De novo proteins integrate cellular systems using ancient protein targeting and degradation pathways

Carly J. Houghton[1*], Nelson Castilho Coelho[1*], Annette Chiang[2*], Stefanie Hedayati[2], Saurin B. Parikh[1], Nejla Ozbaki-Yagan[2], Aaron Wacholder[1], John Iannotta[1], Alexis Berger[1], Matthew L. Wohlever[3], Anne-Ruxandra Carvunis[1¶] and Allyson F. O'Donnell[1,2¶]

[1]Pittsburgh Center for Evolutionary Biology and Medicine (CEBaM), Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15213, United States

[2]Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15260, United States

[3]Department of Cell Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15261, United States

*** These authors contributed equally**

[¶]Address correspondence to: allyod@pitt.edu or anc201@pitt.edu

Short title: Beneficial *de novo* emerging proteins preferentially localize to the ER

Keywords: *de novo* gene birth, *de novo* proteins, microproteins, protein targeting, protein degradation, protein localization, proto-genes, molecular evolution

Abbreviations:
BEP – Beneficial Emerging Proteins
ER – Endoplasmic Reticulum
SND - Srp-iNDependent targeting
GET – Guided Entry of Tail-anchored proteins
TA – Tail-Anchored
CTE – C-Terminal Extension
CHX – Cycloheximide
EGFP – Enhanced Green Fluorescent Protein
mNG – m-NeonGreen
TMD – Transmembrane Domain
Ub - Ubiquitination
ERAD – ER-associated degradation
ESCRT - Endosomal-sorting complex required for transport
TRC – Transmembrane Recognition Complex

EMC – ER Membrane Complex

Abstract

Recent evidence demonstrates that eukaryotic genomes encode thousands of evolutionarily novel proteins that originate *de novo* from non-coding DNA and can contribute to species-specific adaptations. Yet, it remains unclear how these incipient proteins—whose sequences are entirely new to nature—navigate the cellular environment to bring about beneficial traits. Here, we conduct a systematic *in vivo* investigation of *de novo* protein cell biology in yeast, revealing the early stages of *de novo* protein cellular integration. We show that beneficial *de novo* proteins exploit conserved membrane targeting, trafficking, and degradation pathways to integrate cellular systems. We find that beneficial *de novo* proteins frequently localize to membrane-bound compartments, particularly the endoplasmic reticulum (ER). Despite having unrelated sequences, ER-localized *de novo* proteins converge on a shared molecular signature: a C-terminal transmembrane domain that enables recognition by conserved post-translational ER insertion pathways. Subsequently, ER-localized *de novo* proteins traffic from the ER and their homeostasis is regulated by conserved proteasomal and vacuolar degradation pathways. Our findings demonstrate that ancient targeting and degradation pathways can accommodate young *de novo* proteins bearing a convergent molecular signature. These pathways could act as selective filters, biasing which young *de novo* proteins persist. The ER membrane may offer a "safe harbor" for certain classes of *de novo* proteins to acquire a selected function over time, thereby serving as a cradle for evolutionary innovation.

Main

New protein-coding genes can evolve *de novo* from sequences that were previously non-genic. Once considered rare, *de novo* gene birth is now recognized as a major force of molecular innovation and adaptation across species[1-4]. However, the mechanisms by which these *de novo* proteins integrate into cellular systems to provide fitness benefits remain poorly understood[4,5]. Ancient proteins have been coevolving for millions of years with the systems that help them fold and localize correctly, regulate their homeostasis, and enable their beneficial activities[6-9]. How are *de novo* proteins — naïve to these systems — recognized and processed by the cell such that their expression is not only tolerated, but beneficial (**Fig 1A**)?

We and others have hypothesized that membranes may offer a privileged entry point for *de novo* proteins to gain access to the cellular environment[10-14]. However, this hypothesis remains largely untested experimentally. Prior studies demonstrated that hydrophobic protein motifs arise frequently from non-coding DNA[10,15,16], which might facilitate *de novo* protein localization to membrane-bound organelles[10-14]. Membranes could allow hydrophobic *de novo* proteins to evolve local beneficial interactions while limiting potentially deleterious aggregation within the cytoplasm[10,11,14]. Several lines of evidence support this idea. In yeast, beneficial *de novo* emerging proteins (BEPs) that promote growth under nutrient stress are enriched in predicted transmembrane domains (TMDs)[10]. Although the subcellular localization of most *de novo* proteins is unknown, we found that one BEP inserts into the endoplasmic reticulum (ER) membrane[10]. A disproportionate fraction of *de novo* open-reading frames is transcriptionally co-expressed with membrane-associated genes[17]. In humans, newly identified microproteins—many of which have evolved *de novo*[18]—exhibit a high prevalence of predicted TMDs[14]. In arctic fish, a *de novo* protein exerts life-saving anti-freeze activity thanks to its ability to traffic through secretory pathway membranes[19]. While these are compelling observations across diverse organisms, no study has experimentally mapped how *de novo* proteins interact with the cell's machinery to navigate the cellular landscape; Many cell biology questions remain unanswered for these young proteins. Do TMD-containing *de novo* proteins in fact localize to cellular membranes as computational predictions suggest, and if so, which membranes? Are *de novo* proteins capable of utilizing the same systems that regulate targeting of the ancient proteome or do they employ other, yet-unknown mechanisms? Do *de novo* proteins undergo regulated degradation to modulate their protein levels, as occurs for ancient proteins? To answer these questions, systematic investigations into how *de novo* emerged proteins interact with the cellular machinery are needed.

Here, we sought to determine the role of membranes in the emergence and cellular integration of *de novo* proteins in the early stages of their transition from non-coding to coding DNA. To this aim, we leveraged yeast BEPs as a model set of *de novo* proteins and defined their specific subcellular localizations, targeting, and degradation mechanisms using a combination of cell biological and biochemical experiments with evolutionary sequence analyses. We find that BEPs preferentially localize to membrane-bound compartments, with a strong enrichment at the ER relative to the conserved proteome. BEP targeting to the ER depends on the ancient post-translational targeting machinery (Guided Entry of Tail-anchored proteins (GET) and Srp-iNDependent (SND) pathways, while BEP homeostasis at the ER is regulated by the conserved ER-associated degradation pathway (ERAD). We further show that ER-localized BEPs can exit the ER through the secretory pathway and reach the vacuole (the yeast lysosome equivalent). Remarkably, despite being entirely distinct at the sequence level, all ER-localized BEPs share a predicted C-terminal TMD that appears to enable engagement with conserved targeting pathways. This convergence on similar structural features, targeting mechanisms, subcellular localization and modes of degradation point to a common

evolutionary route for *de novo* proteins to integrate into the cell: through membranes and by harnessing ancient regulatory pathways. Over evolutionary timescales, these ancient pathways may act as a filter to constrain which *de novo* emerging sequences persist and contribute to molecular innovation.

BEPs preferentially localize to the ER and are susceptible to degradation

We first addressed how BEPs integrate into the cell by asking where they localize. To this aim, we made 26 C-terminal BEP-EGFP fusions and expressed them on plasmids under the control of an inducible promoter[20-22]. Microscopy revealed that a majority of BEPs (18/26) localized to membranous compartments (ER, mitochondria, nucleus, and plasma membrane) in at least some cells (**Fig 1B**). The ER was overrepresented among these compartments (**Fig 1B and Fig S1**). Seven out of twenty-six BEPs (27%) showed enriched ER localization in contrast to only ~ 8% of the ancient yeast proteome (Fisher exact test, odds ratio= 4.4, p= 3.1×10^{-3}; **Fig 1C**)[23]. Together these results demonstrate that while BEPs localize to a wide variety of cellular compartments, they are enriched in the ER.

In some cells, BEPs displayed diffuse, punctate, or no measurable fluorescence (**Fig 1B**). This could be due to plasmid loss, reduced transcript stability, reduced translation efficiency, or poor protein stability. Immunoblotting revealed a broad distribution of protein abundances and breakdown products (**Fig S2**), suggesting that BEPs are highly susceptible to degradation.

Given the unexpectedly high number of BEPs observed at the ER, we compared the phenotypic impacts of overexpressing ER-localized BEPs and other BEPs. ER-localized BEPs provided growth benefits across a broader array of nutrient stress conditions[10] than other BEPs (Mann Whitney U test, p= 0.025, n= 26; **Fig 1D**). These results reveal a strong association between ER localization and growth benefits among *de novo* emerging proteins in yeast. In what follows, we investigated how BEPs can be successfully targeted to the ER, as well as how their protein homeostasis can be regulated, despite possessing novel sequences that have had minimal opportunity to co-evolve with the cellular machinery.

ER-localized BEPs possess C-terminal transmembrane domains

First, we searched for potential molecular determinants within BEPs' amino acid sequences that may distinguish ER-localized BEPs from other BEPs. All BEPs are shorter, than 200 amino acids; 22/26 are under 150 amino acids, and 10/26 are shorter than 100, a common threshold for microprotein classification[24]. Most BEPs (21/26), whether they are ER localized or not, are predicted to encode TMDs by both Phobius[25] and TMHMM[26] (**Fig S3**)[10]. Structural predictions by Alpha-Fold2[27] were consistent with these predictions revealing widespread alpha helices throughout BEP structures (**Fig S4**). Thus, all the ER-localized BEPs display the biophysical potential to integrate into membranes due to the presence of predicted TMDs, but this feature alone cannot explain their specific ER localization, as it is also present in BEPs that localize elsewhere.

We therefore asked whether the protein sequences of ER-localized BEPs possessed unique biochemical properties distinguishing them from other BEPs. Proteins that access the ER co-translationally often contain a signal sequence to help direct them to the translocon[28]. However, no signal sequences were predicted in the ER-localized BEPs by TargetP 2.0[29] or SignalP 6.0[30]. Alternatively, ER proteins can be synthesized in the cytoplasm and then post-translationally inserted in the ER using the SND and/or GET pathways[31]. These pathways recognize TMDs in the middle or the C-terminus of proteins, respectively[31]. In addition,

membrane proteins targeted to the ER tend to have more hydrophobic TMDs and less positively charged C-tail extensions (CTEs) than proteins targeted to alternative membranes like the mitochondria[32,33]. To look for potential molecular determinants of ER localization in BEP sequences, we therefore measured the length, hydrophobicity, and positive charge in different sequence regions from all BEPs with predicted TMDs. We considered separately their full-length sequences, regions corresponding to their N or C terminal ends, or their predicted TMDs (**Fig 2A**).

We found that all 3 features cleanly distinguished ER-localized BEPs from the other TMD-containing BEPs. Consistent with a post-translational ER-insertion mechanism, ER-localized BEPs were predicted to have TMDs near their C-terminus, with short CTEs less than 26 amino acids[32,34] beyond the last predicted TMD (**Fig S3**). The predicted CTEs of ER-localized BEPs were significantly shorter than those of other TMD-containing BEPs (Mann-Whitney U test, $p = 0.036$, effect size $r = Z/\sqrt{N} = 0.48$; **Fig 2B, S5A**). In contrast, the overall lengths of ER-localized BEPs, as well as the lengths of their N-termini or TMDs, were statistically indistinguishable from those of other TMD-containing BEPs (**Fig 2B, S5A**). In addition to being shorter, the CTEs after the final TMD of ER-localized BEPs were also significantly less positively charged and more hydrophobic than those of other TMD-containing BEPs (**Fig 2C, Fig S5B-E**). The relative paucity of positively charged residues agrees with previous work on distinguishing characteristics of post-translationally targeted ER tail-anchored (TA) proteins in yeast[33]. The increased CTE hydrophobicity of ER-localized BEPs (**Fig 2C**, **Fig S5D-F**) is unexpected.

Upon closer inspection of the predicted structures (**Fig S4**), it becomes apparent that while several BEPs have unstructured CTEs, all ER-localized BEPs' CTEs end with either just a few hydrophobic residues after the TMD or an alpha-helical motif. Therefore, given the uncertainty in TMD boundary predictions[35] (**Fig S3**), some of the residues in the predicted CTEs of ER-localized BEPs could in fact belong to the their C-terminal TMDs. If so, the true TMDs of ER-localized BEPs would be even closer to their C termini than suggested by sequence-based predictions—which are already highly compatible with post-translational insertion mechanisms[34] (**Fig 2C**, **Fig S5F**).

Together our analyses show that all ER-localized BEPs are predicted to contain TMDs near their C-termini. Our results demonstrate that these evolutionarily novel proteins with unique primary sequences nevertheless share common protein features in their C-terminal regions. We thus hypothesized that these C-terminal regions serve as molecular determinants for BEP ER targeting and experimentally tested this hypothesis in a model ER-localized BEP.

The C-terminal region of a model BEP is sufficient for ER targeting

To experimentally test our hypothesis, we focused on the ER-localized BEP Ybr196c-a. This protein is a good model BEP for ER targeting as we previously showed that it is a bona-fide integral ER membrane protein[10]. Ybr196c-a has one of the longest predicted CTEs among all ER-localized BEPs (26 residues; **Fig S3**). We suspected that this predicted CTE might mask a second TMD that failed to be predicted since the C-terminus of Ybr196c-a was predicted to be alpha-helical (**Fig S4**). Evolutionary analyses of the Ybr196c-a locus across budding yeast species are also consistent with a second TMD[10], which is predicted in the homologous region of Ybr196c-a's orthologue in *S. paradoxus*[10].

To map the sequence determinants that dictate Ybr196c-a's ER localization, we assessed the localization of the full length and two discrete fragments of each orthologous Ybr196c-a sequence from *S. cerevisiae* and *S. paradoxus* spanning from: 1) the N-terminus to the end of the *S. cerevisiae*'s predicted TMD (Ybr196c-a 1.1), and 2) the end of the *S.*

cerevisiae's predicted TMD to the C-terminus of either species' sequences (Ybr196c-a 1.2) (**Fig 2D**). We integrated the full-length and partial sequences (1.1 and 1.2) fused with mNG into the *S. cerevisiae* genome under the control of an inducible promoter[22]. The full-length Ybr196c-a of both *S. cerevisiae* and *S. paradoxus* localized to the ER in 100% of cells (**Fig 2E**, left). In both species, the N-terminal sequence, Ybr196c-a 1.1, localized to the mitochondria, while the C-terminal sequence, Ybr196c-a 1.2, displayed clear localization to the ER (**Fig 2E**, right). These results suggest that the CTE of Ybr196c-a in *S. cerevisiae* contains a TMD that was missed by computational prediction yet enables integration into the ER membrane[10]. The fact that the C-terminal region of our model ER-localized BEP dictates its targeting to the ER is consistent with known mechanisms of post-translational insertion for conserved proteins[31]. Therefore, *de novo* emerging proteins possess features enabling ER localization that are akin to those of conserved proteins, suggesting they may use similar strategies for ER targeting.

ER targeting of BEPs depends on conserved post-translational machinery

To determine if our model ER-localized BEP, Ybr196c-a, reaches the ER via a canonical co- or post-translational targeting mechanism, we evaluated whether its localization was altered upon perturbing the three major eukaryotic ER targeting pathways (SRP-dependent, GET, and SND)[31]. To this aim, we engineered relevant gene deletions into strains expressing a chromosomally integrated Ybr196c-a-mNG fusion under the control of an inducible promoter, as in **Fig 2E**. Upon disruption of the core GET or SND machinery (Get1-3 or Snd2-3, respectively), fluorescence signal at the ER was reduced and an elevated number of cytosolic puncta was observed (**Fig 3A-C**, **Fig S6A**). This punctate localization pattern is expected for protein clients of these post-translational insertion pathways, as disruption of these targeting mechanisms results in cytosolic aggregation of client proteins[36,37]. In contrast, disruption of GET pathway accessory proteins (*e.g.,* Get4, Get5 and Sgt2) or the co-translational SRP-dependent targeting components (*e.g.,* Sbh1, Sec71, and Sec72) did not alter Ybr196c-a localization (**Fig S6A**). Ybr196c-a's reliance on operational GET and SND systems for its ER targeting was more apparent with the N-terminal than the C-terminal mNG-tagged forms (**Fig 3A-C**, **Fig S6A**). Perturbing GET or SND components had no impact on free mNG localization, ruling out effects of the fluorescent tag on Ybr196c-a's targeting (**Fig S6B**). Together, these findings demonstrate that Ybr196c-a depends on functional GET or SND machinery to access the ER.

To determine if these findings generalize to the other ER-localized BEPs, we examined their localization in the absence of Get3 and Snd3 using chromosomally integrated C and N-terminal mNG-fusions under the control of an inducible promoter. We found that, like Ybr196c-a, all the ER-localized BEPs were dependent on the GET or SND core machinery for their targeting to the ER (**Fig 3B-C** and **Fig S6C-D**). This effect was more pronounced when the BEPs were N-terminally tagged and their C-termini were accessible (**Fig 3B-C**). We hypothesize that this asymmetry is caused by the C-terminal mNG tags masking the TMDs needed for efficient post-translational ER insertion. Indeed, a similar effect as has been reported for other proteins[38] and linked to a favored topology of insertion where the C terminal end more efficiently translocates through the membrane[34]. Together these results reveal that all ER-localized BEPs depend on the core components of the GET and SND post-translational targeting pathways to reach the ER. While the BEPs are evolutionarily novel, the GET and SND pathways are conserved across eukarya[39].

Conserved ERAD pathways regulate BEP stability at the ER

Having established that BEPs depend on conserved pathways for ER localization, we next assessed whether they also rely on conserved pathways for their homeostasis. To this aim we first asked whether the abundance of our model ER-localized BEP, Ybr196c-a, depends on the two major ER-resident E3 ubiquitin (Ub) ligases involved in ERAD that specifically manage the surveillance of membrane proteins, Doa10 and Hrd1[40]. Immunoblotting for Ybr196c-a tagged either N- or C-terminally with mNG revealed a large increase in Ybr196c-a levels upon *DOA10* deletion (**Fig 4A-B**). This increase was not driven by the tag: free mNG was not stabilized by the loss of *DOA10* (**Fig S7A-B**), and a similar increase upon deletion of *DOA10* was observed when Ybr196c-a was fused to an HA tag (**Fig S7C-D**). We then examined the protein turnover of Ybr196c-a by blocking translation with the addition of cycloheximide (CHX)[41] and monitoring its protein abundance over time. We performed these CHX-chase assays using: 1) C-terminal mNG-tagged Ybr196c-a (**Fig 4C-D**), 2) free mNG (**Fig S7E-F**), 3) C-terminal HA-tagged Ybr196c-a (**Fig S7G-H**), and 4) N-terminal mNG-tagged Ybr196c-a (**Fig S7I-J**). From these assays, we confirmed that all versions of Ybr196c-a (N- or C-terminally-tagged with either mNG or HA) were stabilized by loss of Doa10. Doa10 ubiquitination of ER-resident proteins often results in their retrotranslocation from the ER and subsequent proteasomal degradation in the cytosoplasm[42]. Accordingly, we found that Ybr196c-a was more stable in the presence of a proteasome inhibitor than in vehicle control-treated cells (**Fig 4E-F**), unlike free mNG (**Fig S7K**) and irrespective of the positioning of the mNG tag (**Fig S7L**). Together, these experiments demonstrate that the protein stability of our model ER-localized BEP depends on the ER-resident Ub ligase Doa10 and proteasome-dependent degradation.

To confirm our findings and determine if they generalize to the other ER-localized BEPs, we examined their abundance using microscopy in the absence of Doa10 and/or Hrd1 using chromosomally integrated C- and N-terminal mNG fusions under the control of an inducible promoter. Using this approach, we confirmed that Ybr196c-a abundance and ER localization increases in the absence of Doa10, irrespective of tag location (**Fig 5A-B**). Quantification of whole-cell fluorescence from these images revealed Ybr196c-a's abundance also significantly increased upon loss of Hrd1, but less so than in the Doa10 deletion cells (**Fig 5C**). Like Ybr196c-a, we found that all the ER-localized BEPs increased in abundance and ER localization upon loss of at least one of the Ub ligases; immunoblotting revealed distinct bands (**Fig S8A-E**), and microscopy revealed increased whole cell fluorescence and ER localization (**Fig S9**). Based on immunoblotting and microscopy, protein stabilization was more pronounced for N-terminally tagged BEPs, which would make their C-termini accessible (**Fig 5D**, **Fig S8-S9**). While a larger fraction of cells with distinguishable ER localization was observed for all C-terminally tagged BEPs upon Ub ligase deletion, whole-cell fluorescence increased for only three of these BEPs (**Fig 5D**, **Fig S9**). Collectively these results demonstrate that the homeostasis of all ER-localized BEPs depends on the major ER-resident Ub ligases, which are widely conserved in eukarya[43,44].

Vacuolar protein sorting machinery mediates downstream BEP homeostasis

Finally, we explored the possibility that the homeostasis of ER-localized BEPs might also be regulated through proteasome-independent mechanisms in the vacuole, the yeast equivalent of the lysosome[45]. In the absence of the master vacuolar protease Pep4, fluorescently-tagged proteins targeted to this organelle accumulate vacuolar fluorescence because the fluorophore is not degraded efficiently[46-49]. Fluorescence microscopy revealed that all ER-localized BEPs accumulate in the vacuole when Pep4 is deleted, suggesting they are all able to transit the secretory pathway to the vacuole (**Fig 6A-B** and **Fig S10A-G**). The two known transit routes

from the ER to the vacuole are via ER-phagy or vesicle-mediated trafficking[50]. In the case of vesicular trafficking, disruption of vacuolar protein sorting via loss of Vps4, an ESCRT-III (endosomal-sorting complex required for transport) component[51], will prevent sorting of membrane proteins into multi-vesicular bodies and block their delivery into the vacuole lumen. If ER-localized BEPs rely on vesicle-mediated trafficking to the vacuole, then loss of Vps4 will block their accumulation in the vacuolar lumen, instead leading to their accumulation in endosomes adjacent to the vacuole[52]. Consistently, deletion of Vps4 in cells lacking Pep4 prevented accumulation of mNG signal in the vacuole and led to the formation of large fluorescent puncta adjacent to the vacuole (**Fig 6C-D**, **Fig S10H-L**). These findings imply that ER-localized BEPs traffic through the secretory and/or endocytic pathways beyond the ER. Altogether our results demonstrate that the homeostasis of ER-localized BEPs is not only dependent on the ERAD pathway and the proteasome, but also on the vacuolar protein sorting machinery (**Fig 6E**), which predates the origin of eukaryotes[53].

Discussion

In this study, we report the first experimental investigation of the cellular processing of *de novo* proteins, revealing the early stages of *de novo* protein cellular integration. Our results demonstrate that ancient targeting, trafficking and degradation pathways control the homeostasis of yeast beneficial emerging proteins (BEPs), a sub class of *de novo* proteins. Specifically, the GET and SND post-translational membrane insertion pathways, which are required for BEP insertion into the ER, as well as the ERAD pathway, which controls BEP levels, are conserved across eukarya[39,43,44]. The ESCRT-mediated mechanism that enables BEP trafficking to the vacuole predates the last universal eukaryotic ancestor[53]. Our discovery that BEPs co-opt cellular pathways with deep conservation suggests that these same pathways, or similar mechanisms, may facilitate the integration of *de novo* proteins into pre-existing cellular systems across the diverse taxa where *de novo* gene birth occurs[2]. Furthermore, since eukaryotic genomes encode thousands of *de novo* proteins[54-56], our findings raise the question of whether the *de novo* proteome may have shaped the early evolution of targeting systems and the molecular signals they recognize.

We find a majority of BEPs are capable of localizing to membrane-bound organelles and are four times as likely to localize to the ER than ancient proteins. The strong association between ER localization and growth benefit that we uncovered suggests new hypotheses regarding the mechanisms of *de novo* protein adaptability. For example, conserved human ER-localized microproteins, also short in length, often regulate larger protein complexes and engage in protein interactions[57]. Evolutionarily novel ER proteins may function through similar mechanisms. Targeting to the ER is somewhat promiscuous, relying on the recognition of crude molecular signatures (TMDs)[58,59] that are common in non-coding DNA[10,15,16] and accessible without extensive prior adaptation. Generally, non-genic sequences enriched in thymine such as poly dA/dT tracks tend to encode hydrophobic amino acids and have a high propensity to form TMDs[10,15,16]. It is plausible that *de novo* proteins regularly emerge with biochemical properties that support their membrane targeting. For example, Ybr196c-a emerged from a thymine-rich promoter sequence that already possessed the capacity to form a C-terminal TMD and the efficiency of its ER targeting may have increased over time[10]. Future studies will be needed to determine the precise evolutionary mechanism by which BEPs acquired their shared capability to access the ER, whether spontaneously upon *de novo* emergence, under the action of natural selection, or a combination of both[60].

We find all ER-localized BEPs are subject to degradation by ERAD and vacuolar proteases. The significance of this finding warrants careful interpretation in the context of *de novo*

emergence. Degradation could be construed as a barrier to the emergence of *de novo* genes, but we propose an alternative view: regulated degradation may facilitate *de novo* gene birth by mitigating the potentially negative effects of high levels of *de novo* protein expression before cells have adapted to their function. Indeed, regulated degradation modulates protein abundance without eliminating all molecules from the cell. As expression levels of incipient *de novo* proteins increase over evolutionary time[61], regulated degradation could reduce the likelihood of toxicity while also allowing selection to act on potential adaptive effects of the intact proteins. We speculate that such balance between degradation rates and expression levels likely explains how BEPs are simultaneously highly susceptible to degradation and capable of mediating adaptive phenotypes when overexpressed[10].

Furthermore, an evolutionary model where the ability to co-opt degradation pathways represents a selective advantage for young *de novo* proteins would align mechanistically with our discovery that BEPs tend to localize at the ER, an ancient organelle well-equipped for stringent protein quality control of nascent transmembrane proteins. The ER membranes may serve as a "safe harbor"[11] for *de novo* emerging TMD proteins, facilitating their evolutionary persistence by offering a prospective adaptive niche while ensuring that excess proteins are efficiently cleared to mitigate toxicity and reduce purifying selection. Our data show that BEPs' hydrophobic C-terminal TMDs are involved in their ER targeting. It is possible that the same sequences, when exposed in the cytoplasm, also act as degrons. Indeed, recent screens and computational analyses suggested that hydrophobic C-termini can act as signals for proteasomal degradation and influence the evolution of human noncanonical translated sequences[12,13,62]. BEPs' hydrophobic C-terminal TMDs may therefore disproportionately impact the cellular and evolutionary fate of incipient *de novo* TMD proteins by mediating their recognition by both targeting pathways, which enable navigating the cellular environment, and degradation pathways, which tune protein levels to allow safe increases in protein abundance over evolutionary time.

More generally, our findings offer new directions of research at the intersection of cell biology and *de novo* gene birth. For example, we show that some BEPs escape the ER and enter the secretory pathway. Accordingly, ER-localized BEPs lack the canonical KDEL retrieval sequence[63] and so will not be retrieved if they are incorporated into vesicles leaving the ER. Their passive incorporation into ER-derived vesicles may be more likely under conditions of cargo crowding, which improves sorting fidelity and increases bulk flow[64]. Both the specific trafficking route BEPs take to the vacuole as well as the cellular context could ultimately facilitate the evolution of novel activities for these proteins outside of the ER. As another example, we discovered two distinct targeting potentials exist within Ybr196c-a for the ER and the mitochondria. Consistent with this observation, ER and mitochondrial targeting pathways can actively compete for clients and the biochemical properties of amino acids surrounding the TMD profoundly influence membrane sorting [33,59,65]. Therefore, mechanistic research on recently-emerged *de novo* proteins, which are unburdened by evolutionary history and homology, may prove useful in defining the biochemical parameters that dictate cargo recognition and targeting to distinct organelles and improve our understanding and predictive ability in diverse areas of cellular biology.

Finally, our findings reveal that *de novo* proteins can display molecular convergence despite having unique primary sequences that are new to nature. ER-localized BEPs converge on similar molecular signatures, targeting mechanisms, modes of degradation, and phenotypic outcomes, pointing to a common, favored route of integration in the cell. Over evolutionary timescales, such favored routes could act as filters for *de novo* emerging sequences, ultimately constraining molecular innovation. By channeling proteins through specific pathways, cells may bias the molecular features—and thus the functional capacities—of proteins that successfully emerge *de novo*. As a result, the functional repertoire of young genes may

represent a biased subset of the molecular activities conserved proteins can perform. Our study paves the way towards predictive models linking molecular signatures, cellular integration, and phenotypic impacts of *de novo* emerging proteins across species and environmental contexts.

Materials and Methods

Plasmids and DNA manipulations

Gateway entry clones

To create a collection of entry clones for cloning with the Gateway System (Thermo Fisher, Waltham, MA), primers were designed containing the attB1 and attB2 sequences to amplify each ORF with a stop codon for the N' terminal tagging and without, for the C' terminal tagging. The primers were synthesized by Integrated DNA Technologies (IDT, Coralville, IA) and used to amplify each ORF using Q5 High-Fidelity 2x Master Mix (M0492L, New England BioLabs, Ipswich, MA). Genomic DNA from the strain FY4 was extracted with a Yeast DNA Extraction kit (78870, Thermo Fisher) and used as a template for the PCR. The PCR conditions were as follows: 98°C for 30 s, followed by 25 cycles of 98°C for 10 s, 55°C for 15 s and 72°C for 30 s and a final elongation step of 2 min at 72°C. The PCR products were purified with NucleoFast 96 PCR Plates (743100.10, Macherey Nagel, Allentown, PA) and quantified on a Nanodrop (Thermo Fisher) prior to use for recombination with the donor plasmid pDONR223, using BP Clonase II Enzyme mix (11789100, ThermoFisher). The recombination reactions were used to transform DH5α competent cells (C2987U, New England BioLabs) and positive clones were selected in LB media[66] supplemented with 50 µg/ml of spectinomycin (158993, MP Biomedicals, Santa Ana, CA). The NucleoSpin 96 Plasmid kit (740625.4, Macherey Nagel) was used to extract the plasmids and DNA quantification was done using the SpectraMax M4 plate reader (Molecular Devices, San Jose, CA).

Destination plasmids

Two sets of destination plasmids were used in this study. The first set from the Yeast Gateway kit (1000000011, Addgene, Watertown, MA[67]) includes the plasmids pAG415-GAL-ccdB-EGFP and pAG415-GAL-EGFP-ccdB. The second set was built to swap the GAL promoter with the Z_3EV promoter and the ccdB-EGFP was replaced with ccdB-mNG-Tadh1 and a hygromycin cassette (for the C terminally tag) or mNG-ccdB-Tadh1and a hygromycin cassette (for the N terminally tag), creating the destination plasmids pARC0031 and pARC0152.

Expression plasmids

Expression plasmids were made by LR recombination between the entry clones and the destination plasmids described above, using the LR Clonase II enzyme mix (11791020, Thermo Fisher). The recombination reactions were used to transform DH5α competent cells (C2987U, New England BioLabs) and positive clones were selected in LB media supplemented with 100 µg/ml of ampicillin (J60977, Alfa Aesar, Haverfill, MA). The NucleoSpin 96 Plasmid kit (740625.4, Macherey Nagel) was used to extract the plasmids and DNA quantification was

done using the SpectraMax M4 plate reader (Molecular Devices). We attempted to create expression plasmids for all 28 BEPs from Vakirlis et al[10], but only succeeded in constructing a final plasmid containing a BEP-EGFP fusion for 26 of the BEPs (YGR296W and YOR011W-A failed). Plasmids used in this work are described in **Supplemental Table 1**.

Yeast strains and growth conditions

The yeast strains used are described in **Supplemental Table 2** and are all derived from S288c genetic backgrounds of *S. cerevisiae*.

Yeast cells were grown in either synthetic complete medium (SC) lacking the appropriate amino acids for plasmid selection, prepared as described in[68] and using ammonium sulfate as a nitrogen source, or YPD[66] medium where indicated. Liquid medium was filter-sterilized and solid plate medium had 2% agar (w/v) added before autoclaving. When necessary for selection, hygromycin B or G-418 (H75020-1.0 and G64000-5.0, Research Products International, Mt Prospect, IL) was added to the media to a final concentration of 200 µg/ml. Yeast cells were grown at 30°C and where appropriate, 10-20 µM of β-estradiol (E2758-1G, Millipore-Sigma, St. Louis, MO) was added to cultures for 3 hours to induce expression from the *GEV* or *Z_3EV* promoters[22].

Yeast Transformation

Strains containing plasmids overexpressing fusions with EGFP were transformed in the BY4741 background with or without *ACT1pr*-Z_3EV-NatMX to be used with β-estradiol or galactose induction, respectively. All other strains were built based on the DBY12394 containing the *ACT1pr*-Z_3EV-NatMX expression system generously provided by the Noyes lab[22].

The constructs or DNA cassettes amplified from plasmids were integrated in the genome using the lithium acetate (LiAc), polyethyleneglycol (PEG) and salmon sperm DNA transformation protocol[69] with an adaptation for high-throughput scale. The yeast strain was grown in 1ml of YPD media in 96-deep-well plates and used for transformation with the liquid-handler EVO 150 (Tecan Group Ltd, Switzerland). Cells were washed in 750 µl of water, followed by 1 ml of LiAc tris-EDTA (TE) buffer and finally resuspended in 500 µl of the same buffer. Transformations were carried out using 5 µl of ssDNA (2 mg/ml) with 1 µg of purified PCR product or 500 ng of plasmid and 50 µl of cells. 250 µl of PLATE (40% PEG, 0.1M LiAc, 10mM Tris-HCl, 1mM EDTA) was added to each cell mixture and incubated at 30°C for 15 min, followed by 60 min at 42°C. Cells were then pelleted, resuspended in 200 µl of YPD or SC media for a recovery step at 30°C for 3 h, prior to being plated as 3 µl drops on selective medium. Plates were incubated at 30°C for 2-4 days until transformants grew. Agar plates were then used to pin into liquid media with the Singer RoToR (Singer Instruments, Somerset, UK) and then used to make glycerol stocks.

The constructs containing the *Z_3EVpr* followed by the coding sequences for individual BEPs tagged with mNG at the N' or C' terminus were amplified using the expression plasmids as template and integrated at the *HIS3* chromosomal locus. Amplification was carried out using Q5 High-Fidelity 2x Master Mix (M0492L, New England BioLabs) and primers ARC338 (TCTATATTTTTTTATGCCTCGGTAATGATTTTCATTTTTTTTTTTCCACCTAGCGGA TGACTCTTTTTTTTTCTTAGCGATTGGCATTATCACATAATGAATTATACATTATA TAAAGTAATGTGATTTCTTCGAAGAATATACTAAAAAATGAGCAGGCAAGATAA ACGAAGGCAAAGacaaaagctggagctctagta) and ARC339

(AAAGAAAAAGCAAAAAGAAAAAAGGAAAGCGCGCCTCGTTCAGAATGACACGT
ATAGAATGATGCATTACCTTGTCATCTTCAGTATCATACTGTTCGTATACATACTT
ACTGACATTCATAGGTATACATATATACACATGTATATATATCGTATGCTGCAGC
TTTAAATAATCGGTGTCAgcgaattgggtaccggcc) in a 50 µl reaction. The PCR conditions were as follows: 98°C for 30 s, followed by 25 cycles of 98°C for 10 s, 56°C for 15 s and 72°C for 2 min and a final elongation step of 5 min at 72°C. The PCR products were incubated at 37°C with 1µl of *Dpn*I (R0176S, New England BioLabs) and then purified with NucleoFast 96 PCR Plates (743100.10, Macherey Nagel) and the DNA was quantified on a Nanodrop (Thermo Fisher) prior to being used for yeast transformation. We were not successful in integrating YLL006W with either an N- or C-terminal mNG, nor for YMR151W with a C-terminal mNG, leaving 6 BEPs with mNG tagged on the N terminus and 5 with a C-terminal mNG tag.

Gene deletions were performed with the same transformation protocol and using DNA cassettes (*URA3* or *KANMX*) targeted to the region of interest using primers containing sequences homologous to the genomic locus.

Yeast Protein Extraction and Immunoblot Analyses

Whole cell extracts of yeast proteins were generated using trichloroacetic acid (TCA) method as described in[70] and modified from[71]. In brief, cells were grown in SC medium to mid-exponential log phase at 30°C (A_{600} = 0.6-1.0) and an equal density of cells was harvested by centrifugation. Cell pellets were frozen in liquid nitrogen and stored at -80°C until processing. Cells were lysed using sodium hydroxide, precipitated with 50% TCA, solubilized in SDS/Urea buffer [8 M Urea, 200 mM Tris-HCl (pH 6.8), 0.1 mM EDTA (pH 8.0), 100 mM DTT, 100 mM Tris (not pH adjusted)] and heated to 37°C for 15 min prior to loading on an SDS-PAGE gel. Immunoblotted proteins were detected using mouse anti-green fluorescent protein (GFP) (Santa Cruz Biotechnology, Santa Cruz, CA), mNeonGreen (Cell Signaling Technology, Danvers, MA) or HA antibodies (Roche, Indianapolis, IN). Anti-mouse secondary antibodies conjugated to IRDye-800 or IRDye-680 were used to detect primary GFP or mNG antibodies on an Odyssey CLx infrared imaging system (LI-COR Biosciences, Lincoln, NE). The HRP-conjugated anti-HA antibody was detected on a ChemiDoc (Bio-Rad, Hercules, CA). As a loading and transfer control, membranes were stained with Revert (LI-COR Biosciences) and detected using the Odyssey CLx.

Extracts containing HA-tagged BEPs were loaded on 16.5% Tris-tricine gels and run in the cold using 1X tricine running buffer (Bio-Rad). Proteins were blotted to Immobilon-PSQ (0.2 micron) PVDF membrane using the Criterion system (Bio-Rad). Membranes were stained with REVERT total protein stain (LI-COR Biosciences), followed by blocking with TBST with 3% BSA and overnight incubation at 4°C on a platform rocker with an HRP-conjugated anti-HA antibody in TBST containing 1 % BSA (1:5000 dilution, Roche). Membranes were then washed 3 times with TBST and detected using the SuperSignal West Pico PLUS Chemiluminescent Substrate (Thermo Fisher) on the Bio-Rad Chemidoc XRS+ Imaging System (Bio-Rad).

Protein Stability Assays

The stability of mNG- or HA-tagged YBR196c-a (as an N-terminal or C-terminal fusion) or mNG alone expressed as a chromosomal integration from the *Z_3EVpr* was assessed by cycloheximide (CHX) chase assay as described in[41]. Cells were grown to mid-exponential

phase and induced to express the tagged YBR196c-a or mNG using β-estradiol. Cells were next treated with 150 μg/ml CHX (Gold Bio, St. Louis, MO) and equal densities of cells were harvested at the indicated times. Proteins were extracted from cell pellets and subjected to SDS-PAGE and immunoblotting, as described above.

For assays that employed the proteasome inhibitor MG-132 (Thermo Fisher), cells were incubated with 10 μM of MG-132 (stock 10 mM in DMSO) or an equivalent volume of DMSO (vehicle control) for 1 h prior to CHX addition. When CHX was added to block new protein synthesis the t=0 timepoint was harvested to initiate the time course.

Fluorescence microscopy

For imaging experiments, cells were grown and induced to express mNG-tagged BEPs or free mNG as indicated above. Fluorescent proteins were localized using: 1) epifluorescence microscopy, 2) confocal microscopy in low-throughput, or 3) confocal microscopy in high-throughput. For epifluorescence microscopy and low-throughput confocal microscopy, cells were stained with 70 μM Cell Tracker Blue CMAC dye (Life Technologies, Carlsbad, CA) and 10 μM trypan blue (Gibco, Dublin, Ireland) and plated onto 35 mm glass bottom microwell dishes that were concanavalin A (MP Biomedicals, Solon, OH) or poly-D-lysine coated (MatTek Corporation, Ashland, MA). For epifluorescence microscopy, cells were imaged using a Nikon Ti2 inverted microscope (Nikon, Chiyoda, Tokyo, Japan) outfitted with an Orca Flask 4.0 cMOS camera (Hammamatsu, Bridgewater, NJ) and a 100x objective (NA 1.49). For low-throughput confocal microscopy, images were captured using a Nikon Ti2 A1R inverted microscope (Nikon) equipped with a 100x objective (NA 1.49) and detected with GaAsP or multi-alkali photomultiplier tube detectors. For high-throughput confocal microscopy, imaging was done as described in Bowman *et al.*[72].

In all cases, image acquisition was controlled using NIS-Elements software (Nikon) and all images within an experiment were captured using identical settings. Images were cropped and adjusted equivalently using NIS-Elements (Nikon), and where added adjusted images are needed to capture the range of fluorescence intensities in an experiment, additional images are presented that are also evenly adjusted and the change in this adjustment is indicated in the figure panel.

Fluorescence microscopy image analysis and statistical tests

To define the localizations represented for both EGFP- and mNG-tagged BEPs, all images were analyzed using Image J software (National Institutes of Health, Bethesda, MD) with the Cell Counter plugin to categorize the localization for every cell in a field. These manually defined localization patterns are summarized in the pie charts presented in **Figures 1B, 2E, 5A-B, S9A-B,D-E,G,I-J, and L-M**.

To quantify the whole cell fluorescence intensity for mNG-tagged BEPs, we used Nikon NIS-Elements *.ai* (Artificial Intelligence) and Nikon General Analysis 3 (GA3) software packages. We trained imaging on a ground-truth set of samples, where cells were segmented using images acquired with the trypan blue-stained cell surfaces or the DIC images. Next, the NIS.*ai* software performed iterative training until a training loss threshold of <0.2 was obtained, which is indicative of a high degree of agreement between the initial ground truth and the output generated by NIS.*ai*. Fields of images were then processed, using the automated segmentation, and the mean fluorescence in the green channel (480 nm) was measured for each cell. Any partial cells at the edges of images were removed. Fluorescence intensities are plotted

as scatter plots using Prism (GraphPad Software, San Diego, CA). We performed one-way ANOVA statistical tests with Dunnette's post hoc correction for multiple comparisons, and or Mann-Whitney tests when only two samples were compared. In all cases, significant p-values from these tests are represented as: * <0.05; ** <0.0005, *** <0.0005; ns >0.05.

Biochemical properties, transmembrane domain prediction, and statistics

Amino acid sequences of the BEPs were analyzed for their biochemical properties using Python. A custom script that used the packages Biopython and 'peptides' was used to calculate hydrophobicity, percent charged residues, and length in different sequence regions. Transmembrane domain prediction was conducted by accessing the Phobius (https://phobius.sbc.su.se) and TMHMM 2.0 (https://services.healthtech.dtu.dk/services/TMHMM-2.0/) online servers[26]. A FASTA file containing all the amino acid sequences was uploaded and both analyses were run using default parameters. All statistical tests were performed in Python using scipy.stats.

Description and re-analysis of published datasets

Localization annotations for ancient proteins (shown in **Fig 1C**) were obtained from a published study using the yeast GFP collection[23]. The localizations assignments from replicate "WT1" were parsed and proteins were grouped into "ER" or "other" categories. For proteins that exhibited more than one localization, if one was ER, than that protein was counted as "ER", otherwise it was placed in the "other".

Fitness measurements for BEPs (shown in **Fig 1D**) were obtained and reanalyzed from[10]. For each BEP we counted the number of conditions that it was beneficial relative to a control, with the minimum and maximum number being 1 and 5, respectively.

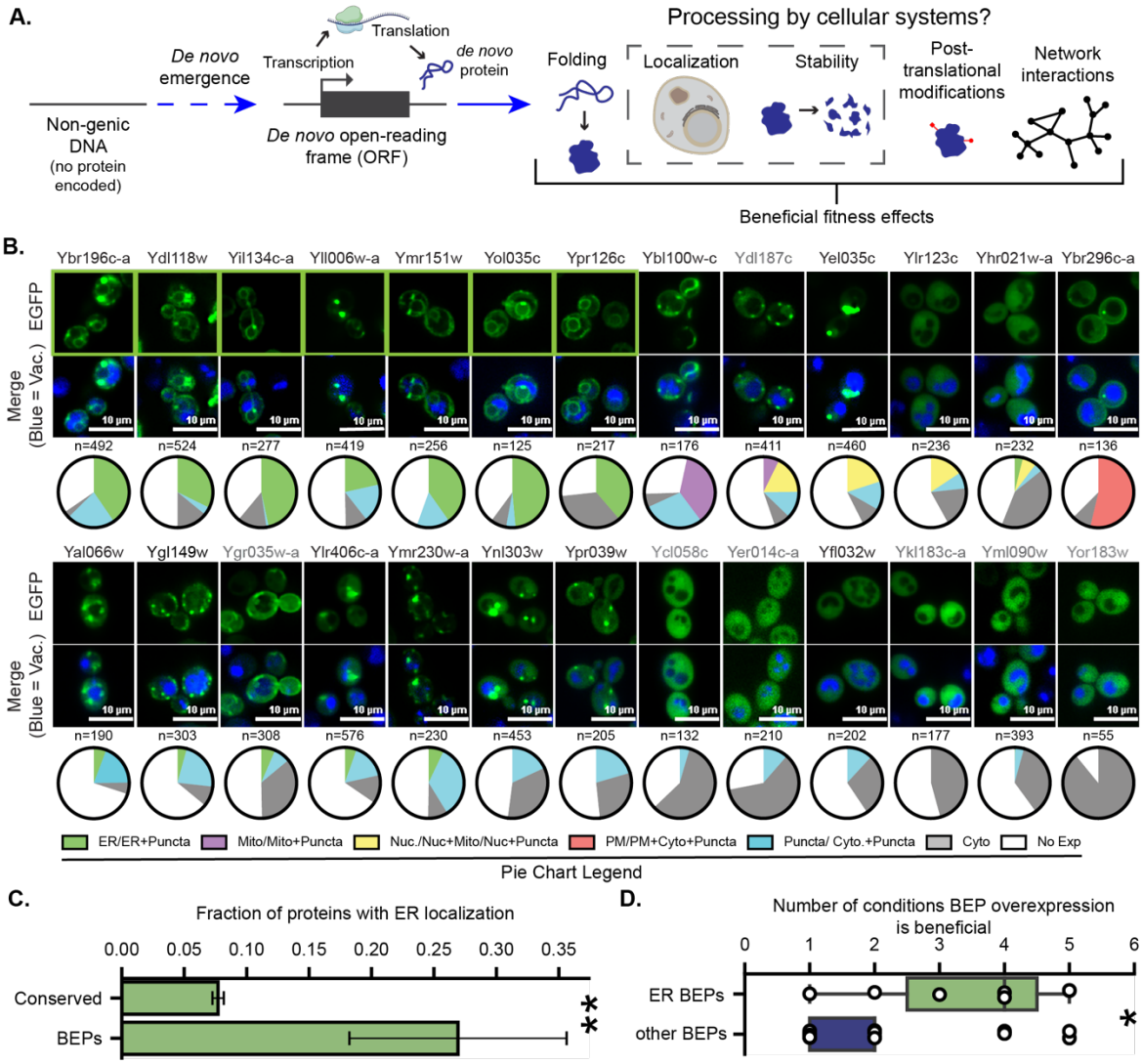


Figure 1. BEPs preferentially localize at the ER.

A. Schematic representation of BEP emergence and the various processing by cellular systems that ultimately gives rise to their beneficial fitness effects. BEPs are encoded by *de novo* ORFs which are created by mutations within non-genic DNA. The specific processes and mechanisms that BEPs use to integrate into the existing cellular systems remain unknown. The processes of localization and stability (boxed by dashed outline) are the focus of the current study. **B.** Seven out of 26 BEPs localize to the ER. Fluorescence microscopy of cells induced to express BEP-EGFP fusions from a 2µ plasmid are shown. Images of ER-localized BEPs are outlined in green. Cells were stained with CMAC blue to mark the vacuoles. The distributions of localizations observed across individual cells (n) are provided as pie-chart graphs below each image. Mito = mitochondria, Nuc = nucleus, PM = plasma membrane, Cyto = cytoplasm, No Exp = no expression. The BEPs highlighted in grey correspond to those not presenting a band at the expected size on immunoblotting analysis as shown on Figure S2. **C.** BEPs are enriched in the ER relative to conserved proteins. The fraction of all BEP-EGFP fusions in this study (n=26) that localize to the ER upon induced expression *vs* the fraction of the whole *S. cerevisiae* proteome fused C-terminally to GFP and natively expressed (n=3488, GFP collection[23]) that stably localize to the ER are compared. Error bars indicate the standard error of the proportion. **: Fisher exact test, odds ratio= 4.4, p= 3.1 x 10^{-3}. **D.** ER-localized BEPs are beneficial across more conditions than other BEPs. The distribution of the number of experimental conditions

in which ER-localized and other BEPs were found to be beneficial when overexpressed by Vakirlis and colleagues[10] is presented. *: Mann Whitney U test, p= 0.025, n= 26.

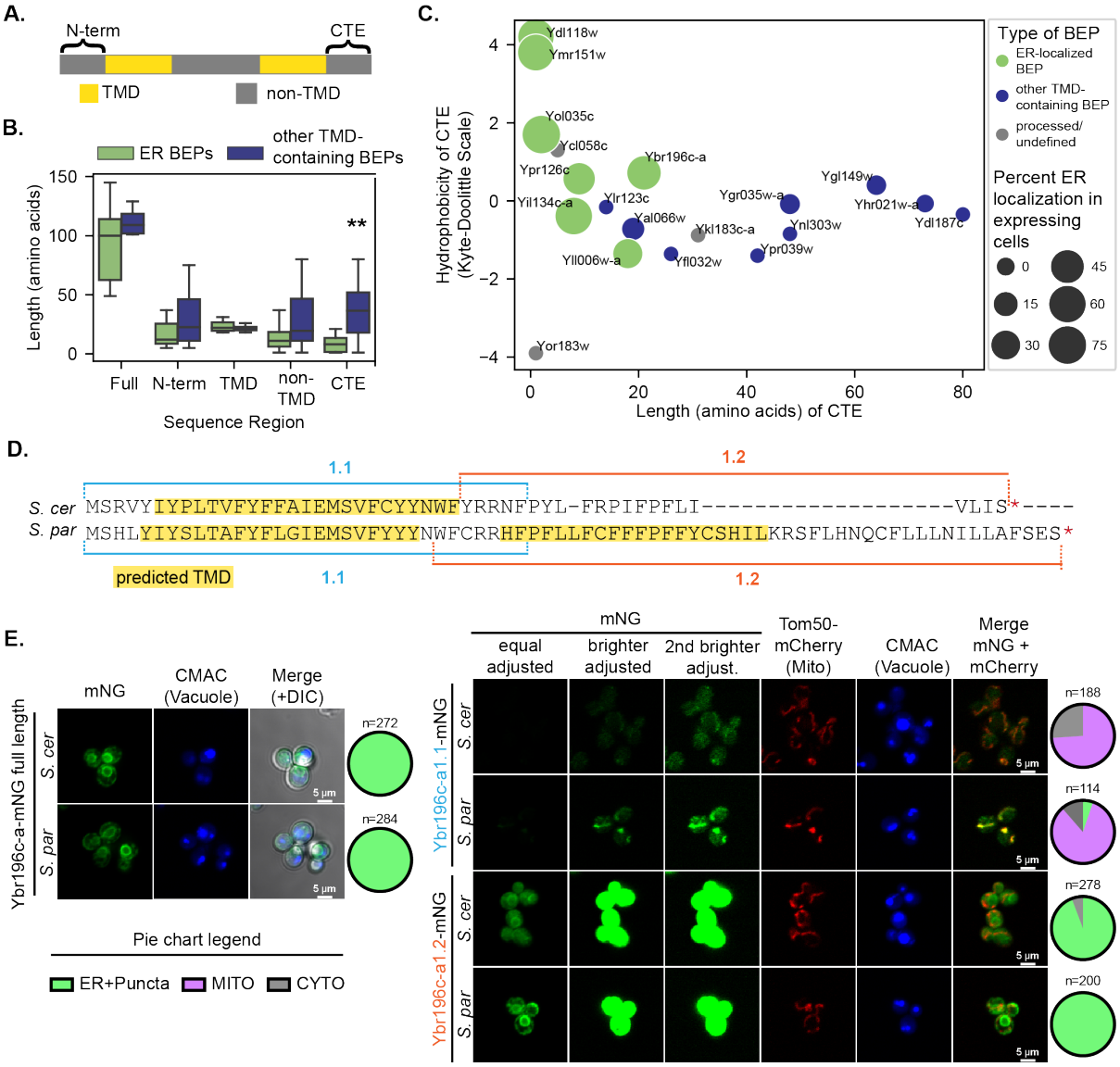


Figure 2. ER-localized BEPs have C-terminal TMDs that drive localization.
A. Cartoon diagram of a TMD-containing BEP sequence. Each sequence was separated into regions based on TMD predictions by either Phobius[25] or TMHMM[26] in Fig S5. These regions included 1) the N term: all residues from the start of the sequence to the start of a predicted TMD; 2) TMD: regions predicted to be TMDs; 3) non-TMD: regions including all residues not predicted to be in a TMD; and 4) the CTE: all residues following the most distal TMD in the sequence. **B.** ER-localized BEPs have significantly shorter CTEs than other TMD-containing BEPs. Length in amino acid residues for different sequence regions in ER-localized and other TMD-containing BEPs is plotted (Mann-Whitney U test. *: $p<0.05$; Full=Full sequence. For Full, N-term, and CTE, n=19. For TMD, n=30. For non-TMD n=49. **C.** Hydrophobicity versus length in amino acid residues for the CTEs of all TMD containing BEPs. Processed/undefined: BEPs that showed some evidence of breakdown in western blots (see Fig S2). The size of the dots represent the percent of expressing cells that had ER localization in Figure 1B. Corresponding systematic gene names from *Saccharomyces* Genome Database shown next to each dot. **D.** Sequence alignment shown for Ybr196c-a between *S. cerevisiae* (*S. cer*) and its sister species *S. paradoxus* (*S. par*) as obtained from Vakirlis et al 2020[10]. TMD predictions

made with Phobius[25] are shown in yellow. Brackets indicate the sequences used for each subclone, 1.1 and 1.2 in either species. Full length and partial sequences for both species were fused C-terminally to mNG and genomically integrated into *S. cerevisiae* under the control of an inducible promoter. **E.** The C-terminal region of Ybr196c-a and its orthologous sequence in *S. paradoxus* drive localization to the ER. Fluorescent micrographs are shown of the full length (left) and two partial sequences (right) of the Ybr196c-a proteins from both species. An mCherry-tagged Tom50 was genomically integrated into these strains to allow visualization of the mitochondria and cells were stained with CMAC-blue to mark the vacuoles. The distributions of localizations in single cells (n) observed for each construct are represented in the pie-chart graphs to the right of each image. For simplicity, in partial sequences (right) the merge shows the co-localization between the mNG and the mCherry signal only.

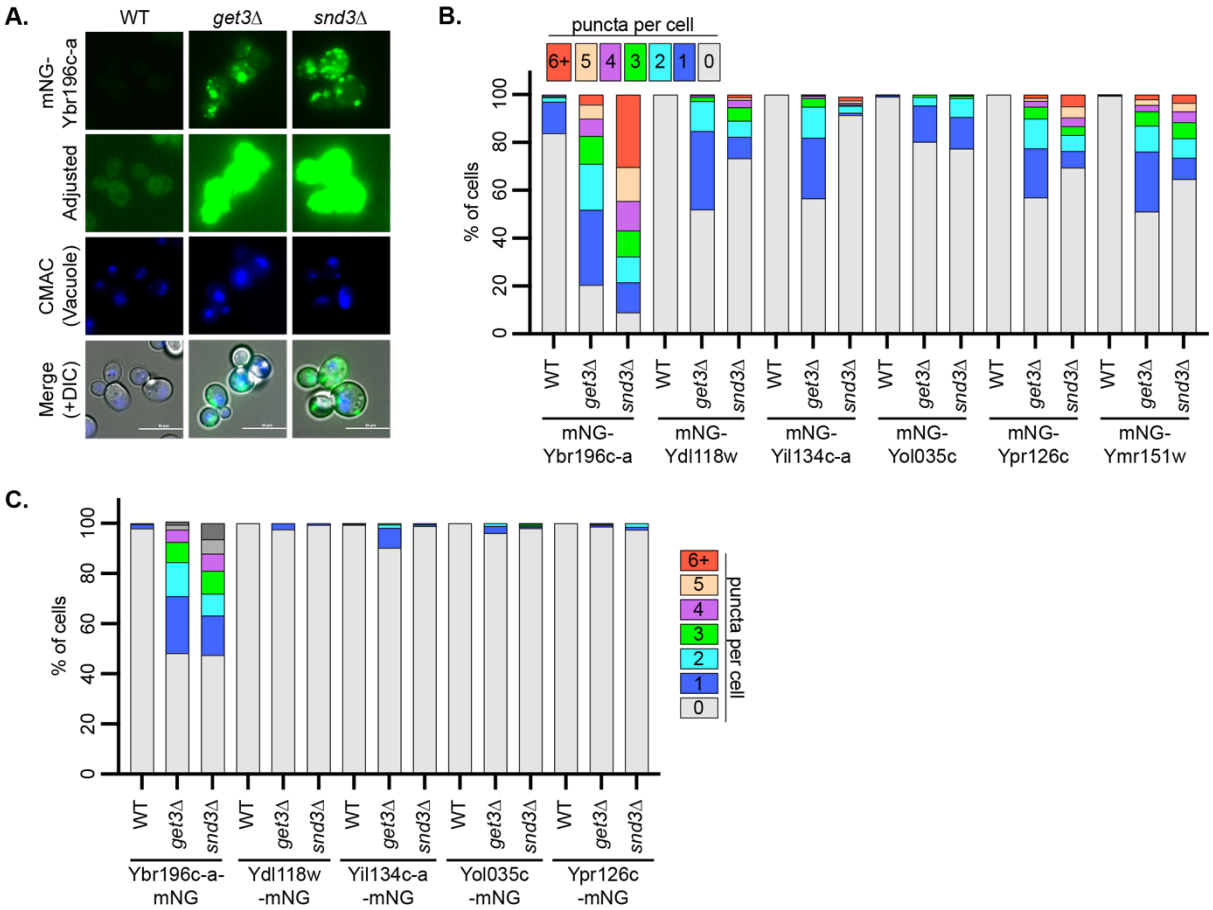


Figure 3. The GET and SND posttranslational targeting pathways are needed for BEPs to access the ER.

A. Fluorescence micrographs of mNG-Ybr196c-a in cells with the indicated deletions of GET or SND components (Get3 or Snd3, respectively) are shown. **B and C**. Increasing numbers of puncta per cell, and concomitant loss of clear ER localization, are associated with disruption of the ER-targeting pathways needed for BEP targeting. The percentage of cells (n=305-881) with the number of puncta per cell (based on images shown in panel A and Figure S6C-D), are indicated as a stacked distribution.

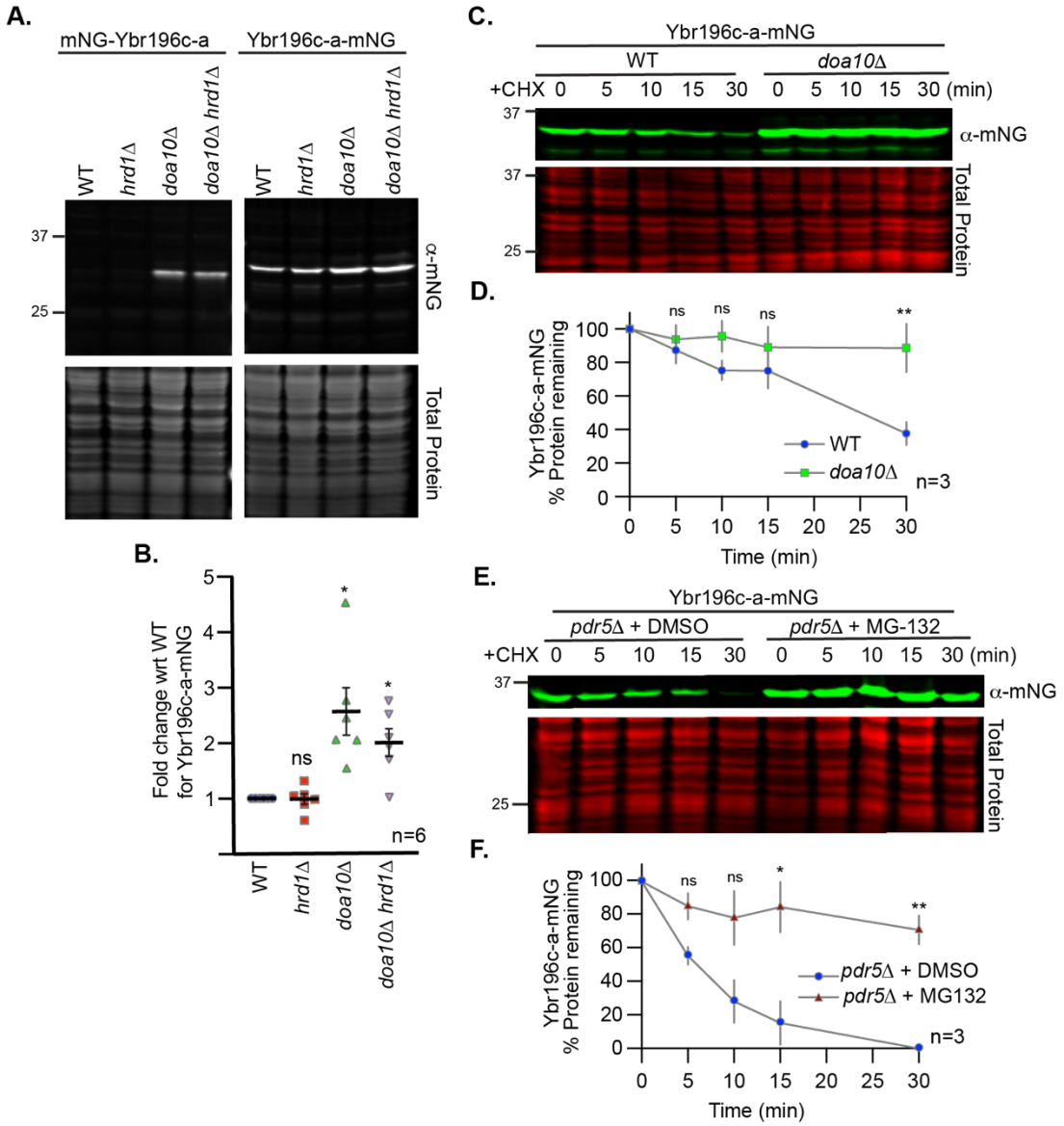


Figure 4. Ybr196c-a's stability is dependent upon Doa10 and the proteasome.
A. A representative immunoblot of protein extracts made for N- or C-terminally mNG-tagged Ybr196c-a expressed in the indicated cells. **B.** Quantification of the relative abundance of Ybr196c-a-mNG from six biological replicates of immunoblots, as in **A,** for the cells indicated. A Student's t-test comparing the WT to the Ub ligase deletions assessed statistical differences. (ns: not significant; *: p-value < 0.05). **C** and **E.** Representative immunoblot of protein extracts expressing Ybr196c-a-mNG from cells post cycloheximide (CHX) (in panel C), or CHX plus MG-132 or DMSO (in panel E) addition for the times indicated (in minutes) When the proteasomal inhibitor MG-132 was added, *pdr5*Δ cells were used to improve retention of MG-132. **D** and **F.** Quantification of the percent of Ybr196c-a-mNG remaining post-CHX addition in the indicated backgrounds. The dots represent the mean of the biological replicates (n), and the error bars show the standard error of the mean. A Student's t-test comparing the **D**, WT to the *doa10*Δ cells and **F**, DMSO- to the MG-132-treated cells at each time point was used to assess statistical differences (ns: not significant; *: p-value < 0.05; **: p-value <0.005; ***: p-value <0.0005). Similar analysis to those presented in **C** and **D** for an mNG control and the N-terminally mNG-tagged Ybr196c-a are presented in Figures S7E-F, S7I-J, respectively. Similar analysis to those presented in **E** for an mNG control and the N-terminally mNG-tagged

Ybr196c-a are presented in Figures S7K and S7L, respectively. For all immunoblots, REVERT total protein stain serves as the loading control.

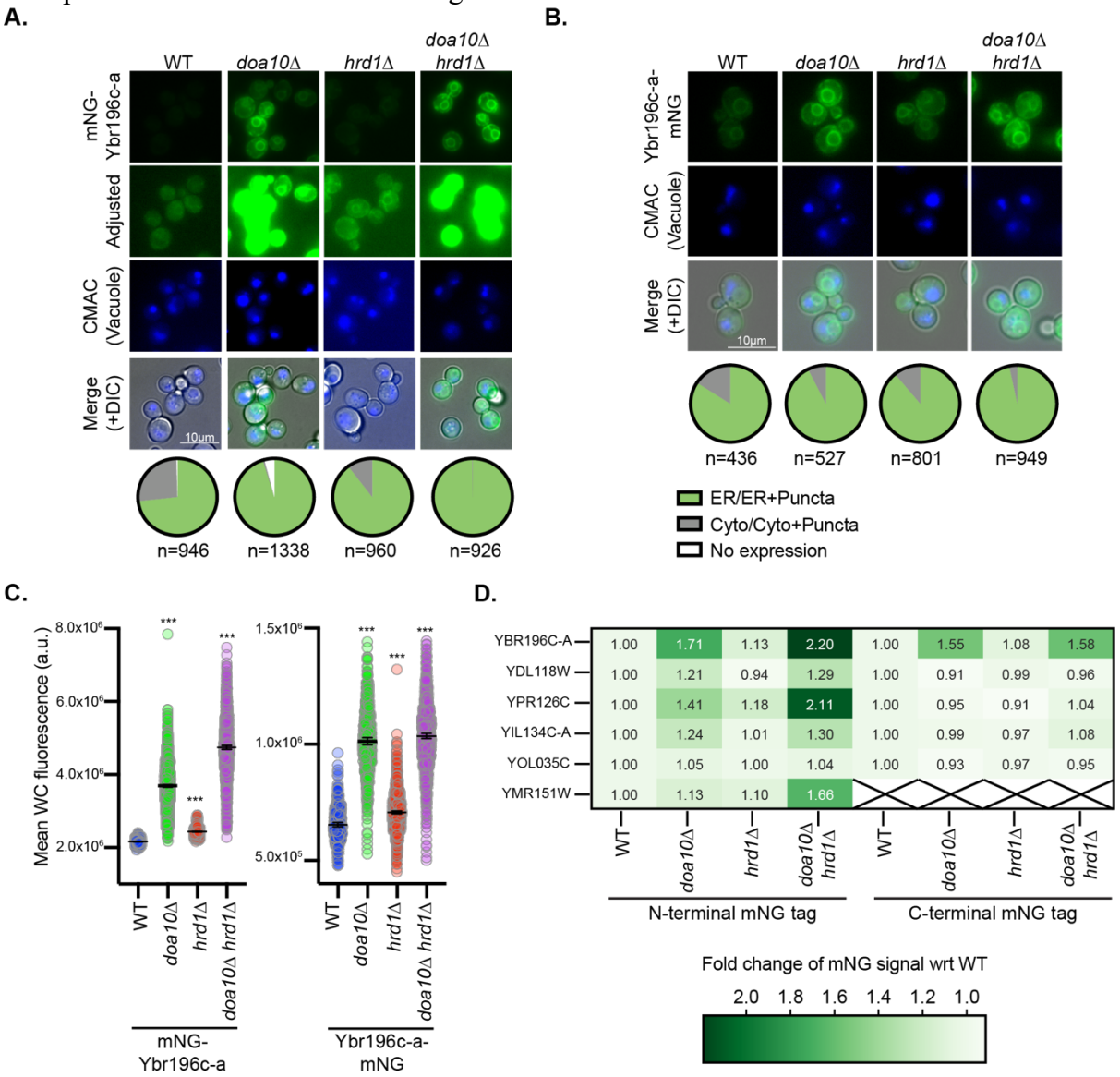


Figure 5. ER-localized BEPs' abundance depends on Hrd1 and Doa10.
A-B. Representative fluorescence microscopy of **A**, N-terminally- or **B**, C-terminally-mNG tagged Ybr196c-a. The distribution of fluorescence observed across individual cells (n) is indicated in the pie chart below each image panel. Cells are stained with CMAC blue to mark vacuoles and imaged via DIC, which is shown in the merge. Analogous imaging data for all ER-localized BEPs is provided in Figures S9. **C.** Quantification of the whole cell fluorescence intensities for the cells (n=180-270) shown in panels **A** and **B** is shown. A one-way ANOVA with Dunnett's multiple comparisons test was performed relative to the fluorescence in WT cells to identify statistically significant changes (p-value < 0.0005 = ***). **D.** A heatmap indicating the fold change in the mNG fluorescent signal for ER-localized BEPs in cells lacking the indicated Ub ligase relative to the WT controls (based on data in C, Figs S9C, S9F, S9H, S9K, S9N).

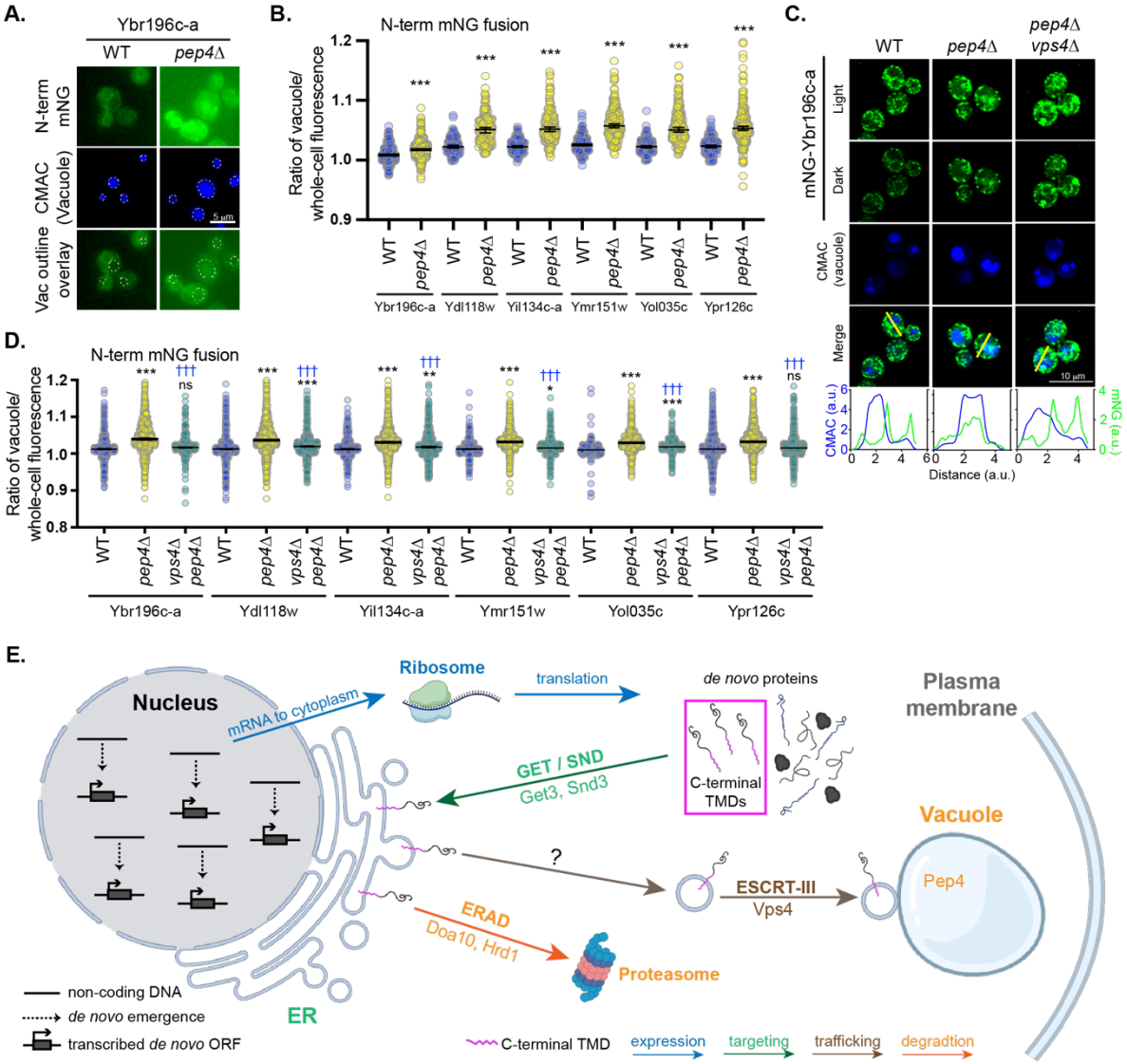


Figure 6. ER-localized BEPs traffic beyond the ER.
A and **C.** Fluorescence micrographs of cells expressing N-terminal mNG-tagged Ybr196c-a in WT cells or those lacking *PEP4* and/or *VPS4* are shown. Vacuoles are stained with CMAC-blue. **A**. Vacuoles are indicated on the green channel images using a white dashed line. **C.** Plots indicate the fluorescence intensity from the blue and green channels along a line scan through the vacuole (indicated in yellow on merge). Overlapping maximum peaks of blue and green lines indicate mNG presence in the vacuolar lumen. Similar results and analysis are shown for all other N-terminal mNG-tagged BEPs in Figures S10H-L. **B** and **D.** Scatterplots of the ratio of the vacuolar fluorescence over the whole cell fluorescence for the cells imaged in panel A and Figure S10A are shown in **B** (num of cells=193-924) and those imaged in C and Figures S10H-L are shown in panel **D** (num of cells=206-342). A Mann-Whitney two-sided test assessed statistical difference between the WT and *pep4*Δ (denoted by asterisks) in panel **B** and a one-way ANOVA test with Dunnett's multiple comparison correction between WT and *pep4*Δ (denoted by asterisks) and *vps4*Δ *pep4*Δ (denoted by blue daggers) in panel **D** cell populations (ns: not significant, $p < 0.05$ = *, $p < 0.0005$ = ***, $p < 0.0005$ = †††). The median fluorescence intensity and 95% confidence interval are represented by horizontal lines. Similar analyses to those shown in panel B for C-terminal mNG-tagged BEPs is provided in Figure S10C. **E.** Model of ER-localized BEP cellular integration. *De novo* originated ORFs arising

from non-coding DNA are transcribed and translated into BEPs. ER-localized BEPs all share a common C-terminal TMD (outlined in pink) and depend on the post-translational targeting pathways GET and SND to localize to the ER. Once at the ER, BEPs are either degraded in the proteasome via ERAD, or traffic out of the ER using an unknown mechanism (indicated by a question mark) to eventually arrive at the vacuole via an ESCRT-III-mediated pathway where they are degraded by vacuolar proteases. Specific components of each major pathway found to be involved in BEP targeting, trafficking, and degradation listed below corresponding arrows.

Supplemental Figures

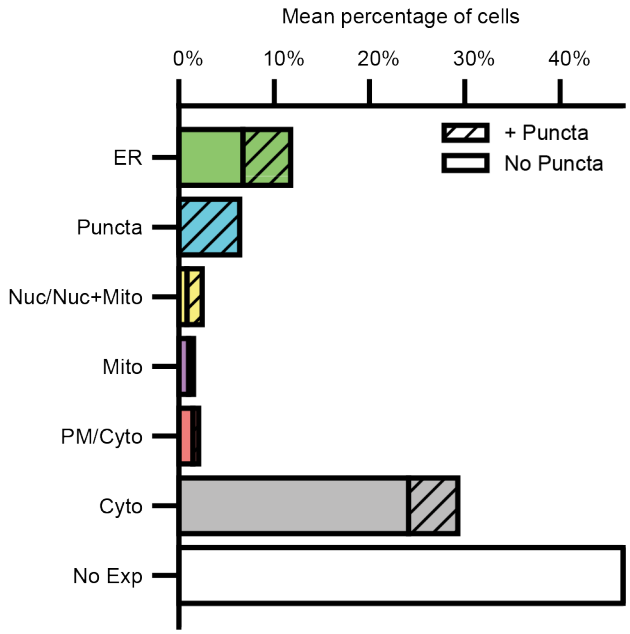


Figure S1. The ER is the predominant membrane localization for BEPs. (accompanies **Figure 1**)
The percentages for each localization from Figure 1B (pie charts) are aggregated here to display the mean percentage of individual localizations in the BEP-EGFPs fusions. Membranous compartments on the plot include ER, Nuc (nucleus), Mito (mitochondria), PM (plasma membrane).

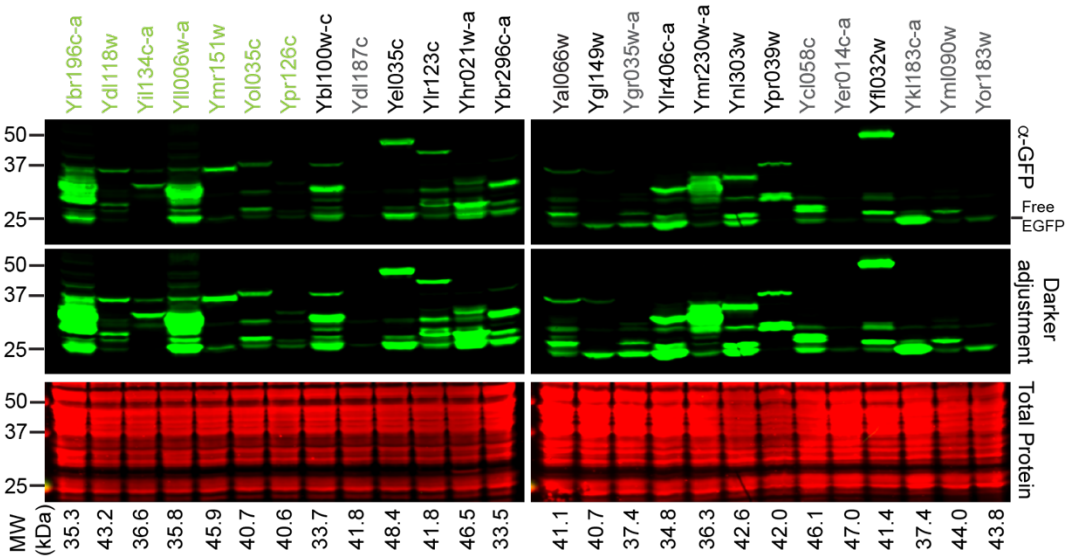


Figure S2. Immunoblot analyses of BEPs. (accompanies **Figure 1**)
Representative immunoblot of protein extracts made for EGFP C-terminally tagged BEPs expressed from plasmids as shown in Figure 1B. Labeled in green text are the ER-localized BEPs, and this designation is based on the microscopy shown in Figure 1B. Expected sizes of the BEP-EGFP fusions indicated below immunoblot. The BEPs in gray are the cases identified as potentially being break down/processed products because the only bands observed are close to the size of free EGFP (~26.9kDa). REVERT total protein stain serves as the loading control.

Ybr196c-a
MSRVYIYPLTVFYFFAIEMSVFCYYNWFYRRNFPYLFRPIFPFLIVLIS

Ydl118w
MFESFVNEGTAFLLFAKDERIRFKHDKYKALPIDVRNAEGNVPLHSCRGLSISFKFFHNVAFLSCWILVFKRSKGCNATADVSPPKKPPIKCDEFLGLVACSVMYAASQIVYLLVVPAVLFAFLLV

Yil134c-a
MVVGSEFNNTADVDVVAILLTLLNADYFVIKQRLVSACFTKRIKWYFIYAIALLSLFSSRNTVCAPY

Yll006w-a
MTQLSQSNVERVALVYISVYFFSCISLIVYFFTFCLSVSIPKNRQHPIKIIYNNKCPS

Ymr151w
MGSVPTVIPGIVHCWMIHCLIGPFLFAIIYWSIFIEGMFFPQKYLDCQSVINLLCGSNQQKRIFKEKLRIYQVHIQEKIGCSKFNTLLLIYQLGSRHKINRQINDLLCNLCSKNYMIMLLYAEIIIYSSFSCVLSATIFLRGFLL

Yol035c
MRVYHHIYVYTYILSAVIYSQDLFPSWVVQLSTVKSDIIRPYLIHGNSFLFQILQVLITAPSTRCKFSFQNSIPFIFLALLLSQDFHVFLGIELHAFFVS

Ypr126c
MLARVAESVSCGLMGQVKTGLLLFDGSGFSDRLGVMRFYVWSSRIYVLVLVVQAQLILDAHNGVLFLLLFFLHNFFLLPQLFQFLLSGCLIFLNDVYFNLMV

Ydl187c
MQVASILQVRPPFFPQIHFLLFFFCFFVSKSRVASQRGTHFLVRICFEPPLQGCRMGKLCRPHLNHLLYLCYTAYFSNISWLSVSKSTSRNRVHKYIPYKSTELRSVAN

Ylr123c
MIMQKRLLNRNSSVVKSRNCLARHQHESSISKLHVSLRLHRCLRPIVRFDFLHNAGSKVCSQNSCPSPQLHVGTMALFHTVFILWPHFCGILWTVHEKLYNYLLSIEVY

Yhr021w-a
MNRSFHFLIKYIYIHVLLVFYFHIKQQAIMPFFIFFFSSFDGLSFDLRVVAFLAKHVFVGVCSPFFVVGFFGSSRVVVTEWLSKLVLPPPPVSITQVFSLSRKRGEFSSGYILIINPYKSFLRSLLDFSIFNNTAKNKSSTFTLNLEDVSK

Yal066w
MLSLVKRSILHSIPITRHILPIQLILVKMNHVQIRNIKLYHFISYGFMLTKLTVFLFNLFFYRLRILCRLTLLILSLPVQIYIKEIQTKMLEKHTASDTSCI

Ygl149w
MSINALLYVLSLALLIWTGSVVTLLLLLFFCLFLLFFSLHFFCFTREHVHYTLPPKCHSLKFQFDSIPSSSLSLSPFPFLFFPRLRAVAFASPTLSFFFPI

Ygr035w-a
MLLYLYILYIALKTVSHLFFCNPCFRNLSVGDMLNPRLSLSFFTNHLLCPEAPLIIRGKGSYSAVGNFFSRKK

Ylr406c-a
MIQKPILLSSFLFLYIRALLHSIHPYIRTSVHLYTKKITSYNFLGVPFK

Ynl303w
MTTLQIINQFVAFSLRQIIQQDSQAAIVQYANVIICTNIDALYNPTIYIIVKLFTTTYSSMVFLMFCRGKLKSCHFYSDRNPWFDTFYLKKTKSIYFGFGYIDDDECTYKMAKFK

Ypr039w
MASFDYLFHPFVPCTICPDFPLYKSPAFPSSCLHHPRLLFNDKAFCPLFLVPFPASFTRWLTFLFHLVIYNNKMHHHTYAPHIHDLRAALDTTAPQKKCPKETLHRSDHQG

Ycl058c
MQYHSALYVYIYVTFTTIPYKEKPDIISICFSMLSFVFDFSVRICSRTLESFSWSLISSSAFKVVSAFSLAGSCVLASRSSVGIIVSLLLFNFSTCNFVLFLSAVLIDLFFCTFLPTPTFLPTPFFFMLHLPIFSLLNALELLYLIIAGLHI

Yfl032w
MRVVRQSGSALSRAKKSRKYTSYRVMNVVLNTLFSFVLAPYIHYILEEISPQWTTREMNTEICFLAKFSFLLVFLFYLNFQGFNSVSNITTSSSPTYDNNRHYGND

Ykl183c-a
MYSKILLYRSNVLFMNFFSVFVCTIGTLFLVFADVYVLASAFFQSKKEKETKFKHLHYQKRSCFFLANIH

Yml090w
MLVFSFLFVVVSINLNALIFLCKKSWASYLFLYLYNFFFCEDEYKLTKNCVRVEAIAPFMMCLGSLGAILGKQRTANFLLLSYNVINNPVVLVYYVENFSRINFIKHTTKEKSVIWTNERQLNPWICN

Yor183w
MRLLHHGEYGTKLIGGKCSIDGKLGHPCPLSRRRKKHLREKEMGPQYVRMYGPKRKAIIRTGNPDDGINLPDTGRGTLTAATIWSRAYHSNYSYLVRPKVVTLSKHRELMTTFLLYVLYVCIYISAFIK

Transmembrane Domain Prediction Method: Phobius TMHMM

Figure S3. BEPs amino acid sequences and predicted TMDs. (accompanies **Figure 2**) Amino acid sequences of 21 of the 26 BEPs and those that have at least one predicted TMD, as determined by Phobius[25] or TMHMM[26] are highlighted in yellow or underlined, respectively. The first seven sequences boxed in green correspond to the ER-localized BEPs, which were defined by the microscopy in Figure 1B.

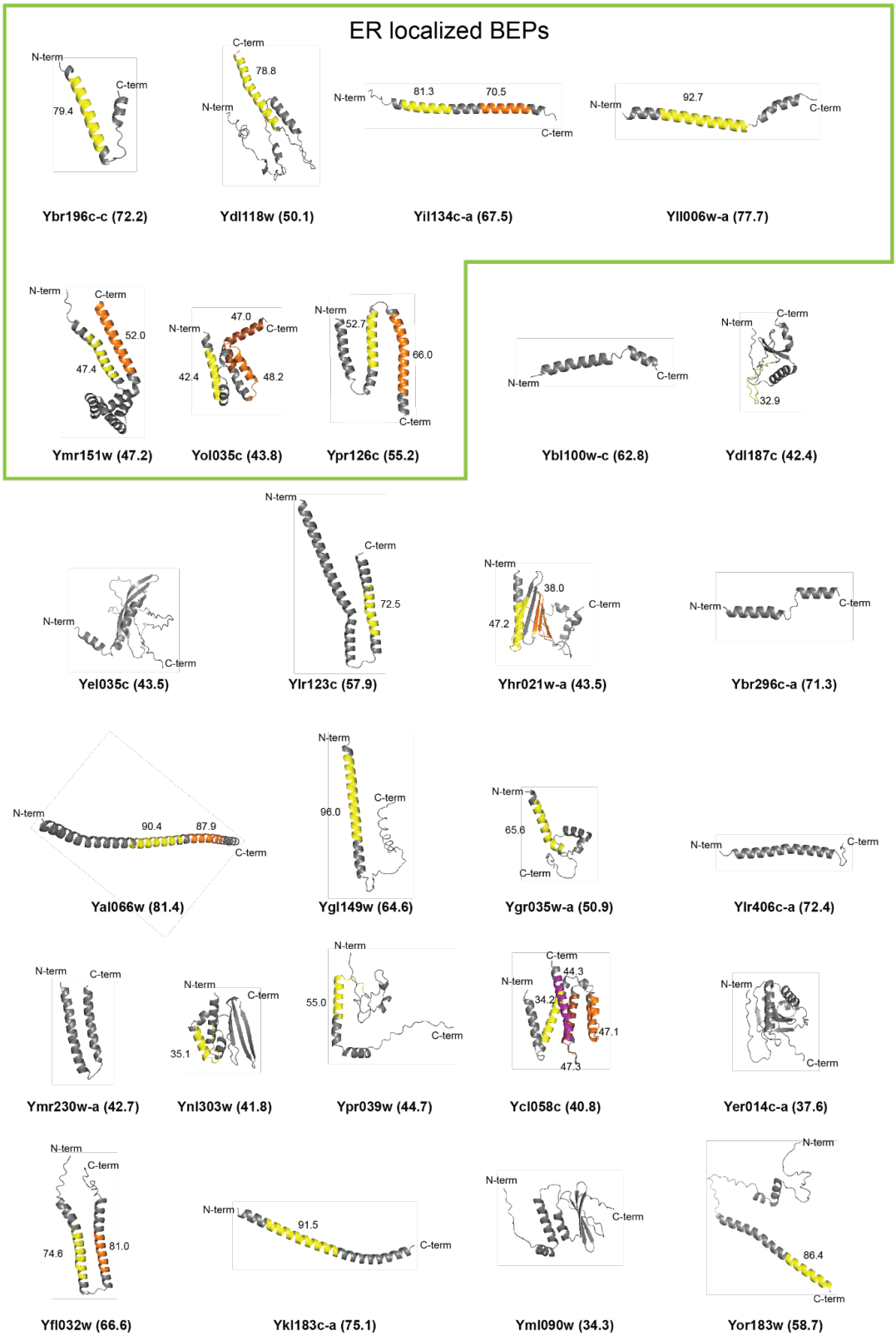


Figure S4. BEPs are rich in alpha helices that correspond with predicted TMDs. (accompanies **Figure 2**)

Alpha-fold 2.0[27,73] predicted tertiary structures for the ER-localized (outlined at top) and other BEPs. The N- and C-termini for each BEP are indicated. The confidence of the Alpha-fold 2.0 prediction for each BEP sequence is presented in parenthesis. Note that some Alpha-Fold 2.0 predictions have low confidence due to BEPs' lack of homology with known proteins[74,75].

TMDs predicted by Phobius[25] are indicated by yellow, orange, and magenta regions and the prediction confidence is presented next to each putative TMD.

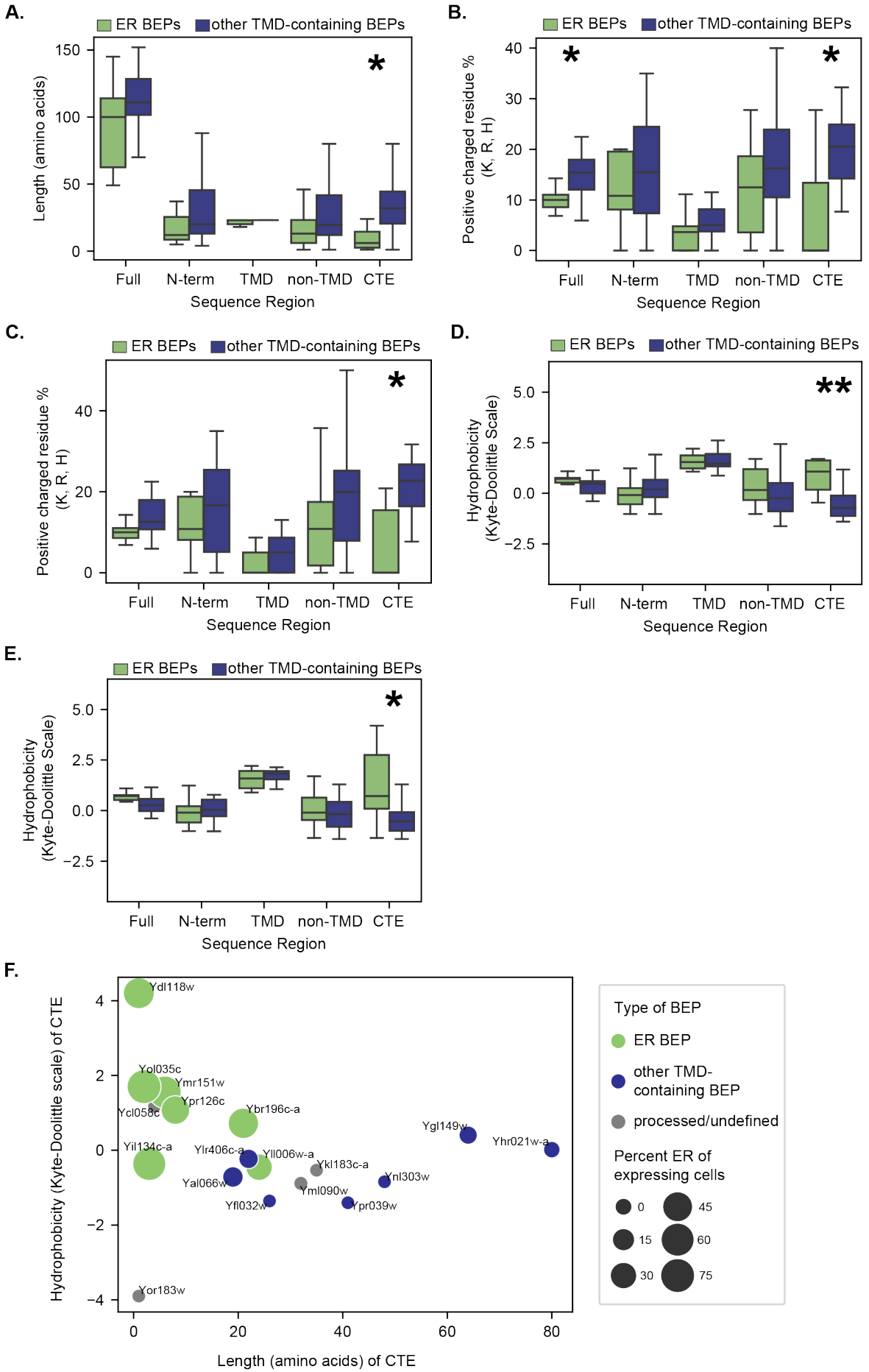

A.
ER BEPs
other TMD-containing BEPs
Length (amino acids)
150
100
50
0
*
Full
N-term
TMD
non-TMD
CTE
Sequence Region
B.
ER BEPs
other TMD-containing BEPs
Positive charged residue %
(K, R, H)
40
30
20
10
0
*
*
Full
N-term
TMD
non-TMD
CTE
Sequence Region
C.
ER BEPs
other TMD-containing BEPs
Positive charged residue %
(K, R, H)
40
20
0
*
Full
N-term
TMD
non-TMD
CTE
Sequence Region
D.
ER BEPs
other TMD-containing BEPs
Hydrophobicity
(Kyte-Doolittle Scale)
5.0
2.5
0.0
−2.5
**
Full
N-term
TMD
non-TMD
CTE
Sequence Region
E.
ER BEPs
other TMD-containing BEPs
Hydrophobicity
(Kyte-Doolittle Scale)
5.0
2.5
0.0
−2.5
*
Full
N-term
TMD
non-TMD
CTE
Sequence Region
F.
Hydrophobicity (Kyte-Doolittle scale) of CTE
4
2
0
−2
−4
Ydl118w
Yol035c
Ymr151w
Ypr126c
Ycl058c
Ybr196c-a
Ygl149w
Yhr021w-a
Yil134c-a
Ylr406c-a
Yll006w-a
Ykl183c-a
Yal066w
Ynl303w
Yml090w
Yfl032w
Ypr039w
Yor183w
0
20
40
60
80
Length (amino acids) of CTE
Type of BEP
ER BEP
other TMD-containing BEP
processed/undefined
Percent ER of expressing cells
0
15
30
45
60
75

Figure S5. Sequence analysis of TMD-containing BEP protein sequences using Phobius or TMHMM reveals ER-localized BEPs have short, hydrophobic, and less positively charged CTEs. (accompanies **Figure 2**).

A. ER-localized BEPs have significantly shorter CTEs than other TMD-containing BEPs. Comparison of the distributions of length in amino acid residues for different sequence regions in ER-localized and other TMD-containing BEPs as predicted by TMHMM[26] are shown. Similar results using Phobius[25] predications are shown in Figure 2B. **B-C.** ER-localized BEPs have less positively charged CTEs than other TMD-containing BEPs. Percent of positively charged residues for sequence regions in both ER-localized and other TMD-containing BEPs as predicted by **B**, Phobius[25] or **C**, TMHMM[26]. **D-E.** ER-localized BEPs have significantly more hydrophobic CTEs than other TMD-containing BEPs. Comparison of the distributions of the mean hydrophobicity for different sequence regions in ER-localized and other TMD-containing BEPs as predicted by TMHMM[26] and Phobius[25] are shown in **D** and **E**, respectively. **A-E.** All P values were calculated from Mann-Whitney U test. **: $p < 0.005$, *: $p < 0.05$ Full=Full sequence, N-term=N-terminal region, CTE=C-terminal extension. Full, N-term, and CTE **A, C,** and **D,** n=18 and **B** and **E**, n=19. For TMD **A, C,** and **D**, n=28 and **B** and E, n=30. For non-TMD **A, C,** and **D**, n=46 and **B** and **E**, n=49. **F.** Hydrophobicity versus length in amino acid residues for the CTE of all TMD containing BEPs as predicted by TMHMM[26]. Processed/undefined: BEPs that showed some evidence of breakdown in western blots (see Figure S2). Size of the dots represent percent of expressing cells that had ER localization in Figure 1B. Corresponding systematic gene names from *Saccharomyces* Genome Database are shown next to each dot.

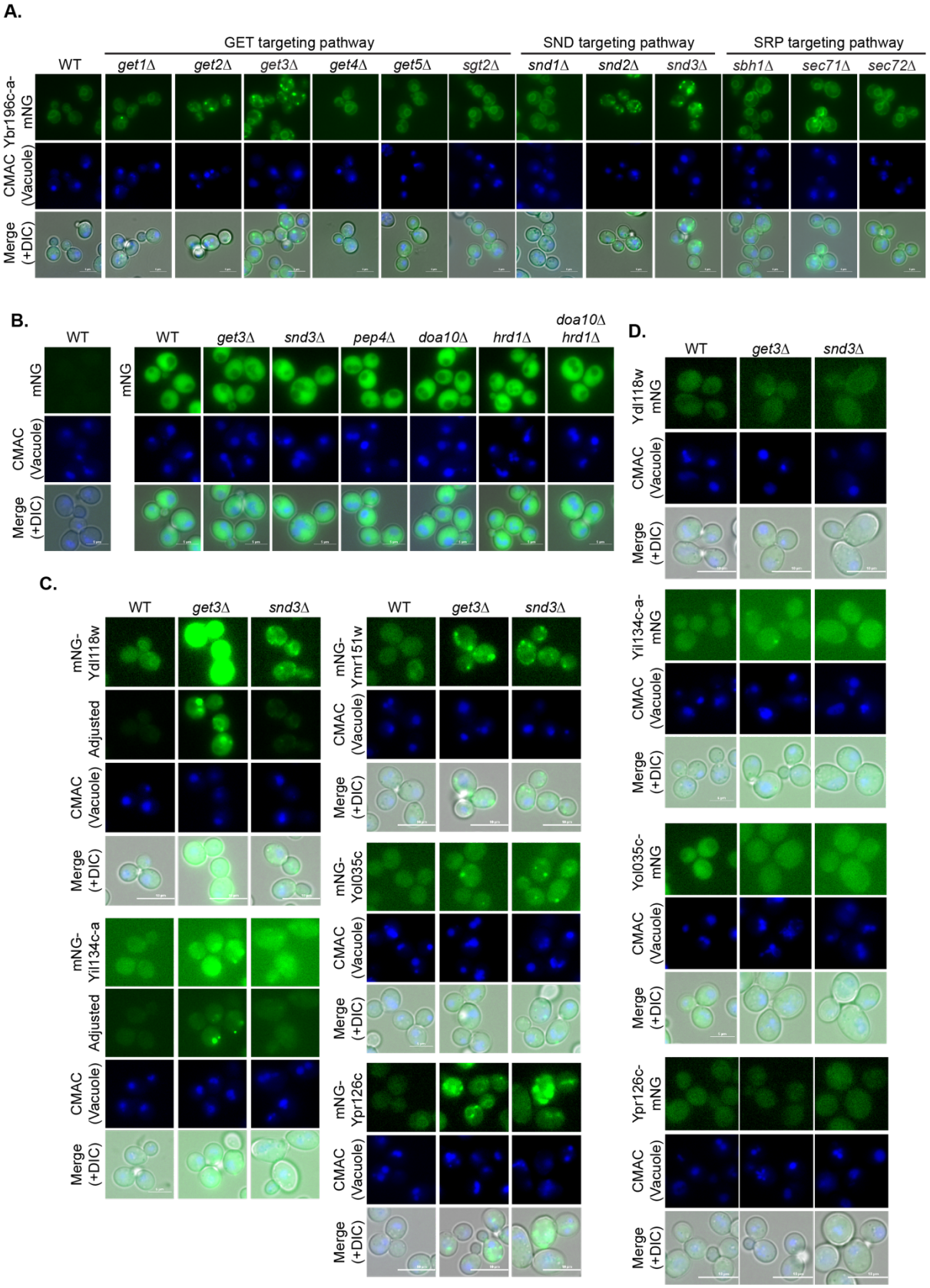


Figure S6. Free mNG and ER-localized BEPs in cells lacking functional GET or SND pathways. (accompanies **Figure 3**)

Fluorescence micrographs of **A.** Ybr196c-a-mNG or **B.** free mNG expressing cells that lack core components of the GET (Get1-3) or SND (Snd1-3) pathways, or GET pathway accessory proteins (*e.g.,* Get4, Get5 and Sgt2) or SRP-dependent targeting components (Sbh1, Sec71, and

Sec72) are shown. **C-D.** Fluorescence micrographs of ER-localized BEPs that are N- or C-terminally tagged with mNG, in panels C and D, respectively, are shown. The percentage of cells with the indicated number of puncta per cell are indicated as a stacked distribution in Figure 3B-C, respectively, are based on the images shown in these panels.

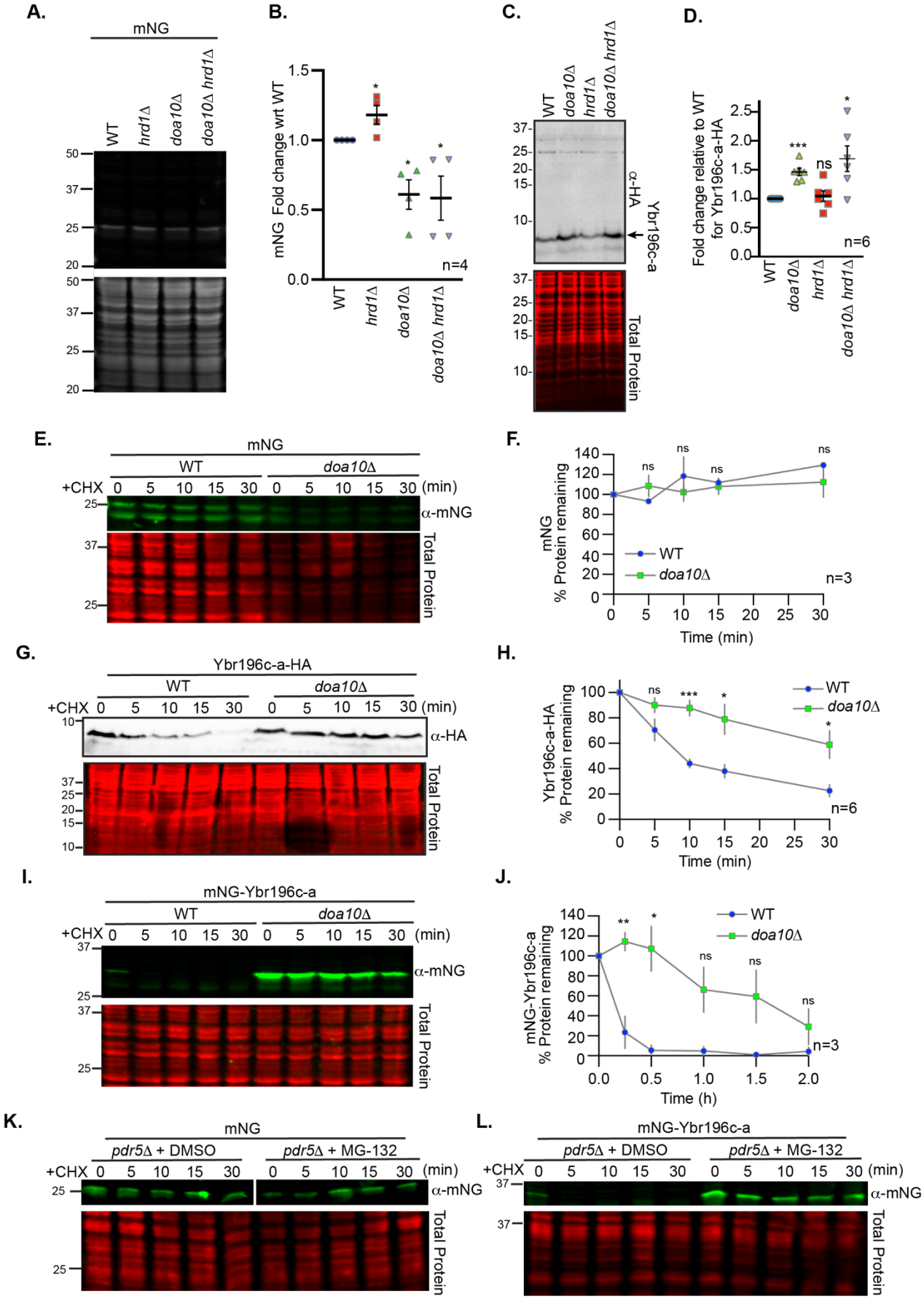

A.
mNG
WT
hrd1Δ
doa10Δ
doa10Δ hrd1Δ
50
37
25
20
B.
mNG Fold change wrt WT
1.5
1.0
0.5
0.0
n=4
C.
WT
doa10Δ
hrd1Δ
doa10Δ hrd1Δ
α-HA
Ybr196c-a
Total Protein
D.
Fold change relative to WT for Ybr196c-a-HA
2.5
2.0
1.5
1.0
0.5
0.0
***
ns
n=6
E.
mNG
WT
doa10Δ
+CHX 0 5 10 15 30 0 5 10 15 30 (min)
α-mNG
Total Protein
F.
mNG % Protein remaining
140
120
100
80
60
40
20
0
ns
WT
doa10Δ
n=3
Time (min)
G.
Ybr196c-a-HA
WT
doa10Δ
+CHX 0 5 10 15 30 0 5 10 15 30 (min)
α-HA
Total Protein
H.
Ybr196c-a-HA % Protein remaining
ns
***
*
WT
doa10Δ
n=6
Time (min)
I.
mNG-Ybr196c-a
WT
doa10Δ
+CHX 0 5 10 15 30 0 5 10 15 30 (min)
α-mNG
Total Protein
J.
mNG-Ybr196c-a % Protein remaining
**
*
ns
WT
doa10Δ
n=3
Time (h)
K.
mNG
pdr5Δ + DMSO
pdr5Δ + MG-132
+CHX 0 5 10 15 30 0 5 10 15 30 (min)
α-mNG
Total Protein
L.
mNG-Ybr196c-a
pdr5Δ + DMSO
pdr5Δ + MG-132
+CHX 0 5 10 15 30 0 5 10 15 30 (min)
α-mNG
Total Protein

Figure S7. mNG stability is not dependent upon Doa10 or the proteasome, but N-terminally, mNG-tagged or C-terminally, HA-tagged Ybr196c-a stability remains dependent on both these factors. (accompanies **Figure 4**)
A and **C.** Representative immunoblots of protein extracts made for **A,** free mNG, and **C**, C-terminally HA-tagged Ybr196c-a in cells with the indicated gene deletions. **B** and **D.** The relative abundance of **B,** free mNG, or **D**, Ybr196c-a-HA from biological replicates (n) of the immunoblots in **A** and **C**, respectively. A Student's t-test comparing the WT to the Ub ligase deletions was used to assess statistical differences (ns: not significant; *:$p < 0.05$; ***:$p < 0.0005$). **E, G,** and **I.** Representative immunoblots of protein extracts expressing **E**, free mNG, **G**, Ybr196c-a-HA, and **I,** mNG-Ybr196c-a in WT or *doa10*Δ cells post cycloheximide (CHX) addition for the times indicated (in minutes) are shown. **F, H,** and **J.** Quantification of the percent of **F**, free mNG, **H**, Ybr196-c-a-HA, or **J**, mNG-Ybr196c-a remaining post-CHX addition is plotted. The symbols represent the mean of the biological replicates (n), and the error bars show the standard error of the mean. A Student's t-test comparing the WT to the *doa10*Δ cells at each time point was used to assess statistical differences (ns: not significant; *:$p < 0.05$; ***:$p < 0.0005$). **K-L.** Representative immunoblots of protein extracts expressing **K**, free mNG or **L**, mNG-Ybr196c-a in *pdr5*Δ cells that were untreated (+DMSO) or treated with the proteasomal inhibitor MG-132 post-CHX addition are shown. Deletion of Pdr5 allows for better retention of MG-132 in cells.

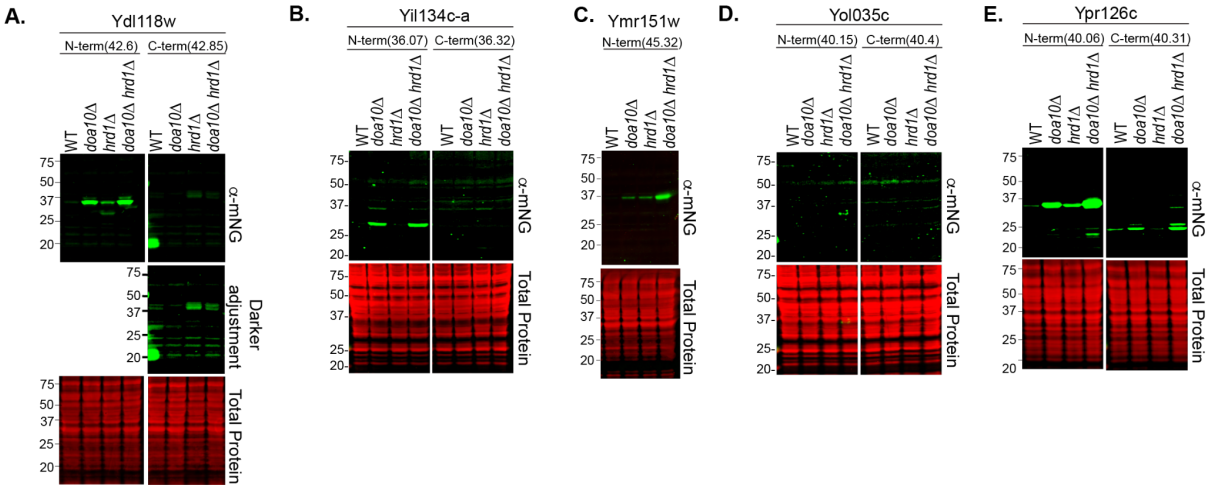


Figure S8. Most ER-localized BEPs are stabilized by loss of E3 Ub ligases Doa10 and/or Hrd1. (accompanies **Figure 4**)

A-E. Representative immunoblots of protein extracts made for each ER-localized BEP expressing the BEP fused N or C-terminally with mNG in the indicated cells are shown, as was done for Ybr196c-a in Figure 4A. Expected sizes (MW) of the BEP fusions that are N-terminally (N-term) or C-terminally (C-term)-tagged with mNG are listed in parenthesis below each BEP name.

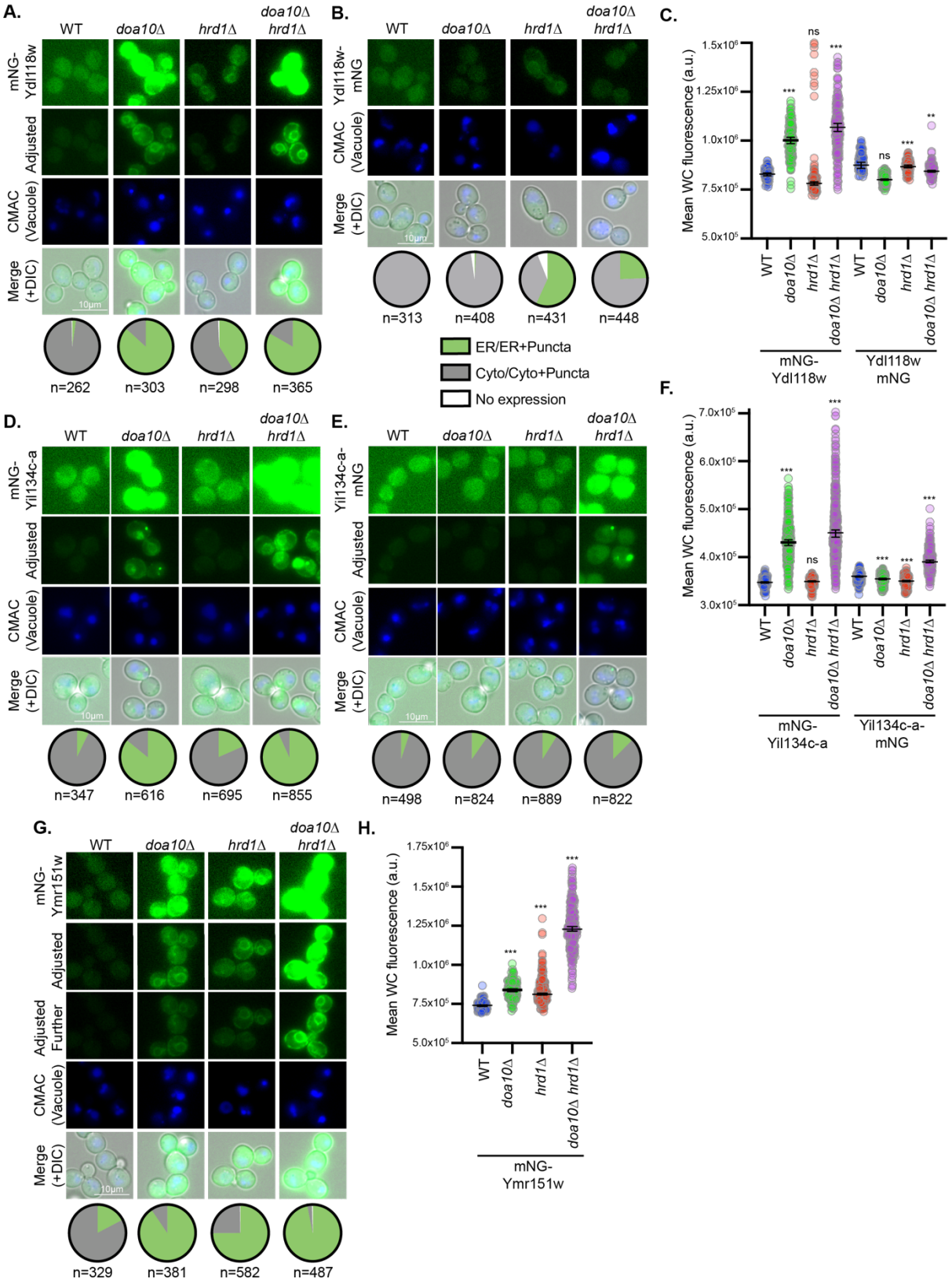
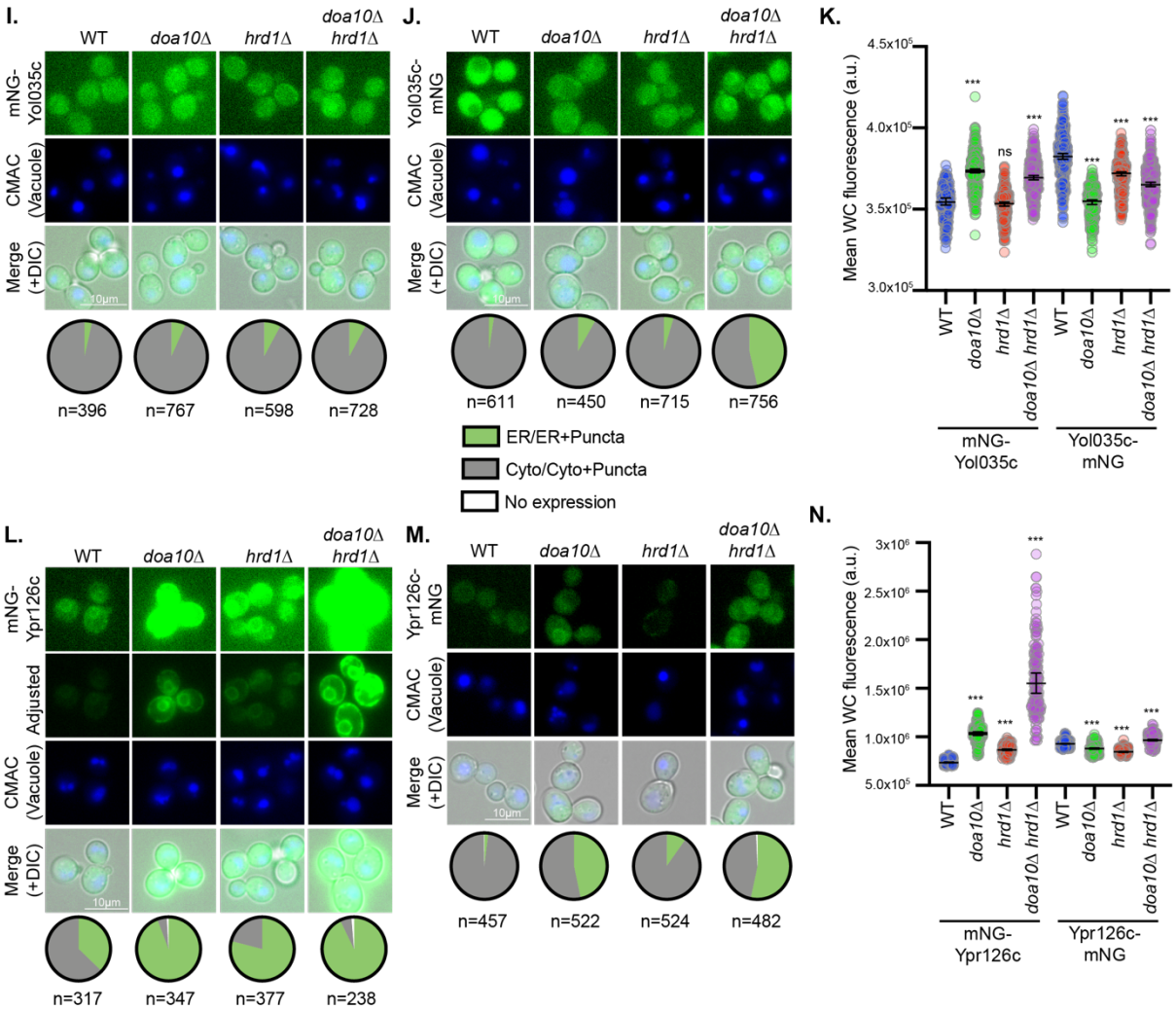


Figure S9. ER-localized BEPs increase in abundance when ER-resident Ub ligases are lost. (accompanies **Figure 5**)

Representative fluorescence microscopy of N- or C-terminally tagged BEPs with mNG are shown for **A-B,** Ydl118w; **D-E,** Yil134c-a **G,** Ymr151w, **I-J,** Yol035c, and **L-M,** Ypr126c in the cells with the indicated genotypes. These are similar to the data shown for Ybr196c-a in Figure 5A and 5B. The distribution of fluorescence observed across the cell populations is

indicated in the pie chart below each diagram (n=number of cells assessed). Cells are stained with CMAC blue to mark vacuoles and imaged via DIC, which is shown in the merge. **C, F, H, K,** and **N.** Quantification of the whole cell (WC) fluorescence intensities from individual cells (n=130-479) determined using the NIS.*ai* and GA3 software (see methods), for the cells shown in panels **A-B**, **D-E**, **G**, **I-J,** and **L-M,** respectively. A one-way ANOVA with Dunnett's multiple comparisons test was performed relative to the fluorescence in WC cells to identify statistically significant changes (p-value < 0.0005 = ***).



Continued Figure S9. ER-localized BEPs increase in abundance when ER-resident Ub ligases are lost (accompanies **Figure 5**).

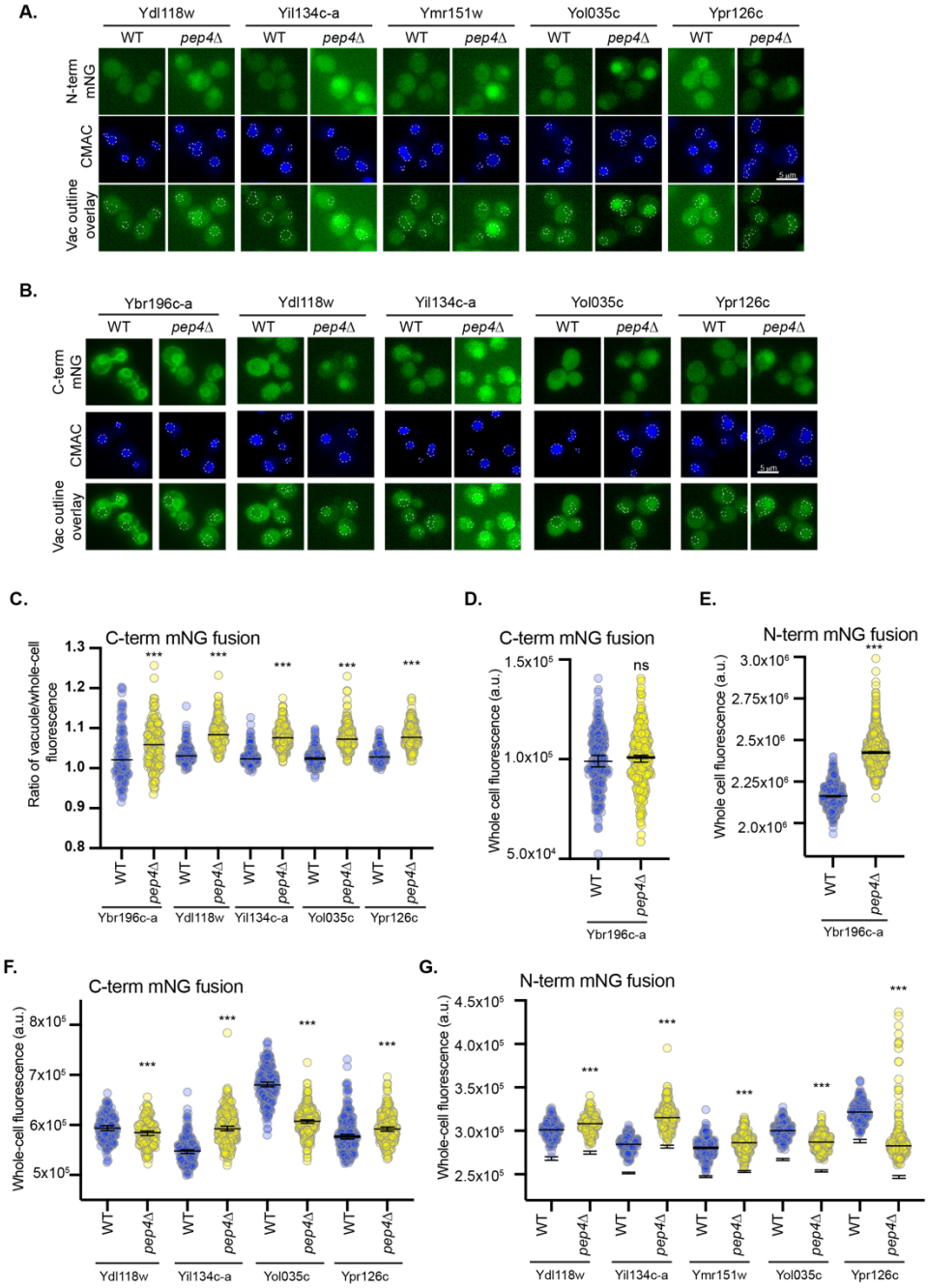
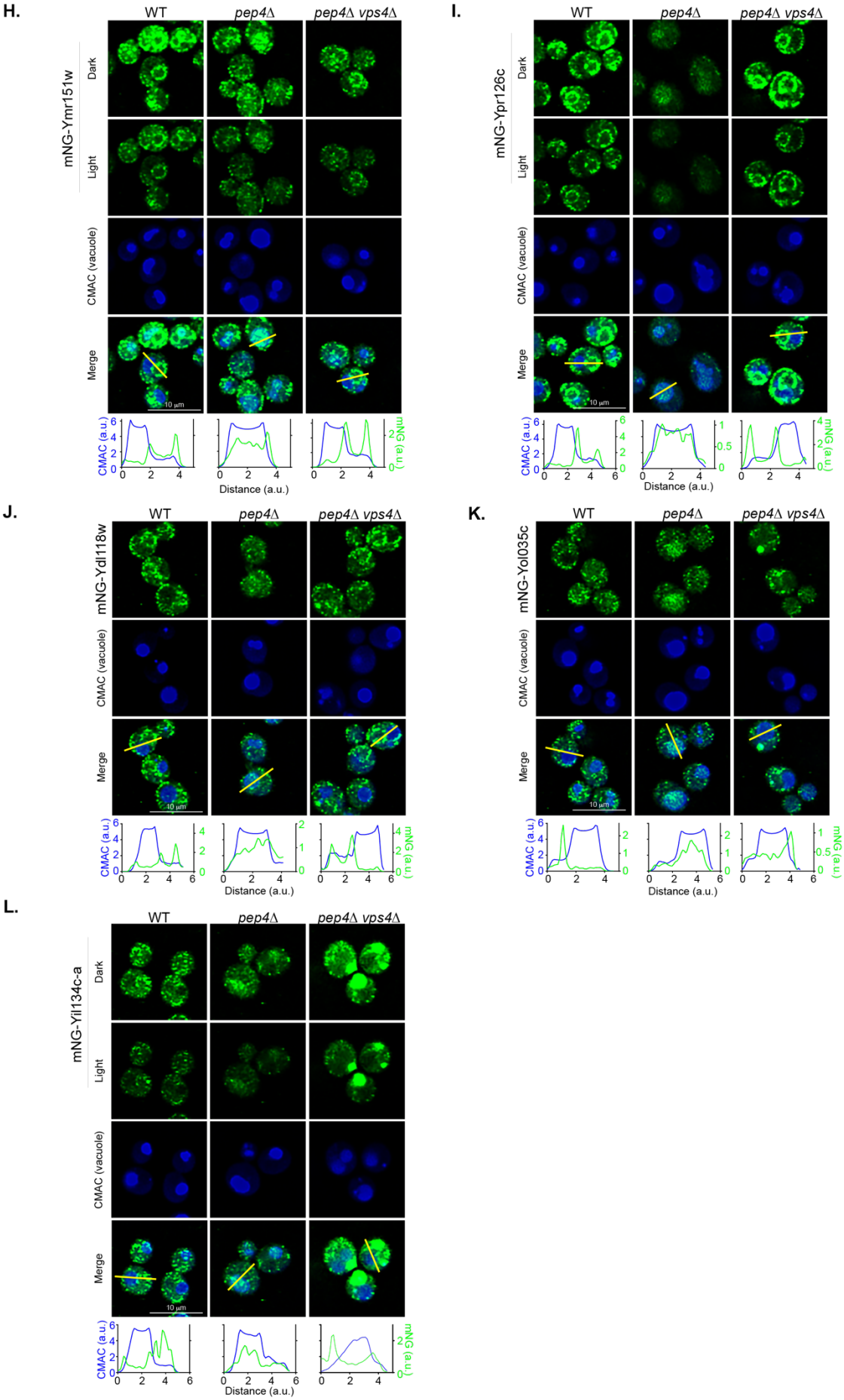


Figure S10. ER-localized BEPs are stabilized by loss of vacuolar proteases and depend on ESCRT III to reach the vacuolar lumen. (accompanies **Figure 6**)

A-B. Fluorescence micrographs of cells expressing N-terminal in panel **A** or C-terminal mNG-tagged BEPs in panel **B** in WT or *pep4*Δ cells are shown. Vacuoles are stained with CMAC blue. Vacuoles are indicated on the green channel images using a white dashed line. **C.** Scatter plot graphing the ratio of the vacuolar fluorescence over the whole cell fluorescence for the cells imaged in panel B is provided. **D-G.** The scatter plots indicate the whole-cell fluorescence intensities, as determined using NIS.*ai* and GA3 software, for imaging shown in B (panel **D** and **F**), Figure 6A (panel **E**), and A (panel **G**). The median fluorescence intensity and 95% confidence interval are represented by the horizontal line. **C-G.** A Mann-Whitney two-sided

test assessed statistical difference between the WT and *pep4*Δ cell populations (n=136-1294; $p < 0.0005$ = ***). **H-L.** Fluorescence micrographs of cells expressing mNG-BEPs in cells of the indicated genetic background are shown. Vacuoles are shown using CMAC blue. Plots of the fluorescence intensity from the blue and green channels along a line scan that runs through the vacuole (indicated in yellow on merge) are shown below each set of images. Overlapping maximum peaks of blue and green lines indicate mNG presence in the vacuolar lumen.


H.
WT
pep4Δ
pep4Δ vps4Δ
mNG-Ymr151w
Dark
Light
CMAC (vacuole)
Merge
10 µm
CMAC (a.u.)
mNG (a.u.)
Distance (a.u.)
I.
mNG-Ypr126c
J.
mNG-Ydl118w
K.
mNG-Yol035c
L.
mNG-Yil134c-a

Continued Figure S10. ER-localized BEPs are stabilized by loss of vacuolar proteases and depend on ESCRT III to reach the vacuolar lumen. (accompanies **Figure 6**).

Data Availability

All raw values and statistics for imaging in this study and Supplementary table 1 and 2 are available for download: **https://github.com/CarvunisLab/ER_BEPS_2025**

Code Availability

All custom scripts to run statistical analysis and analyze BEP sequences is available for download: **https://github.com/CarvunisLab/ER_BEPS_2025**

Acknowledgments

We thank the Dr. Jeff Brodsky and his lab (Univ. of Pittsburgh) for their discussions of this work. We further acknowledge the technical assistance of Dr. Simon Watkins (Director of the Center for Biologic Imaging, Dept of Cell Biology, Univ. of Pittsburgh), Christina Goldbach (Nikon Instruments Inc.) and Dr. Chaowei Shang (Director of Microscopy and Imaging, Dietrich School of Arts and Sciences, Univ. of Pittsburgh) to establish the high-content imaging approaches used in this work. Additionally, we would like to thank all the members of the Carvunis and O'Donnell labs, as well as, the Pittsburgh Area Yeast Meeting group (PAYM) and Molecular Evolution Laboratory Discussion group (MELD), for their feedback and suggestions on various stages of this work. Figures 1A and 6E were created using BioRender.com. This work was supported by funds provided by the National Institute of General Medical Sciences of the National Institutes of Health grant DP2GM137422 awarded to A.-R.C. and the National Science Foundation grant MCB-2144349 awarded to A.-R.C, a subcontract from this grant awarded to A.F.O. and start-up funds provided by the Dept. of Biological Sciences, Univ of Pittsburgh to A.F.O.

Author contributions

Conceptualization, NCC, CJH, SBP, ARC, AFO.; methodology, CJH, NCC, AC, SH, NO, SBP, AFO.; investigation, CJH, NCC, AC, SH, SBP, NO, JI, AB, AFO; writing – original draft, AFO.; writing – review & editing, CJH, NCC, AC, AW, MW, ARC, AFO.; supervision, ARC and AFO.

Declaration of interests

A.-R.C. is a member of the Scientific Advisory Board for Flagship Labs 69, Inc. (ProFound Therapeutics).

References

1 Weisman, C. M. The Origins and Functions of De Novo Genes: Against All Odds? *J Mol Evol* **90**, 244-257, doi:10.1007/s00239-022-10055-3 (2022).

2 Van Oss, S. B. & Carvunis, A. R. De novo gene birth. *PLoS Genet* **15**, e1008160, doi:10.1371/journal.pgen.1008160 (2019).

3 Broeils, L. A., Ruiz-Orera, J., Snel, B., Hubner, N. & van Heesch, S. Evolution and implications of de novo genes in humans. *Nat Ecol Evol* **7**, 804-815, doi:10.1038/s41559-023-02014-y (2023).

4 Zhao, L., Svetec, N. & Begun, D. J. De Novo Genes. *Annu Rev Genet*, doi:10.1146/annurev-genet-111523-102413 (2024).

5 Parikh, S. B., Houghton, C., Van Oss, S. B., Wacholder, A. & Carvunis, A. R. Origins, evolution, and physiological implications of de novo genes in yeast. *Yeast* **39**, 471-481, doi:10.1002/yea.3810 (2022).

6 Bohnsack, M. T. & Schleiff, E. The evolution of protein targeting and translocation systems. *Biochim Biophys Acta* **1803**, 1115-1130, doi:10.1016/j.bbamcr.2010.06.005 (2010).

7 Gabaldon, T. & Pittis, A. A. Origin and evolution of metabolic sub-cellular compartmentalization in eukaryotes. *Biochimie* **119**, 262-268, doi:10.1016/j.biochi.2015.03.021 (2015).

8 Powers, E. T. & Balch, W. E. Diversity in the origins of proteostasis networks--a driver for protein function in evolution. *Nat Rev Mol Cell Biol* **14**, 237-248, doi:10.1038/nrm3542 (2013).

9 Rebeaud, M. E., Mallik, S., Goloubinoff, P. & Tawfik, D. S. On the evolution of chaperones and cochaperones and the expansion of proteomes across the Tree of Life. *Proc Natl Acad Sci U S A* **118**, doi:10.1073/pnas.2020885118 (2021).

10 Vakirlis, N. *et al.* De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nat Commun* **11**, 781, doi:10.1038/s41467-020-14500-z (2020).

11 Dunn, C. D. & Paavilainen, V. O. Wherever I may roam: organellar protein targeting and evolvability. *Curr Opin Genet Dev* **58-59**, 9-16, doi:10.1016/j.gde.2019.07.012 (2019).

12 Kesner, J. S. *et al.* Noncoding translation mitigation. *Nature* **617**, 395-402, doi:10.1038/s41586-023-05946-4 (2023).

13 Kesner, J. S. & Wu, X. Mechanisms suppressing noncoding translation. *Trends Cell Biol*, doi:10.1016/j.tcb.2024.09.004 (2024).

14 Makarewich, C. A. The hidden world of membrane microproteins. *Exp Cell Res* **388**, 111853, doi:10.1016/j.yexcr.2020.111853 (2020).

15 Tassios, E., Nikolaou, C. & Vakirlis, N. Intergenic Regions of Saccharomycotina Yeasts are Enriched in Potential to Encode Transmembrane Domains. *Mol Biol Evol* **40**, doi:10.1093/molbev/msad059 (2023).

16 Vakirlis, N. & Fuqua, T. Intergenic polyA/T tracts explain the propensity of yeast de novo genes to encode transmembrane domains. *bioRxiv* (2025).

17 Rich, A., Acar, O. & Carvunis, A. R. Massively integrated coexpression analysis reveals transcriptional regulation, evolution and cellular implications of the yeast noncanonical translatome. *Genome Biol* **25**, 183, doi:10.1186/s13059-024-03287-7 (2024).

18 Sandmann, C. L. *et al.* Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. *Mol Cell* **83**, 994-1011 e1018, doi:10.1016/j.molcel.2023.01.023 (2023).

19 Zhuang, X., Yang, C., Murphy, K. R. & Cheng, C. C. Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. *Proc Natl Acad Sci U S A* **116**, 4400-4405, doi:10.1073/pnas.1817138116 (2019).

20 Veatch, J. R., McMurray, M. A., Nelson, Z. W. & Gottschling, D. E. Mitochondrial dysfunction leads to nuclear genome instability via an iron-sulfur cluster defect. *Cell* **137**, 1247-1258, doi:10.1016/j.cell.2009.04.014 (2009).

21 McIsaac, R. S. *et al.* Fast-acting and nearly gratuitous induction of gene expression and protein depletion in Saccharomyces cerevisiae. *Mol Biol Cell* **22**, 4447-4459, doi:10.1091/mbc.E11-05-0466 (2011).

22 McIsaac, R. S. *et al.* Synthetic gene expression perturbation systems with rapid, tunable, single-gene specificity in yeast. *Nucleic Acids Res* **41**, e57, doi:10.1093/nar/gks1313 (2013).

23 Chong, Y. T. *et al.* Yeast Proteome Dynamics from Single Cell Imaging and Automated Analysis. *Cell* **161**, 1413-1424, doi:10.1016/j.cell.2015.04.051 (2015).

24 Andrews, S. J. & Rothnagel, J. A. Emerging evidence for functional peptides encoded by short open reading frames. *Nat Rev Genet* **15**, 193-204, doi:10.1038/nrg3520 (2014).

25 Kall, L., Krogh, A. & Sonnhammer, E. L. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* **338**, 1027-1036, doi:10.1016/j.jmb.2004.03.016 (2004).

26 Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, 567-580, doi:10.1006/jmbi.2000.4315 (2001).

27 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589, doi:10.1038/s41586-021-03819-2 (2021).

28 Akopian, D., Shen, K., Zhang, X. & Shan, S. O. Signal recognition particle: an essential protein-targeting machine. *Annu Rev Biochem* **82**, 693-721, doi:10.1146/annurev-biochem-072711-164732 (2013).

29 Almagro Armenteros, J. J. *et al.* Detecting sequence signals in targeting peptides using deep learning. *Life Sci Alliance* **2**, doi:10.26508/lsa.201900429 (2019).

30 Teufel, F. *et al.* SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat Biotechnol* **40**, 1023-1025, doi:10.1038/s41587-021-01156-3 (2022).

31 Aviram, N. & Schuldiner, M. Targeting and translocation of proteins to the endoplasmic reticulum at a glance. *J Cell Sci* **130**, 4079-4085, doi:10.1242/jcs.204396 (2017).

32 Borgese, N., Coy-Vergara, J., Colombo, S. F. & Schwappach, B. The Ways of Tails: the GET Pathway and more. *Protein J* **38**, 289-305, doi:10.1007/s10930-019-09845-4 (2019).

33 Rao, M. *et al.* Multiple selection filters ensure accurate tail-anchored membrane protein targeting. *Elife* **5**, doi:10.7554/eLife.21301 (2016).

34 Borgese, N., Brambillasca, S. & Colombo, S. How tails guide tail-anchored proteins to their destinations. *Curr Opin Cell Biol* **19**, 368-375, doi:10.1016/j.ceb.2007.04.019 (2007).

35 Singh, S. & Mittal, A. Transmembrane Domain Lengths Serve as Signatures of Organismal Complexity and Viral Transport Mechanisms. *Sci Rep* **6**, 22352, doi:10.1038/srep22352 (2016).

36 Schuldiner, M. *et al.* The GET complex mediates insertion of tail-anchored proteins into the ER membrane. *Cell* **134**, 634-645, doi:10.1016/j.cell.2008.06.025 (2008).

37 Aviram, N. *et al.* The SND proteins constitute an alternative targeting route to the endoplasmic reticulum. *Nature* **540**, 134-138, doi:10.1038/nature20169 (2016).

38 Weill, U. *et al.* Assessment of GFP Tag Position on Protein Localization and Growth Fitness in Yeast. *J Mol Biol* **431**, 636-641, doi:10.1016/j.jmb.2018.12.004 (2019).

39 Tirincsi, A., Sicking, M., Hadzibeganovic, D., Hassdenteufel, S. & Lang, S. The Molecular Biodiversity of Protein Targeting and Protein Transport Related to the Endoplasmic Reticulum. *Int J Mol Sci* **23**, doi:10.3390/ijms23010143 (2021).

40 Swanson, R., Locher, M. & Hochstrasser, M. A conserved ubiquitin ligase of the nuclear envelope/endoplasmic reticulum that functions in both ER-associated and Matalpha2 repressor degradation. *Genes Dev* **15**, 2660-2674, doi:10.1101/gad.933301 (2001).

41 Schneider-Poetsch, T. *et al.* Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. *Nat Chem Biol* **6**, 209-217, doi:10.1038/nchembio.304 (2010).

42 Nakatsukasa, K. & Brodsky, J. L. The recognition and retrotranslocation of misfolded proteins from the endoplasmic reticulum. *Traffic* **9**, 861-870, doi:10.1111/j.1600-0854.2008.00729.x (2008).

43 Lemberg, M. K. & Strisovsky, K. Maintenance of organellar protein homeostasis by ER-associated degradation and related mechanisms. *Mol Cell* **81**, 2507-2519, doi:10.1016/j.molcel.2021.05.004 (2021).

44 Wu, K. *et al.* Substrate recognition mechanism of the endoplasmic reticulum-associated ubiquitin ligase Doa10. *Nat Commun* **15**, 2182, doi:10.1038/s41467-024-46409-2 (2024).

45 Li, S. C. & Kane, P. M. The yeast lysosome-like vacuole: endpoint and crossroads. *Biochim Biophys Acta* **1793**, 650-663, doi:10.1016/j.bbamcr.2008.08.003 (2009).

46 Hecht, K. A., O'Donnell, A. F. & Brodsky, J. L. The proteolytic landscape of the yeast vacuole. *Cell Logist* **4**, e28023, doi:10.4161/cl.28023 (2014).

47 Woolford, C. A. *et al.* The PEP4 gene encodes an aspartyl protease implicated in the posttranslational regulation of Saccharomyces cerevisiae vacuolar hydrolases. *Mol Cell Biol* **6**, 2500-2510, doi:10.1128/mcb.6.7.2500-2510.1986 (1986).

48 Li, Y., Kane, T., Tipper, C., Spatrick, P. & Jenness, D. D. Yeast mutants affecting possible quality control of plasma membrane proteins. *Mol Cell Biol* **19**, 3588-3599, doi:10.1128/MCB.19.5.3588 (1999).

49 Nikko, E. & Pelham, H. R. Arrestin-mediated endocytosis of yeast plasma membrane transporters. *Traffic* **10**, 1856-1867, doi:10.1111/j.1600-0854.2009.00990.x (2009).

50 Knupp, J., Pletan, M. L., Arvan, P. & Tsai, B. Autophagy of the ER: the secretome finds the lysosome. *FEBS J* **290**, 5656-5673, doi:10.1111/febs.16986 (2023).

51 Henne, W. M., Buchkovich, N. J. & Emr, S. D. The ESCRT pathway. *Dev Cell* **21**, 77-91, doi:10.1016/j.devcel.2011.05.015 (2011).

52 Bowers, K. & Stevens, T. H. Protein transport from the late Golgi to the vacuole in the yeast Saccharomyces cerevisiae. *Biochim Biophys Acta* **1744**, 438-454, doi:10.1016/j.bbamcr.2005.04.004 (2005).

53 Makarova, K. S. *et al.* Diversity, origin, and evolution of the ESCRT systems. *mBio* **15**, e0033524, doi:10.1128/mbio.00335-24 (2024).

54 Wacholder, A. *et al.* A vast evolutionarily transient translatome contributes to phenotype and fitness. *Cell Syst* **14**, 363-381 e368, doi:10.1016/j.cels.2023.04.002 (2023).

55 Heames, B., Schmitz, J. & Bornberg-Bauer, E. A Continuum of Evolving De Novo Genes Drives Protein-Coding Novelty in Drosophila. *J Mol Evol* **88**, 382-398, doi:10.1007/s00239-020-09939-z (2020).

56 Dowling, D., Schmitz, J. F. & Bornberg-Bauer, E. Stochastic Gain and Loss of Novel Transcribed Open Reading Frames in the Human Lineage. *Genome Biol Evol* **12**, 2183-2195, doi:10.1093/gbe/evaa194 (2020).

57 Coughlin, T. M. & Makarewich, C. A. Emerging roles for microproteins as critical regulators of endoplasmic reticulum function and cellular homeostasis. *Semin Cell Dev Biol* **170**, 103608, doi:10.1016/j.semcdb.2025.103608 (2025).

58 Chen, Y., Shanmugam, S. K. & Dalbey, R. E. The Principles of Protein Targeting and Transport Across Cell Membranes. *Protein J* **38**, 236-248, doi:10.1007/s10930-019-09847-2 (2019).

59 Mehlhorn, D. G., Asseck, L. Y. & Grefen, C. Looking for a safe haven: tail-anchored proteins and their membrane insertion pathways. *Plant Physiol* **187**, 1916-1928, doi:10.1093/plphys/kiab298 (2021).

60 Stayton, C. T. What does convergent evolution mean? The interpretation of convergence and its implications in the search for limits to evolution. *Interface Focus* **5**, 20150039, doi:10.1098/rsfs.2015.0039 (2015).

61 Carvunis, A. R. *et al.* Proto-genes and de novo gene birth. *Nature* **487**, 370-374, doi:10.1038/nature11184 (2012).

62 Casola, C., Owoyemi, A. & Vakirlis, N. Degradation determinants are abundant in human noncanonical proteins. *bioRxiv*, 2024.2005.2001.592071, doi:10.1101/2024.05.01.592071 (2024).

63 Newstead, S. & Barr, F. Molecular basis for KDEL-mediated retrieval of escaped ER-resident proteins - SWEET talking the COPs. *J Cell Sci* **133**, doi:10.1242/jcs.250100 (2020).

64 Gomez-Navarro, N. *et al.* Cargo crowding contributes to sorting stringency in COPII vesicles. *J Cell Biol* **219**, doi:10.1083/jcb.201806038 (2020).

65 Vitali, D. G. *et al.* The GET pathway can increase the risk of mitochondrial outer membrane proteins to be mistargeted to the ER. *J Cell Sci* **131**, doi:10.1242/jcs.211110 (2018).

66 Amberg, D. C., Burke, D., Strathern, J. N. & Laboratory, C. S. H. *Methods in Yeast Genetics: A Cold Spring Harbor Laboratory Course Manual*. (Cold Spring Harbor Laboratory Press, 2005).

67 Alberti, S., Gitler, A. D. & Lindquist, S. A suite of Gateway cloning vectors for high-throughput genetic analysis in Saccharomyces cerevisiae. *Yeast* **24**, 913-919, doi:10.1002/yea.1502 (2007).

68 Amberg, D. C. Methods in yeast genetics: a Cold Spring Harbor Laboratory course manual. *(No Title)* (2005).

69 Dunham MJ, D. M., Gartenberg MR, Brown GW. Methods in yeast genetics and genomics : a Cold Spring Harbor Laboratory course manual. (2015).

70 Hager, N. A. *et al.* Select alpha-arrestins control cell-surface abundance of the mammalian Kir2.1 potassium channel in a yeast model. *J Biol Chem* **293**, 11006-11021, doi:10.1074/jbc.RA117.001293 (2018).

71 Volland, C., Urban-Grimal, D., Geraud, G. & Haguenauer-Tsapis, R. Endocytosis and degradation of the yeast uracil permease under adverse conditions. *J Biol Chem* **269**, 9833-9841 (1994).

72 Bowman, R. W., 2nd *et al.* TORC1 Signaling Controls the Stability and Function of alpha-Arrestins Aly1 and Aly2. *Biomolecules* **12**, doi:10.3390/biom12040533 (2022).

73 Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* **50**, D439-D444, doi:10.1093/nar/gkab1061 (2022).

74 Peng, J., Svetec, N., Molina, H. & Zhao, L. The Origin and Evolution of Sex Peptide and Sex Peptide Receptor Interactions. *Mol Biol Evol* **41**, doi:10.1093/molbev/msae065 (2024).

75 Terwilliger, T. C. *et al.* AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nat Methods* **21**, 110-116, doi:10.1038/s41592-023-02087-4 (2024).