Rapidly reversible mutations generate subclonal genetic diversity and reversible drug resistance

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 a single clonal population, and determine using machine learning 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(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). Pre-existing tandem repeats (satellite DNA) undergo frequent expansion and contraction (4–6), but repeats are rare inside of coding sequences and most functional elements. Chromatin-based epigenetic states have been associated with transient drug resistance in cancer cells(7, 8), and transiently resistant states have been characterized by differences in organelle state, growth rate, and gene expression in budding yeast(9, 10). In bacteria, copy-number gain and subsequent loss can result in transient antibiotic resistance(11). 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(12–14), and it is challenging to distinguish most types of low-frequency mutations from sequencing errors(15), 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 caused 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*(16)) **(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*(17) 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(18).

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, Table S7)**. We postulate these Microhomology-mediated Tandem Duplications (MTDs)(19–21) 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 randomly selected drug-sensitive revertants of the MTD alleles had the wild-type *ssp1* sequence. Finally, *ssp1Δ* and $ssp1^{\mathrm{MTD}}$ strains are temperature sensitive. 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(22), and therefore 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* **(Fig. S1B, S2, Table S7)**. 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.

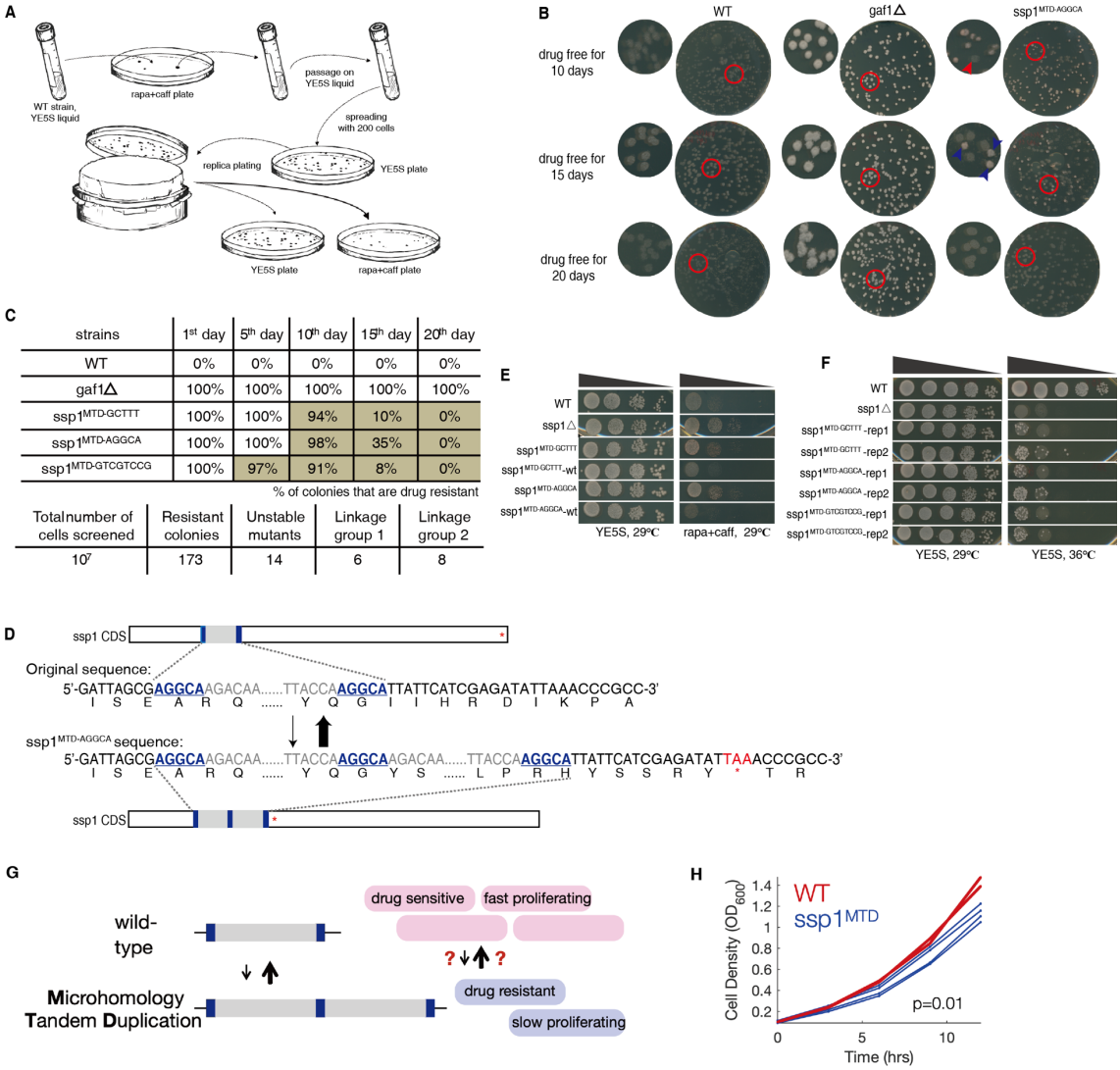

Figure 1. Screen for mutants with unstable inherited resistance by rapamycin plus caffeine and identify highly reversible mutations in *ssp1*. (A) Procedure to screen mutants with unstable rapa+caff resistance using sensitive wild-type strains in S. *p*ombe. **(B)** Unstable phenotype for one of screened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

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(12–14).

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 (23) 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. It is unclear why the plasmid-borne copy of *ssp1* lacks MTD but it 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(24). In comparison to the diploid, subclonal MTDs were depleted in essential genes in haploids (p=0.0023) **(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

mutations in *ura4*(25). Our results also emphasize that high-coverage sequencing is necessary to identify sufficient numbers of MTDs; one billion reads would be required to identify half of the 25 million possible MTDs in the *S. pombe* genome **(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(26) suppressors of $htb1^{G52D}$ identified MTDs in *gcn5*, as well as in *ubp8*, where we also observed an MTD in our high-coverage sequencing data **(Fig**. **S1B)**. 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.

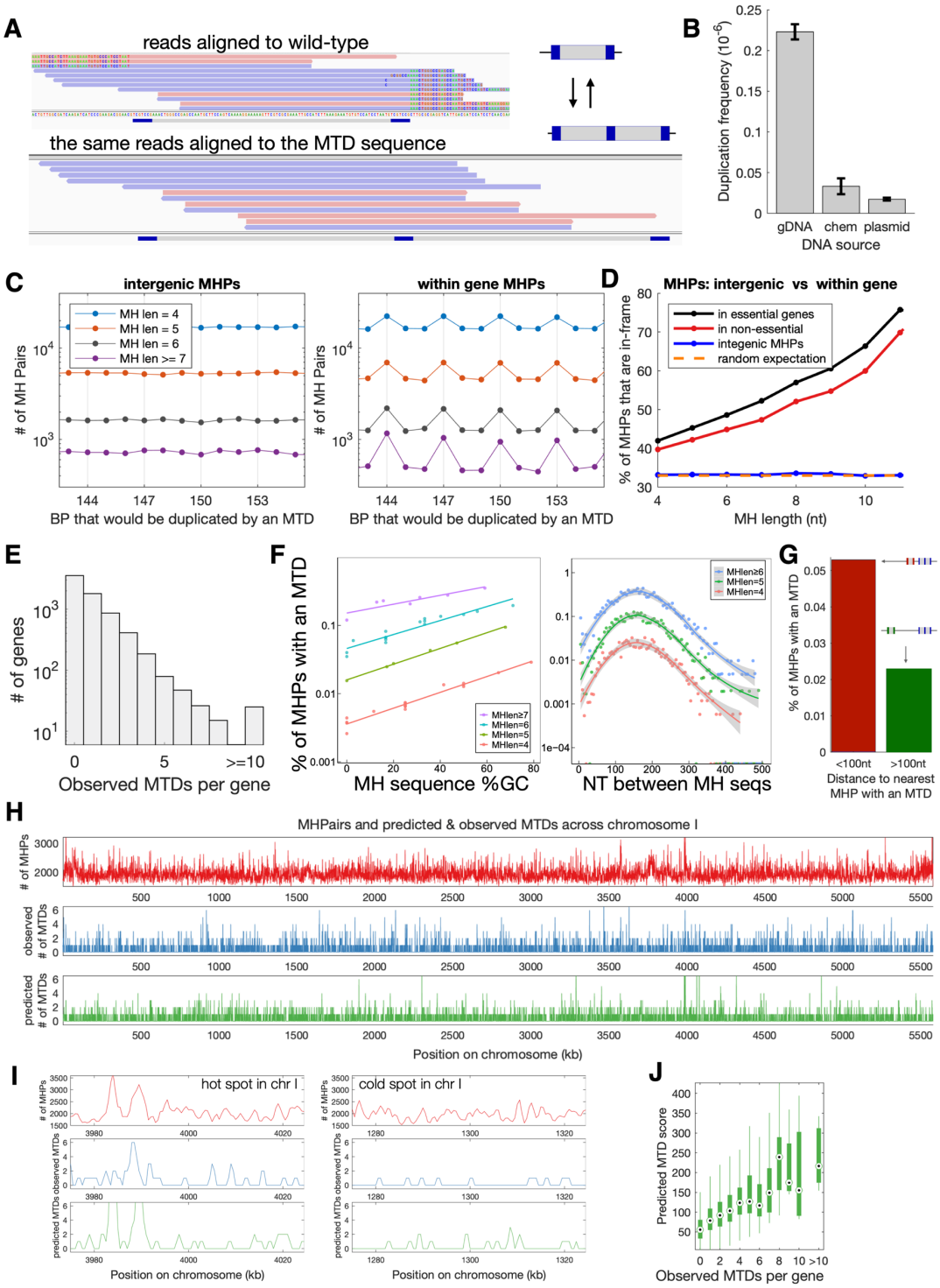

Figure 2. Identification of the *cis*-determinants of MTD through ultra-deep sequencing and identification of subclonal duplications. (A) The computational pipeline finds all sequencing reads that whose ends do not match the reference genome, and checks if the reads instead match the sequence that would 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. The cold spot has fewer MTDs after taking into account the number of MHPs, (Fisher's exact test, p=2.76e-09, odds ratio = 3.843). **(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.

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

238

239 Having established that local *cis*-encoded features determine the frequency with which
240 tandem duplications arise from microhomology-pairs, we next sought to identify the *trans*-
241 genes that affect MTD process. $ssp1^{MTD}$ alleles fail to grow at 36ºC, and their reversion back
242 to wild-type suppresses the temperature sensitivity, providing way to measure the effects of
243 mutations on reversion frequency. We screened a panel of 360 strains with mutations in DNA
244 replication, repair, recombination or chromatin organization genes for mutants that affect the
245 rate of $ssp1^{MTD}$ reversion back to wild-type **(Table S6)**, and found three mutants that
246 significantly increased and eight that significantly decreased the frequency of $ssp1^{WT}$ revertants
247 **(Fig. 3A,B,C)**.

248

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

257

258

259 Replication stresses activate a checkpoint that promotes DNA repair and recovery of stalled
260 or collapsed replication forks, and delays entry into mitosis(28, 29). The inactivation of
261 replication checkpoint kinase *cds1* or its regulator *mrc1* may thus result in a failure to restore
262 the replication fork, causing increased genome instability and MTD reversion. The replication
263 checkpoint is required for the stability of MTDs. Consistently, we found that deletion of the
264 DNA damage checkpoint kinase *cds1* or its regulator *mrc1* increased the frequency of $ssp1^{WT}$
265 revertants. Deletion of the single-stranded DNA binding A (RPA) subunit *ssb3* (RPA3/RFA3)
266 or the multifunctional 5'-flap endonuclease *rad2* also increased the frequency of revertants
267 **(Fig. 3C)**.

268

269 Many genes identified in the screen are multifunctional, and play roles in both replication
270 and repair. We therefore performed quantitative epistasis analysis to determine the relation
271 between six of the identified genes and the Mediator of the Replication Checkpoint, *mrc1*,
272 which interacts with and stabilizes Pol2 at stalled replication forks. In addition to the
273 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 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 we performed 10^6x coverage sequencing of the hotspot gene *SPCC1235.01*. We observe MTDs at fewer MHPairs, and an overall decrease in the number of MTDs in both mutants **(Fig. 3E,F)**.

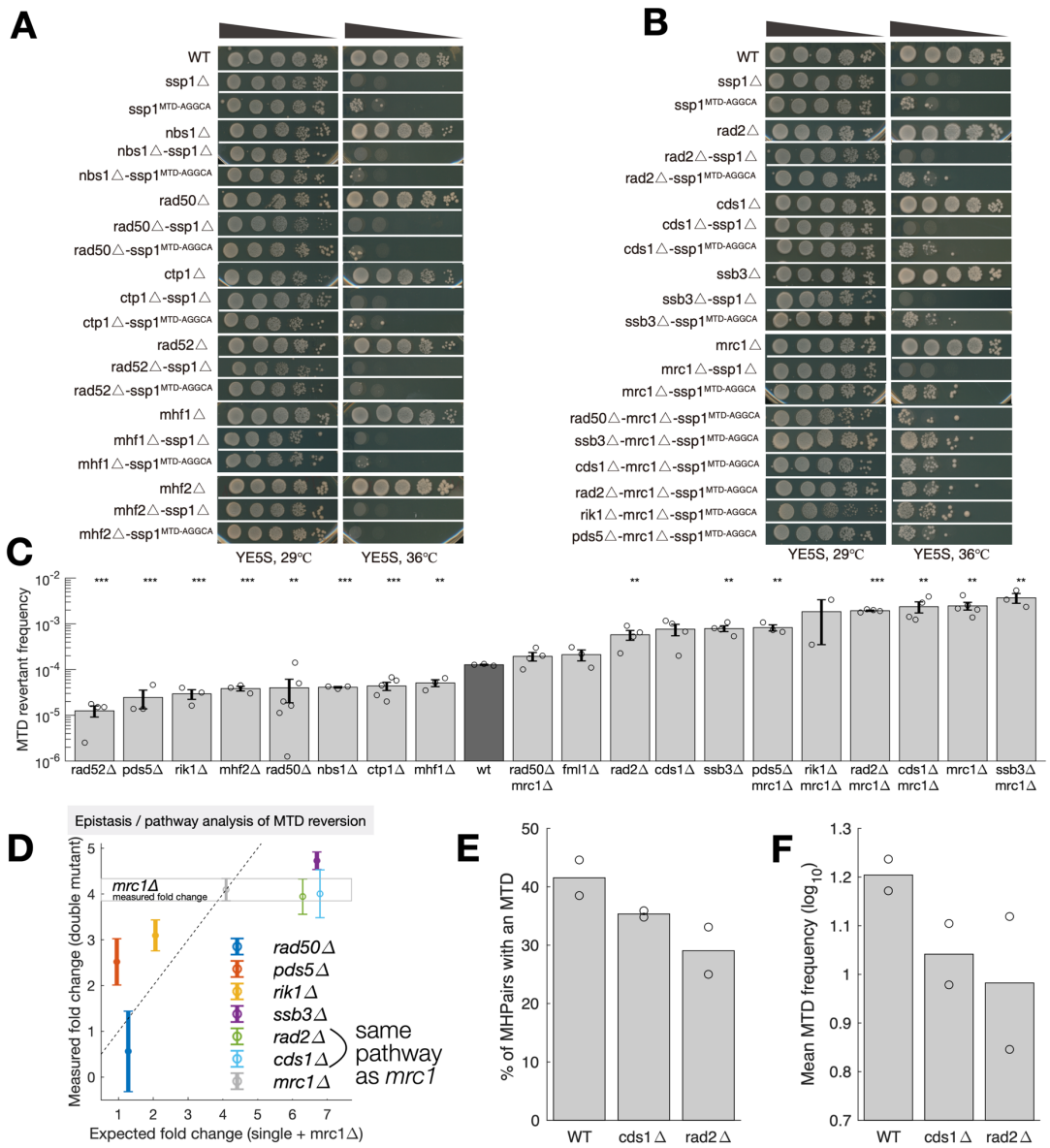

Figure 3. A genetic screen to identify the regulators of MTD reversion. (A,B). Surveyed mutants showed reduced ssp1^{MTD} reversion frequency represented by TS recovery phenotype. The non-TS phenotype of single mutation and *ssp1Δ* alone or combined with other mutants retained severe temperature sensitive 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 MHPs. **(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.

302 **Half of insertions and tandem duplications in natural isolates are MH-mediated**
303
304 It was baffling that MTDs are prevalent within populations, and that the first theoretical
305 proposal for microhomology-mediated processes in the generation of tandem duplications is
306 twenty years old(5), yet, relatively little is known about the forward process, and even less
307 about the reversion, suggesting that these events are not often encountered, or at least not
308 identified as such. To better understand the dynamics of MTDs within a population we used a
309 simple model of neutral mutations within a growing population that takes into account both
310 forward and reverse mutation rates and began with 100% of individuals as wild-type (see
311 Methods). The mutant frequency always increases, and over short timescales **(Fig. 4A**, left**)**
312 increasing the reverse rate from being equal to the forward mutation rate (grey) to being 10,000
313 times higher (yellow) has little effect.
314
315 Over longer timescales, high reversion rates cause the mutant frequency to plateau and
316 remain subclonal **(Fig. 4A**, right), reducing the fraction of neutral MTDs within a population.
317 However, in spite of the high reversion rate, both drift and selection enable fixation of MTDs
318 within a population. To identify fixed microhomology mediated insertions we searched the
319 genome sequences of 57 wild *S. pombe* isolates (30), and found that 50% of insertions larger
320 than 10bp involve microhomology repeats **(Fig 4B,C)**. Among these were 158 microhomology
321 mediated insertions that did not contain an obvious duplication, and 113 MTDs with a
322 microhomology mediated tandem duplication.
323
324 To test if the propensity of MTD formation within the lab strain is predictive of extant
325 sequence variation observed in natural isolates, we tested if the MTD score predicted for each
326 gene predicts the likelihood of microhomology mediated insertions in that gene. We found that
327 genes with microhomology mediated insertions in natural isolates tend to have higher predicted
328 MTD scores, and more experimentally observed MTDs **(Fig. 4D)**, suggesting that the local
329 features that affect MTD formation in the lab also shape evolution in nature.
330
331 Taken together, our results demonstrate that MTDs occur frequently and broadly throughout
332 the genome within a clonal population. This indicates that high levels of subclonal genetic
333 divergence may be prevalent but are under-detected using conventional sequencing approaches
334 that tend to disfavor the detection of low abundance subclonal variants. As many MTDs create
335 large insertions, they are more likely to be deleterious. Nonetheless, MTDs provide plasticity
336 to the genome and its functionality, for example, by allowing cells to become drug resistant,
337 while allowing the resistant cell lineage to revert back to wild-type and regain high fitness once
338 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[5], which are evolutionarily fixed in extant genomes and traditionally regarded as a major source of genome divergence. While previous work has shown that pre-existing repeats undergo rapidly reversible changes, the sequence-encoded rules regulating the birth and death of such sequences remained unknown. This work reveals that numerous sites throughout the genome have the potential of evolving into such repetitive elements.

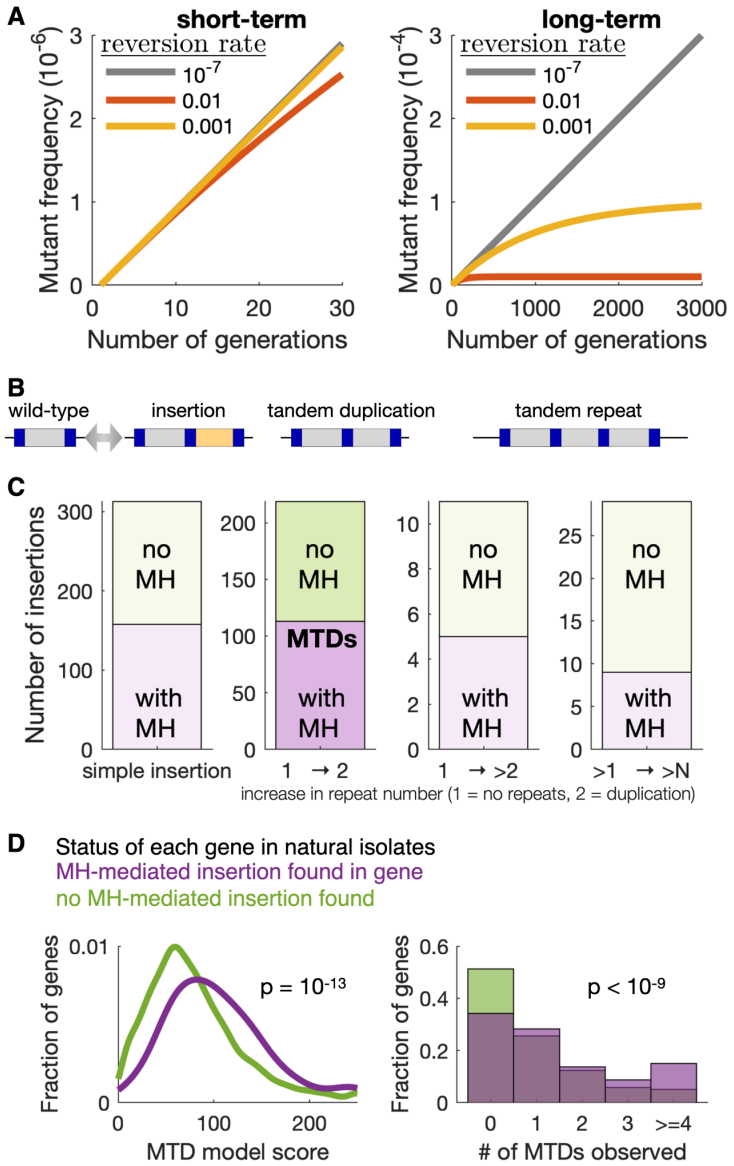

Figure 4. MTDs remain subclonal due to high reversion rates, yet half of insertions and de-novo tandem duplications in natural populations arise at microhomology sequence pairs. (A) Simulations showing the frequency of a neutral mutation (forward mutation rate = 10^{-7}) within a growing population at 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. **(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). The right-most column (>1x -> >Nx) 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.

365

366 **References and Notes:**

367 1. O. J. Rando, K. J. Verstrepen, Timescales of genetic and epigenetic inheritance. *Cell* **128**,
368 655–668 (2007).

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

371 3. R. E. Lenski, Bacterial evolution and the cost of antibiotic resistance. *Int. Microbiol.* **1**,
372 265–270 (1998).

373 4. R. Gemayel, M. D. Vinces, M. Legendre, K. J. Verstrepen, Variable tandem repeats
374 accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.* **44**, 445–477
375 (2010).

376 5. J. E. Haber, E. J. Louis, Minisatellite Origins in Yeast and Humans. *Genomics* **48**, 132–
377 135 (1998).

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

380 7. S. M. Shaffer, *et al.*, Rare cell variability and drug-induced reprogramming as a mode of
381 cancer drug resistance. *Nature* **546**, 431–435 (2017).

382 8. S. V. Sharma, *et al.*, A Chromatin-Mediated Reversible Drug-Tolerant State in Cancer
383 Cell Subpopulations. *Cell* **141**, 69–80 (2010).

384 9. R. Dhar, A. M. Missarova, B. Lehner, L. B. Carey, Single cell functional genomics
385 reveals the importance of mitochondria in cell-to-cell phenotypic variation. *Elife* **8**
386 (2019).

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

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

392 12. D. L. Hartl, E. W. Jones, *Genetics: principles and analysis*, 4th ed (Jones and Bartlett
393 Publishers, 1998).

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

14. 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).

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

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

17. 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).

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

19. 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).

20. 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).

21. N. A. Willis, *et al.*, Mechanism of tandem duplication formation in BRCA1-mutant cells. *Nature* **551**, 590–595 (2017).

22. A. Farlow, *et al.*, The Spontaneous Mutation Rate in the Fission Yeast Schizosaccharomyces pombe. *Genetics* **201**, 737–744 (2015).

23. S. R. Head, *et al.*, Library construction for next-generation sequencing: overviews and challenges. *BioTechniques* **56**, 61–64, 66, 68, passim (2014).

24. 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 USA* **115**, E5046–E5055 (2018).

25. S. Gangloff, *et al.*, Quiescence unveils a novel mutational force in fission yeast. *eLife* **6**, e27469 (2017).

26. 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* **8**, 1031–1038 (2018).

429 27. I. Iraqui, *et al.*, Recovery of arrested replication forks by homologous recombination is
430 error-prone. *PLoS Genet.* **8**, e1002976 (2012).

431 28. A. A. Alcasabas, *et al.*, Mrc1 transduces signals of DNA replication stress to activate
432 Rad53. *Nat. Cell Biol.* **3**, 958–965 (2001).

433 29. K. Myung, R. D. Kolodner, Suppression of genome instability by redundant S-phase
434 checkpoint pathways in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U.S.A.* **99**,
435 4500–4507 (2002).

436 30. D. C. Jeffares, *et al.*, The genomic and phenotypic diversity of Schizosaccharomyces
437 pombe. *Nat Genet* **47**, 235–241 (2015).

438 31. D.-U. Kim, *et al.*, Analysis of a genome-wide set of gene deletions in the fission yeast
439 Schizosaccharomyces pombe. *Nat. Biotechnol.* **28**, 617–623 (2010).

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

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

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

447 35. P. Cingolani, *et al.*, A program for annotating and predicting the effects of single
448 nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster
449 strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).

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

452 **Acknowledgments:** We thank Lilin Du, Aaron New, and Wenfeng Qian for insightful
453 discussions and for comments on the manuscript. We thank Qi Zhou for assistance with
454 rapamycin and caffeine resistance screen.

455 **Funding:** L.B.C. was supported by the Peking-Tsinghua Center for Life Sciences. X.H. was
456 supported by National 973 Plan for Basic Research Grant 2015CB910602 and National Natural
457 Science Foundation of China Grant 31628012.

458

459 **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 SRA BioProject accession PRJNA631756.

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(31) or constructed by overlap PCR strategy and gene-specific homologous recombination using standard procedures(32).

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 prepared by adding 100mg rapamycin to 1ml DMSO (100mg/ml) and diluting by 1000 folds. 1.942g power caffeine was dissolved in 60-80°C 20 ml sterile ddH_2O and added into 1L YE5S medium to final concentration 10mM.

Unstable drug-resistant mutants screen

A fresh single colony of wild-type cells was picked and grown to mid-log phase culture. Cultivated cells were then spread on YE5S agar plates containing 100ng/ml rapamycin and 10mM caffeine (hereafter called YE5S+drug plates) at the density of 1×10^5 cells per plate, and

incubated at 29°C for 10 days.. To test the stability of the drug resistance, each strain is grown continuously in YE5S liquid media in the absence of the drugs at 29°C by refreshing the culture with YE5S liquid media daily for up to 20 days. Every five days, cell samples were taken and spread to the YE5S plate at the density of 200 cells per plate. After 3 day incubation at 29°C, each plate was replica plated to fresh YE5S and YE5S+drugs plates, respectively, incubated for two days at 29°C. Plates were visually examined for colonies that grow on YE5S but fail to grow on YE5S+drugs plates. The stability test was repeated at least two times for identified unstable drug-resistant stains. The *gaf1-d* mutant was used as the control for stable and robust drug resistance.

Genetic linkage test

Identified unstable drug-resistant strains were backcrossed with wild-type cells or crossed with each other on the ME plate. After 24-48h sporulation at 29°C, tetrad-dissection was performed on the YE5S plate following the standard procedure(Escorcia and Forsburg, 2018). After 3 days incubation at 29°C, YE5S plates are replica plated to the YE5S+drug plate and incubated at 29°C for 2 days to identify drug-resistant colonies among the four progeny originated from one ascus. The segregation pattern of the drug-resistant and drug-sensitive phenotypes is analyzed and used to determine the genetic linkage of the tested mutation alleles.

Whole-genome sequencing and datasets analysis

Genomic DNA was extracted using phenol-chloroform, mechanically sheared to ~200bp using ultrasonicator. Sheared genome DNA was used to build the library using NEBNext® Ultra™ DNA Library Prep Kit for Illumina® (E7370/7335, NEB) and Illumina sequenced by Ribobio in Wuhan, China.

For data analysis, adapter-trimmed FASTQ clean data were mapped to the ENSEMBL Fungi's *S. pombe* genome version ASM294v2 with the BWA mem aligner(Li and Durbin, 2009) (version 0.7.17, with -M flag on). After removing PCR duplicates(Li et al., 2009), alignment maps (BAM files) were fed to the GATK's HaplotypeCaller for a first run. The output variants are used to recalibrate base quality scores in the BAM files using GATK's BaseRecalibrator. Recalibrated BAM files were then inputted to the HaplotypeCaller to generate raw mutation callings(McKenna et al., 2010), which were filtered and annotated using the ENSEMBL's variant effect predictor (VEP, version 93.3).

Double mutant construction and MTD reversion regulators survey

Double mutants which combine MTD mutation at *ssp1* ($ssp1^{MTD\text{-}AGGCA}$) and each deletion within the mutant panel were created by genetic crossing following the standard procedures

522 (Roguev et al., 2018; Schuldiner et al., 2006) using a high throughput robotic apparatus (Peking
523 university, F. Li lab. Protocol for high throughput manipulation is available upon request).
524 To assess MTD reversion rates semi-quantitatively, a single colony of each double mutant
525 was used to inoculate 3ml YE5S liquid culture, incubated at 29°C overnight, refreshed by 1:10
526 dilution in 20ml YE5S liquid medium, and grown to mid-log phase. Serial 1:10 dilutions of the
527 culture were prepared using fresh YE5S liquid medium, spotted on YE5S plates (5□l per spot,
528 corresponding to 10^5 to 10 cell per spot), incubated for 4-5 days at 36°C or 29°C.
529 To quantify the reversion frequency, double mutant cells were spread on YE5S plates at the
530 density of 2×10^4 or 2×10^5 cells per plate, incubated 4-5 days at 36°C. The number of non-ts
531 revertant colonies was scored in three biological repeats for each double mutant strains. Two
532 non-ts revertants were picked for each strain and the *ssp1* locus in these revertants is PCR
533 amplified/Sanger sequenced to verify true reversion of $ssp1^{MTD\text{-}AGGCA}$ to wild type.
534 We also did a fluctuation test for some double mutant strains to quantify the reversion
535 frequency by growing a single overnight culture in YE5S broth for the strain to be tested,
536 diluting with fresh YE5S broth to obtain 10^2 yeast cells /ml. For each strain, the diluted
537 suspension was divided into 48 of 100μL and incubated at 29°C. Then 40 replica cultures with
538 100μL were plated in their entirety onto YE5S agar plates and incubated at 36°C for 3~5 days.
539 For the rest 10 100μL replica cultures, the average number of cells per culture (N) was
540 calculated using a blood counting chamber. Then counted the number of 36°C survival cells
541 (reverted wt cells) per culture, and calculated the mutation rate with the p0 method or the MSS-
542 maximum likelihood method (33).
543
544 **Identification of $yox1^{MTD}$ and $lsk1^{MTD}$ among $cnp1^{H100M}$ suppressors and stability test for**
545 **$yox1^{MTD}$**
546 Haploid $cnp1^{H100M}$ cells derived from heterozygous $cnp1^{H100M}$ diploid by tetrad dissection were
547 spread on YE5S plates, incubated at 29ºC for 5 days. Rare large colonies (~$1/10^4$) were isolated
548 as spontaneous $cnp1^{H100M}$ suppressors (FigS2A). Whole-genome sequencing was performed
549 on isolated $cnp1^{H100M}$ suppressors to identify the target gene. With the analysis process in
550 "Whole-genome sequencing and datasets analysis" part, MTD events in *yox1* and *lsk1* gene
551 were identified and verified by Sanger sequencing.
552 To verify the genetic stability of $yox1^{MTD}$ alleles, $cnp1^{H100M}$ suppressors were backcrossed
553 with wild type, $yox1^{MTD}$ were separated from $cnp1^{H100M}$ mutation. *yox1*-GFP, $yox1^{MTD}$-GFP
554 strains were constructed by fusing a GFP tag in the endogenous *yox1* locus (FigS2B). MTD
555 (20bp tandem duplication) in *yox1* disrupts the open reading frame and generates a premature
556 stop codon (TAG) at 523nt loci, resulting in inactivation of GFP fluorescence, while the
557 reversion of $yox1^{MTD}$ would recover the GFP fluorescence. In the stability test, *yox1*-GFP and
558 $yox1^{MTD}$-GFP cells were grown continuously at 29°C by refreshing the culture with YE5S

559 liquid media daily for up to 60 days. Every ten days, cell samples were taken and subjected to
560 microscopical observation for GFP fluorescence. The percentage of progenies exhibiting the
561 nuclear GFP signal was scored in three individual biological repeats. To verify *yox1*^{MTD}-GFP
562 reversion, *yox1* locus of @ single colonies derived from *yox1*^{MTD}-GFP 40 day culture was
563 amplified by PCR and subjected for Sanger sequencing
564
565
566 **Finding microhomology pairs on genome**
567 A fast algorithm is implemented to find micro-homology pairs across the S. pombe's genome
568 sequence (or any given DNA sequence). First, the input sequence is scanned one-time for initial
569 k-mer homology pairs with pre-set limitations. Here we arbitrarily set limitations to 1) the size
570 of the homology should be no smaller than 4 bps and no greater than 12 bps, 2) the homology
571 should not be a mononucleotide repeat, 3) space between two homologies in a pair should be
572 greater than 3 bps, and 4) the INDEL size (the length of a homology plus the inter-space) should
573 not exceed 100 bps. Then, the initial homology pairs are forth scanned for one run to merge
574 adjacent homology pairs to longer pairs. The current implementation would only report the
575 left-most pair of tandem repeats with micro-homology pairs on repeat junctions
576
577
578 **Annotating insertions and tandem repeats flanked by micro-homology pairs in natural**
579 **isolates and in the reference genome.** To identify MH-flanked tandem repeats in the reference
580 genome we used the Tandem Repeat Finder(34) to generate an initial tandem repeat candidate
581 list. All parameters were set to the default value except the INDEL penalty, which was set to
582 1000 to avoid reporting tandem repeats with non-uniform unit sizes. After removing candidates
583 with the reported unit size smaller than 10nt, self-information smaller than 1.5 bits, and repeat
584 number smaller than 2, remaining tandem repeats were verified by three steps: 1) finding if
585 there were still internal repeats within the reported repeat unit, 2) finding if there were still
586 repeat units on the left and right wings to the reported length, and 3) sliding the whole frame
587 to the left-most base while the repeats' consistency did not drop. Finally, we checked the
588 junctions for the existence of a micro-homology of at least 2nt. If homology size is long enough
589 (longer than 75% of the unit size and longer than unit size–4 bps), we considered it as repeat
590 number plus 1 and start over for finding junction micro-homologies.
591 To identify MH-flanked insertions in natural isolates we used the indels .vcf file from Jeffares
592 et al.(30) and used SnpEff (35) to predict the impact of each indel. We extracted the left and
593 right flanking sequences from the reference genome to determine the presence of
594 microhomology and to identify the repeat unit.
595

596

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

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

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

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

619

620 **Simulations of mutation frequency at different reversion rates**

621 Let A_{wt} be the wild-type allele, and A_{mut} be the mutant allele. Let k_{fwd} be the forward (A_{wt} to
622 A_{mut}) mutation rate, and k_{rev} be the reverse (A_{mut} to A_{wt}) mutation rate. Let p_t be the frequency
623 of A_{wt} and q_t the frequency of A_{mut} at time t. Then, if we assume that mutations are neutral, the
624 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
625 $k_{fwd} >= k_{rev}$ or, as is the case for subclonal MTDs, when q is small, the reverse mutation can
626 mostly be ignored. However, when $k_{rev} >> k_{fwd}$ or $q == 1$, as is the case for clonal fixed MTDs,
627 k_{rev} has a large impact on dynamics. For simulations, the initial conditions were set to p=1,q=0,
628 $k_{fwd} = 10^{-7}$, and k_{rev} was varied as is shown in the figure.

629

630 **Logistic regression to predict MTD frequency from local features**

631 To predict the likelihood of a duplication event in each micro-homology pair (MHP), we used
632 a logistic regression model (the function glm() from R) with 10-fold cross-validation. The data

are highly imbalanced; MTDs were detected at fewer than 0.1% of MHPs. We therefore trained and tested the model using a balanced dataset consisting of all MHPs with an MTD, plus a randomly chosen subset MHPs with no MTD of the same size, so that half of MHPs had an MTD. We first trained a model using three features: MHlength, GC-content-of-the-MH-sequence, and inter-MH-distance, which has an AUC of 0.876. This is the "top 3 features" model, and all three of these features are predictive by visual inspection (e.g., Fig 2). To determine which additional features to add we continuously added features, and kept only those that increased the AUC over this base 3-feature model. The additional predictive features were: MHPlength (MHlen), nucleotides between two repeats (interMH), interMH (interGCcon), nucleosome occupancy (entire_nucle) and gene expression (entire_gene) of the entire MHP, and nucleotides to the closest MHR which has duplication event(ntclosestMHR).

To perform whole-genome predictions using the model trained on the balanced data, we used the model to score all 25 million MHPs in the genome, and either used the sum of predicted scores for all MHPs in a single gene, or selected the top 6234 MHPs, the same number of duplication events as observed experimentally, to be predicted duplication events.

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'

Figure 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.

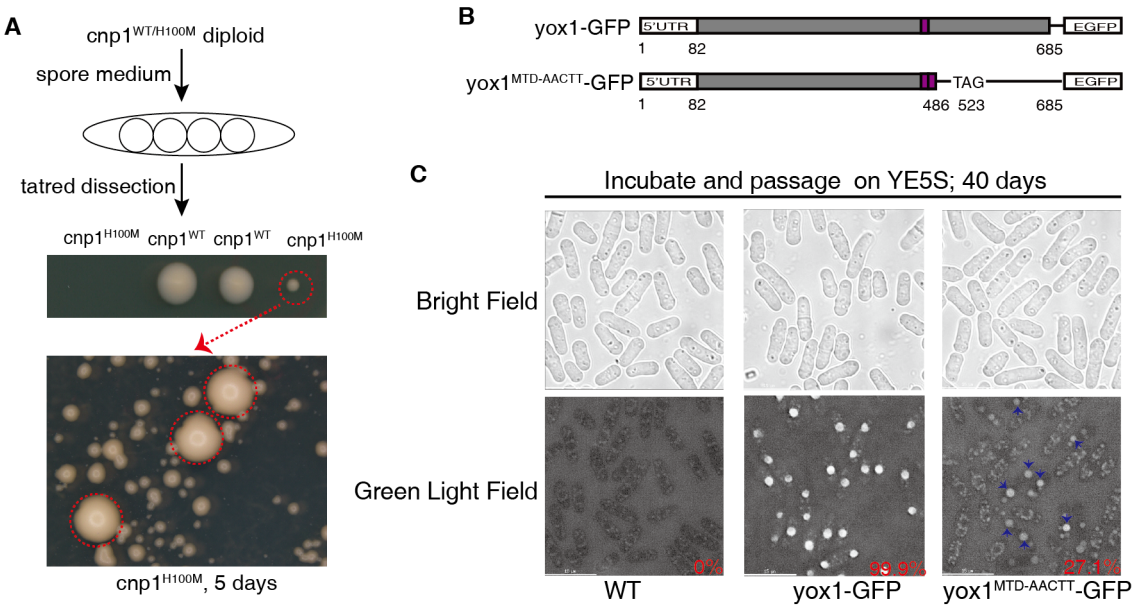

Figure S2. Identification of MTDs in *cnp1^{H100M}* suppressor screen. (A). Process to isolate suppressors rescuing severe growth defect of *cnp1^{H100M}*. Suppressors occurred after 5 days cultivation of cnp1^{H100M} mini-clones on YE5S plate, and marked with red dotted circle. **(B)**. Construction 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.

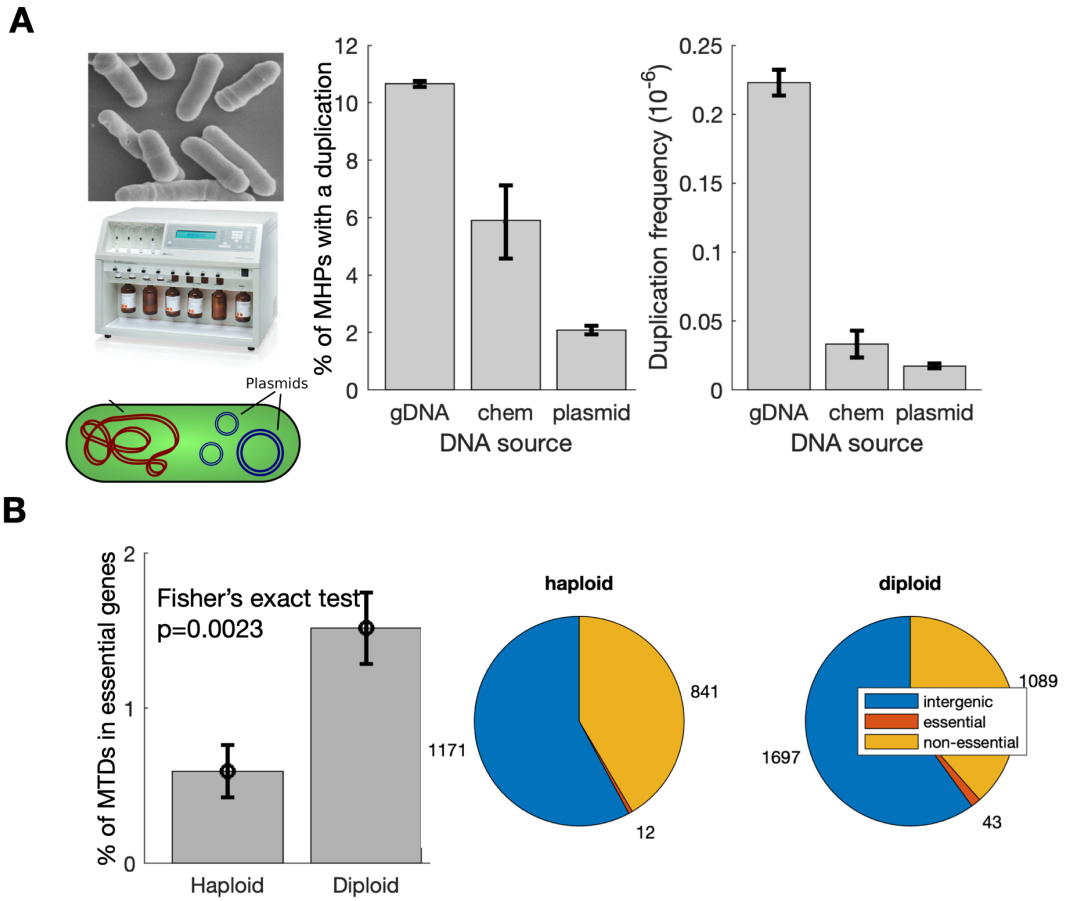

Figure S3. MTDs are more commonly observed in genomic DNA and subclonal MTDs in essential genes are more common in diploids in S. cerevisiae. (A) The *ssp1* gene was cloned into a plasmid in E. coli, and the gene was amplified by PCR from either *S. pombe* genomic DNA or miniprepped plasmid, or 200nt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

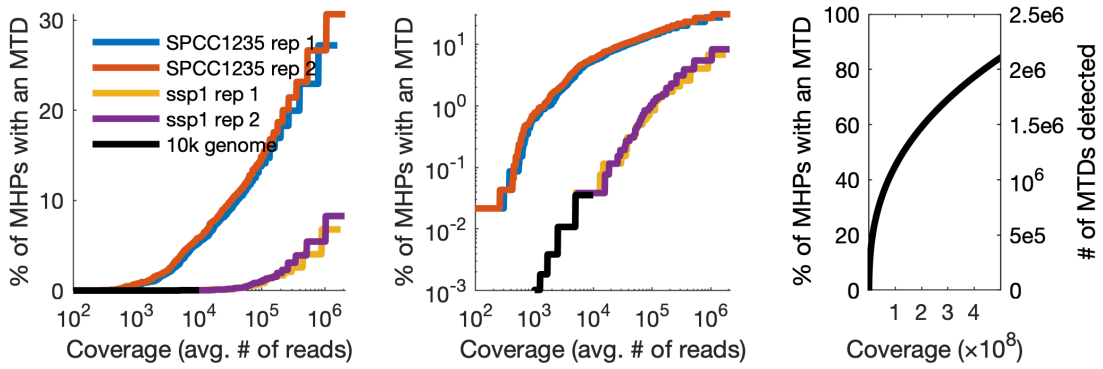

Figure S4. The measured (left & middle) and estimated (right) sequencing coverage required to observe all of the possible MTDs in the genome. Shown are the % of MHPs with an observed MTD in ultra-deep amplicon sequencing (single genes) and for 10k whole-genome sequencing (black) as a function of the 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.

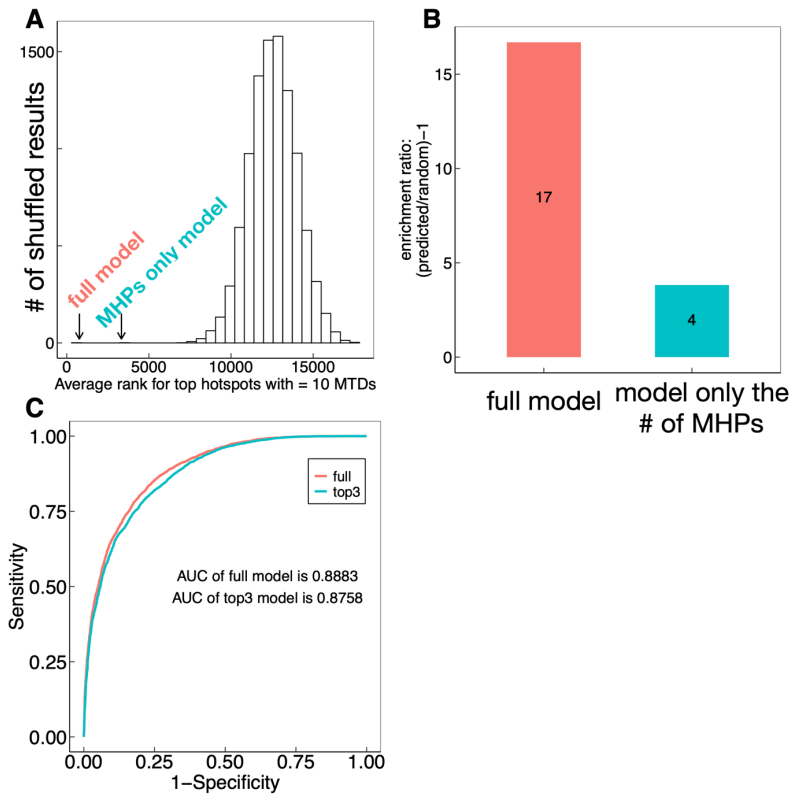

Figure S5. Characterization of the logistic regression model for predicting MTDs and hot-spots from *cis* MHP features. (A,B) Hotspots were defined as 1kb windows with more than 10 observed MTDs in the 10k whole-genome sequencing data. To determine if hotspots are solely due to MHP density, or are due 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.

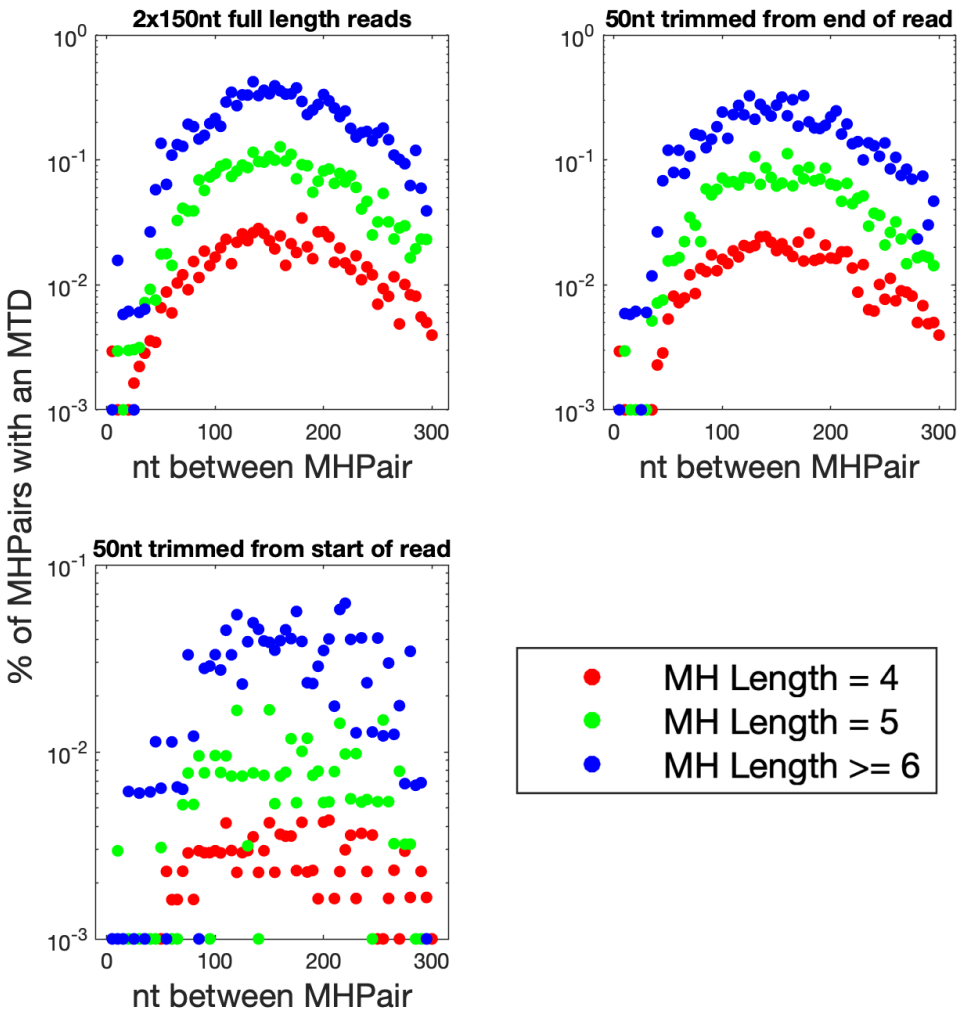

Figure S6. The relation between inter-MH spacing and MTD frequency is independent of read length. Trimmomatic(36) was used to remove either the first 50nt or the last 50nt from the end of each read, resulting in 2x100nt reads instead of 2x150nt reads; the peak at 150 remains unchanged. The higher noise 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.

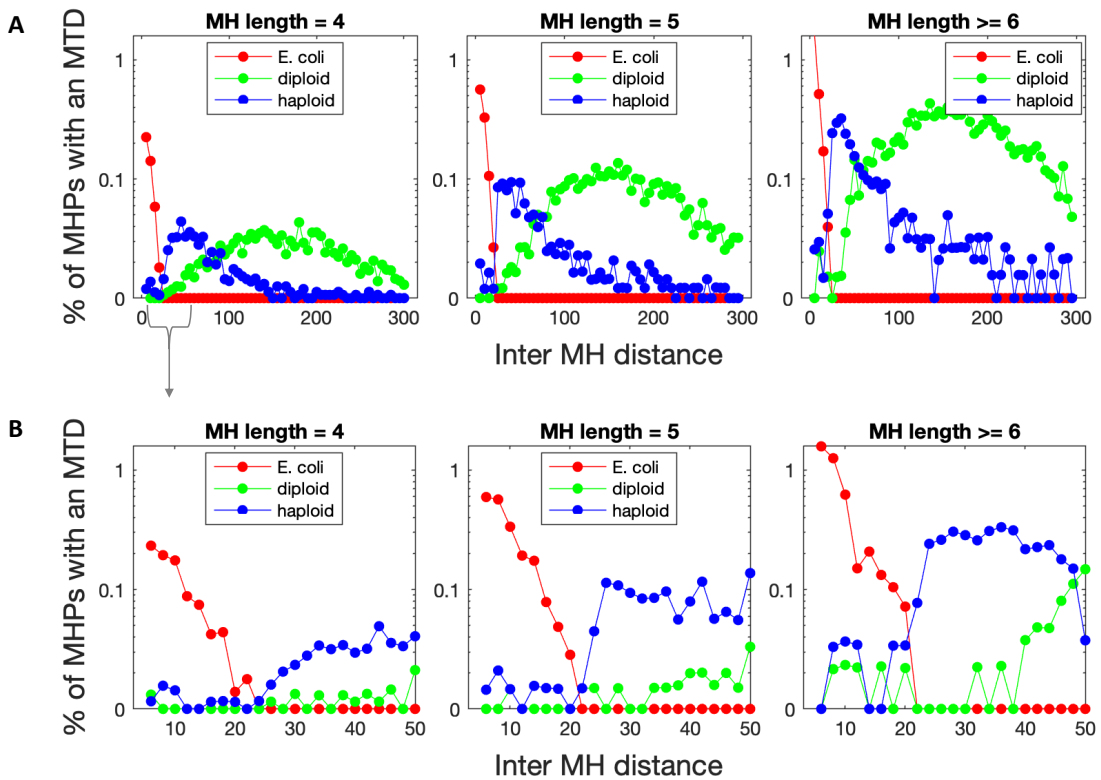

Figure S7. Characterization the relation between MH sequence length, inter-MH distance, and observed MTD frequency across different ultra-deep whole-genome sequencing datasets. (A) The relation between MTD frequency and inter-MH distance are shown for diploid *S. pombe* (green, this study), an isogenic 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.

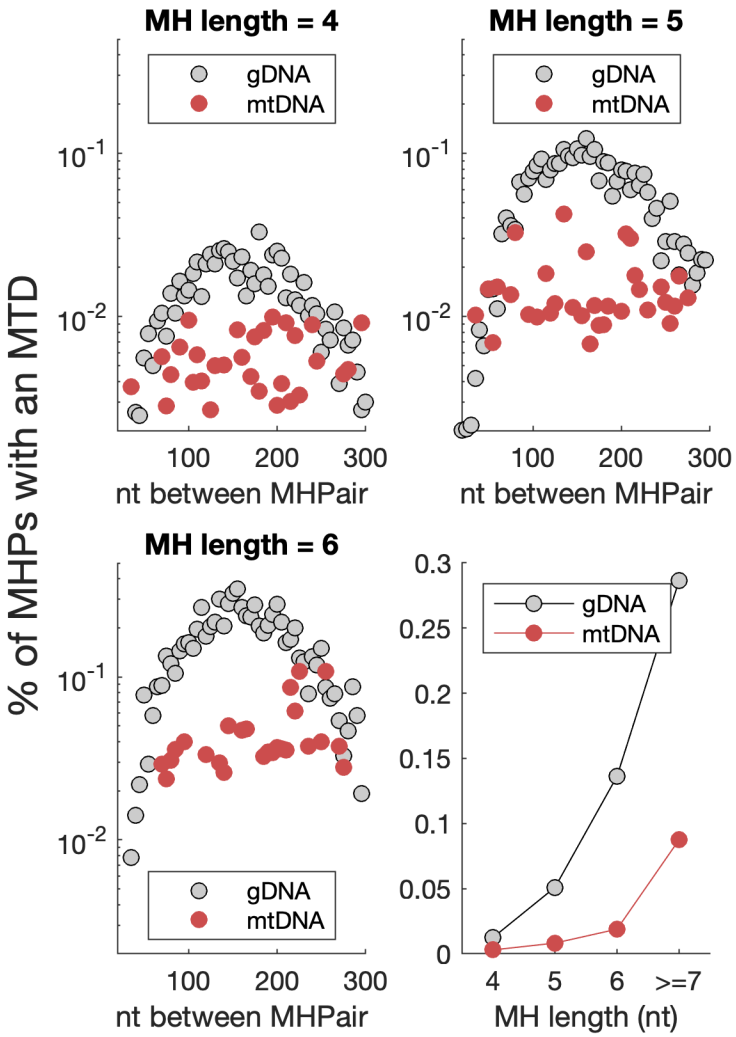

Figure S8. MTDs are less common in the mitochondria, and do not exhibit a peak at 150nt. MTDs in the mitochondrial DNA were downsampled so that the median sequencing coverage was identical to that of the gDNA. Downsampling was repeated 5000 times to increase the statistical power.