A rapidly reversible mutation generates subclonal genetic diversity and unstable drug resistance

Lufeng Dan[a], Yuze Li[a], Shuhua Chen[b], Jingbo Liu[a], Yu Wang[c], Fangting Li[b], Xiangwei He[a,1], and Lucas B. Carey[b,1]

[a]The MOE Key Laboratory of Biosystems Homeostasis & Protection and Innovation Center for Cell Signaling Network, Life Sciences Institute, Zhejiang University, Hangzhou 310058, China; [b]Center for Quantitative Biology, Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China; and[c]State Key Laboratory of Plant Physiology and Biochemistry, China Agricultural University, Beijing 100193, China

Edited by Rodney Rothstein, Columbia University Irving Medical Center, New York, NY, and approved August 11, 2021 (received for review September 16, 2020)

Most genetic changes have negligible reversion rates. As most mutations that confer resistance to an adverse condition (e.g., drug treatment) also confer a growth defect in its absence, it is challenging for cells to genetically adapt to transient environmental changes. Here, we identify a set of rapidly reversible drug-resistance mutations in *Schizosaccharomyces pombe* that are caused by microhomology-mediated tandem duplication (MTD) and reversion back to the wild-type sequence. Using 10,000× coverage whole-genome sequencing, we identify nearly 6,000 subclonal MTDs in a single clonal population and determine, using machine learning, how MTD frequency is encoded in the genome. We find that sequences with the highest-predicted MTD rates tend to generate insertions that maintain the correct reading frame, suggesting that MTD formation has shaped the evolution of coding sequences. Our study reveals a common mechanism of reversible genetic variation that is beneficial for adaptation to environmental fluctuations and facilitates evolutionary divergence.

mutations | genome evolution | sequencing | yeast | drug resistance

Different mechanisms of adaptation have different timescales. Epigenetic changes are often rapid and reversible, while most genetic changes have nearly negligible rates of reversion (1). This poses a challenge for genetic adaptation to transient conditions such as drug treatment; mutations that confer drug resistance are often deleterious in the absence of drug, and the second-site suppressor mutations are required to restore fitness (2, 3). Preexisting tandem repeats (satellite DNA) undergo frequent expansion and contraction (4–6). While repeats are rare inside of most coding sequences and functional elements, there is some evidence for conserved repetitive regions that undergo expansion and contraction to regulate protein functions or expression (6–8). RNA interference– or Chromatin-based epigenetic states have been associated with transient drug resistance in fungi (9) and cancer cells (10, 11), and transient resistant states have been characterized by differences in organelle state, growth rate, and gene expression in budding yeast (12, 13). In bacteria and in fungi, copy-number gain and subsequent loss can result in reversible drug resistance (14–18). However, all genetic systems developed so far for studying unstable genotypes rely on reporter genes and thus investigate only one genetic locus and only one type of genetic change.

Unbiased, next-generation sequencing-based approaches could give a more global view, allowing us to understand the rules that govern unstable genotypes at a genome-wide scale. However, genetic changes with high rates of reversion tend to remain subclonal (19–21), and it is challenging to distinguish most types of low-frequency mutations from sequencing errors (22), especially in complex genomes with large amount of repetitive DNA or de novo duplicated genes. Thus, fast-growing organisms with relatively small and simple genomes are particularly well suited for determining whether transient mutations exist, for the genome-wide characterization of such mutations, and for identification of the underlying mechanisms.

Results

Microhomology-Mediated Tandem Duplications in Specific Genes Caused Reversible Phenotypes in *Schizosaccharomyces pombe*. To discover transient drug-resistance mechanisms in a eukaryote, we performed a genetic screen in the fission yeast *Schizosaccharomyces pombe* for spontaneous mutants that are reversibly resistant to rapamycin plus caffeine (caffeine is required for rapamycin to inhibit growth in *S. pombe*) (23) (Fig. 1*A*). We plated 10^7 cells from each of two independent wild-type strains to YE5S+rapamycin+caffeine plates and obtained 173 drug-resistant colonies, 14 (7%) of which exhibited reversible drug resistance following serial passage in no-drug media (Fig. 1 *B* and *C*). In contrast, resistance for deletion mutants such as *gaf1*Δ (24) is irreversible, suggesting the existence of a type of genetic or epigenetic alteration allowing for reversible drug resistance in the newly isolated strains (Fig. 1 *B* and *C*).

We used genetic linkage mapping and whole-genome sequencing to identify the molecular basis of reversible rapamycin+caffeine resistance. We identified two linkage groups (*SI Appendix*, Fig. S1*A*); we could not identify any common mutations in the first linkage group, suggesting an epigenetic or

Significance

Mutations that confer drug resistance often confer a growth defect in the absence of drug. Mechanisms that enable temporary mutations—mutations that provide drug resistance but frequently revert back to the wild-type genomic DNA sequence—would therefore be advantageous for organisms forced to adapt in changing environments. Here, we show that rapidly reversible mutations are frequently generated by microhomology-mediated tandem duplications (MTDs) in the gene ssp1, causing rapamycin resistance and a growth defect, and reversal back to wild type restores fitness and drug sensitivity. We also found that genomes have evolved to minimize the number of potentially deleterious MTDs and used machine learning to determine the sequence-encoded rules that govern the formation and collapse of MTDs.

Author contributions: L.D. performed research; L.D., Y.L., S.C., and L.B.C. analyzed data; J.L., Y.W., and F.L. contributed new reagents/analytic tools; X.H. and L.B.C. designed research; and X.H. and L.B.C. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: lucas.carey@pku.edu.cn or xhe@zju.edu.cn.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2019060118/-/DCSupplemental.

Published October 21, 2021.

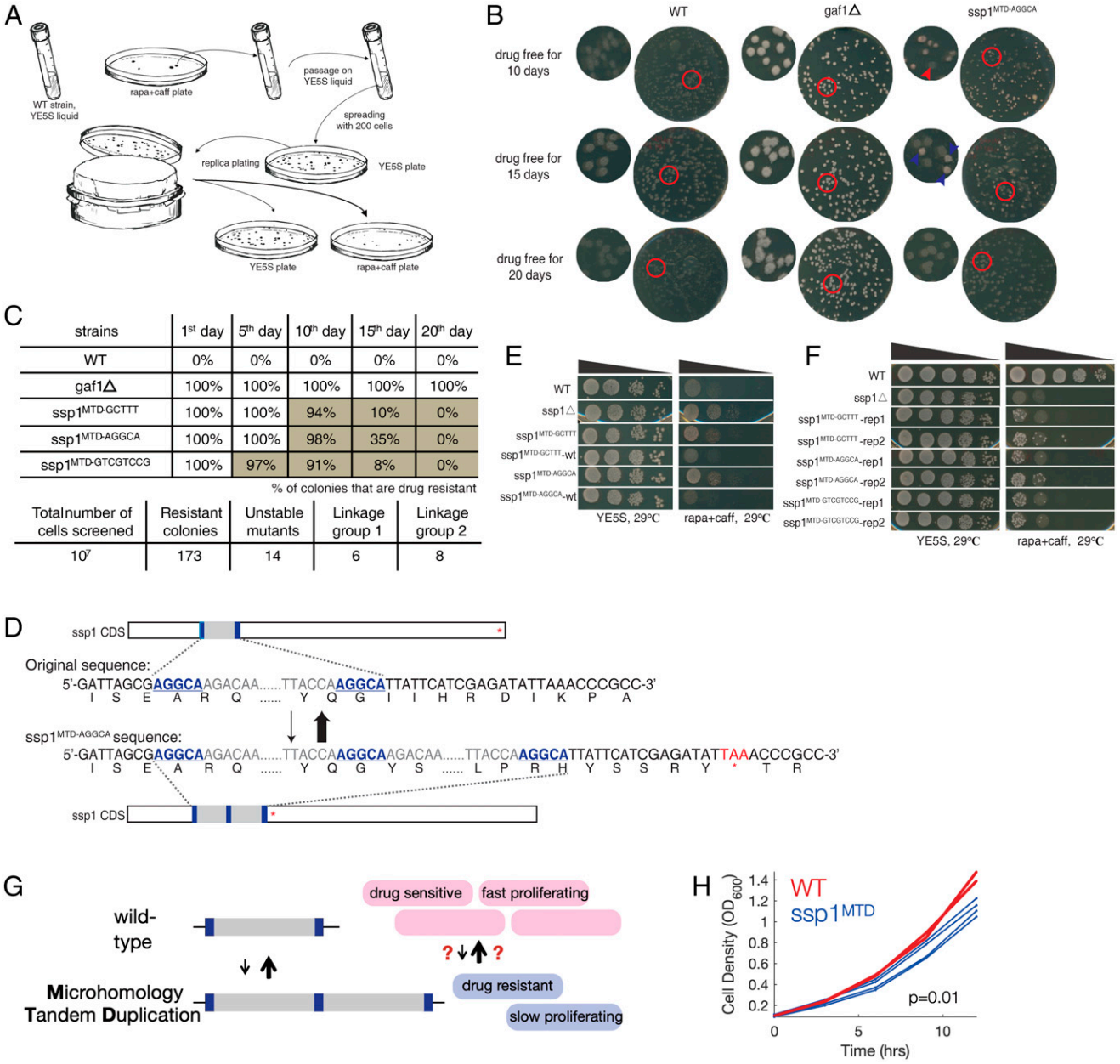

Fig. 1. Screen for mutants with unstable inherited resistance by rapamycin plus caffeine and identify highly reversible mutations in *ssp1*. (*A*) Procedure to screen mutants with unstable rapa+caff resistance using sensitive wild-type (WT) strains in *S. pombe*. (*B*) Unstable phenotype for one of screened mutants on rapa+caff plates after replica plating. *gaf1ρ* as positive control shows strong and stable resistance. The days represent for incubation time on drug free condition, allowing the growth of resistance degenerated progeny. The red arrows point to sensitive progenies, while the blue to resistant ones. (*C*) Dynamics of reversion among identified reversibly drug-resistant colonies. (*D*) Identification of tandem segment duplication in *ssp1* for drug resistance progenies by whole-genome sequencing and reconfirmation by locus-specific PCR/Sanger sequencing. Underlined and bold bases stand for the MHP. The prematured stop codon is marked with red. (*E*) *ssp1* inactivation caused rapamycin resistance and the replacement of ssp1^{MTD} sequence to wt-ssp1 rescue the drug resistance to WT level. (*F*) Heat -resistant isolates are frequently obtained in ssp1^{MTD} strains. (*G*) A cartoon of reversible MTDs that cause drug resistance and a proliferation defect. (*H*) Growth curves of WT (red, two replicates) and ssp1MTD^{AGGCA} (blue, four replicates).

nonnuclear genetic mutation or an inheritable variation that remains to be detected. In contrast, all eight strains in the second linkage group contained tandem duplications in the gene *ssp1*, a Ca^{2+}/calmodulin-dependent protein kinase (human ortholog: CAMKK1/2) which negatively regulates TORC1 signaling, the pathway inhibited by rapamycin, suggesting that mutations in *ssp1* were causal for drug resistance (25).

The *ssp1* linkage group contained three insertion alleles, all of which were tandem duplications of a short DNA segment (55/68/92 base pairs [bp] in length) and had 5 to 8 bp of identical sequence (microhomology pairs, MHPs) at each end (Fig. 1*D* and *SI Appendix*, Fig. S1*B* and Dataset 7). We postulate that these microhomology-mediated tandem duplications (MTDs) (26–28) are important for de novo generation of reversible mutations.

All three MTDs resulted in frameshifts and inactivation of *ssp1*. A similar level of drug resistance was found in the *ssp1*Δ, and replacement of the MTD alleles by transformation with wild-type *ssp1* restored sensitivity (Fig. 1*E*). Sanger sequencing showed that all 16 randomly selected drug-sensitive revertants

of the MTD alleles had the wild-type *ssp1* sequence. Finally, *ssp1*Δ and $ssp1^{MTD}$ strains are temperature sensitive. Spontaneous drug-sensitive nonts revertants were frequently recovered for all the $ssp1^{MTD}$ alleles at a frequency of roughly 1/10,000 cells but not for the *ssp1* deletion (Fig. 1*F*). The frequency of revertants is thus 100× higher than the forward MTD frequency ($8/10^7$), and MTDs in *ssp1* are causal for reversible temperature sensitivity and drug resistance.

Supporting the notion that MTDs may not be specific to rapamycin/caffeine treatment and/or the target gene *ssp1*, in an unrelated genetic screen for suppressors of the slow growth defect of *cnp1-H100M*, a point mutation in the centromere-specific histone gene, we identified MTDs in the transcription repressor genes *yox1* and *lsk1* (*SI Appendix*, Figs. S1*B* and S2 and Dataset 7). These MTDs increase fitness in the *cnp1-H100M* background, and therefore, unlike $ssp1^{MTDs}$, revertants do not increase in abundance in the mutant background. However, in the $ssp1^{wt}$ background, these MTDs are deleterious, and revertants accumulate (*SI Appendix*, Figs. S1 and S2). Thus, MTDs are not gene specific and likely occur throughout the genome.

10,000× Whole-Genome Sequencing Identified Thousands of Subclonal MTDs within a Clonal Population. Based on the scale of the initial genetic screen and assuming drug resistance is not induced by rapamycin, the frequency of cells with any protein-inactivating MTD in *ssp1* in an exponentially growing nonselected wild-type population is estimated $\sim 8 \times 10^{-5}$. This result suggests that a clonal, presumed "isogenic" population contains a wide variety of subclonal MTDs at multiple loci throughout the genome. The frequency of any single MTD will depend on the rate of MTD formation, the rate of reversion, and the fitness cost of the MTD (19–21). The fitness defect imposed by the MTD can be due to altered gene expression or protein function or from the fitness cost of ∼0.025% per kb of additional DNA (29, 30).

To identify the *cis*-encoded determinants of MTD frequency, we developed a computational pipeline for detecting subclonal MTDs in high-coverage Illumina sequencing data (see *Materials and Methods* for details). This method first identifies all MHPs in a DNA segment or genome and generates "signatures" for sequences that would be created by each possible MTD. It then identifies sequencing reads that match these signatures and thus provides experimental support for the existence of a particular MTD within the population (Fig. 2*A*). This method is capable of identifying subclonal MTDs present at very low frequencies in the population.

To determine whether subclonal MTDs captured by sequencing represent the true genetic variation or are technical artifacts (31), we performed two orthogonal tests. In the first, we tested whether MTDs are specific to genomic DNA or also exist in chemically synthesized DNA. We performed 10^5x – 10^6x coverage sequencing of *ssp1* DNA fragments PCR-amplified from genomic DNA from a cloned copy of the gene in a plasmid in *Escherichia coli* or chemically synthesized 150-nt and 500-nt fragments of the gene as well as direct sequencing of chemically synthesized short DNA fragment and plasmid-borne fragment without PCR amplification. We observed far more MTDs in the *pombe* genomic DNA than in the chemically synthesized or plasmid borne controls (Fig. 2*B* and *SI Appendix*, Fig. S3), suggesting that MTDs are largely not caused by PCR or an artifact of Illumina sequencing. It is unclear why the plasmid-borne

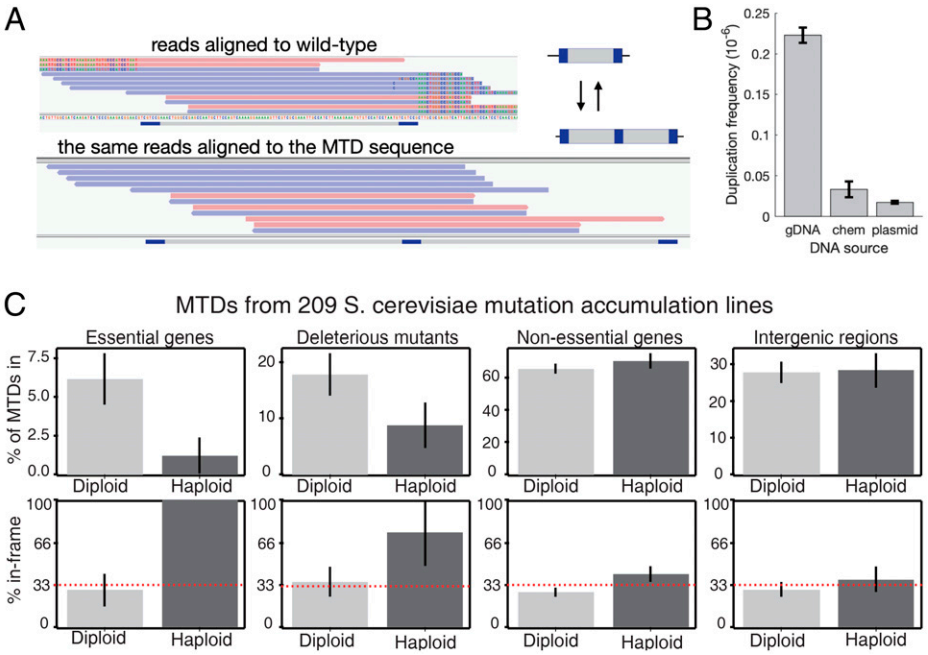

Fig. 2. Identification and verification of subclonal MTDs from ultra-deep sequencing data. (*A*) The computational pipeline finds all sequencing reads whose ends do not match the reference genome and checks whether the reads instead match the sequence that would exist due to an MTD. Shown are reads identified in the pipeline, aligned to either the reference genome (*Top*) or to a synthetic genome with the MTD (*Bottom*). Red and blue mark reads that map to opposite strands are shown. The MHPairs are shown in dark blue, and positions in each read that do not match the reference are colored according to the base in the read. (*B*) The average frequency of sequencing reads that support each MTD in ssp1 from 10^6 coverage sequencing of the gene from *S. pombe*, from a plasmid-borne *ssp1* in *E. coli*, or from a chemically synthesized fragment of the *ssp1* gene. Error bars are SEM across replicates. (*C*) The genomic locations of 314 MTDs that occur in only one single MA line and reanalyzed raw sequencing data from ref. 32; error bars are SD from bootstrapping. The *Top* row shows the percent of MTDs found in essential genes, in genes whose haploid deletion mutant is viable but shows a fitness defect (33), in all nonessential genes, or in intergenic regions, separately for haploid and diploid MA lines. The *Bottom* row shows the percent of MTDs in each category that create a tandem duplication that has a length divisible by three and, for intragenic MTDs, do not disrupt the reading frame. The red dashed line shows the random expectation (one-third of MTDs).

copy of *ssp1* has fewer MTDs than the chemically synthesized DNA, but it raises the possibility that MTDs may be more frequent in eukaryotes (Fig. 2*B* and *SI Appendix*, Figs. S6–S8). We detect MTDs in the *E. coli* genome 1/20th as often as in *S. pombe* and 1/60th as often as in *Saccharomyces cerevisiae* (*SI Appendix*, Fig. S9).

As a second test, we hypothesized that most MTDs in essential genes should be deleterious and recessive. We therefore analyzed raw sequencing data from 209 *S. cerevisiae* haploid and diploid mutation accumulation (MA) lines (32) and identified all MTDs that occur in only one MA line. In haploids, MTDs were depleted in both essential genes and nonessential genes whose deletion causes a fitness defect (Fig. 2*C*). In addition, the MTDs that did occur were more likely to maintain the correct reading frame; the single MTD in an essential gene in a haploid was subclonal, maintained the correct reading frame, and was just 112 bp from the 3′ end of CCT7, a 1,652-bp gene. In contrast, there was no such reading-frame bias in diploids, nonessential genes, or intergenic regions (Fig. 2*C*). Therefore, rare subclonal MTDs identified by ultra-deep sequencing are likely real biological events mostly not experimental artifacts.

To assess the prevalence of MTDs and to identify the sequence-based rules that determine the probability of formation of each tandem duplication, we grew a single diploid fission yeast cell up to ~10^8 cells (25 generations) and performed whole-genome sequencing to an average coverage of 10,000× the diploidy relaxed selection, allowing mutations to accumulate throughout the *S. pombe* genome.

With 10,000× genome sequencing, we identified 5,968 (0.02%) MHPs in which one or more sequencing reads supported an MTD. We observed zero MTDs in most genes, likely due to under-sampling (*SI Appendix*, Fig. S4). However, 20 genes contained more than 10 different MTDs in a single "clonal" population (Fig. 3*A*). To understand this heterogeneity across the genome, we used a logistic regression machine-learning model to predict the probability of duplication at each MHP. MH length, guanine and cytosine (GC) content, inter-MH distance, measured nucleosome occupancy, transcription level, and a local clustering on the scale of 100 nt were able to predict which MHPs give rise to duplications with an area under the curve score of 0.9 with 10-fold cross validation (Fig. 3 *B* and *C* and *SI Appendix*, Fig. S5 and Dataset 5). We note that the peak at 150-nt inter-MH spacing is independent of read length, was not found in *E. coli* or in mitochondrial DNA, and varies between haploid and diploid (*SI Appendix*, Figs. S5–S8). This analysis revealed that properties of MHPs significantly affect the likelihood of MTD formation; for example, and consistent with previous work in *E. coli* (34), long GC-rich MHP is 1,000× more likely to generate a tandem duplication than a short AT-rich one.

While MHPs are spread roughly uniformly throughout the genome (Fig. 3*D*, red), we observed both hot spots, in which MH-mediated generation of tandem duplications are common, and cold spots, in which they are rare (Fig. 3*E*). Local differences in MHP density can only explain some of the hotspots, while our logistic regression model explains the vast majority, suggesting that hotspots with frequent formation of tandem duplications are mostly determined by the local DNA sequence features in addition to microhomologies. The consequence is that duplications are more than 10× more likely to occur in some genes than others, and this variation is correctly predicted by our model (Fig. 3*F*). We detected no MTDs in *ura4*, which has a score of 52, placing it in the bottom third of genes (*SI Appendix*, Fig. S10 and Dataset 4) and providing a possible explanation why MTDs have not been noticed in 5-FOA-based screens of mutations in *ura4* (35). Our results also emphasize that high-coverage sequencing is necessary to identify sufficient numbers of MTDs; 1 billion reads would be required to identify half of the 25 million possible MTDs in the *S. pombe* genome (*SI Appendix*, Fig. S4).

We identified three different subclonal MTDs in the SAGA complex histone acetyltransferase catalytic subunit *gcn5*, placing *gcn5* in the top 5% of genes for both observed and predicted MTDs, suggesting that MTDs in *gcn5* should be found frequently in a genetic screen. Indeed, examination of 16 previously identified (36) suppressors of $htb1^{G52D}$ identified MTDs in *gcn5*, as well as in *ubp8*, in which we also observed an MTD in our high-coverage sequencing data (*SI Appendix*, Fig. S1*B*). These results suggest that MTDs arise in most genes at a high-enough frequency within populations in order to be the raw material on which natural selection acts.

Replication Slippage Modulates the Rate of MTD Reversion at *ssp1*. Having established that local *cis*-encoded features determine the frequency with which tandem duplications arise from MHPs, we next sought to identify the *trans*-genes that affect the MTD process. $ssp1^{MTD}$ alleles fail to grow at 36 °C, and their reversion back to wild type suppresses the temperature sensitivity, providing a way to measure the effects of mutations on reversion frequency. We screened a panel of 364 strains with mutations in DNA replication, repair, recombination, or chromatin organization genes for mutants that affect the rate of $ssp1^{MTD}$ reversion back to wild type (Dataset 6) and found three mutants that significantly increased and eight that significantly decreased the frequency of $ssp1^{WT}$ revertants (Fig. 4 *A*–*C*).

Replication fork collapse is a major source of double-stranded breaks (DSBs), and the ensuing homologous recombination (HR)–related restarting process is error prone and is known to generate microhomology-flanked insertions and deletions via replication slippage, a process in which, when replication resumes for a stalled or collapsed fork, the unwound nascent strand may anneal with a homologous segment on the template, either at the vicinity (37, 38) or at a distance (34) of the paused site, with ensuing replication on noncontinuous template. Inactivation of Rad50, Rad52, or Ctp1 results in decreased replication slippage and decreased MTD reversion (37, 39) (Fig. 4 *A*–*C*). Deletions of *mhf1* and *mhf2*, two subunits of the FANCM–MHF complex, which is involved in the stabilization and remodeling of blocked replication forks, also decreased the frequency of MTD revertants. It is therefore likely that replication slippage during HR-mediated fork recovery causes reversion of MTDs in these mutants and could be one contributing factor in wild-type background.

Replication stresses activate a checkpoint that promotes DNA repair and recovery of stalled or collapsed replication forks and delays entry into mitosis (40, 41). The inactivation of replication checkpoint kinase *cds1* or its regulator *mrc1* may thus result in a failure to restore the replication fork, causing increased genome instability and MTD reversion. The replication checkpoint would thus be required for the stability of MTDs. Consistently, we found that deletion of the DNA damage checkpoint kinase *cds1* or its regulator *mrc1* increased the frequency of $ssp1^{WT}$ revertants. Deletion of the single-stranded DNA binding A (RPA) subunit *ssb3* (RPA3/RFA3) or the multifunctional 5′-flap endonuclease *rad2* also increased the frequency of revertants (Fig. 4*C*).

Many genes identified in the screen are multifunctional and play roles in both replication and repair. We therefore performed quantitative epistasis analysis to determine the relation between six of the identified genes and the Mediator of the Replication Checkpoint, *mrc1*, which interacts with and stabilizes Pol2 at stalled replication forks. In addition to the checkpoint activator *cds1*, deletion of *rad2* had no effect in an *mrc1*Δ background, suggesting that all three of these genes act in the same pathway (Fig. 4*D*). In contrast, deletion of *ssb3* increased the frequency of revertants in both wild-type and *mrc1*Δ

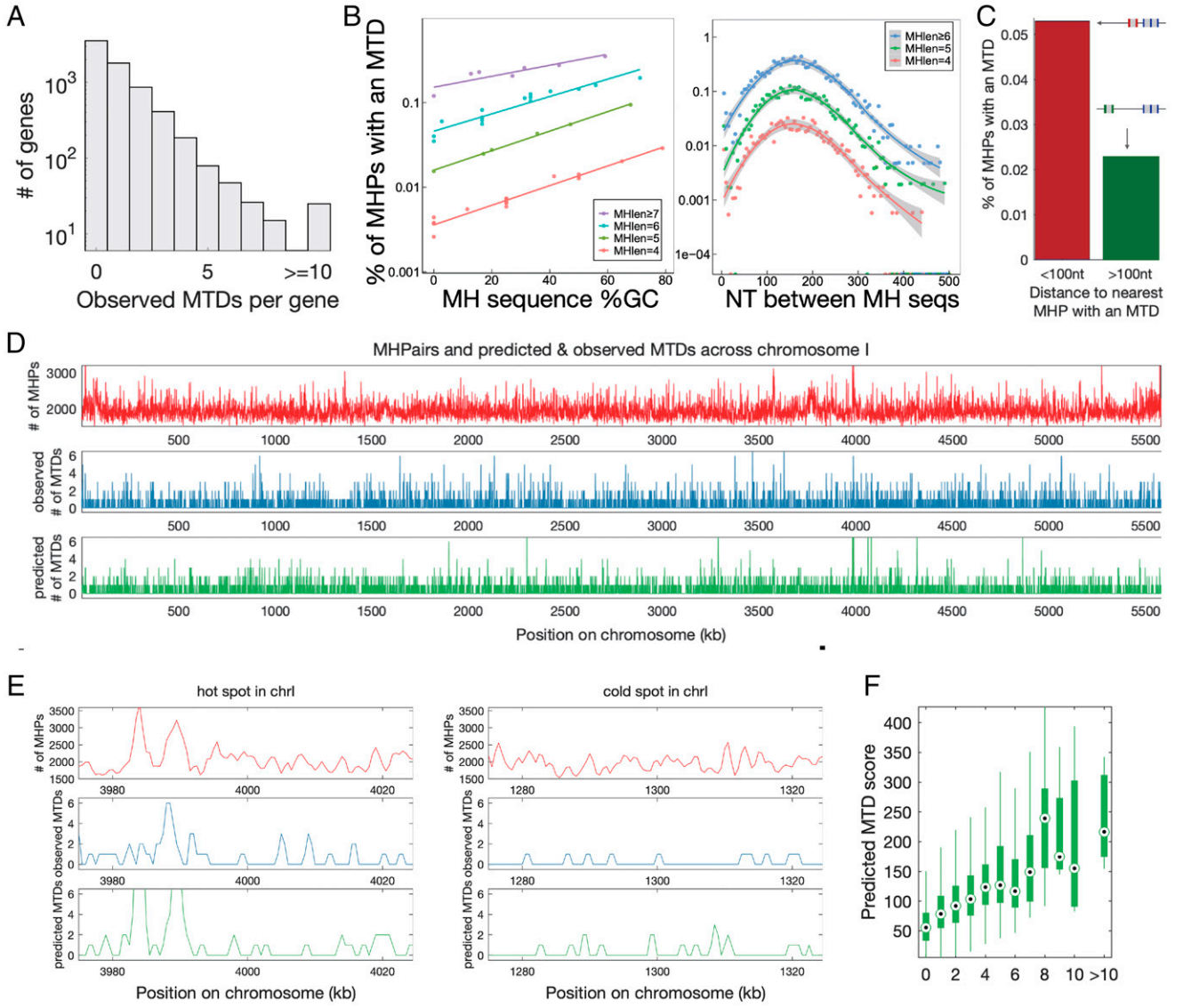

Fig. 3. Identification of the *cis*-determinants of MTD through ultra-deep sequencing and identification of subclonal duplications. (*A*) A histogram of the number of MTDs found in each gene from 10,000× whole-genome sequencing. (*B*) The 25 million MHPs in the genome were binned in groups of 10,000 with the same MH sequence length and similar GC content (*Left*) or inter-MHPair distance (*Right*) and the percentage of MHPs in each group with an observed MTD was calculated. A logistic regression model was trained with 10-fold cross-validation to predict the probability of observing an MTD at each MHPair. (*C*) The distance from each MHP to the nearest MHP with an MTD was calculated, and the percentage of MHPs with an MTD was calculated for MHPs less than (red) or farther than (green) 100 nt from the closest MHP. (*D*) For each 1-kb window in the genome, shown are the number of MHPairs (red), the number of observed MTDs (blue), and the predicted number of MTDs from the logistic regression model (green). (*E*) An example cold spot (0.2MTDs/kb) and hot spot (0.7 MTDs/kb) in chromosome I. The cold spot has fewer MTDs after taking into account the number of MHPs, (Fisher's exact test, $P = 2.76 \times 10^{-9}$, odds ratio = 3.843). (*F*) The sum of scores from the logistic regression model for each MHP in each gene, with the genes grouped by the observed number of MTDs in the 10,000× coverage data.

backgrounds, and deletion of *pds5* or *rik1* decreased the frequency of revertants in both wild-type and *mrc1*Δ backgrounds, though not to the extent expected for genetic independence, suggesting partial epistasis. In contrast, the effects of *rad50* deletion were completely independent of *mrc1* (Fig. 4*D*).

While the observed numbers of MTDs in ultra-deep sequencing experiments are a function of both duplication and reversion rates and all of the above genes may play a role in both processes, the above results suggested that due to increased reversion rates, the number and frequency of MTDs would be reduced in *cds1*Δ and *rad2*Δ strains. To test this idea, we performed 10^6x coverage sequencing of the hotspot gene *SPCC1235.01*. We observe MTDs at fewer MHPs and an overall decrease in the number of MTDs in both mutants (Fig. 4 *E* and *F*).

Half of Insertions and Tandem Duplications in Natural Isolates Are MH Mediated. It was baffling that MTDs are prevalent within populations and that the first theoretical proposal for microhomology-mediated processes in the generation of tandem duplications is 20 y old (5); yet, relatively little is known about the forward process and even less about the reversion, suggesting that these events are not often encountered or identified as such. To better understand the dynamics of MTDs within a population, we used a simple model of neutral mutations within a growing population that takes into account both forward and reverse mutation rates and began with 100% of individuals as wild type (see *Materials and Methods*). The mutant frequency always increases and over short timescales (Fig. 5 *A*, *Left*), increasing the reverse rate from being equal to

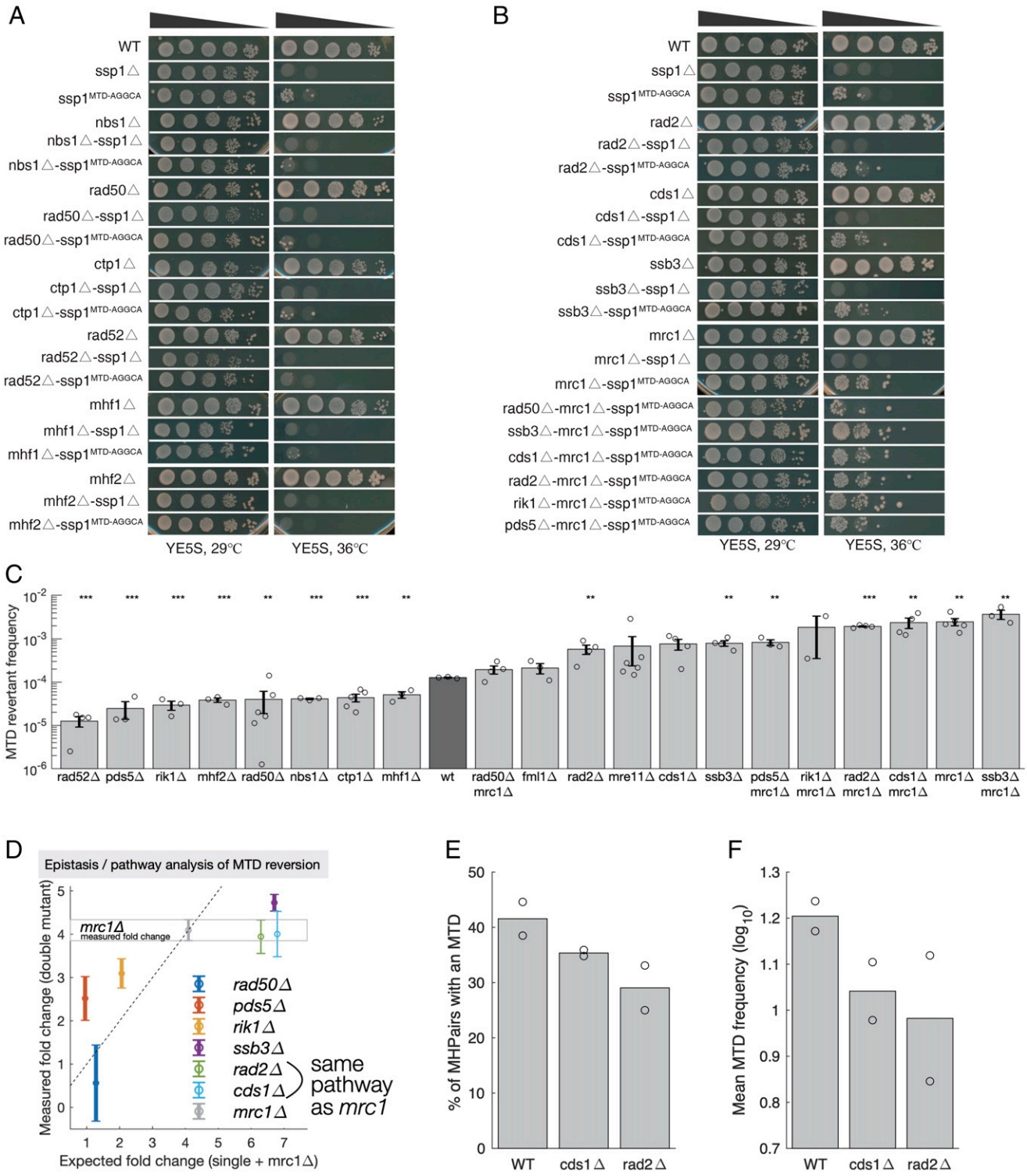

Fig. 4. A genetic screen to identify the regulators of MTD reversion. (*A* and *B*) Surveyed mutants showed reduced *ssp1*^{MTD} reversion frequency represented by TS recovery phenotype. The non-TS phenotype of single mutation and *ssp1ρ* alone or combined with other mutants retained severe temperature-sensitive phenotype at 36 °C should be established. The number of TS revertants under 36 °C indicate the reversion frequency of ssp1^{MTD}. The initial gradient for spotting assay was 10^5 cells and diluted with 10-fold gradient (cell number: 10^5, 10^4, 10^3, 10^2, and 10^1). (*C*). Quantification of *ssp1*^{MTD} reversion frequency in mutants ($n \geq 3$ biological repeats, error bars are s.e.m., ***$P < 0.001$, **$P < 0.01$, and *$P < 0.05$ *t* test compared to wt). (*D*) Two colonies of WT and two of each mutant were picked and *SPCC1235.01* amplified by PCR and sequenced to 10^6 coverage. Show is the average across the two replicates of the MTD frequency at each of the 3,002 MHPs. (*E*) The percentage of MHPairs with one or more reads in support of an MTD in *SPCC1235.01*. (*F*) For all MHPairs with an MTD, the frequency of reads supporting that MTD per 10^6 reads that map to that MHPair.

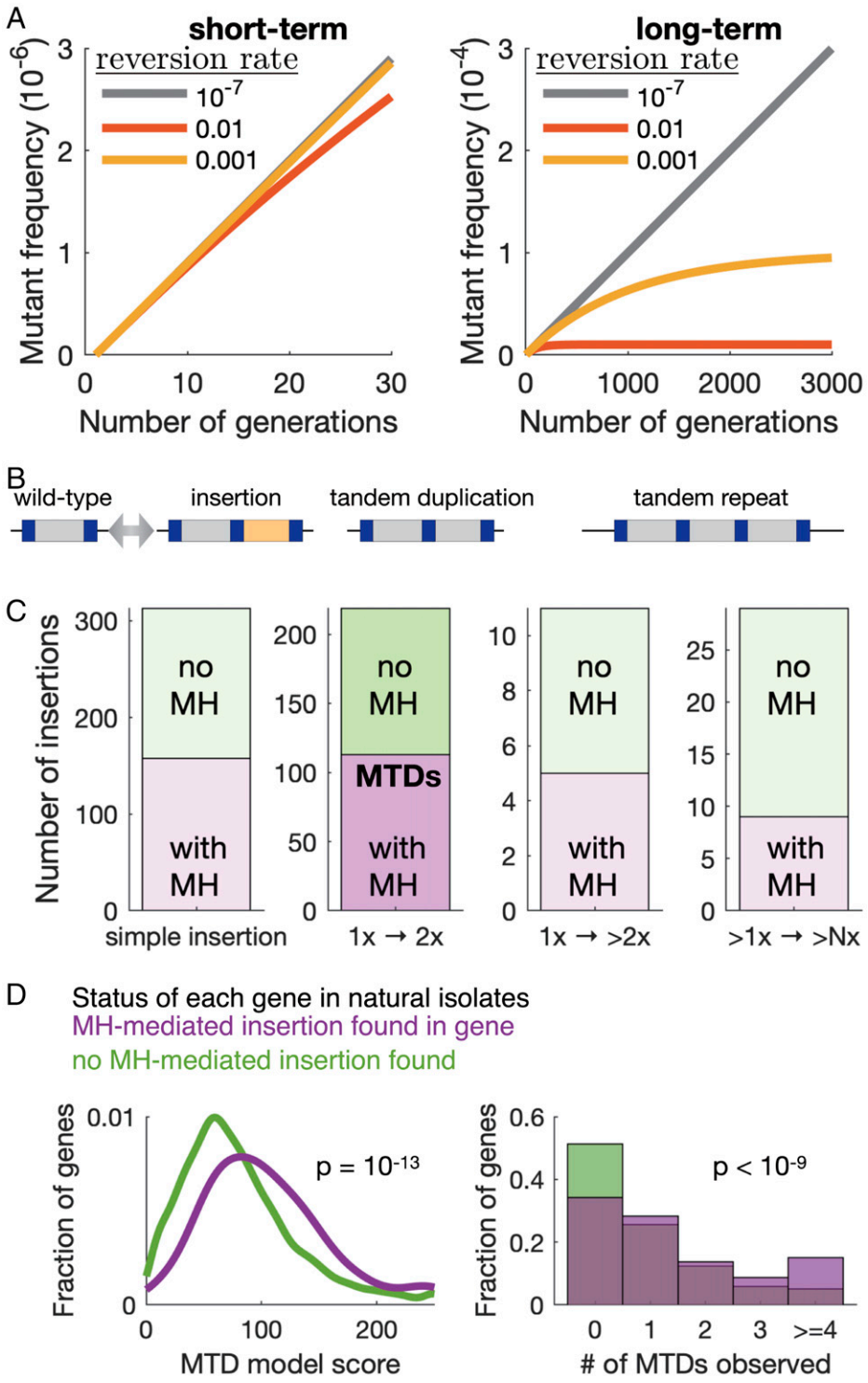

Fig. 5. MTDs remain subclonal because of high reversion rates, yet half of insertions and de novo tandem duplications in natural populations arise at microhomology sequence pairs. (*A*) Simulations showing the frequency of a neutral mutation (forward mutation rate = 10^{-7}) within a growing population at three different reversion rates (colors). *Left* and *Right* show the same simulates at different timescales, with the effect of reversion only apparent at long timescales. (*B*) A cartoon showing three possible types of microhomlogy-mediated insertions: simple insertion, tandem duplication, and higher copy repeat. (*C*) Quantification of all insertions of at least 10 bp fixed in any of the 57 natural *S. pombe* isolates that represent most of the genetic diversity within the species, relative to the reference genome. Insertions were classified according to the presence (purple) or absence (green) of exact MHPs on either side of the insert and to the type of insert. There are 113 MTDs in wild *pombe* strains (second column). The right-most column (>1× -> >N×) refers to the expansion of repeats present in the reference genome. (*D*) Distributions of the predicted MTD score from the logistic regression model (*Left*) and the number of experimentally observed subclonal MTDs (*Right*) for genes with one or more microhomology-mediated insertions (purple) or for genes with no MH-mediated insertions (green) in any of the natural isolates. *P* values are from a Mann–Whitney *U* test.

the forward mutation rate (gray) to being 10,000 times higher (yellow) has little effect.

Over longer timescales, high reversion rates cause the mutant frequency to plateau and remain subclonal (Fig. 5 *A*, *Right*), reducing the fraction of neutral MTDs within a population. However, despite the high reversion rate, both drift and selection enable fixation of MTDs within a population. To identify fixed microhomology-mediated insertions, we searched the genome sequences of 57 wild *S. pombe* isolates (42) and found that 50% of insertions larger than 10 bp involve microhomology repeats (Fig. 5 *B* and *C*). Among these were 158 microhomology-mediated insertions that did not contain an obvious duplication and 113 MTDs with an MTD.

To test whether the propensity of MTD formation within the laboratory strain is predictive of extant sequence variation observed in natural isolates, we tested whether the MTD score for each gene predicts the likelihood of microhomology-mediated insertions in that gene. We found that genes with microhomology-mediated insertions in natural isolates tend to have higher predicted MTD scores and more experimentally observed MTDs (Fig. 5*D*), suggesting that the local features that affect MTD formation in the laboratory also shape evolution in nature.

MHPs with Longer MH Sequences Are More Likely to Generate MTDs that Maintain the Correct Reading Frame. We found that MHPs with longer MH sequences are more likely to form MTDs. If the high propensity to generate MTDs has shaped the *S. pombe* genome, any signature of selection should be stronger at MHPs with longer MH sequences and should also be stronger in essential genes versus nonessential genes. We therefore divided the 25 million MHPs with an MH length of 4 to 25 nt and an inter-MH distance of 3 to 500 nt into those fully contained within in intergenic regions or fully contained within essential or nonessential genes and split them by MH sequence length. Specifically in coding sequences, MHPs at which an MTD would not disrupt the reading frame are more common than expected by chance, and this enrichment is higher in essential genes and at longer MH sequences (Fig. 6). At the same time, MHPs within genes are more common that expected by chance (*SI Appendix*, Fig. S12). Therefore, natural selection has acted to decrease the number of MHPs that would create potentially deleterious MTDs, and this selection is weaker for MHPs that would create an in-frame MTD.

Taken together, our results demonstrate that MTDs occur frequently and broadly throughout the genome within a clonal population. These findings indicate that high levels of subclonal genetic divergence are prevalent but are under detected using conventional sequencing approaches that tend to disfavor the detection of low-abundance subclonal variants. As many MTDs create large insertions, they are more likely to be deleterious. Nonetheless, MTDs provide plasticity to the genome and its functionality, for example, by allowing cells to become drug resistant, while allowing the resistant cell lineage to revert back to wild type and regain high fitness once the drug is removed. Selection can act on this genetic diversity for its reversibility or by using the tandem duplications as the initial step for the generation of higher copy number repeats, which are evolutionarily fixed in extant genomes and traditionally regarded as a major source of genome divergence. While previous work has shown that preexisting repeats undergo rapidly reversible changes, the sequence-encoded rules regulating the birth and death of such sequences were less studied (34). This work reveals that numerous sites throughout the genome have the potential of evolving into such repetitive elements. Furthermore, MH sequence length–dependent depletion of frame-shifting MHPs in essential genes shows that natural selection has shaped the genome to avoid MHPs that would frequently generate deleterious MTDs. Finally, much in the same way as repetitive DNA may have been positively selected for as a regulatory element to maintain reversible genetic diversity (7, 8), the large number of MHPs that would result in in-frame MTDs raises the possibility that some genes may maintain MHPs to generation functional genetic diversity, creating a dynamic protein-coding genome.

Discussion

Why haven't MTDs been identified more frequently in genetic screens and MA assays? There are several possible reasons. Mutation callers are ineffective in detecting long insertions from short reads: on simulated data, both Mutect2 and HaplotypeCaller often fail to detect tandem duplications longer than 85 bp. Also, MTDs are often identified as insertions but not specifically as MTDs (*SI Appendix*, Figs. S1*B* and S11), suggesting the need for computational tools for identifying MTDs. *URA4* and *URA3* have relatively few MHPs (*SI Appendix*, Fig. S10). In many 5-FOA–based mutation spectra papers, only substitutions were analyzed in detail, as indels do not occur with high-enough frequency to generate good statistics (43). Due to the high rate of reversion, it is likely that tandem duplications in *URA4* would yield URA+ colonies when restruck onto -URA plates; this high reversion rate may lead to MTD-

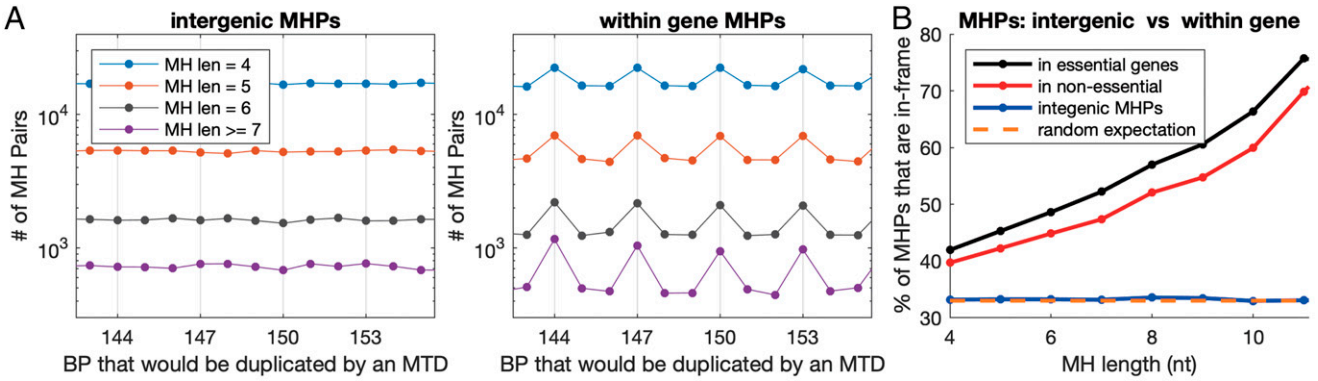

Fig. 6. Frame-shifting MHPs are depleted from essential genes in an MH sequence length–dependent manner. (*A*) The number of MHPs in the *S. pombe* genome with different MH sequence lengths (colors) for which an MTD would generate varying insert sizes (*x*-axis). *X*-axis grid lines mark MTDs with insertion sizes divisible by three. *Left* shows MHPs that are intergenic and *Right* MHPs that are fully contained with a coding sequence of a gene. (*B*) The percentage of MHPs with lengths evenly divisible by three (*y*-axis) for each MH sequence length (*x*-axis) that are found in intergenic regions (blue), fully contained within essential genes (black) or within nonessential genes (red). Random expectation is that one-third of MHPs will have an insert size evenly divisible by three (orange).

containing colonies being discarded in many different types of genetic screens. In our reanalysis of *S. cerevisiae* MA lines, almost all de novo MTDs were subclonal (*SI Appendix*, Fig. S11*B*). With new computational tools for identifying MTDs plus third-generation sequencing platforms with improved ability to detect long indels, it is likely that MTDs will be implicated in more phenotypes.

Unbiased genome-scale approaches have been very informative for the mechanisms that generate point mutations in both wild-type and mutant cells (44). It is clear that multiple molecular mechanisms can give rise to tandem duplications in microhomology-dependent and independent manners, and the mechanisms may differ in mutants and between species. In plant mitochondria genomes, longer MH sequences are associated with longer tandem duplications, suggesting that microhomology is involved in the generation of tandem duplications, likely via microhomology-mediated repairing of DSBs (45, 46) or slippage strand replication (47). In contrast, tandem duplications in the rice nuclear genome tend to have no or shorter microhomology sequences, suggesting that in the rice nuclear genome, tandem duplications likely form via patch-mediated DSB creation followed by NHEJ (48). In *E. coli*, the lagging-strand processing activity of Pol I is required for stress-induced, MH-mediated amplification of 7- to 32-kb segments (34). However, simple models cannot account for all features observed across studies, and it is clear that multiple mechanisms play a role (49–51).

All current methods of measuring microhomology-mediated duplications and deletions impose artificial length scales, whereas genetic screens require the entire gene or a specific region be duplicated. Thus, while microhomologies are associated with deletions on the 500-bp to 1-kb range in *E. coli* (52) and with unstable amplifications of 7 to 37 kb (16) and selection for increased gene expression often enriches for ∼12-bp MH-mediated amplification of ∼10kb (53), these length scales are determined by the locations of MH sequences in the particular genomic region required to be duplicated in the genetic screen and the size of the region required to be duplicated or amplified.

Genome-wide sequencing-based approaches are less biased but still not bias free. We limit the length-scale to 500 bp and set lower and upper bounds of 3 bp and 25 bp for the MH sequences. While the MTD frequency relative to the number of 1-bp or 2-bp MH sequence pairs in the genome is likely to be low, the number of 2-bp and 1-bp MH sequences is high. Preferential flanking of tandem duplications by 2-bp and even 1-bp sequence identities have been reported (27, 28, 45, 54), suggesting that subclonal MTDs generated by short MH sequences may be common. While it is computationally intractable to apply the directed "signatures" method presented here to search MH pairs separated by more than 500 bp, this method could be extended to >500 bp if the search is limited to longer MH sequences, which are rare but drive high rates of MTD formation (28, 34). Similarly, the method could be extended to broken microhomologies (55) but with an even higher computational cost. Computational approaches using third-generation (Nanopore and PacBio) sequencing have the potential to provide a truly unbiased measure of duplication and deletion frequencies as well as answer questions about how often amplifications are extrachromosomal versus intrachromosomal (17, 56) and tandem versus inverted amplifications (57). Verification of these algorithms will take some work, as Nanopore and Pac-Bio sometimes give different results when sequencing tandem repeats (58).

Because different molecular mechanisms have different sequence requirements, replication strand biases, and length scales (34, 50), genome-wide unbiased methods are necessary to understand the relative contribution of each mechanism to MTD formation and collapse. As they can, in theory, measure events across any distance, from single base pair to interchromosomal, and at an unlimited number of different loci, with variation in chromatin contexts, transcription, and other genome-architecture features, ultra-deep sequencing is likely the best way to quantitatively understand the various biological mechanisms that contribute to the dynamic genome.

GENETICS

Materials and Methods

Details of the materials and methods, including cell growth, mutants screen, genetic linkage test, MTD reversion regulators survey, find MHPs, ultra-deep next-generation sequencing, and logistic regression are presented in *SI Appendix, Materials and Methods*.

Data Availability. All processed data and code are available in GitHub at https://github.com/carey-lab/MicroHomologyMediatedTandemDuplications and raw sequencing data at NCBI SRA BioProject accession no. PRJNA631756.

ACKNOWLEDGMENTS. We thank Lilin Du, Aaron New, and Wenfeng Qian for insightful discussions and for comments on the manuscript. We thank Qi Zhou for assistance with rapamycin and caffeine resistance screen. L.B.C. was supported by the Peking-Tsinghua Center for Life Sciences. X.H. was supported by National 973 Plan for Basic Research Grant 2015CB910602 and National Natural Science Foundation of China Grants 31628012, 31671396, 31871253, and 31801131.

1. O. J. Rando, K. J. Verstrepen, Timescales of genetic and epigenetic inheritance. *Cell* **128**, 655–668 (2007).
2. D. I. Andersson, B. R. Levin, The biological cost of antibiotic resistance. *Curr. Opin. Microbiol.* **2**, 489–493 (1999).
3. R. E. Lenski, Bacterial evolution and the cost of antibiotic resistance. *Int. Microbiol.* **1**, 265–270 (1998).
4. R. Gemayel, M. D. Vinces, M. Legendre, K. J. Verstrepen, Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.* **44**, 445–477 (2010).
5. J. E. Haber, E. J. Louis, Minisatellite origins in yeast and humans. *Genomics* **48**, 132–135 (1998).
6. K. J. Verstrepen, A. Jansen, F. Lewitter, G. R. Fink, Intragenic tandem repeats generate functional variability. *Nat. Genet.* **37**, 986–990 (2005).
7. E. R. Moxon, P. B. Rainey, M. A. Nowak, R. E. Lenski, Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol.* **4**, 24–33 (1994).
8. R. Moxon, C. Bayliss, D. Hood, Bacterial contingency loci: The role of simple sequence DNA repeats in bacterial adaptation. *Annu. Rev. Genet.* **40**, 307–333 (2006).
9. S. Calo *et al.*, Antifungal drug resistance evoked via RNAi-dependent epimutations. *Nature* **513**, 555–558 (2014).
10. S. M. Shaffer *et al.*, Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546**, 431–435 (2017).
11. S. V. Sharma *et al.*, A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell* **141**, 69–80 (2010).
12. R. Dhar, A. M. Missarova, B. Lehner, L. B. Carey, Single cell functional genomics reveals the importance of mitochondria in cell-to-cell phenotypic variation. *eLife* **8**, e38904 (2019).
13. S. F. Levy, N. Ziv, M. L. Siegal, Bet hedging in yeast by heterogeneous, age-correlated expression of a stress protectant. *PLoS Biol.* **10**, e1001325 (2012).
14. H. Nicoloff, K. Hjort, B. R. Levin, D. I. Andersson, The high prevalence of antibiotic heteroresistance in pathogenic bacteria is mainly caused by gene amplification. *Nat. Microbiol.* **4**, 504–514 (2019).
15. R. T. Todd, A. Selmecki, Expandable and reversible copy number amplification drives rapid adaptation to antifungal drugs. *eLife* **9**, e58349 (2020).
16. T. D. Tlsty, A. M. Albertini, J. H. Miller, Gene amplification in the lac region of E. coli. *Cell* **37**, 217–224 (1984).
17. T. Huang, J. L. Campbell, Amplification of a circular episome carrying an inverted repeat of the *DFR1* locus and adjacent autonomously replicating sequence element of *Saccharomyces cerevisiae*. *J. Biol. Chem.* **270**, 9607–9614 (1995).
18. P. J. Hastings, H. J. Bull, J. R. Klump, S. M. Rosenberg, Adaptive amplification: An inducible chromosomal instability mechanism. *Cell* **103**, 723–731 (2000).
19. D. L. Hartl, E. W. Jones, *Genetics: Principles and Analysis* (Jones and Bartlett Publishers, ed. 4, 1998).

20. R. Lande, Risk of population extinction from fixation of deleterious and reverse mutations. *Genetica* **102-103**, 21–27 (1998).
21. T. Maruyama, M. Kimura, A note on the speed of gene frequency changes in reverse directions in a finite population. *Evolution* **28**, 161–163 (1974).
22. L. B. Carey, RNA polymerase errors cause splicing defects and can be regulated by differential expression of RNA polymerase subunits. *eLife* **4**, e09945 (2015).
23. R. Weisman, M. Choder, Y. Koltin, Rapamycin specifically interferes with the developmental response of fission yeast to starvation. *J. Bacteriol.* **179**, 6325–6334 (1997).
24. D. Laor, A. Cohen, M. Kupiec, R. Weisman, TORC1 regulates developmental responses to nitrogen stress via regulation of the GATA transcription factor Gaf1. *MBio* **6**, e00959 (2015).
25. E. Davie, G. M. A. Forte, J. Petersen, Nitrogen regulates AMPK to control TORC1 signaling. *Curr. Biol.* **25**, 445–454 (2015).
26. A. R. J. Lawson *et al.*, RAF gene fusion breakpoints in pediatric brain tumors are characterized by significant enrichment of sequence microhomology. *Genome Res.* **21**, 505–514 (2011).
27. L. E. L. M. Vissers *et al.*, Rare pathogenic microdeletions and tandem duplications are microhomology-mediated and stimulated by local genomic architecture. *Hum. Mol. Genet.* **18**, 3579–3593 (2009).
28. N. A. Willis *et al.*, Mechanism of tandem duplication formation in BRCA1-mutant cells. *Nature* **551**, 590–595 (2017).
29. E. Harrison, V. Koufopanou, A. Burt, R. C. MacLean, The cost of copy number in a selfish genetic element: The 2-μm plasmid of Saccharomyces cerevisiae. *J. Evol. Biol.* **25**, 2348–2356 (2012).
30. E. M. Torres *et al.*, Effects of aneuploidy on cellular physiology and cell division in haploid yeast. *Science* **317**, 916–924 (2007).
31. S. R. Head *et al.*, Library construction for next-generation sequencing: Overviews and challenges. *Biotechniques* **56**, 61–77 (2014).
32. N. P. Sharp, L. Sandell, C. G. James, S. P. Otto, The genome-wide rate and spectrum of spontaneous mutations differ between haploid and diploid yeast. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E5046–E5055 (2018).
33. A. Baryshnikova *et al.*, Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nat. Methods* **7**, 1017–1024 (2010).
34. A. Slack, P. C. Thornton, D. B. Magner, S. M. Rosenberg, P. J. Hastings, On the mechanism of gene amplification induced under stress in Escherichia coli. *PLoS Genet.* **2**, e48 (2006).
35. S. Gangloff *et al.*, Quiescence unveils a novel mutational force in fission yeast. *eLife* **6**, e27469 (2017).
36. X. Xu, L. Wang, M. Yanagida, Whole-genome sequencing of suppressor dna mixtures identifies pathways that compensate for chromosome segregation defects in *Schizosaccharomyces pombe*. *G3 (Bethesda)* **8**, 1031–1038 (2018).
37. I. Iraqui *et al.*, Recovery of arrested replication forks by homologous recombination is error-prone. *PLoS Genet.* **8**, e1002976 (2012).
38. S. T. Lovett, Encoded errors: Mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Mol. Microbiol.* **52**, 1243–1253 (2004).
39. A. Teixeira-Silva *et al.*, The end-joining factor Ku acts in the end-resection of double strand break-free arrested replication forks. *Nat. Commun.* **8**, 1982 (2017).
40. A. A. Alcasabas *et al.*, Mrc1 transduces signals of DNA replication stress to activate Rad53. *Nat. Cell Biol.* **3**, 958–965 (2001).
41. K. Myung, R. D. Kolodner, Suppression of genome instability by redundant S-phase checkpoint pathways in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 4500–4507 (2002).
42. D. C. Jeffares *et al.*, The genomic and phenotypic diversity of Schizosaccharomyces pombe. *Nat. Genet.* **47**, 235–241 (2015).
43. C. Duan *et al.*, Reduced intrinsic DNA curvature leads to increased mutation rate. *Genome Biol.* **19**, 132 (2018).
44. F. Supek, B. Lehner, Scales and mechanisms of somatic mutation rate variation across the human genome. *DNA Repair (Amst.)* **81**, 102647 (2019).
45. D. Ottaviani, M. LeCain, D. Sheer, The role of microhomology in genomic structural variation. *Trends Genet.* **30**, 85–94 (2014).
46. M. McVey, S. E. Lee, MMEJ repair of double-strand breaks (director's cut): Deleted sequences and alternative endings. *Trends Genet.* **24**, 529–538 (2008).
47. E. Darmon, D. R. F. Leach, Bacterial genome instability. *Microbiol. Mol. Biol. Rev.* **78**, 1–39 (2014).
48. J. N. Vaughn, J. L. Bennetzen, Natural insertions in rice commonly form tandem duplications indicative of patch-mediated double-strand break induction and repair. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 6684–6689 (2014).
49. M. Bzymek, S. T. Lovett, Instability of repetitive DNA sequences: The role of replication in multiple mechanisms. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 8319–8325 (2001).
50. P. J. Hastings, G. Ira, J. R. Lupski, A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.* **5**, e1000327 (2009).
51. D. X. Tishkoff, N. Filosi, G. M. Gaida, R. D. Kolodner, A novel mutation avoidance mechanism dependent on S. cerevisiae RAD27 is distinct from DNA mismatch repair. *Cell* **88**, 253–263 (1997).
52. A. M. Albertini, M. Hofer, M. P. Calos, J. H. Miller, On the formation of spontaneous deletions: The importance of short sequence homologies in the generation of large deletions. *Cell* **29**, 319–328 (1982).
53. T. Edlund, S. Normark, Recombination between short DNA homologies causes tandem duplication. *Nature* **292**, 269–271 (1981).
54. H. Xia, W. Zhao, Y. Shi, X.-R. Wang, B. Wang, Microhomologies are associated with tandem duplications and structural variation in plant mitochondrial genomes. *Genome Biol. Evol.* **12**, 1965–1974 (2020).
55. S. K. Whoriskey, V. H. Nghiem, P. M. Leong, J. M. Masson, J. H. Miller, Genetic rearrangements and gene amplification in Escherichia coli: DNA sequences at the junctures of amplified gene fusions. *Genes Dev.* **1**, 227–237 (1987).
56. R. J. Kaufman, P. C. Brown, R. T. Schimke, Amplified dihydrofolate reductase genes in unstably methotrexate-resistant cells are associated with double minute chromosomes. *Proc. Natl. Acad. Sci. U.S.A.* **76**, 5669–5673 (1979).
57. C. Payen *et al.*, The dynamics of diverse segmental amplifications in populations of Saccharomyces cerevisiae adapting to strong selection. *G3 (Bethesda)* **4**, 399–409 (2014).
58. S. Mitsuhashi *et al.*, Tandem-genotypes: Robust detection of tandem repeat expansions from long DNA reads. *Genome Biol.* **20**, 58 (2019).