The role of break-induced replication in large-scale expansions of (CAG)$_n$/(CTG)$_n$ repeats

Jane C Kim$^{1,2}$, Samantha T Harris$^1$, Teresa Dinter$^1$, Kartik A Shah$^{1,2}$ & Sergei M Mirkin$^1$

Expansions of (CAG)$_n$/(CTG)$_n$ trinucleotide repeats are responsible for over a dozen neuromuscular and neurodegenerative disorders. Large-scale expansions are commonly observed in human pedigrees and may be explained by iterative small-scale events such as strand slippage during replication or repair DNA synthesis. Alternatively, a distinct mechanism may lead to a large-scale repeat expansion as a single step. To distinguish between these possibilities, we developed a novel experimental system specifically tuned to analyze large-scale expansions of (CAG)$_n$/(CTG)$_n$ repeats in Saccharomyces cerevisiae. The median size of repeat expansions was ~60 triplets, although we also observed additions of more than 150 triplets. Genetic analysis revealed that Rad51, Rad52, Mre11, Pol32, Pif1, and Mus81 and/or Yen1 proteins are required for large-scale expansions, whereas proteins previously implicated in small-scale expansions are not involved. From these results, we propose a new model for large-scale expansions, which is based on the recovery of replication forks broken at (CAG)$_n$/(CTG)$_n$ repeats via break-induced replication.

Expansions of (CAG)$_n$/(CTG)$_n$ repeats are responsible for over a dozen neuromuscular and neurodegenerative disorders in humans, including Huntington's disease (HD), myotonic dystrophy (DM1), and numerous forms of spinocerebellar ataxia$^{1,2}$. Individuals with adult-onset HD typically have 40–80 (CAG)$_n$ repeats in the coding region of the HTT gene. Longer CAG tracts do occur but are rare and are associated with juvenile onset$^6$. In contrast, individuals with DM1 commonly have hundreds of (CTG)$_n$ repeats in the 3’ untranslated region of the DMPK gene, and up to 4,000 copies may be present in severe cases$^6$. The molecular mechanisms of (CAG)$_n$/(CTG)$_n$ (hereafter abbreviated CAG)-repeat expansions have been intensively studied in model organisms and human cells, and the results recapitulate many properties observed in human patients and pedigrees, such as a length-dependent increase in repeat instability$^{2,6}$. Both in vitro and in vivo, CAG sequences have been shown to form stable hairpins and slipped-strand DNA structures, which stall replication forks, promote replication-fork reversal, and cause chromosomal breakage in a length-dependent manner$^{7-9}$.

All models of CAG-repeat expansions suggest deleterious effects of their secondary structures on DNA replication, transcription, and repair processes$^{10,11}$. DNA polymerase slippage followed by hairpin formation on the nascent DNA strand can lead to small-scale expansions if the hairpin persists to the next round of replication$^{6,11}$. Strand slippage and hairpin formation can also occur during repair DNA synthesis in the course of base excision repair$^{12,13}$, nucleotide excision repair$^{14}$, and transcription-coupled repair$^{15}$. In all of the above scenarios, expansion size is limited by slippage events that are normally small in scale. Thus, these models can explain large-scale expansions through the iterative succession of independent small-scale events. For example, oxidized DNA bases can lead to subsequent base excision repair, in which strand displacement creates a DNA hairpin that is refractory to cleavage by flap endonuclease. This hairpin would then result in a single expansion, and many rounds of oxidation, repair, and expansions would create a ‘toxic oxidation cycle’ that generates large-scale expansions$^{13}$.

Most experimental systems to study CAG-repeat expansions focus on relatively small-scale events$^{16-20}$, which we define as an increase of up to 20 repeats. The first selectable system described in budding yeast deliberately examined the instability of a short (i.e., (CAG)$_{25}$) starting tract, which simulated a change from normal- to premutation-length alleles, as in HD16. In this system, approximately ten repeats were added at a rate of ~10$^5$. Yeast studies of longer CAG repeats (45–155 units) have consistently detected small-scale expansions occurring at a level of ~1% (ref. 17). Altogether, these yeast systems have enabled powerful genetics analysis of small-scale repeat expansion, thus establishing the importance of replication-fork integrity, chromatin remodeling, specialized helicases, and nuclear localization of the repeats$^{8,21-25}$.

In a Drosophila experimental system, the scale of repeat expansions has been found to be even smaller: the majority of events are additions of only one or two repeats to a long (CAG)$_{270}$ tract$^{18}$. In mice, compared with humans, much longer CAG repeats are required to produce disease phenotypes. Similarly to yeast, mice display predominantly small-scale expansions during intergenerational transmission and in somatic tissues$^{13,26}$. An exception is the case of the remarkably small humanized DM1 mouse, which bear 430 to >1,000 CAG repeats exhibiting jumps in excess of hundreds of repeats during intergenerational transmission$^{27}$. Hairpin formation and the role of replication in CAG-repeat instability has been confirmed in human cells$^{28,29}$.

1Department of Biology, Tufts University, Medford, Massachusetts, USA. 2Present addresses: Department of Biological Sciences, California State University San Marcos, San Marcos, California, USA (J.C.K.) and Amgen, Cambridge, Massachusetts, USA (K.A.S.). Correspondence should be addressed to S.M.M. (sergei.mirkin@tufts.edu).

Received 12 August; accepted 31 October; published online 5 December 2016; doi:10.1038/nsmb.3334
Additionally, large-scale expansions have been recovered from a very long starting tract of 800 repeats (ref. 30). However, in such experimental systems, extensive genetic analyses remain challenging.

Given the lack of experimental systems available to detect large-scale CAG expansions, it is impossible to ascertain whether these expansions occur via a distinct mechanism or result from the sequential accumulation of small-scale expansions. Previous studies of large-scale expansions of (GAA)\(_n\)/(TTC)\(_n\) and (ATTCT)\(_n\)/(AGAAT)\(_n\) repeats in a yeast system have led us to propose a template-switching model for large-scale expansion during DNA replication\(^3\,31,32\). Genetic analysis of these large-scale events has revealed dramatic differences compared with small-scale CAG-repeat expansions studied by others. However, it has been unclear whether differences in the scale of expansions, the repeat sequences, or the experimental model systems account for these differences.

To address this problem, we established a new system to detect and analyze large-scale CAG expansions in S. cerevisiae. The median size of repeat expansions in this system was ~60 triplets, although we also observed additions of more than 150 triplets. Our genetic analysis revealed that Rad51, Rad52, Mre11, Pol32, Pif1 and Mus81 or Yen1 proteins are required for large-scale expansions, whereas proteins previously implicated in small-scale expansions are not involved. These results suggest a mechanism that is distinct from that of small-scale CAG expansions and is based on the recovery of replication forks broken at (CAG)\(_n\)/(CTG)\(_n\) repeats through the interplay between break-induced replication and broken-fork repair (BFR). Thus, large-scale CAG expansions may provide a striking example of genome instability arising from the break-induced-replication machinery.

RESULTS
Large-scale CAG-repeat expansions can be recovered in budding yeast
To investigate large-scale CAG expansions, we used a system previously developed to study GAA repeats\(^3\,33\). This system relies on the well-characterized GAL1 promoter, in which the distance (i.e., the spacer) between the upstream activating sequence (UAS GAL) and TATA box (P\(_{GAL}\)) is constrained such that transcriptional activation no longer occurs when the spacer is too long\(^3\,34\). We cloned CAG repeats into this spacer, upstream of the forward-selection marker CAN1. We then integrated these constructs into chromosome III, ~1 kb away from the replication origin ARS306 (Fig. 1a and Supplementary Figs. 1 and 2).

Our starting strain contained a selectable cassette with 140 CAG repeats, which correspond to a number associated with mild DM1 and undergo large-scale expansions during intergenerational transmission. We positioned the CAG sequence on the lagging-strand template for DNA replication because the CTG orientation is known to be deletion prone\(^3\,35,36\). We reasoned that if a large-scale expansion (which we designate as >20 repeats) occurred during nonselective growth, the increased spacer distance would preclude the expression of CAN1 and consequently permit colony formation on plates containing canavanine, a toxic analog of arginine, in the presence of galactose (Fig. 1a).

We determined the rate of large-scale CAG expansions in fluctuation test experiments in which nonselective growth occurred on either glucose or galactose. We then determined the length of CAG repeats in individual canavanine-resistant (Can\(^R\)) clones through single-colony PCR (Fig. 1b and Supplementary Fig. 3). The rate of large-scale (CAG)\(_{140}\) expansion corresponded to 1.4 × 10\(^{-5}\) per replication in cells pregrown on galactose and was ten-fold lower (1.0 × 10\(^{-6}\)) in cells pregrown on glucose (Fig. 1c).

We determined the number of added (CAG)\(_n\) tracts for each Can\(^R\) clone (Fig. 1d). For pregrowth on both galactose and glucose, the median size of large-scale expansions corresponded to ~60 repeats.
Chromosomal fragility and expansions of (CAG)_{13} or (CAG)_{25} repeats^{16,21}. Chromosomal fragility and expansions of (CAG)_{70} runs are enhanced by loss of these recombination proteins^{40}. However, in mutant backgrounds in which (CAG)_{70} expansions are elevated, this increase is dependent on HR proteins^{8,40}. In contrast, large-scale expansions of GAA and ATTCT repeats are not affected by loss of Rad51 or Rad52. Notably, we found that the rate of large-scale CAG expansion was decreased 32-fold in the rad52Δ mutant and five-fold in the rad51Δ mutant compared with the wild-type strain (Fig. 2a), thus suggesting a role of HR in this process.

HR encompasses double-strand break (DSB) repair, synthesis-dependent strand annealing, and break-induced replication (BIR), all of which require end resection by the MRX (Mre11–Rad50–Xrs2) complex^{41}. Inactivation of Mre11 diminished large-scale expansions of CAG repeats ten-fold (Fig. 2a). DSB repair and BIR act on two-ended or one-ended DNA breaks whose formation may require resolution of a double or single Holliday junction, respectively. We tested single and double knockouts of the resolvase genes MUS81 and YEN1 and found that only the double mutant showed a decrease in expansion rate (Fig. 2b), thus indicating an overlapping role of these proteins in resolving Holliday junctions associated with the formation of large-scale CAG-repeat expansions.

To discriminate among the distinct pathways of HR, we examined the role of Pol32, a nonessential subunit of polymerase δ that is required for BIR^{42} but is less important in other branches of HR. In BIR, DNA synthesis occurs in the context of a displacement loop (D loop), potentially to the end of the chromosome. Notably, eliminating POL32 decreased the expansion rate 15-fold (Fig. 2a). After end resection and strand invasion during BIR, Pif1 helicase stimulates DNA synthesis. We observed a strong (ten-fold) decrease in the expansion rate in the pif1Δ mutant and a more modest decrease in the pif1-m2 mutant^{43}, in which Pif1 nuclear activity is eliminated while function in the mitochondria is maintained (Fig. 2a and Supplementary Fig. 4). The lesser effect of the pif1-m2 allele may be due to residual Pif1 activity in the nucleus, as has previously been reported for other BIR assays^{44,45}. Knockout of the REV3 gene, encoding DNA polymerase ζ, showed only a two-fold decrease in the CAG-expansion rate, thus indicating that the role of Pol32 and Pif1 in large-scale CAG expansion is not primarily dependent on translesion synthesis (Supplementary Fig. 4). Large-scale CAG-repeat expansions are replication-dependent events associated with replication-fork stalling

In eukaryotes, BIR has been most extensively characterized in budding yeast in the context of an irreparable one-ended DSB generated by HO endonuclease^{41,46} or chromosome fragmentation^{47}. Because these events were almost exclusively repaired in G2 phase, the D loop may be involved in DNA synthesis to the end of the chromosome. Though a remarkable feat, repair of the broken chromosome comes at a cost, given the high mutagenicity of BIR synthesis^{48}. BIR has also been proposed to repair one-ended DSBs resulting from replication-fork breakage. However, recent work investigating this subject has concluded that whereas repair of one-ended DSBs initially uses error-prone Pol32-dependent synthesis, its scope is limited, owing to the arrival of a converging replication fork followed by Mus81- or Yen1-dependent D-loop cleavage, referred to as BFR^{49}. Thus, the tradeoff between BIR and BFR pathways depends on the proximity or activity of a convergent replication origin.

To test whether large-scale CAG-repeat expansions occurred during S phase in our system, we treated cells with low doses of camptothecin, a topoisomerase I inhibitor, which triggers replication-fork breakage

**Figure 2** Genetic control of large-scale CAG-repeat expansions.
Effect of the indicated gene knockouts on the large-scale expansion rate (per cell per division) of (CAG)_{140} dregrown on galactose. (a, b) 60 μg/mL canavanine concentration (a) and 200 μg/mL canavanine concentration (b). WT, wild type. Expansion rates ± 95% confidence intervals are shown, as calculated on the basis of the distribution of expanded clones in at least nine independent cultures using the Ma–Sandri–Sarkar maximum-likelihood estimator with a correction for sampling and plating efficiency. The dashed line designates a three-fold decrease relative to the wild-type rate. Numbers above the dashed line show the fold decrease relative to the wild-type rate. Source data are available online.

Notably, the rate of large-scale expansions of the (CAG)_{140} run was considerably lower than the frequency of small-scale expansions previously determined for similarly sized CAG tracts^{17}. To determine how frequently small-scale expansions occurred in our experimental system, we grew our strain with (CAG)_{140} repeats under nonselective conditions and then analyzed the repeat lengths by single-colony PCR. We observed that, similarly to the previous data, relatively high frequencies of small-scale expansions were present in strains grown on either glucose or galactose (~1.0%), and even higher frequencies of contractions were present, particularly on galactose (17.5% versus 5.3% on glucose) (Supplementary Table 2).

**Genetic control of large-scale CAG-repeat expansions is distinct from that of small-scale expansion**

Using our selectable system, we used a candidate-gene approach to identify genes involved in large-scale CAG expansions. Srs2 is a DNA helicase that has been shown to unwind CAG repeats *in vitro*^{37}. In addition, eliminating Srs2 function results in an increased rate of small-scale expansions^{8,21}. We reasoned that if the large-scale expansions that we observed resulted from multiple small-scale events, Srs2 deletion would similarly show an increased expansion rate in our experimental system. However, we observed no effect of srs2 deletion on large-scale CAG expansions (Fig. 2a).

Mismatch repair (MMR) proteins have been shown to affect CAG expansions in several experimental systems^{38}. Msh2–Msh3 (MutSβ) typically binds to long insertion or deletion loops and promotes repair, but in the context of CAG repeats, Msh2–Msh3 promotes expansions. Analysis of MMR proteins in a yeast system for detecting small-scale expansions has indicated that msh3 deletion decreases, whereas msh6 deletion increases, the CAG-expansion rate^{39}. In contrast, we observed no differences in the rates of large-scale expansions in various MMR-deficient strains compared with the wild type (Fig. 2b).

**Genetic control of large-scale CAG-repeat expansions implicates genes required for homologous recombination and specifically the break-induced replication pathway**

The role of homologous recombination (HR) in CAG-repeat expansion has been studied in several yeast systems, which have yielded differing results that may reflect distinct aspects of CAG-repeat instability. Elimination of Rad51 and Rad52 proteins has no effect on the expansion rate of short tracts of (CAG)_{13} or (CAG)_{25} repeats^{16,21}.
in S phase, as well as hydroxyurea to increase replication-fork stalling and collapse. We found that both treatments increased (over threefold) the rate of large-scale CAG expansion (Fig. 3a).

Previous studies have convincingly demonstrated that long CAG repeats promote replication-fork stalling, breakage, and the formation of joint molecules consistent with recombination intermediates. To confirm whether replication-fork stalling and breakage occurred in our CAG system, we looked for accumulation of the γ-H2AX histone variant—a marker of both fork stalling and DSB repair—in our repetitive run during S phase. Using chromatin immunoprecipitation (ChIP) analysis, we indeed observed an enrichment in γ-H2AX at the (CAG)140 repeat, which peaked 40 min after release into S phase from α-factor arrest (Fig. 3b).

**DISCUSSION**

Our results clearly show that the genetic control of large-scale CAG expansions is different from the genetic control of small-scale CAG expansions. Most importantly, the roles of principal players in BIR—the Pol32 subunit of DNA polymerase δ and Pif1 helicase—in large-scale CAG expansions had not been previously observed for CAG or other expandable repeats, to our knowledge.

From these results, we propose a comprehensive model of large-scale CAG-repeat expansions. We believe that large-scale expansions of long CAG tracts are rooted in their ability to form stable hairpin structures during DNA synthesis and ultimately lead to replication-fork stalling (Fig. 4a,b). Such stalling, including that at various triplet repeats, has previously been shown to cause fork reversal (Fig. 4c). Isomerization of the resulting four-way junction (chicken-foot structure) leads to the formation of a Holliday junction (Fig. 4d), whose resolution by proteins such as Mus81 or Yen1 results in a one-ended DSB (Fig. 4e). Alternatively, endonucleases may act directly on the hairpin structure of the repetitive sequence (Fig. 4f). To reinitiate replication, the one-ended DSB invades the sister chromatid and creates a D loop (Fig. 4g). Because this one-ended DSB occurs within a long repetitive run, it would tend to invade its repetitive counterpart out of register. Indeed, out-of-register invasion has previously been proposed as a mechanism of CAG-repeat instability. However, these studies have used artificially induced two-ended DSBs or DSBs generated during meiotic recombination. Notably, in our model, hairpin formation on the ssDNA portion of the repeat tract would exacerbate this out-of-register invasion, thus potentially explaining the bias toward repeat expansions observed during intergenerational transmissions in human pedigrees. The convergent replication origin (ARS307) is ~30 kb away in our experimental system. Thus, only ~10 min would be required for the converging fork to reach the stall site. Consequently, BIR would progress over only a relatively short distance past the break (Fig. 4h). After its collision with the converging fork, a single Holliday junction would need to be resolved to separate the newly synthesized DNA molecule, which would have accumulated extra CAG repeats in the nascent DNA strand at a length equivalent to that of the out-of-register invasion step. Mus81, Yen1, or the two proteins together could potentially act at this step in addition to their earlier actions in Holliday-junction resolution of the isomerized four-way junction. Thus, the genetic control of large-scale expansions of CAG repeats has characteristics of both BIR and BFR.

Intriguingly, the proposed mechanism of large-scale CAG-repeat expansion is distinct from mechanisms described for small-scale expansions of short CAG tracts. In those cases, loss of RAD51 or RAD52 genes has little, if any, effect on the rate of (CAG)25 expansions. A probable explanation for this difference is that long CAG tracts are more susceptible to fork stalling and DNA breakage than shorter tracts. In further support of this reasoning, it has recently been found that expanded CAG repeats are more likely than unexpanded repeats to localize to the nuclear periphery during the S phase of the cell cycle. This observation may suggest the repair of one-ended DSBs, as we have proposed, at nuclear pores.

Additionally, our genetic analysis of large-scale expansions of a different trinucleotide repeat, (GAA)ₙ, was inconsistent with HR, BIR, and BFR pathways, but suggested template-switching during DNA replication as a mechanism for expansions. Although fork stalling at long GAA runs, similarly to stalling at CAG runs, results in DSB formation, we believe that the differences in expansion pathways may be due to the triplex-forming potential of the GAA repeat, which would hide ssDNA formed after end resection from the HR machinery.

Notably, the proposed mechanism has underlying similarities with the microhomology-mediated break-induced-replication mechanism.
pathway, which was originally proposed to explain copy-number variation and chromosomal rearrangements in humans and budding yeast. Our data add large-scale expansions of CAG repeats as another striking example of genome instability arising from BIR and BFR, thus highlighting the fundamental importance of these processes in DNA-damage repair, albeit at the expense of repeat instability and mutagenesis.

Could BIR account for the CAG-repeat expansions observed in human patients? Although it is too early to say, we believe that this possibility would provide an attractive model for large-scale expansions during cell divisions. In contrast, somatic instability has been reported to result from cumulative small mutations, thus making the intermediate to proceed directly to f. (c) Replication-fork reversal. (d) Four-way junction from c isomerizes to form a Holliday junction, which can be cleaved by resolvase proteins (gray arrows). (e) A one-ended DSB is subject to 5′-to-3′ resection (dotted line), thus resulting in a single-stranded repetitive sequence. (f) The single-stranded repetitive sequence is coated by replication protein A, then Rad51 (dark gray circles) and forms stable hairpin structure(s). (g) Out-of-register invasion results in formation of a D loop. A leftward-moving convergent replication fork shows the leading strand (green) and lagging strand (yellow) emanating from the adjacent replication origin. (h) Pol32-dependent DNA synthesis continues, thus further extending D-loop progression. (i) The D loop converges with the leftward-moving replication fork, thus resulting in a Holliday junction, which must be resolved (gray arrows). (j) The nascent DNA strand (bottom) would have accumulated extra CAG repeats at a length equivalent to that of the out-of-register invasion step, thus resulting in a large-scale expansion.

**METHODS**

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

**ACKNOWLEDGMENTS**

Research in the laboratory of S.M.M. is supported by NIH grants R01GM60987 and P01GM105473 and by a generous contribution from the White family. J.C.K. was supported by the NIH Training in Education and Critical Research Skills postdoctoral program (K12GM074869) and by Tufts University. We thank C. Freudenreich for valuable comments on the manuscript, G. Ira and S. Jinks-Robertson for helpful discussions, K. Spivakovsky-Gonzalez for help analyzing MSH mutants, and D. Padmanhaban for help with strain construction. S.T.H. received support from REU award NSF DBI 1263030. T.D. received support from Tufts Summer Scholars.

**AUTHOR CONTRIBUTIONS**

J.C.K. and S.M.M. designed the study; J.C.K., S.T.H., and T.D. performed experiments; J.C.K., S.T.H., T.D., and S.M.M. analyzed data; K.A.S. provided reagents; J.C.K. and S.M.M. wrote the manuscript.
COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. La Spada, A.R. & Taylor, J.P. Repeat expansion disease: progress and puzzles in disease pathogenesis. Nat. Rev. Genet. 11, 247–258 (2010).

2. McMurray, C.T. Mechanisms of trinucleotide repeat instability during human development. Nat. Rev. Genet. 11, 786–799 (2010).

3. Nance, M.A. & Myers, R.H. Juvenile onset Huntington’s disease: clinical and research perspectives. Ment. Retard. Dev. Disabil. Res. Rev. 7, 153–157 (2001).

4. Thornton, C.A. Myotonic dystrophy. Neurol. Clin. 32, 705–719 (2014).

5. Kim, J.C. & Mirkin, S.M. The balancing act of DNA repeat expansions. Trends Biochem. Sci. 37, 162–172 (2012).

6. Liu, Y. & Wilson, S.H. DNA base excision repair: a mechanism of trinucleotide repeat instability. Trends Biochem. Sci. 37, 42–49 (2012).

7. Gacy, A.M., Goellner, G., Juranic, N., Macura, S. & McMurray, C.T. Trinucleotide repeats by restraining recombination. Mol. Cell. Biol. 26, 351–358 (2006).

8. Møllersen, L., Rowe, A.D., Larsen, E., Rognes, T. & Klungland, A. Continuous and periodic expansion of CAG repeats in Huntington’s disease R6/1 mice. Proc. Natl. Acad. Sci. USA 108, 2843–2848 (2011).

9. Shah, K.A., McGirr, R.J., Egorova, V.I. & Mirkin, S.M. Coupling transcriptional state to large-scale repeat expansions in yeast. Cell Rep. 9, 1594–1602 (2014).

10. Mirkin, S.M. DNA structures, repeat expansions and human hereditary disorders. Nature 475, 581–587 (2013).

11. Samadashwily, G.M., Raca, G. & Mirkin, S.M. Trinucleotide repeats affect DNA repair. Insights from model systems. Crit. Rev. Biochem. Mol. Biol. 50, 142–167 (2015).

12. Gac, A.M., Goelinger, G., Juranic, N., Macura, S. & McMurray, C.T. Trinucleotide repeats that expand in human disease form hairpin structures in vitro. Cell 81, 533–543 (2000).
**ONLINE METHODS**

**Plasmids.** CAG repeats were obtained from pGL2-CTG40 (ref. 60). The repeats were excised with AvrII and SfoI and initially cloned into a pYES3-TET644 derivative (described as GAA) (ref. 61), which had been cut with BsaBI and an engineered AvrII site. This plasmid was used as template for generating a PCR product (primers JK109 and JK110) containing CAG repeats with Ncol and NotI handles. The PCR product was cloned into pYes3-G4G1C1-T150-GAA100 (ref. 33), which had been cut with Ncol and NotI to remove the GAA repeats. Because this construct did not allow selection of large-scale CAG-repeat expansions, the length of the ‘space’ was increased with PCR products with Ncol handles generated from the bacterial tetracycline-resistance gene (forward JK161 with reverse JK162, JK163, or JK164). The plasmid used in the present study for analyzing large-scale expansion of (CAG)140 is pYes3-G4G1C1-Fori-CAG140-tetbal1-rev (799 bp). For analyzing (CAG)93, a plasmid containing contracted CAG repeats and a longer fragment from the tetracycline-resistance gene was used, pYes3-G4G1C1-Fori-CAG93-tetbal2-rev (708 bp). The repeat’s integrity in these constructs was verified by plasmid sequencing with primers FlankL and CanF.

A no-repeat control plasmid was constructed by cloning a PCR fragment from the tetracycline-resistance gene (primers JK161 and JK165) into the Ncol and SphI sites of pYes3-G4G1C1-T150-GAA100, which is called pYes3-G4G1C1-Fori_tet340. All bacterial cloning steps were carried out in the E. coli strain (coli2) SURE2 strain (Agilent). Primers are listed in **Supplementary Table 3**.

**Yeast strains.** All strains in this study are isogenic to *S. cerevisiae* wild-type (WT) strain CH1585 (MATα, leu2-3,112, trp1-Δ63, ura3-52, his3-200), an S288c-related haploid strain used in our previous studies. For transformation into the CH1585 strain: KanMX strain 35, the construct was digested from the plasmid with SwaI. Transformants were selected on synthetic complete medium lacking tryptophan. The cassette was positioned ~1 kb downstream of CAN1::KanMX. Transformants were selected on synthetic complete medium lacking tryptophan.

**Transformants** were selected on synthetic complete medium lacking tryptophan. Nonselective growth of expanded clones by PCR or the number of CanR clones multiplied by the total efficiency determined as \((z – 1)/(z \times \ln(z))\), where \(z\) is the fraction of the culture (described as (GAA)0 (ref. 61), which had been cut with BsaBI and engineered AvrII site. This plasmid was used as template for generating a PCR product (primers JK161 and JK165) into the Ncol and SphI sites of pYes3-G4G1C1-T150-GAA100, which is called pYes3-G4G1C1-Fori_tet340. All bacterial cloning steps were carried out in the *E. coli* strain (coli2) SURE2 strain (Agilent). Primers are listed in **Supplementary Table 3**.

**Fluctuation assay and calculation of rates.** Colonies were suspended in 0.2–1 mL of sterile water, serially diluted, and plated Fluctuation assay and calculation of rates. The standard concentration of canavanine used was 60 µM as on YPD medium for determination of total cell number. Nonselective growth of expanded clones by PCR or the number of CanR clones multiplied by the total efficiency determined as \((z – 1)/(z \times \ln(z))\), where \(z\) is the fraction of the culture (described as (GAA)0 (ref. 61), which had been cut with BsaBI and engineered AvrII site. This plasmid was used as template for generating a PCR product (primers JK161 and JK165) into the Ncol and SphI sites of pYes3-G4G1C1-T150-GAA100, which is called pYes3-G4G1C1-Fori_tet340. All bacterial cloning steps were carried out in the *E. coli* strain (coli2) SURE2 strain (Agilent). Primers are listed in **Supplementary Table 3**.

**Antibody-bound DNA was immunoprecipitated (IP) with Protein G Dynabeads (Invitrogen).** Samples were analyzed by quantitative PCR with QuantiStudio 6 (Applied Biosystems). Relative quantities of IP and input DNA were determined with a standard curve for primers at the CAG locus as well as the nonenriched control ACT1. IP/input values for the CAG locus were normalized to ACT1. The graph shows the fold enrichment of the CAG strain compared with the no-repeat control strain.

**Chromatin immunoprecipitation.** Strain YJK146 containing (CAG)140 and the no-repeat control YJK154 were analyzed by ChIP with antibodies recognizing phosphorylated histone H2A Ser129 (ab15083, Abcam; validation on manufacturer’s website and ref. 50). This modification has been referred to as γ H2AX in yeast60. Yeast were grown to an OD of ~0.8 in YPGal, arrested in G1 with α-factor, then washed twice with water and released into YPGal. 50-µL cultures of 0-, 20-, 40-, 60-, and 90-min time points were cross-linked with 1% formaldehyde for 15 min at room temperature and quenched with glycine. Cells were lysed mechanically with 0.5-mm glass beads. The chromatin-containing cell suspension was sonicated, thus yielding sheared DNA in the range of approximately 100–1,500 bp. Antibody-bound DNA was immunoprecipitated (IP) with Protein G Dynabeads (Invitrogen). Samples were analyzed by quantitative PCR with QuantiStudio 6 (Applied Biosystems). Relative quantities of IP and input DNA were determined with a standard curve for primers at the CAG locus as well as the nonenriched control ACT1. IP/input values for the CAG locus were normalized to ACT1.

**Statistical methods.** Rates of expansion and 95% confidence intervals were calculated on the basis of the distribution of expanded clones in at least nine independent cultures (number depends on clones found to have the correct starting repeat length) with the Ma–Sandri–Sarkar maximum-likelihood estimator, with a correction for sampling and plating efficiency. Median and interquartile values are reported in Figure 1d. ChIP time points were analyzed by one-way ANOVA.

**Data availability.** Source data for Figures 1–3 are available with the paper online. All other data is available upon request.

60. Raca, G., Siyanova, E.Y., McMurray, C.T. & Mirkin, S.M. Expansion of the (CTG)n repeat in the 5′-UTR of a reporter gene impedes translation. *Nucleic Acids Res.* 28, 3943–3949 (2000).
61. Shah, K.A. et al. Role of DNA polymerases in repeat-mediated genome instability. *Cell Rep.* 2, 1088–1095 (2012).
62. Goldstein, A.L. & McCusker, J.H. Three new dominant drug resistance cassettes for the haploid strain used in our previous studies. For transformation into the CH1585 strain: KanMX strain 35, the construct was digested from the plasmid with SwaI. Transformants were selected on synthetic complete medium lacking tryptophan. The cassette was positioned ~1 kb downstream of CAN1::KanMX. Transformants were selected on synthetic complete medium lacking tryptophan.

Transformants were selected on synthetic complete medium lacking tryptophan. Nonselective growth of expanded clones by PCR or the number of CanR clones multiplied by the total efficiency determined as \((z – 1)/(z \times \ln(z))\), where \(z\) is the fraction of the culture (described as (GAA)0 (ref. 61), which had been cut with BsaBI and engineered AvrII site. This plasmid was used as template for generating a PCR product (primers JK161 and JK165) into the Ncol and SphI sites of pYes3-G4G1C1-T150-GAA100, which is called pYes3-G4G1C1-Fori_tet340. All bacterial cloning steps were carried out in the *E. coli* strain (coli2) SURE2 strain (Agilent). Primers are listed in **Supplementary Table 3**.

**Antibody-bound DNA was immunoprecipitated (IP) with Protein G Dynabeads (Invitrogen).** Samples were analyzed by quantitative PCR with QuantiStudio 6 (Applied Biosystems). Relative quantities of IP and input DNA were determined with a standard curve for primers at the CAG locus as well as the nonenriched control ACT1. IP/input values for the CAG locus were normalized to ACT1. The graph shows the fold enrichment of the CAG strain compared with the no-repeat control strain.

**Statistical methods.** Rates of expansion and 95% confidence intervals were calculated on the basis of the distribution of expanded clones in at least nine independent cultures (number depends on clones found to have the correct starting repeat length) with the Ma–Sandri–Sarkar maximum-likelihood estimator, with a correction for sampling and plating efficiency. Median and interquartile values are reported in Figure 1d. ChIP time points were analyzed by one-way ANOVA.

**Data availability.** Source data for Figures 1–3 are available with the paper online. All other data is available upon request.