The evolution of short inverted repeats

Bar Lavi, May Abraham, and Einat Hazkani-Covo einatco@openu.ac.il

Department of Natural and Life Sciences, The Open University of Israel, Ra'anana, Israel

Abstract

Inverted repeats (IRs) are sequences with internal symmetry that form non-canonical DNA structures and can induce genome instability. Diverged IRs frequently undergo template switches, in which one arm of the repeat serves as a template for synthesis of the second arm resulting in erasing of variation between arms. In contrast to other mechanisms that resulted in the correction of IR arms during evolution, the evolutionary impact of template switching was previously neglected. If template switching occurs in genomes, then we expect it to contribute to the correction of imperfect IRs to perfect ones, resulting in the conservation of IRs through evolution. In addition, Template switching is a non-conservative mutation mechanism that introduces multi nucleotide mutations at once. Thus, it has the potential to introduce functional changes into genomes. Our analysis shows that short IRs are conserved during the evolution of *Saccharomyces cerevisiae and E. coli*, supporting the model of IR arm corrections by template switching.

Template switching

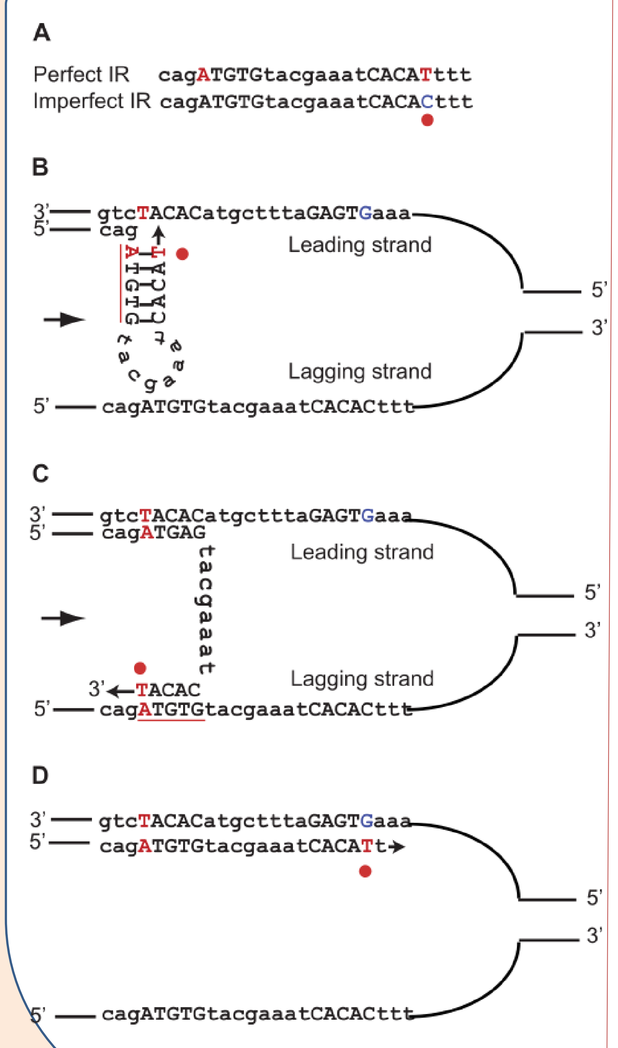
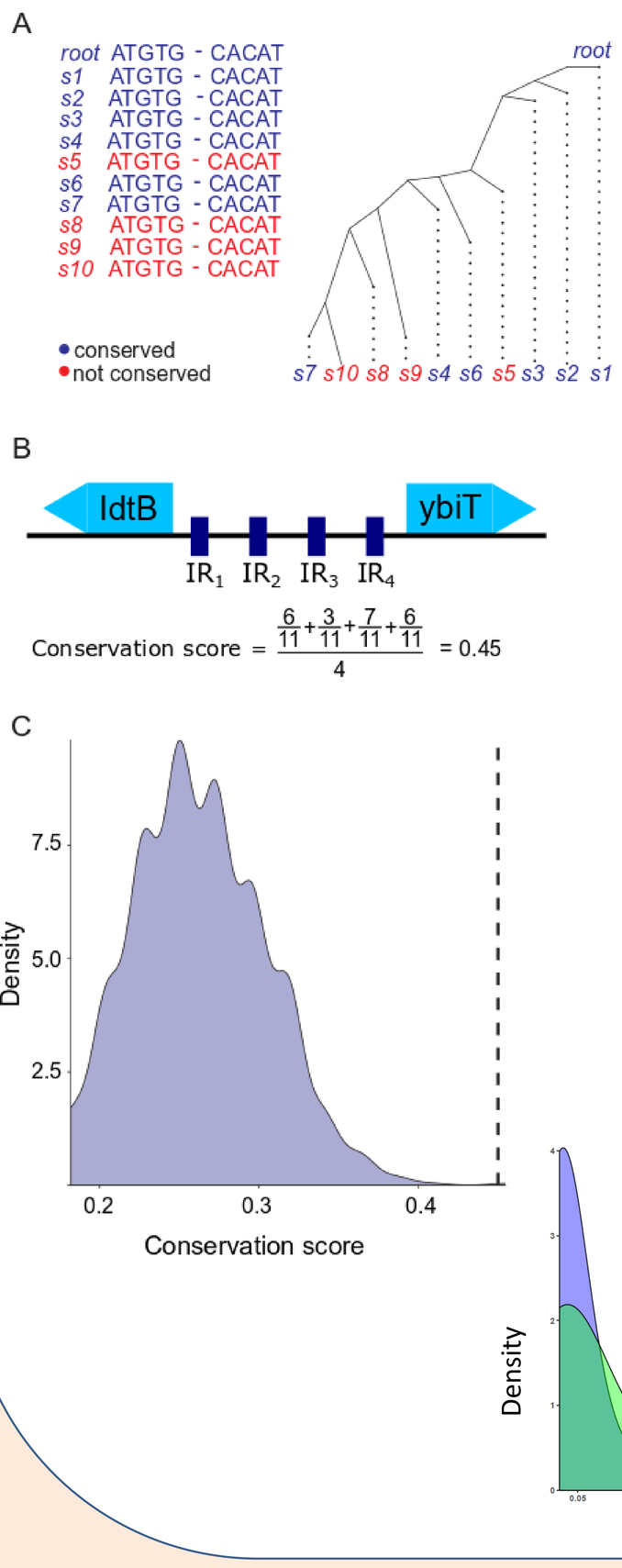


Fig 1. Template switching converts an imperfect IR to a perfect one. (A) The upper and lower sequences represent a perfect and an imperfect IR, respectively, located in an orthologous locus in two genomes. (B) The first switch under intramolecular template switching. Here the nascent strand is used as template. (C) The first switch under intermolecular template switching. Here the strand across the fork is used as template. (D) The second switch returns the nascent strand into the original template, resulting in a perfect IR as represented by the upper sequence in A. Upper case letters represent the IR arms while red dots represent mismatches between the arms. The noncanonical template is marked with a red line. The direction of the replication fork is indicated with an arrow.

IRs appear more conserved than their immediate environment in proteobacteria non coding regions



To test the conservation of IRs we collected orthologous sequence units for the *E. coli* MG1655 genome across 20 additional proteobacteria. Each of the orthologous sets of non coding regions was aligned using MAFFT and a maximum likelihood tree was reconstructed with PhyML.

We searched MG1655 for IRs with an arm length of at least 5 bp using the EMBOSS palindrome package. In our search, we allowed a spacer of up to 70 bp between the two IR arms. Out of 27,678 IRs in non coding regions of MG1655 we were able to identify 914 IRs that appear in at least 10 species. These orthologs reside in 234 non coding regions. For each of these 234 regions, we computed a conservation score (as shown on left).

We observed high conservation of perfect IRs: out of the 234 examined orthologous regions, 145 were more conserved than expected, which is statistically significant even after correcting for multiple testing. Our results together with previous experimental findings support a model in which imperfect IRs are corrected to perfect IRs in a preferential manner via a template switching mechanism.

Fig 2. Conservation analysis of IRs in shown on left. (A) An example alignment of an IR and its mapping onto its corresponding phylogenetic tree, with the IR of the MG1655 as the root sequence. The IR conservation score is 7/11, since 7 out of 11 sequences are identical to the root IR. (B) Conservation score computation for the entire NC region, located between the ldtB and the yblT genes in theMG1655 genome, which contains three additional IRs. (C) Analysis of conservation of IRs in the region located between the ldtB and the yblT genes. The distribution of 1,000 conservation scores computed using simulated data is shown in blue and the conservation score value computed from real data is shown as a bold dashed line. The figure explained the mechanism is shown on the top of the poster.

Fig 3. Comparison of conservation significance between IRs and control regions that reside 20bp apart. An empirical p-value was computed for each region based on its 1,000 corresponding simulated datasets. Shown in blue are the P-values for the IR regions and in green for the control regions.

Multiple IR variants are observed in orthologous loci in yeast proteins

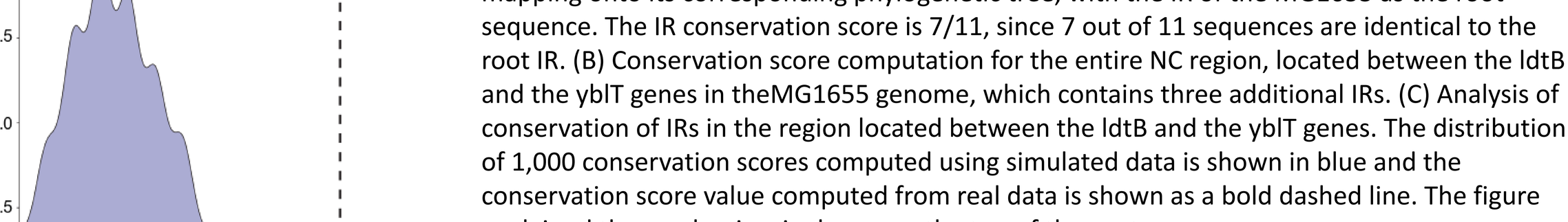


Fig 4. Evidence for template switching during evolution of *S. cerevisiae*. (A) A phylogenetic tree of seven Saccharomyces genomes. (B) Three perfect IRs appear in S288c three coding regions. Different perfect IRs appear at the same loci in three additional genomes. Each locus is shown in a column and different colours show different palindrome forms. Loci that show non perfect IRs are shown in grey. Cases in which the loci could not be identified are shown in black. Positions identical to S288c are shown with a dot and the spacer between arms is represented with a tilde sign. (C) Ninety-two S288c loci that appear in two additional genomes.

Summary Template-switching mechanism that causes short perfect IRs occurs through evolution of *S. cerevisiae* and *E. coli.*

Omer S, Lavi B, Mieczkowski PA, Covo S, Hazkani-Covo E (2017). G3 7(11):3775-3787; Lavi B, Levy Karin E, Pupko T, Hazkani-Covo E (2018). GBE 3(10):918-927