Noncanonical usage of stop codons in ciliates expands proteins with 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; stop codon reassignment; ciliates; S/T-Q cluster domain (SCD); polyglutamine (polyQ) and polyglutamine/asparagine (polyQ/N); intrinsically disordered region.

This PDF file includes:
Main text file, twelve figures and two tables

The supplemental information includes four tables (Table S1-S4), three figures (Figure S1-S3)

34 excel tables are also provided (Datasets DS1-DS34).

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 report that due to their high percentages of STQ or STQN amino acid content, four SCDs and three prion-causing Q/N-rich motifs of yeast proteins possess autonomous protein expression-enhancing activities. Since these Q-rich motifs can endow proteins with structural and functional plasticity, we suggest that they represent useful toolkits for evolutionary novelty. 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 and pseudohyphal growth, *Candida albicans* filamentous growth, ciliate peptidyl-glutamic acid modification and microtubule-based movement, *Tetrahymena thermophila* xylan catabolism and meiosis, *Dictyostelium discoideum* development and sexual cycles, *Plasmodium falciparum* infection, and the nervous systems of *Drosophila melanogaster, Mus musculus* and *Homo sapiens*. We also show that Q-rich-motif proteins are expanded massively in ten ciliates with reassigned TAA^Q and TAG^Q codons. Notably, the usage frequency of CAG^Q is much lower in ciliates with reassigned TAA^Q and TAG^Q codons than in organisms with expanded and unstable Q runs (e.g., *D. melanogaster* and *H. sapiens*), indicating that the use of noncanonical stop codons in ciliates may have coevolved with codon usage biases to avoid triplet repeat disorders mediated by CAG/GTC replication slippage.

Introduction

We reported previously that the NH_2-terminal domain (NTD; residues 1–66) of budding yeast *Saccharomyces cerevisiae* Rad51 protein contains three SQ motifs (S^2Q, $S^{12}Q$and $S^{30}Q$) (1). The S/T-Q motifs, comprising S or T followed by Q, are the target sites of DNA damage sensor protein kinases, i.e., ATM (ataxia-telangiectasia mutated), ATR (RAD3-related) (2, 3) and DNA-dependent protein kinase (DNA-PK) (4, 5). 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 (2, 3, 6). This clustering of three SQ motifs within a stretch of 31 amino acids in Rad51-NTD fulfills the criteria to define an S/T-Q cluster domain (SCD) (4, 5). The three SQ motifs of Rad51-NTD are phosphorylated in a Mec1- and Tel1-dependent manner 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 (1), supporting the notion that Mec1 and Tel1 exhibit an essential function in regulating protein homeostasis (proteostasis) in *S. cerevisiae* (7, 8).

A unifying definition of an SCD is having ≥3 S/T-Q sites within a stretch of 50-100 amino acids (4, 5). One of the best-understood mechanisms of SCD phosphorylation involves the association of SCDs with their binding partners containing a forkhead-associated (FHA) domain. For example, Mec1/Tel1-dependent phosphorylation of Rad53-SCD1 (residues 1-29) and Hop1-SCD (residues 258-324) specifically recruits and activates their downstream DNA damage checkpoint kinases Dun1 and Mek1, respectively (9-11). Dun1 phosphorylates three serine residues (S56, S58 and S60) of the ribonucleotide reductase inhibitor Sml1, subsequently promoting Sml1 ubiquitination by the E2 ubiquitin-conjugating enzyme Rad6 and the E3

ubiquitin ligase Ubr2, as well as promoting Sml1 degradation via the 26S proteasome (12-14). Mek1 phosphorylates two Rad51 accessory factors, Rad54 and Hed1 (a meiosis-specific inhibitor of Rad51), suppressing Rad51's strand-exchange activity and preventing Rad51-mediated DSB repair, respectively (15, 16).

There are many other SCD-containing proteins that are neither targets of ATM/Tel1 or ATR/Mec1 nor functionally linked to DNA Damage response or DNA repair (4, 5), indicating that SCDs might possess previously uncharacterized biochemical properties or physiological functions. Interestingly, due to their high percentages of STQ amino acid content, SCDs often are intrinsically disordered regions (IDRs) in their native states rather than adopting stable secondary and/or tertiary structures (4). A common feature of IDRs is their high content of serine (S), threonine (T), glutamine (Q), asparanine (N), proline (P), glycine (G) or charged amino acids [arginine (R), lysine (K) and histidine (H)] (17-20). Functionally, IDRs are key components of subcellular machineries and signaling pathways because they have the potential to associate with many partners due to their multiple possible metastable conformations. Many IDRs are regulated by alternative splicing and post-translational modifications. Some IDRs are involved in the formation of various membraneless organelles via intracellular liquid-liquid phase separation (21, 22). Highly-charged IDRs can act as entropic bristles that, when translationally fused to their partner proteins, only enhance the water solubility but not the steady-state levels of their partner proteins (23).

In this study, we first report that seven Q-rich motifs of *S. cerevisiae* proteins, including Rad51-NTD (1), have high STQN or STQ amino acid contents and exhibit autonomous expression-enhancing activity for high-level production of native protein and when fused to exogenous target proteins, e.g., β-galactosidase (LacZ), *in vivo*. We also reveal structural and

genetic requirements for the "nanny" function of these Q-rich motifs in regulating protein homeostasis, leading to the hypothesis that Q-rich motifs are useful toolkits for structural and functional plasticity, as well as evolutionary novelty. 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. Notably, relative to most other eukaryotes, 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, 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. *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 (24-36). Although it has been reported previously that Q is used more frequently in *Tetrahymena thermophila* and *Paramecium tetraurelia* than in other species (29, 30), 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? Finally, the codons that code for the polyQ motifs are prone to a CAG/GTC-slippage mechanism during DNA replication, and thus many Q-runs are unstable and expanded in some eukaryotic organisms, leading to polyQ-associated diseases (e.g., Huntington's Disease) (37-39). Accordingly, polyQ tracts in proteins associated with disease are more enriched in the CAG codon, becoming almost CAG exclusive (39, 40). In contrast, CAA insertions in a *Drosophila* model of a polyQ-associated disease revealed that even though the resulting polyQ tract is of the same length as the disease-associated tract, the proteins display reduced toxicity (41). In this study, we determine and compare the usage frequency of TAA^{Q}, TAG^{Q}, CAA^{Q}, and CAG^{Q} in ciliates and non-ciliate eukaryotes.

Results

SCDs provide versatile functionalities in proteins

We have shown previously in *S. cerevisiae* that Rad51-NTD autonomously promotes high-level production of native Rad51 and its COOH-terminal fusion protein LacZ (β-galactosidase) *in vivo* (1). To do so, in brief, we expressed Rad51-NTD-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 (1). We confirmed that the 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). Here, we further report that yeast Rad53-SCD1, Hop1-SCD, Sml1-$NTD^{1–50}$ (residues 1–50) and Sml1-$NTD^{1–27}$ (residues 1–27) also exhibit protein expression-enhancing (PEE) activities (Figure 1A and 1B, Table 1).

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 $T^{47}M$).

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.

Accordingly, 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 dysfunctional 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 (48), the Ure2 prion domain (UPD) (residues 1-91) (49, 50), and the New1 prion domain (NPD; residue 1-146) (51) (Figure 1A and 1B, Table S1). Sup35-PND containing 3S, 12Q, 18N and an $S^{17}Q$ motif exerts critical functions in

promoting [*PSI*+] prion nucleation (52). 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 (53). New1 is a non-essential ATP-binding cassette type F protein that fine-tunes the efficiency of translation termination or ribosome recycling (54). The NPD of New1 supports [*NU*$^{+}$] and is susceptible to [*PSI*+] prion induction (55, 56). 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 (1), removal of the NTD from Rad51 reduced by ~97% the 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* (Table S2) by N-terminal fusion of Rad53-SCD1, Rad53-SCD1-5STA (all five S/T-Q motifs changed to AQs) or Sup35-PND, respectively. Rad53-SCD1-5STA is a mutant protein defective in Mec1- and Tel1-mediated phosphorylation. Compared to wild-type yeast cells, the three corresponding yeast mutants (*rad53-SCD1-rad51-ΔN*, *rad53-SCD1-5STA-rad51-*

ΔN and *sup35-PND-rad51-ΔN*) 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 during DSB repair 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).

In conclusion, our results indicate that the quaternary structures of the target proteins (i.e., GFP, GSTnd, GST, LacZ and Rad51-ΔN) are irrelevant to the autonomous PEE activity. We assert that our use of a nuclear localization signal on the C-terminal VHN tag was unlikely to influence protein degradation kinetics or to sequester the reporter, leading to their accumulation and the appearance of enhanced expression for two reasons. First, the negative control LacZ-NV also possesses the same nuclear localization signal (Figure 1A, lane 2). Second, as an endogenous fusion target, Rad51-ΔN does not harbor the NVH tag (Figure 1D, lanes 3-4). Compared to wild-type Rad51, Rad51-ΔN is highly labile. In our previous study, removal of the NTD from Rad51 reduced by ~97% the protein levels of corresponding Rad51-ΔN proteins relative to wild-type (1).

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

The PEE function is unlikely to operate at the transcriptional level, as revealed by genomic and reverse-transcription quantitative polymerase chain reaction analyses (i.e., g-qPCR and RT-

qPCR, respectively) (Figure 2, Table S3 and DS1). We found that the 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, the 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 (57). 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 (58, 59). New1 is a translation factor that fine-tunes ribosome recycling and the efficiency of translation termination (54). Doa1 (also called Ufd3) is an ubiquitin- and Cdc48-binding protein with a role in ubiquitin homeostasis and/or protein degradation (60, 61). The *doa1*Δ mutant exhibits diminished formation of [*PSI*+] prion (62). Doa4 is a deubiquitinating enzyme required for recycling ubiquitin from proteasome-bound ubiquitinated intermediates (63). The *doa4*Δ mutant exhibits increased sensitivity to the protein synthesis inhibitor cycloheximide (64). San1 is an ubiquitin-protein ligase that targets highly aggregation-prone proteins (65, 66). Oaz1 (ornithine decarboxylase antizyme) stimulates ubiquitin-independent degradation of Spe1 ornithine decarboxylase by the proteasome (67). 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- to 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 the addition of Rad51-NTD to LacZ-NVH can abrogate the protein homeostasis defects caused by the loss of each of these six genes. For example, Rad51-NTD might compensate for the ribosome assembly and translation defects in *new1*Δ (54), as well as the cycloheximide-hypersensitive phenotype displayed by *doa4*Δ (64). 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 (62), the steady-state levels of Rad51-NTD-LacZ-NVH in the *doa1*Δ line are also slightly higher than those in WT.

The N-end rule is not likely relevant to the PEE function of Q-rich motifs

The N-end rule links the *in vivo* half-life of a protein to the identity of its N-terminal residues. In *S. cerevisiae*, the N-end rule operates as part of the ubiquitin system and comprises two pathways. First, the Arg/N-end rule pathway, involving a single N-terminal amidohydrolase Nta1, mediates deamidation of N-terminal asparagine (N) and glutamine (Q) into aspartate (D) and glutamate (E), which in turn are arginylated by a single Ate1 R-transferase, generating the Arg/N degron. N-terminal R and other primary degrons are recognized by a single N-recognin Ubr1 in concert with ubiquitin-conjugating Ubc2/Rad6. Ubr1 can also recognize several other N-terminal residues, including lysine (K), histidine (H), phenylalanine (F), tryptophan (W), leucine

(L) and isoleucine (I) (68-70). Second, the Ac/N-end rule pathway targets proteins containing N-terminally acetylated (Ac) residues. Prior to acetylation, the first amino acid methionine (M) is catalytically removed by Met-aminopeptidases (MetAPs), unless a residue at position 2 is non-permissive (too large) for MetAPs. If a retained N-terminal M or otherwise a valine (V), cysteine (C), alanine (A), serine (S) or threonine (T) residue is followed by residues that allow N-terminal acetylation, the proteins containing these AcN degrons are targeted for ubiquitylation and proteasome-mediated degradation by the Doa10 E3 ligase (71).

For two reasons, the PEE activities of these Q-rich domains are unlikely to arise from counteracting the N-end rule. First, the first two amino acid residues of Rad51-NTD, Hop1-SCD, Rad53-SCD1, Sup35-PND, Rad51-ΔN, and LacZ-NVH are MS, ME, ME, MS, ME, and MI, respectively, where M is methionine, S is serine, E is glutamic acid and I is isoleucine. Second, Sml1-NTD behaves similarly to these N-terminal fusion tags, despite its methionine and glutamine (MQ) amino acid signature at the N-terminus.

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

We applied an alanine scanning mutagenesis approach to reduce the percentages of S, T, Q or N in Rad51-NTD, Rad53-SCD1, and Sup35-NPD, respectively. These three Q-rich motifs exhibit a very strong positive relationship between STQ and STQN amino acid percentages and β-galactosidase activities (Figure 4 and Figure 5). IUPred2A (https://iupred2a.elte.hu/plot_new), a web-server for identifying disordered protein regions (72), also revealed that Rad51-NTD, Rad53-SCD1 and Sup35-NPD are structurally flexible peptides. 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 (4), and that a common feature of IDRs is their high content of S, T, Q, N, proline (P), glycine (G) and charged amino acids (17, 19, 20).

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. Given that IDRs can endow proteins with structural and functional plasticity (73, 74), we hypothesized that Q-rich motifs (e.g., SCD, polyQ and polyQ/N) represent useful toolkits for creating new diversity during protein evolution.

Comparative proteome-wide analyses of amino acid contents, SCDs and polyX motifs

Next, we designed four JavaScript software programs (AS-aa-content, AS-codon-usage, AS-Finder-SCD and AS-Finder-polyX) for proteome-wide analyses (Table S4). AS-aa-content and AS-codon-usage determine the proteome-wide average contents of 20 different amino acids and the proteome-wide usage frequency of 64 genetic codons, respectively. ASFinder-SCD and ASFinder-polyX were applied to search for amino acid sequences that contain ≥3 S/T-Q motifs within a stretch of ≤100 residues (5) and for 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 (39, 75-77). Based on a previous study (39), we applied seven different thresholds to seek both short and long, as well as pure and impure, polyX strings in 20 different representative near-complete proteomes, including 4X (4/4), 5X (4/5-5/5), 6X (4/6-6/6), 7X (4/7-7/7), 8-10X (≥50%X), 11-10X (≥50%X) and ≥21X (≥50%X). The lowest threshold was

≥4/7, i.e., a minimum number of four identical X amino acid residues in a localized region of seven amino acid residues (Figures 6, Figures S1-S3 and Datasets DS7-DS32).

We then searched and compared the near-complete proteomes of 26 different eukaryotes (Table 1), 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 slime 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 1). Genome or protein matrix scores >95% for model organisms are generally deemed complete reference genomes or proteomes (78).

It was reported previously that SCDs are overrepresented in the yeast and human proteomes (5, 79), and that polyX prevalence differs among species (43, 80-82). 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 DS3). Among the 6050 proteins in the most recently updated *S. cerevisiae* reference proteome (https://www.uniprot.org/proteomes/UP000002311), we identified 1016 SCD-hosting proteins

(DS3), including all 436 SCD-harboring proteins previously revealed by ScanProsite (5). ScanProsite is a publicly available database of protein families, domains and motifs (83).

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 unicellular ciliates (i.e., *T. thermophila*, *P. tetraurelia*, *O. trifallax*, *S. lemnae* and *P. persalinus*) than in eukaryotes with standard genetic codons, including the unicellular ciliate *S. coeruleus* and all of the 20 non-ciliate species we examined herein (Figures S1 and DS2). Hereafter, we refer to the five unicellular ciliates with reassigned stop codons as 'group I' ciliates. Due to higher proteome-wide Q contents, there are higher percentages of SCD, polyQ and polyQ/N in the five group I ciliates than in *S. coeruleus* (Figures 6, S2, S3, and DS2-DS4).

Next, we analyzed 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) (84), which were then reassembled and reannotated by Brown and colleagues (85). All transcripts are publicly available from Zendo (https://zenodo.org/record/1212585#.Y79zoi2l3PA) (85). We applied TransDecoder (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 *Mesodinium pulex*, *Blepharisma japonicum* and *Euplotes focardii*, each of which utilizes the reassigned codons UAA^Y, UAG^Y, UGA^W and UGA^C, respectively (Table 1). For two reasons, the proteins encoded by these

MMESTP transcripts are unlikely to represent the entire protein complement of all 11 ciliate species. First, many MMETSP transcripts are not intact (i.e., broken mRNAs) and thus encode incomplete protein sequences. Second, 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 S1 and DS2). 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. Since N is used more frequently in *L. pictus* than the other ciliates in groups II-IV, it has higher percentages of polyN and polyQ/N proteins (Figures 6, S2, S3, and DS2-DS4). 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 S1 and DS2). 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 S3).

To further confirm the above-described results, we normalized the runs of amino acids and created a null expectation from each proteome by determining the ratios of the overall number of X residues for each of the seven polyX motifs relative to those in the entire proteome of each species, respectively. The results for four different polyX motifs, i.e., polyQ, polyN, polyS and polyT, are presented in Figures 7-10 and DS5. The results summarized in Figures 7-10 support that polyX prevalence differs among species and that the overall X contents of polyX motifs often but not always correlate with the X usage frequencies in entire proteomes (43). Most importantly, our results reveal that, compared to *S. coeruleus* or several non-ciliate eukaryotic organisms (e.g., *P. falciparum, C. elegans*, *D. rerio*, *M. musculus* and *H. sapiens*), the

five ciliates with reassigned TAA^{Q} and TAG^{Q} codons not only have higher Q usage frequency but also more polyQ motifs in their proteomes (Figure 7 and DS5). In contrast, polyQ motifs prevail in *C. albicans, C. tropicalis, D. discoideum, C. reinhardtii, D. melanogaster* and *A. aegypti*, though the Q usage frequencies in their entire proteomes are not significantly higher than those of other eukaryotes (Figure 7 and DS5). Due to their higher N usage frequencies, *D. discoideum, P. falciparum* and *P. persalinus* have more polyN motifs than the other 23 eukaryotes we examined here (Figure 8 and DS5). Generally speaking, all 26 eukaryotes we assessed have similar S usage frequencies and percentages of S contents in polyS motifs (Figure 9 and DS5). Among these 26 eukaryotes, *D. discoideum* possesses many more polyT motifs, though its T usage frequency is similar to that of the other 25 eukaryotes (Figure 10 and DS5). Several other polyX motifs are particularly enriched in specific eukaryotes, for example, polyK and polyY in *P. falciparum*, polyK and polyF in *D. discoideum*, polyG, polyA, polyP and polyW in *C. reinhardtii*, as well as the longest polyC (i.e., ≥21C and ≥50%C) in *C. tropicalis* (DS5). Further investigations will decipher the structural and functional relevance of those polyX motif proteins. In conclusion, these normalized results further confirm that reassignment of stop codons to Q indeed results in both higher Q usage frequencies and more polyQ motifs in ciliates.

The frequency of TAA^{Q} and TAG^{Q}, CAA^{Q} and CAG^{Q} usage in 26 different organisms

PolyQ motifs have a particular length-dependent codon usage that relates to strand slippage in CAG/CTG trinucleotide repeat regions during DNA replication (37, 39). In most organisms having standard genetic codons, Q is encoded by CAG^{Q} and CAA^{Q}. We applied AS-Xcontent-7polyX, a JavaScript software program (Table S4), to determine and compare proteome-wide Q contents, as well as CAG^{Q} usage frequencies (i.e., the ratio between CAG^{Q} and

the sum of CAG^Q, CAG^Q, TAA^Q, and TAG^Q) (Table 2 and DS6). Our results reveal that the likelihood of forming long CAG/CTG trinucleotide repeats is higher in five eukaryotes due to their higher CAG^Q usage frequencies, including *D. melanogaster* (86.6% Q), *D. rerio* (74.0% Q), *M. musculus* (74.0% Q), *H. sapiens* (73.5% Q), and *C. reinhardtii* (87.3% Q) (orange background, Table 2). In contrast, another five eukaryotes that possess high numbers of polyQ motifs (i.e., *D. discoideum, C. albicans*, *C. tropicalis, P. falciparum* and *S. coeruleus*) (Figure 7) utilize more CAA^Q (96.2%, 84.6%, 84.5%, 86.7% and 75.7%) than CAG^Q (3.8%, 15.4%, 15.5%, 13.3% and 24.3%), respectively, to avoid forming long CAG/CTG trinucleotide repeats (green background, Table 2). Similarly, all five ciliates with reassigned stop codons (TAA^Q and TAG^Q) display low CAG^Q usage frequencies (i.e., ranging from 3.8% Q in *P. persalinus* to 12.6% Q in *O. trifallax*) (Table 2). Accordingly, the CAG-slippage mechanism may operate more frequently in *C. reinhardtii*, *D. melanogaster*, *D. rerio*, *M. musculus* and *H. sapiens* than in *D. discoideum, C. albicans*, *C. tropicalis*, *P. falciparum*, *S. coeruleus* and the five ciliates with reassigned stop codons (TAA^Q and TAG^Q).

Q-rich-motif proteins are overrepresented in specialized biological processes of various eukaryotic proteomes

To determine the biological impacts of Q-rich-motif proteins, we designed a JavaScript software tool AS-GOfuncR-FWER (Table S4) to carry out comparative Gene Ontology (GO) enrichment analyses using information on the functions of genes provided by the GO knowledgebase (http://geneontology.org). Rigorous statistical testing for overrepresentation or underrepresentation of SCD and polyX proteins was performed using GOfuncR (https://bioconductor.org/packages/release/bioc/html/GOfuncR.html), an R package program that conducts standard candidate vs. background enrichment analysis employing the hypergeometric

test. The raw p-values were adjusted according to the Family-Wise Error Rate (FWER). The same method was applied to the GO enrichment analysis of human genomes (86). The results presented in Figure 11 and Figure 12 (DS7-DS32) support the hypothesis that Q-rich motifs prevail in proteins involved in specialized biological processes, including *S. cerevisiae* RNA-mediated transposition, *C. albicans* filamentous growth, peptidyl-glutamic acid modification in ciliates with reassigned stop codons (TAA^Q and TAG^Q), *T. thermophila* xylan catabolism, *D. discoideum* sexual reproduction, *P. falciparum* infection, as well as the nervous systems of *D. melanogaster, M. musculus,* and *H. sapiens*. In contrast, peptidyl-glutamic acid modification is not overrepresented with Q-rich-motif proteins in *S. coeruleus,* a ciliate with standard stop codons.

Our results are also consistent with a previous report that there is an overrepresentation of conserved Q-rich-motif proteins in pathways related to the human nervous system (79). 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 (87, 88). 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 a high amino acid identity with the N-terminus (1-640 amino acid residues) of human CTTNBP2. IUPred2A (https://iupred2a.elte.hu/plot_new) also reveals 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 (89).

Q-rich-motif proteins prevail in the *T. thermophila* xylan catabolic process

The proteome of *T. thermophila* contains 58 proteins involved in the xylan catabolic process (GO ID: 45493), of which 56 (97%), 55 (95%), 58 (100%), and 49 (84%) proteins harbor SCD, polyQ, polyQ/N, and polyN tracts, respectively (DS19-DS24). Using the NCBI BLASTP search tool with an expect value (E-value) ≤ 10e-5 to search for homologs of these 58 proteins among all the other 16 ciliates analyzed in this study, we only identified 144 proteins with amino acid identity >60% and a raw alignment score of >150 (Dataset DS33). Thus, *T. thermophila* has more abundant xylan catabolic proteins than all other ciliates we examined herein.

Most proteins involved in *T. thermophila* meiosis harbor one or more Q-rich motif(s)

Ciliate meiosis is remarkable relative to that of other studied sexual eukaryotes. 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. The MAC is generated from the MIC by massive amplification, editing and rearrangement of the genome (see reviews in (90, 91)). In *T. 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 (92)). 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

(93). Meiosis-specific CYC2 and CYC17 cyclins, as well as cyclin-dependent kinase CDK3, are required to initiate meiosis and for crescent assembly (94-96). CYC2/CDK2 promotes bouquet formation in MICs by controlling microtubule-directed elongation (96) and it also controls the gene expression of proteins involved in DSB formation (SPO11), DNA repair (COM1, EXO1, DMC1), and crossover formation (HOP2, MND1, MSH4, MSH5, ZPH3, BIM1, and BIM2) (97). 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 (97). Nevertheless, the molecular mechanisms underlying crescent assembly and disassembly remain poorly understood.

Among the 124 *T. thermophila* meiotic proteins (DS34, see review in (92)), 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 contain 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 (97). 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 (98). 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 (99). GIW1 (gentlemen-in-waiting 1) physically directs a mature Argonaute-siRNA complex to the MIC nucleus, thus promoting programmed IES elimination (100). GIW1 has 9 SCDs and 2 polyQ/N tracts (DS34).

Using IUPred2A (https://iupred2a.elte.hu/plot_new), we found that the Q-rich motifs in most (if not all) of these meiotic proteins are intrinsically disordered. Accordingly, we 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.

Discussion

We present three 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. Further investigations would prove illuminating as to 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, the reassignment of stop codons to Q in the group I and group II ciliates significantly increases 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 (17, 19, 20). Third, the results in Table 2 support that a decrease or increase of CAG^{Q} usage frequency in different eukaryotes is responsible for a reduction or augmentation, respectively, of polyQ instability (or expansion) caused by DNA strand slippage in CAG/CTG trinucleotide repeat regions during DNA replication (37, 39). Accordingly, it would be interesting to decipher how the molecular mechanism(s) that controls the codon usage bias of TAA^{Q}, TAG^{Q}, CAA^{Q} and CAG^{Q} evolved.

Due to their structural flexibility, Q-rich motifs can endow proteins with structural and functional plasticity. Based on previous reports (73, 74) and our findings from this study, Q-rich motifs, such as IDRs, 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 DS33). Second, our GO enrichment results reveal that Q-rich motifs prevail in proteins involved in specialized biological processes (Figure 11 and Figure 12). 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, among other processes (21, 22). A typical example is that the C-terminus of mouse CTTNBP2 facilitates co-condensation of CTTNBP2 with SHANK3 at the postsynaptic density in a Zn^{2+}-dependent manner (89). Third, Borgs are long and linear extrachromosomal DNA sequences in methane-oxidizing *Methanoperedens archaea*, which display the potential to augment methane oxidation (101). A striking feature of Borgs is pervasive tandem direct repeat (TR) regions. TRs in open reading frames (ORFs) are under very strong selective pressure, leading to perfect amino acid TRs (aaTRs) that are commonly IDRs. Notably, aaTRs often contain disorder-promoting amino acids, including Q, N, P, T, E, K, V, D and S (102). Accordingly, distinct evolutionary strategies are employed in different species to alter 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). The impacts of such genome-wide alternations had not been well understood. Here, 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 the massive expansion of proteins harboring

structurally flexible or even intrinsically disordered Q-rich motifs. Since Q-rich motifs can endow proteins with structural and functional plasticity and Q-rich-motif proteins are overrepresented in several species-specific or even phylum-specific biological processes, we suggest that Q-rich motifs are useful toolkits for evolutionary novelty.

5 **Methods**

All plasmids, yeast strains, and PCR primers used in this study are listed in Tables S1-S3, 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 (1, 11). The mouse anti-V5 antibody was purchased from BioRad (CA, USA). The
10 rabbit anti-Hsp104 antiserum was kindly provided by Chung Wang (Institute of Molecular Biology, 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
15 described previously (1, 11). Quantitative β-galactosidase activity assays were carried out as previously described (1, 103). The sources of proteome and transcript datasets are described in Table 1. All six JavaScript software programs used in this study are listed in Table S4 and publicly available on Github (https://github.com/tfwangasimb/AS-Q-rich-motif). The GO enrichment analyses were performed using publicly available data in the GO Resource
20 (http://geneontology.org). The GO identities (ID) of different biological processes, cellular components, and molecular functions, as well as the names of all SCD and polyX proteins in the 26 near-completed eukaryotic proteomes, are listed in Dataset DS7-DS32, respectively.

GOfuncR was applied for rigorous statistical testing by conducting standard candidate vs. background enrichment analysis using the hypergeometric test. The raw p-values of over-represented and under-represented GO groups were adjusted according to the Family-wise error rate (FWER).

Supporting Datasets

Dataset DS1: The raw qPCR data of cDNA and gDNA in Figure 2.

Dataset DS2: The average usages of 20 different amino acids in 17 ciliate and 20 non-ciliate species.

Dataset DS3: The number of proteins containing different types of polyQ, polyQ/N and polyN tracts in 17 ciliate and 20 non-ciliate species.

Dataset DS4: The numbers and percentages of SCD and polyX proteins in 17 ciliate and 20 non-ciliate species.

Dataset DS5: The ratios of the overall number of X residues for each of the seven polyX motifs relative to those in the entire proteome of each species, respectively.

Dataset DS6: The codon usage frequency in 26 near-complete proteomes and 11 ciliate proteomes encoded by the transcripts generated as part of the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP).

Dataset DS7-DS32: GO enrichment analyses revealing the SCD and polyX proteins involved in different biological processes in 6 ciliate and 20 non-ciliate species. The percentages and numbers of SCD and polyX proteins in our search that belong to each indicated Gene Ontology (GO) group are shown. GOfuncR (86) was applied for GO enrichment and statistical analysis. The *p* values adjusted according to the Family-wise error rate (FWER) are shown. The overrepresented GO groups (adjusted *p* values ≤0.001) are highlighted in red font.

Dataset DS33: The results of BLASTP searches using the 58 *Tetrahymena thermophila* proteins involved in xylan catabolysis.

Dataset DS34: The list of 124 *Tetrahymena thermophila* proteins involved in meiosis (kindly provided by Josef Loidl). The numbers of SCD and polyX tracts in each protein are indicated.

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 are listed in the supporting information (Tables S1-S4 and Figures S1-S3).

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 legends:

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 (1). Western blots for visualization of LacZ-NVH fusion proteins (A) and quantitative β-galactosidase assays (B) were carried out as described previously (1). 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).

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 intrinsically disordered regions (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 using 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. Q contents in 7 different types of polyQ motifs in 20 near-complete proteomes. The five ciliates with reassigned stops codon (TAA^Q and TAG^Q) are indicated in red. *Stentor coeruleus,* a ciliate with standard stop codons, is indicated in green.

Figure 8. N contents in 7 different types of polyN motifs in 20 near-complete proteomes. The five ciliates with reassigned stops codon (TAA^Q and TAG^Q) are indicated in red. *Stentor coeruleus,* a ciliate with standard stop codons, is indicated in green.

Figure 9. S contents in 7 different types of polyS motifs in 20 near-complete proteomes. The five ciliates with reassigned stops codon (TAA^Q and TAG^Q) are indicated in red. *Stentor coeruleus,* a ciliate with standard stop codons, is indicated in green.

Figure 10. T contents in 7 different types of polyT motifs in 20 near-complete proteomes. The five ciliates with reassigned stops codon (TAA^Q and TAG^Q) are indicated in red. *Stentor coeruleus,* a ciliate with standard stop codons, is indicated in green.

Figure 11. Selection of biological processes with overrepresented SCD-containing proteins in different eukaryotes. The percentages and number of SCD-containing proteins in our search that belong to each indicated Gene Ontology (GO) group are shown. GOfuncR (86) was applied for GO enrichment and statistical analysis. The p values adjusted according to the Family-wise error rate (FWER) are shown.

Figure 12. Selection of biological processes with overrepresented polyQ-containing proteins in different eukaryotes. The percentages and numbers of polyQ-containing proteins in our search that belong to each indicated Gene Ontology (GO) group are shown. GOfuncR (86) was applied for GO enrichment and statistical analysis. The *p* values adjusted according to the Family-wise error rate (FWER) are shown. The five ciliates with reassigned stops codons (TAA^Q and TAG^Q) are indicated in red. *Stentor coeruleus,* a ciliate with standard stop codons, is indicated in green.

Table 1. 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	UniProt	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	UniProt	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	UniProt	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	UniProt	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	UniProt	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	UniProt	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	UniProt	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	UniProt	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	UniProt	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	UniProt	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	UniProt	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	UniProt	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	UniProt	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	UniProt	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	UniProt	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	UniProt	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	UniProt	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	UniProt	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	UniProt	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	UniProt	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	UniProt	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	UniProt	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	UniProt	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 ciliates				→ 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 ciliates				*	*	*	Q	Q	C	C	Y	Y	W
Stentor coeruleus	UniProt	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 2. Usage frequencies of TAA*, TAG*, TAA^{Q}, TAG^{Q}, CAA^{Q} and CAG^{Q} codons in the entire proteomes of 20 different organisms.

Species	CAA	CAG	TAA	TAG
Saccharomyces cerevisiae S288c	2.73 (62.6%Q)	1.21 (37.4%Q)	0.11	0.05
Candida albicans	3.57 (84.6%Q)	0.65 (15.4%Q)	0.1	0.05
Candida auris	1.81 (46.1%Q)	2.12 (53.9%Q)	0.08	0.06
Candida tropicalis	3.61 (84.5%Q)	0.66 (15.5%Q)	0.1	0.07
Neurospora crassa	1.70 (39.5%Q)	2.60 (60.5%Q)	0.06	0.05
Magnaporthe oryzae	1.37 (33.7%Q)	2.69 (66.3%Q)	0.06	0.07
Trichoderma reesei	1.17 (28.4%Q)	2.95 (71.6%Q)	0.06	0.06
Cryptococcus neoformans	2.06 (53.5%Q)	1.79 (46.5%Q)	0.07	0.06
Ustilago maydis	1.82 (41.3%Q)	2.61 (58.7%Q)	0.04	0.05
Taiwanofungus camphoratus	1.57 (41.8%Q)	2.19 (58.2%Q)	0.05	0.06
Dictyostelium discoideum	4.86 (96.2%Q)	0.19 (3.8%Q)	0.16	0.01
Plasmodium falciparum	2.42 (86.7%Q)	0.37 (13.3%Q)	0.09	0.01
Drosophila melanogaster	1.56 (13.4%Q)	3.61 (86.6%Q)	0.08	0.07
Aedes aegypti	1.76 (40.6%Q)	2.58 (59.4%Q)	0.11	0.07
Caenorhabditis elegans	2.74 (65.6%Q)	1.44 (34.4%Q)	0.16	0.06
Danio rerio	1.18 (26.0%Q)	3.35 (74.0%Q)	0.11	0.06
Mus musculus	1.20 (26.0%Q)	3.41 (74.0%Q)	0.1	0.08
Homo sapiens	1.23 (26.5%Q)	3.42 (73.5%Q)	0.1	0.08
Arabidopsis thaliana	1.94 (56.1%Q)	1.52 (43.9%Q)	0.09	0.05
Chlamydomonas reinhardtii	0.59 (12.7%Q)	4.05 (87.3%Q)	0.03	0.04
Tetrahymena thermophila	2.04 (21.2%Q)	0.48 (5.0%Q)	5.46 (56.8%Q)	1.63 (17.0%Q)
Paramecium tetraurelia	2.54 (27.9%Q)	0.57 (6.3%Q)	4.53 (46.7%Q)	1.48 (16.2%Q)
Oxytricha trifallax	2.68 (29.9%Q)	1.07 (12.0%Q)	3.63 (40.6%Q)	1.57 (17.5%Q)
Stylonychia lemnae	2.26 (21.1%Q)	1.05 (12.6%Q)	3.22 (38.6%Q)	1.81 (21.7%Q)
Pseudocohnilembus persalinus	1.76 (18.0%Q)	0.37 (3.8%Q)	7.36 (76.0%Q)	1.39 (14.4%Q)
Stentor coeruleus	2.77 (75.7%Q)	0.89 (24.3%Q)	0.16	0.08

References

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. Zhao X, Rothstein R. The Dun1 checkpoint kinase phosphorylates and regulates the ribonucleotide reductase inhibitor Sml1. Proc Natl Acad Sci U S A. 2002;99(6):3746-51.
13. Uchiki T, Dice LT, Hettich RL, Dealwis C. Identification of phosphorylation sites on the yeast ribonucleotide reductase inhibitor Sml1. J Biol Chem. 2004;279(12):11293-303.
14. Andreson BL, Gupta A, Georgieva BP, Rothstein R. The ribonucleotide reductase inhibitor, Sml1, is sequentially phosphorylated, ubiquitylated and degraded in response to DNA damage. Nucleic Acids Res. 2010;38(19):6490-501.
15. Niu H, Wan L, Busygina V, Kwon Y, Allen JA, Li X, et al. Regulation of meiotic recombination via Mek1-mediated Rad54 phosphorylation. Mol Cell. 2009;36(3):393-404.
16. Callender TL, Laureau R, Wan L, Chen X, Sandhu R, Laljee S, et al. Mek1 down regulates Rad51 activity during yeast meiosis by phosphorylation of Hed1. PLoS Genet. 2016;12(8):e1006226.

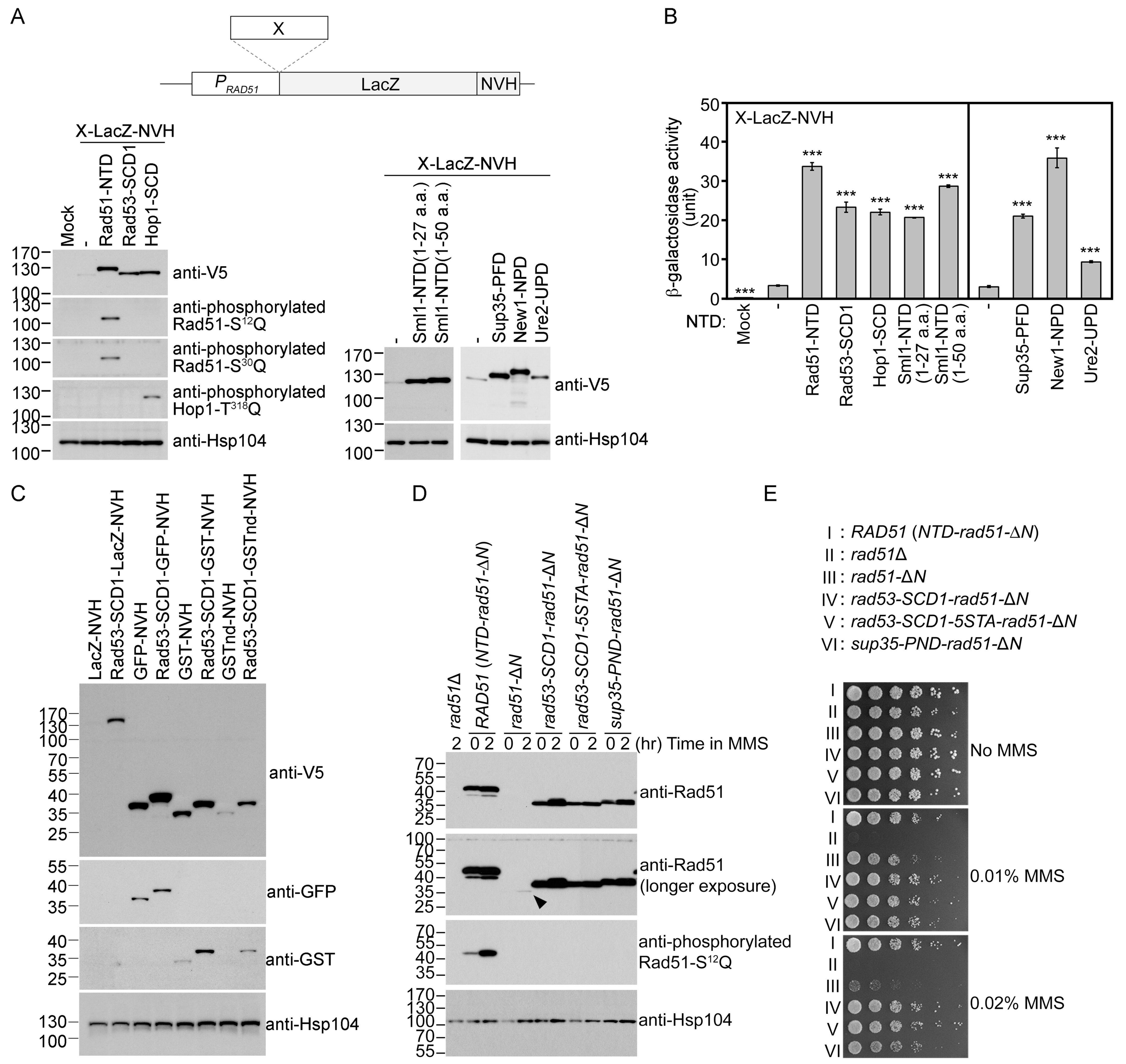
17. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. Proteins. 2001;42(1):38-48.
18. Uversky VN. Intrinsically disordered proteins and their "mysterious" (meta)physics. Frontiers in Physics. 2019;7:10.
19. Macossay-Castillo M, Marvelli G, Guharoy 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.
20. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins. 2000;41(3):415-27.
21. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol. 1999;293(2):321-31.
22. Posey AE, Holehouse AS, Pappu RV. Phase separation of intrinsically disordered proteins. . Methods Enzymol. 2018;611:1-30.
23. Santner AA, Croy CH, Vasanwala FH, Uversky VN, Van YY, Dunker AK. Sweeping away protein aggregation with entropic bristles: intrinsically disordered protein fusions enhance soluble expression. Biochemistry. 2012;51(37):7250-62.
24. Caron F, Meyer E. Does *Paramecium primaurelia* use a different genetic code in its macronucleus? Nature. 1985;314(6007):185-8.
25. 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.
26. 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.
27. 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.
28. Lozupone CA, Knight RD, Landweber LF. The molecular basis of nuclear genetic code change in ciliates. Curr Biol. 2001;11(2):65-74.
29. 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.
30. 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.
31. 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.
32. 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.
33. 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.
34. 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.
35. 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.

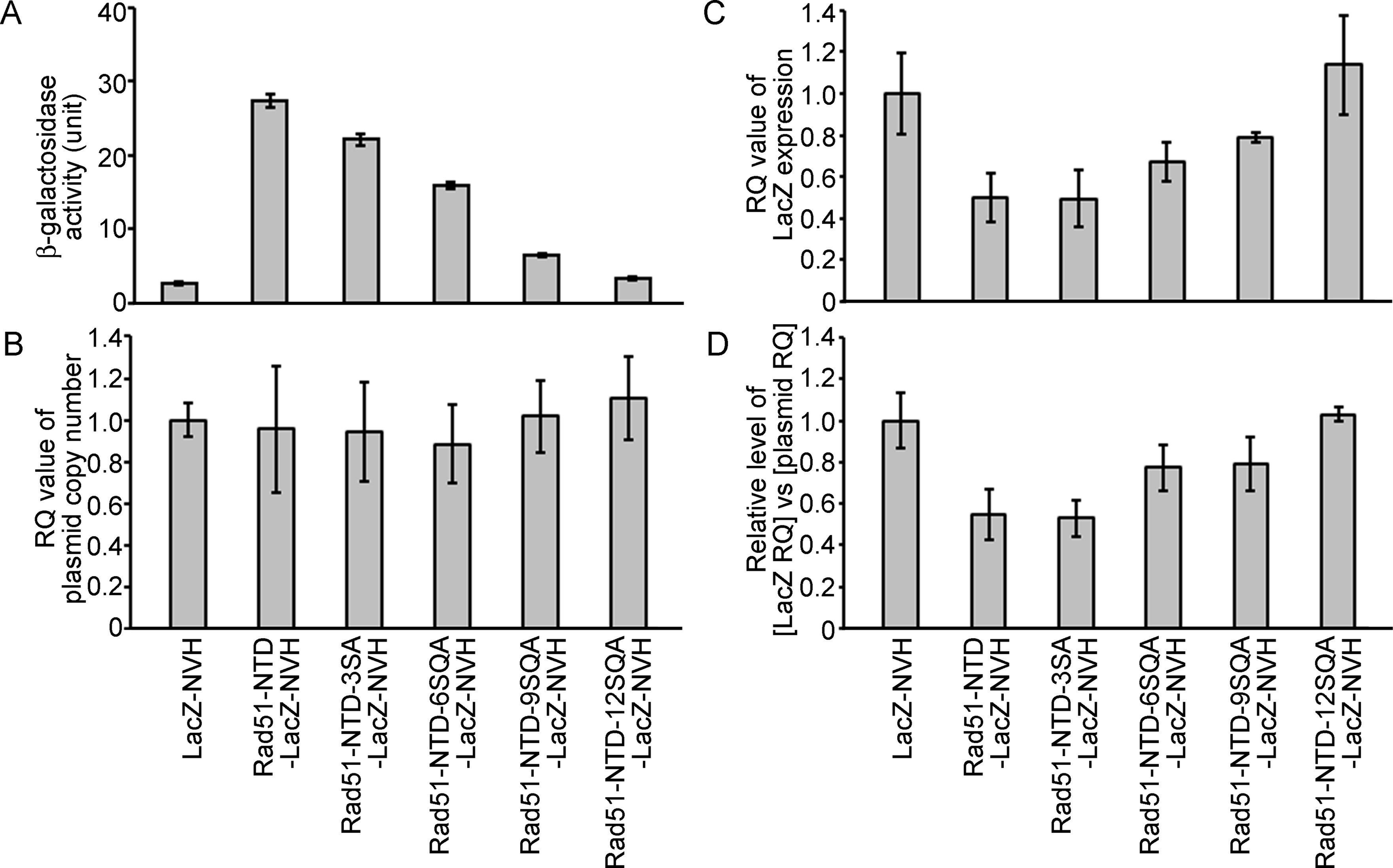
36. Kollmar M, Muhlhausen S. Nuclear codon reassignments in the genomics era and mechanisms behind their evolution. Bioessays. 2017;39(5).
37. Petruska J, Hartenstine MJ, Goodman MF. Analysis of strand slippage in DNA polymerase expansions of CAG/CTG triplet repeats associated with neurodegenerative disease. J Biol Chem. 1998;273(9):5204-10.
38. Ruff KM, Warner JB, Posey AE, Tan PS, Lemke EA, Pappu RV, et al. Polyglutamine length dependent structural properties and phase behavior of Huntingtin exon 1. Biophysical Journal. 2017;112(3):511a-a.
39. Mier P, Andrade-Navarro MA. Between interactions and aggregates: The polyQ balance. Genome Biol Evol. 2021;13(11).
40. Nalavade R, Griesche N, Ryan DP, Hildebrand S, Krauss S. Mechanisms of RNA-induced toxicity in CAG repeat disorders. Cell Death Dis. 2013;4(8):e752.
41. Li LB, Yu Z, Teng X, Bonini NM. RNA toxicity is a component of ataxin-3 degeneration in Drosophila. Nature. 2008;453(7198):1107-11.
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. Tuite MF. Yeast prions and their prion-forming domain. Cell. 2000;100(3):289-92.
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. 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.
53. 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.
54. 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.
55. Santoso A, Chien P, Osherovich LZ, Weissman JS. Molecular basis of a yeast prion species barrier. Cell. 2000;100(2):277-88.

56. 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.
57. Chen B, Retzlaff M, Roos T, Frydman J. Cellular strategies of protein quality control. Cold Spring Harb Perspect Biol. 2011;3(8):a004374.
58. Shorter J, Southworth DR. Spiraling in Control: Structures and mechanisms of the Hsp104 disaggregase. Cold Spring Harb Perspect Biol. 2019;11(8).
59. 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.
60. 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.
61. 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.
62. Tyedmers J, Madariaga ML, Lindquist S. Prion switching in response to environmental stress. PLoS Biol. 2008;6(11):e294.
63. 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.
64. 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.
65. Dasgupta A, Ramsey KL, Smith JS, Auble DT. Sir Antagonist 1 (San1) is a ubiquitin ligase. J Biol Chem. 2004;279(26):26830-8.
66. 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.
67. 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.
68. Bachmair A, Finley D, Varshavsky A. In vivo half-life of a protein is a function of its amino-terminal residue. Science. 1986;234(4773):179-86.
69. Tasaki T, Sriram SM, Park KS, Kwon YT. The N-end rule pathway. Annu Rev Biochem. 2012;81:261-89.
70. Varshavsky A. N-degron and C-degron pathways of protein degradation. Proc Natl Acad Sci U S A. 2019;116(2):358-66.
71. Hwang CS, Shemorry A, Auerbach D, Varshavsky A. The N-end rule pathway is mediated by a complex of the RING-type Ubr1 and HECT-type Ufd4 ubiquitin ligases. Nat Cell Biol. 2010;12(12):1177-85.
72. 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.
73. Zhou J, Oldfield CJ, Yan W, Shen B, Dunker AK. Intrinsically disordered domains: Sequence ➔ disorder ➔ function relationships. Protein Sci. 2019;28(9):1652-63.
74. Bondos SE, Dunker AK, Uversky VN. Intrinsically disordered proteins play diverse roles in cell signaling. Cell Commun Signal. 2022;20(1):20.

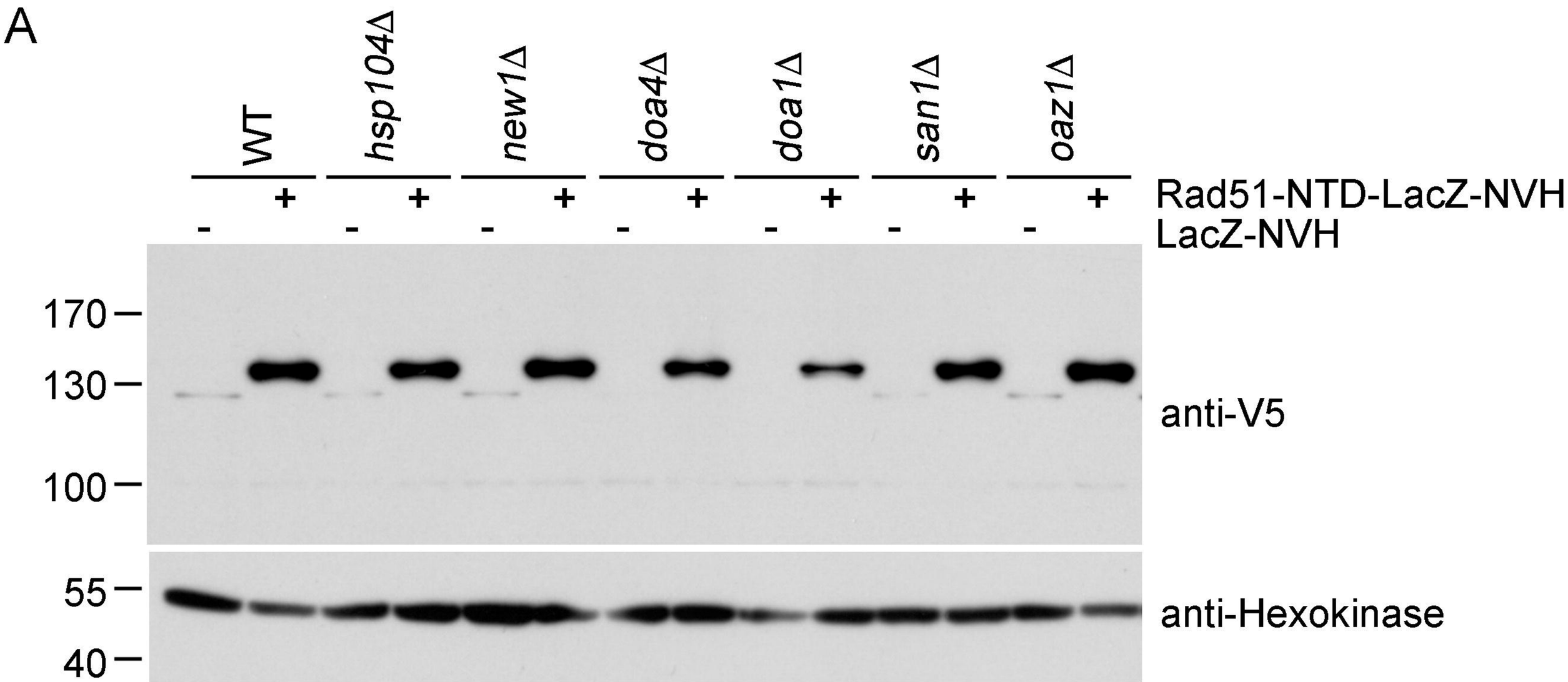
75. Ramazzotti M, Monsellier E, Kamoun C, Degl'Innocenti D, Melki R. Polyglutamine repeats are associated to specific sequence biases that are conserved among eukaryotes. PLoS One. 2012;7(2):e30824.
76. Li C, Nagel J, Androulakis S, Song J, Buckle AM. PolyQ 2.0: an improved version of PolyQ, a database of human polyglutamine proteins. Database (Oxford). 2016;2016.
77. Totzeck F, Andrade-Navarro MA, Mier P. The protein structure context of polyQ regions. PLoS One. 2017;12(1):e0170801.
78. Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness. Methods Mol Biol. 2019;1962:227-45.
79. 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.
80. Kuspa A, Loomis WF. The genome of *Dictyostelium discoideum*. Methods Mol Biol. 2006;346:15-30.
81. Davies HM, Nofal SD, McLaughlin EJ, Osborne AR. Repetitive sequences in malaria parasite proteins. FEMS Microbiol Rev. 2017;41(6):923-40.
82. 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.
83. 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.
84. 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.
85. Johnson LK, Alexander H, Brown CT. Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes. Gigascience. 2019;8(4).
86. Huttenhower C, Haley EM, Hibbs MA, Dumeaux V, Barrett DR, Coller HA, et al. Exploring the human genome with functional maps. Genome Res. 2009;19(6):1093-106.
87. 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.
88. 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.
89. 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.
90. Prescott DM. The DNA of ciliated protozoa. Microbiol Rev. 1994;58(2):233-67.
91. Chalker DL, Yao MC. DNA elimination in ciliates: transposon domestication and genome surveillance. Annu Rev Genet. 2011;45:227-46.
92. Loidl J. *Tetrahymena* meiosis: Simple yet ingenious. PLoS Genet. 2021;17(7):e1009627.
93. 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.
94. 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.

95. 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.
96. 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.
97. 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.
98. 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.
99. 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.
100. 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.
101. Al-Shayeb B, Schoelmerich MC, West-Roberts J, Valentin-Alvarado LE, Sachdeva R, Mullen S, et al. Borgs are giant genetic elements with potential to expand metabolic capacity. Nature. 2022;610(7933):731-6.
102. 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.
103. 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.

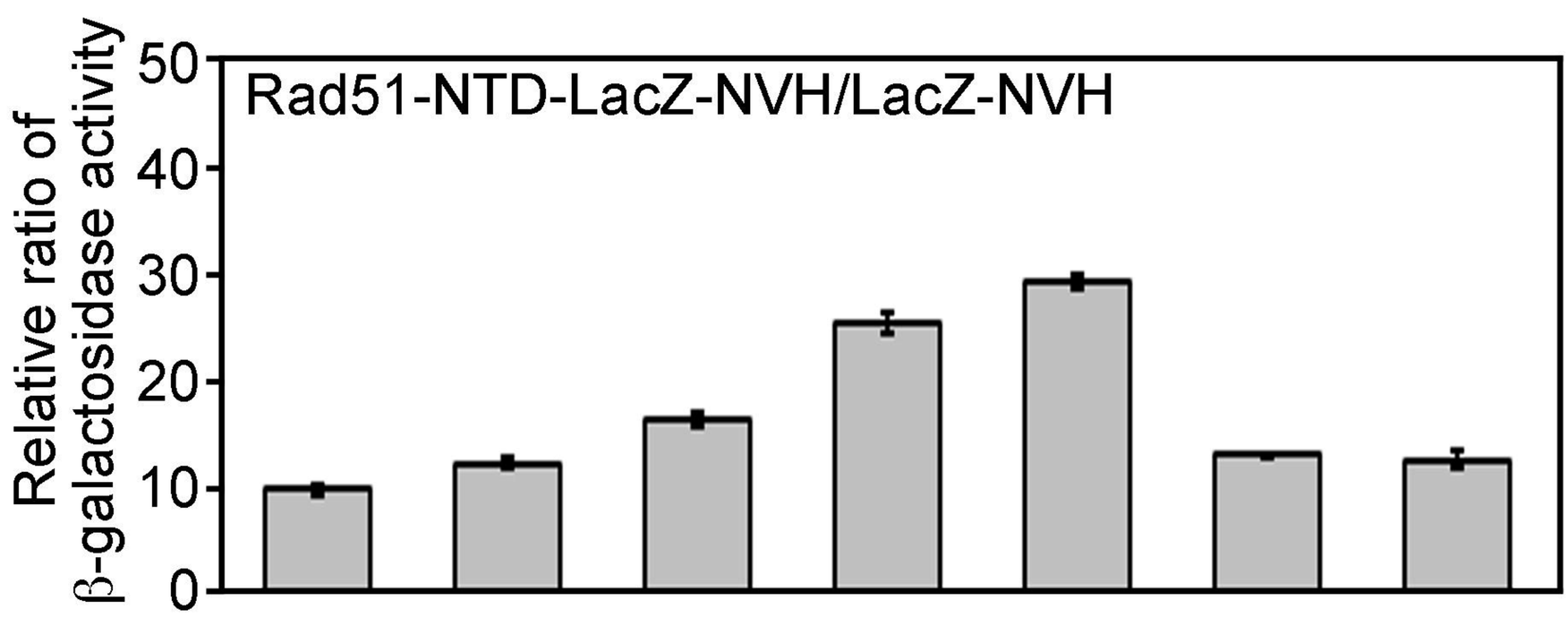

A
X
P_{RAD51}
LacZ
NVH
X-LacZ-NVH
Mock
-
Rad51-NTD
Rad53-SCD1
Hop1-SCD
anti-V5
anti-phosphorylated Rad51-S12Q
anti-phosphorylated Rad51-S30Q
anti-phosphorylated Hop1-T318Q
anti-Hsp104
Sml1-NTD(1-27 a.a.)
Sml1-NTD(1-50 a.a.)
Sup35-PFD
New1-NPD
Ure2-UPD
B
β-galactosidase activity (unit)
NTD:
C
LacZ-NVH
Rad53-SCD1-LacZ-NVH
GFP-NVH
Rad53-SCD1-GFP-NVH
GST-NVH
Rad53-SCD1-GST-NVH
GSTnd-NVH
Rad53-SCD1-GSTnd-NVH
anti-GFP
anti-GST
D
rad51Δ
RAD51 (NTD-rad51-ΔN)
rad51-ΔN
rad53-SCD1-rad51-ΔN
rad53-SCD1-5STA-rad51-ΔN
sup35-PND-rad51-ΔN
(hr) Time in MMS
anti-Rad51
anti-Rad51 (longer exposure)
E
I : RAD51 (NTD-rad51-ΔN)
II : rad51Δ
III : rad51-ΔN
IV : rad53-SCD1-rad51-ΔN
V : rad53-SCD1-5STA-rad51-ΔN
VI : sup35-PND-rad51-ΔN
No MMS
0.01% MMS
0.02% MMS


A
β-galactosidase activity (unit)
0
10
20
30
40
B
RQ value of plasmid copy number
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
C
RQ value of LacZ expression
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
D
Relative level of [LacZ RQ] vs [plasmid RQ]
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
LacZ-NVH
Rad51-NTD -LacZ-NVH
Rad51-NTD-3SA -LacZ-NVH
Rad51-NTD-6SQA -LacZ-NVH
Rad51-NTD-9SQA -LacZ-NVH
Rad51-NTD-12SQA -LacZ-NVH

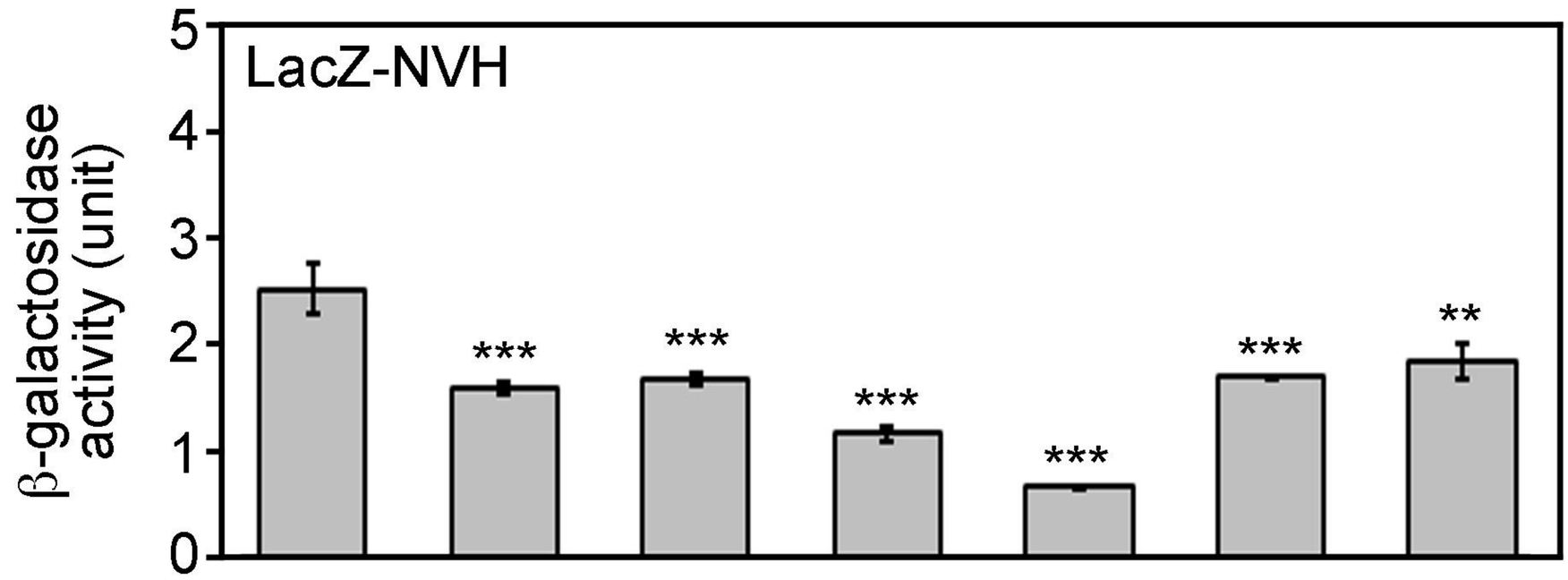
A


WT
hsp104Δ
new1Δ
doa4Δ
doa1Δ
san1Δ
oaz1Δ
Rad51-NTD-LacZ-NVH
LacZ-NVH
170
130
100
anti-V5
55
40
anti-Hexokinase

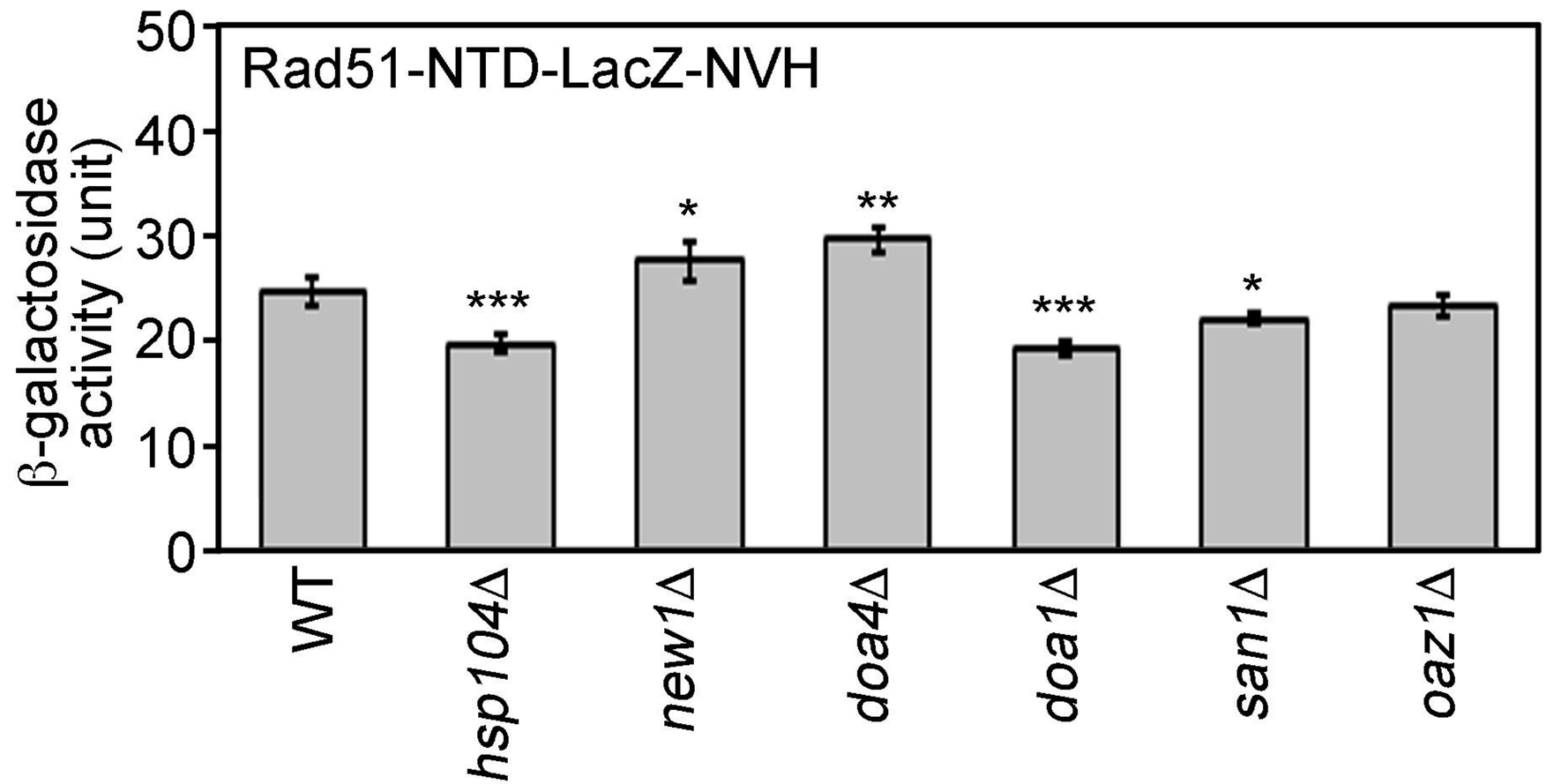
B


Rad51-NTD-LacZ-NVH/LacZ-NVH
Relative ratio of β-galactosidase activity
50
40
30
20
10
0

C


LacZ-NVH
β-galactosidase activity (unit)
5
4
3
2
1
0
***
***
***
***
***
**

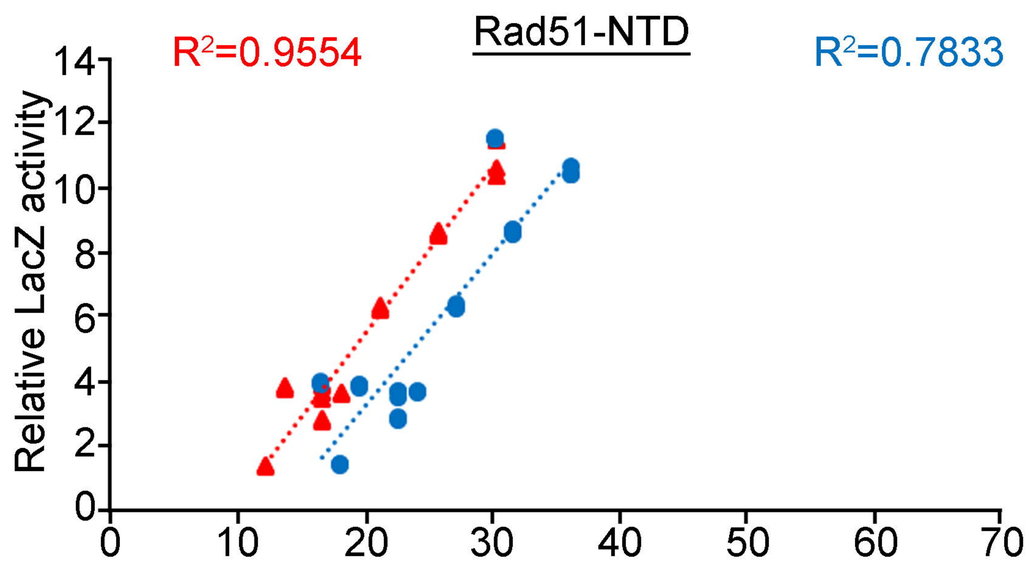
D


Rad51-NTD-LacZ-NVH
β-galactosidase activity (unit)
50
40
30
20
10
0
***
*
**
***
*
WT
hsp104Δ
new1Δ
doa4Δ
doa1Δ
san1Δ
oaz1Δ

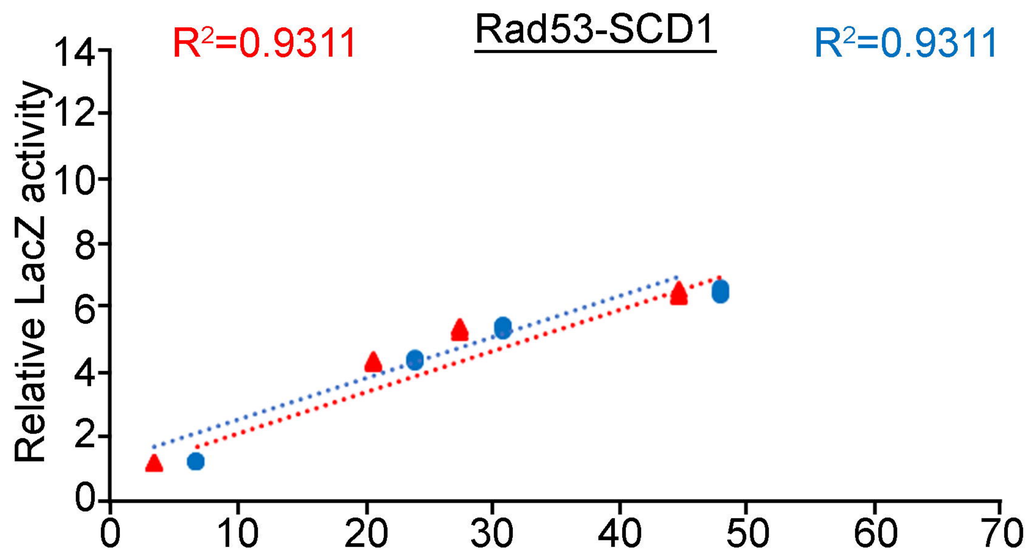
A

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-PND	39	3	0	12	9	38.46	61.54	8.34±0.20
Sup35-PND-S^{17}A	39	2	0	12	9	35.90	58.97	7.80±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
Vps64-NTD	152	21	10	15	8	30.26	35.53	4.86±0.18
Ssk2-NTD	191	40	13	11	13	33.50	40.31	5.99±0.43
Kel1-NTD	69	11	6	5	3	31.88	36.23	10.75±0.40
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
Sup35-PND	39	3	0	12	9	38.46	61.54	8.30±0.08
Sup35-PND-S^{17}A	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
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

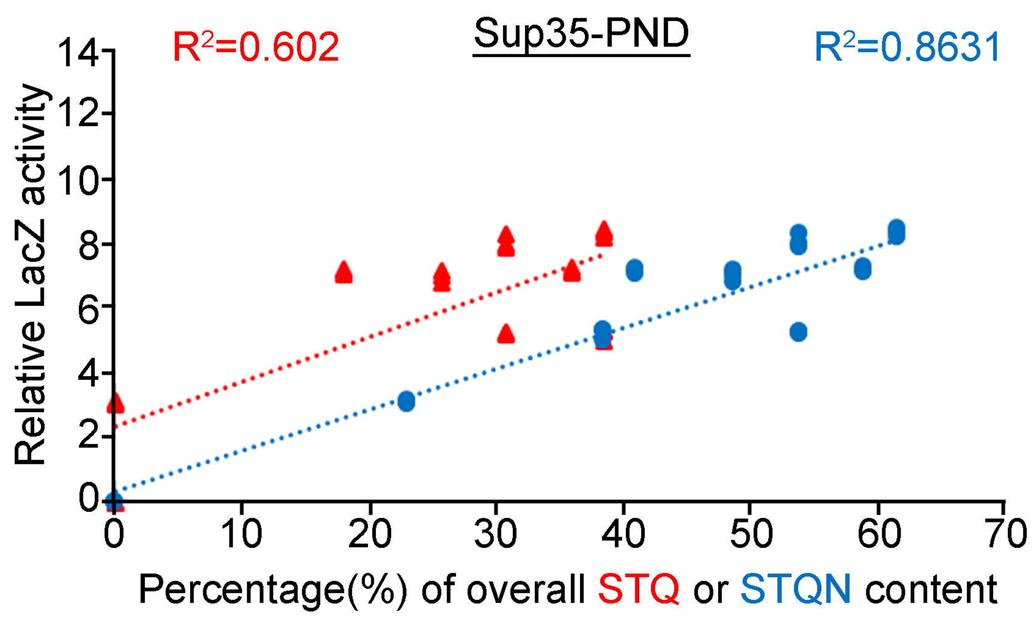
B



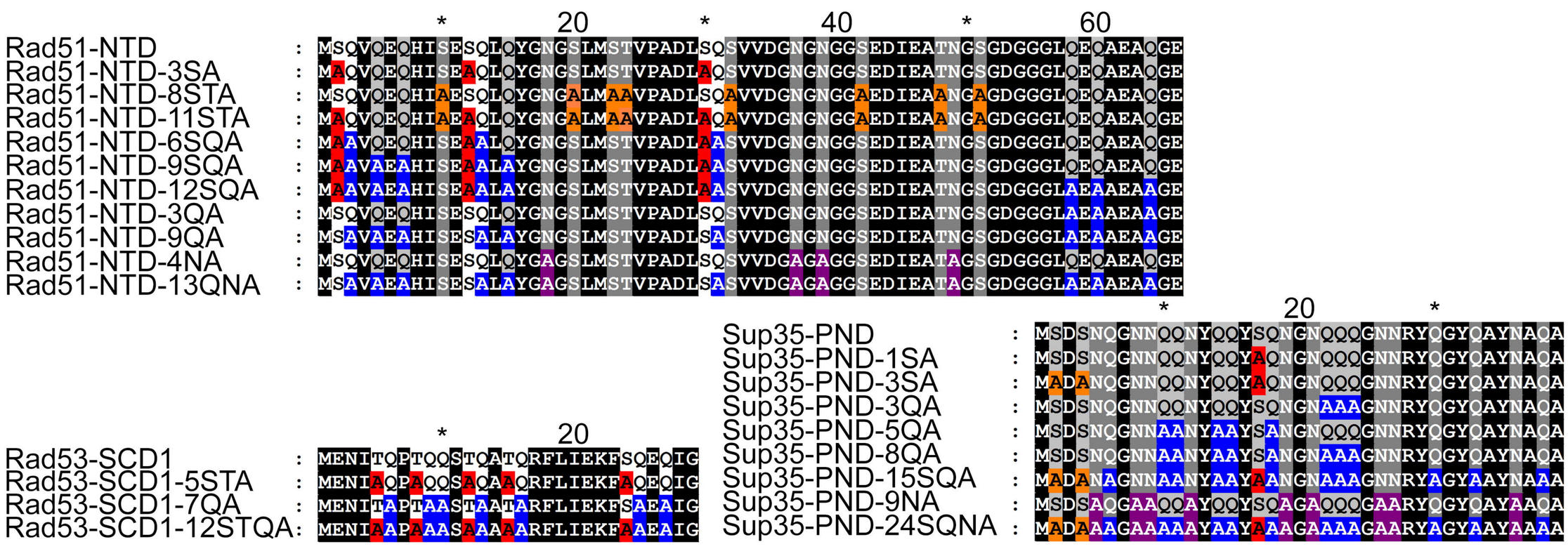
C

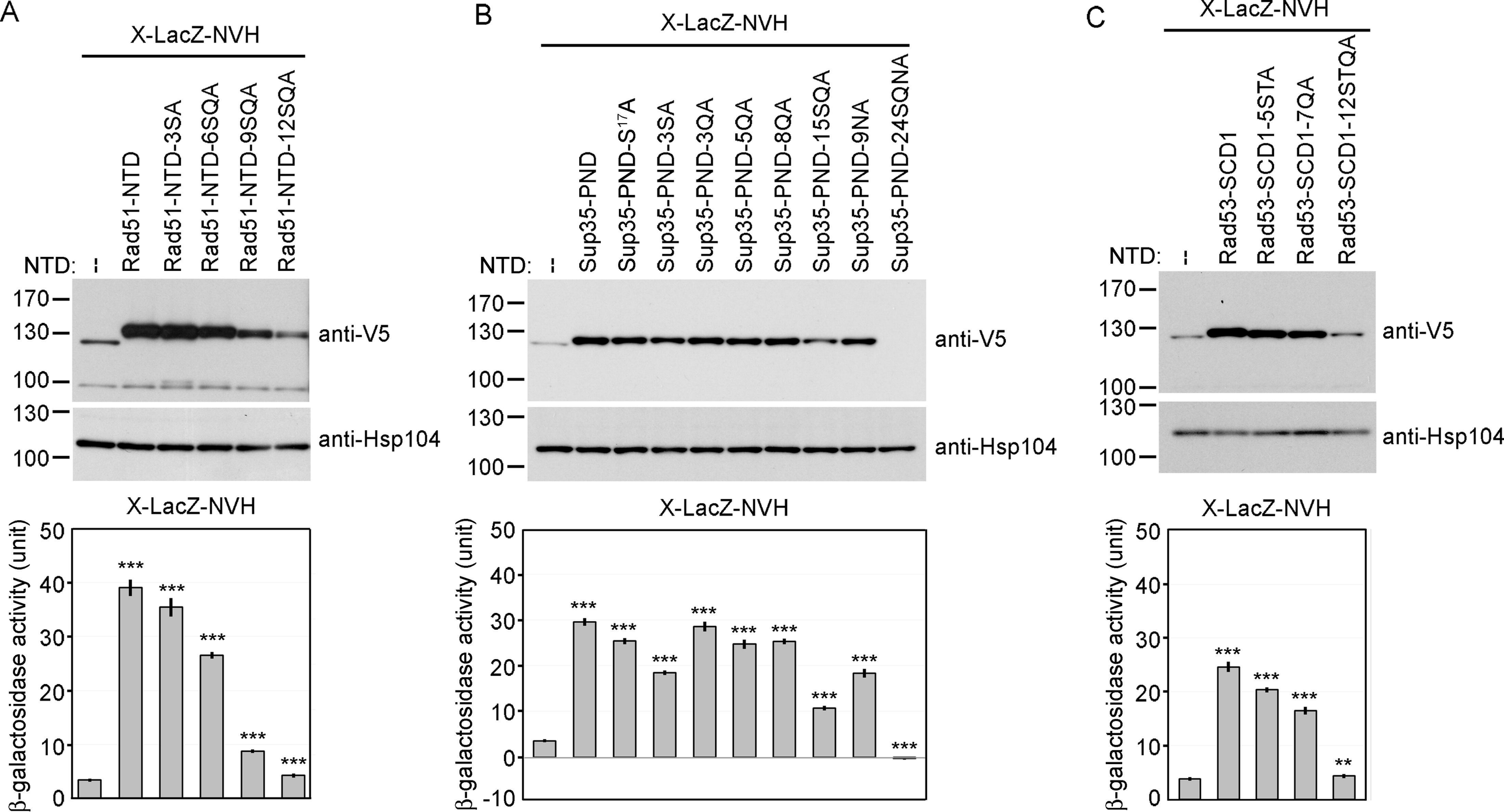


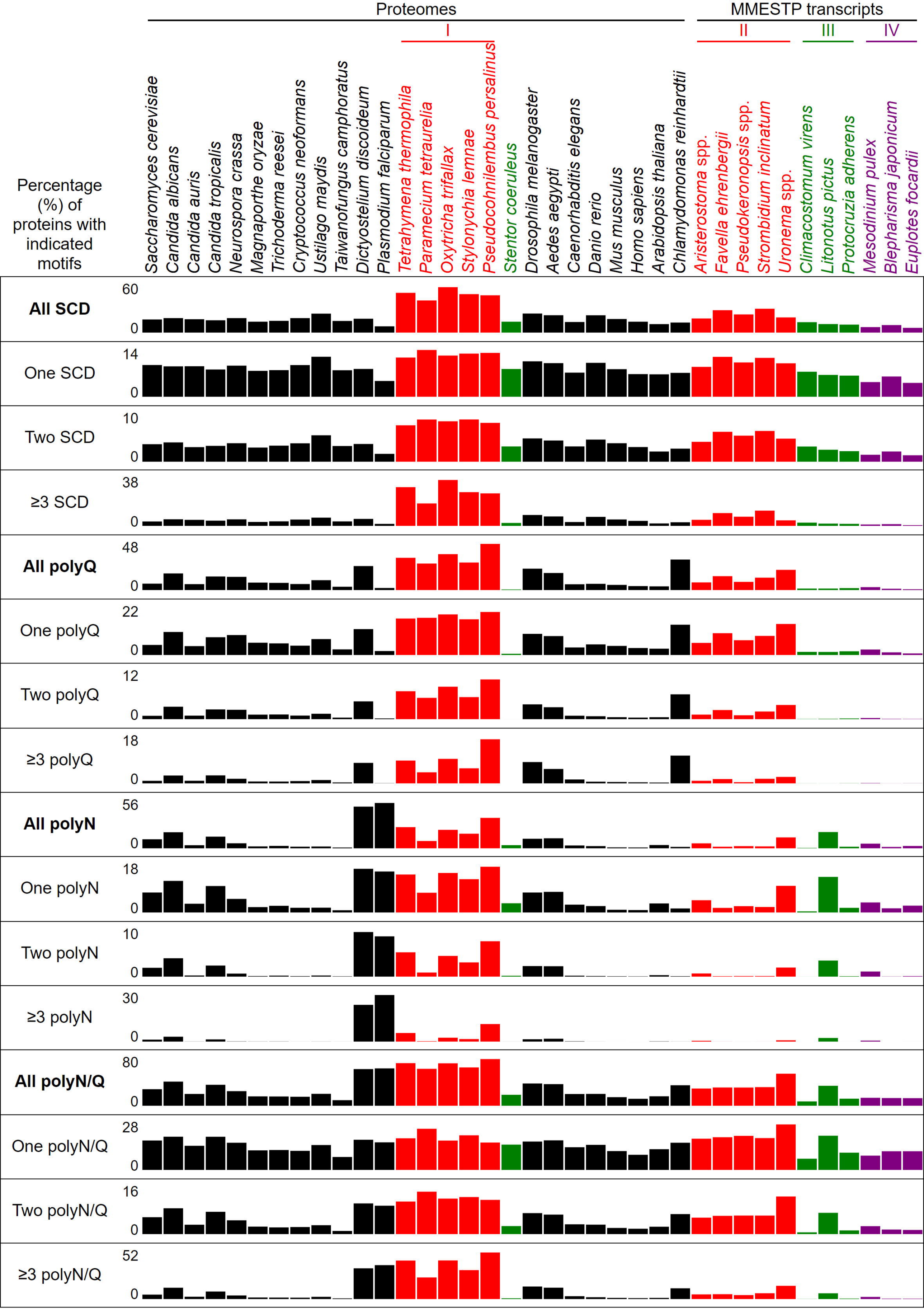
D

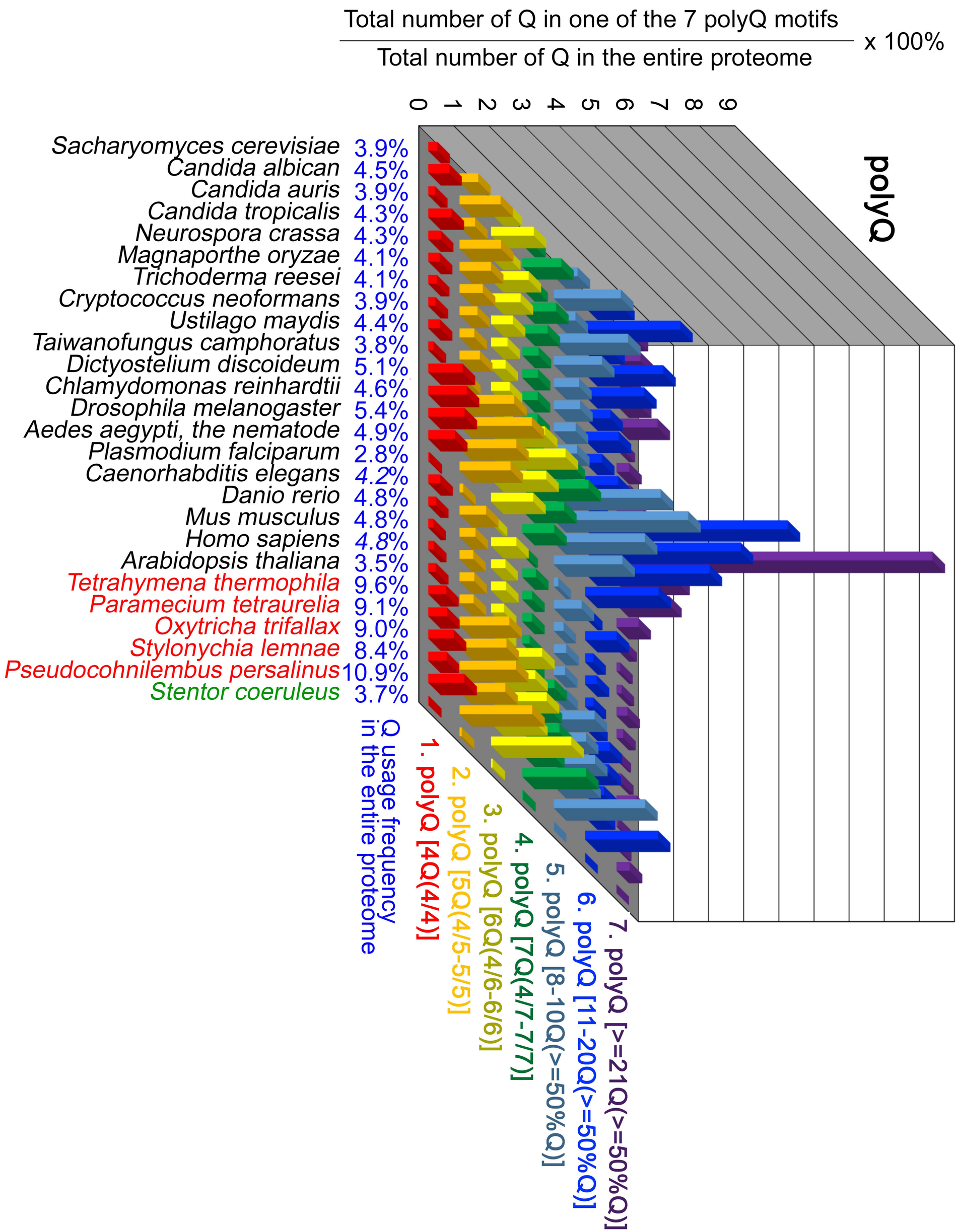


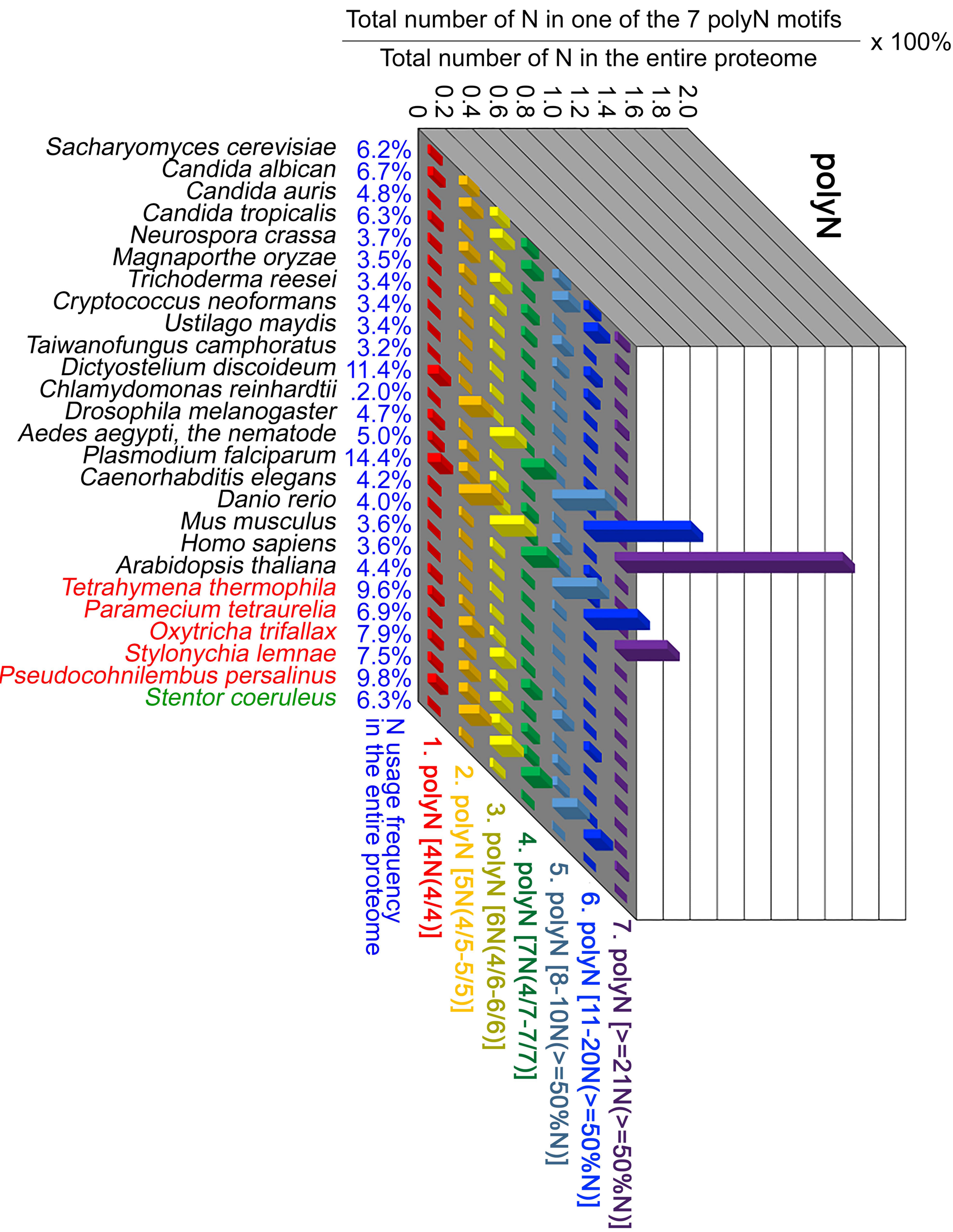
E

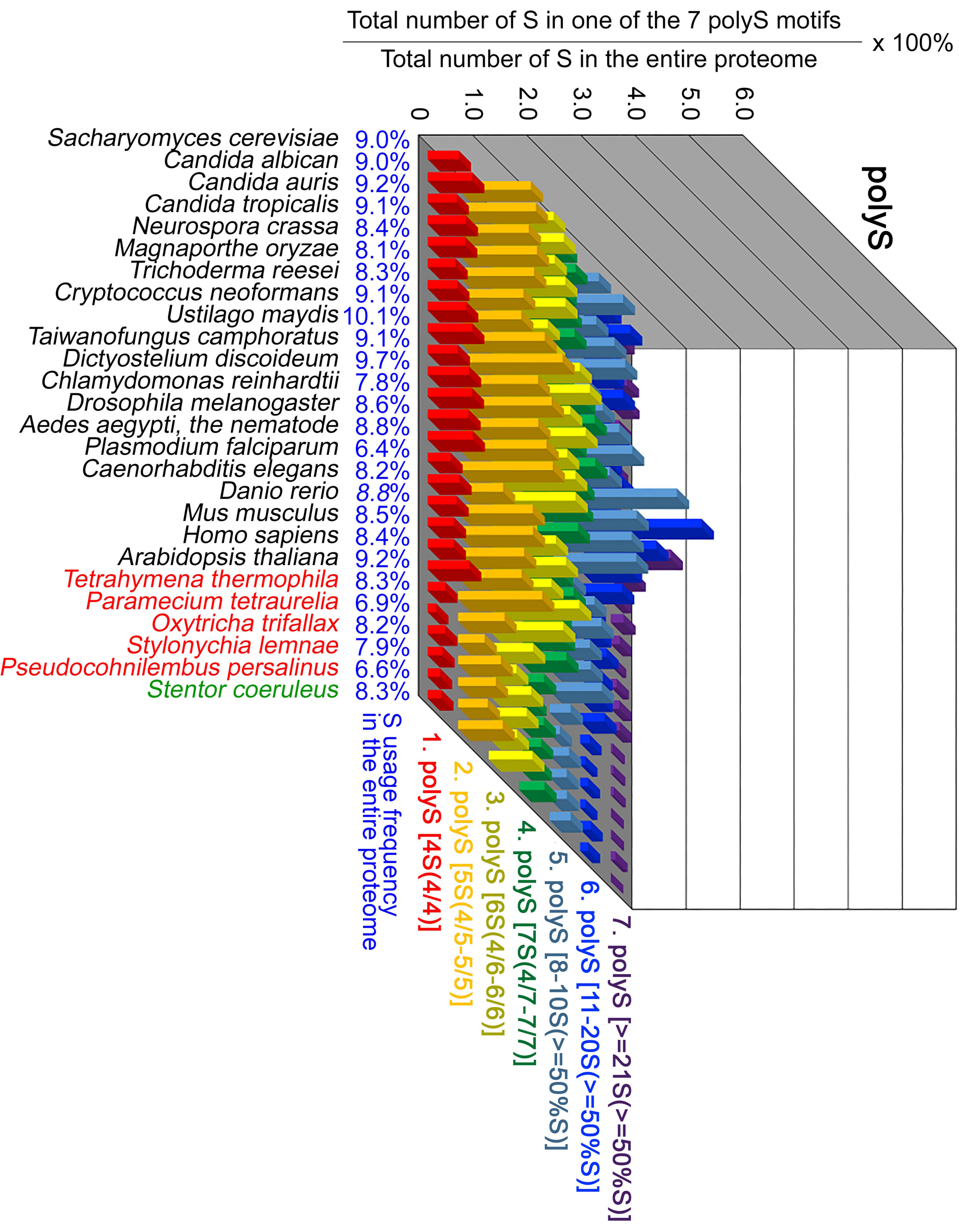


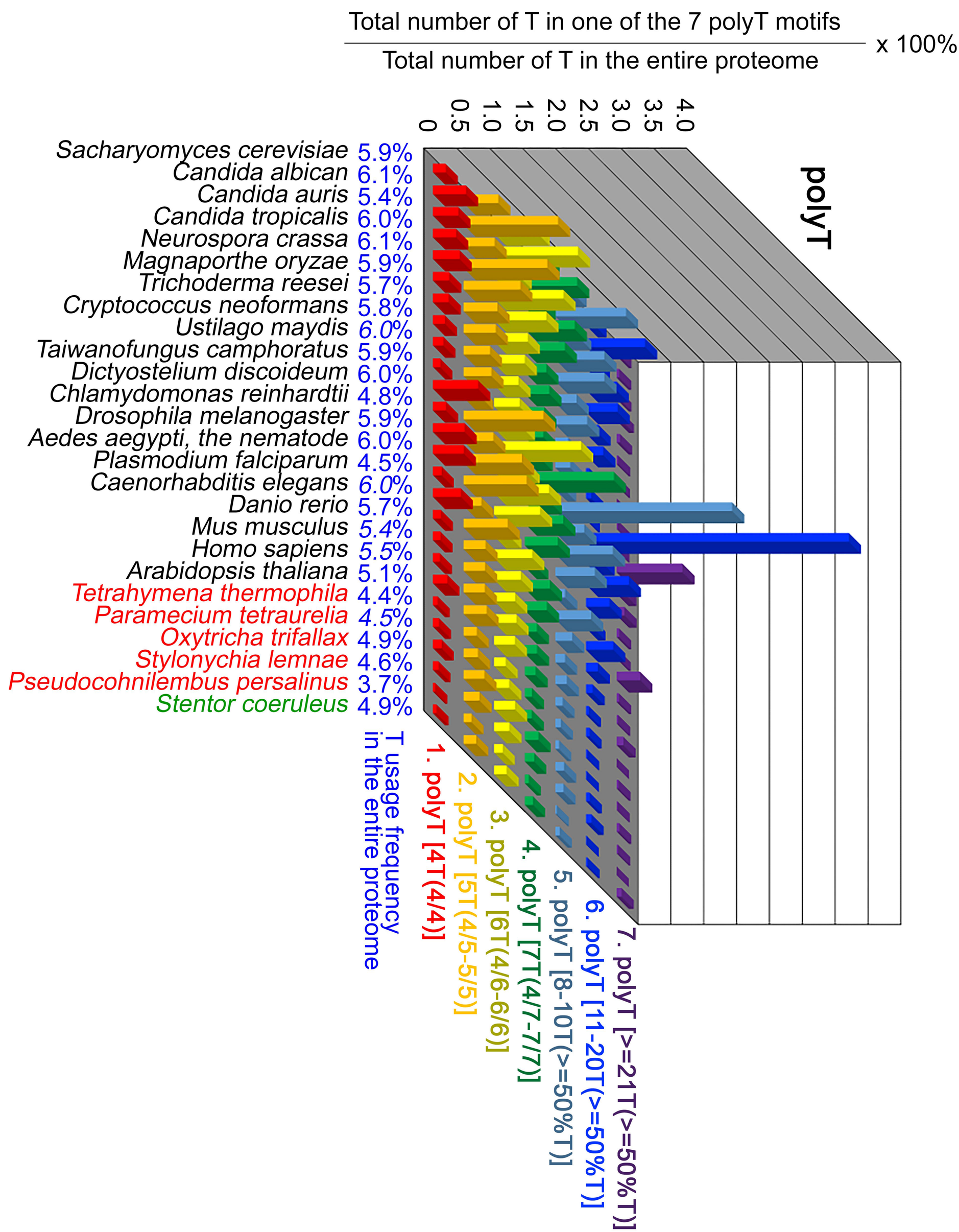

A
X-LacZ-NVH
NTD:
-
Rad51-NTD
Rad51-NTD-3SA
Rad51-NTD-6SQA
Rad51-NTD-9SQA
Rad51-NTD-12SQA
170
130
100
anti-V5
130
100
anti-Hsp104
X-LacZ-NVH
β-galactosidase activity (unit)
50
40
30
20
10
0
***
B
X-LacZ-NVH
NTD:
-
Sup35-PND
Sup35-PND-S17A
Sup35-PND-3SA
Sup35-PND-3QA
Sup35-PND-5QA
Sup35-PND-8QA
Sup35-PND-15SQA
Sup35-PND-9NA
Sup35-PND-24SQNA
170
130
100
anti-V5
130
100
anti-Hsp104
X-LacZ-NVH
β-galactosidase activity (unit)
50
40
30
20
10
0
-10
***
C
X-LacZ-NVH
NTD:
-
Rad53-SCD1
Rad53-SCD1-5STA
Rad53-SCD1-7QA
Rad53-SCD1-12STQA
170
130
100
anti-V5
130
100
anti-Hsp104
X-LacZ-NVH
β-galactosidase activity (unit)
50
40
30
20
10
0
***
**

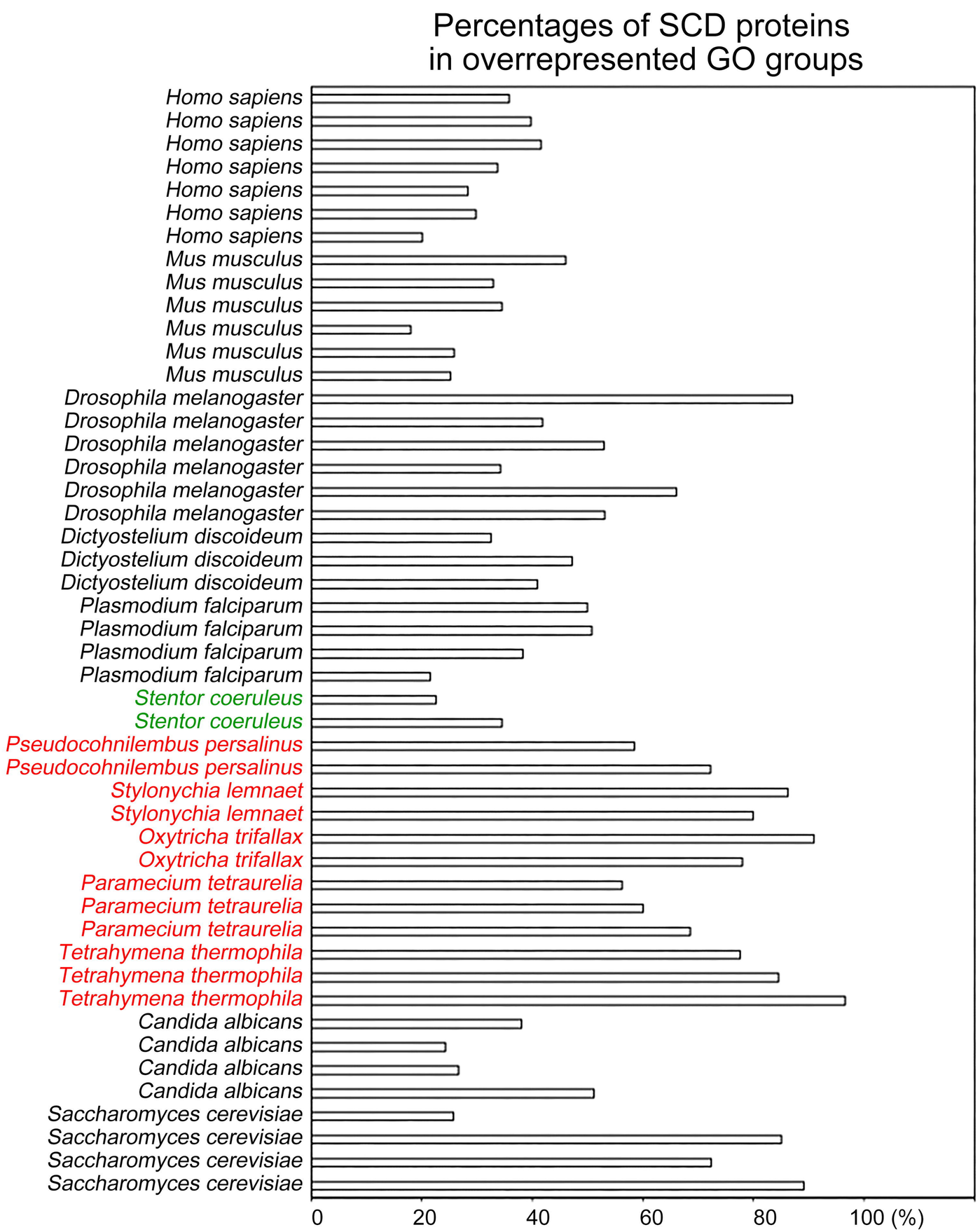

Proteomes
MMESTP transcripts
I
II
III
IV
Percentage (%) of proteins with indicated motifs
Saccharomyces cerevisiae
Candida albicans
Candida auris
Candida tropicalis
Neurospora crassa
Magnaporthe oryzae
Trichoderma reesei
Cryptococcus neoformans
Ustilago maydis
Taiwanofungus camphoratus
Dictyostelium discoideum
Plasmodium falciparum
Tetrahymena thermophila
Paramecium tetraurelia
Oxytricha trifallax
Stylonychia lemnae
Pseudocohnilembus persalinus
Stentor coeruleus
Drosophila melanogaster
Aedes aegypti
Caenorhabditis elegans
Danio rerio
Mus musculus
Homo sapiens
Arabidopsis thaliana
Chlamydomonas reinhardtii
Aristerostoma spp.
Favella ehrenbergii
Pseudokeronopsis spp.
Strombidium inclinatum
Uronema spp.
Climacostomum virens
Litonotus pictus
Protocruzia adherens
Mesodinium pulex
Blepharisma japonicum
Euplotes focardii
All SCD
60
0
One SCD
14
0
Two SCD
10
0
≥3 SCD
38
0
All polyQ
48
0
One polyQ
22
0
Two polyQ
12
0
≥3 polyQ
18
0
All polyN
56
0
One polyN
18
0
Two polyN
10
0
≥3 polyN
30
0
All polyN/Q
80
0
One polyN/Q
28
0
Two polyN/Q
16
0
≥3 polyN/Q
52
0


Total number of Q in one of the 7 polyQ motifs
Total number of Q in the entire proteome
x 100%
0
1
2
3
4
5
6
7
8
9
polyQ
Sacharyomyces cerevisiae 3.9%
Candida albican 4.5%
Candida auris 3.9%
Candida tropicalis 4.3%
Neurospora crassa 4.3%
Magnaporthe oryzae 4.1%
Trichoderma reesei 4.1%
Cryptococcus neoformans 3.9%
Ustilago maydis 4.4%
Taiwanofungus camphoratus 3.8%
Dictyostelium discoideum 5.1%
Chlamydomonas reinhardtii 4.6%
Drosophila melanogaster 5.4%
Aedes aegypti, the nematode 4.9%
Plasmodium falciparum 2.8%
Caenorhabditis elegans 4.2%
Danio rerio 4.8%
Mus musculus 4.8%
Homo sapiens 4.8%
Arabidopsis thaliana 3.5%
Tetrahymena thermophila 9.6%
Paramecium tetraurelia 9.1%
Oxytricha trifallax 9.0%
Stylonychia lemnae 8.4%
Pseudocohnilembus persalinus 10.9%
Stentor coeruleus 3.7%
Q usage frequency in the entire proteome
1. polyQ [4Q(4/4)]
2. polyQ [5Q(4/5-5/5)]
3. polyQ [6Q(4/6-6/6)]
4. polyQ [7Q(4/7-7/7)]
5. polyQ [8-10Q(>=50%Q)]
6. polyQ [11-20Q(>=50%Q)]
7. polyQ [>=21Q(>=50%Q)]

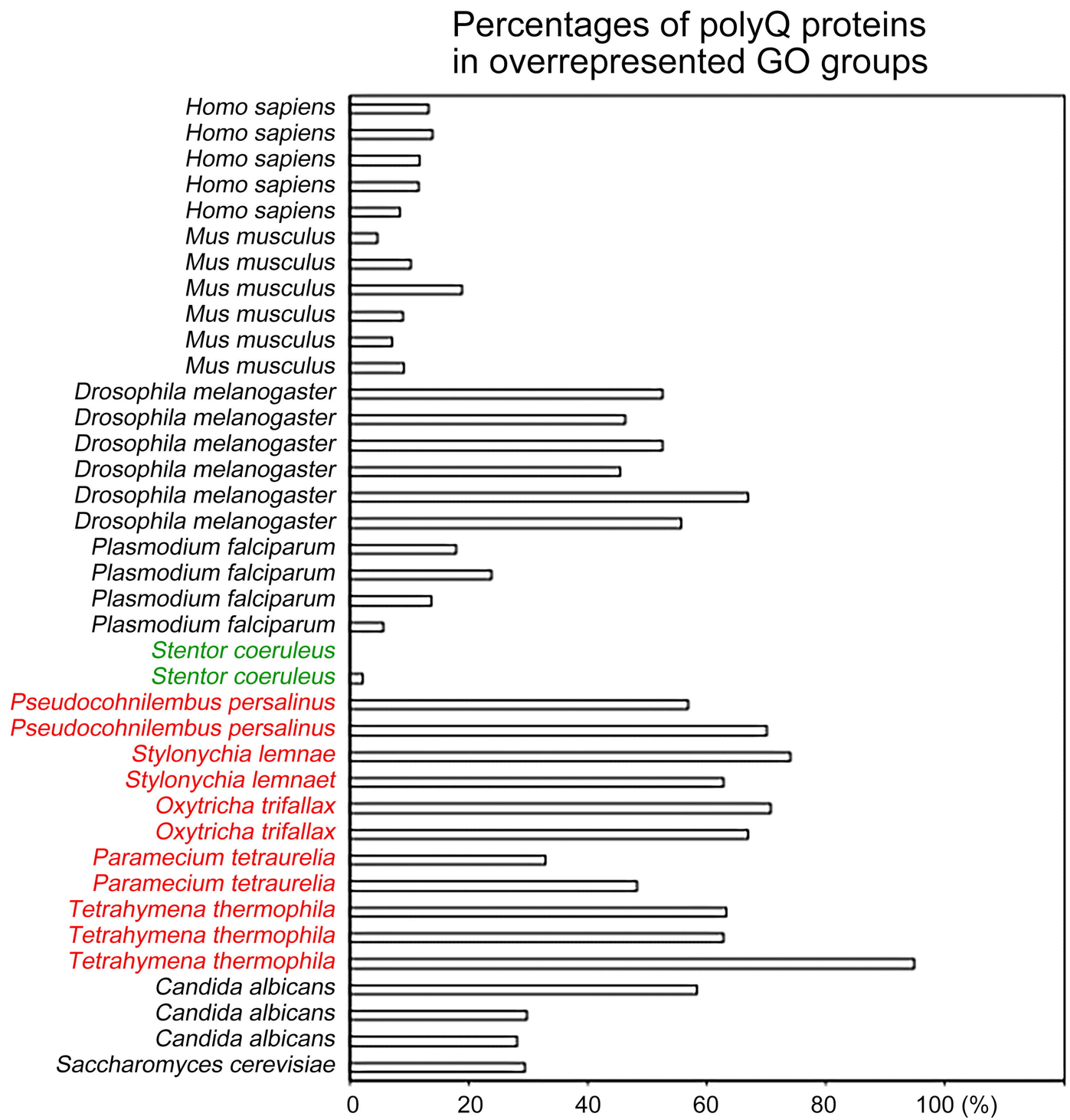

Total number of N in one of the 7 polyN motifs
Total number of N in the entire proteome
x 100%
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
polyN
Sacharyomyces cerevisiae 6.2%
Candida albican 6.7%
Candida auris 4.8%
Candida tropicalis 6.3%
Neurospora crassa 3.7%
Magnaporthe oryzae 3.5%
Trichoderma reesei 3.4%
Cryptococcus neoformans 3.4%
Ustilago maydis 3.4%
Taiwanofungus camphoratus 3.2%
Dictyostelium discoideum 11.4%
Chlamydomonas reinhardtii .2.0%
Drosophila melanogaster 4.7%
Aedes aegypti, the nematode 5.0%
Plasmodium falciparum 14.4%
Caenorhabditis elegans 4.2%
Danio rerio 4.0%
Mus musculus 3.6%
Homo sapiens 3.6%
Arabidopsis thaliana 4.4%
Tetrahymena thermophila 9.6%
Paramecium tetraurelia 6.9%
Oxytricha trifallax 7.9%
Stylonychia lemnae 7.5%
Pseudocohnilembus persalinus 9.8%
Stentor coeruleus 6.3%
N usage frequency in the entire proteome
1. polyN [4N(4/4)]
2. polyN [5N(4/5-5/5)]
3. polyN [6N(4/6-6/6)]
4. polyN [7N(4/7-7/7)]
5. polyN [8-10N(>=50%N)]
6. polyN [11-20N(>=50%N)]
7. polyN [>=21N(>=50%N)]


Total number of S in one of the 7 polyS motifs
Total number of S in the entire proteome
x 100%
0
1.0
2.0
3.0
4.0
5.0
6.0
polyS
Sacharyomyces cerevisiae 9.0%
Candida albican 9.0%
Candida auris 9.2%
Candida tropicalis 9.1%
Neurospora crassa 8.4%
Magnaporthe oryzae 8.1%
Trichoderma reesei 8.3%
Cryptococcus neoformans 9.1%
Ustilago maydis 10.1%
Taiwanofungus camphoratus 9.1%
Dictyostelium discoideum 9.7%
Chlamydomonas reinhardtii 7.8%
Drosophila melanogaster 8.6%
Aedes aegypti, the nematode 8.8%
Plasmodium falciparum 6.4%
Caenorhabditis elegans 8.2%
Danio rerio 8.8%
Mus musculus 8.5%
Homo sapiens 8.4%
Arabidopsis thaliana 9.2%
Tetrahymena thermophila 8.3%
Paramecium tetraurelia 6.9%
Oxytricha trifallax 8.2%
Stylonychia lemnae 7.9%
Pseudocohnilembus persalinus 6.6%
Stentor coeruleus 8.3%
S usage frequency in the entire proteome
1. polyS [4S(4/4)]
2. polyS [5S(4/5-5/5)]
3. polyS [6S(4/6-6/6)]
4. polyS [7S(4/7-7/7)]
5. polyS [8-10S(>=50%S)]
6. polyS [11-20S(>=50%S)]
7. polyS [>=21S(>=50%S)]


Total number of T in one of the 7 polyT motifs
Total number of T in the entire proteome
x 100%
polyT
0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Sacharyomyces cerevisiae 5.9%
Candida albican 6.1%
Candida auris 5.4%
Candida tropicalis 6.0%
Neurospora crassa 6.1%
Magnaporthe oryzae 5.9%
Trichoderma reesei 5.7%
Cryptococcus neoformans 5.8%
Ustilago maydis 6.0%
Taiwanofungus camphoratus 5.9%
Dictyostelium discoideum 6.0%
Chlamydomonas reinhardtii 4.8%
Drosophila melanogaster 5.9%
Aedes aegypti, the nematode 6.0%
Plasmodium falciparum 4.5%
Caenorhabditis elegans 6.0%
Danio rerio 5.7%
Mus musculus 5.4%
Homo sapiens 5.5%
Arabidopsis thaliana 5.1%
Tetrahymena thermophila 4.4%
Paramecium tetraurelia 4.5%
Oxytricha trifallax 4.9%
Stylonychia lemnae 4.6%
Pseudocohnilembus persalinus 3.7%
Stentor coeruleus 4.9%
T usage frequency
in the entire proteome
1. polyT [4T(4/4)]
2. polyT [5T(4/5-5/5)]
3. polyT [6T(4/6-6/6)]
4. polyT [7T(4/7-7/7)]
5. polyT [8-10T(>=50%T)]
6. polyT [11-20T(>=50%T)]
7. polyT [>=21T(>=50%T)]



Adjusted *p* vales	SCD protein number	Number of GO	GO terms
0	24	50808	Synapse organization
0	27	48812	Neuron projection morphogenesis
0	27	22008	Axonogenesis
0	28	7409	CNS development
0	43	7417	Brain development
0	77	7420	Nervous system development
0	97	7399	Dendrite development
0	29	16358	Neuron projection morphogenesis
0	29	48812	Neurogenesis
0	31	22008	Neuron projection morphogenesis
0	40	30182	Neuron differeniation
0	51	31175	Neuron projection development
0	86	7399	Nervous system development
0	101	50770	Regulation of axonogenesis
0	74	7399	Nervous system development
0	96	48813	Dendrite morphogenesis
0	67	7476	Imaginal disc morphogeneis
0	140	16319	Mushroom body development
0	199	7411	Axon guidance
0.154	63	30587	Sorocarp development
0	43	48870	Cell mobility
0	56	19953	Sexual reproduction
0	28	20035	Symbiont protein meidated cytoadherence
0	32	98609	Cell-cell adhesion
0	28	20013	Symbiont-modulated erythrocyte aggregaion
0	38	20033	Antigen variation
0.814	30	18200	Peptidyl-glutamic acid modification
0	78	7018	Microtubule-based movement
0.241	38	18200	Peptidyl-glutamic acid modification
0	65	7018	Microtubule-based movemen
0	63	18200	Peptidyl-glutamic acid modification
0	84	7018	Microtubule-based movementt
0	81	18200	Peptidyl-glutamic acid modification
0	99	7018	Microtubule-based movement
0.010	72	7131	Reciprocal meiotic recombination
0.027	42	18200	Peptidyl-glutamic acid modification
0	155	7018	Microtubule-based movement
0	38	18200	Peptidyl-glutamic acid modification
0	93	7018	Microtubule-based movement
0	56	45493	Xylan catabolic process
0.003	8	1900445	Filamentous growth (3)
0.018	47	44182	Filamentous growth (2)
0.007	103	30447	Filamentous growth (1)
0	22	7155	Cell adheion
0.037	27	6281	DNA repair
0	40	15074	DNA integration
0	55	6310	DNA recombinaiton
0	82	32197	RNA-mediated transpostion



Adjusted *p* vales	PolyQ protein number	Number of GO	GO terms
0	9	18812	Neuron projection morphogenesis
0	9	22006	Neurogenesis
0	18	7417	CNS development
0.001	30	7420	Brain development
0	41	7399	Nervous system development
0.003	3	16358	Dendrite development
0	9	18812	Neuron projection morphogenesis
0	17	22006	Neurogenesis
0	20	30182	Neuron differentiation
0.004	14	31175	Neuron projection development
0	31	7399	Nervous system development
0	95	60770	Regulation of axonogenesis
0	82	7399	Nervous system development
0	95	48813	Dendrite morphogenesis
0	89	7476	Imaginal disc morphogeneis
0	142	16319	Mushroom body development
0	209	7411	Axon guidance
0	10	20035	Symbiont protein meidated cytoadherence
0	15	98609	Cell-cell adhesion
0.001	10	20013	Symbiont-modulated erythrocyte aggregaion
0.716	10	20033	Antigen variation
1	0	18200	Peptidyl-glutamic acid modification
0.124	5	7018	Microtubule-based movement
0.791	37	18200	Peptidyl-glutamic acid modification
0	63	7018	Microtubule-based movemen
0	54	18200	Peptidyl-glutamic acid modification
0	66	7018	Microtubule-based movementt
0	63	18200	Peptidyl-glutamic acid modification
0	85	7018	Microtubule-based movement
0.027	23	18200	Peptidyl-glutamic acid modification
0	109	7018	Microtubule-based movement
0	31	18200	Peptidyl-glutamic acid modification
0	55	7018	Xylan catabolic process
0	69	45493	Microtubule-based movement
0	21	1900445	Filamentous growth (3)
0	48	44182	Filamentous growth (2)
0	109	30447	Filamentous growth (1)
0	15	7124	Pseudohyphal growth