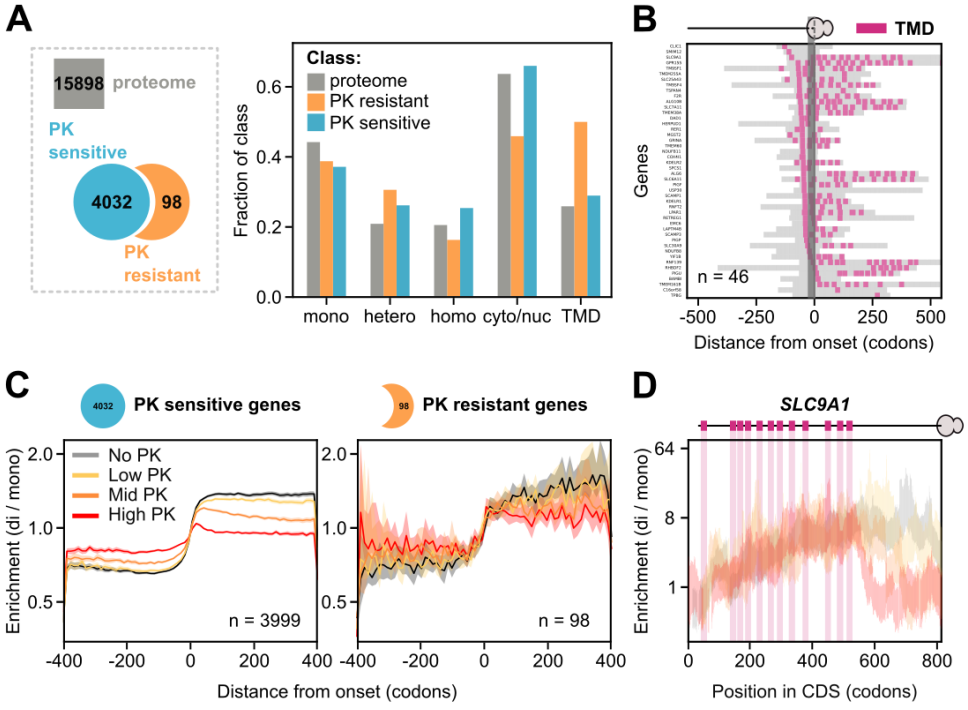


***Figure 26.*** *Features of PK-resistant low confidence candidates*

*A) Proteins that were sensitive to puromycin but not to PK (PK resistant) were compared to proteins that were sensitive to PK (PK sensitive) and to the general proteome. The frequency of general features (annotation of complex composition and subcellular localization) within each class was calculated.*

*B) Heatmap of PK-resistant proteins containing annotated transmembrane domains (TMD) aligned to assembly onset. The ribosome exit tunnel is depicted as a grey bar covering a 30 amino acids-long C-terminal segment of nascent proteins at the time point of assembly onset.*

*C) Metagene profiles of disome over monosome enrichment aligned to assembly onset comparing DiSP samples treated with increasing PK concentrations. PK sensitive (left) and resistant (right) genes are analysed separately. One replicate shown.*

*D) Representative example of PK resistant gene (SLC9A1). Enrichment profiles (disome / monosome) of DiSP samples treated with increasing PK concentrations. Purple boxes and vertical bars indicate exposure of annotated TMD domains dueing translation. One replicate shown.*

the translocon for co-translational insertion into the membrane (Fig. 26B). In most instances, ribosomes translating this group of proteins shifted to the disome fraction closely after the first TMD had emerged out of the ribosome exit tunnel, suggesting co-translational interactions between neighbouring TMDs.

Contrarily to puromycin resistant candidates, disome formation on PK resistant genes mostly persisted with increasing PK concentrations, as revealed by enrichment metagene profiles aligned to assembly onset (Fig. 26C). The average disome enrichment was completely unaffected by PK treatment immediately after the onset of assembly, but was moderately reduced shortly after, indicating a higher accessibility of PK to nascent chains after TMD insertion. One example is *SLC9A1*, encoding for a subunit of a sodium/hydrogen membrane transporter. Disome enrichment on this gene was not affected by the PK treatment during translation of its 12 TMDs; only after exposure of the last transmembrane region, a PK-dependent loss of disome footprints was observed (Fig. 26D).

Besides genes that failed to respond to either of our controls, the low confidence class

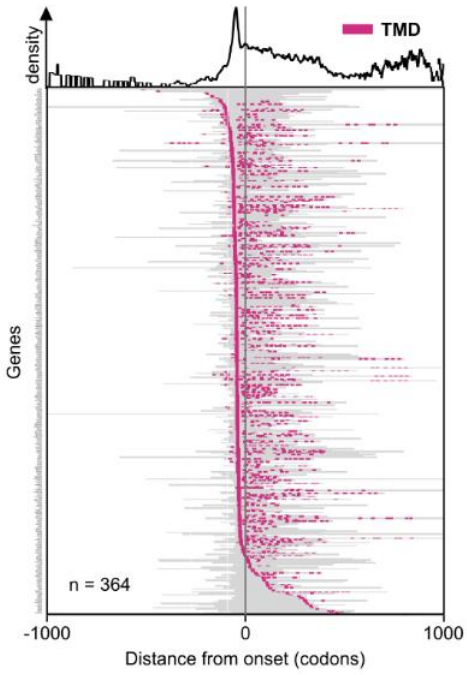



***Figure 27.*** *Transmembrane proteins in the low confidence class often expose a TMD right before assembly onset*

*Heatmap showing the position of annotated TMDs aligned to the onset of assembly, including all low confidence transmembrane proteins. Corresponding domain density profile aligned on top. Proteins in the heatmap are arranged based on the distance between onset and the closest TMD to onset.*

includes 1404 DiSP candidates that have a final subcellular localization other than the cytosol or the nucleus, among which 364 are inserted into the membrane. Analysis of TMD positions aligned to the onset of disome formation confirmed that co-co assembly mostly initiated shortly after emergence of an N-terminal TMD, suggesting interactions of two TMDs shortly after their synthesis and lateral insertion into the ER membrane (Fig. 27).

### 3.4. Efficiency of co-co assembly

DiSP revealed hundreds of high confidence and thousands of low confidence co-co assembly candidates. However, it remains unclear what fraction of the ribosomes translating the same mRNA molecule are engaged in co-co assembly interactions after assembly onset (from hereon named efficiency of co-co assembly).

The extent of disome enrichment is not a good measure of efficiency for two main reasons: (i) it suffers from the inevitable and fluctuating contamination from the more prominent monosome peak and (ii) it is defined by the number of co-co assembly candidates (i.e. the higher the number of candidates, the lower the disome enrichment on each of them). Therefore, analysis of disome enrichments can only be employed for qualitative claims, namely to reveal the identity and onset position of co-co assembly candidates.

Since the majority of elongating ribosomes are in the monosome state, this fraction is much less influenced by the variables mentioned above. Metagene analysis including high confidence genes aligned to the assembly onset revealed a substantial reduction of monosome footprints which was strongly correlated with disome formation and largely independent of the overall ribosome density in this region (total translatome, Fig. 28A). This confirmed that monosome depletion was caused by a shift of ribosomes to the disome fraction. Hence, the extent of monosome depletion downstream of assembly onset relative to

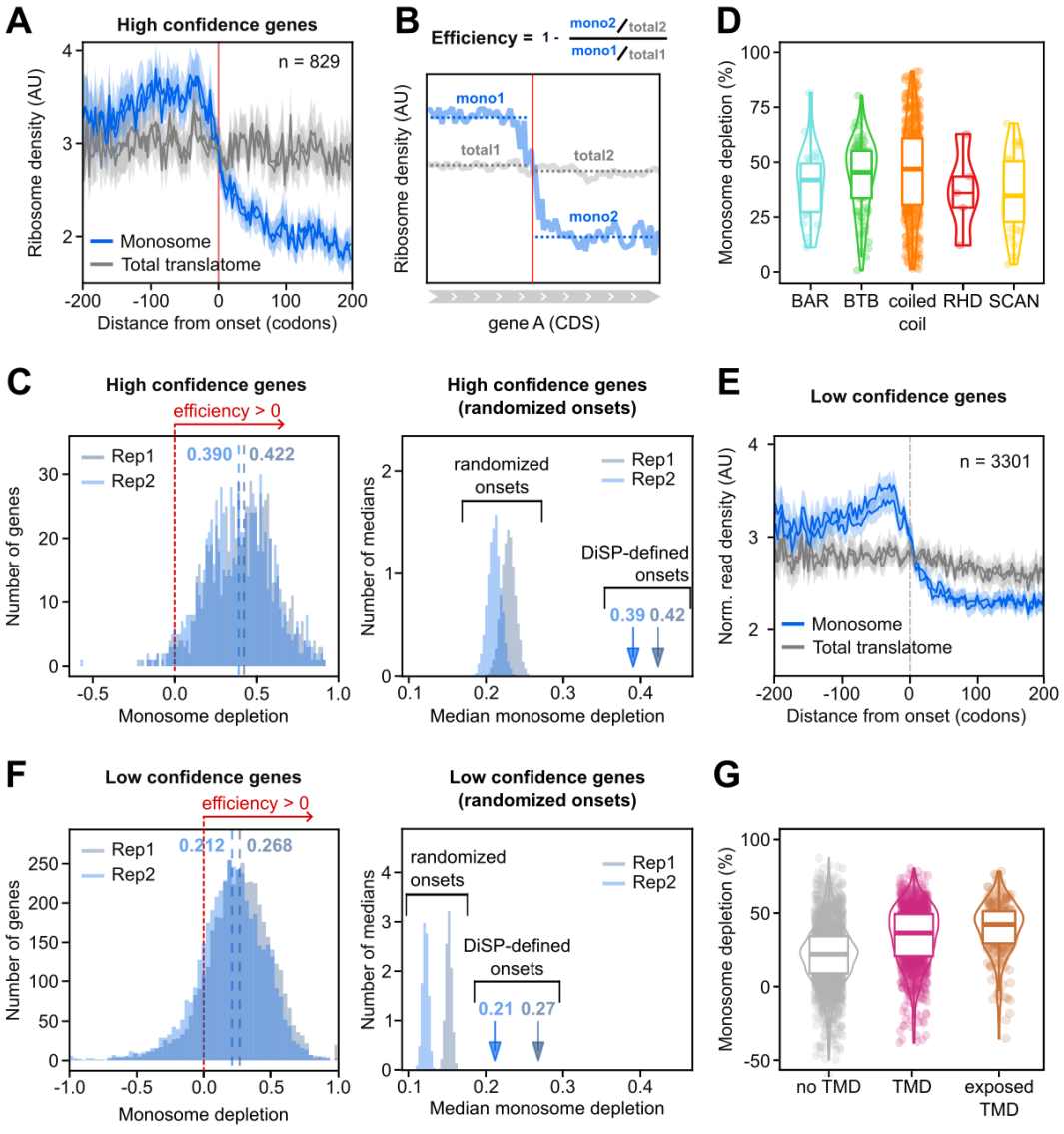


***Figure 28.*** *Monosome depletion after onset reveals the co-co assembly efficiency of each gene*

*A) Metagene analysis of high confidence candidates showing the footprint densities in the monosome fraction and the total translatome aligned to assembly onset.*

*B) Schematic illustration for the calculation of monosome depletion (efficiency) on a single gene profile. First, the residual monosome is calculated by comparing the monosome density after onset (mono2) to the monosome density before onset (mono1) normalized by the corresponding densities in the total translatome (total 2 and 1). Note that "mono" and "total" are sum of RPM values before or after assembly onset. Second, monosome depletion is calculated as the complement of the residual monosome (Efficiency = 1 – residual monosome).*

*C) (Left) Monosome depletion is quantified for each high confidence gene separately as illustrated in panel B. Median monosome depletions of two biological replicates are indicated by blue dashed lines. Depletion values ranging between 0 and 1 are directly proportional to varying co-co assembly efficiencies. (Right) Distribution of median depletions obtained in $10^5$ iterations of onsets scrambling across high confidence genes. Median monosome depletions obtained using DiSP onsets of high confidence candidates are indicated by blue arrows.*

*D) Monosome depletion of high confidence candidates that employ different dimerization domains for co-co assembly. Depletion values are averages of two biological replicates.*

*E) Metagene analysis of low confidence candidates showing the footprint densities in the monosome fraction and the total translatome aligned to assembly onset.*

*F) (Left) Monosome depletion is quantified for each low confidence gene separately as illustrated in panel B. Median monosome depletions of two biological replicates are indicated by blue dashed lines. Depletion values ranging between 0 and 1 are directly proportional to varying co-co assembly efficiencies. (Right) Distribution of median depletions obtained in $10^5$ iterations of onsets scrambling across low confidence genes. Median monosome depletions obtained using DiSP onsets of low confidence candidates are indicated by blue arrows.*

*G) Monosome depletion values are compared for low confidence candidates that either do not contain annotated TMDs, contain at least one TMD or expose a TMD before initiation of co-co assembly. Depletion values are averages of two biological replicates.*

the total translatome should approximately reflect the fraction of ribosomes that were shifted to disomes (Fig. 28B).

We went on to calculate efficiency values of individual high confidence candidates (analysis developed by formed postdoc Dr. Ilia Kats). The median efficiency of co-co assembly was about 40%, and for some proteins even exceeded 90%, indicating that the majority of these proteins' nascent chains assembled co-translationally (Fig. 28C, left). As a control, assembly onsets were scrambled $10^5$ times among high confidence candidates and the median monosome depletion was calculated for each iteration (Fig. 28C, right). None of the medians from random sampling was equal or higher than the median efficiencies from DiSP data, indicating that the observed monosome depletion was tightly dependent on DiSP onset positions and was not observed by chance (i.e. statistically significant).

The measured monosome depletion values were very variable. It is reasonable to assume that proteins with high efficiency employ co-co assembly as the main route to complex formation, while proteins with lower efficiencies may partially form by co-post and post-translational assembly inside cells, which we do not test by DiSP. However, monosome depletion detected by DiSP likely provides an underestimation of the *in vivo* efficiency of co-co assembly. The reasons are (i) the smaller yet inevitable contamination of the monosome

fraction with disomes and (ii) the partial dissociation of disomes into monosomes that may occur during the ribosome purification procedure of DiSP. Therefore, the extent of monosome depletion could also partially report on the stability of the nascent dimer interaction. To investigate this question, we analysed monosome depletion levels conferred by the five major co-co assembly dimerization domains (Fig. 28D). Coiled coils generally conferred the highest, although very variable stability to the co-co assembly interactions, followed by BTB, BAR, RHD and SCAN interfaces. Interestingly, the three proteins with the strongest co-co assembly efficiencies (over 90%, namely *TPR*, *CLIP1* and *EEA1*), were all characterized by extremely long coiled coil domains, suggesting high stability of the nascent chain interactions (1000-1500 amino acids, compared to a median coiled coil length of 66 amino acids in the cellular proteome).

Although less pronounced, a certain level of monosome reduction was observed also for many candidates of the low confidence list, with a median depletion of about 24% (Fig. 28E and F, left). Again, scrambling onset positions among low confidence candidates showed that monosome depletion was not observed by chance (Fig. 28F, right), further indicating that a number of low confidence candidates may indeed assemble co-translationally. Among low confidence candidates, we observed a generally higher co-co assembly efficiency of transmembrane proteins, especially the ones exposing a TMD just before assembly onset (Fig. 28G). This may reflect the need of TMDs to efficiently interact with their partner transmembrane helices in order to bury the hydrophilic dimerization interfaces and allow proper insertion into the lipid bilayer (*121*).

## 3.5. Investigation of *cis* and *trans* mechanisms of co-co assembly

A central question for a mechanistic understanding of co-co assembly is whether it involves nascent chains translated from the same or different transcripts (i.e. *in cis* or *in trans*). Several studies provided indirect evidence for both mechanisms (*67*, *71*, *75–79*, *81*). Whether both mechanisms exist in human cells, when and how often each of them is employed has still to be determined.

### 3.5.1. Investigation of a heterodimeric candidate's assembly *in trans*

We first aimed to identify and validate heteromeric co-co assembly interactions, which would directly prove *trans* assembly. Speaking against frequent co-co assembly of heteromeric complexes, analysis of DiSP data indicated that heteromer subunits were not significantly enriched in either of the co-co assembly lists, and that, unlike homomers, they did not generally expose interaction interfaces on nascent chains at the initiation of co-co assembly (Fig. 19). Screening for protein pairs forming heteromeric candidate complexes, we found that, in most cases, only one of the annotated subunits of a heteromer was included in our co-co assembly lists, suggesting a so far undescribed homo-dimerization or interaction with an unknown partner subunit. Hence, we could not select any well-established, structurally characterized protein complex, whose subunits were both co-co assembly candidates and formed exclusively heteromeric (and not homomeric) interactions. Among the most promising candidates was a heteromeric complex composed of TCOF1 (Treacle protein) and NOLC1 (nucleolar and coiled-body phosphoprotein 1). These proteins are intrinsically disordered and enclose an N-terminal LisH domain (with proposed dimerization function) and a highly negatively charged central domain. A recent study reported that TCOF1 and NOLC1 central domains are extensively ubiquitinated during differentiation of pluripotent cells into neural crest cells, leading to a conformational change that allows dimerization and keeps the proteins soluble (*122*). Our data suggest that in HEK293-T and U2OS cells, dimerization may occur co-translationally, right after emergence of the LisH domain from the ribosome exit tunnel, suggesting a constitutive biogenesis pathway distinct from the ubiquitination-dependent post-translational dimerization which occurs during differentiation. Indeed, co-translational dimerization, before emergence of the charged domain, may not require ubiquitination, and could favour early stabilization of the two proteins.

To test co-translational dimerization of the TCOF1-NOLC1 complex, we first asked if targeted nascent chain cleavage of one subunit could suppress disome enrichment of the partner subunit. Therefore, we introduced TEV cleavage sites at two different positions of each gene at their endogenous loci and performed DiSP of untreated or TEV-treated lysates. All four CRISPR experiments resulted in widespread cell death, suggesting toxicity of the *TCOF1* and *NOLC1* edits. The only vital clones were heterozygous and the modified alleles were only weakly expressed, hindering interpretation of DiSP data (data not shown).

In parallel, we performed single molecule Fluorescence In Situ Hybridization (smFISH) to test the co-localization of the mRNAs encoding both TCOF1 and NOLC1 subunits, which is a direct prerequisite for co-co assembly *in trans*. As positive control we tested the co-localization of probes hybridizing to the 5' or 3' halves of the *TCOF1* mRNA and as negative control we measured the co-localization of unrelated transcripts (i.e. *LMNA* and *TCOF1,* Fig. 29). Quantifying the fraction of co-localized mRNAs per cell indicated that *TCOF1* and *NOLC1* did not co-localize in U2OS cells, suggesting that the disome shifts were not caused by *trans* assembly of these proteins (Fig. 29).

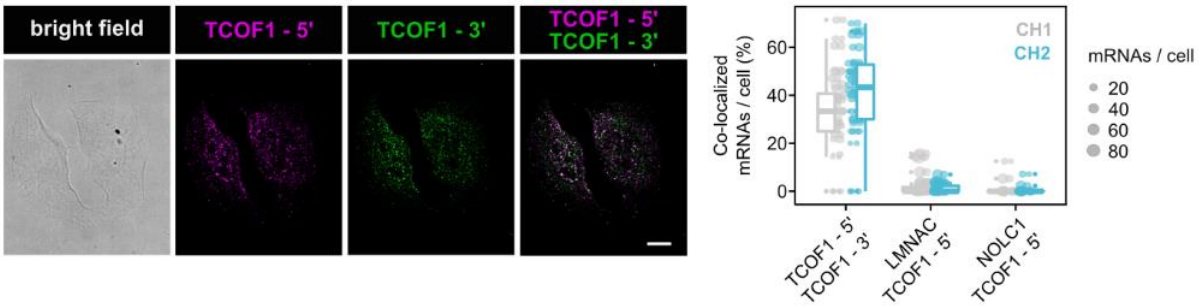



***Figure 29.*** *TCOF1 and NOLC1 mRNAs are not co-localized in human U2OS cells*

*(Left) Representative image of U2OS cells simultaneously hybridized with smFISH probes targeting the 5' or the 3' halves of TCOF1 mRNA (TCOF1 – 5', magenta, Atto647N fluorescence and TCOF1 – 3', green, Atto565 fluorescence), employed as positive control. Overlapping spots in the composite image are white. Scale bar = 10 µm.*
*(Right) Quantification of co-localization of indicated mRNA pairs was performed with the FISH-quant toolbox. The fraction of co-localized mRNAs per cell over the total number of mRNAs detected in each channel (CH1 and CH2) was calculated. Each plotted dot represents one analysed cell, the size of the dot is proportional to the total number of spots (single mRNAs) detected in that cell.*

In summary, a targeted validation of selected heteromeric candidates is complicated by the low abundance of promising heteromeric candidates in our co-co assembly lists and our efforts in this direction could not provide supportive evidence of the *trans* assembly mechanism.

### 3.5.2. A proteome-wide screen of co-co assembly *in trans*

We next attempted to globally assess the existence of *trans* assembly events on a proteome-wide scale using a combination of polysome profiling and sequencing (PP-seq). This strategy relies on the slower sedimentation of two nascent chain-coupled polysomes in sucrose gradients as compared to a single polysome. Nascent chain degradation by PK treatment should split *trans* assembling polysomes, and shift them to an apparent smaller size in sucrose gradients (Fig. 30A). Importantly, nascent chain degradation should not affect the sedimentation properties of polysomes undergoing assembly *in cis*. Hence, *trans* assembly may be identified by RNA-sequencing of polysome profiling fractions isolated from untreated or PK-treated lysates (PP-seq). Candidates identified in this screen may be subsequently validated by quantitative methods (i.e. qRT-PCR).
We went on to perform a series of sucrose gradient optimizations and generated an initial PP-seq dataset. The gathered data did not reveal any promising *trans* assembly candidate. One possibility is that the mechanism does not exist. Another possibility is that of technical and experimental limitations that prevented its detection, including the following:

1. Detection of *trans* assembly requires that each nascent chain connection between two polysomes is digested and that no other PK-resistant connection exists between them (such as RNA-binding proteins).
2. The limited resolution of sucrose gradients. Under the conditions employed, polysome peaks 1 to 8 could be collected separately, while higher polysomes could only be pooled into two samples: one including peaks 9 to 11 and one including any polysome size equal

or higher than 12 (Fig. 30B). Therefore, a PK-dependent shift of two polysomes each loaded with 12 or more ribosomes would be missed.

3. Interpretation of the sequencing data. A necessary assumption when comparing the levels of each transcript across RNA-seq libraries is that those libraries should have similar complexities. Because polysome fractions likely contain a variable number of transcripts, it is theoretically not possible to compare transcript levels among different fractions. Despite this limitation, a conservative approach would still allow to compare the fold change of transcript abundance upon PK treatment within the same polysome fraction. Performing this analysis on DiSP candidates however did not reveal any obvious shift from higher to lower polysome positions upon PK treatment (data not shown).

Given all the limitations explained above, we could not support nor exclude the existence of *trans* assembly in human cells.

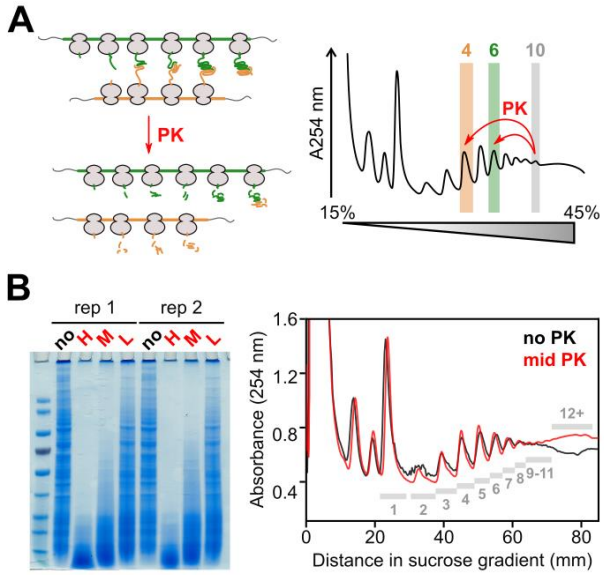



***Figure 30.*** *Polysome Profiling and sequencing (PPseq)*

*A) Polysomes connected by interactions of nascent chains in trans will co-sediment in sucrose gradients to fractions corresponding to the sum of the individual ribosomal loads (6 + 4 = 10 in this example). PK treatment should disconnect the two polysomes and shift their sedimentation to lighter polysome fractions of sucrose gradients (4 and 6 in this example).*

*B) (Left) PK was titrated on HEK293-T cell lysates and protein degradation was monitored by SDS-PAGE. Samples are labelled: no = no PK, H = high PK (1:200, PK to total protein amount), M = mid PK (1:2000), L = low PK (1:20000). Two biological replicates shown.*

*(Right) The mid PK concentration was employed for the PPseq experiment. Polysome profiles of untreated and PK-treated PPseq samples are shown, collected fractions are indicated by grey bars and labels.*

### 3.5.3. Co-co assembly *in cis* as a strategy for isoform-specific homodimer formation

Many human proteins do not mix with structurally related variants even if they contain highly conserved interaction domains. Such assembly specificity is particularly surprising when it comes to protein isoforms, which are highly prevalent in the human genome (*123*). Among them are human lamin A and C, two splicing-derived isoforms encoded by the *LMNA* gene, that share a substantial part of their N-terminus, including the long coiled coil dimerization domain (rod), and yet never form mixed dimers inside cells.

As presented in previous chapters, *LMNA* is among our high confidence co-co assembly candidates. Given the large overlap of A and C coding sequences, *LMNA* DiSP profile shows the cumulative read density of both transcripts for the majority of the gene, and only the last ~100 codons are specific to lamin A (Fig. 31A). While lamin A is certainly translated at lower

levels than lamin C (as revealed by the substantial drop of total translatome, monosome and disome read densities at codon 567), analysis of the DiSP enrichment profile indicates that both isoforms employ co-co assembly for generation of lamin homodimers, the building blocks of the nuclear lamina (Fig. 31A, right).

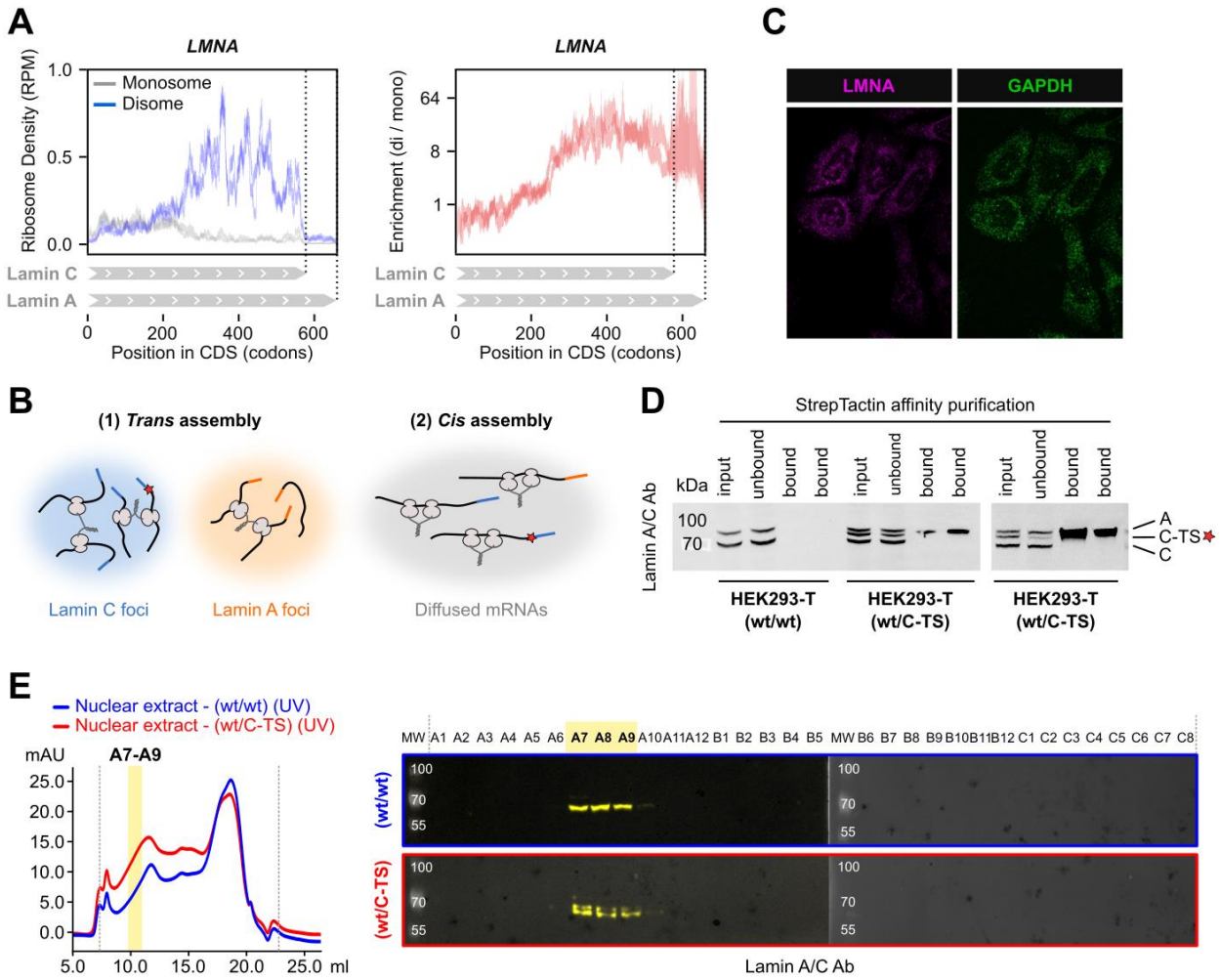



***Figure 31.*** *Transcript-templated assembly of Lamin A/C homodimers ensures isoform specificity*

*A) Density (left) and enrichment (right) DiSP profiles of LMNA in HEK293-T cells. Grey bars indicate the coding sequences of alternatively spliced Lamin A and Lamin C transcripts derived from the LMNA gene.*

*B) Possible co-co assembly scenarios that would favour isoform-specific lamin homodimerization. Lamin A and C transcripts are distinguished by their unique 3' UTRs (coloured orange and blue, respectively). Red stars indicate the TwinStrep tag at the C-terminus of lamin C in the heterozygous cell line employed for the experiment in panels D and E of this figure. Tagged and untagged versions of Lamin C would be expected to mix in the trans assembly but not the cis assembly scenario.*

*C) Representative image of U2OS cells hybridized with smFISH probes targeting LMNA or GAPDH transcripts (LMNA, magenta, Atto647N fluorescence and GAPDH, green, Atto565 fluorescence).*

*D) Nuclear lysates were prepared from wild type (wt, left) or heterozygous (wt/C-TS) HEK293-T cells and subjected to affinity purifications with StrepTactin-coupled beads. All*

*LMNA products are immuno-detected with Lamin A/C antibody. Bands are labeled: A (lamin A), C (lamin C), C-TS (lamin C – TwinStrep). Two technical replicates shown.*

*E) Size exclusion chromatography elution profiles of nuclear lysates prepared from wild type and heterozygous (wt/C-TS) HEK293-T cells (left). Grey dashed lines indicate the beginning and end of fractionation. Collected fractions were analysed by western blot using anti-lamin A/C antibody. Under the employed extraction conditions only lamin C and not lamin A dimers are efficiently extracted from the nuclear lamina (Schirmer and Gerace (JBC, 2004)). Two bands, corresponding to wild type and tagged lamin C are detected in the same fractions of heterozygous cell lysates.*

Co-co assembly may be employed to ensure biogenesis of isoform-specific lamin homodimers by either assembly *in trans* involving co-localized transcripts of the same type (for example employing putative localization elements in the unique 3' UTR sequences of the A and C mRNAs) or simply by assembly *in cis*, with the polysome itself providing the scaffold for interactions between neighboring nascent proteins (Fig. 31B).

Analysis of lamin A/C transcripts distribution by smFISH revealed that they were not clustered in specific subcellular locations of human U2OS cells, but were rather diffused in the cytosol, similarly to transcripts of *GAPDH* (a non co-co assembly gene, Fig. 31C). Still, co-localization of only two transcripts of the same kind would be sufficient for assembly *in trans* and would not require a specific subcellular localization of lamin-encoding mRNAs.

To discriminate between the *cis* and *trans* scenarios, we generated a heterozygous HEK293-T cell line by inserting a TwinStrep tag at the C-terminus of one lamin C allele and performed a series of affinity purifications using StrepTactin-coupled beads. We found that the tagged lamin C variant ("C-TS" in Fig. 31D) never co-purified its untagged counterpart ("C" in Fig. 31D), even if both proteins originated from identically spliced mRNAs sharing the same UTRs, suggesting that lamin dimerization involved proteins originated from the same mRNA molecule. Importantly, tagged and wild type lamin C showed identical gel filtration elution profiles, indicating that the presence of the tag did not impair the dimerization propensity of lamin C (Fig. 31E).

These results are in agreement with additional observations that co-co assembly of human lamin C can be recapitulated in bacteria, in the absence of any eukaryotic-specific assembly or mRNA localization machineries (data shown in the PhD Thesis of Kai Fenzl and in (*115*)). Together, our evidence supports a model of assembly *in cis*, driven exclusively by the close vicinity of N-terminal nascent chains with dimerization properties, providing a simple solution for cells to warrant isoform-specific homodimer formation.

## 3.6. Exploration of possible mechanisms facilitating co-co assembly

### 3.6.1. Correlation with ribosomal load

Assembly of nascent homomer subunits emerging from the same polysome would likely be facilitated by high ribosomal load to ensure that enough nascent subunits are available for interaction and to increase their proximity.
To test this possibility, we compared our co-co assembly dataset with translation efficiency measurements. Translation efficiency (TE) for each gene was defined as the ratio between ribosome density, derived from our total translatome dataset of HEK293-T cells (in RPKM), and mRNA levels, derived from published RNA-seq data of the same cell line (*124*) (in FPKM).

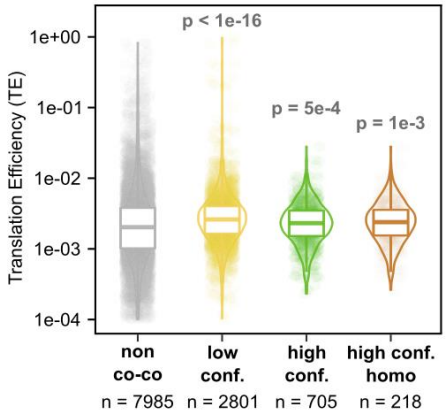



***Figure 32.*** *Translation efficiencies of co-co assembly classes*

*Violin plots of translation efficiency values (TE) of genes included in each assembly class. TE levels of each assembly class were compared to the non co-co assembly proteome, significance was tested by pairwise Wilcoxon tests with Bonferroni correction for multiple comparisons (adjusted p-values shown).*

We found that the median TE levels were significantly higher for transcripts in both high and low confidence co-co assembly classes as compared to the non co-co assembly proteome (Fig. 32). This was also true for the subgroup of high confidence homomeric proteins, which are expected to be mostly affected by ribosomal load on mRNAs. Importantly, the statistical power of our co-co assembly detection method increases as a function of gene expression, causing a bias towards highly expressed genes in our co-co assembly classes (as already discussed, see Fig. 19B). Since gene expression, in the form of ribosome densities (RPKM), is employed for TE calculation, we cannot distinguish between a true biological effect and a bias introduced by our methodology. However, the small quantitative differences in Fig. 32 seem to suggest the absence of a meaningful biological effect.

### 3.6.2. Correlation with ribosome stalling

An additional regulatory mechanism may be the coordination of co-co assembly with stalling of elongating ribosomes. Timely assembly between nascent chains could be facilitated by the slow-down of one ribosome that exposes a dimerization-competent nascent chain segment; this in turn may allow the trailing ribosome on the same mRNA to catch up (*in cis*) or provide enough time for the nascent chain synthetized on a separate polysome to establish the interaction (*in trans*, Fig.33, top). To explore this possibility, we isolated 60 nt-long footprints from the disome fraction, which are signatures of ribosome collisions caused by extended stalling (*47–50*) (experiment performed in collaboration with my colleague Kai Fenzl).

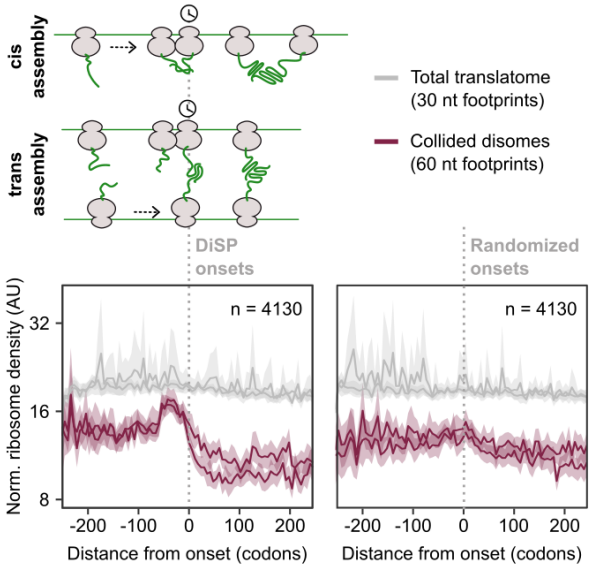



***Figure 33.*** *Ribosomes stall at the onset of co-co assembly*

*(Top) Cartoon showing the possible mechanisms by which translation stalling may favour co-co assembly in cis or in trans. In both cases, stalling of one ribosome (indicated by the clock) increases the frequency of collisions with the trailing ribosome and hence formation of 60 nt footprints.*
*(Bottom) Metagene analysis of disome 60 nt footprints (indicating ribosome collisions) and total translatome 30 nt footprints (indicating general ribosome density) including all co-co assembly proteins (high and low confidence candidates). Metagene profiles are aligned to assembly onsets determined by DiSP (left) or to randomized onset positions as a control (right).*

Metagene analysis of collided disomes aligned to the onsets of all co-co assembly candidates revealed an increased frequency of ribosome stalling shortly before and overlapping assembly onset (Fig. 33, bottom, left). This enrichment is not due to a generally higher ribosome density around the onset, as indicated by the uniform distribution of the total translatome and is largely lost if density profiles are aligned to randomized assembly onsets, suggesting a specific correlation to co-co assembly onsets (Fig. 33, bottom, right). Disome collisions seemed to be generally less frequent in the 200 codons downstream assembly onset, suggesting that ribosomes run in close proximity before onset and separate again after initiation of co-co assembly.
Inspecting single gene profiles suggested the existence of two classes of high confidence genes characterized by different patterns of stalling correlated to assembly. The first class showed repeated ribosome collisions upstream but not downstream the assembly onset and included genes with higher translation efficiency values, indicative of a higher ribosomal load on mRNAs (Fig. 34 and 35A). Examples of this class were *CAPRIN1*, *AMOT* and *VIM* (which assemble through coiled coils) and *ZNF24* (which assembles through a SCAN domain). The second class mostly showed only one stalling site right upstream of assembly onset and included genes with lower translation efficiencies, such as *ERC1* and *KIF5B* (which assemble through coiled coils), *PATZ1* and *ZNF131* (assembling via BTB domains, Fig. 34 and 35B). On one hand, repeated collisions induced by a generally high ribosome density upstream of assembly onset may facilitate early interaction of relatively short nascent chains. On the other hand, transcripts with lower translation initiation rates may have evolved deterministic stalling sites to allow the timely encounter of co-co assembly-competent nascent chains.
Additional experiments as well as the development of dedicated bioinformatics tools will be required to support the here presented findings.

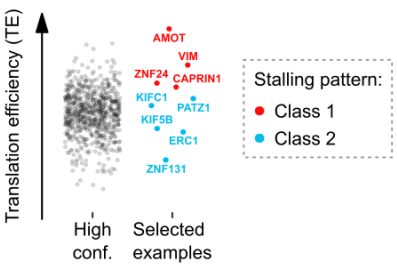



***Figure 34.*** *Two classes of ribosome stalling patterns related to co-co assembly*

*Among high confidence candidates, selected examples showing translation stalling upstream of assembly onset are highlighted. Two different classes of stalling patterns can be identified, which correlate with different ribosomal loads on mRNAs (estimated by translation efficiency measurements, TE).*

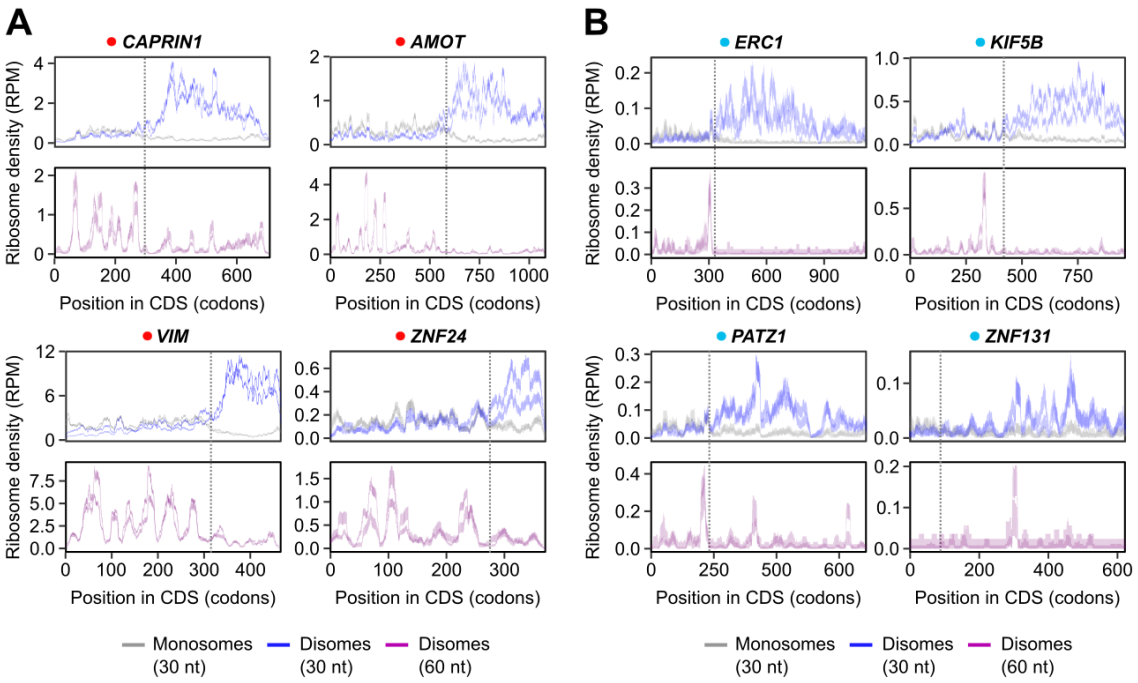



***Figure 35.*** *Repeated or deterministic stalling events correlate with co-co assembly onsets*

*Selected examples of genes showing repeated (A) or deterministic (B) stalling events upstream but not downstream assembly onset (a red or a blue dot besides the gene name indicates whether they are included in class 1 or 2 as in Fig. 34). DiSP profiles (30 nt monosome and disome footprints, top) and disome collision profiles (60 nt disome footprints, bottom) are shown. Positions of assembly onsets derived from DiSP data are indicated by dashed grey lines.*

# 4. DISCUSSION AND OUTLOOK

## 4.1. Co-co assembly is a general route to complex formation

### 4.1.1. DiSP detects nascent chain dimerization on a proteome-wide scale

Current knowledge on co-co assembly is based on mostly indirect evidence and limited to a small number of individually tested complexes (*67*, *71*, *75–79*). In this study we developed a ribosome profiling based technique, named Disome Selective Profiling (DiSP), and employed it on human cells to achieve a comprehensive picture of complex formation via interaction of two nascent subunits (*115*). In contrast to previously employed (low-throughput) approaches, DiSP identifies the nascent chains that undergo co-co assembly across the entire proteome in a single experiment. To demonstrate the validity of the detected co-co interactions, we showed that they are effectively suppressed if nascent chains are degraded by limited proteolysis or released by puromycin treatment. Similarly, RIp-chip studies analysing the co-translational assembly of selected proteins commonly rely on the perturbation of polysome integrity to discriminate between association to nascent chains and other types of interactions with the mRNA (*65*). One limitation of these controls (both of DiSP and RIp-chip) is that they cannot exclude that the detected co-translational interactions are mediated by additional “bridging” factors. To our knowledge however, there is no described mechanism which causes indirect association of ribosome pairs, is at least partially mediated by nascent chains and involves a large fraction of ribosomes synthetizing a broad spectrum of proteins. Moreover, additional investigations showed that co-co assembly of the two high confidence candidates *LMNA* and *DCTN1* can be reconstituted in *E. coli* (*115*) and, for *LMNA*, *in vitro* by optical tweezer experiments (unpublished data in collaboration with the Tans Laboratory at AMOLF). This strongly suggests that co-co assembly does not rely on eukaryotic-specific factors. Moreover, perturbing the periodicity of hydrophobic residues in heptad repeats of the *LMNA* coiled coil dimerization domain completely suppressed disome formation in *E. coli* (*115*). This modification did not alter the physical properties of lamin nascent chains, nor their predicted ability to form α-helices. Therefore, this proof of principle experiment indicates that co-co assembly minimally relies on the ability of nascent chains to directly interact with each other for productive complex formation.
Aiming to systematically identify proteins that employ co-co assembly, we further established a new bioinformatics regime which relies on fitting the DiSP ratio distributions to sigmoidal models. This approach does not bear the limitations of arbitrary threshold-based analyses and is not biased towards the length or magnitude of the disome enrichment. Thus, combining DiSP with the bioinformatics analysis developed here should allow to estimate the fraction of all proteins that employ co-co assembly (i.e. its prevalence).

### 4.1.2. Prevalence of co-co assembly across the proteome

A conservative identification of co-co assembly included genes that (i) reproducibly showed a sigmoidal shape of disome enrichment in five independent DiSP experiments (including biological replicates of the main and control datasets), (ii) lost the sigmoidal DiSP enrichment upon both PK and puromycin treatments and (iii) had a cytosolic or nuclear subcellular

localization. This high confidence class enclosed 829 proteins, among which 512 were annotated subunits of protein complexes, corresponding to about 11% of all complex subunits residing in the cytosol or nuclear compartment.

Notably, this number is likely an underestimation of the *in vivo* frequency of co-co assembly. First, interactions involving more than two nascent chains are not analysed by DiSP. Second, DiSP can only detect nascent dimers that remain connected throughout the sample preparation procedure. In an attempt to identify putative nascent chain interactions that are missed in the standard screening, we repeated DiSP on crosslinked lysates (Fig. 11). This experiment did not reveal any additional candidate; however, it also did not increase detection of early co-co assembly intermediates which would be expected to be stabilized by crosslinking, suggesting that additional experiments will be required to make definitive statements. Third, lowly expressed genes are less likely to satisfy the stringent statistical requirements of the sigmoid fitting algorithm and are largely excluded from the co-co assembly class. Fourth, for some candidates, disomes may be resistant to either PK or puromycin treatments but still be connected by nascent chains. This is likely the case for short, multi-spanning transmembrane proteins, where nascent chains are inaccessible to PK (Fig. 26) and for genes with less pronounced disome enrichments whose identification is often missed in DiSP samples of the puromycin control (Fig. 25). Finally, co-co assembly may also be employed by proteins that are translocated through the ER membrane, in agreement with previous experimental indications (*72*, *77*, *81*, *83*, *84*). Our screening identified a large number of in particular transmembrane proteins. On these transcripts, disome formation often initiates upon full exposure of N-terminal transmembrane helices and generally occurs with higher efficiency. These findings may be interpreted in two ways. The first is that the highly efficient ribosome shift to the disome fraction is caused by the co-translational translocation process itself, where N-terminal TMDs on nascent chains interact with membrane components of the translocation machinery. The second is that co-co assembly of transmembrane proteins often involves interactions of two TMDs in the ER membrane. These interactions may be particularly efficient to allow insertion of marginally hydrophobic transmembrane helices into the membrane. Furthermore, the ER membrane may facilitate highly efficient co-co assembly during translocation, by organizing the local proximity of translocons through which the interacting nascent chains are translated and by directing the orientation of interacting subunits.

To account for all these possibilities, we additionally implemented a low confidence class which relied on a more permissive selection of co-co assembly candidates. This list included proteins from all cellular compartments and minimally required a significant response to only one between PK and puromycin controls. The low confidence class enclosed 3301 proteins, among which 1615 were annotated subunits of protein complexes. Taken together, the high and low confidence co-co assembly classes comprise about 32% of all annotated complex subunits in the human proteome.

Considering the limitations described above, we believe that the real fraction of proteins employing co-co assembly in human cells lies in between 11% and 32%.

A previous study attempted to estimate the prevalence of co-post assembly in yeast, and showed that 38% of tested proteins associated with their partner subunits co-translationally (*70*). However, this measure is based on a relatively low number of pre-selected complex subunits (31 in total), suggesting it may be overestimated. Thus, unbiased and higher throughput studies on co-post assembly will be required to allow comparing the relative occurrence of the two co-translational assembly events. Nevertheless, it is now apparent that

co-co assembly, similarly to co-post assembly, is a widespread phenomenon inside cells, mediating assembly of a large and heterogeneous group of proteins.

### 4.1.3. Co-co assembly is not an *all-or-nothing* process

In the previous section, we asked how many different proteins could be detected as co-co assembly candidates in order to estimate the prevalence of this assembly mode. However, a variable fraction of ribosomes translating the same mRNA species may be involved in co-co assembly (termed efficiency in this study). For example, some proteins have moonlighting functions, they can be incorporated in different types of protein complexes or be translated at different cellular locations; hence, translation of only a sub-population of mRNAs encoding for one protein may be coupled to co-co assembly. In addition, multiple assembly mechanisms may act in parallel, so that a fraction of nascent chains undergo co-post or post-translational assembly in addition to co-co assembly. Therefore, a different question around the prevalence of co-co assembly may be: what fraction of each candidate's nascent chains employ this mechanism?

One strength of DiSP data is that they contain additional information about the efficiency of co-co assembly, i.e. the fraction of ribosomes that convert to disomes on each mRNA. Indeed, the footprint density in the monosome fraction is reduced after assembly onset in the same way the flow-through of an immunoprecipitation experiment would be depleted of bait protein to a variable extent depending on the purification efficiency. The observation of monosome depletion in DiSP data is presumably allowed by the negligible background from the disome fraction which includes a comparably small amount of 30 nt footprints. This is different in Selective Ribosome Profiling (SeRP) experiments analysing co-post assembly by affinity purification of a fully synthetized protein and sequencing of the co-purified ribosome-protected footprints. Analysis of the unbound fraction of SeRP experiments does not commonly reveal a depletion of ribosomes coinciding with the onset of co-post assembly, likely due to the limited efficiency of affinity purification (based on experience in the Bukau lab). Therefore, to our knowledge, this is the first time that the prevalence of a co-translational assembly mechanism is measured at the gene-level.

Our analysis revealed a very variable efficiency across co-co assembly candidates, with a median of about 40% in the high confidence class and some cases exceeding 90%. It is reasonable to assume that, for proteins with the highest efficiencies, co-co assembly is the main route to complex formation and failure to use this mechanism may impair correct protein biogenesis. Interestingly, among proteins with the highest efficiencies are candidates that employ especially long coiled coils for co-translational dimerization. Supporting the importance of co-co interactions for productive coiled coil formation, our unpublished data analysing co-translational folding and assembly of nascent *LMNA* by single molecule optical tweezers (in collaboration with the Tans laboratory at AMOLF) revealed that timely nascent chain interactions are required to avoid misfolding and ensure correct biogenesis of lamin dimers.

In summary, DiSP provides comprehensive and detailed information on the prevalence of co-co assembly. Our study in human cells revealed that co-co assembly is a general route to complex formation, which involves a variable fraction of ribosomes translating a wide spectrum of proteins.

## 4.2. Features of the co-co assembly proteome

### 4.2.1. Homotypic interactions drive most co-co assembly events

DiSP of human cells revealed that co-co assembly mostly involves self-association of nascent chains. This result is supported by a significant enrichment of homomeric but not heteromeric subunits of protein complexes in both high and low confidence classes (Fig. 19). Moreover, a clear correlation between the onset of disome formation and exposure of assembly interface residues on the nascent chain was observed only for homomeric and not heteromeric subunits (Fig. 19). Our observation that homotypic interactions are a predominant feature of co-co assembly is also in line with previous studies which mostly identified co-co assembly of homomeric subunits (*67*, *76–79*).

Efficient assembly of homomeric complexes may be particularly important inside cells, as homomers typically form larger interaction interfaces than heteromers (*125*). Therefore, a higher propensity for misinteraction and aggregation of unpaired homomeric subunits may have imposed evolutionary pressure to assemble early. In addition, co-co assembly of homomeric complexes can mitigate the effects of dominant negative mutations, suggesting that some homotypic co-co interactions may have evolved to dilute the impact of such mutations (*79*, *91*). Importantly, co-co assembly *in cis* is inherently regulated in time and space by the polysome itself. Dimerization is facilitated by the high local concentration of homodimerization interfaces, especially N-terminally ones, which simultaneously emerge from one polysome. In addition, co-translational folding, which is employed by proteins to protect themselves from misfolding in the first place, also has the secondary effect of facilitating native interactions for complex assembly. The higher frequency of homomer co-co assembly may thus simply reflect the intrinsic ease and low energetic cost of establishing interactions *in cis* compared to assembly *in trans*. In fact, impeding interactions of proximal interfaces is likely a more demanding task than letting them happen. To avoid unwanted interactions *in cis*, homomerization domains may be actively protected by ribosome-associated chaperones or dedicated assembly factors. Furthermore, the spacing between ribosomes translating the same mRNA could be modulated in order to avoid the close proximity of interaction-prone nascent chains. For example, ribosome stalling prior to synthesis of the interaction domain would create a translation ramp, producing an increased distance between elongating ribosomes.

### 4.2.2. Do heterotypic interactions also mediate co-co assembly?

We have made several attempts to support the possibility of co-co heteromer assembly. Identification and validation of heteromeric candidates would be of great value, as it would allow to easily test the specificity of co-co interactions, the regulatory mechanisms that drive co-localized translation of complex subunits, the possible effects of the lack of one subunit on the stability and translation kinetics of the second subunit and, importantly, it would enable to test the functional relevance of co-co assembly. The latter point is particularly challenging, as it requires perturbation of the assembly pathway without directly interfering with the formation of the final complex. This can be achieved in case of heteromeric complexes, where translation and assembly can be uncoupled by perturbing the subcellular localization of the interacting polysomes. A similar experiment was performed in bacteria, where the efficiency of the heterodimeric luciferase complex formation is reduced by ~40% if the two subunits are

expressed from distant sites of the genome compared to co-post assembly from a single polycistronic mRNA (*63*).
Identification of promising heteromeric candidates was however problematic, as the high and low confidence lists often included only one subunit of a heteromeric complex. Even when two annotated subunits of a heteromeric complex were both DiSP candidates, their interaction was often poorly characterized. This was the case of the TCOF1-NOLC1 complex which was only studied in a specific cellular context so far, namely during differentiation of pluripotent cells into neural crest cells (*122*). Our attempts to validate co-co assembly of this putative heterodimeric complex in HEK293-T cells were all unsuccessful. In particular, the lack of co-localization of the mRNAs encoding both complex subunits observed by smFISH suggests that the disome shifts were either due to co-co assembly with an unknown partner subunit or, possibly, to homo-dimer formation (Fig. 29).
We also did not find clear evidence for co-co assembly of two recently described heterodimeric complexes (*71*, *75*).
In conclusion, this study could not support nor exclude a co-co assembly mechanism for heteromeric complexes. Given the difficulties of selecting individual heteromers for validation experiments, the development of proteome-wide screens that systematically and selectively identify heteromer co-co assembly is an important future direction (including further optimization of the PP-seq approach described in this Thesis).

### 4.2.3. N-terminal interfaces: a risky necessity for co-co assembly

Co-co assembly typically initiates during translation of the N-terminal halves of coding sequences (Fig. 20). The onsets of assembly mostly correlate with exposure of homomeric interfaces on the nascent proteins (Fig. 19). Accordingly, all major dimerization folds mediating co-co assembly are located towards the N-terminus of the human proteome, except for coiled coils which can be found at any position (Fig. 36).

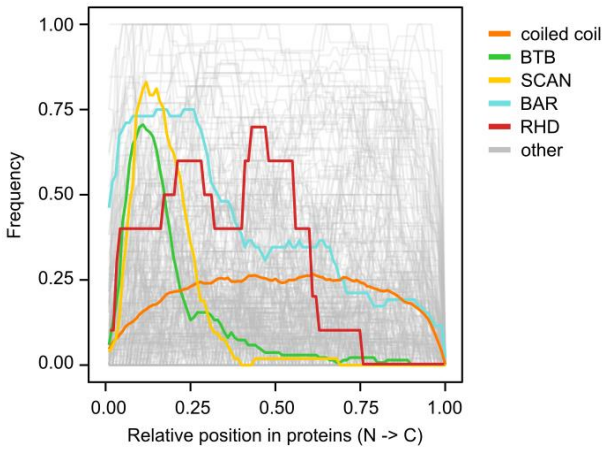



***Figure 36.*** *N-terminal bias of domains mediating co-co assembly*

*Relative position of protein domains across the human proteome. The frequency of each domain is plotted against their relative postion along coding sequences. All protein domains annotated in Uniprot KB which occur at least 5 times in the human proteome are included (grey). The five major co-co dimerization domains are colored differently.*

A previous study showed that homomerization domains are evolutionarily selected to be enriched in C-terminal halves of proteins across all organisms, from bacteria to humans (*73*). Assembly of nascent N-terminal domains on one polysome forces the yet to be synthetized C-terminal parts in close vicinity, which are initially unfolded or partially folded. The authors suggest that this results in an increased risk of nascent chain misfolding which poses constraints to the evolution of homomeric proteins with N-terminal dimerization domains. Our analysis confirms a general C-terminal biasing of homomeric contacts in the human proteome, except for the subset of high confidence candidates, where this tendency is

reversed (Fig. 20). This result indicates that a specific subset of proteins escaped the counter-selection described by Natan et al., and suggests that, for these proteins, the benefits of early assembly predominate over the risks of co-translational misfolding. A number of features and regulatory mechanisms may have enabled evolution of co-co assembly, for example a rapid co-translational folding of the protein segments C-terminal to the co-co dimerization domain could avoid misfolding and entanglement of nascent chains beyond co-co assembly (*73*). Accordingly, the folds downstream of co-co dimerization domains would be expected to be dominated by short-range interactions and by simple, α-helical topologies that can be rapidly formed during translation. Furthermore, the folding rate of nascent proteins can be actively modulated by the ribosome itself (*32*, *33*). Finally, chaperones are likely to play an important role in shielding the sensitive nascent chain segments until they are properly folded.

### 4.2.4. How do co-co assembly polysomes look like?

In the future, a structural approach may allow to discriminate whether polysomes translating individual co-co assembly candidates associate with each other in a configuration that would allow nascent chains to assemble *in trans* or if they are mostly found in isolation suggesting assembly *in cis*. In the latter case, it will be important to elucidate the relative orientation of ribosomes within a polysome that facilitate interactions of neighbouring nascent chains. It was argued that the three-dimensional organization of polysomes has evolved to avoid proximal nascent proteins to interact during translation, when they are most vulnerable to misfolding and aggregation (*13–17*). This notion does not contrast with our findings. Indeed, maximising distance between exit tunnels is likely to be important for most proteins, including all monomeric and heteromeric complexes, and the homomers which do not employ co-co assembly. Notably, the described polysome topologies additionally allow chaperones, modifying enzymes and targeting factors to easily access their binding sites close to the ribosome exit tunnel and on nascent chains (*14*). Therefore, a more specific investigation of polysomes translating co-co assembly candidates will be required to reveal whether they acquire a distinct organization and, if so, whether engagement of ribosome-associated factors is affected.

### 4.2.5. Conformational basis of co-co assembly

Folding of individual polypeptides is generally considered a prerequisite for productive assembly. Our study revealed five major interaction domains mediating co-co assembly. The globular BTB, RHD and SCAN domains are fully exposed on the ribosome surface when assembly starts, supporting the notion that folding of the full dimerization domains is required prior to dimerization, as already reported for co-post assembly (*66*). In these cases, the dimerization interface is formed only upon acquisition of the tertiary structure and, for BTB and SCAN, stable dimer formation requires cross-interactions of the most C-terminal residues of the domains (*118*, *126*). Similar to co-post assembly (*63*, *66*), ribosome-associated chaperones may play a role in keeping these globular domains monomeric until all required residues have emerged into the cytosol and acquired the native fold.
Coiled coils and BAR domains are all-helical structural motives which are often only partially exposed when co-co assembly starts. In these cases, α-helices may be formed already inside the ribosome tunnel so that the segments required for assembly would be directly

available in an assembly-competent form as they progressively emerge into the cytosol. This may allow such domains to assemble in a continuous, zipper-like process, until the interaction is strong enough to allow stable dimer formation. A structural feature of coiled coils is that they do not possess a tertiary structure independent of their quaternary structure, suggesting a gradual assembly may be needed for their biogenesis. To test this hypothesis, co-co assembly of the coiled coil candidate *LMNA* was reconstituted in an *in vitro* single molecule optical tweezer setup (unpublished data in collaboration with the Tans Laboratory at AMOLF). In this approach, the interaction between two lamin nascent chains emerging from stalled ribosomes is probed at different translation stages. The results showed that lamin nascent chains gradually fold and “zip-up” as translation progresses, forming stable dimers. When co-translational folding is monitored in a monomeric setup, compact structures – compatible with nascent chain misfolding – are observed at high frequency. Strikingly, these misfolding events are largely suppressed when a second lamin nascent chain is brought in close proximity and allowed to interact, indicating that early assembly of coiled coil nascent chains is required for native folding.
These results challenge the prevailing notion that folding of individual subunits is prerequisite to their productive interaction and support the concept that folding and assembly are highly intertwined processes which can mutually influence each other.

## 4.3. Co-co assembly as a mechanism to control complex composition

### 4.3.1. Implications of the *cis* and *trans* topologies of co-co assembly

A fundamental question is whether nascent chain interactions detected by DiSP involve proximal ribosomes on the same (*in cis*) or different mRNAs (*in trans*). Understanding which of these two interaction modes is mostly employed inside cells and for individual candidates would provide important insights into the molecular mechanisms and functional implications of co-co assembly.
An obvious starting point is the identification and validation of individual heteromeric candidates, which necessarily follow a *trans* assembly pathway. We performed initial experiments in this direction, including DiSP combined with targeted nascent chain cleavage and the monitoring of mRNA co-localization by smFISH (Fig. 29). As discussed above, these attempts were however not successful.
A more convenient approach would be to screen for *trans* assembly events on a proteome-wide scale to first identify promising candidates that can be then individually studied. We reasoned that assembly *in trans* can be identified based on the larger size of polysome pairs that are connected through nascent chain interactions. Such polysomes should dissociate upon degradation of nascent chains by PK treatment. Therefore, polysomes that undergo co-co assembly *in trans* should be found at “lighter” positions in sucrose gradients in PK-treated compared to untreated samples (an approach we named here PP-seq). Applying PP-seq in HEK293-T cells, we could not find any clear evidence of assembly *in trans*. This may be due to the substantial technical difficulties of this approach, described in detail in chapter 3.5.2. However, it is also possible that *trans* assembly is rare or even not possible inside cells, as this mechanism would require a remarkably high level of coordination between translation and folding of two independent proteins. Another possibility is that mRNAs involved in *trans* assembly interactions are additionally linked by RBPs, which would be less efficiently digested than nascent chains by PK treatment and thus would be missed by PP-seq. A direct coupling of two mRNAs via RBPs may organize the co-localized translation of mRNAs prior

to assembly. This mechanism was already shown to facilitate co-translational assembly of the voltage-gated potassium channel hERG heterodimer (*81*) and is the basis of the RNA operon hypothesis (see chapter 1.3.2) (*93*). Thus, we could not provide clear evidence for co-co assembly *in trans* in this study. Whether this mechanism is employed and if so, to what extent, remain open questions for future investigation.
The repertoire of suitable experimental approaches aimed to identify co-co assembly *in cis* is limited. Previously, assembly *in cis* of tested complexes was supported by the observation that two variants of a protein subunit did not segregate in a random (diffusion-driven) process, but preferentially formed homomeric assemblies. The two variants were either co-translated *in vitro* (*67*, *76*, *78*, *79*) or *in vivo* from a plasmid-encoded bicistronic mRNA (*67*). Here, we employed a similar approach to test the segregation of a wild-type and a C-terminally tagged variant of lamin C. In this experiment however the two variants are expressed from their endogenous genomic loci in human cells and the transcripts share the same nucleotide sequence (expect for the short C-terminal tag), so that the final complexes are the result of a physiological biogenesis pathway. If any regulatory mechanism exists (based on nuclear or cytosolic RBPs) that organises the co-localization of C-type transcripts in the cytosol to facilitate isoform-specific homo-dimerization, it should be preserved in this setup (see chapter 1.3.2). Our observation that tagged lamin C does not mix with the untagged counterpart reveals that lamin assembly is not only isoform-specific, but also allele-specific, and suggests it employs a *cis* assembly mechanism. Another possible explanation for the allele-specificity of lamin C dimerization is that the products of different alleles are regulated separately starting from their genomic loci, enabling transcripts of each allele to segregate in the cytosol and undergo assembly *in trans*. Such a mechanism may have evolved to mitigate the deleterious effects of dominant negative mutations which are highly prevalent in subunits of homomeric complexes. However, we note that assembly *in cis* represents a much simpler and cost-effective mechanism to achieve the same result, relying solely on the polysome scaffold itself.

### 4.3.2. Co-co assembly minimizes promiscuous interactions

Until now, the absence of A-C lamin heterodimers was hypothesized to result from a putative post-translational surveillance mechanism that would recognize the different tail domains of the two isoforms and eliminate mixed dimers (*127*). However, the *cis* co-co assembly mechanism proposed here provides a more convenient solution to directly assemble homomers in a specific manner. The isoform-specificity of lamin A/C dimerization is an extreme example of a broader cellular problem, namely the control of complex composition in the extremely heterogeneous and dense cytosolic environment. Indeed, protein isoforms and, more generally, structurally-related proteins with similar interfaces are highly abundant in the human proteome (*123*, *128*, *129*). Therefore, co-co assembly *in cis* could be a general mechanism relevant for many of these proteins that function as homomers. Accordingly, DiSP revealed that co-co assembly is mostly employed to assemble homomeric complexes and that it is largely mediated by modular dimerization domains that are conserved and highly redundant in the human proteome (i.e. coiled coils, BTB, SCAN, RHD and BAR domains). These domains bear a significant potential to establish promiscuous interactions (*118*, *129*, *130*), yet their association is surprisingly specific inside the cell (*127*, *128*, *131*). We speculate that interactions in *cis* could drive most co-co assembly events of homomeric complexes, thereby providing an easy solution for co-localized synthesis of complex subunits

and avoiding promiscuous interactions with homologous domains in other proteins or with alternatively spliced variants. A more sophisticated and energy demanding mechanism based on concerted regulation and co-localized translation of complex subunits may be employed to ensure specificity of heteromeric complex formation *in trans* (*93*).
A new pathway has been recently discovered, that specifically recognizes and degrades complexes of aberrant composition, while leaving the native counterparts intact (Dimerization Quality Control, DQC) (*126*, *131*). Interestingly, DQC was described as a quality control mechanism for BTB dimers but the authors suggested that a similar dedicated molecular machinery is likely to exist to monitor the composition of another highly redundant dimerization unit, namely coiled-coils (*131*).
We therefore speculate that co-co assembly is a general strategy to secure the specificity of subunits association in the first place, thereby preventing excessive burden on the downstream quality control machineries, including the recently discovered DQC.

## 4.4. Correlation of co-co assembly with ribosome stalling: a mutual regulatory mechanism?

During co-co assembly, folding and interaction of two nascent proteins must be coordinated with their simultaneous synthesis, making this a non-trivial process that likely requires fine regulation. Analysing ribosome stalling events by 60 nt disome footprinting, we show that the onset of co-co assembly is often immediately preceded by a local slowdown of translation elongation. Importantly, a more detailed analysis of the 60 nt disome footprinting data presented in this study should provide initial insights on the determinants of co-co assembly-associated stalling events. This will require development of dedicated bioinformatics tools to reliably and systematically annotate stalling sites on a genome-wide scale.
In particular, whether ribosome stalling plays a functional role for co-co assembly is a key question that remains unresolved. The coordination of translation kinetics and assembly may enhance the efficiency of both possible co-co interaction modes. On one hand, assembly *in cis* could be facilitated by a higher local density of the interacting nascent chains on the same mRNA obtained by stalling. On the other hand, ribosomes exposing assembly-competent nascent chain segments may be stalled on two different mRNAs to provide an extended timeframe to establish the interaction *in trans*. Another possibility is that stalling indirectly facilitates assembly of nascent chains (both *in cis* and *in trans*) by enhancing folding of the ribosome-exposed dimerization domain, a feature supported by several studies (*44*, *46*, *50*). Distinguishing between a direct effect of translation stalling in optimizing the timing of co-co interactions and the indirect effect of facilitating proper folding of the involved dimerization domain will be very challenging.
Finally, an intriguing possibility is that, similar to co-translational folding, assembly in proximity of the ribosome surface exerts a mechanical force on the nascent chain which could affect translation elongation. In such scenario, a stalling event that is required to facilitate the encounter of two nascent chains could be rescued by a successful co-co interaction. If assembly does not occur in time, prolonged stalling and ribosome collisions would trigger the ribosome-associated quality control (RQC) pathways, which target the mRNA and nascent peptide for degradation and allow ribosome recycling (*132*). This hypothetical feedback model would enable proteins to regulate their own biogenesis based on productive co-translational folding and assembly (Fig. 37).

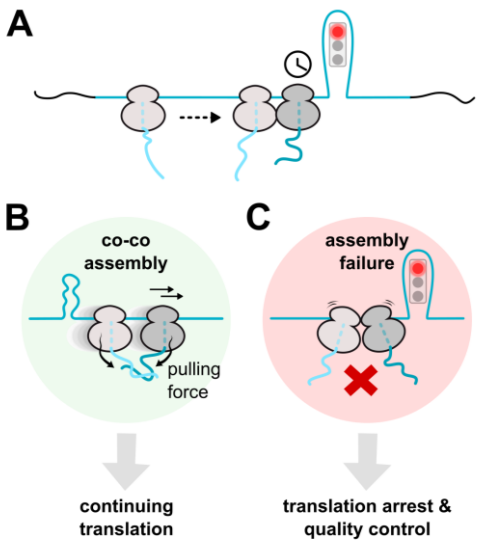



***Figure 37.*** *Model illustrating the hypothetical crosstalk between co-co assembly and quality control through ribosome stalling.*

*A) During translation of a co-co assembly protein, the leading ribosome (dark grey) encounters a roadblock (here an mRNA secondary structure). Stalling of the leading ribosome allows a second ribosome (in this example on the same mRNA) to catch up.*

*B) If the nascent chains of the two proximal ribosomes assemble, the co-co interaction generates a mechanical force that enables the leading ribosome to overcome the roadblock and continue productive translation. The mRNA hairpin in this example would be allowed to re-form behind the ribosome couple and mediate stalling of the next ribosome translating the same mRNA.*

*C) If for any reason assembly cannot occur, prolonged stalling induces translation arrest and the activation of the ribosome-associated quality control machineries.*

# 5. MATERIALS AND METHODS

## 5.1. Materials

***Table 3.*** *General laboratory equipment*

| Instrument | Model | Company |
|---|---|---|
| Bioanalyzer | 2100 Bioanalyzer Instruments | Agilent Technologies |
| Table top centrifuges | 5424 / 5424R | Eppendorf AG |
| Blue Light LED Transilluminator | UVT-22 BE-LED | Herolab |
| FACS | BD FACS Canto | BD Biosciences |
| Spinning Disk microscope | DMi8 | Leica |
| Confocal Scanner Unit | CSU-X1 | Yokogawa |
| Gel Image System | E-BOX CX5 | Vilber |
| Gel Electrophoresis Systems | XCell SureLock Mini-Cell System, Mini-PROTEAN Tetra Vertical Electrophoresis Cell, Criterion Vertical Electrophoresis Cell | Invitrogen, Bio-Rad |
| Agarose gel chamber and trays | // | ZMBH workshop |
| Image Analyzer | ImageQuant LAS 4000 | GE Healthcare |
| Incubator | MIR-254, MIR-154-PE, ISF1-X (Climo Shaker) | Panasonic, Kuhner |
| Nanodrop Spectrophotometer | ND 2000 UV-VIS | Thermo Fisher Scientific |
| Photometer | NovaSpec UV/VIS Spectrophotometer | GE Healthcare |
| Qubit | Qubit 3.0 Fluorometer | Thermo Fisher Scientific |
| Sucrose Gradient Forming Station | Gradient Master Model 109 | BioComp |
| Sucrose Gradient Fractionator | Gradient Station Model 153, Piston Gradient Fractionator Model 152 | BioComp |
| Turbo Blotter | Trans-Blot Turbo Transfer System | Bio-Rad |
| Magnetic separation stand | MagneSphere | Promega |
| Thermal Cycler System | MyCycler | Bio-Rad |
| Magnetic stirrer | MR 3001 K | Heidolph |
| Thermomixer | Thermomixer comfort | Eppendorf AG |
| Ultracentrifuges | Sorvall Discovery 100SE | Thermo Fisher Scientific |

| | | |
|---|---|---|
| | Sorvall Discovery M120SE | |
| Ultracentrifuge Rotors | S120AT2, SW 40 Ti | Beckman, Sorvall Discovery |
| Cell culture Incubator | 150i CO2 Incubator | Heracell |
| Laminar flow cabinet | SAFE2020 Class II | Thermo Fisher Scientific |
| Neubauer Counting Chamber | // | Hecht-Assistant |
| Inverted Microscope | ECLIPSE TS100 | Nikon |
| Freezing Container | Mr. Frosty | Thermo Fisher Scientific |
| Sequencing Systems | HiSeq 2000, NextSeq 550 | Illumina |

***Table 4.*** *Expendable items*

| **Item** | **Company** |
|---|---|
| Whatman Paper, 3 mm Schleicher & Schuell | BioScience Inc. |
| TBE-Urea polyacrylamide gel, 8, 10 or 15% | Invitrogen |
| Syringe, 1, 2, 5, 10 and 50 ml | Becton Dickinson BD |
| Streptavidin magnetic beads | New England BioLabs |
| MagStrep "type3" XT beads (5% suspension) | Iba |
| Sterile filters, 0.2 µM | Sarstedt AG & Co. |
| Sterile bottle filters | Sarstedt AG & Co. |
| Syringe filters, 0.22 µm | Sarstedt AG & Co |
| RunBlue SDS-PAGE Precast Gels 8x10 cm | Expedeon Ltd. |
| PVDF membrane (0.2 µM) | Carl Roth GmbH |
| Polypropylene conical centrifuge tubes (15 ml / 50 ml) | Sarstedt AG & Co. |
| Petri dishes | Greiner |
| PCR tubes (200 µl) | Sarstedt AG & Co. |
| Serological pipette (5 ml, 10 ml, 25 ml) | Greiner |
| Non-stick RNase-free tubes (1.5 ml) | Ambion |
| Spin-X-cellulose acetate columns | Sigma-Aldrich Co. |
| Microtiter plate, 6, 24 or 96 well | Greiner |
| Microcentrifuge tubes (1.5 ml / 2 ml) | Sarstedt AG & Co. |
| Gel breaker tubes | IST Engineering |
| Filter tips (P10, P20, P200 and P1000) | Steinbrenner |
| Cuvettes | Sarstedt AG & Co. |
| Criterion TGX Precast Gels | Bio-Rad |
| Cover slides | Carl Roth GmbH + Co. KG |
| Scalpel, 5518016 | Braun |
| Open-Top_polyclear centrifuge tubes 14x95 | Seton |

| | |
|---|---|
| mm | |
| Nunc Lab-Tec II chamber slides (8 well, 0.7 $cm^2$ surface) | Thermo Fisher Scientific |
| Polycarbonate centrifuge tubes 11x34 mm | Beckman Coulter |
| Cell Culture Flasks (T25 and T75) | CellStar, Greiner Bio-one |
| Gel filtration column Superdex200 10/100 GL Increase | GE Life Sciences |
| Cell Culture Dishes (10 cm and 15 cm) | CellStar, Greiner Bio-one |
| Cryo-S vials | Greiner Bio-one |

***Table 5.*** *Enzymes*

| **Enzyme** | **Company** |
|---|---|
| RNase-free DNase I (10 U/µl) | Roche |
| TEV protease | Produced in house |
| Restriction enzymes | New England BioLabs<br>Thermo Fisher Scientific |
| RNase I (100 U/µl) | Ambion |
| T4 Polynucleotide Kinase (10 U/µl) | New England BioLabs |
| Superscript III Reverse Transcriptase (200 U/µl) | Invitrogen |
| CircLigase ssDNA Ligase (100 U/µl) | Epicentre |
| T4 RNA Ligase 2, truncated (100 U/µl) | New England BioLabs |
| HF Phusion Polymerase (2 U/µl) | New England BioLabs |
| Taq DNA polymerase | Produced in house |
| TrueCut Cas9 protein v2 | Invitrogen |
| Proteinase K from Tritirachium album | Sigma |
| Benzonase | E1014 Millipore |

***Table 6.*** *Size Standards*

| **Size Standard** | **Company** |
|---|---|
| 10 bp DNA Ladder | Invitrogen |
| GeneRuler 1 kb DNA Ladder | Thermo Fisher Scientific |
| PageRuler Prestained Protein Ladder | Thermo Fisher Scientific |

***Table 7.*** *Commercial Kits*

| **Kit** | **Company** |
|---|---|
| HuluFISH kit | PixelBiotech GmbH |
| QIAquick Gel Extraction Kit | Qiagen |

| | |
|---|---|
| QIAprep Spin Miniprep Kit | Qiagen |
| Z-Competent *E. coli* Transformation Kit | Zymo Research |
| Agilent High Sensitivity DNA kit | Agilent Technologies |
| Agilent RNA 600 NANO kit | Agilent Technologies |
| Agilent Small RNA kit | Agilent Technologies |
| Qubit dsDNA and RNA HS Assay Kits | Thermo Fisher Scientific |
| SuperScript III first-strand synthesis Kit | Thermo Fisher Scientific |
| SMARTer smRNA-Seq Kit for Illumina | TaKaRa |
| NEXTflex Rapid RNA-Seq Kit | PerkinElmer |
| NextSeq 500/550 Mid Output Kit v2.5 (150 Cycles) | Illumina |
| NextSeq 500/550 High Output Kit v2.5 (75 Cycles) | Illumina |

***Table 8.*** *Chemicals and supplements*

| **Chemical** | **Company** |
|---|---|
| Murine RNase Inhibitor (40,000 U/ml) | New England BioLabs |
| cOmplete, EDTA-free Protease Inhibitor Cocktail | Sigma-Aldrich |
| GlycoBlue (15 mg/ml) | Ambion |
| Bio-Rad Protein Assay Dye Reagent Concentrate (5 x Bradford reagent) | Bio-Rad |
| Trypsin-EDTA (0.25%), phenol red | Gibco |
| Poly-L-lysine | Sigma-Aldrich Co. |
| Lipofectamine Cas9 Plus Reagent | Invitrogen |
| Lipofectamine CRISPRMAX reagent | Invitrogen |
| Lipofectamin 2000 reagent | Invitrogen |
| Cycloheximide | BioChemica |
| Puromycin Dihydrochloride | Gibco |
| Formaldehyde 16% (w/v), Methanol-free | Pierce |
| EDC (1-ethyl-3-(3-dimethylaminopropyl)-carbodiimide hydrochloride) | Thermo Fisher Scientific |
| BS3 (bis(sulfosuccinimidyl)suberate) | Thermo Fisher Scientific |
| ProLong Gold antifade mounting medium | Thermo Fisher Scientific |
| 37% Formaldehyde | Merck |
| Hoechst 33342 | Sigma-Aldrich |
| 20X SSC | Gibco |
| Denhardt's solution, 50x concentrate | Serva |
| Buffer BXT | Iba |

***Table 9.*** *Media and media components*

| Media / media component | Company |
|---|---|
| High glucose DMEM with GlutaMAX™ and pyruvate | Gibco |
| Heat-inactivated FCS | Gibco |
| Penicillin-Streptomycin (10,000 U/mL) | Gibco |
| Opti-MEMTM medium | Gibco |

***Table 10.*** *Antibodies*

| Antibody | Source |
|---|---|
| Lamin A/C Antikörper (E-1) | Santa Cruz Biotechnology, Cat# sc-376248, RRID: AB_10991536 |
| anti-Puromycin antibody (12D10) | Millipore, Cat# MABE343, RRID: AB_2566826 |
| StrepTactin-AP conjugate | Iba, Cat# 2-1503-001 |
| Mouse IgG kappa binding protein conjugated with HRP (m-IgGκ BP-HRP) | Santa Cruz Biotechnology, Cat# sc-516102, RRID:AB_2687626 |

***Table 11.*** *Plasmids and genes*

| Name | Application | Source |
|---|---|---|
| pcDNA3.1-GFP(1-10) | Clone selection for CRISPR cell line generation | Addgene (Plasmid #70219) |
| pcDNA3.1- LMNA-TEV-TS-eGFP | DiSP with TEV experiment | Invitrogen synthetic gene cloned into pcDNA3.1 |

## 5.2. Standard molecular biology and biochemistry methods

Common molecular biology and biochemistry methods employed in this thesis followed standard procedures and manuals. Microbiology and cloning techniques included: agarose gel electrophoresis of DNA, polymerase chain reaction (PCR) for preparative and analytical purposes, DNA digestion with restriction enzymes and ligation, plasmid isolation, *E. coli* cell growth and selection on solid and liquid media, preparation of chemically competent *E. coli* cells and transformation of plasmid DNA into chemically competent cells.

The *LMNA-TEV-TS-eGFP* construct employed for the experiments in chapter 3.2.1 was chemically synthetized (GeneArt Gene Synthesis, Invitrogen) and cloned by restriction (KpnI-HF: NEB and XhoI: Thermo Fisher Scientific) and ligation into the pcDNA3.1 plasmid for expression in mammalian cells. Sequences encoding TEV-TwinStrep and GFP11-TwinStrep were obtained by direct oligonucleotides annealing.

Additional standard methods included extraction of genomic DNA from human cells, protein precipitation with trichloroacetic acid (TCA), immuno-precipitations, protein separation by

SDS PAGE, Coomassie staining and Western Blotting. All reagents, enzymes and kits employed are listed in Tables 3 to 11.

## 5.3. Cell culture

All cell lines employed in this study (Table 12) were cultivated in high glucose DMEM media containing GlutaMAX™ and pyruvate (Gibco), which was freshly supplemented with 10% heat-inactivated FCS (Gibco), 100 units/mL penicillin and 100 µg/mL streptomycin (Gibco) and were grown in a humidified incubator with 5% CO2 at 37°C (HERAcell 150i). They were cultivated in flasks (Greiner) and checked approximately every 6 months for possible mycoplasma contamination (myGATC mycoplasmacheck). Cells were passaged regularly to avoid reaching confluency until a maximum passage number of ~35. For passaging, cells were first rinsed with pre-warmed 1x PBS and then trypsinized (Gibco) and diluted in pre-warmed media.
For all profiling experiments, cells were seeded one day in advance in 15 $cm^2$ dishes to obtain a confluency of about 70-90% on the experiment day. To this end, ~3.5 million U2OS and ~6 million HEK293-T cells were seeded per dish. Under these conditions, a single dish of cells was typically enough for one DiSP sample set.

***Table 12.*** *Cell lines*

| Cell lines | Source |
|---|---|
| U2OS cells (Homo sapiens osteosarcoma) | ATCC Cat# HTB-96, RRID: CVCL_0042 |
| HEK293-T cells (Homo sapiens embryonal kidney) | DSMZ Cat# ACC 635 |
| *TCOF1(wt/TEV-TS)* HEK293-T cell line | This study |
| *NOLC1(wt/TEV-TS)* HEK293-T cell line | This study |
| *LMNA(wt/gfp11-TS)* HEK293-T cell line | This study |

## 5.4. Cell line generation

The *TCOF1(wt/TEV-TS)* and *NOLC1(wt/TEV-TS)* cell lines employed for DiSP with targeted nascent chain cleavage (chapter 3.5.1) and the *LMNA(wt/gfp11-TS)* HEK293-T cell line employed for affinity purification experiments (chapter 3.5.3) were generated from wild type HEK293-T cells by CRISPR knock-in. Genome editing was performed by transfecting Ribo-Nucleo-Proteins (RNPs) comprising the 2-piece guide RNA (gRNA, composed of a crRNA and a tracrRNA) and the purified Cas9 nuclease followed by transfection of a single-stranded donor oligonucleotide (ssODN) for homology directed repair. All CRISPR experiments were performed using Invitrogen TrueGuide Synthetic gRNA reagents and user manual.

### Design of gRNAs and templates for CRISPR

The sequence encoding for the TEV cleavage site followed by a TwinStrep tag (TEV-TS) was inserted in-frame into two different sites of both *TCOF1* and *NOLC1* coding sequences, 3’ of the sequence encoding for their putative dimerization domain (LisH). Since both proteins are largely disordered, target sites were chosen as the least conserved protein

regions (using the ClustalOmega multiple sequence alignment program and comparing sequences of *Homo sapiens*, *Rattus norvegicus*, *Bos taurus* and *Mus musculus*).
The sequence encoding for gfp11 followed by a TwinStrep tag (GFP11-TS) was inserted in-frame upstream of lamin C stop codon in the *LMNA* gene (note that the insert is incorporated in the lamin C transcript but it is spliced-out in the pre-lamin A transcript). The sequence encoding for gfp11 was implemented to facilitate selection of positive clones by complementation with gfp(1-10) expressed from plasmid and sorting by FACS.
Guide RNAs (gRNAs) were optimized using the Dharmacon online design tool (http://dharmacon.horizondiscovery.com/gene-editing/crispr-cas9/crispr-design-tool/), (Table 13). The ssODN templates were designed as previously described (*133*) with 35-49 nt homology arms at each side of the insert (based on template length). For all templates, a "minimal" TwinStrep tag including a shortened linker was employed to limit template size (*134*). Synonymous mutations were introduced in templates to destroy the PAM sequence in case the HDR product contained an intact protospacer + PAM sequence. Sense or anti-sense sequences were employed as templates based on previously described rules (*133*). Finally, ssODNs were purchased as ultramer oligos (desalted, IDT), (Table 13).

***Table 13.*** *CRISPR gRNAs and templates.*
*Annotation and sequences of gRNAs and ssODN templates employed in this study. Templates include homology arms (underlined) and insert sequences (**bold**).*

| ID | Application | Cut site | Sequence |
|---|---|---|---|
| gRNA13 | targeting non-conserved region 1 of *TCOF1* | *TCOF1* exon 13 (AAS\|VPV) | GGACCCCTTGACAGGCACTG (PAM: AGG) |
| gRNA14 | targeting non-conserved region 2 of *TCOF1* | *TCOF1* exon 16 (AVT\|SAQ) | AGGCCTGGCAAGTCTTACCT (PAM: GGG) |
| gRNA15 | targeting non-conserved region 1 of *NOLC1* | *NOLC1* exon 10 (PAA\|APK) | ACCGCCCACAGGTTGCTTGG (PAM: GGG) |
| gRNA16 | targeting non-conserved region 2 of *NOLC1* | *NOLC1* exon 9 (KAV\|VSK) | AGAGAAGAAACCCCCAACTA (PAM: AGG) |
| gRNA5 | CRISPR cell line generation | *LMNA* exon 10 (exactly upstream of lamin C stop codon) | GTGAGTGGTAGCCGCCGCTG (PAM: AGG) |
| T13 | Donor template for insertion into non-conserved region 1 of *TCOF1* | | ATCCCAGGCAAAGTCTGTGGGGAAAGGCCTCCAGGTGAAAGCAGCCTCA**TCCGGCGAAAACCTGTACTTCCAAGCGGGTTCGGCGTGGTCTCACCCACAGTTCGAGAAAGGAGGTAGCGCCTGGAGCCACCCGCAGTTCGAAAAAGCGGGT**GTGCCTGTCAAGGGGTCCTTGGGGCAAGGGACTGCTCCAGTACTCCCTG |
| T14 | Donor template for insertion into non-conserved region 2 of *TCOF1* | | CAGCAGCGAGGAATCAGACAGTGATGGGGAGGCACCGGCAGCTGTGACC**TCCGGCGAAAACCTGTACTTCCAAGCGGGTTCGGCGTGG** |

| | | |
|---|---|---|
| | | **TCTCACCCACAGTTCGAGAAAG GAGGTAGCGCCTGGAGCCACCC GCAGTTCGAAAAAGCGGGT**TCT GCUCAGGTAAGACTTGCCAGGC CTCTGAGCCACCAACACTCACT CC |
| T15 | Donor template for insertion into non-conserved region 1 of *NOLC1* | AAATAAGCCAGCTGTCACCACC AAGTCACCTGCAGTGAAGCCAG CTGCA**TCCGGCGAAAACCTGTA CTTCCAAGCGGGTTCGGCGTGG TCTCACCCACAGTTCGAGAAAG GAGGTAGCGCCTGGAGCCACCC GCAGTTCGAAAAAGCGGGT**GCU CCCAAGCAACCTGTGGGCGGTG GCCAGAAGCTTCTGACGAGAAA GG |
| T16 | Donor template for insertion into non-conserved region 2 of *NOLC1* | TCTCTGCTGCTTTCTTTGCTGG AGGTGGTTTAGTGGTTGCTTTA GAGAC**ACCCGCTTTTTCGAACT GCGGGTGGCTCCAGGCGCTACC TCCTTTCTCGAACTGTGGGTGA GACCACGCCGAACCCGCTTGGA AGTACAGGTTTTCGCCGGA**TAC TGCTTTAGTTGGGGGTTTCTTC TCTTCTTCAGAACTTGAATCTG AG |
| T5 | Donor template for insertion into *LMNA* (encoding for lamin C) | CTGCTCCATCACCACCACGTGA GTGGTAGCCGCCG**CAGCGCCAG AGACCACATGGTGCTGCATGAG TACGTAAACGCCGCCGGGATTA CAGGTAGCGCCTGGTCTCACCC CCAGTTCGAGAAAGGAGGTAGC GCCTGGAGCCACCCCCAGTTCG AAAAA**TGAGGCCGAGCCTGCAC TGGGGCCACCCAGCCAGG |

**Gene tagging by CRISPR**

gRNAs were prepared in advance by annealing tracrRNA and crRNAs according to the TrueGuide Synthetic gRNA user guide (Invitrogen). One day before transfection, $6x10^4$ HEK293-T cells were seeded on poly-L-Lysine coated 24-well plates (Greiner).
On the following day, a transfection mix (20 µl per well) was prepared in RNase-free tubes, including 7.5 pmol TrueCut Cas9 protein v2, 7.5 pmol crRNA:tracrRNA duplex and 1:10 v/v Lipofectamine Cas9 Plus Reagent in Opti-MEMTM medium. The mix was incubated for 5 min at room temperature and further supplemented with 5.5 pmol of ssODN template. Diluted Lipofectamine CRISPRMAX reagent (1.5 µl in 25 µl opti-MEMTM / well) was added to the transfection RNP mix and 55 µl final transfection complex was distributed on each well.
Cells were incubated with the RNP mix for 24 hours, then they were passed to poly-L-lysine coated 6-well (Greiner) with fresh DMEM + 10% FBS.

After ~24 hours, cells were sorted at the ZMBH Flow Cytometry and FACS facility and single clones were grown in 96-well plates.
If edits included a GFP11 tag, cells were first transfected with 1.5 µg of pcDNA3.1-GFP(1-10) plasmid, 4.5 µl Lipofectamine 2000 reagent in opti-MEMTM (Invitrogen) (180 µl transfection mix per well) and then FACS-sorted to enrich for positive edits.
Genome edits were validated by genome extraction from single clones, PCR (using primers in Table 14) and sequencing. Additional validation was performed by western blotting using StrepTactin-AP (iba).

***Table 14.*** *Primers employed for validation of CRISPR edits*

| ID | Name | Application | Sequence |
|---|---|---|---|
| MB176 | TCOF1_gRNA13_fw | Validation of *TCOF1* CRISPR edit 1 | AACTGCGACTTCTCCAGCAGGC |
| MB177 | TCOF1_gRNA13_rev | Validation of *TCOF1* CRISPR edit 1 | TGCTGGAGATGCAGCTGCTTC |
| MB178 | TCOF1_gRNA14_fw | Validation of *TCOF1* CRISPR edit 2 | CAGGTGAAGCCTTCAGGGAAGAC |
| MB179 | TCOF1_gRNA14_rev | Validation of *TCOF1* CRISPR edit 2 | TCCCAGGTTCTAGGGATTCTCCTG |
| MB180 | NOLC1_gRNA15_fw | Validation of *NOLC1* CRISPR edit 1 | CTGACAGCTCTGAGGATGATGAAGC |
| MB181 | NOLC1_gRNA15_rev | Validation of *NOLC1* CRISPR edit 1 | CTGCACCTCCTGCTACCTTCTGTG |
| MB182 | NOLC1_gRNA16_fw | Validation of *NOLC1* CRISPR edit 2 | AGACAGCAGTGATGAGTCTGGTGAG |
| MB183 | NOLC1_gRNA16_rev | Validation of *NOLC1* CRISPR edit 2 | AGCAGCTGCCTGAGAGCATCTC |
| MB132 | LMNA_gRNA5_fw | Validation of *LMNA* CRISPR edit | GTAGACATGCTGTACAACCCTTCC |
| MB133 | LMNA_gRNA5_rev | Validation of *LMNA* CRISPR edit | GGTATAGGGAGGAGAGAGAAGAAAGG |

## 5.5. Ribosome profiling and RNA-seq based methods

All steps of Ribosome profiling and RNA-seq based protocols were performed on ice, using RNase-free and ice-cold solutions and tools.

### 5.5.1. DiSP of HEK293-T cells

**Cell harvesting**

To keep conditions consistent among different datasets, all DiSP experiments of HEK293-T cells were performed with high salt (500 mM KCl) in lysis and sucrose buffers, which was a technical requirement for DiSP with Puromycin (*114*) and DiSP with PK (data shown in the PhD Thesis of Kai Fenzl).

A maximum of three cell dishes (15 $cm^2$ dishes with 70-90% confluent cells) were removed from the incubator and rapidly harvested per time. Growth medium was discarded completely by inversion and dishes were quickly placed on ice. Cells were detached by pipetting ~10 ml ice-cold PBS solution (1x PBS, 10 mM $MgCl_2$, 100 µg/ml CHX) with a serological pipette and collected in falcon tubes. Cells were then pelleted for 3 min at 2000 ×g, 4°C and pellets derived from each dish were resuspended in 200 µl 1x high salt lysis buffer (50 mM HEPES pH 7.0, 10 mM $MgCl_2$, 500 mM KCl, 1% NP40, 10 mM DTT, 100 µg/ml CHX, 25 U/ml recombinant Dnase1 (Roche) and protease inhibitor (complete EDTA free, Roche)). Lysates were incubated for 15 min in ice, triturated five times using a 26-G needle and cleared by centrifugation for 5 min at 20,000 ×g at 4°C.

**RNA digestion**

The RNA concentration in each sample (1:100 dilutions of the cleared lysates in water) was determined by Qubit HS RNA assay. Next, lysates were supplemented with 150 U RNase1 (Ambion) / 40 µg RNA and incubated at 4°C on a thermomixer under continuous shaking.

**Sucrose Density Gradients**

For optimal separation of monosome and disome peaks, we employed either 5-45% or 10-25% sucrose gradients with 3.5 hour or 3 hour centrifugation, respectively, at 35,000 rpm, 4°C (SW40-rotor, Sorvall Discovery 100SE Ultracentrifuge).
Briefly, sucrose was dissolved in sucrose buffer (50 mM HEPES pH 7.0, 5 mM $MgCl_2$, 500 mM KCl, 100 µg/ml Cycloheximide, EDTA Free protease inhibitor tablet Roche) at the right concentrations (e.g. 5% and 45% or 10% and 25%) and the obtained solutions were filtered through 0.2 µm filters. Next, gradients were prepared in SW40 centrifugation tubes (SETON) with short caps using the Gradient Station (BioComp). Following custom programs were used for gradient formation:

- 5 - 45% gradients: M#1: 09 sec/83.0°/30 rpm M#2: 09 sec/83.0°/0 rpm M#3: 01 sec/86.0°/40 rpm M#4: 7 min/90.0°/0 rpm, sequence 12121212121234.
- 10-25% gradients: 2:19 min/81.5°/14 rpm.

Gradients were cooled down to 4°C before use. The RNase-treated lysates were gently loaded on gradients (up to 300 µg total RNA). After centrifugation, gradients were analysed and fractionated with the Piston Gradient Fractionator (Biocomp). For maximal resolution, 60 fractions of 200 µl were collected and flash-frozen in liquid nitrogen. Monosome and disome fractions were pooled separately, omitting 5 to 8 fractions in between to minimize cross-contamination.

**DiSP of HEK293-T cells with TEV treatment**

For DiSP with TEV treatment (chapters 3.2.1 and 3.5.1), protein concentration in the cleared lysate was determined using the Bradford protein assay. Next, TEV protease (lab collection) was supplemented to the lysate in a 1:50 ratio (TEV to total protein content). RNase 1 was added in parallel and lysates were incubated for 30 min at 4°C under constant shaking to allow simultaneous digestion of RNA and the TEV-targeted nascent chains.
All downstream operations were performed as described for HEK293-T cells.

### 5.5.2. DiSP with Puromycin treatment

As described in detail in chapter 3.2.2, experimental conditions suited to avoid dissociation of ribosomes from mRNAs while allowing efficient nascent chain release by Puromycin were based on (*114*). Extra care was taken to maintain samples at a temperature of ~0°C (keeping them in an ice bath). Cycloheximide was omitted from all buffers because incompatible with Puromycin and high salt buffers were employed as described above. Different from standard DiSP, cells were seeded on poly-L-lysine coated 15 $cm^2$ dishes and lysis was performed directly on dish, to avoid the longer detachment and pelleting procedure which is employed for DiSP of HEK293-T cells and exposes samples to higher temperature (~4°C).
Briefly, after removing the medium by inversion, cells were gently rinsed by pouring ~10 ml ice-cold PBS solution (1x PBS, 10 mM $MgCl_2$) on the dish, taking care to cover the whole surface. Next, the PBS solution was completely discarded by repeatedly tapping the dish on tissue paper. The dish surface covered by cells retains some of the PBS solution, therefore we implemented the use of a 5x concentrated lysis buffer lacking cycloheximide (250 mM HEPES pH 7.0, 50 mM $MgCl_2$, 750 mM KCl, 5% NP40, 50 mM DTT, 500 µg/ml CHX, 125 U/ml recombinant Dnase1 (Roche) and protease inhibitor (complete EDTA free, Roche),). 100 µl of 5x concentrated lysis buffer were distributed on the dish surface and cells were scraped off. This operation resulted in about 500 µl of lysate per dish. The lysate was transferred to RNase-free tubes (Ambion), supplemented with KCl to a final concentration of 500 mM and incubated for 15 min in an ice bath. After centrifugation for 5 min at 20,000 ×g at 4°C to remove cell debris, treated samples were supplemented with 2 mM Puromycin (Gibco™ Puromycin Dihydrochloride) and untreated samples with an equivalent volume of 1x lysis buffer. Because RNase 1 is less active at 0°C than at 4°C, lysates were supplemented with 750U RNase1 (Ambion) / 40 µg RNA and incubated in an ice bath for 25 min with occasional shaking. Finally, lysates were either directly loaded on sucrose gradients (EXP1, Fig. 15) or crosslinked with 0.5% formaldehyde (Pierce™ 16% Formaldehyde (w/v), Methanol-free), incubated for 30 min in the ice-bath and then loaded on gradients (EXP2, Fig. 15). DiSP samples were loaded on 5-45% sucrose gradients and all downstream steps were carried out as described for standard DiSP of HEK293-T cells.
For polysome analysis (Fig. 15B, C), RNase1 was omitted and the rest of the protocol was performed as described above. Control lysates were loaded on 5-45% sucrose gradients and centrifuged for 2.5 hours at 35,000 rpm, 4°C (SW40-rotor, Sorvall Discovery 100SE Ultracentrifuge). Sucrose fractions corresponding to the supernatant and polysomes were collected separately and proteins were precipitated with trichloroacetic acid (TCA). Equal sample volumes were loaded on polyacrylamide gels and proteins were separated SDS PAGE. Detection of puromycilated nascent chains was performed by wester blotting with anti-Puromycin antibody (Millipore Cat# MABE343).

### 5.5.3. Classical Ribosome Profiling

Total translatome samples of HEK293-T cells were prepared as described above, except that lysates were loaded on 25% sucrose cushions (homogeneous sucrose solution prepared in sucrose buffer as above) instead of sucrose gradients, following published protocols (*135*, *136*). Each polycarbonate centrifuge tube (11x34 mm) was filled with 700 µl filtered cushion solution. RNase-treated lysates were loaded on cushions and centrifuged for ≥1h at 100k rpm at 4°C (S120AT2-rotor, Sorvall Discovery M120SE Ultracentrifuge). Ribosome pellets

were resuspended in ribosome buffer (50 mM HEPES pH 7.0, 10 mM $MgCl_2$, 150 mM KCl) and flash frozen in liquid nitrogen.

### 5.5.4. DiSP of U2OS cells

DiSP experiments of U2OS cells were performed under standard salt conditions, i.e. all buffers have the same composition as for DiSP of HEK293-T cells, except for the lower KCl concentration (150 mM KCl). Moreover, because U2OS cells adhere more tightly to the growth surface, cell lysis had to be performed directly on dish. Finally, we employed a lysis-coupled RNA digestion procedure for all U2OS DiSP experiments, where RNase 1 was directly supplemented in the lysis buffer.
Briefly, after removing the medium by inversion, cells were gently rinsed by pouring ~10 ml ice-cold PBS solution (1x PBS, 10 mM $MgCl_2$, 100 µg/ml CHX) on the dish, taking care to cover the whole surface. Next, the PBS solution was completely discarded by repeatedly tapping the dish on tissue paper. The dish surface covered by cells retains some of the PBS solution, therefore we implemented the use of a 5x concentrated lysis buffer which also contained RNase 1 (250 mM HEPES pH 7.0, 50 mM $MgCl_2$, 750 mM KCl, 5% NP40, 50 mM DTT, 500 µg/ml CHX, 125 U/ml recombinant Dnase1 (Roche) and protease inhibitor (complete EDTA free, Roche), 6.6 units/µl RNase 1). Thus, 100 µl of 5x concentrated lysis buffer were distributed on the dish surface and cells were simultaneously lysed and collected by means of a cell scraper. This operation usually resulted in 500 µl of lysate for each dish. The resulting lysates were transferred to RNase-free tubes (Ambion) and incubated for 15 min on ice. All downstream operations were performed as described for HEK293-T cells.

**DiSP of U2OS cells with chemical crosslinking**

DiSP with chemical crosslinking was performed as described above (DiSP of U2OS cells), except that the lysis buffer was additionally supplemented with 2.5 mM BS3 and 20 mM EDC. Stocks of both crosslinkers were prepared freshly starting from powder on the day of the experiment. BS3 was dissolved in water, as a more concentrated buffer salt can interfere with initial solubility of the reagent. EDC stock was prepared directly in 5x lysis buffer.

### 5.5.5. Polysome profiling and sequencing (PP-seq)

For maximal resolution of polysome peaks, we employed 15-45% sucrose gradients (formed using the same program as for 5-45% gradients, see “DiSP of HEK293-T cells”). Cell harvesting was performed as for DiSP of HEK293-T cells. The protein concentration in cleared lysates was determined by Bradford assay. PK treated samples (Proteinase K from Tritirachium album, Sigma) were incubated for 30 min at 4°C under continuous shaking with the following PK amounts (defined as ratios of PK to total protein): (i) low PK = 1:20000; (ii) mid PK = 1:2000; (iii) high PK = 1:200. Untreated samples were supplemented with an equal volume of PK storage buffer (50 mM Tris pH 7.5, 5 mM CaCl2, 40% glycerol). 20 µl aliquots of each untreated and treated lysate were loaded on a polyacrylamide gel for SDS PAGE and protein degradation was monitored by Coomassie staining.
All samples were loaded on 15-45% gradients and centrifuged for 2.5 hours at 35,000 rpm, 4°C (SW40-rotor, Sorvall Discovery 100SE Ultracentrifuge). 60 fractions of 200 µl were collected from each gradient and flash-frozen in liquid nitrogen.

### 5.5.6. Library preparation and sequencing

All samples were subjected to acid phenol RNA extraction as previously described (*136*).
Ribosome profiling libraries of HEK293-T cells were prepared according to published protocols (*135*, *136*) in combination with a custom rRNA depletion step (see below). The only exceptions were ribosome profiling libraries of DiSP with targeted nascent chain cleavage of plasmid-expressed lamin (chapter 3.2.1), which were prepared using the SMARTer smRNA-Seq Kit (TaKaRa). These libraries were sequenced on our in-house NextSeq550 (Illumina) using high-output 75 cycles sequencing kits (Illumina) according to the manufacturer's protocol (single end sequencing, Read1 = 51 cycles, Index = 6 cycles). Libraries prepared from 60nt disome footprints (to detect ribosome stalling sites) were sequenced using a mid-output 150 cycles sequencing kit (double end sequencing, Read1 = 47 cycles, Read2 = 37 cycles, Index = 6 cycles).
All ribosome profiling libraries of U2OS cells were prepared according to an older protocol version as described in (*12*) and sequenced on a HiSeq 2000 (Illumina) at the DKFZ Core Facility for Sequencing.
RNA-seq libraries for the PP-seq experiment were prepared using the NEXTflex Rapid RNA-Seq Kit (PerkinElmer) and sequenced on our in-house NextSeq550 (Illumina) using a high-output 75 cycles sequencing kit (Illumina) according to the manufacturer's protocol (double end sequencing, Read1 = 35 cycles, Read2 = 35 cycles, Index = 6 cycles).

**Custom rRNA depletion**

A custom rRNA depletion tool was developed in collaboration with siTOOLs Biotech (Table 15). The tool consists in a mix of biotinylated DNA oligonucleotides that are reverse-complement to the most abundant rRNA contamination fragments in our ribosome profiling libraries. Typically, the depletion oligos were hybridized to the adaptor-ligated RNA footprints (L1-footprint) and subsequently removed by a pull-down via magnetic Streptavidin beads (NEB). For maximal depletion, this step was optionally repeated on the circularized cDNA libraries (before amplification by PCR) using a reverse-complement pool of depletion oligos.
A 4-fold molar excess of rRNA depletion oligos were mixed with each library sample (at the RNA or cDNA stage) and the mix was supplemented with 25 µl of 2x wash/binding buffer (40 mM Tris pH7, 1 M NaCl, 2 mM EDTA, 0.1% Tween 20 supplemented with 2 µl murine RNase inhibitor) and DEPC water to a final volume of 50 µl. The nucleic acids were first denatured at 99°C for 90 s. Next, annealing was facilitated by a temperature ramp (-0.1°C per second) until 37°C, followed by a 15 min incubation at 37°C.
A 2-fold excess of Streptavidin Magnetic Beads (NEB, calculated based on the theoretical beads binding capacity) over the employed biotinylated oligos were employed to remove the annealed rRNA fragments. Beads were first washed three times with 1x wash/binding buffer and incubated for 15 min with each hybridized library at room temperature. Finally, Streptavidin beads were magnetized and the depleted libraries were precipitated as previously described (*12*).

***Table 15.*** *Custom biotinylated rRNA depletion oligos (source: siTOOLs Biotech)*

| ID | Sequence for depletion at the RNA stage* |
|---|---|
| 1 | ACCGGCTATCCGAGGCCAAC |
| 2 | GACCGGCTATCCGAGGCCAA |
| 3 | CGGCTATCCGAGGCCAACCG |

| | |
|---|---|
| 4 | CCGGCTATCCGAGGCCAACC |
| 5 | CGGGCGCTTGGCGCCAGAAG |
| 6 | CCGGGCGCTTGGCGCCAGAA |
| 7 | CAGACAGGCGTAGCCCCGGG |
| 8 | GACGCTCAGACAGGCGTAGC |
| 9 | CGACGCTCAGACAGGCGTAG |
| 10 | GCGACGCTCAGACAGGCGTA |
| 11 | AGCGACGCTCAGACAGGCGT |
| 12 | GACAGGCGTAGCCCCGGGAG |
| 13 | GCCGGGCGCTTGGCGCCAGA |
| 14 | CCTCGATCAGAAGGACTTGG |
| 15 | GCCTCGATCAGAAGGACTTG |
| 16 | TGCGATCGGCCCGAGGTTAT |
| 17 | CGATCGGCCCGAGGTTATCT |
| 18 | GCGATCGGCCCGAGGTTATC |
| 19 | GGGCCGGTGGTGCGCCCTCG |
| 20 | CGGGCCGGTGGTGCGCCCTC |
| 21 | GACGGCGCGACCCGCCCGGG |
| 22 | ACCGGGTCAGTGAAAAAACG |
| 23 | ACTCCGCACCGGACCCCGGT |
| 24 | ACAGGCGTAGCCCCGGGAGG |
| 25 | ACAGGCGTAGCCCCGGGAGA |
| 26 | CGACGGCGCGACCCGCCCGG |
| 27 | AGGACTTGGGCCCCCCACGA |
| 28 | CCGGGTCAGTGAAAAAACGA |
| 29 | CGGGTCGACTCCGTGTACAT |
| 30 | AGGCCTCGGGATCCCACCTC |

*Note: For depletion at the circularized cDNA stage of library preparation, a reverse-complement pool of the oligos in this Table was employed.

### 5.6. Single molecule Fluorescence In Situ Hybridization (smFISH)

Probes for smFISH were purchased from PixelBiotech GmbH (HuluFISH technology) and the protocol employed for smFISH experiments is based on provider's instructions (https://www.pixelbiotech.com).

U2OS cells were seeded on 8 well Nunc Lab-Tec II chamber slides ($4.5 \times 10^4$ cells per well in 0.5 ml culture media). On the next day, cells were about 80-90% confluent in each well and the fixation and hybridization procedure were started.

All following operations were performed by never touching the wells bottom with any instrument and by always carefully pipetting solutions and reagents to the walls of the wells. All equipment and solutions employed in this protocol were filter-sterilized and RNase-free.

**Cell fixation**

The medium was removed by careful aspiration and cells were gently washed three times with PBSM solution (1x PBS, 5mM MgCl2 in DEPC water). Next, cells were fixed in 300 µl fixation buffer (PBSM + 4% PFA) for 10 min at RT. The fixation buffer was removed and cells were incubated in 300 µl quenching buffer (PBSM + 100mM glycine) 10 min at RT. Cells were washed twice in PBSM for 5 min.

**Cell permeabilization**

Cells were incubated with 300 µl permeabilization buffer (1x PBS + 0.1% Triton X-100) for 10 min at room temperature. The permeabilization buffer was aspirated and cells were washed twice by incubating for 10 min at room temperature with 300 µl Hulu wash buffer (2xSSC, 2M Urea).

**Hybridization with HuluFISH probes**

From here on, direct light exposure was avoided.
1.25 µl of each HuluFISH probe was diluted in 250 µl hybridization buffer (2xSSC, 2M Urea, 10% dextran (0.5g for 5ml), 5x Denhardt's solution (Serva)) and the resulting working solutions were mixed several times by pipetting. The wash buffer was removed completely from the wells and the whole volume of respective hybridization working solution was added to each well. The chamber slides were placed in the humidified incubator at 37°C overnight.

**Mounting slides**

The hybridization solution was removed and cells were washed twice by incubating in wash buffer in the humidified incubator at 37°C for 15 min. Cells were then quickly rinsed three times with 2x SSC at room temperature. Hoechst stain solution (Sigma) was diluted to 5ug/ml in 2x SSC, added to each well and incubated for > 1 min at room temperature. The Hoechst solution was removed and cells were incubated for 5 min in 2x SSC.
Finally, the solution was aspirated and wells were removed from the chamber slides. A drop of ProLong Gold Antifade Mountant (Thermo Fisher Scientific) was added to each well and coverslips nr. #1.5 were mounted carefully. Slides were cured at room temperature for at least 24 hours in the dark. Slides were stored at 4°C for several days.

**Microscopy**

Imaging was performed on a Leica DMi8 spinning disk system equipped with a CSU-X1 confocal scanner unit (Yokogawa), an Orca Flash 4.0 LT digital camera (Hamamatsu) and a HC PLAPO 63×/1.40 NA oil objective lens (Leica Microsystems). Images were acquired using lasers at 60% power and 150 ms exposure. Z stacks were acquired with 0.15 um steps.

**Co-localization analysis**

Single transcript positions were annotated using the FISH-quant toolbox and co-localization of transcript pairs was quantified using the dedicate user interface FQ_DualColor according to developer's instructions (*137*). Co-localization analysis was corrected by average drift and employed a threshold of 200 nm as maximum three-dimensional distance between two spots to be considered co-localized.

## 5.7. Affinity purification of endogenously tagged lamin C

Pellets of wild type and heterozygous *LMNA(wt/gfp11-TS)* HEK293-T cells were prepared from ~90% confluent T75 flasks and stored at -80°C. Cells were broken by resuspending each pellet in 0.5 ml hypotonic buffer (10 mM HEPES pH 7, 1.5 mM MgCl2, 10 mM KCl, 1 mM EDTA, 0.05% NP-40) and nuclei were pellet by centrifugation at 3,300 ×g for 10 min. Nuclear pellets were washed once by resuspension in 0.5 ml hypotonic buffer and pellet again. Next, nuclei were lysed by resuspension in 200 µl lamin extraction buffer (25 mM Tris pH 8.6, 1% NP-40, 0.5% DOC, 0.1% SDS, 500 mM NaCl, 1 µl Benzonase (E1014 Millipore), EDTA Free protease inhibitor tablet Roche). This extraction buffer is an adapted version of the RIPA buffer which was optimized based on previously described conditions that facilitate extraction of lamin dimers from the nuclear lamina (*138*). The lysates were incubated for 10 min in ice and debris were removed by centrifugation for 10 min at 20,000 ×g.
Affinity purification was performed using 40 µl MagStrep "type3" XT beads (5% suspension, iba) for each sample, following the provider's user manual. Bound proteins were eluted by incubating the beads in 20 µl lamin extraction buffer supplemented with 1x Buffer BXT (iba) for ≥ 10 min at room temperature with occasional shaking. Input, flow-through and elution samples were analysed by western blotting. Immuno-detection was performed with anti-Lamin A/C antibody (Santa Cruz Biotechnology, Table 10).
To rule out a possible effect of the gfp11-TS C-terminal insert on the ability of lamin C to form dimers, whole nuclear lysates were extracted in urea buffer (25mM HEPES 8.0, 500mM NaCl, 5mM β-mercaptoethanol, 7M urea (filtered and degased) and subjected to gel filtration on a S200 Increase G10/300 (GE Life Sciences) with the same buffer (at room temperature). Extraction of human lamins with high urea concentrations was previously shown to allow efficient break-down of the lamina polymer without dissociating dimers (*139*). In this study, a urea concentration of ~7M was shown to solubilize ~50% of the total lamin C dimers from polymers (while a higher concentration was required to elute lamin A dimers).

## 5.8. Bioinformatics methods

***Table 16.*** *Software*

| Software | Version | Company / Reference |
|---|---|---|
| Cutadapt | 2.3 | (*140*) |
| Bowtie2 | 2.3.4.1 | (*141*) |
| STAR | 2.7.1a | (*142*) |
| SAMtools | 1.5 | (*143*) |
| FastQC | 0.11.5 | (*143*) |
| Python | 3.6.8 | Simon Andrews |
| Spyder | 3.3.4 | Pierre Raybout |
| R | 4.0.3 | The R Foundation |
| RStudio | 1.2.1335 | RStudio |
| Pymol - Molecular Graphics System | 1.7.4.5 Edu | Schrödinger, LLC |
| Mendeley | 1.19.4 | Mendeley Ltd |
| Inkscape | 0.91 | Inkscape Project |

| | | |
|---|---|---|
| Salmon | 1.4.0 | (*144*) |
| MetaMorph Advanced Acquisition | // | Molecular Devices |
| ImageJ | 1.519 | (*145*) |

### 5.8.1. Processing of ribosome profiling sequencing data

Commands for processing and quality assessment of ribosome profiling and DiSP data were piped with R Markdown. A representative pipeline employed for processing of DiSP data of HEK293-T cells (main dataset of this study, replicate 2) is displayed below.

## Processing of Ribosome Profiling data
## DiSP in HEK cells Replicate 2

Matilde Bertolini
June 9, 2019

### Pipeline information:

This pipeline runs the data analysis and quality assessment of Ribosome Profiling Data. It analyzes short unpaired raw reads, requires FASTQ files as input and provides e-, p- and a-site assigned reads as output (HDF5 files). Custom Julia scripts evoked in this analysis (steps 2 and 6) are available as supplementary information in (*115*).
The pipeline performs following operations:

(i) Trimming of 3' adapter with Cutadapt
(ii) Isolation of Unique Molecular Identifiers (UMIs) with custom Julia script (developed by Dr. Ilia Kats), to remove PCR-duplicates
(iii) Removal of rRNA sequences with Bowtie2
(iv) Alignment of reads to the human genome with soft-clipping using STAR (performed on the NextSeq computer of the Bukau lab)
(v) Quality assessment with FastQC
(vi) Reads assignment (5'-end) with custom Julia script (developed by Dr. Ilia Kats), including removal of duplicated UMIs.

*Prelims*

```
knitr::opts_chunk$set(echo = TRUE)
library(reticulate)
library(purrr)
library(fastqcr)
use_python('/home/matilde/anaconda3/bin/python3', required=T)
Sys.setenv(params1 = params$path)
Sys.setenv(params2 = paste(params$samples, collapse=" "))
path <- params$path
samples <- params$samples
```

*Create necessary folder structure if it doesn't exist.*

```
cd $params1

declare -a StringArray=("01_Combined_Data" "02_Cutadapt_Trimmed_Data"
"03_Bowtie2_rRNA_Depleted_Data"
```

```
"04_STAR_Alignment_Data" "05_FastQC_Analysis" "06_Reads_Assignment")
for val in ${StringArray[@]}; do
  if [ ! -d $val ]; then
    mkdir $val;
  fi;
done

declare -a StringArray1=$params2
for val in ${StringArray1[@]}; do
  if [ ! -d "04_STAR_Alignment_Data/"$val ]; then
    mkdir "04_STAR_Alignment_Data/"$val;
  fi;
done

for val in ${StringArray1[@]}; do
  if [ ! -d "06_Reads_Assignment/"$val ]; then
    mkdir "06_Reads_Assignment/"$val;
  fi;
done
```

*Combine raw data from different lanes or sequencing runs*

```
cd $params1

declare -a StringArray=$params2
for val in ${StringArray[@]}; do
  if [ ! -f "01_Combined_Data/"$val'.fastq.gz' ]; then
    cat '00_Raw_Data/'$val*'.fastq.gz' > "01_Combined_Data/"$val'.fastq.gz';
    echo 'Raw data for '$val 'merged'
  else
    echo "Sample "$val": merged data already present";
  fi;
done

## Sample mono: merged data already present
## Sample dis: merged data already present
```

### (i) Trim 3' Adaptor with Cutadapt

*The adaptor is trimmed from the 3'-end of reads if at least 6 nt align to the given adaptor sequence. After trimming, reads are filtered based on sequencing quality and length.*

```
cd $params1

declare -a StringArray=$params2
for val in ${StringArray[@]}; do
  if [ ! -f "02_Cutadapt_Trimmed_Data/"$val*'.fastq.gz' ]; then
    echo "Sample "$val ": Running Cutadapt"
    /home/matilde/anaconda3/bin/cutadapt --cores=4 -q20 -m23 --discard-untrimmed -
O6 -a ATCGTAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -o
'02_Cutadapt_Trimmed_Data/'$val'_Cutadapt-Trimmed.fastq.gz'
'01_Combined_Data/'$val'.fastq.gz' 1>
'02_Cutadapt_Trimmed_Data/'$val'_Cutadapt_report.txt'
  else
    echo "Sample "$val ": Trimmed data already present"
  fi
done

## bash: line 4: [: 02_Cutadapt_Trimmed_Data/mono_Cutadapt-Trimmed.fastq.gz:
binary operator expected
```

```
## Sample mono : Trimmed data already present
## bash: line 4: [: 02_Cutadapt_Trimmed_Data/dis_Cutadapt-Trimmed.fastq.gz: binary
operator expected
## Sample dis : Trimmed data already present

cd $params1

declare -a StringArray=$params2
for val in ${StringArray[@]}; do
  report='02_Cutadapt_Trimmed_Data/'$val'_Cutadapt_report.txt'
  echo "Cutadapt report for sample $val"
  sed -n '7,20p' < "$report"
done


## Cutadapt report for sample mono
##
## Total reads processed:              82,025,457
## Reads with adapters:                59,491,583 (72.5%)
## Reads that were too short:              77,076 (0.1%)
## Reads written (passing filters):    59,439,821 (72.5%)
##
## Total basepairs processed: 4,173,993,725 bp
## Quality-trimmed:               31,049,154 bp (0.7%)
## Total written (filtered):  2,320,043,733 bp (55.6%)
##
## === Adapter 1 ===
##
## Sequence: ATCGTAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC; Type: regular 3'; Length:
39; Trimmed: 59491583 times.
##
## Cutadapt report for sample dis
##
## Total reads processed:             107,989,057
## Reads with adapters:                92,878,003 (86.0%)
## Reads that were too short:              78,931 (0.1%)
## Reads written (passing filters):    92,815,960 (85.9%)
##
## Total basepairs processed: 5,810,514,454 bp
## Quality-trimmed:               34,950,820 bp (0.6%)
## Total written (filtered):  3,621,390,408 bp (62.3%)
##
## === Adapter 1 ===
##
## Sequence: ATCGTAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC; Type: regular 3'; Length:
39; Trimmed: 92878003 times.
```

### (ii) Create UMI-aware trimmed data

*Unique Molecular Identifiers (UMIs) are random nucleotides (five in the linker adaptor and two in the reverse primer) that allow to identify individual footprints. Trimmed reads are parsed to extract the seven nucleotides (5 at the footprint's 3' end and 2 at the footprint's 5' end). The extracted UMIs are written next to each read's name in the output trimmed_umi-aware.fastq file and will be used in the final step (reads assignment) to identify and discard duplicates.*

```
cd $params1
declare -a StringArray=$params2
```

```
for val in ${StringArray[@]}; do
 if [ ! -f "02_Cutadapt_Trimmed_Data/"$val*'_trimmed_umi-aware.fastq.gz' ]; then
    echo "Sample "$val ": Creating umi-aware trimmed data"
    /usr/bin/julia/bin/julia
'/media/matilde/Matys_Disk/Data_Analysis_new/scripts/Julia_scripts/01_trimmed_read
s.jl' '02_Cutadapt_Trimmed_Data/'$val'_Cutadapt-Trimmed.fastq.gz'
'02_Cutadapt_Trimmed_Data/'$val'_trimmed_umi-aware.fastq.gz' --umi3 5 --umi5 2
  else
    echo "Sample "$val ": Umi-aware trimmed data already present"
  fi
done

## Sample mono : Umi-aware trimmed data already present
## Sample dis : Umi-aware trimmed data already present
```

### (iii) Remove rRNA sequences

*Trimmed reads are aligned to human rRNA sequences to assess the level of rRNA contamination. The unaligned sequences are processed further while the aligned rRNA reads are discarded.*

```
cd $params1

declare -a StringArray=$params2
for val in ${StringArray[@]}; do
  if [ ! -f "03_Bowtie2_rRNA_Depleted_Data/"$val*'_norRNA.fastq' ]; then
    echo "Sample "$val ": Removing rRNA sequences"
    bowtie2 -p3 -t -x
'/media/matilde/Matys_Disk/Data_Analysis_new/data_files/indexed_rRNA_23.06.19/inde
xed_rRNA' -q '02_Cutadapt_Trimmed_Data/'$val'_trimmed_umi-aware.fastq.gz' --un
'03_Bowtie2_rRNA_Depleted_Data/'$val'_norRNA.fastq' -L 13 -S /dev/null 2>
'03_Bowtie2_rRNA_Depleted_Data/'$val'_Bowtie2_report.txt'
  else
    echo "Sample "$val ": rRNA depleted data already present"
  fi
done

## Sample mono : rRNA depleted data already present
## Sample dis : rRNA depleted data already present

cd $params1

declare -a StringArray=$params2
for val in ${StringArray[@]}; do
  report='03_Bowtie2_rRNA_Depleted_Data/'$val'_Bowtie2_report.txt'
  echo "Bowtie2 report for sample $val
  "
  sed -n '5,10p' < "$report"
  echo "
  "
done

## Bowtie2 report for sample mono
##
## 59439821 reads; of these:
##   59439821 (100.00%) were unpaired; of these:
##     40389413 (67.95%) aligned 0 times
##     341380 (0.57%) aligned exactly 1 time
##     18709028 (31.48%) aligned >1 times
## 32.05% overall alignment rate
```

```
## Bowtie2 report for sample dis
##
## 92815960 reads; of these:
##   92815960 (100.00%) were unpaired; of these:
##     32483240 (35.00%) aligned 0 times
##     6192775 (6.67%) aligned exactly 1 time
##     54139945 (58.33%) aligned >1 times
## 65.00% overall alignment rate
##
```

### (iv) Reads alignment with STAR

*Reads alignment to the human genome (GRCh38.p12) was performed on the NextSeq computer using the STAR aligned:*

```
ssh kai@129.206.92.193 cd STAR-2.7.1a/source
```

```
./STAR –runThreadN 24 –genomeDir
/mnt/DataKramer/ilia/umi_mammalian/grch38_primary/star/ –readFilesIn
/mnt/DataKramer/Matilde_Kai_backup/star_maty/mono_norRNA.fastq –
outFilterMultimapNmax 1 –outFilterType BySJout –alignIntronMin 5 –
outFileNamePrefix /mnt/DataKramer/Matilde_Kai_backup/star_maty/mono/ –
outReadsUnmapped Fastx –outSAMtype BAM SortedByCoordinate –outSAMattributes All XS
–quantMode GeneCounts –twopassMode Basic
```

```
./STAR –runThreadN 24 –genomeDir
/mnt/DataKramer/ilia/umi_mammalian/grch38_primary/star/ –readFilesIn
/mnt/DataKramer/Matilde_Kai_backup/star_maty/dis_norRNA.fastq –
outFilterMultimapNmax 1 –outFilterType BySJout –alignIntronMin 5 –
outFileNamePrefix /mnt/DataKramer/Matilde_Kai_backup/star_maty/dis/ –
outReadsUnmapped Fastx –outSAMtype BAM SortedByCoordinate –outSAMattributes All XS
–quantMode GeneCounts –twopassMode Basic
```

### (v) Quality check with FastQC

*The quality of genome-aligned sequences was assessed by FastQC. First, BAM files are converted to SAM files and moved to the same (parent) folder*

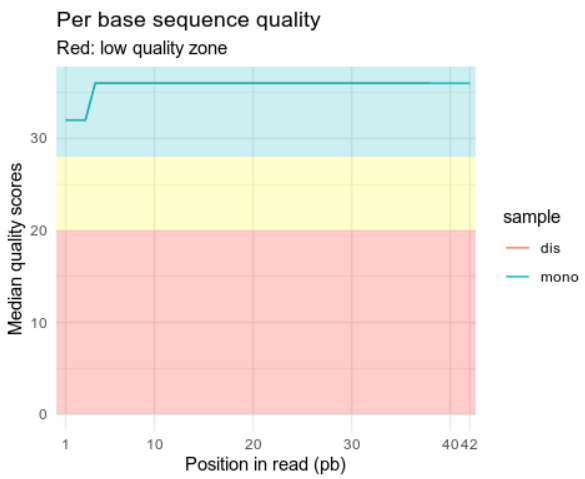
```
cd $params1

declare -a StringArray=$params2
for val in ${StringArray[@]}; do
  if [ ! -f "04_STAR_Alignment_Data/"$val"_Aligned.sam" ]; then
    echo "Sample "$val ": Converting to SAM"
    samtools view -h -o
"04_STAR_Alignment_Data/"$val"/Aligned.sortedByCoord.out.sam"
"04_STAR_Alignment_Data/"$val"/Aligned.sortedByCoord.out.bam"
  else
    echo "Sample "$val ": SAM files already present"
  fi
done

## Sample mono : SAM files already present
## Sample dis : SAM files already present

cd $params1

declare -a StringArray=$params2
for val in ${StringArray[@]}; do
  if [ ! -f "04_STAR_Alignment_Data/"$val'_Aligned.sam' ]; then
```

```
    mv "04_STAR_Alignment_Data/"$val"/Aligned.sortedByCoord.out.sam"
"04_STAR_Alignment_Data/Aligned.sortedByCoord.out.sam"
    mv "04_STAR_Alignment_Data/Aligned.sortedByCoord.out.sam"
"04_STAR_Alignment_Data/"$val'_Aligned.sam'
  else
    echo "Nothing to move"
  fi
done

## Nothing to move
## Nothing to move
```

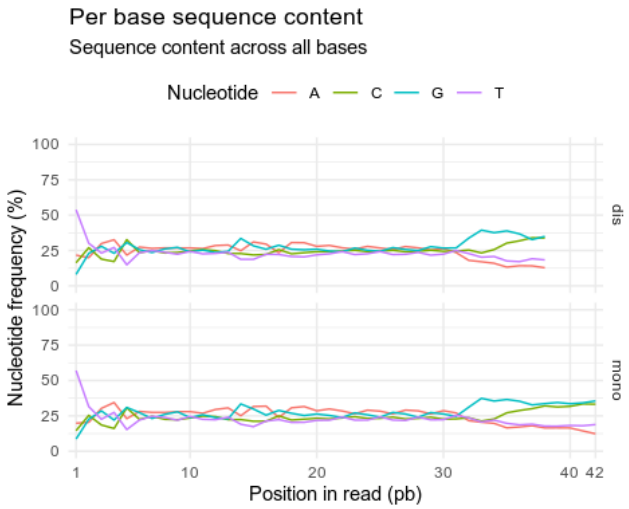
*Second, fastQC is invoked for all SAM files in the parent folder and results are displayed*

```
indir <- paste0(params$path, '/04_STAR_Alignment_Data/')
destfile <- paste0(params$path, '/05_FastQC_Analysis/', params$samples[2],
"_Aligned_fastqc.zip")
if(!file.exists(destfile)) {
  fastqc(fq.dir = indir, qc.dir = paste0(params$path, '/05_FastQC_Analysis'),
threads = 8, fastqc.path = "/usr/local/bin/fastqc")
} else {
print("FastQC was already run")
}

## [1] "FastQC was already run"

library(DT)
a <- paste0(params$path, '/05_FastQC_Analysis/')
qc <- qc_aggregate(a)
datatable(qc_stats(qc))
datatable(qc_problems(qc, "sample", compact = FALSE))

datatable(qc)
path <- params$path
samples <- params$samples
qc.dir <- paste0(path, '/05_FastQC_Analysis')
qc.files <- list.files(qc.dir, full.names = TRUE, pattern = ".zip")
l = length(qc.files)
qc <- qc_read_collection(qc.files, sample_names = params$samples)
qc_plot_collection(qc, c("Per sequence GC content", "Sequence Length
Distribution", "Per base sequence quality", "Per base sequence content", "Sequence
duplication levels"))
```

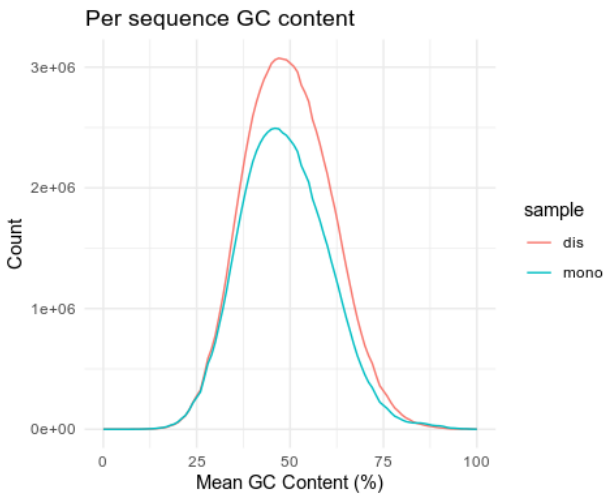
```
## $per_base_sequence_quality
```



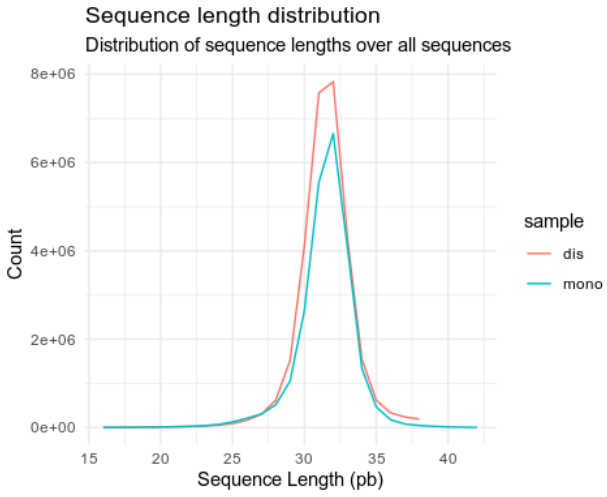
```
## $per_base_sequence_content
```



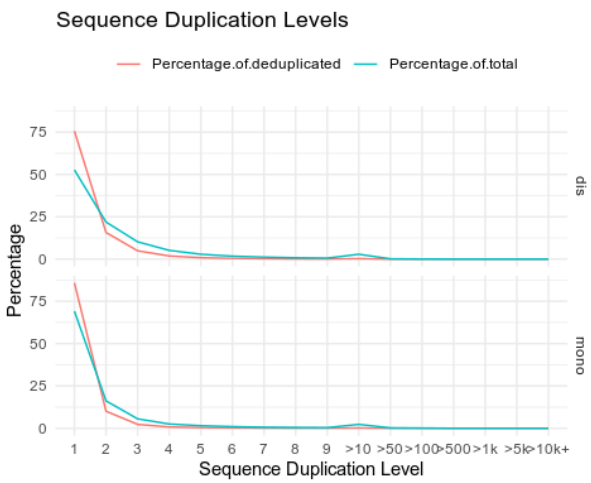

```
## $per_sequence_gc_content
```




```
## $sequence_length_distribution
```



```
## $sequence_duplication_levels
```




## (vi) Reads assignment

*Reads assignment is performed with a custom Julia script (developed by Dr. Ilia Kats). For each gene, the transcript with the longest coding sequence was selected and reads were assigned (a-, p-, e-site). We employed the script with following specifications: 5' end assignment, umi-collapsed output and soft-clipping.*

```
cd $params1

declare -a StringArray=$params2
for val in ${StringArray[@]}; do
  if [ ! -f "06_Reads_Assignment/"$val'/counts_umi.h5' ]; then
    /usr/bin/julia/bin/julia
/media/matilde/Matys_Disk/Data_Analysis_new/scripts/Julia_scripts/02_assign_count_
streaming_11.10.19.jl -g
/media/matilde/Matys_Disk/Data_Analysis_new/data_files/GCF_000001405.38_GRCh38.p12
_genomic.gff -u -c 1 -o '06_Reads_Assignment/'$val
'04_STAR_Alignment_Data/'$val'/Aligned.sortedByCoord.out.bam'
    echo "Running reads 5' assignment for sample: "$val
  else
    echo "Reads were already 5' assigned for sample: "$val
  fi
done

## Reads were already 5' assigned for sample: mono
## Reads were already 5' assigned for sample: dis
```

*The number of reads assigned to coding sequences and the fraction thereof with unique UMIs is displayed.*

```
import h5py

## /home/matilde/R/x86_64-pc-linux-gnu-
library/4.0/reticulate/python/rpytools/loader.py:24: FutureWarning: Conversion of
the second argument of issubdtype from `float` to `np.floating` is deprecated. In
future, it will be treated as `np.float64 == np.dtype(float).type`.
##   level=level

import numpy as np

for sample in r.samples:
```

```
  cds_all_psite = r.path + '/06_Reads_Assignment/' + sample + '/cds_all_psite.h5'
  cds_umi_psite = r.path + '/06_Reads_Assignment/' + sample + '/cds_umi_psite.h5'
  infiles = [cds_all_psite, cds_umi_psite]
  i = 0
  for f in infiles:
      with h5py.File(f, 'r') as hdf:
          ls = list(hdf.keys())
          total = 0
          for l in ls:
              data = hdf.get(l)
              dataset1 = np.array(data)
              reads = np.sum(dataset1[1])
              total += reads
      if i == 0:
          all_reads = total
          print(sample + ' - all reads in cds: ' + '{:.2e}'.format(all_reads))
      if i == 1:
          umi_reads = total
          percentage = (umi_reads / all_reads) * 100
          print(sample + ' - umi-collapsed reads in cds: ' +
'{:.2e}'.format(umi_reads) +
                ' (' + str(round(percentage, 2)) + '%)')
      i += 1

## mono - all reads in cds: 2.81e+07
## mono - umi-collapsed reads in cds: 2.05e+07 (73.13%)
## dis - all reads in cds: 2.24e+07
## dis - umi-collapsed reads in cds: 2.00e+07 (89.23%)
```

### 5.8.2. Standard analysis of Ribosome Profiling data

All downstream analyses were performed on the p-site assigned reads aligned to coding sequences only (output HDF5 files of our pipeline). Each HDF5 file contains gene-specific datasets, each consisting of a 2-row matrix (row 1: 1-based position along the coding sequence of that gene; row 2: number of p-site assigned reads at that position). The gene-specific dataset contains additional attributes, such as gene and protein identifiers, transcript isoform used for assignment, genomic location and length of the coding sequence.

**Single gene profiles**

Single gene density and enrichment profiles were plotted using the R package RiboSeqTools (available at: https://github.com/ilia-kats/RiboSeqTools and (*146*)).
This tool allows to calculate and plot position-wise density or enrichment confidence intervals along genes' coding sequences. Contrarily to common practice, where the position-wise average read density or enrichment are plotted, this method implements direct visualization of the statistical confidence of the depicted single gene profiles: densities or enrichments calculated from only few reads carry a high level of uncertainty and are indicated by large confidence intervals (vertical bars at each codon position).
Therefore, all single gene plots show the position-wise 95% Poisson confidence interval corrected for library size, smoothed with a 15-codon wide sliding window. The calculation of confidence intervals is described in more detail in (*115*).

**Metagene profiles**

Genes were first filtered based on sequencing coverage (summed coverage in monosome and disome raw data in two replicates should be ≥ 0.5 reads/codon, i.e. a coverage ≥ 0.25 reads/codon in average in single replicates). The contribution of each gene was expression normalized, by dividing the position-wise reads density of each gene to its normalized read density derived from total translatomes (in RPKM). Finally, metagene profiles were calculated and plotted using RiboSeqTools (available at: https://github.com/ilia-kats/RiboSeqTools and (*146*)). This tool calculates average (density) metagene profiles as the arithmetic mean at each nucleotide or codon position for each sample and replicate separately. Alternatively, metagene enrichment profiles are calculated as the position-wise enrichment of disome over monosome. Profiles are calculated from all genes (solid line in metagene plots) as well as from bootstrapping samples of genes (shading in metagene plots indicates the 95% bootstrapping confidence interval).

### 5.8.3. Identification of high confidence candidates

**Sigmoid fitting**

Three models of single gene enrichment profiles were defined: (i) constant enrichment, (ii) single sigmoidal enrichment and (iii) double sigmoidal enrichment (see chapter 3.3.1). The mathematical definition of these models can be found as Supplementary information in (*115*). For each gene, the best fitting model was selected using the Bayesian Information Criterion (BIC). For models (ii) and (iii), assembly onset was defined as the inflection point of the sigmoid curve. Genes fitting best to models (ii) or (iii) were considered “DiSP candidates”. Genes with onsets included in the first 30 or the last codons of the coding sequence were assigned to the “non co-co assembly proteome”.
This analysis was performed using a sigmoid fitting script developed by former postdoc Dr. Ilia Kats (which can be found as Supplementary material of (*115*)).
The script was invoked as follows:

```
julia <path_to_script>/Script3.jl <path_to_input>.hdf5
```

**Definition of co-co assembly classes**

DiSP candidates identified from the main DiSP experiment of HEK293-T cells are expected to display a significantly less sigmoidal enrichment profile in PK and Puromycin treated samples.
Within the Puromycin dataset, treated samples were compared to untreated ones and the effect of Puromycin treatment on the read counts before versus after assembly onset was determined (DiSP candidates and onsets determined from the main DiSP experiment of HEK293-T cells were used, see (*115*)).
For the PK dataset, a dose-response model was employed. Predictor values were determined for each PK concentration (dose) such that these values had a linear relationship with the response (loss of sigmoidal enrichment); the determined predictor values were then used in place of the PK concentrations and the effect of PK treatment on the read counts before and after onset was determined (as for the Puromycin analysis, DiSP candidates and onsets determined from the main DiSP experiment of HEK293-T cells were used, see (*115*)).
DiSP candidates were defined as Puromycin or PK sensitive if they showed a significant response (at FDR ≤ 0.01) to the respective treatment and if the calculated “effect” value was

negative (indicating that the sigmoidal enrichment was more flat in the treated compared to the untreated sample).
High confidence genes were defined as DiSP candidates that were sensitive to both Puromycin and PK controls according to this analysis and that encoded for cytosolic or nuclear proteins according to our custom annotation which combines information from several databases (see below).

### 5.8.4. Calculation of monosome depletion

To calculate the gene-wise monosome depletion after onset of co-co assembly, we first calculated the fraction of residual monosome density after compared to before onset and then defined monosome depletion as the complement of the residual monosome density (1 - residual monosome). To account for possible variations in the general ribosome occupancy along coding sequences (independent of monosome depletion), we normalized monosome reads to total translatome reads.

The complete formula, employed to compute monosome depletion *Depl(mono)* for each gene reads as follows:

$$Depl(mono) = 1 - \frac{\frac{mono2}{\text{total2}}}{\frac{mono1}{\text{total1}}}$$

***Equation 2.*** *Calculation of the gene-wise monosome depletion (efficiency of co-co assembly).*

*Mono2 and mono1 are the sum of reads in the monosome sample before and after onset, respectively; similarly, total2 and total1 are the sum of reads in the total translatome sample before and after onset, respectively. Note that "after onset" includes reads mapping downstream of the determined onset, until the end of the coding sequence (for single sigmoidal profiles) or until the end of co-co assembly (for double sigmoidal profiles). Total1 and total2 are calculated from averages of RPMs over replicates.*

**Randomized onsets**

To test significance of the observed monosome depletion values, the same analysis can be repeated with randomized onset positions.
A linear regression was fitted using the logarithm of the CDS length as predictor and the logarithm of the onset as response variable. A Normal distribution truncated to 1 and the CDS length can be defined using the linear regression prediction as mean and the regression's residual standard deviation as standard deviation. A new onset for each gene was drawn from this distribution for $10^4$ times. The median depletion of each iteration was calculated and the distribution of medians was plotted (Fig. 28C and F). If none of the median values from the randomized control is equal or higher the median value from the real data, the observed monosome depletion is statistically significant ($p < 10^{-4}$).

### 5.8.5. Analysis of the features of co-co assembly nascent proteins

**Annotation of the human proteome**

To achieve a complete annotation, the information derived from several databases has been employed (curated by Dr. Frank Tippmann).
For annotation of protein subcellular localization, following databases were merged: Human Proteome Atlas (*147*), UniProtKB (*148*), LOCATE (*149*), and the benchmark dataset of iLoc-Euk (*150*). Localization annotation of mouse and rat homologues were retrieved in case no information was present for the human protein. The resulting merged database was screened for following terms: *'cytosol', 'nucleoplasm', 'nucleus', 'cytoplasm', 'nucleoli', 'nucleolus', 'perinuclear region of cytoplasm'*. Proteins annotated with at least one of these keywords and with no transmembrane domain (TMD) annotated in UniprotKB were defined "cyto/nuclear".
For annotation of protein oligomeric state, we employed following databases: UniprotKB (*148*), PDB (*151*), Corum (*152*) and Swissmodel (*117*). A single annotation was selected for each protein based on the following hierarchical rules:

(i) Among different organisms we selected *human* > *mouse* > *rat*;
(ii) Among annotations from different databases we selected *UniprotKB* > *PDB* > *Corum* > *Swissmodel*;
(iii) Among different oligomeric states annotated within the same database, we selected *homomer* > *heteromer* > *monomer*.

Proteins that were annotated as heteromers of homomers were excluded from analysis (defined as NA) to minimize noise.

**Enrichment of annotated complex subunits**

The enrichment of complex subunits was determined by dividing the frequency of annotated subunit type (*homomer*, *heteromer* or *monomer*) in each assembly class (*high confidence* or *low confidence*) by their frequency in the respective background proteome. The background proteome of the high confidence class included all human cyto/nuclear proteins, while the background proteome of the low confidence class included all human proteins from any subcellular location. This ratio is the "Frequency Enrichment" in Fig. 19A and Table 2.
Because our definition of high- and low confidence proteomes relies on statistical methods, these classes are biased towards genes with higher sequencing cover (i.e. higher expression). Given that the same bias is intrinsic to annotation databases, the enrichment of annotated subunits in the co-co assembly classes would be artificially enhanced. To correct for this artefact, we performed bias-corrected significance calculation of enrichment, with the goseq package (*153*).

**Enrichment of annotated domains**

Annotation of protein domains was retrieved from UniprotKB ("Domain[FT]" and "Coiled coil" fields) and their annotated position within proteins was compared to the onset of co-co assembly (in codon-coordinates) detected by DiSP. Briefly, a domain was considered "exposed" at the onset of co-co assembly if *onset – 30 > 0* (i.e. if its N-terminal boundary is expected to have emerged from the ribosomal exit tunnel at the onset of assembly). To avoid identifying domains that are generally positioned in N-terminal protein regions and reveal meaningful co-co assembly-driving domains, we compared the domains that are "exposed" at the onset of high confidence candidates to similar N-terminal regions of the cyto/nuclear

proteome (by assigning randomized onsets as described above). Thus, following resampling approach was employed:

(i) The fraction of proteins in the high confidence class exposing domain *D* at the onset of assembly is determined (named *Fraction_highconf_D*);

(ii) A sample of proteins is drawn from the background proteome (all cyto/nuclear proteins) of the same size as the high confidence class (829 proteins) and a randomized onset is assigned to each protein of the sample (as described above);

(iii) The fraction of proteins in the background sample defined in step (ii) exposing domain *D* at the randomized onset of assembly is determined (named *Fraction_background_D*);

(iv) Steps (ii) and (iii) are repeated for $10^5$ iterations;

The median enrichment of domain *D* (Fold-change in Fig. 22A) was calculated as *Fraction_highconf_D* divided by the median of *Fraction_background_D* (across all iterations). The significance of enrichment (p-value) of domain *D* was defined as $(N+1)/(10^5+1)$, where N is the number of iterations in which *Fraction_background_D* ≥ *Fraction_highconf_D*. This analysis was performed for all annotated domains in parallel and p-values were adjusted for multiple comparisons with the Benjamini & Yekutieli method (*154*) ("adj. p-value" in Fig. 22A).

**Coiled coil prediction**

The DeepCoil tool (*120*) was employed to predict coiled coils based on amino acid sequences, by following developer's instructions (https://github.com/labstructbioinf/DeepCoil).

A custom python script was employed to extract protein sequences of 500 residues centred around assembly onsets of high confidence proteins (output: onset_aligned.fa). As a control, the same operation was repeated for cyto/nuclear proteins in the non co-co assembly proteome aligned to randomized assembly onsets (calculated as described above).

Finally, coiled coils were predicted on both high confidence and non co-co assembly sequences using following command:

```
python <path_to_script>/deepcoil.py -i
<path_to_infile>/onset_aligned.fa -out_path
<path_to_outfolder>/predictions_out/
```

**Analysis of crystal structures**

All crystal structures containing human proteins were retrieved from PDB (*151*). For each protein we selected the structure with highest sequence coverage and resolution. Structures of proteins shorter than 10 amino acids or that are not based on the 20 proteinogenic amino acids were excluded from analysis. For each residue in each structure, the solvent accessible surface area was computed using FreeSASA (https://freesasa.github.io) and structure interface analysis was conducted as previously described (Fig. 20B) (*73*). The same analysis was repeated to calculate onset-aligned inter-subunit interface enrichment in high confidence candidates (Fig. 19C). In this case, only interfaces located in the 250 amino acids upstream or downstream assembly onset were analysed. Homomeric interface analysis included only exclusively homomeric structures and considered contacts between proteins with the same Uniprot ID. Heteromeric interface analysis included only structures where none of the subunits was present multiple times and considered contacts between a high confidence protein and other proteins within the same structure with different Uniprot

IDs. Interface enrichment values were normalized by the arithmetic mean of all values and plotted (Interface enrichment in Fig. 19C).

**Other analyses and plots**

All analyses and plots employed custom Python and R scripts.

### 5.8.6. Analysis of RNA-seq data

Transcript abundance from RNA-seq reads (of the PP-seq dataset) was quantified using Salmon (*144*).
The latest Refseq transcriptome was downloaded from:
ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_latest/refseq_identifiers/
Coding sequences were extracted (GRCh38_rna_coding.fa) and indexed with following command:

```
salmon index -t <path_to_transcriptome>/GRCh38_rna_coding.fa -i
<path_to_index>/GRCh38_rna_coding_indexed
```

Raw reads were directly mapped to the indexed transcriptome with following command:

```
salmon quant -i <path_to_index>/GRCh38_rna_coding_indexed -l A \
-1 <path_to_raw_reads>/Read1.fastq.gz \
-2 <path_to_raw_reads>/Read2.fastq.gz \
-p 8 --validateMappings -o <path_to_outfolder>/
```

Finally, the Bioconductor package tximport (*155*) was employed to import and summarize transcript abundance to the gene level following developer's instructions (https://bioconductor.org/packages/release/bioc/vignettes/tximport/inst/doc/tximport.html).

# BIBLIOGRAPHY


1. C. A. Waudby, C. M. Dobson, J. Christodoulou, Nature and Regulation of Protein Folding on the Ribosome (2019), doi:10.1016/j.tibs.2019.06.008.

2. P. Ciryam, R. I. Morimoto, M. Vendruscolo, C. M. Dobson, E. P. O'brien, In vivo translation rates can substantially delay the cotranslational folding of the Escherichia coli cytosolic proteome, doi:10.1073/pnas.1213624110.

3. M. Liutkute, E. Samatova, M. V. Rodnina, Cotranslational folding of proteins on the ribosome. *Biomolecules*. **10** (2020), , doi:10.3390/biom10010097.

4. A. J. Samelson, E. Bolin, S. M. Costello, A. K. Sharma, E. P. O'Brien, S. Marqusee, Kinetic and structural comparison of a protein's cotranslational folding and refolding pathways. *Sci. Adv.* **4**, eaas9098 (2018).

5. M. S. Evans, I. M. Sander, P. L. Clark, Cotranslational Folding Promotes β-Helix Formation and Avoids Aggregation In Vivo. *J. Mol. Biol.* **383**, 683–692 (2008).

6. W. J. Netzer, F. U. Hartl, Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature*. **388**, 343–349 (1997).

7. G. Kramer, A. Shiber, B. Bukau, Mechanisms of Cotranslational Maturation of Newly Synthesized Proteins. *Annu. Rev. Biochem.* **88**, 337–364 (2019).

8. A. Karniel, D. Mrusek, W. Steinchen, O. Dym, G. Bange, E. Bibi, Co-translational Folding Intermediate Dictates Membrane Targeting of the Signal Recognition Particle Receptor. *J. Mol. Biol.* **430**, 1607–1620 (2018).

9. J. L. Chaney, P. L. Clark, Roles for Synonymous Codon Usage in Protein Biogenesis. *Annu. Rev. Biophys.* **44**, 143–166 (2015).

10. C. J. Tsai, Z. E. Sauna, C. Kimchi-Sarfaty, S. V. Ambudkar, M. M. Gottesman, R. Nussinov, Synonymous Mutations and Ribosome Stalling Can Lead to Altered Folding Pathways and Distinct Minima. *J. Mol. Biol.* **383** (2008), pp. 281–291.

11. G. Zhang, Z. Ignatova, Folding at the birth of the nascent chain: Coordinating translation with co-translational folding. *Curr. Opin. Struct. Biol.* **21** (2011), pp. 25–31.

12. K. Döring, N. Ahmed, T. Riemer, H. G. Suresh, Y. Vainshtein, M. Habich, J. Riemer, M. P. Mayer, E. P. O'Brien, G. Kramer, B. Bukau, Profiling Ssb-Nascent Chain Interactions Reveals Principles of Hsp70-Assisted Folding. *Cell*. **170**, 298-311.e20 (2017).

13. F. Brandt, S. A. Etchells, J. O. Ortiz, A. H. Elcock, F. U. Hartl, The Native 3D Organization of Bacterial Polysomes. *Cell*. **136**, 261–271 (2009).

14. F. Brandt, L. A. Carlson, F. U. Hartl, W. Baumeister, K. Grünewald, The Three-Dimensional Organization of Polyribosomes in Intact Human Cells. *Mol. Cell*. **39**, 560–569 (2010).

15. Z. A. Afonina, A. G. Myasnikov, N. F. Khabibullina, A. Y. Belorusova, J. F. Menetret, V. D. Vasiliev, B. P. Klaholz, V. A. Shirokov, A. S. Spirin, Topology of mRNA chain in isolated eukaryotic double-row polyribosomes. *Biochem.* **78**, 445–454 (2013).

16. Z. A. Afonina, A. G. Myasnikov, V. A. Shirokov, B. P. Klaholz, A. S. Spirin, Conformation transitions of eukaryotic polyribosomes during multi-round translation.

*Nucleic Acids Res.* **43**, 618–628 (2015).

17. A. G. Myasnikov, Z. A. Afonina, J. F. Ménétret, V. A. Shirokov, A. S. Spirin, B. P. Klaholz, The molecular structure of the left-handed supra-molecular helix of eukaryotic polyribosomes. *Nat. Commun.* **5** (2014), doi:10.1038/ncomms6294.

18. S. Duttler, S. Pechmann, J. Frydman, Principles of cotranslational ubiquitination and quality control at the ribosome. *Mol. Cell*. **50**, 379–393 (2013).

19. F. Wang, L. A. Durfee, J. M. Huibregtse, A cotranslational ubiquitination pathway for quality control of misfolded proteins. *Mol. Cell*. **50**, 368–378 (2013).

20. M. Bañó-Polo, C. Baeza-Delgado, S. Tamborero, A. Hazel, B. Grau, I. M. Nilsson, P. Whitley, J. C. Gumbart, G. von Heijne, I. Mingarro, Transmembrane but not soluble helices fold inside the ribosome tunnel. *Nat. Commun.* **9** (2018), doi:10.1038/s41467-018-07554-7.

21. A. Kosolapov, C. Deutsch, Tertiary interactions within the ribosomal exit tunnel. *Nat. Struct. Mol. Biol.* **16**, 405–411 (2009).

22. C. A. Woolhead, P. J. McCormick, A. E. Johnson, Nascent membrane and secretory proteins differ in FRET-detected folding far inside the ribosome and in their exposure to ribosomal proteins. *Cell*. **116**, 725–736 (2004).

23. W. Holtkamp, G. Kokic, M. Jäger, J. Mittelstaet, A. A. Komar, M. V. Rodnina, Cotranslational protein folding on the ribosome monitored in real time. *Science (80-. ).* **350**, 1104–1107 (2015).

24. D. Lucent, C. D. Snow, C. E. Aitken, V. S. Pande, Non-Bulk-Like Solvent Behavior in the Ribosome Exit Tunnel. *PLoS Comput Biol*. **6**, 1000963 (2010).

25. J. A. Farías-Rico, F. R. Selin, I. Myronidi, M. Frühauf, G. Von Heijne, Effects of protein size, thermodynamic stability, and net charge on cotranslational folding on the ribosome. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E9280–E9287 (2018).

26. D. N. Wilson, S. Arenz, R. Beckmann, Translation regulation via nascent polypeptide-mediated ribosome stalling. *Curr. Opin. Struct. Biol.* **37** (2016), pp. 123–133.

27. K. Ito, S. Chiba, Arrest peptides: Cis-acting modulators of translation. *Annu. Rev. Biochem.* **82** (2013), pp. 171–202.

28. D. A. Kelkar, A. Khushoo, Z. Yang, W. R. Skach, Kinetic analysis of ribosome-bound fluorescent proteins reveals an early, stable, cotranslational folding intermediate. *J. Biol. Chem.* **287**, 2568–2578 (2012).

29. A. M. Knight, P. H. Culviner, N. Kurt-Yilmaz, T. Zou, S. B. Ozkan, S. Cavagnero, Electrostatic effect of the ribosomal surface on nascent polypeptide dynamics. *ACS Chem. Biol.* **8**, 1195–1204 (2013).

30. E. P. O'Brien, J. Christodoulou, M. Vendruscolo, C. M. Dobson, New scenarios of protein folding can occur on the ribosome. *J. Am. Chem. Soc.* **133**, 513–526 (2011).

31. A. Hoffmann, A. H. Becker, B. Zachmann-Brand, E. Deuerling, B. Bukau, G. Kramer, Concerted Action of the Ribosome and the Associated Chaperone Trigger Factor Confines Nascent Polypeptide Folding. *Mol. Cell*. **48**, 63–74 (2012).

32. K. Liu, J. E. Rehfus, E. Mattson, C. M. Kaiser, The ribosome destabilizes native and non-native structures in a nascent multidomain protein. *Protein Sci.* **26**, 1439–1451 (2017).

33. K. Liu, K. Maciuba, C. M. Kaiser, The Ribosome Cooperates with a Chaperone to Guide Multi-domain Protein Folding. *Mol. Cell*. **74**, 310-319.e7 (2019).

34. K. Peisker, D. Braun, T. Wölfle, J. Hentschel, U. Fünfschilling, G. Fischer, A. Sickmann, S. Rospert, Ribosome-associated complex binds to ribosomes in close proximity of Rpl31 at the exit of the polypeptide tunnel in yeast. *Mol. Biol. Cell*. **19**, 5279–5288 (2008).

35. M. A. Hanebuth, R. Kityk, S. J. Fries, A. Jain, A. Kriel, V. Albanese, T. Frickey, C. Peter, M. P. Mayer, J. Frydman, E. Deuerling, Multivalent contacts of the Hsp70 Ssb contribute to its architecture on ribosomes and nascent chain interaction. *Nat. Commun.* **7**, 1–13 (2016).

36. P. Huang, M. Gautschi, W. Walter, S. Rospert, E. A. Craig, The Hsp70 Ssz1 modulates the function of the ribosome-associated J-protein Zuol. *Nat. Struct. Mol. Biol.* **12**, 497–504 (2005).

37. A. Koplin, S. Preissler, Y. Llina, M. Koch, A. Scior, M. Erhardt, E. Deuerling, A dual function for chaperones SSB-RAC and the NAC nascent polypeptide-associated complex on ribosomes. *J. Cell Biol.* **189**, 57–68 (2010).

38. J. Kirstein-Miles, A. Scior, E. Deuerling, R. I. Morimoto, The nascent polypeptide-associated complex is a key regulator of proteostasis. *EMBO J.* **32**, 1451–1468 (2013).

39. B. Wiedmann, H. Sakai, T. A. Davis, M. Wiedmann, A protein complex required for signal-sequence-specific sorting and translocation. *Nature*. **370**, 434–440 (1994).

40. K. C. Stein, J. Frydman, The stop-and-go traffic regulating protein biogenesis: How translation kinetics controls proteostasis. *J. Biol. Chem.* **294** (2019), pp. 2076–2084.

41. I. M. Sander, J. L. Chaney, P. L. Clark, Expanding anfinsen's principle: Contributions of synonymous codon selection to rational protein design. *J. Am. Chem. Soc.* **136**, 858–861 (2014).

42. I. M. Walsh, M. A. Bowman, I. F. Soto Santarriaga, A. Rodriguez, P. L. Clark, Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 3528–3534 (2020).

43. A. M. E. Cassaignau, L. D. Cabrita, J. Christodoulou, How Does the Ribosome Fold the Proteome? *Annu. Rev. Biochem.* **89** (2020), pp. 389–415.

44. W. M. Jacobs, E. I. Shakhnovich, Evidence of evolutionary selection for cotranslational folding. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 11434–11439 (2017).

45. J. L. Chaney, A. Steele, R. Carmichael, A. Rodriguez, A. T. Specht, K. Ngo, J. Li, S. Emrich, P. L. Clark, Widespread position-specific conservation of synonymous rare codons within coding sequences. *PLoS Comput. Biol.* **13** (2017), doi:10.1371/journal.pcbi.1005531.

46. S. Pechmann, J. Frydman, Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.* **20**, 237–243 (2013).

47. P. Han, Y. Shichino, T. Schneider-, Y. Mishima, T. Inada, Genome-wide Survey of Ribosome Collision. *CellReports*. **31**, 107610 (2020).

48. T. Zhao, Y.-M. Chen, Y. Li, J. Wang, S. Chen, N. Gao, W. Qian, Disome-seq reveals widespread ribosome collisions that recruit co-translational chaperones. *bioRxiv*, 746875 (2019).

49. N. R. Guydosh, R. Green, Dom34 rescues ribosomes in 3′ untranslated regions. *Cell*. **156**, 950–962 (2014).

50. A. B. Arpat, A. Liechti, M. De Matos, R. Dreos, P. Janich, D. Gatfield, Transcriptome-

wide sites of collided ribosomes reveal principles of translational pausing. *Genome Res.* **30**, 985–999 (2020).

51. A. S. Wolf, E. J. Grayhack, Asc1, homolog of human RACK1, prevents frameshifting in yeast by ribosomes stalled at CGA codon repeats. *RNA*. **21**, 935–945 (2015).

52. D. A. Drummond, C. O. Wilke, Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell*. **134**, 341–352 (2008).

53. H. Nakatogawa, K. Ito, The ribosomal exit tunnel functions as a discriminating gate. *Cell*. **108**, 629–636 (2002).

54. D. H. Goldman, C. M. Kaiser, A. Milin, M. Righini, I. Tinoco, C. Bustamante, Mechanical force releases nascent chain-mediated ribosome arrest in vitro and in vivo. *Science (80-. ).* **348**, 457–460 (2015).

55. F. Cymer, G. Von Heijne, Cotranslational folding of membrane proteins probed by arrest-peptide- mediated force measurements. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 14640–14645 (2013).

56. S. E. Leininger, F. Trovato, D. A. Nissley, E. P. O'Brien, Domain topology, stability, and translation speed determine mechanical force generation on the ribosome. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 5523–5532 (2019).

57. N. Ismail, R. Hedman, N. Schiller, G. Von Heijne, A biphasic pulling force acts on transmembrane helices during translocon-mediated membrane integration. *Nat. Struct. Mol. Biol.* **19**, 1018–1023 (2012).

58. F. Wruck, A. Katranidis, K. H. Nierhaus, G. Büldt, M. Hegner, Translation and folding of single proteins in real time. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E4399–E4407 (2017).

59. H. R. Harrington, M. H. Zimmer, L. M. Chamness, V. Nash, W. D. Penn, T. F. Miller, S. Mukhopadhyay, J. P. Schlebach, Cotranslational folding stimulates programmed ribosomal frameshifting in the alphavirus structural polyprotein. *J. Biol. Chem.* **295** (2020), pp. 6798–6808.

60. A. Schwarz, M. Beck, The Benefits of Cotranslational Assembly: A Structural Perspective. *Trends Cell Biol.* **29**, 791–803 (2019).

61. D. Zipser, Studies on the ribosome-bound β-galactosidase of Escherichia coli. *J. Mol. Biol.* **7**, 739–751 (1963).

62. A. B. Fulton, T. L'ecuyer, Cotranslational assembly of some cytoskeletal proteins: implications and prospects. *J. Cell Sci.* **105**, 867–871 (1993).

63. Y.-W. Shieh, P. Minguez, P. Bork, J. J. Auburger, D. L. Guilbride, G. Kramer, B. Bukau, Operon structure and cotranslational subunit association direct protein assembly in bacteria | Science. *Science*. **350**, 678–680 (2015).

64. J. N. Wells, L. T. Bergendahl, J. A. Marsh, Operon Gene Order Is Optimized for Ordered Protein Complex Assembly. *Cell Rep.* **14**, 679–685 (2016).

65. C. D. S. Duncan, J. Mata, Widespread cotranslational formation of protein complexes. *PLoS Genet.* **7** (2011), doi:10.1371/journal.pgen.1002398.

66. A. Shiber, K. Döring, U. Friedrich, K. Klann, D. Merker, M. Zedan, F. Tippmann, G. Kramer, B. Bukau, Cotranslational assembly of protein complexes in eukaryotes revealed by ribosome profiling. *Nature*. **561**, 268–272 (2018).

67. L. Lin, G. N. DeMartino, W. C. Greene, Cotranslational dimerization of the Rel homology domain of NF-κB1 generates p50–p105 heterodimers and is required for

effective p50 production. *EMBO J.* **19**, 4712–4722 (2000).

68. J. A. Marsh, S. A. Teichmann, J. D. Forman-Kay, Probing the diverse landscape of protein flexibility and binding. *Curr. Opin. Struct. Biol.* **22** (2012), pp. 643–650.

69. J. Marino, K. J. Buholzer, F. Zosel, D. Nettels, B. Schuler, Charge Interactions Can Dominate Coupled Folding and Binding on the Ribosome. *Biophys. J.* **115**, 996–1006 (2018).

70. C. D. S. Duncan, J. Mata, Widespread Cotranslational Formation of Protein Complexes. *PLoS Genet.* **7**, e1002398 (2011).

71. O. O. Panasenko, S. P. Somasekharan, Z. Villanyi, M. Zagatti, F. Bezrukov, R. Rashpa, J. Cornut, J. Iqbal, M. Longis, S. H. Carl, C. Peña, V. G. Panse, M. A. Collart, Co-translational assembly of proteasome subunits in NOT1-containing assemblysomes. *Nat. Struct. Mol. Biol.* **26**, 110–120 (2019).

72. L. Zhang, V. Paakkarinen, K. J. Van Wijk, E. M. Aro, Co-translational assembly of the D1 protein into photosystem II. *J. Biol. Chem.* **274**, 16062–16067 (1999).

73. E. Natan, T. Endoh, L. Haim-Vilmovsky, T. Flock, G. Chalancon, J. T. S. Hopper, B. Kintses, P. Horvath, L. Daruka, G. Fekete, C. Pál, B. Papp, E. Őszi, Z. Magyar, J. A. Marsh, A. H. Elcock, M. M. Babu, C. V Robinson, N. Sugimoto, S. A. Teichmann, E. Oszi, Z. Magyar, J. A. Marsh, A. H. Elcock, M. M. Babu, C. V Robinson, N. Sugimoto, S. A. Teichmann, Cotranslational protein assembly imposes evolutionary constraints on homomeric proteins. *Nat. Struct. Mol. Biol.* **25**, 279–288 (2018).

74. A. Halbach, H. Zhang, A. Wengi, Z. Jablonska, I. M. L. Gruber, R. E. Halbeisen, P. M. Dehé, P. Kemmeren, F. Holstege, V. Géli, A. P. Gerber, B. Dichtl, Cotranslational assembly of the yeast SET1C histone methyltransferase complex. *EMBO J.* **28**, 2959–2970 (2009).

75. I. Kamenova, P. Mukherjee, S. Conic, F. Mueller, F. El-Saafin, P. Bardot, J.-M. Garnier, D. Dembele, S. Capponi, H. T. M. Timmers, S. D. Vincent, L. Tora, Co-translational assembly of mammalian nuclear multisubunit complexes. *Nat. Commun.* **10**, 1–15 (2019).

76. R. Gilmore, M. C. Coffey, G. Leone, K. McLure, P. W. Lee, Co-translational trimerization of the reovirus cell attachment protein. *EMBO J.* **15**, 2651–2658 (1996).

77. S. D. Redick, J. E. Schwarzbauer, Rapid intracellular assembly of tenascin hexabrachions suggests a novel cotranslational process. *J. Cell Sci.* **108**, 1761–1769 (1995).

78. J. Mrazek, D. Toso, S. Ryazantsev, X. Zhang, Z. H. Zhou, B. C. Fernandez, V. A. Kickhoefer, L. H. Rome, Polyribosomes are molecular 3D nanoprinters that orchestrate the assembly of vault particles. *ACS Nano*. **8**, 11552–11559 (2014).

79. C. D. Nicholls, K. G. McLure, M. A. Shields, P. W. K. Lee, Biogenesis of p53 Involves Cotranslational Dimerization of Monomers and Posttranslational Dimerization of Dimers: IMPLICATIONS ON THE DOMINANT NEGATIVE EFFECT. *J. Biol. Chem.* **277**, 12937–12945 (2002).

80. L. Lin, G. N. DeMartino, W. C. Greene, Cotranslational biogenesis of NF-κB p50 by the 26S proteasome. *Cell*. **92**, 819–828 (1998).

81. F. Liu, D. K. Jones, W. J. de Lange, G. A. Robertson, Cotranslational association of mRNA encoding subunits of heteromeric ion channels. *Proc. Natl. Acad. Sci.* **113**, 4859–4864 (2016).

82. F. Cymer, G. Von Heijne, S. H. White, Mechanisms of integral membrane protein

insertion and folding. *J. Mol. Biol.* **427** (2015), pp. 999–1022.

83. J. Lu, J. M. Robinson, D. Edwards, C. Deutsch, T1−T1 Interactions Occur in ER Membranes while Nascent Kv Peptides Are Still Attached to Ribosomes †. *Biochemistry*. **40**, 10934–10946 (2001).

84. M. J. Feige, L. M. Hendershot, Quality control of integral membrane proteins by assembly-dependent membrane integration. *Mol. Cell*. **51**, 297–309 (2013).

85. G.-W. Li, D. Burkhardt, C. Gross, J. S. Weissman, Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. *Cell*. **157**, 624–635 (2014).

86. A. Rousseau, A. Bertolotti, Regulation of proteasome assembly and activity in health and disease. *Nat. Rev. Mol. Cell Biol.* **19**, 697–712 (2018).

87. L. A. Mingle, Localization of all seven messenger RNAs for the actin-polymerization nucleator Arp2/3 complex in the protrusions of fibroblasts. *J. Cell Sci.* **118**, 2425–2433 (2005).

88. W. R. Jeffery, C. R. Tomlinson, R. D. Brodeur, Localization of actin messenger RNA during early ascidian development. *Dev. Biol.* **99**, 408–417 (1983).

89. K. C. Martin, A. Ephrussi, mRNA Localization: Gene Expression in the Spatial Dimension. *Cell*. **136**, 719–730 (2009).

90. S. Kassem, Z. Villanyi, M. A. Collart, Not5-dependent co-translational assembly of Ada2 and Spt20 is essential for functional integrity of SAGA. *Nucleic Acids Res.* **45**, 1186–1199 (2017).

91. E. Natan, J. N. Wells, S. A. Teichmann, J. A. Marsh, Regulation, evolution and consequences of cotranslational protein complex assembly. *Curr. Opin. Struct. Biol.* **42**, 90–97 (2017).

92. R. Smith, Moving molecules: mRNA trafficking in mammalian oligodendrocytes and neurons. *Neuroscientist*. **10** (2004), pp. 495–500.

93. J. D. Keene, RNA regulons: Coordination of post-transcriptional events. *Nat. Rev. Genet.* **8** (2007), pp. 533–543.

94. J. Lui, L. M. Castelli, M. Pizzinga, C. E. Simpson, N. P. Hoyle, K. L. Bailey, S. G. Campbell, M. P. Ashe, Granules harboring translationally active mRNAs provide a platform for P-body formation following stress. *Cell Rep.* **9**, 944–954 (2014).

95. N. N. Batada, L. A. Shepp, D. O. Siegmund, Stochastic model of protein-protein interaction: Why signaling proteins need to be colocalized. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 6445–6449 (2004).

96. L. Chang, Y. Shav-Tal, T. Trcek, R. H. Singer, R. D. Goldman, Assembling an intermediate filament network by dynamic cotranslation. *J. Cell Biol.* **172**, 747–758 (2006).

97. B. Hampoelz, A. Schwarz, P. Ronchi, H. Bragulat-Teixidor, C. Tischer, I. Gaspar, A. Ephrussi, Y. Schwab, M. Beck, Nuclear Pores Assemble from Nucleoporin Condensates During Oogenesis. *Cell*. **179**, 671-686.e17 (2019).

98. N. K. Williams, B. Dichtl, Co-translational control of protein complex formation: a fundamental pathway of cellular organization? *Biochem. Soc. Trans.* **46**, 197–206 (2018).

99. E. Lécuyer, H. Yoshida, N. Parthasarathy, C. Alm, T. Babak, T. Cerovina, T. R. Hughes, P. Tomancak, H. M. Krause, Global Analysis of mRNA Localization Reveals

a Prominent Role in Organizing Cellular Architecture and Function. *Cell*. **131**, 174–187 (2007).

100. D. W. Reid, C. V. Nicchitta, Diversity and selectivity in mRNA translation on the endoplasmic reticulum. *Nat. Rev. Mol. Cell Biol.* **16** (2015), pp. 221–231.

101. A. M. Hoffman, C. V. Nicchitta, Meta-organization of Translation Centers Revealed by Proximity Mapping of Endoplasmic Reticulum Ribosome Interactors. *bioRxiv* (2018), p. 398669.

102. W. Ma, C. Mayr, A Membraneless Organelle Associated with the Endoplasmic Reticulum Enables 3′UTR-Mediated Protein-Protein Interactions. *Cell*. **175**, 1492-1506.e19 (2018).

103. K. L. Engel, A. Arora, R. Goering, H. Y. G. Lo, J. M. Taliaferro, Mechanisms and consequences of subcellular RNA localization across diverse cell types. *Traffic*. **21** (2020), pp. 404–418.

104. A. Bashirullah, S. R. Halsell, R. L. Cooperstock, M. Kloc, A. Karaiskakis, W. W. Fisher, F. Weili, J. K. Hamilton, L. D. Etkin, H. D. Lipshitz, Joint action of two RNA degradation pathways controls the timing of maternal transcript elimination at the midblastula transition in Drosophila melanogaster. *EMBO J.* **18**, 2610–2620 (1999).

105. M. Pizzinga, C. Bates, J. Lui, G. Forte, F. Morales-Polanco, E. Linney, B. Knotkova, B. Wilson, C. A. Solari, L. E. Berchowitz, P. Portela, M. P. Ashe, Translation factor mRNA granules direct protein synthetic capacity to regions of polarized growth. *J. Cell Biol.* **218**, 1564–1581 (2019).

106. M. Gama-Carvalho, N. L. Barbosa-Morais, A. S. Brodsky, P. A. Silver, M. Carmo-Fonseca, Genome-wide identification of functionally distinct subsets of cellular mRNAs associated with two nucleocytoplasmic-shuttling mammalian splicing factors. *Genome Biol.* **7** (2006), doi:10.1186/gb-2006-7-11-r113.

107. H. Hieronymus, P. A. Silver, Genome-wide analysis of RNA-protein interactions illustrates specificity of the mRNA export machinery. *Nat. Genet.* **33**, 155–161 (2003).

108. M. Tarbier, S. D. Mackowiak, J. Frade, S. Catuara-Solarz, I. Biryukova, E. Gelali, D. B. Menéndez, L. Zapata, S. Ossowski, M. Bienko, C. J. Gallant, M. R. Friedländer, Nuclear gene proximity and protein interactions shape transcript covariations in mammalian single cells. *Nat. Commun.* **11** (2020), doi:10.1038/s41467-020-19011-5.

109. W. B. Isaacs, A. B. Fulton, Cotranslational assembly of myosin heavy chain in developing cultured skeletal muscle. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 6174–6178 (1987).

110. W. B. Isaacs, I. S. Kim, A. Struve, A. B. Fulton, Biosynthesis of titin in cultured skeletal muscle cells. *J. Cell Biol.* **109**, 2189–2195 (1989).

111. W. B. Isaacs, R. K. Cook, J. C. Van Atta, C. M. Redmond, A. B. Fulton, Assembly of vimentin in cultured cells varies with cell type. *J. Biol. Chem.* **264**, 17953–17960 (1989).

112. T. J. L'Ecuyer, J. A. Noller, A. B. Fulton, Assembly of tropomyosin isoforms into the cytoskeleton of avian muscle cells. *Pediatr. Res.* **43**, 813–822 (1998).

113. E. Oh, A. H. Becker, A. Sandikci, D. Huber, R. Chaba, F. Gloge, R. J. Nichols, A. Typas, C. A. Gross, G. Kramer, J. S. Weissman, B. Bukau, Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell*. **147**, 1295–1308 (2011).

114. G. Blobel, D. Sabatini, Dissociation of Mammalian Polyribosomes into Subunits by

Puromycin. *Proc. Natl. Acad. Sci. U. S. A.* **68**, 390–394 (1971).

115. M. Bertolini, K. Fenzl, I. Kats, F. Wruck, F. Tippmann, J. Schmitt, J. J. Auburger, S. Tans, B. Bukau, G. Kramer, Interactions between nascent proteins translated by adjacent ribosomes drive homomer assembly. *Science (80-. ).* **in press** (2021), doi:10.1126/science.abc7151.

116. T. B. Stanishneva-Konovalova, N. I. Derkacheva, S. V Polevova, O. S. Sokolova, The Role of BAR Domain Proteins in the Regulation of Membrane Dynamics. *Acta Naturae*. **8**, 60–69 (2016).

117. A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. P. de Beer, C. Rempfer, L. Bordoli, R. Lepore, T. Schwede, SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).

118. A. J. Williams, S. C. Blacklow, T. Collins, The Zinc Finger-Associated SCAN Box Is a Conserved Oligomerization Domain. *Mol. Cell. Biol.* **19**, 8526–8535 (1999).

119. R. Rose, M. Weyand, M. Lammers, T. Ishizaki, M. R. Ahmadian, A. Wittinghofer, Structural and mechanistic insights into the interaction between Rho and mammalian Dia. *Nature*. **435**, 513–518 (2005).

120. J. Ludwiczak, A. Winski, K. Szczepaniak, V. Alva, S. Dunin-Horkawicz, DeepCoil-a fast and accurate prediction of coiled-coil domains in protein sequences. *Bioinformatics*. **35**, 2790–2795 (2019).

121. D. Thévenin, T. Lazarova, Stable interactions between the transmembrane domains of the adenosine A 2A receptor . *Protein Sci.* **17**, 1188–1199 (2008).

122. A. Werner, S. Iwasaki, C. A. McGourty, S. Medina-Ruiz, N. Teerikorpi, I. Fedrigo, N. T. Ingolia, M. Rape, Cell-fate determination by ubiquitin-dependent regulation of translation. *Nature*. **525**, 523–527 (2015).

123. J. Y. Ryu, H. U. Kim, S. Y. Lee, Human genes with a greater number of transcript variants tend to show biological features of housekeeping and essential genes. *Mol. Biosyst.* **11**, 2798–2807 (2015).

124. M. Schueler, M. Munschauer, L. H. Gregersen, A. Finzel, A. Loewer, W. Chen, M. Landthaler, C. Dieterich, Differential protein occupancy profiling of the mRNA transcriptome. *Genome Biol.* **15** (2014), doi:10.1186/gb-2014-15-1-r15.

125. E. D. Levy, S. Teichmann, in *Progress in Molecular Biology and Translational Science* (Elsevier B.V., 2013), vol. 117, pp. 25–51.

126. E. L. Mena, P. Jevtić, B. J. Greber, C. L. Gee, B. G. Lew, D. Akopian, E. Nogales, J. Kuriyan, M. Rape, Structural basis for dimerization quality control. *Nature*. **586**, 452–456 (2020).

127. T. Kolb, K. Maass, M. Hergt, U. Aebi, H. Herrmann, Lamin A and lamin C form homodimers and coexist in higher complex forms both in the nucleoplasmic fraction and in the lamina of cultured human cells. *Nucleus*. **2**, 425–433 (2011).

128. G. Schreiber, A. E. Keating, Protein binding specificity versus promiscuity. *Curr. Opin. Struct. Biol.* **21**, 50–61 (2011).

129. E. L. Huttlin, R. J. Bruckner, J. A. Paulo, J. R. Cannon, L. Ting, K. Baltier, G. Colby, F. Gebreab, M. P. Gygi, H. Parzen, J. Szpyt, S. Tam, G. Zarraga, L. Pontano-Vaites, S. Swarup, A. E. White, D. K. Schweppe, R. Rad, B. K. Erickson, R. A. Obar, K. G. Guruharsha, K. Li, S. Artavanis-Tsakonas, S. P. Gygi, J. W. Harper, Architecture of the human interactome defines protein communities and disease networks. *Nature*.

**545**, 505–509 (2017).

130. Q. Ye, H. J. Worman, Protein-protein interactions between human nuclear lamins expressed in yeast. *Exp. Cell Res.* **219**, 292–298 (1995).

131. E. L. Mena, R. A. S. Kjolby, R. A. Saxton, A. Werner, B. G. Lew, J. M. Boyle, R. Harland, M. Rape, Dimerization quality control ensures neuronal development and survival. *Science (80-. ).* **362**, eaap8236 (2018).

132. C. A. P. Joazeiro, Mechanisms and functions of ribosome-associated protein quality control. *Nat. Rev. Mol. Cell Biol.* **20** (2019), pp. 368–383.

133. A. Paix, A. Folkmann, D. H. Goldman, H. Kulaga, M. J. Grzelak, D. Rasoloson, S. Paidemarry, R. Green, R. R. Reed, G. Seydoux, Precision genome editing using synthesis-dependent repair of Cas9-induced DNA breaks. *Proc. Natl. Acad. Sci.* **114**, E10745–E10754 (2017).

134. A. Yeliseev, L. Zoubak, K. Gawrisch, Use of dual affinity tags for expression and purification of functional peripheral cannabinoid receptor. *Protein Expr. Purif.* **53**, 153–163 (2007).

135. N. J. MGlincy, N. T. Ingolia, Transcriptome-wide measurement of translation by ribosome profiling. *Methods*. **126**, 112–129 (2017).

136. C. V Galmozzi, D. Merker, U. A. Friedrich, K. Döring, G. Kramer, Selective ribosome profiling to study interactions of translating ribosomes in yeast. *Nat. Protoc.* **14**, 2279–2317 (2019).

137. F. Mueller, A. Senecal, K. Tantale, H. Marie-Nelly, N. Ly, O. Collin, E. Basyuk, E. Bertrand, X. Darzacq, C. Zimmer, FISH-quant: Automatic counting of transcripts in 3D FISH images. *Nat. Methods*. **10** (2013), pp. 277–278.

138. G. V Shivashankar, *"Immunoprecipitation of Nucleoskeletal Proteins" in Nuclear Mechanics and Genome Regulation, pages 111-112* (Academic Press, 2010; https://books.google.de/books?id=bz0s2xFedHkC), vol. 98.

139. E. C. Schirmer, L. Gerace, The Stability of the Nuclear Lamina Polymer Changes with the Composition of Lamin Subtypes According to Their Individual Binding Strengths. *J. Biol. Chem.* **279**, 42811–42817 (2004).

140. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. **17**, 10 (2011).

141. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. **9**, 357–359 (2012).

142. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2013).

143. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

144. R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*. **14**, 417–419 (2017).

145. C. A. Schneider, W. S. Rasband, K. W. Eliceiri, NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods*. **9** (2012), pp. 671–675.

146. ilia-kats. (2020, September 5). ilia-kats/RiboSeqTools: v0.1 (Version v0.1). Zenodo. http://doi.org/10.5281/zenodo.4016066.

147. P. J. Thul, L. Akesson, M. Wiking, D. Mahdessian, A. Geladaki, H. Ait Blal, T. Alm, A. Asplund, L. Björk, L. M. Breckels, A. Bäckström, F. Danielsson, L. Fagerberg, J. Fall, L. Gatto, C. Gnann, S. Hober, M. Hjelmare, F. Johansson, S. Lee, C. Lindskog, J. Mulder, C. M. Mulvey, P. Nilsson, P. Oksvold, J. Rockberg, R. Schutten, J. M. Schwenk, A. Sivertsson, E. Sjöstedt, M. Skogs, C. Stadler, D. P. Sullivan, H. Tegel, C. Winsnes, C. Zhang, M. Zwahlen, A. Mardinoglu, F. Pontén, K. Von Feilitzen, K. S. Lilley, M. Uhlén, E. Lundberg, A subcellular map of the human proteome. *Science (80-. ).* **356**, eaal3321 (2017).

148. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

149. J. Sprenger, J. L. Fink, S. Karunaratne, K. Hanson, N. A. Hamilton, R. D. Teasdale, LOCATE: a mammalian protein subcellular localization database. *Nucleic Acids Res.* **36**, D230–D233 (2008).

150. K.-C. Chou, Z.-C. Wu, X. Xiao, iLoc-Euk: A Multi-Label Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Eukaryotic Proteins. *PLoS One*. **6**, e18258 (2011).

151. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

152. M. Giurgiu, J. Reinhard, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, A. Ruepp, CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res.* **47**, D559–D563 (2019).

153. M. D. Young, M. J. Wakefield, G. K. Smyth, A. Oshlack, Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11**, R14 (2010).

154. Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).

155. C. Soneson, M. I. Love, M. D. Robinson, Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. *F1000Research*. **4** (2016), doi:10.12688/F1000RESEARCH.7563.2.

# PUBLICATIONS

M. Bertolini*, K. Fenzl*, I. Kats, F. Wruck, F. Tippmann, J. Schmitt, J. J. Auburger, S. Tans, B. Bukau†, G. Kramer†, **Interactions between nascent proteins translated by adjacent ribosomes drive homomer assembly**. Science (80-. ). in press (2021), doi:10.1126/science.abc7151.

## In preparation:

F. Wruck, J. Schmitt*, K. Fenzl*, M. Bertolini*, A. Katranidis, B. Bukau, G. Kramer, S. Tans†, **The conformational basis of Lamin co-translational assembly**.

J. Trendel, M. Bertolini, M. Aleksić, Jochem, S. Pfeffer, G. Kramer, B. Bukau, J. Krijgsveld†, **Translational Activity Controls Ribophagic Flux and Turnover of Distinct Ribosome Pools**.

* Contributed equally
† Corresponding author

# ACKNOWLEDGMENTS

Throughout my PhD studies I have received a great deal of scientific assistance and personal support from many people, without whom this work would have not been possible.

First, I am deeply grateful to my supervisor Prof. Dr. Bernd Bukau, for giving me the opportunity to work on this fascinating project and be part of such an exciting and diverse scientific environment as the Bukau lab. Your enthusiasm towards science and your support have inspired me and were of great importance.

I am very thankful to Prof. Dr. Claudio Joazeiro, the second referee of this Dissertation, for always engaging with excitement in scientific discussions and for providing invaluable personal and career advice.
I would also like to thank Prof. Dr. Walter Nickel for his helpful suggestions during the thesis advisory committee meetings and for being always available and friendly.
Many thanks to Prof. Dr. Irmgard Sinning for readily accepting to be part of my examination committee.

I deeply thank Dr. Ilia Kats, for his conceptual and practical contributions to this Thesis. Thank you for your patience and for teaching me how to think differently about data.
Thanks also to Dr. Frank Tippmann, for helpful contributions to this Thesis and for being a funny lab mate.
Thanks to the Heidelberg Biosciences International Graduate School (HBIGS) and Boehringer Ingelheim Fonds (BIF) for training, career support and for funding my PhD project.

I am especially thankful to Dr. Günter Kramer, who supervised this project closely and always with great excitement. Throughout my PhD, our scientific discussions have inspired me and helped me build a critical thinking, but most importantly, they were always fun. Thank you so much for being supportive on both professional and personal matters and for encouraging me to continue on this path.

I will never thank enough my colleague Kai Fenzl, for being a great working partner and a true friend. Thank you for sharing the good and the bad of this journey with me, for remaining a loyal, kind and honest person all the way and for making Germany feel like home to me.
A special “thank you” goes to Dr. Carla Galmozzi, for being a wonderful person and friend, and for her incredible ability to shine a light on any dark moment.
I am also truly grateful to all former and present members of the Bukau laboratory. Thank you for making a fun and never boring working environment!

Grazie di cuore ai miei genitori, Paolo e Donatella, e alle mie sorelle Chiara e Camilla, per essere un punto di riferimento fondamentale e per l’infinito affetto e sostegno con cui accompagnate la mia vita.

Matteo, grazie per aver affrontato questa avventura insieme a me, e per essere pronto ad affrontarne un’altra, insieme.