Rapid & reversable mutations generate subclonal genetic diversity

Lufeng Dan[1], Yuze Li[1], Shuhua Chen[2], Jingbo Liu[1], Fangting Li[2], Yu Wang[3], Xiangwei He[1]*, Lucas B. Carey[2]*

Affiliations:

[1] The Life Sciences Institute and Innovation Center for Cell Signaling Network, Zhejiang University, Hangzhou, Zhejiang 310058, China.

[2] Center for Quantitative Biology and Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China.

[3] State Key Laboratory of Plant Physiology and Biochemistry, China Agricultural University, Beijing, China

*Correspondence to: L.B.C. lucas.carey@pku.edu.cn and X.H. xhe@zju.edu.cn

Abstract:

Most genetic changes have negligible reversion rates. As most mutations that confer resistance to an adversary 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 *S. pombe* that are caused by Microhomology mediated Tandem Duplication (MTD), and reversion back to the wild-type sequence. Using 10,000x coverage whole-genome sequencing we identify over 6000 subclonal MTDs in single a clonal population, and use machine learning to determine how MTD frequency is encoded in DNA. We find that sequences with the highest predicted MTDs rates tend to generate insertions that maintain the correct reading frame; MTD formation has shaped the evolution of coding sequences. Our study reveals a mechanism of reversible genetic variation that is beneficial for adaptation to environmental fluctuations and facilitates evolutionary divergence.

Main Text:

Different mechanisms of adaptation have different timescales. Epigenetic changes are often rapid and reversible, while most genetic changes have nearly negligible rates of reversion(Rando and Verstrepen, 2007). 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(Andersson and Levin, 1999; Lenski, 1998). Pre-existing tandem repeats (satellite DNA) undergo frequent expansion and contraction (Gemayel et al., 2010; Haber and Louis, 1998; Verstrepen et al., 2005), but are rare inside of coding sequences and other functional elements. Chromatin-based epigenetic states have been associated with transient drug resistance in cancer cells(Shaffer et al., 2017; Sharma et al., 2010), and transiently resistant states have been characterized by differences in organelle state, growth rate, and gene expression in budding yeast(Dhar et al., 2019; Levy et al., 2012). In bacteria, copy-number gain and subsequent loss can result in transient antibiotic resistance(Nicoloff et al., 2019). However, no similar transient genetic resistance mechanisms have been identified in eukaryotes.

This is in part because genetic changes with high rates of reversion tend to remain subclonal(Hartl and Jones, 1998; Lande, 1998; Maruyama and Kimura, 1974), and it is challenging to distinguish most types of low-frequency mutations from sequencing errors(Carey, 2015), especially in complex genomes with large amount of repetitive DNA or recently duplicate genes. Thus, fast growing organisms with relatively small and simple genomes are particularly well suited for determining if transient mutations exist, and for identification of the underlying mechanisms.

Results:

Microhomology mediated tandem duplications in specific genes causing reversible phenotypes in *S. pombe*.

To discover novel transient genetic drug resistance mechanisms in a eukaryote we performed a genetic screen in the fission yeast *S. pombe* for spontaneous mutants that are reversibly resistant to rapamycin plus caffeine (caffeine is required for rapamycin to inhibit growth in *S. pombe*(Weisman et al., 1997)) **(Fig. 1A)**.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. 1B,C)**. In contrast, resistance for deletion mutants such as *gaf1Δ*(Laor et al., 2015) is irreversible suggesting, the existence of a novel type of genetic or epigenetic alteration allowing for reversible drug resistance in the newly isolated strains **(Fig. 1B,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 **(Fig.**

S1A); we could not identify any common mutations in the first linkage group, suggesting an epigenetic or non-nuclear genetic mutation, or an inheritable variation that remains to be detected. In contrast, all eight strains in the second linkage group contained novel 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(Davie et al., 2015).

The *ssp1* linkage group contained three insertion alleles, all of which were tandem duplications of a short DNA segment (55/68/92 bps in length) and had 5-8 bp of identical sequence (MicroHomology Pairs, MHPs) at each end **(Fig. 1D, Fig. S1B)**. We postulate these Microhomology-mediated Tandem Duplications (MTDs)(Lawson et al., 2011; Vissers et al., 2009; Willis et al., 2017) 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 *ssp1Δ,* and replacement of the MTD alleles by transformation with wild-type *ssp1* restored sensitivity **(Fig. 1E)**. Sanger sequencing showed that all 16 drug-sensitive revertants of the MTD alleles had the wild-type *ssp1* sequence. Finally, *ssp1Δ* and $ssp1^{\mathrm{MTD}}$ strains are temperature sensitive, and spontaneous drug-sensitive non-ts 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. 1F)**. The frequency of revertants is thus at least two orders of magnitude higher than the forward mutation frequency(Farlow et al., 2015), and therefore MTDs in *ssp1* are causal for reversible temperature sensitivity and drug resistance.

To test if MTDs are specific to the drug treatment and/or *ssp1*, we performed a second screen for suppressors of the slow growth defect of *cnp1-H100M*, a point mutation in the centromere-specific histone gene, and identified MTDs in thte ranscription repressor genes *yox1* and *lsk1* **(Fig. S1B, S2)**. 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 wild-type background, the MTD is deleterious and revertants accumulate **(Fig. S1, S2)**. Thus, MTDs are not gene-specific and likely occur throughout the genome.

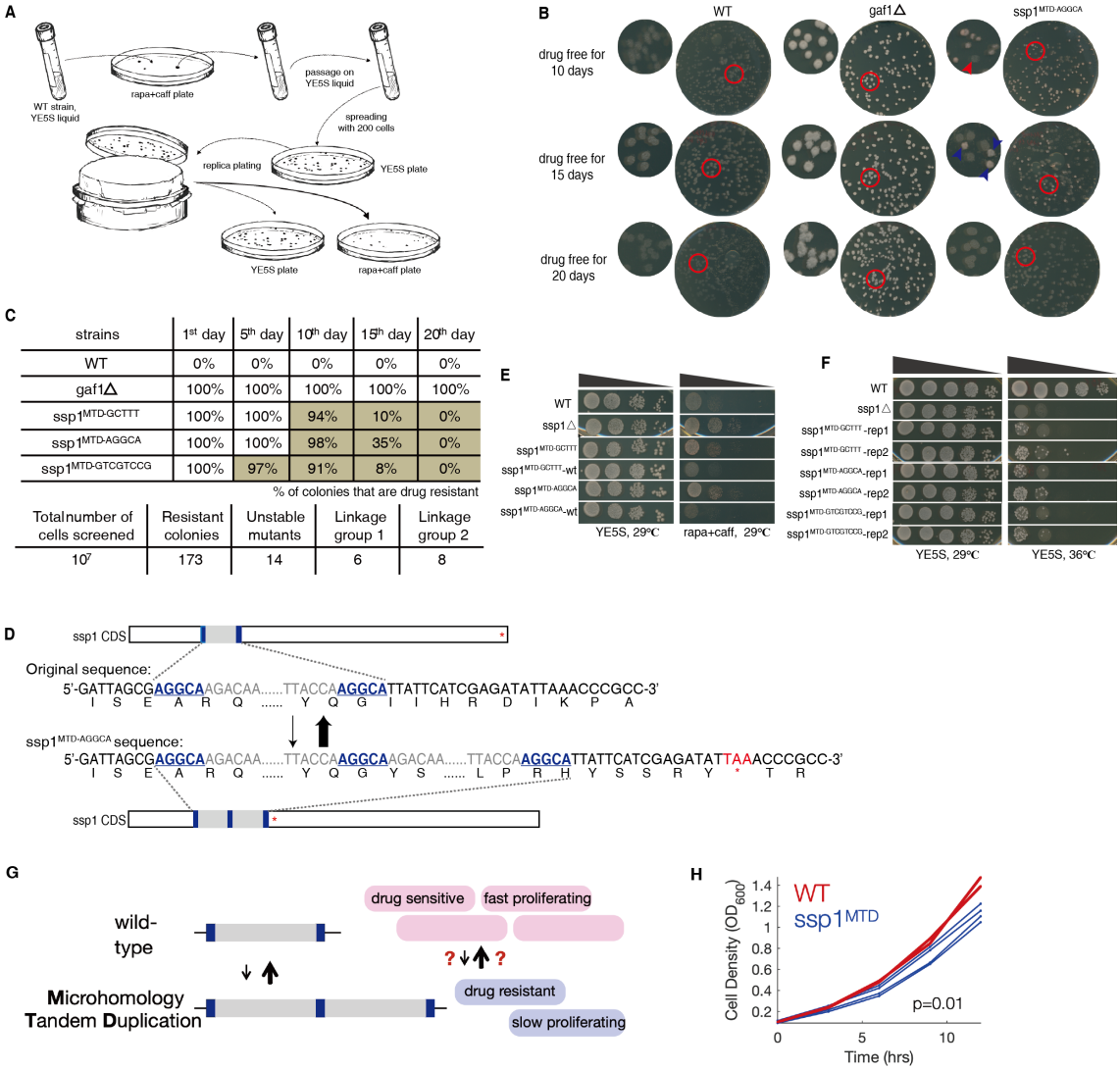


Fig. 1. Screen for mutants with unstable inherited resistance by rapamycin plus caffeine and identify highly reversible mutations in *ssp1*. (A) Procedure to screen mutants with unstable rapa+caff resistance using sensitive wild-type strains in S. *p*ombe. **(B)** Unstable phenotype for one of screened mutants on rapa+caff plates after replica plating. *gaf1Δ* 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 microhomology pair. The pre-matured 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)** Growth curves of wild-type (red, two replicates) and ssp1MTD^{AGGCA} (blue, four replicates). **(H)** A cartoon of reversible MTDs that cause drug resistance and a proliferation defect.

10,000x whole-genome sequencing identified thousands of subclonal MTDs within a clonal population

Based on the scale of the initial genetic screen, the frequency of cells with any protein-inactivating MTD in *ssp1* in an exponentially growing non-selected wild-type population is approximately $8x10^{-5}$. This 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(Hartl and Jones, 1998; Lande, 1998; Maruyama and Kimura, 1974).

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 Methods for details**)**. This method first identifies all MH Pairs (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. 2A)**. This method is capable of identifying subclonal MTDs independent of their frequency in the population.

To determine if subclonal MTDs captured by sequencing represent the true genetic variation, or are technical artifacts (Head et al., 2014) we performed two orthogonal tests. In the first, we tested if 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 *E. coli*, or chemically synthesized 150nt and 500nt fragments of the gene as well as 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. 2B, Fig. S3)**, suggesting that MTDs are largely not caused by PCR or an artifact of Illumina sequencing. The lack of MTDs in the plasmid-borne copy of *ssp1* raises the possibility that MTDs may be eukaryote-specific **(see also Fig. 2D, S6, S7, S8)**.

As a second test, we hypothesized that most MTDs in essential genes should be deleterious and recessive. We therefore analyzed raw sequencing data from 220 *S. cerevisiae* haploid and diploid mutation accumulation lines(Sharp et al., 2018). In comparison to the diploid, subclonal MTDs were depleted in essential genes in haploids in both *S. cerevisiae* (p=0.0023) and in *S. pombe* (p=0.0105) **(Fig. S3)**. Therefore, rare subclonal MTDs identified by ultra-deep sequencing are likely real biological events mostly not experimental artifacts.

To assess the prevalence of MTDs throughout the genome 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 10,000x. The diploidy relaxed selection, allowing recessive mutations to accumulate.

We annotated the *S. pombe* genome and identified 25 million MHPs with an MH length of 4-25nt and an inter-MH distance of 3-500nt. 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, suggesting that natural selection has acted to decrease the occurrence of deleterious MTDs, and that this selection is stronger for longer MH sequences **(Fig 2C,D).**

With 10,000x genome sequencing, we identified 5968 (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 **(Fig. S4)**. However, 20 genes contained more than ten different MTDs in a single 'clonal' population **(Fig. 2E)**. 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, GC content, inter-MH distance, measured nucleosome occupancy, transcription level, and a local clustering on the scale of 100nt, were able to predict which MHPs give rise to duplications with an AUC of 0.9 with 10-fold cross validation **(Fig. 2F,G, S5, Table S5)**. We note that the peak at 150nt inter MH spacing is independent of read length, was not found in *E. coli* or in mitochondrial DNA, and varies between haploid and diploid **(Fig. S5**, **S6, S7, S8)**. This analysis revealed properties of MHPs significantly affect the likelihood of MTD formation; for example, long GC-rich MH Pair is 1000x more likely to generate a tandem duplication than a short AT-rich one.

While MHPs are spread roughly uniformly throughout the genome **(Fig. 2H, 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. 2I)**. Local differences in MHPair 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 10x more likely to occur in some genes than others, and this variation is correctly predicted by our model **(Fig 2J)**. We detected no MTDs in *ura4*, which has a score of 52, placing it in the bottom third of genes **(Table S4)**, and providing a possible explanation why MTDs have not been noticed in 5-FOA based screens of

196 mutations in *ura4*(Gangloff et al., 2017). Our results also emphasize that high-coverage
197 sequencing is necessary to identify sufficient numbers of MTDs; one billion reads would be
198 required to identify half of the 25 million possible MTDs in the *S. pombe* genome **(Fig. S4)**.
199 We identified three different subclonal MTDs in the SAGA complex histone
200 acetyltransferase catalytic subunit *gcn5*, placing *gcn5* in the top 5% of genes for both observed
201 and predicted MTDs, suggesting that MTDs in *gcn5* should be found frequently in a genetic
202 screen. Indeed, examination of 16 previously identified(Xu et al., 2018) suppressors of $htb1^{G52D}$
203 identified MTDs in *gcn5*, as well as in *ubp8*, where we also observed an MTD in our high-
204 coverage sequencing data **(Fig. S1B)**. These results suggest that MTDs arise in most genes at
205 a high enough frequency within populations in order to be the raw material on which natural
206 selection acts.

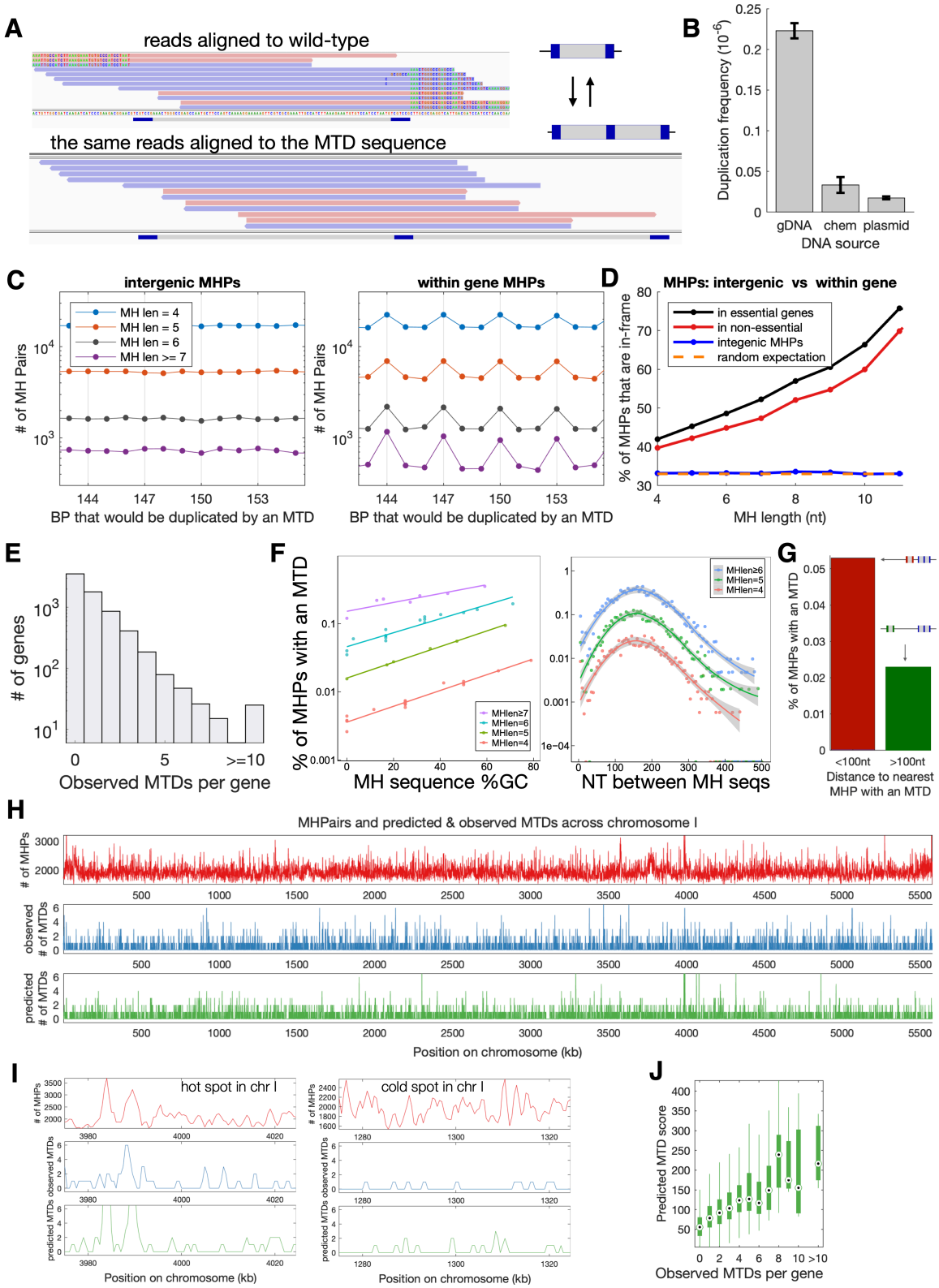


Fig. 2. Identification of the *cis*-determinants of MTD through ultra-deep sequencing and identification of subclonal duplications. (A) The computational pipeline finds all sequencing

reads that whose ends do not match the reference genome, and checks if 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. 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 standard error of the mean across replicates. **(C)** 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. **(D)** The % 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 non-essential genes (red). Random expectation is that 1/3rd of MHPs will have an insert size evenly divisible by three (orange). **(E)** A histogram of the number of MTDs found in each gene from 10,000x whole-genome sequencing. **(F)** 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 % 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. **(G)** The distance from each MHP to the nearest MHP with an MTD was calculated, and the % of MHPs with an MTD was calculated for MHPs less than (red) or farther than (green) 100nt from the closest MHP. **(H)** For each 1kb window in the genome, shown are the number of MHPairs (red), the number of observed MTDs (blue), the predicted number of MTDs from the logistic regression model (green). **(I)** An example cold spot (0.2MTDs/kb) and hot spot (0.7 MTDs/kb) in chromosome I. **(J)** 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 10k coverage data.

240 **Replication slippage modulates the rate of MTD reversion at *ssp1*.**

241

242 Having established that local *cis*-encoded features determine the frequency with which
243 tandem duplications arise from microhomology-pairs, we next sought to identify the *trans*-
244 genes that affect MTD process. $ssp1^{MTD}$ alleles fail to grow at 36ºC, and their reversion back
245 to wild-type suppresses the temperature sensitivity, providing way to measure the effects of
246 mutations on reversion frequency. We screened a panel of 290 strains with mutations in DNA
247 replication, repair, recombination or chromatin organization genes for mutants that affect the
248 rate of $ssp1^{MTD}$ reversion back to wild-type, and found three mutants that significantly
249 increased and eight that significantly decreased the frequency of $ssp1^{WT}$ revertants **(Fig.**
250 **3A,B,C)**.

251

252 Replication fork collapse is a major source of double stranded breaks (DSBs), and the
253 ensuing Homologous Recombination (HR)-related restarting process is error-prone and is
254 known to generate microhomology flanked insertions and deletions via replication slippage
255 (Iraqui et al., 2012). Inactivation of Rad50, Rad52 or Ctp1 results in decreased replication
256 slippage, and decreased MTD reversion **(Fig. 3A,B,C)**. Deletions of *mhf1* and *mhf2*, two
257 subunits of the FANCM-MHF complex, which is involved in the stabilization and remodeling
258 of blocked replication forks, also decreased the frequency of MTD revertants. It is therefore
259 likely that replication slippage during HR-mediated fork recovery contributes to the reversion
260 of MTDs.

261

262

263 Replication stresses activate a checkpoint that promotes DNA repair and recovery of stalled
264 or collapsed replication forks, and delays entry into mitosis(Alcasabas et al., 2001; Myung and
265 Kolodner, 2002). The inactivation of replication checkpoint kinase *cds1* or its regulator *mrc1*
266 may thus result in a failure to restore the replication fork, causing increased genome instability
267 and MTD reversion. The replication checkpoint is required for the stability of MTDs; deletion
268 of the DNA damage checkpoint kinase *cds1* or its regulator *mrc1* increased the frequency of
269 $ssp1^{WT}$ revertants. Deletion of the single-stranded DNA binding A (RPA) subunit *ssb3*
270 (RPA3/RFA3) or the multifunctional 5'-flap endonuclease *rad2* also increased the frequency
271 of revertants **(Fig. 3C)**.

272

273 Many genes identified in the screen are multifunctional, and play roles in both replication
274 and repair. We therefore performed quantitative epistasis analysis to determine the relation
275 between six of the identified genes and the Mediator of the Replication Checkpoint, *mrc1*,
276 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 3D)**. In contrast, deletion of *ssb3* increased the frequency of revertants in both wild-type and *mrc1*Δ 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 3D)**.

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 predicted that the subclonal MTDs would be reduced in *cds1*Δ and *rad2*Δ strains. To test this we performed 10^6x coverage sequencing of the hotspot gene *SPCC1235.01*. Consistent with the prediction, we observe MTDs at fewer MHPairs, and an overall decrease in the number of MTDs in both mutants **(Fig. 3E,F)**.

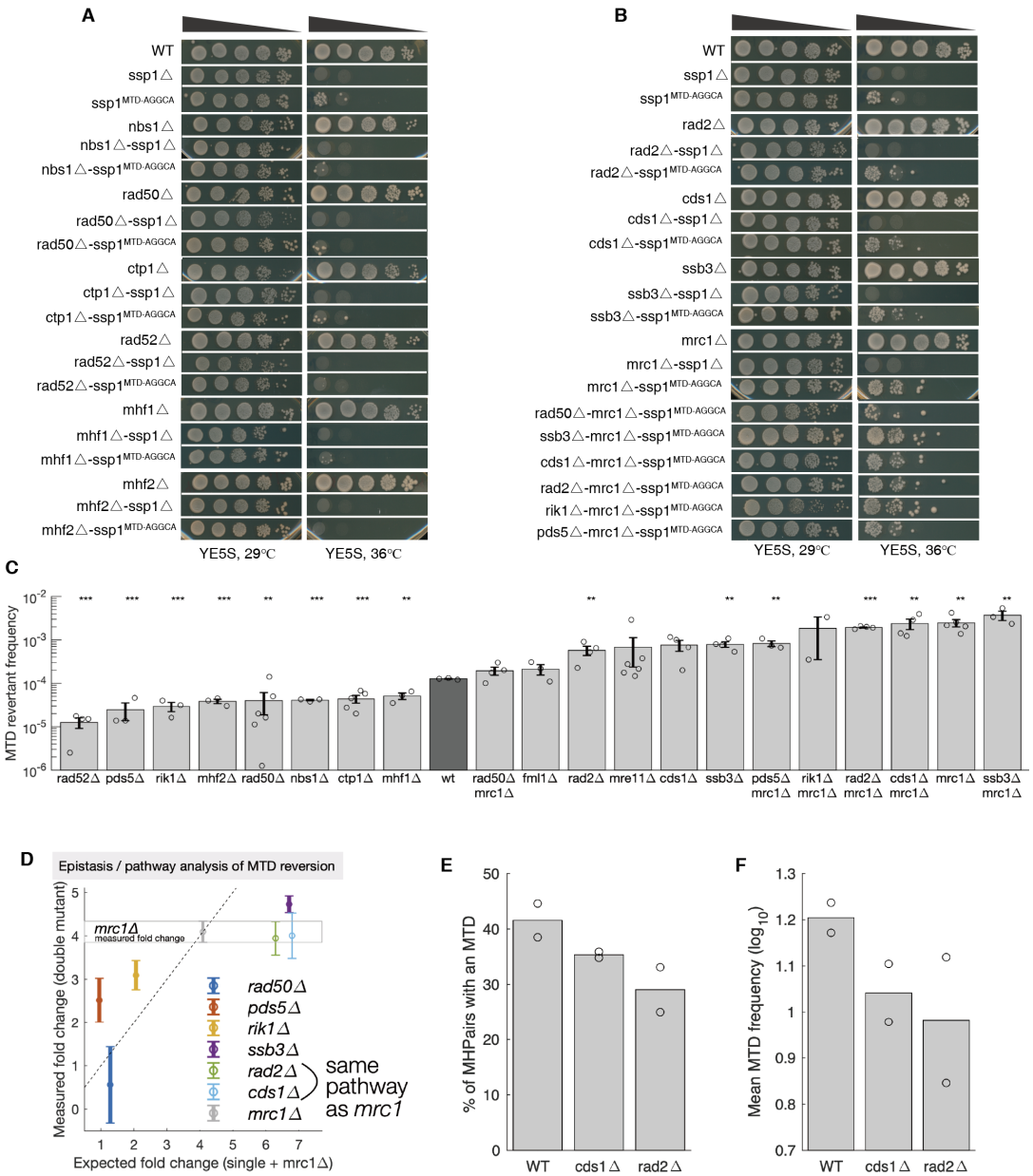
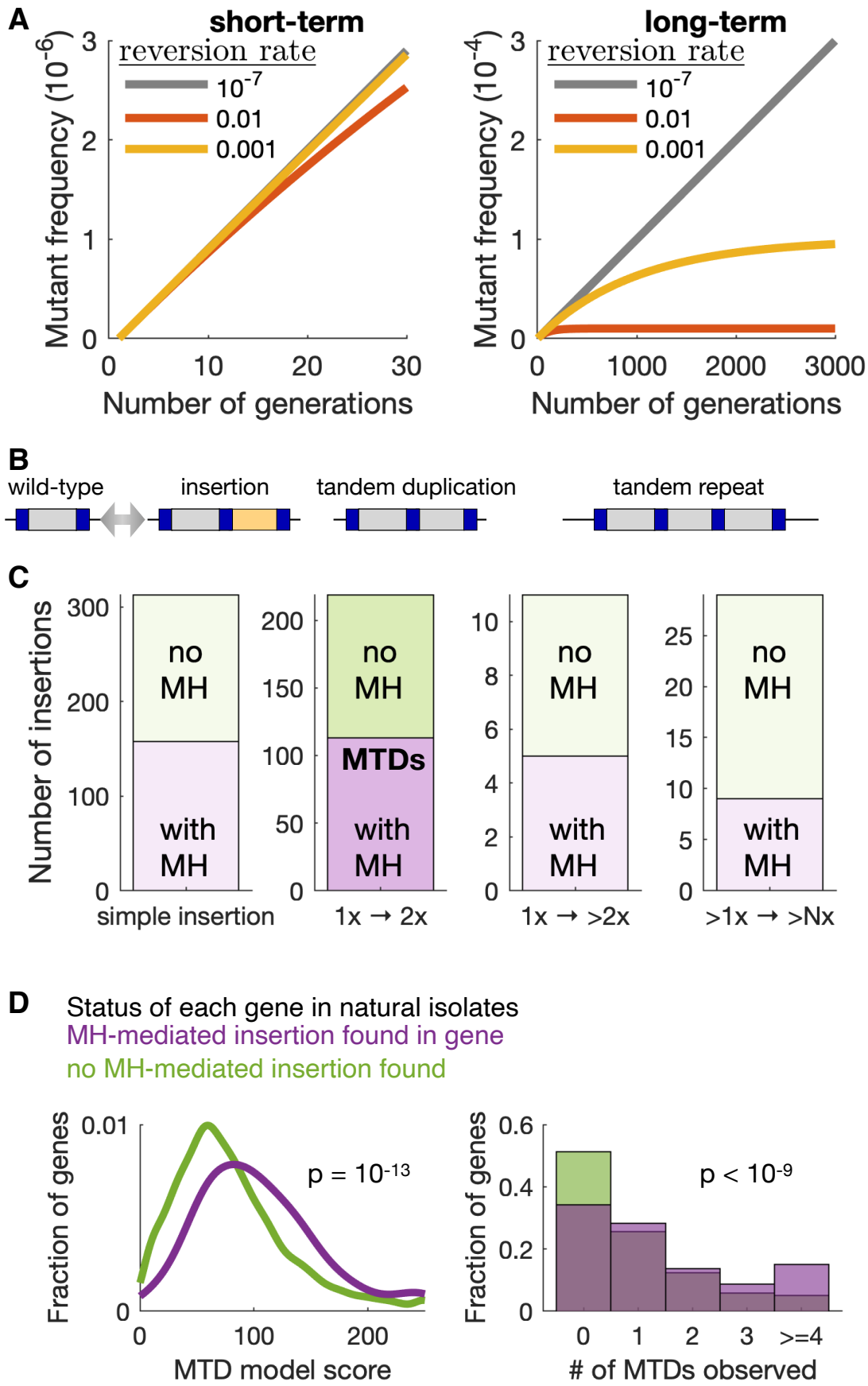


Fig. 3. A genetic screen to identify the regulators of MTD reversion. (A,B). Surveyed mutants showed reduced ssp1^{MTD} reversion frequency represented by TS recovery phenotype. The non-TS phenotype of single mutation and *ssp1Δ* alone or combined with other mutants retained severe temperature sensitive 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 tenfold gradient (cell number: 10^5, 10^4, 10^3, 10^2, 10^1). **(C)**. Quantification of ssp1^{MTD} reversion frequency in mutants (n>=3 biological repeats, error bars are s.e.m., *** = p<0.001, **=p<0.01, *=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 3002 MHPairs. **(E)** The % 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.

306 **Half of insertions and tandem duplications in natural isolates are MH-mediated**

307 It was baffling that MTDs are prevalent within populations, and that the first theoretical
308 proposal for microhomology-mediated processes in the generation of tandem duplications is
309 twenty years old(Haber and Louis, 1998), yet, relatively little is known about the forward
310 process, and even less about the reversion, suggesting that these events are not often
311 encountered, or at least not identified as such. To better understand the dynamics of MTDs
312 within a population we used a simple model of neutral mutations within a growing population
313 that takes into account both forward and reverse mutation rates and began with 100% of
314 individuals as wild-type (see Methods). The mutant frequency always increases, and over short
315 timescales **(Fig. 4A**, left**)** increasing the reverse rate from being equal to the forward mutation
316 rate (grey) to being 10,000 times higher (yellow) has little effect.
317
318 Over longer timescales, high reversion rates cause the mutant frequency to plateau and
319 remain subclonal **(Fig. 4A**, right), reducing the fraction of neutral MTDs within a population.
320 However, in spite of the high reversion rate, both drift and selection enable fixation of MTDs
321 within a population. To identify fixed microhomology mediated insertions we searched the
322 genome sequences of 57 wild *S. pombe* isolates (Jeffares et al., 2015), and found that 50% of
323 insertions larger than 10bp involve microhomology repeats **(Fig 4B,C)**. Among these were 158
324 microhomology mediated insertions that did not contain an obvious duplication, and 113 MTDs
325 with a microhomology mediated tandem duplication.
326
327 To test if the propensity of MTD formation within the lab strain is predictive of extant
328 sequence variation observed in natural isolates, we tested if the MTD score predicted for each
329 gene predicts the likelihood of microhomology mediated insertions in that gene. We found that
330 genes with microhomology mediated insertions in natural isolates tend to have higher predicted
331 MTD scores, and more experimentally observed MTDs **(Fig. 4D)**, suggesting that the local
332 features that affect MTD formation in the lab also shape evolution in nature.
333
334 Taken together, our results demonstrate that MTDs occur frequently and broadly throughout
335 the genome within a clonal population. This indicates that high levels of subclonal genetic
336 divergence may be prevalent but are under-detected using conventional sequencing approaches
337 that tend to disfavor the detection of low abundance subclonal variants. As many MTDs create
338 large insertions, they are more likely to be deleterious. Nonetheless, MTDs provide plasticity
339 to the genome and its functionality, for example, by allowing cells to become drug resistant,
340 while allowing the resistant cell lineage to revert back to wild-type and regain high fitness once
341 the drug is removed. Selection can act on this genetic diversity for its reversibility or by using
342 the tandem duplications as the initial step for the generation of longer repeats.


A
short-term
long-term
reversion rate
10^-7
0.01
0.001
Mutant frequency (10^-6)
Mutant frequency (10^-4)
Number of generations
B
wild-type
insertion
tandem duplication
tandem repeat
C
Number of insertions
no MH
with MH
MTDs
simple insertion
1x → 2x
1x → >2x
>1x → >Nx
D
Status of each gene in natural isolates
MH-mediated insertion found in gene
no MH-mediated insertion found
Fraction of genes
p = 10^-13
MTD model score
p < 10^-9
# of MTDs observed

343

Fig 4. MTDs remain subclonal due to 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. **(B)** Cartoon of three types of microhomology mediated insertions. **(C)** Quantification of all insertions of at least 10bp 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 the presence (purple) or absence (green) of exact microhomology pairs on either side of the insert, and to the type of insert. There are 113 MTDs in wild *pombe* strains (second column). **(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.

361

362 **References and Notes:**

363 Alcasabas, A.A., Osborn, A.J., Bachant, J., Hu, F., Werler, P.J., Bousset, K., Furuya, K.,
364 Diffley, J.F., Carr, A.M., and Elledge, S.J. (2001). Mrc1 transduces signals of DNA
365 replication stress to activate Rad53. Nat. Cell Biol. *3*, 958–965.

366 Andersson, D.I., and Levin, B.R. (1999). The biological cost of antibiotic resistance. Curr.
367 Opin. Microbiol. *2*, 489–493.

368 Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic
369 Acids Res. *27*, 573–580.

370 Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for
371 Illumina sequence data. Bioinformatics *30*, 2114–2120.

372 Carey, L.B. (2015). RNA polymerase errors cause splicing defects and can be regulated by
373 differential expression of RNA polymerase subunits. Elife *4*.

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

378 Davie, E., Forte, G.M.A., and Petersen, J. (2015). Nitrogen regulates AMPK to control
379 TORC1 signaling. Curr. Biol. *25*, 445–454.

380 Dhar, R., Missarova, A.M., Lehner, B., and Carey, L.B. (2019). Single cell functional
381 genomics reveals the importance of mitochondria in cell-to-cell phenotypic variation. Elife *8*.

382 Farlow, A., Long, H., Arnoux, S., Sung, W., Doak, T.G., Nordborg, M., and Lynch, M.
383 (2015). The Spontaneous Mutation Rate in the Fission Yeast Schizosaccharomyces pombe.
384 Genetics *201*, 737–744.

385 Gangloff, S., Achaz, G., Francesconi, S., Villain, A., Miled, S., Denis, C., and Arcangioli, B.
386 (2017). Quiescence unveils a novel mutational force in fission yeast. ELife *6*, e27469.

387 Gemayel, R., Vinces, M.D., Legendre, M., and Verstrepen, K.J. (2010). Variable tandem
388 repeats accelerate evolution of coding and regulatory sequences. Annu. Rev. Genet. *44*, 445–
389 477.

390 Haber, J.E., and Louis, E.J. (1998). Minisatellite Origins in Yeast and Humans. Genomics *48*,
391 132–135.

392 Hartl, D.L., and Jones, E.W. (1998). Genetics: principles and analysis (Sudbury, Mass: Jones
393 and Bartlett Publishers).

Head, S.R., Komori, H.K., LaMere, S.A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D.R., and Ordoukhanian, P. (2014). Library construction for next-generation sequencing: overviews and challenges. BioTechniques *56*, 61–64, 66, 68, passim.

Iraqui, I., Chekkal, Y., Jmari, N., Pietrobon, V., Fréon, K., Costes, A., and Lambert, S.A.E. (2012). Recovery of arrested replication forks by homologous recombination is error-prone. PLoS Genet. *8*, e1002976.

Jeffares, D.C., Rallis, C., Rieux, A., Speed, D., Převorovský, M., Mourier, T., Marsellach, F.X., Iqbal, Z., Lau, W., Cheng, T.M.K., et al. (2015). The genomic and phenotypic diversity of Schizosaccharomyces pombe. Nat Genet *47*, 235–241.

Kim, D.-U., Hayles, J., Kim, D., Wood, V., Park, H.-O., Won, M., Yoo, H.-S., Duhig, T., Nam, M., Palmer, G., et al. (2010). Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. Nat. Biotechnol. *28*, 617–623.

Krawchuk, M.D., and Wahls, W.P. (1999). High-efficiency gene targeting in Schizosaccharomyces pombe using a modular, PCR-based approach with long tracts of flanking homology. Yeast *15*, 1419–1427.

Lande, R. (1998). Risk of population extinction from fixation of deleterious and reverse mutations. Genetica *102–103*, 21–27.

Lang, G.I. (2018). Measuring Mutation Rates Using the Luria-Delbrück Fluctuation Assay. Methods Mol. Biol. *1672*, 21–31.

Laor, D., Cohen, A., Kupiec, M., and Weisman, R. (2015). TORC1 Regulates Developmental Responses to Nitrogen Stress via Regulation of the GATA Transcription Factor Gaf1. MBio *6*, e00959.

Lawson, A.R.J., Hindley, G.F.L., Forshew, T., Tatevossian, R.G., Jamie, G.A., Kelly, G.P., Neale, G.A., Ma, J., Jones, T.A., Ellison, D.W., et al. (2011). RAF gene fusion breakpoints in pediatric brain tumors are characterized by significant enrichment of sequence microhomology. Genome Res. *21*, 505–514.

Lenski, R.E. (1998). Bacterial evolution and the cost of antibiotic resistance. Int. Microbiol. *1*, 265–270.

Levy, S.F., Ziv, N., and Siegal, M.L. (2012). Bet hedging in yeast by heterogeneous, age-correlated expression of a stress protectant. PLoS Biol. *10*, e1001325.

Maruyama, T., and Kimura, M. (1974). A NOTE ON THE SPEED OF GENE FREQUENCY CHANGES IN REVERSE DIRECTIONS IN A FINITE POPULATION. Evolution *28*, 161–163.

Myung, K., and Kolodner, R.D. (2002). Suppression of genome instability by redundant S-phase checkpoint pathways in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U.S.A. *99*, 4500–4507.

Nicoloff, H., Hjort, K., Levin, B.R., and Andersson, D.I. (2019). The high prevalence of antibiotic heteroresistance in pathogenic bacteria is mainly caused by gene amplification. Nat Microbiol *4*, 504–514.

Rando, O.J., and Verstrepen, K.J. (2007). Timescales of genetic and epigenetic inheritance. Cell *128*, 655–668.

Shaffer, S.M., Dunagin, M.C., Torborg, S.R., Torre, E.A., Emert, B., Krepler, C., Beqiri, M., Sproesser, K., Brafford, P.A., Xiao, M., et al. (2017). Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. Nature *546*, 431–435.

Sharma, S.V., Lee, D.Y., Li, B., Quinlan, M.P., Takahashi, F., Maheswaran, S., McDermott, U., Azizian, N., Zou, L., Fischbach, M.A., et al. (2010). A Chromatin-Mediated Reversible Drug-Tolerant State in Cancer Cell Subpopulations. Cell *141*, 69–80.

Sharp, N.P., Sandell, L., James, C.G., and Otto, S.P. (2018). The genome-wide rate and spectrum of spontaneous mutations differ between haploid and diploid yeast. Proc Natl Acad Sci USA *115*, E5046–E5055.

Verstrepen, K.J., Jansen, A., Lewitter, F., and Fink, G.R. (2005). Intragenic tandem repeats generate functional variability. Nat Genet *37*, 986–990.

Vissers, L.E.L.M., Bhatt, S.S., Janssen, I.M., Xia, Z., Lalani, S.R., Pfundt, R., Derwinska, K., de Vries, B.B.A., Gilissen, C., Hoischen, A., et al. (2009). Rare pathogenic microdeletions and tandem duplications are microhomology-mediated and stimulated by local genomic architecture. Hum. Mol. Genet. *18*, 3579–3593.

Weisman, R., Choder, M., and Koltin, Y. (1997). Rapamycin specifically interferes with the developmental response of fission yeast to starvation. J. Bacteriol. *179*, 6325–6334.

Willis, N.A., Frock, R.L., Menghi, F., Duffey, E.E., Panday, A., Camacho, V., Hasty, E.P., Liu, E.T., Alt, F.W., and Scully, R. (2017). Mechanism of tandem duplication formation in BRCA1-mutant cells. Nature *551*, 590–595.

Xu, X., Wang, L., and Yanagida, M. (2018). Whole-Genome Sequencing of Suppressor DNA Mixtures Identifies Pathways That Compensate for Chromosome Segregation Defects in *Schizosaccharomyces pombe*. G3 *8*, 1031–1038.

Acknowledgments: We thank Lilin Du, Aaron New, and Wenfeng Qian for insightful discussions and for comments on the manuscript.

Funding: 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 Grant 31628012.

Author contributions (CRedIT):

AUTHORS	ROLE
X.H.	Conceptualization
L.D., S.C., Y.L., X.H., L.B.C.	Methodology
S.C., Y.L., & L.B.C.	Software
L.D., S.C., Y.L., & L.B.C.	Validation
L.D., S.C., Y.L., & L.B.C.	Formal analysis
L.D., S.C., Y.L., J.L & L.B.C.	Investigation
X.H., F.L., Y.W.	Resources
L.D., S.C. Y.L. & L.B.C.	Data Curation
L.B.C. L.D. & X.H.	Writing
L.D., S.C. X.H. & L.B.C.	Visualization
L.B.C. & X.H.	Supervision
L.B.C. & X.H.	Project administration
L.B.C. & X.H.	Funding acquisition

Competing interests: Authors declare no competing interests.

Data and materials availability: All processed data and code are available at https://github.com/carey-lab/MicroHomologyMediatedTandemDuplications and raw sequencing data at NCBI GEO accession @@@@.

Materials and Methods

Strains

S.pombe strains used in this study are listed in Table S1. The deletion strains and GFP-tagging strain were originated from the genome-wide deletion library(Kim et al., 2010) or constructed by overlap PCR strategy and gene-specific homologous recombination using standard procedures(Krawchuk and Wahls, 1999).

Cell Growth

Fission yeast cells were grown on YE5S liquid or solid medium (5S: supplemented with histidine, uracil, lysine, leucine, adenine), mated or sporulated on specific malt extract (ME) agar medium following standard procedures (Ekwall and Thon, 2017). For the preparation of rapamycin plus caffeine drug plate, 1000X stock solution of rapamycin(100μg/ml) was

484 prepared by adding 100mg rapamycin to 1ml DMSO (100mg/ml) and diluting by 1000 folds.
485 1.942g power caffeine was dissolved in 60-80ºC 20 ml sterile ddH_2O and added into 1L YE5S
486 medium to final concentration 10mM.
487
488 **Unstable drug-resistant mutants screen**
489 A fresh single colony of wild-type cells was picked and grown to mid-log phase culture.
490 Cultivated cells were then spread on YE5S agar plates containing 100ng/ml rapamycin and
491 10mM caffeine (hereafter called YE5S+drug plates) at the density of 1×10^5 cells per plate, and
492 incubated at 29°C for 10 days.. To test the stability of the drug resistance, each strain is grown
493 continuously in YE5S liquid media in the absence of the drugs at 29°C by refreshing the culture
494 with YE5S liquid media daily for up to 20 days. Every five days, cell samples were taken and
495 spread to the YE5S plate at the density of 200 cells per plate. After 3 day incubation at 29°C,
496 each plate was replica plated to fresh YE5S and YE5S+drugs plates, respectively, incubated
497 for two days at 29°C. Plates were visually examined for colonies that grow on YE5S but fail to
498 grow on YE5S+drugs plates. The stability test was repeated at least two times for identified
499 unstable drug-resistant stains. The *gaf1-d* mutant was used as the control for stable and robust
500 drug resistance.
501
502 **Genetic linkage test**
503 Identified unstable drug-resistant strains were backcrossed with wild-type cells or crossed with
504 each other on the ME plate. After 24-48h sporulation at 29°C, tetrad-dissection was performed
505 on the YE5S plate following the standard procedure(Escorcia and Forsburg, 2018). After 3
506 days incubation at 29°C, YE5S plates are replica plated to the YE5S+drug plate and incubated
507 at 29°C for 2 days to identify drug-resistant colonies among the four progeny originated from
508 one ascus. The segregation pattern of the drug-resistant and drug-sensitive phenotypes is
509 analyzed and used to determine the genetic linkage of the tested mutation alleles.
510
511 **Whole-genome sequencing and datasets analysis**
512 Genomic DNA was extracted using phenol-chloroform, mechanically sheared to ~200bp using
513 ultrasonicator. Sheared genome DNA was used to build the library using NEBNext® Ultra™
514 DNA Library Prep Kit for Illumina® (E7370/7335, NEB) and Illumina sequenced by Ribobio
515 in Wuhan, China.
516 For data analysis, adapter-trimmed FASTQ clean data were mapped to the ENSEMBL
517 Fungi's *S. pombe* genome version ASM294v2 with the BWA mem aligner(Li and Durbin,
518 2009) (version 0.7.17, with -M flag on). After removing PCR duplicates(Li et al., 2009),
519 alignment maps (BAM files) were fed to the GATK's HaplotypeCaller for a first run. The
520 output variants are used to recalibrate base quality scores in the BAM files using GATK's

521 BaseRecalibrator. Recalibrated BAM files were then inputted to the HaplotypeCaller to
522 generate raw mutation callings(McKenna et al., 2010), which were filtered and annotated using
523 the ENSEMBL's variant effect predictor (VEP, version 93.3).
524
525 **Double mutant construction and MTD reversion regulators survey**
526 Double mutants which combine MTD mutation at *ssp1* (*ssp1*$^{MTD\text{-}AGGCA}$) and each deletion
527 within the mutant panel were created by genetic crossing following the standard procedures
528 (Roguev et al., 2018; Schuldiner et al., 2006) using a high throughput robotic apparatus (Peking
529 university, F. Li lab. Protocol for high throughput manipulation is available upon request).
530 To assess MTD reversion rates semi-quantitatively, a single colony of each double mutant
531 was used to inoculate 3ml YE5S liquid culture, incubated at 29°C overnight, refreshed by 1:10
532 dilution in 20ml YE5S liquid medium, and grown to mid-log phase. Serial 1:10 dilutions of the
533 culture were prepared using fresh YE5S liquid medium, spotted on YE5S plates (5□l per spot,
534 corresponding to 10^5 to 10 cell per spot), incubated for 4-5 days at 36°C or 29°C.
535 To quantify the reversion frequency, double mutant cells were spread on YE5S plates at the
536 density of 2×10^4 or 2×10^5 cells per plate, incubated 4-5 days at 36°C. The number of non-ts
537 revertant colonies was scored in three biological repeats for each double mutant strains. Two
538 non-ts revertants were picked for each strain and the *ssp1* locus in these revertants is PCR
539 amplified/Sanger sequenced to verify true reversion of *ssp1*$^{MTD\text{-}AGGCA}$ to wild type.
540 We also did a fluctuation test for some double mutant strains to quantify the reversion
541 frequency by growing a single overnight culture in YE5S broth for the strain to be tested,
542 diluting with fresh YE5S broth to obtain 10^2 yeast cells /ml. For each strain, the diluted
543 suspension was divided into 48 of 100μL and incubated at 29°C. Then 40 replica cultures with
544 100μL were plated in their entirety onto YE5S agar plates and incubated at 36°C for 3~5 days.
545 For the rest 10 100μL replica cultures, the average number of cells per culture (N) was
546 calculated using a blood counting chamber. Then counted the number of 36°C survival cells
547 (reverted wt cells) per culture, and calculated the mutation rate with the p0 method or the MSS-
548 maximum likelihood method (Lang, 2018).
549
550 **Identification of *yox1*^{MTD} and *lsk1*^{MTD} among *cnp1*^{H100M} suppressors and stability test for**
551 ***yox1*^{MTD}**
552 Haploid *cnp1*^{H100M} cells derived from heterozygous *cnp1*^{H100M} diploid by tetrad dissection were
553 spread on YE5S plates, incubated at 29ºC for 5 days. Rare large colonies (~1/10^4) were isolated
554 as spontaneous *cnp1*^{H100M} suppressors (FigS2A). Whole-genome sequencing was performed
555 on isolated *cnp1*^{H100M} suppressors to identify the target gene. With the analysis process in
556 "Whole-genome sequencing and datasets analysis" part, MTD events in *yox1* and *lsk1* gene
557 were identified and verified by Sanger sequencing.

558 To verify the genetic stability of *yox1*^{MTD} alleles, *cnp1*^{H100M} suppressors were backcrossed
559 with wild type, *yox1*^{MTD} were separated from *cnp1*^{H100M} mutation. *yox1*-GFP, *yox1*^{MTD}-GFP
560 strains were constructed by fusing a GFP tag in the endogenous *yox1* locus (FigS2B). MTD
561 (20bp tandem duplication) in *yox1* disrupts the open reading frame and generates a premature
562 stop codon (TAG) at 523nt loci, resulting in inactivation of GFP fluorescence, while the
563 reversion of yox1^{MTD} would recover the GFP fluorescence. In the stability test, *yox1*-GFP and
564 *yox1*^{MTD}-GFP cells were grown continuously at 29℃ by refreshing the culture with YE5S
565 liquid media daily for up to 60 days. Every ten days, cell samples were taken and subjected to
566 microscopical observation for GFP fluorescence. The percentage of progenies exhibiting the
567 nuclear GFP signal was scored in three individual biological repeats. To verify *yox1*^{MTD}-GFP
568 reversion, *yox1* locus of @ single colonies derived from *yox1*^{MTD}-GFP 40 day culture was
569 amplified by PCR and subjected for Sanger sequencing
570
571
572 **Finding microhomology pairs on genome**
573 A fast algorithm is implemented to find micro-homology pairs across the S. pombe's genome
574 sequence (or any given DNA sequence). First, the input sequence is scanned one-time for initial
575 k-mer homology pairs with pre-set limitations. Here we arbitrarily set limitations to 1) the size
576 of the homology should be no smaller than 4 bps and no greater than 12 bps, 2) the homology
577 should not be a mononucleotide repeat, 3) space between two homologies in a pair should be
578 greater than 3 bps, and 4) the INDEL size (the length of a homology plus the inter-space) should
579 not exceed 100 bps. Then, the initial homology pairs are forth scanned for one run to merge
580 adjacent homology pairs to longer pairs. The current implementation would only report the
581 left-most pair of tandem repeats with micro-homology pairs on repeat junctions
582
583
584 **Annotating insertions and tandem repeats flanked by micro-homology pairs in natural**
585 **isolates and in the reference genome.** To identify MH-flanked tandem repeats in the reference
586 genome we used the Tandem Repeat Finder(Benson, 1999) to generate an initial tandem repeat
587 candidate list. All parameters were set to the default value except the INDEL penalty, which
588 was set to 1000 to avoid reporting tandem repeats with non-uniform unit sizes. After removing
589 candidates with the reported unit size smaller than 10nt, self-information smaller than 1.5 bits,
590 and repeat number smaller than 2, remaining tandem repeats were verified by three steps: 1)
591 finding if there were still internal repeats within the reported repeat unit, 2) finding if there
592 were still repeat units on the left and right wings to the reported length, and 3) sliding the whole
593 frame to the left-most base while the repeats' consistency did not drop. Finally, we checked
594 the junctions for the existence of a micro-homology of at least 2nt. If homology size is long

enough (longer than 75% of the unit size and longer than unit size–4 bps), we considered it as repeat number plus 1 and start over for finding junction micro-homologies.

To identify MH-flanked insertions in natural isolates we used the indels .vcf file from Jeffares et al.(Jeffares et al., 2015) and used SnpEff (Cingolani et al., 2012) to predict the impact of each indel. We extracted the left and right flanking sequences from the reference genome to determine the presence of microhomology and to identify the repeat unit.

Site-specific PCR amplification and ultra-deep next-generation sequencing

The fresh single colony was picked from the YE5S plate, inoculate 3ml YE5S liquid medium and incubated at 29ºC overnight. The mini-culture was refreshed by 1:10 dilution in 20ml YE5S liquid medium and grown to mid-log phase. Genomic DNA was extracted using phenol-chloroform, used as the template for PCR amplification with high fidelity polymerase (RR006Q, Takara, Tokyo, Japan). Alternatively, a plasmid containing the *ssp1* coding sequence was constructed and amplified in E.coli, extracted, and digested with endonucleases to release the *ssp1* DNA fragments. Chemically synthesized ssp1 DNA fragments were produced by commercial service (Hzykang, Hangzhou, China). ssp1 DNA fragments from various sources described above were subjected to Illumina NGS following standard procedure at the coverage of ~1×10^6 (Bioacme, Wuhan, China).

The sequencing library of wild type diploid cells derived from a single colony was prepared as above and subjected to Illumina NGS by Frasergen in Wuhan, China.

For sequencing data analysis, trimmed FASTQ files are mapped to the reference sequences with BWA mem (with -Y flag on) and only primary alignments are kept. The program described in "Finding micro-homology pairs" is used here to find micro-homology pairs in the library reference. The left and right adjacent bases (here we arbitrarily chose 10 bps) to each of the two homologies in the micro-homology pairs are extracted as "signature sequences". Then the alignment maps are scanned: for a clipped read, in those pairs that are possible to generate duplication/collapse at the clipping position, we test whether the clipped sequence matches with any pair's "signature sequence"; for an INDEL possessing read, we test the opening and ending positions (and as well the inserted sequence for insertion reads).

Simulations of mutation frequency at different reversion rates

Let A_{wt} be the wild-type allele, and A_{mut} be the mutant allele. Let k_{fwd} be the forward (A_{wt} to A_{mut}) mutation rate, and k_{rev} be the reverse (A_{mut} to A_{wt}) mutation rate. Let p_t be the frequency of A_{wt} and q_t the frequency of A_{mut} at time t. Then, if we assume that mutations are neutral, the A_{mut} genotype frequency (q) changes as $q_{(t+1)} = q_t + (k_{fwd} \times p_t - k_{rev} \times q_t)$, and p=(1-q). When $k_{fwd} >= k_{rev}$ or, as is the case for subclonal MTDs, when q is small, the reverse mutation can

632 mostly be ignored. However, when $k_{rev} >> k_{fwd}$ or $q == 1$, as is the case for clonal fixed MTDs,
633 k_{rev} has a large impact on dynamics. For simulations, the initial conditions were set to p=1,q=0,
634 $k_{fwd} = 10^{-7}$, and k_{rev} was varied as is shown in the figure.
635
636 **Logistic regression to predict MTD frequency from local features**
637 To predict the likelihood of a duplication event in each micro-homology pair (MHP), we used
638 a logistic regression model (the function glm() from R) with 10-fold cross-validation. The data
639 are highly imbalanced; MTDs were detected at fewer than 0.1% of MHPs. We therefore trained
640 and tested the model using a balanced dataset consisting of all MHPs with an MTD, plus a
641 randomly chosen subset MHPs with no MTD of the same size, so that half of MHPs had an
642 MTD. We first trained a model using three features: MHlength, GC-content-of-the-MH-
643 sequence, and inter-MH-distance, which has an AUC of 0.876. This is the "top 3 features"
644 model, and all three of these features are predictive by visual inspection (e.g., Fig 2). To
645 determine which additional features to add we continuously added features, and kept only those
646 that increased the AUC over this base 3-feature model. The additional predictive features were:
647 MHPlength (MHlen), nucleotides between two repeats (interMH), interMH (interGCcon),
648 nucleosome occupancy (entire_nucle) and gene expression (entire_gene) of the entire MHP,
649 and nucleotides to the closest MHR which has duplication event(ntclosestMHR).
650 To perform whole-genome predictions using the model trained on the balanced data, we used
651 the model to score all 25 million MHPs in the genome, and either used the sum of predicted
652 scores for all MHPs in a single gene, or selected the top 6234 MHPs, the same number of
653 duplication events as observed experimentally, to be predicted duplication events.
654

A

wt×wt	gaf1△×wt	Epm101×wt	Epm102×wt	Epm101×Epm102
0:4 (17/17)	2:2 (33/34)	2:2 (84/84)	2:2 (55/56)	4:0 (16/16)

Backcross	R:S (pairs/ total pairs)	Cross	R:S (pairs/ total pairs)
Epm201 × wt	2:2 (37/42)	Epm101 × Epm201	4:0 (19/19)
Epm203 × wt	2:2 (38/39)	Epm102 × Epm201	4:0 (25/26)
Epm205 × wt	2:2 (46/46)	Epm101 × Epm205	4:0 (20/20)
Epm206 × wt	2:2 (36/36)	Epm102 × Epm205	4:0 (24/24)
		Epm201 × Epm205	4:0 (36/37)
		Epm203 × Epm205	4:0 (26/26)
Epm103 × wt	2:2 (23/24)	Epm106 × Epm103	4:0 (16/16)
Epm104 × wt	2:2 (37/37)	Epm106 × Epm104	4:0 (26/26)
Epm105 × wt	2:2 (16/17)	Epm106 × Epm105	4:0 (17/17)
Epm106 × wt	2:2 (30/30)	Epm106 × Epm107	4:0 (20/20)
Epm107 × wt	2:2 (26/27)	Epm106 × Epm108	4:0 (15/15)
Epm108 × wt	2:2 (20/20)	Epm104 × Epm105	4:0 (11/11)
		Epm206 × Epm106	3:1 (11/17), 2:2 (6/17)

B

ssp1 original sequence:
5'-GTATGG**GCTTT**AGGCGT.....GTATGA**GCTTT**TTGACA-3'
V W A L G V..... Y E L F D

ssp1$^{MTD\text{-}GCTTT}$ sequence:
5'-GTATGG**GCTTT**AGGCGT.....GTATGA**GCTTT**AGGCGT.....GTATGA**GCTTT**TTGACA-3'
V W A L G V..... Y E L * A ..... S M S F L T

ssp1 original sequence:
5'-ACGGAAC**GTCGTCCG**AAACTG.....CCTAAT**GTCGTCCG**CTTGCG-3'
T E R R P K L ..... P N V V R L R

ssp1$^{MTD\text{-}GTCGTCCG}$ sequence:
5'-ACGGAAC**GTCGTCCG**AAACTG.....CCTAAT**GTCGTCCG**AAACTG.....TCTTAAAGAAATGTGTCCATCCTAAT**GTCGTCCG**CTTGCG-3'
T E R R P K L ..... P N V V R N W..... S * R N V S I L M S S A C

yox1 original sequence:
5'-CTCTGCGA**AACTT**CTACAGGACAGGCGG**AACTT**TTCTTC.....CTGTCGTAGGCGAAA-3'
L C E T S T G Q A E L F F ..... S V V G E

yox1$^{MTD\text{-}AACTT}$ sequence:
5'-CTCTGCGA**AACTT**CTACAGGACAGGCGG**AACTT**CTACAGGACAGGCGG**AACTT**TTCTTC.....CTGTCGTAGGCGAAA-3'
L C E T S T G Q A E L L Q D R R N F S S..... L S * A K

lsk1 original sequence:
5'-ATAAGGC**TGG**AGCAGGAAAAGGA**TGG**TTTCCCTATAACTACTGTCCGTGAAGTCAA-3'
I R L E Q E K D G F P I T T V R E V K

lsk1$^{MTD\text{-}TGG}$ sequence:
5'-ATAAGGC**TGG**AGCAGGAAAAGGA**TGG**AGCAGGAAAAGGA**TGG**TTTCCCTATAACTACTGTCCGTGAAGTCAA-3'
I R L E Q E K D G A G K G W F P Y N Y C P * S Q

ubp8 original sequence:
5'-CTTAAATC**TTATGTGG**AAACTCAGCAAATCT**TTATGTGG**TTACTCCC-3'

ubp8$^{MTD\text{-}TTATGTGG}$ sequence:
5'-CTTAAATC**TTATGTGG**AAACTCAGCAAATCT**TTATGTGG**AAACTCAGCAAATCT**TTATGTGG**TTACTCCC-3'

gcn5 original sequence:
5'-TATGGTTC**TCATTTAA**TGAA**TCATTTAA**AGGATTAT-3'

gcn5$^{MTD\text{-}TCATTTAA}$ sequence:
5'-TATGGTTC**TCATTTAA**TGAA**TCATTTAA**TGAA**TCATTTAA**AGGATTAT-3'

gcn5 original sequence:
5'-AGGACGAA**CAA**TC**CAA**TC**CAA**ATGTTCGCG-3'

gcn5$^{ins\text{-}CAA}$ sequence:
5'-AGGACGAA**CAA**TC**CAA**TC**CAA**TC**CAA**ATGTTCGCG-3'

Fig. S1. Identified highly reversible MTD mutations. (A). Genetic linkage test for isolated reversible mutants in rapamycin plus caffeine screen. The ratio of resistant to sensitive progenies(R:S) is scored. The resistant progeny is labeled with red ellipses in the image panel above the table. And the statistic number in the brackets showed the pairs meeting the indicated R:S ratio/the total calculated pairs. **(B).** Tandem duplication in multiple sites results in frame shift and pre-mature stop codon.

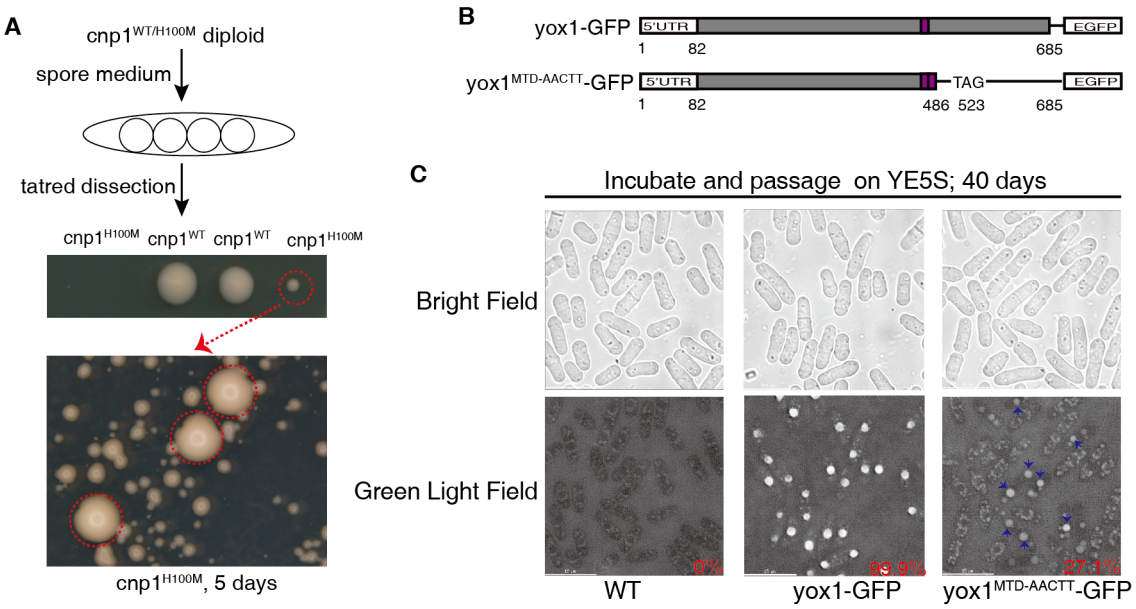
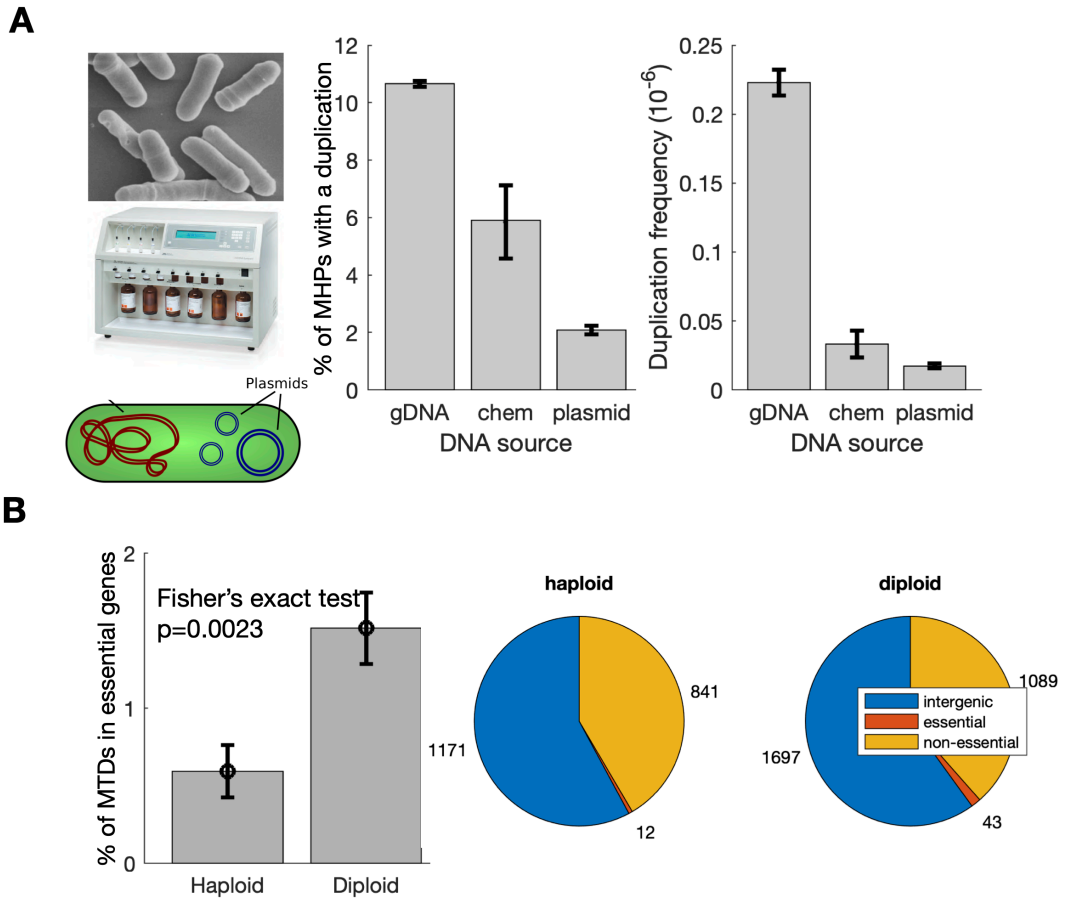


Fig. S2. Identification of MTDs in *cnp1^{H100M}* suppressor screen. (A). Process to isolate suppressors rescuing severe growth defect of *cnp1^{H100M}*. Suppressors occurred after 5 days cultivation of cnp1^{H100M} mini-clones on YE5S plate, and marked with red dotted circle. **(B)**. Construction of yox1-GFP and yox1^{MTD}-GFP strains. "TAG" is the premature stop codon. **(C)**. Genetic instability of yox1^{MTD} mutation is verified by fusing a GFP fluorescence marker. The blue arrows point recovered GFP signal, and the percentage marked with red shows the rate of cells with GFP signal.



674

675 **Fig. S3. MTDs are more commonly observed in genomic DNA and subclonal MTDs in**
676 **essential genes are more common in diploids. (A)** The *ssp1* gene was cloned into a plasmid
677 in E. coli, and the gene was amplified by PCR from either *S. pombe* genomic DNA or
678 miniprepped plasmid, or 200nt or 500nt chemically synthesized fragments, and all PCR
679 amplicons were sequenced together to similar sequencing depths (10^5-10^6x coverage). Shown
680 are the % of MHPs in *ssp1* in which a duplication was observed, as well as the measured
681 duplication frequency (reads per 10^6 coverage at that position). **(B)** Shown are the % of
682 observed MTDs that are fully contained within essential genes in haploid or diploid lines of
683 budding yeast, as well as the distribution of MTDs throughout the genome. Reads from each
684 haploid or diploid mutation accumulation line were mapped and analyzed independently, and
685 the results merged.

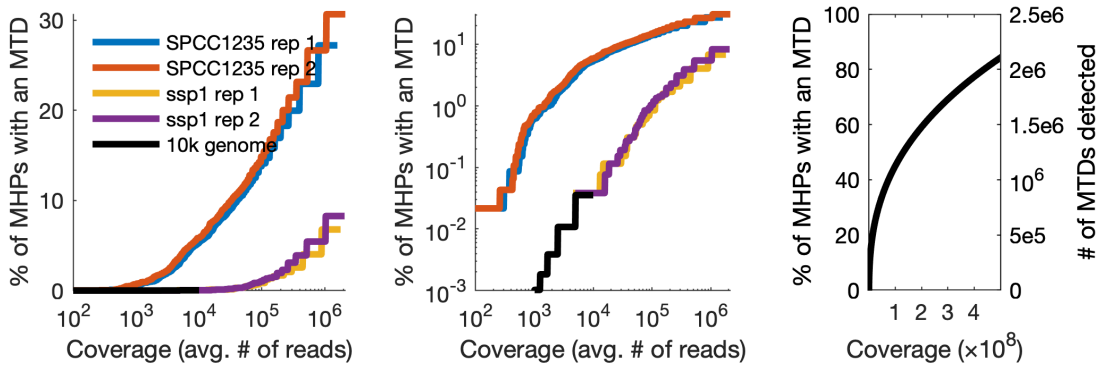


Fig. S4. The measured (left & middle) and estimated (right) sequencing coverage required to observe all of the possible MTDs in the genome. Shown are the % of MHPs with an observed MTD in ultra-deep amplicon sequencing (single genes) and for 10k whole-genome sequencing (black) as a function of the sequencing coverage. SPCC1235 is a hot gene; the same coverage results in far more observed MTDs, while ssp1 is more representative of the genome as a whole. The far right shows simulated data where the 10k data + ssp1 line is extended out to 10^8 coverage.

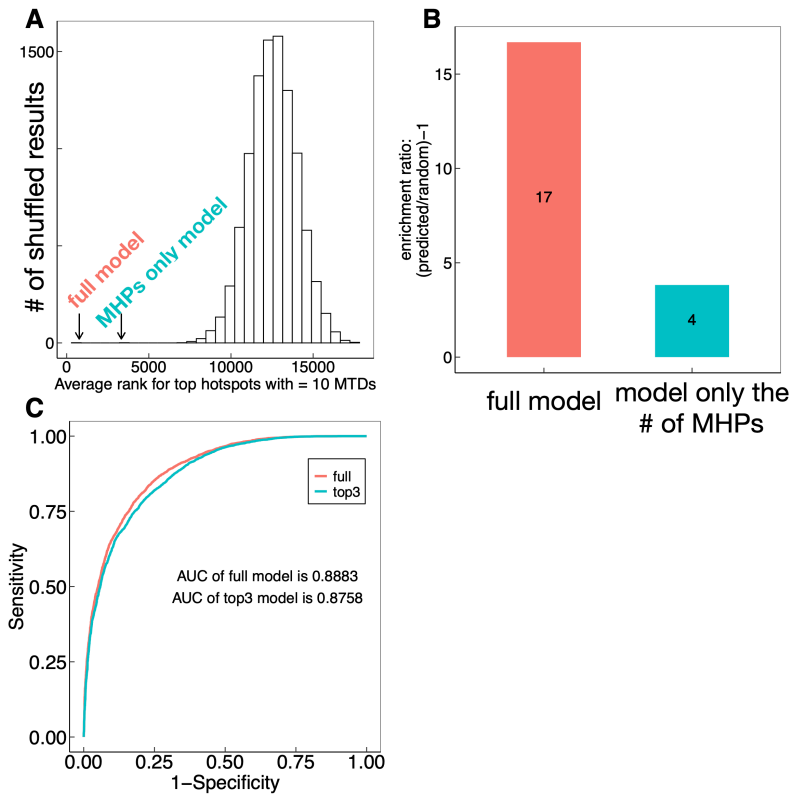


Fig. S5. Characterization of the logistic regression model for predicting MTDs and hot-spots from *cis* MHP features. (A,B) Hotspots were defined as 1kb windows with more than 10 observed MTDs in the 10k whole-genome sequencing data. To determine if hotspots are solely due to MHP density, or are due to other sequence features incorporated into the model, we generate a random background distribution (histogram, white bars). The observed MTDs were shuffled across all MHPs in the genome, and the 1kb windows were ranked by the number of MTDs contained within each window (rank=1 has the most MTDs), and the average rank of the top windows was calculated. The classification model was then used to predict hotspots using all features, or only by counting MHPs. **(C)** Receiver Operating Characteristic (ROC) curve for models with all features, or with only GC content, inter-MH-distance, and MH length. the number of MHPs in each 1kb window. The full classification model outperforms the MHP count; hotspots are determined by more than just MHP density.

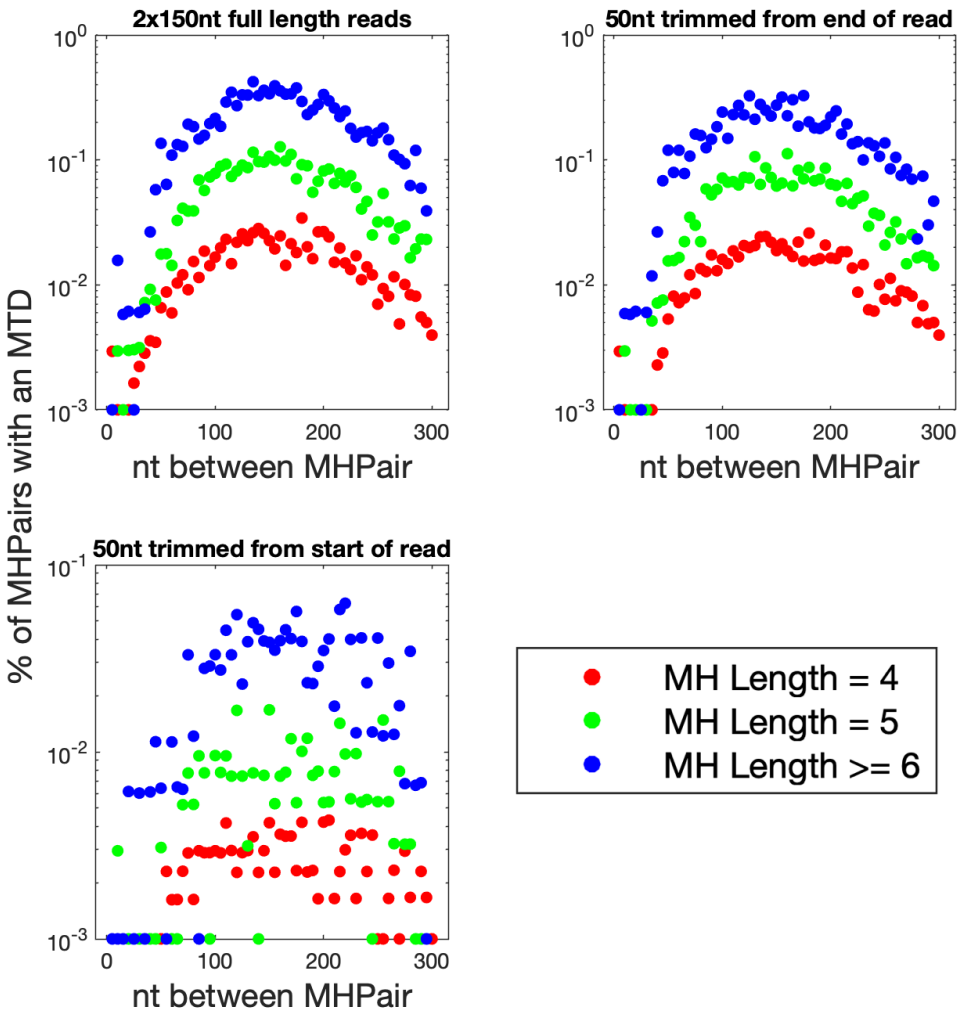


Fig. S6. The relation between inter-MH spacing and MTD frequency is independent of read length. Trimmomatic(Bolger et al., 2014) was used to remove either the first 50nt or the last 50nt from the end of each read, resulting in 2x100nt reads instead of 2x150nt reads; the peak at 150 remains unchanged. The higher noise when removing 50nt from the start is due to fewer identified MTDs, likely due to the higher error rate at the end of the read combined with the requirement for a perfect match to the MTD signature.

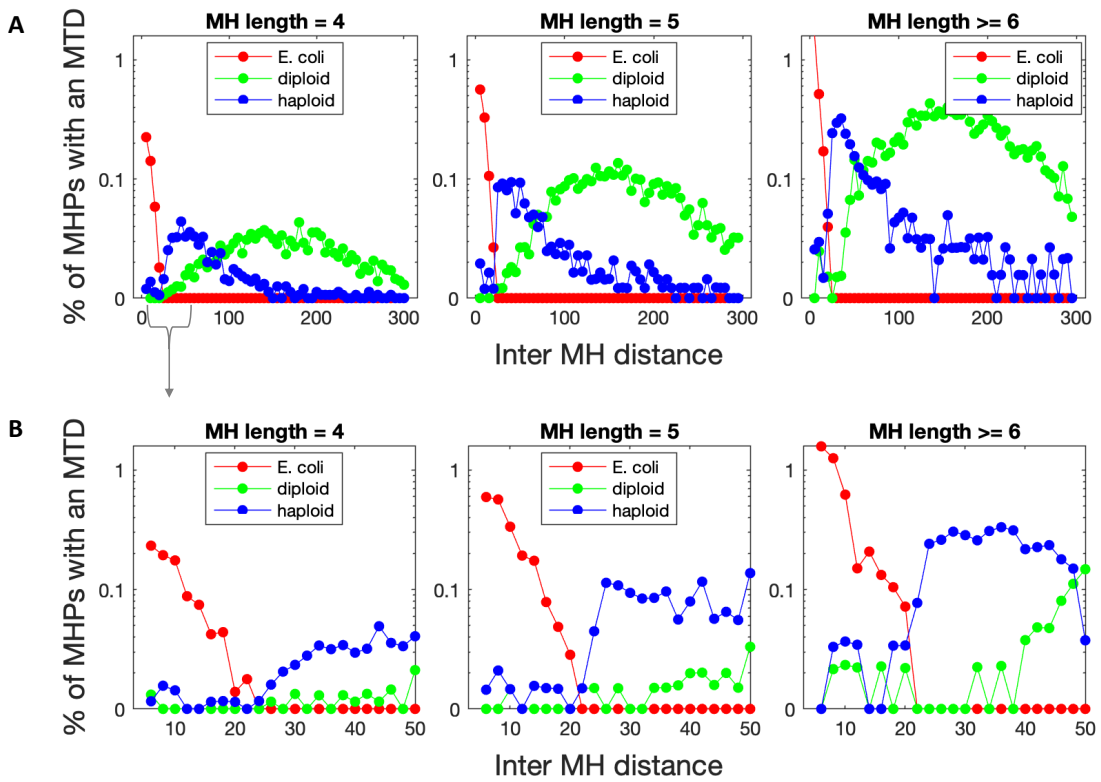


Fig. S7. Characterization the relation between MH sequence length, inter-MH distance, and observed MTD frequency across different ultra-deep whole-genome sequencing datasets. (A) The relation between MTD frequency and inter-MH distance are shown for diploid *S. pombe* (green, this study), an isogenic haploid *S. pombe* (SRR7817502, 1700x coverage, blue), and E. coli (PRJNA329347 , 14000x coverage, red). We note that the shorter haploid *S. pombe* (blue) inter-MH distance distribution is more similar to the insert lengths found in genetic screens, all of which were done in haploid strains. **(B)** Same data as in (A), but only inter-MH distances 3-50nt are shown.

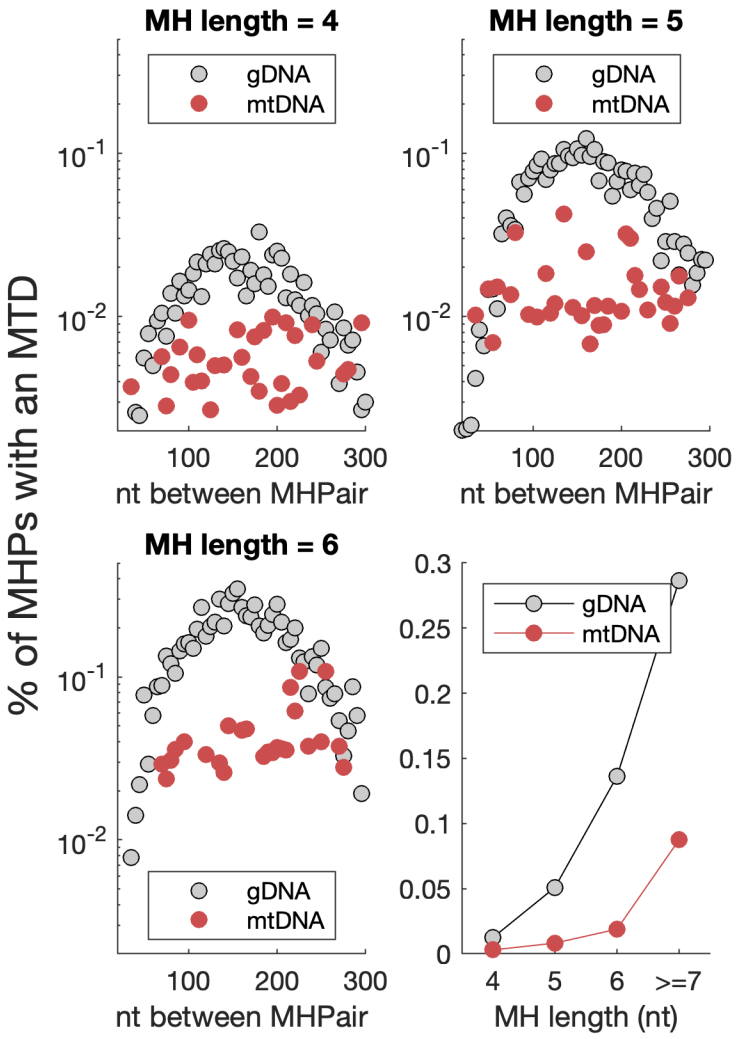


Fig. S8. MTDs are less common in the mitochondria, and do not exhibit a peak at 150nt. MTDs in the mitochondrial DNA were downsampled so that the median sequencing coverage was identical to that of the gDNA. Downsampling was repeated 5000 times to increase the statistical power.