scientific reports

Check for updates

OPEN

Protein innovation through template switching in the *Saccharomyces cerevisiae* lineage

May Abraham & Einat Hazkani-Covo

DNA polymerase template switching between short, non-identical inverted repeats (IRs) is a genetic mechanism that leads to the homogenization of IR arms and to IR spacer inversion, which cause multinucleotide mutations (MNMs). It is unknown if and how template switching affects gene evolution. In this study, we performed a phylogenetic analysis to determine the effect of template switching between IR arms on coding DNA of *Saccharomyces cerevisiae*. To achieve this, perfect IRs that co-occurred with MNMs between a strain and its parental node were identified in *S. cerevisiae* strains. We determined that template switching introduced MNMs into 39 protein-coding genes through *S. cerevisiae* evolution, resulting in both arm homogenization and inversion of the IR spacer. These events in turn resulted in nonsynonymous substitutions and up to five neighboring amino acid replacements in a single gene. The study demonstrates that template switching is a powerful generator of multiple substitutions within codons. Additionally, some template switching events occurred more than once during *S. cerevisiae* evolution. Our findings suggest that template switching constitutes a general mutagenic mechanism that results in both nonsynonymous substitutions and parallel evolution, which are traditionally considered as evidence for positive selection, without the need for adaptive explanations.

Inverted repeats (IRs) are sequences with two copies of a DNA sequence in a reverse-complement orientation (e.g., 5′ATGTGxxxxCACAT 3′). IRs include internal symmetry, enabling them to switch between inter-strand and intra-strand base-pairing, resulting in non-canonical DNA structures such as cruciforms and hairpins. Long IRs lead to genome instability[1,2], either because they are processed to create a double-strand break or because they block the replication fork[1–6]. The resulting genomic instability events are diverse, and include gene amplification[7–9], translocations[10–13], insertions[14], and deletions[2]. Despite the potential of IRs to destabilize genomes, short IRs have several functions in organisms throughout the tree of life, e.g., IRs found in promoters enable the binding of homodimer transcription factors[15]. IRs also have a functional role in the viral origin of replications[16], the CRISPR immune system[17], alternative termination of bacterial genes[18], and immunoglobulin V(D)J rearrangement[19].

A striking characteristic of IRs is their tendency to undergo homogenization, abolishing variation between the two arms of the IR. Numerous examples have been reported of long-IR[20–25] and short-IR homogenization[26–28]. From a mechanistic standpoint, DNA polymerase template switching can eliminate variations as short as a few bases between IR arms[28]. This occurs when one arm of the IR serves as a template for the synthesis of the second arm. Template switching, first suggested by Ripley (1982) in bacteriophages, requires two hops of DNA polymerase between templates and can occur either intramolecularly or intermolecularly[26,29–33].

Since its identification, template switching has been reported throughout the tree of life[33–37]. Short-IR homogenization via template switching is known to occur in numerous organisms[26,27,35,36] and is associated with mutational hotspots[27,34]. The extent of template switching is affected by the directionality of the replication fork[38–40], the level of transcription[40], and the local sequence context[41]. There are several genotypes in which template switching is more common. For example, we have previously shown that template switching occurs in yeast strains lacking Rad27p, a key player in Okazaki fragment maturation[42].

While the evolutionary consequences of homogenization of long IRs are well-studied[21,43–47], the impact of short IR homogenization on genome evolution has received little attention. Our previous evolutionary analysis

Department of Natural and Life Sciences, The Open University of Israel, Ra'anana, Israel. email: einatco@openu.ac.il

of non-coding regions with short IRs in proteobacteria orthologs[48] indicated that these regions are more conserved than their immediate surrounding. This suggests that repeated template switching between IR arms is common during the evolution of proteobacteria. Template switching between IR arms was also recently shown to be abundant in humans[49,50].

To date there have only been a few reports of template switching in genes and these have usually been considered in their mutagenic context. Template switching was identified with the context of loss-of-function in T4 rII gene in T4[51], *E. coli* rpsL[52], and thyA[34] genes, as well as in the *S. cerevisiae CYC1*[35] gene. It was suggested that template switching contributes to the spectrum of mutations that affect the TP53 gene in human cancers[36]. Template switching was also shown to be involved in certain mutagenic processes that lead to several genetic diseases, such as hereditary angioneurotic edema[37]. In our previous analysis of *rad27* mutants in *S. cerevisiae*, we identified nine template switching events in coding genes[42]. It is unknown if and how template switching affects genes during evolution. The present work studied the effect of template switching on coding regions from an evolutionary perspective.

Template switching between short IR arms can cause mutation clusters though arm homogenization[53]. Multinucleotide mutations (MNMs), which comprise ~ 1% of single nucleotide polymorphisms in genomes[54–57], form one type of mutation cluster, in which mutations appear at adjacent sites. MNMs in codons can be the outcome of two entirely different scenarios: mutational mechanisms that simultaneously affect nearby nucleotides or multiple changes that occur via adaptive evolution. Ignoring the contribution of mechanisms that simultaneously affect nearby nucleotides in codons may lead to false identifications of positive selection[54,55,57]. This is because positive selection tests determine nonsynonymous to synonymous ratios, while assuming an independency of mutations. Identifying the mechanisms that cause nonsynonymous replacements through MNMs is essential to the understanding of protein evolution. Template switching between short IR arms is a potential mechanism of MNM formation in genes.

Here, protein-coding genes of 50 closely related wild type *Saccharomyces cerevisiae* strains were analyzed to identify MNMs arising from template switching between IR arms. Such events were identified in 39 yeast proteins and were responsible for nonsynonymous substitutions and, thus, for amino acid replacements. While template switching primarily introduced single amino acid changes, events simultaneously affecting up to five nearby amino acids were also recorded. The presented results indicate that template switching is an important mechanism in protein evolution.

Results

IRs are associated with MNMs on IR arms in *wild type* yeast. To identify the effect of template switching between IR arms on coding genes, we sought out IRs associated with MNMs. To classify IRs as associated with MNMs, we first identified perfect IRs in a *S. cerevisiae* strains. Next, based on the reconstructed phylogenic tree, we identified MNMs that occurred between a strain and its parental node (see "Methods" section). Finally, we looked for cases of MNMs with coordinates overlapping IR arms that mapped to the terminal branch leading to the same strain with the IR (Fig. 1a). Identification of MNMs on the specific branch is based on ancestral sequence reconstruction. To increase the reliability of the analysis and avoid uncertainty resulting from ancestral sequence reconstruction, focus was placed on MNMs associated with IRs on external branches only. IRs with an arm length of ≥ 7 bp and a spacer ≤ 70 that are associated with MNMs were identified in 68 genes (Table 1).

Next, we sought to determine whether the identified events differ from what could be expected when no special mechanism for IR homogenization exists. If no specific mechanism acts on IRs, then IR regions are expected to evolve under the same mutation-selection regime as non-IR regions in the gene. To test the significance of the association between IRs and MNMs in *S. cerevisiae* genes and whether they can be ascribed to template switching between IR arms, we simulated each of the gene multiple sequence alignments (MSAs) until reaching 100 simulations with a similar number of IRs in the real data. All evolutionary parameters used for the simulations mirrored those of the real gene MSA: tree topology, branch lengths, phylogeny model, and proportion of invariant positions. An IR score was computed for each gene and each simulation, and represented the enrichment of MNMs presumably formed by IR homogenization. To account for MNM variation not associated with IRs, the score was calculated by dividing the number of IRs associated with MNMs by the number of non-IRs associated with MNMs (see "Methods" section). An IR score was calculated for each of the real genes and its 100 simulations. The empirical distribution of IR scores in 100 simulations served as a null distribution to which the score of the real gene was compared. If the value of the real IR score fell within 5% of the values of the IR scores of its 100 simulations, a gene was considered to have more MNMs associated with IRs than its simulations. Genes with a statistically significant association between MNMs and IRs were considered to have undergone template switching.

Our simulation revealed that IRs were significantly associated with MNMs in 30 out of 68 yeast genes (Table 1, Supplementary Tables 1, 2). The longer the IR arm, the higher the fraction of genes with a significant real IR score. For IRs with an arm length of 7 bp, only 19% (8/42) of the genes had a higher IR score than their null distribution; while for IRs with an arm length of 8 bp, 69% of the genes (11/16) had a higher IR score than their null distribution. All IRs with arm lengths of 9–11 bp (10/10) had a higher IR score than randomly expected (Table 1). Due to the inability to simulate sufficient sequences with an IR arm of length of 16 bp, no statistical evaluation was performed for this event. However, since the association between IR arm length and recent MNMs was stronger for longer IR arms, the association of a MNM with IR arm length 16 bp is likely real. In conclusion, template switching between IR arms formed MNMs and modified protein-coding DNA during the evolution of 30 *S. cerevisiae* genes.

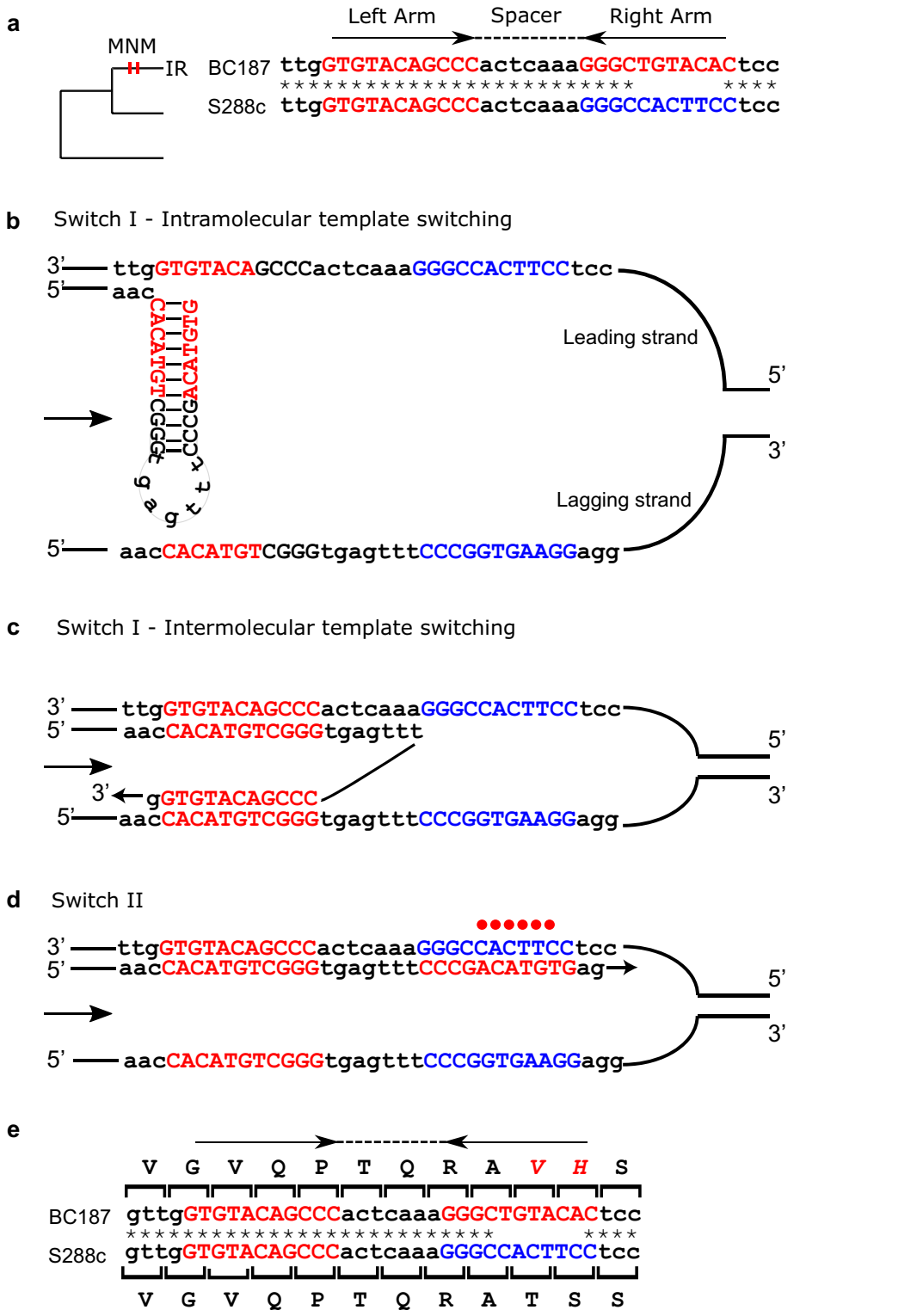

Figure 1. Template switching causes IR arm homogenization. (**a**) A recently occurring perfect IR in *UTP5* overlapped with a MNM on the branch leading to BC187 (shown as vertical red lines on the branch). Alignment of BC187 (perfect IR) and S288c (imperfect IR) is shown, with arrows representing IR arms, the dotted line representing the IR spacer and asterisks representing matches between strains. A very short IR with an arm length of 4 bp and a spacer of 7 bp appears in S288c. In strain BC187, a longer IR with an arm length of 11 bp was formed. (**b**–**d**) Template switching converted an imperfect IR to a perfect IR. (**b**) Assuming intramolecular template switching, the nascent strand folded on itself and served as a template. (**c**) Assuming intermolecular template switching, the nascent strand invades the sister chromatid and uses it as a template. (**d**) The second switch returns the nascent strand into the original template, resulting in a perfect IR, represented by the upper sequence in A. Uppercase letters represent the IR arms, with the same arms colored in the same color. The red dots represent mismatches between the arms. The direction of the replication fork is indicated by an arrow. (**e**) Six substitutions were induced during the template switching, resulting in the replacement of two amino acids (T38V, S39H).

Arm length	Number of genes with MNMs on arms only	Number of genes with a significant IR score compared to 100 random simulations (%)
16	1	*n.d.*
11	2	2 (100%)
10	2	2 (100%)
9	6	6 (100%)
8	16[a]	11 (69%)
7	42[a]	8 (20%)

Table 1. Number of genes that include IRs with an arm length of ≥ 7 bp and are associated with multinucleotide mutations. [a]One gene had multinucleotide mutations (MNMs) in both inverted repeats (IRs), which were 7 bp and 8 bp long. *n.d.* not detected.

The gene *UTP5* in BC187 *S. cerevisiae* strain has a perfect IR with an arm length of 11 bp, while at this locus S228c and other *S. cerevisiae* strains have an imperfect IR, with a continuous arm length of only 4 bp (Fig. 1b). Herein, we describe the template switching event that formed the perfect IR in BC187 from the ancestral form presenting in S288c (Fig. 1b–d). The process was comprised of two switches. The first switch moved the polymerase from its nascent template to the other IR arm, via either an intramolecular or intermolecular mechanism. In case of the intramolecular mechanism (Fig. 1b), the nascent strand folds upon itself using the arm base pairing. Thus the first arm is used as a template for replicating the second arm. In case of the intermolecular mechanism (Fig. 1c), the first switch is achieved when the nascent strand replicating one arm invades the template of the other sister chromatid. In both scenarios, the first switch is followed by a second switch, whereupon the polymerase returns to use the original strand as a template (Fig. 1d). The fork then resolves, leaving one daughter cell with an imperfect IR and one daughter cell with a perfect IR. The template switching in *UTP5* resulted in six base substitutions and two amino acid replacements (T38V, S39H) in strain BC187 (Fig. 1e).

Out of the 30 genes with IRs significantly associated with MNMs, 17 occurred uniquely in a single strain. The additional 13 genes included IRs that appeared in more than one *S. cerevisiae* strain, of which six had multiple IRs associated with MNMs on terminal branches (Supplementary Table 1). For example, in 14 strains, the *MSH4* gene had an IR with an arm length of 8 bp formed by an AA → TT MNM, which resulted in two amino acid changes, L394F and I395F (Fig. 2a). IR homogenization arose on three external branches of the *S. cerevisiae* phylogeny, leading to the strains YJM269, RedStar, and EC9-8. The ancestral sequence reconstruction revealed that one event had also occurred on one internal branch.

MNMs on IR arms formed by template switching caused nonsynonymous substitutions in 28 out of 30 genes. Twenty-five of these led to a single amino acid replacement, while two amino acid changes were observed in three genes. Thus, template switching is a source of parallel events and genetic innovation in proteins.

Template switching causes spacer inversions and parallel evolution. Template switching between IR arms not only homogenizes the arms, but can form inversion of IR spacers[58,59]. This occurs when the first switch takes place through intermolecular template switching, during the replication of the first arm (Fig. 3). A search for IRs on *S. cerevisiae* strains carrying MNMs on the spacer, identified cases in which the MNM on the spacer arose from a complete spacer inversion. Ten IR spacer inversion events were identified within the coding sequence of *wild type* yeasts (Table 2, Supplementary Table 3). All events were statistically significant—none of the 100 simulations of the genes with spacer inversion showed inversions. Spacer inversions were observed only on IRs with arms longer than 9 bp. Inverted spacers ranged between 2 and 5 bp substitutions and resulted in up to two amino acid replacements.

In nine out of the 10 genes with a spacer inversion, the inversion occurred in the case of previous homogenously perfect IR arms (only the spacer sequence changed). *SYG1* has 4 bp spacer inversion (Fig. 3). In this gene, a perfect IR arm of 10 bp was observed both in the derived JAY291 strain and in the ancestral form represented by S288c. The derived JAY291 strain had an inversion on the spacer, forming a 4 bp MNM. First, an intermolecular template switch occurring during the synthesis of the left arm, caused the 4 bp spacer inversion (Fig. 3b). Next, the nascent strand switched back to the original template (Fig. 3c). This inversion resulted in a Y46G replacement in the JAY291 strain (Fig. 3d).

In only one event (*REG2*), did a spacer inversion occur together with conversion of an arm to form a perfect IR (Fig. 4). In this example, the ancestral form is represented by S288c, which has an IR with a 5 bp arm and a 5 bp spacer. This form evolved into a perfect IR with an arm length of 16 bp in the YJM339 strain. This event occurred through an intermolecular mechanism, similar to that presented in Fig. 3, and resulted in seven point mutations. Only two of these mutations were part of a continuous MNM, while the others were in a mutation cluster, 2 bp apart from each other (Fig. 4b, c). A total of five amino acid replacements were observed (Fig. 4d), the highest number of amino acid replacements resulting from template switching between IR arms that we identified in *S. cerevisiae* strains. Three amino acid replacements occurred on the arms (L208Q, D209G, P210R) and two occurred on the spacer (K205D, S206F).

In four of the genes, recent inversion of the spacer occurred independently in several strains, resulting in parallel evolution of this amino acid position. Spacer inversion occurred twice in *AYR1*, three times in *ICT1* and *SPO75*, and five times in *YBZ1*. In the genes *SPO75* and *YBZ1*, a flip inversion to two amino acid forms was observed.

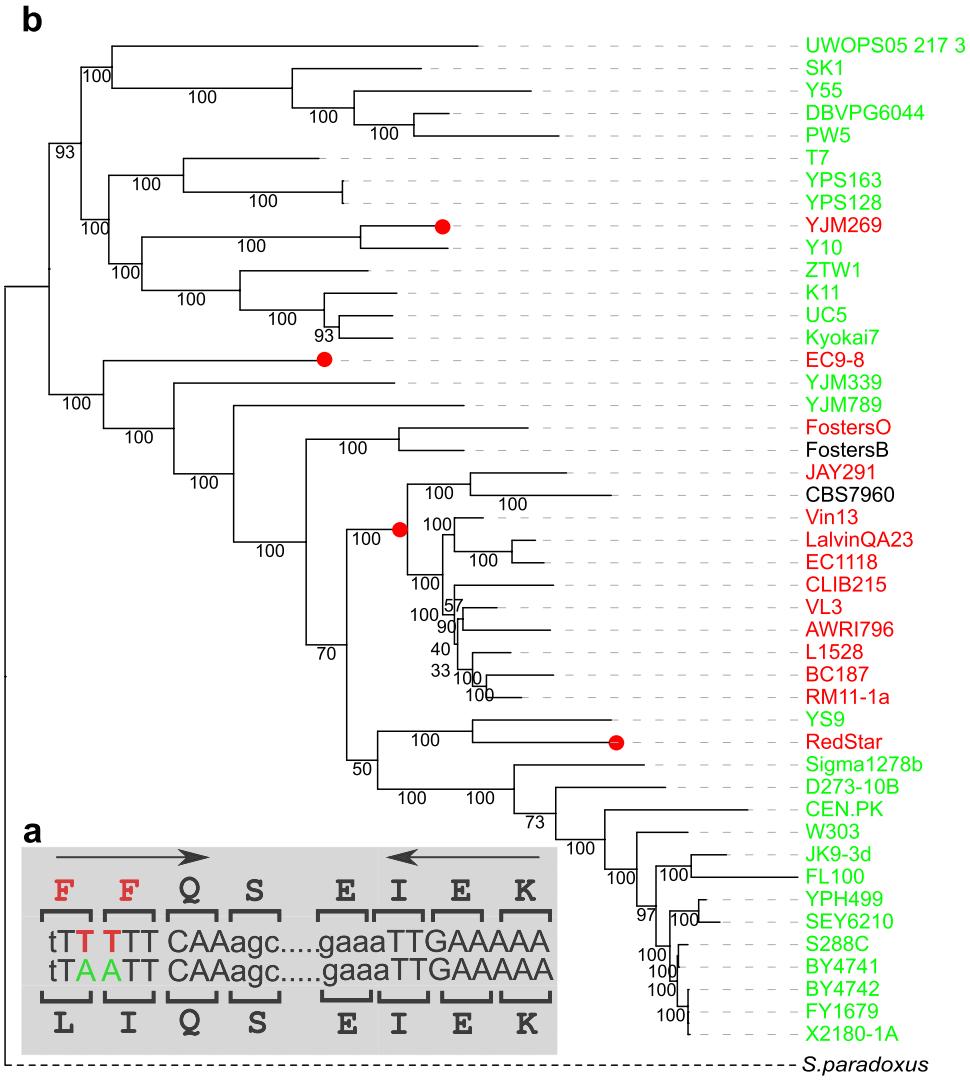

Figure 2. Template switching events forming an IR in the gene *MSH4*. (**a**) The homogenous IR has an arm length of 8 bp (upper sequence). The IR homogenesis was formed by an AA → TT MNM and resulted in two amino acid changes, i.e., L394F, I395F. The spacer is not shown. (**b**) Placing of template switching events on the tree. Fourteen strains with perfect IR and the TT form are shown in red, strains with the AA form are shown in green and other forms are shown in black. Template switching occurred on three terminal branches in *S. cerevisiae* lineage (red circles): on the branches leading to YJM269, RedStar, and EC9-8. While not directly estimated in this study, an ancestral sequence reconstruction revealed one extra event on one internal branch. The best maximum likelihood tree reconstructed by RAxML using DNA sequences from 4304 orthologs and the ancestral sequence reconstruction of codons 394–395 performed by FASTML are presented. (see "Methods" section). Bootstrap values are shown on branches. Tree outgroup is not drawn to scale.

The two-base IR spacer appearing in *SPO75*, where perfect IRs with 11 bp arms appear in all *S. cerevisiae* strains (Fig. 5a), displays two forms: either the ancestral form AA (red), or the derived form TT (blue) in positions 1 and 2 of the alternative codons AAA (K) and TTA (L), respectively. As a result, Spo75 has one of two forms: K or L on codon 409. Three transitions were identified on the terminal branches (two from L to K, and one from K to L), two of which are on highly supported branches. In addition, ancestral sequence reconstruction indicated that one K-to-L reversal event occurred on an internal branch with a high support, and another with a low support (Fig. 5b). We concluded that spacer inversion is an event that can change forms within a short evolutionary time, usually on the basis of a perfect IR.

Spacer length is a key player in the formation of DNA structure and thus influences the frequency of template switching[60,61]. We examined whether spacer length of IRs associated with template switching differs from that of IRs that are not associated with template switching (Supplementary Fig. 1). The spacer length of IRs associated with MNMs with significant IR scores is shorter than that of IRs associated with MNMs with non-significant IR scores (one-tail Wilcoxon rank sum test $p < 0.0011$). Similarly, spacer length of IRs associated with MNMs with significant IR scores is also shorter than that of IRs with only single mutations on IR arms (one-tail Wilcoxon

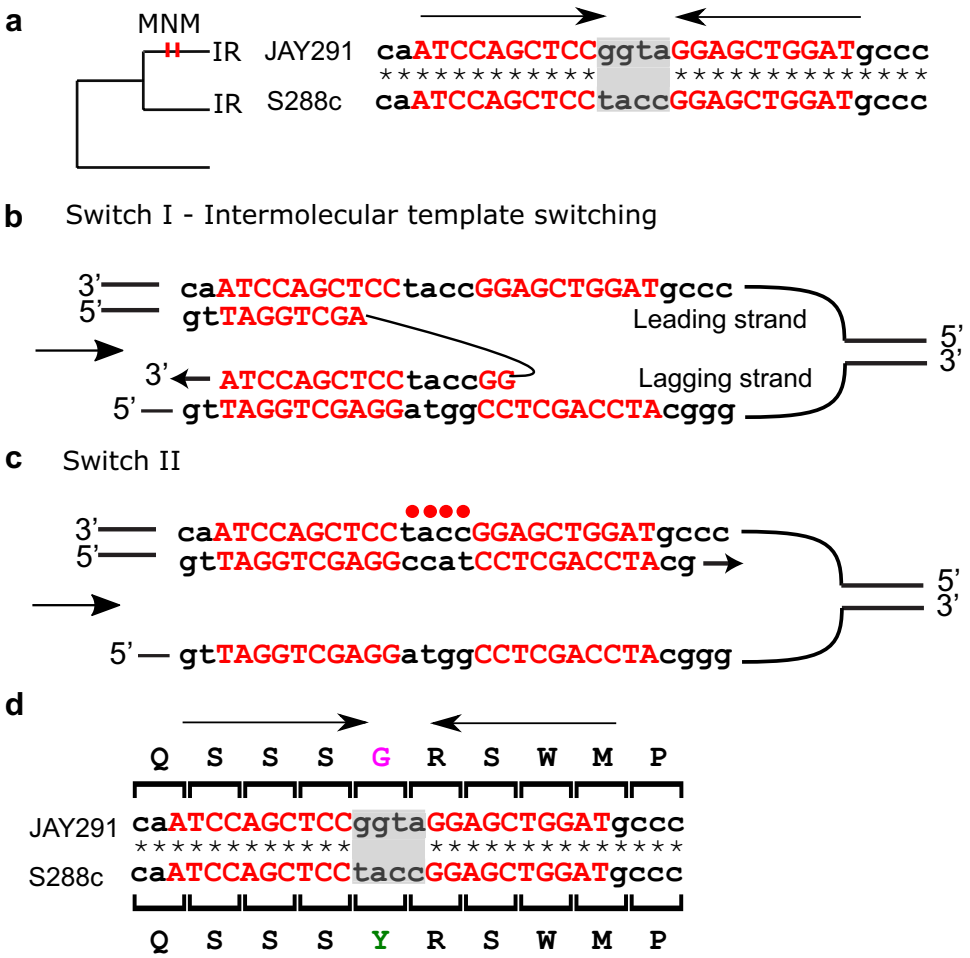

Figure 3. Intermolecular template switching causes spacer inversion. (**a**) A perfect IR in JAY291 on *SYG1* overlapping with a MNM located on the spacer. The MNM occurred in the branch leading to JAY291 (shown as vertical lines on the tree). In this case, the sister taxon, represented here by S288c, also has a perfect IR with the same arms. (**b**) The mechanism that created spacer inversion (AAGTC→GACTT) was intermolecular template switching, with the first switch having occurred during synthesis of the first arm. (**c**) The second switch returned the nascent strand to the original template, resulting in an inverted spacer of JAY291 compared to S288c. (**d**) IR spacer inversion resulted in one amino acid replacement (Y46G) in JAY291. Uppercase letters represent the IR arms, with the same arms shown in the same color. The red dots represent mismatches. The direction of the replication fork is indicated by an arrow.

rank sum test $p < 4.4171 \times 10^{-4}$). Thus, spacers of IRs that undergo template switching to form MNMs are shorter than other IRs in coding genes.

The nonsynonymous substitutions resulted from template switching. MNMs associated with IRs can be located on a single codon or on two neighboring codons. Multiple differences in a single codon will change the amino acid (except for the rare case of serine). MNMs located on two codons will appear on the third position of the first codon and the first position of the second codon, and can result in a change between zero and two amino acids. Out of the 39 events associated with template switching, two resulted in synonymous substitutions only. In contrast, in 37 cases, at least one nonsynonymous substitution occurred (Fig. 6a). Out of these 37 events, in 24 genes, MNMs spanned a single codon and in 13 they spanned two codons. The number of transitions between the strains and their parental nodes in the template switching regions was higher than the number of transversions. The transition/transversion ratio was, however, smaller in template switching regions than in non-IR regions. As previously reported[57], these nonsynonymous substitutions are prone to misidentification as positive selection sites. Indeed, arm and spacer MNMs in 26 out of the 37 genes were inaccurately estimated to have undergone positive selection, 15 of which showed very strong support (Supplementary Table 1).

Next, we sought to determine whether the amino acid replacements have the potential to alter the structure or function of proteins. Figure 6b presents all the amino acid changes we observed, against the background of a Grantham's physicochemical distances table[62]. This table is based on amino acid properties such as composition, polarity, and molecular volume. While most amino acid changes displayed low Grantham physicochemical

Gene	IR arm and spacer length (bp)	Strains with recent inversions	Amino acid replacements due to inversion	Parental and strain sequences
PIM1	10, 3	YJM339	**S → G** **TCC → GGA**	CTCCAGAAGCtccGCTTCTGGAG CTCCAGAAGCggaGCTTCTGGAG
REG2	16, 5	YJM339	**K → D** **AAG → GAC,** **S → F** **TCT → TTT**	ACAGCGCCCTTGTTCAaagtcTGAACT AGATCCCTGT ACAGCGCCCTTGTTCAgacttTGAACA AGGGCGCTGT
SYG1	10, 4	JAY291	**Y → G** **TAC → GGT,** R → R CGG → AGG	ATCCAGCTCCtaccGGAGCTGGAT ATCCAGCTCCggtaGGAGCTGGAT
AYR1	10, 4	UWOPS05, YJM269	P → P CCT → CCA, **D → S** **GAT → TCA**	CTAATTTACCtgatGGTAAATTAG CTAATTTACCatcaGGTAAATTAG
SPO75	11, 2	D273-10B, CEN.PK, Y55	**D273-10B, CEN.PK:** **L → K,** **TTA → AAA** **Y55: K → L,** **AAA → TTA**	TGCCCGACGATttATCGTCGGGCA TGCCCGACGATaaATCGTCGGGCA
ICT1	10, 2	D273-10B, SEY6210, YJM789	L → L CTG → CTT, **K → Q** **AAG → CAG**	TGCAGGGCCTgaAGGCCCTGCA TGCAGGGCCTtcAGGCCCTGCA
YPS1	11, 3	UWOPS05	S → S TCT → TCG, **S → K** **TCG → AAG**	CCATACTGTTCttcGAACAGTATGG CCATACTGTTCgaaGAACAGTATGG
ECM30	10 , 2	W303	**D → S,** **GAT → TCT**	GGAGGACGCAgaTGCGTCCTCC GGAGGACGCAtcTGCGTCCTCC
YBZ1	10, 3	UWOPS05, Kyokai7, RedStar, CLIB215, L1528	**UWOPS05:** **E → F** **GAA → TTC** **Kyokai7, RedStar, CLIB215, L1528:** **F → E,** **TTC → GAA**	TGTCAATGCCttcGGCATTGACA TGTCAATGCCgaaGGCATTGACA
TAT2	11, 3	SEY6210	**E → F,** **GAA → TTC**	TTGGATTTGTAgaaTACAAATCCAA TTGGATTTGTAttcTACAAATCCAA

Table 2. Genes with inverted repeats and spacer inversion. Nonsynonymous substitutions are shown in bold; the spacer are underlined. UWOPS05 is a short for UWOPS05_217_3.

distances, representing similar amino acid properties, 13 amino acid replacements displayed high Grantham's physicochemical distances (above 120).

We then used PredictSNP[63], a classifier that combines several prediction tools, to identify the effect of mutations on protein structure and function. PredictSNP predicted that most amino acid changes had a neutral effect on protein function, while changes in six nonessential genes were considered non-neutral. These changes were L208Q in *REG2*, D98C in *DLD3*, G191K in *RTT10*, P50F in *OM45*, W127Y in *RHO5*, and G166Y in *CCS1*. The EggNOG database[64] was then used to determine whether these novel amino acids are represented in other species during *Saccharomycotina* evolution. While the divergence time of *S. cerevisiae* and *S. paradoxus* is 4.0–5.8 mya, the origin of the budding yeast subphylum *Saccharomycotina* is 317–523 mya[65]. Five of the six genes had an orthologous group in EggNOG. In these genes, the alternative amino acid did not appear in *Saccharomycotina* (*OM45*, *RHO5*, *CCS1, DLD3,* and *RTT10*). In *RTT10*, the novel amino acid was, however, common when considering the entire *Ascomycota* phylum. Only two proteins of the genes *RHO5* and *DLD3* had an amino acid replacement in residues that show conservation throughout evolution. In conclusion, template switching introduced new nonsynonymous mutations into DNA coding genes during *S. cerevisiae* evolution. Most of the selected positions had physicochemical properties that were similar to those of their ancestral counterparts. In addition, amino acids with physicochemical properties distant from those of their ancestors, rarely occurred in conserved protein regions.

Discussion

Template switching events have been previously reported in the context of loss of function in genes[34–37,51,52]. The question of how this affects normal gene evolution, however, has not been addressed. Given that template switching events involve multiple substitutions, our identification of template switching mediated by IRs in 39 *wild type S. cerevisiae* coding genes (~1% of the analyzed genes) was surprising. In most cases we identified in *S. cerevisiae* strains, template switching events yielded nonsynonymous substitutions. Most template switching events resulted in a single nonsynonymous substitution, although one extreme circumstance of five nearby amino acid replacements was also identified (Fig. 4). The influence of template switching on coding genes is probably

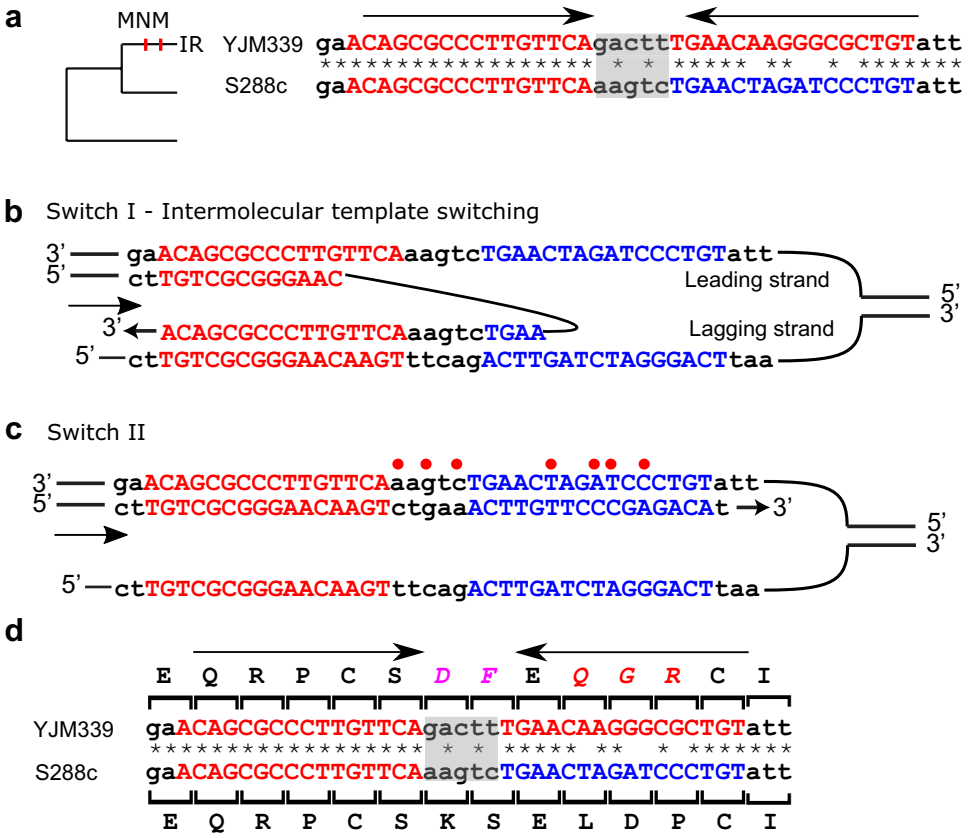

Figure 4. Intermolecular template switching causes arm homogenization and spacer inversion. (**a**) *REG2* has a perfect IR with an arm length of 16 bp in YHM339. The ancestral form represented by S288c has an arm length of 5 bp. The MNM occurred on the IR arm in the branch leading to YHM339 (shown as vertical lines on tree). (**b**) Intermolecular template switching, in which the first switch occurred during the synthesis of the first arm, was responsible for spacer inversion. (**c**) The second switch returned the nascent strand to the original template, resulting in an inverted spacer of YHM339 compared to S288c. (**d**) Seven substitutions were induced by this event (three on the spacer and four on the arm, two of which were part of a continuous MNM) resulting in five modified amino acids (KSELDP → DFEQRR, positions 206–211).

even stronger than reported here, as we ignored cluster mutations that are not MNMs, indels, and MNMs occurring on internal branches.

This work focused on the effect that template switching between IR arms have on gene evolution. What we learned on the genetic mechanism of template switching was heavily affected by selection. We were, however, able to show that IRs involved in template switching have shorter spacers than other IRs, as previously reported[60,61] (Supplementary Fig. 1). In our evolutionary data our findings were also influenced by the selection for short spacers in intermolecular template switching with spacer inversion (Table 2). This is because long inversion in proteins are unlikely to be neutral. There was a preference for spacer inversion when long IR arms were involved (Table 2, Figs. 3, 4).

In contrast to previously published data from bacteria in which template switching favored transversion over transition[66] here we observed more transitions than transversions. The reason for this is probably due to selection over evolutionary time. Based on the structure of the genetic code, transitions are more likely than transversions to be synonymous, making transitions less often selected against than transversions[67,68]. IRs are unique since homogenization can occur when the nascent DNA strand folds and is used as a template. It is also of interest to further study the effect of other forms of repeats on protein evolution. Our preliminary results suggest that direct repeats can also promote the formation of MNMs.

The discovery that template switching between IR arms results in multiple events through evolution and, hence, in cryptic hotspots for parallel evolution, is an important finding. We note that identifying multiple template switching events (Figs. 2, 5) relies on the accuracy of the *S. cerevisiae* tree and the assumption that all genes within the *S. cerevisiae* lineage have the same evolutionary history. The reconstructed concatenated tree had a high bootstrap support.

Mutational hotspots have been previously suggested to be associated with parallel evolution at the amino acid level. For example, an elevated mutation rate at CpGs was shown to underlie hemoglobin adaptation in high-altitude Andean house wrens[69]. Template switching, however, was not previously considered a mechanism that contributes to parallel evolution in proteins. Here, we identified the reoccurrence of template switching over a

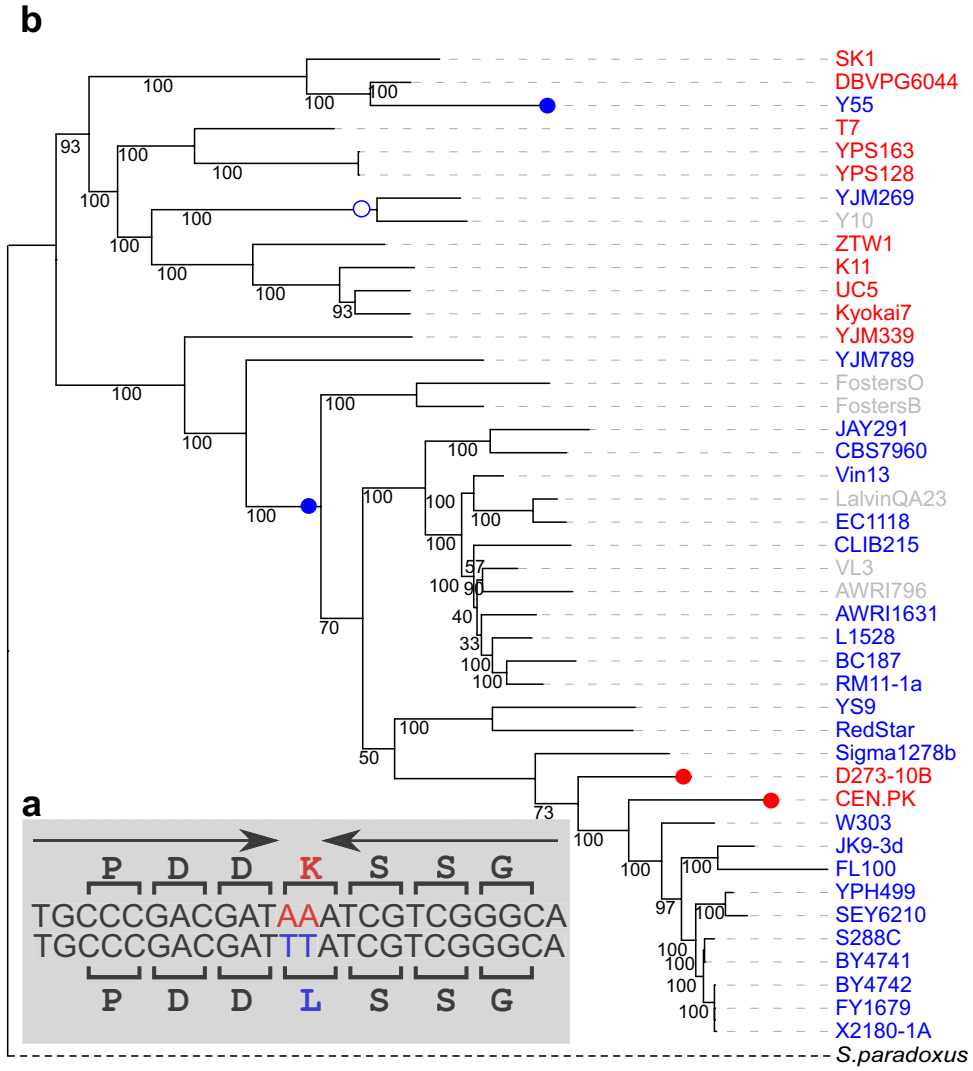

Figure 5. Spacer flip inversion in IR of *SPO75*. (**a**) IR with an arm length of 11 bp shows a perfect form in all *S. cerevisiae* strains. The 2 bp-spacer has either the ancestral form AA (red) or the derived form TT (blue) in positions 1 and 2 of codon 409 coding AAA (K) and TTA (L) respectively. (**b**) Tree presenting the template switching events. Strains in red have AA on the spacer, strains in blue have TT on the spacer, and strains in gray have indels in this region. *S. paradoxus* has AA in this locus but does not share the same IR. Three transitions are shown in full circles (two from L into K, shown in red, and one from K into L, shown in blue) on terminal branches. According to ancestral sequence reconstruction, there were also one to two reversal events from K to L on internal branches, one of which has only low support (empty circle). The best maximum likelihood tree reconstructed by RAxML using DNA sequences of 4304 orthologs and the ancestral sequence reconstruction of codon 409 by FASTML, are presented (see "Methods" section). Bootstrap values are shown on branches. Tree outgroup is not drawn to scale.

short evolutionary period, yielding both arm homogenization and IR spacer inversion. Spacer inversions were reversible through sequential template switching events.

Codons with multiple changes between similar species can be the outcome of either a mechanism that simultaneously affects adjacent nucleotides, or of positive selection. Because substitutions are assumed to be independent, MNMs are sometimes considered evidence of multistep adaptive changes. Similarly, parallel evolution is usually considered evidence of adaptive selection. Such interpretations can be incorrect when the mutations occur together[54,55,57]. Indeed, when tested, 70% of what we recognized as template switching events on arms were inaccurately estimated as positive selection (Supplementary Table 1). Thus, our results suggest that template switching is a general mutagenic mechanism that causes MNMs, as well as parallel evolution, eliminating the need for adaptive explanations. Identification of the complex mechanisms that cause MNMs, such as template switching, error-prone translesion DNA synthesis[70], gene conversion[71], and probably other yet to be discovered processes, is essential in order to prevent overestimation of adaptive selection.

Although adaptive processes are not needed to explain MNMs caused by template switching and other complex mechanisms, this does not rule out the option that mutations formed by template switching can be a

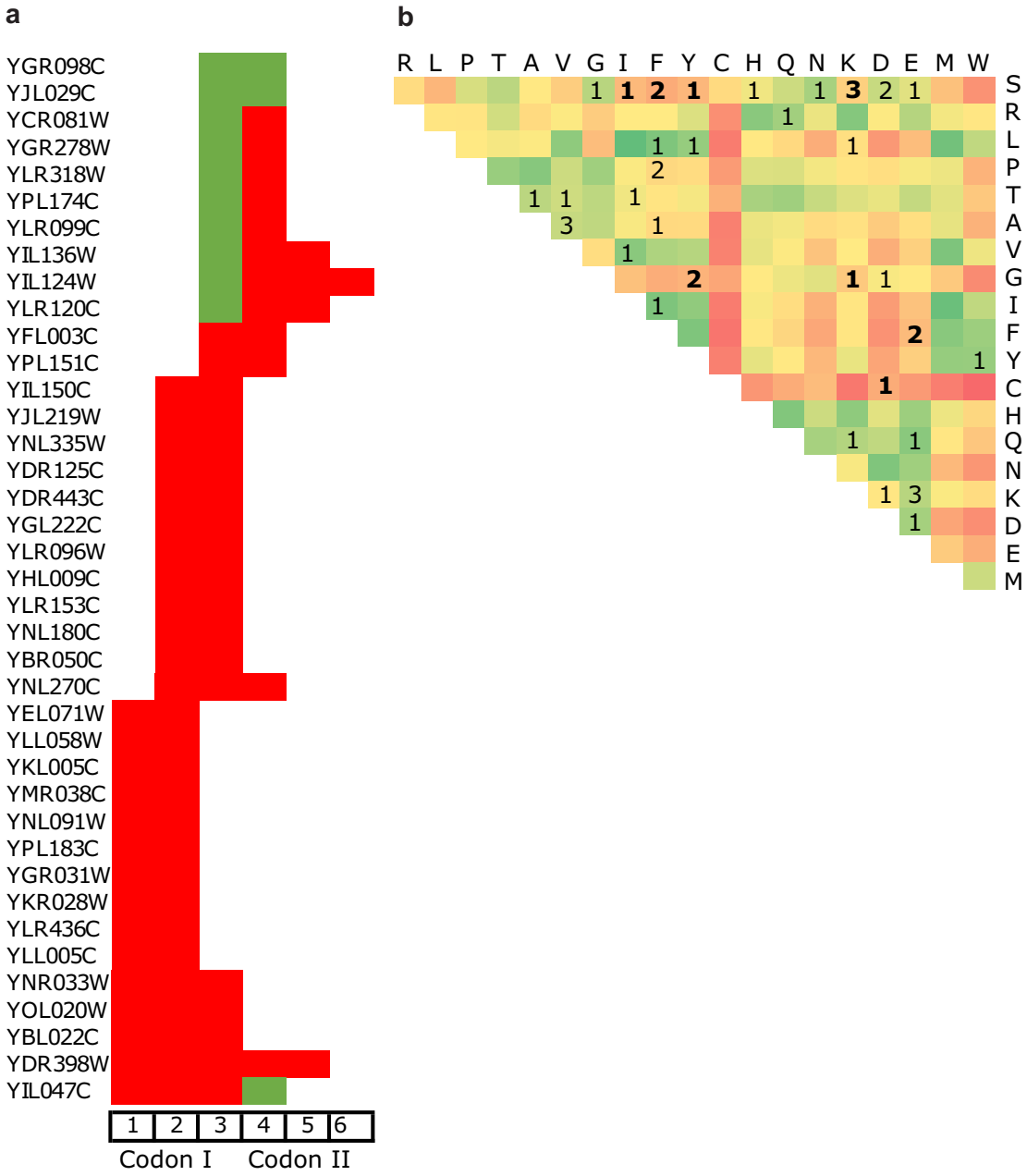

Figure 6. Effect of template switching on codons and amino acids. (**a**) Location of substitutions on two neighboring codons (six nucleotide positions). Synonymous substitutions are shown in green, and nonsynonymous substitutions are shown in red. For example, the MNM of gene YGR098C (top) are located in the 3rd position of the first codon and on the 1st position of the second codon. Both substitutions caused the synonymous changes shown in green. Similarly, YIL047C (bottom) has four neighboring substitutions spanning two codons. The first codon has three substitutions that replace the amino acid, shown in red, and the second codon has one synonymous substitution, shown in green. (**b**) The amino acid changes induced by MNMs in 39 proteins are shown; one event from its type per gene locus. The matrix is colored by Grantham's physicochemical distance table[62] which has values of 5–215, with a mean distance of 100. Green represents similar amino acid pairs, while red represents distant amino acid pairs. Thirteen events had an amino acid with a physicochemical distance of 120 or higher (in bold).

target for positive selection[54]. Most nonsynonymous mutations are eliminated by purifying selection and those that are fixed are usually replaced by physicochemically similar ones. However, nonsynonymous substitutions are still the ones with a small chance of improving function. By causing multiple nonsynonymous substitutions, template switching can, therefore, enable hopping between adaptive peaks without the crossing of low-fitness valleys in the adaptive landscape[54,72,73]. Based on the reported effect of synonymous mutations on RNA stability and protein translation efficiency[74], even a combination of synonymous and nonsynonymous substitutions (Fig. 6) can promote a similar outcome.

Methods

Data collection and IR detection. The sequence of 6569 orthologous sets of coding DNA and protein sequences from 50 wild type *S. cerevisiae* strains were downloaded from the *Saccharomyces* Genome Database[75]. The *Saccharomyces paradoxus* ortholog of each gene was assigned based on the Fungal Orthogroups Repository[76]. Sequences with more than 20 *Ns* and sequences found in fewer than four strains were removed from the orthologous sets. These screenings resulted in a total of 4304 genes. For each orthologous group, we searched for IRs with an arm length of at least 7 bp and a spacer of up to 70 bp, in each of the available strains, using the EMBOSS palindrome package[77]. Each IR arm length was analyzed separately.

Phylogeny. Each of the 4304 orthologous sets was aligned using MAFFT V3.705[78] with default parameters. DNA MSAs were concatenated, and the best maximum likelihood (ML) tree was reconstructed by RAxML version 8.2.11[79] under the GTR replacement matrix[80], with among-site-rate-variation accounted for by assuming a discrete gamma distribution[81] and with rapid bootstraps. This was the species tree used in this study.

Ancestral tree reconstruction. Each orthologous set was also aligned by codon alignment. The phylogenetic species tree was pruned and used together with each orthologous codon MSA to estimate branch length and reconstruct ancestral codon sequences using FASTML[82] with the M5 codon model[83]. In this step, orthologous sets with immature stop codon were eliminated, resulting in the 4252 orthologous sets that were used in this study.

Identifying IRs that overlap MNMs. An MNM was defined when two or more adjacent substitutions were observed between a *S. cerevisiae* strain and its immediate parent node, as determined from FASTML ancestral reconstruction output. Thus, neighboring mutations mapping to two different branches were not identified as MNMs[54]. Insertions and deletions (indels) were not considered for MNM classification. IRs that mapped to a *S. cerevisiae* strain with overlapping MNMs in the terminal branch leading to this strain, were further analyzed. For each IR arm length, overlapping IRs on the same strain were excluded from the analysis. However, when an IR was fully nested in another IR, they were both analyzed.

Elimination of false MSAs. To ensure MSA accuracy and avoid false MNMs, we used GUIDANCE2[84] to score alignment regions with IRs in a codon model, using default parameters. We looked for IR regions with a cutoff higher than 0.95. In addition, codon MSA might not give the best MSA if insertion/deletions (indels) are not a multiple of three. We therefore used the similarity between nucleotide and codon comparisons in the IR region to identify high-quality MSAs. Codon and nucleotide pairwise alignments of the IR with a MNM and its sister taxa (plus a tail of 50 bp) were scored with a scoring matrix of match = 1, mismatch = − 1, and gap = 0. IRs were ignored if the difference between nucleotide and codon MSA scores was higher than 15, indicating a problematic codon MSA. Moreover, IR regions that included indels between the *S. cerevisiae* strain and their immediate parent node were ignored. Finally, to avoid mutation saturation, IR corresponding to branch lengths longer than 0.2 were ignored.

Simulation. To determine the significance of association between IRs and MNMs, sequences were simulated along rooted phylogenetic trees using INDELible[85] with the M5 model[83]. Each orthologous group codon MSA was simulated according to the FASTML phylogenetic tree using the inferred M5 evolutionary model parameters (kappa: transition/transversion ratio, and omega: dN/dS ratio) and PAML codeml[86]. In each simulation, the root was set to *S. paradoxus*. The sequence length for the INDELible simulation was set to four-five times the *S. paradoxus* sequence length in order to yield sufficient IRs with the exact arm length in each simulation. The length factor needed was selected based on the empirical evaluation of simulations. We continued the simulations until we obtained 100 MSAs in which the number of IRs in each was equal to or greater than the number of IRs in the real MSA. For each simulation in each orthologous set, the analysis performed for the real codon MSA was repeated. Each IR length was simulated separately.

Control sequences. To correct for MNM enrichment that was not associated with IRs in the real MSA compared to the simulation, we searched for IR-less alignment segments of the same length as the IR. A control sequence was matched for each IR arm of the same length, in a non-IR region with the closest proximity to the IR. Thus, the number of IRs and control regions for each orthologous group was the same. If identical IRs were identified in multiple strains, the control regions were also chosen in the same positions and the same strains. Exact control analysis was also carried out for all simulations of the orthologous group.

IR score. For each orthologous group and for each of its 100 simulations, the IR score was calculated as follows:

$$IR\ score = \frac{IRs\ with\ MNMs + 1}{IRs\ without\ MNMs + 1} : \frac{Controls\ with\ MNMs + 1}{Controls\ without\ MNMs + 1}$$

Namely, we calculated the ratio between the number of IRs with and without MNMs divided by the number of controls with and without MNMs. A pseudo count of 1 was added to all elements to prevent division by zero. The analysis was carried out separately for each IR arm length.

Estimation of positive selection. Positive selection was estimated by PAML codeml[86]. MNMs that overlapped IRs were estimated with positive selection if Bayes empirical Bayes[87] posterior probability under the positive selection model was above 0.5. Strong support for positive selection was identified if posterior probability under the positive selection model was above 0.95.

Novelty of amino acid replacements. PredictSNP[63] classifier with default parameters was used to identify amino acid replacements with a potential affect of protein structure. Proteins with a predicted affect on structure were manually inspected in the EggNOG database of orthology relationships[64]. For each protein we identified whether the replaced amino acid was represented in the MSA of *Saccharomycotina*[64].

Received: 20 June 2021; Accepted: 27 October 2021
Published online: 19 November 2021

References

1. Gordenin, D. A. *et al.* Inverted DNA repeats: A source of eukaryotic genomic instability. *Mol. Cell Biol.* **13**, 5315–5322 (1993).
2. Lobachev, K. S. *et al.* Factors affecting inverted repeat stimulation of recombination and deletion in *Saccharomyces cerevisiae*. *Genetics* **148**, 1507–1524 (1998).
3. Connelly, J. C. & Leach, D. R. The sbcC and sbcD genes of *Escherichia coli* encode a nuclease involved in palindrome inviability and genetic recombination. *Genes Cells* **1**, 285–291 (1996).
4. Leach, D. R. Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair. *BioEssays* **16**, 893–900. https://doi.org/10.1002/bies.950161207 (1994).
5. Leach, D. R., Okely, E. A. & Pinder, D. J. Repair by recombination of DNA containing a palindromic sequence. *Mol. Microbiol.* **26**, 597–606 (1997).
6. Lewis, S. M. Palindromy is eliminated through a structure-specific recombination process in rodent cells. *Nucl. Acids Res.* **27**, 2521–2528 (1999).
7. Narayanan, V., Mieczkowski, P. A., Kim, H. M., Petes, T. D. & Lobachev, K. S. The pattern of gene amplification is determined by the chromosomal location of hairpin-capped breaks. *Cell* **125**, 1283–1296. https://doi.org/10.1016/j.cell.2006.04.042 (2006).
8. Tanaka, H. & Yao, M. C. Palindromic gene amplification—An evolutionarily conserved role for DNA inverted repeats in the genome. *Nat. Rev. Cancer* **9**, 216–224. https://doi.org/10.1038/nrc2591 (2009).
9. Tanaka, H. *et al.* Intrastrand annealing leads to the formation of a large DNA palindrome and determines the boundaries of genomic amplification in human cancer. *Mol. Cell Biol.* **27**, 1993–2002. https://doi.org/10.1128/MCB.01313-06 (2007).
10. Zackai, E. H. & Emanuel, B. S. Site-specific reciprocal translocation, t(11;22) (q23;q11), in several unrelated families with 3:1 meiotic disjunction. *Am. J. Med. Genet.* **7**, 507–521. https://doi.org/10.1002/ajmg.1320070412 (1980).
11. Mizuno, K., Miyabe, I., Schalbetter, S. A., Carr, A. M. & Murray, J. M. Recombination-restarted replication makes inverted chromosome fusions at inverted repeats. *Nature* **493**, 246–249. https://doi.org/10.1038/nature11676 (2013).
12. Kurahashi, H. *et al.* Palindrome-mediated chromosomal translocations in humans. *DNA Repair (Amst.)* **5**, 1136–1145. https://doi.org/10.1016/j.dnarep.2006.05.035 (2006).
13. Bzymek, M. & Lovett, S. T. Instability of repetitive DNA sequences: The role of replication in multiple mechanisms. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 8319–8325. https://doi.org/10.1073/pnas.111008398 (2001).
14. Montgomery, S. B. *et al.* The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. *Genome Res.* **23**, 749–761. https://doi.org/10.1101/gr.148718.112 (2013).
15. Weingarten-Gabbay, S. & Segal, E. The grammar of transcriptional regulation. *Hum. Genet.* **133**, 701–711. https://doi.org/10.1007/s00439-013-1413-1 (2014).
16. Leung, M. Y., Choi, K. P., Xia, A. & Chen, L. H. Nonrandom clusters of palindromes in herpesvirus genomes. *J. Comput. Biol.* **12**, 331–354. https://doi.org/10.1089/cmb.2005.12.331 (2005).
17. Horvath, P. & Barrangou, R. CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**, 167–170. https://doi.org/10.1126/science.1179555 (2010).
18. Li, X., Lindahl, L., Sha, Y. & Zengel, J. M. Analysis of the *Bacillus subtilis* S10 ribosomal protein gene cluster identifies two promoters that may be responsible for transcription of the entire 15-kilobase S10-spc-alpha cluster. *J. Bacteriol.* **179**, 7046–7054 (1997).
19. Cuomo, C. A., Mundy, C. L. & Oettinger, M. A. DNA sequence and structure requirements for cleavage of V(D)J recombination signal sequences. *Mol. Cell Biol.* **16**, 5683–5690 (1996).
20. Willis, K. K. & Klein, H. L. Intrachromosomal recombination in *Saccharomyces cerevisiae*: Reciprocal exchange in an inverted repeat and associated gene conversion. *Genetics* **117**, 633–643 (1987).
21. Rozen, S. *et al.* Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423**, 873–876. https://doi.org/10.1038/nature01723 (2003).
22. Kolodner, R. & Tewari, K. K. Inverted repeats in chloroplast DNA from higher plants. *Proc. Natl. Acad. Sci. U. S. A.* **76**, 41–45. https://doi.org/10.1073/pnas.76.1.41 (1979).
23. Rattray, A. J. & Symington, L. S. Use of a chromosomal inverted repeat to demonstrate that the RAD51 and RAD52 genes of *Saccharomyces cerevisiae* have different roles in mitotic recombination. *Genetics* **138**, 587–595 (1994).
24. Waldman, A. S., Tran, H., Goldsmith, E. C. & Resnick, M. A. Long inverted repeats are an at-risk motif for recombination in mammalian cells. *Genetics* **153**, 1873–1883 (1999).
25. Tran, H., Degtyareva, N., Gordenin, D. & Resnick, M. A. Altered replication and inverted repeats induce mismatch repair-independent recombination between highly diverged DNAs in yeast. *Mol. Cell Biol.* **17**, 1027–1036. https://doi.org/10.1128/mcb.17.2.1027 (1997).
26. Ripley, L. S. Model for the participation of quasi-palindromic DNA sequences in frameshift mutation. *Proc. Natl. Acad. Sci. U. S. A.* **79**, 4128–4132 (1982).
27. Rosche, W. A., Ripley, L. S. & Sinden, R. R. Primer-template misalignments during leading strand DNA synthesis account for the most frequent spontaneous mutations in a quasipalindromic region in *Escherichia coli*. *J. Mol. Biol.* **284**, 633–646. https://doi.org/10.1006/jmbi.1998.2193 (1998).
28. Lovett, S. T. Template-switching during replication fork repair in bacteria. *DNA Repair (Amst.)* **56**, 118–128. https://doi.org/10.1016/j.dnarep.2017.06.014 (2017).
29. Strawbridge, E. M., Benson, G., Gelfand, Y. & Benham, C. J. The distribution of inverted repeat sequences in the *Saccharomyces cerevisiae* genome. *Curr. Genet.* **56**, 321–340. https://doi.org/10.1007/s00294-010-0302-6 (2010).
30. Lisnić, B., Svetec, I. K., Sarić, H., Nikolić, I. & Zgaga, Z. Palindrome content of the yeast *Saccharomyces cerevisiae* genome. *Curr. Genet.* **47**, 289–297. https://doi.org/10.1007/s00294-005-0573-5 (2005).

31. van Noort, V., Worning, P., Ussery, D. W., Rosche, W. A. & Sinden, R. R. Strand misalignments lead to quasipalindrome correction. *Trends Genet.* **19**, 365–369 (2003).
32. Bzymek, M. & Lovett, S. T. Evidence for two mechanisms of palindrome-stimulated deletion in *Escherichia coli*: Single-strand annealing and replication slipped mispairing. *Genetics* **158**, 527–540 (2001).
33. Rosche, W. A., Trinh, T. Q. & Sinden, R. R. Leading strand specific spontaneous mutation corrects a quasipalindrome by an intermolecular strand switch mechanism. *J. Mol. Biol.* **269**, 176–187. https://doi.org/10.1006/jmbi.1997.1034 (1997).
34. Viswanathan, M., Lacirignola, J. J., Hurley, R. L. & Lovett, S. T. A novel mutational hotspot in a natural quasipalindrome in *Escherichia coli*. *J. Mol. Biol.* **302**, 553–564. https://doi.org/10.1006/jmbi.2000.4088 (2000).
35. Hampsey, D. M., Ernst, J. F., Stewart, J. W. & Sherman, F. Multiple base-pair mutations in yeast. *J. Mol. Biol.* **201**, 471–486. https://doi.org/10.1016/0022-2836(88)90629-8 (1988).
36. Greenblatt, M. S., Grollman, A. P. & Harris, C. C. Deletions and insertions in the p53 tumor suppressor gene in human cancers: Confirmation of the DNA polymerase slippage/misalignment model. *Cancer Res.* **56**, 2130–2136 (1996).
37. Bissler, J. J. DNA inverted repeats and human disease. *Front. Biosci.* **3**, d408-418 (1998).
38. Seier, T. *et al.* Insights into mutagenesis using *Escherichia coli* chromosomal lacZ strains that enable detection of a wide spectrum of mutational events. *Genetics* **188**, 247–262. https://doi.org/10.1534/genetics.111.127746 (2011).
39. Yoshiyama, K., Higuchi, K., Matsumura, H. & Maki, H. Directionality of DNA replication fork movement strongly affects the generation of spontaneous mutations in *Escherichia coli*. *J. Mol. Biol.* **307**, 1195–1206. https://doi.org/10.1006/jmbi.2001.4557 (2001).
40. Kim, N., Cho, J. E., Li, Y. C. & Jinks-Robertson, S. RNA:DNA hybrids initiate quasi-palindrome-associated mutations in highly transcribed yeast DNA. *PLoS Genet.* **9**, e1003924. https://doi.org/10.1371/journal.pgen.1003924 (2013).
41. Schultz, G. E. & Drake, J. W. Templated mutagenesis in bacteriophage T4 involving imperfect direct or indirect sequence repeats. *Genetics* **178**, 661–673. https://doi.org/10.1534/genetics.107.083444 (2008).
42. Omer, S., Lavi, B., Mieczkowski, P. A., Covo, S. & Hazkani-Covo, E. Whole genome sequence analysis of mutations accumulated in. *G3 (Bethesda)* **7**, 3775–3787. https://doi.org/10.1534/g3.117.300262 (2017).
43. Zhao, G., Chang, K. Y., Varley, K. & Stormo, G. D. Evidence for active maintenance of inverted repeat structures identified by a comparative genomic approach. *PLoS ONE* **2**, e262. https://doi.org/10.1371/journal.pone.0000262 (2007).
44. Warburton, P. E., Giordano, J., Cheung, F., Gelfand, Y. & Benson, G. Inverted repeat structure of the human genome: The X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome Res.* **14**, 1861–1869. https://doi.org/10.1101/gr.2542904 (2004).
45. Batzer, M. A. & Deininger, P. L. Alu repeats and human genomic diversity. *Nat. Rev. Genet.* **3**, 370–379. https://doi.org/10.1038/nrg798 (2002).
46. Aygun, N. Correlations between long inverted repeat (LIR) features, deletion size and distance from breakpoint in human gross gene deletions. *Sci. Rep.* **5**, 8300. https://doi.org/10.1038/srep08300 (2015).
47. Cook, G. W. *et al.* Alu pair exclusions in the human genome. *Mob. DNA* **2**, 10. https://doi.org/10.1186/1759-8753-2-10 (2011).
48. Lavi, B., Levy Karin, E., Pupko, T. & Hazkani-Covo, E. The prevalence and evolutionary conservation of inverted repeats in proteobacteria. *Genome Biol. Evol.* **10**, 918–927. https://doi.org/10.1093/gbe/evy044 (2018).
49. Löytynoja, A. & Goldman, N. Short template switch events explain mutation clusters in the human genome. *Genome Res.* **27**, 1039–1049. https://doi.org/10.1101/gr.214973.116 (2017).
50. Walker, C. R., Scally, A., De Maio, N. & Goldman, N. Short-range template switching in great ape genomes explored using pair hidden Markov models. *PLoS Genet.* **17**, e1009221. https://doi.org/10.1371/journal.pgen.1009221 (2021).
51. de Boer, J. G. & Ripley, L. S. Demonstration of the production of frameshift and base-substitution mutations by quasipalindromic DNA sequences. *Proc. Natl. Acad. Sci. U. S. A.* **81**, 5528–5531. https://doi.org/10.1073/pnas.81.17.5528 (1984).
52. Mo, J. Y., Maki, H. & Sekiguchi, M. Mutational specificity of the dnaE173 mutator associated with a defect in the catalytic subunit of DNA polymerase III of *Escherichia coli*. *J. Mol. Biol.* **222**, 925–936. https://doi.org/10.1016/0022-2836(91)90586-u (1991).
53. Chan, K. & Gordenin, D. A. Clusters of multiple mutations: Incidence and molecular mechanisms. *Annu. Rev. Genet.* **49**, 243–267. https://doi.org/10.1146/annurev-genet-112414-054714 (2015).
54. Schrider, D. R., Hourmozdi, J. N. & Hahn, M. W. Pervasive multinucleotide mutational events in eukaryotes. *Curr. Biol.* **21**, 1051–1054. https://doi.org/10.1016/j.cub.2011.05.013 (2011).
55. Besenbacher, S. *et al.* Multi-nucleotide de novo mutations in humans. *PLoS Genet.* **12**, e1006315. https://doi.org/10.1371/journal.pgen.1006315 (2016).
56. Hodgkinson, A. & Eyre-Walker, A. Human triallelic sites: Evidence for a new mutational mechanism?. *Genetics* **184**, 233–241. https://doi.org/10.1534/genetics.109.110510 (2010).
57. Venkat, A., Hahn, M. W. & Thornton, J. W. Multinucleotide mutations cause false inferences of lineage-specific positive selection. *Nat. Ecol. Evol.* **2**, 1280–1288. https://doi.org/10.1038/s41559-018-0584-5 (2018).
58. Schofield, M. A., Agbunag, R. & Miller, J. H. DNA inversions between short inverted repeats in *Escherichia coli*. *Genetics* **132**, 295–302 (1992).
59. Lovett, S. T. Encoded errors: Mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Mol. Microbiol.* **52**, 1243–1253. https://doi.org/10.1111/j.1365-2958.2004.04076.x (2004).
60. Voineagu, I., Narayanan, V., Lobachev, K. S. & Mirkin, S. M. Replication stalling at unstable inverted repeats: Interplay between DNA hairpins and fork stabilizing proteins. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 9936–9941. https://doi.org/10.1073/pnas.0804510105 (2008).
61. Sinden, R. R., Zheng, G. X., Brankamp, R. G. & Allen, K. N. On the deletion of inverted repeated DNA in *Escherichia coli*: Effects of length, thermal stability, and cruciform formation in vivo. *Genetics* **129**, 991–1005 (1991).
62. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864. https://doi.org/10.1126/science.185.4154.862 (1974).
63. Bendl, J. *et al.* PredictSNP: Robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput. Biol.* **10**, e1003440. https://doi.org/10.1371/journal.pcbi.1003440 (2014).
64. Huerta-Cepas, J. *et al.* eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucl. Acids Res.* **47**, D309–D314. https://doi.org/10.1093/nar/gky1085 (2019).
65. Shen, X. X. *et al.* Tempo and mode of genome evolution in the budding yeast subphylum. *Cell* **175**, 1533.e1520-1545.e1520. https://doi.org/10.1016/j.cell.2018.10.023 (2018).
66. Dutra, B. E. & Lovett, S. T. Cis and trans-acting effects on a mutational hotspot involving a replication template switch. *J. Mol. Biol.* **356**, 300–311. https://doi.org/10.1016/j.jmb.2005.11.071 (2006).
67. Zhang, J. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *J. Mol. Evol.* **50**, 56–68. https://doi.org/10.1007/s002399910007 (2000).
68. Schrider, D. R., Houle, D., Lynch, M. & Hahn, M. W. Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. *Genetics* **194**, 937–954. https://doi.org/10.1534/genetics.113.151670 (2013).
69. Galen, S. C. *et al.* Contribution of a mutational hot spot to hemoglobin adaptation in high-altitude Andean house wrens. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 13958–13963. https://doi.org/10.1073/pnas.1507300112 (2015).
70. Harris, K. & Nielsen, R. Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome Res.* **24**, 1445–1454. https://doi.org/10.1101/gr.170696.113 (2014).

71. Ji, X., Griffing, A. & Thorne, J. L. A phylogenetic approach finds abundant interlocus gene conversion in yeast. *Mol. Biol. Evol.* **33**, 2469–2476. https://doi.org/10.1093/molbev/msw114 (2016).
72. Weinreich, D. M., Delaney, N. F., Depristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114. https://doi.org/10.1126/science.1123539 (2006).
73. Wright, S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc. Sixth Int. Congr. Genet.* **1**, 356–366 (1932).
74. Sauna, Z. E. & Kimchi-Sarfaty, C. Understanding the contribution of synonymous mutations to human disease. *Nat. Rev. Genet.* **12**, 683–691. https://doi.org/10.1038/nrg3051 (2011).
75. Cherry, J. M. *et al.* Saccharomyces genome database: The genomics resource of budding yeast. *Nucl. Acids Res.* **40**, D700–D705. https://doi.org/10.1093/nar/gkr1029 (2012).
76. Wapinski, I., Pfeffer, A., Friedman, N. & Regev, A. Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449**, 54–61. https://doi.org/10.1038/nature06107 (2007).
77. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends Genet.* **16**, 276–277 (2000).
78. Katoh, K., Asimenos, G. & Toh, H. Multiple alignment of DNA sequences with MAFFT. *Methods Mol. Biol.* **537**, 39–64. https://doi.org/10.1007/978-1-59745-251-9_3 (2009).
79. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033 (2014).
80. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320. https://doi.org/10.1093/molbev/msn067 (2008).
81. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306–314. https://doi.org/10.1007/BF00160154 (1994).
82. Ashkenazy, H. *et al.* FastML: A web server for probabilistic reconstruction of ancestral sequences. *Nucl. Acids Res.* **40**, W580–W584. https://doi.org/10.1093/nar/gks498 (2012).
83. Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A. M. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431–449 (2000).
84. Sela, I., Ashkenazy, H., Katoh, K. & Pupko, T. GUIDANCE2: Accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucl. Acids Res.* **43**, W7–W14. https://doi.org/10.1093/nar/gkv318 (2015).
85. Fletcher, W. & Yang, Z. INDELible: a flexible simulator of biological sequence evolution. *Mol. Biol. Evol.* **26**, 1879–1888. https://doi.org/10.1093/molbev/msp098 (2009).
86. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591. https://doi.org/10.1093/molbev/msm088 (2007).
87. Yang, Z., Wong, W. S. & Nielsen, R. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* **22**, 1107–1118. https://doi.org/10.1093/molbev/msi097 (2005).

Acknowledgements

We thank Tal Pupko and Haim Ashkenazy for discussions. E.H-C. is supported by the Israel Science Foundation Grant 605/20 and by the Open University of Israel Research Fund.

Author contributions

E.H.-C. conceived the study. E.H.-C. and M.A. designed the analysis. M.A. performed analysis. E.H.-C. finalized the manuscript. E.H.-C. managed the project and recruited the funding. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-01736-y.

Correspondence and requests for materials should be addressed to E.H.-C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.