Title

Systematic overexpression reveals a conserved spectrum of harmful and beneficial amino acid homorepeats shaped by evolution.

Authors

Yukihiro Murase[1], Naoki Kitamura[2], Shotaro Namba[1], Ayano Satoh[3], Takashi Makino[4], Ayako Moriya[5], and Hisao Moriya[5]

Affiliations

1. Graduate School of Life, Environmental, Natural Science and Technology, Okayama University
2. Graduate School of Interdisciplinary Science and Engineering in Health Systems, Okayama University
3. Faculty of Interdisciplinary Science and Engineering in Health Systems, Okayama University
4. Graduate School of Life Sciences, Tohoku University
5. Faculty of Life, Environmental, Natural Science and Technology, Okayama University

Conflict of interest statement

The authors declare that there is no conflict of interest accompanying this paper.

Author contribution

YM and NK performed the experiments. TM conducted the bioinformatics analyses, and SN carried out the data analysis. AM achieved technical support. AS supervised the research. YM and HM prepared the figures and wrote the manuscript.

Funding

This work was partly supported by the Nagase Science and Technology Foundation, Institute for Fermentation, Osaka (IFO) General Research Funding, JSPS KAKENHI Grant Numbers 25K22437, and 24K02313, the New Energy and Industrial Technology Development Organization (NEDO) Grant Number P20011.

Keywords

- amino acid homorepeat
- overexpression
- protein cytotoxicity
- protein evolution

Abstract

Proteins are composed of combinations of 20 amino acids, and the sequences selected through evolution reflect a balance between harmful and beneficial properties that together define their compatibility. However, the general principles governing the compatibility of amino acid sequences in living cells remain unclear. In this study, we systematically evaluated the neutrality—a measure of harmful versus beneficial properties—of simple amino acid homorepeats composed of ten identical residues ($Poly_{10}X$), using yeast as an experimental model system. EGFP-$Poly_{10}X$ fusions were analyzed using the genetic tug-of-war (gTOW) method to quantify maximum expression levels, revealing amino-acid-specific neutrality values. Hydrophobic repeats (I, V, W, F, Y) were strongly harmful and aggregation-prone, whereas hydrophilic repeats (E, S, N, Q) alleviated harmfulness. These trends were conserved across protein backgrounds and structural contexts. Consistently, hydrophobic repeats such as $Poly_{10}I$ formed aggregates and were intrinsically harmful, whereas $Poly_{10}E$ suppressed protein aggregation and prevented activation of proteotoxic stress responses. Furthermore, analysis of PolyX occurrence across the proteomes of multiple species revealed a strong correlation between experimentally measured $Poly_{10}X$ neutrality and both the length and frequency of homorepeats in natural proteins. Hydrophobic amino acids were rarely observed as homorepeats, whereas residues such as E, N, Q, and S appeared frequently and were beneficial in our assays. Taken together, $Poly_{10}X$ neutrality serves as an experimental measure of amino acid tolerance in cells and defines the compatible sequence space of proteins. It explains the absence of certain homorepeats in nature and has implications for protein design and mutation-induced harmfulness.

Introduction

Most proteins produced by extant organisms are composed of combinations of 20 amino acids with diverse physicochemical properties. The possible combinations of amino acid sequences—and thus their physicochemical characteristics—are virtually infinite. Among them, sequences that are compatible with living organisms have been selectively retained through the course of evolution [1]. This compatibility is determined by a balance between harms and benefits to biological functions, where the harms include various factors that decrease fitness, such as cytotoxicity or excessive consumption of cellular resources [2].

One effective way to understand the compatibility of amino acid sequences is to identify patterns commonly present in amino acid sequences (natural sequences) used by extant organisms. With the vast amount of protein sequence data now available, attempts have been made to extract features of natural sequences using artificial intelligence and to apply them to three-dimensional structure prediction and de novo protein design [3–6]. Although this approach has achieved success in extracting, predicting, and designing functional sequences, it does not directly reveal the compatibility of sequences that do not exist in nature, nor the characteristics of sequences that are not compatible due to harmful effects. "Nullpeptides," which are peptides absent from the proteome of a given organism, may represent sequences that have been eliminated during evolution [7,8]. However, such inferences still depend on the limited sequence space of natural proteins, and thus have inherent limitations in capturing the full landscape of non-compatible sequences.

Another approach to investigating sequence compatibility is the synthetic exploration of compatible combinations among amino acid sequences. By constructing a comprehensive sequence library and performing functional screening, it is possible to obtain functional sequences that do not exist in nature [9–11]. Furthermore, the use of Deep Mutational Scanning (DMS) enables systematic and quantitative evaluation of sequences within the library, allowing the identification of both the diversity and the characteristics of compatible sequences [12,13]. The resulting data can also serve as training datasets for

83 machine learning, and with sufficient data density, highly accurate predictive models can be developed
84 [14,15]. Neme *et al.* demonstrated through DMS in *Escherichia coli* that, while many random sequences
85 are harmful, a considerable number of beneficial sequences also exist [16,17]. However, this exploration
86 still covers only a limited portion of the vast sequence space, and the specific features and physiological
87 effects of harmful or beneficial sequences remain largely unknown.

88 Therefore, in this study, we aimed to extract harmful and beneficial features from extremely simple
89 model sequences. Of particular interest are continuous stretches of identical amino acids, known as
90 amino acid homorepeats [18–24] or PolyX. PolyX sequences are also referred to as homopolymeric amino
91 acid stretches/repeats [25–29], amino acid runs [30], or homopeptide repeats [31], and they are present in a wide
92 range of natural proteins. PolyX motifs exhibit distinct physiological functions and have been exploited
93 in various biotechnological applications [19,25,32]. For example, PolyP (~10 residues) segments of formin
94 act as nucleation sites for actin filament formation [33], and the arginine-rich region of protamine
95 containing PolyR tracts (~6 residues) binds strongly to DNA, inducing chromatin condensation [34].
96 Translation of the poly(A) tail resulting from termination failure produces PolyK, which serves as a
97 signal for protein degradation [35,36]. The metal-chelating property of PolyH (~6 residues) is utilized as a
98 purification tag [37], while the membrane-penetrating ability of PolyR (~8 residues) is exploited for
99 molecular delivery [38]. Conversely, PolyX sequences can also exert harmful effects and are associated
100 with diseases. Abnormal expansion of PolyQ beyond ~40 residues leads to aggregate formation and
101 neurodegenerative disorders [39]. Positively charged PolyR and PolyK (~10 residues) interact with the
102 negatively charged ribosomal tunnel, causing translational stalling [40], and PolyW (~10 residues) induces
103 ribosome stalling that prevents CAT-tail addition [41]. Negatively charged PolyD and PolyE (~10
104 residues) destabilize ribosomes and induce translational arrest [42,43].In natural protein sequences, both
105 the frequency and the length of PolyX motifs show amino acid–specific biases [18,29,31,44], which are
106 thought to reflect a balance between harmful and beneficial effects. PolyX sequences are considered
107 evolutionary hotspots, as they can readily arise through mutations [29,45–47]. When present in natural
108 proteins, they may confer benefits, whereas their absence could indicate harms that led to their

elimination. Although the cytotoxicity of certain PolyX sequences has been examined in mammalian cells [26,27,48], general trends that transcend species and expression conditions remain unexplored.

In this study, we systematically evaluated the harmful and beneficial effects of peptide sequences composed of ten identical amino acids ($Poly_{10}X$) by individually overexpressing them in yeast. A length of ten residues was chosen as it is sufficiently long to manifest amino-acid-specific physicochemical effects, such as charge accumulation or hydrophobic interactions, while remaining short enough to minimize sequence context and secondary-structure complexity. By fixing sequence length and repeating a single residue, this approach allows direct comparison of how intrinsic physicochemical properties—such as charge and hydrophobicity—affect cellular tolerance. Moreover, comparison with PolyX occurrence in natural proteomes enables us to infer which sequence features are favored or disfavored by evolutionary selection, thereby testing whether their harmful or beneficial effects contribute to determining their abundance in extant proteomes.

121 # Results

122 ## The harmful effects of each Poly$_{10}$X are generally conserved across species.

123 In this study, we began by investigating the expression limits of enhanced green fluorescent protein
124 (EGFP) fused at its C-terminus with 20 different Poly$_{10}$X peptides (Figure 1A, S1, and S2). The fusion
125 was made at the C-terminus rather than the N-terminus to minimize the influence of N-terminal
126 sequences, which are known to strongly affect translation efficiency [49]. For harmfulness assessment in
127 yeast, we expressed EGFP and EGFP-Poly$_{10}$X from the strong constitutive *TDH3* promoter (*TDH3$_{pro}$*)
128 and increased plasmid copy numbers using the genetic tug-of-war (gTOW) method [50,51]. During this
129 process, we measured both growth inhibition and EGFP expression levels (cellular fluorescence). In the
130 gTOW assay, plasmid copy number can be increased by culturing cells with or without leucine
131 supplementation. Under both conditions, we evaluated cell growth and fluorescence to calculate the
132 maximum growth rate (MGR) and maximum fluorescence intensity (MFI). To enable one-dimensional
133 comparison of the harmfulness of each Poly$_{10}$X variant, we defined relative neutrality based on growth
134 and expression data as follows: Relative neutrality = %(MGR_Poly$_{10}$X / MGR_Δ) × %(MFI_Poly$_{10}$X /
135 MFI_Δ). Where, Δ refers to the control EGFP without any Poly$_{10}$X. Thus, the relative neutrality values
136 were normalized such that the control (Δ) equals 10,000 (Figure 1A) .

137 As a result, the growth rate and fluorescence intensity varied greatly depending on the type of expressed
138 EGFP-Poly$_{10}$X (Figure 1B). Under leucine-depleted conditions, strong growth inhibition was observed
139 for about half of the Poly$_{10}$X variants, and yeast cells showed almost no growth. The harmfulness of
140 each Poly$_{10}$X was quantified as relative neutrality (Figure 1C), revealing that the degree of harmfulness
141 differed substantially among amino acids. The addition of some Poly$_{10}$X sequences markedly reduced
142 the neutrality of EGFP (increasing harmful effects), whereas the addition of other Poly$_{10}$X sequences
143 resulted in neutrality comparable to, or even higher than, that of EGFP.

144 Next, to examine whether the neutrality of $Poly_{10}X$ is conserved across species, we performed
145 equivalent experiments in *Escherichia coli*. Since the gTOW method cannot be applied to *E. coli*, we
146 constructed multicopy plasmids expressing EGFP-$Poly_{10}X$ under the control of the *lac* promoter (lac_{pro})
147 and induced their expression with isopropyl β-D-1-thiogalactopyranoside (IPTG) (Figure S3). The
148 resulting data were also converted into relative neutrality values (Figures 1D and 1E). In *E. coli*,
149 although the expression levels of EGFP-$Poly_{10}X$ varied considerably among the different $Poly_{10}X$
150 variants, the effects on growth were limited (Figure 1D). This is likely due to the weaker selection
151 pressure on plasmid maintenance compared with the yeast gTOW system—plasmids expressing highly
152 harmful $Poly_{10}X$ variants are thought to be lost from the population during cultivation (Figure S3E).
153 Previously, the cytotoxicity of PolyX (~30 residues) in mammalian cells was evaluated by Oma *et*
154 *al.* [26]. They assessed PolyX cytotoxicity in African green monkey kidney–derived COS-7 cells by
155 strongly expressing YFP-PolyX under the CMV promoter and measuring cell viability. Because
156 numeric cell viability data were not provided, we relied on reported statistical significance and
157 converted it into four ranks (1–4 corresponding to $p < 0.001$, $p < 0.01$, $p < 0.05$, and $p > 0.05$,
158 respectively). We then compared the $Poly_{10}X$ neutrality data obtained in this study for yeast and *E. coli*
159 with those reported by Oma *et al.* for mammalian cells. The results showed a generally high correlation
160 among the three systems (Spearman's correlation $\rho > 0.6$; Figures 1F-H), indicating that $Poly_{10}X$
161 harmfulness exhibits a conserved trend across species.

162 To further refine the analysis, we evaluated the neutrality of highly harmful $Poly_{10}X$ variants by using
163 the WTC_{846} promoter, which allows tunable expression [52]. By gradually modulating expression levels,
164 we assessed the neutrality of each $Poly_{10}X$ variant (Figure S4). The neutrality values (Figure 1I) showed
165 a stronger correlation with mammalian cell cytotoxicity ($\rho = 0.72$, Figure 1J), thereby reinforcing the
166 conclusion that $Poly_{10}X$ harmful exhibits a conserved trend across species. Figure 1K compares the
167 neutrality of EGFP-$Poly_{10}X$ variants in yeast across a range of expression strengths. Some amino acids
168 were consistently harmful across all expression levels, whereas others were more neutral than EGFP
169 (i.e., beneficial) at specific expression ranges.

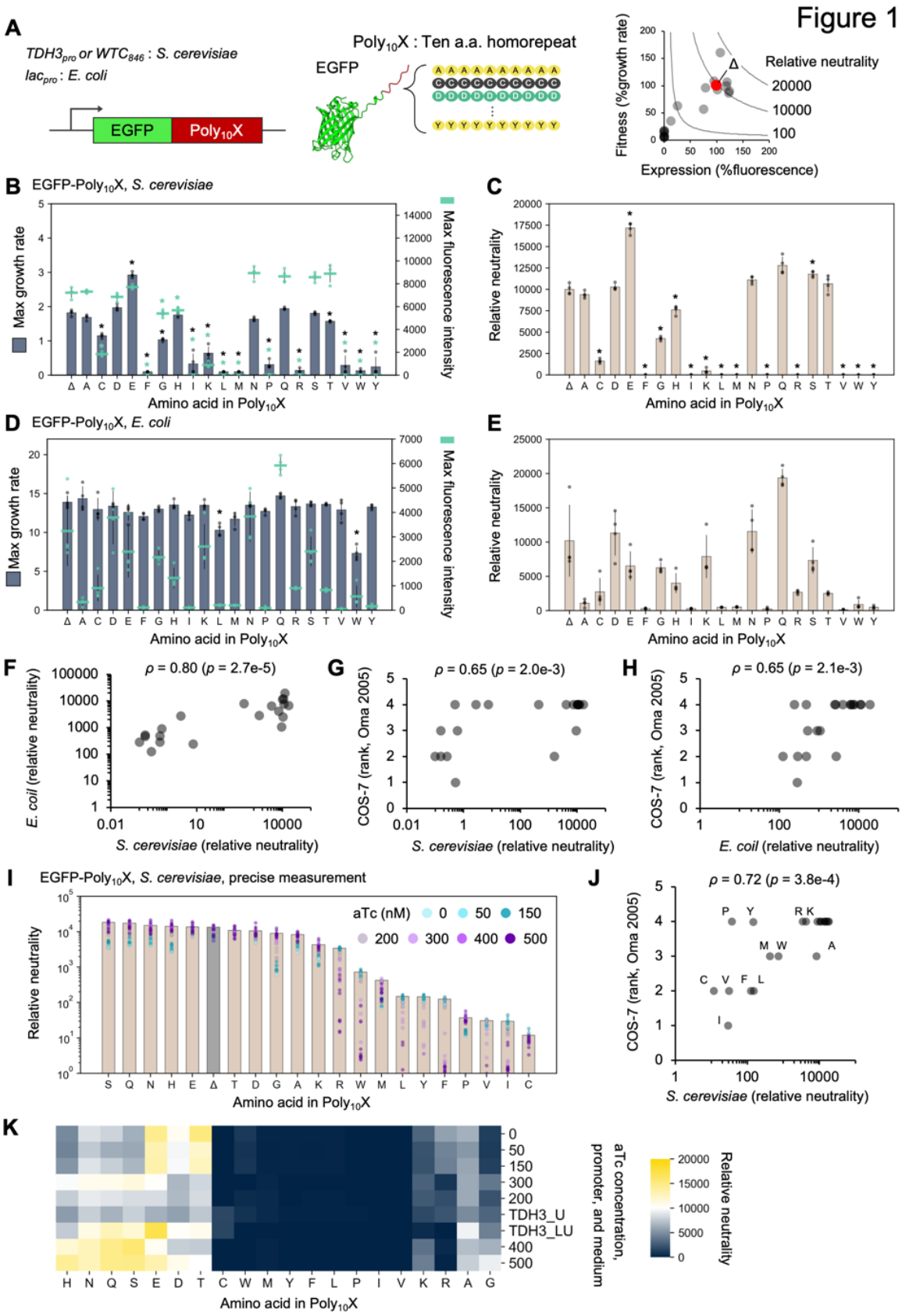


Figure 1
A
TDH3pro or WTC846 : S. cerevisiae
lacpro : E. coli
EGFP
Poly10X
Poly10X : Ten a.a. homorepeat
EGFP
Fitness (%growth rate)
Expression (%fluorescence)
Relative neutrality
20000
10000
100
B EGFP-Poly10X, S. cerevisiae
Max growth rate
Max fluorescence intensity
Amino acid in Poly10X
C
Relative neutrality
Amino acid in Poly10X
D EGFP-Poly10X, E. coli
Max growth rate
Max fluorescence intensity
Amino acid in Poly10X
E
Relative neutrality
Amino acid in Poly10X
F
ρ = 0.80 (p = 2.7e-5)
E. coil (relative neutrality)
S. cerevisiae (relative neutrality)
G
ρ = 0.65 (p = 2.0e-3)
COS-7 (rank, Oma 2005)
S. cerevisiae (relative neutrality)
H
ρ = 0.65 (p = 2.1e-3)
COS-7 (rank, Oma 2005)
E. coil (relative neutrality)
I EGFP-Poly10X, S. cerevisiae, precise measurement
aTc (nM) 0 50 150 200 300 400 500
Relative neutrality
Amino acid in Poly10X
J
ρ = 0.72 (p = 3.8e-4)
COS-7 (rank, Oma 2005)
S. cerevisiae (relative neutrality)
K
aTc concentration, promoter, and medium
Relative neutrality
Amino acid in Poly10X

170

171 **Figure 1.** The harmful effects of each $Poly_{10}X$ are generally conserved across
172 species.

173 **A)** Experimental setup. $Poly_{10}X$ sequences were fused to the C-terminus of EGFP and expressed in *S.*
174 *cerevisiae* under the *TDH3* or WTC_{846} promoter, or in *E. coli* under the *lac* promoter. Cellular fitness
175 (growth rate) and expression level (fluorescence intensity) of cells overexpressing each construct were
176 measured. Relative neutrality was calculated by comparison with the corresponding control EGFP
177 construct lacking $Poly_{10}X$ (Δ).

178 **B, C)** Maximum growth rate and maximum fluorescence intensity of *S. cerevisiae* cells overexpressing
179 EGFP–$Poly_{10}X$ in SC–LU medium at 30 °C, and the calculated relative neutrality (**C**). Growth and
180 fluorescence curves are shown in Figure S2B.

181 **D, E)** Maximum growth rate and maximum fluorescence intensity of *E. coli* expressing EGFP–$Poly_{10}X$
182 in LB + ampicillin medium supplemented with 1 mM IPTG at 37 °C, and the calculated relative
183 neutrality (**E**). Growth and fluorescence curves are shown in Figure S3B.

184 **F–H)** Cross-species comparison of $Poly_{10}X$ harmful effects. Average relative neutrality values obtained
185 in *S. cerevisiae* (**C**) and *E. coli* (**E**) were compared with each other and with previously reported $Poly_{10}X$
186 cytotoxicity ranks in COS-7 cells [26]. Cytotoxicity ranks were calculated as described in *Materials and*
187 *Methods*.

188 **I)** Precise determination of relative neutrality in *S. cerevisiae*. EGFP–$Poly_{10}X$ expression was titrated
189 across seven concentrations of aTc, and the maximum neutrality point was used for final normalization
190 to the EGFP control (Δ). Growth and fluorescence curves are shown in Figure S4B and C.

191 **J)** Correlation between the maximum relative neutrality of $Poly_{10}X$ in *S. cerevisiae* (**I**) and the $Poly_{10}X$
192 cytotoxicity ranking in COS-7 cells [26]. Single-letter codes indicate the amino acid repeated in PolyX.

193 **K)** Hierarchical clustering analysis of relative neutrality values across varying expression levels. On the
194 y-axis, TDH3_U and TDH3_LU indicate experiments using $TDH3_{pro}$ under SC-U (Figure S1) and SC-

195 LU (**C** and Figure S2) conditions, respectively. The numbers indicate the concentration of aTc in
196 experiments using WTC_{846} (**I** and Figure S4). Each dataset was normalized to the neutrality of the EGFP
197 control. Euclidean distance and average linkage were used. The heatmap represents relative neutrality
198 levels.

199 In **B**–**E** and **I**, bars, dots, and error bars represent the mean, individual values, and standard deviation
200 from at least three biological replicates. In **B**, **C**, and **D**, asterisks indicate significant differences in
201 maximum growth rate, maximum fluorescence intensity, and relative neutrality, respectively, compared
202 with the control (Δ) ($p < 0.05$, Student's *t*-test with Bonferroni correction). In **F**–**H** and **J,** Spearman's
203 rank correlation coefficient (ρ) and its *p*-value are shown.

204

205 Poly$_{10}$X harmful and beneficial effects are associated with amino acid
206 polarity and hydrophobicity

207 Next, to examine whether the harmful or beneficial effects of Poly$_{10}$X depend on the physicochemical
208 properties of the fused protein, we fused Poly$_{10}$X to the C-terminus of six different fluorescent proteins
209 and evaluated their neutrality in the same manner (Figure 2A, S5–S9). The obtained neutrality values
210 showed similar trends, with correlations above 0.85 between datasets (Figure 2C). Based on this, we
211 analyzed the relationship between amino acid biochemical and physicochemical parameters and
212 neutrality. Correlation analysis with 28 amino acid indices [2,53] (Figure S10A) revealed that biosynthetic
213 energy cost showed the highest correlation ($\rho = -0.76$, $p = 1.8e\text{-}23$), followed by polar requirement (ρ
214 $= 0.66$, $p = 2.2e\text{-}16$) and hydrophobicity ($\rho = -0.60$, $p = 4.0e\text{-}13$) (Figure 2D). Because these parameters
215 mutually showed high intercorrelation (Figure S10B), neutrality may be largely explained by these
216 characteristics.

217 To further characterize the nature of Poly$_{10}$X, we performed cluster analysis and classified amino acids
218 into three groups (Figure 2E). Cluster 1 included amino acids that increased the relative neutrality above

219 10,000 for most fluorescent proteins when added to the C-terminus, indicating beneficial effects that
220 reduced protein cytotoxicity. Cluster 2 consisted of amino acids that exhibited harmful effects
221 regardless of the fluorescent protein to which they were attached. Cluster 3 contained amino acids
222 whose effects varied depending on the fused fluorescent protein. Notably, we did not detect any
223 beneficial amino acids in Gamillus, probably because the expression level was insufficient to cause
224 overexpression-dependent growth inhibition, even at maximum induction (Figure S7). Figure 2F shows
225 the representative biochemical and physicochemical properties of amino acids in each cluster. Cluster
226 1 included E, S, N, and Q, which are low-cost, have high polar requirements, and low hydrophobicity.
227 Cluster 2 mainly contains amino acids with high cost, low polarity, and high hydrophobicity (C, P, M,
228 V, F, I, W, L, Y), but also includes positively charged ones (K, R). Cluster 3 contains G, T, H, and A,
229 which exhibit intermediate cost, polar requirement, and hydrophobicity, while D is somewhat
230 exceptional.

231 To examine whether the strong harmfulness of the four $Poly_{10}X$ variants (F, I, W, and Y) was caused
232 by the high amino acid synthesis cost, we re-evaluated the harmfulness of EGFP (Δ) and four EGFP–
233 $Poly_{10}X$ constructs after supplementing the medium with excess corresponding amino acids. As a result,
234 in all cases, the harmfulness was not alleviated (Figure 2G and S11), suggesting that amino acid
235 depletion due to high biosynthetic cost was not the cause of harmfulness. As mentioned above, since
236 the biosynthetic cost correlates well with polarity and hydrophobicity, the apparent correlation with
237 synthesis cost likely reflects a pseudo-correlation arising from these physicochemical properties.

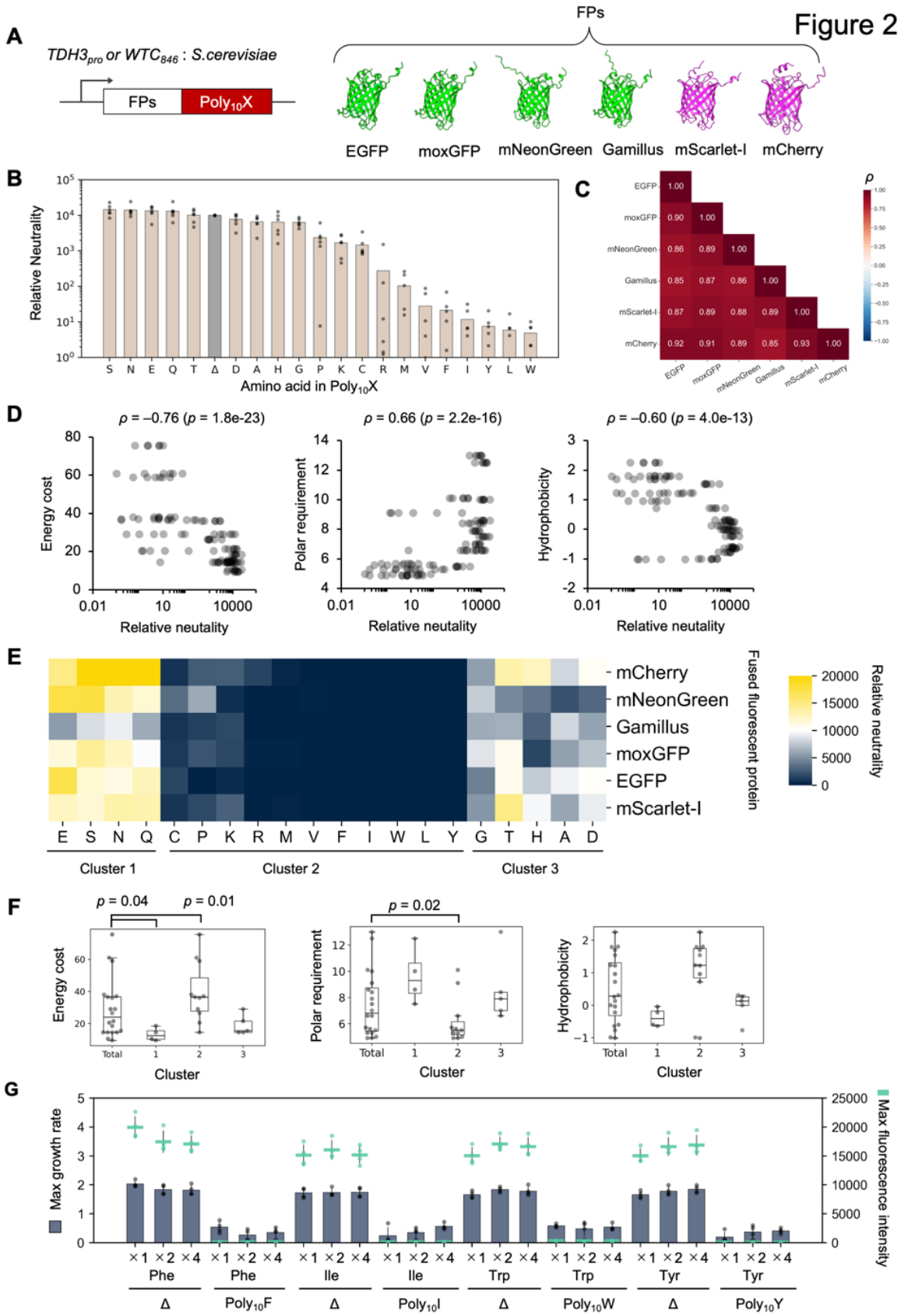


Figure 2
A
TDH3pro or WTC846 : S.cerevisiae
FPs
Poly10X
FPs
EGFP
moxGFP
mNeonGreen
Gamillus
mScarlet-I
mCherry
B
Relative Neutrality
Amino acid in Poly10X
S N E Q T Δ D A H G P K C R M V F I Y L W
C
ρ
EGFP 1.00
moxGFP 0.90 1.00
mNeonGreen 0.86 0.89 1.00
Gamillus 0.85 0.87 0.86 1.00
mScarlet-I 0.87 0.89 0.88 0.89 1.00
mCherry 0.92 0.91 0.89 0.85 0.93 1.00
D
ρ = –0.76 (p = 1.8e-23)
ρ = 0.66 (p = 2.2e-16)
ρ = –0.60 (p = 4.0e-13)
Energy cost
Polar requirement
Hydrophobicity
Relative neutality
E
mCherry
mNeonGreen
Gamillus
moxGFP
EGFP
mScarlet-I
Fused fluorescent protein
Relative neutrality
E S N Q C P K R M V F I W L Y G T H A D
Cluster 1
Cluster 2
Cluster 3
F
p = 0.04
p = 0.01
p = 0.02
Energy cost
Polar requirement
Hydrophobicity
Total 1 2 3
Cluster
G
Max growth rate
Max fluorescence intensity
×1 ×2 ×4
Phe Phe Ile Ile Trp Trp Tyr Tyr
Δ Poly10F Δ Poly10I Δ Poly10W Δ Poly10Y

238

239 **Figure 2.** Poly$_{10}$X harmful and beneficial effects are associated with amino acid
240 polarity and hydrophobicity.

241 **A**) Schematic of the constructs used in this study. Each Poly$_{10}$X was attached to the C-terminal of each
242 fluorescent protein (FP), and expressed in *S. cerevisiae* under the control of the *WTC$_{846}$* promoter. EGFP
243 constructs were expressed under the *TDH3* promoter (see Figure 1B, C).

244 **B**) Relative neutrality of Poly$_{10}$X measured using various fluorescent proteins in SC–LU medium at
245 30 °C. Bars represent the mean relative neutrality for each amino acid, and dots correspond to
246 measurements across different fluorescent proteins. Each dot represents the mean of at least three
247 biological replicates, and the bars indicate the mean relative neutrality obtained across six different
248 fluorescent proteins. Growth and fluorescence curves are shown in Figure S2B, S5B–S9B.

249 **C**) Correlation matrix comparing the trends of Poly$_{10}$X neutrality across different fluorescent proteins.
250 Each value represents the Spearman's rank correlation coefficient based on relative neutrality values
251 measured for each FPs–Poly$_{10}$X.

252 **D**) Relationships between Poly$_{10}$X neutrality and amino acid physicochemical properties (energy cost,
253 polar requirement and hydrophobicity). Spearman's rank correlation coefficient (ρ) and its p-values are
254 shown.

255 **E**) Hierarchical clustering of relative neutrality values obtained from different fluorescent proteins.
256 Clustering was performed using Euclidean distance and average linkage. Single-letter codes indicate
257 the amino acid repeated in Poly$_{10}$X.

258 **F**) Comparison of the three Poly$_{10}$X clusters identified in **E** with amino acid physicochemical properties
259 (energy cost, polar requirement and hydrophobicity). Statistical analyses were performed using the
260 Mann–Whitney U test between each cluster and the remaining total samples, followed by Bonferroni
261 correction for multiple comparisons.

262 **G**) Maximum growth rate and maximum fluorescence intensity of *S. cerevisiae* cells overexpressing
263 EGFP without $Poly_{10}X$ (Δ) or with C-terminal $Poly_{10}F$, $Poly_{10}I$, $Poly_{10}W$, or $Poly_{10}Y$ fusions in SC–LU
264 medium at 30 °C. Indicated amino acid was supplemented to the medium at standard (×1), ×2, or ×4
265 concentrations. Bars, dots, and error bars represent the mean, individual values, and standard deviation
266 from four biological replicates. Growth and fluorescence curves are shown in Figure S11B.

267

268 Structural context modulates the effect of $Poly_{10}X$, while its overall
269 neutrality trend is conserved.

270 So far, we have examined the effects of C-terminally fused $Poly_{10}X$. Next, to investigate the effects of
271 $Poly_{10}X$ in different structural contexts, we constructed three types of variants (Figures 3A and S12–
272 14): $Poly_{10}X$ was inserted between two fluorescent proteins (EGFP and mCherry); $Poly_{10}X$ was inserted
273 into an internal loop of EGFP; and the C-terminally fused $Poly_{10}X$ was detached from the fluorescent
274 protein by a P2A sequence [54]. These designs allowed us to test whether the effects of $Poly_{10}X$ depend
275 on the presence of a downstream protein domain, its position within a protein, or its physical coupling
276 to the fluorescent protein. As a result, all constructs showed distinct growth rates, fluorescence levels,
277 and neutrality indices depending on the $Poly_{10}X$ sequence (Figures 3B–G and S12–14). It should be
278 noted that in the detached construct, the fluorescence intensity reflects the amount of free EGFP released
279 by the P2A cleavage, but not the amount of the $Poly_{10}X$ peptide itself. In contrast, internal $Poly_{10}X$
280 insertions may influence not only protein neutrality but also the fluorescence properties of EGFP,
281 meaning that fluorescence does not necessarily reflect expression levels directly. The neutrality indices
282 among the three structural contexts and the C-terminal EGFP–$Poly_{10}X$ construct showed rank
283 correlations above 0.74 (p = 2.2e-4) (Figure 3H), indicating that $Poly_{10}X$ exhibits a generally conserved
284 trend in neutrality regardless of structural context. Among them, K and P showed markedly reduced
285 harmfulness when placed between two fluorescent proteins compared with when fused to the C-
286 terminus (Figure 3I, upper left). I, V, and W maintained strong harmfulness even in the detached

287 $Poly_{10}X$ construct (Figure 3I, lower right), suggesting that the $Poly_{10}X$ peptides themselves can exert
288 harmfulness independently. Conversely, the significant harmfulness of C and H (Figure 1C) were lost
289 in the detached $Poly_{10}X$ construct (Figure 3G), indicating that these $Poly_{10}X$ sequences exhibit
290 harmfulness only when fused to EGFP. No beneficial effect of $Poly_{10}X$ was observed when the $Poly_{10}X$
291 were detached (Figure 3G). Therefore, the beneficial effects might arise from mitigating the
292 overexpression-induced harmfulness of fluorescent proteins.

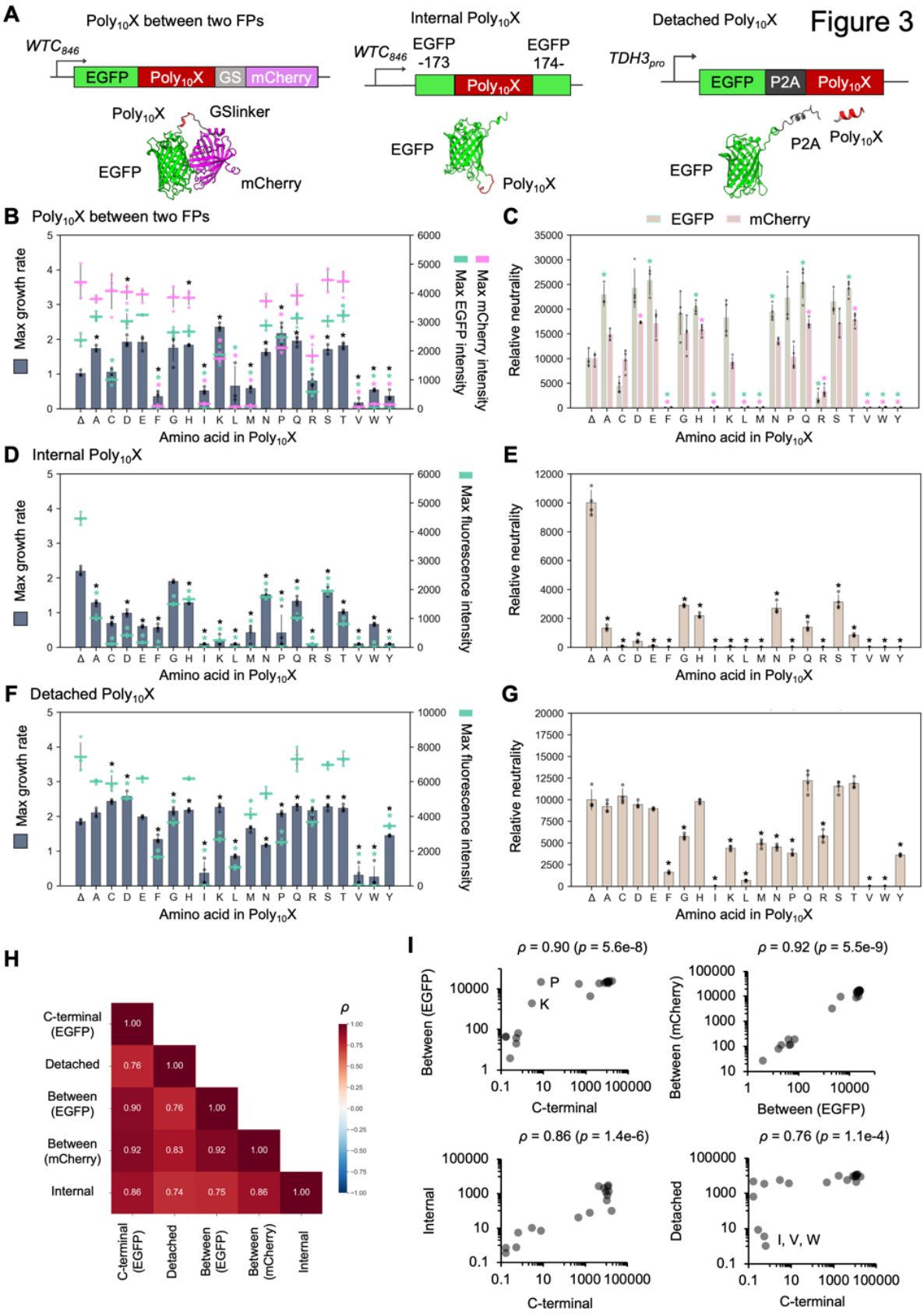


Figure 3
A
Poly10X between two FPs
WTC846
EGFP
Poly10X
GS
mCherry
Poly10X
GSlinker
EGFP
mCherry
Internal Poly10X
WTC846
EGFP -173
EGFP 174-
Poly10X
EGFP
Poly10X
Detached Poly10X
TDH3pro
EGFP
P2A
Poly10X
P2A
Poly10X
EGFP
B Poly10X between two FPs
Max growth rate
Max EGFP intensity
Max mCherry intensity
Amino acid in Poly10X
C
EGFP
mCherry
Relative neutrality
Amino acid in Poly10X
D Internal Poly10X
Max growth rate
Max fluorescence intensity
Amino acid in Poly10X
E
Relative neutrality
Amino acid in Poly10X
F Detached Poly10X
Max growth rate
Max fluorescence intensity
Amino acid in Poly10X
G
Relative neutrality
Amino acid in Poly10X
H
C-terminal (EGFP)
Detached
Between (EGFP)
Between (mCherry)
Internal
ρ
I
ρ = 0.90 (p = 5.6e-8)
Between (EGFP)
C-terminal
ρ = 0.92 (p = 5.5e-9)
Between (mCherry)
Between (EGFP)
ρ = 0.86 (p = 1.4e-6)
Internal
C-terminal
ρ = 0.76 (p = 1.1e-4)
Detached
I, V, W
C-terminal

293

294 **Figure 3**. Structural context modulates the effect of $Poly_{10}X$, while its overall
295 neutrality trend is conserved.

296 **A**) Schematic representation of the constructs used to examine the effect of $Poly_{10}X$ in different
297 structural contexts. In the "$Poly_{10}X$ between two FPs" construct, a $Poly_{10}X$–GS linker sequence was
298 inserted between EGFP and mCherry. In the "internal $Poly_{10}X$" construct, $Poly_{10}X$ was inserted between
299 residues 173 and 174 of EGFP. In the "detached $Poly_{10}X$" construct, a P2A self-cleaving sequence was
300 placed between EGFP and $Poly_{10}X$.

301 **B, C**) Maximum growth rate and maximum fluorescence intensity (**B**), and relative neutrality (**C**) of *S.*
302 *cerevisiae* cells overexpressing the $Poly_{10}X$ between two FPs construct in SC–LU medium at 30 °C.
303 Growth and fluorescence curves are shown in Figure S12B.

304 **D**, **E**) Maximum growth rate and maximum fluorescence intensity (**D**), and relative neutrality (**E**) of
305 cells expressing the internal $Poly_{10}X$ construct in SC–LU medium at 30 °C. Growth and fluorescence
306 curves are shown in Figure S13B.

307 **F**, **G**) Maximum growth rate and maximum fluorescence intensity (**F**), and relative neutrality (**G**) of
308 cells overexpressing the detached $Poly_{10}X$ construct in SC–LU medium at 30 °C. Growth and
309 fluorescence curves are shown in Figure S14B.

310 **H**) Cross-comparison of $Poly_{10}X$ neutrality profiles among different structural contexts. Each value
311 represents the Spearman's rank correlation coefficient calculated from the relative neutrality indices
312 obtained in each construct.

313 **I**) Pairwise comparison of $Poly_{10}X$ neutrality trends between different structural contexts. Scatter plots
314 show correlations of average relative neutrality values across constructs. Spearman's rank correlation
315 coefficient (ρ) and its p-values are shown. Single-letter codes indicate the amino acid repeated in
316 $Poly_{10}X$.

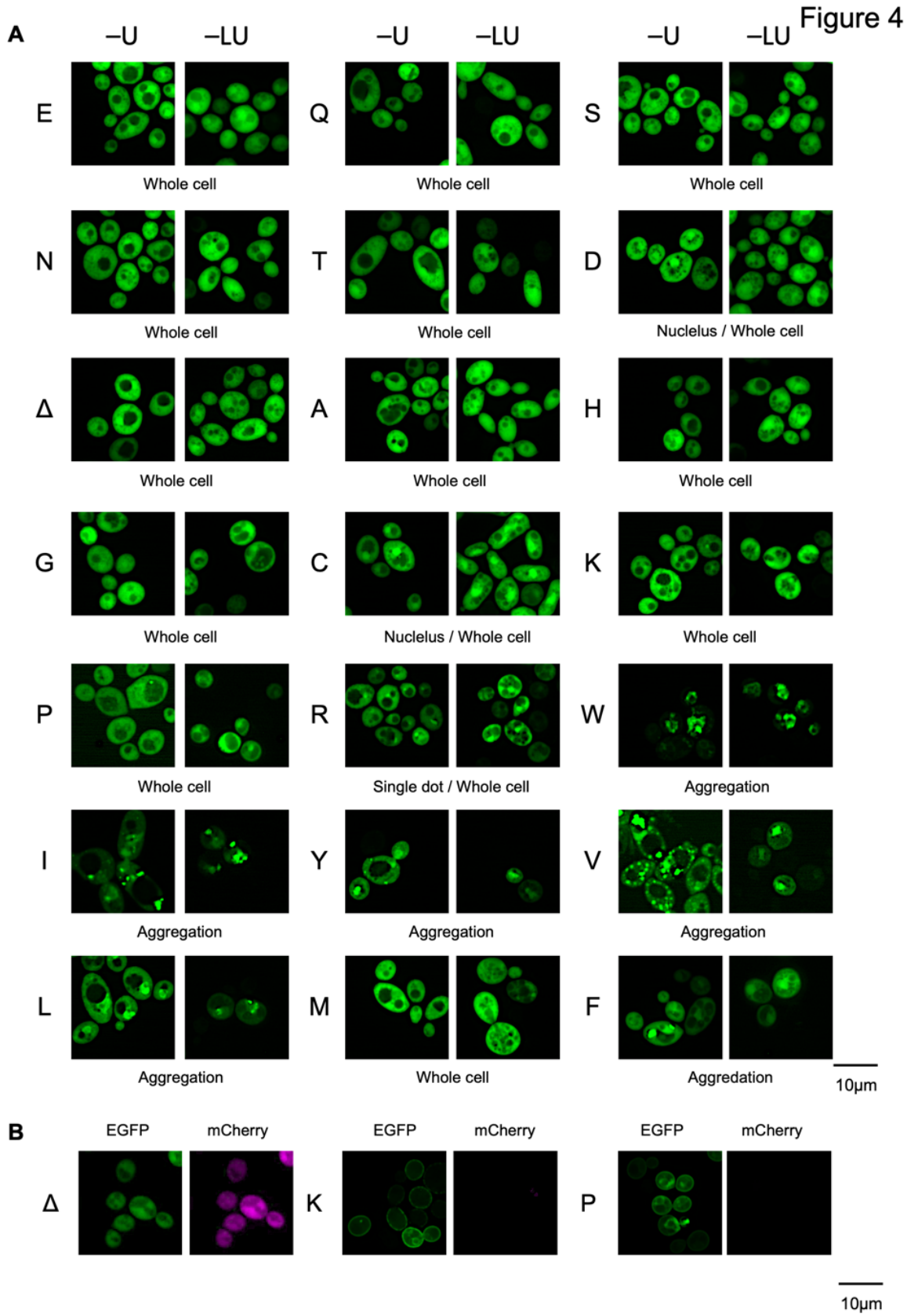
317 In **B**–**G** bars, dots, and error bars represent the mean, individual values, and standard deviation from at
318 least three biological replicates.

319

320 Poly$_{10}$X induces protein relocalization and aggregate formation

321 Next, to investigate the effects of Poly$_{10}$X on cellular functions, we observed the intracellular
322 localization of EGFP–Poly$_{10}$X using fluorescence microscopy (Figures 4A and S15–S16). As a result,
323 highly harmful and hydrophobic amino acids—W, I, Y, V, L, and F, except for M—were observed as
324 intracellular aggregates. The appearance of these aggregates varied among amino acids, suggesting that
325 they form distinct types of intracellular assemblies. D and C were enriched in the nucleus, whereas R
326 formed a single bright punctum. Other amino acids showed diffuse cytoplasmic localization similar to
327 the control. These results suggest that highly harmful Poly$_{10}$X sequences exert their harmfulness through
328 aggregation propensity and interactions with cellular structures. Notably, these localization patterns
329 were largely consistent with the subcellular localization of YFP–PolyX observed in COS-7 cells by
330 Oma *et al.* [27], indicating that the intracellular behavior of PolyX is conserved among eukaryotic cells.

331 We also examined the localization of constructs in which Poly$_{10}$X was inserted between EGFP and
332 mCherry (Figures S17–S19). For most amino acids, EGFP and mCherry showed identical localization
333 patterns, consistent with those observed for EGFP–Poly$_{10}$X. In contrast, in the Poly$_{10}$K and Poly$_{10}$P
334 constructs, EGFP localized to the cell surface, while mCherry fluorescence was barely detectable
335 (Figure 4B). Because such behavior was not observed in EGFP–Poly$_{10}$K or EGFP–Poly$_{10}$P (Figure 4A),
336 some Poly$_{10}$X can exhibit distinct behaviors depending on the structural context.



337

338 **Figure 4.** Poly$_{10}$X induces protein relocalization and aggregate formation

339 **A**) Fluorescence microscopy images of *S. cerevisiae* cells expressing EGFP–Poly$_{10}$X under the control
340 of *TDH3$_{pro}$*. Cells were pre-cultured in SC–U medium and then cultured overnight in SC–U or SC–LU
341 medium before imaging shown as –U and –LU. Brightness and contrast were adjusted to allow clear
342 visualization of cell morphology. Indicated subcellular localization and aggregation were categorized
343 by visual inspection.

344 **B**) Fluorescence microscopy images of *S. cerevisiae* cells expressing EGFP–mCherry, EGFP–Poly$_{10}$K–
345 mCherry, and EGFP–Poly$_{10}$P–mCherry under the control of the *WTC$_{846}$* promoter, observed 6 hours
346 after aTc induction. No mCherry fluorescence was detected in cells expressing EGFP–Poly$_{10}$K–
347 mCherry, and EGFP–Poly$_{10}$P–mCherry.

348 Single-letter codes indicate the amino acid repeated in Poly$_{10}$X. Δ represents the protein without a
349 Poly$_{10}$X fusion.

350

351

352 Poly$_{10}$E reduces the harmful effects of protein overexpression through
353 aggregation suppression.

354 As described above, beneficial Poly$_{10}$X sequences are thought to alleviate the harmful effects caused
355 by the overexpression of fluorescent proteins. Overexpressed fluorescent proteins can exert toxicity
356 through multiple mechanisms [55]. In the case of EGFP, overexpression leads to aggregate formation due
357 to misfolding and imposes a burden on proteostasis [56]. Since these adverse effects become more
358 pronounced at high temperatures, we evaluated the neutrality of EGFP-Poly$_{10}$X at 38 °C. As a result,
359 the growth of most Poly$_{10}$X-overexpressing strains, including EGFP (Δ), which had been able to grow
360 at 30 °C, was markedly reduced (Figure S20). In contrast, appending Poly$_{10}$E, Poly$_{10}$G, Poly$_{10}$Q, or
361 Poly$_{10}$S maintained growth even under this condition, resulting in a substantial increase in relative

362 neutrality (Figure 5A and Figure S20), with $Poly_{10}E$ showing the most pronounced increase in relative
363 neutrality (Figure 5B).

364 The adverse effects on proteostasis caused by overexpression are limited in the case of moxGFP, whose
365 folding properties have been improved compared to EGFP [56]. Indeed, all moxGFP-$Poly_{10}X$
366 overexpression strains that were able to grow at 30 °C (Figure S5) also grew at 38 °C, and their relative
367 neutrality did not change significantly (Figure 5C, 5D, and S21). In contrast, although $Poly_{10}E$ did not
368 show any beneficial effect in the 30 °C experiments using moxGFP (Figure S6), it became beneficial at
369 38 °C (Figure 5C and 5D). It is likely that moxGFP, which has stable folding at 30 °C, does not benefit
370 from $Poly_{10}E$ under that condition, but as the structure becomes less stable at 38 °C, the beneficial effect
371 of $Poly_{10}E$ emerges.

372 Next, we examined whether the addition of $Poly_{10}E$ mitigates the formation of protein aggregates,
373 Hsp70 aggregates, morphological abnormalities in mutant strains, and the induction of the heat shock
374 response that occur upon EGFP overexpression [56]. The amount of insoluble EGFP-$Poly_{10}E$ in
375 overexpressing cells was significantly lower than that of insoluble EGFP (Figure 5E, 5F, and S22). The
376 number of Hsp70 aggregates in EGFP-$Poly_{10}E$–overexpressing cells showed a decreasing trend, though
377 not statistically significant, compared to EGFP-overexpressing cells (Figure 5G), and images showing
378 more dispersed aggregates were frequently observed (Figure 5H and S23). Furthermore, the
379 morphological abnormalities of the *cdc24-5* and *rpl19aΔ* strains induced by EGFP overexpression were
380 not observed in EGFP-$Poly_{10}E$–overexpressing cells (Figure 5I, S24, and S25). Finally, transcriptome
381 analysis by RNA-seq revealed that the heat shock response and the induction of the proteasome-related
382 transcription factor *RPN4*, both triggered by EGFP overexpression, were markedly reduced in EGFP-
383 $Poly_{10}E$ (Figure 5J and S26A). In contrast, EGFP-$Poly_{10}D$, which does not show a growth advantage
384 over EGFP (Figure 1C), exhibited only limited suppression of the heat shock response and *RPN4*
385 induction (Figure 5K and S26A). Collectively, all these results support the conclusion that the addition
386 of $Poly_{10}E$ to EGFP alleviates the proteostatic burden caused by misfolding during EGFP
387 overexpression.

388 Notably, when one of the most harmful $Poly_{10}X$ variants, $Poly_{10}I$, was appended to EGFP, it induced a
389 strong heat shock response and *RPN4* activation, opposite to the effect observed with $Poly_{10}E$ (Figure
390 5L and S26B). Therefore, $Poly_{10}I$ is suggested to either promote misfolding of EGFP more strongly or
391 aggregate on its own, thereby disrupting intracellular proteostasis. These contrasting effects suggest that
392 the physicochemical properties of the appended poly-amino acid strongly influence protein folding and
393 proteostasis.

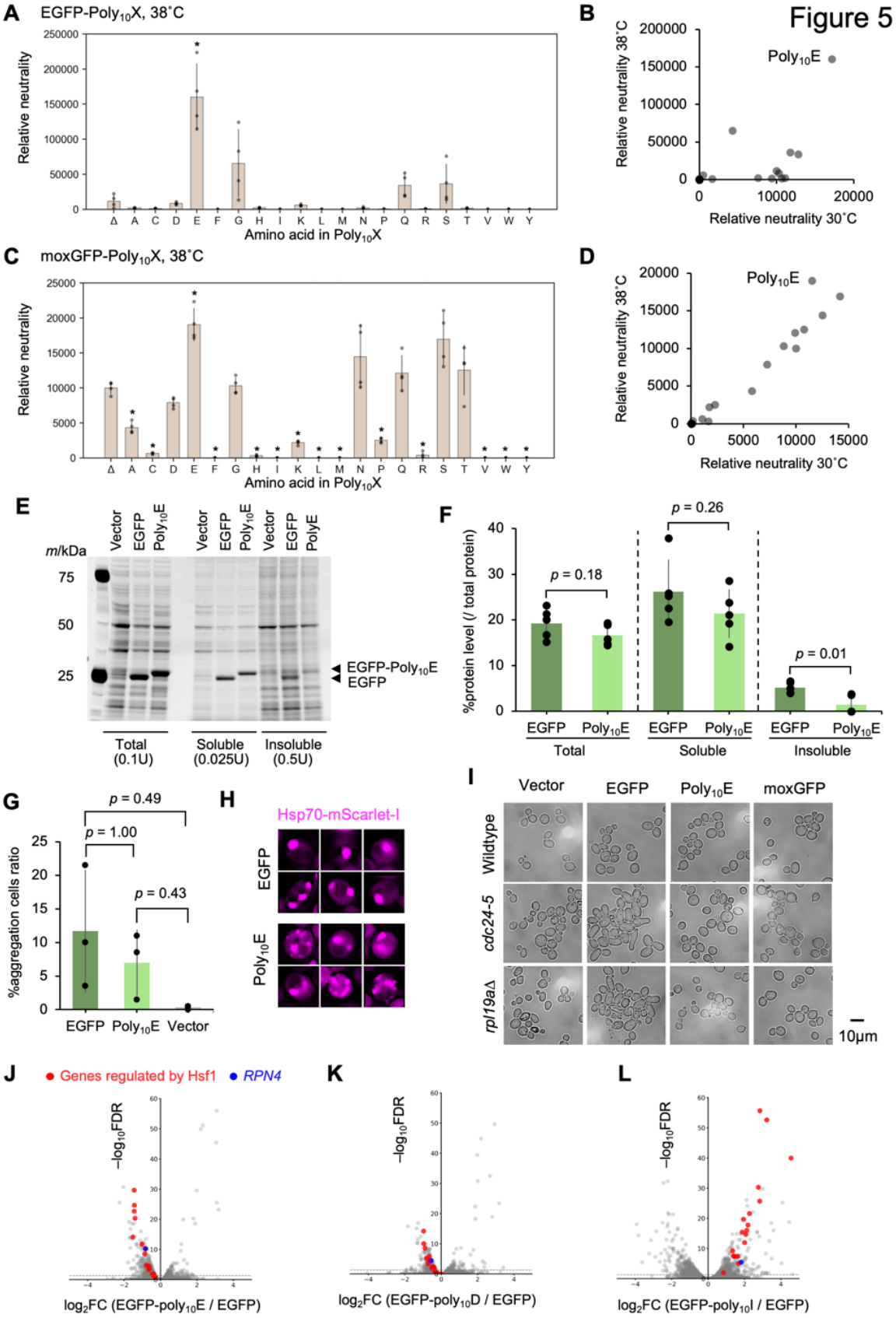


Figure 5
A
EGFP-Poly10X, 38˚C
Relative neutrality
Amino acid in Poly10X
B
Poly10E
Relative neutrality 38˚C
Relative neutrality 30˚C
C
moxGFP-Poly10X, 38˚C
D
E
m/kDa
Vector
EGFP
Poly10E
PolyE
EGFP-Poly10E
EGFP
Total (0.1U)
Soluble (0.025U)
Insoluble (0.5U)
F
%protein level (/ total protein)
p = 0.18
p = 0.26
p = 0.01
Total
Soluble
Insoluble
G
%aggregation cells ratio
p = 0.49
p = 1.00
p = 0.43
H
Hsp70-mScarlet-I
I
Wildtype
cdc24-5
rpl19aΔ
moxGFP
10μm
J
Genes regulated by Hsf1
RPN4
–log10FDR
log2FC (EGFP-poly10E / EGFP)
K
log2FC (EGFP-poly10D / EGFP)
L
log2FC (EGFP-poly10I / EGFP)

394

395 **Figure 5.** Poly$_{10}$E reduces the harmful effects of protein overexpression through
396 aggregation suppression.

397 **A**) Relative neutrality of EGFP-Poly$_{10}$X under heat stress. Cells were cultured in SC–LU medium at
398 38 °C. Growth and fluorescence data are shown in Figure S20.

399 **B**) Comparison of the relative neutrality of EGFP–Poly$_{10}$X at 30 °C (from Figure 1C) and 38 °C (from
400 **A**). Average values are used, and the plot of EGFP-Poly$_{10}$E is shown as Poly$_{10}$E.

401 **C**) Relative neutrality of moxGFP-Poly$_{10}$X under heat stress. Cells were cultured in SC–LU medium at
402 38 °C. Growth and fluorescence data are shown in Figure S21.

403 **D**) Comparison of the relative neutrality of moxGFP–Poly$_{10}$X at 30 °C (from Figure S5D) and 38 °C
404 (from **C**). Average values are used, and the plot of moxGFP-Poly$_{10}$E is shown as Poly$_{10}$E.

405 **E**) Representative image of SDS–PAGE of total, soluble, and insoluble protein fractions from cells
406 overexpressing vector, EGFP, and EGFP–Poly$_{10}$E cultured in SC–LU medium. Proteins were visualized
407 by fluorescent dye staining. The arrowhead indicates the band corresponding to EGFP or EGFP–Poly$_{10}$E.

408 **F**) Quantification of target protein levels (% of total protein) based on the band intensity in **E** and Figure
409 S22. Only monomeric bands were quantified; therefore, polymerized EGFP or EGFP–Poly$_{10}$E may not
410 be accurately represented. Bars, dots, and error bars represent the mean, individual values, and standard
411 deviation from five biological replicates. Statistical comparisons were performed by Welch's t-test. The
412 raw gel images used for the quantification are shown in Figure S22.

413 **G**) Percentage of cells forming Hsp70 aggregates among those overexpressing vector, EGFP, or EGFP–
414 Poly$_{10}$E. The *p*-values were calculated by Welch's t-test with Bonferroni correction. For each of three
415 biological replicates, 200 cells per sample were randomly selected, and aggregate formation was
416 visually assessed under blinded conditions.

417 **H**) Fluorescent microscopic images of Hsp70 foci in cells overexpressing EGFP or EGFP–Poly$_{10}$E.
418 Hsp70/Ssa1–mScarlet-I was genomically integrated to monitor Hsp70 aggregate formation. Image
419 brightness and contrast were adjusted to enhance visibility of aggregates. Enlarged images are shown
420 in Figure S23.

421 **I**) Microscopic observation of morphological phenotypes in various yeast strains overexpressing vector,
422 EGFP, EGFP–Poly$_{10}$E, or moxGFP. Wild-type strain BY4741 and mutant strains (*cdc24-5*, *rpl19aΔ*)
423 were cultured in SC–LU medium for 18 h before imaging. Representative images, the morphological
424 quantification method, and the quantitative results for all seven mutants are shown in Figures S24 and
425 S25.

426 **J**–**L**) Transcriptomic analysis of *S. cerevisiae* cells overexpressing EGFP or EGFP–Poly$_{10}$X constructs
427 by RNA-seq. Volcano plots show differential gene expression between cells expressing EGFP and
428 EGFP–Poly$_{10}$E (**J**), EGFP–Poly$_{10}$D (**K**), and EGFP–Poly$_{10}$I (**L**). Genes regulated by the heat-shock
429 transcription factor Hsf1 and the proteasome stress-response gene *RPN4* are highlighted. Cells
430 expressing EGFP–Poly$_{10}$E and EGFP–Poly$_{10}$D were cultured in SC–LU medium, while those
431 expressing EGFP–Poly$_{10}$I were cultured in SC–U together with the EGFP control.

432

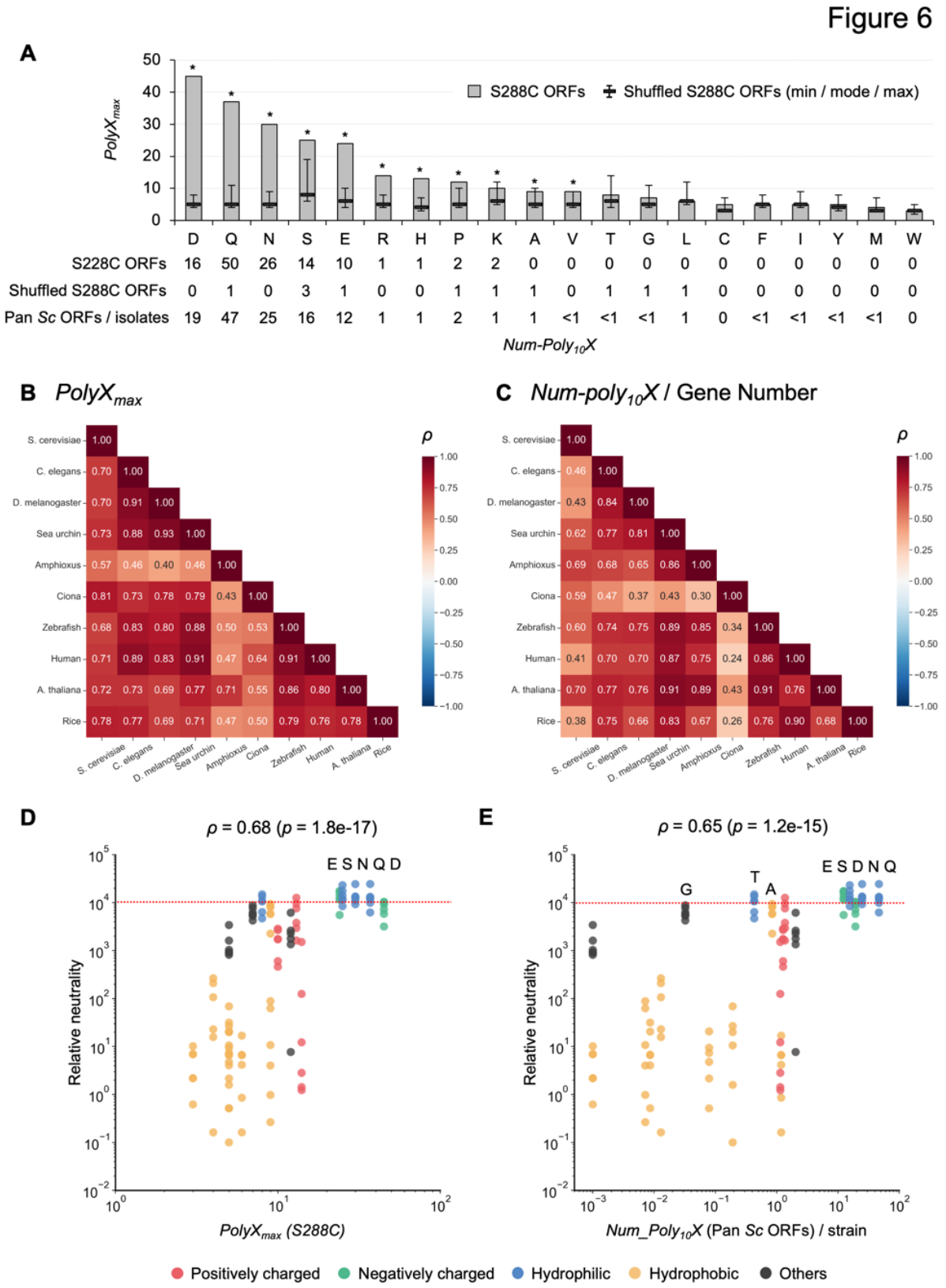
433 The neutrality of Poly$_{10}$X mirrors its evolutionary usage in proteomes

434 To explore whether the experimentally observed neutrality of Poly$_{10}$X is reflected in natural proteomes,
435 we first analyzed the maximum number of consecutive residues for each of the 20 amino acids
436 (PolyX$_{max}$) and the number of sequences containing ten or more consecutive identical residues (Num-
437 Poly$_{10}$X) in the *S. cerevisiae* S288C reference ORFs. We found that both PolyX$_{max}$ and Num-Poly$_{10}$X
438 showed clear biases (Figure 6A). About half of the amino acids never appeared as stretches longer than
439 ten residues. To test whether the observed PolyX$_{max}$ values are maintained by natural selection, we
440 compared them with those obtained from a simulation using randomized ORF sequences (Shuffled

441 S288C ORFs). This analysis revealed that D, Q, N, S, E, R, H, P, K, A, and V displayed significantly
442 longer consecutive runs than expected ($q < 0.001$, Monte Carlo method with FDR correction). The other
443 amino acids fell within the expected range, and no clear bias toward increasing or decreasing repeat
444 length was detected. Next, using ORFs from 1,392 *S. cerevisiae* isolates (pan *Sc* ORFs) [57], we calculated
445 $PolyX_{max}$ and Num-$Poly_{10}X$ with higher resolution. In the *S. cerevisiae* S288C reference strain, the
446 amino acids A, V, T, G, L, F, I, Y, and M, which showed zero Num-$Poly_{10}X$ values, appeared as $Poly_{10}X$
447 sequences at very low frequencies (≤ 1); $Poly_{10}X$ of C and W were never observed (Figure 6A).

448 To examine whether the PolyX occurrence patterns observed in *S. cerevisiae* are shared across other organisms,
449 we analyzed nine species (*C. elegans*, *D. melanogaster*, sea urchin, amphioxus, *Ciona*, zebrafish, human, *A.*
450 *thaliana*, and rice) for $PolyX_{max}$ and Num-$Poly_{10}X$ normalized by gene number, and compared their frequencies
451 among species. A comparison of $PolyX_{max}$ between *S. cerevisiae* and the other species showed a high correlation
452 coefficient of 0.57 or greater, suggesting that the overall occurrence pattern of $PolyX_{max}$ is largely conserved
453 across species (Figure 6B). In contrast, a similar comparison using Num-$Poly_{10}X$ yielded a lower correlation of
454 0.38 or higher (Figure 6C). Whether this lower correlation reflects species-specific characteristics or results from
455 annotation errors in ORF datasets remains to be determined.

456 Finally, we compared the occurrence patterns of PolyX observed in ORFs with the neutrality of $Poly_{10}X$
457 determined in our experiments, in which each $Poly_{10}X$ was fused to the C-terminus of a fluorescent protein (Figure
458 2). As shown in Figure 6D and 6E, both $PolyX_{max}$ and Num-$Poly_{10}X$ in the pan-ORF dataset showed strong
459 correlations with neutrality. Five amino acids—E, S, N, Q, and D—were clearly separated from the others,
460 exhibiting longer $PolyX_{max}$ values and higher occurrence numbers in $Poly_{10}X$. These amino acids also displayed
461 beneficial effects when fused to the C-terminus of certain fluorescent proteins (neutrality >10,000). In contrast,
462 hydrophobic amino acids with low occurrence frequencies generally showed low neutrality and high harmfulness.
463 Exceptions were A, G, and T, which exhibited high neutrality despite having an average Num-$Poly_{10}X$ of less
464 than one in the pan *Sc* ORF dataset. The probability that these amino acids (A, G, and T) would randomly appear
465 as ten consecutive residues is low (Figure 6A). Therefore, their behavior may represent that of amino acids that
466 are neutral and lack inherent beneficial effects. Taken together, these findings indicate that $Poly_{10}X$ neutrality, as
467 determined experimentally, closely explains the occurrence patterns of amino acid homorepeats in natural
468 proteomes.



469

470 **Figure 6.** The neutrality of $Poly_{10}X$ mirrors its evolutionary usage in proteomes.

A）Analysis of PolyX occurrence patterns ($PolyX_{max}$ and Num–$Poly_{10}X$) in *S. cerevisiae*. The bar graph shows the maximum number of consecutive identical residues ($PolyX_{max}$) for each amino acid in S288C ORFs. Shuffled S288C ORFs represent the minimum, mode, and maximum values obtained from 10,000 random simulations. Asterisks indicate amino acids that exhibit significantly longer homorepeat lengths than expected from simulation ($q < 0.001$, Monte Carlo test with FDR correction). Numbers below the graph indicate the counts of proteins containing ≥10 consecutive identical residues (Num–$Poly_{10}X$) for S288C ORFs, shuffled S288C ORFs, and pan *Sc* ORFs/isolates (from top to bottom). Values for pan-*Sc* ORFs/isolates are rounded to one decimal place.

B）Spearman's rank correlation coefficient (ρ) of $PolyX_{max}$ values among different species. The underlying numerical data used for the calculation are provided in Figure S27.

C）Spearman's rank correlation coefficients (ρ) of Num–$Poly_{10}X$/Gene_number across species. The underlying numerical data used for the calculation are provided in Figure S27.

D）Relationship between $PolyX_{max}$ in S288C ORFs and experimentally measured relative neutrality. Colors indicate the chemical properties of each amino acid.

E）Relationship between Num–$Poly_{10}X$ in Pan *Sc* ORFs/isolates and experimentally measured relative neutrality. For cysteine (C) and tryptophan (W), values of 0 were plotted as 0.001 for visualization purposes.

In **D** and **E**, Spearman's rank correlation coefficient (ρ) and its p-value are shown. Colors indicate the physicochemical properties of each amino acid.

Discussion

In this study, we systematically evaluated the neutrality (harmful and beneficial properties) of simple amino acid sequences in which a single residue is repeated ten times ($Poly_{10}X$). We found that neutrality varied markedly among amino acids, with hydrophobic residues exhibiting pronounced harmfulness (Figures 1 and 2). In contrast, several hydrophilic or negatively charged amino acids (e.g., E, S, N, Q) showed beneficial effects on cellular tolerance to protein overexpression, which was particularly intriguing. Notably, the harmfulness of $Poly_{10}I$, $Poly_{10}V$, and $Poly_{10}W$ persisted even when detached from EGFP (Figure 3), indicating that these repeats possess intrinsic harmfulness. Conversely, the beneficial effects of $Poly_{10}E$, $Poly_{10}Q$, $Poly_{10}N$, and $Poly_{10}S$ disappeared upon detachment from EGFP (Figure 3), indicating that these repeats reduce cellular burden by mitigating the toxicity of the host protein. Mechanistically, the harmful effect of $Poly_{10}I$ appears to stem from its intrinsic aggregation propensity, whereas the beneficial effect of $Poly_{10}E$ is likely due to its ability to prevent aggregation (Figures 4 and 5).

Furthermore, when we compared the occurrence frequencies of PolyX across the proteomes of diverse organisms, we found that their cross-species occurrence profiles were remarkably similar across species and showed strong correlation with the experimentally determined neutrality values (Figure 6). Notably, hydrophobic residues, which rarely form homorepeats in natural proteins, exhibited high harmfulness in our assays, suggesting that such sequences are likely eliminated immediately upon emergence during evolution. In contrast, amino acids such as E, N, Q, and S—whose homorepeats are widely observed in natural proteins—showed low harmfulness and, in some cases, even beneficial effects, implying that their appearance may confer an advantage to cells from the outset. PolyQ, for example, can contribute to normal protein function by modulating protein–protein interactions when maintained at moderate lengths, but become pathogenic upon excessive expansion [39]. Thus, the evolutionarily permissible range of PolyQ length is likely determined by a balance between harm and benefit.

Taken together, the occurrence frequencies of amino acid sequences in natural proteomes are well explained by experimentally determined neutrality, reflecting a balance between harmful and beneficial effects. Our assays were performed by placing simple homorepeat sequences at the C-terminus or within internal loops of fluorescent proteins and expressing them in the cytosol. Because homorepeats found in natural proteins predominantly occur in intrinsically disordered regions (Figure S29), our experimental design provides an appropriate framework for

evaluating their harmful and beneficial effects. In contrast, whether similar principles apply to more complex sequences buried within membranes or folded protein cores remains an open question. For structured or interaction-dependent motifs, intrinsic harmfulness might be masked—or even neutralized—by specific binding partners. This idea raises the possibility that "eliminating harmfulness" itself acts as an evolutionary pressure driving the formation of higher-order structures or interaction modules. Interestingly, hydrophobic homorepeats—among the most harmful sequences in our assays—are also utilized as secretion signals [58]. It is tempting to speculate that what was originally a "dangerous" sequence requiring removal from the cytosol may have been evolutionarily co-opted into the secretory pathway.

The design of artificial proteins is rapidly advancing. However, it remains difficult to determine whether sequences absent in nature are missing because they offer no benefit or because they are eliminated due to harmfulness. The parameter we identified in this study—neutrality, which integrates both harmful and beneficial effects—may represent a fundamental evolutionary design principle that shapes the compatible amino acid sequence space in living organisms. A deeper understanding of this parameter will aid in rational protein design and may also shed light on the molecular mechanisms underlying neurodegenerative diseases and aging, in which mutations elevate protein harmfulness.

Materials and Methods

Strains, plasmids, growth conditions

S. cerevisiae and *E. coli* strains and plasmids used in this study are listed in the Key Resources Table. Synthetic complete (SC) medium lacking uracil (U) or leucine (L) was used for yeast cultures. LB medium was used for *E. coli* cultures.

Genetic tug-of-war (gTOW) method

The gTOW method [50,51] enables artificial elevation of plasmid copy number for a target protein by exploiting the combination of a multi-copy plasmid replication origin (2μ ORI) and the auxotrophic selection marker *leu2-89*. In SC–U medium, where no copy-number–increasing selection pressure is present, the plasmid remains at a low copy number (approximately 30 copies for an empty vector). In contrast, in SC–LU medium, strong selection pressure increases the plasmid copy number to approximately 150 copies. As plasmid copy number increases, the expression level of the target protein also rises. If expression becomes high enough to inhibit cell growth before the plasmid reaches ~150 copies, this expression level is defined as the protein's maximum tolerable expression level. Thus, gTOW allows quantitative determination of how much of a given protein a yeast cell can tolerate before growth inhibition occurs. By combining the measured expression level with the degree of growth inhibition, protein harmfulness can be represented as a one-dimensional parameter, relative neutrality, enabling direct comparison of harmfulness across different proteins [55]. In this study, modified fluorescent proteins were expressed under either the strong *TDH3* promoter or the WTC_{846} ($P_{7.tet1}$) promoter, a *TDH3*-derived variant whose expression can be induced by anhydrotetracycline (aTc) [52].

556 Measurement of growth rate, fluorescence intensity, and calculation of the
557 relative neutrality in yeast.

558 Cells of the BY4741 or BYW2 strain carrying the expression plasmids were pre-cultured in SC–U and
559 then transferred to either SC–LU (for constructs expressed under $TDH3_{pro}$ in BY4741) or SC–LU
560 supplemented with aTc (for constructs expressed under the WTC_{846} promoter in BYW2). Cells were
561 cultured under these conditions while OD_{595} and fluorescence signals were monitored every 10 or 30
562 minutes using an Infinite F Nano+ microplate reader. Fluorescence of EGFP, moxGFP, mNeonGreen, and
563 Gamillus was measured using an excitation/emission filter set of 485/535 nm, whereas fluorescence of mScarlet-
564 I and mCherry was measured using a 535/590 nm filter set. From these time-course measurements, the
565 maximum growth rate (MGR) and maximum fluorescence intensity (MFI) were determined using a
566 custom Python script. The relative neutrality index was calculated by multiplying the percentage of
567 MGR relative to the control vector (Δ) by the percentage of MFI relative to the control fluorescent
568 protein as follows: Relative neutrality = %(MGR_$Poly_{10}$X / MGR_Δ) × %(MFI_$Poly_{10}$X / MFI_Δ).

569

570 Measurement of growth rate, fluorescence intensity, and calculation of the
571 relative neutrality in *E. coli*.

572 Cells of the BW25113 strain carrying the expression plasmids were pre-cultured in LB + ampicillin
573 medium and then transferred to fresh LB + ampicillin medium. Cells were cultured under these
574 conditions while OD_{595} and fluorescence (485/535 nm) signals were monitored every 5 minutes using an
575 Infinite F Nano+ microplate reader. From these time-course measurements, the maximum growth rate
576 (MGR) and maximum fluorescence intensity (MFI) were determined using a custom Python script. The
577 relative neutrality index was calculated by multiplying the percentage of MGR relative to the control
578 vector (Δ) by the percentage of MFI relative to the control fluorescent protein as follows: Relative
579 neutrality = %(MGR_$Poly_{10}$X / MGR_Δ) × %(MFI_$Poly_{10}$X / MFI_Δ).

580

581 Clustering analysis

582 Clustering analysis was performed using a custom Python script (pandas, seaborn, and matplotlib) after
583 organizing the relative neutrality data in an Excel file. Hierarchical clustering and heatmap visualization
584 were conducted using the clustermap function in seaborn. Euclidean distance (metric = "euclidean")
585 was used as the distance metric, and average linkage (UPGMA; method = "average") was applied for
586 clustering. The resulting clustering patterns were visualized as row- and column-wise heatmaps, with
587 dendrograms shown or hidden depending on the purpose of the analysis. A custom continuous colormap
588 defined by three reference points was used, and the display range was fixed from 0 to 20,000. This
589 scaling was chosen because relative neutrality values were normalized such that the control (Δ)
590 corresponded to 10,000, representing the non-toxic state; accordingly, the color scale was centered at
591 10,000.

592

593 Amino acid parameters

594 For the correlation analysis in Figure S10, a total of 28 amino acid parameters were used, consisting of
595 18 nonredundant indices from the AAindex database (https://www.genome.jp/aaindex/; see
596 https://doi.org/10.1002/pro.5239) and 10 additional parameters: hydrophobicity [59], hydropathy index [60],
597 isoelectric point, side-chain molecular weight, amino acid usage (%) (this study), metabolic cost [61],
598 biosynthetic cost (ATPs) [62], biosynthetic steps (enzymes) [62], and total energy costs of amino acids and
599 nucleotide precursors under fermentative or respiratory conditions [63].

600

601 Microscopic observation

602 Yeast cells overexpressing the target protein were cultured in either SC–U or SC–LU medium, and
603 imaged using a DMI6000B microscope (Leica Microsystems), and images were processed using the

604 Leica Application Suite X software. GFP and RFP fluorescence signals were acquired using a GFP
605 filter cube and RFP filter cube, respectively.

606 For the morphology analysis in Figure S24 and S25, microscopic images were processed using Cellpose
607 [64] for cell segmentation. The segmented images were subsequently analyzed using the
608 *MeasureObjectSizeShape* module in CellProfiler [65] to quantify the major and minor axes of individual
609 cells. Elongation ratio values were calculated for each cell using the major and minor axis lengths
610 obtained from the *MeasureObjectSizeShape* module, either with a custom Python script or in Excel,
611 according to the following formula: Elongation Ratio = MajorAxisLength / MinorAxisLength. Cells
612 with an elongation ratio ≥ 1.5 were defined as morphologically abnormal.

613

614 Protein analysis

615 BY4741 strains expressing the target proteins were cultured overnight in 6 mL or 25 mL of SC–LU
616 medium. Total protein was extracted from cells in the logarithmic growth phase (OD_{660} = 0.9–1.0) using
617 0.2 mol/L NaOH, followed by solubilization in NuPAGE sample buffer (Thermo Fisher Scientific). For
618 each analysis, total protein was extracted from a cell mass corresponding to 1 OD unit at OD_{660}.

619 For the cell lysate analysis, cells were washed with PBST [10 mM phosphate-buffered saline (pH 7.4),
620 0.001% Tween 20] supplemented with Halt protease inhibitor cocktail (Thermo Fisher Scientific), then
621 disrupted with glass beads using a bead-beating homogenizer (Micro Smash MS-100, TOMY) at 5000
622 rpm for 30 s, repeated five times. Cell lysates were centrifuged at 15,000 rpm for 10 min to separate
623 soluble and insoluble fractions. The insoluble fraction was washed with 1 mL PBST and centrifuged
624 again at 15,000 rpm for 10 min.

625 Extracted proteins were fluorescently labeled using EZLabel Fluoro NEO (ATTO) according to the
626 manufacturer's instructions and separated by SDS–PAGE on 4–12% gradient gels. Fluorescent signals
627 were detected using the SYBR Green fluorescence mode of a LAS-4000 image analyzer (GE

628 Healthcare). Proteins separated by SDS–PAGE were transferred onto a PVDF membrane (Thermo
629 Fisher Scientific). GFP signals were detected using an anti-GFP primary antibody (Roche), an
630 peroxidase-conjugated secondary antibody (Nichirei Bioscience), and chemiluminescent substrate
631 (Thermo Fisher Scientific). Chemiluminescent images were acquired with the chemiluminescence
632 mode of the LAS-4000 image analyzer.
633

634 RNA sequencing analysis

635 *S cerevisiae* strain BY4741 with the control vector or overexpressing EGFP, EGFP–$Poly_{10}D$, or EGFP–
636 $Poly_{10}E$ was cultured in SC–LU medium, and strain BY4743 with the vector or overexpressing EGFP,
637 or EGFP–$Poly_{10}I$ was cultured in SC–U medium. Cells were collected at the logarithmic growth phase
638 ($OD_{660} \approx 1.0$). Total RNA was extracted according to a previously described protocol [66]. The
639 concentration of purified RNA was initially measured using a Qubit fluorometer (Thermo Fisher
640 Scientific), and samples were stored at −80 °C until further use. Before library preparation, RNA
641 concentrations were re-measured using the Quant-iT RiboGreen RNA Assay Kit (Invitrogen) on an
642 ARVO Multimode Microplate Reader (PerkinElmer), and RNA quality was assessed with a 2100
643 Bioanalyzer (Agilent Technologies). cDNA libraries were prepared using the TruSeq Stranded mRNA
644 Library Prep Kit (Illumina) and sequenced on a NovaSeq X Plus platform (Illumina) with 100-bp
645 paired-end reads. Raw sequencing data were quality-checked using FastP [67] and aligned to the reference
646 genome with Hisat2 [68] . Aligned data were converted to BAM format with Samtools [69] and quantified
647 using StringTie [70]. Differential gene expression analysis was performed using EdgeR [71]. Gene Ontology
648 (GO) enrichment analysis was conducted with the clusterProfiler [72] and org.Sc.sgd.db packages (DOI:
649 10.18129/B9.bioc.org.Sc.sgd.db) in R to identify significantly enriched GO terms. All analyses were
650 performed using three biological replicates for each strain. The raw data were deposited into DDBJ
651 (accession number: PRJDB39951).
652

PolyX analysis in the *S. cerevisiae* proteome and other species

Proteome-wide PolyX analysis was performed using all verified ORFs registered in the *Saccharomyces* Genome Database (SGD), covering all 20 amino acids. As of October 2024, 6,585 ORFs are annotated in the *S. cerevisiae* (S288C) genome. From this set, 5,911 verified ORFs (Sc ORFs) were retained for analysis after removing those annotated as "Dubious." A custom Python script first loaded the ORF list and excluded all entries whose Qualifier field was labeled "Dubious." Regular expressions were then used to identify, for each ORF, the longest homorepeat for each amino acid. The resulting maximum repeat lengths were compiled into a dataframe, and two metrics were calculated as follows: $PolyX_{max}$ – the maximum length of consecutive identical amino acids for each amino acid across all ORFs, and $Num_Poly_{10}X$ – the number of ORFs containing a homorepeat of length ≥ 10 for each amino acid. For pan *Sc* ORFs, the same analysis was conducted using FASTA assemblies downloaded from Zenodo (https://zenodo.org/records/3407352), which include 1,392 *S. cerevisiae* genome assemblies [57].

Simulated (shuffled) S288C ORFs were generated as follows using a custom Python script. Each sequence of above 5,911 ORFs was randomly shuffled while preserving amino acid composition. This produced randomized sequences with identical amino acid compositions but altered residue order. $PolyX_{max}$ and $Num_Poly_{10}X$ values were computed for each shuffled ORF set using the same regular expression pipeline described above. This simulation was repeated 10,000 times, using seed values from 0 to 9,999. For each iteration, $PolyX_{max}$ and $Num_Poly_{10}X$ results were stored in a dataframe. For every amino acid, the minimum, mode, mean, maximum, and standard deviation across the 10,000 trials were calculated and summarized.

For cross-species comparisons, proteomes of nine eukaryotic species (*Caenorhabditis elegans, Drosophila melanogaster, Strongylocentrotus purpuratus, Branchiostoma lanceolatum, Ciona intestinalis, Danio rerio, Homo sapiens, Arabidopsis thaliana*, and *Oryza sativa*) were downloaded from Ensembl databases (https://asia.ensembl.org). When multiple isoforms were annotated for a given gene, the isoform with the longest amino acid sequence was selected as the representative sequence for analysis. For each amino acid sequence, PolyX of identical amino acids were detected based on regular-

expression matching. Two metrics were calculated for each of the 20 amino acids: (i) PolyXmax, defined as the maximum length of consecutive identical residues observed across all proteins, and (ii) Num–Poly$_{10}$X, defined as the number of proteins containing a homorepeat of length ≥10 residues. For interspecies comparisons, Num–Poly$_{10}$X values were normalized by the total number of genes analyzed for each species. Correlations between PolyX occurrence metrics (PolyXmax and Num–Poly$_{10}$X) and experimentally determined Poly$_{10}$X neutrality values were evaluated using Spearman's rank correlation coefficients.

Data availability statement

All source data used for figure generation, and all analysis scripts have been deposited on GitHub (https://github.com/hisaomlab/Murase_PolyX). RNA-seq data have been deposited in a public repository (DDBJ accession number: PRJDB39951). All plasmids used in this study are available from the National BioResource Project (NBRP)-yeast (https://yeast.nig.ac.jp/yeast/).

Acknowledgements

We thank the members of the Moriya laboratory (Okayama University) for helpful discussions, Yuhei Chadani (Okayama University) for providing the *E. coli* strain, and Christian Landry (Université Laval) for valuable comments and suggestions on the manuscript.

References

1. Dryden, D. T. F., Thomson, A. R. & White, J. H. How much of protein sequence space has been explored by life on Earth? *J. R. Soc. Interface* **5**, 953–956 (2008).
2. Barton, M. D., Delneri, D., Oliver, S. G., Rattray, M. & Bergman, C. M. Evolutionary systems biology of amino acid biosynthetic cost in yeast. *PLoS One* **5**, e11935 (2010).
3. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
4. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
5. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
6. Hayes, T. *et al.* Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).
7. Georgakopoulos-Soares, I., Yizhar-Barnea, O., Mouratidis, I., Hemberg, M. & Ahituv, N. Absent from DNA and protein: genomic characterization of nullomers and nullpeptides across functional categories and evolution. *Genome Biol.* **22**, 245 (2021).
8. Koulouras, G. & Frith, M. C. Significant non-existence of sequences in genomes and proteomes. *Nucleic Acids Res.* **49**, 3139–3155 (2021).
9. Keefe, A. D. & Szostak, J. W. Functional proteins from a random-sequence library. *Nature* **410**, 715–718 (2001).
10. Devlin, J. J., Panganiban, L. C. & Devlin, P. E. Random peptide libraries: a source of specific protein binding molecules. *Science* **249**, 404–406 (1990).
11. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10**, 866–876 (2009).
12. Fowler, D. M., Stephany, J. J. & Fields, S. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat. Protoc.* **9**, 2267–2284 (2014).

723 13. Diss, G. & Lehner, B. The genetic landscape of a physical interaction. *Elife* **7**, (2018).
724 14. Vaishnav, E. D. *et al.* The evolution, evolvability and engineering of gene regulatory DNA.
725 *Nature* **603**, 455–463 (2022).
726 15. Thompson, M. *et al.* Massive experimental quantification allows interpretable deep learning of
727 protein aggregation. *Sci. Adv.* **11**, eadt5111 (2025).
728 16. Castro, J. F. & Tautz, D. The effects of sequence length and composition of random sequence
729 peptides on the growth of E. coli cells. *Genes (Basel)* **12**, 1913 (2021).
730 17. Neme, R., Amador, C., Yildirim, B., McConnell, E. & Tautz, D. Random sequences are an
731 abundant source of bioactive RNAs or peptides. *Nat. Ecol. Evol.* **1**, 0217 (2017).
732 18. Chavali, S. *et al.* Constraints and consequences of the emergence of amino acid repeats in
733 eukaryotic proteins. *Nat. Struct. Mol. Biol.* **24**, 765–777 (2017).
734 19. Chavali, S., Singh, A. K., Santhanam, B. & Babu, M. M. Amino acid homorepeats in proteins.
735 *Nat. Rev. Chem.* **4**, 420–434 (2020).
736 20. Jorda, J. & Kajava, A. V. Protein homorepeats sequences, structures, evolution, and functions.
737 *Adv. Protein Chem. Struct. Biol.* **79**, 59–88 (2010).
738 21. Lobanov, M. Y., Sokolovskiy, I. V. & Galzitskaya, O. V. HRaP: database of occurrence of
739 HomoRepeats and patterns in proteomes. *Nucleic Acids Res.* **42**, D273–8 (2014).
740 22. Lobanov, M. Y. & Galzitskaya, O. V. Occurrence of disordered patterns and homorepeats in
741 eukaryotic and bacterial proteomes. *Mol. Biosyst.* **8**, 327–337 (2012).
742 23. Mier, P. & Andrade-Navarro, M. A. dAPE: a web server to detect homorepeats and follow their
743 evolution. *Bioinformatics* **33**, 1221–1223 (2017).
744 24. Mier, P. & Andrade-Navarro, M. A. PolyX2: Fast detection of homorepeats in large protein
745 datasets. *Genes (Basel)* **13**, 758 (2022).
746 25. Cascarina, S. M. & Ross, E. D. Proteome-scale relationships between local amino acid
747 composition and protein fates and functions. *PLoS Comput. Biol.* **14**, e1006256 (2018).
748 26. Oma, Y., Kino, Y., Sasagawa, N. & Ishiura, S. Comparative analysis of the cytotoxicity of
749 homopolymeric amino acids. *Biochim. Biophys. Acta* **1748**, 174–179 (2005).

750 27. Oma, Y., Kino, Y., Sasagawa, N. & Ishiura, S. Intracellular localization of homopolymeric
751 amino acid-containing proteins expressed in mammalian cells. *J. Biol. Chem.* **279**, 21217–21222
752 (2004).
753 28. Oma, Y., Kino, Y., Toriumi, K., Sasagawa, N. & Ishiura, S. Interactions between
754 homopolymeric amino acids (HPAAs). *Protein Sci.* **16**, 2195–2204 (2007).
755 29. Gojobori, J. & Ueda, S. Elevated evolutionary rate in genes with homopolymeric amino acid
756 repeats constituting nondisordered structure. *Mol. Biol. Evol.* **28**, 543–550 (2011).
757 30. Karlin, S., Brocchieri, L., Bergman, A., Mrazek, J. & Gentles, A. J. Amino acid runs in
758 eukaryotic proteomes and disease associations. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 333–338
759 (2002).
760 31. Faux, N. G. *et al.* Functional insights from the distribution and role of homopeptide repeat-
761 containing proteins. *Genome Res.* **15**, 537–551 (2005).
762 32. Qiu, Y., Xu, D., Lei, P., Li, S. & Xu, H. Engineering functional homopolymeric amino acids:
763 from biosynthesis to design. *Trends Biotechnol.* **42**, 310–325 (2024).
764 33. Courtemanche, N. Mechanisms of formin-mediated actin assembly and dynamics. *Biophys. Rev.*
765 **10**, 1553–1569 (2018).
766 34. Balhorn, R. The protamine family of sperm nuclear proteins. *Genome Biol* **8**, 227 (2007).
767 35. Guhaniyogi, J. & Brewer, G. Regulation of mRNA stability in mammalian cells. *Gene* **265**, 11–
768 23 (2001).
769 36. Ito-Harashima, S., Kuroha, K., Tatematsu, T. & Inada, T. Translation of the poly(A) tail plays
770 crucial roles in nonstop mRNA surveillance via translation repression and protein destabilization
771 by proteasome in yeast. *Genes Dev.* **21**, 519–524 (2007).
772 37. Terpe, K. Overview of tag protein fusions: from molecular and biochemical fundamentals to
773 commercial systems. *Appl. Microbiol. Biotechnol.* **60**, 523–533 (2003).
774 38. Futaki, S. *et al.* Arginine-rich peptides. *J. Biol. Chem.* **276**, 5836–5840 (2001).
775 39. Orr, H. T. & Zoghbi, H. Y. Trinucleotide repeat disorders. *Annu. Rev. Neurosci.* **30**, 575–621
776 (2007).

777 40. Dimitrova, L. N., Kuroha, K., Tatematsu, T. & Inada, T. Nascent peptide-dependent translation
778 arrest leads to Not4p-mediated protein degradation by the proteasome. *J. Biol. Chem.* **284**,
779 10343–10352 (2009).
780 41. Mizuno, M. *et al.* The nascent polypeptide in the 60S subunit determines the Rqc2-dependency
781 of ribosomal quality control. *Nucleic Acids Res.* **49**, 2102–2113 (2021).
782 42. Chadani, Y. *et al.* Intrinsic ribosome destabilization underlies translation and provides an
783 organism with a strategy of environmental sensing. *Mol. Cell* **68**, 528–539.e5 (2017).
784 43. Ito, Y. *et al.* Nascent peptide-induced translation discontinuation in eukaryotes impacts biased
785 amino acid usage in proteomes. *Nat. Commun.* **13**, (2022).
786 44. Haerty, W. & Golding, G. B. Low-complexity sequences and single amino acid repeats: not just
787 'junk' peptide sequences. *Genome* **53**, 753–762 (2010).
788 45. Ntountoumi, C. *et al.* Low complexity regions in the proteins of prokaryotes perform important
789 functional roles and are highly conserved. *Nucleic Acids Res.* **47**, 9998–10009 (2019).
790 46. Kashi, Y. & King, D. G. Simple sequence repeats as advantageous mutators in evolution. *Trends*
791 *Genet.* **22**, 253–259 (2006).
792 47. Pelassa, I. *et al.* Compound dynamics and combinatorial patterns of amino acid repeats encode a
793 system of evolutionary and developmental markers. *Genome Biol. Evol.* **11**, 3159–3178 (2019).
794 48. Dorsman, J. C. *et al.* Strong aggregation and increased toxicity of polyleucine over
795 polyglutamine stretches in mammalian cells. *Hum. Mol. Genet.* **11**, 1487–1496 (2002).
796 49. Moriya, H. The expression level and cytotoxicity of green fluorescent protein are modulated by
797 an additional N-terminal sequence. *AIMS Biophys.* **7**, 121–132 (2020).
798 50. Moriya, H., Makanae, K., Watanabe, K., Chino, A. & Shimizu-Yoshida, Y. Robustness analysis
799 of cellular systems using the genetic tug-of-war method. *Mol. Biosyst.* **8**, 2513–2522 (2012).
800 51. Moriya, H., Shimizu-Yoshida, Y. & Kitano, H. In vivo robustness analysis of cell division cycle
801 genes in Saccharomyces cerevisiae. *PLoS Genet.* **2**, e111 (2006).
802 52. Azizoglu, A., Brent, R. & Rudolf, F. A precisely adjustable, variation-suppressed eukaryotic
803 transcriptional controller to enable genetic discovery. *Elife* **10**, (2021).

53. Buric, F. *et al.* Amino acid sequence encodes protein abundance shaped by protein stability at reduced synthesis cost. *Protein Sci.* **34**, e5239 (2025).

54. Liu, Z. *et al.* Systematic comparison of 2A peptides for cloning multi-genes in a polycistronic vector. *Sci. Rep.* **7**, 2193 (2017).

55. Fujita, Y., Namba, S. & Moriya, H. Impact of maximal Overexpression of a non-toxic protein on yeast cell physiology. (2024) doi:10.7554/elife.99572.

56. Namba, S., Kato, H., Shigenobu, S., Makino, T. & Moriya, H. Massive expression of cysteine-containing proteins causes abnormal elongation of yeast cells by perturbing the proteasome. *G3* (2022) doi:10.1093/g3journal/jkac106.

57. Li, G., Ji, B. & Nielsen, J. The pan-genome of Saccharomyces cerevisiae. *FEMS Yeast Res.* **19**, (2019).

58. Hatsuzawa, K., Tagaya, M. & Mizushima, S. The hydrophobic region of signal peptides is a determinant for SRP recognition and protein translocation across the ER membrane. *J. Biochem.* **121**, 270–277 (1997).

59. Fauchère, J. L., Charton, M., Kier, L. B., Verloop, A. & Pliska, V. Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int. J. Pept. Protein Res.* **32**, 269–278 (1988).

60. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).

61. Akashi, H. & Gojobori, T. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 3695–3700 (2002).

62. Craig, C. L. & Weber, R. S. Selection costs of amino acid substitutions in ColE1 and ColIa gene clusters harbored by Escherichia coli. *Mol. Biol. Evol.* **15**, 774–776 (1998).

63. Wagner, A. Energy constraints on the evolution of gene expression. *Mol. Biol. Evol.* **22**, 1365–1374 (2005).

64. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2021).

65. Lamprecht, M. R., Sabatini, D. M. & Carpenter, A. E. CellProfiler: free, versatile software for automated biological image analysis. *Biotechniques* **42**, 71–75 (2007).

66. Köhrer, K. & Domdey, H. [27] Preparation of high molecular weight RNA. in *Methods in Enzymology* vol. 194 398–405 (Academic Press, 1991).

67. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

68. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

69. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

70. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

71. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

72. Xu, S. *et al.* Using clusterProfiler to characterize multiomics data. *Nat. Protoc.* **19**, 3292–3320 (2024).

Supplementary Figures

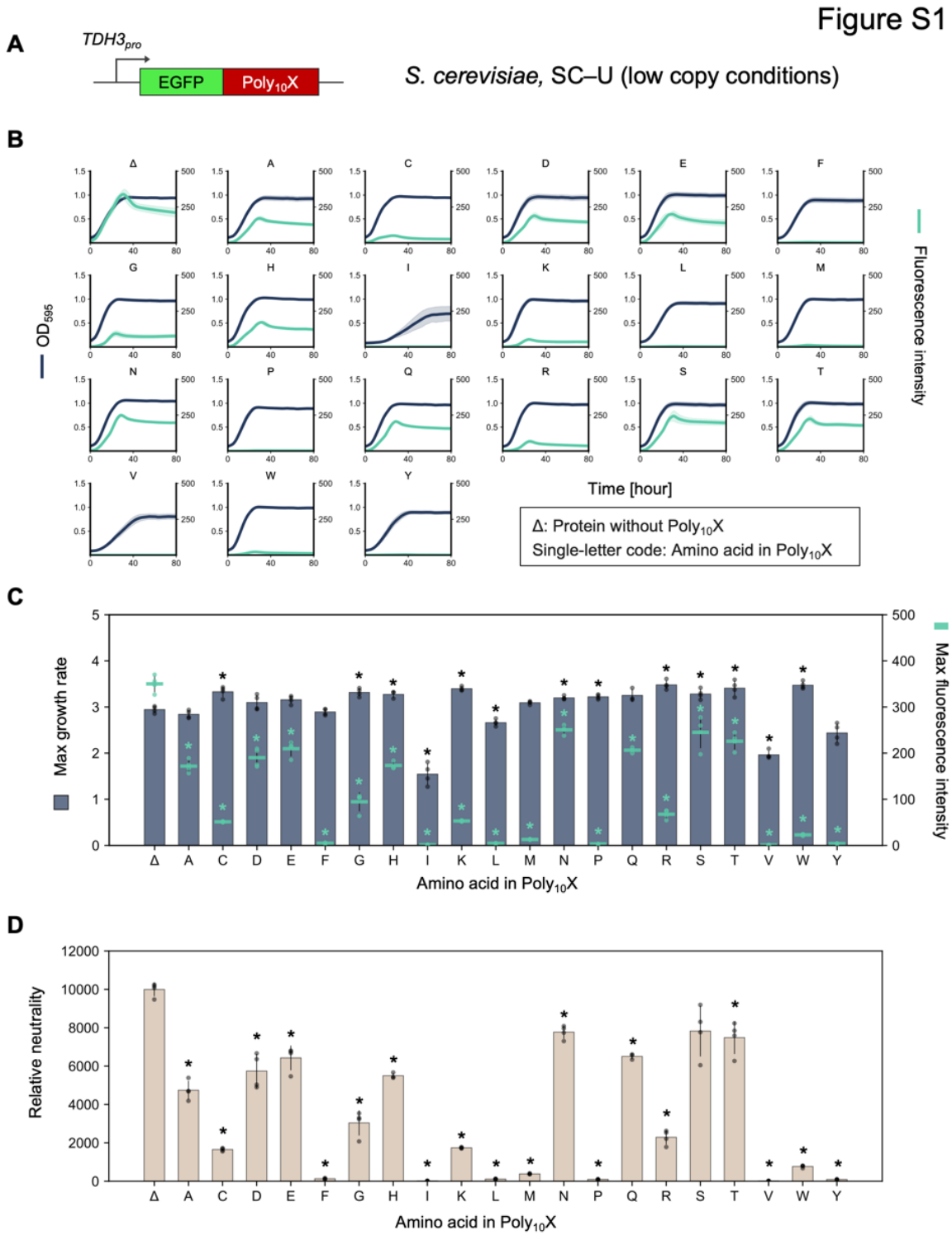


Figure S1. Neutrality of C-terminal Poly10X fusions to EGFP in yeast (low-copy conditions).

854 **A)** Schematic representation of the expression constructs. $Poly_{10}X$ was fused to the C-terminus of EGFP
855 and expressed under the *TDH3* promoter in *S. cerevisiae*.

856 **B)** Growth and fluorescence curves of *S. cerevisiae* cells expressing EGFP–$Poly_{10}X$, measured using
857 the gTOW method in SC–U medium at 30 °C. Curves represent the mean values from at least three
858 biological replicates, and shaded regions indicate the standard deviation (SD).

859 **C, D)** Maximum growth rate and maximum fluorescence intensity of *S. cerevisiae* cells low-level
860 overexpressing EGFP–$Poly_{10}X$ in SC–U medium at 30 °C, and the calculated relative neutrality (**D**).
861 Bars, dots, and error bars represent the mean, individual data points, and standard deviation from at
862 least three biological replicates. Asterisks indicate significant differences in maximum growth rate,
863 maximum fluorescence intensity, and relative neutrality, respectively, compared with the control (Δ) (p
864 < 0.05, Student's t-test with Bonferroni correction).

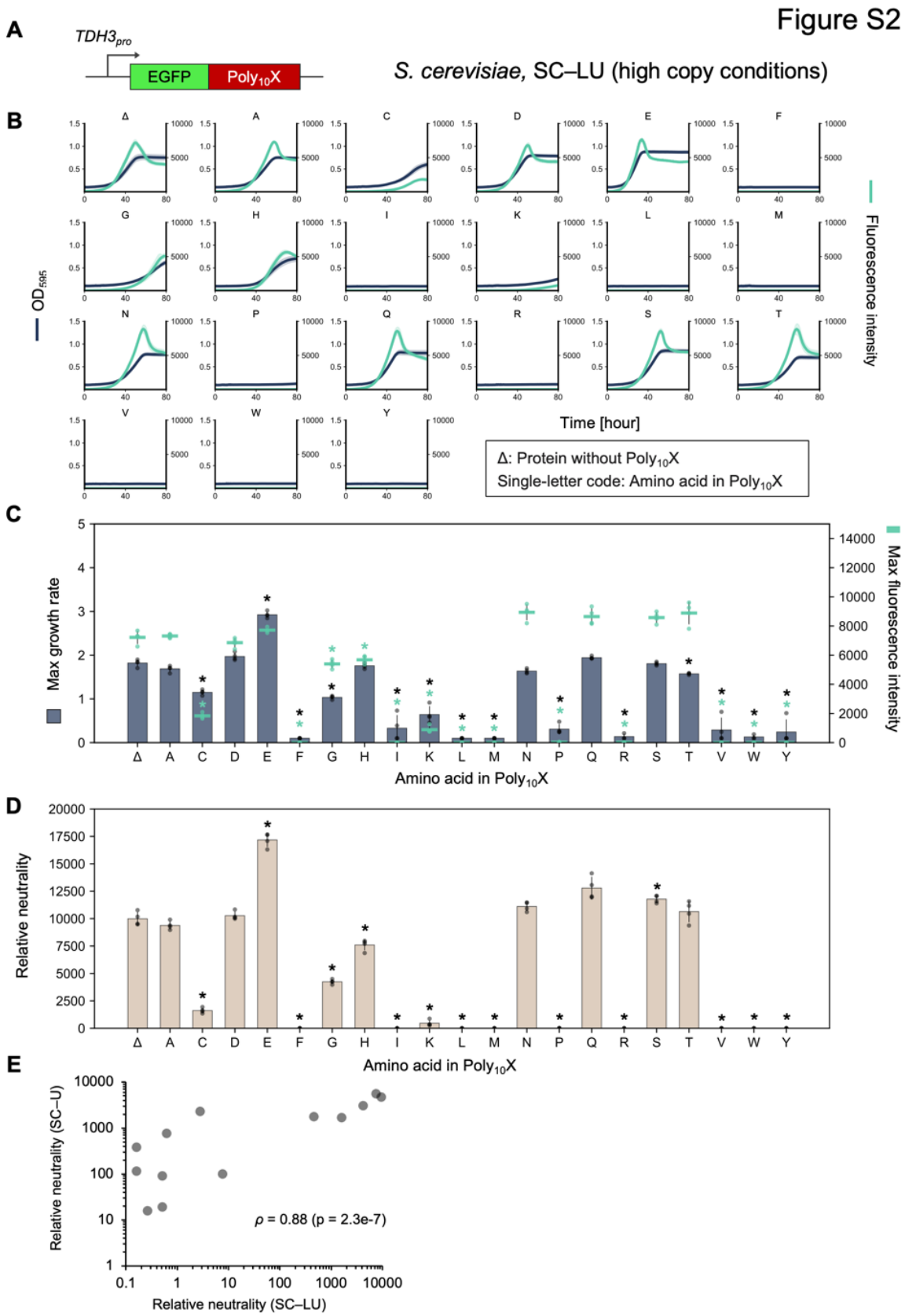


Figure S2
A
TDH3pro
EGFP
Poly10X
S. cerevisiae, SC–LU (high copy conditions)
B
OD595
Fluorescence intensity
Time [hour]
Δ: Protein without Poly10X
Single-letter code: Amino acid in Poly10X
C
Max growth rate
Max fluorescence intensity
Amino acid in Poly10X
D
Relative neutrality
Amino acid in Poly10X
E
Relative neutrality (SC–U)
Relative neutrality (SC–LU)
ρ = 0.88 (p = 2.3e-7)

865

866 **Figure S2.** Neutrality of C-terminal $Poly_{10}X$ fusions to EGFP in yeast (High-
867 copy conditions).

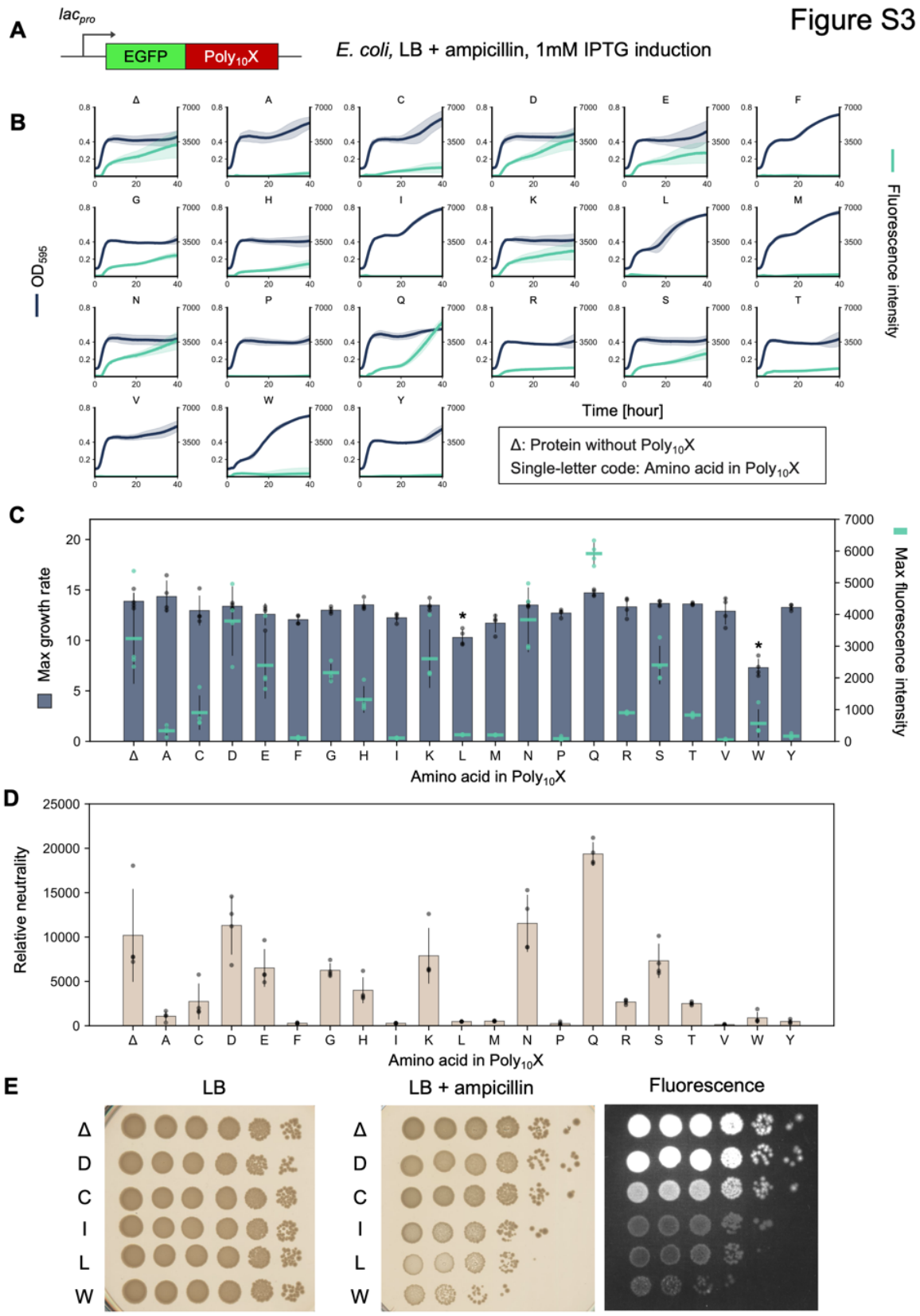
868 **A)** Schematic representation of the expression constructs. $Poly_{10}X$ was fused to the C-terminus of EGFP
869 and expressed under the *TDH3* promoter in *S. cerevisiae*.

870 **B)** Growth and fluorescence curves of *S. cerevisiae* cells expressing EGFP–$Poly_{10}X$, measured using
871 the gTOW method in SC–LU medium at 30 °C. Curves represent the mean values from at least three
872 biological replicates, and shaded regions indicate the standard deviation (SD).

873 **C, D)** Maximum growth rate and maximum fluorescence intensity of *S. cerevisiae* cells overexpressing
874 EGFP–$Poly_{10}X$ in SC–LU medium at 30 °C, and the calculated relative neutrality (**D**). Bars, dots, and
875 error bars represent the mean, individual data points, and standard deviation from at least three
876 biological replicates. Asterisks indicate significant differences in maximum growth rate, maximum
877 fluorescence intensity, and relative neutrality, respectively, compared with the control (Δ) ($p < 0.05$,
878 Student's *t*-test with Bonferroni correction).

879 **E**) Comparison of $Poly_{10}X$ harmfulness trends under low-level (SC–U) and high-level (SC–LU)
880 overexpression conditions. Spearman's rank correlation coefficient (ρ) is shown.

881



882

883 **Figure S3.** Neutrality of C-terminal $Poly_{10}X$ fusions to EGFP in *E. coli*.

884 **A)** Schematic representation of the expression constructs. $Poly_{10}X$ was fused to the C-terminus of EGFP
885 and expressed under the *lac* promoter in *E. coli*.

886 **B)** Growth and fluorescence curves of *E. coli* cells expressing EGFP–$Poly_{10}X$, measured in LB +
887 ampicillin medium at 37 °C. Curves represent the mean values from four biological replicates, and
888 shaded regions indicate the standard deviation (SD).

889 **C, D)** Maximum growth rate and maximum fluorescence intensity of *E. coli* cells expressing EGFP–
890 $Poly_{10}X$ in LB + ampicillin medium at 37 °C, and the calculated relative neutrality (**D**). Bars, dots, and
891 error bars represent the mean, individual data points, and standard deviation from four biological
892 replicates. Asterisks indicate significant differences in maximum growth rate, maximum fluorescence
893 intensity, and relative neutrality, respectively, compared with the control (Δ) ($p < 0.05$, Student's *t*-test
894 with Bonferroni correction).

895 **E**) Several *E. coli* strains expressing EGFP–$Poly_{10}X$ were collected after cultivation in **B**, serially
896 diluted 10-fold, and spotted (5 μl each) onto LB or LB + ampicillin agar plates. Plates were incubated
897 overnight at 37 °C and photographed the following day. These data suggest plasmid loss after
898 cultivation of cells expressing harmful Poly10X variants. Single-letter codes indicate the amino acid
899 repeated in $Poly_{10}X$. Δ represents the protein without a $Poly_{10}X$ fusion.

Figure S4

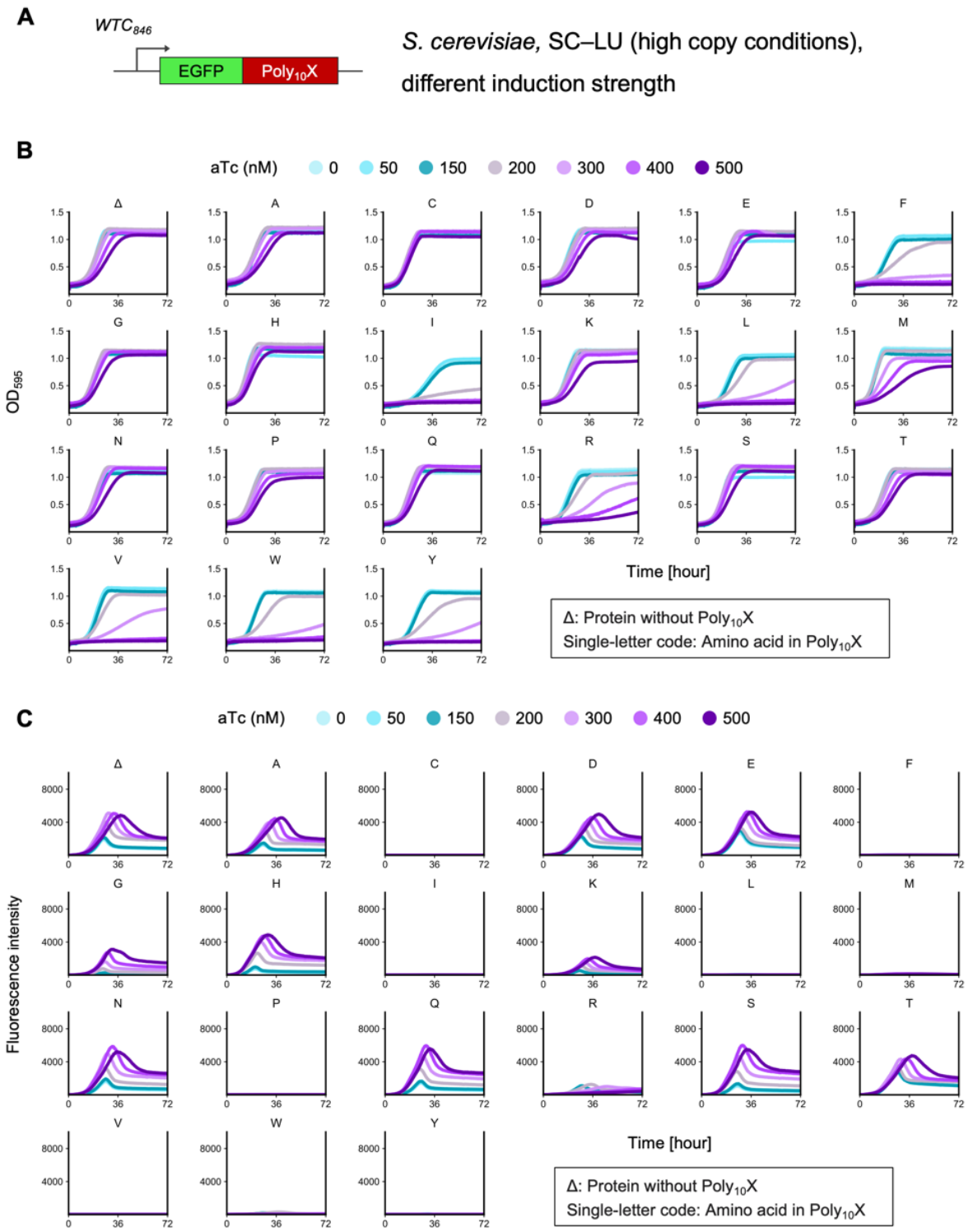


Figure S4. Neutrality of C-terminal $Poly_{10}X$ fusions to EGFP under different induction levels in yeast.

903 **A)** Schematic representation of the expression constructs. $Poly_{10}X$ was fused to the C-terminus of EGFP
904 and expressed under the control of the WTC_{846} promoter in *S. cerevisiae*.

905 **B, C)** Stepwise growth (**B**) and fluorescence (**C**) curves of *S. cerevisiae* cells expressing EGFP–$Poly_{10}X$
906 measured using the gTOW method in SC–LU medium at 30 °C. Gradual induction of expression was
907 achieved under the control of the WTC_{846} promoter by stepwise adjustment of aTc concentration. Curves
908 represent the mean values from four biological replicates.

Figure S5

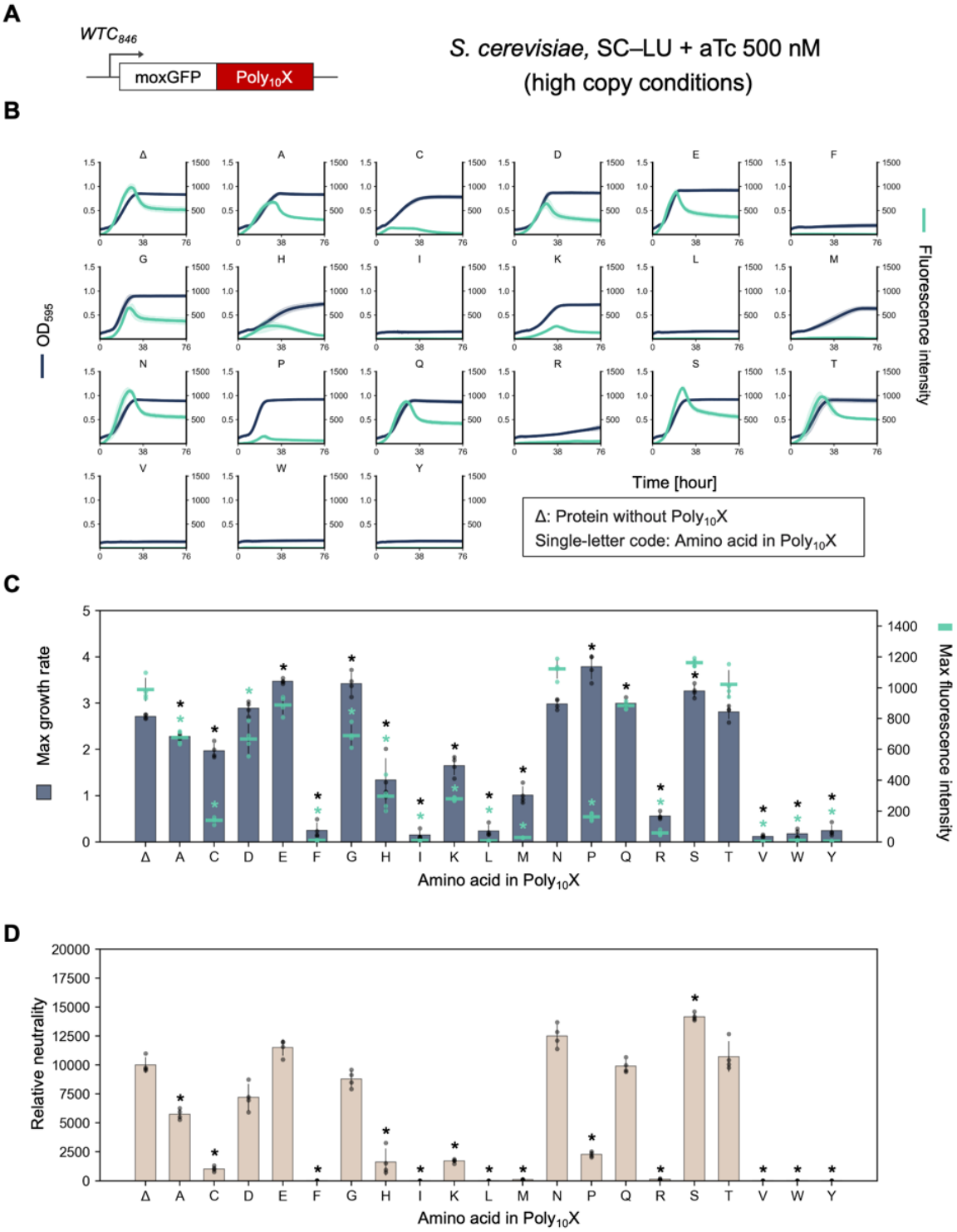
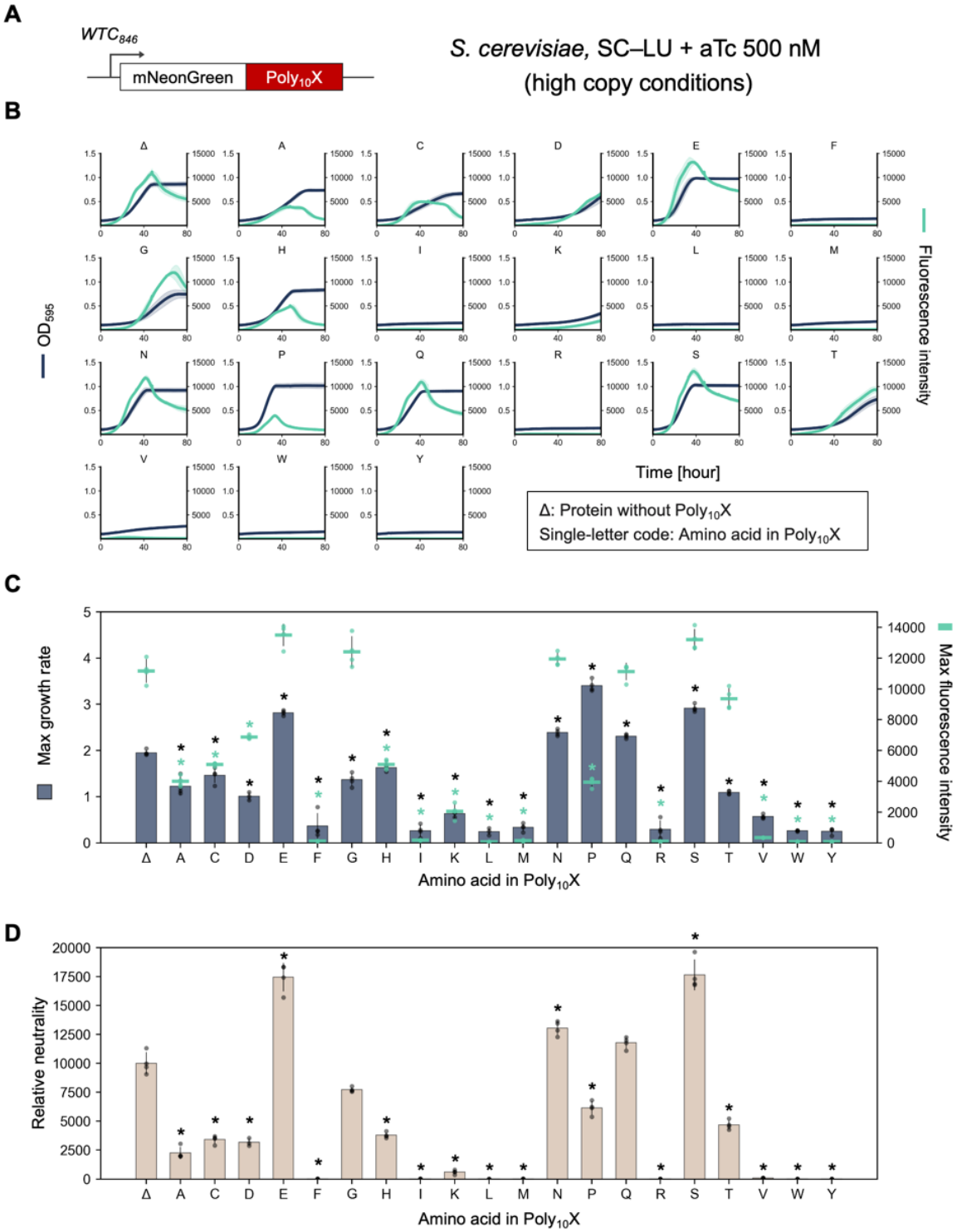


Figure S5. Neutrality of C-terminal $Poly_{10}X$ fusions to moxGFP in yeast.

911 **A)** Schematic representation of the expression constructs. $Poly_{10}X$ was fused to the C-terminus of
912 moxGFP and expressed under the control of the WTC_{846} promoter in *S. cerevisiae*.

913 **B)** Growth and fluorescence curves of *S. cerevisiae* cells expressing moxGFP–$Poly_{10}X$, measured using
914 the gTOW method in SC–LU medium at 30 °C. Curves represent the mean values from at least three
915 biological replicates, and shaded regions indicate the standard deviation (SD).

916 **C, D)** Maximum growth rate and maximum fluorescence intensity of *S. cerevisiae* cells overexpressing
917 moxGFP–$Poly_{10}X$ in SC–LU medium at 30 °C, and the calculated relative neutrality (**D**). Bars, dots,
918 and error bars represent the mean, individual data points, and standard deviation from at least three
919 biological replicates. Asterisks indicate significant differences in maximum growth rate, maximum
920 fluorescence intensity, and relative neutrality, respectively, compared with the control (Δ) ($p < 0.05$,
921 Student's *t*-test with Bonferroni correction).



922

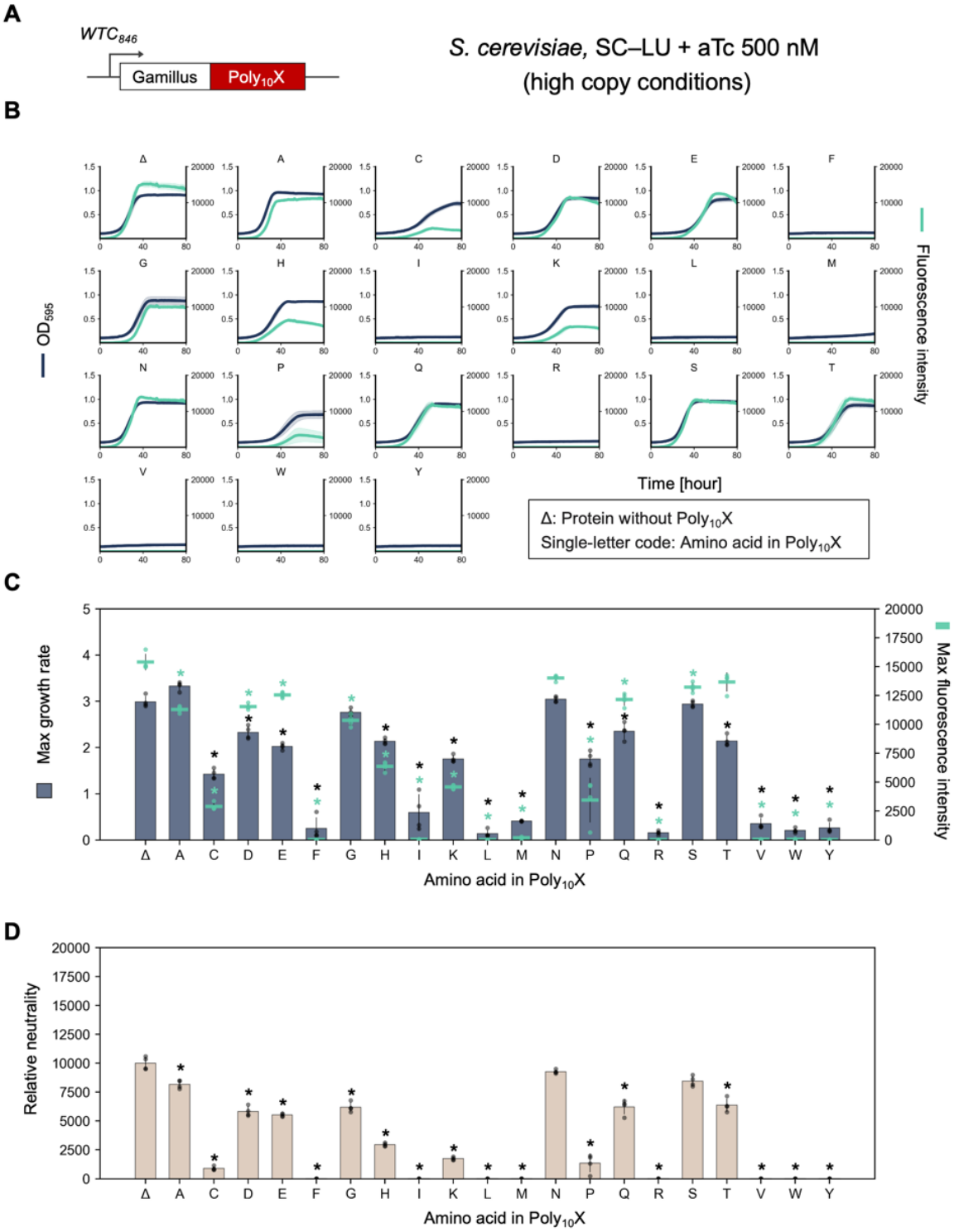
923 **Figure S6.** Neutrality of C-terminal $Poly_{10}X$ fusions to mNeonGreen in yeast.

924 **A)** Schematic representation of the expression constructs. $Poly_{10}X$ was fused to the C-terminus of
925 mNeonGreen and expressed under the control of the WTC_{846} promoter in *S. cerevisiae*.

926 **B)** Growth and fluorescence curves of *S. cerevisiae* cells expressing mNeonGreen–$Poly_{10}X$, measured
927 using the gTOW method in SC–LU medium at 30 °C. Curves represent the mean values from at least
928 three biological replicates, and shaded regions indicate the standard deviation (SD).

929 **C, D)** Maximum growth rate and maximum fluorescence intensity of *S. cerevisiae* cells overexpressing
930 mNeonGreen–$Poly_{10}X$ in SC–LU medium at 30 °C, and the calculated relative neutrality (**D**). Bars, dots,
931 and error bars represent the mean, individual data points, and standard deviation from at least three
932 biological replicates. Asterisks indicate significant differences in maximum growth rate, maximum
933 fluorescence intensity, and relative neutrality, respectively, compared with the control (Δ) ($p < 0.05$,
934 Student's *t*-test with Bonferroni correction).

Figure S7



935

936 **Figure S7.** Neutrality of C-terminal $Poly_{10}X$ fusions to Gamillus in yeast.

937 **A)** Schematic representation of the expression constructs. $Poly_{10}X$ were fused to the C-terminus of
938 Gamillus and expressed under the control of the WTC_{846} promoter in *S. cerevisiae*.

939 **B)** Growth and fluorescence curves of *S. cerevisiae* cells expressing Gamillus–$Poly_{10}X$, measured using
940 the gTOW method in SC–LU medium at 30 °C. Curves represent the mean values from at least three
941 biological replicates, and shaded regions indicate the standard deviation (SD).

942 **C, D)** Maximum growth rate and maximum fluorescence intensity of *S. cerevisiae* cells overexpressing
943 Gamillus–$Poly_{10}X$ in SC–LU medium at 30 °C, and the calculated relative neutrality (**D**). Bars, dots,
944 and error bars represent the mean, individual data points, and standard deviation from at least three
945 biological replicates. Asterisks indicate significant differences in maximum growth rate, maximum
946 fluorescence intensity, and relative neutrality, respectively, compared with the control (Δ) ($p < 0.05$,
947 Student's *t*-test with Bonferroni correction).

Figure S8

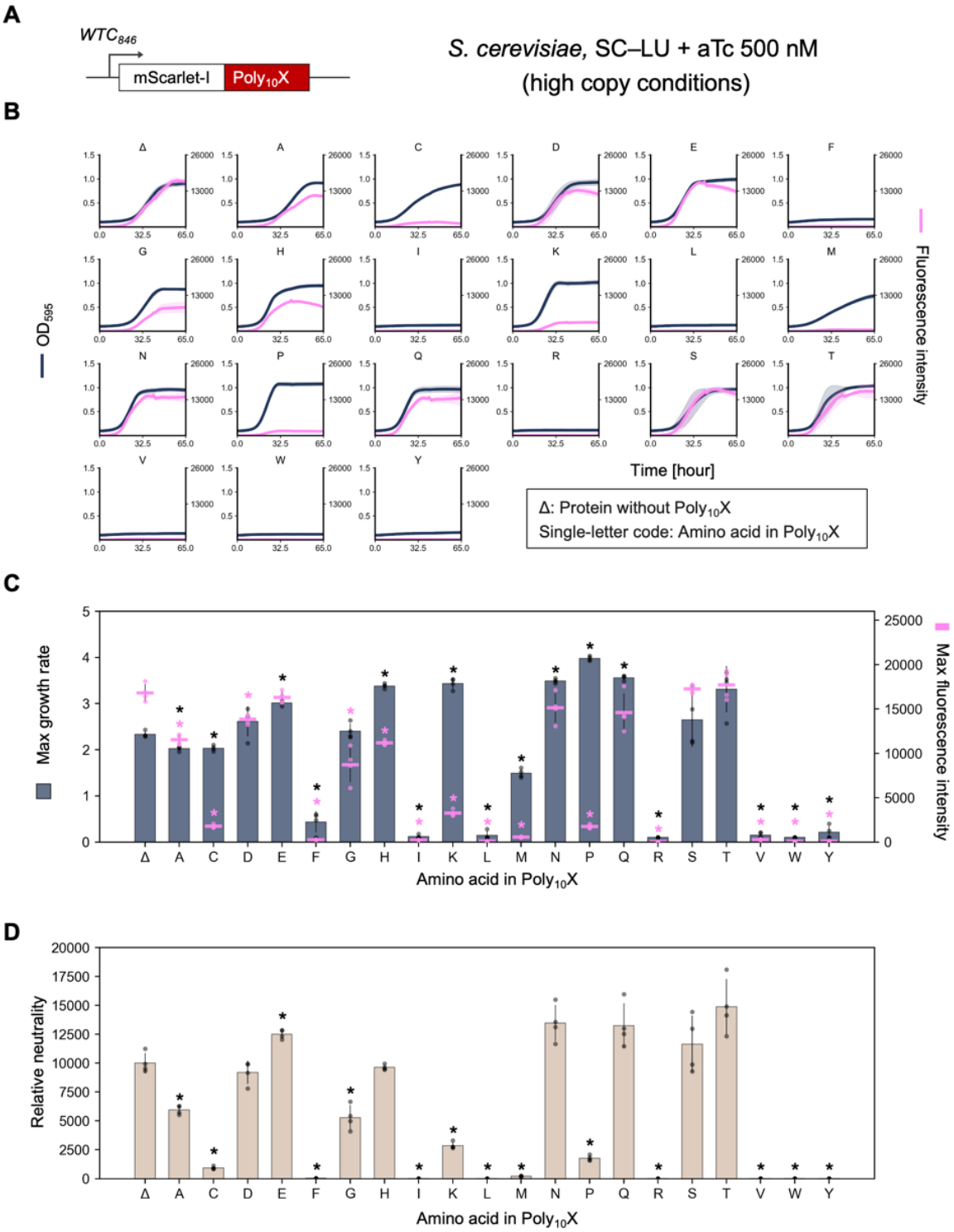


Figure S8. Neutrality of C-terminal $Poly_{10}X$ fusions to mScarlet-I in yeast.

950 **A)** Schematic representation of the expression constructs. $Poly_{10}X$ was fused to the C-terminus of
951 mScarlet-I and expressed under the control of the WTC_{846} promoter in *S. cerevisiae*.

952 **B)** Growth and fluorescence curves of *S. cerevisiae* cells expressing mScarlet-I–$Poly_{10}X$, measured
953 using the gTOW method in SC–LU medium at 30 °C. Curves represent the mean values from at least
954 three biological replicates, and shaded regions indicate the standard deviation (SD).

955 **C, D)** Maximum growth rate and maximum fluorescence intensity of *S. cerevisiae* cells overexpressing
956 mScarlet-I–$Poly_{10}X$ in SC–LU medium at 30 °C, and the calculated relative neutrality (**D**). Bars, dots,
957 and error bars represent the mean, individual data points, and standard deviation from at least three
958 biological replicates. Asterisks indicate significant differences in maximum growth rate, maximum
959 fluorescence intensity, and relative neutrality, respectively, compared with the control (Δ) ($p < 0.05$,
960 Student's *t*-test with Bonferroni correction).

Figure S9

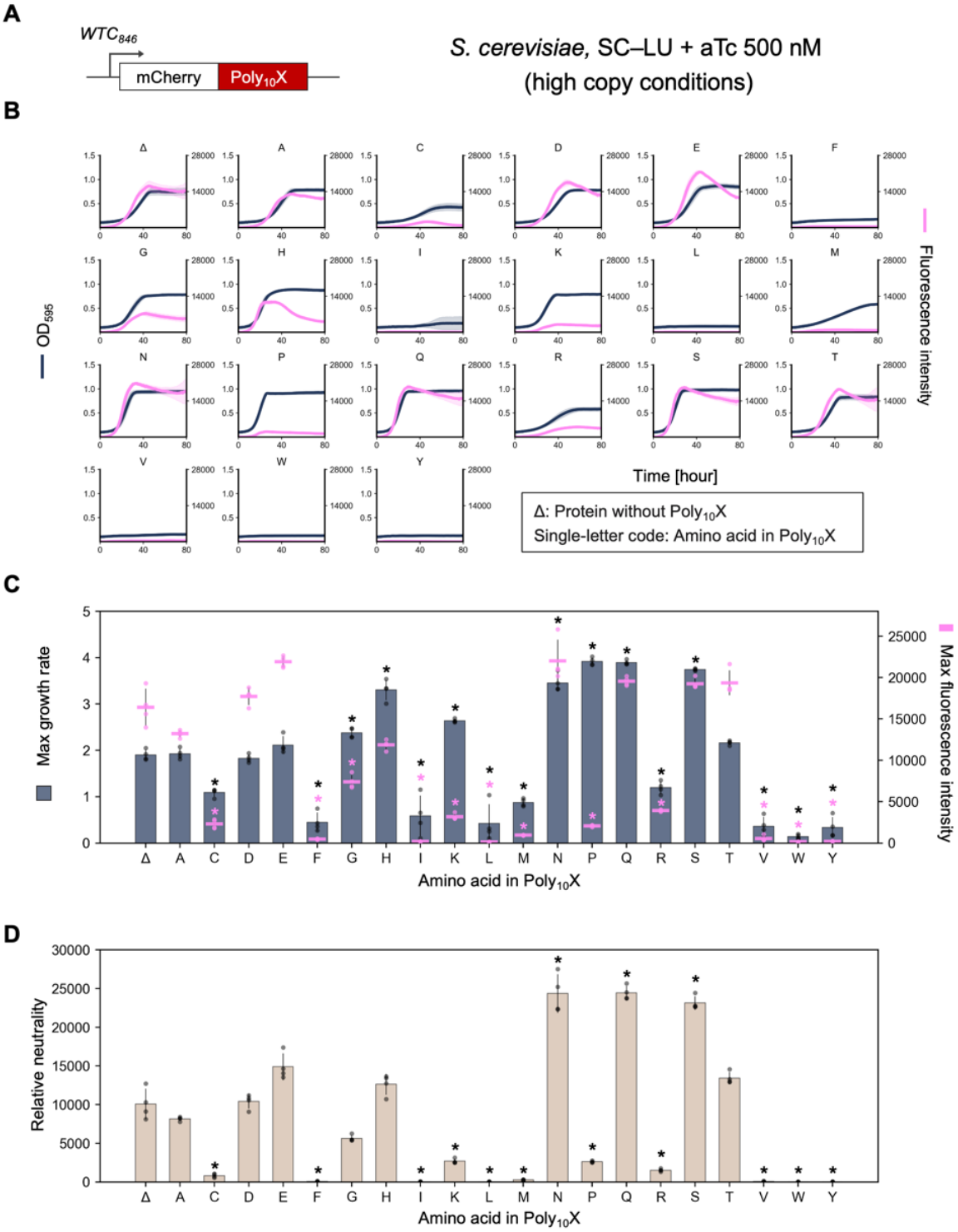


Figure S9. Neutrality of C-terminal $Poly_{10}X$ fusions to mCherry in yeast.

963 **A)** Schematic representation of the expression constructs. $Poly_{10}X$ was fused to the C-terminus of
964 mCherry and expressed under the control of the WTC_{846} promoter in *S. cerevisiae*.

965 **B)** Growth and fluorescence curves of *S. cerevisiae* cells expressing mCherry–$Poly_{10}X$, measured using
966 the gTOW method in SC–LU medium at 30 °C. Curves represent the mean values from at least three
967 biological replicates, and shaded regions indicate the standard deviation (SD).

968 **C, D)** Maximum growth rate and maximum fluorescence intensity of *S. cerevisiae* cells overexpressing
969 mCherry–$Poly_{10}X$ in SC–LU medium at 30 °C, and the calculated relative neutrality (**D**). Bars, dots,
970 and error bars represent the mean, individual data points, and standard deviation from at least three
971 biological replicates. Asterisks indicate significant differences in maximum growth rate, maximum
972 fluorescence intensity, and relative neutrality, respectively, compared with the control (Δ) ($p < 0.05$,
973 Student's *t*-test with Bonferroni correction).

Figure S10

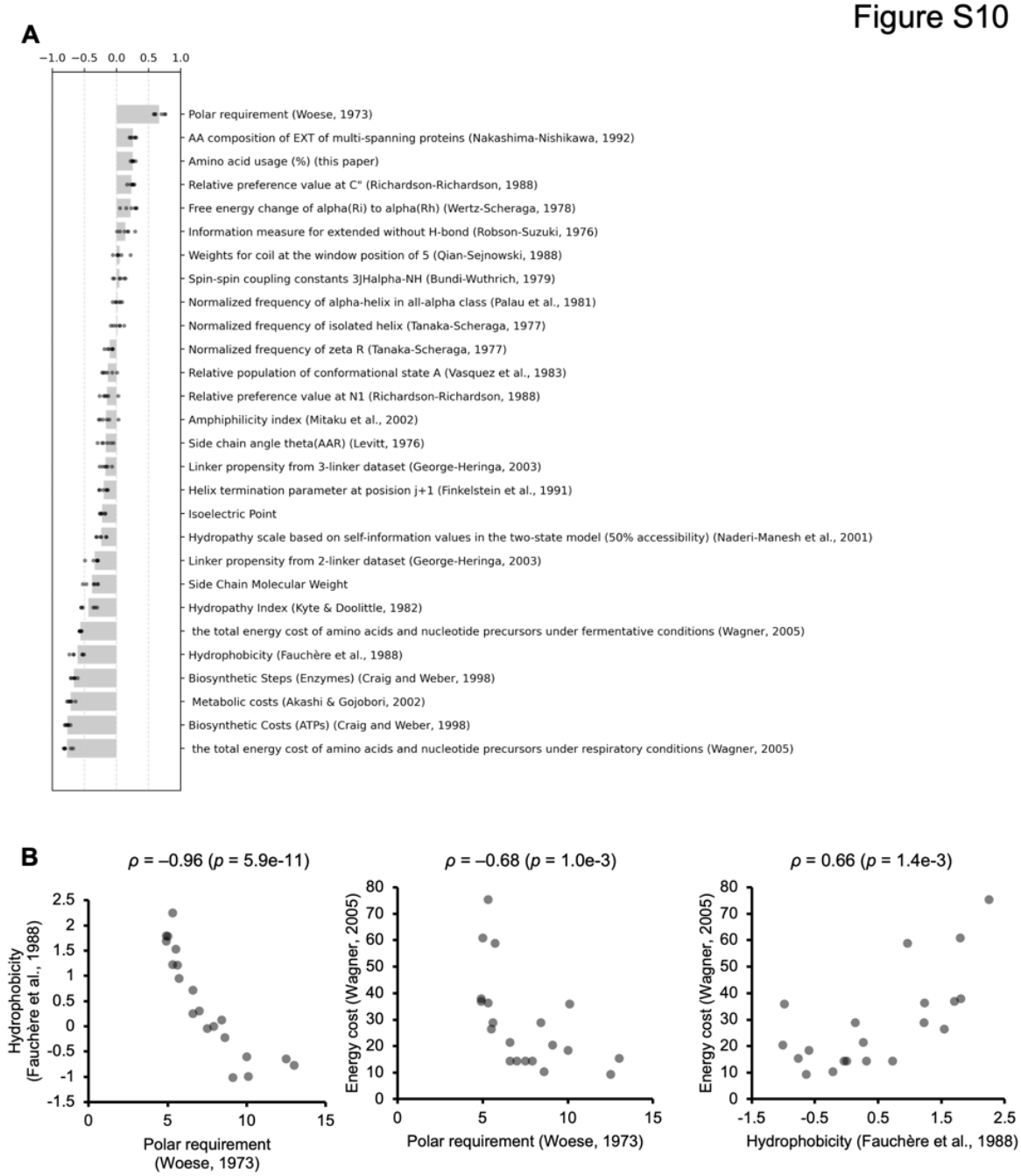
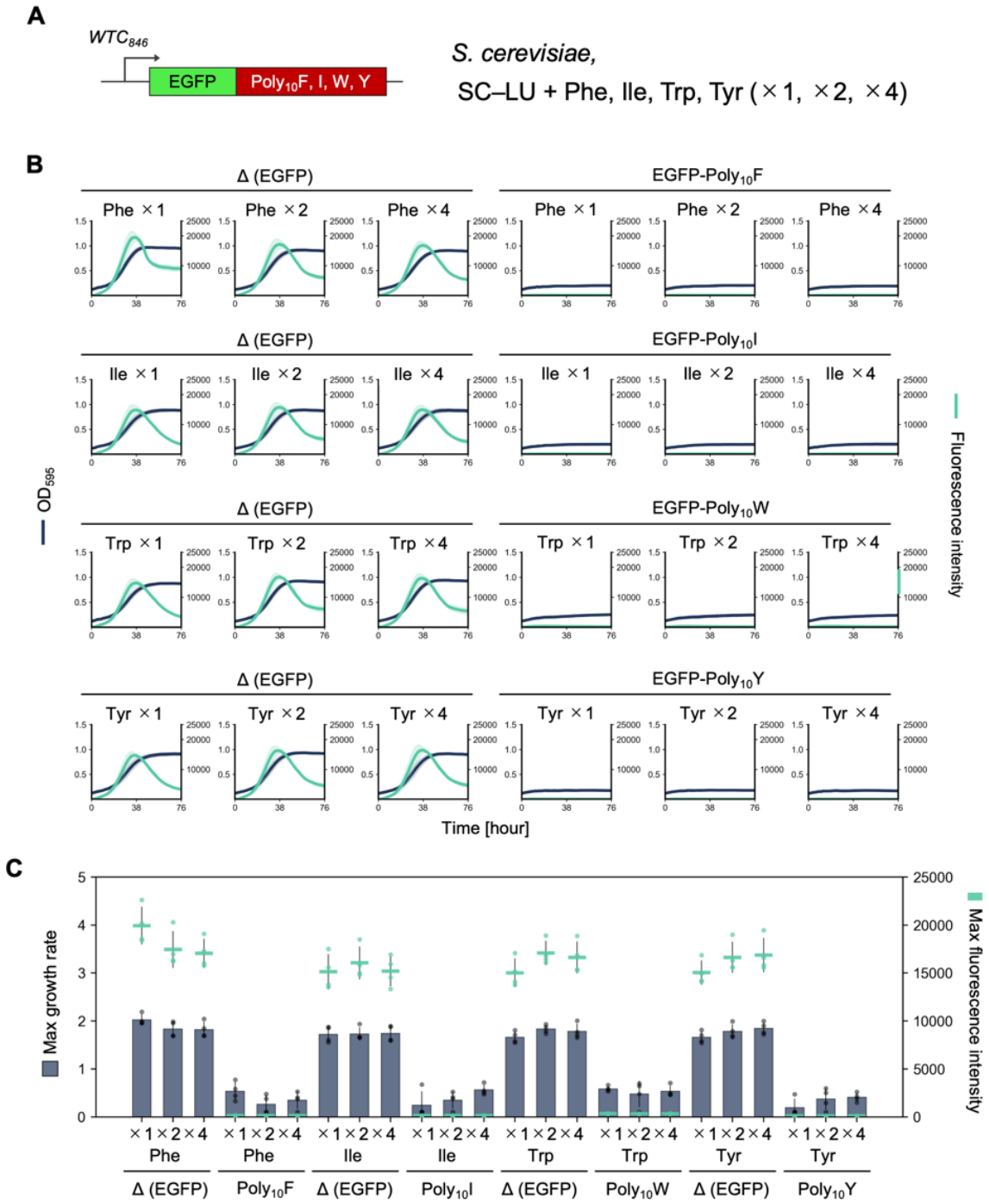


Figure S10. Correlation analysis between $Poly_{10}X$ neutrality and amino acid properties.

977 **A)** Correlation between $Poly_{10}X$ neutrality and various amino acid indices (physicochemical and usage-
978 related properties). Each dot and bar represents individual Spearman's rank correlation coefficients and
979 their mean values calculated across six fluorescent proteins.

980 **B)** Relationships among amino acid indices that showed strong correlations with $Poly_{10}X$ relative
981 neutrality. Spearman's rank correlation coefficient (ρ) and its p-value are shown.

Figure S11



982

983 **Figure S11.** Effects of supplemented amino acids on the harmfulness of high-

984 biosynthetic-cost $Poly_{10}X$ repeats.

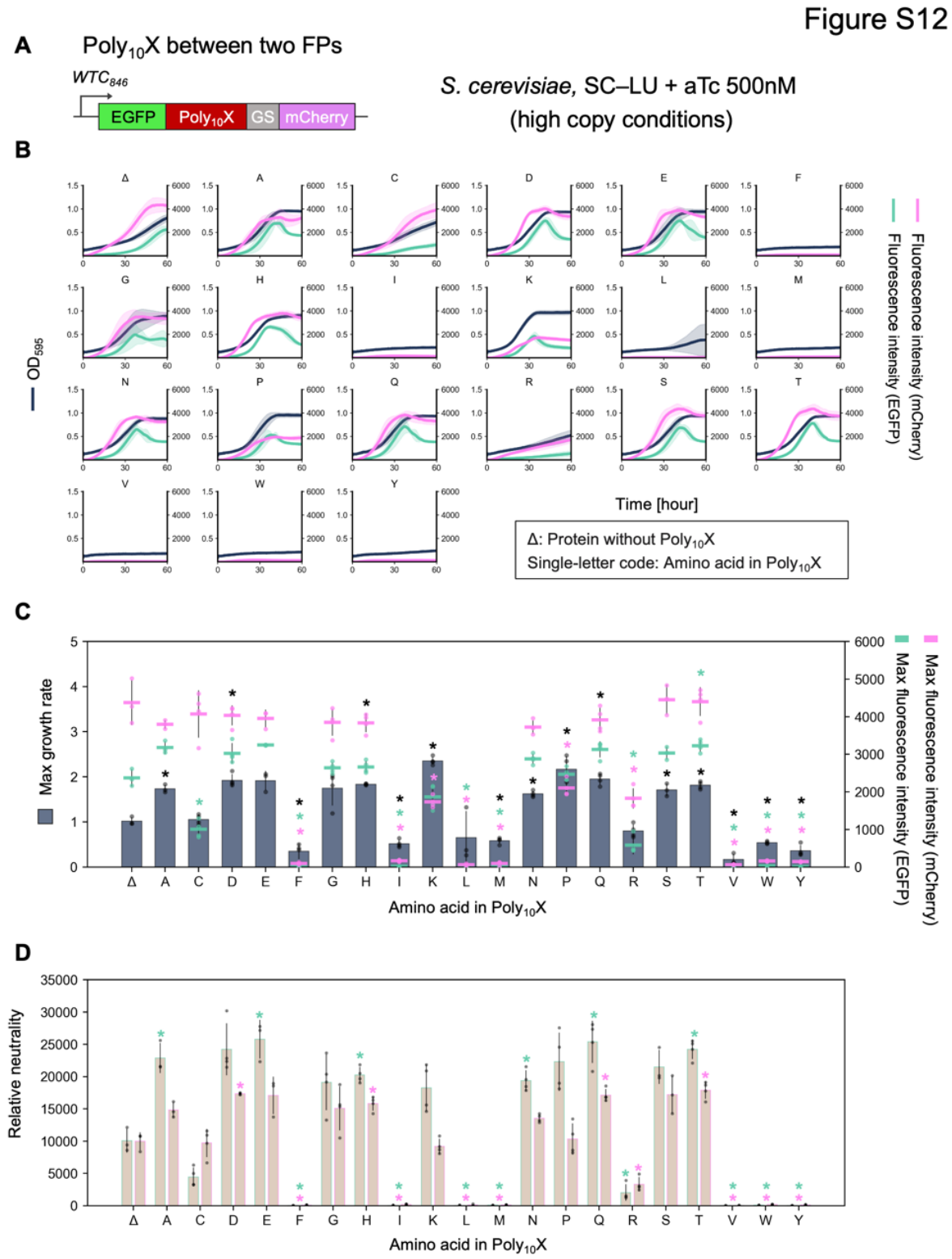
985 **A)** Schematic representation of the expression constructs. $Poly_{10}F$, $Poly_{10}I$, $Poly_{10}W$, and $Poly_{10}Y$ were
986 fused to the C-terminus of EGFP and expressed under the control of the WTC_{846} promoter in *S.*
987 *cerevisiae*.

988 **B)** Growth and fluorescence curves of *S. cerevisiae* cells expressing EGFP–$Poly_{10}X$, measured using
989 the gTOW method in SC–LU medium supplemented with additional amino acids (×1, ×2, ×4) at 30 °C.
990 Curves represent the mean values from four biological replicates, and shaded regions indicate the
991 standard deviation (SD).

992 **C)** Maximum growth rate and maximum fluorescence intensity of *S. cerevisiae* cells overexpressing
993 EGFP–$Poly_{10}X$ in SC–LU medium supplemented with additional amino acids (×1, ×2, ×4) at 30 °C.
994 Bars, dots, and error bars represent the mean, individual data points, and standard deviation from four
995 biological replicates.

996

997 Maximum growth rate and maximum fluorescence intensity of *S. cerevisiae* cells overexpressing EGFP
998 without $Poly_{10}X$ (Δ) or with C-terminal $Poly_{10}F$, $Poly_{10}I$, $Poly_{10}W$, or $Poly_{10}Y$ fusions in SC–LU
999 medium at 30 °C. Indicated amino acid was supplemented to the medium at standard (×1), ×2, or ×4
1000 concentrations. Bars, dots, and error bars represent the mean, individual values, and standard deviation
1001 from four biological replicates.



1002

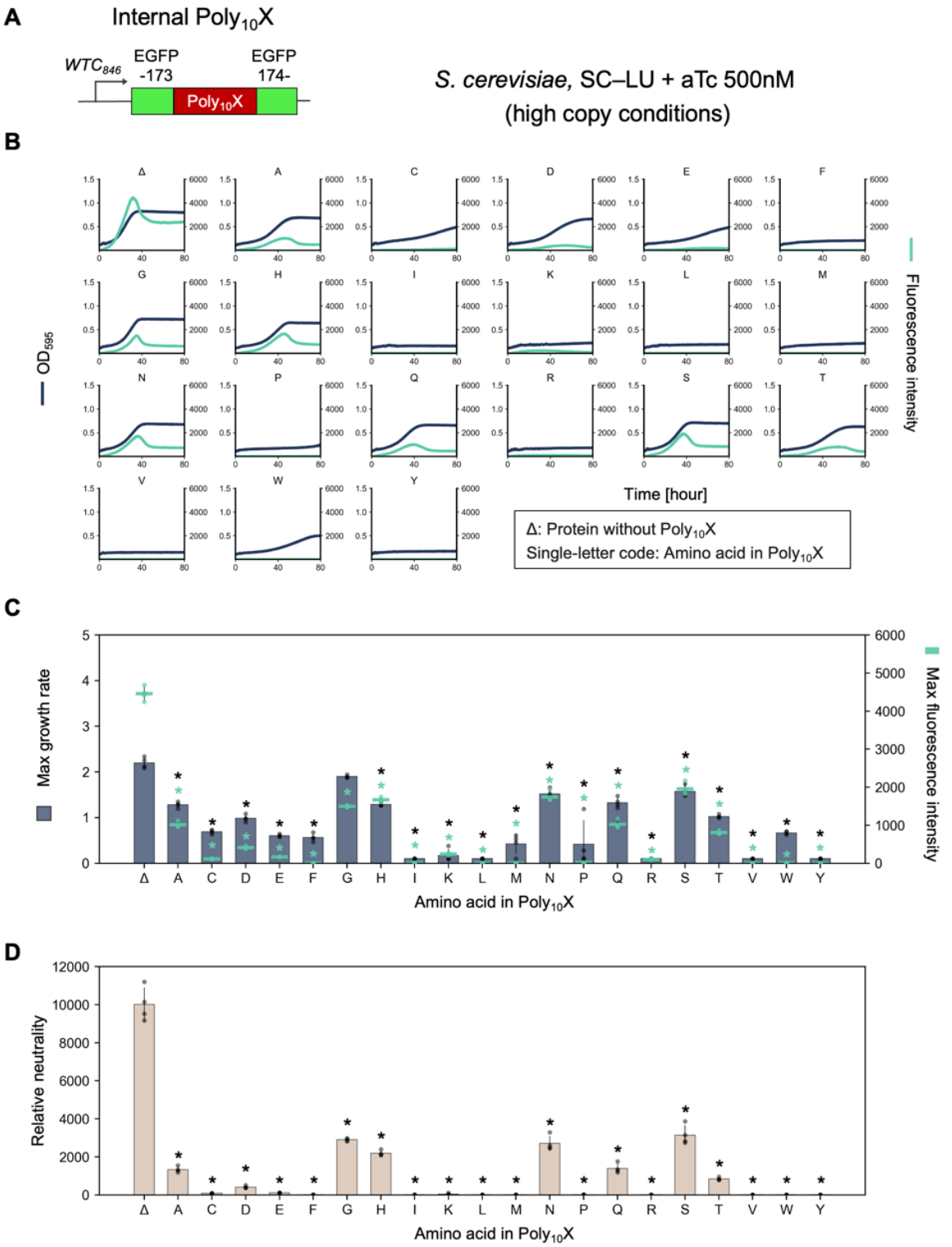
1003 **Figure S12.** Neutrality of $Poly_{10}X$ Insertions between two FPs in yeast.

1004 **A)** Schematic representation of the expression constructs. $Poly_{10}X$ were inserted between EGFP and
1005 mCherry and expressed under the control of the WTC_{846} promoter in *S. cerevisiae*.

1006 **B)** Growth and fluorescence curves of *S. cerevisiae* cells expressing EGFP–$Poly_{10}X$–GSlinker–
1007 mCherry, measured using the gTOW method in SC–LU medium at 30 °C. Curves represent the mean
1008 values from at least three biological replicates, and shaded regions indicate the standard deviation (SD).

1009 **C, D)** Maximum growth rate and maximum fluorescence intensity of *S. cerevisiae* cells overexpressing
1010 EGFP–$Poly_{10}X$–GSlinker–mCherry in SC–LU medium at 30 °C, and the calculated relative neutrality
1011 (**D**). Bars, dots, and error bars represent the mean, individual data points, and standard deviation from
1012 at least three biological replicates. Asterisks indicate significant differences in maximum growth rate,
1013 maximum fluorescence intensity, and relative neutrality, respectively, compared with the control (Δ) (p
1014 < 0.05, Student's t-test with Bonferroni correction).

Figure S13



1015

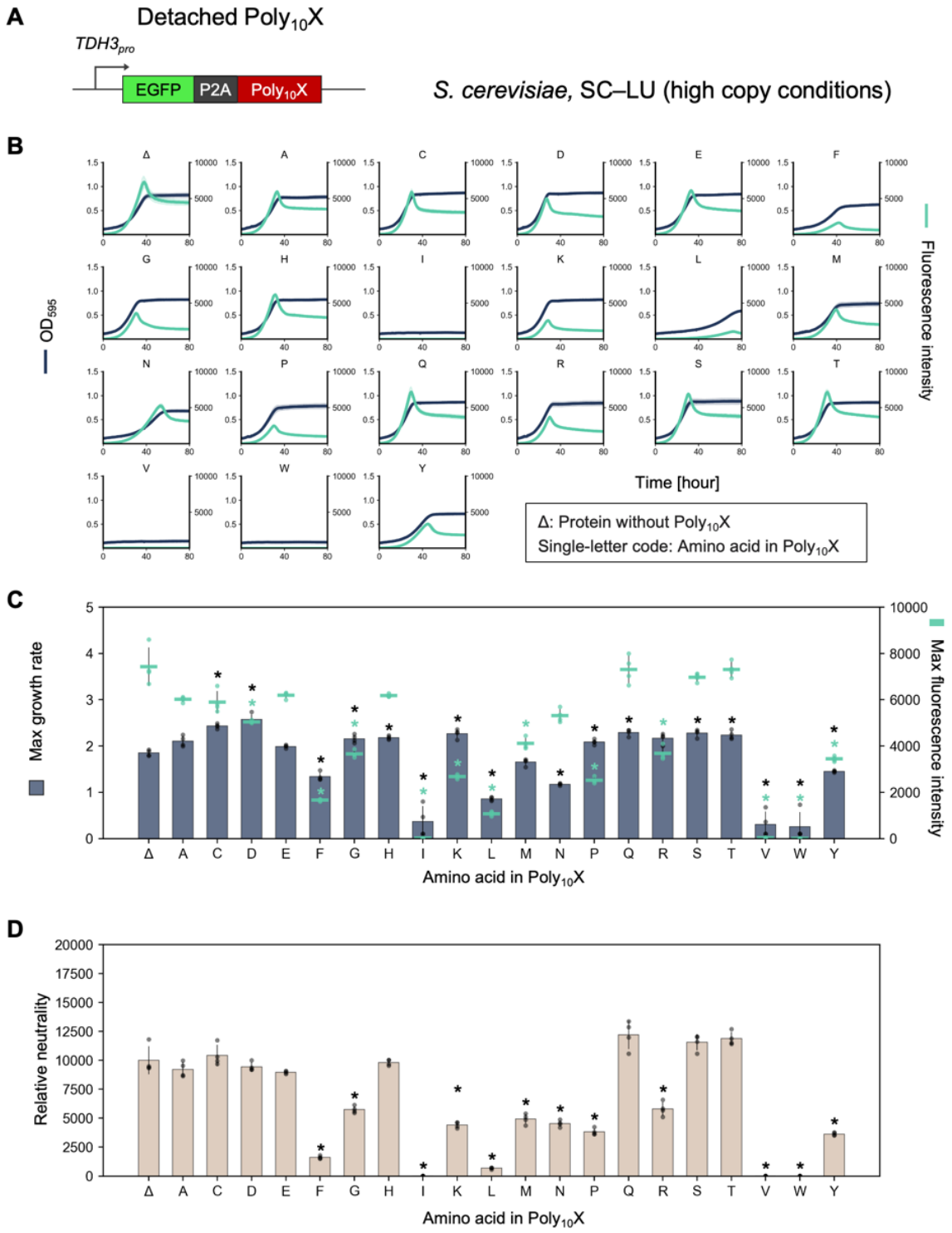
1016 **Figure S13.** Neutrality of Poly$_{10}$X Insertions within EGFP in yeast.

1017 **A)** Schematic representation of the expression constructs. $Poly_{10}X$ was inserted into an internal loop of
1018 EGFP between residues 173 and 174, and expressed under the control of the WTC_{846} promoter in *S.*
1019 *cerevisiae*.

1020 **B)** Growth and fluorescence curves of *S. cerevisiae* cells expressing $EGFP_{173}$–$Poly_{10}X$–$_{174}EGFP$,
1021 measured using the gTOW method in SC–LU medium at 30 °C. Curves represent the mean values from
1022 at least three biological replicates, and shaded regions indicate the standard deviation (SD).

1023 **C, D)** Maximum growth rate and maximum fluorescence intensity of *S. cerevisiae* cells overexpressing
1024 $EGFP_{173}$–$Poly_{10}X$–$_{174}EGFP$ in SC–LU medium at 30 °C, and the calculated relative neutrality (**D**). Bars,
1025 dots, and error bars represent the mean, individual data points, and standard deviation from at least three
1026 biological replicates. Asterisks indicate significant differences in maximum growth rate, maximum
1027 fluorescence intensity, and relative neutrality, respectively, compared with the control (Δ) ($p < 0.05$,
1028 Student's *t*-test with Bonferroni correction).

Figure S14



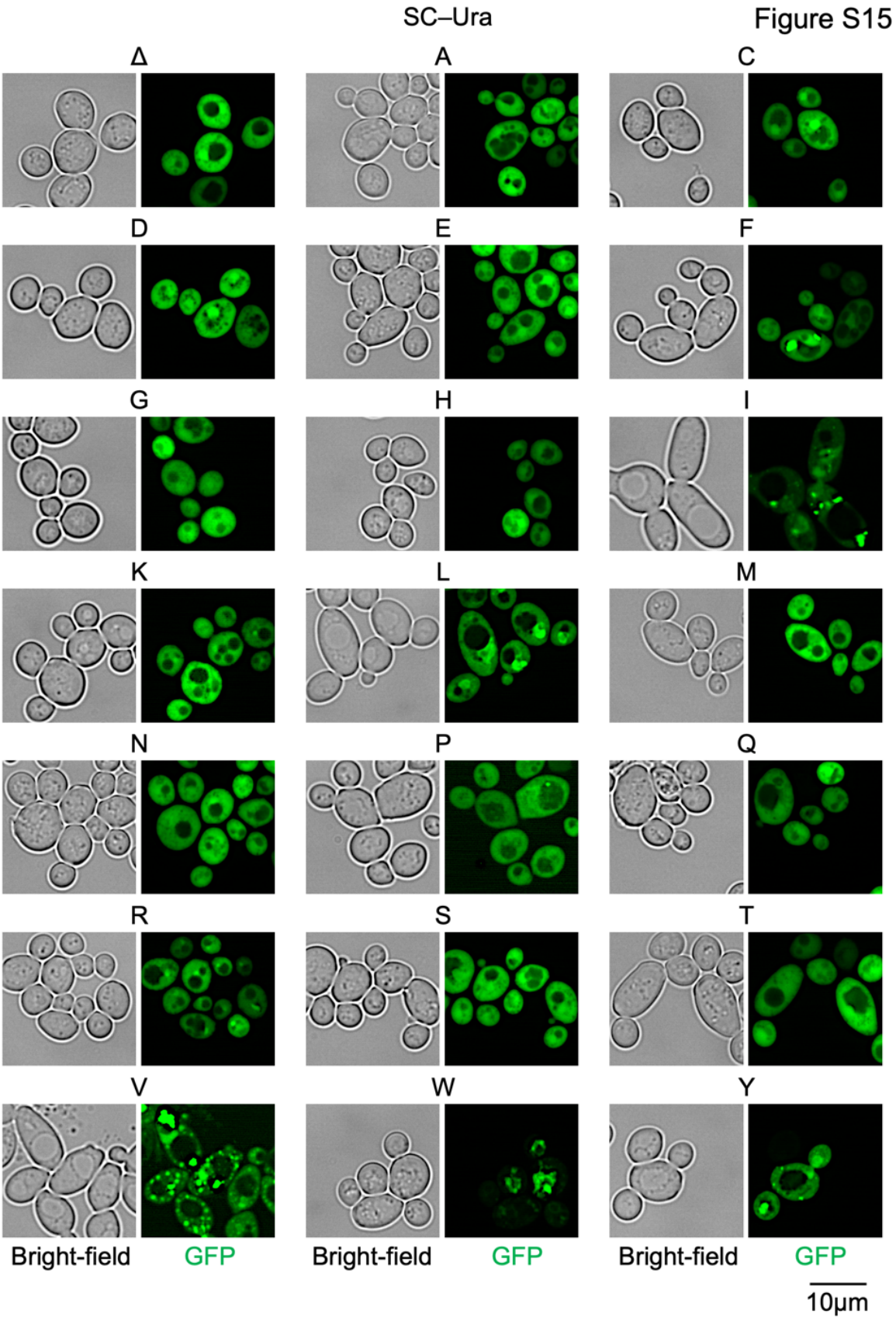
1029

1030 **Figure S14.** Neutrality of $Poly_{10}X$ detached from EGFP via P2A in Yeast.

1031 **A)** Schematic representation of the expression constructs. A self-cleaving P2A sequence was inserted
1032 between EGFP and $Poly_{10}X$, allowing $Poly_{10}X$ to be detached from EGFP during translation. The
1033 construct was expressed under the control of the *TDH3* promoter in *S. cerevisiae*.

1034 **B)** Growth and fluorescence curves of *S. cerevisiae* cells expressing EGFP–P2A–$Poly_{10}X$, measured
1035 using the gTOW method in SC–LU medium at 30 °C. Curves represent the mean values from four
1036 biological replicates, and shaded regions indicate the standard deviation (SD).

1037 **C, D)** Maximum growth rate and maximum fluorescence intensity of *S. cerevisiae* cells overexpressing
1038 EGFP–P2A–$Poly_{10}X$ in SC–LU medium at 30 °C, and the calculated relative neutrality (**D**). Bars, dots,
1039 and error bars represent the mean, individual data points, and standard deviation from four biological
1040 replicates. Asterisks indicate significant differences in maximum growth rate, maximum fluorescence
1041 intensity, and relative neutrality, respectively, compared with the control (Δ) ($p < 0.05$, Student's *t*-test
1042 with Bonferroni correction).


SC–Ura
Figure S15
Δ
A
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
V
W
Y
Bright-field
GFP
Bright-field
GFP
Bright-field
GFP
10μm

1043

1044 **Figure S15.** Fluorescence microscopy of yeast cells expressing EGFP–Poly$_{10}$X

1045 (low-copy conditions).

1046 Fluorescence microscopy images of *S. cerevisiae* cells expressing EGFP–Poly$_{10}$X under the control of

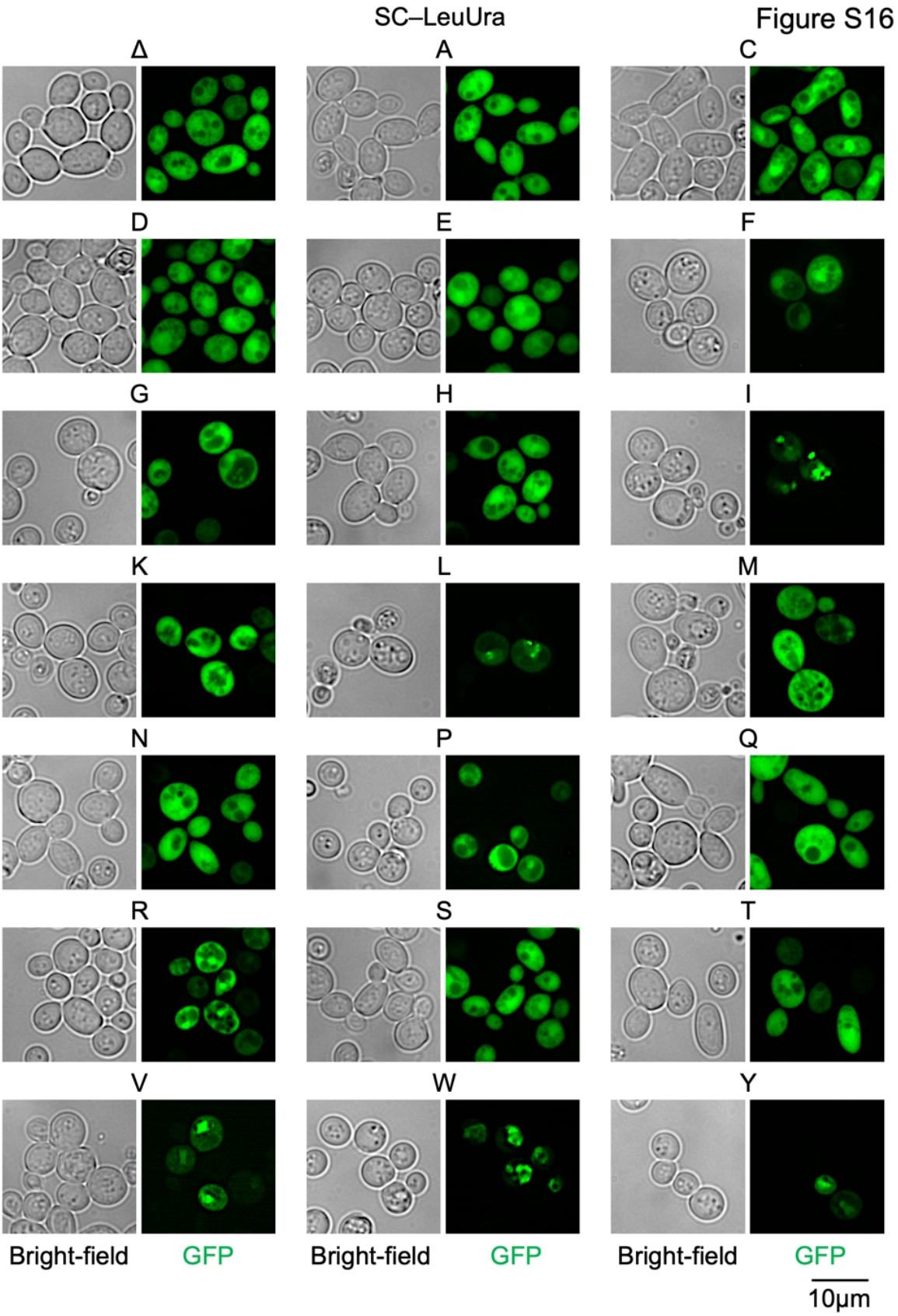
1047 *TDH3$_{pro}$*. Cells were pre-cultured in SC–U medium and subsequently cultured overnight in the same

1048 medium before imaging. Brightness and contrast were adjusted to clearly visualize cell morphology.

1049 For each sample, the bright-field image (left) and the corresponding GFP fluorescence image (right) are

1050 shown. Single-letter codes indicate the amino acid repeated in Poly$_{10}$X. Δ represents the protein without

1051 a Poly$_{10}$X fusion.


SC–LeuUra
Figure S16
Δ
A
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
V
W
Y
Bright-field
GFP
Bright-field
GFP
Bright-field
GFP
10μm

1052

1053 **Figure S16.** Fluorescence microscopy of yeast cells expressing EGFP–Poly$_{10}$X

1054 (High-copy expression).

1055 Fluorescence microscopy images of *S. cerevisiae* cells expressing EGFP–Poly$_{10}$X under the control of

1056 *TDH3$_{pro}$*. Cells were pre-cultured in SC–U medium and subsequently cultured overnight in SC–LU

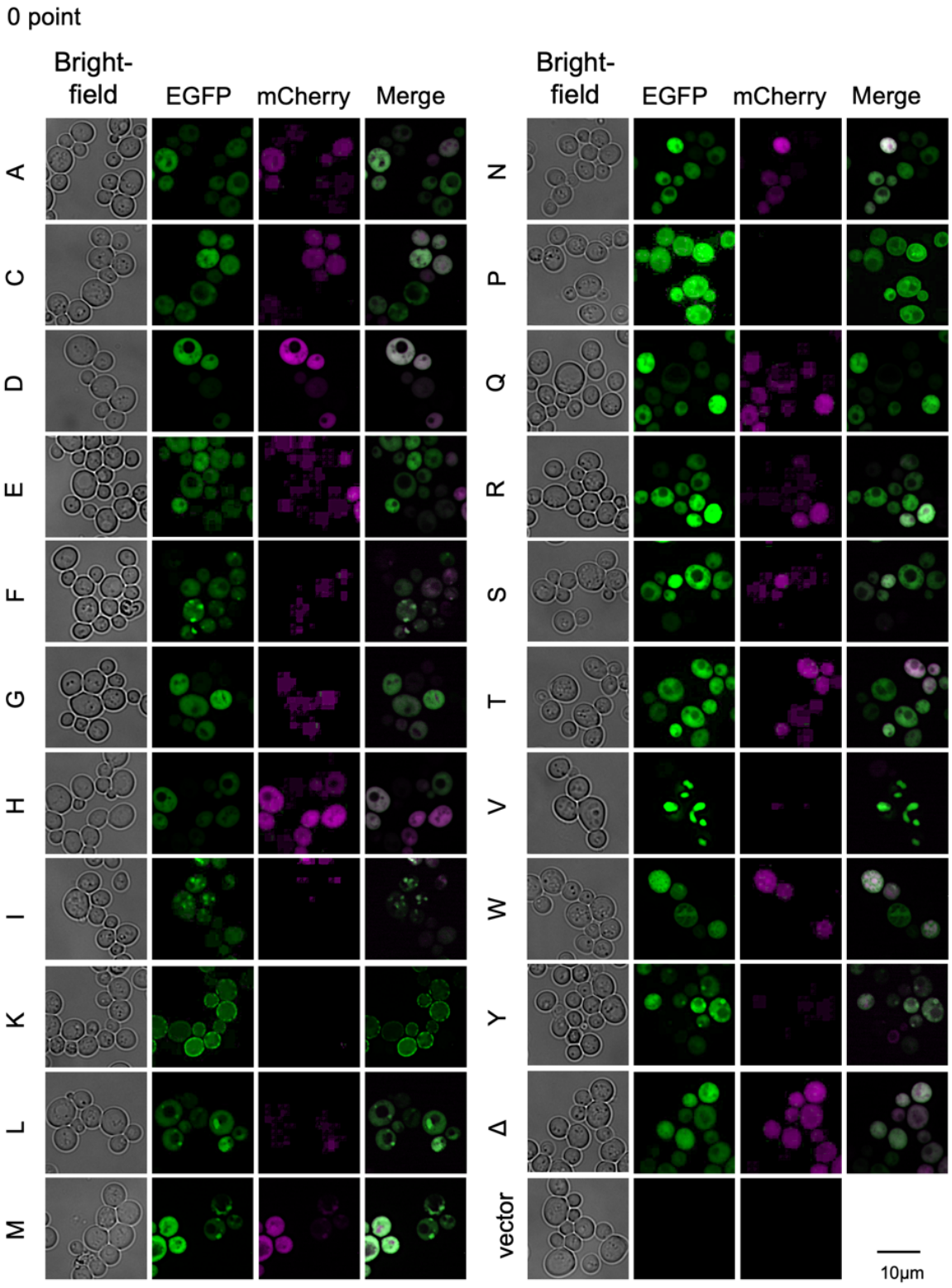
1057 medium before imaging. Brightness and contrast were adjusted to clearly visualize cell morphology.

1058 For each sample, the bright-field image (left) and the corresponding GFP fluorescence image (right) are

1059 shown. Single-letter codes indicate the amino acid repeated in Poly$_{10}$X. Δ represents the protein without

1060 a Poly$_{10}$X fusion.

Figure S17



1061

1062 **Figure S17.** Fluorescence microscopy of EGFP–Poly$_{10}$X–mCherry expression

1063 immediately after induction in yeast.

1064 Fluorescence microscopy images of *S. cerevisiae* cells expressing EGFP–Poly$_{10}$X–GSlinker–mCherry

1065 under the control of the *WTC$_{846}$* promoter immediately after aTc induction. Cells were pre-cultured in

1066 SC–U medium and subsequently cultured overnight in SC–LU medium before imaging. Brightness and

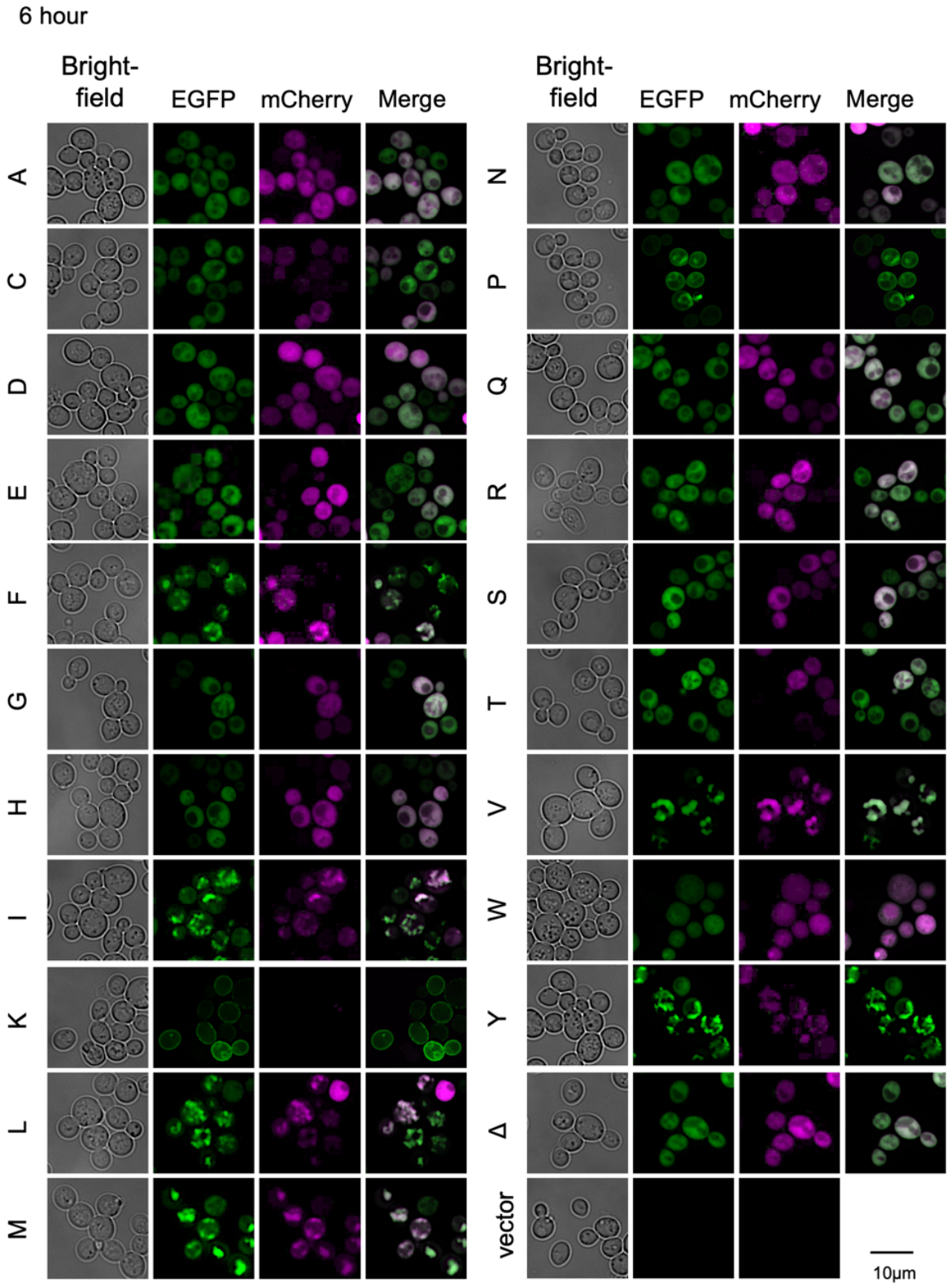
1067 contrast were adjusted to clearly visualize cell morphology. For each sample, the bright-field image

1068 (left), GFP fluorescence image (center left), RFP fluorescence image (center right), and the merged

1069 GFP/RFP image (right) are shown. Single-letter codes indicate the amino acid repeated in Poly$_{10}$X. Δ

1070 represents the protein without a Poly$_{10}$X fusion.

Figure S18



1071

1072 **Figure S18.** Fluorescence microscopy of EGFP–Poly$_{10}$X–mCherry expression

1073 six hours after aTc induction in yeast

1074 Fluorescence microscopy images of *S. cerevisiae* cells expressing EGFP–Poly$_{10}$X–GSlinker–mCherry

1075 under the control of the *WTC$_{846}$* promoter six hours after aTc induction. Cells were pre-cultured in SC–

1076 U medium and subsequently cultured overnight in SC–LU medium before imaging. Brightness and

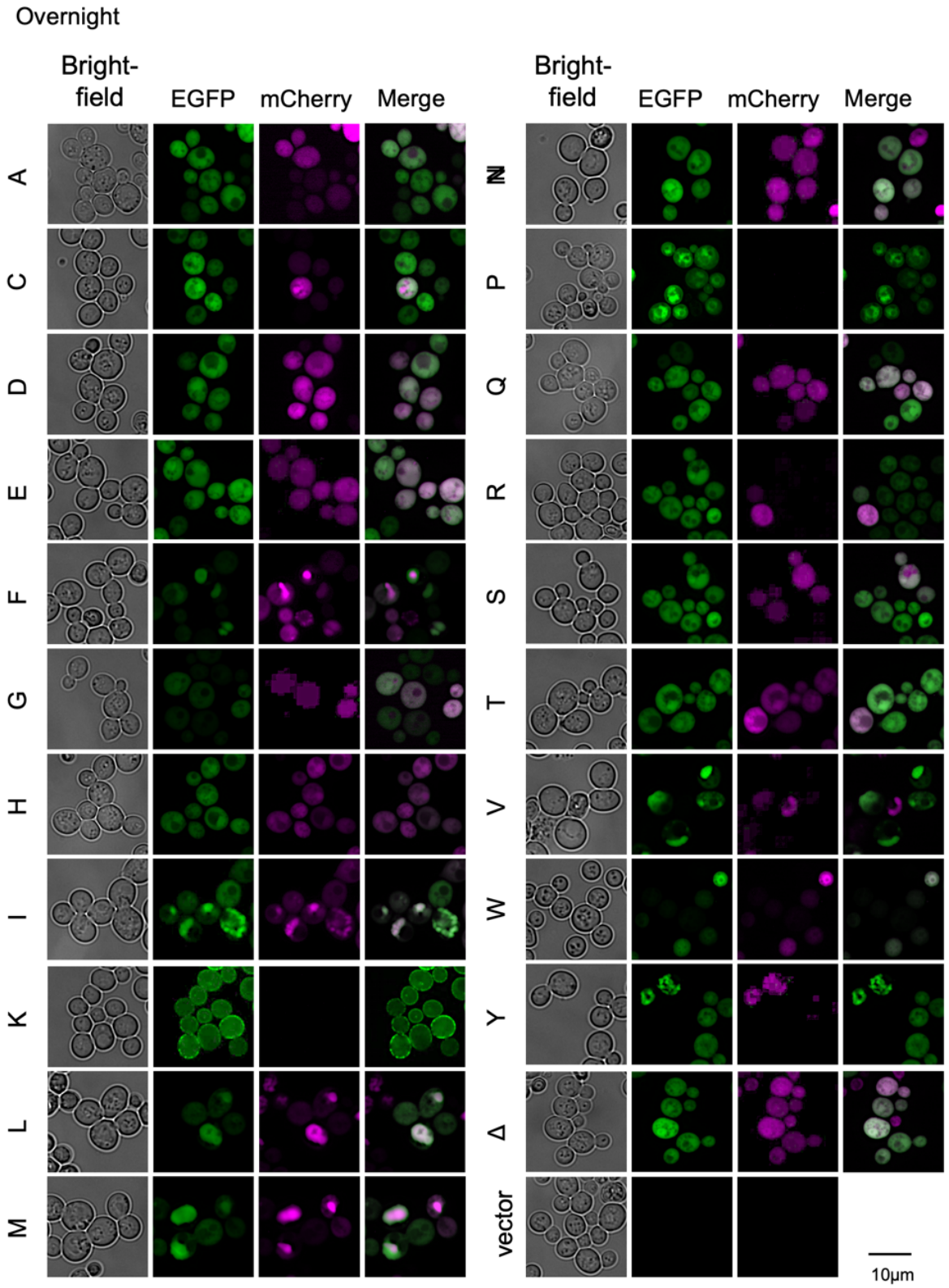
1077 contrast were adjusted to clearly visualize cell morphology. For each sample, the bright-field image

1078 (left), GFP fluorescence image (center left), RFP fluorescence image (center right), and the merged

1079 GFP/RFP image (right) are shown. Single-letter codes indicate the amino acid repeated in Poly$_{10}$X. Δ

1080 represents the protein without a Poly$_{10}$X fusion.

Figure S19



1081

1082 **Figure S19.** Fluorescence microscopy of EGFP–Poly$_{10}$X–mCherry expression

1083 overnight after aTc induction in yeast.

1084 Fluorescence microscopy images of *S. cerevisiae* cells expressing EGFP–Poly$_{10}$X–GSlinker–mCherry

1085 under the control of the *WTC$_{846}$* promoter overnight after aTc induction. Cells were pre-cultured in SC–

1086 U medium and subsequently cultured overnight in SC–LU medium before imaging. Brightness and

1087 contrast were adjusted to clearly visualize cell morphology. For each sample, the bright-field image

1088 (left), GFP fluorescence image (center left), RFP fluorescence image (center right), and the merged

1089 GFP/RFP image (right) are shown. Single-letter codes indicate the amino acid repeated in Poly$_{10}$X. Δ

1090 represents the protein without a Poly$_{10}$X fusion.

Figure S20

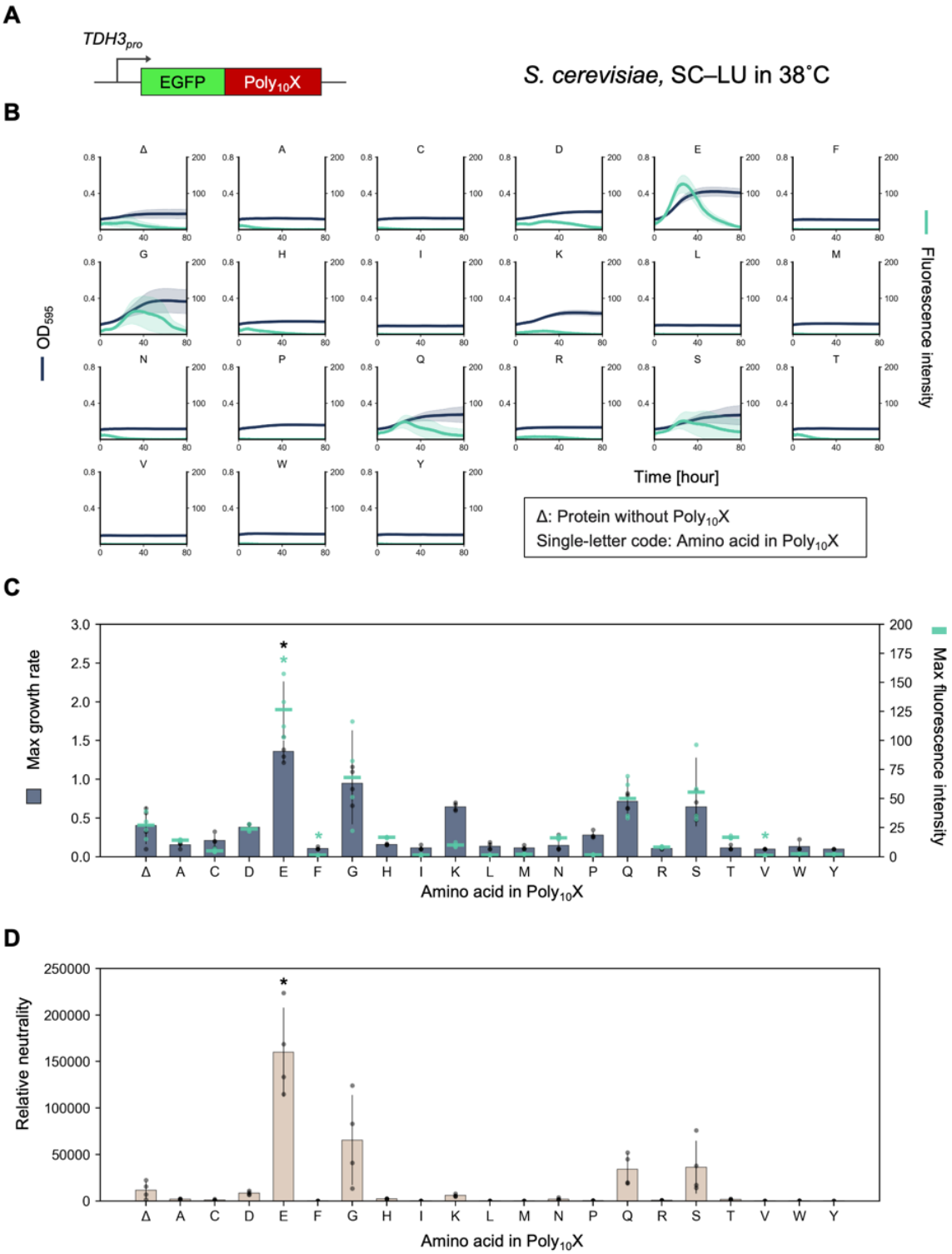


Figure S20. Neutrality of C-terminal $Poly_{10}X$ fusions to EGFP in yeast under heat-stress conditions.

1094 **A)** Schematic representation of the expression constructs. $Poly_{10}X$ was fused to the C-terminus of EGFP
1095 and expressed under the *TDH3* promoter in *S. cerevisiae*.

1096 **B)** Growth and fluorescence curves of *S. cerevisiae* cells expressing EGFP–$Poly_{10}X$, measured using
1097 the gTOW method in SC–LU medium at 38 °C. Curves represent the mean values from at least three
1098 biological replicates, and shaded regions indicate the standard deviation (SD).

1099 **C, D)** Maximum growth rate and maximum fluorescence intensity of *S. cerevisiae* cells overexpressing
1100 EGFP–$Poly_{10}X$ in SC–LU medium at 38 °C, and the calculated relative neutrality (**D**). Bars, dots, and
1101 error bars represent the mean, individual data points, and standard deviation from at least three
1102 biological replicates. Asterisks indicate significant differences in maximum growth rate, maximum
1103 fluorescence intensity, and relative neutrality, respectively, compared with the control (Δ) ($p < 0.05$,
1104 Student's *t*-test with Bonferroni correction).

Figure S21

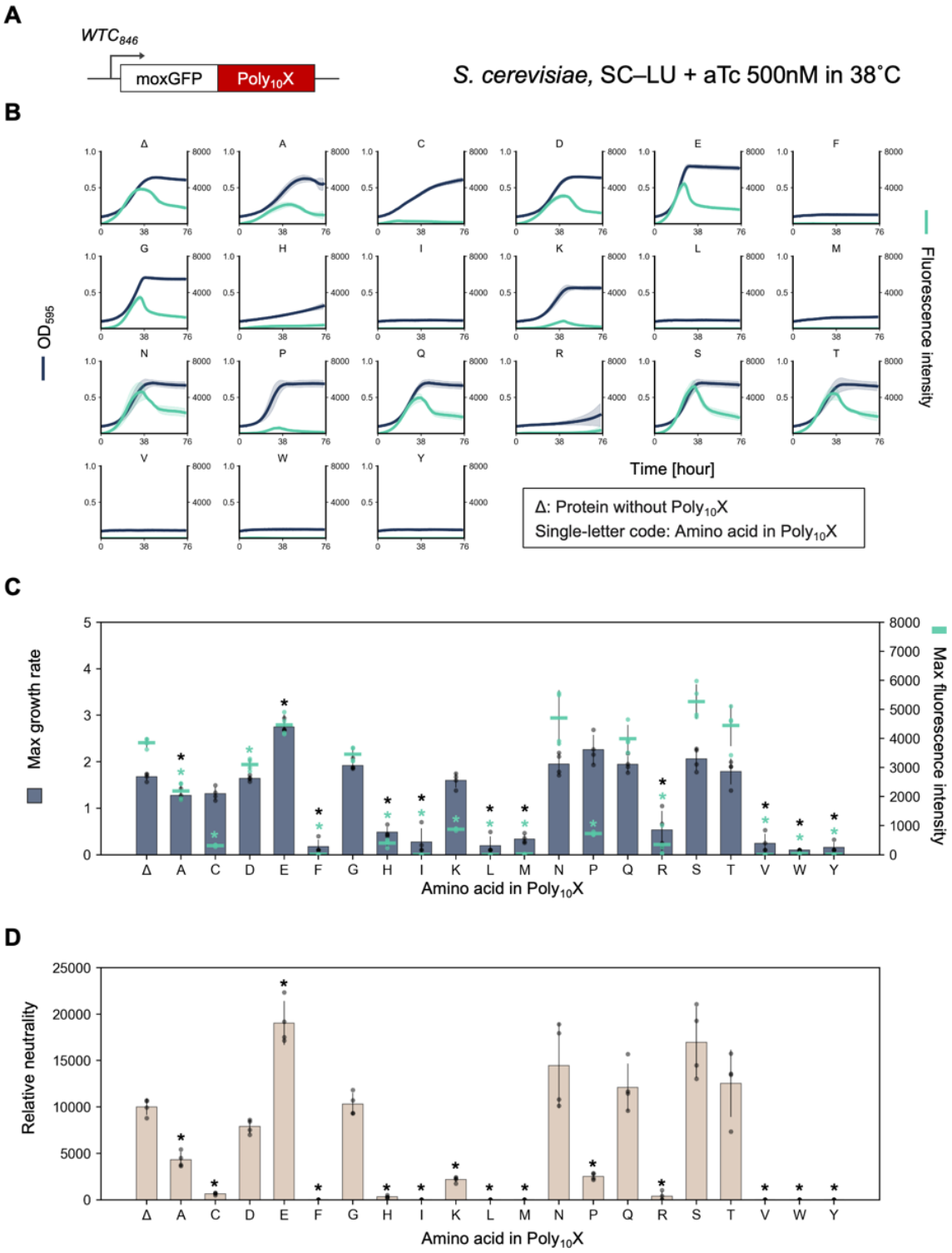


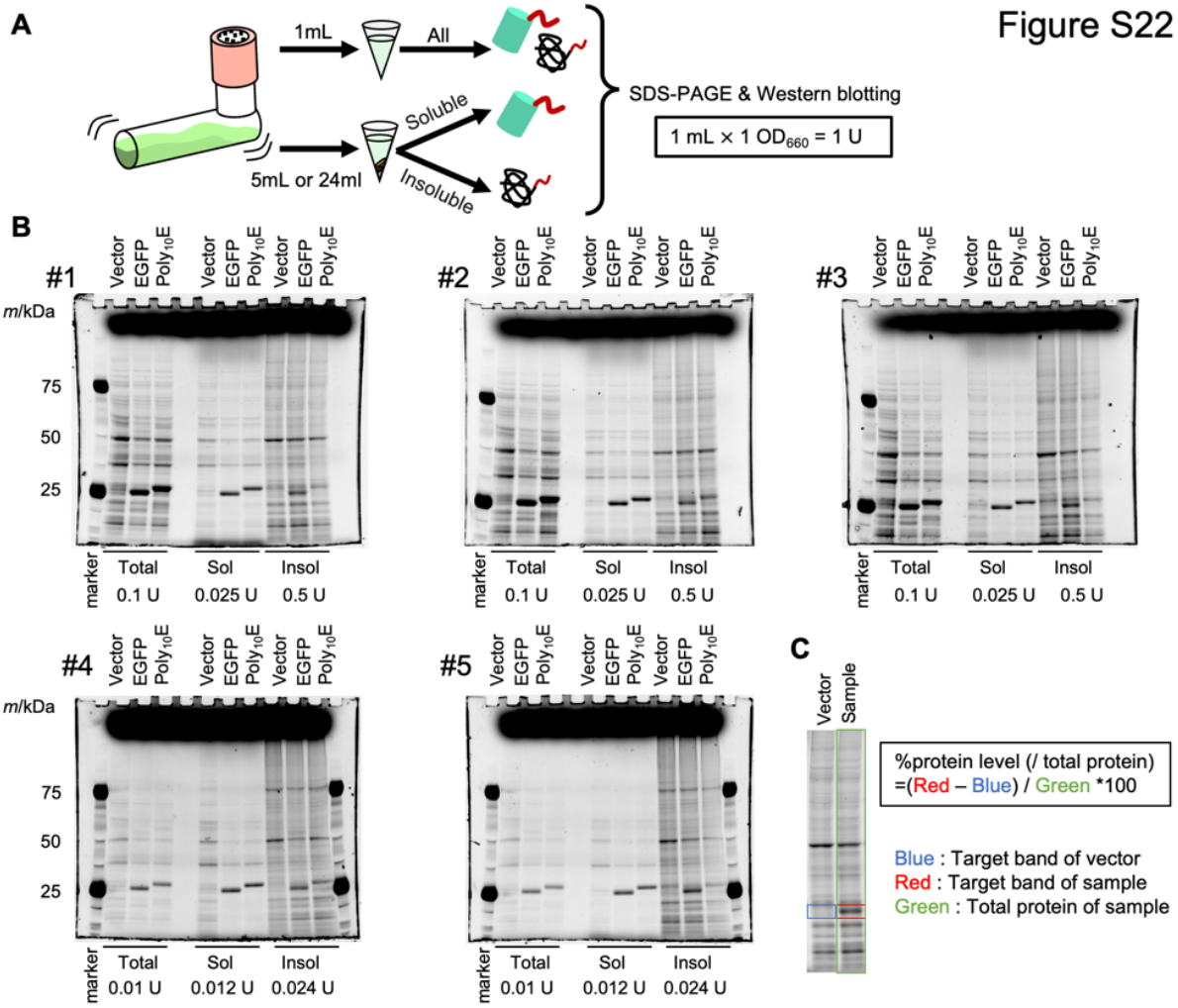
Figure S21. Neutrality of C-terminal Poly$_{10}$X fusions to moxGFP in yeast under heat-stress conditions.

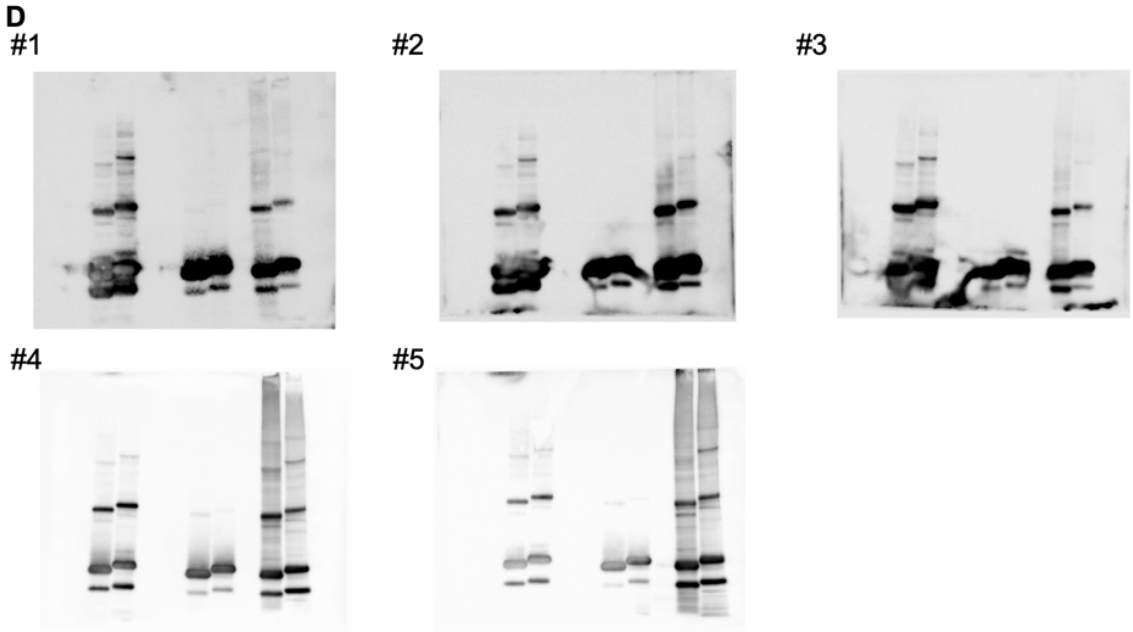
1108 **A)** Schematic representation of the expression constructs. $Poly_{10}X$ were fused to the C-terminus of
1109 moxGFP and expressed under the WTC_{846} promoter in *S. cerevisiae*.

1110 **B)** Growth and fluorescence curves of *S. cerevisiae* cells expressing moxGFP–$Poly_{10}X$, measured using
1111 the gTOW method in SC–LU medium at 38 °C. Curves represent the mean values from at least three
1112 biological replicates, and shaded regions indicate the standard deviation (SD).

1113 **C, D)** Maximum growth rate and maximum fluorescence intensity of *S. cerevisiae* cells overexpressing
1114 moxGFP–$Poly_{10}X$ in SC–LU medium at 38 °C, and the calculated relative neutrality (**D**). Bars, dots,
1115 and error bars represent the mean, individual data points, and standard deviation from at least three
1116 biological replicates. Asterisks indicate significant differences in maximum growth rate, maximum
1117 fluorescence intensity, and relative neutrality, respectively, compared with the control (Δ) ($p < 0.05$,
1118 Student's *t*-test with Bonferroni correction).

Figure S22





1119

1120 **Figure S22.** SDS–PAGE and Western blot analysis of EGFP and EGFP–

1121 $Poly_{10}E$ proteins in yeast

1122 **A)** Schematic overview of the protein extraction procedure. A 25 mL culture was collected at

1123 approximately OD_{660} = 1. One milliliter was used for total protein extraction, while the remaining 5 or

1124 24 mL was fractionated into soluble and insoluble fractions. Each fraction was analyzed by SDS–PAGE,

1125 followed by western blotting. Protein concentrations were normalized based on the total protein amount

1126 extracted from cells in 1 mL of culture at OD_{660} = 1, which was defined as 1 unit (1 U).

1127 **B)** SDS–PAGE images of proteins extracted from cells with the control vector (Vector) or

1128 overexpressing EGFP or EGFP–$Poly_{10}E$ ($Poly_{10}E$). Cells were cultured in SC–LU medium and

1129 collected at approximately OD_{660} = 1. Lanes correspond to total, soluble (Sol), and insoluble (Insol)

1130 fractions (from left to right). Protein bands were visualized using a chemiluminescent detection reagent.

1131 Five biological replicates were analyzed. Protein loading concentrations are indicated in the figure.

1132 **C)** Quantification method for %protein level from SDS–PAGE images shown in **B**. Signal intensity

1133 corresponding to the EGFP band (Red) was corrected by subtracting the background intensity from the

1134 same region (Blue), and the resulting value was normalized to the total protein amount in the sample

1135 (green) to calculate the percentage.

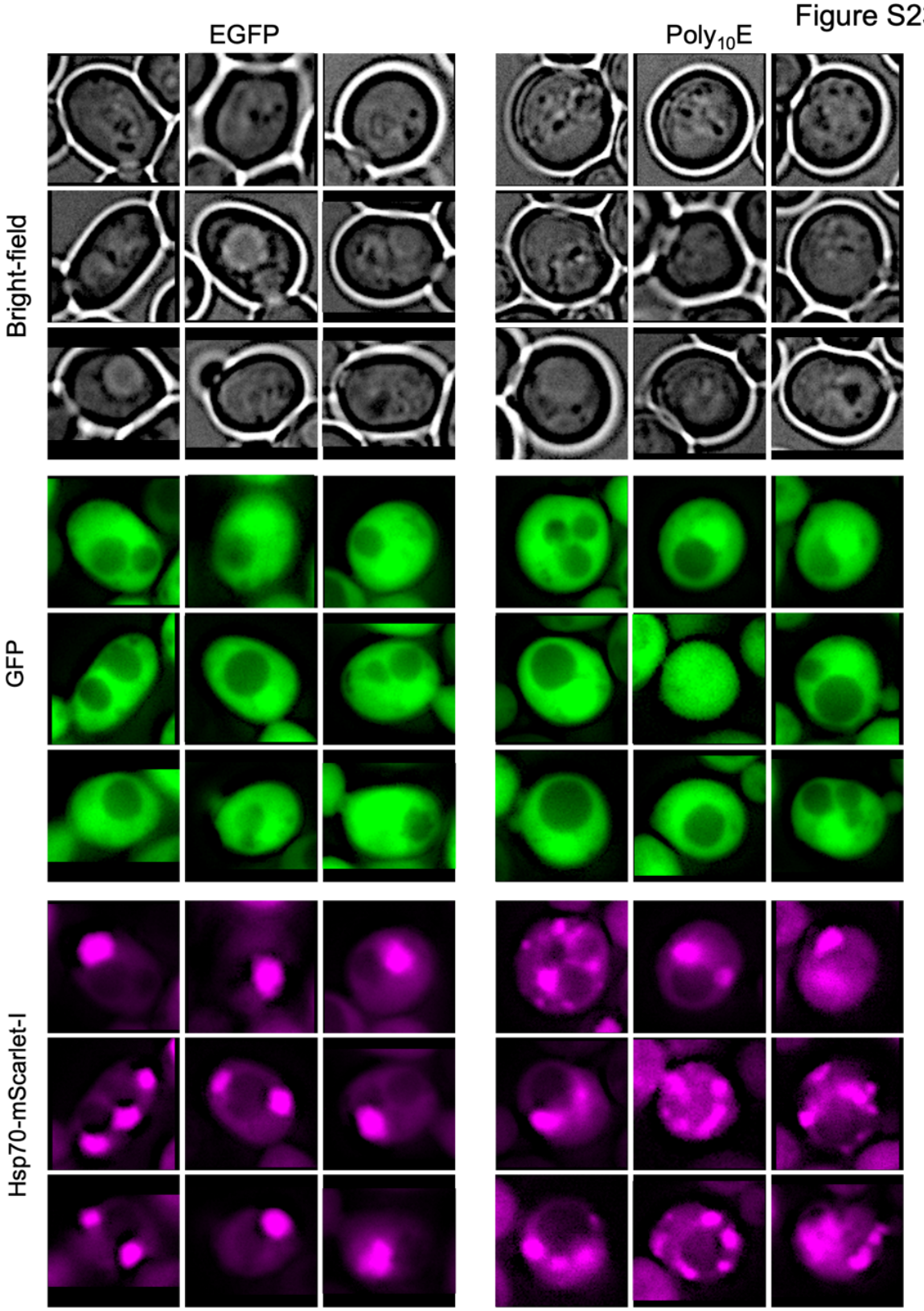
1136 **D)** Western blot images of proteins extracted from cells with the control vector (Vector) or

1137 overexpressing EGFP or EGFP–$Poly_{10}E$ ($Poly_{10}E$). The same gels used in **B** were transferred to

1138 membranes and probed with anti-GFP antibodies. Lanes correspond to total, soluble, and insoluble

1139 fractions (from left to right).

Figure S23



1140

1141 **Figure S23.** Fluorescence imaging of Hsp70 foci in yeast cells overexpressing

1142 EGFP or EGFP–Poly$_{10}$E.

1143 Fluorescence microscopy images of Hsp70 foci in cells overexpressing EGFP or EGFP–Poly$_{10}$E under

1144 the control of *TDH3*$_{pro}$ in SC–LU conditions. Hsp70–mScarlet-I was genomically integrated to visualize

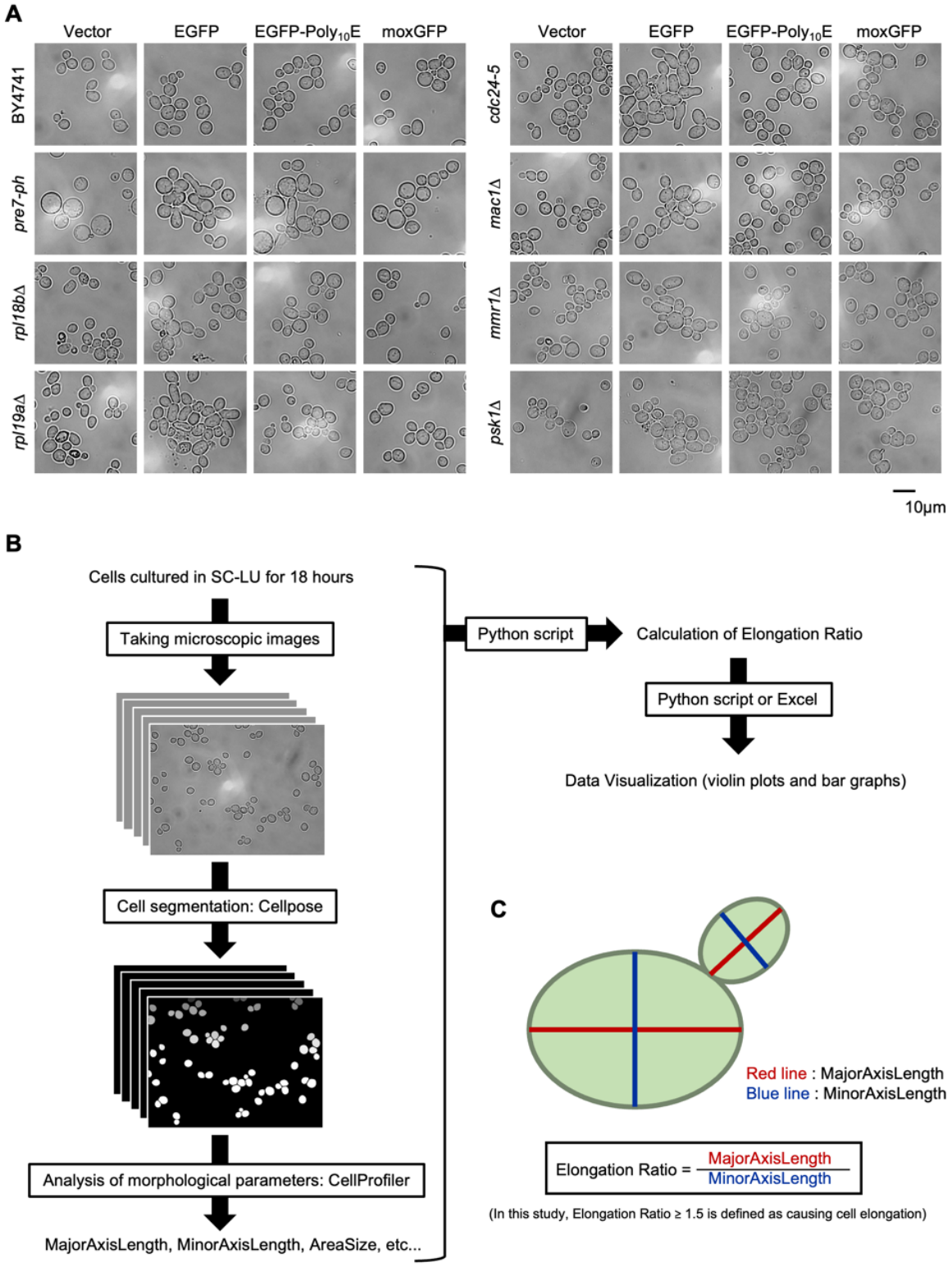
1145 Hsp70 aggregate formation. Representative images of nine individual cells are shown, along with the

1146 corresponding GFP and RFP fluorescence images. Green indicates EGFP or EGFP–Poly$_{10}$E, and

1147 magenta represents the distribution of Hsp70. Image brightness and contrast were adjusted to enhance

1148 the visibility of aggregates.

Figure S24



1149

1150 **Figure S24.** Morphological analysis of yeast strains overexpressing EGFP or

1151 EGFP–Poly$_{10}$E

1152 **A)** Microscopy images of *S. cerevisiae* strain BY4741 and seven mutant strains (*pre7-ph*, *rpl18bΔ*,

1153 *rpl19aΔ*, *cdc24-5*, *mac1Δ*, *mmr1Δ*, and *psk1Δ*) with the control vector (Vector) or overexpressing EGFP,

1154 EGFP–Poly$_{10}$E, or moxGFP under the control of *TDH3$_{pro}$*. Cells were pre-cultured in SC–U medium

1155 and then cultured for 18 h in SC–LU medium before imaging. Brightness and contrast were adjusted to

1156 clearly visualize cell morphology.

1157 **B)** Schematic workflow of image analysis. After culturing the cells for 18 h in SC–LU medium,

1158 microscopic images were acquired and processed using Cellpose 2.2.3 [64] for cell segmentation. The

1159 segmented images were then analyzed with the *MeasureObjectSizeShape* module in CellProfiler 4.2.6

1160 [65] to measure the major and minor axes of each cell. Elongation ratio (major axis/minor axis) values

1161 were calculated using a Python script or Excel and visualized.

1162 **C)** Schematic representation of cell elongation analysis. In this study, cells with an elongation ratio ≥

1163 1.5 were defined as morphologically abnormal.

Figure S25

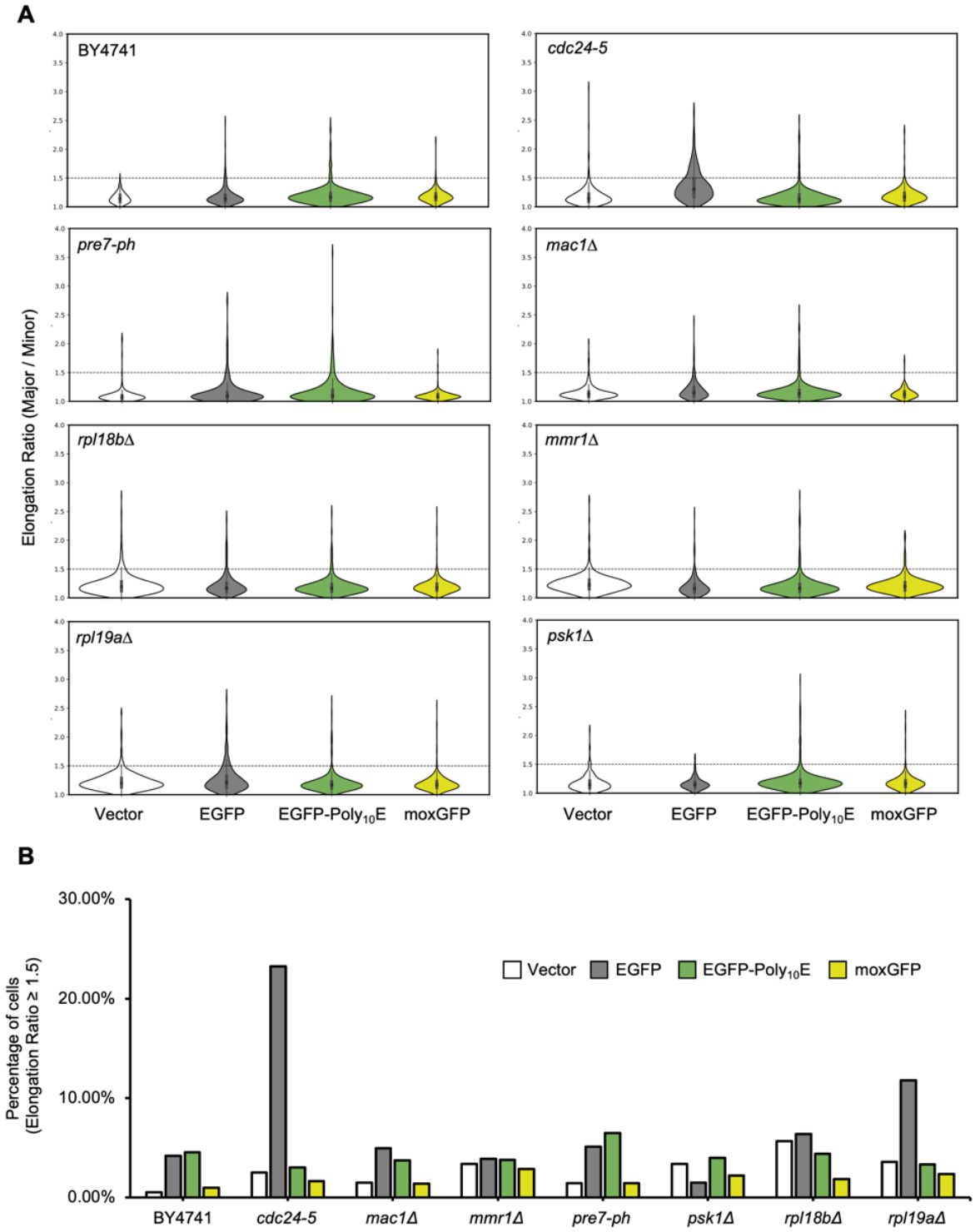


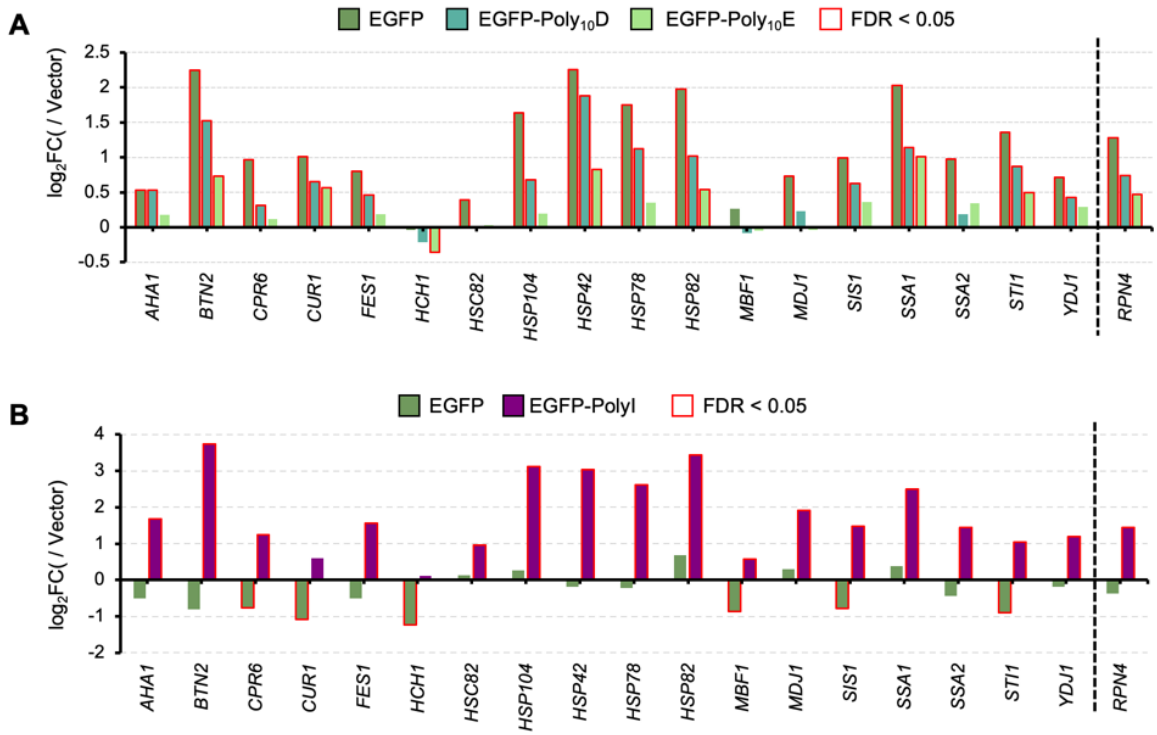
Figure S25. Quantification of cell elongation defects in yeast strains overexpressing EGFP or EGFP–Poly$_{10}$E.

1167 **A)** Distribution of elongation ratios in *S. cerevisiae* strain BY4741 and seven mutant strains (*pre7-ph*,
1168 *rpl18bΔ*, *rpl19aΔ*, *cdc24-5*, *mac1Δ*, *mmr1Δ*, and *psk1Δ*) with the control vector (Vector) or
1169 overexpressing EGFP, EGFP–$Poly_{10}E$, or moxGFP under the control of $TDH3_{pro}$. Each violin plot
1170 represents the elongation ratio (major axis/minor axis) of individual cells. Wider regions indicate a
1171 higher frequency of cells with the corresponding elongation ratio. The dashed horizontal line represents
1172 the threshold for morphological abnormality (elongation ratio ≥ 1.5). Morphological parameters were
1173 calculated from microscopic images analyzed using Cellpose and CellProfiler as described in Figure
1174 S24B and C.

1175 **B)** Proportion of morphologically abnormal cells (elongation ratio ≥ 1.5) in each strain. The analysis
1176 was performed on a single biological replicate, with at least 197 cells analyzed per condition. Bars
1177 indicate the percentage of abnormal cells relative to the total number of segmented cells.

1178

Figure S26



1179

1180 **Figure S26.** Expression changes of Hsf1-regulated genes and *RPN4* in yeast

1181 expressing EGFP-Poly$_{10}$D, EGFP-Poly$_{10}$E, or EGFP-Poly$_{10}$I.

1182 **A, B)** Comparison of expression ratios (vs. Vector) for genes regulated by Hsf1 and for *RPN4*,

1183 quantified by RNA-seq. In **A**, bars indicate the expression ratios of cells overexpressing EGFP, EGFP–

1184 Poly$_{10}$D, and EGFP–Poly$_{10}$E, respectively. In **B**, bars indicate the expression ratios of cells with low-

1185 level overexpressing EGFP and EGFP–Poly$_{10}$I, respectively. Outlines denote genes showing significant

1186 differential expression compared with the vector control (FDR < 0.05).

Figure S27

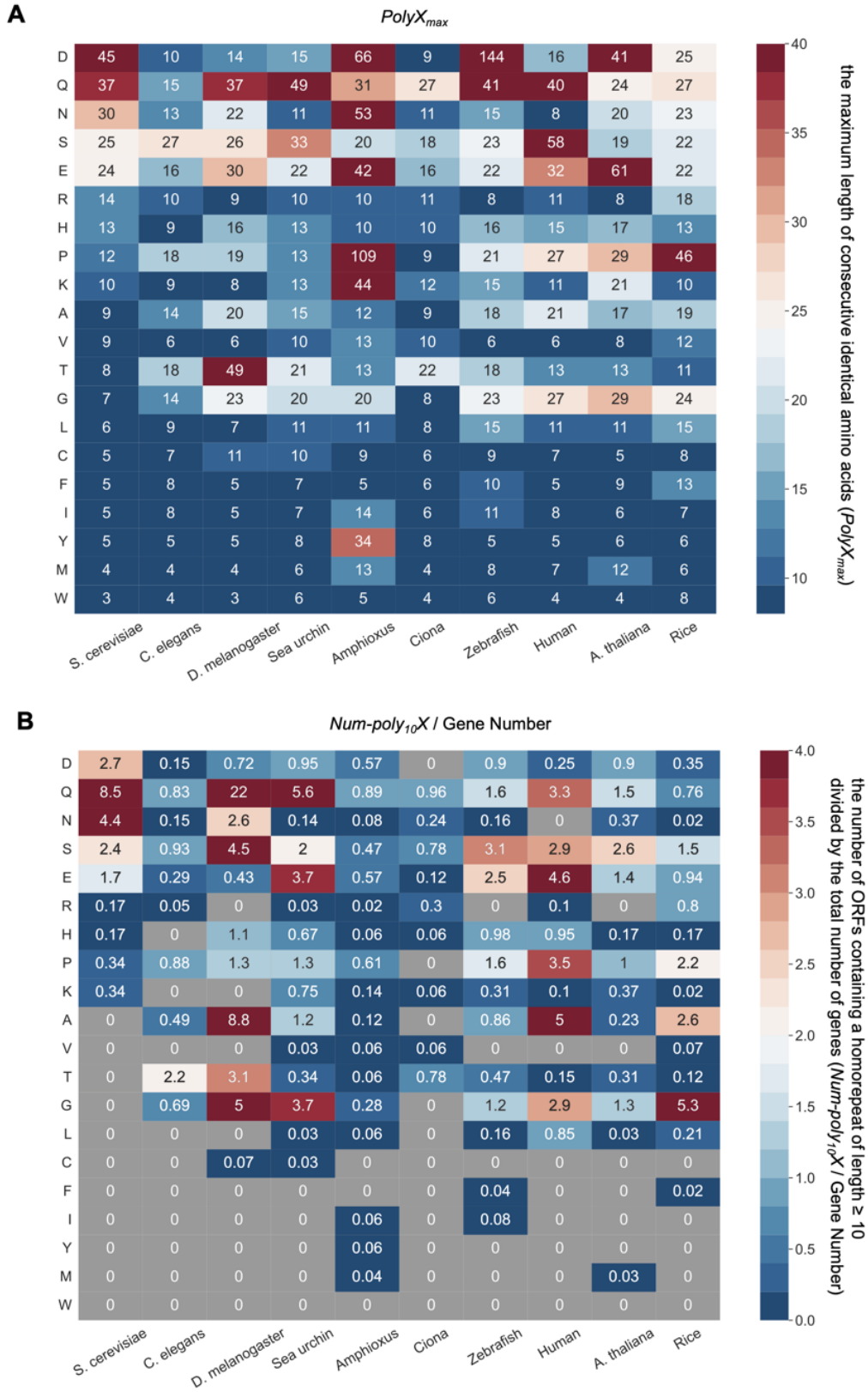


Figure S27. PolyX occurrence across diverse species.

1189 **A)** Heatmap of $PolyX_{max}$ values (maximum homorepeat length for each amino acid) across multiple
1190 species.

1191 **B)** Heatmap of $Num\text{-}Poly_{10}X$ values across the same species. Each value represents the number of
1192 proteins containing ≥10-residue homorepeats per amino acid, normalized by the total number of genes
1193 in the species and multiplied by 1000. Gray boxes indicate amino acids for which no homorepeats of
1194 length ≥10 were detected.

Figure S28

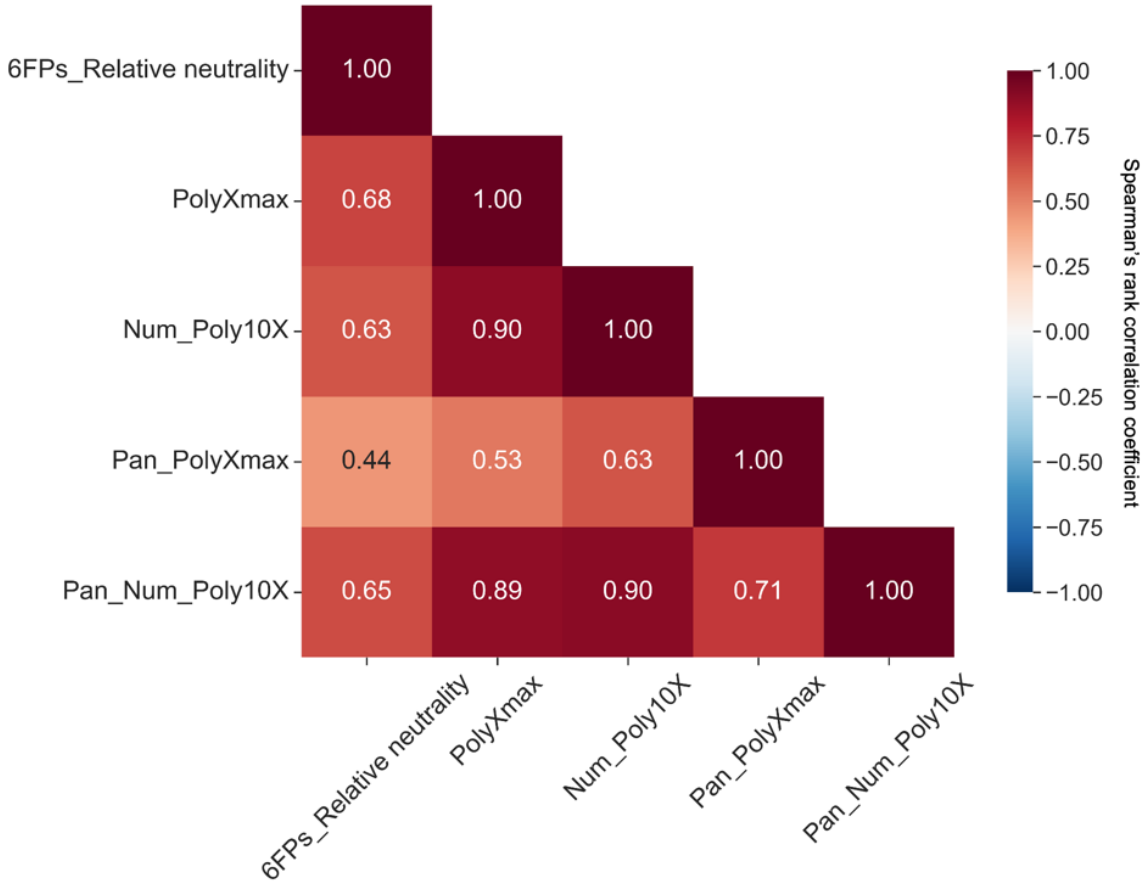


Figure S28. Correlation analysis between experimental $Poly_{10}X$ harmfulness patterns and proteome-wide PolyX occurrence trends.

Correlation matrix comparing $Poly_{10}X$ harmfulness patterns observed across different fluorescent proteins with amino acid–level $PolyX_{max}$ and Num-$Poly_{10}X$ values obtained from proteome-wide computational analyses. Each value represents the Spearman's rank correlation coefficient for the corresponding pairwise comparison.

Figure S29

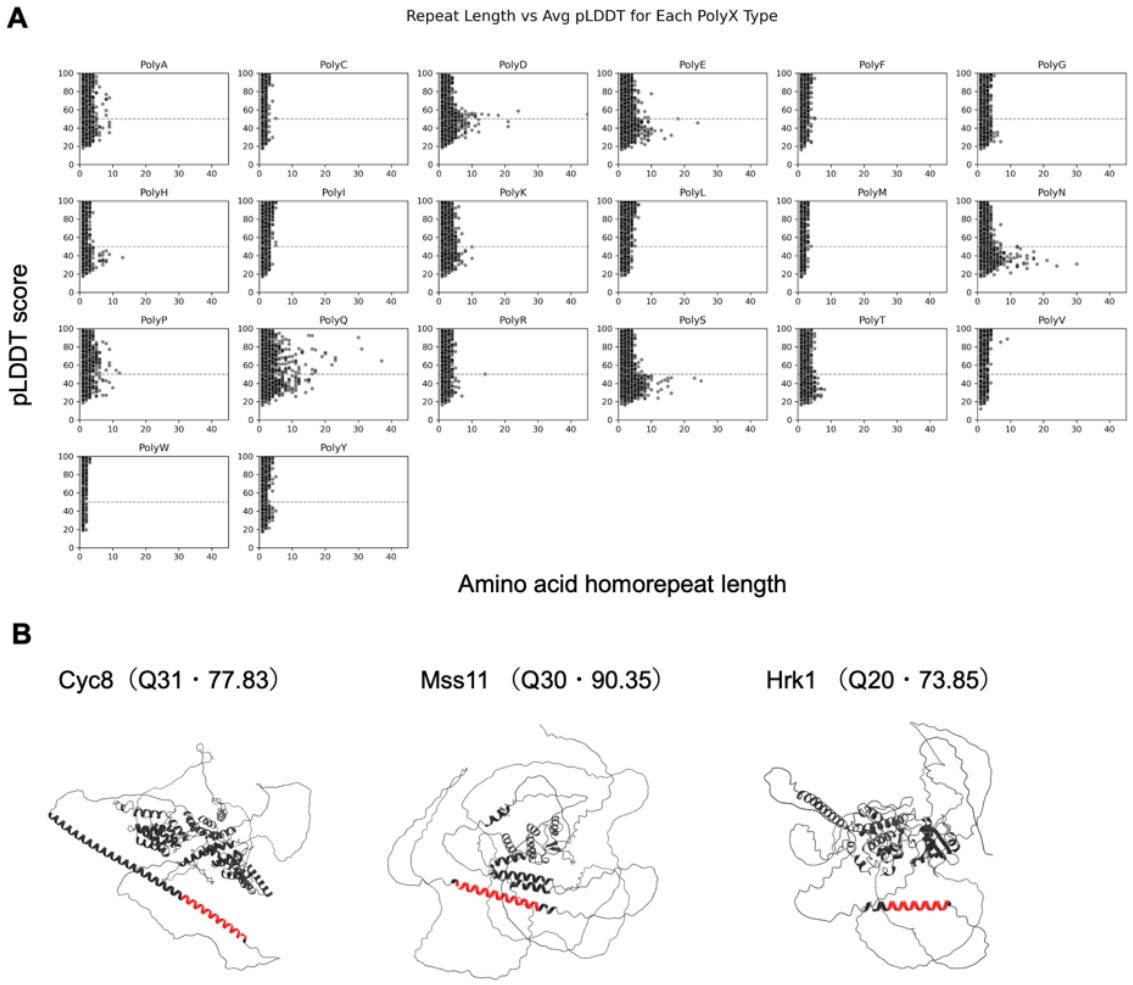


Figure S29. Structural context of PolyX homorepeats in the *S. cerevisiae* proteome based on AlphaFold pLDDT scores.

A) Distribution plots showing the relationship between amino acid homorepeat length in the *S. cerevisiae* proteome and the corresponding pLDDT scores obtained from the AlphaFold structural prediction dataset (https://alphafold.ebi.ac.uk/). Each dot represents a sequence region within a protein. The pLDDT score indicates prediction confidence (100–90: very high; 90–70: confident; 70–50: low; 50–0: very low). Regions with low pLDDT scores likely correspond not only to low model confidence but also to intrinsically disordered regions (IDRs), which do not adopt a fixed 3D structure. This analysis was performed to examine whether naturally occurring PolyX regions tend to reside within

1212 well-structured protein cores or within flexible, disordered regions. For most amino acids, longer
1213 homorepeats were associated with lower pLDDT scores, indicating enrichment in unstructured regions.
1214 In contrast, PolyQ (glutamine) repeats maintained relatively high pLDDT scores even at longer lengths.

1215 **B)** Representative structures of PolyQ-containing proteins, with PolyQ segments that exhibit pLDDT ≥
1216 70 highlighted in red. Although these PolyQ regions show high prediction confidence, they appear to
1217 be located away from the structured protein core. This suggests that PolyX tracts, including PolyQ, are
1218 generally enriched in flexible or intrinsically disordered regions rather than in tightly structured core
1219 domains.