# Noncanonical usage of stop codons in ciliates expands proteins with structurally flexible Q-rich motifs

Chi-Ning Chuang[1, #], Hou-Cheng Liu[1, #], Tai-Ting Woo[1, ‡], Ju-Lan Chao[1], Chiung-Ya Chen[1], Hisao-Tang Hu[1], Yi-Ping Hsueh[1,2], and Ting-Fang Wang[*,1,2]

[1]Institute of Molecular Biology, Academia Sinica, Taipei 115, Taiwan
[2]Department of Biochemical Science and Technology, National Chiayi University, Chiayi, Taiwan

*corresponding author: Ting-Fang Wang
Email: tfwang@gate.sinica.edu.tw

[‡]Current affiliation: Department of Cell and Developmental Biology, University of Michigan Medical School, Ann Arbor, Michigan, United States of America

[#]These two authors contributed equally to this work.

**Classification:**

**Keywords:** noncanonical genetic codons; S/T-Q cluster domain (SCD); polyglutamine (polyQ) and polyglutamine/asparagine (polyQ/N); intrinsically disordered region (IDR)

**This PDF file includes:**
Main text file, nine figures and seven tables

**The supplemental datasets include 32 excel tables (Datasets DS1-DS32).**

## Abstract

Serine(S)/threonine(T)-glutamine(Q) cluster domains (SCDs), polyglutamine (polyQ) tracts and polyglutamine/asparagine (polyQ/N) tracts are Q-rich motifs found in many proteins. SCDs often are intrinsically disordered regions that mediate protein phosphorylation and protein-protein interactions. PolyQ and polyQ/N tracts are structurally flexible sequences that trigger protein aggregation. We show that four SCDs and three prion-causing Q/N-rich motifs of yeast proteins possess autonomous protein expression-enhancing activities. Comparative Gene Ontology (GO) analyses of the near-complete proteomes of 27 representative model eukaryotes reveal that Q-rich motifs prevail in proteins involved in specialized biological processes, including *Saccharomyces cerevisiae* RNA-mediated transposition, *Candida albicans* filamentous growth, ciliate peptidyl-glutamic acid modification, *Tetrahymena thermophila* xylan catabolism and meiosis, *Dictyostelium discoideum* development and sexual cycles, *Plasmodium falciparum* infection, and the *Drosophila melanogaster* nervous system. We also show that Q-rich motifs are expanded massively in ten ciliates with reassigned $TAA^{Q}$ and $TAG^{Q}$ codons. Our results provide new insights to explain why many ciliates reassign their nuclear stop codons into glutamine (Q). The consequence of this preponderance of Q is massive expansion of proteins harboring three structurally flexible or even intrinsically disordered Q-rich motifs. Since these Q-rich motifs can endow proteins with structural and functional plasticity, we suggest that they represent useful toolkits for evolutionary novelty.

## Introduction

Ciliated protists (ciliates) are a large and diverse group of microbial eukaryotes that are important to a wide range of biological studies and applications. Relative to most other eukaryotes, the nuclear genomes of many ciliates possess three interesting features. First, ciliates often have two types of nuclei. Their diploid micronucleus (MIC) carries the cell germline, the genetic material of which is inherited via sexual reproduction and meiosis. The polyploid macronucleus (MAC) or vegetative nucleus provides nuclear RNA for vegetative growth. MAC is generated from the MIC by massive amplification, editing and rearrangement of the genome (see reviews in (1, 2)). Second, ciliate meiosis is remarkable relative to that of other studied sexual eukaryotes. A hallmark of meiosis is programmed initiation of DNA double-strand breaks (DSBs) by Spo11, a meiosis-specific topoisomerase VII subunit A endonuclease (3, 4). Meiotic DSBs are then repaired in a homology-dependent manner via searching, pairing (or synapsis) and exchange of homologous parental chromosomes, leading to the formation of interhomolog crossovers (COs). Ultimately, the resulting recombinant chromosomes equally segregate into haploid gametes. To facilitate pairing and recombination between homologous chromosomes, telomeres attach to the nuclear envelope (NE) at the beginning of meiosis and then cluster to one side of the NE to form a transient structure called the "bouquet". In many sexual eukaryotes, homology pairing usually culminates in the synaptonemal complex (SC), a zipper-like proteinaceous structure that appears to regulate CO formation. The SC behaves like a liquid crystal, tightly connecting paired homologous chromosomes from telomere to telomere and also incorporating lateral component loops in interstitial chromosome segments (5, 6). Most studied ciliates lack SC or only have degenerate SCs (7). In

*Tetrahymena thermophila*, the most intensively studied ciliate, meiotic MICs undergo extreme elongation (by ~50-fold) and form proteinaceous condensates called "crescents". Within these elongated crescents, telomeres and centromeres of all meiotic chromosomes are rearranged at opposing ends in a stretched bouquet-like manner. Meiotic pairing and recombination take place within the crescents (see review in (8)). It has been reported that ATR1 (Ataxia Telangiectasia Mutated 1), an evolutionarily conserved DNA damage senor protein kinase, senses Spo11-induced DSBs and triggers the elongation of MICs (9). Meiosis-specific CYC2 and CYC17 cyclins, as well as cyclin-dependent kinase CDK3, are required to initiate meiosis and for crescent assembly (10-12). CYC2/CDK2 promotes bouquet formation in MIC via controlling microtubule-directed elongation (12) and also controls gene expression of proteins involved in DSB formation (SPO11), DNA repair (COM1, EXO1, DMC1), and CO formation (HOP2, MND1, MSH4, MSH5, ZPH3, BIM1, and BIM2) (13). The DPL2/E2fl1 complex, a meiosis-specific transcription factor, promotes transcriptional induction of DNA repair proteins and chromosome-associated structural proteins, including MRE11, COM1, EXO1, RAD50, RAD51, SMC1, SMC2, SMC3, SMC4, REC8, ESP1, and CNA1, among others (13). Nevertheless, the molecular mechanisms underlying crescent assembly and disassembly remain poorly understood. Third, many ciliates reassign their standard stop codons into amino acids (Table 1). For example, several ciliates possess two noncanonical nuclear genetic codes ($UAA^{Q}$ and $UAG^{Q}$), in which the UAA and UAG stop codons have been reassigned to glutamine (Q) so that UGA is the sole functional stop codon in many ciliates, including *Tetrahymena thermophila*, *Paramecium tetraurelia*, *Paramecium bursaria*, *Oxytricha trifallax, Stylonychia lemnae*, *Pseudocohnilembus persalinus*, *Aristerostoma* sp., *Favella*

*ehrenbergii, Pseudokeronopsis* spp., *Strombidium inclinatum,* and *Uronema* spp. Both the UAA and UAG stop codons are reassigned to tyrosine (Y) in *Favella ehrenbergii*, whereas the UGA stop codon is translated to tryptophan (W) or cysteine (C) in *Blepharisma japonicum* and *Euplotes focardii,* respectively. In contrast, *Stentor coeruleus, Climacostomum virens, Litonotus pictus* and *Protocruzia adherens* utilize the universal set of genetic codons. Notably, *Condylostoma magnum* and *Parduzcia* sp. have no dedicated genetic codes. Their UAA, UAG and UGA codons can be stop codons or translated to Q, C and W, respectively. Translation termination at the mRNA 3′ end occurs in a context-dependent manner to distinguish stop from sense (14-26).

Although it has been reported previously that Q is used more frequently in *Tetrahymena thermophila* and *Paramecium tetraurelia* than in other species (19, 20), many important questions regarding stop codon reassignment in ciliates remain unresolved. For instance, fundamentally, it is unclear if Q, Y, W and C are used more frequently in other ciliates in which stop codons are reassigned. Moreover, whether there are common or specific structural motif(s) in proteins arising from stop codon reassignment is not clear. Furthermore, what are the structural and functional impacts of such genome-wide alterations? Since high-quality transcriptomic and gene annotation data are now available for many ciliate and non-ciliate eukaryotes and genetic codons have also been properly defined for a wide variety of them (Table 1), we were confident we could provide important insights into these interesting topics.

In this report, first we present an unexpected result that the Q-rich motifs of several *Saccharomyces cerevisiae* proteins possess autonomous protein expression-enhancing

(PEE) activities (27), including four serine(S)/threonine(T)-Q cluster domains (SCDs) and three prion-forming domains (PFDs) with polyglutamine (polyQ) or polyglutamine/asparagine (polyQ/N) tracts. Notably, all seven of these Q- or Q/N-rich motifs display high Q, S, T, and/or N contents. A common feature of intrinsically disordered regions (IDRs) is high content of Q, N, S, T, proline (P), glycine (G) and charged amino acids. Rather than adopting stable secondary and/or tertiary structures, IDRs are structurally flexible polypeptides in their native states (28-30). Next, we performed Gene Ontology (GO) enrichment analyses on all proteins having Q-rich motifs (i.e., SCDs, polyQ and polyQ/N), as well as those with the homorepeat (polyX) motifs of other amino acid residues, in 20 non-ciliate and 17 ciliate species. Our results support that Q-rich motifs are not only massively expanded in the 10 ciliate species we examined with reassigned $TAA^{Q}$ and $TAG^{Q}$ codons, but they also prevail in proteins involved in several species-specific or even phylum-specific biological processes.

## Results

### SCDs provide versatile functionalities in proteins

We reported previously that the $NH_2$-terminal domain (NTD; residues 1–66) of budding yeast *Saccharomyces cerevisiae* Rad51 protein (Rad51-NTD) autonomously promotes high-level production of native Rad51 and its COOH-terminal fusion protein LacZ (β-galactosidase) *in vivo* (27). In brief, we expressed Rad51-LacZ-NVH fusion proteins using a *CEN-ARS* plasmid (low-copy number) under the control of the native *RAD51* gene promoter ($P_{RAD51}$) (Table 1). The NVH tag contains an SV40 nuclear localization signal (NLS) peptide preceding a V5 epitope tag and a hexahistidine ($His_6$) affinity tag (27). We confirmed that N-terminal addition of Rad51-NTD to LacZ-NVH increased both steady-

state levels of LacZ-NVH fusion proteins (Figure 1A) and β-galactosidase activities *in vivo* (Figure 1B).

Rad51-NTD is a SCD because it has three SQ motifs ($S^2Q$, $S^{12}Q$ and $S^{30}Q$) (27). The S/T-Q motifs, comprising S or T followed by Q, are the target sites of DNA damage senor protein kinases, i.e., ATM, ATR (ATM and RAD3-related) and DNA-dependent protein kinase (DNA-PK) (31, 32). Mec1 (Mitotic Entry Checkpoint 1) and Tel1 (TELomere maintenance 1) are the budding yeast homologs of mammalian ATR and ATM, respectively. Budding yeast lacks a DNA-PK homolog (33-35). We showed that the three SQ motifs of Rad51-NTD are phosphorylated by Mec1 and Tel1 during vegetative growth and meiosis. Mec1/Tel1-dependent NTD phosphorylation antagonizes Rad51 degradation via the proteasomal pathway, increasing the half-life of Rad51 from ~30 min to ≥180 min (27). This outcome is consistent with the notion that Mec1 exhibits an essential function in regulating protein homeostasis (proteostasis) in *S. cerevisiae* (36, 37).

A unifying definition of an SCD is possession of three or more S/T-Q sites within a stretch of ≤50 or ≤100 amino acids in *S. cerevisiae* and human, respectively (31, 32). One of the best-understood mechanisms of SCD phosphorylation involves their association with binding partners hosting a forkhead-associated (FHA) domain. For example, Mec1 and Tel1 phosphorylate the SCD1 of Rad53 (residues 1-29) and the SCD of Hop1 (residues 258-324), respectively, specifically recruiting and activating their downstream DNA damage checkpoint kinases Dun1 and Mek1 during vegetative growth and meiosis (38-40). Sml1, an inhibitor of ribonucleotide reductase, is involved in regulating Rad53 and the production of deoxynucleoside triphosphate (dNTP) during DNA replication and the DNA damage response (DDR) (41). The Sml1 protein in the SK1 strain harbors three S/T-Q

motifs ($S^4Q$, $S^{14}Q$ and $T^{47}Q$), whereas that in the S288c strain only has one SQ motif ($S^4Q$, $C^{14}Q$ and $T^{47}M$). Here, we further show that yeast Rad53-SCD1, Hop1-SCD, Sml1-NTD$^{1–50}$ (residues 1–50) and Sml1-NTD$^{1–27}$ (residues 1–27) also exhibit PEE activities (Figure 1A and 1B, Table 1).

**The Q-rich motifs of three yeast prion-causing proteins also exhibit PEE activities**

Since Sml1-NTD$^{1–27}$ in the SK1 strain only harbors two S/T-Q motifs ($S^4Q$ and $S^{14}Q$), the number of S/T-Q motifs alone could not account for PEE activity. Notably, Rad51-NTD, Rad53-SCD1, Hop1-SCD, Sml1-NTD$^{1-27}$ and Sml1-NTD$^{1-50}$ all represent Q- or Q/N-rich motifs. Rad51-NTD contains 9 serines (S), 2 threonines (T), 9 glutamines (Q) and 4 asparagines (N). Rad53-SCD1 has 2S, 4T, 7Q and 1N. Hop1-SCD has 6S, 6T, 8Q and 9N. Sml1-NTD$^{1–27}$ and Sml1-NTD$^{1–50}$ in SK1 possess 3S and 5S, 2T and 3T, 6Q and 7Q, as well as 2N and 3N, respectively.

We investigated if other Q- or Q/N-rich motifs in yeast can also promote protein expression *in vivo*. PolyQ and polyQ/N tracts are the most common homorepeats acting as structurally flexible motifs for protein aggregation or protein–protein interactions in many eukaryotes (42, 43). PolyN is not as structurally flexible as polyQ due to a stronger propensity for β-turn formation in polyN than in polyQ (44). In so-called polyQ-associated diseases, long Q-, Q/N- or even N-rich motifs result in an excess of interactions, resulting in malfunctional or pathogenic protein aggregates (45). Many prion-causing proteins contain Q/N-rich prion forming domains (PFDs). In *S. cerevisiae*, the best-characterized prion-causing proteins are Sup35 (or translation terminator eRF35), New1 ([*NU*+] prion formation protein 1), Ure2 (uridosuccinate transport 2), Rnq1 (rich in N and Q 1) and Swi1 (switching deficient 1) (46, 47). We found that the Q/N-rich NTDs of Sup35, Ure2

and New1 also display PEE activities, i.e., the prion nucleation domain (PND; residues 1-39) of Sup35, the Ure2 prion domain (UPD) (residues 1-91) and the New1 prion domain (NPD; residue 1-146) (Figure 1A and 1B, Table 1). Sup35-PND containing 3S, 12Q, 18N and an $S^{17}Q$ motif exerts critical functions in promoting [*PSI*+] prion nucleation (48). The UPD of the Ure2 nitrogen catabolite repression transcriptional regulator is the basis of the prion [$URE3^{+}$] (49, 50). The UPD is critical for Ure2's function *in vivo* because its removal in the corresponding Ure2-ΔUPD mutants elicits reduced protein stability and steady-state protein levels (but not transcript levels) (51). Ure2-UPD contains 10S, 5T, 10Q and 33N, adopting a completely disordered structure (52). New1 is a non-essential ATP-binding cassette type F protein that fine-tunes the efficiency of translation termination or ribosome recycling (53). The NPD of New1 supports [$NU^{+}$] and is susceptible to [*PSI*+] prion induction (54, 55). New1-NPD contains 19S, 8T, 14Q, 28N and an $S^{145}Q$ motif. Here, we applied the LacZ-NVH fusion protein approach to show that N-terminal fusion of Sup35-PND, Ure2-UPD or New1-NPD to LacZ-NVH all increased steady-state protein levels (Figure 1A) and β-galactosidase activities *in vivo* (Figure 1B).

**The PEE function is not affected by the quaternary structures of target proteins**

We found that N-terminal fusion of Rad53-SCD1 to four different NVH-tagged target proteins (Figure 1C) or Rad51-ΔN (Figure 1D) all resulted in higher protein production *in vivo*. LacZ is a tetrameric protein, glutathione S-transferase (GST) is dimeric, and non-dimerizing GST (GSTnd) and GFP are monomeric proteins. As reported recently (27), removal of the NTD from Rad51 reduced by ~97% the protein levels of corresponding Rad51-ΔN proteins relative to wild type (WT) (Figure 1D), leading to lower resistance to the DNA damage agent methyl methanesulfonate (MMS) (Figure 1E). Interestingly, the

autonomous PEE function of Rad51-NTD could be fully rescued in Rad51-ΔN by N-terminal fusion of Rad53-SCD1, Rad53-SCD1-5STA (all five S/T-Q motifs changed to AQs) or Sup35-PND, respectively. Compared to wild-type yeast cells, the three corresponding yeast mutants (*rad51-SCD1-rad51-ΔN*, *rad51-SCD1-5STA-rad51-ΔN* and *sup35-PND-rad51-ΔN*) (Table 2) not only produced similar steady-state levels of Rad51-ΔN fusion proteins (Figure 1D), but they also exhibited high MMS resistance (Figure 1E). During homology directed repair of DNA double strand breaks (DSBs), Rad51 polymerizes into helical filaments on DSB-associated single-stranded DNA (ssDNA) and then promotes homologous search and strand exchange of the ssDNA-protein filament with a second double-stranded DNA (dsDNA). We inferred that the catalytic activity of Rad51-ΔN in HDRR is likely similar to that of wild-type Rad51 because the weak MMS-resistant phenotype of *rad51-ΔN* is mainly due to very low steady-state levels of Rad51-ΔN (Figure 1D). Accordingly, we concluded that the quaternary structures of the target proteins (i.e., GFP, GSTnd, GST, LacZ and Rad51-ΔN) are irrelevant to the autonomous PEE activity.

**The autonomous PEE function is not likely controlled by plasmid copy number or its transcription**

We also show that the PEE function is unlikely to operate at the transcriptional level, as revealed by genomic and reverse-transcription quantitative polymerase chain reactions (i.e., g-qPCR and RT-qPCR, respectively) (Figure 2 and Table 3). We found that addition of WT and mutant Rad51-NTD to LacZ-NVH not only did not affect the average copy number of the corresponding *CEN-ARS* plasmids in exponentially growing *S. cerevisiae* cells (Figure 2A), but also even reduced the steady-state transcript levels of the

corresponding LacZ-NVH fusion protein genes (Figure 2B). Therefore, addition of Rad51-NTD to LacZ-NVH did not result in a significant increase in transcription.

**The protein quality control system moderately regulates autonomous PEE activities**

The protein quality control system is a mechanism by which cells monitor proteins to ensure that they are appropriately folded (56). In the current study, we compared the protein steady-state levels (Figure 3A) and β-galactosidase activities (Figure 3B-3D) of Rad51-NTD-LacZ-NVH and LacZ-NVH in WT, *hsp104*Δ, *new1*Δ, *doa1*Δ, *doa4*Δ, *san1*Δ and *oaz1*Δ yeast cell lines. The protein products encoded by each of the six genes deleted from the latter mutant lines are all functionally relevant to protein homeostasis or prion propagation. Hsp104 is a heat-shock protein with disaggregase activities that disrupts protein aggregation (57, 58). New1 is a translation factor that fine-tunes ribosome recycling and the efficiency of translation termination (53). Doa1 (also called Ufd3) is an ubiquitin- and Cdc48-binding protein with a role in ubiquitin homeostasis and/or protein degradation (59, 60). The *doa1*Δ mutant exhibits diminished formation of [*PSI*+] prion (61). Doa4 is a deubiquitinating enzyme required for recycling ubiquitin from proteasome-bound ubiquitinated intermediates (62). The *doa4*Δ mutant exhibits increased sensitivity to the protein synthesis inhibitor cycloheximide (63). San1 is an ubiquitin-protein ligase that targets highly aggregation-prone proteins (64, 65). Oaz1 (ornithine decarboxylase antizyme) stimulates ubiquitin-independent degradation of Spe1 ornithine decarboxylase by the proteasome (66). We found that the β-galactosidase activities of Rad51-NTD-LacZ-NVH in WT and all six of the gene-knockout strains we examined were 10-29-fold higher than those of LacZ-NVH (Figure 3B). Intriguingly, the β-galactosidase activities of LacZ-NVH in the six gene-knockout mutants are all lower (30-70%) than those in WT (Figure

3C). In contrast, the β-galactosidase activities of Rad51-NTD-LacZ-NVH in WT are either slightly higher or lower than those in the six null mutants (Figure 3D). These results indicate that addition of Rad51-NTD to LacZ-NVH can abrogate the protein homeostasis defects caused by loss of each of these six genes. For example, Rad51-NTD might compensate for the ribosome assembly and translation defects in *new1*Δ (53), as well as the cycloheximide-hypersensitive phenotype displayed by *doa4*Δ (63). Accordingly, the β-galactosidase activities of Rad51-NTD-LacZ-NVH in the *new1*Δ and *doa4*Δ lines are higher than those in the WT, respectively. In contrast, the β-galactosidase activities of LacZ-NVH in the *new1*Δ and *doa4*Δ lines are lower, respectively, than those of WT. Finally, although the *doa1*Δ mutant is defective in [*PSI*+] prion formation (61), the steady-state levels of Rad51-NTD-LacZ-NVH in the *doa1*Δ line are also slightly higher than those in WT.

**The relationship between PEE function, amino acid contents and structural flexibility**

Using three Q-rich motifs (i.e., Rad51-NTD, Rad53-SCD1 and Sup35-NPD) as examples, we observed a very strong positive relationship between STQ and STQN amino acid percentages and β-galactosidase activities (Figure 4 and Figure 5). These results are consistent with the notion that, due to high STQ or STQN content, SCDs or Q-rich motifs are intrinsically disordered regions (IDRs) in their native states, rather than adopting stable secondary and/or tertiary structures (31). A common feature of IDRs is their high content of S, T, Q, N, proline (P), glycine (G) and charged amino acids (28-30).

It is important to note that the threshold of STQ or STQN content varies in the three cases presented herein (Figure 4B). Thus, the percentage of STQ or STQN residues is not

likely the only factor contributing to protein expression levels. Since G, P and glutamate (E) are enriched by >10% in Rad51-NTD, Rad53-SCD1 and Sup35-NPD, these three amino acids may also contribute to the PEE activities and structural flexibility of these three Q-rich motifs.

Functionally, IDRs are key components of subcellular machineries and signaling pathways because they have the potential to associate with many partners due to their ability to adopt multiple metastable conformations. Many IDRs are regulated by alternative splicing and post-translational modifications, and some are involved in the formation of various membraneless organelles or proteinaceous condensates via intracellular liquid-liquid phase separation (67, 68). Since IDRs can endow proteins with structural and functional plasticity, we hypothesized that Q-rich motifs (e.g., SCD, polyQ and polyQ/N) represent useful toolkits for creating new diversity during protein evolution.

**Comparative GO enrichment analyses of SCDs and polyX motifs in different eukaryotes**

To verify our hypothesis, we designed two JavaScript software programs (ASFinder-SCD and ASFinder-polyX) for proteome-wide searches of amino acid sequences that contain ≥3 S/T-Q motifs within a stretch of ≤100 residues (32) and the polyX motifs of 20 different amino acids, respectively. In the latter case, diverse thresholds have been used in different studies or databases to define and detect polyX motifs. We applied the lowest threshold (≥4/7), i.e., a minimum number of four identical X amino acid residues in a localized region of seven amino acid residues (69-72) (Figures 6-9 and Dataset DS1-DS29). We searched and compared the near-complete proteomes of 27 different eukaryotes (Table 4), including the budding yeast *S. cerevisiae*, three pathogenic species of *Candida,* three filamentous

ascomycete fungi (*Neurospora crassa*, *Magnaporthe oryzae* and *Trichoderma reesei*), three basidiomycete fungi (*Cryptococcus neoformans*, *Ustilago maydis* and *Taiwanofungus camphoratus*), the slim mold *Dictyostelium discoideum*, the malaria-causing unicellular protozoan parasite *Plasmodium falciparum*, six unicellular ciliates (*Tetrahymena thermophila*, *Paramecium tetraurelia*, *Oxytricha trifallax*, *Stylonychia lemnae*, *Pseudocohnilembus persalinus* and *Stentor coeruleus*), the fly *Drosophila melanogaster,* the mosquito *Aedes aegypti,* the nematode *Caenorhabditis elegans,* the zebrafish *Danio rerio,* the mouse *Mus musculus*, *Homo sapiens*, the higher plant *Arabidopsis thaliana*, and the single-celled green alga *Chlamydomonas reinhardtii*. The Benchmarking Universal Single-Copy Ortholog (BUSCO) scores of the near-universal single-copy gene orthologs of all 27 proteomes are 92.4-100% (Table 4). Genome or protein matrix scores >95% for model organisms are generally deemed complete reference genomes or proteomes (73).

It was reported previously that SCDs are overrepresented in the yeast and human proteomes (32, 74), and that polyX prevalence differs among species (43, 75-77). Our results reveal that the percentages of SCD proteins in the near-complete proteomes of 21 non-ciliate species and 6 ciliates range from 8.0% in *P. falciparum*, 13.9% in *H. sapiens*, 16.8% in *S. cerevisiae*, 24.2% in *U. maydis*, to a maximum of 58.0% in *O. trifallax* (Figure 6 and Dataset DS2). Among the 6050 proteins in the most recently modified *S. cerevisiae* reference proteome (https://www.uniprot.org/proteomes/UP000002311), we identified 1016 SCD-hosting proteins (Dataset DS4), including all 436 SCD-harboring proteins previously revealed by ScanProsite (32). ScanProsite is a publicly available database of protein families, domains and motifs (78).

The percentages of polyQ proteins in the near-complete proteomes of the 21 non-ciliate species and 6 ciliates range from 0.6% (*S. coeruleus*), 2.1% (*P. falciparum*), 4.1% (*S. cerevisiae*), 6.5% (*H. sapiens*), 21.8% (*D. melanogaster*), 24.5% (*D. discoideum*) to a maximum of 47.2% in *Pseudocohnilembus persalinus*. Those of polyQ/N proteins range from 9.5% in *T. camphoratus*, 11.9% in *H. sapiens*, 28.2% in *S. cerevisiae*, 62.5% in *D discoideum*, 63.7% in *P. falciparum*, to a maximum of 79.4% in *P. persalinus*. Percentages of polyN proteins range from 0.9% in *T. camphoratus*, 1.0% in *H. sapiens*, 1.1% in *M. musculus*, 3.7% in *S. coeruleus*, 49.7% in *D. discoideum*, to a maximum of 54.0% in *P falciparum* (Dataset DS2). Compared to other species, *D. melanogaster* and *A. aegypti* possess higher percentages of polyH proteins; *P. falciparum* has higher percentages of poly-phenylalanine (polyF) and polyY proteins; and *C. reinhardtii* has higher percentages of polyA, polyP, polyG and polyH proteins.

Notably, average usage of a given amino acid (X) is not always positively correlated with the percentages of polyX proteins in different species. For instance, proteome-wide Q usages are 2.8% in *P. falciparum*, 3.7% in *S. coeruleus*, 5.4% in *D. melanogaster* and 10.9% in *P. persalinus.* The proteome-wide usage of F in *P. falciparum* (4.4%) is not significantly different from that of other species, ranging from 2.0% in *C. reinhardtii* to 5.2% in *T. thermophila*. Average usages of P (7.7%), A (17.1%) and G (11.5%) in *C. reinhardtii*, as well as the average usages of N (14.4%) and Y (5.7%) in *P. falciparum*, are higher than those in 6 ciliate and the 20 other non-ciliate species, respectively (Figure 6 and Dataset DS1). These results indicate that distinct evolutionary strategies have evolved to enable specific polyX motifs to accumulate in different species.

Next, we performed comparative GO enrichment analyses using information on the functions of genes provided by GO knowledgebase (http://geneontology.org). Our results provide two intriguing insights into evolutionary biology. First, SCDs (Table 2) and polyQ/N motifs (Table 6) exist in many proteins involved in various evolutionarily conserved processes (Dataset DS4-DS29). It was reported previously that many SCD proteins in yeast and human are neither targets of ATR or ATM nor are related to pathways of functions relevant to the DDR or DNA replication, including telophase, cytokinesis, protein transport, endocytosis and cytoskeleton (32). Second, SCDs (Table 5) and polyQ/N motifs (Table 6) are overrepresented in the specialized biological processes of different species. Of particular note is an overrepresentation of conserved SCD proteins (74) and polyQ/N proteins (this study) in pathways related to vertebrate neural development and neurodegeneration. For instance, human CTTNBP2 (1663 amino acid residues) is a neuron-specific F-actin associated SCD-protein that is involved in the formation and maintenance of dendritic spines and it is associated with autism spectrum disorders (79, 80). Human CTTNBP2 possesses ten S/T-Q motifs ($T^{466}Q$, $T^{493}Q$, $S^{1580}Q$, $S^{553}Q$, $S^{634}Q$, $S^{994}Q$, $S^{1392}Q$, $S^{1580}Q$, $T^{1621}Q$ and $S^{1624}Q$). Mouse CTTNBP2 has 630 amino acid residues, four S/T-Q motifs ($S^{419}Q$, $T^{463}Q$, $S^{550}Q$, $S^{624}Q$), and it shares high amino acid identity to the N-terminus (1-640 amino acid residues) of human CTTNBP2. Using IUPred2A (https://iupred2a.elte.hu/plot_new), a web-server for identifying disordered protein regions (81), we found that both human CTTNBP2 (220-1633 residues) and mouse CTTNBP2 (220-630 residues) are Q/N-rich IDRs with high percentages of S, T, Q, N, G, P, R and K. We reported recently that mouse CTTNBP2 forms self-assembled condensates through its C-terminal IDR and it facilitates co-condensation of an abundant excitatory postsynaptic

scaffold protein SHANK3 at dendritic spines in a $Zn^{2+}$-dependent manner (82). Our results also reveal that SCDs, as well as polyQ, polyQ/N, polyS and polyP tracts, are overrepresented in the nervous system of *D. melanogaster* (Tables 5 and 6). The numbers (and percentages) of SCD-, polyQ-, polyQ/N-, polyS- and polyP-containing proteins, respectively, in different nerve-related biological processes are: regulation of axonogenesis [101 (87%), 93 (80%), 98 (84%), 103 (89%) and 100 (86%)], neuron development [97 (80%), 99 (82%), 102 (84%), 103 (85%) and 88 (73%)], axon guidance [199 (53%), 209 (56%), 278 (74%), 280 (75%), and 186 (50%)], mushroom body development [140 (66%), 142 (67%), 167 (79%), 161 (76%) and 109 (51%)], dendritic morphogenesis [96 (53%), 95 (52%), 132 (73%), 131 (72%) and 61 (34%)], and peripheral nervous system development [94 (73%), 100 (78%), 111 (86%), 106 (82%) and 90 (70%)] (Table 5, Table 6 and Dataset DS22).

We also found SCDs, polyQ/N tracts and a few other structurally flexible polyX tracts that are overrepresented in specialized biological processes of various species. First, in *S. cerevisiae*, 92 proteins are involved in RNA-mediated transposition (GO ID: 32197), among which 82 and 88 are SCD and polyQ/N proteins, respectively (Table 5, Table 6 and Dataset DS4). The 82 SCD (also polyQ/N) proteins are all structural components of the virus-like particles (VLPs), which form the shell that encapsulates the Ty1 (Transposon-yeast 1) and Ty2 retrotransposon dimeric RNA genomes. These capsid proteins also possess nucleocapsid-like chaperone activity, promoting primer tRNA(i)-Met annealing to the multipartite primer-binding site, dimerization of Ty RNA, and initiation of reverse transcription (83). Another example is the developmental and sexual cycles of *D discoideum*, which are induced by starvation and/or other stimuli. During the

developmental cycle, these amoebae secrete small molecules (e.g., cyclic adenosine monophosphate or cAMP) to induce intercellular communication (chemotaxis), ultimately resulting in aggregation of hundreds to hundreds of thousands of amoebae into multicellular fruiting bodies (sorocarps). These sorocaps consist of ~80% asexual spores and ~20% dead stalk cells. During the sexual cycle, two amoebae of complementary mating types fuse to form a zygote. The zygote and the surrounding amoebae form a thick multilayered structure called the macrocyst, representing the site of meiosis and interhomolog DNA recombination, enabling the generation of hundreds of recombinant haploid amoebae (84, 85). We found that SCDs (Table 5) and polyQ/N tracts (Table 6) are overrepresented in many proteins involved in a diversity of biological processes during the *D. discoideum* development and sexual cycles, including chemotaxis to cAMP, amoebae aggregation, sorocarp morphogenesis, sorocarp development, and sexual reproduction (Dataset DS14). A third example is the human malaria-causing parasite *P. falciparum*. Infection in humans begins upon being bitten by an infected female *Anopheles* mosquito. After entering the human bloodstream, sporozoites from the mosquito's salivary gland quickly invade hepatocytes and undergo asexual multiplication, leading to tens of thousands of merozoites bursting from the liver cells. These merozoites then invade red blood cells (erythrocytes) and undergo an additional round of multiplication. The mature merozoites can change the surface properties of infected red blood cells, causing them to stick to blood vessels (a process called cytoadherence), resulting in obstructed microcirculation and multiple organ dysfunction. The success of *P. falciparum* in establishing persistent infections is mainly attributable to immune evasion through antigenic variation (86). Our GO enrichment results reveal that SCDs, as well as polyQ/N,

polyN, polyS, polyT, polyD, polyG and polyK tracts, are overrepresented in many of these specialized infection processes, including antigen variation, modulation of host erythrocyte aggregation, cytoadherence to the microvasculature, and cell-cell adhesion (Table 5, Table 6 and Dataset DS15). For example, the numbers of proteins involved in antigen variation are 38 (SCDs), 82 (polyQ/N), 37 (polyN), 51 (polyS), 49 (polyT), 42 (polP), 69 (polyG), 40 (polyD), 68 (polyE) and 144 (polyK), respectively. These examples are consistent with the notion that S, T, Q, N, P, G and charged amino acids are used more frequently in IDRs (28-30).

**Massive expansion of Q-rich motifs in ciliates having two reassigned stop codons ($UAA^Q$ and $UAG^Q$)**

The most striking finding in our study is that, due to their usage of the two noncanonical codons ($UAA^Q$ and $UAG^Q$), Q (but not S, T or N) is used more frequently in five free-moving unicellular ciliates (*T. thermophila*, *P. tetraurelia*, *O. trifallax*, *S. lemnae* and *P. persalinus*) than in the sessile ciliate *S. coeruleus* or in any of the 21 non-ciliate species we examined herein (Figures 6 and Dataset DS1). We refer hereafter to these five free-moving unicellular ciliates as 'group I' ciliates. The proteome-wide Q and N contents, as well as the percentages of SCD, polyQ, polyQ/N and polyN proteins, respectively, in the different ciliates are: *T. thermophila* (9.6%, 9.2%, 50.8%, 33.0%, 72.6%, 25.2%), *P. tetraurelia* (9.1%, 6.9%, 41.1%, 27.1%, 62.9% and 8.7%), *O. trifallax* (9.0%, 7.9%, 58.0%, 36.7%, 72.3%, 22.0%), *S. lemnae* (8.4%, 7.5%, 49.2%, 28,1%, 65.3%, 17.4%), *P. persalinus* (10.9%, 9.8%, 47.7%, 47.2%, 79.5%, 36.3%), and *S. coeruleus* (3.7%, 6.3%, 13.9%, 0.6%, 18.7% and 3.7%] (Figure 6-9, Dataset DS2 and DS3).

Our GO enrichment results also reveal that the percentages of SCD, polyQ and polyQ/N (but not polyN) proteins in many biological processes of the five group I ciliates are significantly higher than those in *S. coeruleus*. For example, the numbers (and percentages) of SCD, polyQ, polyQ/N and polyN proteins involved, respectively, in DNA repair (GO ID: 06281) are: *T. thermophila* [28 (61%), 18 (39%), 38 (83%) and 13 (28%)] (DS16); *P. tetraurelia* [27 (44%), 21 (34%), 49 (79%) and 3 (5%)] (DS17); *O. trifallax* [50 (61%), 35 (43%), 62 (76%) and 22 (27%)] (DS18); *S. lemnae* [32 (53%), 21 (35%), 44 (73%) and 19 (17%)] (DS19); *P. persalinus* [19 (53%), 19 (53%), 31 (86%) and 10 (28%)] (DS20); and *S. coeruleus* [7 (17%), 1 (2%), 10 (24%) and 2 (5%)] (DS22). The respective numbers (and percentages) of SCD, polyQ, polyQ/N and polyN proteins involved in microtubule-based movement (GO ID: 07018) are: *T. thermophila* [93 (85%), 69 (63%), 107 (97%) and 63 (57%)] (DS16); *P. tetraurelia* [155 (69%), 109 (48%), 195 (86%) and 47 (21%)] (DS17); *O. trifallax* [99 (78%), 85 (66%), 121 (95%) and 47 (37%)] (DS18); *S. lemnae* [84 (80%), 66 (63%), 100 (95%) and 50 (48%)] (DS19); *P. persalinus* [65 (72%), 63 (70%), 85 (94%) and 41 (46%)] (DS20); and *S. coeruleus* [78 (34%), 5 (2%), 84 (37%) and 13 (6%)] (DS21).

Compared to 20 non-ciliate species, all five group I ciliates and the sessile *S. coeruleus* possess proteins involved in peptidyl-glutamic acid modification (GO ID:18200), with respective numbers (and percentages) of SCD, polyQ, polyQ/N and polyN proteins being: *T. thermophila* [38 (78%), 31 (63%), 47 (96%) and 23(47%)] (Dataset DS16); *P. tetraurelia* [42 (60%), 23 (33%), 58 (83%) and 5 (7%)] (Dataset DS17); *O. trifallax* [81 (91%), 63 (71%), 84 (94%) and 50 (56%)] (Dataset DS18); *S. lemnae* [63 (86%), 54 (74%), 68 (93%) and 34 (47%)] (Dataset DS19); *P. persalinus* [38 (58%), 37

(57%), 55 (85%) and 28 (43%)] (Dataset DS20); and *S. coeruleus* [30 (23%), 0 (0%), 50 (38%) and 23 (17%)] (Dataset DS21).

*T. thermophila* has 58 proteins involved in the xylan catabolic process (GO ID: 45493) (Table 7), of which 56 (97%) are SCDs, 55 (95%) are polyQ proteins, 58 (100%) are polyQ/N proteins, and 49 (84%) are polyN proteins (Dataset DS16). Using the NCBI BLASTP search tool with an expect value (E-value) ≤ 10e-5 to search for homologs of these 58 proteins among all other 16 ciliates analyzed in this study, only 144 proteins with amino acid identity <60% and a raw alignment score of <150 were found (Dataset DS30).

In *T. thermophila*, there are 124 annotated proteins involved in various events during meiosis, including meiosis initiation and regulation, homologous pairing, formation of DNA DSBs, formation of CO and chiasma (a structure that forms between a pair of homologous chromosomes by CO recombination and physically links the homologous chromosomes during meiosis), as well as meiotic nuclear divisions (DS31) (see review in (8)). Among the 124 *T. thermophila* meiotic proteins, we identified 85 SCD proteins, 54 polyQ proteins, 106 polyQ/N proteins, 32 polyN proteins, 32 polyS proteins and 48 polyK proteins, respectively. Notably, there are 48 and 59 meiotic proteins that contain ≥4 SCDs and/or ≥4 polyQ/N tracts, respectively. For instance, DPL2, CYC2, CYC17 and ATR1 each contains 15, 6, 8 and 5 SCDs, 2, 0, 1 and 2 polyQ tracts, as well as 11, 4, 4 and 5 polyQ/N tracts, respectively. Pars11, a chromatin-associated protein required for inducing and limiting Spo11-induced DSBs, has 14 SCDs, 2 polyQ tracts and 6 polyQ/N tracts. Spo11-induced DSBs promote ATR1-dependent Pars11 phosphorylation and its removal from chromatin (13). Many *T. thermophila* meiotic DSB repair proteins also harbor several SCDs, polyQ tracts and/or polyQ/N tracts, including MSH4, MSH5, SGS1, FANCM,

REC8, and ZPH3, among others. Several *T. thermophila* proteins involved in editing and rearrangement of the MIC genome also harbor multiple Q-rich motifs. PDD1 (programmed DNA degradation 1), a conjugation-specific HP1-like protein, has 13 SCDs, 3 polyQ tracts and 6 polyQ/N tracts. Mutations in the chromodomain or the chromoshadow domain of PDD1 were found previously to elicit PDD1 mislocalization, prevented histone H3 dimethylation on K9, abolished removal of internal eliminated sequences (IES), and/or resulted in the production of inviable progeny (87). DCL1 (Dicer-like 1) has 5 SCDs, 1 polyQ tract and 8 polyQ/N tracts. DCL1 is required for processing the MIC transcripts to siRNA-like scan (scn) RNAs, as well as for methylation of histone H3 at Lys 9. This latter modification occurs specifically on sequences (IESs) to be eliminated (88). GIW1 (gentlemen-in-waiting 1) physically directs a mature Argonaute-siRNA complex to the MIC nucleus, thus promoting programmed IES elimination (89). GIW1 has 9 SCDs and 2 polyQ/N tracts (Dataset DS31). Using IUPred2A (https://iupred2a.elte.hu/plot_new), we found that the Q-rich motifs in all these meiotic proteins are intrinsically disordered. Accordingly, it is of interest to speculate that, like the C-terminal IDR of mammalian CTTNBP2, the Q-rich motifs in *T. thermophila* meiotic proteins might form tunable proteinaceous condensates to regulate assembly and disassembly of the "crescents" in ciliate MICs.

Finally, we performed comparative GO analyses for the SCD and polyX proteins encoded by the transcriptomes of 11 different ciliate species. These transcripts were originally generated as part of the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (90), which were then reassembled and reannotated by Brown and colleagues (91). All transcripts are publicly available at Zendo

(https://zenodo.org/record/1212585#.Y79zoi2l3PA) (91). We applied TransDecoder (Find Coding Regions Within Transcripts; https://github.com/TransDecoder/TransDecoder/wiki) to identify candidate coding regions within the transcript sequences. Five of those eleven ciliates have reassigned $UAA^{Q}$ and $UAG^{Q}$ codons (hereafter termed "group II ciliates"), i.e., *Aristerostoma* spp*., Favella ehrenbergii, Pseudokeronopsis* spp., *Strombidium inclinatum,* and *Uronema* spp.. Like sessile *S. coeruleus,* group III ciliates (*Climacostomum virens*, *Litonotus pictus* and *Protocruzia adherens*) possess the standard genetic codes. Group IV ciliates encompass *Favella ehrenbergii*, *Blepharisma japonicum* and *Euplotes focardii*, each of which utilizes the reassigned codons $UAA^{Y}$, $UAG^{Y}$, $UGA^{W}$ and $UGA^{C}$, respectively (Table 1). The proteins encoded by these MMESTP transcripts are unlikely to represent the entire protein complement of all 11 ciliate species because 1) many MMETSP transcripts are not intact (i.e., broken mRNAs) and thus encode incomplete protein sequences, and 2) except for *C. virens* (94.7%), the BUSCO protein scores of these MMESTP transcripts only range from 52.6% to 89.9% (Table 1). Nevertheless, our results indicate that Q is used more frequently in group I and group II ciliates than in group III and group IV ciliates or in a further 20 non-ciliate species (Figure 3 and DS1). Accordingly, proportions of SCD, polyQ and polyQ/N proteins in all group I and group II ciliates are higher than they are in the three group III ciliates (except *L. pictus*) and the three group IV ciliates, respectively (Figure 4-6). Since N is used more frequently in *L. pictus* than in group II ciliates, as well as in the other group III and group IV ciliates, it has higher percentages of polyN and polyQ/N proteins (Figures 6-9 and Dataset DS1).

Our data also indicates that Y, W or C are not used more frequently in the three group IV ciliates than in the other 14 ciliate or 20 non-ciliate species (Figure 7 and Dataset DS1). Reassignments of stop codons to Y, W or C also do not result in higher percentages of polyY, polyW or polyC proteins in the three group IV ciliates, respectively (Figure 9).

## Discussion

We present two unexpected results in this report. First, the Q-rich motifs of several yeast proteins (Rad51-NTD, Rad53-SCD1, Hop1-SCD, Sml1-NTD, Sup35-PND, Ure2-UPD and New1-NPD) all exhibit autonomous PEE activities. These structurally flexible Q-rich motifs have useful potential for applications in both basic research (e.g., synthetic biology) and biotechnology. It is of interest to investigate further how these Q-rich motifs exert this PEE function in yeast and whether Q-rich motifs in other eukaryotes also possess similar PEE activities. Second, reassignment of stop codons to Q in the group I and group II ciliates significantly increase proteome-wide Q usage, leading to massive expansion of structurally flexible or even intrinsically disordered Q-rich motifs. In contrast, reassignments of stop codons to Y, W or C do not result in higher usages of these three amino acid residues nor higher percentages of W-, Y- or C-rich proteins in the three group IV ciliates, respectively. These results are consistent with the notion that, unlike for Q, the Y, W and C residues are not common in IDRs (28-30).

Due to their structural flexibility, Q-rich motifs can endow proteins with structural and functional plasticity. Based on our findings from this study, we suggest that Q-rich motifs are useful toolkits for generating novel diversity during protein evolution, including by enabling greater protein expression, protein-protein interactions, posttranslational modifications, increased solubility, and tunable stability, among other important traits. This

speculation may explain three intriguing phenomena. First, due to higher Q usage, many proteins involved in evolutionarily conserved biological processes in group I and group II ciliates display more diverse amino acid sequences than the respective proteins in other ciliate or non-ciliate species. Accordingly, it is sometimes difficult to identify authentic protein homologs among different ciliates, particularly for group I and group II ciliates. We highlight the example of the 58 proteins involved in xylan catabolysis in *T. thermophila* (Dataset DS16). Second, our GO enrichment results reveal that Q-rich motifs or a few other structurally flexible polyX tracts prevail in proteins involved in specialized biological processes, including *S. cerevisiae* RNA-mediated transposition, *C. albicans* filamentous growth, ciliate peptidyl-glutamic acid modification, *T. thermophila* meiosis and xylan catabolism, *D. discoideum* sorocarp (fruiting body) development and sexual reproduction, *P. falciparum* blood cell infection and antigen variation, and development of the *D. melanogaster* nervous system. In theory, structurally flexible Q-rich motifs might form various membraneless organelles or proteinaceous condensates via intracellular liquid-liquid phase separation in tunable manners, including protein posttranslational modification, protein-protein interaction, protein-ligand binding, *et al*., (67, 68). A typical example is that the C-termini of mouse CTTNBP2 facilitates co-condensation of CTTNBP2 with SHANK3 at the postsynaptic density in a $Zn^{2+}$-dependent manner (82). Third, the Borg genomes of anaerobic methane-oxidizing archaea are huge, linear extrachromosomal elements. A striking feature of Borgs is pervasive tandem direct repeat (TR) regions. While we were preparing this manuscript, it was reported that TRs in ORFs are under very strong selective pressure, leading to perfect amino acid TRs (aaTRs) that are commonly IDRs. aaTRs particularly often contain disorder-promoting amino acids,

including Q, N, P, T, E, K, V, D and S, respectively (92). Accordingly, distinct evolutionary strategies have evolved in different species to evolve protein regions that are structurally flexible or even intrinsically disordered. Further investigations are needed to determine if liquid-liquid phase separation prevails in ciliates with reassigned $TAA^{Q}$ and $TAG^{Q}$ codons and/or in the specialized biological processes in various species we have described herein.

## Conclusions

One of the most interesting questions in genome diversity is why many ciliates reassign their nuclear stop codons into amino acids, e.g., glutamine (Q), tyrosine (Y), tryptophan (W) or cysteine (C). However, the impacts of such genome-wide alternations are not well understood. We show that glutamine (Q) is used more frequently in all 10 ciliate species possessing reassigned $TAA^{Q}$ and $TAG^{Q}$ codons than in other ciliates and non-ciliate species. The consequence of this preponderance of Q is massive expansion of proteins harboring three structurally flexible or even intrinsically disordered Q-rich motifs. Since Q-rich motifs can endow proteins with structural and functional plasticity, we also show that Q-rich motifs prevail in proteins involved in several species-specific or even phylum-specific biological processes. Our results indicate that Q-rich motifs are useful toolkits for evolutionary novelty.

## Methods

All plasmids, yeast strains and PCR primers used in this study are listed in the Tables 1-3, respectively. Guinea pig antisera against Rad51, and rabbit antisera against phosphorylated Rad51-$S^{12}$Q, phosphorylated Rad51-$S^{30}$Q, and phosphorylated Hop1-$T^{318}$Q were

described previously (27, 40). The mouse anti-V5 antibody was purchased from BioRad (CA, USA). The rabbit anti-Hsp104 antiserum was kindly provided by Chung Wang (Academia Sinica, Taiwan). Rabbit antisera against phosphorylated Sup35-$S^{17}$Q were raised using the synthetic phosphopeptide $N^{12}$YQQYS$^{(P)}$QNGNQQQGNNR$^{28}$ as an antigen, where $S^{(P)}$ is phosphorylated serine. Phosphopeptide synthesis and animal immunization were conducted by LTK BioLaboratories, Taiwan. Western blotting analyses were performed as described previously (27, 40). Quantitative β-galactosidase activity assays were carried out as previously described (27, 93). The ASFinder-SCD and ASFinder-polyX software are publicly available at Github (https://github.com/tfwangasimb/AS-SCD-finder). The sources of proteome and transcript datasets are described in Table 4. The proteome-wide amino acid contents, the numbers of protein with different types of Q-rich motifs and polyN, as well as the SCD and polyX proteins in all 37 different eukaryotes are listed in Datasets DS1-DS3, respectively. The Gene Ontology (GO) enrichment analyses were performed using publicly available data in the GO Resource (http://geneontology.org). The GO identities (ID) of different biological processes and the names of all SCD and polyX proteins in the 24 near-completed eukaryotic proteomes are listed in Dataset DS4-DS29, respectively.

## Supporting Datasets

Dataset DS1: The average usage of 20 amino acids in 17 ciliate and 20 non-ciliate species.

Dataset DS2: The number proteins containing different types of polyQ, polyQ/N and polyN tracts in 17 ciliate and 20 non-ciliate species.

Dataset DS3: The numbers and percentages of SCD and polyX proteins in 17 ciliate and 20 non-ciliate species.

Dataset DS4-DS29: GO enrichment analyses revealing the SCD and polyX proteins involved in different biological processes in 6 ciliate and 20 non-ciliate species.

Dataset DS30: The results of BLASTP searches using the 58 *Tetrahymena thermophila* proteins involved in xylan catabolysis.

Dataset DS31: The list of 124 *Tetrahymena thermophila* proteins involved in meiosis (kindly provided by Josef Loidl). The numbers of SCD and polyX tract in each protein are indicated.

Dataset DS32: The raw qPCR data of cDNA and gDNA in Figure 2.

## Declarations

### Ethics approval and consent to participate

All experiments were approved by the Ethics Committee of Academia Sinica, Taiwan.

### Consent for publication

All authors have read and approved of its submission to this journal.

**Availability data and materials**

The data analyzed in this study were listed in Table 4. Other data generated in this study were available within the supporting information.

**Competing interests**

The authors declare no conflicts of interests.

**Funding**

Institute of Molecular Biology, Academia Sinica, Taiwan, Republic of China, and National Science and Technology Council, Taiwan, Republic of China [NSTC 110-2811-B-001-542 to TFW], Taiwan, Republic of China.

**Authors' contributions**

HCL, CNC, TTW, YPH, HTH, JLC and TFW performed the experiments and analyzed the data. TFW and YPH conceived and designed the experiments. TFW wrote the paper. All of the authors read and approved the manuscript.

**Acknowledgements**

We thank John O′Brien for English editing, G. Titus Brown (0000-0001-6001-2677) for his help in accessing the reassembled transcriptomic dataset in Zendo, Yu-Tang Huang (IMB Computer Room) for maintaining the computer workstation, Josef Loid (Max Perutz

Labs, University of Vienna, Austria) for providing the list of 124 proteins involved in *T. thermophila* meiosis, Meng-Chao Yao (IMB, Academia Sinica, Taiwan) for his suggestion to include the MMESTP transcripts of 11 different ciliates in this study.

**Authors' information**

Hou-Cheng Liu: hc666@gate.sinica.edu.tw

Chi-Ning Chuang: chining@gate.sinica.edu.tw

Tai-Ting Woo: asanwoo@gmail.com

Hisao-Tang Hu: annebest999@gmail.com

Chiung-Ya Chen: chiungya@gate.sinica.edu.tw

Ju-Lan Chao: julan@gate.sinica.edu.tw

Yi-Ping Hsueh: yph@gate.sinica.edu.tw

Ting-Fang Wang: tfwang@gate.sinica.edu.tw

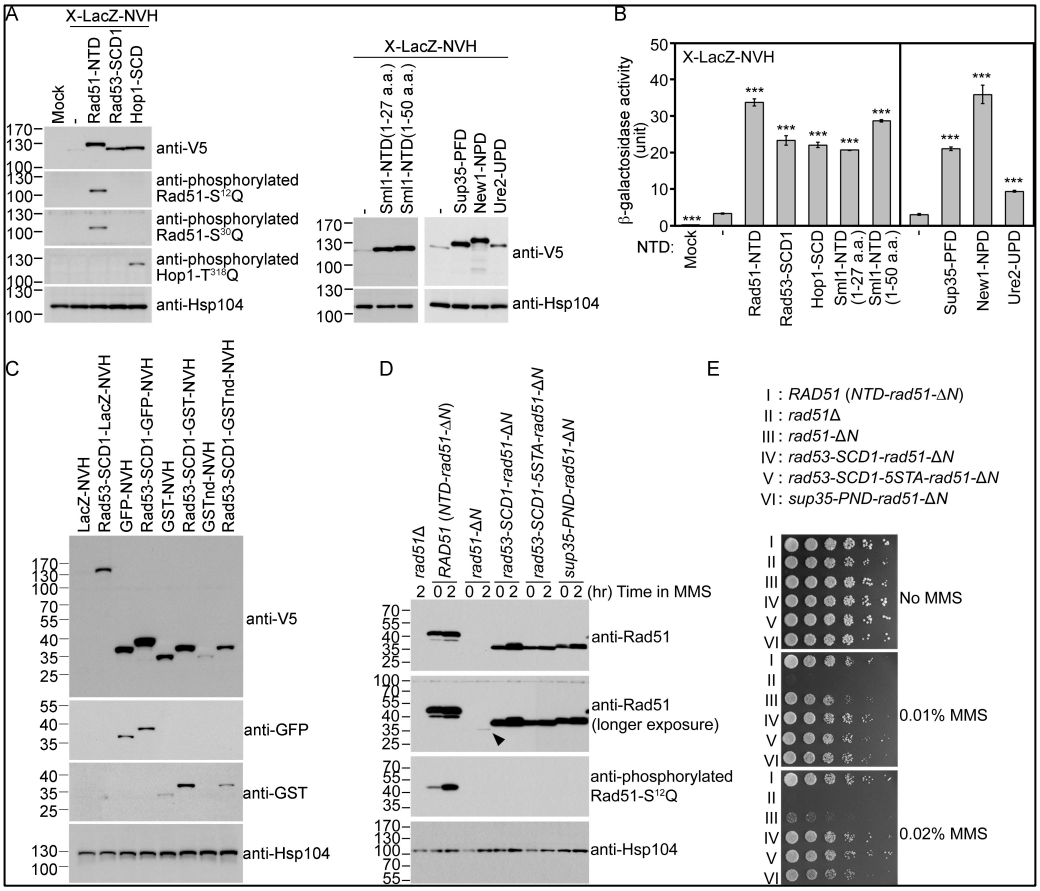



**Figure 1. The Q-rich domains of seven different yeast proteins possess autonomous expression-enhancing (PEE) activities.** (A-B) N-terminal fusion of Rad51-NTD/SCD, Rad53-SCD1, Hop1-SCD, Sml1-NTD, Sup35-PND, Ure2-UPD and New1-NPD promotes high-level expression of LacZ-NVH, respectively. The NVH tag contains an SV40 nuclear localization signal (NLS) peptide preceding a V5 epitope tag and a hexahistidine ($His_6$) affinity tag (27). Western blots for visualization of LacZ-NVH fusion proteins (A) and quantitative β-galactosidase assays (B) were carried out as described previously (27). Error bars indicate standard deviation between experiments ($n \geq 3$). Asterisks indicate significant differences relative to wild type (WT) in A or lacking an NTD in B, with $P$ values calculated using a two-tailed $t$-test (***, $P$ value <0.001; **, $P$ value <0.01). (C-D) The PEE activities of S/T/Q/N-rich domains are independent of the quaternary structures of target proteins. (C) Rad53-SCD1 can be used as an N-terminal fusion tag to enhance production of four different target proteins: LacZ-NVH, GST-NVH, GSTnd-NVH and GFP-NVH. (D) Visualization of native Rad51 (NTD-Rad51-ΔN), Rad51-ΔN, and the Rad51-ΔN fusion proteins by immunoblotting. Hsp104 was used as a loading control. Size in kilodaltons of standard protein markers is labeled to the left of the blots. The black arrowhead indicates the protein band of Rad51-ΔN. (E) MMS sensitivity. Spot assay showing five-fold serial dilutions of indicated strains grown on YPD plates with or without MMS at the indicated concentrations (w/v).

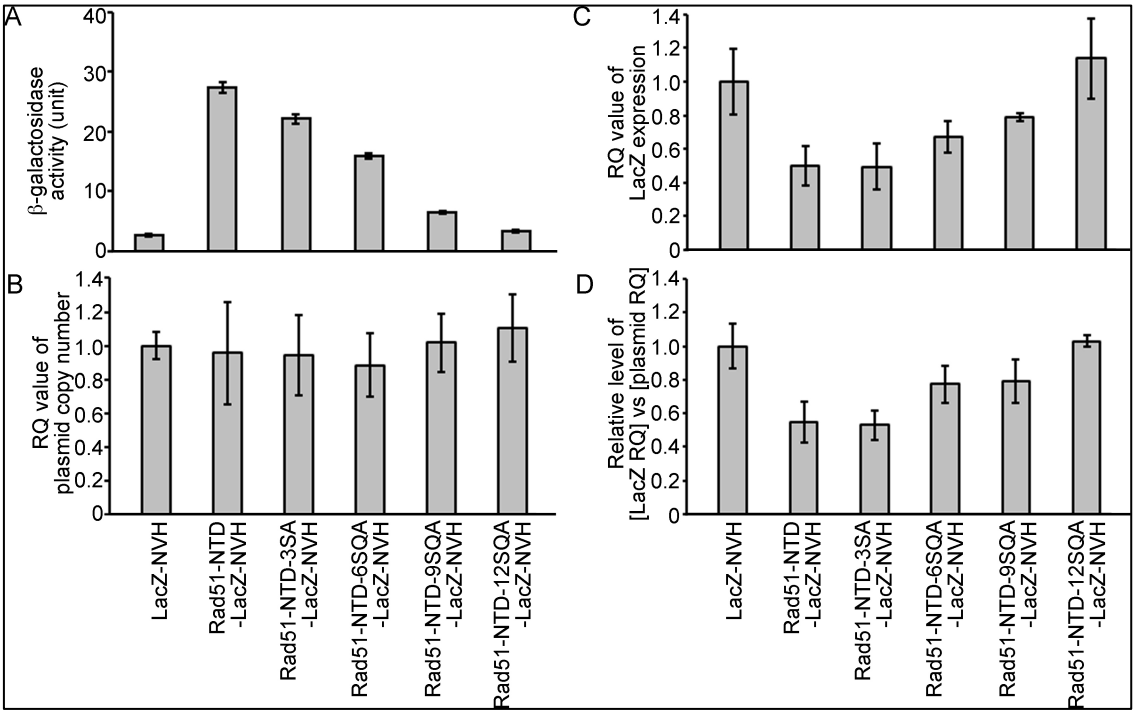



**Figure 2. The autonomous protein-expression-enhancing function of Rad51-NTD is unlikely to be controlled during transcription or simply arise from plasmid copy number differences.** The effects of WT and mutant Rad51-NTD on β-galactosidase activities (A), plasmid DNA copy numbers (B), relative steady-state levels of LacZ-NVH mRNA normalized to *ACT1* (actin) mRNA (C), and relative ratios of LacZ-NVH mRNA *versus* plasmid DNA copy number (D). The wild-type yeast cells were transformed with indicated CEN-ARS plasmids, respectively, to express WT and mutant Rad51-NTD-LacZ-NVH fusion proteins or LacZ-NVH alone under the control of the native RAD51 gene promoter ($P_{RAD51}$). The relative quantification (RQ = $2^{-\Delta\Delta C}$T) values were determined to reveal the plasmid DNA copy number and steady-state levels of LacZ-NVH mRNA by g-qPCR and RT-qPCR, respectively. LacZ and *ACT1* were selected as target and reference protein-encoding genes, respectively, in both g-qPCR and RT-qPCR. The data shown represent mean ± SD from three independent biological data-points.

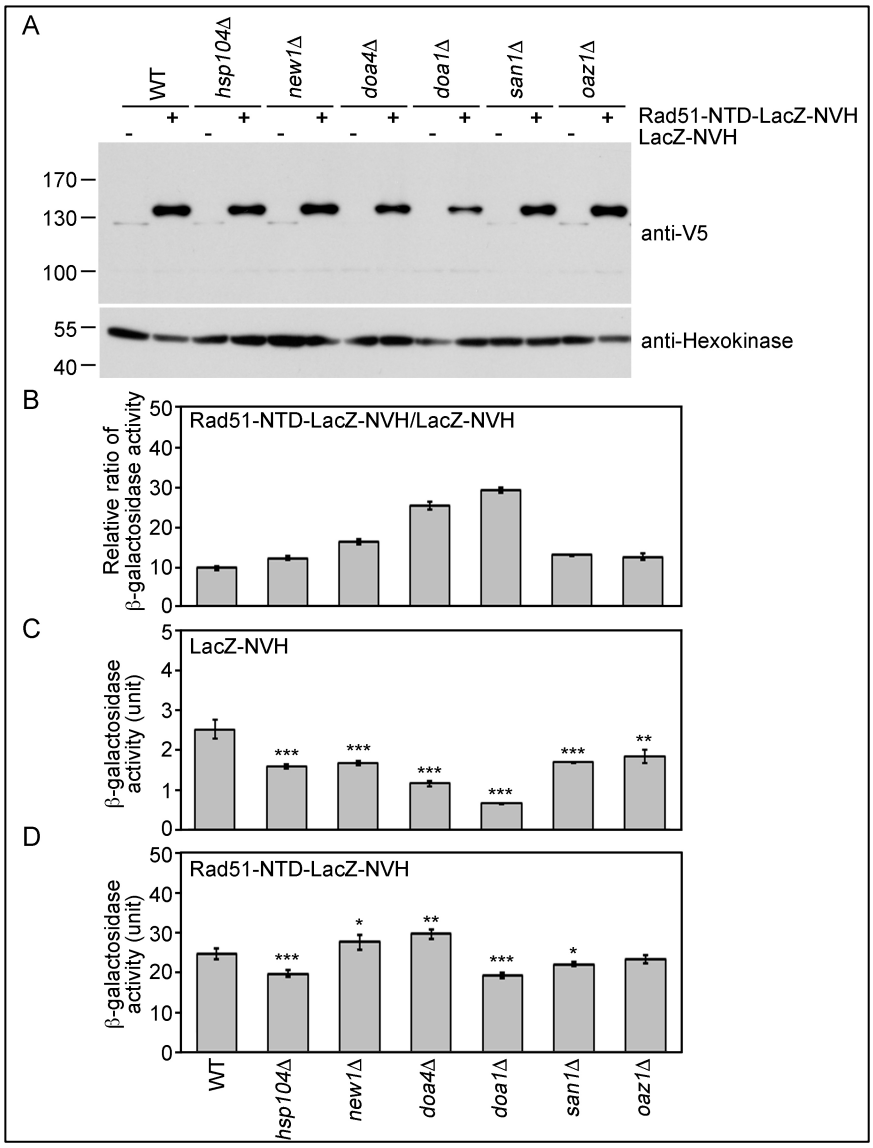
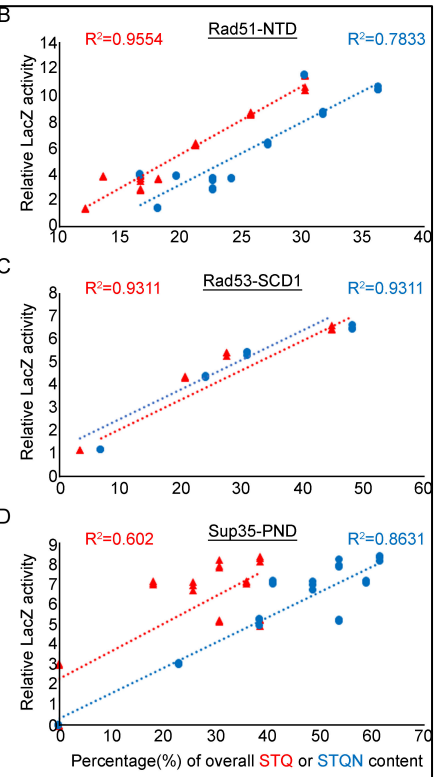



**Figure 3. The expression-promoting function of Rad51-NTD is controlled during protein translation and does not affect ubiquitin-mediated protein degradation.** (A) The steady-state protein levels of Rad51-NTD-LacZ-NVH and LacZ-NVH in WT and six protein homeostasis gene knockout mutants. (B-D) The impact of six protein homeostasis genes on the β-galactosidase activity ratios of Rad51-NTD-LacZ-NVH to LacZ-NVH in WT and the six gene knockout mutants (B). The β-galactosidase activities of LacZ-NVH (C) and Rad51-NTD-LacZ-NVH (D) in WT and the six gene knockout mutants are shown. Asterisks indicate significant differences, with *P* values calculated using a two-tailed *t*-test (***, *P* value <0.001; **, *P* value <0.01; *, *P* value <0.05).

A

| Name of SCDs or IDRs | Length | S | T | Q | N | STQ(%) | STQN(%) | Relative LacZ activity |
|---|---|---|---|---|---|---|---|---|
| Rad51-NTD | 66 | 9 | 2 | 9 | 4 | 30.30 | 36.36 | 10.65±0.27 |
| Rad53-SCD1 | 29 | 2 | 4 | 7 | 1 | 44.83 | 48.28 | 7.35±0.27 |
| Hop1-SCD | 67 | 6 | 6 | 8 | 9 | 29.85 | 43.28 | 7.00±0.19 |
| Sml1-NTD(1-27aa) | 27 | 3 | 3 | 7 | 2 | 48.15 | 55.56 | 6.52±0.12 |
| Sml1-NTD(1-50aa) | 50 | 5 | 4 | 8 | 3 | 34.00 | 40.00 | 9.05±0.18 |
| Sup35-PFD | 114 | 5 | 0 | 32 | 19 | 32.46 | 49.12 | 7.74±0.33 |
| New1-NPD | 156 | 19 | 8 | 14 | 28 | 26.28 | 44.23 | 12.19±0.65 |
| Ure2-UPD | 91 | 10 | 5 | 10 | 33 | 27.47 | 63.74 | 3.16±0.12 |
| Rad51-NTD | 66 | 9 | 2 | 9 | 4 | 30.30 | 36.36 | 10.44±0.20 |
| Rad51-NTD-3SA | 66 | 6 | 2 | 9 | 4 | 25.76 | 31.82 | 8.58±0.11 |
| Rad51-NTD-8STA | 66 | 3 | 0 | 9 | 4 | 18.18 | 24.24 | 3.64±0.05 |
| Rad51-NTD-11STA | 66 | 0 | 0 | 9 | 4 | 13.64 | 19.70 | 3.84±0.08 |
| Rad51-NTD-6SQA | 66 | 6 | 2 | 6 | 4 | 21.21 | 27.27 | 6.28±0.10 |
| Rad51-NTD-9SQA | 66 | 6 | 2 | 3 | 4 | 16.67 | 22.73 | 2.81±0.06 |
| Rad51-NTD-12SQA | 66 | 6 | 2 | 0 | 4 | 12.12 | 18.18 | 1.40±0.02 |
| Rad51-NTD-3QA | 66 | 9 | 2 | 6 | 4 | 25.76 | 31.82 | 8.63±0.12 |
| Rad51-NTD-9QA | 66 | 9 | 2 | 0 | 4 | 16.67 | 22.73 | 3.60±0.06 |
| Rad51-NTD-4NA | 66 | 9 | 2 | 9 | 0 | 30.30 | 30.30 | 11.48±0.17 |
| Rad51-NTD-13QNA | 66 | 9 | 2 | 0 | 0 | 16.67 | 16.67 | 3.92±0.07 |
| Rad53-SCD1 | 29 | 2 | 4 | 7 | 1 | 44.83 | 48.28 | 6.46±0.12 |
| Rad53-SCD1-5STA | 29 | 1 | 0 | 7 | 1 | 27.59 | 31.03 | 5.34±0.06 |
| Rad53-SCD1-7QA | 29 | 2 | 4 | 0 | 1 | 20.69 | 24.14 | 4.32±0.07 |
| Rad53-SCD1-12STQA | 29 | 1 | 0 | 0 | 1 | 3.45 | 6.90 | 1.14±0.01 |
| Sup35-PND | 39 | 3 | 0 | 12 | 9 | 38.46 | 61.54 | 8.30±0.08 |
| Sup35-PND-1SA | 39 | 2 | 0 | 12 | 9 | 35.90 | 58.97 | 7.13±0.06 |
| Sup35-PND-3SA | 39 | 0 | 0 | 12 | 9 | 30.77 | 53.85 | 5.18±0.03 |
| Sup35-PND-3QA | 39 | 3 | 0 | 9 | 9 | 30.77 | 53.85 | 8.02±0.12 |
| Sup35-PND-5QA | 39 | 3 | 0 | 7 | 9 | 25.64 | 48.72 | 6.93±0.11 |
| Sup35-PND-8QA | 39 | 3 | 0 | 4 | 9 | 17.95 | 41.03 | 7.09±0.05 |
| Sup35-PND-15SQA | 39 | 0 | 0 | 0 | 9 | 0 | 23.08 | 3.02±0.03 |
| Sup35-PND-9NA | 39 | 3 | 0 | 12 | 0 | 38.46 | 38.46 | 5.14±0.10 |
| Sup35-PND-24SQNA | 39 | 0 | 0 | 0 | 0 | 0 | 0 | -0.07±0.00 |



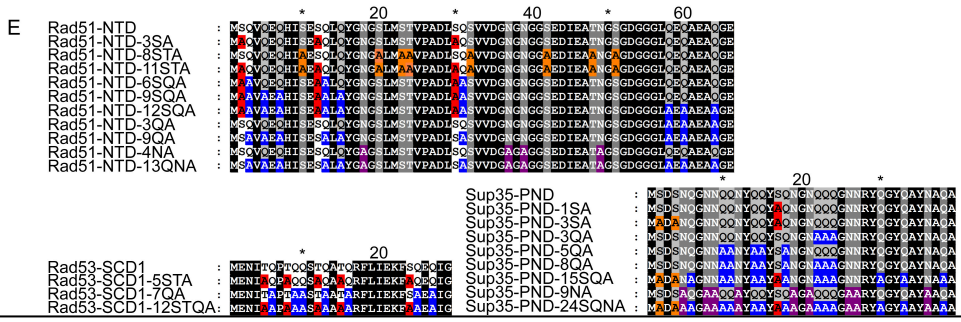




**Figure 4. Relative β-galactosidase (LacZ) activities are correlated with the percentage STQ or STQN amino acid content of three Q-rich motifs.** (A) List of N-terminal tags with their respective length, numbers of S/T/Q/N amino acids, overall STQ or STQN percentages, and relative β-galactosidase activities. (B-D) Linear regressions between relative β-galactosidase activities and overall STQ or STQN percentages for Rad51-NTD (B), Rad53-SCD1 (C) and Sup35-PND (D). The coefficients of determination ($R^2$) are indicated for each simple linear regression. (E) The amino acid sequences of wild-type and mutant Rad51-NTD, Rad51-SCD1 and Sup35-PND, respectively.

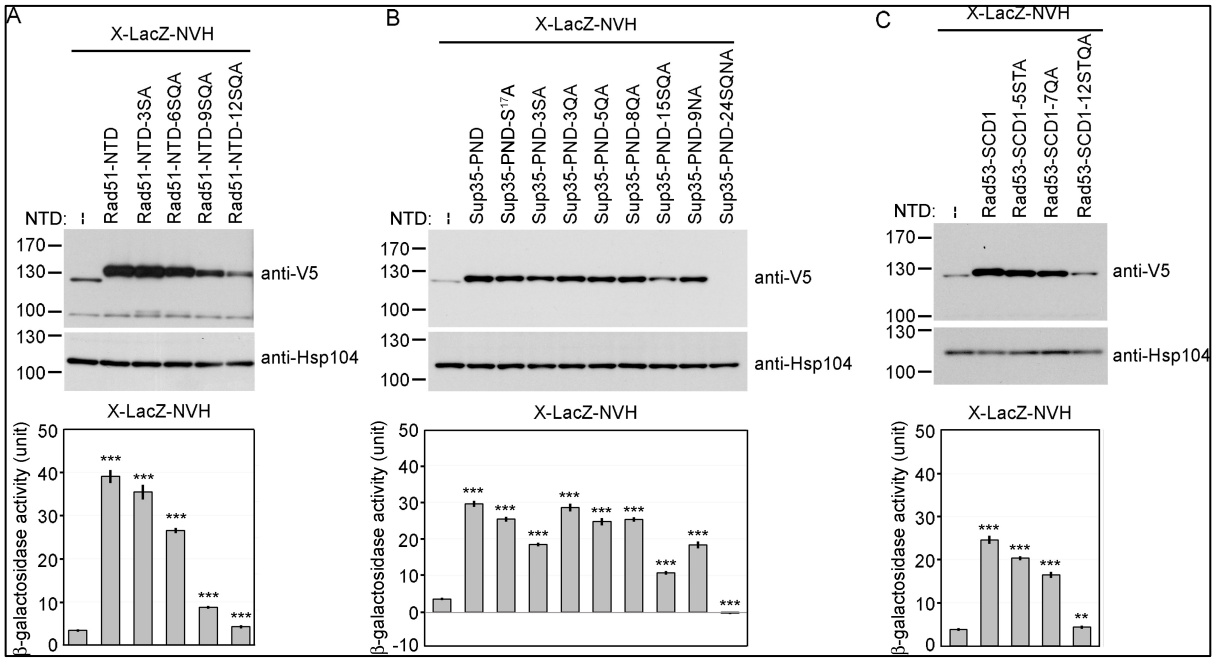



**Figure 5.** Alanine scanning mutagenesis of IDRs. The amino acid sequences of WT and mutant IDRs are listed in Table S1. Total protein lysates prepared from yeast cells expressing Rad51-NTD-LacZ-NVH (A), Sup35-PND-LacZ-NVH (B) or Rad53-SCD1-LacZ-NVH (C) were visualized by immunoblotting with anti-V5 antisera. Hsp104 was used as a loading control. Quantitative yeast β-galactosidase (LacZ) assays were carried out as described in Figure 1. Error bars indicate standard deviation between experiments ($n$ = 3). Asterisks indicate significant differences when compared to LacZ-NVH with $p$ values calculated with a two-tailed $t$-test (**, $P$ value <0.01 and ***, $P$ value <0.001).

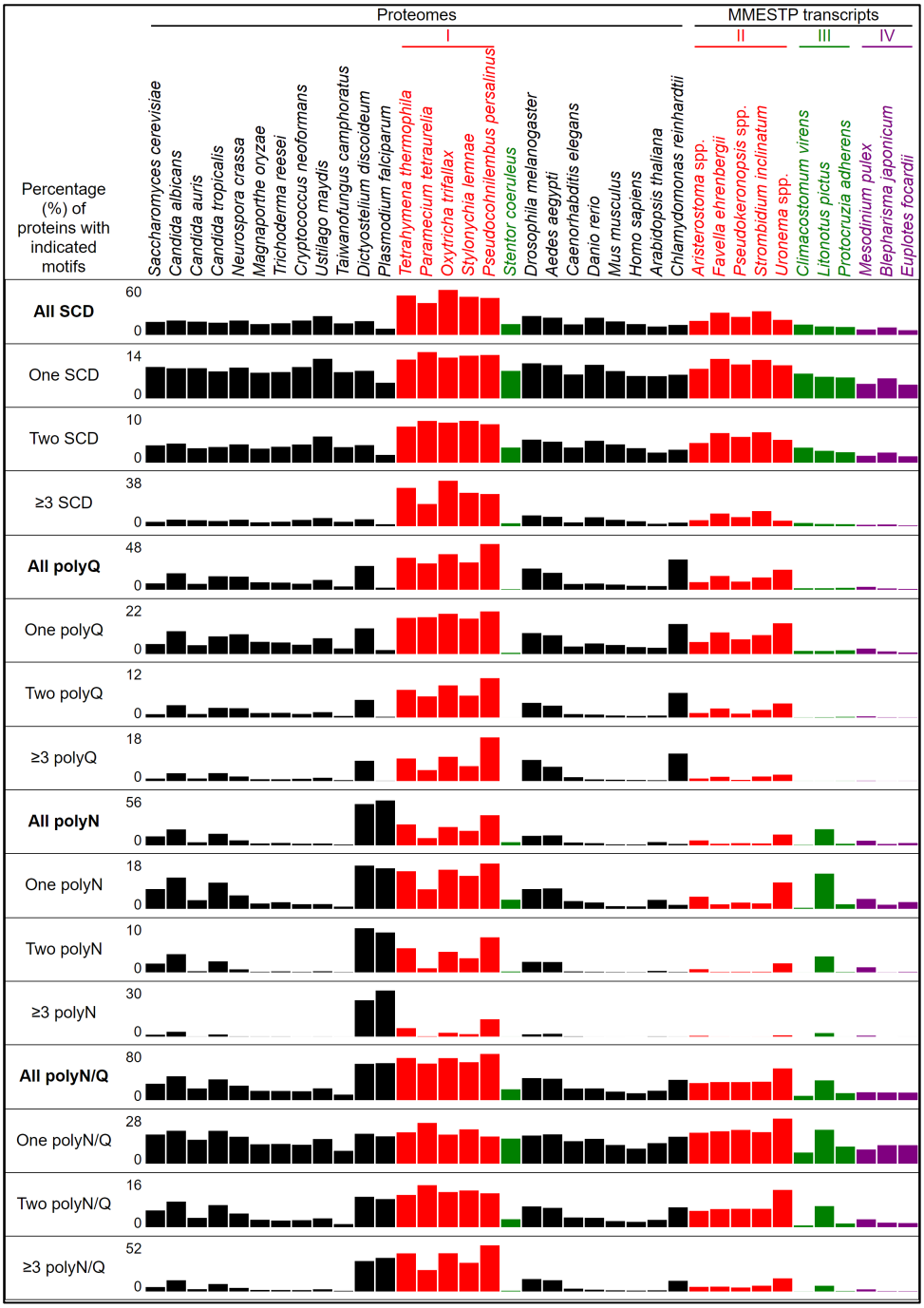



**Figure 6. Percentages of proteins with different numbers of SCDs, and polyQ, polyQ/N or polyN tracts in 37 different eukaryotes.**

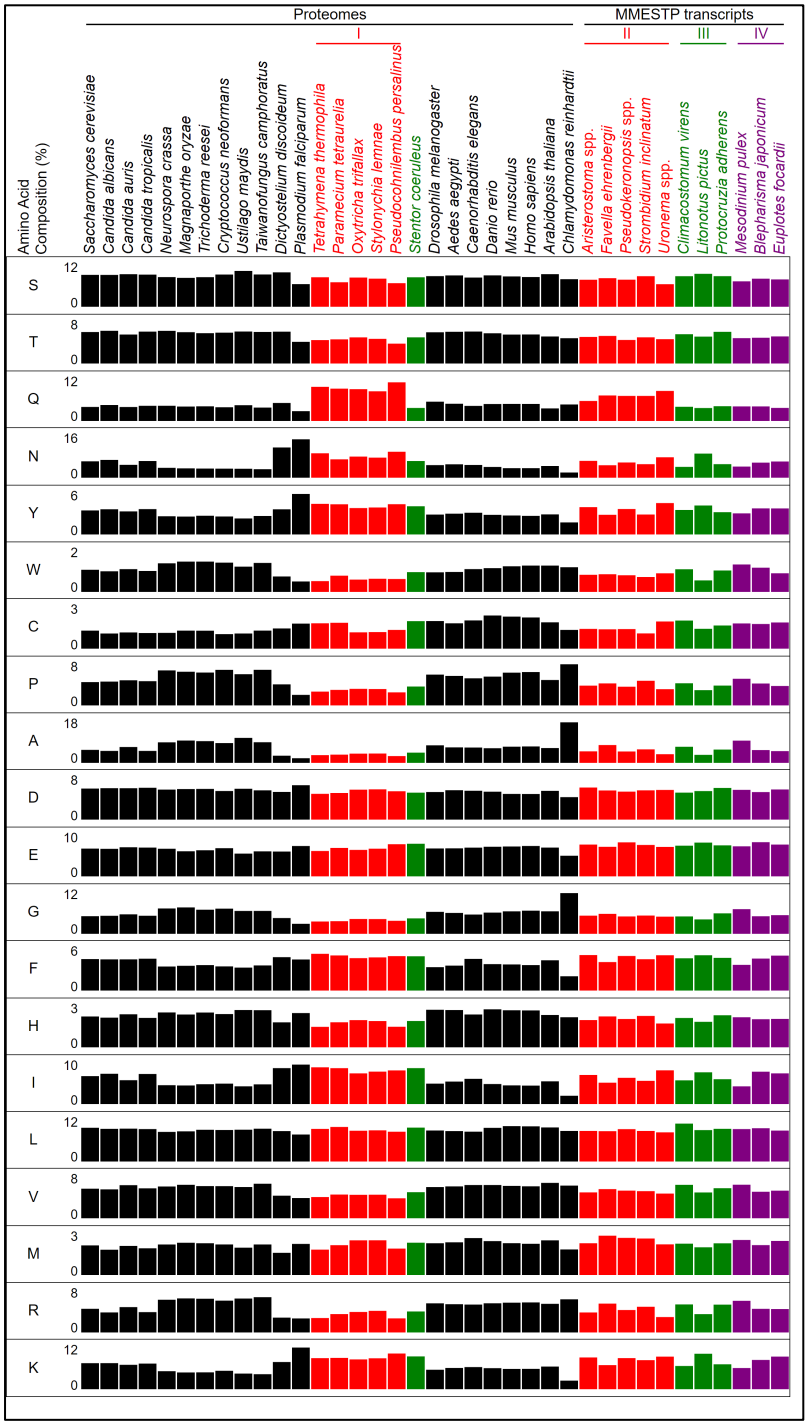



**Figure 7. Proteome-wide contents of 20 different amino acids in 37 different eukaryotes.**

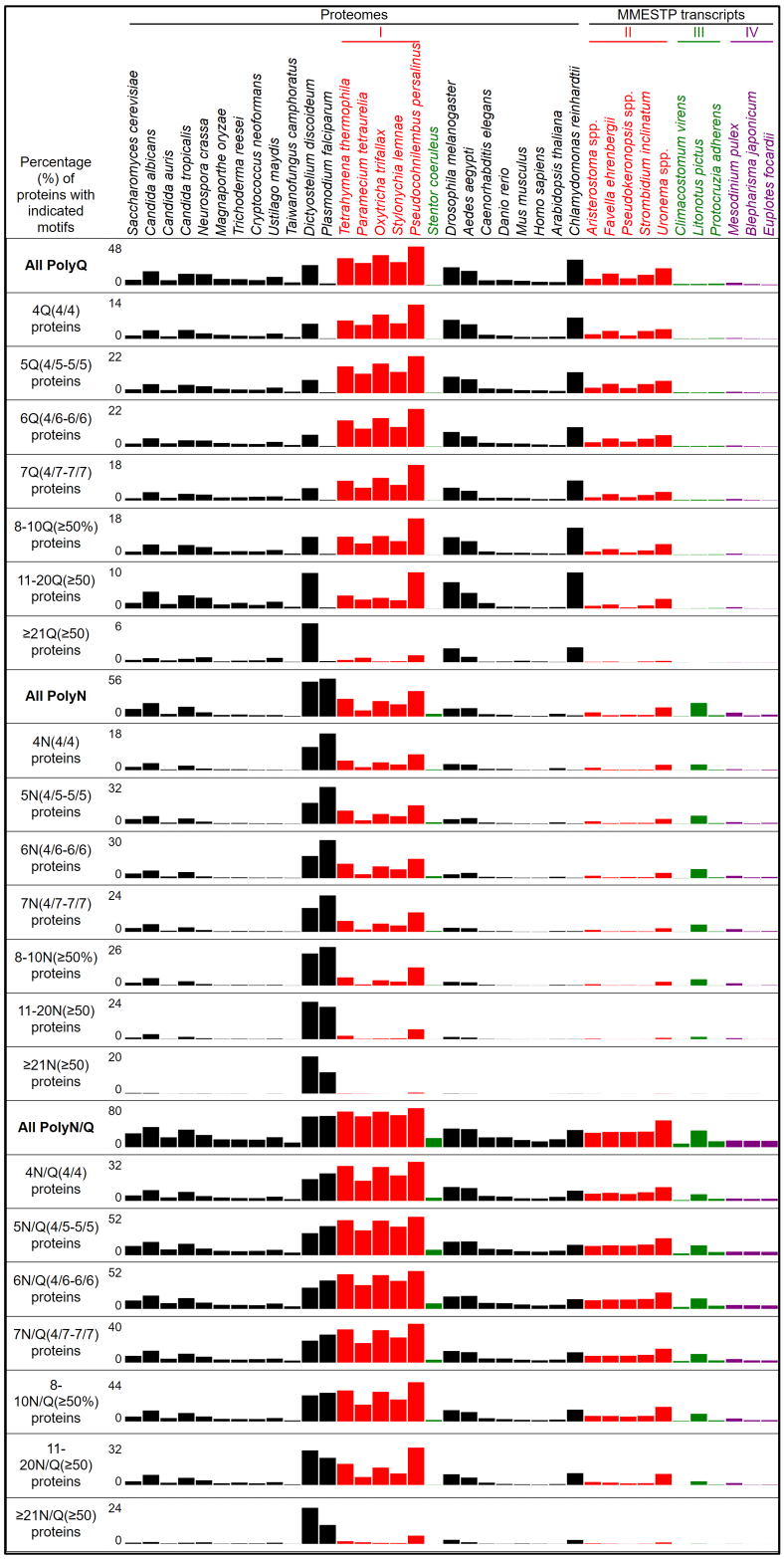



**Figure 8. Percentages of proteins with indicated polyQ and polyQ/N tracts in 37 different eukaryotes.**

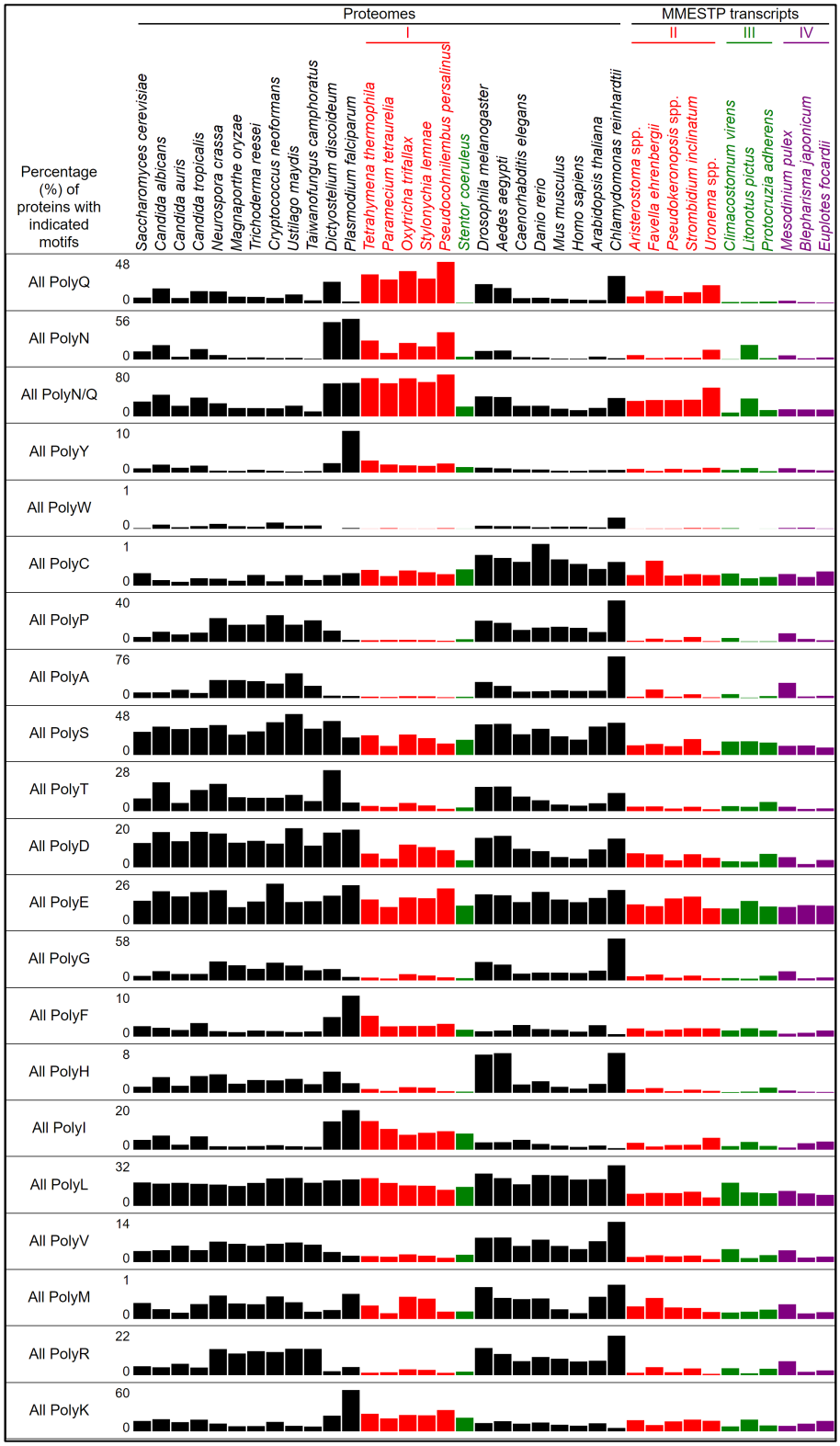



**Figure 9. Percentages of proteins with indicated polyX motifs in 37 different eukaryotes.**

Table 1. Plasmids used in this study

| Plasmid number | Description |
|---|---|
| pTFW2592 | pYC2/NT-C, The mock control. |
| pTFW9957 | *$P_{RAD51}$-LacZ-NVH* |
| pTFW9958 | *$P_{RAD51}$-RAD51-NTD(1-66 a.a.)- LacZ-NVH* |
| pTFW9962 | *$P_{RAD51}$-RAD53-SCD1(1-29 a.a.)-LacZ-NVH* |
| pTFW9963 | *$P_{RAD51}$-HOP1-SCD(258-324 a.a.)-LacZ-NVH* |
| pTFW10027 | *$P_{RAD51}$-SML1-NTD(1-27 a.a.)-LacZ-NVH* |
| pTFW10028 | *$P_{RAD51}$-SML1-NTD(1-50 a.a.)-LacZ-NVH* |
| pTFW10005 | *$P_{RAD51}$-SUP35-PND(1-39 a.a.)-LacZ-NVH* |
| pTFW10034 | *$P_{RAD51}$-sup35-PND-$S^{17}$A-LacZ-NVH* |
| pTFW10221 | *$P_{RAD51}$-SUP35-PFD(1-114 a.a.)-LacZ-NVH* |
| pTFW10169 | *$P_{RAD51}$-NEW1-NPD(1-156 a.a.)-LacZ-NVH* |
| pTFW10177 | *$P_{RAD51}$-URE2-UPD(1-91 a.a.)-LacZ-NVH* |
| pTFW10178 | *$P_{RAD51}$-VPS64-NTD(1-152 a.a.)-LacZ-NVH* |
| pTFW10180 | *$P_{RAD51}$-SSK2-NTD(1-191 a.a.)-LacZ-NVH* |
| pTFW10179 | *$P_{RAD51}$-KEL1-NTD(1-69 a.a.)-LacZ-NVH* |
| pTFW9959 | *$P_{RAD51}$-rad51-NTD-3SA-LacZ-NVH* |
| pTFW10095 | *$P_{RAD51}$-rad51-NTD-8STA-LacZ-NVH* |
| pTFW10096 | *$P_{RAD51}$-rad51-NTD-11STA-LacZ-NVH* |
| pTFW9993 | *$P_{RAD51}$-rad51-NTD-6SQA-LacZ-NVH* |
| pTFW9994 | *$P_{RAD51}$-rad51-NTD-9SQA-LacZ-NVH* |
| pTFW9995 | *$P_{RAD51}$-rad51-NTD-12SQA-LacZ-NVH* |
| pTFW10010 | *$P_{RAD51}$-rad51-NTD-3QA -LacZ-NVH* |
| pTFW10254 | *$P_{RAD51}$-rad51-NTD-9QA-LacZ-NVH* |
| pTFW10255 | *$P_{RAD51}$-rad51-NTD-4NA-LacZ-NVH* |
| pTFW10256 | *$P_{RAD51}$-rad51-NTD-13QNA-LacZ-NVH* |
| pTFW10000 | *$P_{RAD51}$-rad53-SCD1-5STA-LacZ-NVH* |
| pTFW10001 | *$P_{RAD51}$-rad53-SCD1-7QA-LacZ-NVH* |
| pTFW10002 | *$P_{RAD51}$-rad53-SCD1-12STQA-LacZ-NVH* |
| pTFW10101 | *$P_{RAD51}$-sup35-PND-3SA-LacZ-NVH* |
| pTFW10081 | *$P_{RAD51}$-sup35-PND-3QA-LacZ-NVH* |
| pTFW10080 | *$P_{RAD51}$-sup35-PND-5QA-LacZ-NVH* |
| pTFW10082 | *$P_{RAD51}$-sup35-PND-8QA-LacZ-NVH* |
| pTFW10141 | *$P_{RAD51}$-sup35-PND-15SQA-LacZ-NVH* |
| pTFW10142 | *$P_{RAD51}$-sup35-PND-9NA-LacZ-NVH* |
| pTFW10143 | *$P_{RAD51}$-sup35-PND-24SQNA-LacZ-NVH* |
| pTFW9974 | *$P_{RAD51}$-GFP-NVH* |
| pTFW9975 | *$P_{RAD51}$-RAD53-SCD1-GFP-NVH* |
| pTFW9976 | *$P_{RAD51}$-GST-NVH* |
| pTFW9977 | *$P_{RAD51}$-RAD53-SCD1-GST-NVH* |
| pTFW10003 | *$P_{RAD51}$- GSTnd-NVH* |
| pTFW10004 | *$P_{RAD51}$-RAD53-SCD1-GSTnd-NVH* |

Table 2. *S. cerevisiae* strains used in this study

| Name | Genotype |
|---|---|
| WHY13008 | *MATa, ho, leu2, ura3, his4-X::LEU2-(NgoMIV;+ori)-URA3, ERG1(SpeI), RAD51::hphMX4* |
| WHY13283 | *MATa, ho, leu2, ura3, his4-X::LEU2-(NgoMIV;+ori)-URA3, ERG1(SpeI), rad51Δ::hphMX4* |
| WHY13416 | *MATa, ho, leu2, ura3, his4-X::LEU2-(NgoMIV;+ori)-URA3, ERG1(SpeI), rad51ΔN::hphMX4* |
| WHY13744 | *MATα, ho::LYS2, leu2, ura3, lys2, HIS4::LEU2-(BamHI;+ori), Erg1(SalI), dmc1::kanMX4, RAD53$^{SCD1}$-rad51ΔN::hphMX4* |
| WHY13743 | *MATa, ho, leu2, ura3, his4-X::LEU2-(NgoMIV;+ori)-URA3, ERG1(SpeI), dmc1::kanMX4, rad53$^{SCD1-5STA}$-rad51ΔN::hphMX4* |
| WHY13741 | *MATα, ho::LYS2, leu2, ura3, lys2, HIS4::LEU2-(BamHI;+ori), ERG1(SalI), SUP35$^{PND}$-rad51ΔN::hphMX4* |

Table 3. The oligonucleotide primers used for g-qPCR and RT-qPCR

| Gene | Primer Polarity | Name | Sequence |
|---|---|---|---|
| *ACT1* | Sense | PA11240 | 5'-CCACCACTGCTGAAAGAGAAATTGT-3' |
| | Antisense | PA11241 | 5'-CTTGACCATCTGGAAGTTCGTAGGA-3' |
| *LacZ* | Sense | PA11246 | 5'-CCGCCGTTTGTTCCCACGGA-3' |
| | Antisense | PA11247 | 5'-CCATCACCGCGAGGCGGTTT-3' |

Table 4. Usage frequency (%) of standard codons [stop codon (*), Q, C, Y and W] and reassigned stop codons (→ Q, → C or → W) in 37 different eukaryotes

| Species | Source | ID | BUSCO Protein (%) | Protein # | TAA | TAG | TGA | CAA | CAG | TGC | TGT | TAC | TAT | TGG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NCBI genetic code: 1 | Non-ciliate eukaryotes | | | | * | * | * | Q | Q | C | C | Y | Y | W |
| *Saccharomyces cerevisiae* | Uniport | UP000002311 | 99.6 | 6062 | 0.16 | 0.08 | 0.01 | 2.77 | 0.89 | 0.63 | 1.03 | 0.86 | 3.10 | 0.93 |
| *Candida albicans* | Uniport | UP000000559 | 98.8 | 6035 | 0.10 | 0.05 | 0.03 | 3.57 | 0.65 | 0.18 | 0.94 | 1.04 | 2.54 | 1.09 |
| *Candida auris* | Uniport | UP000230249 | 97.4 | 5409 | 0.08 | 0.06 | 0.06 | 1.81 | 2.12 | 0.55 | 0.59 | 2.09 | 1.16 | 1.07 |
| *Candida tropicalis* | Uniport | UP000002037 | 94.6 | 6226 | 0.10 | 0.07 | 0.03 | 3.61 | 0.66 | 0.14 | 0.96 | 0.95 | 2.62 | 0.98 |
| *Neurospora crassa* | Uniport | UP000001805 | 99.2 | 10257 | 0.06 | 0.05 | 0.08 | 1.70 | 2.60 | 0.77 | 0.34 | 1.75 | 0.85 | 1.31 |
| *Magnaporthe oryzae* | Uniport | UP000009058 | 98.6 | 12794 | 0.06 | 0.07 | 0.10 | 1.37 | 2.69 | 0.92 | 0.35 | 1.80 | 0.71 | 1.42 |
| *Trichoderma reesei* | PMID: 34908 505 | PRJNA382020 | 99.2 | 13735 | 0.06 | 0.06 | 0.11 | 1.17 | 2.95 | 0.95 | 0.32 | 1.80 | 0.83 | 1.42 |
| *Cryptococcus neoformans* | Uniport | UP000002149 | 99.5 | 6743 | 0.07 | 0.06 | 0.05 | 2.06 | 1.79 | 0.48 | 0.55 | 1.39 | 1.14 | 1.37 |
| *Ustilago maydis* | Uniport | UP000000561 | 99.4 | 6806 | 0.04 | 0.05 | 0.07 | 1.82 | 2.61 | 0.72 | 0.35 | 1.59 | 0.65 | 1.18 |
| *Taiwanofungus camphoratus* | PMID: 35196 809 | PRJNA615295 | 94.6 | 14019 | 0.05 | 0.06 | 0.11 | 1.57 | 2.19 | 0.70 | 0.57 | 1.38 | 1.22 | 1.36 |
| *Dictyostelium discoideum* | Uniport | UP000002195 | 93.7 | 12734 | 0.16 | 0.01 | 0.01 | 4.86 | 0.19 | 0.15 | 1.27 | 0.52 | 3.02 | 0.73 |
| *Plasmodium falciparum* | Uniport | UP000001450 | 99.1 | 5376 | 0.09 | 0.01 | 0.03 | 2.42 | 0.37 | 0.23 | 1.52 | 0.61 | 5.05 | 0.49 |
| *Drosophila melanogaster* | Uniport | UP000000803 | 100 | 22088 | 0.08 | 0.07 | 0.05 | 1.56 | 3.61 | 1.32 | 0.54 | 1.84 | 1.08 | 0.99 |
| *Aedes aegypti* | Uniport | UP000008820 | 99.4 | 18998 | 0.11 | 0.07 | 0.08 | 1.76 | 2.58 | 1.11 | 0.79 | 2.16 | 1.14 | 1.06 |
| *Caenorhabditis elegans* | Uniport | UP000001940 | 100 | 26548 | 0.16 | 0.06 | 0.14 | 2.74 | 1.44 | 0.91 | 1.12 | 1.37 | 1.75 | 1.11 |
| *Danio rerio* | Uniport | UP000000437 | 95.5 | 46844 | 0.11 | 0.06 | 0.14 | 1.18 | 3.35 | 1.12 | 1.13 | 1.70 | 1.26 | 1.16 |
| *Mus musculus* | Uniport | UP000000589 | 99.7 | 55341 | 0.10 | 0.08 | 0.16 | 1.20 | 3.41 | 1.23 | 1.14 | 1.61 | 1.22 | 1.25 |
| *Homo sapiens* | Uniport | UP000005640 | 99.5 | 79038 | 0.10 | 0.08 | 0.16 | 1.23 | 3.42 | 1.26 | 1.06 | 1.53 | 1.22 | 1.32 |
| *Arabidopsis thaliana* | Uniport | UP000006548 | 100 | 39334 | 0.09 | 0.05 | 0.12 | 1.94 | 1.52 | 0.72 | 1.05 | 1.37 | 1.46 | 1.25 |
| *Chlamydomonas reinhardtii* | Uniport | UP000006906 | 98.9 | 18829 | 0.03 | 0.04 | 0.06 | 0.59 | 4.05 | 1.1 | 0.22 | 1.45 | 0.24 | 1.16 |
| NCBI genetic code: 6 | group I ciliates | | | | → Q | → Q | * | Q | Q | C | C | Y | Y | W |
| *Tetrahymena thermophila* | Uniport | UP000009168 | 98.9 | 26972 | 5.46 | 1.63 | 0.16 | 2.04 | 0.48 | 0.79 | 0.99 | 1.22 | 3.09 | 0.51 |
| *Paramecium tetraurelia* | Uniport | UP000000600 | 98.8 | 39461 | 4.53 | 1.48 | 0.22 | 2.54 | 0.57 | 0.61 | 1.21 | 1.12 | 3.14 | 0.76 |
| *Oxytricha trifallax* | Uniport | UP000006077 | 97.1 | 23559 | 3.63 | 1.57 | 0.15 | 2.68 | 1.07 | 0.59 | 0.56 | 1.44 | 2.27 | 0.58 |
| *Stylonychia lemnae* | Uniport | UP000039865 | 97.1 | 20720 | 3.22 | 1.81 | 0.17 | 2.26 | 1.05 | 0.62 | 0.55 | 1.31 | 2.49 | 0.62 |
| *Pseudocohnilembus persalinus* | Uniport | UP000054937 | 92.4 | 13175 | 7.36 | 1.39 | 0.18 | 1.76 | 0.37 | 0.32 | 1.00 | 1.00 | 3.26 | 0.61 |
| NCBI genetic code: 6 | group II ciliate | | | | → Q | → Q | * | Q | Q | C | C | Y | Y | W |
| *Aristerostoma* | MMETSP | MMETSP0125 | 62.5 | 27868 | 0.96 | 1.04 | 0.15 | 2.65 | 0.97 | 0.71 | 0.68 | 1.35 | 2.49 | 0.8 |
| *Favella ehrenbergii* | MMETSP | MMETSP0123 | 85.4 | 26477 | 0.72 | 1.51 | 0.16 | 1.88 | 3.06 | 1.11 | 0.25 | 2.06 | 0.71 | 0.83 |
| *Pseudokeronopsis* | MMETSP | MMETSP0211 MMETSP1396 | 87.2 | 62574 | 1.04 | 1.37 | 0.16 | 2.05 | 2.58 | 0.94 | 0.44 | 2.18 | 1.40 | 0.78 |
| *Strombidium inclinatum* | MMETSP | MMETSP0208 | 83.6 | 32210 | 0.64 | 1.28 | 0.11 | 1.63 | 3.50 | 0.83 | 0.24 | 2.12 | 0.69 | 0.7 |
| *Uronema spp.* | MMETSP | MMETSP0018 | 52.6 | 13887 | 6.90 | 0.66 | 0.17 | 0.80 | 0.08 | 0.28 | 1.63 | 0.80 | 3.62 | 0.87 |
| NCBI genetic code: 1 | group III ciliate | | | | * | * | * | Q | Q | C | C | Y | Y | W |
| *Stentor coeruleus* | Uniport | UP000187209 | 92.4 | 30969 | 0.16 | 0.08 | 0.01 | 2.77 | 0.89 | 0.63 | 1.03 | 0.86 | 3.1 | 0.93 |
| *Climacostomum virens* | MMETSP | MMETSP1397 | 94.7 | 33899 | 0.11 | 0.09 | 0.04 | 1.79 | 2.20 | 1.38 | 0.60 | 2.60 | 0.85 | 1.06 |
| *Litonotus pictus* | MMETSP | MMETSP0209 | 65.5 | 30222 | 0.08 | 0.03 | 0.01 | 2.12 | 1.52 | 0.63 | 0.77 | 1.83 | 2.25 | 0.54 |
| *Protocruzia adherens* | MMETSP | MMETSP0216 | 74.9 | 40577 | 0.07 | 0.04 | 0.04 | 2.91 | 1.24 | 0.69 | 0.94 | 1.30 | 1.83 | 1.00 |
| NCBI genetic code: 10 | group IV ciliate | | | | * | * | → C | Q | Q | C | C | Y | Y | W |
| *Euplotes focardii* | MMETSP | MMETSP0205 MMETSP0206 | 60.8 | 36659 | 0.23 | 0.06 | 0.51 | 2.43 | 1.23 | 0.49 | 0.84 | 1.28 | 2.38 | 0.87 |
| NCBI genetic code: 4 | group IV ciliate | | | | * | * | → W | Q | Q | C | C | Y | Y | W |
| *Blepharisma japonicum* | MMETSP | MMETSP1395 | 81.9 | 22714 | 0.13 | 0.03 | 0.30 | 2.85 | 1.24 | 0.94 | 0.80 | 0.94 | 2.72 | 0.84 |
| NCBI genetic code: 29 | group IV ciliate | | | | → Y | → Y | * | Q | Q | C | C | Y | Y | W |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Mesodinium pulex* | MMETSP | MMETSP0467 | 88.9 | 61058 | 0.29 | 0.56 | 0.13 | 0.77 | 3.33 | 1.53 | 0.25 | 1.78 | 0.34 | 1.29 |

Table 5. The numbers of SCD proteins involved in different biological processes of six representative eukaryotes

| Organism | GO Identity | GO (Biological Process) | SCD protein number | Total protein number |
|---|---|---|---|---|
| *Saccharomyces cerevisiae* | GO:0032197 | RNA-mediated transposition | 82 | 92 |
| | GO:0045944 | positive regulation of transcription by RNA polymerase II | 77 | 229 |
| | GO:0006357 | regulation of transcription by RNA polymerase II | 62 | 177 |
| | GO:0006310 | DNA recombination | 55 | 76 |
| | GO:0006468 | protein phosphorylation | 45 | 138 |
| | GO:0015031 | protein transport | 42 | 153 |
| | GO:0015074 | DNA integration | 40 | 47 |
| | GO:0051301 | cell division | 38 | 167 |
| | GO:0006897 | endocytosis | 33 | 100 |
| | GO:0031505 | fungal-type cell wall organization | 31 | 135 |
| | GO:0000122 | negative regulation of transcription by RNA polymerase II | 30 | 118 |
| | GO:0006281 | DNA repair | 27 | 105 |
| *Candida albicans* | GO:0030447 | filamentous growth | 103 | 387 |
| | GO:0036180 | filamentous growth of a population of unicellular organisms in response to biotic stimulus | 58 | 233 |
| | GO:0006357 | regulation of transcription by RNA polymerase II | 57 | 144 |
| | GO:0009267 | cellular response to starvation | 47 | 197 |
| | GO:0036170 | filamentous growth of a population of unicellular organisms in response to starvation | 47 | 194 |
| | GO:0045944 | positive regulation of transcription by RNA polymerase II | 46 | 105 |
| | GO:0044182 | filamentous growth of a population of unicellular organisms | 44 | 161 |
| | GO:0006355 | regulation of transcription, DNA-templated | 25 | 66 |
| | GO:0031505 | fungal-type cell wall organization | 24 | 124 |
| | GO:0006281 | DNA repair | 24 | 71 |
| | GO:0044011 | single-species biofilm formation on inanimate substrate | 23 | 89 |
| | GO:0030448 | hyphal growth | 22 | 69 |
| *Tetrahymena thermophila* | GO:0007018 | microtubule-based movement | 93 | 110 |
| | GO:0045493 | xylan catabolic process | 56 | 58 |
| | GO:0007155 | cell adhesion | 55 | 89 |
| | GO:0006511 | ubiquitin-dependent protein catabolic process | 45 | 72 |
| | GO:0016579 | protein deubiquitination | 41 | 51 |
| | GO:0018200 | peptidyl-glutamic acid modification | 38 | 49 |
| | GO:0006886 | intracellular protein transport | 35 | 70 |
| | GO:0016192 | vesicle-mediated transport | 33 | 82 |
| | GO:0005975 | carbohydrate metabolic process | 28 | 88 |
| | GO:0015031 | protein transport | 28 | 68 |
| | GO:0006281 | DNA repair | 28 | 46 |
| | GO:0035556 | intracellular signal transduction | 28 | 32 |
| *Dictyostelium discoideum* | GO:0030587 | sorocarp development | 63 | 194 |
| | GO:0019953 | sexual reproduction | 56 | 137 |
| | GO:0006357 | regulation of transcription by RNA polymerase II | 52 | 126 |

| | | | | |
|---|---|---|---|---|
| | GO:0007165 | signal transduction | 52 | 165 |
| | GO:0006357 | regulation of transcription by RNA polymerase II | 52 | 126 |
| | GO:0048870 | cell motility | 43 | 91 |
| | GO:0031152 | aggregation involved in sorocarp development | 42 | 130 |
| | GO:0006886 | intracellular protein transport | 41 | 148 |
| | GO:0043327 | chemotaxis to cAMP | 37 | 100 |
| | GO:0000281 | mitotic cytokinesis | 30 | 101 |
| | GO:0035556 | intracellular signal transduction | 28 | 77 |
| | GO:0006281 | DNA repair | 27 | 73 |
| | GO:0015031 | protein transport | 25 | 103 |
| | GO:0031288 | sorocarp morphogenesis | 23 | 80 |
| *Plasmodium falciparum* | GO:0020033 | antigenic variation | 38 | 177 |
| | GO:0098609 | cell-cell adhesion | 32 | 63 |
| | GO:0020013 | modulation by symbiont of host erythrocyte aggregation | 28 | 73 |
| | GO:0020035 | cytoadherence to microvasculature, mediated by symbiont protein | 28 | 56 |
| *Drosophila melanogaster* | GO:0006357 | regulation of transcription by RNA polymerase II | 364 | 937 |
| | GO:0045944 | positive regulation of transcription by RNA polymerase II | 213 | 457 |
| | GO:0007411 | axon guidance | 199 | 375 |
| | GO:0006355 | regulation of transcription, DNA-templated | 147 | 413 |
| | GO:0007165 | signal transduction | 142 | 490 |
| | GO:0016319 | mushroom body development | 140 | 212 |
| | GO:0035556 | intracellular signal transduction | 129 | 280 |
| | GO:0007155 | cell adhesion | 127 | 196 |
| | GO:0007419 | ventral cord development | 125 | 179 |
| | GO:0030036 | actin cytoskeleton organization | 115 | 217 |
| | GO:0000122 | negative regulation of transcription by RNA polymerase II | 110 | 263 |
| | GO:0006468 | protein phosphorylation | 109 | 288 |
| | GO:0050770 | regulation of axonogenesis | 101 | 116 |
| | GO:0048666 | neuron development | 97 | 121 |
| | GO:0048813 | dendrite morphogenesis | 96 | 181 |

Table 6. The numbers of polyQ/N proteins involved in different biological processes of six representative eukaryotes

| Organism | GO Identity | GO (Biological Process) | polyQ/N protein number | Total protein number |
|---|---|---|---|---|
| *Saccharomyces cerevisiae* | GO:0045944 | positive regulation of transcription by RNA polymerase II | 140 | 229 |
| | GO:0006357 | regulation of transcription by RNA polymerase II | 101 | 177 |
| | GO:0032197 | RNA-mediated transposition | 88 | 92 |
| | GO:0006468 | protein phosphorylation | 74 | 138 |
| | GO:0051301 | cell division | 71 | 167 |
| | GO:0000122 | negative regulation of transcription by RNA polymerase II | 68 | 118 |
| | GO:0015031 | protein transport | 54 | 153 |
| | GO:0006310 | DNA recombination | 53 | 76 |
| | GO:0006338 | chromatin remodeling | 47 | 91 |
| | GO:0015074 | DNA integration | 45 | 47 |
| | GO:0006897 | endocytosis | 43 | 100 |
| | GO:0007049 | cell cycle | 43 | 95 |
| *Candida albicans* | GO:0030447 | filamentous growth | 218 | 387 |
| | GO:0036180 | filamentous growth of a population of unicellular organisms in response to biotic stimulus | 130 | 233 |
| | GO:0009267 | cellular response to starvation | 109 | 197 |
| | GO:0036170 | filamentous growth of a population of unicellular organisms in response to starvation | 104 | 194 |
| | GO:0006357 | regulation of transcription by RNA polymerase II | 103 | 144 |
| | GO:0044182 | filamentous growth of a population of unicellular organisms | 89 | 161 |
| | GO:0045944 | positive regulation of transcription by RNA polymerase II | 80 | 105 |
| | GO:0031505 | fungal-type cell wall organization | 53 | 124 |
| | GO:0044011 | single-species biofilm formation on inanimate substrate | 51 | 89 |
| | GO:0006355 | regulation of transcription, DNA-templated | 51 | 66 |
| | GO:0015031 | protein transport | 43 | 82 |
| | GO:0030448 | hyphal growth | 42 | 69 |
| *Tetrahymena thermophila* | GO:0007018 | microtubule-based movement | 107 | 110 |
| | GO:0007155 | cell adhesion | 84 | 89 |
| | GO:0045493 | xylan catabolic process | 58 | 58 |
| | GO:0006511 | ubiquitin-dependent protein catabolic process | 56 | 72 |
| | GO:0016192 | vesicle-mediated transport | 55 | 82 |
| | GO:0006886 | intracellular protein transport | 54 | 70 |
| | GO:0005975 | carbohydrate metabolic process | 48 | 88 |
| | GO:0018200 | peptidyl-glutamic acid modification | 47 | 49 |
| | GO:0016579 | protein deubiquitination | 46 | 51 |
| | GO:0015031 | protein transport | 45 | 68 |
| | GO:0032259 | methylation | 44 | 68 |
| | GO:0006281 | DNA repair | 38 | 46 |
| *Dictyostelium discoideum* | GO:0007165 | signal transduction | 139 | 165 |
| | GO:0030587 | sorocarp development | 132 | 194 |
| | GO:0006357 | regulation of transcription by RNA polymerase II | 113 | 126 |

| | | | | |
|---|---|---|---|---|
| | GO:0019953 | sexual reproduction | 100 | 137 |
| | GO:0006886 | intracellular protein transport | 97 | 148 |
| | GO:0031152 | aggregation involved in sorocarp development | 85 | 130 |
| | GO:0006355 | regulation of transcription, DNA-templated | 75 | 87 |
| | GO:0043327 | chemotaxis to cAMP | 72 | 100 |
| | GO:0000398 | mRNA splicing, via spliceosome | 72 | 88 |
| | GO:0035556 | intracellular signal transduction | 72 | 77 |
| | GO:0006468 | protein phosphorylation | 65 | 77 |
| | GO:0015031 | protein transport | 63 | 103 |
| *Plasmodium falciparum* | GO:0020033 | antigenic variation | 82 | 177 |
| | GO:0098609 | cell-cell adhesion | 62 | 63 |
| | GO:0020013 | modulation by symbiont of host erythrocyte aggregation | 58 | 73 |
| | GO:0006364 | rRNA processing | 57 | 73 |
| | GO:0020035 | cytoadherence to microvasculature, mediated by symbiont protein | 55 | 56 |
| | GO:0006468 | protein phosphorylation | 50 | 73 |
| | GO:0000398 | mRNA splicing, via spliceosome | 41 | 66 |
| | GO:0006508 | proteolysis | 32 | 41 |
| | GO:0035556 | intracellular signal transduction | 29 | 48 |
| | GO:0044409 | entry into host | 27 | 35 |
| | GO:0006355 | regulation of transcription, DNA-templated | 26 | 35 |
| | GO:0006281 | DNA repair | 24 | 37 |
| *Drosophila melanogaster* | GO:0006357 | regulation of transcription by RNA polymerase II | 620 | 937 |
| | GO:0045944 | positive regulation of transcription by RNA polymerase II | 319 | 457 |
| | GO:0007411 | axon guidance | 278 | 375 |
| | GO:0000122 | negative regulation of transcription by RNA polymerase II | 190 | 263 |
| | GO:0035556 | intracellular signal transduction | 178 | 280 |
| | GO:0016319 | mushroom body development | 167 | 212 |
| | GO:0007155 | cell adhesion | 161 | 196 |
| | GO:0030036 | actin cytoskeleton organization | 160 | 217 |
| | GO:0006468 | protein phosphorylation | 158 | 288 |
| | GO:0007419 | ventral cord development | 154 | 179 |
| | GO:0048813 | dendrite morphogenesis | 132 | 181 |
| | GO:0007476 | imaginal disc-derived wing morphogenesis | 131 | 196 |

Table 7. The numbers of polyQ proteins of different biological processes in six ciliate species

| Organism | GO Identity | GO (Biological Process) | polyQ protein number | Total protein number |
|---|---|---|---|---|
| *Tetrahymena thermophila* | GO:0007018 | microtubule-based movement | 69 | 110 |
| | GO:0045493 | xylan catabolic process | 55 | 58 |
| | GO:0018200 | peptidyl-glutamic acid modification | 31 | 49 |
| | GO:0006886 | intracellular protein transport | 27 | 70 |
| | GO:0006511 | ubiquitin-dependent protein catabolic process | 26 | 72 |
| | GO:0016579 | protein deubiquitination | 24 | 51 |
| | GO:0016192 | vesicle-mediated transport | 22 | 82 |
| | GO:0015031 | protein transport | 22 | 68 |
| | GO:0007155 | cell adhesion | 21 | 89 |
| | GO:0035556 | intracellular signal transduction | 21 | 32 |
| | GO:0006281 | DNA repair | 18 | 46 |
| | GO:0009190 | cyclic nucleotide biosynthetic process | 17 | 22 |
| *Paramecium tetraurelia* | GO:0018105 | peptidyl-serine phosphorylation | 276 | 1134 |
| | GO:0035556 | intracellular signal transduction | 212 | 722 |
| | GO:0006511 | ubiquitin-dependent protein catabolic process | 121 | 363 |
| | GO:0006355 | regulation of transcription, DNA-templated | 119 | 351 |
| | GO:0046777 | protein autophosphorylation | 115 | 474 |
| | GO:0007018 | microtubule-based movement | 109 | 226 |
| | GO:0016567 | protein ubiquitination | 92 | 286 |
| | GO:0006886 | intracellular protein transport | 85 | 405 |
| | GO:0000122 | negative regulation of transcription by RNA polymerase II | 84 | 233 |
| | GO:1905037 | autophagosome organization | 77 | 225 |
| | GO:0007165 | signal transduction | 62 | 236 |
| | GO:0046999 | regulation of conjugation | 60 | 118 |
| *Oxytricha trifallax* | GO:0007018 | microtubule-based movement | 85 | 127 |
| | GO:0018200 | peptidyl-glutamic acid modification | 63 | 89 |
| | GO:0006281 | DNA repair | 35 | 82 |
| | GO:0016579 | protein deubiquitination | 32 | 66 |
| | GO:0006511 | ubiquitin-dependent protein catabolic process | 30 | 77 |
| | GO:0015031 | protein transport | 25 | 74 |
| | GO:0006260 | DNA replication | 24 | 66 |
| | GO:0006886 | intracellular protein transport | 23 | 60 |
| | GO:0006355 | regulation of transcription, DNA-templated | 20 | 38 |
| | GO:0016192 | vesicle-mediated transport | 19 | 74 |
| | GO:0006397 | mRNA processing | 19 | 45 |
| | GO:0005975 | carbohydrate metabolic process | 18 | 84 |
| *Stylonychia lemnae* | GO:0007018 | microtubule-based movement | 66 | 105 |
| | GO:0018200 | peptidyl-glutamic acid modification | 54 | 73 |
| | GO:0051301 | cell division | 22 | 38 |

| | | | | |
|---|---|---|---|---|
| | GO:0006281 | DNA repair | 21 | 60 |
| | GO:0006886 | intracellular protein transport | 20 | 59 |
| | GO:0015031 | protein transport | 19 | 64 |
| | GO:0016192 | vesicle-mediated transport | 19 | 64 |
| | GO:0006511 | ubiquitin-dependent protein catabolic process | 17 | 69 |
| | GO:0006397 | mRNA processing | 16 | 34 |
| | GO:0006355 | regulation of transcription, DNA-templated | 16 | 33 |
| | GO:0016579 | protein deubiquitination | 15 | 54 |
| | GO:0007049 | cell cycle | 15 | 37 |
| *Pseudocohnilembus persalinus* | GO:0007018 | microtubule-based movement | 66 | 105 |
| | GO:0018200 | peptidyl-glutamic acid modification | 54 | 73 |
| | GO:0051301 | cell division | 22 | 38 |
| | GO:0006281 | DNA repair | 21 | 60 |
| | GO:0006886 | intracellular protein transport | 20 | 59 |
| | GO:0015031 | protein transport | 19 | 64 |
| | GO:0016192 | vesicle-mediated transport | 19 | 64 |
| | GO:0006511 | ubiquitin-dependent protein catabolic process | 17 | 69 |
| | GO:0006397 | mRNA processing | 16 | 34 |
| | GO:0006355 | regulation of transcription, DNA-templated | 16 | 33 |
| | GO:0016579 | protein deubiquitination | 15 | 54 |
| | GO:0007049 | cell cycle | 15 | 37 |
| *Stentor coeruleus* | GO:0007018 | microtubule-based movement | 5 | 226 |
| | GO:0006886 | intracellular protein transport | 4 | 124 |
| | GO:0016192 | vesicle-mediated transport | 4 | 123 |
| | GO:0006355 | regulation of transcription, DNA-templated | 3 | 46 |
| | GO:0007165 | signal transduction | 2 | 118 |
| | GO:0060271 | cilium assembly | 2 | 31 |
| | GO:0030705 | cytoskeleton-dependent intracellular transport | 2 | 14 |
| | GO:0006897 | endocytosis | 2 | 11 |
| | GO:0042073 | intraciliary transport | 2 | 8 |
| | GO:0043087 | regulation of GTPase activity | 2 | 6 |
| | GO:0045162 | clustering of voltage-gated sodium channels | 2 | 2 |
| | GO:0006511 | ubiquitin-dependent protein catabolic process | 1 | 159 |

## References


1. Prescott DM. The DNA of ciliated protozoa. Microbiol Rev. 1994;58(2):233-67.
2. Chalker DL, Yao MC. DNA elimination in ciliates: transposon domestication and genome surveillance. Annu Rev Genet. 2011;45:227-46.
3. Keeney S, Giroux CN, Kleckner N. Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. Cell. 1997;88(3):375-84.
4. de Massy B. Initiation of meiotic recombination: how and where? Conservation and specificities among eukaryotes. Annu Rev Genet. 2013;47:563-99.
5. Zickler D. From early homologue recognition to synaptonemal complex formation. Chromosoma. 2006;115(3):158-74.
6. Rog O, Kohler S, Dernburg AF. The synaptonemal complex has liquid crystalline properties and spatially regulates meiotic recombination factors. Elife. 2017;6.
7. Chi JY, Mahe F, Loidl J, Logsdon J, Dunthorn M. Meiosis gene inventory of four ciliates reveals the prevalence of a synaptonemal complex-independent crossover pathway. Molecular Biology and Evolution. 2014;31(3):660-72.
8. Loidl J. *Tetrahymena* meiosis: Simple yet ingenious. PLoS Genet. 2021;17(7):e1009627.
9. Loidl J, Mochizuki K. *Tetrahymena* meiotic nuclear reorganization is induced by a checkpoint kinase-dependent response to DNA damage. Mol Biol Cell. 2009;20(9):2428-37.

10. Yan GX, Dang H, Tian M, Zhang J, Shodhan A, Ning YZ, et al. Cyc17, a meiosis-specific cyclin, is essential for anaphase initiation and chromosome segregation in *Tetrahymena thermophila*. Cell Cycle. 2016;15(14):1855-64.
11. Yan GX, Zhang J, Shodhan A, Tian M, Miao W. Cdk3, a conjugation-specific cyclin-dependent kinase, is essential for the initiation of meiosis in *Tetrahymena thermophila*. Cell Cycle. 2016;15(18):2506-14.
12. Xu J, Li X, Song W, Wang W, Gao S. Cyclin Cyc2p is required for micronuclear bouquet formation in *Tetrahymena thermophila*. Sci China Life Sci. 2019;62(5):668-80.
13. Zhang J, Yan G, Tian M, Ma Y, Xiong J, Miao W. A DP-like transcription factor protein interacts with E2fl1 to regulate meiosis in *Tetrahymena thermophila*. Cell Cycle. 2018;17(5):634-42.
14. Caron F, Meyer E. Does *Paramecium primaurelia* use a different genetic code in its macronucleus? Nature. 1985;314(6007):185-8.
15. Helftenbein E. Nucleotide sequence of a macronuclear DNA molecule coding for alpha-tubulin from the ciliate *Stylonychia lemnae*. Special codon usage: TAA is not a translation termination codon. Nucleic Acids Res. 1985;13(2):415-33.
16. Horowitz S, Gorovsky MA. An unusual genetic code in nuclear genes of *Tetrahymena*. Proc Natl Acad Sci U S A. 1985;82(8):2452-5.
17. Preer JR, Jr., Preer LB, Rudman BM, Barnett AJ. Deviation from the universal code shown by the gene for surface protein 51A in *Paramecium*. Nature. 1985;314(6007):188-90.

18. Lozupone CA, Knight RD, Landweber LF. The molecular basis of nuclear genetic code change in ciliates. Curr Biol. 2001;11(2):65-74.
19. Ring KL, Cavalcanti AR. Consequences of stop codon reassignment on protein evolution in ciliates with alternative genetic codes. Mol Biol Evol. 2008;25(1):179-86.
20. Salim HM, Ring KL, Cavalcanti AR. Patterns of codon usage in two ciliates that reassign the genetic code: *Tetrahymena thermophila* and *Paramecium tetraurelia*. Protist. 2008;159(2):283-98.
21. Dohra H, Fujishima M, Suzuki H. Analysis of amino acid and codon usage in *Paramecium bursaria*. FEBS Lett. 2015;589(20 Pt B):3113-8.
22. Xiong J, Wang G, Cheng J, Tian M, Pan X, Warren A, et al. Genome of the facultative scuticociliatosis pathogen *Pseudocohnilembus persalinus* provides insight into its virulence through horizontal gene transfer. Sci Rep. 2015;5:15470.
23. Swart EC, Serra V, Petroni G, Nowacki M. Genetic codes with no dedicated stop codon: Context-dependent translation termination. Cell. 2016;166(3):691-702.
24. Heaphy SM, Mariotti M, Gladyshev VN, Atkins JF, Baranov PV. Novel ciliate genetic code variants including the reassignment of all three stop codons to sense codons in *Condylostoma magnum*. Mol Biol Evol. 2016;33(11):2885-9.
25. Slabodnick MM, Ruby JG, Reiff SB, Swart EC, Gosai S, Prabakaran S, et al. The macronuclear genome of *Stentor coeruleus* reveals tiny introns in a giant cell. Curr Biol. 2017;27(4):569-75.
26. Kollmar M, Muhlhausen S. Nuclear codon reassignments in the genomics era and mechanisms behind their evolution. Bioessays. 2017;39(5).

27. Woo TT, Chuang CN, Higashide M, Shinohara A, Wang TF. Dual roles of yeast Rad51 N-terminal domain in repairing DNA double-strand breaks. Nucleic Acids Res. 2020;48(15):8474-89.
28. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. Proteins. 2001;42(1):38-48.
29. Macossay-Castillo M, Marvelli G, Guharoy M, Jain A, Kihara D, Tompa P, et al. The balancing act of intrinsically disordered proteins: enabling functional diversity while minimizing promiscuity. . J Mol Biol. 2019;431(8):1650-70.
30. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins. 2000;41(3):415-27.
31. Traven A, Heierhorst J. SQ/TQ cluster domains: concentrated ATM/ATR kinase phosphorylation site regions in DNA-damage-response proteins. Bioessays. 2005;27(4):397-407.
32. Cheung HC, San Lucas FA, Hicks S, Chang K, Bertuch AA, Ribes-Zamora A. An S/T-Q cluster domain census unveils new putative targets under Tel1/Mec1 control. BMC Genomics. 2012;13:664.
33. Kim ST, Lim DS, Canman CE, Kastan MB. Substrate specificities and identification of putative substrates of ATM kinase family members. J Biol Chem. 1999;274(53):37538-43.
34. Craven RJ, Greenwell PW, Dominska M, Petes TD. Regulation of genome stability by *TEL1* and *MEC1*, yeast homologs of the mammalian ATM and ATR genes. Genetics. 2002;161(2):493-507.

35. Menolfi D, Zha S. ATM, ATR and DNA-PKcs kinases-the lessons from the mouse models: inhibition not equal deletion. Cell Biosci. 2020;10:8.

36. Corcoles-Saez I, Dong K, Johnson AL, Waskiewicz E, Costanzo M, Boone C, et al. Essential function of Mec1, the budding yeast ATM/ATR checkpoint-response kinase, in protein homeostasis. . Dev Cell. 2018;46(4):495-503 e2.

37. Corcoles-Saez I, Dong K, Cha RS. Versatility of the Mec1(ATM/ATR) signaling network in mediating resistance to replication, genotoxic, and proteotoxic stresses. Curr Genet. 2019;65(3):657-61.

38. Lee H, Yuan C, Hammet A, Mahajan A, Chen ES, Wu MR, et al. Diphosphothreonine-specific interaction between an SQ/TQ cluster and an FHA domain in the Rad53-Dun1 kinase cascade. Mol Cell. 2008;30(6):767-78.

39. Carballo JA, Johnson AL, Sedgwick SG, Cha RS. Phosphorylation of the axial element protein Hop1 by Mec1/Tel1 ensures meiotic interhomolog recombination. Cell. 2008;132(5):758-70.

40. Chuang CN, Cheng YH, Wang TF. Mek1 stabilizes Hop1-Thr318 phosphorylation to promote interhomolog recombination and checkpoint responses during yeast meiosis. Nucleic Acids Res. 2012;40(22):11416-27.

41. Zhao X, Muller EG, Rothstein R. A suppressor of two essential checkpoint genes identifies a novel protein that negatively affects dNTP pools. Mol Cell. 1998;2(3):329-40.

42. Chavali S, Chavali PL, Chalancon G, de Groot NS, Gemayel R, Latysheva NS, et al. Constraints and consequences of the emergence of amino acid repeats in eukaryotic proteins. Nat Struct Mol Biol. 2017;24(9):765-77.

43. Mier P, Elena-Real C, Urbanek A, Bernado P, Andrade-Navarro MA. The importance of definitions in the study of polyQ regions: A tale of thresholds, impurities and sequence context. Comput Struct Biotechnol J. 2020;18:306-13.

44. Lu X, Murphy RM. Synthesis and disaggregation of asparagine repeat-containing peptides. J Pept Sci. 2014;20(11):860-7.

45. Zoghbi HY, Orr HT. Glutamine repeats and neurodegeneration. Annu Rev Neurosci. 2000;23:217-47.

46. Michelitsch MD, Weissman JS. A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions. Proc Natl Acad Sci U S A. 2000;97(22):11910-5.

47. Uptain SM, Lindquist S. Prions as protein-based genetic elements. Annu Rev Microbiol. 2002;56:703-41.

48. Toombs JA, Liss NM, Cobble KR, Ben-Musa Z, Ross ED. [PSI+] maintenance is dependent on the composition, not primary sequence, of the oligopeptide repeat domain. PLoS One. 2011;6(7):e21953.

49. Wickner RB, Edskes HK, Roberts BT, Baxa U, Pierce MM, Ross ED, et al. Prions: proteins as genes and infectious entities. Genes Dev. 2004;18(5):470-85.

50. Wickner RB. [*URE3*] as an altered URE2 protein: evidence for a prion analog in *Saccharomyces cerevisiae*. Science. 1994;264(5158):566-9.

51. Shewmaker F, Mull L, Nakayashiki T, Masison DC, Wickner RB. Ure2p function is enhanced by its prion domain in *Saccharomyces cerevisiae*. Genetics. 2007;176(3):1557-65.

52. Ngo S, Chiang V, Ho E, Le L, Guo Z. Prion domain of yeast Ure2 protein adopts a completely disordered structure: a solid-support EPR study. PLoS One. 2012;7(10):e47248.
53. Kasari V, Pochopien AA, Margus T, Murina V, Turnbull K, Zhou Y, et al. A role for the *Saccharomyces cerevisiae* ABCF protein New1 in translation termination/recycling. Nucleic Acids Res. 2019;47(16):8807-20.
54. Santoso A, Chien P, Osherovich LZ, Weissman JS. Molecular basis of a yeast prion species barrier. Cell. 2000;100(2):277-88.
55. Osherovich LZ, Weissman JS. Multiple Gln/Asn-rich prion domains confer susceptibility to induction of the yeast [PSI(+)] prion. Cell. 2001;106(2):183-94.
56. Chen B, Retzlaff M, Roos T, Frydman J. Cellular strategies of protein quality control. Cold Spring Harb Perspect Biol. 2011;3(8):a004374.
57. Shorter J, Southworth DR. Spiraling in Control: Structures and Mechanisms of the Hsp104 Disaggregase. Cold Spring Harb Perspect Biol. 2019;11(8).
58. Ye X, Lin J, Mayne L, Shorter J, Englander SW. Structural and kinetic basis for the regulation and potentiation of Hsp104 function. Proc Natl Acad Sci U S A. 2020;117(17):9384-92.
59. Mullally JE, Chernova T, Wilkinson KD. Doa1 is a Cdc48 adapter that possesses a novel ubiquitin binding domain. Mol Cell Biol. 2006;26(3):822-30.
60. Zhao G, Li G, Schindelin H, Lennarz WJ. An Armadillo motif in Ufd3 interacts with Cdc48 and is involved in ubiquitin homeostasis and protein degradation. Proc Natl Acad Sci U S A. 2009;106(38):16197-202.

61. Tyedmers J, Madariaga ML, Lindquist S. Prion switching in response to environmental stress. PLoS Biol. 2008;6(11):e294.
62. Swaminathan S, Amerik AY, Hochstrasser M. The Doa4 deubiquitinating enzyme is required for ubiquitin homeostasis in yeast. Mol Biol Cell. 1999;10(8):2583-94.
63. Dudley AM, Janse DM, Tanay A, Shamir R, Church GM. A global view of pleiotropy and phenotypically derived gene function in yeast. Mol Syst Biol. 2005;1:2005 0001.
64. Dasgupta A, Ramsey KL, Smith JS, Auble DT. Sir Antagonist 1 (San1) is a ubiquitin ligase. J Biol Chem. 2004;279(26):26830-8.
65. Fredrickson EK, Gallagher PS, Clowes Candadai SV, Gardner RG. Substrate recognition in nuclear protein quality control degradation is governed by exposed hydrophobicity that correlates with aggregation and insolubility. J Biol Chem. 2013;288(9):6130-9.
66. Porat Z, Landau G, Bercovich Z, Krutauz D, Glickman M, Kahana C. Yeast antizyme mediates degradation of yeast ornithine decarboxylase by yeast but not by mammalian proteasome: new insights on yeast antizyme. J Biol Chem. 2008;283(8):4528-34.
67. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol. 1999;293(2):321-31.
68. Posey AE, Holehouse AS, Pappu RV. Phase separation of intrinsically disordered proteins. . Methods Enzymol. 2018;611:1-30.
69. Ramazzotti M, Monsellier E, Kamoun C, Degl'Innocenti D, Melki R. Polyglutamine repeats are associated to specific sequence biases that are conserved among eukaryotes. PLoS One. 2012;7(2):e30824.

70. Li C, Nagel J, Androulakis S, Song J, Buckle AM. PolyQ 2.0: an improved version of PolyQ, a database of human polyglutamine proteins. Database (Oxford). 2016;2016.
71. Totzeck F, Andrade-Navarro MA, Mier P. The protein structure context of polyQ regions. PLoS One. 2017;12(1):e0170801.
72. Mier P, Andrade-Navarro MA. Between interactions and aggregates: The polyQ balance. Genome Biol Evol. 2021;13(11).
73. Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness. Methods Mol Biol. 2019;1962:227-45.
74. Cara L, Baitemirova M, Follis J, Larios-Sanz M, Ribes-Zamora A. The ATM- and ATR-related SCD domain is over-represented in proteins involved in nervous system development. Sci Rep. 2016;6:19050.
75. Kuspa A, Loomis WF. The genome of *Dictyostelium discoideum*. Methods Mol Biol. 2006;346:15-30.
76. Davies HM, Nofal SD, McLaughlin EJ, Osborne AR. Repetitive sequences in malaria parasite proteins. FEMS Microbiol Rev. 2017;41(6):923-40.
77. Mier P, Alanis-Lobato G, Andrade-Navarro MA. Context characterization of amino acid homorepeats using evolution, position, and order. Proteins. 2017;85(4):709-19.
78. de Castro E, Sigrist CJ, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, et al. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res. 2006;34(Web Server issue):W362-5.

79. Chen YK, Hsueh YP. Cortactin-binding protein 2 modulates the mobility of cortactin and regulates dendritic spine formation and maintenance. J Neurosci. 2012;32(3):1043-55.

80. Hsueh YP. Neuron-specific regulation on F-actin cytoskeletons: The role of CTTNBP2 in dendritic spinogenesis and maintenance. Commun Integr Biol. 2012;5(4):334-6.

81. Meszaros B, Erdos G, Dosztanyi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res. 2018;46(W1):W329-W37.

82. Shih PY, Fang YL, Shankar S, Lee SP, Hu HT, Chen H, et al. Phase separation and zinc-induced transition modulate synaptic distribution and association of autism-linked CTTNBP2 and SHANK3. Nat Commun. 2022;13(1):2664.

83. Curcio MJ, Lutz S, Lesage P. The Ty1 LTR-retrotransposon of budding yeast, *Saccharomyces cerevisiae*. Microbiol Spectr. 2015;3(2):1-35.

84. Eichinger L, Pachebat JA, Glockner G, Rajandream MA, Sucgang R, Berriman M, et al. The genome of the social amoeba *Dictyostelium discoideum*. Nature. 2005;435(7038):43-57.

85. Martin-Gonzalez J, Montero-Bullon JF, Lacal J. *Dictyostelium discoideum* as a non-mammalian biomedical model. Microb Biotechnol. 2021;14(1):111-25.

86. Ralph SA, Scheidig-Benatar C, Scherf A. Antigenic variation in *Plasmodium falciparum* is associated with movement of var loci between subnuclear locations. Proc Natl Acad Sci U S A. 2005;102(15):5414-9.

87. Schwope RM, Chalker DL. Mutations in Pdd1 reveal distinct requirements for its chromodomain and chromoshadow domain in directing histone methylation and heterochromatin elimination. Eukaryot Cell. 2014;13(2):190-201.

88. Mochizuki K, Gorovsky MA. Conjugation-specific small RNAs in *Tetrahymena* have predicted properties of scan (scn) RNAs involved in genome rearrangement. Gene Dev. 2004;18(17):2068-73.

89. Noto T, Kurth HM, Kataoka K, Aronica L, DeSouza LV, Siu KWM, et al. The *Tetrahymena* Argonaute-binding protein Giw1p directs a mature Argonaute-siRNA complex to the nucleus. Cell. 2010;140(5):692-703.

90. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, et al. The marine microbial eukaryote transcriptome sequencing project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol. 2014;12(6):e1001889.

91. Johnson LK, Alexander H, Brown CT. Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes. Gigascience. 2019;8(4).

92. Schoelmerich MC, Sachdeva R, West-Roberts J, Waldburger L, Banfield JF. Tandem repeats in giant archaeal Borg elements undergo rapid evolution and create new intrinsically disordered regions in proteins. PLoS Biol. 2023;21(1):e3001980.

93. Lin FM, Lai YJ, Shen HJ, Cheng YH, Wang TF. Yeast axial-element protein, Red1, binds SUMO chains to promote meiotic interhomologue recombination and chromosome synapsis. EMBO J. 2010;29(3):586-96.