Next-generation DNA sequencing technologies and the study of trinucleotide repeat instability

A dissertation
submitted by

Ryan J. McGinty

In partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in

Biology

TUFTS UNIVERSITY

February 2018

Advisor: Dr. Sergei M. Mirkin

Abstract

Microsatellite repeats are the source of multiple hereditary diseases in humans. Certain repetitive sequences form unusual DNA structures that differ from standard B-form DNA. As the length of a repeat increases, structures form more frequently and stably. Secondary structures can interfere with numerous cellular processes, leading to expansions and contractions of the repetitive tract, double-strand breaks (DSBs) and complex genomic rearrangements (CGRs).

DNA sequencing technologies have developed in leaps and bounds in the previous decade. Large-scale DNA sequencing is now fast and inexpensive. More recently, it has become possible to sequence long stretches of DNA in a single, contiguous read. This has exciting possibilities for the study of microsatellites and CGRs, both of which are difficult to resolve using short-read sequencing technologies.

I present here two innovative applications of DNA sequencing technologies used to uncover new mechanisms of microsatellite instability. In Chapter 2, short-read DNA sequencing is central to a novel screening method to identify genes involved in $(GAA)_n$ repeat expansion using yeast as a model system. This led to identification of mutants of the polyadenylation gene *YSH1,* later found to affect transcription-replication collisions that cause DSBs. Chapter 3 also concerns the role of transcription and the effect of nucleosome positioning on $(GAA)_n$ repeat expansions. Chapter 1 consists of a review highlighting the

importance of using model systems to identify characterize modifiers of microsatellites, revealing molecular mechanisms and potentially informing human health.

In Chapter 4, long-read Nanopore sequencing was used to uncover mechanisms leading to CGRs. This approach demonstrates that DNA rearrangements can be captured within individual sequencing reads, revealing the mechanisms by which DNA breakage and repair occur, including the invasion of $(GAA)_n$ repeats into other areas of the genome, recombination between repetitive transposable elements, and the use of break-induced replication. This approach can be extended not only to the further study of microsatellite instability and DNA repair, but also to cancer, where CGRs are frequent.

Acknowledgements

This work would not be possible without mentorship from Sergei Mirkin and Anna Aksenova, support from all members of the Mirkin lab (Jane Kim, Kartik Shah, Alex Shishkin, Tom Ebersole, Elina Radchenko, Alex Neil, Artem Kononenko, Jorge Cebrian-Castillo, Katia Spivakovsky-Gonzalez, Sasha Khristich), the Freudenreich lab (Catherine Freudenreich, Simran Kaushal, Jennifer Nguyen, Allen Su, Nealia House, Lionel Gellon) and additional faculty mentors and committee members (Mitch McVey, Susan Ernst, Shamil Sunyaev). Extensive experimental contributions were made by the Moore lab (Claire Moore, Franco Puleo, Erika Pearson), the Petes lab (Tom Petes, Denis Kiktev, Margaret Dominska) the Housman lab (David Housman, Eric Wang) and two Tufts undergraduates, Julia Hisey and Rachel Rubinstein. Support also came from the Biology Department and staff (including Mike Grossi, Sanjukta Ghosh, Michael Doire, and Deirdre McCann). Friends and family know who they are and what they did.

Table of Contents

Chapter 1:
Cis- and *trans*-modifiers of repeat expansions: Blending model systems with human genetics

Chapter 2:
A defective mRNA cleavage and polyadenylation complex facilitates expansions of transcribed $(GAA)_n$ repeats associated with Friedreich's ataxia

Chapter 3:
Coupling transcriptional state to large-scale repeat expansions in yeast

Chapter 4:
Nanopore sequencing of complex genomic rearrangements in yeast reveals mechanisms of repeat-mediated double-strand break repair

Chapter 1

Cis- and *trans*-modifiers of repeat expansions: Blending model systems with human genetics

Ryan J. McGinty and Sergei M. Mirkin

Department of Biology, Tufts University, Medford MA 02155, U.S.A.

[*] The author to whom all correspondence should be addressed.

E-mail: sergei.mirkin@tufts.edu

Trends in Genetics (under submission)

Abstract

Over thirty hereditary diseases are caused by the expansion of microsatellite repeats. The length of the expandable repeat is the main hereditary determinant of these disorders. They are also affected by numerous genomic variants that are nearby (*cis*) and physically separated from (*trans*) the repetitive locus, which we review here. These genetic variants have largely been elucidated in model systems using gene knockouts, while a few have been directly observed as single nucleotide polymorphisms (SNPs) in patients. There is a notable disconnect between these two bodies of knowledge: knockouts poorly approximate the SNP-level variation in human populations that gives rise to medically-relevant *cis*- and *trans*-modifiers, while the rarity of these diseases limits the statistical power of SNP-based analysis in humans. We propose that high-throughput SNP-based screening in model systems could become a useful approach to quickly identify and characterize modifiers with medical relevance for patients.

Introduction

Microsatellites consist of tandem repeated units of 1-to-9 DNA base pairs that can extend from a few repeats to thousands. Trinucleotide repeats in particular are linked to a number of human genetic disorders [1–3], including $(CAG)_n$ repeats in Huntington's disease (HD) and various spinocerebellar ataxias, $(CTG)_n$ repeats in myotonic dystrophy type 1 (DM1), $(CGG)_n$ repeats in fragile-X syndrome (FXS), $(GAA)_n$ repeats in Friedrich's ataxia (FRDA) and many others. Repetitive sequences are subject to expansions and contractions, a unique class of mutations arising from a variety of distinct mechanisms. In each case, disease occurs in individuals who have inherited a repeat tract that has expanded beyond a certain length. Nearly all microsatellite expansion diseases are neurological or neurodegenerative, with progressive symptoms coinciding with continued somatic expansion throughout life [4–6]. Understanding the nature of repeat expansion should therefore help to explain how individuals inherit and develop microsatellite expansion diseases.

For a subset of micro- and minisatellites (slightly longer 10-100 bp repeat units), the repetition of complementary base pairs leads to stable intra-strand base-pairing, resulting in non-B-form DNA secondary structures **(Fig. 1)**. For $(AT)_n$ dinucleotide repeats, A-T base pairs on each strand can nucleate into stable hairpin (one strand) or cruciform (both strands) structures upon unwinding [7,8]. $(CAG/CTG)_n$ repeats can form imperfect hairpins, with the strong C-G base pairs stabilizing the structure [9]. Hairpins also contain short unpaired caps at the point of symmetry, due to the limited bending angle of DNA strands. In the case of $(GAA)_n$ repeats, one strand consists entirely of purines, while the other contains only pyrimidines. This creates conditions favorable for the formation of triple-helical H-DNA, in which the third strand is bound to the duplex via Hoogsteen (H-y) or reverse Hoogsteen (H-r) base-pairing, rather than Watson-Crick base-pairing [10,11]. $(CGG)_n$ repeats and human telomeric $(TTAGGG)_n$ repeats contains regularly phased guanines in one strand, promoting the formation of G-quadruplex DNA (although the importance of G-quadruplex formation in $(CGG)_n$ repeats remains a subject of debate) [12–14]. All microsatellite repeats also have the potential to form slipped-strand DNA, where the DNA has unwound and reannealed out of register, leaving an unpaired or hairpin-stabilized loop on each strand [15,16]. In most cases, these structures form transiently during processes involving DNA strand separation, such as replication, repair, recombination and transcription. Longer repetitive tracts and high levels of DNA negative supercoiling lead to more energetically stable non-B DNAstructures [17–19].These dynamic structure-forming properties of repetitive DNA can be linked to many of their functional and detrimental consequences.

Repeats may be added or lost a few at a time, or in large jumps [20]. The rate of this instability increases exponentially with repeat length, in accordance with structure-forming potential, with long repeat tracts expanding and contracting at rates that are orders-of-magnitude higher than the rate of point mutations [21–24]. Large repetitive tracts are also frequent sources of double-strand breaks. In particular, fragile-X syndrome is named for the tendency of expanded $(CGG)_n$ repeats to break, and long $(CAG)_n$ and $(GAA)_n$ tracts also serve as fragile sites [25–27]. In addition, $(AT)_n$-rich repeats appear to play a role in the fragility of common fragile sites, and are frequently associated with translocation breakpoints in cancer [28–30]. Repeat-containing broken DNA ends can promote homologous recombination (HR) into non-allelic genomic loci that also contain repeats, illustrated by the frequency of microsatellites appearing at the sites of complex genomic rearrangements (CGRs) in cancer, as well by repeat-mediated CGRs in model organisms [26,30–34].

In this review, we discuss the complex progression of human microsatellite diseases through the lens of *cis*- and *trans*-acting modifiers of microsatellite instability, as well as strategies for identifying new modifiers. Because microsatellite instability is intrinsically linked to basic properties of DNA, the genetic modifiers of instability include core molecular machineries that have been conserved throughout evolution. The molecular mechanisms of microsatellite instability are largely understood due to efforts in model systems, including the eukaryotic baker's yeast, mice and cultured human cells [2,20,35–37]. Studies in human patients have uncovered a few modifiers of somatic instability that align with results from model systems. Due to the rarity of each microsatellite disease, the latter population-based studies are limited in the power to detect all but the most common and powerful modifiers. Sequencing of patient genomes could potentially reveal rare mutations in the same genes implicated in model systems, suggesting which individuals might be most subject to high rates of somatic expansion. However, human genetic variation within genes rarely resembles the most common tool used to investigate gene function in simple model systems, namely gene knockouts. While extensive structural variation exists in humans, the vast majority of large-scale deletions and insertions appears within introns and intergenic regions [38]. Variation in protein-coding regions is more likely to take the form of single nucleotide variants / polymorphisms (SNVs, SNPs) and short, non-frame-shift indels, as these less-severe variants have been tolerated through evolution. Unfortunately, it is not always straightforward to predict the effect of a SNV on a gene's function, even when the gene knockout is well characterized. Furthermore, many gene variants may only become important in the context of a particular genetic background. We discuss here a particular strategy to aid in translating patient genome sequences into actionable medical information, namely the use of model systems for the large-scale identification of SNVs affecting conserved genes involved in microsatellite instability

***Cis*-modifiers of repeat expansions**

Cis-acting modifiers of repeat expansion are those genomic variants that are found within or in the immediate vicinity of the repetitive locus. The first discovered cis-modifiers of repeat expansions were interruptions within the repeats themselves, such as $(AGG)_n$ triplets within FXS $(CGG)_n$ runs [9], $(CAT)_n$ triplets within the SCA1 $(CAG)_n$ runs [39], or $(GGA)_n$ triplets in the FRDA $(GAA)_n$ repeats [40]. These mutations disrupt the stability of secondary structures, leading to drastic reductions in expansions [41,42]. Numerous data from model systems show that microsatellite instability also depends on the orientation relative to replication origins [43–48]. This led to the "ori-switch" model, in which expansions arise frequently in the presence of a genetic or epigenetic *cis*-modifier that results in a switch in the direction of replication through the repeat [49] **(Fig. 2)**. Recently, this hypothesis was confirmed by the discovery of a *cis*-modifier for FXS. SNP genotyping of the region surrounding *FMR1* revealed a single SNP ~50 kb upstream of the $(CGG)_n$ repeats that was present in nearly all individuals with an expanded repeat, but was present in only half of normal-length individuals [50]. Furthermore, using a powerful technique known as single molecule analysis of replicating DNA (SMARD), it was shown that this SNP is correlated with the activity of an underlying replication origin [51,52]. Normally, the *FMR1* locus is replicated from both directions, originating from the aforementioned upstream site and from another downstream origin. In the presence of this SNP, however, the upstream origin gets inactivated, and so replication proceeds through *FMR1* in only one direction. This change places the $(CGG)_n$ repeats exclusively on the leading strand template and the nascent lagging strand of the replication fork. Formation of a secondary structure by the $(CGG)_n$ run might result in nascent strand slippage and realignment out of register with the template strand, leading to repeat expansions **(Fig. 3)**. Note that while both $(CCG)_n$ and $(CGG)_n$ strands of the fragile X repeat can form hairpins, the stability of the two structures differ due to the C-C vs. G-G mismatches. Additionally, only the $(CGG)_n$ strand is potentially capable of forming G-quadruplex structures, though it remains unclear under what conditions this structure may be stable *in vivo* [13].

Similar phenomena have been observed for other repeats. $(CTG)_n$ repeats in DM1 are flanked by CTCF insulator sites, which differ in methylation status between different tissue types, and may affect replication direction [53]. $(GAA)_n$ repeats in cells derived from FRDA patients are replicated in a single orientation, whereas the same locus in cells from healthy individuals replicates from both directions [54]. In the orientation which places $(GAA)_n$ repeats on the leading strand template, replication fork progression was stalled more severely. Treatment of FRDA cells with a polyamide compound, which had been shown to prevent triplex formation and reduce expansions [55], here rescued replication fork stalling at the $(GAA)_n$ repeats as well. This implies that triplex formation leads to fork stalling, a process linked to expansions in model systems (see below).

While it is not known whether a genetic or epigenetic change is responsible for the observed change in replication direction in FRDA cells, these data together are indicative that “ori-switch” may be a common theme for repeat expansions in humans. Why does replication direction matter? Possible explanations include differences in the activity of the leading and lagging strand polymerases, Pol ε and Pol δ, respectively, and their differing responses to replication stress [56]. Okazaki fragment maturation on the lagging strand is another step that is vulnerable to mistakes within microsatellites (see below). In addition to changing replication direction, an “ori-switch” may also alter replication timing and/or position the repeat tract farther away from the next-closest origin. This may impact the stability of the replication fork when it reaches the repeats, or may involve changes to the chromatin environment [46,47,53].

***Trans*-modifiers of repeat expansions identified in model systems**

Trans-modifiers of microsatellite instability may occur in any part of the genome. Most *trans*-modifiers have been uncovered in genetically tractable model systems including yeast, mice and cultured human cells [2,20,35–37]. *A priori*, these modifiers can contribute to one or more of the following broadly characterized mechanisms: secondary structure formation, processes inhibited by secondary structures, recognition and processing of secondary structures, and processes that are invoked in response to the previous categories. The misalignment of repeats represents an additional route to instability, because the repetitive sequence itself poses a problem apart from secondary-structure formation.

DNA replication is vulnerable at multiple points to each of these categories of instability mechanisms. Nearly every replication protein has been implicated in triplet repeat instability. The single-stranded binding protein complex RPA appears important for preventing expansions. This strong protection may encompass multiple mechanisms, including binding to unpaired regions within secondary structures or extensive single-stranded regions which can be produced following uncoupling of the replisome [56,57]. The latter mechanism would explain the increase in expansions observed upon knocking down the replicative DNA helicase gene *MCM4* [57]. Accessory DNA helicases such as the yeast Srs2 and Sgs1, and the human BLM, WRN and RTEL1, were also found to stabilize expandable repeats, likely due to their secondary-structure-unwinding capabilities [23,35,58–61]. This activity may be coordinated via PCNA [60]. Mutations in the replicative DNA polymerases delta and epsilon slow replication, which exacerbates replication blockage by secondary structures and leads to an increased rate of instability, in some cases via translesion synthesis [62]. Expansions increase in the absence the replication fork stabilizer Tof1, or Timeless in yeast and humans, respectively [23,24,63,64].

Restarting replication can proceed by several mechanisms. One of these is template switch, in which replication temporarily switches to use the sister chromatid as a template, then switches back to bypass the secondary structure **(Fig. 4)**. Work in yeast has shown that template switch uses DNA polymerase alpha to synthesize an Okazaki fragment-length segment, and that modifiers in DNA polymerase alpha can result in a larger step size of repeat expansions [62]. Deletion of yeast Rad5, which promotes template switching, results in fewer large-scale $(GAA)_n$ repeat expansions [23]. However, the opposite is true for expansions of short $(CAG)_n$ tracts, where Rad5 acts in a separate pathway [20,65,66]. The flap endonuclease Fen1/Rad27 also appears to protect against large-scale $(GAA)_n$ expansions via template switching, in addition to its role in the strand-slippage that leads to small-scale repeat expansions [67–69]. This may occur even in the absence of fork stalling during post-replicative repair, as Fen1 cleaves repeat-containing flaps left by the lagging strand synthesis. These flaps may otherwise fold back into a triplex and require bypass by template-switch [37].

Replication can also restart by fork reversal, which can generate one-ended DSBs, the repair of which can proceed via break-induced replication (BIR). BIR uses HR to restart replication from the sister chromatid, and conservative DNA synthesis proceeds until interrupted, potentially only when reaching the end of the chromosome. [37,70,71]. Large-scale $(CAG)_n$ repeat expansions in yeast have been shown to require the Pol32 subunit of DNA polymerase delta, as well as Pif1 helicase, two proteins central to BIR [71]. BIR was also recently implicated in expansions of $(CGG)_n$ repeats in cultured mammalian cells [72]. If BIR initiates within a microsatellite tract, repeat instability is possible **(Fig. 5)**. Surprisingly, nuclear pore components have been found to affect $(CAG)_n$ instability in yeast, through a mechanism involving relocalization of stalled replication forks to the pore for efficient fork restart [73]. This relocalization may be important for restricting the homology search during BIR initiation, helping to prevent translocations [74–76].

Nearly every form of DNA repair, in addition to post-replicative repair and BIR (see above), has been implicated in microsatellite instability. Secondary structures, which contain regions of single-strandedness, may be particularly vulnerable to DNA damage, including cytosine deamination and oxidative damage. This damage is then repaired by base excision repair (BER), wherein the damaged base is removed to create an abasic site that is then cleaved by AP endonuclease, leading to a single-strand nick. The nick can be processed into a single strand gap, and the resulting fill-in synthesis is vulnerable to strand slippage, leading to small-scale instability, as well as further secondary structure formation **(Fig. 6)**. This process has been shown to involve Fcy1 and Ung1 in yeast, as well as OGG1 and NEIL1 in mouse models [77–82]. Mismatch repair (MMR) has also been found to promote expansions in numerous systems [82–85], as well as in human genetic studies (see below). It is thought that MMR components mistakenly recognize hairpins as mismatches, either at the capped ends or at actual mismatches within imperfect hairpins, and either stabilize the secondary structure or unnecessarily initiate repair [86–89]. As with BER, the resulting repair generates a nick, which can lead to strand slippage during fill-in synthesis **(Fig. 6)**. Inappropriate MMR also appears to act on $(GAA)_n$ repeats in yeast, perhaps at the single-stranded portion of a triplex or at the unpaired regions of slipped-strand structures [26]. Single-strand nicks occurring nearby on opposite strands ultimately lead to DSBs. In such cases, repair can occur by non-homologous end joining (NHEJ), resulting in a contraction or larger deletion, or by one or more branches of HR **(Fig. 6)**. Importantly, DNA repair can occur during all cell cycle stages. Thus, these modifiers may be particularly important in diseases originating from non-dividing cells.

Homologous recombination genes implicated in repeat instability include recombinases Rad52 and Rad51, the end-resection complex MRX/Sae2, as well as Mus81/Yen1 resolvases [31,57,71,90–93].The process of HR is intrinsically vulnerable to destabilizing repetitive DNA. Realignment of broken repeat ends can occur between any two segments of the repeat, assuring that a length change will frequently occur. This can potentially happen during BIR, as mentioned above, as well as during single-strand annealing (SSA) or synthesis-dependent strand annealing (SDSA) [94–99]. That being said, sister chromatid exchange may have a stabilizing role in repeat maintenance [92,100,101]. Broken ends within repeats may be resected beyond the repetitive tract, allowing the homology search to take advantage of non-repetitive sequence.

Transcription: *cis*- and *trans*-modifiers of repeat expansions

Genes involved in various stages of the transcription process have also been shown to affect microsatellite instability, and may be very important in accounting for expansions that occur in non-dividing cells [35,36,102,103]. This could be due to several mechanistic reasons. Transcription necessarily involves accessing single-stranded DNA to generate the complementary RNA. DNA unwinding generates negative supercoils behind the RNA polymerase, creating favorable conditions for secondary structure formation [17–19]. Nucleosome remodeling and/or removal also accompanies active transcription, leading to secondary structure formation and thus instability [104]. Secondary structure can then inhibit the movement of RNA polymerase in further rounds of transcription [105,106]. These mechanisms likely explain the involvement of chromatin modifiers and transcription initiation factors in repeat expansion observed in yeast, flies and human cells [57,101,107–109]. This also suggests that mutations within the promoter or enhancer of a repeat-containing gene might serve as *cis*-modifiers by increasing or decreasing transcription levels. *Trans*-modifiers of genome stability affecting other stages of transcription have also been uncovered, namely a polyadenylation and 3’-end processing factor [93] (see below), as well as multiple mRNA packaging and export factors in yeast and human cells [110–116]. The former appears to lead to instability by preventing transcript cleavage and detachment of RNA polymerase, thus promoting collisions between RNA and DNA polymerases that can lead to DSBs. The latter appears to lead to increased R-loop formation (extended RNA/DNA hybrids). R-loops may lead to instability by stalling RNA polymerase while also leaving the non-template strand unpaired. This can allow access by DNA damaging agents, leading to BER **(Fig. 6)**, or transcription-coupled nucleotide excision repair (NER), which has separately been implicated in instability [82,109,117–120]. Secondary structure formation on the unpaired strand may further stabilize the R-loop [117]. $(GAA)_n$ repeats may be particularly susceptible to R-loop formation, as purine-pyrimidine DNA-RNA bonds are stronger than purine-pyrimidine DNA-DNA bonds [119].

Transcription in FRDA and FXS: consequences and considerations

Considering the above ways in which transcription of unstable microsatellites is a risky proposition, it may ultimately be beneficial for a cell to inhibit transcription of repeat-containing genes. In fact, *FMR1* and *FXN* genes in FXS and FRDA patients, respectively, typically show chromatin and/or DNA methylation changes that reduce transcription, and the resulting insufficient protein levels lead to disease phenotypes. While this is disastrous for the individual, from the perspective of the cell, this may be the best option for maintaining the stability of the repetitive tract. The alternative is to risk repeat expansion, exponentially increasing the risk of subsequent instability. Note that expansions are not the most detrimental possible outcome when cellular machinery encounters long microsatellite tracts. Chromosomal fragility has been observed for nearly all structure-prone repeats that reach a certain length [29,35]. Interestingly, significant overlap has been observed between genes affecting fragility and instability of repeats [57,121], and repair of DSBs can lead to further repeat expansions (see above). Unrepaired DSBs can lead to the loss of entire chromosome arms and numerous essential genes, likely leading to cell death. Misrepaired DSBs can lead to deletions of various sizes, as well as chromosomal translocations and copy number changes, all potential drivers of cancer. Triplex-forming repeats as well as AT-rich repeats have been found to be prevalent at the breakpoints of translocation in cancer [30]. As mentioned above, DSB repair involving repeats is highly prone to errors, which can also include non-allelic recombination due to the multitude of microsatellites throughout the genome [26,33,34,98,121]. BIR is also prone to generating frequent point mutations, which may account for the phenomenon of repeat-induced mutagenesis (RIM) [98,122–124].

Thus, a somewhat paradoxical situation arises, in that repeat-containing genes must be re-activated to prevent disease symptoms, but higher transcription of the repeats may increase the likelihood of further genomic instability.

Treatments designed to reactivate *FMR1* or *FXN*, such as histone deacetylase (HDAC) inhibitors, appear not to have been successful in clinical trials, however more approaches along this line are being developed [125,126]. One might speculate that this could be due to the side effect of promoting further somatic expansions. In this light, two other proposed therapies for FRDA may have greater success. The first involves using $(UUC)_n$ synthetic oligonucleotides to bind $(GAA)_n$ transcripts, thus preventing R-loop formation [127]. This has the intended effect of promoting efficient FXN transcription, but may also help to prevent further R-loop-dependent somatic expansions. A second therapy involves a synthetic peptide targeted to the mitochondria that promotes efficient translation of FXN transcripts [125]. This has the benefit of increasing FXN protein expression without increasing risky FXN transcription. However, the lasting effectiveness of this therapy may still be limited by ongoing somatic expansions. Perhaps this therapy will prove effective in combination with treatments designed to disrupt secondary structures[128], slow somatic expansions [80] or promote somatic contractions [129,130].

Complex *cis*- and *trans*-modifier interactions

Although it may appear that "everything but the kitchen sink" in DNA-related processes affects microsatellite instability, the difficulty in translating this knowledge to a disease treatment is that each of these factors interacts in a complex network, which can change drastically based on the presence of particular modifiers and the particular microsatellite sequence in question. A few examples of this have been worked out in yeast: The Rad5 helicase/ubiquitin ligase protects against small-scale $(CAG)_n$ expansions during replication slippage, but promotes $(GAA)_n$ expansions via template switching at a stalled replication fork [20]. Another example is the role of HR factor Rad52. In studies of very short $(CAG)_n$ repeat tracts, knocking out Rad52 had little to no effect [20]. However, Rad52 has been shown to have a role in instability of longer $(CAG)_n$ repeat tracts, particularly in conditions of elevated DNA fragility [92,131]. Large-scale $(CAG)_n$ expansions in yeast were also found to occur through a Rad52- and Pol32-dependent BIR mechanism, rather than by replication slippage [71]. For $(GAA)_n$ repeats in yeast, knockout of Rad52 normally does not greatly affect the rate of expansions, which occur by template switch during replication [23]. However, deletion of Rad52 reduces expansions when in the presence of a mutation in the RNA 3' end-processing gene *YSH1.* In this genetic background, $(GAA)_n$ expansions are frequently generated via a Rad52-dependent response to DSBs [93]. Overall, a pattern emerges where the role of HR components in microsatellite instability becomes evident in situations where DNA breaks occur frequently, which may itself be dependent upon any number of *cis*- and *trans*-modifiers [121]. In humans, the picture is further complicated by the relative prominence of HR compared with non-homologous end-joining in particular cell types. Similarly, it has been shown in mouse models that tissue-specific differences in $(CAG)_n$ instability can be linked to differing expression levels of various replication and repair genes [132,133].

The action of various *trans*-modifiers also appears to depend on the *cis*-conditions of the repeat, including the direction of replication through the repeats, chromatin conditions and whether the region is actively transcribed [64,93,134]. In multi-cellular organisms, different cell- and tissue-types present an array of *cis*- and *trans*-conditions that are known to result in striking differences in instability rates within the same individuals [5,135–138]. The challenge, therefore, is how to fully assay the complexity of the human disease, which will require the integration of knowledge generated in simple, often-times unicellular model systems, as well as more closely-related mouse and human cell models, into actionable medical genetics.

Modifiers of microsatellite disorders in human genetics data

Inherited repeat length is the greatest determinant of disease onset and severity, but does not tell the complete picture. The progressive, late-onset characteristics of most microsatellite diseases were initially attributed to a low toxicity of the repetitive RNA or polyQ proteins expressed from expanded microsatellites [139,140]. However, it is becoming increasingly clear that this may also be due to somatic expansions that occur throughout life [80,141].

Indeed, due to somatic instability, it is difficult to accurately determine the inherited repeat length that an individual starts with. PCR or Southern blot analysis of a typical genomic DNA preparation typically produces a smear, rather than a clear band. A number of studies have used small-pool PCR, in which the input DNA is diluted to a single copy per reaction, and a large number of reactions are performed per individual. In a study of DM1 pioneering this analysis, Monckton and colleagues observed a sharp lower limit for $(CTG)_n$ length in each individual, interpreted as the inherited repeat length [4]. Contractions below this length were rare, while expansions beyond it were common and highly variable. This technique was also used to measure somatic instability in FRDA, which showed that contractions of $(GAA)_n$ repeats were more frequent than expansions in most tissue types. However, expansions predominated in dorsal root ganglia, consistent with the phenotypic degeneration [5]. FRDA has also been seen to develop with a late age of onset in individuals carrying one expanded allele and one pre-mutation-length allele. Small-pool PCR revealed many somatic expansions in the shorter allele that reached disease length, suggesting an explanation for the eventual development of symptoms [6]. FXS mouse models display somatic variability in repeat length, with clear tissue-specific differences [137]. In Huntington's disease, higher levels of somatic instability measured in the brain cortex was found to be associated with earlier age of onset [142]. This reflects additional work from mouse models [83,85].

Furthermore, in mouse models, both the knockout of Msh2, as well as chemical compounds that suppressed somatic expansions, were able to substantially delay the onset of symptoms of neurodegeneration [80,134,143]. Thus, it follows that for the various microsatellite disorders, the rate at which somatic expansions occur will critically impact disease progression. As discussed above, this appears to be a highly complex characteristic affected by numerous genetic modifiers.

In a pivotal study of DM1 patients, Morales and colleagues used the small-pool PCR approach along with extensive statistical analysis to examine the relationship between somatic $(CTG)_n$ instability and age of onset [144]. While confirming that the inherited repeat length is the largest contributor to age of onset, they also found that the amount of instability observed in each individual also accounted for some of the variation in age of onset. Furthermore, they were able to show that the amount of instability was a heritable trait within families, demonstrating that *cis*- and/or *trans*-modifiers in the genome were altering expansion rates and meaningfully contributing to progression of DM1. Evidence of the existence of familial risk factors for increased repeat instability has also been found for fragile X [42].

Following up with a larger DM1 cohort, Morales and colleagues also checked for the presence of polymorphisms within several candidate genes, and found that a non-synonymous SNP in the DNA mismatch repair gene *MSH3* was correlated with increased $(CTG)_n$ instability [145]. Another variant in *MSH3* was also identified via genome-wide association study (GWAS) as contributing to the progression of Huntington's disease [141]. This reflects earlier work in mouse models, showing that variations in MSH3 occurring naturally between various strain backgrounds contributed to differences in somatic $(CAG)_n$ instability [146]. A larger GWAS involving HD patients resulted in two variants reaching statistical significance, implicating the DNA repair-related factors *FAN1* and *RRM2B*, as well as a variant in the mismatch repair gene *MLH1* that reached significance upon incorporating data from an additional patient cohort [147,148]. Pathway analysis of this data set, which aggregates the effects of genes that fit into various categories, also implicated DNA repair. These studies confirm a link, long established in model systems (see above), between DNA repair and repeat instability.

The rarity of microsatellite disorders is a significant obstacle in identifying modifier genes directly from human genetic data. The above-mentioned studies were limited to only a few hundred to a few thousand individuals. In contrast, the most successful GWAS discoveries have involved cohorts numbering in the tens to hundreds of thousands, focused on common diseases such as type II diabetes [149]. These massive studies have greater statistical power to detect rare and/or low-impact variants. It has been suggested that more statistical power can be gained by combining cohorts from different microsatellite disorders [150]. This was demonstrated in a study testing a panel of candidate variants among patients of various polyglutamine disorders, including HD and multiple forms of spinocerebellar ataxia (SCA) [151]. *FAN1* and *RRM2B*, found previously in the GeM-HD study, were significantly associated with age of onset in the collective polyglutamine cohort, along with an additional DNA repair factor, *PMS2*. However, it is not a given that this approach can be extended to the non-polyglutamine microsatellite disorders. In comparing the above-mentioned studies, we can see consistencies between HD and DM1, such as in *MSH3*, but also differences, as in *PMS2* and *MLH1*, in which no associations were found in DM1 patients [145]. While these particular examples could possibly be due to statistical power or differences in the underlying populations, i.e. whether these SNPs actually appear in both patient populations, it is also quite clear from work in model systems that not all repeats behave alike [20]. Differences between HD and DM1 may be explained by the orientation of the repeats, their placement in the carrier genes, the chromatin environment surrounding the repeats, as well as a number of other factors. Grouping together of other diseases, such as FRDA and FXS, may be even more likely to turn up differences rather than similarities, as the repeats form different types of secondary structures that may involve different molecular processes, or may even respond in different directions to the same *trans*-modifiers. Thus, more inclusive combinatorial studies should be approached with some caution. Certainly, some modifiers of repeat instability will be specific to certain diseases. In such cases, it may not be possible to uncover modifiers purely through population genetics. Thus, much of what has been found in model systems is likely to be important in understanding rare microsatellite diseases, predicting progression and pointing toward therapies.

Strategies for the characterization of modifiers of microsatellite instability

In the era of personal genomics, one might envision the sequencing of patient genomes to reveal risk factors for high levels of somatic instability, or to reveal a particular pathway that may be therapeutically targetable. There are several obstacles to this goal. The complexity described above suggests that the interpretation of patient genomes will benefit from knowing the status of numerous modifier genes. It is highly likely that we still do not know all of the genes involved in microsatellite instability. Furthermore, it is difficult to know whether or how a particular SNP or other mutation affects each gene. And finally, patient genomes contain combinations of SNPs that may behave in unexpected ways.

As discussed above, human population genetics has revealed a handful of modifiers, and model system studies have revealed many more. Much of the work described above was done via the candidate-gene approach: knocking out or knocking down genes suspected to be involved in some aspect of instability. However, gene knockouts or knock-downs can behave differently than mutations in the same gene. A mutation may alter or eliminate only one of a gene’s multiple functions, and this may further affect how the protein behaves as a part of a complex. Conversely, the appearance of a mutation in a gene known to affect instability is not a guarantee that the mutation is biologically significant. This problem is common to many fields, where patient genome sequences reveal numerous variants of unknown significance (VUS) [152]. In each case, further work is required in order to know whether or not these mutations may be medically relevant, and this can be difficult to accomplish in a time frame that benefits the patient. Far better would be to characterize numerous mutations ahead of time using model systems.

The gene candidate approach to modifier discovery also suffers from issues of scope and bias. A favorite unbiased tool of yeast geneticists is deletion library screening. Each individual knockout of a non-essential gene is represented in the library, and an accompanying library alters the expression of each essential gene. A similar approach involves the random insertion of a plasmid to disrupt genes. This type of screening has been applied to microsatellite instability, as well as other related phenomenon, leading to several unexpected discoveries [57,63,131,153–156]. However, there are several gaps in this screening method. In addition to the above-mentioned issues with using gene knockouts and knock-downs, epistatic interactions are not assayed. Many modifiers may only appear when a redundant pathway cannot rescue their effects. An impressive study generated more than 23 million yeast double knockouts to uncover genetic interactions affecting overall fitness [157], though it would not be feasible to apply this approach to more specific questions like microsatellite instability, or to interactions of more than two genes.

Our lab recently developed a novel high-throughput screening method to begin to address some of these shortcomings [93]. Yeast strains containing $(GAA)_n$ repeats were mutated with UV, generating mostly single nucleotide variants in an otherwise uniform genetic background. Strains with elevated rates of repeat expansion underwent whole-genome sequencing and bioinformatic approaches to identify the causal variants. This approach identified a new gene, *YSH1*, which affects repeat expansion in an unexpected manner (see above). This initial success demonstrated several key points: Deletion library screening was not comprehensive in finding all modifiers of $(GAA)_n$ expansion, and many more genes may remain to be found. Not only is *YSH1* conserved from yeast to humans, where it is known as *CPSF-73*, but even the affected amino acids themselves are conserved. This suggests that individual mutations characterized in a yeast model system may be directly interpretable in patient genomes. Such high levels of conservation have been used to predict the severity of mutations [158]. Due to this relationship, it is likely that further screening will produce many more variants at locations conserved in humans. We suggest that this approach can be carried out to the point of saturation, in order to collect a comprehensive list of conserved variants that affect microsatellite instability. In addition, it is feasible to conduct additional screening in various mutant backgrounds, in order to begin to address combinatorial effects.

Our study took advantage of the low cost and dense genome of *S. cerevisiae*, but future approaches may take advantage of CRISPR-Cas9-based techniques to perform screens directly in human cells, although at greater expense [159]. Finally, after a modifier has been identified, it is valuable to understand how that modifier leads to microsatellite instability, including whether it reveals a new mechanism or contributes indirectly to a known pathway. Here too, genetic manipulation in model systems, including CRISPR-based approaches in human cell culture systems, will be a key tool for characterizing *trans*-modifiers, as has already been demonstrated in studies of MSH3 variants affecting $(CAG)_n$ instability [160]. Given that human population genetics does not have the same power to uncover and characterize modifiers of rare microsatellite diseases, such approaches may be the key to understanding and treating these diseases.

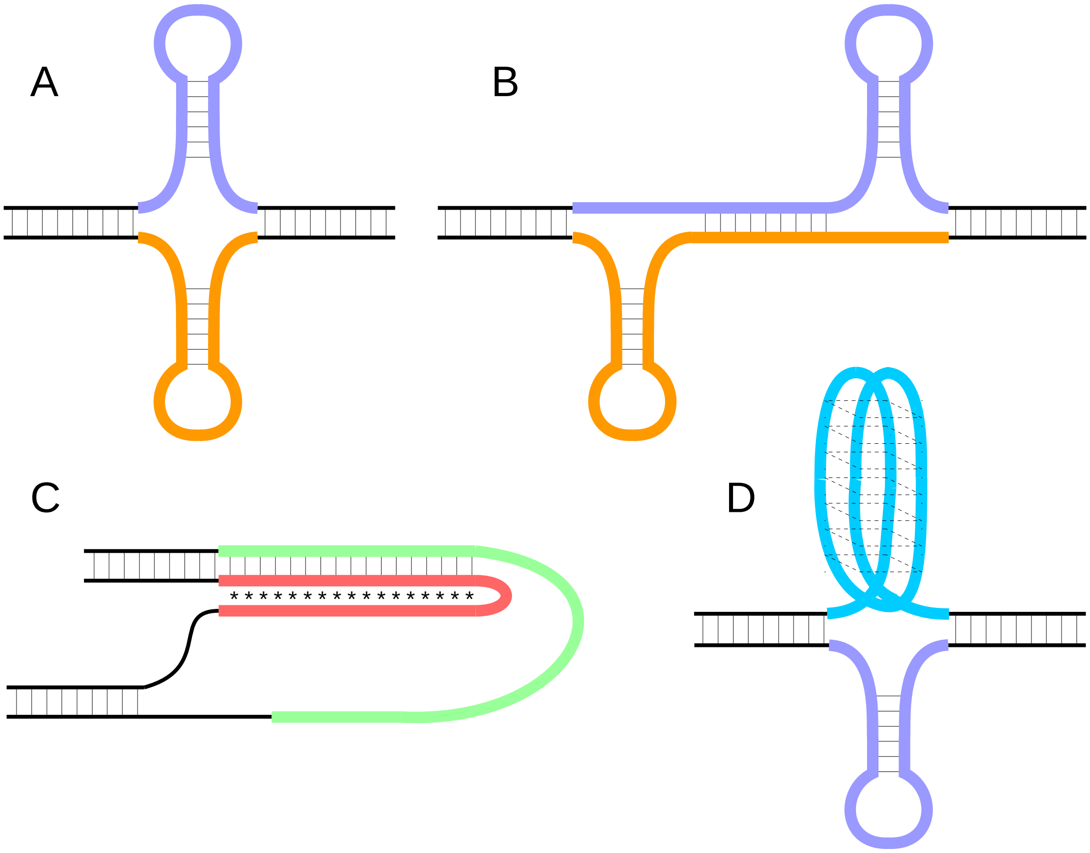


Fig. 1. DNA secondary structures
All panels: repetitive DNA portions pictured in color, non-repetitive DNA pictured in black. A) Cruciform structure, consisting of hairpin structures on the top and bottom strands. B) Slip-stranded DNA, here shown with loop-outs on either end stabilized by hairpin structures. The loop-outs can also remain unpaired, or can be stabilized by a different secondary structure. C) H-DNA, one of several potential secondary structures involving triplex or triple-helical DNA. The triplex is stabilized by Hoogsteen or Reverse-Hoogsteen basepairs (illustrated as *). The fourth strand can remain unpaired, as shown, or can potentially incorporate into further secondary structures. D) G-quadruplex DNA (top strand), also known as G4 or tetrahelical DNA. Several different folding patterns are possible, in addition to the one shown here, involving different arrangements of parallel or anti-parallel strand orientations. Typically, only one of the two strands will contain the regularly-spaced Gs that permit G4 folding, while the other strand can remain unpaired or fold into a different secondary structure, such as a hairpin (bottom strand as pictured).

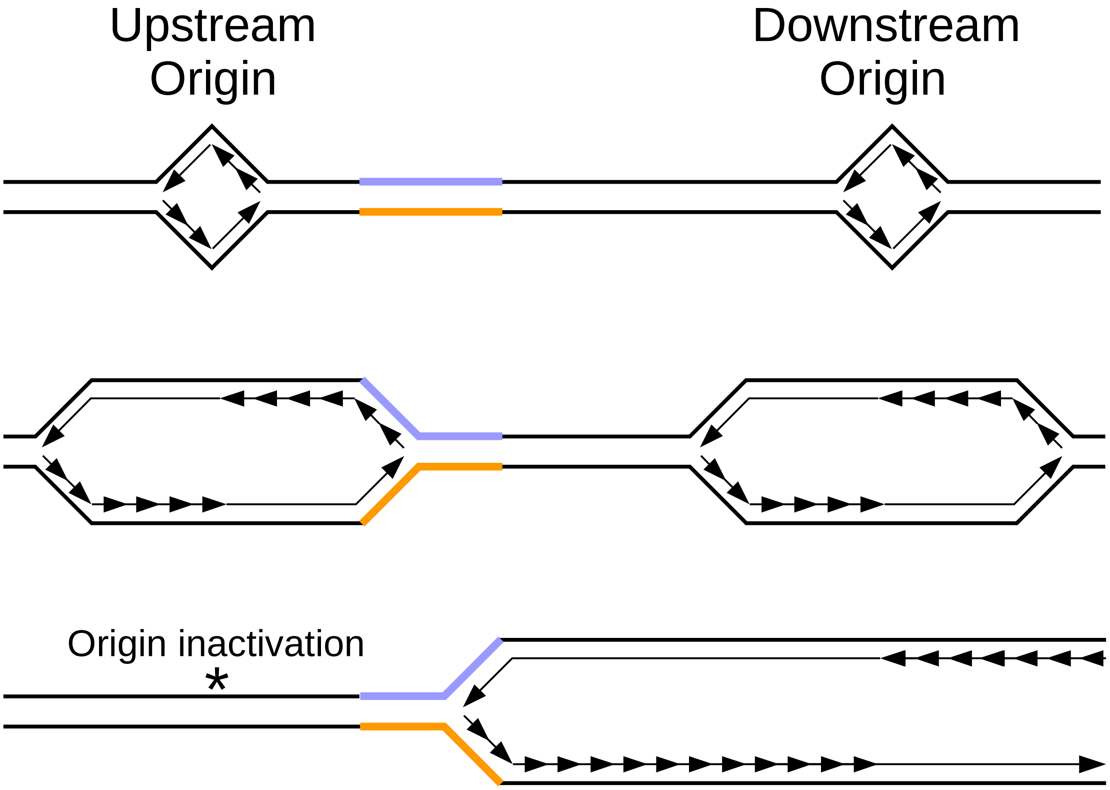


Fig. 2. "Ori-switch" hypothesis
Top panel: Repetitive DNA (colored) sits between two different replication origins.
Middle panel: Replication proceeds bi-directionally from each origin. In this case, the upstream origin reaches the repeats first. The top strand serves as the lagging strand template, while the bottom strand serves as the leading strand template.
Bottom panel: The upstream origin is inactivated, either by a mutation of the binding site, or due to epigenetic changes such as DNA methylation. As a consequence, the downstream origin replicates through the repeats, flipping the orientation such that the top strand now serves as the leading strand template, while the bottom strand serves as the lagging strand template.

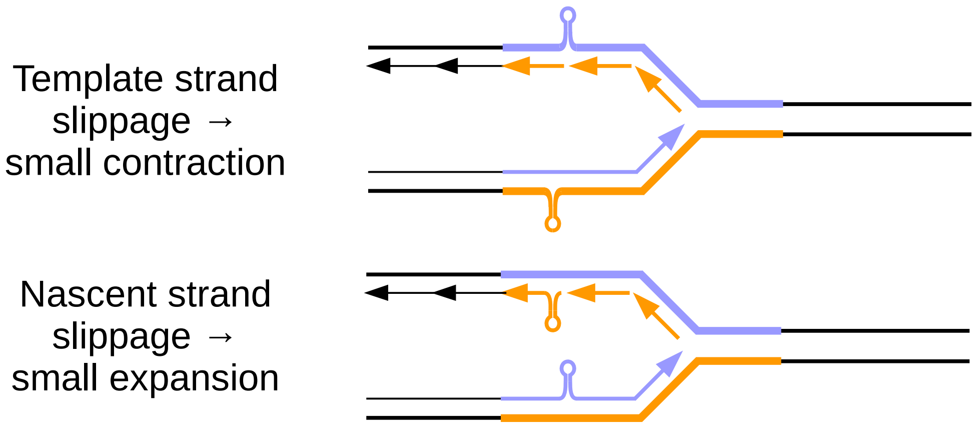

Template strand slippage → small contraction
Nascent strand slippage → small expansion

Fig. 3. Small-scale instability due to replication slippage
Top panel: Secondary structures, here shown as hairpins, form within a repetitive region on either template strand during replication. As a result, the nascent strand skips a small portion of the template, leading to a contraction.
Bottom panel: Secondary structures form on either nascent strand during replication, leading to a small expansion in the newly-generated DNA.
Both panels: Lagging strand synthesis is discontinuous by nature, providing regular opportunities for slippage. Leading strand synthesis is generally continuous, but may occasionally slip while encountering DNA lesions or previously-formed secondary structures.

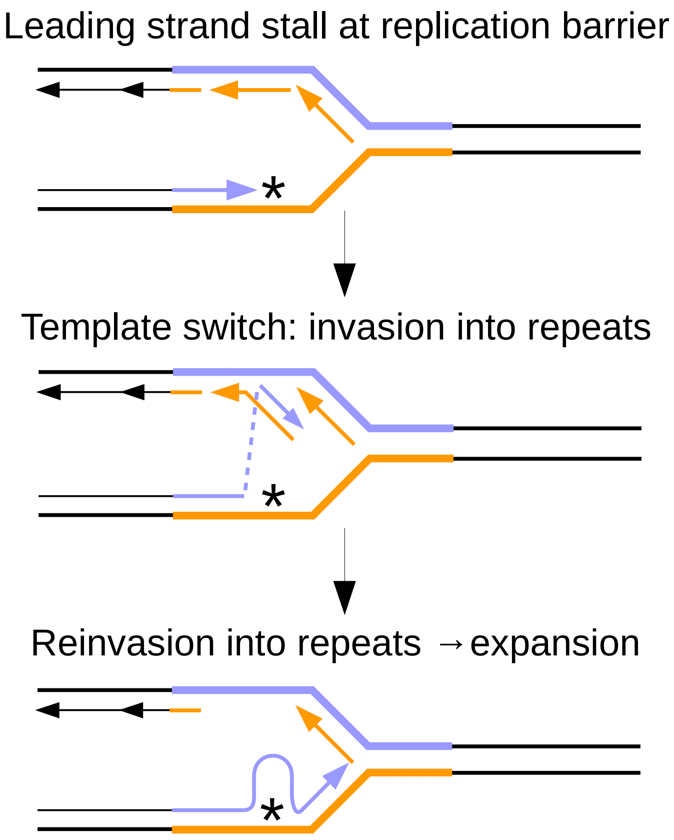
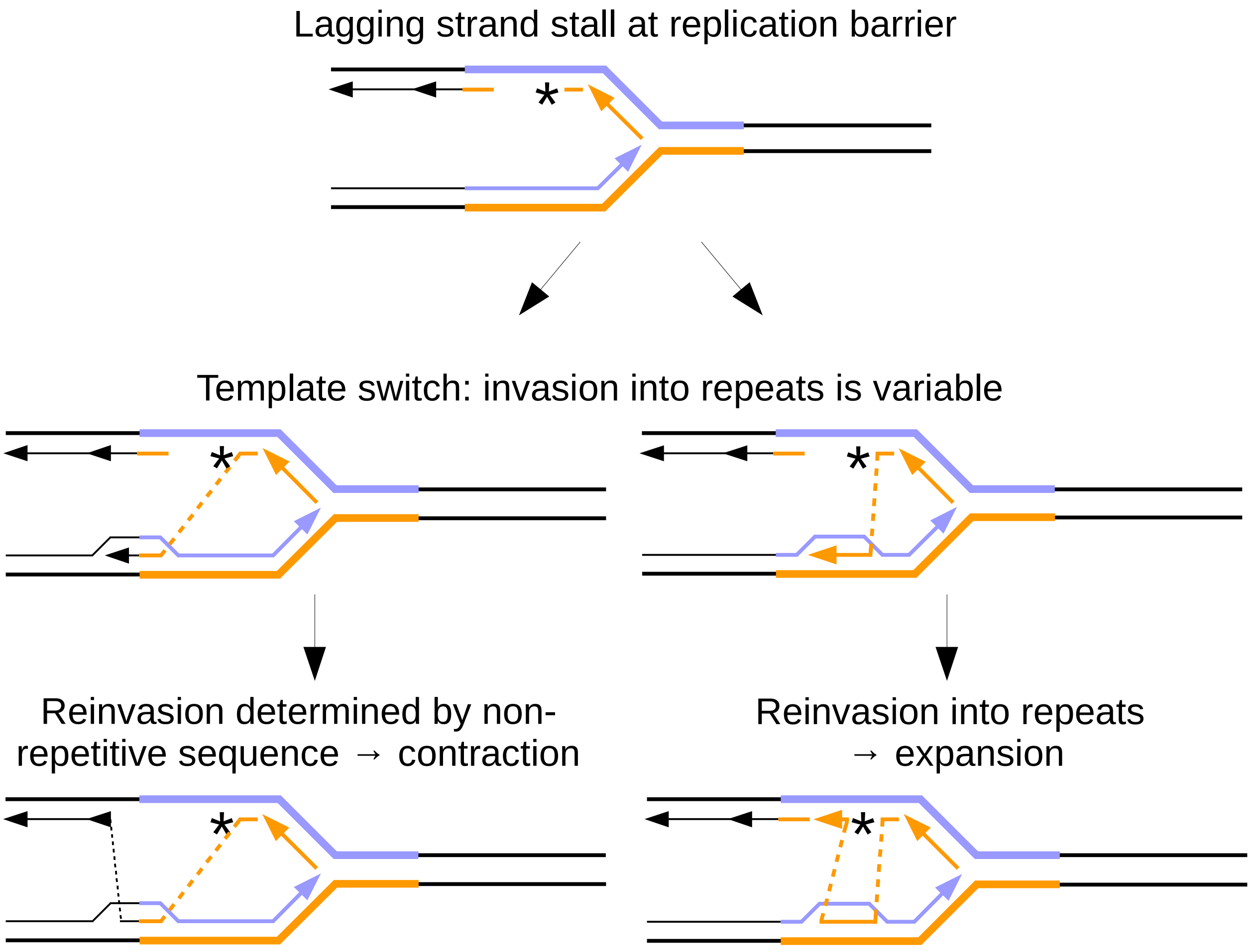

Leading strand stall at replication barrier
*
Template switch: invasion into repeats
*
Reinvasion into repeats → expansion
*

Fig. 4. Instability due to template-switch events
Top panel: During replication, the leading strand may stall after encountering a barrier, including DNA lesions, secondary structures or bound proteins. To bypass the barrier, replication may temporarily switch to use the nascent lagging strand as a template. After reaching the end of the Okazaki fragment, replication re-invades the leading strand template ahead of the lesion. However, within a repetitive region, this re-invasion can occur out-of-register, potentially leading to a large-scale expansion.
Bottom panel: The lagging strand can also encounter a barrier to replication, leading to use of the nascent leading strand as a template. Within a repetitive region, this invasion step can be variable. If it occurs close to the border of the repeat (left panel) the Okazaki fragment will contain non-repetitive sequence, leading to a contraction after re-invasion. If the Okazaki fragment contains only repetitive DNA, reinvasion can occur at any point within the repeat, potentially leading to a large-scale expansion.


Lagging strand stall at replication barrier
*
Template switch: invasion into repeats is variable
*
*
Reinvasion determined by non-repetitive sequence → contraction
*
Reinvasion into repeats → expansion
*

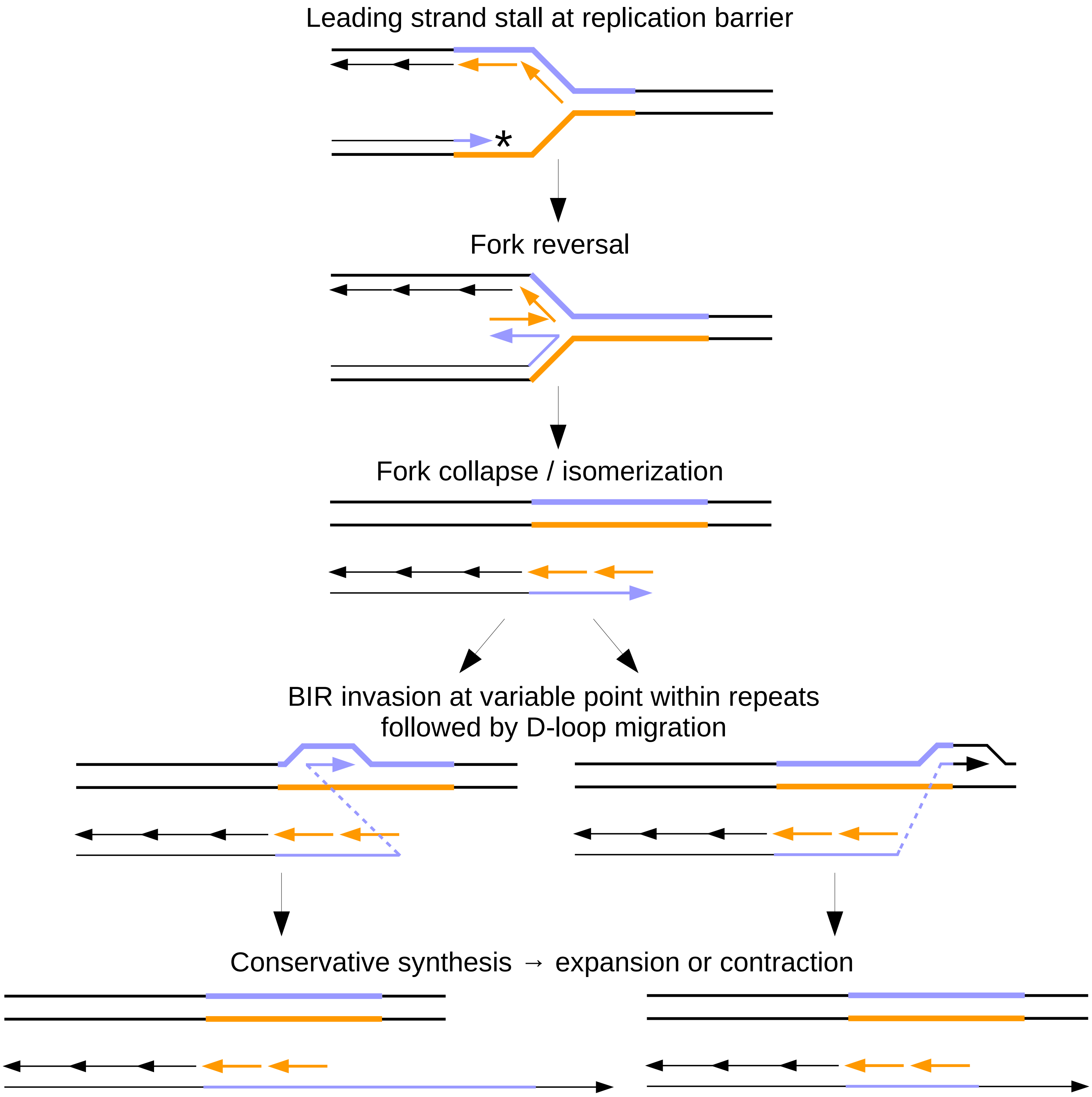

Leading strand stall at replication barrier
*
Fork reversal
Fork collapse / isomerization
BIR invasion at variable point within repeats followed by D-loop migration
Conservative synthesis → expansion or contraction

Fig. 5. Instability due to break-induced replication (BIR)
A stalled replication fork (top panel) that cannot be restarted by other means may lead to fork reversal, resulting in a chicken-foot structure (second panel). Together, the two template strands are intact, while the two nascent strands make up a one-ended double strand break (third panel). The nascent leading strand can then invade the template via homologous recombination to initiate BIR. If the invading end consists of repeats, invasion can occur anywhere within the repetitive tract. This can lead to a large-scale expansion – as much as a doubling of the repeat tract – if invasion occurs near the beginning of the repeats (left). In the opposite case (right), invasion can occur towards the end of the repetive tract, skipping a large portion of the repeat and leading to a large-scale contraction. Synthesis of this strand continues (bottom panel), potentially until reaching the end of the chromosome, before the remaining strand is filled in, resulting in conservative DNA replication.

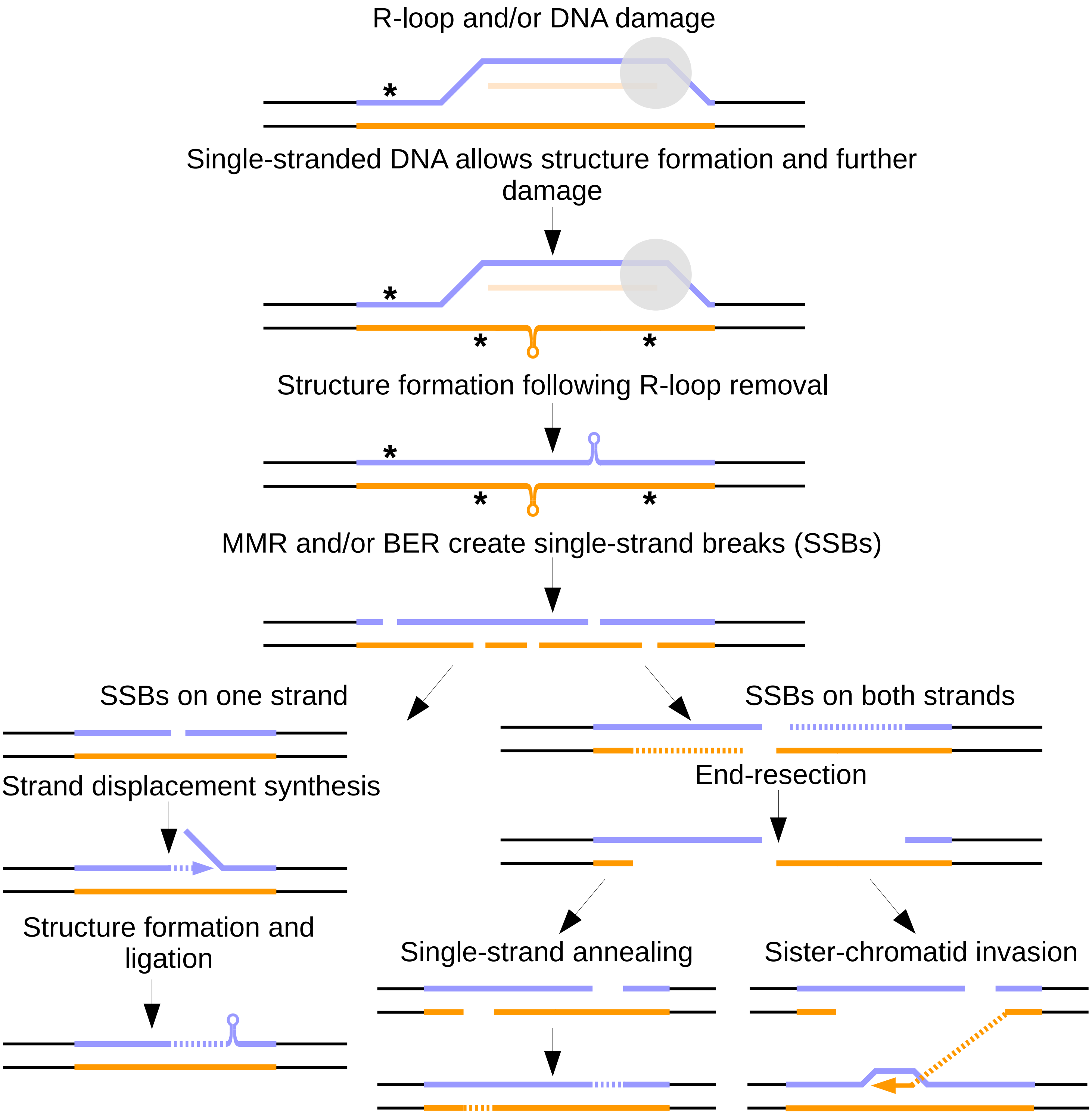

R-loop and/or DNA damage
*
Single-stranded DNA allows structure formation and further damage
*
*
*
Structure formation following R-loop removal
*
*
*
MMR and/or BER create single-strand breaks (SSBs)
SSBs on one strand
SSBs on both strands
Strand displacement synthesis
End-resection
Structure formation and ligation
Single-strand annealing
Sister-chromatid invasion

Fig. 6. Instability due to DNA damage and repair
R-loops (extended RNA-DNA hybrids) can expose long stretches of single-stranded DNA, which can increase the rate of DNA damage (*), including oxidative damage and cytosine deamination. DNA damage undergoes base excision repair, leading to single-strand breaks (SSBs). Secondary structure formation in repetitive tracts occurs in ssDNA exposed on the non-template strand while the R-loop is present, and/or on the template strand following removal of the R-loop. Secondary structures can be recognized and cleaved by various enzymes, potentially leading to contractions, and also leading to SSBs. If SSBs occur only on one strand, repair can occur via strand displacement synthesis, creating a flap that can form secondary structures. The flap can be stabilized by mismatch repair enzymes and incorporated into the repaired DNA strand, causing a repeat expansion. If SSBs occur on both strands, this results in a double strand break. DSB repair can occur by a number of mechanisms, including non-homologous end joining (not shown), BIR (see Fig. 5), single-strand annealing between repetitive tracts, which can result in contractions, and sister chromatid invasion and recombination, which can potentially result in expansions or contractions.

References:

1 Pearson, C.E. *et al.* (2005) Repeat instability: Mechanisms of dynamic mutations. *Nat. Rev. Genet.* 6, 729–742

2 Mirkin, S.M. (2007) Expandable DNA repeats and human disease. *Nature* 447, 932–940

3 Orr, H.T. and Zoghbi, H.Y. (2007) Trinucleotide Repeat Disorders. *Annu. Rev. Neurosci.* 30, 575–621

4 Monckton, D.G. *et al.* (1995) Somatic mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analyses. *Hum. Mol. Genet.* 4, 1–8

5 De Biase, I. *et al.* (2007) Progressive GAA expansions in dorsal root ganglia of Friedreich's ataxia patients. *Ann. Neurol.* 61, 55–60

6 Sharma, R. *et al.* (2002) The GAA triplet-repeat sequence in Friedreich ataxia shows a high level of somatic instability in vivo, with a significant predilection for large contractions. *Hum. Mol. Genet.* 11, 2175–87

7 McClellan, J.A. *et al.* (1990) Superhelical torsion in cellular DNA responds directly to environmental and genetic factors. *Proc. Natl. Acad. Sci. U. S. A.* 87, 8373–8377

8 Dayn, A. *et al.* (1991) Formation of (dA-dT)n cruciforms in Escherichia coli cells under different environmental conditions. *J Bacteriol* 173, 2658–2664

9 Marquis Gacy, A. *et al.* (1995) Trinucleotide repeats that expand in human disease form hairpin structures in vitro. *Cell* 81, 533–540

10 Dayn, a *et al.* (1992) Intramolecular DNA triplexes: unusual sequence requirements and influence on DNA polymerization. *Proc. Natl. Acad. Sci. U. S. A.* 89, 11406–11410

11 Sakamoto, N. *et al.* (1999) Sticky DNA: Self-association properties of long GAA·TTC repeats in R·R·Y triplex structures from Friedreich's ataxia. *Mol. Cell* 3, 465–475

12 Fry, M. and Loeb, L. a (1994) The fragile X syndrome d(CGG)n nucleotide repeats form a stable tetrahelical structure. *Proc. Natl. Acad. Sci. U. S. A.* 91, 4950–4954

13 Renciuk, D. *et al.* (2009) Quadruplex-forming properties of FRAXA (CGG) repeats interrupted by (AGG) triplets. *Biochimie* 91, 416–422

14 Kwok, C.K. and Merrick, C.J. (2017) G-Quadruplexes: Prediction, Characterization, and Biological Application. *Trends Biotechnol.* 35, 997–1013

15 Kunkel, T. a (1993) Nucleotide repeats. Slippery DNA and diseases. *Nature* 365, 207–208

16 Pearson, C.E. *et al.* (2002) Slipped-strand DNAs formed by long (CAG)*(CTG) repeats: slipped-out repeats and slip-out junctions. *Nucleic Acids Res.* 30, 4534–47

17 Mirkin, S.M. and Frank-Kamenetskii, M.D. (1994) H-DNA and Related Structures. *Annu. Rev. Biophys. Biomol. Struct.* 23, 541–576

18 Grabczyk, E. (2000) The GAATTC triplet repeat expanded in Friedreich's ataxia impedes transcription elongation by T7 RNA polymerase in a length and supercoil dependent manner. *Nucleic Acids Res.* 28, 2815–2822

19 Napierala, M. *et al.* (2005) Increased negative superhelical density in vivo enhances the genetic instability of triplet repeat sequences. *J. Biol. Chem.* 280, 37366–37376

20 Kim, J.C. and Mirkin, S.M. (2013) The balancing act of DNA repeat expansions. *Curr. Opin. Genet. Dev.* 23, 280–288

21 Rolfsmeier, M.L. *et al.* (2001) Cis-elements governing trinucleotide repeat instability in Saccharomyces cerevisiae. *Genetics* 157, 1569–1579

22 Dixon, M.J. and Lahue, R.S. (2004) DNA elements important for CAG?? CTG repeat thresholds in Saccharomyces cerevisiae. *Nucleic Acids Res.* 32, 1289–1297

23 Shishkin, A.A. *et al.* (2009) Large-Scale Expansions of Friedreich's Ataxia GAA Repeats in Yeast. *Mol. Cell* 35, 82–92

24 Cherng, N. *et al.* (2011) Expansions, contractions, and fragility of the spinocerebellar ataxia type 10 pentanucleotide repeat in yeast. *Proc. Natl. Acad. Sci.* 108, 2843–2848

25 Freudenreich, C.H. (1998) Expansion and Length-Dependent Fragility of CTG Repeats in Yeast. *Science.* 279, 853–856

26 Kim, H.M. *et al.* (2008) Chromosome fragility at GAA tracts in yeast depends on repeat orientation and requires mismatch repair. *EMBO J.* 27, 2896–2906

27 Kumari, D. *et al.* (2015) Evidence for chromosome fragility at the frataxin locus in Friedreich ataxia. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* 781, 14–21

28 Zhang, H. and Freudenreich, C.H. (2007) An AT-Rich Sequence in Human Common Fragile Site FRA16D Causes Fork Stalling and Chromosome Breakage in S. cerevisiae. *Mol. Cell* 27, 367–379

29 Thys, R.G. *et al.* (2015) DNA Secondary Structure at Chromosomal Fragile Sites in Human Disease DNA Secondary Structure in Human Disease. *Curr. Genomics* 16, 60–70

30 Bacolla, A. *et al.* (2016) Translocation and deletion breakpoints in cancer genomes are associated with potential non-B DNA-forming sequences. *Nucleic Acids Res.* 44, 5673–5688

31 Meservy, J.L. *et al.* (2003) Long CTG tracts from the myotonic dystrophy gene induce deletions and rearrangements during recombination at the APRT locus in CHO cells. *Mol. Cell. Biol.* 23, 3152–62

32 Bacolla, A. *et al.* (2006) The involvement of non-B DNA structures in gross chromosomal rearrangements. *DNA Repair (Amst).* 5, 1161–1170

33 Aksenova, A.Y. *et al.* (2013) Genome rearrangements caused by interstitial telomeric sequences in yeast. *Proc. Natl. Acad. Sci. U. S. A.* 110, 19866–71

34 McGinty, R.J. *et al.* (2017) Nanopore sequencing of complex genomic rearrangements in yeast reveals mechanisms of repeat-mediated double-strand break repair. *Genome Res.* DOI: 10.1101/gr.228148.117

35 Usdin, K. *et al.* (2015) Repeat instability during DNA repair: Insights from model systems. *Crit. Rev. Biochem. Mol. Biol.* 50, 142–167

36 Polyzos, A.A. and McMurray, C.T. (2017) Close encounters: Moving along bumps, breaks, and bubbles on expanded trinucleotide tracts. *DNA Repair (Amst).* 56, 144–155

37 Neil, A.J. *et al.* (2017) Precarious maintenance of simple DNA repeats in eukaryotes. *BioEssays* 39, 1–10

38 Huddleston, J. *et al.* (2017) Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* 27, 677–685

39 Menon, R.P. *et al.* (2013) The Role of Interruptions in polyQ in the Pathology of SCA1. *PLoS Genet.* 9,

40 Sakamoto, N. *et al.* (2001) GGA·TCC-interrupted Triplets in Long GAA·TTC Repeats Inhibit the Formation of Triplex and Sticky DNA Structures, Alleviate Transcription Inhibition, and Reduce Genetic Instabilities. *J. Biol. Chem.* 276, 27178–27187

41 Yrigollen, C.M. *et al.* (2012) AGG interruptions within the maternal FMR1 gene reduce the risk of offspring with fragile X syndrome. *Genet. Med.* 14, 729–736

42 Nolin, S.L. *et al.* (2013) Fragile X AGG analysis provides new risk predictions for 45-69 repeat alleles. *Am. J. Med. Genet. Part A* 161, 771–778

43 Kang, S. *et al.* (1995) Expansion and deletion of CTG repeats from human disease genes are determined by the direction of replication in E. coli. *Nat. Genet.* 10, 213–218

44 Freudenreich, C.H. *et al.* (1997) Stability of a CTG/CAG Trinucleotide Repeat in Yeast Is Dependent on Its Orientation in the Genome. *Mol. Cell. Biol.*

45 Miret, J.J. *et al.* (1998) Orientation-dependent and sequence-specific expansions of CTG/CAG trinucleotide repeats in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U. S. A.* 95, 12438–43

46 Cleary, J.D. *et al.* (2002) Evidence of cis-acting factors in replication-mediated trinucleotide repeat instability in primate cells. *Nat. Genet.* 31, 37–46

47 Rindler, P.M. *et al.* (2006) Replication in mammalian cells recapitulates the locus-specific differences in somatic instability of genomic GAA triplet-repeats. *Nucleic Acids Res.* 34, 6352–6361

48 Liu, G. *et al.* (2010) Replication-dependent instability at (CTG)·(CAG) repeat hairpins in human cells. *Nat. Chem. Biol.* 6, 652–659

49 Mirkin, S.M. and Smirnova, E. V. (2002) Positioned to expand. *Nat. Genet.* 31, 5–6

50 Ennis, S. *et al.* (2007) Closely linked cis-acting modifier of expansion of the CGG repeat in high risk FMR1 haplotypes. *Hum. Mutat.* 28, 1216–1224

51 Gerhardt, J. *et al.* (2014) Cis-acting DNA sequence at a replication origin promotes repeat expansion to fragile X full mutation. *J. Cell Biol.* 206, 599–607

52 Gerhardt, J. *et al.* (2014) The DNA Replication Program Is Altered at the FMR1 Locus in Fragile X Embryonic Stem Cells. *Mol. Cell* 53, 19–31

53 Cleary, J.D. *et al.* (2010) Tissue-and age-specific DNA replication patterns at the CTG/CAG-expanded human myotonic dystrophy type 1 locus. *Nat. Struct. Mol. Biol.* 17, 1079–1087

54 Gerhardt, J. *et al.* (2016) Stalled DNA Replication Forks at the Endogenous GAA Repeats Drive Repeat Expansion in Friedreich's Ataxia Cells. *Cell Rep.* 16, 1218–1227

55 Du, J. *et al.* (2012) Role of mismatch repair enzymes in GAA·TTC triplet-repeat expansion in friedreich ataxia induced pluripotent stem cells. *J. Biol. Chem.* 287, 29861–29872

56 Gan, H. *et al.* (2017) Checkpoint Kinase Rad53 Couples Leading- and Lagging-Strand DNA Synthesis under Replication Short Article Checkpoint Kinase Rad53 Couples DNA Synthesis under Replication Stress. *Mol. Cell* 68, 446–455.e3

57 Zhang, Y. *et al.* (2012) Genome-wide screen identifies pathways that govern GAA/TTC repeat fragility and expansions in dividing and nondividing yeast cells. *Mol. Cell* 48, 254–265

58 Bhattacharyya, S. and Lahue, R.S. (2004) Saccharomyces cerevisiae Srs2 DNA Helicase Selectively Blocks Expansions of Trinucleotide Repeats. *Mol. Cell. Biol.* 24, 7324–7330

59 Frizzell, A. *et al.* (2014) RTEL1 inhibits trinucleotide repeat expansions and fragility. *Cell Rep.* 6, 827–835

60 Nguyen, J.H.G. *et al.* (2017) Differential requirement of Srs2 helicase and Rad51 displacement activities in replication of hairpin-forming CAG/CTG repeats. *Nucleic Acids Res.* 45, 4519–4531

61 Anand, R.P. *et al.* (2012) Overcoming natural replication barriers: Differential helicase requirements. *Nucleic Acids Res.* 40, 1091–1105

62 Shah, K.A. *et al.* (2012) Role of DNA polymerases in repeat-mediated genome instability. *Cell Rep* 2, 1088–1095

63 Razidlo, D.F. and Lahue, R.S. (2008) Mrc1, Tof1 and Csm3 inhibit CAG·CTG repeat instability by at least two mechanisms. *DNA Repair (Amst).* 7, 633–640

64 Liu, G. *et al.* (2012) Altered Replication in Human Cells Promotes DMPK (CTG)n-(CAG)n Repeat Instability. *Mol. Cell. Biol.* 32, 1618–1632

65 Daee, D.L. *et al.* (2007) Postreplication repair inhibits CAG.CTG repeat expansions in Saccharomyces cerevisiae. *Mol. Cell. Biol.* 27, 102–10

66 Ye, Y. *et al.* (2016) The Saccharomyces cerevisiae Mre11-Rad50-Xrs2 complex promotes trinucleotide repeat expansions independently of homologous recombination. *DNA Repair (Amst).* 43, 1–8

67 Spiro, C. *et al.* (1999) Inhibition of FEN-1 processing by DNA secondary structure at trinucleotide repeats. *Mol. Cell* 4, 1079–1085

68 Liu, Y. *et al.* (2004) Saccharomyces cerevisiae flap endonuclease 1 uses flap equilibration to maintain triplet repeat stability. *Mol Cell Biol* 24, 4049–4064

69 Tsutakawa, S.E. *et al.* (2017) Phosphate steering by Flap Endonuclease 1 promotes 5′-flap specificity and incision to prevent genome instability. *Nat. Commun.* 8, 1–14

70 Fouché, N. *et al.* (2006) Replication fork regression in repetitive DNAs. *Nucleic Acids Res.* 34, 6044–6050

71 Kim, J.C. *et al.* (2017) The role of break-induced replication in large-scale expansions of (CAG)n/(CTG)n repeats. *Nat. Struct. Mol. Biol.* 24, 55–60

72 Ebersole, T. *et al.* Mechanisms of genetic instabilities caused by the (CGG)n repeats in an experimental mammalian system. *(in submission)*

73 Su, X.A. *et al.* (2015) Regulation of recombination at yeast nuclear pores controls repair and triplet repeat stability. *Genes Dev.* 29, 1006–1017

74 Miné-Hattab, J. and Rothstein, R. (2012) Increased chromosome mobility facilitates homology search during recombination. *Nat. Cell Biol.* 14, 510–517

75 Dion, V. and Gasser, S.M. (2013) Chromatin movement in the maintenance of genome stability. *Cell* 152, 1355–1364

76 Becker, A. *et al.* (2014) ATM alters the otherwise robust chromatin mobility at sites of DNA Double-Strand Breaks (DSBs) in human cells. *PLoS One* 9, 1–10

77 Kovtun, I. V. and McMurray, C.T. (2001) Trinucleotide expansion in haploid germ cells by gap repair. *Nat. Genet.* 27, 407–411

78 Kovtun, I. V. *et al.* (2007) OGG1 initiates age-dependent CAG trinucleotide expansion in somatic cells. *Nature* 447, 447–452

79 Møllersen, L. *et al.* (2017) Neil1 is a genetic modifier of somatic and germline CAG trinucleotide repeat instability in R6/1 mice. *Hum. Mol. Genet.* 21, 4939–4947

80 Budworth, H. *et al.* (2015) Suppression of Somatic Expansion Delays the Onset of Pathophysiology in a Mouse Model of Huntington's Disease. *PLoS Genet.* 11, 1–22

81 Lokanga, R.A. *et al.* (2015) Heterozygosity for a Hypomorphic Pol?? Mutation Reduces the Expansion Frequency in a Mouse Model of the Fragile X-Related Disorders. *PLoS Genet.* 11, 1–16

82 Su, X.A. and Freudenreich, C.H. (2017) Cytosine deamination and base excision repair cause R-loop–induced CAG repeat fragility and instability in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci.* 114, E8392–E8401

83 Kovalenko, M. *et al.* (2012) Msh2 Acts in Medium-Spiny Striatal Neurons as an Enhancer of CAG Instability and Mutant Huntingtin Phenotypes in Huntington ' s Disease Knock-In Mice. *PLoS One* 7, 1–10

84 Pinto, R.M. *et al.* (2013) Mismatch Repair Genes Mlh1 and Mlh3 Modify CAG Instability in Huntington's Disease Mice: Genome-Wide and Candidate Approaches. *PLoS Genet.* 9,

85 Iyer, R.R. *et al.* (2015) DNA Triplet Repeat Expansion and Mismatch Repair. *Annu. Rev. Biochem.* 84, 199–226

86 Manley, K. *et al.* (1999) Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. *Nat. Genet.* 23, 471–473

87 van den Broek, W.J.A.A. *et al.* (2002) Somatic expansion behaviour of the (CTG)n repeat in myotonic dystrophy knock-in mice is differentially affected by Msh3 and Msh6 mismatch-repair proteins. *Hum. Mol. Genet.* 11, 191–8

88 Owen, B.A.L. *et al.* (2005) (CAG)n-hairpin DNA binds to Msh2-Msh3 and changes properties of mismatch recognition. *Nat. Struct. Mol. Biol.* 12, 663–670

89 McMurray, C.T. (2008) Hijacking of the mismatch repair system to cause CAG expansion and cell death in neurodegenerative disease. *DNA Repair (Amst).* 7, 1121–1134

90 Pluciennik, A. *et al.* (2002) Long CTG·CAG repeats from myotonic dystrophy are preferred sites for intermolecular recombination. *J. Biol. Chem.* 277, 34074–34086

91 Napierala, M. *et al.* (2002) Long CTG·CAG repeat sequences markedly stimulate intramolecular recombination. *J. Biol. Chem.* 277, 34087–34100

92 Sundararajan, R. *et al.* (2010) Double-strand break repair pathways protect against CAG/CTG repeat expansions, contractions and repeat-mediated chromosomal fragility in Saccharomyces cerevisiae. *Genetics* 184, 65–77

93 McGinty, R.J. *et al.* (2017) A Defective mRNA Cleavage and Polyadenylation Complex Facilitates Expansions of Transcribed (GAA)nRepeats Associated with Friedreich's Ataxia. *Cell Rep.* 20, 2490–2500

94 Pollard, L.M. *et al.* (2008) Repair of DNA double-strand breaks within the (GAA·TTC)n sequence results in frequent deletion of the triplet-repeat sequence. *Nucleic Acids Res.* 36, 489–500

95 Mott, C. and Symington, L.S. (2011) RAD51-independent inverted-repeat recombination by a strand-annealing mechanism. *DNA Repair (Amst).* 10, 408–415

96 Crespan, E. *et al.* (2012) Microhomology-mediated DNA strand annealing and elongation by human DNA polymerases λ and β on normal and repetitive DNA sequences. *Nucleic Acids Res.* 40, 5577–5590

97 Finn, K.J. and Li, J.J. (2013) Single-Stranded Annealing Induced by Re-Initiation of Replication Origins Provides a Novel and Efficient Mechanism for Generating Copy Number Expansion via Non-Allelic Homologous Recombination. *PLoS Genet.* 9,

98 Tang, W. *et al.* (2013) Genomic deletions and point mutations induced in Saccharomyces cerevisiae by the trinucleotide repeats (GAA??TTC) associated with Friedreich's ataxia. *DNA Repair (Amst).* 12, 10–17

99 Chumki, S.A. *et al.* (2016) Remarkably long-tract gene conversion induced by fragile site instability in Saccharomyces cerevisiae. *Genetics* 204, 115–128

100 Nag, D.K. *et al.* (2004) Both CAG repeats and inverted DNA repeats stimulate spontaneous unequal sister-chromatid exchange in Saccharomyces cerevisiae. *Nucleic Acids Res.* 32, 5677–5684

101 House, N.C.M. *et al.* (2014) NuA4 Initiates Dynamic Histone H4 Acetylation to Promote High-Fidelity Sister Chromatid Recombination at Postreplication Gaps. *Mol. Cell* 55, 818–828

102 Lin, Y. *et al.* (2006) Transcription promotes contraction of CAG repeat tracts in human cells. *Nat. Struct. Mol. Biol.* 13, 179–180

103 Groh, M. *et al.* (2014) R-loops Associated with Triplet Repeat Expansions Promote Gene Silencing in Friedreich Ataxia and Fragile X Syndrome. *PLoS Genet.* 10,

104 Shah, K.A. *et al.* (2014) Coupling Transcriptional State to Large-Scale Repeat Expansions in Yeast. *Cell Rep.* 9, 1594–1603

105 Belotserkovskii, B.P. *et al.* (2007) A triplex-forming sequence from the human c-MYC promoter interferes with DNA transcription. *J. Biol. Chem.* 282, 32433–32441

106 Pandey, S. *et al.* (2015) Transcription blockage by stable H-DNA analogs in vitro. *Nucleic Acids Res.* 43, 6994–7004

107 Jung, J. and Bonini, N. (2007) CREB-Binding Protein Modulates Repeat Instability in a Drosophila Model for PolyQ Disease. *Science (80-.).* 315, 1857–1859

108 Debacker, K. *et al.* (2012) Histone deacetylase complexes promote trinucleotide repeat expansions. *PLoS Biol.* 10,

109 Koch, M.R. *et al.* The Isw1 chromatin remodeler prevents excision repair-induced CAG repeat expansions during transcription in Saccharomyces cerevisiae. *Genetics* In press,

110 Prado, F. *et al.* (1997) Recombination between DNA repeats in yeast hpr1delta cells is linked to transcription elongation. *EMBO J.* 16, 2826–35

111 Dominguez-Sanchez, M.S. *et al.* (2011) Genome Instability and Transcription Elongation Impairment in Human Cells Depleted of THO / TREX. *PLoS Genet.* 7, 19–22

112 Gómez-González, B. *et al.* (2011) Genome-wide function of THO/TREX in active genes prevents R-loop-dependent replication obstacles. *EMBO J.* 30, 3106–3119

113 Gavaldá, S. *et al.* (2013) R-Loop Mediated Transcription-Associated Recombination in trf4Δ Mutants Reveals New Links between RNA Surveillance and Genome Integrity. *PLoS One* 8, 1–8

114 Santos-Pereira, J.M. *et al.* (2014) Npl3, a new link between RNA-binding proteins and the maintenance of genome integrity. *Cell Cycle* 13, 1524–1529

115 Gavaldá, S. *et al.* (2016) Excess of Yra1 RNA-Binding Factor Causes Transcription-Dependent Genome Instability, Replication Impairment and Telomere Shortening. *PLoS Genet.* 12,

116 Salas-Armenteros, I. *et al.* (2017) Human THO – Sin 3 A interaction reveals new mechanisms to prevent R-loops that cause genome instability. *EMBO J.* 36, 3532–3547

117 Duquette, M.L. *et al.* (2004) Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. *Genes Dev.* 18, 1618–1629

118 Lin, Y. and Wilson, J.H. (2007) Transcription-Induced CAG Repeat Contraction in Human Cells Is Mediated in Part by Transcription-Coupled Nucleotide Excision Repair. *Mol. Cell. Biol.* 27, 6209–6217

119 Belotserkovskii, B.P. *et al.* (2013) Transcription blockage by homopurine DNA sequences: Role of sequence composition and single-strand breaks. *Nucleic Acids Res.* 41, 1817–1828

120 Sollier, J. and Cimprich, K.A. (2015) Breaking bad: R-loops and genome integrity. *Trends Cell Biol.* 25, 514–522

121 Polleys, E.J. *et al.* (2017) Role of recombination and replication fork restart in repeat instability. *DNA Repair (Amst).* 56, 156–165

122 Saini, N. *et al.* (2013) Fragile DNA Motifs Trigger Mutagenesis at Distant Chromosomal Loci in Saccharomyces cerevisiae. *PLoS Genet.* 9, 1–12

123 Shah, K.A. and Mirkin, S.M. (2015) The hidden side of unstable DNA repeats: Mutagenesis at a distance. *DNA Repair (Amst).* 32, 106–112

124 Sakofsky, C.J. and Malkova, A. (2017) Break induced replication in eukaryotes: mechanisms, functions, and consequences. *Crit. Rev. Biochem. Mol. Biol.* 52, 395–413

125 Zhao, H. *et al.* (2017) Peptide SS-31 upregulates frataxin expression and improves the quality of mitochondria: Implications in the treatment of Friedreich ataxia. *Sci. Rep.* 7, 1–11

126 Erwin, G. *et al.* (2017) Synthetic transcription elongation factors license transcription across repressive chromatin. *Science (80-.).* 6414,

127 Li, L. *et al.* (2016) Activating frataxin expression by repeat-targeted nucleic acids. *Nat. Commun.* 7, 1–8

128 Bergquist, H. *et al.* (2016) Disruption of higher order DNA structures in friedreich's ataxia (GAA)n Repeats by PNA or LNA Targeting. *PLoS One* 11, 1–24

129 Richard, G. (2015) Shortening trinucleotide repeats using highly specific endonucleases: a possible approach to gene therapy? *Trends Genet.* 31, 177–186

130 Cinesi, C. *et al.* (2016) Contracting CAG/CTG repeats using the CRISPR-Cas9 nickase. *Nat. Commun.* 7,

131 Gellon, L. *et al.* (2011) New Functions of Ctf18-RFC in Preserving Genome Stability outside Its Role in Sister Chromatid Cohesion. *PLoS Genet.* 7,

132 Goula, A. *et al.* (2009) Stoichiometry of Base Excision Repair Proteins Correlates with Increased Somatic CAG Instability in Striatum over Cerebellum in Huntington's Disease Transgenic Mice. *PLoS Genet.* 5,

133 Mason, A.G. *et al.* (2014) Expression levels of DNA replication and repair genes predict regional somatic repeat instability in the brain but are not altered by polyglutamine disease protein expression or age. *Hum. Mol. Genet.* 23, 1606–1618

134 Wheeler, V.C. *et al.* (2003) Mismatch repair gene Msh2 modifies the timing of early disease in HdhQ111 striatum. *Hum. Mol. Genet.* 12, 273–281

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

136 Goula, A.V. *et al.* (2012) Transcription Elongation and Tissue-Specific Somatic CAG Instability. *PLoS Genet.* 8,

137 Lokanga, R.A. *et al.* (2013) Somatic Expansion in Mouse and Human Carriers of Fragile X Premutation Alleles. *Hum. Mutat.* 34, 157–166

138 Dion, V. (2014) Tissue specificity in DNA repair: Lessons from trinucleotide repeat instability. *Trends Genet.* 30, 220–229

139 Amiri, K. *et al.* (2008) Fragile X–Associated Tremor/Ataxia Syndrome. *Neurol. Rev.* 65, 19–25

140 Zoghbi, H.Y. and Orr, H.T. (2009) Pathogenic mechanisms of a polyglutamine-mediated neurodegenerative disease, Spinocerebellar ataxia type 1. *J. Biol. Chem.* 284, 7425–7429

141 Moss, D.J.H. *et al.* (2017) Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study. *Lancet Neurol.* 16, 701–711

142 Swami, M. *et al.* (2009) Somatic expansion of the Huntington's disease CAG repeat in the brain is associated with an earlier age of disease onset. *Hum. Mol. Genet.* 18, 3039–3047

143 Suelves, N. *et al.* (2017) A selective inhibitor of histone deacetylase 3 prevents cognitive deficits and suppresses striatal CAG repeat expansions in Huntington's disease mice. *Nat. Sci. Reports* 7, 1–15

144 Morales, F. *et al.* (2012) Somatic instability of the expanded CTG triplet repeat in myotonic dystrophy type 1 is a heritable quantitative trait and modifier of disease severity. *Hum. Mol. Genet.* 21, 3558–3567

145 Morales, F. *et al.* (2016) A polymorphism in the MSH3 mismatch repair gene is associated with the levels of somatic instability of the expanded CTG repeat in the blood DNA of myotonic dystrophy type 1 patients. *DNA Repair (Amst).* 40, 57–66

146 Tome, S. *et al.* (2013) MSH3 Polymorphisms and Protein Levels Affect CAG Repeat Instability in Huntington ' s Disease Mice. *PLoS Genet.* 9,

147 Lee, J.M. *et al.* (2015) Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. *Cell* 162, 516–526

148 Lee, J. *et al.* (2017) A modifier of Huntington ' s disease onset at the MLH1 locus. 26, 3859–3867

149 Fuchsberger, C. *et al.* (2016) The genetic architecture of type 2 diabetes. *Nature* 536, 41–47

150 Holmans, P.A. *et al.* (2017) Genetic modifiers of Mendelian disease: Huntington's disease and the trinucleotide repeat disorders. *Hum. Mol. Genet.* 26, R83–R90

151 Bettencourt, C. *et al.* (2016) DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. *Ann. Neurol.* 79, 983–990

152 Richards, S. *et al.* (2015) Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* 17, 405–424

153 Bhattacharyya, S. *et al.* (2002) Identification of RTG2 as a Modifier Gene for CTG • CAG Repeat Instability in Saccharomyces cerevisiae. *Genetics* 589, 579–589

154 Zhang, Y. *et al.* (2013) Genome-Wide Screen Reveals Replication Pathway for Quasi-Palindrome Fragility Dependent on Homologous Recombination. *PLoS Genet.* 9,

155 Pan, X. *et al.* (2012) Identification of novel genes involved in DNA damage response by screening a genome-wide Schizosaccharomyces pombe deletion library. *BMC Genomics* 13, 662

156 Saka, K. *et al.* (2016) More than 10% of yeast genes are related to genome stability and influence cellular senescence via rDNA maintenance. *Nucleic Acids Res.* 44, 4211–4221

157 Costanzo, M. *et al.* (2016) A global genetic interaction network maps a wiring diagram of cellular function. *Science (80-.).* 353,

158 Adzhubei, I. *et al.* (2013) *Predicting functional effect of human missense mutations using PolyPhen-2*,

159 Wang, T. *et al.* (2014) Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science (80-.).* 80, 80–85

160 Keogh, N. *et al.* (2017) MutSBeta abundance and Msh3 ATP hydrolysis activity are important drivers of CTG•CAG repeat expansions. *Nucleic Acids Res.* 45, 10068–10078

Chapter 2

A defective mRNA cleavage and polyadenylation complex facilitates expansions of transcribed $(GAA)_n$ repeats associated with Friedreich's ataxia

Ryan J. McGinty[1], Franco Puleo[2], Anna Y. Aksenova[1,3], Julia A. Hisey[1], Alexander A. Shishkin[1,4], Erika L. Pearson[2], Eric T. Wang[5,6], David E. Housman[5], Claire Moore[2] and Sergei M. Mirkin[1*]

[1] Department of Biology, Tufts University, Medford, MA 02421, USA

[2] Department of Developmental, Molecular & Chemical Biology, Tufts University School of Medicine, Boston, MA 02111, USA

[3] Laboratory of Amyloid Biology, St. Petersburg State University, St. Petersburg 199034, Russia

[4] The Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA

[5] The David H. Koch Institute for Integrative Cancer Research at MIT, Cambridge, MA 02139, USA

[6] Center for Neurogenetics, University of Florida, Gainesville, FL 32610, USA

[*] The author to whom all correspondence should be addressed.

E-mail: sergei.mirkin@tufts.edu

Cell Reports, Volume 20, Issue 10, 2017, Pages 2490-2500

doi: 10.1016/j.celrep.2017.08.051.

Summary

Expansions of microsatellite repeats are responsible for numerous hereditary diseases in humans, including myotonic dystrophy and Friedreich's ataxia. While the length of an expandable repeat is the main factor determining disease inheritance, recent data point to genomic *trans*-modifiers that can impact the likelihood of expansions and disease progression. Detection of these modifiers may lead to understanding and treating repeat expansion diseases. Here we describe a method for the rapid, genome-wide identification of *trans*-modifiers for repeat expansion in a yeast experimental system. Using this method, we found that missense mutations in the endoribonuclease subunit (Ysh1) of the mRNA cleavage and polyadenylation complex dramatically increase the rate of $(GAA)_n$ repeat expansions, but only when they are actively transcribed. These expansions correlate with slower transcription elongation caused by the *ysh1* mutation. These results reveal a previously unsuspected interplay between RNA processing and repeat-mediated genome instability, confirming the validity of our approach.

Introduction

Expansions of DNA microsatellites are responsible for several dozens of hereditary diseases in humans, including fragile X syndrome (FXS), myotonic dystrophy (DM1 and DM2), Huntington's disease (HD), Friedreich's ataxia (FRDA), many spinocerebellar ataxias (SCA), the familial form of amyotrophic lateral sclerosis and frontotemporal dementia (ALS), and others (Lopez Castel et al., 2010; McMurray, 2010; Mirkin, 2007). The scale of expansions differs depending on the location of the DNA repeat: they are relatively small-scale when positioned in the protein-coding part of a gene, or very large-scale when in the non-coding parts of a gene, such as 5'- and 3'-UTRs, or introns (Mirkin, 2007). Repeat expansions readily occur during intergenerational transmissions in human pedigrees, which accounts for the phenomenon of genetic anticipation that is characteristic for these diseases. In some somatic tissues, repeats continue expanding throughout life, which affects age of onset and disease severity (Kovtun and McMurray, 2008).

It is generally believed that the length of an expandable repeat is the key factor determining disease inheritance and development. Significant amounts of data, however, point to the existence of *trans*-modifiers that can affect the likelihood of repeat expansions, and thus, disease progression. While most of these data came from studying repeat expansions in model experimental systems (Usdin et al., 2015), the idea is also supported by human genetics data (Morales et al., 2012).

Expansions of $(CAG)_n$, $(CGG)_n$, $(GAA)_n$ and $(ATTCT)_n$ repeats have been extensively studied in yeast experimental systems. These studies revealed that knocking out genes involved in DNA replication, repair, recombination and transcription machineries can strongly elevate or decrease the rate of repeat expansions in dividing cells (Kim and Mirkin, 2013). Studies of $(CAG)_n$ repeat expansions in a *Drosophila* system showed that repeat instability was decreased when a fly homolog of the nucleotide excision repair gene XPG, *mus201* was mutated (Yu et al., 2011). Mice models for repeat expansions demonstrated the critical role of mismatch repair genes in promoting repeat expansions during both intergenerational transmission and in somatic cells (Kovtun and McMurray, 2001; McMurray, 2008; Savouret et al., 2003; Savouret et al., 2004). At the same time, mutations in the base excision repair machinery specifically prevented repeat expansions in somatic tissues (Kovtun et al., 2007). In a humanized mouse model of fragile X syndrome, the loss of the transcription-coupled DNA repair factor CSB led to a lower frequency of germ-line expansions and a reduction in the scale of somatic expansions (Zhao and Usdin, 2014). In cultured human cells, fork stabilizing proteins Claspin, Timeless, and Tipin were shown to counteract $(CAG)_n$ repeat expansions (Liu et al., 2012), while knockdown of the FANCJ protein resulted in the accumulation of DSBs and ectopic rearrangements at those repeats (Barthelemy et al., 2017). Finally, transcription-coupled repair was shown to trigger $(CAG)_n$ repeat contractions in human cells (Lin et al., 2010; Lin and Wilson, 2007).

Clinical genetics data, while more fragmentary and limited in scope, are generally in line with the conclusions of the model systems studies. In case of

DM1, it was found that the rate of $(CTG)_n$ repeat expansions in the DMPK gene is a heritable trait in itself, pointing to the existence of *trans*-modifiers throughout the genome (Morales et al., 2012). More recently, a polymorphism in the MSH3 mismatch repair gene was specifically associated with the extent of somatic instability of $(CTG)_n$ repeats in the blood of DM1 patients (Morales et al., 2016). Single nucleotide polymorphisms (SNPs) in genes involved in DNA replication, repair and recombination have been associated with increased risk of repeat expansions in Huntington's disease (HD) and spinocerebellar ataxia type 3 (SCA3) families (Genetic Modifiers of Huntington's Disease, 2015; Martins et al., 2014). It was also suggested that differential expression levels for replication and repair genes in various parts of the HD patient brains might determine the extent of somatic instability in the corresponding brain regions (Mason et al., 2014).

There exists, however, a serious gap between the model systems and human genetics data. The former primarily describe the effect of gene knockouts, *i.e.* an all-or-none scenario, while the latter deal with SNPs, *i.e.* much more subtle changes in gene functioning. In this study, we attempted to fill this gap by conducting a genetic screening to detect *trans*-modifiers of repeat expansions in our yeast experimental system (Shah et al., 2012; Shishkin et al., 2009). The screening strategy involves mutagenesis and selection for repeat expansions, followed by whole-genome sequencing and identification of causal SNPs in the expansion process. Totally unexpectedly, this screening revealed mutations in *YSH1*, a gene central for RNA processing.

YSH1 encodes a component of the cleavage and polyadenylation specificity factor complex (CPSF or CPF), which in concert with cleavage stimulation factor (CstF or CFIA) and cleavage factor I (CFI or CFIB) cleaves mRNA transcript at poly(A) signals (Chan et al., 2011; Millevoi and Vagner, 2010). This subsequently allows both the addition of the poly-A tail to the 3' end of the mRNA via Pap1 (poly-A polymerase), as well as the loading of Rat1 exonuclease to the 5' end of the transcript, leading to the transcription termination (Porrua et al., 2016).

We found that in Ysh1 mutants that came from our screen, inefficient transcript cleavage is accompanied by slowed transcription elongation and accumulation of double stranded breaks within transcribed $(GAA)_n$ repeats followed by their expansions in a homologous recombination (HR)-dependent manner. These results reveal a totally unsuspected interplay between RNA processing and repeat-mediated genome instability, hence confirming the validity of our whole-genome screening approach. In the future, this approach can be used to identify trans-factors for large-scale expansions of other repeats, such as $(CGG)_n$ repeats responsible for fragile X syndrome or $(CTG)_n$ repeats responsible for myotonic dystrophy type 1.

RESULTS

Screen Design and Implementation

In yeast, large-scale repeat expansions are rare events as opposed to repeat contractions (Kim and Mirkin, 2013). To detect rare large-scale expansion events, we have developed an experimental system, in which expansions of the

$(GAA)_{100}$ repeat within an artificial intron of the *URA3* gene (Fig. 1A) inhibits its splicing, resulting in yeast growth on 5-FOA-containing media (Shah et al., 2012; Shishkin et al., 2009). Crossing this reporter cassette into various yeast knockout libraries helped us to identify numerous genes involved in DNA replication, repair and transcription that affect repeat expansions (Zhang et al., 2012).

We were concerned, however, that gene knockouts are too blunt of a tool, particularly when it comes to essential genes, and, thus, wanted to assess the effect of subtler genetic changes on repeat expansions. To this end, we chose mild UV mutagenesis to induce point substitutions as opposed to gene deletions or gross-chromosomal rearrangements. While we expected this approach to generate point mutations in genes affecting repeat expansion in our system, we were acutely aware that it might also lead to the accumulation of mutations in the body of our reporter, or in other proteins involved in uracil biosynthesis. To minimize the latter prospect, we added another selectable cassette to make our screening a two-stage process. The second cassette, which contained the *ADE2* reporter with a $(GAA)_{100}$ repeat within its artificial intron, replaced the endogenous *ADE2* gene on chromosome XV (Fig. 1A). Unexpectedly, however, the presence of even the starting length repeat within this intron completely inactivated the *ADE2* gene, making yeast colonies red. Thus, we shortened the $(GAA)_n$ run down to 63 repeats (Fig. 1A) to keep the reporter active. Notably, the presence of 63 repeats in the *ADE2* intron already decreased the reporter's expression sufficiently that the resultant strain had a borderline ADE+ phenotype and pink colonies (see also below).

The strain carrying both selectable cassettes was irradiated with UV light followed by a two-step selection protocol (Fig. 1B): identification of red colonies (step 1 - *ADE2* inactivation), which were then analyzed individually for 5-FOA-resistance (step 2 - *URA3* inactivation). Nearly half of mutagenized red colonies gave rise to augmented papillae growth on 5-FOA-containing media. For roughly half of them, PCR analysis of those 5-FOA-resistant colonies revealed large-scale expansions in the *URA3* cassette. Unexpectedly, however, repeat expansions in the *ADE2* cassette were not detected in any of them (Fig. 1C). As shown later, the *ADE2* inactivation is likely due to reduced expression of the *ADE2* mRNA in these mutants. In summary, our screen revealed new genetic trans-modifiers that repress the reporter gene carrying a short $(GAA)_{63}$ repeat, while simultaneously promoting expansions of longer $(GAA)_{100}$ repeats.

Identification of mutations in the *YSH1* gene

We conducted whole-genome sequencing of sixteen UV-mutagenized strains that simultaneously showed *ADE2* inactivation and high rate of repeat expansions in the *URA3* gene. In brief, genomic DNA was isolated from these strains, barcoded libraries were generated and sequenced using Illumina GAII with 100 bp Paired-End reads. This gave an average coverage of ~80x per strain. Reads were then aligned to the S288C reference genome using Bowtie (Langmead et al., 2009), and mutant variants were called using the SAMtools software (Li et al., 2009). This analysis revealed that our mutagenesis strategy resulted in the accumulation of ~10 mutations per yeast strain.

To assess which of these multiple mutations could potentially be causative, they were further analyzed using snpEFF (Cingolani et al., 2012) and PolyPhen2 (Adzhubei et al., 2013) tools. Remarkably, two out of sixteen sequenced strains contained missense mutations in the same essential gene, *YSH1*, which encodes a cleavage and polyadenylation factor subunit (Garas et al., 2008; Zhao et al., 1997). Furthermore, these mutations (*ysh1-L439S* and *ysh1-L14S)* affected highly conserved amino acids. The yeast L14S and L439S substitutions correspond to L17S and L427S in the human cleavage and polyadenylation factor CPSF-73 (Chan et al., 2011; Millevoi and Vagner, 2010). Both mutations are outside of the enzyme's catalytic center (Fig. S4). The L14 residue appears to reside on the surface of the protein and could potentially affect the stability of the CPF complex, while the L439 residue resides internally, but not in the active site.

Since two independent mutational hits appeared in conserved parts of the *YSH1* gene, we hypothesized that these mutations could be causative for the observed phenotype of increased repeat expansions and gene inactivation. To validate this hypothesis, we made two strains containing individual *ysh1-L439S* and *ysh1-L14S* mutations (see Methods) along with the two repeat-containing cassettes.

<u>Characterization of the *YSH1* mutant strains</u>

We first looked at the growth characteristics of the strains with individual *ysh1-L439S* and *ysh1-L14S* mutations. These strains readily turned red, indicating that the *ysh1* mutations are indeed responsible for inactivating the

ADE2 cassette. Both mutants grew more slowly than the wildtype, and this slow growth was exacerbated at higher temperatures and rescued at lower temperatures. The *ysh1-L14S* mutant appeared to be the stronger of the two mutants in each test we conducted, and it had a clear-cut temperature-sensitive growth phenotype (Fig. 2). Consequently, this mutant was chosen for all further analyses.

Ysh1 is the endonuclease responsible for cleavage of the nascent mRNA transcript during 3' end processing (Mandel et al., 2006; Ryan et al., 2004). It has no other known enzymatic functions, though its presence in the CPF complex facilitates related processes, including polyadenylation and splicing (Chanfreau et al., 1996; Garas et al., 2008; Zhao et al., 1999a). Ysh1 mutants were shown to be defective in both cleavage and polyadenylation *in vitro* (Chanfreau et al., 1996; Garas et al., 2008; Zhao et al., 1999a). Therefore, we performed *in vitro* cleavage and polyadenylation assays for the wild-type and *ysh1-L14S* mutant as described (Zhao et al., 1999b). In brief, cell extracts from both strains were incubated with the full-length ^{32}P-labeled GAL7-1 RNA in the presence of ATP, and the reaction products were separated on a denaturing polyacrylamide gel and visualized via phosphorimager (Fig. 3A). The *ysh1-L14S* mutation causes a strong decrease in the efficiency of RNA cleavage and polyadenylation at the non-permissive temperature. The individual steps of cleavage and poly(A) addition are also compromised in the mutant when uncoupled from each other (Fig. S1).

It was previously reported that mutants defective in the CF IA cleavage/polyadenylation factor are characterized by a slower rate of transcription elongation (Tous et al., 2011), but such effects from mutation of CPF, the complex in which Ysh1 resides, have not been reported. We were curious whether the same is true for the *ysh1-L14S* mutant. To address this question, we studied its sensitivity to mycophenolic acid (MPA), an inhibitor of inosine monophosphate dehydrogenase (IMPDH), which catalyzes the first committed step in GMP biosynthesis. Transcription elongation mutants are hypersensitive to the depletion of GTP pools in the presence of MPA (Desmoucelles et al., 2002). Fig. 3B shows that the *ysh1-L14S* strain is hypersensitive to MPA as compared to the wild-type strain.

Since inactivation of the *ADE2* cassette in the *ysh1-L14S* mutant was not caused by the repeat expansions in its intron, we sought to determine to what extent 3'-end-processing defects of *ysh1-L14S* affect its expression. We first compared the steady-state levels of mRNA for the normal and split *ADE2* gene in the wild-type and mutant strain using RT-qPCR. Owing to the concern that a polyadenylation mutant might affect any transcript used for normalization, we extracted DNA and RNA in parallel from an equal volume of cells, which allowed us to normalize RT-qPCR products to the total DNA. Fig. 4A shows that the presence of a repeat-bearing intron within the *ADE2* gene decreased its expression 6-fold compared to the intron-less gene even in the wild-type strain. This result explains the border-line ADE+ phenotype in our starting strain used for mutagenesis. In the *ysh1-L14S* mutant, we observe an additional drop in the mRNA level in the selectable *ADE2* cassette ranging from 2-fold at 30^{o}C to 5-fold at 37^{o}C.

We then analyzed the usage of the main *ADE2* poly(A) site in the wild type and mutant strain using RT-qPCR analysis with primers upstream or downstream from this site (Fig. 4B). Read-through of the poly(A) site is drastically increased in the *ysh1-L14S* mutant, reaching ~20-fold more than WT at 37ºC. We conclude, therefore, that the *ysh1-L14S* mutant is also defective for mRNA 3' end processing *in vivo* and cells with this mutation likely turn red due to decreased production of polyadenylated *ADE2* mRNA.

In contrast to the *ADE2* cassette, the *ysh1-L14S* mutation did not decrease the RNA level for the repeat-bearing *URA3* cassette (Fig. S2A). While we do not know why the *URA3* cassette behaves differently from the *ADE2* cassette, our preliminary data are indicative of a peculiar interplay between slower transcription elongation (see Fig. 6A below) and higher splicing efficiency of the long repeat-containing intron in the URA3 cassette (Fig. S2), similarly to what was discussed in (Moehle et al., 2014). Whatever the reason, the lack of *URA3* repression necessitated that 5-FOA-resistant clones originating in the *ysh1-L14S* mutant background arose as a result of expansions of the $(GAA)_{100}$ repeat.

Effects of the *ysh1-L14S* mutation on $(GAA)_n$ repeat expansions

To study the effects of *ysh1-L14S* mutation on repeat instability, we first compared the expansion rates for the $(GAA)_{100}$ repeat within the *URA3* cassette (Fig. 5A) in the wild-type and mutant strain using the fluctuation test approach conducted as described previously (Shah et al., 2012). The results shown in Fig. 5B show that even at the semi-permissive temperature (30ºC), *ysh1-L14S* mutation elevates the expansion rate ~4-fold, while in cells pre-grown at 37°C, it was up ~10-fold as compared to the WT.

A mutation in the cleavage and polyadenylation factor complex likely affects expression of numerous yeast genes. We were concerned, therefore, whether its effect on repeat expansions could be mediated by a change in expression of a gene(s) involved in repeat expansions. If this were the case, one would expect *ysh1-L14S* to affect expansions of both transcribed and non-transcribed repeats to a similar extent.

To distinguish between these possibilities, we studied the influence of the *ysh1-L14S* mutation on expansions of $(GAA)_n$ repeats within a different selection cassette, in which they are located between the galactose promoter and its upstream activating sequence (UAS_{GAL}) (Fig. 5C), a region that is practically non-transcribed. Large-scale repeat expansions shut off transcription of the *CAN1* reporter, which results in the appearance of canavanine-resistant colonies (Shah et al., 2014). Fig. 5D shows that, in contrast to transcribed repeats, *ysh1-L14S* mutation has no effect on the expansion of the non-transcribed $(GAA)_{100}$ repeat at either at 30°C or 37°C. We conclude, therefore, that transcription is required for expansion of the repeat in the *ysh1* mutant. Furthermore, the mutation is probably not affecting activity of a protein that directly represses repeat expansion. There remains the possibility that the mutation of Ysh1 affects the expression of a gene that promotes expansions solely within transcribed regions. However, the results below suggest a direct role for Ysh1.

Ysh1p plays a critical role in co-transcriptional 3' end formation and in

RNA polymerase II (RNAP II) transcription termination (Garas et al., 2008; Schaughency et al., 2014). In addition, the CPF factor in which Ysh1 resides is affiliated with actively transcribed chromatin (Kim et al., 2004) and the *ysh1-L14S* mutant is sensitive to the MPA inhibitor of elongation (Fig. 3B). These observations raise the possibility that transcription of the $(GAA)_n$ repeats might be important for expansion induced by the *ysh1* mutation.

Given that *ysh1-L14S* mutation specifically elevates instability of transcribed DNA repeats, we next compared transcription elongation through the *URA3* cassette in this mutant compared to the wildtype strain using an RNA polymerase clearance assay (Mason and Struhl, 2005). To this end, we replaced the *URA3* promoter in our selectable cassette with the inducible *GAL1-10* promoter. To analyze the transcription elongation rate, cells were grown in the presence of galactose, transcription was shut down by the addition of glucose, and RNAP II distribution along the body of the cassette was measured by ChIP. Fig. 6A shows the normalized (glucose/galactose) values for Pol II occupancy, *i.e.* the fraction of Pol II, which failed to clear the cassette following glucose repression. One can see that only 20% of RNA Pol II remains associated with promoter-distal parts of the *URA3* cassette in the wild-type strain, which is indicative of a robust transcription elongation and efficient cassette clearance. In mutant cells, in contrast, the clearance rate appears to be much slower: up to 50% of all RNAP II remain bound to the cassette after glucose repression. Elongation defects were also observed in the *ysh1-L14S* mutant for the *YLR454*, *GAL10* and *GAL1* genes that do not have $(GAA)_n$ repeats (Fig. S3). We conclude, therefore, that transcription elongation rate is strongly decreased in the *ysh1-L14S* mutant.

Slow transcription elongation is known to stimulate R-loop formation at various sequences, including $(GAA)_n$ repeats (Butler and Napierala, 2015; Groh et al., 2014). It was foreseeable, therefore, that increased R-loop formation at $(GAA)_n$ repeats in the *ysh1-L14S* mutant could ultimately promote repeat expansions. RNase H is known to efficiently resolve R-loops by hydrolyzing their RNA component (Hamperl and Cimprich, 2014). Thus, to evaluate a possible role of R-loop formation in our mutant, we knocked out both RNase H1 and RNase H2 in the *ysh1-L14S* strain and measured the rate of $(GAA)_n$ repeat expansions in the *URA3* selectable cassette described above (Fig. 5A). We found that double RNase H knockout has no effect on the rate of repeat expansions in the *ysh1-L14S* mutant (Fig. 6B). Our alternative approach was to overexpress RNase H1, which is known to counteract R-loop formation *in vivo* (Wahba et al., 2011). We first introduced the plasmid overexpressing human RNase H1 described in (Wahba et al., 2011) into our *ysh1-L14S* strain followed by measuring $(GAA)_n$ repeat instability. It appeared that RNase H1 overexpression had little if any effect on the repeat expansion rates. The caveat of these experiments, however, was that strains carrying the RNase H1-expressing plasmid appeared to be fairly sick. Thus, we used a different approach based on the regulation of RNase H1 expression under the control of the inducible *MET25* promoter (Janke et al., 2004). To this end, the promoter of the endogenous *RNH1* gene in our *ysh1-L14S* strain was replaced with the *MET25* promoter as described in the Supplemental Experimental Methods. Fig. 6B shows that the rate of repeat expansions was quantitatively the same whether the expression of RNase H1 was low (in the presence of methionine) or high (in the absence of methionine). Altogether, we conclude that the elevated expansion rate of transcribed $(GAA)_n$ repeats in the *ysh1* mutant is unlikely to be caused by R-loop formation.

We have previously shown that $(GAA)_n$ repeats cause chromosomal fragility in yeast (Kim et al., 2008). Compromised transcription elongation is also known to promote the formation of double-strand breaks (Dutta et al., 2011; Nudler, 2012). It is foreseeable, therefore, that slow transcription through the repeat in the *ysh1-L14S* mutant could result in the formation of double-strand breaks, ultimately resulting in expansions. To test this hypothesis, we moved our *URA3* selectable cassette to the non-essential arm of chromosome V, centromere-proximal to the endogenous *CAN1* marker gene (Chen and Kolodner, 1999). In this setting, breakage at the $(GAA)_n$ repeats could lead to a loss of the whole chromosomal arm containing both *CAN1* and the *URA3* reporters, - an event which is easily detectable on selective media containing canavanine and 5-FOA. Fig. 7A shows that the rate of arm loss is indeed significantly elevated in the *ysh1-L14S* mutant at 37°C. Thus, *ysh1-L14S* mutation indeed promotes breakage of the $(GAA)_n$ repeat.

In yeast, double-strand breaks are preferably repaired via homologous recombination (HR). Misalignment of the repetitive runs in the process of recombination could ultimately result in repeat expansions (Kim et al., 2017). Thus, we decided to assess the role of the key HR proteins, Rad51 and Rad52 (Symington, 2002) on repeat expansions in the *ysh1-L14S* genetic background. To this end, we compared repeat expansions between a double *ysh1-L14S*, *rad52*Δ mutant and a single *rad52*Δ mutant, as well a double *ysh1-L14S*, *rad51*Δ mutant and a single *rad51*Δ mutant. Since the *ysh1-L14S*, *rad52*Δ double mutant grew very slowly at 37°C, we were only able to generate reliable expansion data at the semi-permissive temperature (Fig. 7B). Clearly, knocking down Rad52 brings the rate of repeat expansions in L14S mutant down to the wild-type level. In contrast to *rad52*Δ, knocking out *rad51*Δ did not affect the rate of repeat expansions in the WT or *ysh1-L14S* genetic backgrounds (Fig. 7B). We believe, therefore, that a Rad51-independent sub-pathway of homologous recombination for DSB-repair might be responsible for the elevated rate of $(GAA)_n$ repeat expansions in the *ysh1* mutant.

Discussion

Our screen revealed an unanticipated connection between RNA cleavage/polyadenylation and large-scale expansions of triplet DNA repeats in *S. cerevisiae*. The mechanisms responsible for this link are intriguing, as repeat expansions occur in the course of DNA, rather than RNA synthesis. That being said, there exists a substantial literature showing that transcription elevates triplet repeat instability. To give just a few examples: Transcription of $(CAG)_n$ repeats increased their instability in cultured human cells in a transcription-coupled repair dependent manner (Lin et al., 2009; Lin and Wilson, 2007). Changes in the chromatin structure during repeat transcription were also shown to promote expansions by making repeats more susceptible to inherent and external damage

(Debacker et al., 2012; House et al., 2014; Shah et al., 2014; Yang and Freudenreich, 2010). Additionally, a number of studies implicated R-loops in triplet repeat instability. R-loops detected at triplet repeats (Groh and Gromak, 2014; Groh et al., 2014) were proposed to account for transcription-mediated repeat instability (Lin et al., 2010; Reddy et al., 2014; Reddy et al., 2011). Similarly, R-loop formation and transcription-coupled repair protein ERCC6/CSB were implicated in CGG repeat expansions in a mouse model of the fragile X-syndrome (Zhao and Usdin, 2014). None of these studies, however, investigated the role of co-transcriptional RNA processing.

In a separate development, recent genetic and molecular analyses began to identify RNA-binding proteins (RBPs) as important players in maintaining genome stability by preventing accumulation of harmful RNA/DNA hybrids and by regulating the DNA damage response (DDR) (Dutertre et al., 2014). In *S. cerevisiae*, seven essential subunits of the mRNA cleavage and polyadenylation machinery were implicated in DDR triggered by R loops (Stirling et al., 2012). Knockout of the *TRF4* gene, encoding a non-canonical polyA-polymerase involved in RNA surveillance, gave rise to a transcription-associated recombination phenotype (Gavalda et al., 2013). Cleavage Factor I was shown to contribute to genome integrity by preventing replication hindrance (Gaillard and Aguilera, 2014). Similarly, *S. pombe* cleavage and polyadenylation factor Rna14 was implicated in the maintenance of genomic integrity (Sonkar et al., 2016). None of these studies, however, looked at triplet repeat expansions and/or fragility.

Contrary to the above examples, we found that RNA/DNA hybrids are not likely to be involved in elevated repeat instability in the *ysh1-L14S* mutant background (Fig. 6B). This difference may be due to the unique role that Ysh1 protein plays in RNA processing. Aguilera's group has proposed that mutations in RNA binding proteins lead to their absence from the nascent RNA during transcription, which, in turn, allows this naked RNA to stably pair with its DNA template (Dominguez-Sanchez et al., 2011). We don't think that mutations in *Ysh1* protein would result in the presence of naked RNA during transcription, as other members of the CPSF complex are still expected to be bound to RNA. At the same time, we have demonstrated that mutations in the *Ysh1* protein significantly slow down RNA polymerase progression (Fig. 6A), likely because it remains bound to the transcript, but cannot cleave efficiently. It was also demonstrated by others that transient depletion of Ysh1p triggers transcriptional pausing downstream of known polyadenylation sites (Schaughency et al., 2014).

Our working model combines the above observations with the data from this study. A mutation in the Ysh1 protein, which was isolated from our repeat expansion screen, cause defects in transcript cleavage and polyadenylation (Fig. 3). As this process occurs co-transcriptionally, we reasoned that the entire RNA Pol II elongating complex may slow or stall at potential poly(A) sites on the DNA template when Ysh1 is not efficient. In our mutants, transcription elongation is significantly slowed down across the whole *URA3* cassette (Fig. 6A). Transcription stalling and backsliding is known to trigger the formation of double-stranded breaks in DNA, owing to their collisions with replication machinery or other mechanisms (Mirkin et al., 2006; Nudler, 2012). We do see elevated fragility of the $(GAA)_n$ run in the Ysh1 mutant, which is consistent with DSB formation. When homologous recombination machinery attempts to repair the broken DNA ends, repetitive DNA strands can align out of register, resulting in repeat expansions after the next round of replication (Fig. 7C). Supporting this reasoning, the increase in repeat expansions in the *ysh1-L14S* mutant fades when homologous recombination is completely shut down in the *RAD52* knockout. At the same time, repeat expansions in the *ysh1-L14S* mutant were not diminished in the *RAD51* knockout, indicating that a Rad51-independent sub-pathway of HR is either responsible for the expansions, or can compensate in the absence of canonical Rad51-dependent HR. One possibility is the involvement of the single-strand annealing pathway (SSA), which is known to act within repetitive regions and is not dependent on Rad51 protein (Downing et al., 2008). Another possibility is a Rad51-independent wing of the break-induced replication pathway (Ira and Haber, 2002).

While our studies were performed in *S. cerevisiae*, they may have implications for Friedrich's ataxia in humans. An interesting repercussion from the transcription repeat breakage model is that expansions may pre-nucleate outside of the S-phase. This phenomenon might therefore shed light on how repeat expansions can occur in non-dividing neural and cardiac cells (McMurray, 2010). It would be of prime interest to investigate whether Friedreich's ataxia patients carrying mutations in the *YSH1* homolog *CPSF-73* or other RNA processing genes might be at higher risk for repeat expansions, accounting for the variation in disease severity and age of onset between different individuals. Even in the absence of germline mutations in CPSF complex, transcription may already proceed more slowly through $(GAA)_n$ repeats (Krasilnikova et al., 2007**.** Transcriptional blocks at the $(GAA)_n$ repeat within the FRDA locus could become prominent in specific cell lineages or arise transiently to produce large-scale expansions in non-dividing cells. This can hint at a potential therapy, if it becomes possible to prevent RNA polymerase stalling at the repeat (Gottesfeld et al., 2013; Soragni et al., 2014). Reducing transcription pausing at $(GAA)_n$ repeats may both reduce DNA breakage and rescue the poorly expressed mutant allele of the *FXN* gene.

Experimental Procedures

Yeast strain construction. The list of our strains is presented in Table S1. See Supplemental Experimental Methods for further details.

Fluctuation assays. Fluctuation assays were performed as previously described (Shah et al., 2014; Shishkin et al., 2009). See Supplemental Experimental Methods for further details.

In Vitro 3' End processing. Processing extracts were prepared as described (Zhao et al., 1999) using strains SMY732 and RMG89, which were grown at 30ºC and then shifted to 37ºC for 1.5h. Extracts were incubated with ATP and full-length or pre-cleaved ^{32}P-labeled GAL7-1 RNA. Reaction products were run on a polyacrylimide urea gel and visualized via phosphoimager.

Quantitative RNA analysis. RNA levels were measured via qRT-PCR, employing a strategy wherein gDNA was extracted from an equal portion of the yeast culture used to extract RNA. See Supplemental Experimental Methods for further details.

RNA Pol II elongation assays. Assays were performed as previously described (Mason and Struhl, 2005). See Supplemental Experimental Methods for further details.

Author contributions

Conceptualization, R.J.M., A.Y.A., E.T.W. and S.M.M. Software, R.J.M. Investigation, R.J.M., F.P., A.Y.A, J.A.H., E.L.P. and E.T.W. Resources, A.A.S. Writing – Original Draft, R.J.M. and S.M.M. Writing – Review and Editing, R.J.M., C.M. and S.M.M. Visualization, R.J.M. and F.P. Supervision, D.E.H., C.M. and S.M.M. Funding acquisition, A.Y.A., S.M.M., C.M.

Acknowledgments

We thank Kartik Shah for sharing his experience with the *CAN1* selectable cassette, Allen Su for sharing his experience with RNaseH1 overexpression and Doug Koshland for plasmids overexpressing RNase H1. Research in the lab of S.M.M. is supported by NIH grants R01GM60987 and P01GM105473 and generous contribution from the White family. Research in the lab of C.M. was supported by grant MCB-1244043 from NSF and an NIGMS Post-Baccalaureate Research Education Program Award 2R25GM066567. Anna Aksenova was supported by RFBR grant #15-04-08658 and research project in the Centre for Molecular and Cell Technologies (Research Park, Saint-Petersburg State University, Russia).

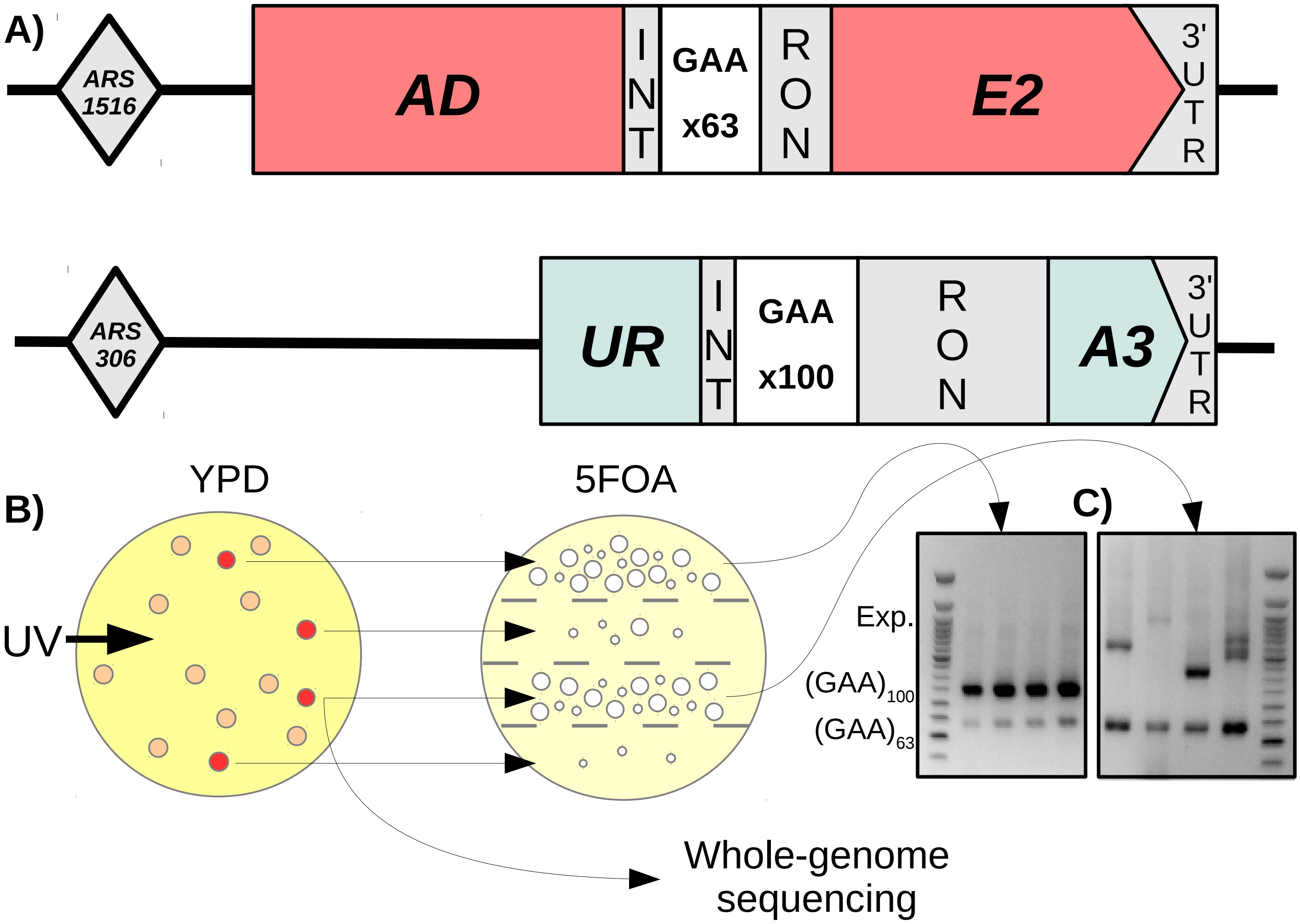


Fig. 1. Overview of screening method.
Top Panel: Diagram of selectable *ADE2* and *URA3* cassettes. The *ADE2* marker contains a short artificial intron with only 63 (GAA) repeats, while the *URA3* marker contains a longer artificial intron with 100 (GAA) repeats.
Middle Panel: Screening procedure: Cells are mutagenized and grown on complete (YPD) media. Colonies form, and those that turn red (*ADE2* inactivation) are spread on sections of a plate containing the selective drug 5FOA. For each strain with a high number of 5FOA-resistant colonies (URA3 inactivation), four individual 5FOA colonies were tested via PCR for repeat length.
Bottom Panel: Example PCRs for amplification of $(GAA)_n$ repeats in both cassettes. The *URA3* $(GAA)_{100}$ repeat consistently expands in strains containing genuine repeat expansion trans-modifiers (right), while remaining at wild-type length in strains containing off-target modifiers (left). The *ADE2* $(GAA)_{63}$ repeat does not appear to expand in any strains.

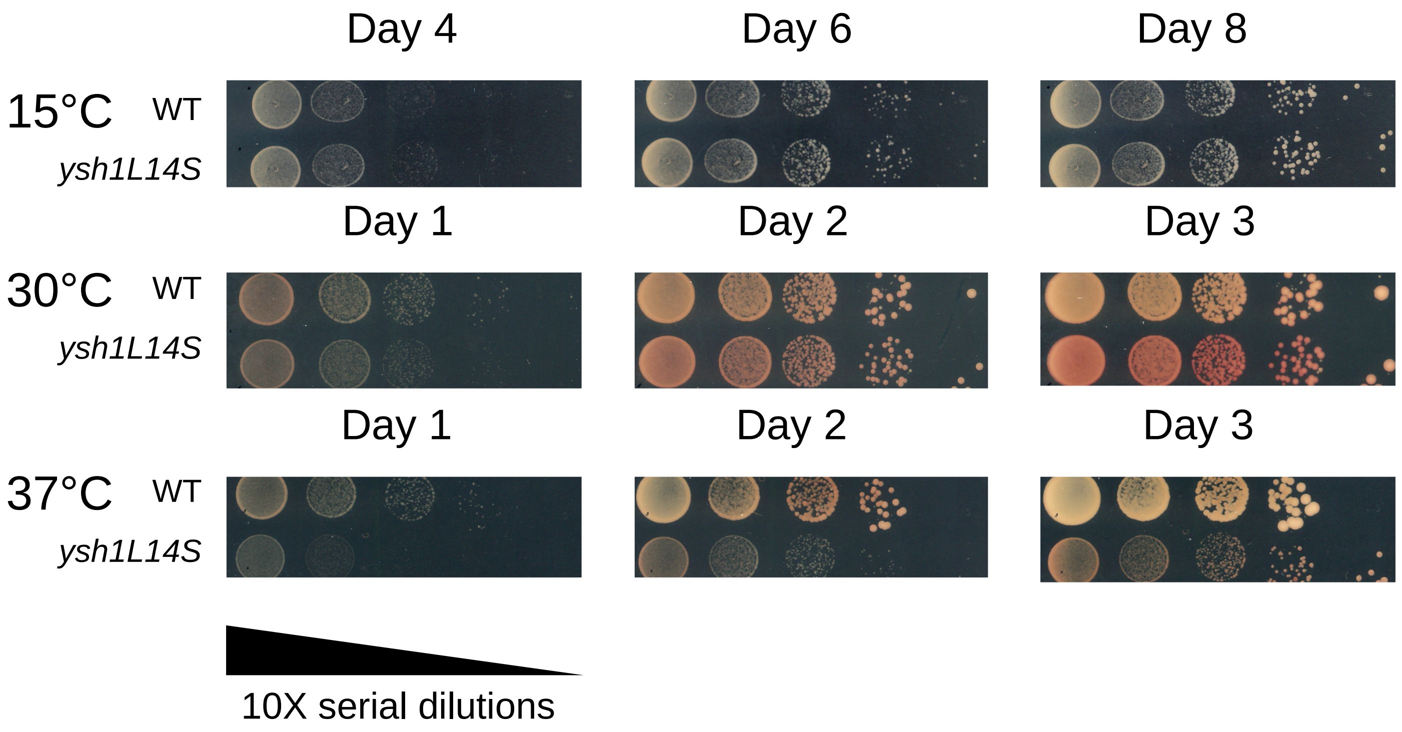


Fig. 2. Mutant *ysh1-L14S* is temperature sensitive for growth.
WT and *ysh1-L14S* mutant strains were serially diluted and grown on complete media at three different temperatures. No difference in growth rate is observable at 15°C, which is below the optimal temperature for wild type yeast. At the optimal temperature of 30°C, the *ysh1* mutant displays slightly reduced growth, best observable after 1 day of growth. Red pigment is observable after 3 days, due to inactivation of the *ADE2* cassette. Incubation at 37°C severely slows the growth of the *ysh1* mutant.

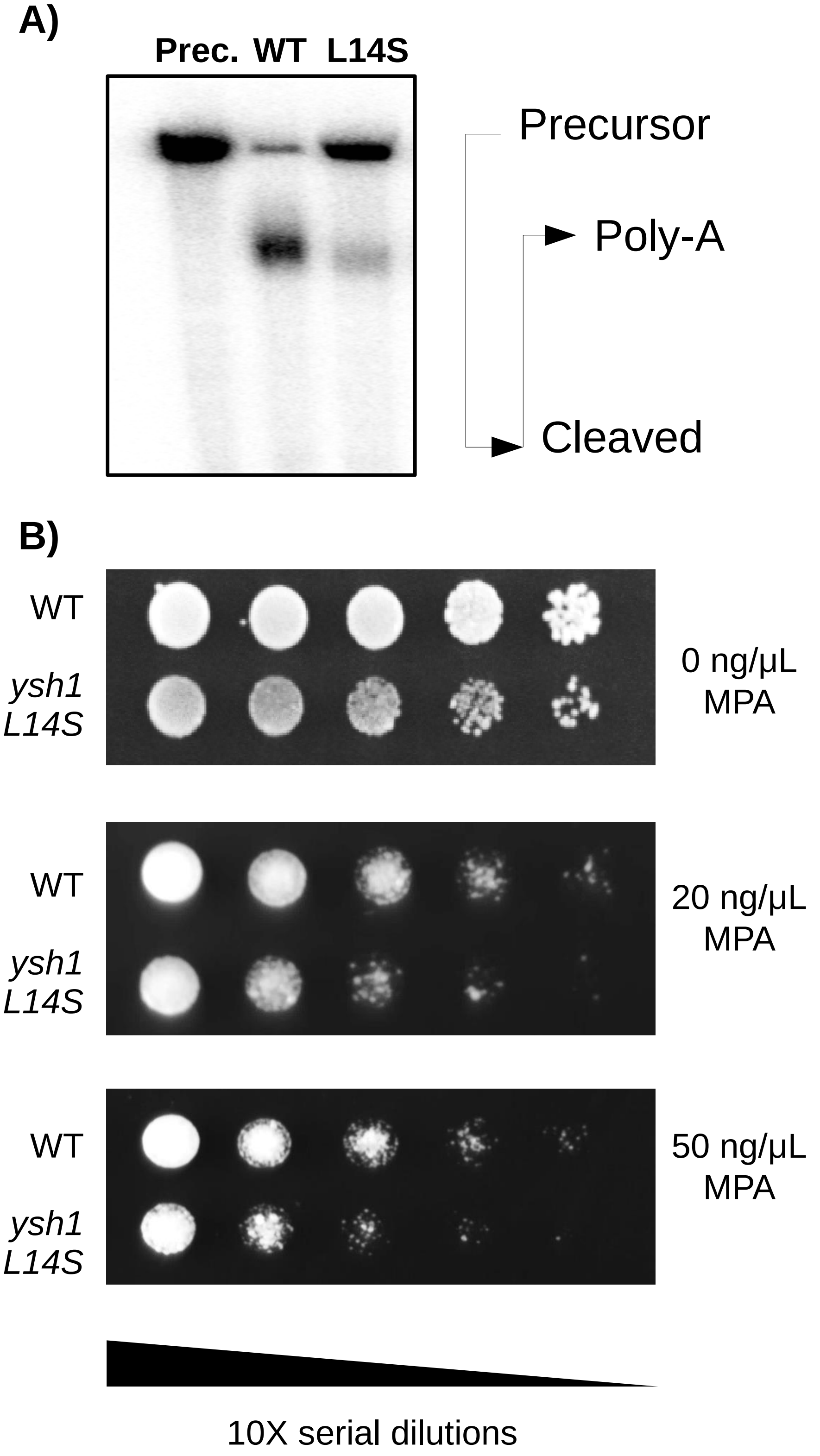


Fig. 3. The *ysh1-L14S* mutant is defective for mRNA 3’ end-processing and transcription elongation.
A) In vitro 3’ end processing reaction. A precursor RNA is combined with cell extracts derived from WT or *ysh-L14S* mutant yeast, which were grown at 30^{o}C and shifted to 37^{o}C for 1.5 hours. The precursor RNA is shortened by the cleavage reaction, and then lengthened by the addition of the poly-A tail. (Positions indicated.) For *ysh1-L14S* mutants, less of the precursor RNA is converted to the polyadenylated form. Cleaved products are not observed, suggesting that the cleavage step is rate-limiting. See also Figure S1.
B) *ysh1-L14S* is sensitive to the transcription elongation inhibitor mycophenolic acid (MPA). WT and *ysh1* L14S mutant strains were serially diluted and grown on synthetic media lacking uracil and containing the indicated MPA concentrations. Plates were incubated for 3 days at 30°C. *ysh1-L14S* strains display a pronounced growth inhibition under MPA treatment.

Figure contributed by Franco Puleo.

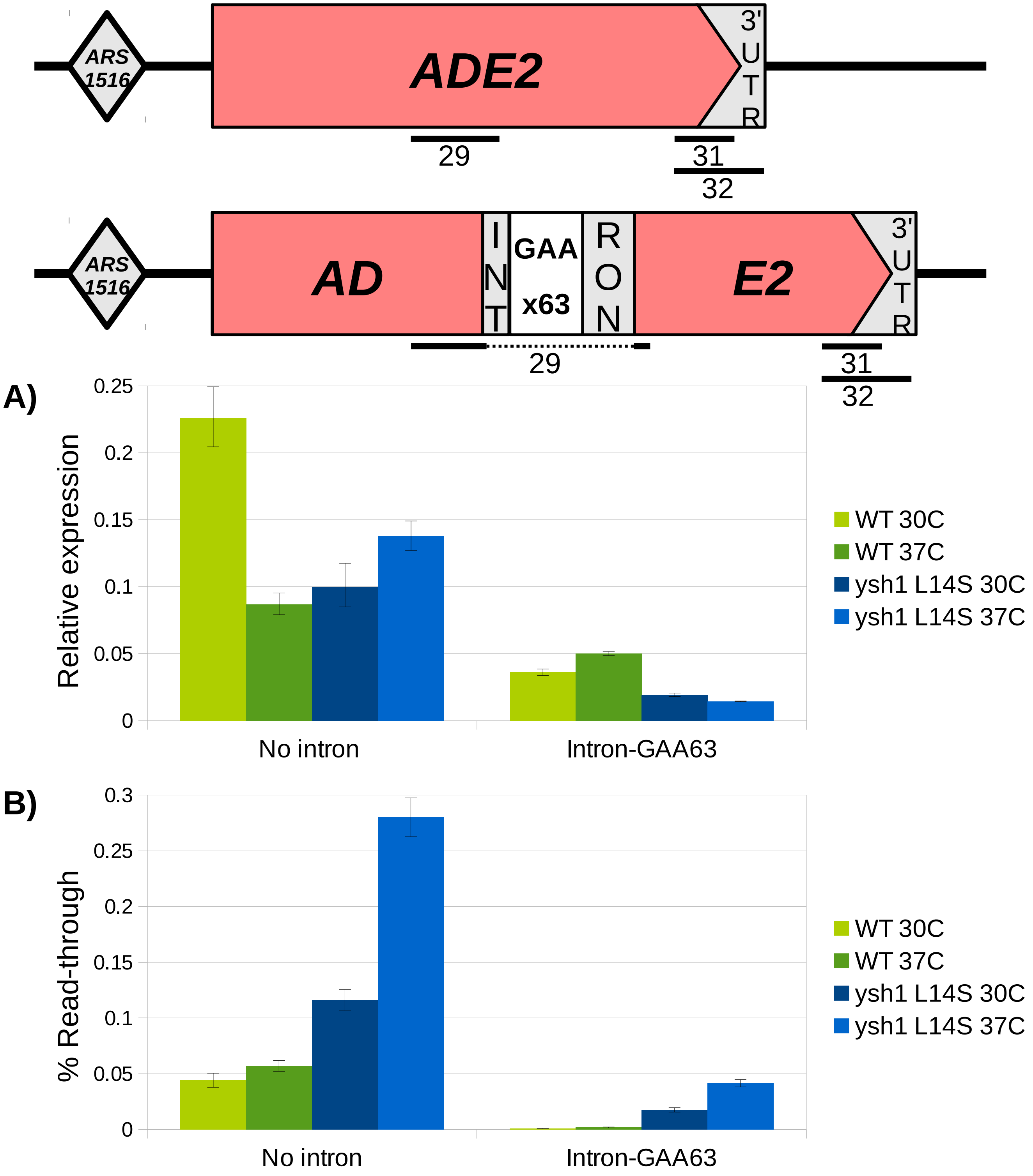


Fig. 4. RNA analysis of *ADE2* cassette transcripts.
Top panel: Diagram of the *ADE2* gene and *ADE2-(GAA)*$_{63}$ cassette, indicating the position of primer pairs used for RNA analysis.
A) Results of RT-qPCR using primer pair #29, which is specific to spliced mRNA in the split *ADE2* cassette. Comparing the two versions of *ADE2*, the presence of the intron reduces mRNA expression in both the WT and *ysh1-L14S* mutant background. With the split *ADE2* cassette, the *ysh1-L14S* mutant shows decreased levels of spliced *ADE2* mRNA at both 30°C and 37°C. Reverse transcription was performed using oligo-dT primers. Error bars represent the SD of qPCR technical replicates. See also Figure S2.
B) Calculation of read-through transcription levels based on qRTPCR using primer pairs before (pair #31) and after the annotated poly-A site (pair #32). In both versions of the *ADE2* gene, *ysh1-L14S* mutants show increased levels of read-through at 30°C, with a further increase at 37°C. Reverse transcription was performed using random hexamer primers. Error bars represent the SD of four qPCR technical replicates.

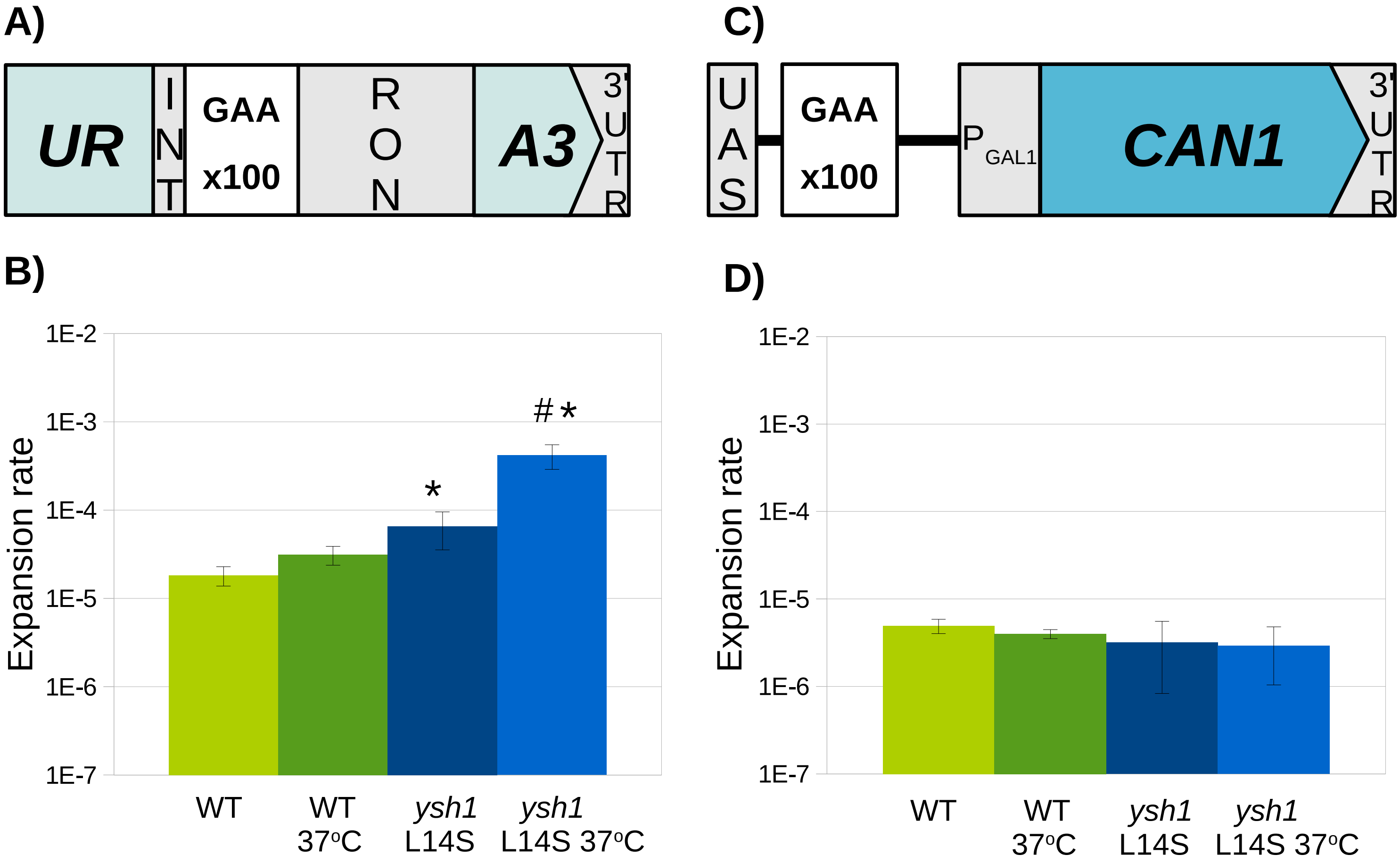


Fig. 5. Mutation in *YSH1* gene increases expansions of transcribed (GAA)$_n$ repeats.

A) Selective system to assess large-scale (GAA)$_n$ repeat expansions in a transcribed setting. Repeats are placed within an artificial intron in the *URA3* counterselectable marker. This distance is length constrained, with an expansion inhibiting splicing of the intron. Fluctuation tests were performed to determine the large-scale (GAA)$_n$ expansion rate for WT and *ysh1-L14S* mutant strains.

B) The *ysh1-L14S* mutant shows increased rates of repeat expansion, which increase further under temperature-sensitive conditions. (left side of graph).

C) Selective system to assess expansions in a non-transcribed setting. Repeats are placed between the galactose promoter and its upstream activating sequence. This distance is length constrained, with an expansion shutting off expression of the *CAN1* marker.

D) The *ysh1-L14S* mutant shows no change in the rate of repeat expansion in the non-transcribed setting. Error bars represent 95% confidence intervals of two trials. * Significantly different from WT. # Significantly different from *ysh1-L14S*.

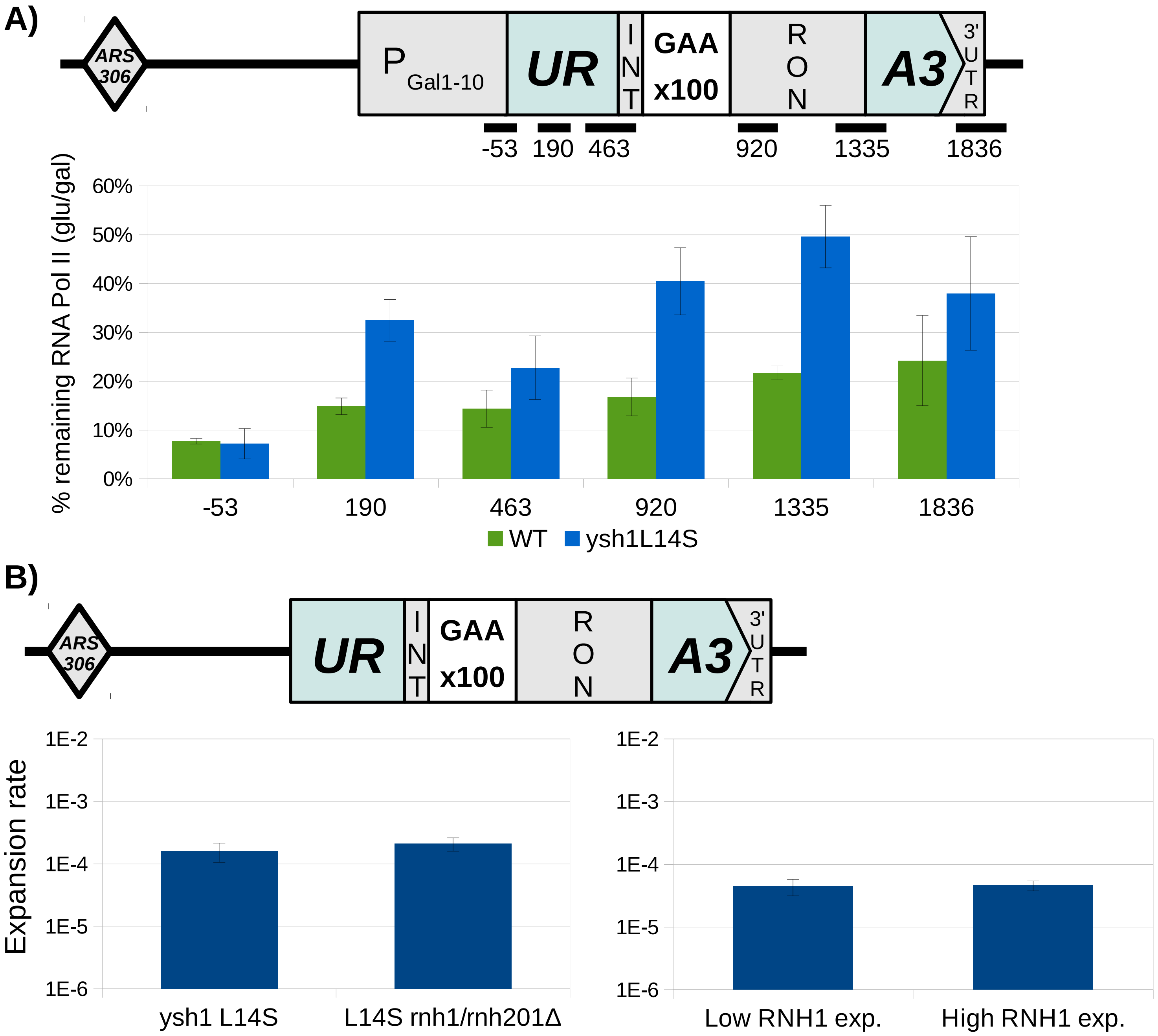


Fig. 6. *Ysh1-L14S* mutant exhibits slow transcription elongation, but expansions are not affected by R-loop processing enzymes.

A) Diagram of the modified *URA3*-(GAA)$_{100}$ cassette placed under control of the *GAL1-10* promoter. This modified cassette was used to measure transcription elongation speed via RNA polymerase clearance assays. The *ysh1-L14S* mutant displays markedly slower elongation speed, especially downstream of the repeat tract, as indicated by a greater fraction of RNA Pol II remaining two minutes after glucose inhibition. Error bars represent standard error of two trials. Primer pairs are numbered by the midpoint of the PCR product, with respect to the ORF start position. See also Figure S3.

B) Knockout of RNaseH1 (*rnh1*Δ) and RNaseH2 (*rnh201*Δ) (left), which remove R-loops, or overexpression of RNaseH1 (right) do not affect expansions in a *ysh1-L14S* mutant background. Fluctuation assays were performed using the *URA3* cassette located at ARS306, with cells grown at 30^{o}C.

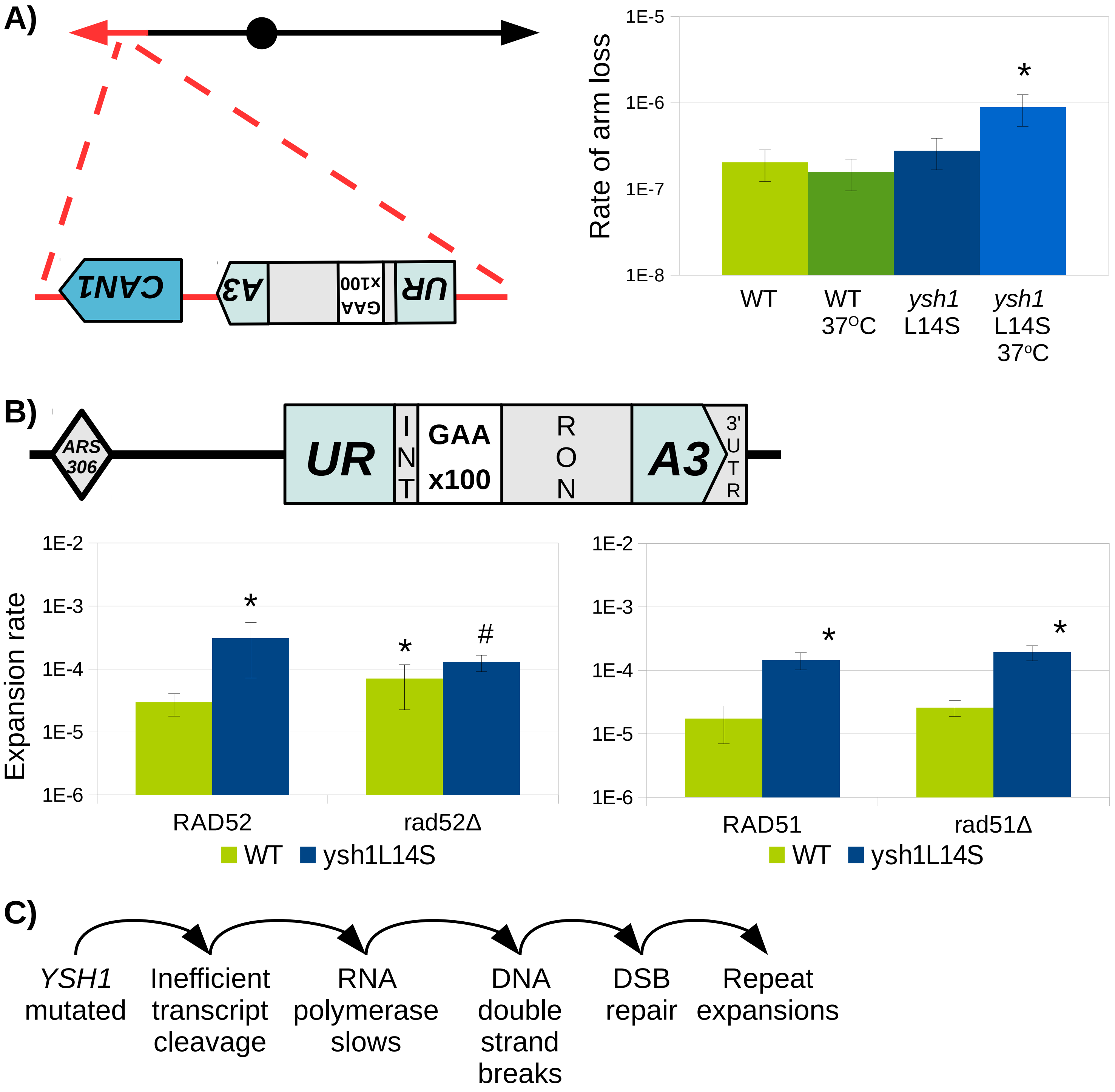


Fig. 7. *Ysh1-L14S* mutation leads to double strand breaks, which may be processed by HR into repeat expansions.

A) Selective system for chromosomal arm loss at $(GAA)_n$ repeats. The original *URA3*-$(GAA)_{100}$ cassette was moved to the non-essential arm (marked in red) of chromosome V, just upstream of the endogenous *CAN1* marker gene. An unrepaired double strand break at the repeats will confer resistance to both canavanine and 5FOA. Fluctuation assay shows an increase in the arm loss rate for the *ysh1-L14S* mutant, which becomes significant under temperature-sensitive conditions. Error bars represent 95% confidence intervals after two trials.

B) Knockout of *RAD52* (left) reduces expansions in a *ysh1-L14S* mutant background, while knockout of *RAD51* (right) does not affect expansions. Fluctuation assays were performed using the *URA3* cassette located at ARS306, with cells grown at 30ºC. * Significantly different from WT. # Significantly different from *ysh1-L14S*

C) Proposed chain of events leading to *ysh1-L14S*-driven repeat expansion.

References

Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet *Chapter 7*, Unit7 20.

Barthelemy, J., Hanenberg, H., and Leffak, M. (2016). FANCJ is essential to maintain microsatellite structure genome-wide during replication stress. Nucleic Acids Res *44*, 6803–6816.

Butler, J.S., and Napierala, M. (2015). Friedreich's ataxia--a case of aberrant transcription termination? Transcription *6*, 33-36.

Chan, S., Choi, E.A., and Shi, Y. (2011). Pre-mRNA 3'-end processing complex assembly and function. Wiley Interdiscip Rev RNA *2*, 321-335.

Chanfreau, G., Noble, S.M., and Guthrie, C. (1996). Essential yeast protein with unexpected similarity to subunits of mammalian cleavage and polyadenylation specificity factor (CPSF). Science *274*, 1511-1514.

Chen, C., and Kolodner, R.D. (1999). Gross chromosomal rearrangements in Saccharomyces cerevisiae replication and recombination defective mutants. Nat Genet *23*, 81-85.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80-92.

Debacker, K., Frizzell, A., Gleeson, O., Kirkham-McCarthy, L., Mertz, T., and Lahue, R.S. (2012). Histone deacetylase complexes promote trinucleotide repeat expansions. PLoS Biol *10*, e1001257.

Desmoucelles, C., Pinson, B., Saint-Marc, C., and Daignan-Fornier, B. (2002). Screening the yeast "disruptome" for mutants affecting resistance to the immunosuppressive drug, mycophenolic acid. J Biol Chem *277*, 27036-27044.

Dominguez-Sanchez, M.S., Barroso, S., Gomez-Gonzalez, B., Luna, R., and Aguilera, A. (2011). Genome instability and transcription elongation impairment in human cells depleted of THO/TREX. PLoS Genet *7*, e1002386.

Downing, B., Morgan, R., VanHulle, K., Deem, A., and Malkova, A. (2008). Large inverted repeats in the vicinity of a single double-strand break strongly affect repair in yeast diploids lacking Rad51. Mutat Res *645*, 9-18.

Dutertre, M., Lambert, S., Carreira, A., Amor-Gueret, M., and Vagner, S. (2014). DNA damage: RNA-binding proteins protect from near and far. Trends Biochem Sci *39*, 141-149.

Dutta, D., Shatalin, K., Epshtein, V., Gottesman, M.E., and Nudler, E. (2011). Linking RNA polymerase backtracking to genome instability in E. coli. Cell *146*, 533-543.

Gaillard, H., and Aguilera, A. (2014). Cleavage factor I links transcription termination to DNA damage response and genome integrity maintenance in Saccharomyces cerevisiae. PLoS Genet *10*, e1004203.
Garas, M., Dichtl, B., and Keller, W. (2008). The role of the putative 3' end processing endonuclease Ysh1p in mRNA and snoRNA synthesis. RNA *14*, 2671-2684.

Gavalda, S., Gallardo, M., Luna, R., and Aguilera, A. (2013). R-loop mediated transcription-associated recombination in trf4Delta mutants reveals new links between RNA surveillance and genome integrity. PLoS One *8*, e65541.

Genetic Modifiers of Huntington's Disease, C. (2015). Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. Cell *162*, 516-526.

Gottesfeld, J.M., Rusche, J.R., and Pandolfo, M. (2013). Increasing frataxin gene expression with histone deacetylase inhibitors as a therapeutic approach for Friedreich's ataxia. J Neurochem *126 Suppl 1*, 147-154.

Groh, M., and Gromak, N. (2014). Out of balance: R-loops in human disease. PLoS Genet *10*, e1004630.

Groh, M., Lufino, M.M., Wade-Martins, R., and Gromak, N. (2014). R-loops associated with triplet repeat expansions promote gene silencing in Friedreich ataxia and fragile X syndrome. PLoS Genet *10*, e1004318.

Hamperl, S., and Cimprich, K.A. (2014). The contribution of co-transcriptional RNA:DNA hybrid structures to DNA damage and genome instability. DNA Repair (Amst) *19*, 84-94.

House, N.C., Yang, J.H., Walsh, S.C., Moy, J.M., and Freudenreich, C.H. (2014). NuA4 initiates dynamic histone H4 acetylation to promote high-fidelity sister chromatid recombination at postreplication gaps. Mol Cell *55*, 818-828.

Ira, G., and Haber, J.E. (2002). Characterization of RAD51-independent break-induced replication that acts preferentially with short homologous sequences. Mol Cell Biol *22*, 6384-6392.

Janke, C., Magiera, M.M., Rathfelder, N., Taxis, C., Reber, S., Maekawa, H., Moreno-Borchart, A., Doenges, G., Schwob, E., Schiebel, E., *et al.* (2004). A versatile toolbox for PCR-based tagging of yeast genes: new fluorescent proteins, more markers and promoter substitution cassettes. Yeast *21*, 947-962.

Kim, H.M., Narayanan, V., Mieczkowski, P.A., Petes, T.D., Krasilnikova, M.M., Mirkin, S.M., and Lobachev, K.S. (2008). Chromosome fragility at GAA tracts in yeast depends on repeat orientation and requires mismatch repair. EMBO J *27*, 2896–2906.

Kim, J.C., Harris, S.T., Dinter, T., Shah, K.A., and Mirkin, S.M. (2017). The role of break-induced replication in large-scale expansions of (CAG)n/(CTG)n repeats. Nat Struct Mol Biol *24*, 55-60.

Kim, J.C., and Mirkin, S.M. (2013). The balancing act of DNA repeat expansions. Curr Opin Genet Dev *23*, 280-288.

Kim, M., Ahn, S.H., Krogan, N.J., Greenblatt, J.F., and Buratowski, S. (2004). Transitions in RNA polymerase II elongation complexes at the 3' ends of genes. EMBO J *23*, 354-364.

Kovtun, I.V., Liu, Y., Bjoras, M., Klungland, A., Wilson, S.H., and McMurray, C.T. (2007). OGG1 initiates age-dependent CAG trinucleotide expansion in somatic cells. Nature *447*, 447-452.

Kovtun, I.V., and McMurray, C.T. (2001). Trinucleotide expansion in haploid germ cells by gap repair. Nat Genet *27*, 407-411.

Kovtun, I.V., and McMurray, C.T. (2008). Features of trinucleotide repeat instability in vivo. Cell Res *18*, 198-213.

Krasilnikova, M.M., Kireeva, M.L., Petrovic, V., Knijnikova, N., Kashlev, M., and Mirkin, S.M. (2007). Effects of Friedreich's ataxia (GAA)n*(TTC)n repeats on RNA synthesis and stability. Nucleic Acids Res *35*, 1075-1084.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Lin, Y., Dent, S.Y., Wilson, J.H., Wells, R.D., and Napierala, M. (2010). R loops stimulate genetic instability of CTG.CAG repeats. Proc Natl Acad Sci U S A *107*, 692-697.

Lin, Y., Hubert, L., Jr., and Wilson, J.H. (2009). Transcription destabilizes triplet repeats. Mol Carcinog *48*, 350-361.

Lin, Y., and Wilson, J.H. (2007). Transcription-induced CAG repeat contraction in human cells is mediated in part by transcription-coupled nucleotide excision repair. Mol Cell Biol *27*, 6209-6217.

Liu, G., Chen, X., Gao, Y., Lewis, T., Barthelemy, J., and Leffak, M. (2012). Altered replication in human cells promotes DMPK (CTG)(n) . (CAG)(n) repeat instability. Mol Cell Biol *32*, 1618-1632.

Lopez Castel, A., Cleary, J.D., and Pearson, C.E. (2010). Repeat instability as the basis for human diseases and as a potential target for therapy. Nat Rev Mol Cell Biol *11*, 165-170.

Mandel, C.R., Kaneko, S., Zhang, H., Gebauer, D., Vethantham, V., Manley, J.L., and Tong, L. (2006). Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. Nature *444*, 953-956.

Martins, S., Pearson, C.E., Coutinho, P., Provost, S., Amorim, A., Dube, M.P., Sequeiros, J., and Rouleau, G.A. (2014). Modifiers of (CAG)(n) instability in Machado-Joseph disease (MJD/SCA3) transmissions: an association study with DNA replication, repair and recombination genes. Hum Genet *133*, 1311-1318.

Mason, A.G., Tome, S., Simard, J.P., Libby, R.T., Bammler, T.K., Beyer, R.P., Morton, A.J., Pearson, C.E., and La Spada, A.R. (2014). Expression levels of DNA replication and repair genes predict regional somatic repeat instability in the brain but are not altered by polyglutamine disease protein expression or age. Hum Mol Genet *23*, 1606-1618.
Mirkin, S.M. (2007). Expandable DNA repeats and human disease. Nature *447*, 932-940.

Moehle, E.A., Braberg, H., Krogan, N.J., and Guthrie, C. (2014). Adventures in time and space: splicing efficiency and RNA polymerase II elongation rate. RNA Biol *11*, 313-319.

Morales, F., Couto, J.M., Higham, C.F., Hogg, G., Cuenca, P., Braida, C., Wilson, R.H., Adam, B., del Valle, G., Brian, R*., et al.* (2012). Somatic instability of the expanded CTG triplet repeat in myotonic dystrophy type 1 is a heritable quantitative trait and modifier of disease severity. Human molecular genetics *21*, 3558-3567.

Morales, F., Vasquez, M., Santamaria, C., Cuenca, P., Corrales, E., and Monckton, D.G. (2016). A polymorphism in the MSH3 mismatch repair gene is associated with the levels of somatic instability of the expanded CTG repeat in the blood DNA of myotonic dystrophy type 1 patients. DNA Repair (Amst) *40*, 57-66.

Nudler, E. (2012). RNA polymerase backtracking in gene regulation and genome instability. Cell *149*, 1438-1445.

Porrua, O., Boudvillain, M., and Libri, D. (2016). Transcription Termination: Variations on Common Themes. Trends Genet *32*, 508-522.

Reddy, K., Schmidt, M.H., Geist, J.M., Thakkar, N.P., Panigrahi, G.B., Wang, Y.H., and Pearson, C.E. (2014). Processing of double-R-loops in (CAG).(CTG) and C9orf72 (GGGGCC).(GGCCCC) repeats causes instability. Nucleic Acids Res *42*, 10473-10487.

Reddy, K., Tam, M., Bowater, R.P., Barber, M., Tomlinson, M., Nichol Edamura, K., Wang, Y.H., and Pearson, C.E. (2011). Determinants of R-loop formation at convergent bidirectionally transcribed trinucleotide repeats. Nucleic Acids Res *39*, 1749-1762.

Ryan, K., Calvo, O., and Manley, J.L. (2004). Evidence that polyadenylation factor CPSF-73 is the mRNA 3' processing endonuclease. RNA *10*, 565-573.

Savouret, C., Brisson, E., Essers, J., Kanaar, R., Pastink, A., te Riele, H., Junien, C., and Gourdon, G. (2003). CTG repeat instability and size variation timing in DNA repair-deficient mice. EMBO J *22*, 2264-2273.

Savouret, C., Garcia-Cordier, C., Megret, J., te Riele, H., Junien, C., and Gourdon, G. (2004). MSH2-dependent germinal CTG repeat expansions are produced continuously in spermatogonia from DM1 transgenic mice. Molecular and cellular biology. *24*, 629-637.

Schaughency, P., Merran, J., and Corden, J.L. (2014). Genome-wide mapping of yeast RNA polymerase II termination. PLoS Genet *10*, e1004632.

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

Shah, K.A., Shishkin, A.A., Voineagu, I., Pavlov, Y.I., Shcherbakova, P.V., and Mirkin, S.M. (2012). Role of DNA polymerases in repeat-mediated genome instability. Cell Rep *2*, 1088-1095.

Shishkin, A.A., Voineagu, I., Matera, R., Cherng, N., Chernet, B.T., Krasilnikova, M.M., Narayanan, V., Lobachev, K.S., and Mirkin, S.M. (2009). Large-scale expansions of Friedreich's ataxia GAA repeats in yeast. Mol Cell *35*, 82-92.

Sonkar, A., Yadav, S., and Ahmed, S. (2016). Cleavage and polyadenylation factor, Rna14 is an essential protein required for the maintenance of genomic integrity in fission yeast Schizosaccharomyces pombe. Biochim Biophys Acta

1863, 189-197.

Soragni, E., Miao, W., Iudicello, M., Jacoby, D., De Mercanti, S., Clerico, M., Longo, F., Piga, A., Ku, S., Campau, E.*, et al.* (2014). Epigenetic therapy for Friedreich ataxia. Ann Neurol *76*, 489-508.

Stirling, P.C., Chan, Y.A., Minaker, S.W., Aristizabal, M.J., Barrett, I., Sipahimalani, P., Kobor, M.S., and Hieter, P. (2012). R-loop-mediated genome instability in mRNA cleavage and polyadenylation mutants. Genes Dev *26*, 163-175.

Symington, L.S. (2002). Role of RAD52 epistasis group genes in homologous recombination and double-strand break repair. Microbiol Mol Biol Rev *66*, 630-670.

Tous, C., Rondon, A.G., Garcia-Rubio, M., Gonzalez-Aguilera, C., Luna, R., and Aguilera, A. (2011). A novel assay identifies transcript elongation roles for the Nup84 complex and RNA processing factors. EMBO J *30*, 1953-1964.

Usdin, K., House, N.C., and Freudenreich, C.H. (2015). Repeat instability during DNA repair: Insights from model systems. Crit Rev Biochem Mol Biol *50*, 142-167.

Wahba, L., Amon, J.D., Koshland, D., and Vuica-Ross, M. (2011). RNase H and multiple RNA biogenesis factors cooperate to prevent RNA:DNA hybrids from generating genome instability. Mol Cell *44*, 978-988.

Yang, J.H., and Freudenreich, C.H. (2010). The Rtt109 histone acetyltransferase facilitates error-free replication to prevent CAG/CTG repeat contractions. DNA Repair *9*, 414-420.

Yu, Z., Zhu, Y., Chen-Plotkin, A.S., Clay-Falcone, D., McCluskey, L., Elman, L., Kalb, R.G., Trojanowski, J.Q., Lee, V.M., Van Deerlin, V.M.*, et al.* (2011). PolyQ repeat expansions in ATXN2 associated with ALS are CAA interrupted repeats. PLoS One *6*, e17951.

Zhang, Y., Shishkin, A.A., Nishida, Y., Marcinkowski-Desmond, D., Saini, N., Volkov, K.V., Mirkin, S.M., and Lobachev, K.S. (2012). Genome-wide screen identifies pathways that govern GAA/TTC repeat fragility and expansions in dividing and nondividing yeast cells. Molecular cell *48*, 254-265.

Zhao, J., Hyman, L., and Moore, C. (1999a). Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. Microbiol Mol Biol Rev *63*, 405-445.

Zhao, J., Kessler, M., Helmling, S., O'Connor, J.P., and Moore, C. (1999b). Pta1, a component of yeast CF II, is required for both cleavage and poly(A) addition of mRNA precursor. Mol Cell Biol *19*, 7733-7740.

Zhao, J., Kessler, M.M., and Moore, C.L. (1997). Cleavage factor II of Saccharomyces cerevisiae contains homologues to subunits of the mammalian Cleavage/ polyadenylation specificity factor and exhibits sequence-specific, ATP-dependent interaction with precursor RNA. J Biol Chem *272*, 10831-10838.

Zhao, X.N., and Usdin, K. (2014). Gender and cell-type-specific effects of the transcription-coupled repair protein, ERCC6/CSB, on repeat expansion in a mouse model of the fragile X-related disorders. Hum Mutat *35*, 341-349.

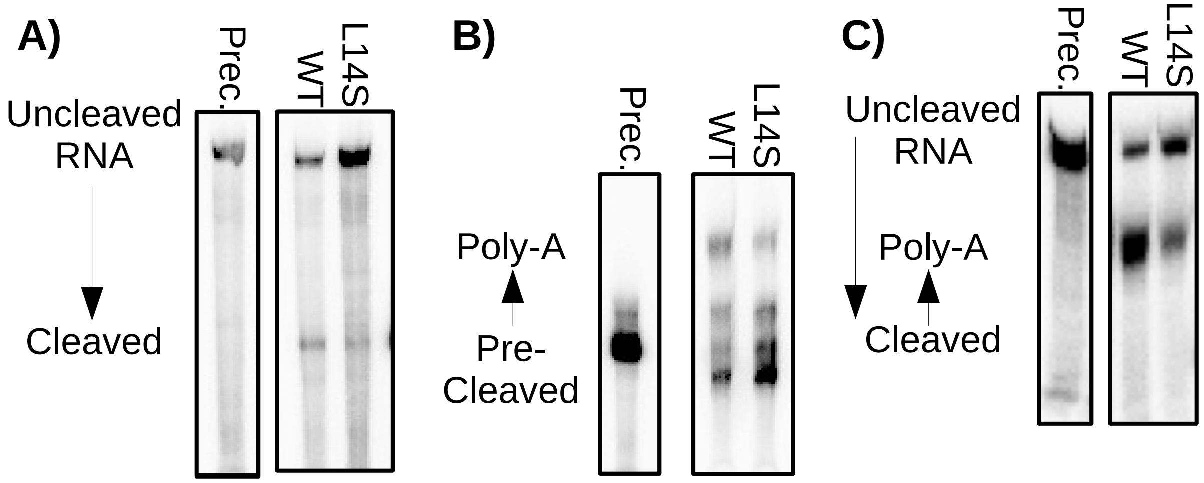


Figure S1. The *ysh1-L14S* mutant is defective for in-vitro mRNA 3' end-processing. Related to Figure 3. *In vitro* assay in which a precursor RNA is combined with cell extracts derived from WT or *ysh1-L14S* mutant yeast, which were grown at 30^{o}C. **(A)** The precursor RNA is shortened by the cleavage reaction without poly(A) addition if the reaction is conducted in the presence of 3' dATP. For *ysh1-L14S* mutants, less of the precursor RNA is converted to the cleaved form. **(B)** Pre-cleaved precursor RNA which ends at the *GAL7* poly(A) site is lengthened by the addition of the poly-A tail. For *ysh1-L14S* mutants, less of the precursor RNA is converted to the polyadenylated form. **(C)** Coupled reaction in the presence of ATP, in which an uncleaved precursor RNA is cleaved and subsequently polyadenylated. For *ysh1-L14S* mutants, less of the precursor RNA is converted to the polyadenylated form.

Figure contributed by Julia Hisey.

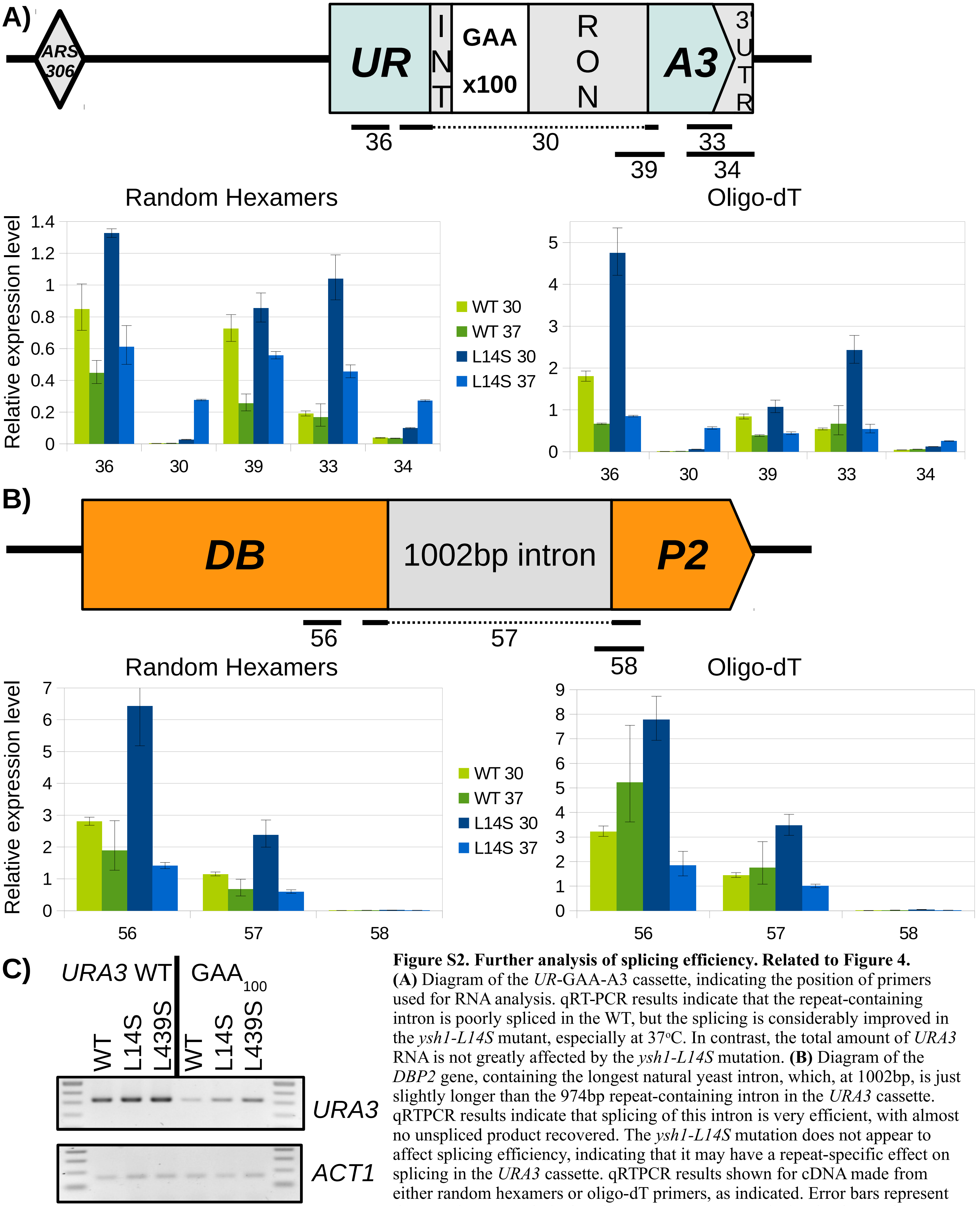


Figure S2. Further analysis of splicing efficiency. Related to Figure 4. **(A)** Diagram of the *UR*-GAA-A3 cassette, indicating the position of primers used for RNA analysis. qRT-PCR results indicate that the repeat-containing intron is poorly spliced in the WT, but the splicing is considerably improved in the *ysh1-L14S* mutant, especially at 37°C. In contrast, the total amount of *URA3* RNA is not greatly affected by the *ysh1-L14S* mutation. **(B)** Diagram of the *DBP2* gene, containing the longest natural yeast intron, which, at 1002bp, is just slightly longer than the 974bp repeat-containing intron in the *URA3* cassette. qRTPCR results indicate that splicing of this intron is very efficient, with almost no unspliced product recovered. The *ysh1-L14S* mutation does not appear to affect splicing efficiency, indicating that it may have a repeat-specific effect on splicing in the *URA3* cassette. qRTPCR results shown for cDNA made from either random hexamers or oligo-dT primers, as indicated. Error bars represent the SD of qPCR technical replicates. **(C)** RTPCR analysis of polysomal RNA, comparing spliced *URA3* mRNA (primer pair #3), and loading control *ACT1* spliced mRNA. Results are shown for both *ysh1-L14S* and *ysh1-L439S* mutants, which show an increase in mRNA relative to the WT for the UR-GAA-A3 cassette (right panel). In contrast, the *URA3* gene lacking an intron is unaffected (left panel). This supports the results in (**A**), suggesting that the increase in splicing efficiency leads to RNA that is stably exported from the nucleus.

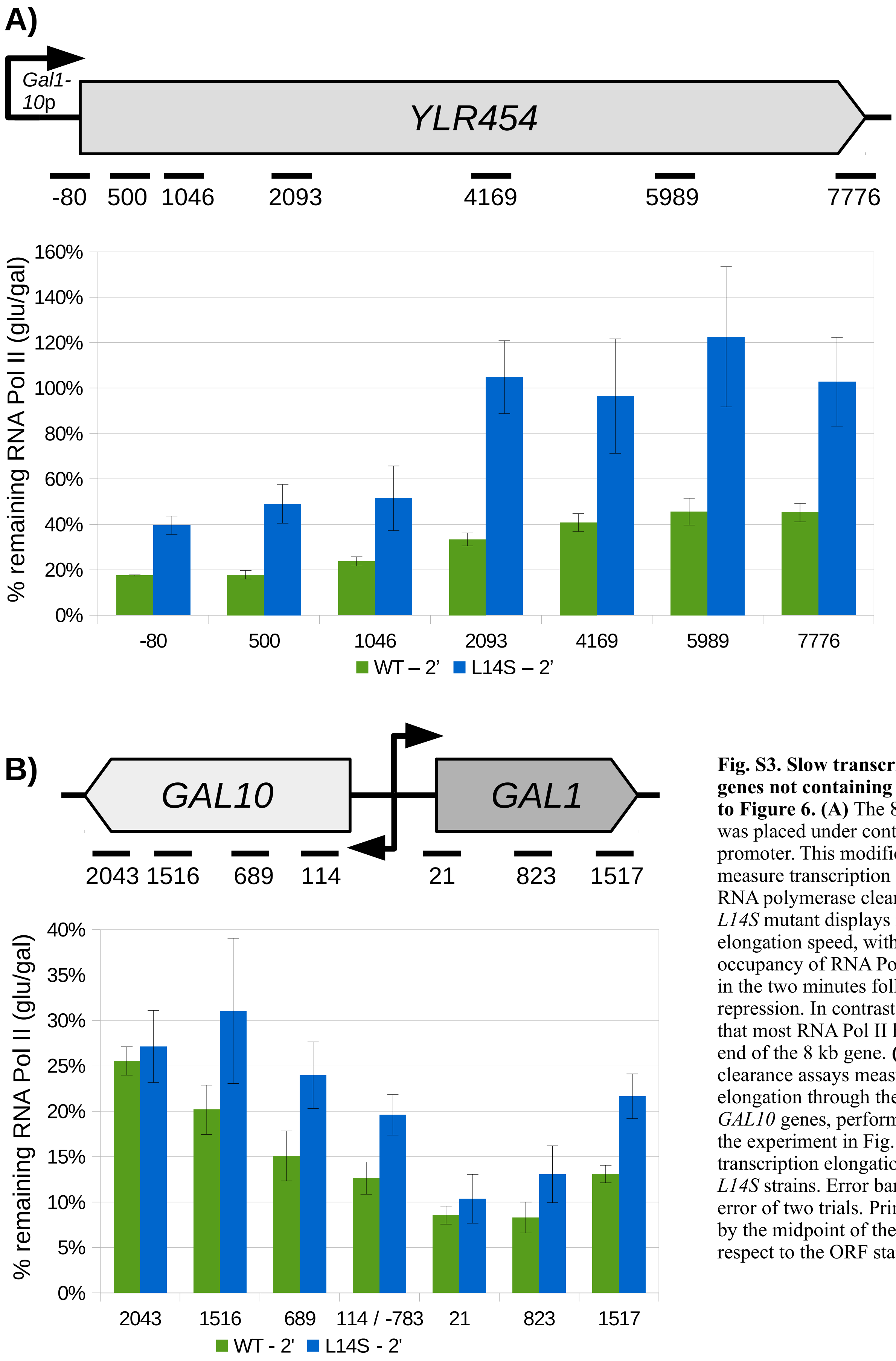


Fig. S3. Slow transcription elongation at genes not containing GAA repeats. Related to Figure 6. (A) The 8 kb-long gene *YLR454w* was placed under control of the *GAL1-10* promoter. This modified gene was used to measure transcription elongation speed via RNA polymerase clearance assays. The *ysh1-L14S* mutant displays markedly slower elongation speed, with ~2.5-fold greater occupancy of RNA Pol II observed past 2 kb in the two minutes following glucose repression. In contrast, the WT strain shows that most RNA Pol II has proceeded past the end of the 8 kb gene. **(B)** RNA polymerase clearance assays measuring transcription elongation through the endogenous *GAL1* and *GAL10* genes, performed concurrently with the experiment in Fig. 6A. Slower transcription elongation is evident in the *ysh1-L14S* strains. Error bars represent standard error of two trials. Primer pairs are numbered by the midpoint of the PCR product, with respect to the ORF start position.

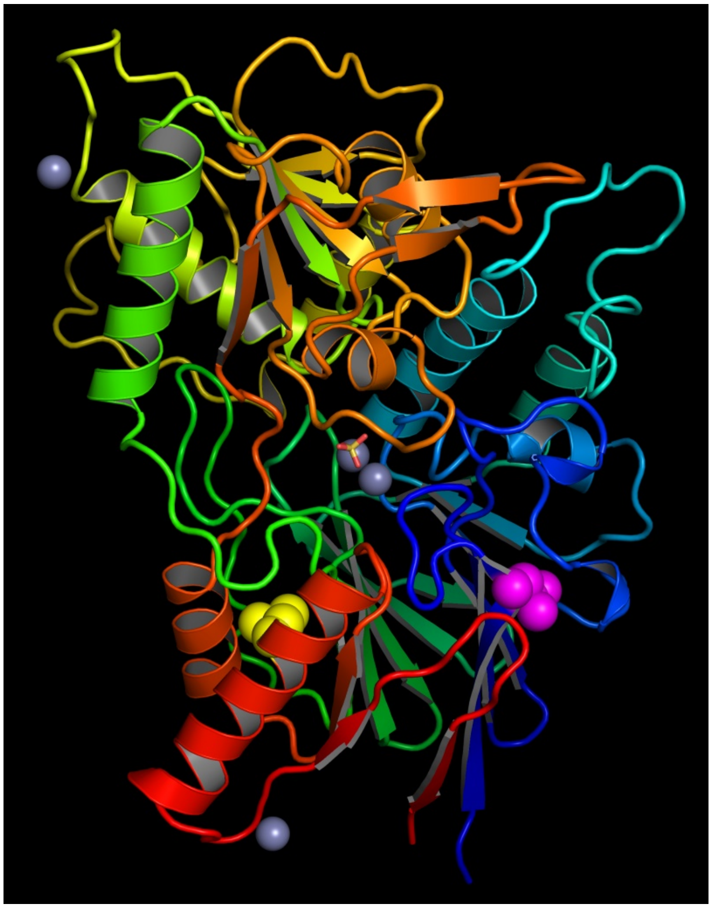


Fig. S4. Location of mutants within Ysh1 protein structure. Related to Figure 3. A portion of the protein structure of the human homolog of Ysh1, CPSF-73, is shown (Mandel et al., 2006). The *ysh1-L14S* mutant in (magenta spheres) is located near the surface and partially exposed, while the *ysh1-L439S* (yellow spheres) is buried within the protein.

Figure contributed by Franco Puleo and Claire Moore.

Table S1. UV Dosage Determination. Related to Experimental Methods – Screening Approach. Determining a suitable UV dosage by measuring the rate of mutation of the *CAN1* gene under various UV doses. μ_g represents mutation rates determined as described in (Drake, 1991), using three independent trials. The relative number of mutations is scaled to the 0 UV dose.

Dose (x100 µJ):	**µg:**	**Relative # Mutations:**
0	0.00192	1
25	0.00193	1.01
50	0.00218	1.14
75	0.00818	4.27
100	0.0632	33.0

Table S2. Yeast strains in this study.

Strain name:	Parent strain:	Cassettes:	Additional genotype info:
SMY706	CH1585	none	leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
SMY720	SMY706	none	ade2::URA3, leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
SMY724	SMY720	AD-GAA63-E2	leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
SMY732	SMY724	AD-GAA63-E2, UR-GAA100-A3 at ARS306 locus	leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
RMG28	SMY732	AD-GAA63-E2, UR-GAA100-A3 at ARS306 locus	ysh1 L439S, ysh1 S59F, leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa, various UV-induced mutations
RMG35	SMY732	AD-GAA63-E2, UR-GAA100-A3 at ARS306 locus	ysh1 L14S, leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa, various UV-induced mutations
RMG87	SMY732	AD-GAA63-E2, UR-GAA100-A3 at ARS306 locus	ysh1 L439S, leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
RMG89	SMY732	AD-GAA63-E2, UR-GAA100-A3 at ARS306 locus	ysh1 L14S, leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
RMG92	SMY732	AD-GAA63-E2, UR-GAA100-A3 at ARS306 locus	leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
RMG108	G4G1C1T150 (Shah et al., 2014)	UAS-GAA100-Galp-CAN1	ysh1 L14S, CAN1::kanMX4, BAR1::URA3, leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
RMG110	G4G1C1T150 (Shah et al., 2014)	UAS-GAA100-Galp-CAN1	CAN1::kanMX4, BAR1::URA3, leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
RMG385	SMY706	none	leu2Δ1, trp1Δ63, ura3::HPH, his3Δ200, MIP1, HAP1, MATa
RMG389	RMG385	UR-GAA100-A3 at Chr. V non-essential arm	leu2Δ1, trp1Δ63, ura3::HPH, his3Δ200, MIP1, HAP1, MATa
RMG391	RMG389	UR-GAA100-A3 at Chr. V non-essential arm	ysh1 L14S, leu2Δ1, trp1Δ63, ura3::HPH, his3Δ200, MIP1, HAP1, MATa
RMG338	SMY724	AD-GAA63-E2	leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
FPY01	SMY724	GAL1-10p-YLR454, AD-GAA63-E2	leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
FPY02	RMG338	GAL1-10p-YLR454, AD-GAA63-E2	leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
RMG431	SMY732	Gal1-10p-UR-GAA100-A3 at ARS306 locus	leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
RMG434	RMG89	Gal1-10p-UR-GAA100-A3 at ARS306 locus	ysh1 L14S, leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
RMG407	SMY732	UR-GAA100-A3 at ARS306 locus	leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
RMG409	RMG89	UR-GAA100-A3 at ARS306 locus	ysh1 L14S, leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
RMG167	SMY732	AD-GAA63-E2, UR-GAA100-A3 at ARS306 locus	RAD52::HPH, leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
RMG169	RMG89	AD-GAA63-E2, UR-GAA100-A3 at ARS306 locus	RAD52::HPH, ysh1 L14S, leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
RMG232	SMY732	AD-GAA63-E2, UR-GAA100-A3 at ARS306 locus	RAD51::HPH, leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
RMG236	RMG89	AD-GAA63-E2, UR-GAA100-A3 at ARS306 locus	RAD51::HPH, ysh1 L14S, leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
RMG114	RMG92	AD-GAA63-E2, UR-GAA100-A3 at ARS306 locus	RNH201::LEU2, leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa

RMG112	RMG89	AD-GAA63-E2, UR-GAA100-A3 at ARS306 locus	RNH201::LEU2, ysh1 L14S, leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
RMG118	RMG114	AD-GAA63-E2, UR-GAA100-A3 at ARS306 locus	RNH1::HPH, RNH201::LEU2, leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
RMG120	RMG112	AD-GAA63-E2, UR-GAA100-A3 at ARS306 locus	RNH1::HPH, RNH201::LEU2, ysh1 L14S, leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
RMG242	SMY732	AD-GAA63-E2, UR-GAA100-A3 at ARS306 locus	pRNH1::Nat-pMet25, leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa
RMG247	RMG89	AD-GAA63-E2, UR-GAA100-A3 at ARS306 locus	pRNH1::Nat-pMet25, ysh1 L14S, leu2Δ1, trp1Δ63, ura3Δ52, his3Δ200, MIP1, HAP1, MATa

Table S3. PCR products in this study

Set #:	Purpose:	Template:	Primer sequence (FWD, REV):
1	Amplifies GAA repeats in all cassettes	colony PCR	CTCGATGTGCAGAACCTGAAGCTTGATCT
			GCTCGAGTGCAGACCTCAAATTCGATGA
2	For replacement of ADE2 gene with URA3 gene	pRS306 (Sikorski and Hieter, 1989)	GAGAAGTGACGCAAGCATCAATGGTATAATGTCCAGAGTTGTGAGGCCTTGACACCGCAGGGTAATAACTGAT
			CTTTTCCCGGTTGTGGTATATTTGGTGTGGAAATGTTCTATTTAGAAACACGGTCACAGCTTGTCTGTAA
3	Amplifies repeats in AD-GAA-E2 cassette for sequencing	colony PCR	CCCGGTTGTGGTATATTTGGTG
			GCATAATGGCGTTCGTTGTAATGG
4	Amplifies YSH1 flanking sequence for plasmid construction	SMY732 gDNA	TATCGTCTCACGCGCTATTCCTACAGTAACTAACGCAGACATCA
			CCAGATTTTGGTTTGGTATTACTTCTATAAAGTAGTCTA
5	Amplifies YSH1 gene for plasmid construction	RMG28 or RMG35 gDNA	TAGAACTGCCTTTTATTGTTTACCTTAATCACT
			ATAGCGCGCCTGTATATCGTCATTTAGGGTT
6	Confirmation of HIS3–YSH1 integration in yeast	colony PCR	GGAGCGAACAAATACAACAAC
			ATAGCGCGCCTGTATATCGTCATTTAGGGTT
7	Sequence confirmation of ysh1 mutations	colony PCR	CTAGCGGATCTTTAACAACT
			GTCTTTGTGTTCCTCTTCAT
8	Amplify HIS3–YSH1 from yeast for propagation into new strains	RMG89 gDNA	GGAGCGAACAAATACAACAAC
			ATAGCGCGCCTGTATATCGTCATTTAGGGTT
9	Replace ura3Δ52 with hphMX4	pCORE-UH (Storici and Resnick, 2003)	TTTTGACCATCAAAGAAGGTTAATGTGGCTGTGGTTTCAGGGTCCATAAAGTTTAGCTTGCCTCGTCCC
			TATTGGATAGTTCCTTTTTATAAAGGCCATGAAGCTTTTTCTTTCCAATTCTTATCCTTCACCATAAATATGCCTCGCAAAAAAGGTAATAATACGCAAACCGCCTCTC
10	Confirmation of ura3Δ52 absence	colony PCR	CATTATCCGCCAAGTACAAT
			GTCTCCCTTGTCATCTAAAC
11	Amplifies UR-GAA-A3 cassette along with TRP1 marker, with flanking homology to the non-essential arm of Chr. V	SMY732 gDNA	ACAGAAACCCTAGATTTCTATAGGGCAAATTTCAGGGTTATACTAAACACTCGCAGACATTGTTATTTCC
			AATTAAACCTATTTCTTTATCATCATATTTACTTATATCTTTAACAGATTCCAGGCGATAAGCTAGCAGGCAA
12	Located outside integration site for UR-GAA-A3 cassette on Chr. V non-essential arm	colony PCR	GATTCCCTGATTCGGTTTACTCT
			CTCTTGTCCCTTATTAGCCTTGA
13	Replacement of URA3 promoter with Gal1-10 promoter	pCORE-Hp53 (Storici and Resnick, 2003)	TAATCAAGTAACACTCGCAGACATTGTTATTTCCGCGGATCCGGAGATCTGTTTAGCTTGCCTCGTCCC
			GCAGCACGTTCCTTATATGTAGCTTTCGACATGATTTATCTTCGTTTCCTTCTTAAAGTTAAACAAAATTATTTCTAGTCGA
14	Confirmation of GAL1-10 promoter integration	colony PCR	ATGACCCACTCAGGTGTTAAA
			TAATGCCTTTAGCGGCTTAACT
15	ADE2 WT restoration	Strain #1 gDNA	TATTAGTGAGAAGCCGAGAA
			TTGAGCCGCCTTATATGAA
16	RAD52 KO integration	pAG32 (Goldstein and McCusker, 1999)	CGAATGGCGTTTTTAAGCTATTTTGCCACTGAGAATCAACAAATGCAAACAAGGAGGTTGCCAGATCTGTTTAGCTTGCCT
			GGTTTCACGCGGTACTTGATTCCCAGCCCCTTCTAGCATATGAGGCCCCAGTTCTTTATCATCGATGAATTCGAGCTCGTT
17	RAD52 KO external confirmation	colony PCR	CTAGAGGATTTTGGAGTAATAAATAATGATG
			ACGTCGCTAAAGATGGTATGGTA
18	RAD52 KO internal confirmation	colony PCR	GCCAAGAAATCTGCCGTTAC
			TGAGCTTTCGCTGATTTCATCC

19	RAD51 KO integration	pAG32 (Goldstein and McCusker, 1999)	AAATGTTGGAAATGCACCACTACCGTTCTTCAACCAATCTAGTTTAGCTATTTAGAACGCGGCTACAATTA
			AAAGAGGAGAATTGAAAGTAAACCTGTGTAAATAAATAGAGACAAGAGACCAAATACCTACCCTGATTCTGTGGATAACC
20	RAD51 KO external confirmation	colony PCR	CAATTCGCAAGAAACGCACT
			AAGTAGTCATCGGGAAGAAGAGTA
21	RAD51 KO internal confirmation	colony PCR	AGATCGGAGCTGATTTGTTTGAC
			CTTCACCGCCACCAATATCC
22	RNH201 KO integration	pRS305 (Sikorski and Hieter, 1989)	ATGGTACCCCCCACGGTAGAAGCATCCTTGGAGTCTCCTTACACTAAGTCGTAGGTGTCGGGGCTGGCTTAA
			AGGTGATCACCGGTACCAATTATCTAGGGTTCTCAGCCTCTTCCTTCTAACTCCTTACGCATCTGTGCGGTA
23	RNH201 KO confirmation	colony PCR	ATTTTGCGACGCCTGCCAAT
			GAAACGGCAAAGCATAGTAGCAGAT
24	RNH1 KO integration	pAG32 (Goldstein and McCusker, 1999)	AATTATGGCAAGGCAAGGGAACTTCTACGCGGTTAGAAAGGGCAGGGAAACAGCTGAAGCTTCGTACGC
			CCTGGATCACCATCGTGTCCTTTTACCCATTCAATCTGAAATTTACCATTCATAGGCCACTAGTGGATCTG
25	RNH1 KO confirmation	colony PCR	GACAGCAGCATCAACAATGAA
			CACGCTTATAGATAGTTATCGGGTA
26	RNH1 promoter replacement with pMet	pYM-N35 (Janke et al., 2004)	TCACTCCTTGCTTATCGAAGGAACTATCGATTCCTAATTACGTACGCTGCAGGTCGAC
			CCCTGCCCTTTCTAACCGCGTAGAAGTTCCCTTGCCTTGCCATCGATGAATTCTCTGTCG
27	pMet-RNH1 confirmation	colony PCR	CAGGGTCGTCAGATACATAGATAC
			ATAACCTGCCCTTGAAGATGAC

Table S4. qPCR primer sets

Set #:	Target:	Primer sequence (FWD, REV):
28	SCR1 gene – RNA normalization	GTGGGATGGGATACGTTGAG
		TTTACGACGGAGGAAAGACG
29	ADE2 cassette (#2) spliced mRNA	TGTAGAGACTATCCACAAGGACA
		TCCAGAGTTGTGAGGCCTT
30	URA3 cassette (#1) spliced mRNA	CTGCTAACATCAAAAGGCCTCTA
		AGTTGAAGCATTAGGTCCCAAA
31	ADE2 cassette (#2) – 3’ end pre-poly-A site	ATTGTGCAAATGCCTAGAGGT
		TAGATAAGCTTCGTAACCGACAGT
32	ADE2 cassette (#2) – 3’ end post-poly-A site	ATTGTGCAAATGCCTAGAGGT
		ACATTTGATGTAATCATAACAAAGCCT
33	URA3 cassette (#1) – 3’ end pre-poly-A site	ATTTGAGAAGATGCGGCCA
		GTCGACGGGTAATAACTGATATAATTAAA
34	URA3 cassette (#1) – 3’ end post-poly-A site	ATTTGAGAAGATGCGGCCA
		GCCTCGTGATACGCCTATTT
35	GAL1-10p-URA3 – promoter-exon junction	TACCTCTATACTTTAACGTCAAGGAGAA
		ATCTTCGTTTCCTTCTTAAAGTTAAACA
36	URA3 cassette – within exon 1	AGTTGAAGCATTAGGTCCCAAA
		TAATGCCTTTAGCGGCTTAACT
37	URA3 cassette – exon-intron junction 1	CCCAGGTATTGTTAGCGGTT
		ATCGAATTTGAGGTCTGCACT
38	URA3 cassette – within intron	AGATCAAGCTTCAGGTTCT
		GAAGGCTCTCAAGGGCATC
39	URA3 cassette – intron-exon junction 2	TTGACTGATCTGTAATAACCACGA
		GAGCCCTTGCATGACAATTC
40	ChrV intergenic region	CAAGTTTGGTGAGATAGTTTACGC
		CCCTGCATTGAGATCGCATC
41	YLR454 promoter (-80bp)	CTGGGGTAATTAATCAGCGAAGCGATG
		CACTTGTACAGTAGAACATTAATCGGAAAC
42	YLR454 (500bp)	CGCAATTAGTCAACAACGATATCACGATTG
		CCTACTTGAAGTCCATCCTTCAGAGG
43	YLR454 –(1046bp)	CAATACCAACAGGTTCAGAAATGAGATGC
		GAGAGAACAAATTGGTTTCGCCAAATATCG
44	YLR454 (2093bp)	CATATCATCCACCCTAGGTGCTAGGTCGG
		GAGCTGACCAGACCTAACCATAGTAGCGTG
45	YLR454 (4169bp)	AGATATTACTCGTTGTTCGTGCCCAG
		AGATATTACTCGTTGTTCGTGCCCAG
46	YLR454 (5989bp)	CGTACTGTTGAAATGGAACGAGGACGC
		ATCGCTTCCATACTCGTTGTATCATCAGTC
47	YLR454 (7776bp)	GAGGGTCACAGATCTATTACTTGCCC
		GTTGTGAGTTGCTTCAGTGGTGAAGTG
48	GAL10 (2043bp)	AATCTGTAGACAATCTTGGACCC
		GCCTGGTAGATACATTGATGCTAT
49	GAL10 (1516bp)	TTCACCAGCAGTCAATTTACCT
		GCCGAGTACATGCTGATAGATAA
50	GAL10 (689bp)	ACCATCTCTGGAATCATAATCGTC
		ATATGGCTCAAGTAGCTGTTGGTA
51	GAL10 (114bp)	TCAAGACCTCTAACCTGGCTA
		TGGTGCTGGATACATTGGTTCA
52	GAL1 (21bp)	TACCTCTATACTTTAACGTCAAGGAGAA
		TCGGCCAATGGTCTTGGTAA
53	GAL1 (823bp)	TTATTGCGAACACCCTTGTTGTA
		CACCGTACGTGGCAGCTAA
54	GAL1 (1517bp)	AAAGAAGCCCTTGCCAATGA
		CTGCCCAATGCTGGTTTAGA

55	RNH1 gene	GGTTAGAAAGGGCAGGGAAA
		TGTCAAGGAGGAATAATGAAGTGG
56	DBP2 5' exon	TACTTAGAAACAGCCTCTCAAGAC
		TCTTTGTCACCATGAATAGCCA
57	DBP2 spliced	CGCCAGAGGTATCGATGTCAA
		AGATATAGCAGTACCAGTAGCACC
58	DBP2 unspliced	ATTACTAACAGTATGCTTTGTAAACGT
		AGATATAGCAGTACCAGTAGCACC

Supplemental Experimental Methods

Strain construction

The list of our strains is presented in Table S1.

The strain used for the screen (SMY732) was constructed in several steps. First, the *ADE2* gene in the starting strain SMY706 was replaced with the *URA3* gene using PCR product #2 by selecting for URA$^+$ red colonies. The resulting strain SMY720 was then transformed with the AD-GAA-E2 cassette (Fig. 1A) by selecting for 5FOA^r white or light-pink colonies. An isolate of this transformation with 63 uninterrupted GAA repeats, verified by Sanger sequencing of PCR product #3, became strain SMY724. This strain was then transformed with cassette UR-GAA-A3, as previously described (Shishkin et al., 2009). The resulting strain SMY732 was used for the mutagenesis screen.

To create *YSH1* point mutant strains, first *YSH1* flanking sequence was amplified from genomic DNA using PCR product #4 (Table S3) and cloned into pRS303 (Sikorski and Hieter, 1989) at the *Bst*Z17I and *Bss*HII sites. Second, the mutant alleles of *YSH1* were created by PCR-amplification of genomic DNA from mutagenized strains RMG28 and RMG35 (PCR product #5, Table S3). They were then cloned into the *Bsm*BI and *Afe*I sites of the aforementioned pRS303 derivative. The resultant plasmids were cut with *Bss*HII. The fragment containing the mutant *YSH1* allele together with the *HIS3* marker was isolated and transformed into strain SMY732 selecting for HIS+ clones. The correct integration of the mutant *YSH1* alleles was confirmed by the presence of PCR product #6 and by Sanger sequencing of PCR product #7, resulting in strains RMG87 (*ysh1-L439S*), RMG89 (*ysh1-L14S*) and RMG92 (*YSH1*).

The *ysh1-L14S* allele from the strain RMG89 was subsequently used as a template for PCR product #8 in order to propagate the *ysh1-L14S* mutation, along with the *HIS3* marker, to further strains. The strain G4G1C1T150 (Shah et al., 2014), which contains the UAS-GAA-CAN1 cassette was transformed in this manner to create the strains RMG108 (*ysh1-L14S*) and RMG110 (*YSH1*). The strain SMY724 was modified in the same manner to create RMG338 (*ysh1-L14S*).

The wild-type strain for the arm loss assay was created by transforming SMY706 with PCR product #9, which directly replaced the *ura3Δ52* allele with the *HPH* selectable marker, in order to discourage intra-chromosomal repair events with the UR-GAA-A3 cassette. The resulting strain, RMG385, was confirmed by the absence of PCR product #10. To create RMG389, RMG385 was then transformed with PCR product #11, consisting of the UR-GAA-A3 cassette together with the *TRP1* selectable marker, along with primer tails directing integration to the non-essential arm of chromosome V, just centromeric to the endogenous *CAN1* gene. Correct integration of the cassette was verified using PCR products #1 and #12. The *ysh1-L14S* mutant strain for the arm-loss assay, RMG391, was then created by transforming RMG389 with the PCR product #8 from RMG89, containing the *ysh1-L14S* allele together with *HIS3* selectable marker.

To replace the *URA3* promoter in the UR-GAA-A3 cassette with the Gal1-10 promoter, strains SMY732 and RMG89 were transformed with the PCR product #13, containing the GAL1-10 promoter together with the *HPH* selectable marker. Correct integration of the GAL1-10 promoter was confirmed by Sanger sequencing of PCR product #14. The resultant strains RMG431 and RMG434 contained the GAL-UR-GAA-A3 hybrid in the *YSH1* or *ysh-L14S* genetic background, respectively.

To create strains with the single UR-GAA-A3 cassette, the AD-GAA-E2 cassette in the SMY732 and RMG89 strains was replaced with the wild-type *ADE2* gene, which was PCR amplified from SMY706 (PCR product #15), resulting in strains RMG407 and RMG409, respectively. Restoration of the WT gene was confirmed by the ADE+ phenotype combined with the absence of the 240bp band in PCR product #1.

To assay elongation on the *YLR454* gene, two strains FPY01 and FPY02 were created from the SMY724 strain and RMG338, respectively. by single-step integration of a *TRP1* plasmid containing the *GAL1* promoter fused to the 5'-most 300 bp of the *YLR454w* open-reading frame at the *YLR454w* locus (Mason and Struhl, 2005).

RAD52 knockouts in strains SMY732 and RMG89 were obtained by direct replacement with the *HPH* gene obtained by PCR amplification from pAG32 (Goldstein and McCusker, 1999) (PCR product #16), resulting in strains RMG167 and RMG169, respectively. The knockouts were confirmed by the presence of the insertion and the absence of the original gene, via PCR primer pairs #17 and #18, respectively. *RAD51* knockouts were obtained in the same manner, using PCR product #19 for transformation and PCR primer pairs #20 and #21 for confirmation, resulting in strains RMG232 and RMG236.

RNH1/201 knockouts were made in two steps. First, RNH201 knockouts in strains RMG92 (wild type) and RMG89 (*ysh1*-L14S) were obtained by direct replacement with the LEU2 gene obtained by PCR amplification from pRS305 (Sikorski and Hieter, 1989) (PCR product #22), resulting in strains RMG114 and RMG112, respectively. RNH201 replacement was confirmed via PCR using primer pair #23. Next, RNH1 was knocked out in strains RMG114 and RMG112 by direct replacement with the HPH gene obtained by PCR amplification from pAG32 (Goldstein and McCusker, 1999) (PCR product #24), resulting in strains RMG118 and RMG120, respectively. RNH1 replacement was confirmed via PCR using primer pair #25.

Strains for RNH1 overexpression were generated by replacing the RNH1 promoter with the inducible MET25 promoter. pMET was amplified from pYM-N35 (Janke et al., 2004), along with the nourseothricin-resistance marker gene, using primer pair #26 containing homology to the pRNH1 region. This PCR product was used to transform strains SMY732 and RMG89, resulting in strains RMG242 and RMG247, respectively. Promoter replacement was confirmed via PCR primer pair #27.

Screening approach

First, a suitable UV dosage was chosen by measuring the rate of mutation of the *CAN1* gene under various UV doses. Strain SMY732 was resuspended in water and subjected to 0J, 0.025J, 0.05J, 0.075J and 0.1J UV radiation using a UV Stratalinker 1800 (Stratagene). Appropriate dilutions were grown on YPD and canavanine-containing media for 3 days at 30C. The resulting colonies were counted and used to estimate mutation rate as described in (Drake, 1991) (Table S1).

For the screen, strain SMY732 was spread on YPD plates and subjected to 0.1J UV radiation followed by growth for 3 days at 30°C. Dark red colonies were isolated and spread on synthetic complete media lacking uracil and containing 0.1% 5-FOA. 5-FOA-resistant colonies were checked for large-scale repeat expansions by PCR using primer set #1, as previously described (Shishkin et al., 2009). Genomic DNA was isolated from strains with high frequencies of repeat expansions, and then used for whole genome sequencing (Illumina Genome Analyzer II, 100bp paired-end reads, barcoded libraries), resulting in an average of ~80X coverage per strain. Reads were aligned to the S288C reference genome using Bowtie (Langmead et al., 2009), and variants were called using Samtools (Li et al., 2009). Variants were further analyzed for potential deleterious effects using snpEFF (Cingolani et al., 2012) and PolyPhen2 (Adzhubei et al., 2013). Strains RMG28 and RMG35, containing *ysh1-L439S* and *ysh1-L14S* variants, respectively, were subjected to further analysis.

Fluctuation Assays

Fluctuation assays were performed as previously described (Shah et al., 2014; Shishkin et al., 2009). Briefly, frozen stocks are spread for single colonies on YPD media supplemented with uracil, to ensure that cells with an inactivated *URA3* cassette will be able to grow. At least eight colonies per strain isolate are picked after ~40 hours and diluted in 200ul dH_2O. Five steps of 10-fold serial dilutions are performed. For the wild-type strain, YPD plates receive 50ul of suspended cells from the 4^h dilution and 5FOA plates receive 50ul of undiluted cells. Volumes and concentrations may be adjusted for the various mutant strains in order to plate an appropriate density of cells. Plates are grown at 30°C for three days, and colonies are counted. For 5FOA plates, colonies are examined for repeat length via PCR using primer set #1. Numbers of colonies on 5FOA are thus adjusted to reflect only expansions, and the expansion rates are calculated (Drake, 1991). Slow growing strains were given extra time to grow at all steps, as compared with the wild type strain. For the arm loss assays, selection for canavanine-resistance preceded selection for 5-FOA resistance. Absence of PCR product #1 was used to distinguish chromosomal arm loss events from *can1ura3* double point mutants.

RNH1 Overexpression

Strains overexpressing *RNH1* were obtained via replacement of the endogenous *RNH1* promoter with the inducible *MET25* promoter. Fluctuation assays were carried out as above, with the following modification: strains RMG242 (wild type) and RMG247 (*ysh1*-L14S) were pre-grown on solid media lacking the amino acids methionine and cysteine and supplemented with uracil (high expression), as well as on synthetic complete media supplemented with uracil (low expression). Concurrently, *RNH1* expression was measured under these conditions by extracting RNA and gDNA from cells on the same solid media plates and performing qPCR on each (see "Quantitative RNA Analysis" below for further details) using PCR primer pair #55. *RNH1* expression analysis was performed as described below, using primer pair #55, and was found to increase ~3-9 fold in media lacking methionine (data not shown), which is in line with published expectations (Mumberg et al., 1994).

Quantitative RNA Analysis

RNA levels were determined using the strategy described in the main text. Strains SMY732, RMG89, RMG407 and RMG409 were grown overnight in 2 ml of YPD media at 30°C, split into two tubes each of 1 ml culture and 9 ml YPD, then grown for 4 hours at either 30°C or 37°C. 1.5 ml aliquots of each were used to extract RNA and genomic DNA using the Zymo YeaStar RNA and Genomic DNA kits, respectively. RNA was then DNase treated (Turbo DNA-free kit – Thermo-Fisher), followed by cDNA generation (Superscript IV First Strand Synthesis system – Thermo-Fisher) using either random hexamers, or poly-dT primers. qPCR was then performed using SYBR Select Master Mix (Thermo-Fisher) and QuantStudio 6 Flex RTPCR system (Applied Biosystems) for both cDNA and gDNA samples with specific primer sets as noted. qPCR values from gDNA for PCR product #28 were then used for normalization of all cDNA results, on the basis that each haploid cell contains a single copy of the genome, and that the number of cells was kept consistent between the RNA and gDNA preparations. This normalization method is therefore robust to strains with RNA processing defects, or any other conditions that might alter the levels of various RNA species typically used for normalization.

RNA Pol II elongation assays

The RNA Pol II transcription elongation assay was performed as previously described (Mason and Struhl, 2005). Details particular to this experiment are described here: Briefly, strains RMG431 and RMG434 were cultured to OD 0.4 in raffinose-containing minimal medium at 30°C. Expression from the Gal1-10 promoter was induced with the addition of 2% galactose for 2.5 hours, and cultures were subsequently shifted to 37°C for 1 hour. Cells were formaldehyde-fixed two minutes after the addition of 2% glucose, or directly from the galactose culture. Fixed cells were lysed by cryo-grinding. Cross-linked chromatin was sheared to an average length of 200 bp. RNA Pol II-bound chromatin was then incubated with anti-Rpb1 CTD antibody 4H8 (Santa Cruz) for 5 hours at 4°C. Immunoprecipitated chromatin was then de-crosslinked, purified and analyzed by qPCR as described above. Pol II occupancy was normalized via PCR product #40, representing an intergenic region on chromosome V that should contain minimal bound RNA Pol II. Strains FPY01 and FPY02 were monitored at the pGAL-YLR454 locus using primer sets #41-47, while strains RMG431 and RMG434 were monitored at the Gal1-10p–UR-GAA-A3 cassette using primer sets #33-39, as well as the endogenous Gal1-10 locus, using primer sets #48-54. Each PCR locus was then normalized to itself by dividing the glucose over the galactose values.

Supplemental References

Drake, J.W. (1991). A constant rate of spontaneous mutation in DNA-based microbes. Proc Natl Acad Sci U S A *88*, 7160-7164.

Goldstein, A.L., and McCusker, J.H. (1999). Three new dominant drug resistance cassettes for gene disruption in Saccharomyces cerevisiae. Yeast *15*, 1541-1553.

Mumberg, D., Muller, R., and Funk, M. (1994). Regulatable promoters of Saccharomyces cerevisiae: comparison of transcriptional activity and their use for heterologous expression. Nucleic Acids Res *22*, 5767-5768.

Sikorski, R.S., and Hieter, P. (1989). A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae. Genetics *122*, 19-27.

Storici, F., and Resnick, M.A. (2003). Delitto perfetto targeted mutagenesis in yeast with oligonucleotides. Genet Eng (N Y) *25*, 189-207.

Addendum to Chapter 2

Further details and reflections on the screen:

As stated in the main text, the screen was originally designed to select for *trans*-modifiers of repeat expansion, and to avoid selecting for point mutations within the body of the *URA3* and *ADE2* reporter genes. It was necessary to shorten the repetitive tract in the *ADE2* reporter system to $(GAA)_{63}$, as colonies with the *ADE2* $(GAA)_{100}$ construct were already fully red-colored. However, we observed, during screening with the *ADE2* $(GAA)_{63}$ reporter, that red colonies did not contain *ADE2* expansions. In retrospect, this was not surprising, as the rate of large-scale $(GAA)_n$ repeat expansions was observed to increase exponentially with repeat length (Shah et al., 2012; Shishkin et al., 2009). $(GAA)_{52}$ expansions were below the detection limit of 1 in 10^{-8}, while $(GAA)_{78}$ expansions occurred at a rate of $4.5x10^{-6}$, and each further addition of 25 repeats resulting in a 3-4-fold increase in expansions (Shah et al., 2012). Given the exponential nature of this relationship, it is likely that the $(GAA)_{63}$ expansion rate is nearer to the rate for $(GAA)_{52}$ than $(GAA)_{78}$. The nature of this threshold may be related to the length of an Okazaki fragment or the length of DNA wrapped around a single nucleosome (Shah et al., 2012, 2014). While $(GAA)_{78}$ easily exceeds both of these lengths, $(GAA)_{63}$ is roughly equivalent or only barely exceeds them. In the future, it may be desirable to incorporate a longer $(GAA)_{100}$ repeat into a different reporter, such as GFP, to avoid these issues while retaining the convenience of a color-based screening method. Color-based reporters allow screening on the basis of a single colony, as opposed to a patch test or spot test used for growth-based reporters, thus requiring less physical labor and resources in the first round of screening.

Nevertheless, when the initial screening was performed, roughly half of the colonies chosen for red color, absent any expansions in *ADE2* $(GAA)_{63}$, were observed to contain expansions of *URA3* $(GAA)_{100}$. This suggested the possibility of a category of *trans*-modifiers that affected both repeat expansion and some form of *ADE2* gene inactivation. Indeed, mutants in *YSH1* turned out to be a perfect example of this type of modifier, with RNA processing defects affecting *ADE2* expression, and transcription elongation defects appearing to lead to DSBs within the *URA3* $(GAA)_{100}$ reporter. Further analysis of the remaining sequenced strains suggests that the *YSH1* finding may have been a very lucky one.

YSH1 mutants were found in two separate strains, suggesting a strong candidate for further analysis. This was unlikely to have happened purely by chance. Given the observed rate of ~8 non-synonymous mutations per strain, and the presence of ~6700 genes in the yeast genome, this overlap should only occur at random once after sequencing over 800 strains. This estimate could be improved by incorporating the length of each gene, the number of possible mutations resulting in non-synonymous changes, and the mutation spectrum of

UV mutagenesis, but serves as a useful rough indicator. For future rounds of screening that might involve hundreds of yeast strains, this statistic could be incorporated into the sequencing analysis. Depending on the number of possible modifiers of the *URA3* and *ADE2* reporters, the finding of two separate *YSH1* mutants in this limited pilot study may also have been a rather unlikely event.

In retrospect, the finding of two *YSH1* modifiers in this small study may have been even more unlikely due to imperfections in the screening method. 14 UV-mutagenized strains were sequenced in the initial round, nine of which showed expansions in the initial screening. Five further strains did not show expansions initially, but were sequenced in order to determine the source of *ADE2*/*URA3* gene inactivations and better understand the screening method. Table A1 contains the full list of non-synonymous SNVs detected in each strain. One further strain, which turned red in the absence of an UV mutagenesis and showed expansions in *URA3*, was also sequenced and found not to contain any SNVs. This could be due to a structural variant that was not detected in the SNV analysis, or could indicate that background fluctuations in the *ADE2*/*URA3* reporters allow the occasional strain to slip through the screening process. Furthermore, full fluctuation assays were performed for strains 75-2-7 and 100-2-9, revealing that they also slipped through the screening process. In this follow-up analysis, strain 75-2-7 showed a very high rate of 5-FOA resistance, but no longer showed any expansions in the PCR analysis **(Fig. A1)**. This is perhaps not surprising, given that the strain contains a severe mutation in the *ISY1* splicing factor, which would inactivate both the *URA3* and *ADE2* genes. Strain 100-2-9 was initially selected for its dark red color, and the observed mutation in the *MSL5* splicing factor is likely responsible for this. Interestingly, this did not seem to affect splicing in the *URA3* reporter, which showed only a slightly elevated 5-FOA resistance rate. This may reflect differences in splicing between the *ADE2* and *URA3* reporters that were also observed in *YSH1* mutant strains (See below). Expansions in 100-2-9 did not rise above the wild-type rate **(Fig. A1)**. These two examples appear to indicate that the initial method of *URA3* reporter screening, involving patching cells onto a plate and testing a few large colonies via PCR, was not stringent enough.

Among those strains that did not show expansions in the initial analysis, the common thread appears to be mutations in splicing factors and other transcriptional modifiers **(Table A1)**. Among the total list of SNVs found across all strains, both expanding and not, Gene Ontology terms involving RNA splicing and transcription were heavily represented **(Table A2)**. Again, this is not surprising in retrospect, given that both the *ADE2* and *URA3* reporters contain introns. To avoid this concentration, future rounds of screening may incorporate the *CAN1* $(GAA)_{100}$ reporter, in which the repeats are not transcribed and the reporter does not contain an intron. However, it should be noted that *YSH1* mutants did not elevate expansions in the *CAN1* reporter, and thus modifiers in this category would be missed in such a screen.

Of those sequenced strains not yet subjected to further analysis, there may be several interesting candidates for *trans*-modifiers. *PIF1*, whose role in break-induced replication is well-studied, is the likeliest culprit for expansions in strain 100-4-7. The mutation W568* truncates the protein to 568/859 amino acids, which appears to remove one NTP-hydrolase domain while leaving the helicase domain intact. It has been shown that the loss of *PIF1* affects large-scale $(CAG)_n$ expansions through its role in BIR. It would be interesting to show the same for $(GAA)_n$ repeats. Additionally, much of *PIF1* function is studied either with a full knockout, which results in sick yeast strains due to the role of *PIF1* in mitochondrial genome stability, or the *PIF1*-m2 mutant allele, which limits the nuclear expression of *PIF1* but typically has a weaker effect than the full knockout. This new mutant allele may be easier to work with, or have unexpected effects due to the partial truncation. Another interesting *trans*-modifier candidate is the *SIN3* E398* mutation found in strain 100-11-4. The histone deacetylase *SIN3* has been shown to have a role in small-scale $(CTG)_n$ expansions in yeast via interaction with the Sae2 nuclease (Debacker et al., 2012), but has not been studied for its role in large-scale $(GAA)_n$ repeats. Additionally, the strain 100-4-8, which did not show expansions in the initial screening, contains a *SIN3* M1I start loss mutation. The next closest start codon in the *SIN3* sequence would produce a short frame-shifted protein, thus the M1I mutation likely serves as a total loss-of-function variant. The existence of two mutants with different behaviors may help to elucidate the role of *SIN3* in $(GAA)_n$ repeat expansion.

Analysis of *PRP19* mutants:

Strain 100-10-11 was found to contain a severe mutation, L84S, in the essential splicing factor gene *PRP19*. This strain did not initially display expansions in the screen. However, this gene was of interest due to a recently uncovered connection to the DNA Damage Response (DDR) found in humans, in which *PRP19* appeared to be acting outside of its role in splicing (Maréchal et al., 2014). Through alignment of the yeast and human sequences, the human Y405A mutation, controlling this response, was determined to be homologous to the Y387 position in yeast. Thus, we generated yeast strains containing either the *PRP19* L84S mutation, found in this study, or the Y387A mutation. Upon full fluctuation analysis, the UV-mutagenized strain 100-10-11 was found to have an extremely high rate of 5-FOA resistance in the *URA3* reporter, but no expansions were found in the subsequent PCR analysis **(Fig. A1)**. Strains with the L84S mutation generated in the *CAN1* reporter background were shown to have an expansion rate equivalent to wild-type **(Fig. A1)**. Finally, strains with the Y387A mutation generated in the *URA3* reporter background were also shown to have an expansion rate equivalent to wild-type **(Fig. A1)**. Interestingly, this mutation did not increase 5-FOA resistance, indicating that this amino acid position, which appears to be conserved through to humans, is not important for splicing. This could potentially indicate that the connection to the DDR also exists in yeast, but is not important for expansions, at least in the given genetic background. This is perhaps not surprising, given that knocking out the DDR factor *MEC1* also did not affect expansions in a wild-type background (See below). Perhaps in a background with high levels of DSBs and/or a deficiency of certain repair factors, both of these factors could be come important for expansions.

Additional analysis of *YSH1* transcriptional effects:

Ysh1 plays a critical role in mRNA processing **(Fig. A2)**. Ysh1 is part of the Cleavage and Polyadenylation Specificity Factor (CPSF) complex, which recognizes one of two polyadenylation signals within the elongating pre-mRNA. A second signal is recognized by the CstF complex, and together they position Ysh1 at the appropriate cut site for poly-A tail addition. The Ysh1 endonuclease then cuts the pre-mRNA, allowing Poly-A Polymerase to add the A-tail to the pre-mRNA. On the other end of the cut, Rat1 exonuclease is loaded onto the RNA, which then chews back towards RNA Pol II, eventually displacing the transcriptional complex from the template DNA. In vitro, Ysh1 is also important for the poly-A addition step, given a pre-cleaved substrate, possibly via a structural role in the CPSF complex. Thus, cleavage of the transcript by Ysh1 is critical for both polyadenylation and transcription termination. (See main text for additional details and references.)

More indirectly, Ysh1 appears also to be involved in both transcription elongation and splicing. The role in elongation appears due to transcriptional pausing that occurs after positioning of the CPSF complex at the signal site and before cleavage by Ysh1 (Nag et al., 2007; Schaughency et al., 2014). (See main text for additional details and references.) We observed slowing or stalling of transcription elongation in our *ysh1* L14S mutant strain by means of a RNA polymerase clearance assay (**Figs. 6A and S3**). This assay uses the *GAL1-10* promoter, which is induced in the presence of galactose and repressed in the presence of glucose. Cells are initially grown in galactose, allowing RNA Pol II transcription initiation at a gene of interest. Addition of glucose halts transcription initiation, but elongating RNA Pol II complexes continue moving along the gene. By using ChIP to measure RNA Pol II occupation at various positions along the gene, this serves as a measurement of elongation speed (Mason and Struhl, 2005). For instance, if Pol II occupancy is low (matching glucose-only levels) at the 500 bp position but high (matching galactose-only levels) at the 1000 bp position after one minute of glucose exposure, RNA polymerase speed can be approximated at 500-1000 bp per minute. The precision of this measurement can be improved by measuring additional positions within the gene. We observed that the *ysh1* L14S mutant strain exhibited slow movement through the *URA3* reporter, as well as the original *GAL1* and *GAL10* genes and an additional 8 kb-long gene, *YLR454* **(Figs. 6A, S3)**. In the latter case, it is possible to see the point where Pol II occupancy reaches 100% of galactose levels, indicating that Pol II in the *ysh1* L14S background traveled 1-2 kb in two minutes on average, as opposed to Pol II in the wild-type background, which traveled >8 kb in the same time. It also appears that a greater portion of Pol II remained in the 0.5-1 kb region in the *ysh1* mutant (~50% occupancy) compared to the wild-type (~20% occupancy). This is similar to the levels observed across the *URA3* reporter, which is only 2 kb in length. This observation likely represents Pol II complexes moving slower than the average, either due to outliers with a constant lower speed, or due to inconsistent movement, ie. transient RNA Pol II stalling. This is apparent as a slow, gradual trend in the wild-type, reaching 40% occupancy towards the end of *YLR454*, but is more severe at earlier positions in the *ysh1* mutant background. Thus, the *ysh1* L14S mutant appears to slow transcription elongation overall, and may also cause Pol II stalling.

The role of Ysh1 in splicing is less clear, but may stem from the same effects on transcription elongation. One theory states that, because splicing and 3' processing both occur co-transcriptionally, they compete with one another (Moehle et al., 2014). For example, an inefficient splicing reaction may not complete before 3' processing occurs, resulting in transcription termination and the dissolution of the splicing complex. As alluded to in the main text, we likely observed the opposite taking place in the *ysh1* mutant strains. Splicing of the *URA3* $(GAA)_{100}$ reporter's intron is normally very inefficient **(Fig. S2)**. However, in combination with the inefficient 3' processing in the *ysh1* L14S mutant background, a greater portion of *URA3* transcripts are spliced. Likely what has occurred is not an increase in splicing efficiency *per se*, but a greater allotment of time for splicing to complete before the process is disrupted by transcription termination. We do not see the same effect in the *ADE2* reporter, which contains a shorter $(GAA)_{63}$ repeat, or *DBP2*, a gene containing a very long natural intron without $(GAA)_n$ repeats **(Figs. 4 and S2)**. In the latter case, splicing efficiency in the wild-type background is very high despite the length of the intron, which is the longest in the yeast genome and slightly longer than the artificial intron in the *URA3* reporter. Thus it is likely that $(GAA)_n$ repeats have a detrimental effect on splicing efficiency, possibly due to secondary structure formation.

Additional analysis of gene knockouts in a *ysh1* mutant background:

In addition to those included in the main text, a number of additional gene knockouts were generated in the *ysh1* L14S mutant background. In some cases, the data was not included in the main text because the results were negative, ie. not able to identify the pathway leading to expansions, or the interpretation of the data was unclear. Some of the following data sets are included with the caveat that they do not have the sufficient number of replicates necessary for a high-confidence analysis. In addition, T-tests were performed for all of the following data, yet even where a significant difference is shown, a change of less than 3-fold is not generally considered biologically significant. This is because the expansion rates can fluctuate by this amount from test to test, as can be seen in the wild-type and *ysh1* L14S strains that appear in multiple figures below. For this same reason, each data set includes a matched control of the wild-type and *ysh1* L14S strains performed at the same time as the mutants, or in close proximity. This may reduce some amount of fluctuation due to slight differences in the media used for each assay, but cannot account for natural rate fluctuations.

TOP1 is the yeast type 1 topoisomerase, which makes a single strand nick to relax DNA supercoils. Knockout of TOP1 on its own appears to lead to a small increase in expansions, which is statistically but not necessarily biologically significant. However, in a *ysh1* L14S background, expansions appear to decrease **(Fig. A3)**. This was an unexpected result. It was initially predicted that the lack of Top1 would increase the severity of transcription-replication collisions, leading to further DNA breaks and repeat expansions. An alternative interpretation is that, without Top1, transcription-replication collisions result in cell death, leading to the apparent loss of expansions. Another possibility is that Top1 is a source of DSBs in the *ysh1* L14S background, if it creates DNA nicks within the $(GAA)_n$ repeats that cannot be repaired due to triplex formation. Further experiments will be required to confirm and uncover Top1's role. Knockout of *MPH1* on its own significantly increased expansions, which increases further in the *ysh1* L14S background **(Fig. A3)**.

MPH1 is the homolog of the human FANCM helicase, and appears to have multiple functions including the suppression of break-induced replication (BIR) (Štafa et al., 2014). Importantly, it appears that the increase in expansions seen in the *ysh1* L14S mutant background is synergistic, rather than additive, in combination with the *MPH1* knockout. This would be consistent with Mph1's role in suppressing BIR. In the absence of Mph1, DSBs resulting from transcription-replication collisions would be more likely to initiate BIR, potentially leading to expansions, rather than non-homologous end joining (NHEJ) or single-strand annealing (SSA), which would tend to generate repeat contractions, or gap repair, which could lead to correct repair if the flanking non-repetitive sequences were used in the strand exchange. However, Mph1 also appears to have a role in promoting Okazaki fragment flap processing by Rad27/FEN1, the knockout of which is also known to greatly increase large scale $(GAA)_n$ expansions (Kang et al., 2009; Tsutakawa et al., 2017) Thus, further investigation will be required to distinguish between these two roles.

In the main text, we show that *ysh1*-driven expansions are dependent on *RAD52* but not *RAD51*, which potentially implicates the involvement of SSA or a *RAD51*-independent branch of BIR. On its own, SSA is more likely to result in repeat contractions, as it does not involve DNA synthesis. However, SSA may also be involved in the initiation of *RAD51*-independent BIR events by connecting short homologous sequences, in this case the $(GAA)_n$ repeats, as opposed to using longer *RAD51*-dependent homologies (Signon et al., 2001). This pathway also appears partially dependent on *RAD50*, part of the MRX end-resection complex (Mre11-Rad50-Xrs2) (Signon et al., 2001). We did not knockout *RAD50* in this study, but we did see that *MRE11* does not appear to affect the rate of *ysh1*-driven expansions **(Fig. A3)**. However, knockout of *MRE11* has also been shown to promote BIR through a *RAD51*-dependent pathway (Krishna et al., 2007). It is possible that the *RAD51*-dependent and -independent BIR pathways can compensate for one another to generate *ysh1*-driven expansions. Altogether, BIR is a strong candidate for the generation of *ysh1*-driven expansions. In the future, testing the effects of a double knockout of *RAD51* and *MRE11,* as well as *POL32* or *PIF1* knockouts, would help to confirm this hypothesis.

Several other knockouts did not produce significant changes in the rate of *ysh1*-driven expansions. Knockout of *MEC1*, the DNA damage checkpoint activator, had no effect on expansions, either on its own or in the *ysh1* L14S background **(Fig. A4)**. Note that *MEC1* knockouts are not viable, except in a *sml1Δ* background. Thus the DNA Damage Response (DDR) is either not involved in expansions, or the multiple effects of *MEC1* absence cancel each other out. Knockout of *RAD5*, involved in template switch during post-replicative repair, showed a small decrease in *ysh1*-driven expansions **(Fig. A5)**. This effect was statistically significant, but is unlikely to be biologically significant given its small effect size. Knockout of *DST1*, the yeast TFIIS transcription factor, which aids transcription elongation by cleaving transcripts at RNA Pol II stalls, did not alter expansion rates on its own, or combined with *ysh1* L14S **(Fig. A6)**. Likewise, knockout of *ELC1*, which is involved in degradation of the RNA Pol II complex during transcriptional stalls, also had no effect on expansions in either background **(Fig. A7)**. These latter two knockouts may suggest that it is primarily slow, rather than stalled, transcription that leads to transcription-replication collisions in the *ysh1* background.

5-FOA resistance in the *ysh1* mutant background

Upon performing fluctuation assays in the *ysh1* mutant strains using the *URA3* $(GAA)_{100}$ reporter, it was observed that there were two major classes of 5-FOA-resistant colonies. The first category consisted of large colonies (~2-3 mm in diameter), the majority of which were found to be caused by large-scale expansions, typical of our previous experience with the *URA3* $(GAA)_{100}$ reporter. The second category consisted of much smaller colonies (>1 mm in diameter). Smaller colonies can be observed in the wild-type background, though they are relatively rare and contain a larger percentage of expansions. However, in the *ysh1* mutant background, they appeared at a very high rate, and with a low percentage of expansions **(Fig. A8)**. The rate of small colony expansions in the wild type background is similar or slightly higher than the expansion rate for large colonies. (See main text for comparison.) We would expect to see this rate of expansion if cells plated on 5-FOA were able to survive for an additional one or rarely two generations. The toxic effects of the 5-FOA prior to acquisition of the expanded $(GAA)_n$ repeats are likely responsible for lagging growth of the colony. In contrast, the small colony expansion rate for the *ysh1* L14S mutant is at least twice as high as in the large colonies. This is indicative that the cells can survive 5-FOA exposure for an additional two or more generations. Additionally, the presence of numerous small colonies lacking any expansion suggests that a large portion of cells can survive indefinitely on 5-FOA, albeit at a reduced growth rate.

We sought to determine the source of this 5-FOA resistance in the *ysh1*-L14S background. As was shown in Figure S2, *URA3* mRNA levels could not explain 5-FOA resistance, and actually showed an increase in splicing efficiency, which should otherwise lead to increased 5-FOA sensitivity. After exploring RNA levels of additional genes in the uracil biosynthesis pathway (data not shown), we instead turned out attention toward the uracil transporter gene, *FUR4*. This gene was named for its fluorouracil-resistance phenotype, yet it has never been determined how 5-FOA, analog of an upstream component in uracil biosynthesis, might enter the cell. We observed a large increase in read-through transcription past the annotated *FUR4* poly-A site, very likely due to the *ysh1* RNA processing defects **(Fig. A9)**. This would account for a large drop in *FUR4* mRNA levels in the *ysh1* L14S mutant strains at the temperature-sensitive 37°C growth condition **(Fig. A9)**. We then knocked out *FUR4* in the wild-type background and observed a large increase in 5-FOA-resistant small colonies, very similar to what was observed in the ysh1 *L14S* mutant **(Fig. A8)**. Thus, lack of polyadenylated *FUR4* is likely responsible for a large portion of 5FOA-resistant small colonies in the *ysh1* mutant background. However, we observed an even higher rate of 5-FOA-resistant small colonies in the *ysh1* L14S mutant when grown at 37°C, exceeding that of the *FUR4* knockout **(Fig. A8)**. This indicates that there are likely additional transporters of 5-FOA that are also affected by RNA processing defects in the *ysh1* mutant background.

Methods:

All experiments were performed as described in the main text.

Acknowledgements:

Rachel Rubinstein generated *PRP19* mutant strains and performed fluctuation assays.

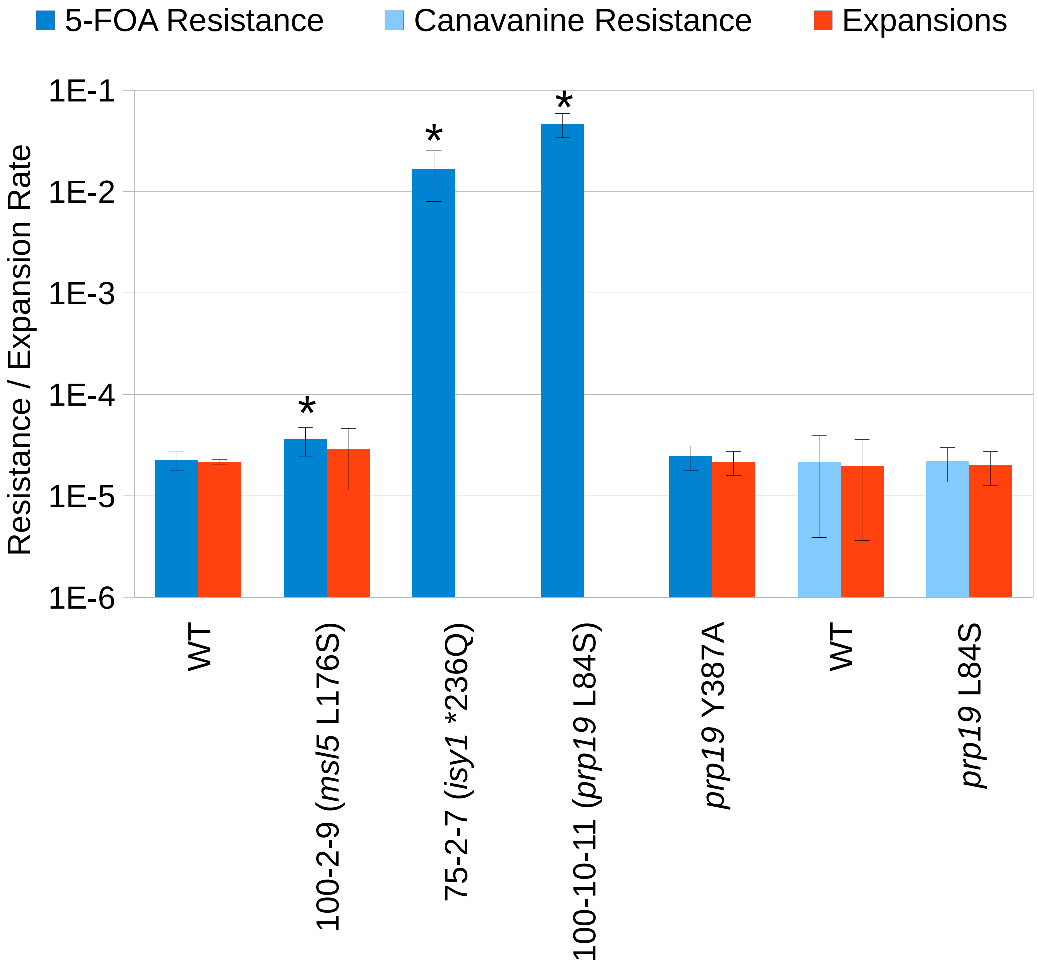


Fig. A1: Fluctuation assays performed for various mutant strains

Fluctuation assays were performed in strains containing either the *URA3* $(GAA)_{100}$ reporter, or the UAS-$(GAA)_{100}$-GAL-*CAN1* reporter, where indicated. UV mutagenized strains are highlighted for their most-likely-damaging mutation, in parentheses. Mutations not listed in parenthesis were created by site-specific mutagenesis in the wild-type background. For strains 75-2-7 and 100-10-11, no expansions were detected upon PCR analysis of >100 5-FOA-resistant colonies. * indicates a statistically significant difference from the wild-type strain (T-test P value of <0.05). Error bars indicate 95% confidence intervals.

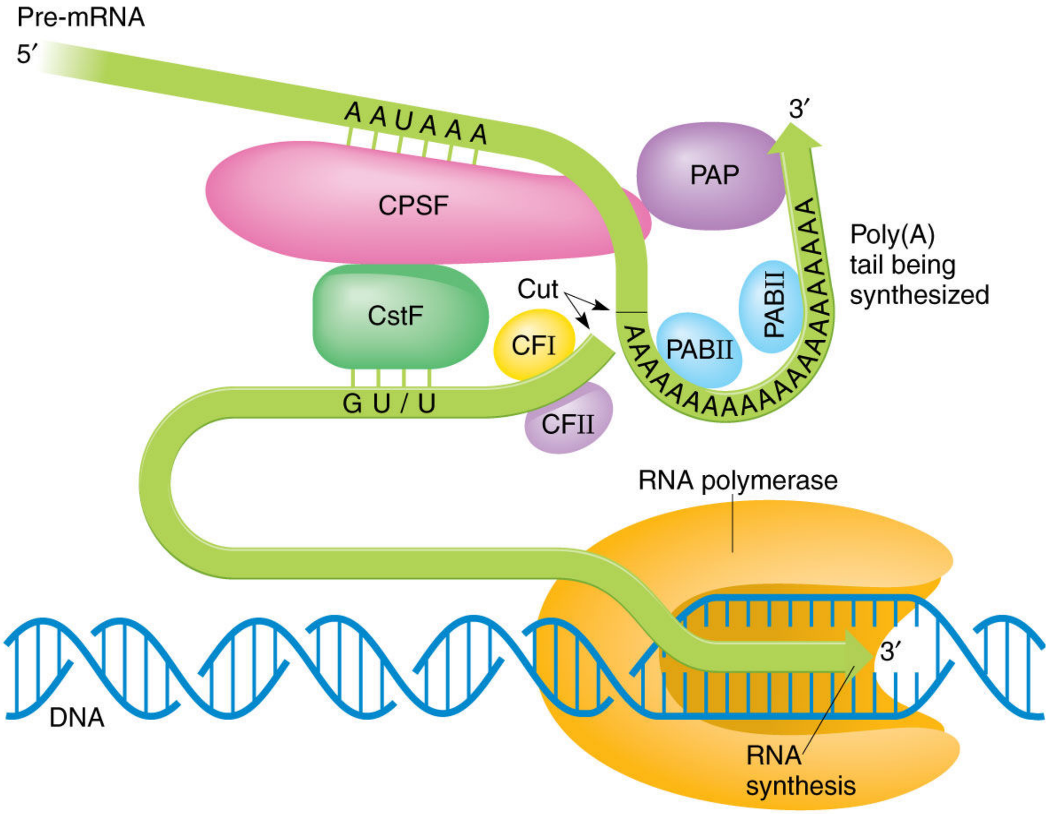


Fig. A2: mRNA 3' Processing Overview

See Addendum text for details. Figure obtained from "iGenetics: A Molecular Approach." 3rd Edition, Pg. 92. Peter J. Russell. 2010 Pearson Education.

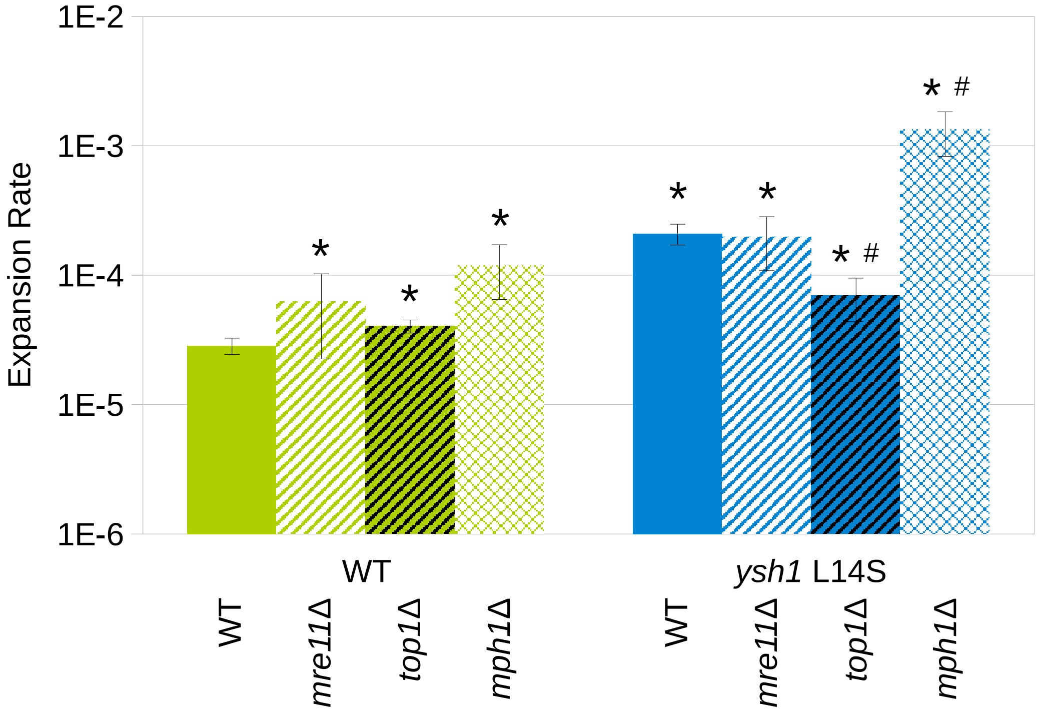


Fig. A3: Fluctuation assays performed for various knockout strains

Fluctuation assays were performed in strains containing the *URA3* (GAA)$_{100}$ reporter. Knockouts were generated in both a wild-type and *ysh1* L14S background. Fluctuation assays were performed concurrently. * indicates a statistically significant difference from the wild-type strain (T-test P value of <0.05). # indicates a statistically significant difference from the *ysh1* L14S strain (T-test P value of <0.05). Error bars indicate 95% confidence intervals.

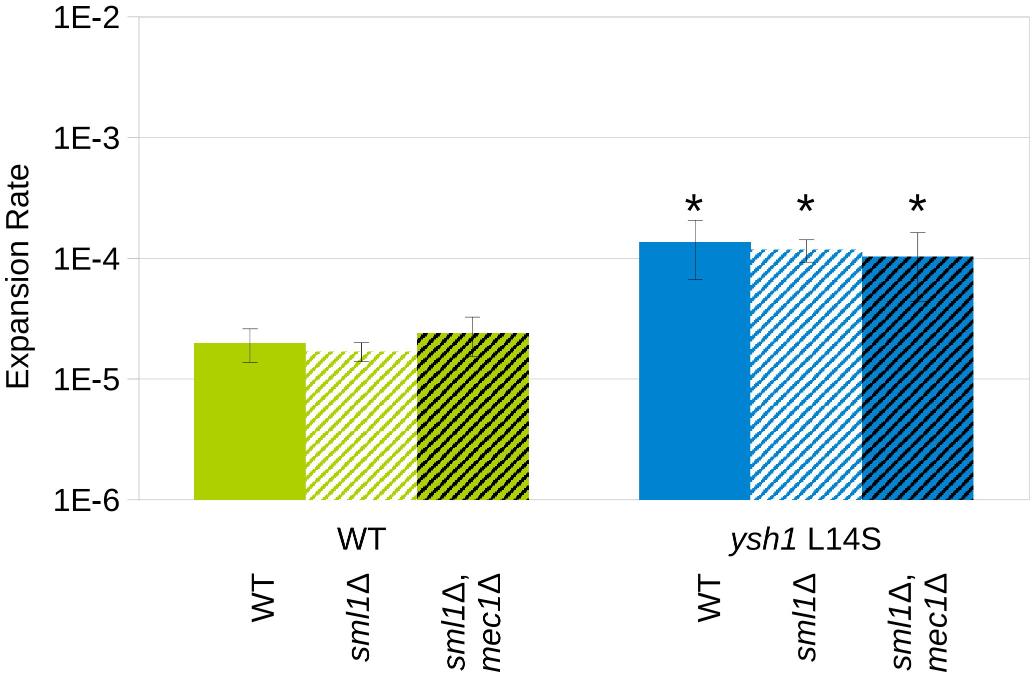


Fig. A4: Fluctuation assays performed for *SML1*/*MEC1* knockout strains

Fluctuation assays were performed in strains containing the *URA3* $(GAA)_{100}$ reporter. Knockouts were generated in both a wild-type and *ysh1* L14S background. Fluctuation assays were performed concurrently. * indicates a statistically significant difference from the wild-type strain (T-test P value of <0.05). Error bars indicate 95% confidence intervals.

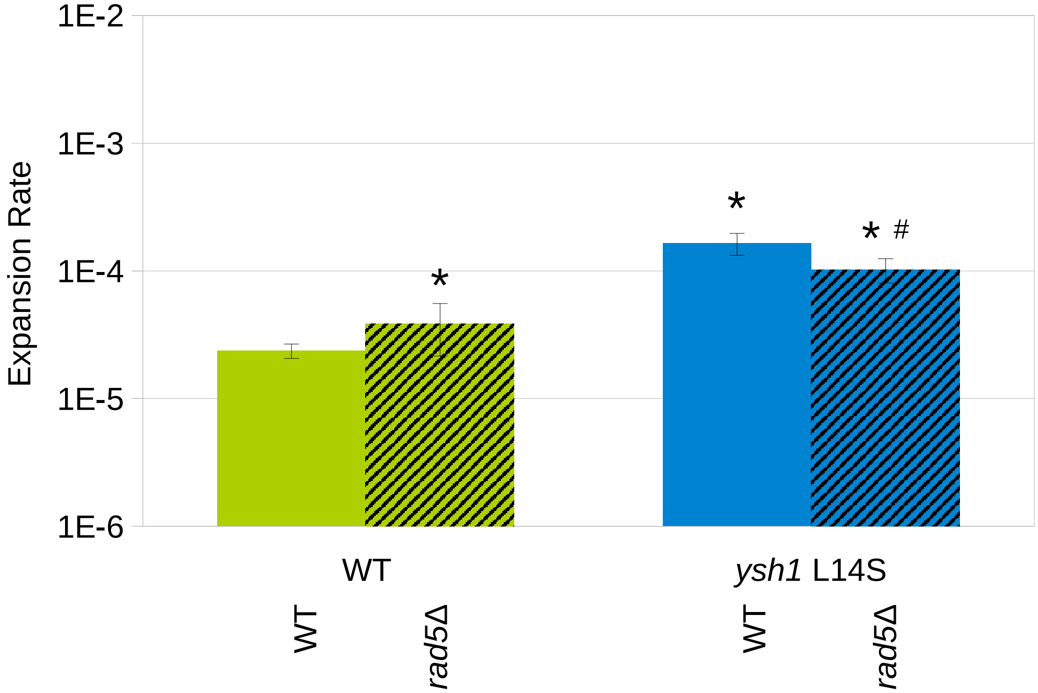


Fig. A5: Fluctuation assays performed for *RAD5* knockout strains

Fluctuation assays were performed in strains containing the *URA3* $(GAA)_{100}$ reporter. Knockouts were generated in both a wild-type and *ysh1* L14S background. Fluctuation assays were performed concurrently. * indicates a statistically significant difference from the wild-type strain (T-test P value of <0.05). # indicates a statistically significant difference from the *ysh1* L14S strain (T-test P value of <0.05). Note that the effect sizes are small, and may not be biologically significant. Error bars indicate 95% confidence intervals.

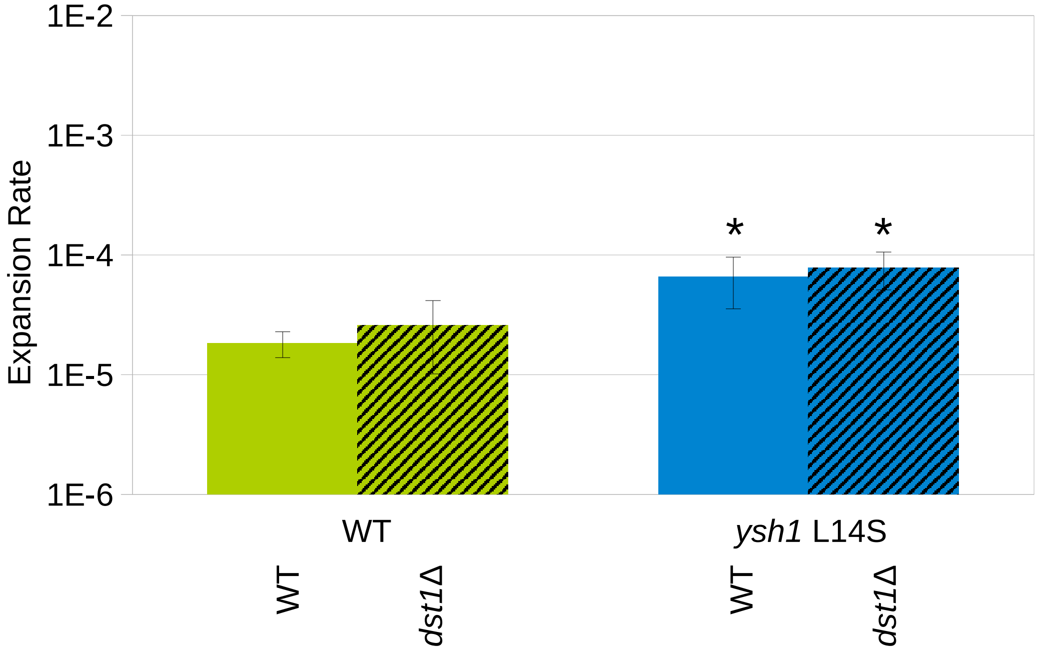


Fig. A6: Fluctuation assays performed for *DST1* knockout strains

Fluctuation assays were performed in strains containing the *URA3* $(GAA)_{100}$ reporter. Knockouts were generated in both a wild-type and *ysh1* L14S background. Fluctuation assays were performed concurrently. * indicates a statistically significant difference from the wild-type strain (T-test P value of <0.05). Error bars indicate 95% confidence intervals.

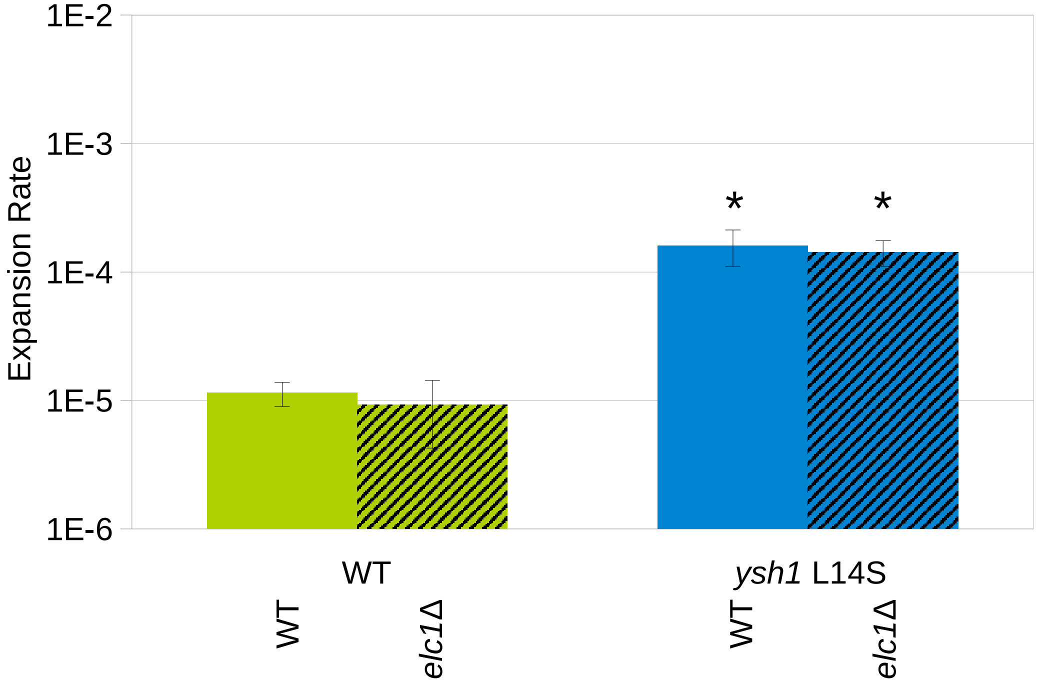


Fig. A7: Fluctuation assays performed for *ELC1* knockout strains

Fluctuation assays were performed in strains containing the *URA3* $(GAA)_{100}$ reporter. Knockouts were generated in both a wild-type and *ysh1* L14S background. Fluctuation assays were performed concurrently. * indicates a statistically significant difference from the wild-type strain (T-test P value of <0.05). Error bars indicate 95% confidence intervals.

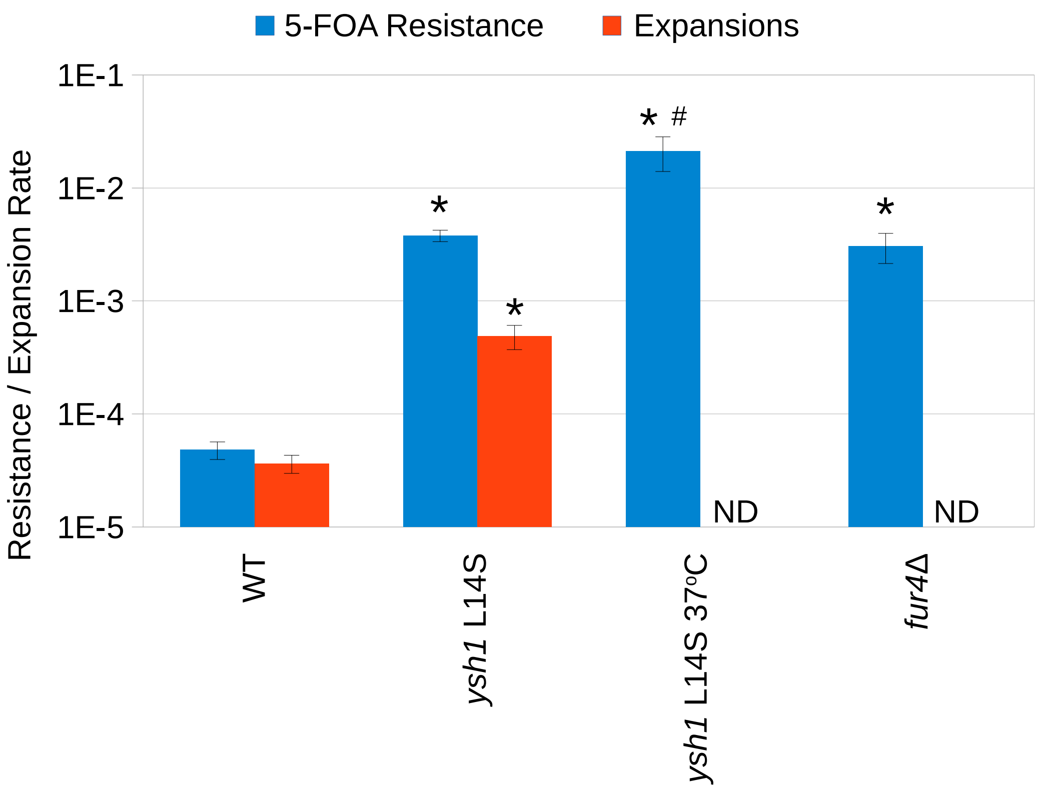


Fig. A8: Analysis of 5-FOA-resistant small colonies

Fluctuation assays were performed in strains containing the *URA3* $(GAA)_{100}$ reporter. However, only small colonies (<1 mm in diameter) were counted and analyzed. Cells were pre-grown at 30°C, except where indicated. For the latter two conditions, PCR analysis of the 5-FOA-resistant colonies was not performed. * indicates a statistically significant difference from the wild-type strain (T-test P value of <0.05). Error bars indicate 95% confidence intervals.

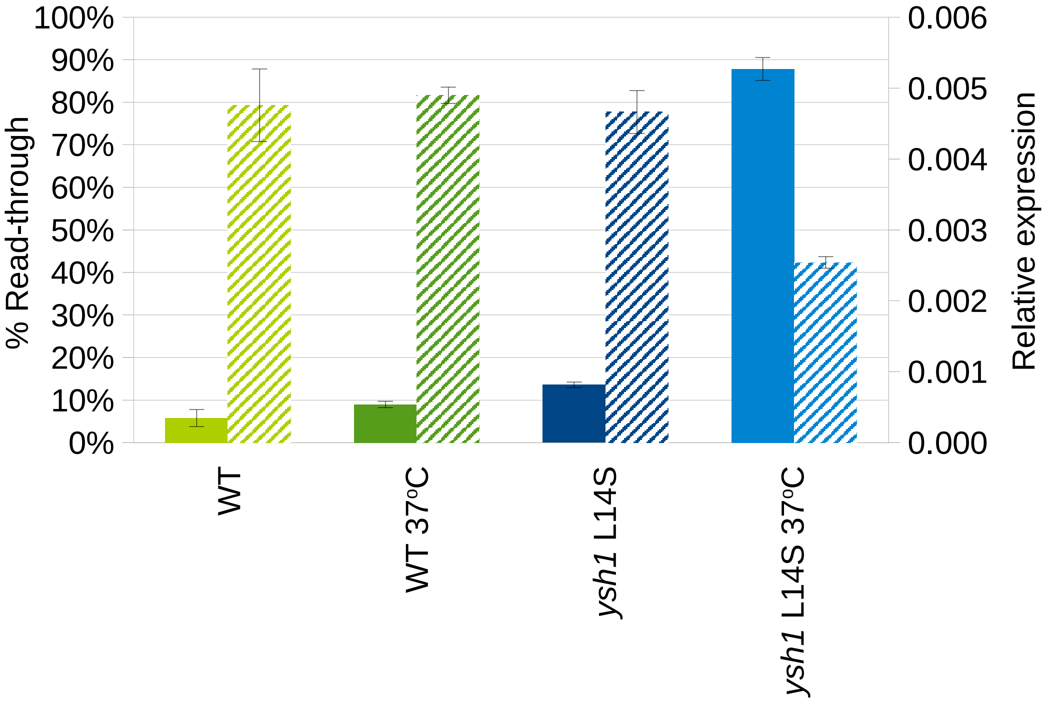


Fig. A9: RNA analysis of the *FUR4* gene

RT-qPCR analysis was performed in strains containing *ysh1* L14S mutation. Cells were pre-grown at 30ºC, except where indicated. Two PCR primer sets were used to measure RNA levels. The first measures overall RNA levels at the 3' end of *FUR4*, prior to the annotated poly-A site, while the second measures overall RNA levels after the annotated poly-A site. The solid bars represent the second primer set divided by the first, to obtain a percentage of transcripts reading past the poly-A site. This percentage was then multiplied by the results of the first primer set, to approximate the amount of functional mRNA, represented by the striped bars. Error bars indicate standard deviation.

Table A1: Non-synonymous mutations in sequenced strains

Strains with expansions seen in initial spot check:

Gene	Variant	GO Process Terms	Human Homologs / Orthologs	Polyphen 2 score	Notes
Strain: 100-11-3					
YSH1	L14S	mRNA polyadenylation, mRNA processing, pre-mRNA cleavage required for polyadenylation, RNA phosphodiester bond hydrolysis - endonucleolytic, RNA splicing, snoRNA 3'-end processing, snoRNA splicing, termination of RNA polymerase II transcription	CPSF3, CPSF3L	1	
VHR2	L265S	regulation of transcription - DNA-templated, transcription – DNA-templated			
AKL1	S389F	actin cortical patch assembly, actin cytoskeleton organization, actin filament organization, phosphorylation, protein phosphorylation, regulation of endocytosis	AAK1, BMP2K		
SPE3	N62D	pantothenate biosynthetic process, polyamine biosynthetic process, polyamine metabolic process, spermidine biosynthetic process	SMS, SRM		
ATG19	L218F	Autophagy, CVT pathway, ER-associated ubiquitin-dependent protein catabolic process, protein complex localization, protein processing, protein transport, transport, vesicle organization			
FOX2	D60Y	fatty acid beta-oxidation, fatty acid metabolic process, lipid metabolic process, metabolic process, oxidation-reduction process	DHRS9, DHRS2, DHRS4, DHRS1, SDR9C7, SCP2D1, RDH12, RDH10, DECR1, SDR16C5, DHRS7C, DHRS7B, DECR2, HSD17B10, DHRS4L2, HPGD, HSD11B1, HSD11B2, HSD17B1, HSD17B3, HSD17B2, HSD17B4, HSD17B13, HSD11B1L, RDH8, RDH11, HSD17B12, HSD17B11, DCXR, HSD17B7, DHRS7, WWOX, RDH5, BDH1, SCP2, SPR, DHRS4L1, DHRS11, HSD17B8, DHRS12, HSDL1, HSDL2, CBR4, RDH16, HSD17B6, CBR1, CBR3, DHRS3		
TOM70	A544V	intracellular protein transport, protein import into mitochondrial inner membrane, protein import into mitochondrial matrix, protein targeting to mitochondrion	STUB1, UNC45B, UNC45A, TTC1, TOMM70A, TOMM70A		
PIL1	L181S	eisosome assembly, endocytosis, negative regulation of protein kinase activity, protein localization, protein localization to eisosome filament, response to heat			
Strain: 100-8-2					
YSH1	S59F, L439S	mRNA polyadenylation, mRNA processing, pre-mRNA cleavage required for polyadenylation, RNA phosphodiester bond hydrolysis - endonucleolytic, RNA splicing, snoRNA 3'-end processing, snoRNA splicing, termination of RNA polymerase II transcription	CPSF3, CPSF3L	1	
SYF2	P83L	mRNA processing, mRNA splicing - via spliceosome, RNA splicing	SYF2		
ELG1	K599N	cell cycle, DNA clamp unloading, DNA-dependent DNA replication, double-strand break repair via homologous recombination, mitotic sister chromatid cohesion, negative regulation of DNA recombination, negative regulation of transposition - RNA-mediated, telomere maintenance	ATAD5		
AHK1	L376S	negative regulation of MAP kinase activity, osmosensory signaling pathway via Sho1 osmosensor, positive regulation of signal transduction			
YAP1801	N387H	clathrin coat assembly, Endocytosis	PICALM, SNAP91		
SIP3	N103D	intracellular sterol transport, positive regulation of transcription from RNA polymerase II promoter	CCNYL1, CCNY, CCNYL2, CCNYL3		
MCP2	I329N, F330L	lipid homeostasis, mitochondrion organization	ADCK5, ADCK1		
YLR236C	E12K	biological process unknown			
YPR078C	D120A	biological process unknown			
Strain: 100-4-7					
PIF1	W568*	cellular response to DNA damage stimulus, chromosome organization, DNA duplex unwinding, DNA recombination, DNA repair, DNA strand renaturation, double-strand break repair via break-induced replication, G-quadruplex DNA unwinding, mitochondrial genome maintenance, negative regulation of telomerase activity, negative regulation of telomere maintenance via telomerase, replication fork reversal, telomere maintenance, telomere maintenance via recombination	PIF1		568/859 amino acid truncation
PAB1	*578Q	mRNA processing, mRNA transport, negative regulation of catalytic activity, regulation of nuclear-transcribed mRNA poly(A) tail shortening, regulation of translation, regulation of translational initiation, Transport	PABPC4L, PABPC5, ELAVL2, ELAVL1, ELAVL3, ELAVL4, PABPC1, RBMS3, PABPC1L2A, PABPC3, RBMS1, RBMS2, PABPC1L2B, PABPC1L, PABPC4		Adds 4 amino acids
MSH2	L416S	cellular response to DNA damage stimulus, chromatin silencing at silent mating-type cassette, DNA recombination, DNA repair, interstrand cross-link repair, maintenance of DNA repeat elements, meiotic gene conversion, meiotic mismatch repair, mismatch repair, mitotic recombination, negative regulation of reciprocal meiotic recombination, postreplication repair, removal of nonhomologous ends, replication fork arrest	MSH2		non-conserved AA in humans
GCN20	L45S	regulation of translation, regulation of translational elongation	ABCF1, ABCF3		
ADE12	K88*	'de novo' AMP biosynthetic process, IMP metabolic process, purine nucleotide biosynthetic process	ADSSL1, ADSS		88/433 amino acid truncation
UTP13	L278P, K279Q	endonucleolytic cleavage in 5'-ETS of tricistronic rRNA transcript (SSU-rRNA - 5.8S rRNA - LSU-rRNA), endonucleolytic cleavage in ITS1 to separate SSU-rRNA from 5.8S rRNA and LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA - 5.8S rRNA - LSU-rRNA), endonucleolytic cleavage to generate mature 5'-end of SSU-rRNA from (SSU-rRNA - 5.8S rRNA - LSU-rRNA), maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA - 5.8S rRNA - LSU-rRNA), ribosome biogenesis, rRNA processing	TBL3		
YNL140C	R166K	biological process unknown			
SKI7	I233F	nonfunctional rRNA decay, nuclear-transcribed mRNA catabolic process - 3'-5' exonucleolytic nonsense-mediated decay, nuclear-transcribed mRNA catabolic process - exonucleolytic - 3'-5', nuclear-transcribed mRNA catabolic process - nonsense-mediated decay, nuclear-transcribed mRNA catabolic process - non-stop decay, protein catabolic process, regulation of translation, Translation	MTIF2		
YDR444W	P564S	biological process unknown, cellular lipid metabolic process, lipid catabolic process, lipid metabolic process			
YLL037W	L20P	biological process unknown			
GEA2	S352L	actin cytoskeleton organization, ER to Golgi vesicle-mediated transport, intra-Golgi vesicle-mediated transport, macroautophagy, positive regulation of GTPase activity, regulation of ARF protein signal transduction, retrograde vesicle-mediated transport - Golgi to ER, secretory granule organization	GBF1		

Gene	Variant	GO Process Terms	Human Homologs / Orthologs	Polyphen 2 score	Notes
ATP2	R370I	ATP biosynthetic process, ATP hydrolysis coupled cation transmembrane transport, ATP metabolic process, ATP synthesis coupled proton transport, ion transport, proton transport, Transport	ATP5B		
Strain: 100-11-4					
SIN3	E398*	cell cycle, cell division, covalent chromatin modification, double-strand break repair via nonhomologous end joining, histone deacetylation, negative regulation of chromatin silencing at rDNA, negative regulation of chromatin silencing at silent mating-type cassette, negative regulation of chromatin silencing at telomere, negative regulation of transcription from RNA polymerase II promoter, negative regulation of transcription from RNA polymerase I promoter, negative regulation of transcription involved in meiotic cell cycle, positive regulation of transcription from RNA polymerase II promoter, positive regulation of transcription from RNA polymerase II promoter in response to heat stress, regulation of DNA-dependent DNA replication initiation, regulation of transcription - DNA-templated, regulation of transcription involved in G2/M transition of mitotic cell cycle, transcription - DNA-templated, transfer RNA gene-mediated silencing	SIN3B, SIN3A		398/1536 amino acid truncation
SAL1	R190I	ADP transport, ATP transport, mitochondrial transport, transmembrane transport, Transport	SLC25A25, SLC25A41, SLC25A24, SLC25A23		
MFG1	S432F	biological process unknown, regulation of transcription - DNA-templated, transcription - DNA-templated			
Strain: 75-2-7					
ISY1	*236Q	generation of catalytic spliceosome for second transesterification step, mRNA 3'-splice site recognition, mRNA processing, RNA splicing	RAB43, ISY1		Adds 32 amino acids
SSH1	L44F	protein transmembrane transport, protein transport, SRP-dependent cotranslational protein targeting to membrane, Transport	SEC61A1, SEC61A2		
CWC27	D277N	biological process unknown, mRNA processing, protein peptidyl-prolyl isomerization, RNA splicing	PPIB		
Strain: 100-2-9					
MSL5	L176S	mRNA processing, mRNA splicing - via spliceosome, RNA splicing	KHDRBS3, KHDRBS1, KHDRBS2, SF1, QKI	1	
MMS22	L95S	cell cycle, cell division, cellular response to DNA damage stimulus, DNA repair, double-strand break repair, double-strand break repair via homologous recombination, meiotic sister chromatid segregation, recombinational repair, replication fork processing	MMS22L		
EBS1	L818S	DNA recombination, mRNA export from nucleus, negative regulation of translation, nuclear-transcribed mRNA catabolic process - nonsense-mediated decay, regulation of RNA stability, RNA phosphodiester bond hydrolysis, telomere maintenance via telomerase	SMG6, SMG5, SMG7		
SPT7	R1122K	cellular protein complex assembly, chromatin organization, conjugation with cellular fusion, histone acetylation, regulation of transcription - DNA-templated, transcription - DNA-templated			
CUE3	E434K	biological process unknown	ASCC2		CUE = “Coupling of Ubiquitin conjugation to ER degradation”
SKN1	G715S	(1->6)-beta-D-glucan biosynthetic process, carbohydrate metabolic process, cell wall organization, fungal-type cell wall organization, sphingolipid biosynthetic process			
ATP17	F3I	ATP biosynthetic process, ATP hydrolysis coupled cation transmembrane transport, ATP synthesis coupled proton transport, ion transport, proton transport, Transport			
ECM3	F413L	biological process unknown, cell wall organization, transmembrane transport			
Strain: 100-9-2					
BRR2	L1022S	mRNA processing, RNA splicing, spliceosome conformational change to release U4 (or U4atac) and U1 (or U11)	SNRNP200	0.999	
ARK1	K26E	actin cortical patch assembly, actin filament organization, phosphorylation, protein phosphorylation, regulation of endocytosis	AAK1, BMP2K		
SSM4	L1231F	ER-associated ubiquitin-dependent protein catabolic process, protein ubiquitination	MARCH6, MARCH3, MARCH8, MARCH11, MARCH2, MARCH1, MARCH4, MARCH9		
BIO5	Q347K	amino acid transmembrane transport, biotin biosynthetic process, L-alpha-amino acid transmembrane transport, L-amino acid transport, transport, vitamin transport	SLC7A9, SLC7A13, SLC7A8, SLC7A11, SLC7A10, SLC7A5, SLC7A7, SLC7A6		
Strain: 100-10-9					
MSL1	L35P	mRNA processing, mRNA splicing - via spliceosome, RNA splicing	SNRPA, SNRPB2	0	
RPE1	D136V	carbohydrate metabolic process, cellular carbohydrate metabolic process, metabolic process, pentose catabolic process, pentose-phosphate shunt, pentose-phosphate shunt - non-oxidative branch	RPE, RPEL1		
LAM1	Y1066H	intracellular sterol transport	CCNYL1, CCNY, CCNYL2, CCNYL3		
RIB7	L168F	oxidation-reduction process, riboflavin biosynthetic process, RNA modification			
Strain: 100-6-5					
SAS10	P232S	cell cycle, endonucleolytic cleavage in 5'-ETS of tricistronic rRNA transcript (SSU-rRNA - 5.8S rRNA - LSU-rRNA), endonucleolytic cleavage in ITS1 to separate SSU-rRNA from 5.8S rRNA and LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA - 5.8S rRNA - LSU-rRNA), endonucleolytic cleavage to generate mature 5'-end of SSU-rRNA from (SSU-rRNA - 5.8S rRNA - LSU-rRNA), maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA - 5.8S rRNA - LSU-rRNA), ribosome biogenesis, rRNA processing	UTP3		non-conserved AA in humans
DUN1	S443F	cellular response to DNA damage stimulus, cellular response to oxidative stress, DNA damage checkpoint, intracellular signal transduction, peptidyl-serine phosphorylation, peptidyl-threonine phosphorylation, phosphorylation, protein phosphorylation, replication fork protection	CHEK2, PNCK, PRKD3, PRKD2, PRKD1, PSKH1, CAMK1D, CAMK1G, CAMKV, CAMK1, PSKH2		
GAB1	P294S	attachment of GPI anchor to protein, cell cycle, cell division, GPI anchor biosynthetic process	PIGU		non-conserved AA in humans
YSA1	L49S	ribose phosphate metabolic process	NUDT5, NUDT14		
FUS1	R404G	cortical protein anchoring, cytogamy, regulation of termination of mating projection growth			
TRM8	P183S	methylation, tRNA methylation, tRNA modification, tRNA processing	METTL1		

Gene	Variant	GO Process Terms	Human Homologs / Orthologs	Polyphen 2 score	Notes
ATG18	Y40F	autophagy, CVT pathway, late endosome to vacuole transport, late nucleophagy, macroautophagy, mitophagy, pexophagy, piecemeal microautophagy of nucleus, protein lipidation, protein localization to pre-autophagosomal structure, protein transport, transport, vacuolar protein processing	WIPI2, WIPI1		
YLF2	L152S	biological process unknown	OLA1		
YJL213W	M302L, N306S	biological process unknown, organic substance catabolic process			
ECM25	L147S	biological process unknown, cell wall organization, regulation of GTPase activity, signal transduction			
BMT6	G294E	methylation, rRNA base methylation, rRNA processing			
ATG26	L484F	ascospore-type prospore membrane assembly, carbohydrate metabolic process, lipid glycosylation, lipid metabolic process, metabolic process, protein transport, steroid biosynthetic process, steroid metabolic process, sterol biosynthetic process, sterol metabolic process, Transport			

Strains with no expansions seen in initial spot check:

Strain: 100-10-11					
PRP19	L84S	cellular response to DNA damage stimulus, DNA repair, generation of catalytic spliceosome for first transesterification step, mRNA processing, protein ubiquitination, RNA splicing	WDR36	0.989	
CDC40	L56P	generation of catalytic spliceosome for second transesterification step, mRNA 3'-splice site recognition, mRNA branch site recognition, mRNA processing, mRNA splicing - via spliceosome, RNA splicing	WDR97, CDC40, WDR25		non-conserved AA in human CDC40
THO2	K1099E	DNA recombination, mRNA 3'-end processing, mRNA export from nucleus, positive regulation of transcription elongation from RNA polymerase I promoter, positive regulation of transcription from RNA polymerase I promoter, regulation of transcription - DNA-templated, transcription-coupled nucleotide-excision repair, transcription - DNA-templated, transcription elongation from RNA polymerase II promoter	THOC2		“homologous substitution” exists in human sequence
RPL19A	D188N	cytoplasmic translation, Translation	RPL19		
RRP7	S47F	ribosomal small subunit assembly, ribosome biogenesis, rRNA processing	RRP7A, RRP7BP		
YER156C	T276I	biological process unknown			
RPS29A	R22C	cytoplasmic translation, Translation	RPS29		
YMR102C	N492S	biological process unknown			
Strain: 100-4-8					
SIN3	start lost M1I	cell cycle, cell division, covalent chromatin modification, double-strand break repair via nonhomologous end joining, histone deacetylation, negative regulation of chromatin silencing at rDNA, negative regulation of chromatin silencing at silent mating-type cassette, negative regulation of chromatin silencing at telomere, negative regulation of transcription from RNA polymerase II promoter, negative regulation of transcription from RNA polymerase I promoter, negative regulation of transcription involved in meiotic cell cycle, positive regulation of transcription from RNA polymerase II promoter, positive regulation of transcription from RNA polymerase II promoter in response to heat stress, regulation of DNA-dependent DNA replication initiation, regulation of transcription - DNA-templated, regulation of transcription involved in G2/M transition of mitotic cell cycle, transcription - DNA-templated, transfer RNA gene-mediated silencing	SIN3B, SIN3A		
SOF1	L294S	maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA - 5.8S rRNA - LSU-rRNA), ribosome biogenesis, rRNA processing	DCAF13	0.977	
ECM5	N783D	cellular response to oxidative stress, cell wall organization	JARID2, KDM5C, KDM5D		
YGL036W	E264K	biological process unknown			
Strain: 100-5-6					
GIS1	L679F	chronological cell aging, histone demethylation, histone H3-K36 demethylation, maintenance of stationary phase in response to starvation, negative regulation of transcription from RNA polymerase II promoter, positive regulation of transcription from RNA polymerase II promoter, regulation of phospholipid biosynthetic process, regulation of transcription - DNA-templated, transcription - DNA-templated, transcription from RNA polymerase II promoter	KDM4B, KDM4C, KDM4E, KDM4D, KDM4A		
HRD1	F147L	endoplasmic reticulum unfolded protein response, ER-associated ubiquitin-dependent protein catabolic process, fungal-type cell wall organization, protein autoubiquitination, protein polyubiquitination, protein ubiquitination, protein ubiquitination involved in ubiquitin-dependent protein catabolic process, retrograde protein transport - ER to cytosol	RNF145, AMFR, SYVN1		non-conserved AA in humans
YDR018C	Y306H	lipid metabolic process, metabolic process, phospholipid biosynthetic process			
PIR1	P229S	cell wall organization, fungal-type cell wall organization, intracellular protein transport			
Strain: 100-5-3					
PRP6	L159S	mRNA processing, mRNA splicing - via spliceosome, RNA processing, RNA splicing, spliceosomal tri-snRNP complex assembly	PRPF6	0.633	
RSC8	L201S	ATP-dependent chromatin remodeling, covalent chromatin modification, double-strand break repair via nonhomologous end joining, nucleosome disassembly, regulation of transcription - DNA-templated, transcription - DNA-templated, transcription elongation from RNA polymerase II promoter	SMARCC1, SMARCC2, MPND		
FUN30	L318S	ATP-dependent chromatin remodeling, cellular response to DNA damage stimulus, chromatin remodeling, chromatin silencing at rDNA, chromatin silencing at silent mating-type cassette, chromatin silencing at telomere, covalent chromatin modification, DNA double-strand break processing, DNA repair, heterochromatin assembly involved in chromatin silencing, heterochromatin maintenance involved in chromatin silencing, negative regulation of transcription from RNA polymerase II promoter, nucleosome mobilization, nucleosome positioning	SMARCAD1		no domain in human homolog
RTT106	F360S	DNA replication-dependent nucleosome assembly, DNA replication-independent nucleosome assembly, heterochromatin assembly involved in chromatin silencing, negative regulation of transcription from RNA polymerase II promoter, regulation of transcription - DNA-templated, transcription - DNA-templated, transcription elongation from RNA polymerase II promoter, Transposition			
SKN1	K377*	(1->6)-beta-D-glucan biosynthetic process, carbohydrate metabolic process, cell wall organization, fungal-type cell wall organization, sphingolipid biosynthetic process			377/771 amino acid truncation

Gene	Variant	GO Process Terms	Human Homologs / Orthologs	Polyphen 2 score	Notes
YNL095C	*643Y	biological process unknown, transmembrane transport, Transport			Adds 2 amino acids
YBT1	N578K	ATP hydrolysis coupled anion transmembrane transport, bile acid and bile salt transport, calcium ion transport, ion transport, transmembrane transport, Transport	ABCC5, ABCC9, ABCC4, CFTR, ABCC2, ABCC6, ABCC1, ABCC8, ABCC11, ABCC3, ABCC10, ABCC12		
CMS1	P99T	biological process unknown	CMSS1		
GCN3	T21I	cellular metabolic process, positive regulation of cellular response to amino acid starvation, positive regulation of GTPase activity, regulation of catalytic activity, regulation of translation, regulation of translational initiation, translation, translational initiation	EIF2B1		
LAT1	N405D	acetyl-CoA biosynthetic process from pyruvate, metabolic process, pyruvate metabolic process	DLAT, PDHX		
NOP12	P357L	maturation of LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA - 5.8S rRNA - LSU-rRNA), ribosome biogenesis, rRNA processing	HNRNPA0, RBM34, HNRNPAB, HNRNPD, HNRNPDL		
MRP17	L82P	mitochondrial translation, Translation	MRPS6		
PAU17	A70V	biological process unknown, response to stress			
YKL165C-A	S73L	biological process unknown			
Strain: 100-11-2					
PRP5	L589S	mRNA branch site recognition, mRNA processing, RNA secondary structure unwinding, RNA splicing, rRNA processing	DDX46	1	
NMD2	L858S	DNA recombination, nuclear-transcribed mRNA catabolic process - 3'-5' exonucleolytic nonsense-mediated decay, nuclear-transcribed mRNA catabolic process - nonsense-mediated decay, translational frameshifting	UPF2		non-conserved AA in humans
USO1	L764S	ER to Golgi vesicle-mediated transport, Golgi vesicle docking, intracellular protein transport, protein transport, SNARE complex assembly, transcytosis, transport, vesicle fusion with Golgi apparatus	RUNDC3A, RUNDC3B, RUFY3, RUFY2, RUFY1, USO1		

Legend:	
	Transcription-related GO terms
	DNA replication / repair / etc. -related GO terms

Table A2: GO Process Term Frequency in sequenced strains

GO term	Frequency in List	% in List	Genome Frequency	% in Genome	Fold Increase	Gene(s)
RNA splicing	10 out of 93 genes	10.80%	144 of 6433 genes	2.20%	4.91	PRP6, PRP5, CDC40, BRR2, SYF2, MSL1, ISY1, PRP19, MSL5, YSH1
vitamin metabolic process	3 out of 93 genes	3.20%	43 of 6433 genes	0.70%	4.57	RIB7, BIO5, SPE3
mRNA processing	11 out of 93 genes	11.80%	170 of 6433 genes	2.60%	4.54	PRP6, PRP5, CDC40, BRR2, SYF2, MSL1, ISY1, PRP19, MSL5, YSH1, THO2
regulation of DNA metabolic process	5 out of 93 genes	5.40%	107 of 6433 genes	1.70%	3.18	DUN1, PIF1, SIN3, MSH2, ELG1
DNA replication	6 out of 93 genes	6.50%	139 of 6433 genes	2.20%	2.95	DUN1, MMS22, PIF1, SIN3, MSH2, ELG1
lipid transport	3 out of 93 genes	3.20%	68 of 6433 genes	1.10%	2.91	LAM1, YBT1, SIP3
endocytosis	4 out of 93 genes	4.30%	97 of 6433 genes	1.50%	2.87	AKL1, PIL1, YAP1801, ARK1
protein maturation	2 out of 93 genes	2.20%	52 of 6433 genes	0.80%	2.75	ATG18, ATG19
regulation of translation	4 out of 93 genes	4.30%	105 of 6433 genes	1.60%	2.69	EBS1, PAB1, GCN20, GCN3
DNA recombination	7 out of 93 genes	7.50%	185 of 6433 genes	2.90%	2.59	EBS1, NMD2, MMS22, PIF1, THO2, MSH2, ELG1
vesicle organization	3 out of 93 genes	3.20%	81 of 6433 genes	1.30%	2.46	USO1, GEA2, ATG19
translational initiation	2 out of 93 genes	2.20%	61 of 6433 genes	0.90%	2.44	PAB1, GCN3
response to starvation	2 out of 93 genes	2.20%	56 of 6433 genes	0.90%	2.44	GIS1, GCN3
cell morphogenesis	1 out of 93 genes	1.10%	29 of 6433 genes	0.50%	2.20	FUS1
DNA repair	8 out of 93 genes	8.60%	256 of 6433 genes	4.00%	2.15	FUN30, RSC8, MMS22, PIF1, THO2, SIN3, MSH2, ELG1
DNA-templated transcription elongation	3 out of 93 genes	3.20%	97 of 6433 genes	1.50%	2.13	RSC8, THO2, RTT106
ribosomal small subunit biogenesis	4 out of 93 genes	4.30%	135 of 6433 genes	2.10%	2.05	RRP7, SAS10, SOF1, UTP13
RNA catabolic process	4 out of 93 genes	4.30%	132 of 6433 genes	2.10%	2.05	EBS1, PAB1, NMD2, SKI7
cellular response to DNA damage stimulus	9 out of 93 genes	9.70%	307 of 6433 genes	4.80%	2.02	FUN30, DUN1, RSC8, MMS22, PIF1, THO2, SIN3, MSH2, ELG1
transcription from RNA polymerase I promoter	2 out of 93 genes	2.20%	68 of 6433 genes	1.10%	2.00	THO2, SIN3
response to heat	2 out of 93 genes	2.20%	73 of 6433 genes	1.10%	2.00	PIL1, SIN3
monocarboxylic acid metabolic process	4 out of 93 genes	4.30%	140 of 6433 genes	2.20%	1.95	FOX2, LAT1, BIO5, SPE3
telomere organization	2 out of 93 genes	2.20%	78 of 6433 genes	1.20%	1.83	PIF1, ELG1
histone modification	3 out of 93 genes	3.20%	114 of 6433 genes	1.80%	1.78	SPT7, GIS1, SIN3
protein phosphorylation	5 out of 93 genes	5.40%	197 of 6433 genes	3.10%	1.74	AKL1, AHK1, DUN1, PIL1, ARK1
regulation of transport	2 out of 93 genes	2.20%	85 of 6433 genes	1.30%	1.69	AKL1, ARK1
DNA-templated transcription termination	1 out of 93 genes	1.10%	43 of 6433 genes	0.70%	1.57	YSH1
protein lipidation	1 out of 93 genes	1.10%	45 of 6433 genes	0.70%	1.57	GAB1
snoRNA processing	1 out of 93 genes	1.10%	45 of 6433 genes	0.70%	1.57	YSH1
cofactor metabolic process	4 out of 93 genes	4.30%	179 of 6433 genes	2.80%	1.54	RPE1, LAT1, BIO5, SPE3
chromatin organization	7 out of 93 genes	7.50%	318 of 6433 genes	4.90%	1.53	FUN30, SPT7, GIS1, RSC8, RTT106, SIN3, MSH2
lipid metabolic process	6 out of 93 genes	6.50%	296 of 6433 genes	4.60%	1.41	YDR018C, GIS1, SKN1, FOX2, ATG26, GAB1
nucleobase-containing small molecule metabolic process	4 out of 93 genes	4.30%	198 of 6433 genes	3.10%	1.39	ATP17, RPE1, ATP2, ADE12
rRNA processing	6 out of 93 genes	6.50%	311 of 6433 genes	4.80%	1.35	RRP7, SAS10, SOF1, BMT6, UTP13, NOP12
regulation of protein modification process	2 out of 93 genes	2.20%	111 of 6433 genes	1.70%	1.29	AHK1, PIL1
ribosome assembly	1 out of 93 genes	1.10%	59 of 6433 genes	0.90%	1.22	RRP7
transcription from RNA polymerase II promoter	8 out of 93 genes	8.60%	476 of 6433 genes	7.40%	1.16	FUN30, GIS1, RSC8, YSH1, THO2, RTT106, SIP3, SIN3
biological process unknown	19 out of 93 genes	20.40%	1133 of 6433 genes	17.60%	1.16	MFG1, YDR444W, YER156C, YGL036W, CUE3, YLF2, ECM25, YJL213W, YKL165C-A, PAU17, YLL037W, CMS1, YLR236C, YMR102C, YNL095C, YNL140C, ECM3, CWC27, YPR078C
conjugation	2 out of 93 genes	2.20%	124 of 6433 genes	1.90%	1.16	SPT7, FUS1
protein acylation	1 out of 93 genes	1.10%	66 of 6433 genes	1.00%	1.10	SPT7
ion transport	4 out of 93 genes	4.30%	263 of 6433 genes	4.10%	1.05	ATP17, ATP2, YBT1, SAL1
cell wall organization or biogenesis	3 out of 93 genes	3.20%	198 of 6433 genes	3.10%	1.03	SKN1, PIR1, HRD1
nucleobase-containing compound transport	2 out of 93 genes	2.20%	143 of 6433 genes	2.20%	1.00	SAL1, THO2
meiotic cell cycle	4 out of 93 genes	4.30%	282 of 6433 genes	4.40%	0.98	ATG26, MMS22, SIN3, MSH2
peptidyl-amino acid modification	2 out of 93 genes	2.20%	151 of 6433 genes	2.30%	0.96	SPT7, DUN1
proteolysis involved in cellular protein catabolic process	3 out of 93 genes	3.20%	216 of 6433 genes	3.40%	0.94	SSM4, HRD1, ATG19
membrane fusion	1 out of 93 genes	1.10%	76 of 6433 genes	1.20%	0.92	USO1
protein targeting	4 out of 93 genes	4.30%	307 of 6433 genes	4.80%	0.90	SSH1, ATG18, TOM70, ATG19
transmembrane transport	3 out of 93 genes	3.20%	234 of 6433 genes	3.60%	0.89	ATP17, ATP2, TOM70
cytoskeleton organization	3 out of 93 genes	3.20%	238 of 6433 genes	3.70%	0.86	AKL1, GEA2, ARK1
RNA modification	2 out of 93 genes	2.20%	172 of 6433 genes	2.70%	0.81	TRM8, BMT6
cytoplasmic translation	2 out of 93 genes	2.20%	180 of 6433 genes	2.80%	0.79	RPL19A, RPS29A
response to osmotic stress	1 out of 93 genes	1.10%	92 of 6433 genes	1.40%	0.79	AHK1
organelle fusion	1 out of 93 genes	1.10%	89 of 6433 genes	1.40%	0.79	USO1
response to chemical	5 out of 93 genes	5.40%	446 of 6433 genes	6.90%	0.78	FUS1, SSM4, ECM5, HRD1, ATG19
mitochondrion organization	3 out of 93 genes	3.20%	265 of 6433 genes	4.10%	0.78	MCP2, PIF1, TOM70
Golgi vesicle transport	2 out of 93 genes	2.20%	191 of 6433 genes	3.00%	0.73	USO1, GEA2

GO term	Frequency in List	% in List	Genome Frequency	% in Genome	Fold Increase	Gene(s)
chromosome segregation	2 out of 93 genes	2.20%	203 of 6433 genes	3.20%	0.69	MMS22, ELG1
endosomal transport	1 out of 93 genes	1.10%	102 of 6433 genes	1.60%	0.69	ATG18
response to oxidative stress	1 out of 93 genes	1.10%	105 of 6433 genes	1.60%	0.69	ECM5
ribosomal large subunit biogenesis	1 out of 93 genes	1.10%	104 of 6433 genes	1.60%	0.69	NOP12
organelle fission	3 out of 93 genes	3.20%	302 of 6433 genes	4.70%	0.68	MMS22, MSH2, ELG1
transposition	1 out of 93 genes	1.10%	110 of 6433 genes	1.70%	0.65	ELG1
tRNA processing	1 out of 93 genes	1.10%	109 of 6433 genes	1.70%	0.65	TRM8
signaling	2 out of 93 genes	2.20%	245 of 6433 genes	3.80%	0.58	AHK1, HRD1
organelle assembly	1 out of 93 genes	1.10%	135 of 6433 genes	2.10%	0.52	RRP7
sporulation	1 out of 93 genes	1.10%	133 of 6433 genes	2.10%	0.52	ATG26
regulation of organelle organization	2 out of 93 genes	2.20%	283 of 6433 genes	4.40%	0.50	PIF1, SIN3
protein complex biogenesis	2 out of 93 genes	2.20%	293 of 6433 genes	4.60%	0.48	SPT7, USO1
mitotic cell cycle	2 out of 93 genes	2.20%	319 of 6433 genes	5.00%	0.44	SIN3, ELG1
mitochondrial translation	1 out of 93 genes	1.10%	167 of 6433 genes	2.60%	0.42	MRP17
protein modification by small protein conjugation or removal	1 out of 93 genes	1.10%	170 of 6433 genes	2.60%	0.42	HRD1
nuclear transport	1 out of 93 genes	1.10%	181 of 6433 genes	2.80%	0.39	THO2
carbohydrate metabolic process	1 out of 93 genes	1.10%	198 of 6433 genes	3.10%	0.35	SKN1
regulation of cell cycle	1 out of 93 genes	1.10%	233 of 6433 genes	3.60%	0.31	DUN1
translational elongation	1 out of 93 genes	1.10%	334 of 6433 genes	5.20%	0.21	GCN20
pseudohyphal growth	0 out of 93 genes	0.00%	74 of 6433 genes	1.20%	0.00	none
carbohydrate transport	0 out of 93 genes	0.00%	33 of 6433 genes	0.50%	0.00	none
cytokinesis	0 out of 93 genes	0.00%	84 of 6433 genes	1.30%	0.00	none
protein glycosylation	0 out of 93 genes	0.00%	62 of 6433 genes	1.00%	0.00	none
protein dephosphorylation	0 out of 93 genes	0.00%	49 of 6433 genes	0.80%	0.00	none
cellular respiration	0 out of 93 genes	0.00%	83 of 6433 genes	1.30%	0.00	none
cellular amino acid metabolic process	0 out of 93 genes	0.00%	199 of 6433 genes	3.10%	0.00	none
protein alkylation	0 out of 93 genes	0.00%	50 of 6433 genes	0.80%	0.00	none
vacuole organization	0 out of 93 genes	0.00%	88 of 6433 genes	1.40%	0.00	none
protein folding	0 out of 93 genes	0.00%	95 of 6433 genes	1.50%	0.00	none
generation of precursor metabolites and energy	0 out of 93 genes	0.00%	154 of 6433 genes	2.40%	0.00	none
transcription from RNA polymerase III promoter	0 out of 93 genes	0.00%	41 of 6433 genes	0.60%	0.00	none
invasive growth in response to glucose limitation	0 out of 93 genes	0.00%	59 of 6433 genes	0.90%	0.00	none
DNA-templated transcription initiation	0 out of 93 genes	0.00%	73 of 6433 genes	1.10%	0.00	none
cell budding	0 out of 93 genes	0.00%	58 of 6433 genes	0.90%	0.00	none
organelle inheritance	0 out of 93 genes	0.00%	60 of 6433 genes	0.90%	0.00	none
peroxisome organization	0 out of 93 genes	0.00%	50 of 6433 genes	0.80%	0.00	none
exocytosis	0 out of 93 genes	0.00%	49 of 6433 genes	0.80%	0.00	none
cellular ion homeostasis	0 out of 93 genes	0.00%	128 of 6433 genes	2.00%	0.00	none
oligosaccharide metabolic process	0 out of 93 genes	0.00%	27 of 6433 genes	0.40%	0.00	none
amino acid transport	0 out of 93 genes	0.00%	42 of 6433 genes	0.70%	0.00	none
tRNA aminoacylation for protein translation	0 out of 93 genes	0.00%	36 of 6433 genes	0.60%	0.00	none
nucleus organization	0 out of 93 genes	0.00%	66 of 6433 genes	1.00%	0.00	none
ribosomal subunit export from nucleus	0 out of 93 genes	0.00%	46 of 6433 genes	0.70%	0.00	none

Legend:	Transcription-related GO terms
	DNA replication / repair / etc. -related GO terms

Addendum References:

Debacker, K., Frizzell, A., Gleeson, O., Kirkham-McCarthy, L., Mertz, T., and Lahue, R.S. (2012). Histone deacetylase complexes promote trinucleotide repeat expansions. PLoS Biol. *10*.

Kang, Y.H., Kang, M.J., Kim, J.H., Lee, C.H., Cho, I.T., Hurwitz, J., and Seo, Y.S. (2009). The MPH1 gene of Saccharomyces cerevisiae functions in Okazaki fragment processing. J. Biol. Chem. *284*, 10376–10386.

Krishna, S., Wagener, B.M., Liu, H.P., Lo, Y.C., Sterk, R., Petrini, J.H.J., and Nickoloff, J.A. (2007). Mre11 and Ku regulation of double-strand break repair by gene conversion and break-induced replication. DNA Repair (Amst). *6*, 797–808.

Maréchal, A., Li, J., Ji, X.Y., Wu, C., Yazinski, S.A., Nguyen, H.D., Liu, S., Jiménez, A.E., Jin, J., and Zou, L. (2014). PRP19 Transforms into a Sensor of RPA-ssDNA after DNA Damage and Drives ATR Activation via a Ubiquitin-Mediated Circuitry. Mol. Cell *53*, 235–246.

Mason, P.B., and Struhl, K. (2005). Distinction and relationship between elongation rate and processivity of RNA polymerase II in vivo. Mol. Cell *17*, 831–840.

Moehle, E.A., Braberg, H., Krogan, N.J., and Guthrie, C. (2014). Adventures in time and space: Splicing efficiency and RNA polymerase II elongation rate. RNA Biol. *11*, 313–319.

Nag, A., Narsinh, K., and Martinson, H.G. (2007). The poly(A)-dependent transcriptional pause is mediated by CPSF acting on the body of the polymerase. Nat. Struct. Mol. Biol. *14*, 662–669.

Schaughency, P., Merran, J., and Corden, J.L. (2014). Genome-Wide Mapping of Yeast RNA Polymerase II Termination. PLoS Genet. *10*.

Shah, K.A., Shishkin, A.A., Voineagu, I., Pavlov, Y.I., Shcherbakova, P. V, and Mirkin, S.M. (2012). Role of DNA polymerases in repeat-mediated genome instability. Cell Rep *2*, 1088–1095.

Shah, K.A., McGinty, R.J., Egorova, V.I., and Mirkin, S.M. (2014). Coupling Transcriptional State to Large-Scale Repeat Expansions in Yeast. Cell Rep. *9*, 1594–1603.

Shishkin, A.A., Voineagu, I., Matera, R., Cherng, N., Chernet, B.T., Krasilnikova, M.M., Narayanan, V., Lobachev, K.S., and Mirkin, S.M. (2009). Large-Scale Expansions of Friedreich's Ataxia GAA Repeats in Yeast. Mol. Cell *35*, 82–92.

Signon, L., Malkova, A., Naylor, M.L., Haber, J.E., and Klein, H. (2001). Genetic Requirements for RAD51 - and Replication Repair of a Chromosomal Double-Strand Break Genetic Requirements for RAD51 - and RAD54 -Independent Break-Induced Replication Repair of a Chromosomal Double-Strand Break. Mol. Cell. Biol. *21*, 2048–2056.

Štafa, A., Donnianni, R.A., Timashev, L.A., Lam, A.F., and Symington, L.S. (2014). Template switching during break-induced replication is promoted by the mph1 helicase in Saccharomyces cerevisiae. Genetics *196*, 1017–1028.

Tsutakawa, S.E., Thompson, M.J., Arvai, A.S., Neil, A.J., Shaw, S.J., Algasaier, S.I., Kim, J.C., Finger, L.D., Jardine, E., Gotham, V.J.B., et al. (2017). Phosphate steering by Flap Endonuclease 1 promotes 5′-flap specificity and incision to prevent genome instability. Nat. Commun. *8*, 1–14.

Chapter 3

Excerpts pertaining to the work of Ryan McGinty from the following publication:

("…" indicates removed portions of the manuscript contributed entirely by the remaining authors.)

Coupling transcriptional state to large-scale repeat expansions in yeast

Kartik A. Shah, Ryan J. McGinty, Vera I. Egorova and Sergei M. Mirkin*

Department of Biology, Tufts University, Medford MA 02155

* The author to whom all correspondence should be addressed.

E-mail: sergei.mirkin@tufts.edu

Cell Reports, Volume 9, Issue 5, 2014, Pages 1594-1602

doi: 10.1016/j.celrep.2014.10.048.

SUMMARY:

Expansions of simple DNA repeats cause numerous hereditary disorders in humans. Replication, repair and transcription are implicated in the expansion process, but their relative contributions are yet to be distinguished. To separate the role of replication and transcription in the expansion of Friedreich's ataxia $(GAA)_n$ repeats, we designed two yeast genetic systems that utilize a galactose-inducible *GAL1* promoter, but contain these repeats in either the transcribed or non-transcribed region of a selectable cassette. We found that large-scale repeat expansions can occur in the lack of transcription. Induction of transcription strongly elevated the rate of expansions in both systems, indicating that active transcriptional state rather than transcription through the repeat *per se* affects this process. Furthermore, replication defects increased the rate of repeat expansions irrespective of transcriptional state. We present a model where transcriptional state, linked to the nucleosomal density of a region, acts as a modulator of large-scale repeat expansions.

RESULTS

A new system to study repeat expansions in a non-transcribed location

In yeast, transcriptional enhancers or upstream activating sequences are usually positioned several hundred base pairs (bp) upstream of the TATA-box in the core promoter (Dobi and Winston, 2007). In the endogenous *GAL1* regulon, an upstream activating sequence (UAS GAL) is situated around 150 bps upstream of the basal promoter P *GAL1* . It was previously found that increasing the distance between the UAS GAL and TATA-box beyond ~800 bps disrupted transcriptional activation of a downstream marker (Dobi and Winston, 2007). Since the DNA region between UAS GAL and basal promoter is strongly occupied by a nucleosome, it inhibits access to RNA polymerase (Lohr and Lopez, 1995; Lohr et al., 1995). We reasoned therefore that a $(GAA)_{100}$ repeat positioned between the UAS GAL and TATA-box is unlikely to be transcribed, particularly in cells grown on a glucose-containing media. Furthermore, if the repeat expands, the distance between the UAS GAL and basal promoter would exceed the 800 bps distance threshold, blocking activation of the downstream marker and thereby allowing us to select for such events on appropriate selection media.

...

To ensure that the repeats positioned between UAS GAL and TATA are not transcribed, we wanted to analyze transcription through this regulatory region by quantitative reverse-transcription PCR (qRT-PCR). Recent studies have shown that transcription in yeast is much more pervasive than earlier believed (Djebali et al., 2012; Nagalakshmi et al., 2008). Many cryptic and/or non-coding transcripts are quickly degraded by a ribonuclease encoded by *RRP6*. Hence, we decided to conduct the RT-PCR analysis in an *rrp6*Δ background (for further rationale, see Discussion). Using primer pair A-B, situated on either side of the repeat, we have previously shown that the repeat in the original P *URA3* -UR-Intron-GAA-A3 cassette is actively transcribed (Shishkin et al., 2009). Using the same primer set here for RT-PCR, we were unable to detect transcription through the repeat situated between the UAS GAL and TATA-box in the new cassette **(Fig. 1B)**, irrespective of whether the cells were grown on glucose or galactose.

We have previously found in a different system (and a different chromosomal setting) that $(GAA)_n$ runs can serve as weak promoters (Zhang et al., 2012). Thus, we wanted to determine if there is transcription downstream of the $(GAA)_{100}$ run in our current construct. To this end, we conducted qRT-PCR with three more primer pairs **(Fig. 1)**: primer pair C-D for the region immediately downstream of the repeat, primer pair E-F for the region located further downstream of the repeat but upstream of the TATA-box and primer pair G-H for the 5' end of the *CAN1* ORF. Traditional RT-PCRs were also carried out separately and run on an agarose gel for comparison. Low levels of C-D and E-F transcript were indeed observed in the region downstream of the repeats **(Fig. 1B)**, consistent with the idea that they can serve as a weak promoter. Note however, that the relative levels of these transcripts were at least 100-fold lower than for the G-H transcript. Overall, our results show that even in an rrp6Δ background, the levels of transcription in the vicinity of the repeat are negligible and likely due to non-canonical or gratuitous transcription. We also carried out traditional RT-PCRs in the WT background (presented in Fig. S1), where primer pairs C–D and E–F did not detect transcription between the repeat and *CAN1* gene, irrespective of whether the strains were grown on glucose or galactose. Altogether, these results indicate that the $(GAA)_{100}$ repeats in our new system are positioned in a non-transcribed location on the chromosome.

Our RT-PCR and qRT-PCR data from Fig. 1 show that when grown on galactose, the P *GAL1*-*CAN1* constructs generated high levels of G-H transcript. Importantly however, this transcript was visible even when the strains were grown on glucose, albeit at a much lower level (~ 150-fold). The latter observation can be attributed to leakage from the *GAL1* promoter, most likely due to the lack of glucose repressor elements (URS GAL) in our system.

Large-scale expansion of repeats positioned in a non-transcribed location

…

Previous studies have shown that the native UAS GAL element is hypersensitive to micrococcal nuclease (MNase) digestion, indicating that it remains free of nucleosomes irrespective of the carbon source (Bryant et al., 2008). In the non-induced state, two nucleosomes are positioned downstream of the UAS GAL and upstream of the ORF. These nucleosomes get disrupted upon galactose induction through a process mediated by Gal4p (Lohr et al., 1995). We decided to investigate if this is indeed the case in our modified constructs through the MNase-qPCR assay (Infante et al., 2012). Strains containing the P *GAL1-CAN1* and P *GAL1*-GAA-*CAN1* constructs were grown in either glucose or galactose, before isolating chromatin and digesting with MNase to extract mononucleosomal-sized DNA fragments. We designed a series of overlapping primers across our constructs (with amplicons ranging from 50 to 100 bps) to perform qPCR analyses on the mononucleosomal DNA. Genomic DNA was also extracted simultaneously and used for normalization. The results of this assay are presented in Fig. 2. In the P *GAL1-CAN1* construct, the UAS GAL element was hypersensitive to MNase digestion **(Fig. 2A)**, consistent with results from previous studies (Bryant et al., 2008). The region immediately downstream of UAS GAL was less sensitive (or relatively enriched), indicating that it was occupied by a nucleosome in the non-induced state (glucose), but not in the induced state (galactose). The region further downstream of UAS GAL and including the transcription start site (TSS) of *CAN1* was highly enriched, suggesting that it was strongly occupied by a nucleosome in either conditions, although to a lower level upon induction. The latter results are consistent with the previous genome-wide analyses of phased nucleosomes positioned at TSSs (Jiang and Pugh, 2009). In the P *GAL1*-GAA-*CAN1* construct, we observed similar results - the UAS GAL element was hypersensitive to MNase digestion and the TSS was highly enriched **(Fig. 2B)**. Note that the distance between UAS GAL and TSS in this construct is longer (~ 650 bp) than in the repeat-less construct. The repetitive nature of the GAA tract precluded us from using overlapping primers across this segment. Hence, we decided to extract trinucleosome-sized fragments from the above MNase digests and use the A-B primer pair to compare relative protection under glucose and galactose conditions. We observed that the $(GAA)_{100}$ repeat-containing region seemed to be relatively enriched in the non-induced (glucose) over the induced condition (galactose) **(Fig. 2B inset)**. Overall, our results suggest that the nucleosomal density in the $(GAA)_{100}$ repeat-containing region changes upon transcriptional activation of UAS GAL , and is higher in the non-induced than in the induced state, as we expected based on the data for the endogenous *GAL1* promoter.

DISCUSSION

…

To investigate the role of transcription further, we developed a new selectable system, utilizing an inducible promoter to drive expression of the selectable marker, but positioned $(GAA)_{100}$ repeats in the region between the UAS GAL and TATA-box of the promoter. This strategy was chosen based on previous studies which demonstrated that this region is strongly bound by nucleosomes, precluding it from being accessed by RNA polymerase (Lohr and Lopez, 1995; Lohr et al., 1995). Since the discovery of pervasive transcription in yeast and higher eukaryotes, many Cryptic Unstable Transcripts (CUTs) and Stable Unannotated Transcripts (SUTs) have been discovered (Marquardt et al., 2011;

Thompson and Parker, 2007). These non-coding RNAs are quickly degraded in the nuclei by various RNA surveillance pathways. In budding yeast, the *RRP6* gene codes for a 3' – 5' ribonuclease involved in the rapid degradation of CUTs and SUTs (Hazelbaker et al., 2013). As a result, isolation of these cryptic transcripts from Rrp6 + strains is almost impossible using standard RNA extraction protocols. Hence, we additionally confirmed that the repeat tract located between UAS GAL and TATA-box in our construct is not actively transcribed in a WT or *rrp6*Δ background. We found large-scale $(GAA)_n$ repeat expansions in this system, effectively blocking expression of the downstream *CAN1* marker. Because these events were observed under conditions where little to no transcription was detected by qRT-PCR, our results strongly suggest that transcription is not necessary for repeat expansions to occur. Remarkably though, the rate of expansions increased 10-fold when transcription was induced, quantitatively similarly to what we observed upon induction when the repeat was located in a transcribed region. We conclude from these data that it is the transcriptional state of a repeat-containing region, rather than transcription elongation through the repeat tract *per se*, that accounts for the elevated rate of expansions.

On one hand, our results suggest that while transcription through a repeat is not required for expansions to occur, the transcriptional state of a repeat-containing region affected its expansion rate. On the other hand, increases in the expansion rate due to replication defects were independent of transcriptional state. The seemingly paradoxical nature of our results prompted us to wonder if the arrangement of nucleosomes at a repeat-containing region was modulating the rate of repeat expansions. Recently, genome-wide analysis of lagging strand DNA synthesis revealed that the size of Okazaki fragments correspond (on average) to a mononucleosomal-sized length of around 155 bps in yeast (Smith and Whitehouse, 2012). In fact, the authors argue that positioning of nascent nucleosomes systematically terminates synthesis of each Okazaki fragment on the lagging strand. Data from our lab and others indicate that repeat expansions are inextricably linked to Okazaki fragment synthesis and likely involve some form of template switching (Shah et al., 2012; Shishkin et al., 2009) or fork reversal (Follonier et al., 2013; Kerrest et al., 2009) at the replication fork.

If this is the case, then our previously described template-switch model generates a testable prediction: the rate of repeat expansions depends on the density of nucleosomes around the repeat-containing region. To investigate, we used MNase-qPCR to analyze the density of nucleosomes in our new system. Previous studies have shown that strong transcriptional induction of the native *GAL1* promoter wipes out nucleosomes from the region downstream of UAS GAL (Angermayr and Bandlow, 2003; Bryant et al., 2008). Consistent with these data, our results show that the nucleosomal density in the repeat-containing region of the new system is lower in the induced than the non-induced state. Taken together, in a transcriptionally non-induced state, the repeat-containing region contained a higher density of nucleosomes and the rate of expansion remained low. Upon transcriptional induction, nucleosomes were disrupted and the rate of expansions was elevated by ~ 10-fold. How exactly could nucleosome density affect template switching? Our model proposes that a high density of nucleosomes limits the portion of the leading strand available for switching, or in other words, the switching-window of nascent strands is limited to an area the size of just one nucleosome (~155 bp). Decreasing the density of nucleosomes increases the size of this switching-window, thereby increasing the probability of a template-switch. Since template switching is the first in a series of steps that ultimately leads to an expanded repeat, affecting its probability affects the overall rate of expansions. Our model could explain why expandable repeats are found only in the transcribed regions of the human genome (Mirkin, 2007). It could also explain the destabilizing effects of nuclear reprogramming on repeat stability (Ku et al., 2010; Mirkin, 2010). While it is replication-centric, it does not exclude the role of mismatch repair (MMR) or transcription-coupled repair (TC-NER), both of which were implicated in the repeat expansion process (Lin and Wilson, 2012). Finally, we want to emphasize that our new system allows one to study the large-scaleexpansions or contractions of any DNA microsatellite, including ones that are prone to deletions when placed in a transcribed location or block transcription elongation even at short lengths (for eg: AT-rich repeats).

EXPERIMENTAL PROCEDURES:

...

RT-PCR analysis

Yeast strains were grown overnight (starting OD 600 = 0.1) in media containing raffinose (YPRaff) followed by growing in either glucose or galactose-containing media (YPD or YPGal, starting OD 600 = 0.1) for about 6 hours at 30°C and 250 rpm shaking. Total RNA was isolated using the YeaSTAR kit (Zymo Research) and DNase treated using TURBO DNA-free. The resulting DNA-free RNA preps were all diluted down to a concentration of 20 ng/µl. From this, 200 ng of total RNA was reverse transcribed (Superscript III RT, Invitrogen) using a combination of olig(dT) and random primers. 1 µl of cDNA was then used as template for PCR amplification (Sequences of primer pairs A-B, C-D, E-F and G-H are listed in Table S1). *ACT1* mRNA was used for normalization (Table S1, primer pair 9 – 10). PCR reactions were carried out in 1x Green GoTaq buffer (Promega) with the following cycles: Initial denaturation 95°C for 1 minute, followed by 25 cycles of 93°C for 30 seconds, 60°C for 20 seconds, 72°C for 1 minute 30 seconds. 5 µl of the PCR product was run on an agarose gel.

For qRT-PCR reactions, 1 µl of the above cDNA was used along with 12.5 µl of SYBRselect Master Mix (Life Technologies), 0.5 µl of each primer (10 µM) and 9.5 µl ultrapure water (Sigma). Reactions were performed in triplicate on an ABI 7300 Real-Time PCR System with the following run conditions: 50°C for 2 minutes, 95°C for 2 minutes, followed by 40 cycles of 95°C for 15 seconds, 58°C for 15 seconds, 72°C for 60 seconds. Relative amplification values were calculated for each primer set by the formula 2^-ddCt, where ddCt represents the difference in cycle threshold between the primer set of interest and the *ACT1* control.

MNase-qPCR analysis

Yeast strains were grown overnight (starting OD 600 = 0.1) in media containing raffinose (YPRaff) followed by growing in either glucose or galactose-containing media (YPD or YPGal, starting OD 600 = 0.1) for about 6 hours at 30°C and 250 rpm shaking. A similar number of cells from each condition were pelleted and lysed by treating with 125 units of zymolyase (Zymo Research) at 37°C for 1 hour. Spheroplasts were pelleted by centrifugation at 3000g for 5 minutes, followed by washing the pellet with 1 ml of 1.2M sorbitol. Spheroplasts were pelleted again and resuspended in 500 µl of Nuclei Prep buffer from the EZ Nucleosomal DNA Prep Kit (Zymo Research). After incubating on ice for 5 minutes, spheroplasts were spun down and washed before resuspending in 500 µl of MN Digestion buffer. 100 µl aliquots of this suspension were treated with 0.1 units of micrococcal nuclease at room temperature for 5 minutes. The remainder of the EZ Nucleosomal DNA Prep Kit protocol was followed unmodified. The resulting MNase-digested nucleosomal DNA was run on an agarose gel and bands corresponding to mono-, di- and tri- nucleosomal-sized fragments were extracted using the GeneJet Gel Extraction kit (Thermo). Genomic DNA was also isolated from these strains using the YeaStar Genomic DNA Kit (Zymo Research).

All resulting DNA preps were diluted down to a concentration of 1 ng/µl and 2 µl of this was used for qPCR. Each qPCR reaction consisted of 12.5 µl of SYBRselect Master Mix (Life Technologies), 1 µl of each primer (10 µM), and 8.5 µl ultrapure water (Sigma). Reactions were performed in triplicate on an ABI 7300 Real-Time PCR System, with the following run conditions: 50°C for 2 minutes, 95°C for 2 minutes, followed by 40 cycles of 95°C for 15 seconds, 54-60°C for 30 seconds (according to primer T m), 60°C for 60 seconds. Relative protection values for each primer pair were calculated by the formula (E^-dCT n)/(E^-dCT g), where E is the primer efficiency (values ranging between 1 and 2), dCT is the mean difference in cycle threshold between the sample and reference, n represents a nucleosomal sample and g represents genomic DNA.

AUTHOR CONTRIBUTIONS

K.A.S. designed and performed experiments, analyzed data and wrote the manuscript. R.J.M. and V.I.E. designed and performed experiments. S.M.M. supervised the project, analyzed data and wrote the manuscript.

ACKNOWLEDGEMENTS

We thank Steve Buratowski, Catherine Freudenreich, Mitch McVey and Melissa Koch for their useful comments and helpful suggestions. This study was funded by NIH grants GM60987and GM105473 to S.M.M. and supported by the Tufts Graduate Student Research Award to K.A.S.

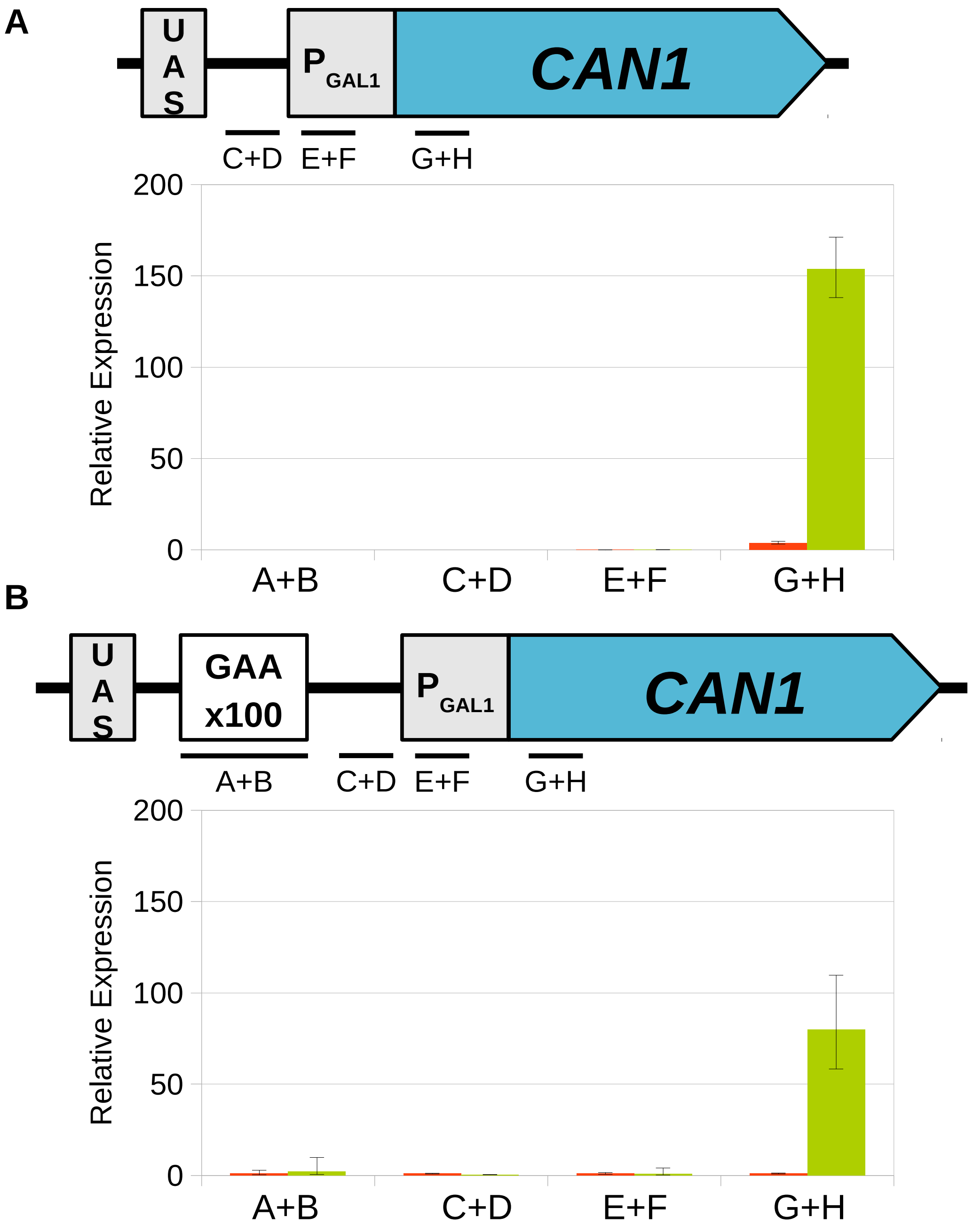


Fig. 1: Analysis of transcription between UAS and GAL promoter.
Black lines show regions amplified by primer pairs used for qRT-PCR analysis in the **(A)** P *GAL1*-*CAN1* and **(B)** P *GAL1*-GAA-*CAN1* constructs. qRT-PCRs of the primer pairs as indicated. Threshold cycle (Ct) values were normalized to *ACT1* expression before plotting. Bars depict mean relative expression from three technical replicates and error bars represent standard deviation.

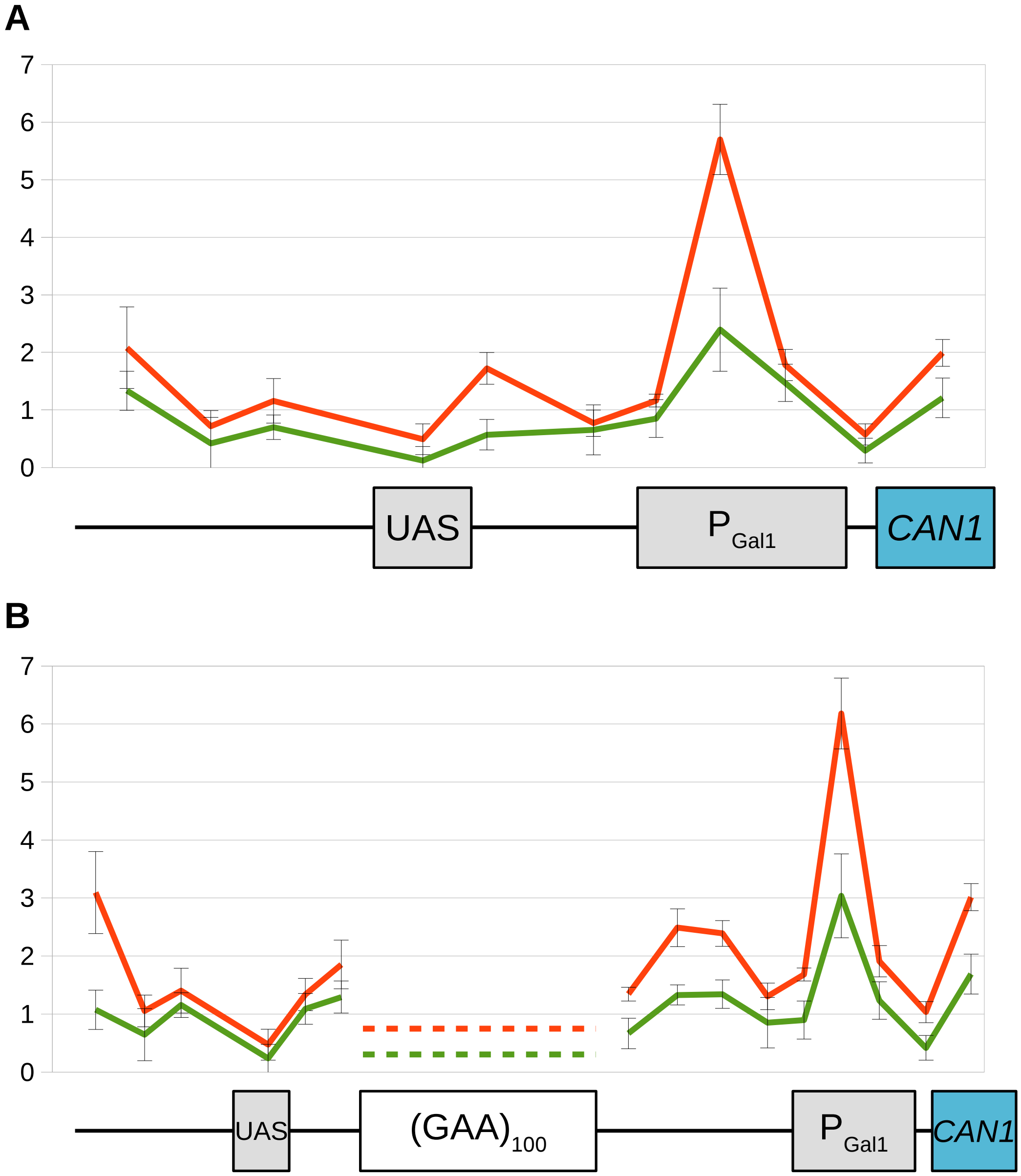


Fig. 2: Link between transcriptional status and nucleosome density. Nucleosome density in the **(A)** P *GAL1-CAN1* and **(B)** P *GAL1*-GAA-*CAN1* constructs under high (galactose) or low (glucose) transcriptional states. X-axis is to scale with the gene diagrams as pictured. Figures plot fold protection values derived from mononucleosome-sized fragments against position of each primer pair. Dashed lines in **(B)** represent trinucleosome-sized fragments.

REFERENCES

Angermayr, M., and Bandlow, W. (2003). Permanent nucleosome exclusion from the Gal4p-inducible yeast GCY1 promoter. J. Biol. Chem. 278, 11026–11031.

Bryant, G.O., Prabhu, V., Floer, M., Wang, X., Spagna, D., Schreiber, D., and Ptashne, M. (2008). Activator Control of Nucleosome Occupancy in Activation and Repression of Transcription. PLoS Biol 6, 2928–2939.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature 489, 101–108.

Dobi, K.C., and Winston, F. (2007). Analysis of transcriptional activation at a distance in Saccharomyces cerevisiae. Mol Cell Biol 27, 5575–5586.

Follonier, C., Oehler, J., Herrador, R., and Lopes, M. (2013). Friedreich's ataxia-associated GAA repeats induce replication-fork reversal and unusual molecular junctions. Nat. Struct. Mol. Biol. 20, 486–494.

Hazelbaker, D.Z., Marquardt, S., Wlotzka, W., and Buratowski, S. (2013). Kinetic competition between RNA Polymerase II and Sen1-dependent transcription termination. Mol. Cell 49, 55–66.

Herman, D., Jenssen, K., Burnett, R., Soragni, E., Perlman, S.L., and Gottesfeld, J.M. (2006). Histone deacetylase inhibitors reverse gene silencing in Friedreich's ataxia. Nat Chem Biol 2, 551–558.

Infante, J.J., Law, G.L., and Young, E.T. (2012). Analysis of nucleosome positioning using a nucleosome-scanning assay. Methods Mol. Biol. 833, 63–87.

Jiang, C., and Pugh, B.F. (2009). Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet 10, 161–172.

Kerrest, A., Anand, R.P., Sundararajan, R., Bermejo, R., Liberi, G., Dujon, B., Freudenreich, C.H., and Richard, G.-F. (2009). SRS2 and SGS1 prevent chromosomal breaks and stabilize triplet repeats by restraining recombination. Nat. Struct. Mol. Biol. 16, 159–167.

Ku, S., Soragni, E., Campau, E., Thomas, E.A., Altun, G., Laurent, L.C., Loring, J.F., Napierala, M., and Gottesfeld, J.M. (2010). Friedreich's ataxia induced pluripotent stem cells model intergenerational GAA·TTC triplet repeat instability. Cell Stem Cell 7, 631–637.

Lohr, D., and Lopez, J. (1995). GAL4/GAL80-dependent nucleosome disruption/deposition on the upstream regions of the yeast GAL1-10 and GAL80 genes. J. Biol. Chem. 270, 27671–27678.

Lohr, D., Venkov, P., and Zlatanova, J. (1995). Transcriptional regulation in the yeast GAL gene family: a complex genetic network. Faseb J. 9, 777–787.

Marquardt, S., Hazelbaker, D.Z., and Buratowski, S. (2011). Distinct RNA degradation pathways and 3' extensions of yeast non-coding RNA species. Transcription 2, 145–154.

Mirkin, S.M. (2007). Expandable DNA repeats and human disease. Nature 447, 932–940.

Mirkin, S.M. (2010). Getting to the core of repeat expansions by cell reprogramming. Cell Stem Cell 7, 545–546.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320, 1344–1349.

Shah, K.A., Shishkin, A.A., Voineagu, I., Pavlov, Y.I., Shcherbakova, P.V., and Mirkin, S.M. (2012). Role of DNA polymerases in repeat-mediated genome instability. Cell Reports 2, 1088–1095.

Shishkin, A.A., Voineagu, I., Matera, R., Cherng, N., Chernet, B.T., Krasilnikova, M.M., Narayanan, V., Lobachev, K.S., and Mirkin, S.M. (2009). Large-scale expansions of Friedreich's ataxia GAA repeats in yeast. Mol. Cell 35, 82–92.

Smith, D.J., and Whitehouse, I. (2012). Intrinsic coupling of lagging-strand synthesis to chromatin assembly. Nature 483, 434–438.

Thompson, D.M., and Parker, R. (2007). Cytoplasmic decay of intergenic transcripts in Saccharomyces cerevisiae. Mol Cell Biol 27, 92–101.

Zhang, Y., Shishkin, A.A., Nishida, Y., Marcinkowski-Desmond, D., Saini, N., Volkov, K.V., Mirkin, S.M., and Lobachev, K.S. (2012). Genome-wide Screen Identifies Pathways that Govern GAA/TTC Repeat Fragility and Expansions in Dividing and Nondividing Yeast Cells. Mol. Cell
48, 254–265.

Chapter 4

Nanopore sequencing of complex genomic rearrangements in yeast reveals mechanisms of repeat-mediated double-strand break repair

Ryan J. McGinty[1], Rachel G. Rubinstein[1], Alexander J. Neil[1], Margaret Dominska[2], Denis Kiktev[2], Thomas D. Petes[2], Sergei M. Mirkin[1*]

[1] Department of Biology, Tufts University, Medford, MA 02155, U.S.A.

[2] Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, NC 27710, U.S.A.

[*] The author to whom all correspondence should be addressed.

E-mail: sergei.mirkin@tufts.edu

Genome Research, Volume 27, Issue 12, 2017, Pages 2072-2082

doi: 10.1101/gr.228148.117

Abstract

Improper DNA double-strand break (DSB) repair results in complex genomic rearrangements (CGRs) in many cancers and various congenital disorders in humans. Trinucleotide repeat sequences, such as $(GAA)_n$ repeats in Friedreich's ataxia, $(CTG)_n$ repeats in myotonic dystrophy and $(CGG)_n$ repeats in fragile X syndrome, are also subject to double strand breaks within the repetitive tract followed by DNA repair. Mapping the outcomes of CGRs is important for understanding their causes and potential phenotypic effects. However, high-resolution mapping of CGRs has traditionally been a laborious and highly-skilled process. Recent advances in long-read DNA sequencing technologies, specifically Nanopore sequencing, have made possible the rapid identification of CGRs with single base pair resolution. Here we have employed whole-genome Nanopore sequencing to characterize several CGRs that originated from naturally occurring DSBs at $(GAA)_n$ microsatellites in *S. cerevisiae*. These data gave us important insights into the mechanisms of DSB repair leading to CGRs.

Introduction

Complex genomic rearrangements (CGRs) mixing together various genome alterations such as insertions, duplications, deletions, inversions and translocations, are important contributors to genome variation in human disease. Loss of genes that protect the integrity of the genome in cancerous cells often results in an extreme degree of CGRs (Lee et al. 2016). Another example of CGRs called chromoanasynthesis (Carvalho and Lupski 2016), which combines chromosomal rearrangements with copy-number gains, leads to various severe congenital disorders, including MECP2 duplication syndrome (Carvalho et al. 2011) and Pelizeaus-Merzbacher disease (Beck et al. 2015). Several molecular mechanisms that could accout for these CGRs were discussed in the literature. They include FoSTeS (**fo**rk **s**talling and **te**mplate **s**witching) (Zhang et al. 2009), BIR (**b**reak-**i**nduced **r**eplication) (Costantino et al. 2014), MMBIR (**m**icrohomology-**m**ediated **b**reak-**i**nduced **r**eplication) (Zhang et al. 2009; Sakofsky et al. 2015) and others.

It was also noticed that DNA repeats that can form various non-B DNA structures (DNA cruciforms, triplex H-DNA, G4-DNA, *etc*) were associated with the locations of break points of such CGRs (Bacolla et al. 2016; Carvalho and Lupski 2016). A particular class of repetitive sequences called trinucleotide repeats was implicated in hereditary human diseases known as repeat-expansion diseases, such as Huntington's disease, Fragile X syndrome and Friedrich's ataxia (Pearson et al. 2005; Mirkin 2007; Orr and Zoghbi 2007). The ability of trinucleotide repeats to form non-B DNA structures was shown to lead to polymerase stalling during DNA replication, transcription and repair, ultimately resulting in their instability (expansions and contractions of the repeat tract) (Usdin et al. 2015; Neil et al. 2017; Polleys et al. 2017; Polyzos and McMurray 2017). We and others have also shown that trinucleotide repeat can induce mutagenesis at a distance (RIM- repeat induced mutagenesis) and trigger CGRs (Shah et al. 2012; Saini et al. 2013; Tang et al. 2013).

While understanding the fine structure of CGRs can shed light on the origin and the mechanisms of human diseases, their detection has never been a straightforward affair. Visual analysis of karyotypes is limited to events that are very large, typically involving entire chromosome arms. Fluorescent in-situ hybridization (FISH) allows detection of particular sequences that appear in unexpected locations (Aten et al. 2008). In *S. cerevisiae*, the relatively short length of chromosomes allows their separation by size via contour-clamped homogeneous electric field (CHEF) gel electrophoresis (Vollrath and Davis 1987). Combined with Southern blotting, this approach allows estimation of medium to large-scale changes in chromosome size, and can indicate whether particular regions have undergone translocation. However, the process is extremely laborious and limited in resolution.

Comparative genomic hybridization (CGH) arrays offer a vast improvement in resolution over visual methods, and can detect specific copy number changes. This approach has been used to map structural variation in the human population (Iafrate et al. 2004; Sebat et al. 2004), as well as to uncover specific CGRs in human genomic disorders (Lee et al. 2007; Potocki et al. 2007; Carvalho et al. 2009). However, inversions and translocations do not appear as copy number changes, and extensive follow-up PCR and Sanger sequencing is required to map CGR junctions with base pair-specificity. Even then, it is not always possible to map the boundaries of CGRs occurring in repetitive regions.

More recently, whole-genome and exome sequencing has been used to detect structural variation in model systems, human populations, cancer and other settings (Kidd et al. 2008; Stephens et al. 2009; Genomes Project et al. 2010; Macintyre et al. 2016; Jeffares et al. 2017). Copy number changes are represented by changes in read-depth, and the sequences themselves can reveal junctions. However, analysis of CGRs has been hindered by the short sequencing reads that are inherent to the most commonly used sequencing platforms, such as Illumina. Typically under 300 bp, relatively few reads will happen to fall on CGR junctions, and may not be distinguishable as such if they fall within repetitive elements. Various experimental and computational approaches have been developed to overcome these hurdles to the extent possible, though many limitations remain (Alkan et al. 2011).

The latest developments in CGR detection have involved long read sequencing technologies. Pacific Biosciences first developed a single-molecule sequencing approach capable of producing reads of more than 20 kilobases (kb). This has been used to identify CGRs in patients with Potocki-Lupski syndrome and Pelizeaus-Merzbacher disease (Wang et al. 2015; Zhang et al. 2017). Due to the relatively high cost compared with Illumina sequencing, these studies used targeted sequence-capture approaches to focus on known regions of interest. Whole genome sequencing has been feasible in *S. cerevisiae*, allowing detection of structural variation between different strains (Yue et al. 2017). Most recently, Oxford Nanopore Technologies has developed the MinION, a single-molecule sequencing approach where DNA strands are unwound and passed through a protein pore. The shape of each nucleotide restricts the flow of ions through the pore to a different degree, allowing identification of the bases. Most importantly, there appears to be nearly no limit on the read length, aside from the length of the DNA polymer itself following purification. In practice, reads can reach hundreds of kilobases (Jain et al. 2016). These extremely long reads have already proved useful in genome assembly and structural variation detection (Loman et al. 2015; Jain et al. 2016; Norris et al. 2016; Debladis et al. 2017; Istace et al. 2017; Jain et al. 2017; Jansen et al. 2017).

Here we decided to explore the potential of Nanopore sequencing as a method for characterizing the DNA repair pathways involved in CGRs caused by unstable microsatellite repeats. Our labs have used *S. cerevisiae* to study the length instability and CGRs caused by $(GAA)_n$ repeats, which are responsible for Friedrich's ataxia, as well as interstitial telomeric sequences (ITS) (Shishkin et al. 2009; Shah et al. 2012; Aksenova et al. 2013). Previously, these CGRs were identified using a combination of CGH arrays, CHEF gels, Southern blotting, PCR and Sanger sequencing (Kim et al. 2008; Shishkin et al. 2009; Aksenova et al. 2013; Tang et al. 2013). It appeared that a number of the events were truly complex, involving various combinations of chromosomal arm inversions, BIR responsible for arm duplications, and/or non-allelic homologous recombination (HR) mediated by microsatellites and transposable elements (Kim et al. 2008; Aksenova et al. 2013). However, these approaches were extremely laborious, limited in resolution, and hindered by the repetitive elements involved. The present study is dedicated to CGR triggered by $(GAA)_n$ repeats. We evaluated whether the ultra-long reads of Nanopore sequencing could effectively identify spontaneous $(GAA)_n$-mediated CGRs in a single step. Because of the potential for CGRs involving chromosome-scale changes, we chose a whole-genome sequencing approach, as opposed to targeted sequence capture. Our results demonstrate that Nanopore sequencing is an effective and efficient method of identifying novel CGRs in *S. cerevisiae,* which provided important insights into the mechanisms of DNA repair.

Results

Initial characterization of CGRs

To generate strains with CGRs, we used a previously-characterized selectable system for repeat instability in *S. cerevisiae* (Shishkin et al. 2009; Shah et al. 2012), in which (GAA)n repeats are located within an intron inside of the counter-selectable marker gene *URA3*. Selecting for inactivation of the URA3 gene most frequently turns up expansions of the repeat tract, which is the type of mutation most commonly associated with the inheritance of Friedrich's ataxia (Pandolfo 2002). However, the same process also selects for large deletions and CGRs that remove or separate the two halves of the split *URA3* gene. Because the selectable cassette is located in a region on Chromosome III (Fig. 1A) that contains essential genes both centromere-proximal and distal to the repeats, this precludes simple chromosomal arm loss, leading to more complex DNA repair events. In this system, probable CGR events are detected by the lack of a PCR product that typically amplifies the repetitive cassette. 23 strains with probable CGRs mediated by $(GAA)_n$ repeats in the *URA3* cassette were analyzed by CHEF gels combined with Southern hybridization (characteristic results are shown in Fig. S1) followed by CGH analysis as previously described (Aksenova et al. 2013). Using this course of analysis, 16 of the strains showed a likely gene conversion event between our *UR*-$(GAA)_{100}$-*A3* cassette on Chromosome III and the *ura3-52* allele, an inactive copy of the *URA3* gene remaining on Chromosome V. This appears similar to what was previously observed as a rare event for ITS (Aksenova et al. 2013). The remaining strains showed evidence of more complex rearrangements that were not fully resolvable from the initial analysis.

Nanopore sequencing approach

To unambiguously characterize the observed GCRs, the CGR strains together with our starting strain SMY502 (Shah et al. 2012) were subjected to Nanopore sequencing. DNA from each strain was purified and used to construct barcoded sequencing libraries. The libraries were then pooled and sequenced together on a single flow cell resulting in roughly 30x coverage per strain. Nearly three gigabases of total sequence were generated, largely by reads with a length of 20-to-30 kb and above (Fig. S2).

Genomic alterations in the parent yeast strain

Our parent yeast strain is closely related, but not identical to S288C, the extremely well-characterized laboratory strain used in the initial systematic sequencing of *S. cerevisiae.* In order to identify CGRs in the Nanopore sequences, it was first necessary to examine the parent strain for changes relative to the S288C reference genome available from the Saccharomyces Genome Database (SGD). To do this, reads were aligned to S288C and examined for potential structural variants. The alignment/variant-calling approach was chosen, as opposed to genome assembly, because it involved significantly fewer compute resources, and because the S288C genome is extremely well-characterized and closely-related to our parent strain. Alignments were visualized using Ribbon, a sequence visualization tool specializing in split reads, or reads that map to multiple genomic locations (Nattestad et al.). In addition, the alignments were visualized using the bioinformatics software UGENE, which can display a pileup of reads for each chromosome (Okonechnikov et al. 2012).

Using this approach, we confirmed the presence of a number of known structural variants in our parents strain, including alterations in selectable marker genes, as well as the insertion of our *UR*-(GAA)$_{100}$-*A3* selectable cassette, and a mating type switch (Fig. 1B). This demonstrates a high success rate in finding relatively simple structural variations. We also found four unexpected Ty element insertions that were not present in the S288C reference genome, three of which appear on Chromosome III (Figs. 1A & S3).

The reference genome was altered to reflect to these observed changes, and used as the reference to which the remaining strains were aligned. We discuss here three independent CGR events observed in the strains #101, #118, and #105, which were analyzed independently by CHEF/CGH and by Nanopore sequencing.

Strain #101 – Gene conversion involving Ty retrotransposon elements

The CHEF/CGH analysis identified strain #101 as containing a *ura3-52* gene conversion event. Specifically, CHEF analysis showed that strain #101 had only one change: Chromosome III was slightly smaller than observed in the starting strain (Fig. 2). This smaller chromosome hybridized to three probes specific to genes on Chromosome III (*CHA1*, *LEU2*, and *RAD18*) as expected (Figs. 1A & 2). By CGH arrays (Fig. 3A), strain #101 had a deletion with a left endpoint located between SGD coordinates 75,142-75,758, which overlaps with the location of the *UR*-(GAA)$_{100}$-*A3* cassette (replacing SGD coordinates 75,594-75,641). The right end of the deletion had a breakpoint between SGD coordinates 82,646-84,263. This region overlaps with a cluster of Ty retrotransposon elements, including an unannotated Ty1 element replacing *YCLWdelta15* (SGD coordinates 82,700-83,036, Fig. S3), a Ty2 element (*YCLWTy2-1* at SGD coordinates 84,811-90,769), and multiple delta sequences (long terminal repeats (LTRs) left behind by ancestral Ty elements). Note that, while strain #101 represents a frequently-observed *ura3-52* gene conversion event (see Fig. S1), it is likely that similar CGRs vary in their interactions with the particular Ty elements in this cluster.

The Nanopore sequencing analysis of strain #101 arrived at the same conclusion, and was able to narrow the breakpoints to single base pair resolution (Figs. 3D & S4C). In particular, the right end of the deletion was shown to extend into the 5’ LTR of *YCLWTy2-1* (Fig. 3B,D). Ribbon displayed a ~6kb insertion at these breakpoints (Figs. 3C & S4A,B), which is the length of a typical Ty element. While the inserted sequence could not be mapped with high confidence in each individual read, the *ura3-52* gene was the most commonly identified (Fig. S4B). The *ura3-52* gene could also be identified by SNPs in the consensus sequence (Fig. S4C). Thus, the rearrangement hypothesized in the CGH and CHEF gel-electrophoresis analyses was confirmed and refined through our analysis of Nanopore whole-genome sequence.

Based on these observations, we suggest that a DSB occurred within the $(GAA)_n$ tract. The centromere-distal side of the break was resected into the 5’ end of *URA3* and the centromere-proximal side of the break was processed into or near *YCLWTy2-1* (Fig. 3E). These broken ends initiated HR with the *ura3-52* gene on Chromosome V, and repair of the resulting gap produced a gene conversion event.

Strain #118 – Gene conversion involving $(GAA)_n$ repeats

The CHEF analysis of strain #118 showed that Chromosome III appeared ~10 kb longer, and no other changes were observed (Figs. 2 and S1). Since Chromosome III is 365 kb-long, a difference in size of 10 kb is often difficult to visualize; this small difference in size is more obvious in Fig. S1. This chromosome hybridized to probes for all three genes along Chromosome III (Fig. 2). By CGH, strain #118 showed a deletion of the second half of the *UR*-(GAA)$_{100}$-*A3* cassette, similar to strain #101 (Fig. 4A). Additionally, CGH analysis showed a duplication of a 15-to-21 kb portion of Chromosome II (corresponding to SGD coordinates 205,204-205,992 and 220,575-226,877) (Fig. 4A). The right end of this duplication overlaps with the Ty element *YBLWTy1-1*. This implied the possibility of a gene conversion event, similar to strain #101, but involving Chromosome II as the donor. Subsequent CHEF analysis (Fig. 2) showed that the longer Chromosome III did not hybridize to a probe for the Chromosome II centromere (*CEN2*), and confirmed that Chromosome II did not change in size, eliminating the possibility of a reciprocal translocation.

Nanopore sequencing of strain #118 arrived at the same conclusion as the CHEF/CGH analysis. We see a deletion on Chromosome III and a duplication on Chromosome II consistent with CGH analysis (Fig. 4B). Further, sequencing revealed that the centromere-distal junction of this gene conversion links the (GAA)$_{100}$ repeat in the cassette with an imperfect (GAA)$_n$ repeat inside the *SCT1* gene on Chromosome II (Figs. 4D & S5C). Nanopore sequencing also revealed single-base resolution of the centromere-proximal junction between the Chromosome II *YBLWTy1-1* and the unannotated Ty1 element located ~11 kb downstream of the *UR*-(GAA)$_{100}$-*A3* cassette (Fig. 4D & S5D). Multiple reads were found crossing each of these junctions (Fig. S5A,B), and indeed, at least two reads were able capture the complete ~17 kb insertion, unambiguously identifying the event as a gene conversion (Fig. 4C). Note that a gene conversion consisting of a ~17 kb insertion and an ~11 kb deletion results in a ~6 kb-longer Chromosome III, which is qualitatively consistent with the CHEF analysis.

Altogether, these results indicate that the CGR in strain #118 likely resulted from a DSB within the $(GAA)_{100}$ repeats in the cassette. The centromere-distal broken end was processed only a short distance, or not at all, depending on the exact location of the DSB within the repeats, and then recombined with the imperfect $(GAA)_n$ repeat within *SCT1*. The centromere-proximal broken end was processed to the unannotated Ty1 element adjacent to *YCLWTy2-1*, which then interacted with the homologous *YBLWTy1-1* on Chromosome II. Invasion of the broken ends into Chromosome II, followed by gap repair, produced the gene conversion (Fig. 4E).

Strain #105 – highly-complex rearrangement involving chromosome-scale duplications

Strain #105 was the most complex strain predicted from the CGH/CHEF analysis. By CHEF analysis, Chromosomes II and III appeared to be replaced by three novel chromosomes with approximate lengths of ~700 kb, ~480 kb, and ~440 kb (Fig. 2). The ~700 kb chromosome hybridized to the *CEN2* probe and the *CHA1* probe from Chromosome III, while the ~480 kb chromosome hybridized to the Chromosome III probes *LEU2* and *RAD18,* and the ~440 kb chromosome hybridized to the *CEN2* and *RAD18* probes (Fig. 2). By CGH analysis (Fig. 5A), we observed a deletion centromere-proximal to the *UR*-$(GAA)_{100}$-*A3* cassette, similar to those deletions in strains #101 and #118. In addition, sequences on

Chromosome III were duplicated from SGD coordinates ~123,000-168,000, and triplicated from ~168,000 to the end of the chromosome. These approximate breakpoints overlap with a solo LTR (*YCRCdelta6* – SGD coordinates 124,134-124,465) and an unannotated pair of Ty elements replacing *YCRWdelta11* (SGD coordinates 169,573-169,888) (Lemoine et al. 2005) (Fig. S3). In addition, the CGH analysis showed a duplication of a centromere-containing ~50 kb region of Chromosome II (SGD coordinates ~205,000-260,000). The left end of this duplication again overlaps the *SCT1* gene, while the right end overlaps a Ty1 element (*YBRWTy1-2*).

Nanopore sequencing of strain #105 revealed the same copy number changes as observed in the CGH analysis, with the ratio of the read depth corresponding to the duplication and triplication regions as predicted (Fig. 5B). Three groups of split reads were observed (Figs. 5C). The first group consists of the left half of the *UR*-$(GAA)_{100}$-*A3* cassette joined to the *SCT1* gene on Chromosome II (Figs. 5C,D & S6A,B). This junction is nearly identical to that in strain #118, except that a larger portion of the $(GAA)_{100}$ repeat appears intact (Fig. S6C). The second group of split reads map from *YBRWTy1-2* on Chromosome II to the second of the two unannotated Ty1 elements located at *YCRWdelta11* on Chromosome III, making up the duplication/triplication border (Fig. 5C,D & S6A,B). Interestingly, SNPs located at this junction indicate the presence of a small portion of *YCLWTy2-1* located in between *YBRWTy1-2* and the unannotated Ty1 (Figs. 5D and S6E). Non-split reads map to the same breakpoint on Chromosome III in an approximate 2:1 ratio with the split reads (Fig. S6A,B).

Both ends of the Chromosome II duplication also show non-split reads mapping across breakpoints in an approximate 1:1 ratio with the split reads (Fig. S6A,B). Finally, the third group of split reads maps the centromere-proximal broken end from *YCLWTy2-1* to *YCRCdelta6* on the opposite side of the Chromosome III centromere (Fig. 5C,D & S6A,B). These two loci are linked in an inverted orientation, and this junction corresponds to the single-copy/duplication border on Chromosome III (Fig. 5D & S6A,B). Non-split reads map to the *YCRCdelta6* loci in an approximate 1:1 ratio with the split reads. (Fig. S6A,B). There is no evidence for any other consistent group of split reads elsewhere in the genome.

The CGH/CHEF analysis combined with the Nanopore sequencing reveal key features of this complex genomic rearrangement with little ambiguity. The main limitation of our Nanopore sequencing analysis is due to the lack of reads spanning the entire ~50kb duplicated region of chromosome II. This unfortunately prevents the distinction between gene conversion and reciprocal translocation involving the duplicated portion of chromosome II. In the scenario in which a reciprocal translocation occurred between chromosomes II and III, the predicted chromosome sizes from the Nanopore analysis match the novel chromosomes observed in the CHEF gel.

Given the sheer complexity of CGRs in #105, one could have imagined multiple possible pathways. The most plausible scenario based on the combination of our results is presented in Fig. 5E. We suggest that after a DSB originated within the $(GAA)_{100}$ tract, the broken chromosome was duplicated, resulting in two copies of Chromosome III with four broken ends. One copy underwent the following rearrangements: its centromere-proximal end was processed to the unannotated Ty1 element adjacent to *YCLWTy2-1*, and this end invaded *YCRCdelta6,* initiating a BIR event that duplicated the whole right arm of Chromosome III distal to *YCRCdelta6*. This intrachromosomal BIR event generated the ~480 kb-long chromosome (III-III in Fig. 5E). The centromere-distal acentric arm of Chromosome III was likely lost during cell division. The second broken Chromosome III was repaired as follows: The centromere-distal end crossed with the imperfect $(GAA)_n$ tract in *SCT1* on Chromosome II. The centromere-proximal end was processed to the *YCLWTy2-1* element and crossed with the *YBRYTy1-2* of Chromosome II. After the gap repair event and crossover resolution, two translocated chromosomes were formed: a ~700 kb-long hybrid chromosome (III-II in Fig. 5E) and an unstable II-III dicentric chromosome. *CEN3* was subsequently lost from this dicentric by recombination between the *YCLWTy2-1* element and the Watson-oriented unannotated Ty element that replaces *YCRWdelta11* on the right arm of the chromosome, resulting in the ~440 kb product (II-III in Fig. 5E). Loss of one centromere in dicentric chromosomes as a consequence of recombination between flanking repeats has been reported previously (Brock and Bloom 1994; Lemoine et al. 2005).

Discussion

Our study is the first to directly compare the use of Nanopore whole-genome sequencing to traditional methods used to identify and map CGRs in *S. cerevisiae*. We show that this approach was able to replicate results from the established but more laborious techniques used in prior studies, and was further able to uncover novel observations of CGRs that were not easily resolvable through prior methods. In addition, this study represents the first extensive investigation of complex genomic rearrangements resulting from spontaneous breakage of a microsatellite sequence located in an essential chromosome region. In a previous study, we examined genetic alterations in a strain in which the $(GAA)_n$ tract was inserted in a non-essential region and in which we selected events that resulted in loss of sequences distal to the tract (Kim et al. 2008). Thus, this analysis was biased toward the recovery of non-reciprocal BIR events. Our current analysis allows an exploration of a more varied spectrum of events, both reciprocal and non-reciprocal.

In strain #101, a DSB generated within the $(GAA)_{100}$ repeat was able to be processed to expose homology in the 5’ end of the *URA3* gene, allowing recombination with the inactive *ura3-52* gene. Events of this type were previously observed in experiments in which the *URA3* reporter gene had interstitial telomeric sequences (ITS) instead of $(GAA)_n$ repeats (Aksenova et al. 2013). In contrast, in strains #118 (Fig. 4) and #105 (Fig. 5), HR was initiated directly from the broken $(GAA)_n$ repeats on Chromosome III, invading an imperfect $(GAA)_n$ repeat within the *SCT1* gene on Chromosome II. $(GAA)_n$ and other homopurine repeats have previously been observed to promote CGRs, both in our previous studies in *S. cerevisiae*, as well as in cancer genomes (Kim et al. 2008; Bacolla et al. 2016). Interestingly, in strain #105, nearly the full $(GAA)_{100}$ repeat is present following the CGR, while strain #118 contains only ~25 repeats following the rearrangement (Figs. S5C, S6C,D). These differences may reflect the tendency of the DSB to form close to the 5’ end vs. the 3’ end of the repeat, or may reflect variability in end-processing efficiency at $(GAA)_n$ repeats. This observation also demonstrates the ability of Nanopore sequencing to measure the length of long $(GAA)_n$ microsatellites. This use of long-read sequencing technologies to uncover difficult-to-measure variations in microsatellite length is another important area of focus that is relatively unexplored (Liu et al. 2017).

Another novel observation is of the numerous and varied rearrangements that occurred in strain #105 as the result of a single originating DSB. Three recombinant chromosomes were produced, likely involving four broken DNA ends (Fig. 5E). One of these ends initiated an intrachromosomal BIR that generated a new chromosome in which the right arm of Chromosome III was duplicated. Two other broken ends of Chromosome III interacted with Chromosome II at loci ~50kb apart. Repair of this gap required extensive DNA synthesis, possibly involving DNA repair or a BIR-like mechanism originating from both broken ends. The meeting of these two synthesis events would result in the formation of a double-Holliday junction that could be processed into two translocated chromosomes. Finally, one of the translocated chromosomes appeared to be an unstable dicentric, which subsequently lost one centromere via an intrachromosomal recombination between two Ty elements. The presence of SNPs from a third Ty element at the junction of this recombination (Figs. 5D & S6E) is strong evidence that this event indeed took place. Altogether, different non-reciprocal mechanisms involving both intra- or inter-chromosomal interactions are involved in the repair of broken DNA ends resulting from a single DSB. Note that this scenario could explain complex de-novo rearrangements called chromoanasynthesis that were observed in several human congenital disorders (Carvalho and Lupski 2016). Importantly, unraveling this CGR required a whole-genome sequencing approach, which was able to identify chromosome-scale duplication and triplication events that would not have been observed via targeted sequencing of the $(GAA)_n$ region.

The question arises of how much sequencing coverage is needed to determine the nature of various CGRs. Ultimately, multiple factors must be considered. In the example of strain #118, our lowest-covered sample, two reads unambiguously showed that the ~20 kb insertion was resolved via gene conversion. While a number of additional reads map separately to each of the III-II and II-III junctions, ultimately our interpretation relies on these two reads that spanned the entire event. Thus, we might consider ~20x coverage to be nearly the minimum coverage required to map a ~20 kb insertion, given a ~25 kb average read length. However, in the example of strain #105, we did not find any reads covering the entire ~50 kb insertion end-to-end, despite a higher coverage of ~40x, which limited our ability to distinguish between gene conversion and crossover resolution. Thus, the ability to unambiguously identify an event is a function of the coverage, the average read length, and the size of the insertion in question. With a greater average read length, less coverage may be required. Nanopore sequencing is capable of ultra-long reads in excess of ~800 kb when careful DNA extraction techniques are employed (Liu et al. 2017). One strategy to reduce the sequencing burden would be to sequence a large number of samples at low coverage, then identify ambiguous CGRs and sequence those samples to greater depth.

Without the crucial component of long-read sequencing, the intricacies of these CGRs would not have been uncovered. Nanopore sequencing brings a rapid and effective new method of analysis for CGRs, allowing single base pair resolution of breakpoints and long reads that span repetitive regions. This analysis can be applied to whole genome sequences for the identification of previously uncharacterized CGRs without *a priori* knowledge of the regions involved. Importantly, the level of detail obtained through this method is sufficient to extensively interrogate the mechanisms of DNA repair involved in CGR formation. For analyzing CGRs in *S. cerevisiae*, this technology is already capable of sequencing large numbers of genomes at relatively low cost. The techniques employed here can be performed in a small lab with minimal specialized equipment and a modest level of expertise. As the output and accuracy of this developing technology continues to improve, similar analysis of human genomes, including cancer genomes with numerous complex rearrangements, will surely be possible.

Methods

Generation and isolation of yeast strains containing CGRs:

The parent strain, SMY502, is a haploid strain of *S. cerevisiae* derived from FY1679 (*ura3-52, his3Δ200, leu2Δ1, trp1Δ63, bar1*::*HIS3MX6*, mat a). It also contains the *UR*-$(GAA)_{100}$-*A3* selectable cassette, located ~1kb downstream of the replication origin ARS306. Fluctuation tests were performed with SMY502 as previously described (Shah et al. 2012). Briefly, strains are grown in the presence of 5-FOA, which selects against an active *URA3* gene. Inactivation of the *UR*-$(GAA)_{100}$-*A3* selectable cassette results in 5-FOA resistant colonies, which were then categorized for mutation type as previously described (Shishkin et al. 2009), via PCR primers located just outside of the repeats. Those colonies that showed a lack of this PCR product were tested with a further PCR primer pair that amplifies the entire selectable cassette, in order to distinguish CGRs from short deletions. Strains with possible CGRs were saved as frozen stocks at -80ºC.

CHEF gel and CGH array analysis:

Experiments and analysis performed as previously described (Aksenova et al. 2013).

DNA extraction:

DNA was extracted via ethanol precipitation. See Supplemental Methods for details. This method of DNA preparation resulted in an average fragment size of

24-48 kb (Fig. S2A). DNA quantity was measured via Qubit (Qubit dsDNA BR Assay kit – Thermo Scientific).

Nanopore sequencing:

1.5ug of purified DNA was used to construct barcoded sequencing libraries, using the ONT (Oxford Nanopore Technologies) Ligation Sequencing Kit 1D (SQK-LSK108) in combination with the Native Barcoding Kit 1D (EXP-NBD103). All procedures recommended in the ONT-provided protocol were followed, including nick repair (NEBNext FFPE Repair mix – New England Biolabs). The libraries were pooled and sequenced together on a single SpotON Flow Cell Mk I R9.4 (FLO-SPOTR9) for 48 hours.

Bioinformatics:

Raw current traces generated by ONT sequencing were basecalled via the Albacore basecalling software (ONT version 2.02). For the parent strain, reads were then aligned to the S288C reference genome (R64-1-1, obtained from ensembl.org) using NGM-LR (Sedlazeck et al. 2017). The output was imported to Ribbon (Nattestad et al.) as well as the bioinformatics software UGENE (Okonechnikov et al. 2012), for visualization. This analysis of the parent strain identified various deletions, insertions and SNPs (Figs. 1B & S3), which were then incorporated into the reference genome. Following this, the above analysis pipeline was repeated for each strain. Single base pair resolution of breakpoints within Ty elements was determined by analysis of SNPs within each Ty element of origin. See Supplemental Methods for more details.

Data access

The sequencing data from this study have been submitted to the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under accession number SRP111355.

Acknowledgements

We thank Oxford Nanopore Technologies for providing us with MinION device and necessary reagents under the MinION Access Program. Supported by the NIH grants R01GM60987 and P01GM105473 to SMM and R35GM118020 to TDP.

Author Contributions

Conceptualization, R.J.M. and S.M.M. Methodology, R.J.M., T.D.P. and S.M.M. Software, R.J.M. Investigation, R.J.M., R.G.R., A.J.N., M.G. and D.K. Writing – original draft, R.J.M. and S.M.M. Writing – review and editing, R.J.M., T.D.P. and S.M.M. Visualization, R.J.M. and T.D.P. Supervision, T.D.P. and S.M.M. Funding acquisition, T.D.P. and S.M.M.

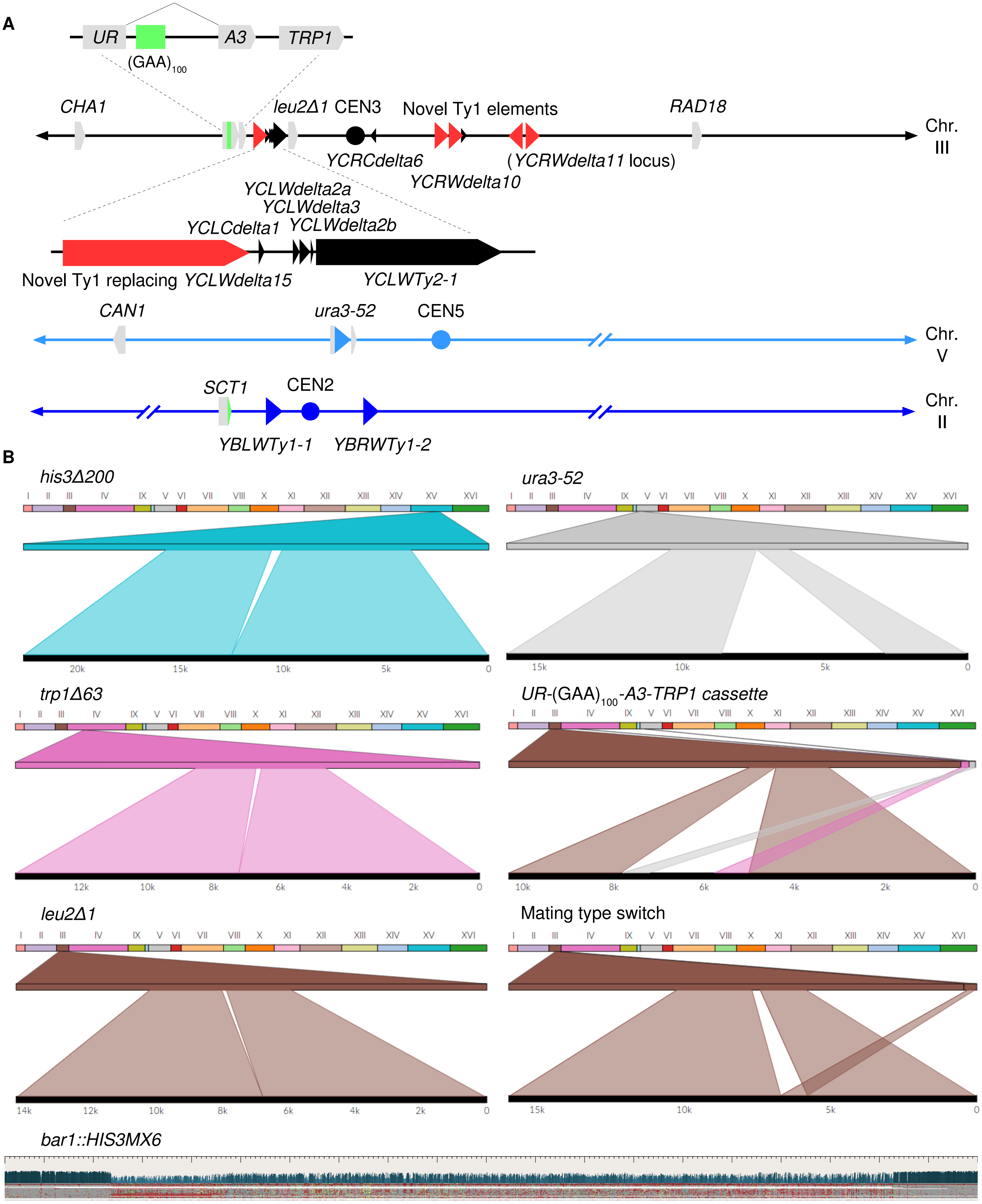

A
UR
A3
TRP1
$(GAA)_{100}$
CHA1
leu2Δ1
CEN3
Novel Ty1 elements
RAD18
Chr. III
YCRCdelta6
(YCRWdelta11 locus)
YCRWdelta10
YCLWdelta2a
YCLWdelta3
YCLCdelta1
YCLWdelta2b
Novel Ty1 replacing YCLWdelta15
YCLWTy2-1
CAN1
ura3-52
CEN5
Chr. V
SCT1
CEN2
Chr. II
YBLWTy1-1
YBRWTy1-2
B
his3Δ200
ura3-52
trp1Δ63
UR-$(GAA)_{100}$-A3-TRP1 cassette
leu2Δ1
Mating type switch
I II III IV IX V VI VII VIII X XI XII XIII XIV XV XVI
20k 15k 10k 5k 0
15k 10k 5k 0
12k 10k 8k 6k 4k 2k 0
10k 8k 6k 4k 2k 0
14k 12k 10k 8k 6k 4k 2k 0
15k 10k 5k 0
bar1::HIS3MX6

Fig. 1. Known deletions and rearrangements in SMY502 vs. S288C.

A) Map of Chromosomes II, III and V, indicating positions of genes, centromeres and Ty elements. Diagonal lines represent contiguous sequences not in display, such that the displayed portions are pictured to scale. Genes are shown by grey boxes with points indicating the orientation. Centromeres are marked by circles. Ty elements are indicated by triangles, with black indicating Ty elements found in the S288C reference genome, and red indicating previously unannotated Ty elements. The point of the triangle indicates orientation. Small black triangles represent solo LTR sequences, also known as delta elements. Zoomed-in portions of Chromosome III show a cluster of Ty elements, as well as the *UR*-$(GAA)_{100}$-*A3* selectable cassette. Bright green portions represent the location of $(GAA)_n$ repeats.

B) Ribbon single-read views highlighting known large-scale genomic changes in the SMY502 parent strain, mapped to the S288C reference genome. For each panel, the top bar contains a color-coded list of chromosomes, while the bottom black bar displays the full sequencing read. Windows connect which portion of the read maps to which chromosomal position. *his3Δ200*, *trp1Δ63* and *leu2Δ1* are 1-2kb deletions, and are observed as split reads in which part of the reference sequence (top) is missing from the read (bottom). *Ura3-52* is an insertion of a ~6kb Ty element, observed as a split read in which sequence not in the reference (top) is found in the read (bottom). Because of high sequence similarity among Ty elements, the insertion is not always associated with a particular part of the reference sequence in each individual read. Our *UR*-$(GAA)_{100}$-*A3*-*TRP1* selectable cassette also maps as an insertion, but the 5' portion of *URA3* and the *TRP1* gene are both matched to their respective genomic locations in the reference sequence. The difference in mating types between S288C and our strain is also observed as a split read, due to the peculiar control of yeast mating type, in which one of two inactive regions on either end of Chromosome III is copied via HR into the active mating loci, located near the middle of Chromosome III. One allele, *bar1*::*HIS3MX6*, was not apparent in the Ribbon analysis, because the *BAR1* gene was replaced with a similarly-sized marker gene. The HIS3MX6 marker was aligned to the reference with a number of mismatches and short gaps, which was readily apparent in the UGENE alignment. In this view, gray boxes indicate bases that match the reference sequence, colored boxes indicate bases that do not match the reference sequence (blue=G, green=C, yellow=A, light red=T, dark red=deletion), and the blue bars above represent read depth at each position.

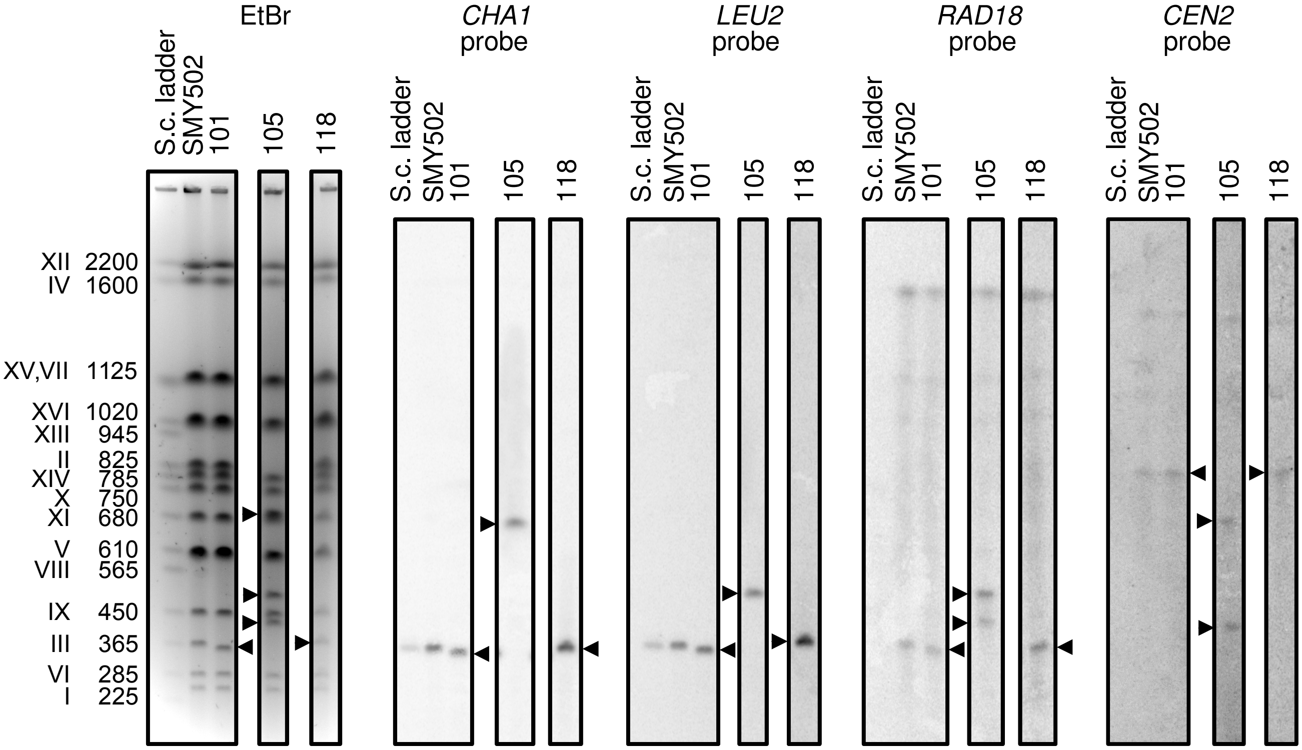


Fig. 2. CHEF gel analysis
CHEF gel electrophoresis was used to separate whole chromosomes by size. The left panel shows the gel stained with ethidium bromide, displaying all chromosomes. Lane 1 is a size ladder of *S. cerevisiae* chromosomes. The following lanes contain DNA prepared from the strains as indicated. Black triangles point to chromosomes with an altered size. The four right panels display Southern blot hybridizations using probes to the indicated genes.

Experiments performed in the lab of Tom Petes. Images contributed by Tom Petes.

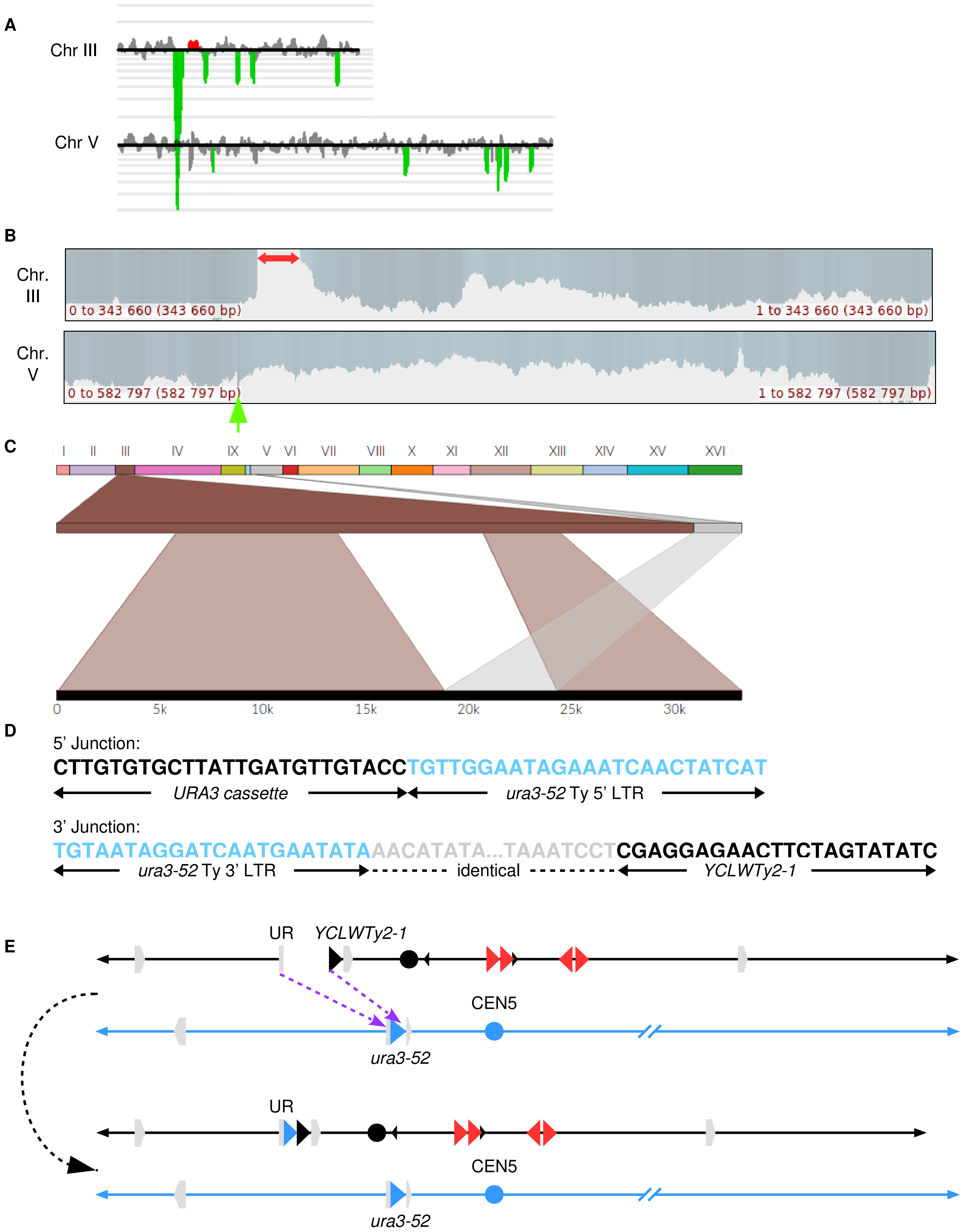

A
Chr III
Chr V
B
Chr. III
0 to 343 660 (343 660 bp)
1 to 343 660 (343 660 bp)
Chr. V
0 to 582 797 (582 797 bp)
1 to 582 797 (582 797 bp)
C
I II III IV IX V VI VII VIII X XI XII XIII XIV XV XVI
0 5k 10k 15k 20k 25k 30k
D
5' Junction:
CTTGTGTGCTTATTGATGTTGTACCTGTTGGAATAGAAATCAACTATCAT
URA3 cassette
ura3-52 Ty 5' LTR
3' Junction:
TGTAATAGGATCAATGAATATAAACATATA...TAAATCCTCGAGGAGAACTTCTAGTATATC
ura3-52 Ty 3' LTR
identical
YCLWTy2-1
E
UR
YCLWTy2-1
CEN5
ura3-52
UR
CEN5
ura3-52

Fig. 3. Identifying genomic rearrangements in strain #101.

A) CGH microarray analysis, displaying results for Chromosomes III and V. The large green region corresponds to the deletion surrounding the repeats. By examining the hybridization values for individual oligonucleotides on the microarrays, we found that the small red and green regions depicted in this figure do not represent true duplications and deletions, respectively. Experiments performed in the lab of Tom Petes. Images contributed by Tom Petes.

B) Nanopore sequencing coverage maps of Chromosomes III and V, generated via UGENE, with a red arrow highlighting the deletion boundaries.

C) Ribbon single-read view highlighting a read mapping the entirety of the gene conversion event in which a Ty element was inserted in place of the 3' half of the *UR*-(GAA)$_{100}$-*A3* cassette on Chromosome III.

D) Single base pair resolution of the 5' and 3' breakpoints of the deletion. The 5' junction connects the 5' portion of the *UR*-(GAA)$_{100}$-*A3* cassette with *ura3-52* on Chromosome V. Note that the crossover could have occurred anywhere in the 341 bp of identity between the cassette and *ura3-52*. The 3' junction consists of the 3' LTR region of *ura3-52* and YCLWTy2-1 on Chromosome III. The gray region represents an 80 bp window of identity between *ura3-52* and *YCLWTy2-1* in which the gene conversion occurred. SNPs are visible in the alignment on each side of this window (Fig. S4C).

E) Diagram of the CGR event resulting in a gene conversion. Chromosome maps have the same format as in Fig. 1A. Relevant features are labeled. Purple arrows indicate sites of HR invasion. The top portion displays the broken Chromosome III, processed to expose ends for HR, and the donor Chromosome V. The bottom portion displays the final chromosome products. See main text for details.

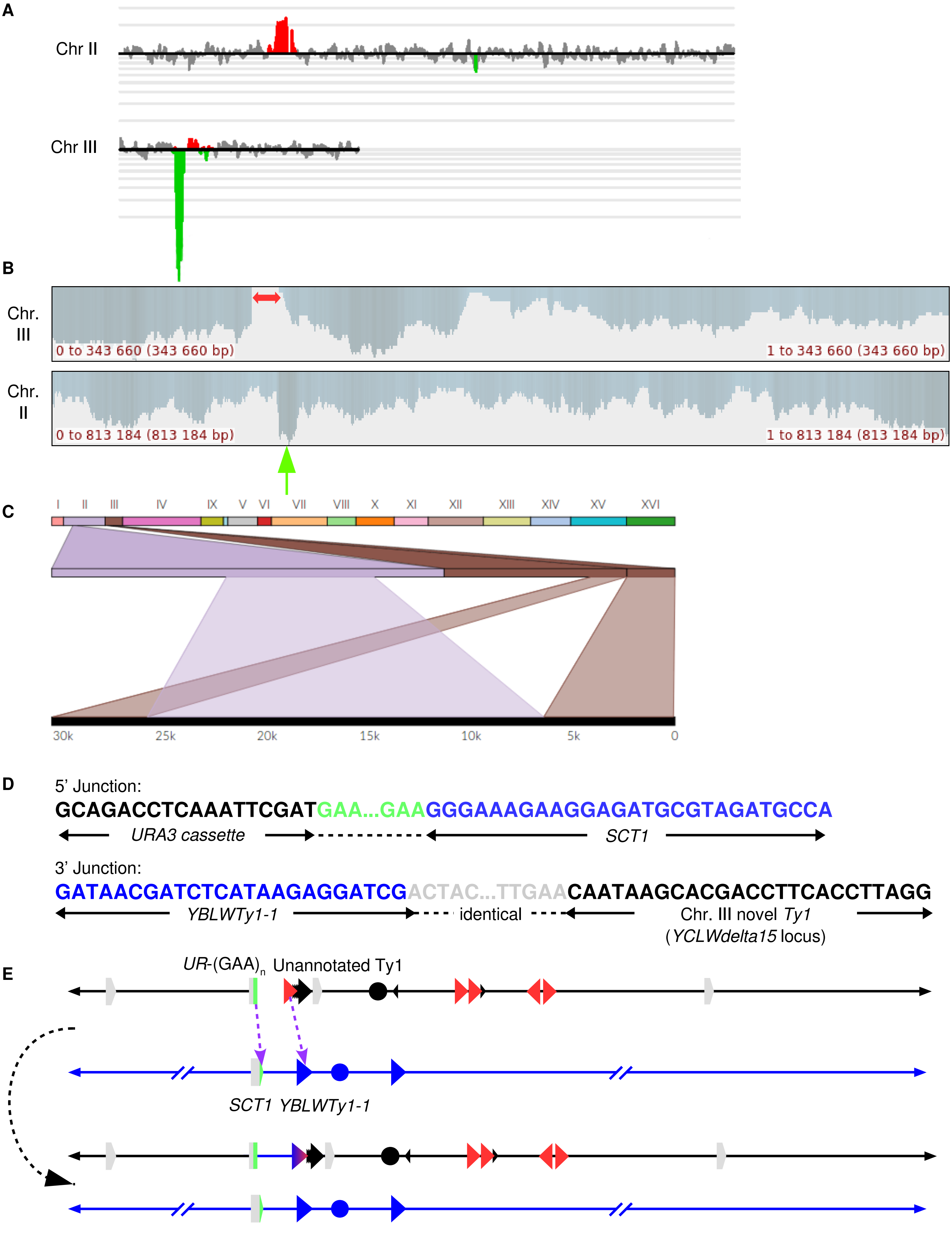

A
Chr II
Chr III
B
Chr. III
0 to 343 660 (343 660 bp)
1 to 343 660 (343 660 bp)
Chr. II
0 to 813 184 (813 184 bp)
1 to 813 184 (813 184 bp)
C
I
II
III
IV
IX
V
VI
VII
VIII
X
XI
XII
XIII
XIV
XV
XVI
30k
25k
20k
15k
10k
5k
0
D
5' Junction:
GCAGACCTCAAATTCGATGAA...GAAGGGAAAGAAGGAGATGCGTAGATGCCA
URA3 cassette
SCT1
3' Junction:
GATAACGATCTCATAAGAGGATCGACTAC...TTGAACAATAAGCACGACCTTCACCTTAGG
YBLWTy1-1
identical
Chr. III novel Ty1
(YCLWdelta15 locus)
E
UR-(GAA)n
Unannotated Ty1
SCT1
YBLWTy1-1

Fig. 4. Identifying CGRs in strain #118.

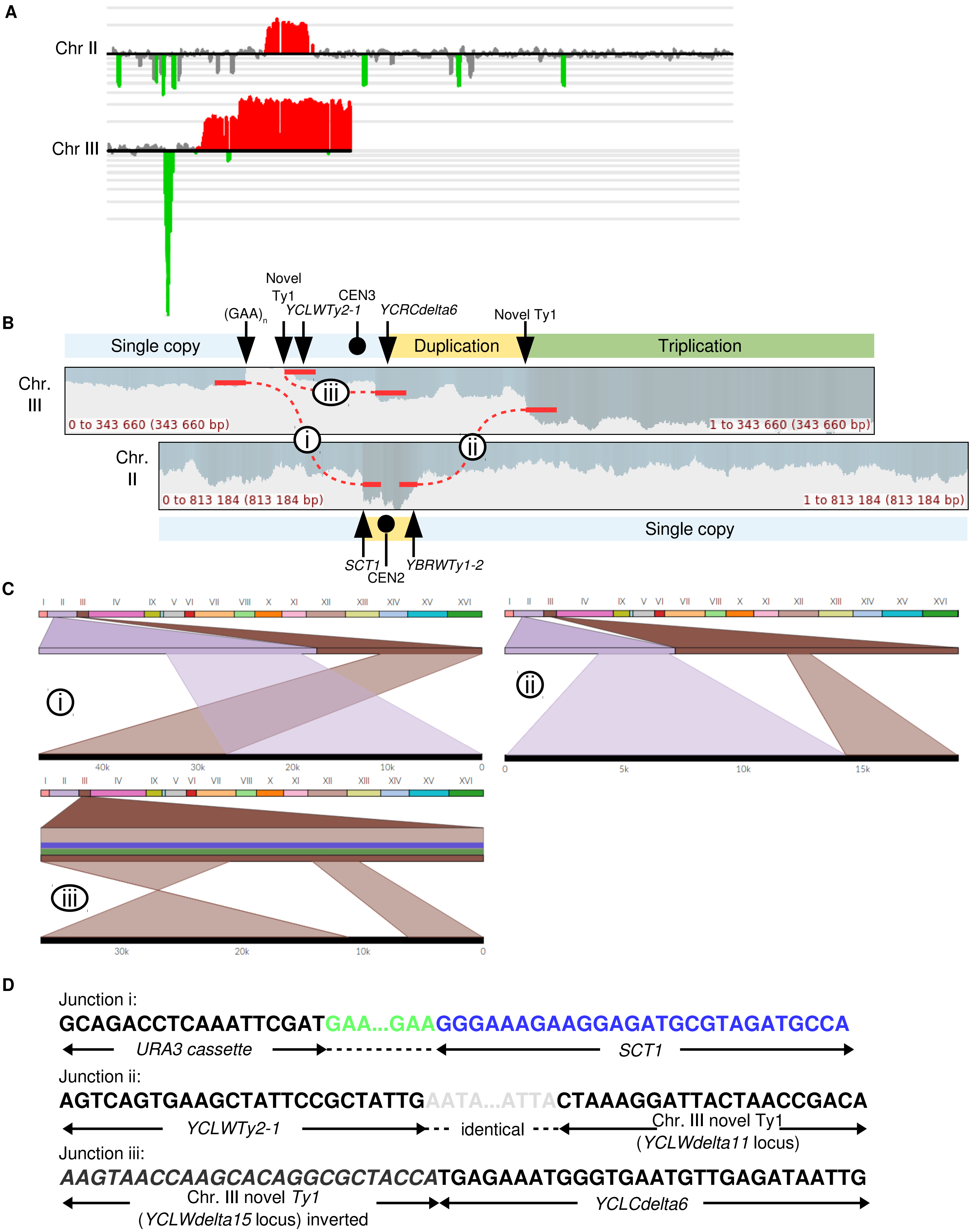
A) CGH microarray analysis, displaying results for Chromosomes II and III. The large green region corresponds to the deletion surrounding the repeats, while the red region corresponds to the duplication surrounding the *SCT1* locus. By examining the hybridization values for individual oligonucleotides on the microarrays, we found that the small red and green regions depicted in this figure do not represent true duplications and deletions, respectively. Experiments performed in the lab of Tom Petes. Images contributed by Tom Petes.

B) Nanopore sequencing coverage maps of Chromosomes III and II, generated via UGENE, with a red arrow highlighting the deletion boundaries, and a green arrow indicating a duplication.

C) Ribbon single-read view highlighting a long read that captured the entire gene conversion event, showing a ~20kb insertion of Chromosome II in place of the deleted region on Chromosome III.

D) Single base pair resolution of the 5' and 3' junctions between Chromosomes III and II. The 5' junction shows that the break and repair occurred within the $(GAA)_n$ repeats. The 3' junction shows that the recombination event occurred within Ty elements on Chromosomes II and III. The gray region represents a 23 bp window of identity, with SNPs on either side identifying the specific Ty element. (Fig. S5D) The unannotated Ty1 element is adjacent to *YCLWTy2-1* (Fig. S3).

E) Diagram of the CGR event resulting in a gene conversion. Chromosome maps have the same format as in Fig. 1A. Relevant features are labeled. Purple arrows indicate sites of HR invasion. The top portion displays the broken Chromosome III, processed to expose ends for HR, and the donor Chromosome II. The bottom portion displays the final chromosome products. See main text for details.


A
Chr II
Chr III
B
(GAA)$_n$
Novel Ty1
YCLWTy2-1
CEN3
YCRCdelta6
Novel Ty1
Single copy
Duplication
Triplication
Chr. III
0 to 343 660 (343 660 bp)
1 to 343 660 (343 660 bp)
iii
i
ii
Chr. II
0 to 813 184 (813 184 bp)
1 to 813 184 (813 184 bp)
Single copy
SCT1
CEN2
YBRWTy1-2
C
I II III IV IX V VI VII VIII X XI XII XIII XIV XV XVI
i
40k 30k 20k 10k 0
ii
0 5k 10k 15k
iii
30k 20k 10k 0
D
Junction i:
GCAGACCTCAAATTCGATGAA...GAAGGGAAAGAAGGAGATGCGTAGATGCCA
URA3 cassette
SCT1
Junction ii:
AGTCAGTGAAGCTATTCCGCTATTGAATA...ATTACTAAAGGATTACTAACCGACA
YCLWTy2-1
identical
Chr. III novel Ty1
(YCLWdelta11 locus)
Junction iii:
AAGTAACCAAGCACAGGCGCTACCATGAGAAATGGGTGAATGTTGAGATAATTG
Chr. III novel Ty1
(YCLWdelta15 locus) inverted
YCLCdelta6

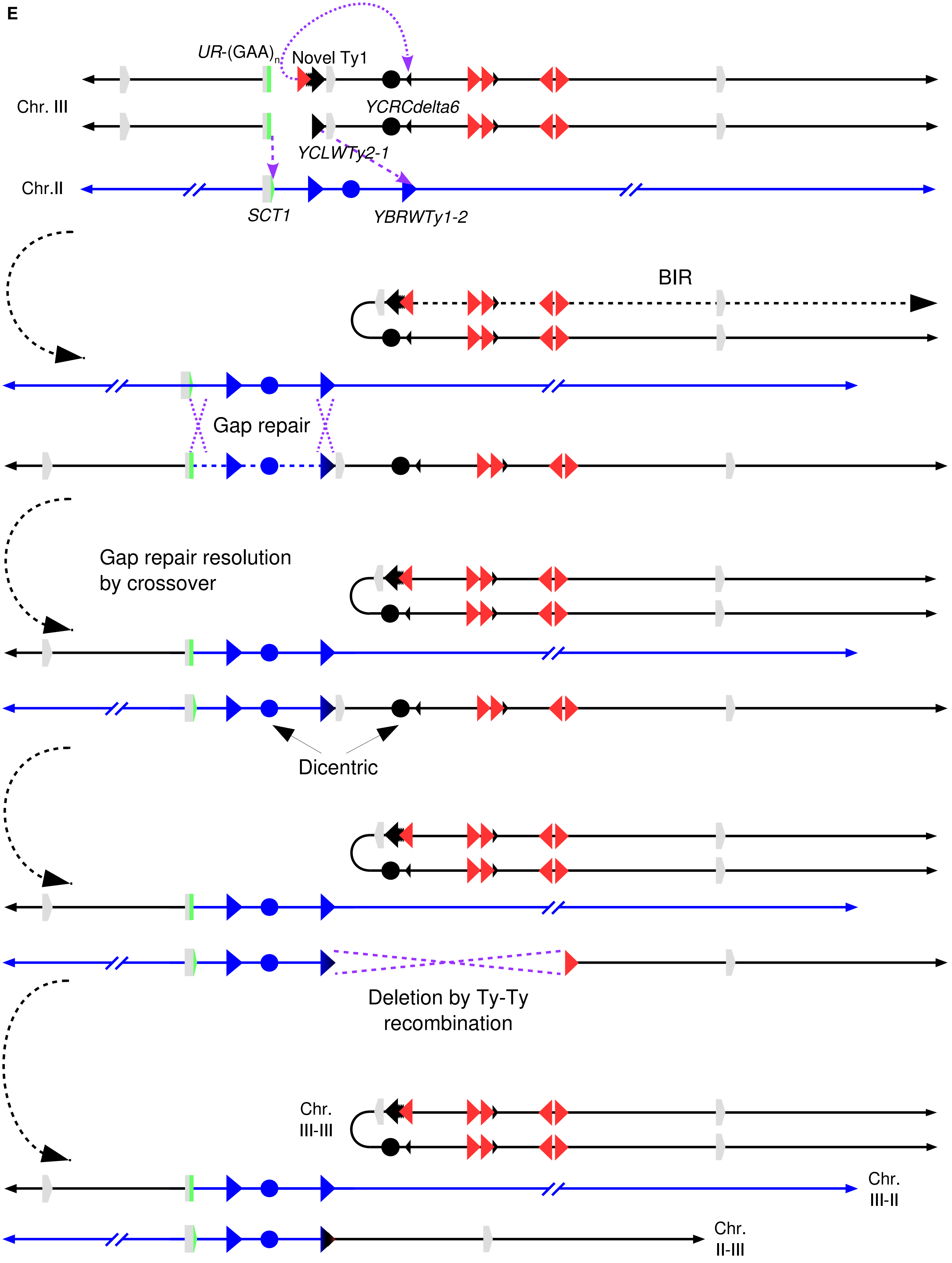

E
UR-(GAA)$_n$
Novel Ty1
Chr. III
YCRCdelta6
YCLWTy2-1
Chr.II
SCT1
YBRWTy1-2
BIR
Gap repair
Gap repair resolution
by crossover
Dicentric
Deletion by Ty-Ty
recombination
Chr.
III-III
Chr.
III-II
Chr.
II-III

Fig. 5. Identifying complex genomic rearrangements in strain #105.

A) CGH microarray analysis, displaying results for Chromosomes II and III. Large green and red regions show regions of the chromosome that are deleted and duplicated, respectively. By examining the hybridization values for individual oligonucleotides on the microarrays, we found that the small green regions depicted in this figure do not represent true deletions. Experiments performed in the lab of Tom Petes. Images contributed by Tom Petes.

B) Nanopore sequencing coverage maps of Chromosomes III and II, generated via UGENE. Positions of relevant sequence features and large-scale copy number changes are indicated above/below the coverage maps. Positions of observed split reads and normal reads at these same junctions are overlayed on the coverage map, and are labeled i-iii.

C) Ribbon single-read view corresponding to the indicated split reads. The x-shaped window in the third panel indicates that this portion of the read maps to an inversion of the chromosome.

D) Single base pair resolution of the indicated junctions. Junction i shows that the break and repair occurred within the $(GAA)_n$ repeats. Junction ii shows that the split read joining *YBRWTy1-2* and the novel Ty1 on the right arm of Chromosome III actually contains sequences matching to *YCLWTy2-1* on the left arm of Chromosome III, suggesting an intermediate step involving a dicentric chromosome (see main text for details). The gray region represents a 46bp window of identity, with SNPs on either side identifying the specific Ty element. (Fig. S6E) The unannotated Ty1 element is the second of two Ty1 elements inserted in place of *YCRWdelta11* (Fig. S3). Junction iii shows the inverted left arm of Chromosome III joining to the beginning of the *YCRCdelta6* LTR on the right arm of Chromosome III.

E) Diagram of the CGR event. Chromosome maps have the same format as in Fig. 1A. Relevant features are labeled. Purple arrows indicate sites of HR invasion. Purple dashed lines indicate sites of Holliday junctions. The top portion displays the broken Chromosome III following duplication and processing to expose ends for HR, as well as the donor Chromosome II. The second portion displays an intermediate step in which two new chromosomes have been formed, one by intrachromosomal BIR and another by gap repair using Chromosome II as a donor, resulting in a dicentric. The third portion shows the previous gap repair resolved as a crossover, resulting in a reciprocal translocation. In the fourth portion, a DSB in the dicentric chromosome is processed to expose homology between the two Ty elements, recombination between which results in deletion of *CEN3*; this recombination event could be a crossover (as shown) or a single-strand annealing event. The bottom portion displays the final chromosome products. See main text for details.

References

Aksenova AY, Greenwell PW, Dominska M, Shishkin AA, Kim JC, Petes TD, Mirkin SM. 2013. Genome rearrangements caused by interstitial telomeric sequences in yeast. *Proc Natl Acad Sci U S A* **110**: 19866-19871.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**: 363-376.

Aten E, White SJ, Kalf ME, Vossen RH, Thygesen HH, Ruivenkamp CA, Kriek M, Breuning MH, den Dunnen JT. 2008. Methods to detect CNVs in the human genome. *Cytogenet Genome Res* **123**: 313-321.

Bacolla A, Tainer JA, Vasquez KM, Cooper DN. 2016. Translocation and deletion breakpoints in cancer genomes are associated with potential non-B DNA-forming sequences. *Nucleic Acids Res* **44**: 5673-5688.

Beck CR, Carvalho CM, Banser L, Gambin T, Stubbolo D, Yuan B, Sperle K, McCahan SM, Henneke M, Seeman P et al. 2015. Complex genomic rearrangements at the PLP1 locus include triplication and quadruplication. *PLoS Genet* **11**: e1005050.

Brock JA, Bloom K. 1994. A chromosome breakage assay to monitor mitotic forces in budding yeast. *J Cell Sci* **107 (Pt 4)**: 891-902.

Carvalho CM, Lupski JR. 2016. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* **17**: 224-238.

Carvalho CM, Ramocki MB, Pehlivan D, Franco LM, Gonzaga-Jauregui C, Fang P, McCall A, Pivnick EK, Hines-Dowell S, Seaver LH et al. 2011. Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nat Genet* **43**: 1074-1081.

Carvalho CM, Zhang F, Liu P, Patel A, Sahoo T, Bacino CA, Shaw C, Peacock S, Pursley A, Tavyev YJ et al. 2009. Complex rearrangements in patients with duplications of MECP2 can occur by fork stalling and template switching. *Hum Mol Genet* **18**: 2188-2203.

Costantino L, Sotiriou SK, Rantala JK, Magin S, Mladenov E, Helleday T, Haber JE, Iliakis G, Kallioniemi OP, Halazonetis TD. 2014. Break-induced replication repair of damaged forks induces genomic duplications in human cells. *Science* **343**: 88-91.

Debladis E, Llauro C, Carpentier MC, Mirouze M, Panaud O. 2017. Detection of active transposable elements in Arabidopsis thaliana using Oxford Nanopore Sequencing technology. *BMC Genomics* **18**: 537.

Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061-1073.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004. Detection of large-scale variation in the human genome. *Nat Genet* **36**: 949-951.

Istace B, Friedrich A, d'Agata L, Faye S, Payen E, Beluche O, Caradec C, Davidas S, Cruaud C, Liti G et al. 2017. de novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *Gigascience* **6**: 1-13.

Jain M, Koren S, Quick J, Rand AG, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S et al. 2017. Nanopore sequencing and assembly of a human genome with ultra-long reads. *BioRxiv*.

Jain M, Olsen HE, Paten B, Akeson M. 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* **17**: 239.

Jansen HJ, Liem M, Jong-Raadsen SA, Dufour S, Weltzien FA, Swinkels W, Koelewijn A, Palstra AP, Pelster B, Spaink HP et al. 2017. Rapid de novo assembly of the European eel genome from nanopore sequencing reads. *Sci Rep* **7**: 7213.

Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bahler J, Sedlazeck FJ. 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun* **8**: 14061.

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**: 56-64.

Kim HM, Narayanan V, Mieczkowski PA, Petes TD, Krasilnikova MM, Mirkin SM, Lobachev KS. 2008. Chromosome fragility at GAA tracts in yeast depends on repeat orientation and requires mismatch repair. *EMBO J* **27**: 2896–2906.

Lee JA, Carvalho CM, Lupski JR. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131**: 1235-1247.

Lee JK, Choi YL, Kwon M, Park PJ. 2016. Mechanisms and Consequences of Cancer Genome Instability: Lessons from Genome Sequencing Studies. *Annu Rev Pathol* **11**: 283-312.

Lemoine FJ, Degtyareva NP, Lobachev K, Petes TD. 2005. Chromosomal translocations in yeast induced by low levels of DNA polymerase a model for chromosome fragile sites.[see comment]. *Cell* **120**: 587-598.

Liu Q, Zhang P, Wang D, Gu W, Wang K. 2017. Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing. *Genome Med* **9**: 65.

Loman NJ, Quick J, Simpson JT. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* **12**: 733-735.

Macintyre G, Ylstra B, Brenton JD. 2016. Sequencing Structural Variants in Cancer for Precision Therapeutics. *Trends Genet* **32**: 530-542.

Mirkin SM. 2007. Expandable DNA repeats and human disease. *Nature* **447**: 932-940.

Nattestad M, Chin C, Schatz MC. Ribbon: Visualizing complex genome alignments and structural variation. *bioRxiv* doi:https://doi.org/10.1101/082123

Neil AJ, Kim JC, Mirkin SM. 2017. Precarious maintenance of simple DNA repeats in eukaryotes. *Bioessays* **39**.

Norris AL, Workman RE, Fan Y, Eshleman JR, Timp W. 2016. Nanopore sequencing detects structural variants in cancer. *Cancer Biol Ther* **17**: 246-253.

Okonechnikov K, Golosova O, Fursov M, team U. 2012. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* **28**: 1166-1167.

Orr HT, Zoghbi HY. 2007. Trinucleotide repeat disorders. *Annual review of neuroscience* **30**: 575-621.

Pandolfo M. 2002. The molecular basis of Friedreich ataxia. *Adv Exp Med Biol* **516**: 99-118.

Pearson CE, Nichol Edamura K, Cleary JD. 2005. Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet* **6**: 729-742.

Polleys EJ, House NCM, Freudenreich CH. 2017. Role of recombination and replication fork restart in repeat instability. *DNA Repair (Amst)* **56**: 156-165.

Polyzos AA, McMurray CT. 2017. Close encounters: Moving along bumps, breaks, and bubbles on expanded trinucleotide tracts. *DNA Repair (Amst)* **56**: 144-155.

Potocki L, Bi W, Treadwell-Deering D, Carvalho CM, Eifert A, Friedman EM, Glaze D, Krull K, Lee JA, Lewis RA et al. 2007. Characterization of Potocki-Lupski syndrome (dup(17)(p11.2p11.2)) and delineation of a dosage-sensitive critical interval that can convey an autism phenotype. *Am J Hum Genet* **80**: 633-649.

Saini N, Zhang Y, Nishida Y, Sheng Z, Choudhury S, Mieczkowski P, Lobachev KS. 2013. Fragile DNA Motifs Trigger Mutagenesis at Distant Chromosomal Loci in Saccharomyces cerevisiae. *PLoS genetics* **9**: e1003551.

Sakofsky CJ, Ayyar S, Deem AK, Chung WH, Ira G, Malkova A. 2015. Translesion Polymerases Drive Microhomology-Mediated Break-Induced Replication Leading to Complex Chromosomal Rearrangements. *Mol Cell* **60**: 860-872.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305**: 525-528.

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz M. 2017. Accurate detection of complex structural variations using single molecule sequencing. *BioRxiv*.

Shah KA, Shishkin AA, Voineagu I, Pavlov YI, Shcherbakova PV, Mirkin SM. 2012. Role of DNA polymerases in repeat-mediated genome instability. *Cell Rep* **2**: 1088-1095.

Shishkin AA, Voineagu I, Matera R, Cherng N, Chernet BT, Krasilnikova MM, Narayanan V, Lobachev KS, Mirkin SM. 2009. Large-scale expansions of Friedreich's ataxia GAA repeats in yeast. *Molecular cell* **35**: 82-92.

Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, Simpson JT, Stebbings LA, Leroy C, Edkins S, Mudie LJ et al. 2009. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**: 1005-1010.

Tang W, Dominska M, Gawel M, Greenwell PW, Petes TD. 2013. Genomic deletions and point mutations induced in Saccharomyces cerevisiae by the trinucleotide repeats (GAA.TTC) associated with Friedreich's ataxia. *DNA Repair* **12**: 10-17.

Usdin K, House NC, Freudenreich CH. 2015. Repeat instability during DNA repair: Insights from model systems. *Crit Rev Biochem Mol Biol* **50**: 142-167.

Vollrath D, Davis RW. 1987. Resolution of DNA molecules greater than 5 megabases by contour-clamped homogeneous electric fields. *Nucleic Acids Res* **15**: 7865-7876.

Wang M, Beck CR, English AC, Meng Q, Buhay C, Han Y, Doddapaneni HV, Yu F, Boerwinkle E, Lupski JR et al. 2015. PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations. *BMC Genomics* **16**: 214.

Yue JX, Li J, Aigrain L, Hallin J, Persson K, Oliver K, Bergstrom A, Coupland P, Warringer J, Lagomarsino MC et al. 2017. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nat Genet* **49**: 913-924.

Zhang F, Khajavi M, Connolly AM, Towne CF, Batish SD, Lupski JR. 2009. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat Genet* **41**: 849-853.

Zhang L, Wang J, Zhang C, Li D, Carvalho CMB, Ji H, Xiao J, Wu Y, Zhou W, Wang H et al. 2017. Efficient CNV breakpoint analysis reveals unexpected structural complexity and correlation of dosage-sensitive genes with clinical severity in genomic disorders. *Hum Mol Genet* **26**: 1927-1941.

Supplemental Material

Nanopore sequencing of complex genomic rearrangements in yeast reveals mechanisms of repeat-mediated double-strand break repair

McGinty, et al.

Contents:

Supplemental Methods

DNA extraction:

The DNA extraction protocol used is a slight modification of a classic ethanol precipitation. First, yeast cultures were grown overnight in 2 ml of complete media (YPD) and refreshed for 4 hours in an additional 8 ml YPD to achieve logarithmic growth. The cultures were spun down, and the pellets were resuspended in 290 µl of solution containing 0.9M Sorbitol and 0.1M EDTA at pH 7.5. 10 µl lyticase enzyme was added, and the mixture was incubated for 30 minutes at 37^{o}C in order to break down the yeast cell wall. The mixture was centrifuged at 8,000 rpm for two minutes, and the pellet was resuspended in 270 µl of solution containing 50mM Tris 20mM EDTA at pH 7.5, and 30 µl of 10% SDS. Following five minutes incubation at room temperature, 150 µl of chilled 5M potassium acetate solution was added. This mixture was incubated for 10 min at 4^{o}C and centrifuged for 10 min at 13,000 rpm. The supernatant was then combined with 900 µl of pure ethanol, which had been chilled on ice. This solution was stored overnight at -20^{o}C, and then centrifuged for 20 minutes at 4,000 rpm in a refrigerated centrifuge. The pellet was then washed twice with 70% ethanol, allowed to dry completely, and then resuspended in 50 µl TE (10mM Tris, 1mM EDTA, pH 8). The resuspended DNA was then treated with 20 µl RNaseA solution, incubated for 10 minutes at 37^{o}C and 30 minutes at room temperature. The above ethanol precipitation was then repeated in order to remove the digested RNA and enzymes. A gel was prepared with 0.4% agarose and 0.5X TBE. A portion of the sample, along with a high-range DNA ladder (GeneRuler High Range DNA Ladder – Thermo Scientific), was run at very low voltage and in a 4ºC cold room overnight. The resulting gel was stained with ethidium bromide and visualized using a BioRad GelDoc XR. This method of DNA preparation resulted in an average fragment size of 24-48 kb. (Fig. S1A) DNA quantity was measured via Qubit (Qubit dsDNA BR Assay kit – Thermo Scientific) and quality assessed via Nanodrop (Thermo Scientific).

Bioinformatics:

Raw current traces generated by ONT sequencing were basecalled via the Albacore basecalling software (ONT version 2.02), which produced barcode-separated FASTQ files. For the parent strain, reads were then aligned to the S288C reference genome (version R64-1-1, obtained from ensembl.org) using NGM-LR (Sedlazeck et al. 2017). This produced a sequence alignment map (SAM) file, which was then processed into a genome-coordinate-sorted BAM file using Samtools (Li et al. 2009). The BAM file was then analyzed by Sniffles (Sedlazeck et al. 2017). All bioinformatics steps were performed on a single Intel i7-based computer, with parameters set to maximize use of all processing cores. Compute times ranged from hours (Albacore) to minutes (NGM-LR, Samtools, Sniffles).

The Sniffles output and the BAM file were then imported to Ribbon (Nattestad et al.) as well as the bioinformatics software UGENE (Okonechnikov et al. 2012), for visualization. To avoid false positives, each structural variant call from Sniffles was examined in both Ribbon and UGENE. To avoid false negatives, regions with readily-apparent copy-number changes seen in UGENE were also closely examined in Ribbon, by typing in specific genome coordinates from which to select the reads. Sharp breakpoints that lined up in multiple reads, with either end mapping to a consistent region, were considered to be the hallmarks of true CGR events.

This analysis of the parent strain identified various deletions and insertions (Figs. 1B & S3), which were then incorporated into the reference genome. Deletion boundaries were identified via the UGENE alignment viewer and the reference sequence trimmed accordingly. Ty element insertions were initially identified via the UGENE alignment viewer. Sequences at the insertion boundaries were extracted and aligned against a database of Ty element sequences (Carr et al. 2012). A closely-related Ty element sequence, obtained from SGD (yeastgenome.org), was inserted into the reference FASTA at the determined boundaries to approximate the insertion. The inserted sequences were then refined by first re-aligning the Nanopore reads to the newly-created FASTA reference, followed by SNP calling in the region using the Samtools ‘mpileup’ command, and then incorporation of high-quality SNPs into the FASTA reference via the Bcftools ‘consensus’ command. This process was repeated several times, until no additional SNPs were detected.

After producing a FASTA reference representative of the parent strain, the above analysis pipeline from NGM-LR to Ribbon was repeated for each CGR-containing strain. Single base pair resolution of breakpoints within Ty elements was determined by analysis of SNPs within each Ty element of origin. Reference sequences of Ty elements involved in CGRs were obtained from SGD (yeastgenome.org) and aligned via the MUSCLE algorithm (Edgar 2004), revealing known SNPs. UGENE alignments were then examined for the presence of expected SNPs. SNPs were distinguished from random sequencing errors by the consistent alignment of SNPs in nearly all of the reads. For junctions involving a copy number change, SNPs were expected to appear in a particular proportion of the reads, and groups of SNPs were expected to consistently appear together in the same individual reads.

The length of $(GAA)_n$ repeats in Nanopore sequencing reads were determined as follows: First, reads were aligned to the reference genome, as described above. Reads aligning to the 5' non-repetitive portion of the *URA3* cassette were then extracted in FASTA format via the UGENE alignment viewer. Note that these FASTA files contained the entire read, rather than just the sequences visible in the alignment viewer. The boundaries of the repeat were then determined by searching the FASTA file for the 5' and 3' surrounding non-repetitive sequences. The total number of base pairs between the 5' and 3' boundaries was then divided by three to approximate the number of $(GAA)_n$ triplets in each read.

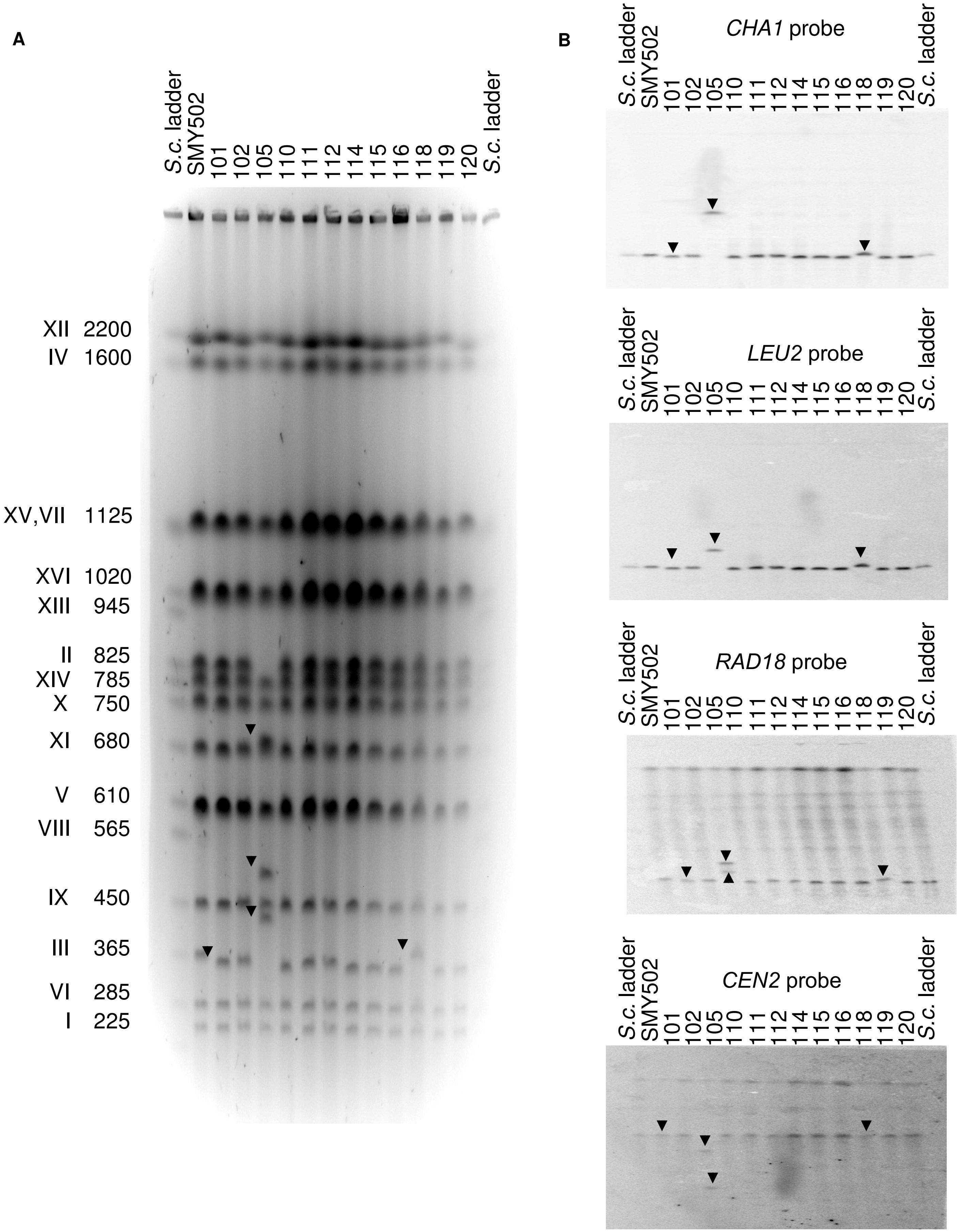

A
S.c. ladder
SMY502
101
102
105
110
111
112
114
115
116
118
119
120
S.c. ladder
XII 2200
IV 1600
XV,VII 1125
XVI 1020
XIII 945
II 825
XIV 785
X 750
XI 680
V 610
VIII 565
IX 450
III 365
VI 285
I 225
B
CHA1 probe
LEU2 probe
RAD18 probe
CEN2 probe

Fig. S1. CHEF gel and Southern analysis of multiple independent strains with $(GAA)_n$-induced chromosome rearrangements. A. EtBr-stained gel. The "*S. c.* ladder" lane contains commercially-available genomic DNA from the strain YY295; the approximate sizes of each chromosome in kb are indicated. SMY502 is the parental strain used in our study, and strains #101-120 are independent Ura$^-$ derivatives. The strains relevant to this analysis are #101, 105, and 118, and the chromosomes relevant to our analysis are shown with arrows. Note that the remaining strains closely resemble strain #101, representing a commonly-observed class of rearrangements. **B.** These portions of the figures represent Southern analyses of the CHEF gels with various hybridization probes. The Southerns probed with *CHA1* and *LEU2* sequences were derived from the gel shown in A.; those probed with *RAD18* and *CEN2* sequences were derived from a different CHEF gel with the same DNA samples. *CHA1* is located on the left arm of Chromosome III, centromere-distal to the location of the $(GAA)_n$ tract. *LEU2* is located on the left arm of Chromosome III, centromere-proximal to the $(GAA)_n$ tract. *RAD18* is located on right arm of Chromosome III. *CEN2* is located at the centromere of Chromosome II.

Experiments performed in the lab of Tom Petes. Images contributed by Tom Petes.

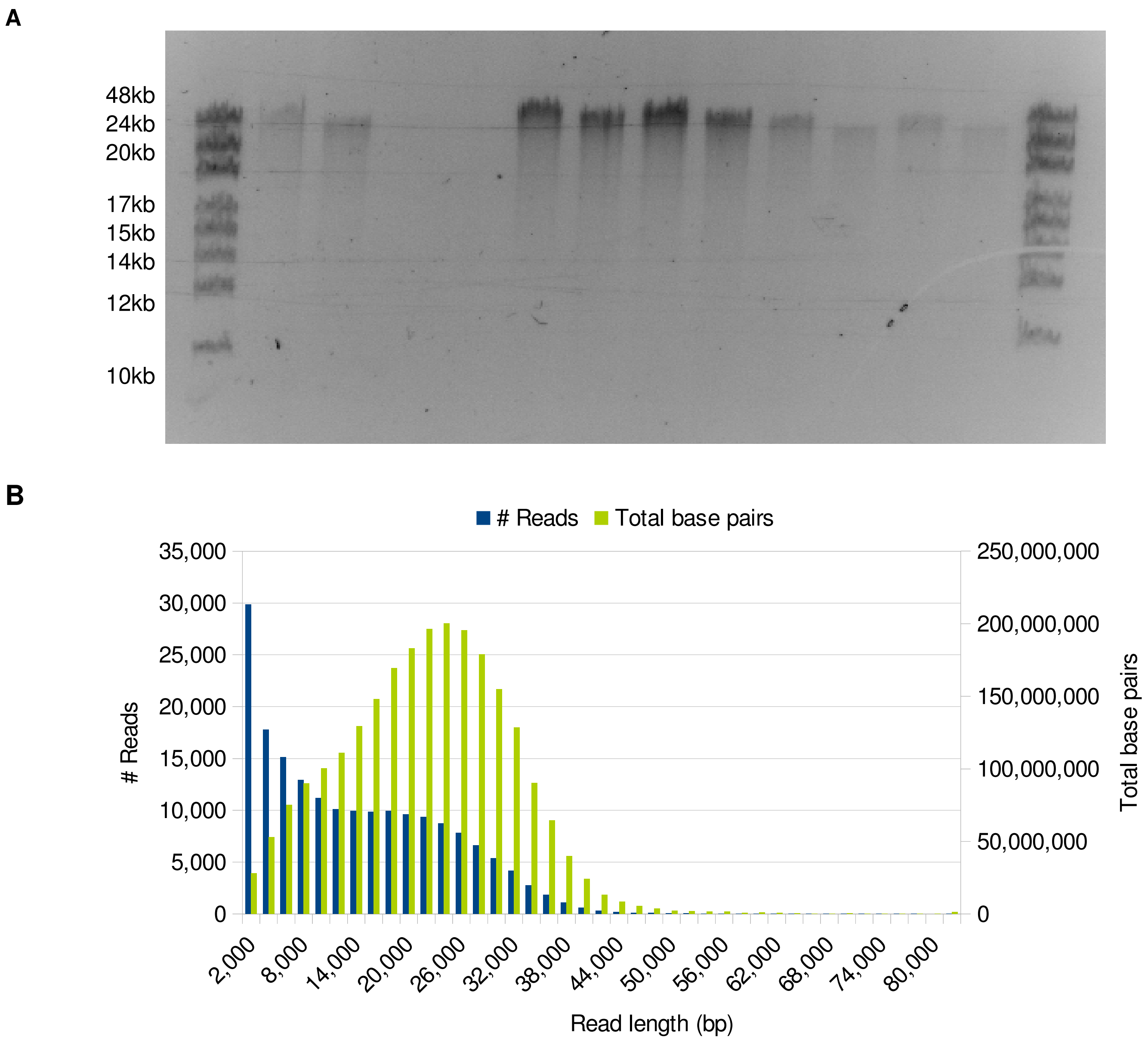
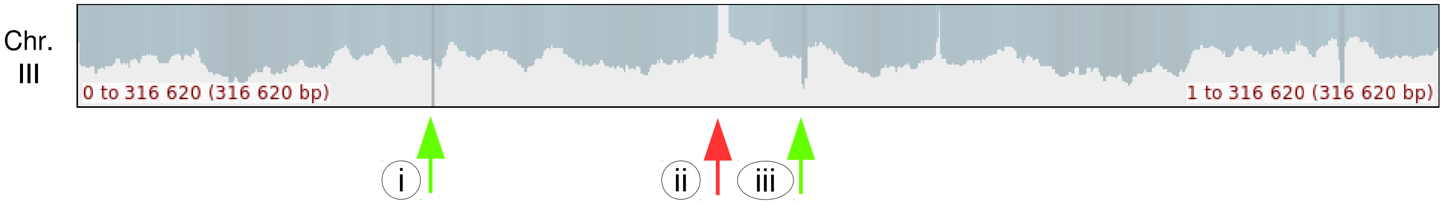
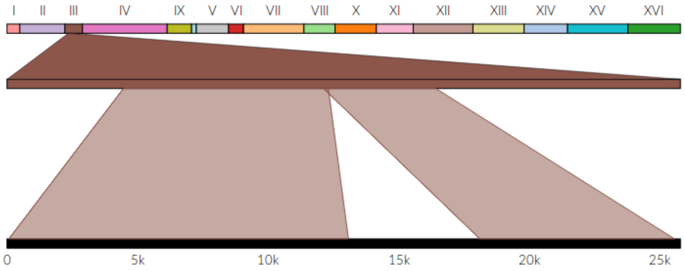
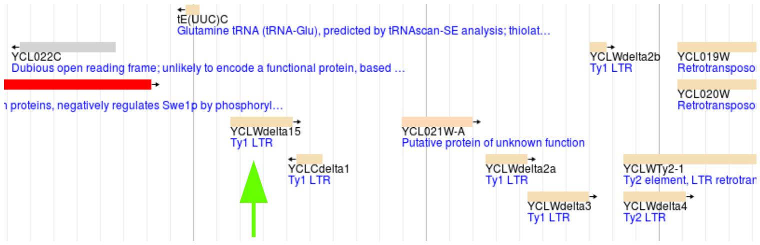


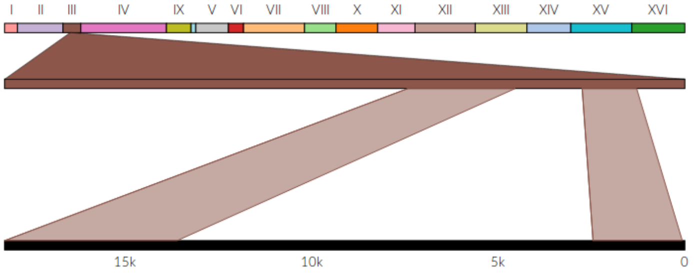
Fig. S2. Read lengths vs. DNA extraction. A. Gel electrophoresis following DNA isolation via ethanol precipitation. The gel contains 0.4% agarose and 0.5X TBE, and was run at low voltage overnight at 4°C. First and last lanes are GeneRuler High Range DNA Ladder (Thermo Scientific). Middle lanes are various DNA samples (not necessarily those that were sequenced). **B.** Read lengths for the MinION sequence output. Blue bars correspond to the number of reads (left Y axis) for each length bin (X axis). Green bars correspond to the total base pair output (right Y axis) for each bin. Note that while shorter reads are common, longer reads contribute more to the overall data set.



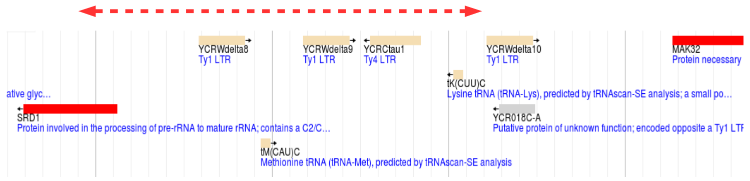


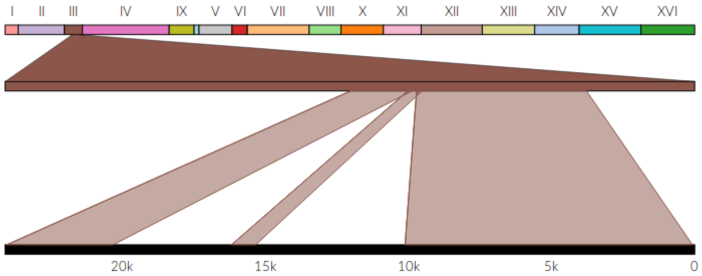
i) Ty element inserted at solo LTR in cluster


tE(UUC)C
Glutamine tRNA (tRNA-Glu), predicted by tRNAscan-SE analysis; thiolat…
YCL022C
Dubious open reading frame; unlikely to encode a functional protein, based …
ı proteins, negatively regulates Swe1p by phosphoryl…
YCLWdelta15
Ty1 LTR
YCLCdelta1
Ty1 LTR
YCL021W-A
Putative protein of unknown function
YCLWdelta2a
Ty1 LTR
YCLWdelta3
Ty1 LTR
YCLWdelta2b
Ty1 LTR
YCL019W
Retrotransposo
YCL020W
Retrotransposo
YCLWTy2-1
Ty2 element, LTR retrotran
YCLWdelta4
Ty2 LTR

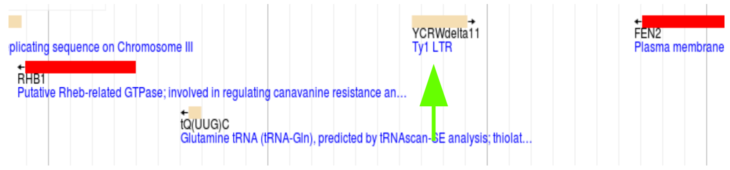


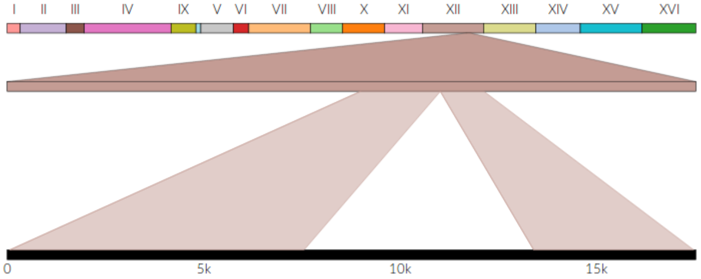
ii) ~10kb Ty insertion at solo LTR cluster replaces surrounding sequence


YCRWdelta8
Ty1 LTR
YCRWdelta9
Ty1 LTR
YCRCtau1
Ty4 LTR
YCRWdelta10
Ty1 LTR
MAK32
Protein necessary
tK(CUU)C
Lysine tRNA (tRNA-Lys), predicted by tRNAscan-SE analysis; a small po…
ative glyc…
SRD1
Protein involved in the processing of pre-rRNA to mature rRNA; contains a C2/C…
YCR018C-A
Putative protein of unknown function; encoded opposite a Ty1 LTI
tM(CAU)C
Methionine tRNA (tRNA-Met), predicted by tRNAscan-SE analysis



iii) 2x Ty insertion at LTR with duplication


plicating sequence on Chromosome III
RHB1
Putative Rheb-related GTPase; involved in regulating canavanine resistance an…
tQ(UUG)C
Glutamine tRNA (tRNA-Gln), predicted by tRNAscan-SE analysis; thiolat…
YCRWdelta11
Ty1 LTR
FEN2
Plasma membrane



iv) Chr. XII – Ty insertion at solo LTR

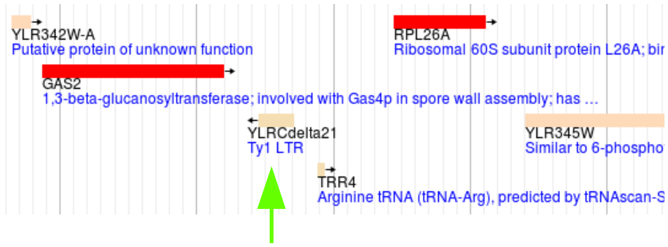

YLR342W-A
Putative protein of unknown function
RPL26A
Ribosomal 60S subunit protein L26A; bir
GAS2
1,3-beta-glucanosyltransferase; involved with Gas4p in spore wall assembly; has …
YLRCdelta21
Ty1 LTR
YLR345W
Similar to 6-phospho
TRR4
Arginine tRNA (tRNA-Arg), predicted by tRNAscan-S

Fig. S3. Identifying novel Ty elements. Top: Nanopore sequencing coverage map of Chromosome III, generated via UGENE, for our starting strain, as aligned to the unaltered S288C reference genome. The chromosome position is represented on the X axis, and the read depth is indicated on the Y axis. Chromosome III contains three out of the four novel Ty elements identified in SMY502 that were not present in S288C. Arrows indicate the position of spikes or gaps in read-depth at the sites of the novel Ty insertions. **Left panels:** Ribbon single-read views highlighting split reads that correspond to each of the indicated regions (i-iii), as well as one additional region on Chromosome XII. Individual reads did not typically identify a particular Ty element as the donor for the insertion, and thus the portion of the read corresponding to the insertion is blank. The error rate of Nanopore sequencing is roughly on par with the divergence of the various Ty elements, some of which are still active and thus nearly identical. All Ty elements begin and end with a long terminal repeat (LTR) sequence of ~340bp, which are generally conserved within each Ty class 1 through 4. The entire Ty element is typically 5-6kb in length. Insertions i and iv are consistent with a single Ty element addition, while insertions ii and iii are consistent with tandem Ty insertions. **Right panels:** SGD Genome Browser views (yeastgenome.org), highlighting the location of each Ty insertion. Green arrows point to the specific LTRs used as insertion points, which were duplicated following the insertion. The red arrow in panel (ii) shows the location that was replaced by the insert, consisting of a cluster of LTR delta elements and some neighboring non-repetitive sequence.

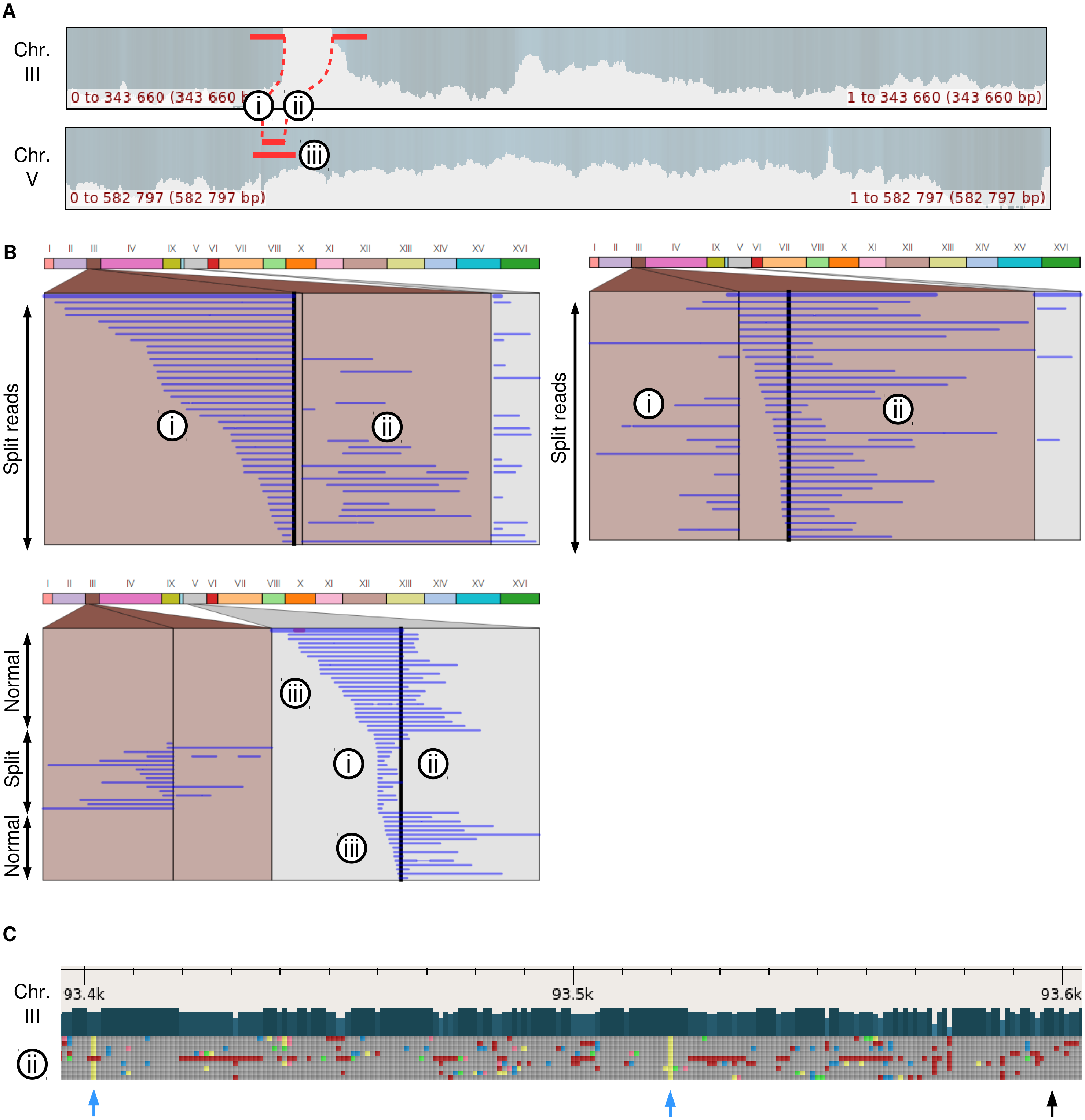

A
Chr. III
0 to 343 660 (343 660 b
1 to 343 660 (343 660 bp)
i
ii
Chr. V
iii
0 to 582 797 (582 797 bp)
1 to 582 797 (582 797 bp)
B
I II III IV IX V VI VII VIII X XI XII XIII XIV XV XVI
Split reads
i
ii
I II III IV IX V VI VII VIII X XI XII XIII XIV XV XVI
Split reads
i
ii
I II III IV IX V VI VII VIII X XI XII XIII XIV XV XVI
Normal
Split
Normal
iii
i
ii
iii
C
Chr. III
93.4k
93.5k
93.6k
ii

Fig. S4. Strain #101: CGR identification. A. Nanopore sequencing coverage maps of Chromosomes III and V, generated via UGENE. Positions of observed split reads and normal reads at these same junctions are overlaid on the coverage map, and are labeled i-iii. **B.** Ribbon multi-read views highlighting reads mapping at each of the labeled junctions. In the multi-read view, the horizontal axis corresponds to genomic locations as indicated by the windows. It is important to note that, unlike the single-read view, the multi-read view does not show how each portion of the reference chromosome fits sequentially into each read. Rather, this view highlights regions found at any point within a read, and stacks multiple reads on the vertical axis. In the top two images, nearly all reads display the same 5' and 3' breakpoints on Chromosome III, indicating a consistent rearrangement. Portions of the reads mapping to Chromosome V at the *ura3-52* locus are seen on the right side. In the bottom panel, many reads are seen that span the *ura3-52* region but do not show a split-read pattern, indicating that the unmodified Chromosome V exists intact. **C.** View of the UGENE alignment zoomed in to the portion of Chromosome III at junction ii. The top portion contains a horizontal scale showing the chromosomal location, along with dark blue vertical bars indicating read depth. Below are individual reads running horizontally, which are stacked vertically. Gray boxes indicate bases that match the reference sequence. Colored boxes indicate bases that do not match the reference sequence (blue=G, green=C, yellow=A, light red=T, dark red=deletion). Blue arrows indicate sites of consistent SNPs which match to *ura3-52* on Chromosome V, rather than the novel Ty1 element adjacent to *YCLWTy2-1* on Chromosome III. The black arrow indicates the location of an expected SNP from *ura3-52* which does not appear on Chromosome III, thus establishing the junction boundaries with single-base pair resolution (Fig. 3D).

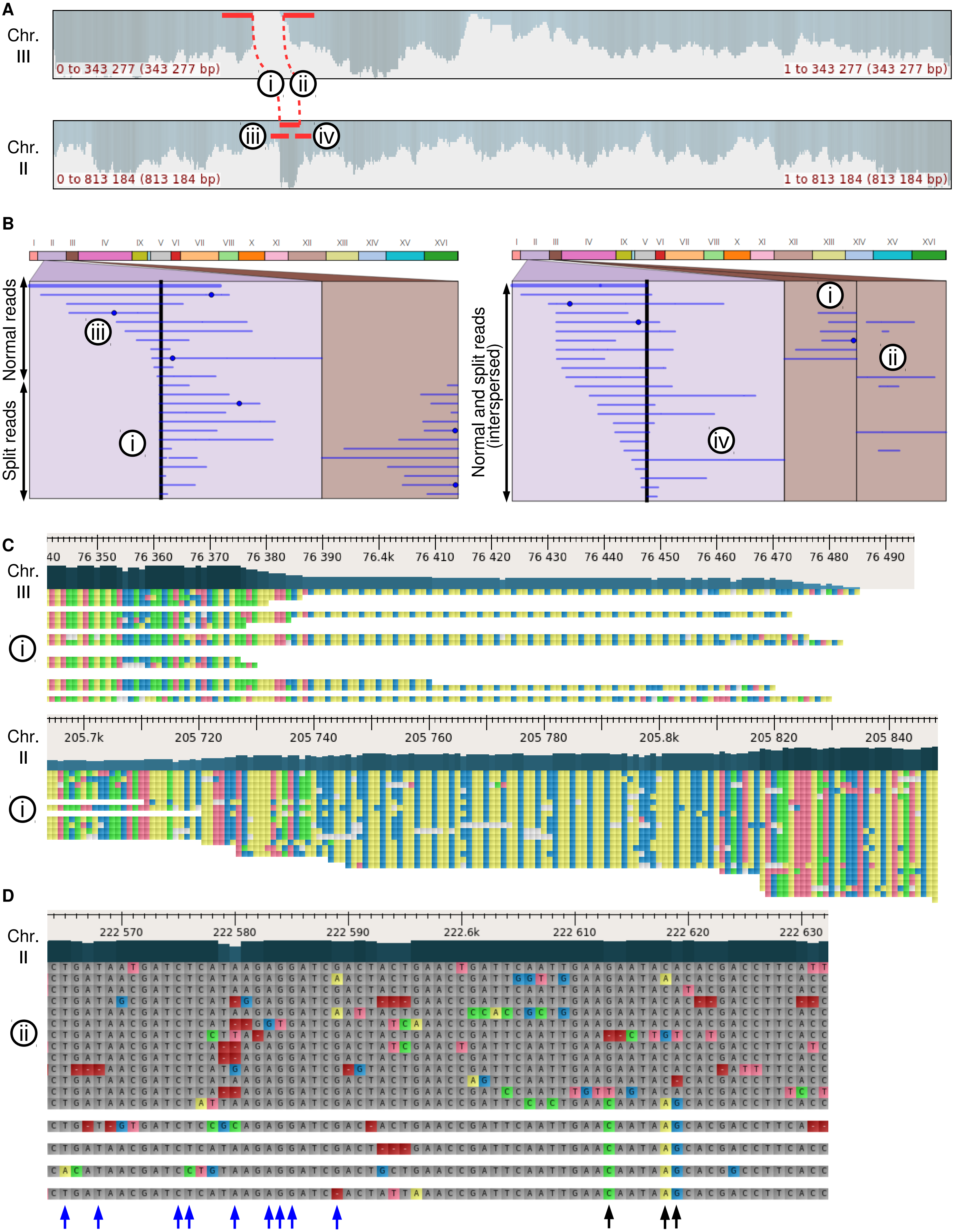

A
Chr. III
0 to 343 277 (343 277 bp)
1 to 343 277 (343 277 bp)
i
ii
iii
iv
Chr. II
0 to 813 184 (813 184 bp)
1 to 813 184 (813 184 bp)
B
I II III IV IX V VI VII VIII X XI XII XIII XIV XV XVI
Normal reads
Split reads
iii
i
Normal and split reads (interspersed)
i
ii
iv
C
40 76 350 76 360 76 370 76 380 76 390 76.4k 76 410 76 420 76 430 76 440 76 450 76 460 76 470 76 480 76 490
Chr. III
i
205.7k 205 720 205 740 205 760 205 780 205.8k 205 820 205 840
Chr. II
i
D
222 570 222 580 222 590 222.6k 222 610 222 620 222 630
Chr. II
ii

Fig. S5. Strain #118: CGR identification. A. Nanopore sequencing coverage maps of Chromosomes III and II, generated via UGENE. Positions of observed split reads and normal reads at these same junctions are overlaid on the coverage map, and are labeled i-iv. **B.** Ribbon multi-read views highlighting reads mapping at each of the labeled junctions. See Fig. S3B for explanation of Ribbon multi-read view. Both split reads and normal reads are observed at each of the junctions, indicating that the unmodified Chromosome II sequence exists intact. **C.** View of the UGENE alignment zoomed in to junction i. See Fig. S3C for explanation of the diagrams. In this particular view, all bases are displayed by color (blue=G, green=C, yellow=A, light red=T, dark red=deletion). Junction i is displayed for chromsomes III and II. The $(GAA)_n$ repeats are clearly visible by the blue-yellow-yellow pattern. Chromosome III shows that the reads stop aligning within the repeats, while Chromosome II shows that the read depth doubles within the repeats. **D.** View of the UGENE alignment zoomed in to the portion of Chromosome II at junction ii. See Fig. S3C for explanation of the diagrams. Black arrows indicate sites of consistent SNPs which match to the novel Ty1 element adjacent to *YCLWTy2-1* on Chromosome III, rather than *YBLWTy1-1* on Chromosome II. The blue arrows indicate the absence of SNPs on *YBLWTy1-1*, indicating that the junction was resolved within this 15bp window.

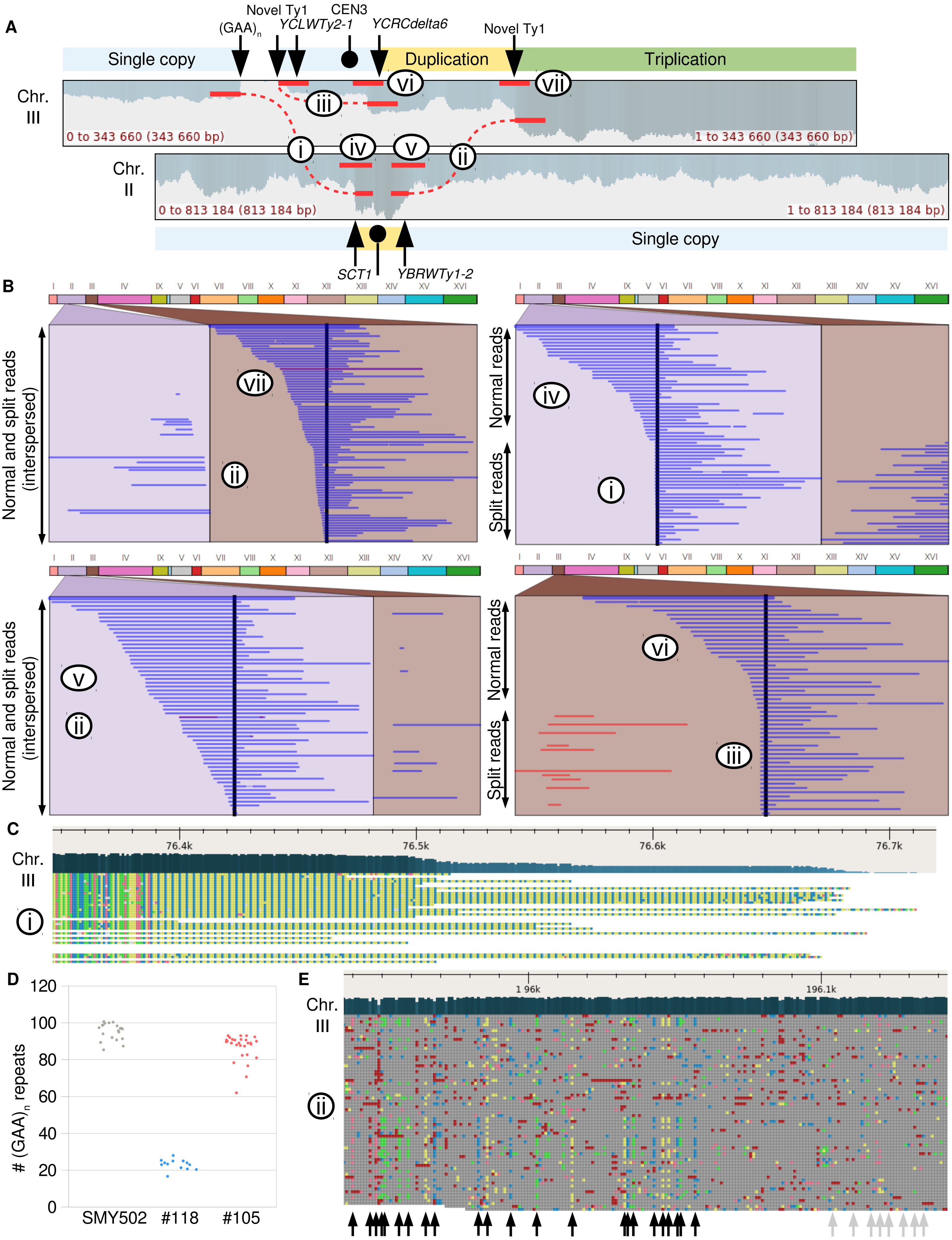

A
Novel Ty1
CEN3
(GAA)$_n$
YCLWTy2-1
YCRCdelta6
Novel Ty1
Single copy
Duplication
Triplication
Chr. III
vi
vii
iii
0 to 343 660 (343 660 bp)
1 to 343 660 (343 660 bp)
i
iv
v
ii
Chr. II
0 to 813 184 (813 184 bp)
1 to 813 184 (813 184 bp)
Single copy
SCT1
YBRWTy1-2
B
I II III IV IX V VI VII VIII X XI XII XIII XIV XV XVI
Normal and split reads (interspersed)
vii
ii
Normal reads
Split reads
iv
i
v
ii
vi
iii
C
76.4k
76.5k
76.6k
76.7k
Chr. III
i
D
120
100
80
60
40
20
0
# (GAA)$_n$ repeats
SMY502
#118
#105
E
1 96k
196.1k
Chr. III
ii

Fig. S6. Identifying genomic rearrangements in strain #105. A. Nanopore sequencing coverage maps of Chromosomes III and II, generated via UGENE. Positions of relevant sequence features and large-scale copy number changes are indicated above/below the coverage maps. Positions of observed split reads and normal reads at these same junctions are overlayed on the coverage map, and are labeled i-vii. **B.** Ribbon multi-read views highlighting reads mapping at each of the labeled junctions. See Fig. S3B for explanation of Ribbon multi-read view. In the lower right image, red lines indicate reads that map in an inverted orientation. Both split reads and normal reads are observed at each of the junctions, indicating that the unmodified Chromosome II sequence exists intact in at least one of the altered chromosomes. However, because no reads span the entire ~50kb duplication on Chromosome II, it could not be determined whether a translocation existed by our sequence data alone. **C.** View of the UGENE alignment zoomed in junction i on Chromosome III. See Fig. S3C for explanation of the diagrams. In this particular view, all bases are displayed by color (blue=G, green=C, yellow=A, light red=T, dark red=deletion). The $(GAA)_n$ repeats are clearly visible by the blue-yellow-yellow pattern. Most reads show that the deletion begins at the very end of the repeat tract. **D.** $(GAA)_n$ repeat length analysis, comparing strains #118 and #105 to the reference strain, which contains 100 GAA repeats. Each dot represents the number of repeats found in an individual Nanopore read. **E.** View of the UGENE alignment zoomed in to the portion of Chromosome III at junction ii. See Fig. S3C for explanation of the diagrams. Black arrows indicate sites of consistent SNPs in approximately 1/3 of the reads (consistent with the triplication junction), which match to *YCLWTy2-1* on Chromosome III, rather than either *YBRWTy2-1* on Chromosome II or the novel Ty1 element replacing *YCRWdelta11* on Chromosome III (which is the reference sequence in this view). Gray arrows indicate the absence of SNPs from *YCLWTy2-1*, establishing the boundaries of the junction.

Supplemental References:

Carr M, Bensasson D, Bergman CM. 2012. Evolutionary genomics of transposable elements in Saccharomyces cerevisiae. *PLoS One* **7**: e50978.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792-1797.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Nattestad M, Chin C, Schatz MC. Ribbon: Visualizing complex genome alignments and structural variation. *bioRxiv* doi:https://doi.org/10.1101/082123

Okonechnikov K, Golosova O, Fursov M, team U. 2012. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* **28**: 1166-1167.

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz M. 2017. Accurate detection of complex structural variations using single molecule sequencing. *BioRxiv*.