

Supporting Information

Lorenzen *et al.* 10.1073/pnas.0800444105

```
1  TATACTAGTTTATCAAAAAAGCGGAACATAGAAATTGTGTAGGTAAAGAACTGAACTTTGTGAAAATGTTGTTACCCACCCAAAAATAAATTGATAA
101  CAGTTTCTTAGCCATTCTCAAAAAACAACCCCTTTAAGAAGGTTTTCTAGTCTTTCTGGGGTTCAGTATATGCTTTTCTGTGGCGATAACTATG
      M P P V R R Q P R Y Q Q L T P F E R G R I
201  AAAAATTTGTTTTCTGTGTTTTATTGGGCACTAACAGTATGCCTCCCGTACGAAGACAACCACGTTACCAGCAGTTAACCCCTTTGAGCGGGGGCGAAT
      V G L R E G G M S V R E I A A R V N R G V A T V L R C I R A W E E
301  GTCGGGCTACGTGAAGTGGCATGTCGGTTCGGGAAATTGCAGCTCGTGTGAATCGTGGAGTGGCTACTGTCCTGCGATGTATTGGGCATGGGAGGAAG
      E G R E H R A R G S G R P R G T T A R Q D R Y L H F L A F R D R H V
401  AAGGGAGAGAACATCGAGCGAGAGGTTCTGGACGGCCACGAGGTACTACCGCAGCCAGGATCGGTACCTCCACTTTTAGCCTTTTCGAGACCGGCATG
      S T R R I G D Q W Y A A K G R P V T M A T V Y R R I R S F G L H S
501  CTCCACACGGCGGATTGGTGACCAGTGGTATGCTGCAAAGGGTTCGACCAGTTACAATGGCAACTGTATATAGACGAATCAGATCTTTTCGGGCTTCACTT
      Y R P H L V L P L T P Q Q R Q H R L D W C R A R E N W D L E W N S
601  TATCGTCCGCATTTGGTGCTCCCTTAACTCCTCAACAACGACAACATCGACTTGACTGGTGTCTGTCGCGAGAGAATTGGGATCTAGAATGGAATTCG
      V V F S D E S R F C L G M H D G R Q R V R R V R G E R * N V A F S V
701  TGGTCTTTTCGGATGAATCGCGGTTTTGTTTGGGCATGCATGATGGTTCGACAAAGAGTTAGAAGGGTACGTGGGGAACGGTGAAACGTGGCCTTTTCTGT
      E L P V A R T V G V M I W G A I A Y D S R S P L V F I E G S M T A
801  GGAATCCCTGTAGCTCGAACAGTGGGTGCATGATTTGGGGCGCATAGCCTATGATAGTAGGTCACCCCTTGTTCATTGAGGGATCCATGACTGG
      Q R Y V Q E V L E P V A V P Y V Q T I E N A S F Q Q D N A T P H S
901  CAGCGCTATGTACAGGAGGTTCTGGAACCTGTCGCAGTACCATATGTGCAAACTATTGAAAACGCGTCGTTTCAACAGGATAACGCCACGCCACACTCAG
      A R F T L R Y L E E V Q V Q V L P W P P R S P D L S P I E H I W D S
1001  CACGCTTACGTTGAGGTACTTAGAAGAAGTTCAGGTGCAAGTCTTCTCTGGCCGCTCGATCGCCTGACCTCTCGCCGATCGAGCATATGGGATT
      I G R R V T N L P Q P P Q T L A D L R R E I L T A W E A L P Q D E
1101  AATAGGTCGGCGAGTGACGAATTTACCCAGCCTCCACAAACGCTAGCGGACCTGCGACGCGAAATTTGACTGCTTGGGAGGCCCTGCCCAAGACGA
      I N H L I R S M P R R V A E C I H A R G G P T H Y *
1201  ATTAATCATTTAATAGAGTATGCCACGGAGGTTGCAGAGTGTATACATGCACGTGGAGGGCCAACCCATTATTAAGTTTTTTTTGTTAAAAATTTCTA
1301  TTTTCAATTACACTTTTTTTCATCATTATGTTTACTCTACCACCATTTTATATTTTTAATTTCATAAATTCACGCATTATTTCTGGGTGTTCCGCT
1401  TTTTTTGATAAGTAGTATA
```

Fig. S1. DNA sequence and conceptual translation of the *Tc1-1* element in the *M'* insertion. The sequence begins and ends with the duplicated TA target motif in bold. The 35-nt inverted terminal repeats (ITRs) are underlined, with palindromic regions shaded. Authentic (TAA) and premature (TGA) stop codons are in bold. The otherwise fully intact ORF is interrupted by a single C→T transition that converts R to the latter stop codon at nucleotide 781. The diagnostic residues of the catalytic DD34E motif are boxed.

```

1 80
Tc1-1 MPPVRRQPRYQQLTPFERGRIVGLREGGMSVREIAARVNRGVATVLRICIRAWEEEGREHRARGSGRPRGTTARQDRYLHF
Tc1-2 -----
Tc1-3 -----
81 160
Tc1-1 LAFRDRHVSTRRIGDQWYAAKGRPVTMATVYRRIRSFGLHSYRPHLVLPLTPQORQHRLDWCRARENWDLEWNSVVFSDIE
Tc1-2 -----E-----
Tc1-3 -----E-----
161 240
Tc1-1 SRFCLGMHDGRQVRVRRVGER*NVAFSVELPVARTVGVMIWGAIAYDSRSPLVFIEGSMTAQRVYVQEVLEPVAVPYVQTI
Tc1-2 -----R-----R-----G-----L-----
Tc1-3 -----R-----R-----G-----L-----
241 320
Tc1-1 ENASFQQDNATPHSARFTRLRYLEEVQVQLPWPPRSPDLSPIHIIWDSIGRRVTNLPQPPQTLADLRREILTAWREALPQD
Tc1-2 -----A-----
Tc1-3 -----T-----
321 346
Tc1-1 EINHLIRSMPPRRVAECIHARGGPTHY*
Tc1-2 -----*
Tc1-3 -----*

```

Fig. S2. Alignment of Tc1 proteins. Tc1-1: translation of transposase gene in the *M*^I-associated *Tc1* insertion (LG3). Tc1-2 and Tc1-3 are closely related elements found in a non-*M*^I (GA2) genome. Tc1-2: translation of transposase gene found within an intron of EST DT789639 (GLEAN.00717; LG2). Tc1-3: translation of transposase gene in the *Tc1* element corresponding to GLEAN.02110 (LG unknown). *, stop codon. Dashes indicate identity with Tc1-1. Note premature stop codon in Tc1-1 at residue 182. Boxed residues: Tc1-DD34E motif [Shao H, Tu Z (2001) Expanding the diversity of the IS630-Tc1-mariner superfamily: Discovery of a unique DD37E transposon and reclassification of the DD37D and DD39D transposons. *Genetics* 159:1103–1115]. Black underline: transposase_5 DNA-binding domain (pfam01498). Blue underline: COG3415 DNA binding domain.

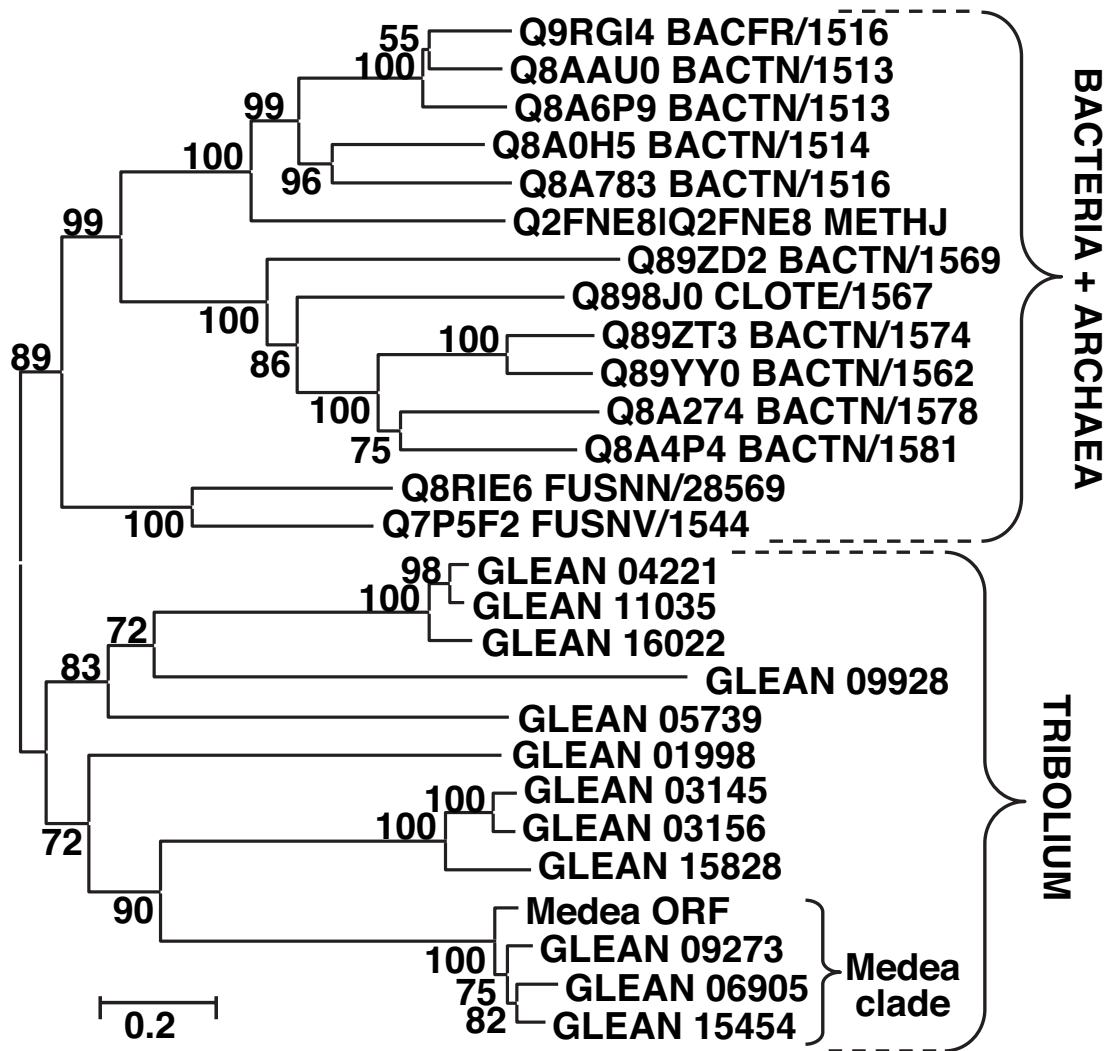


Fig. S4. Phylogenetic analysis of DUF1703 proteins. Shown are unrooted neighbor-joining trees based on multiple alignment of complete amino acid sequences with bootstrap values for 1,000 replications indicated. Accession numbers are given for the representative 14 bacterial/archaeal proteins, and GLEAN numbers are given for *Tribolium* proteins. Scale indicates distance, rate of change per amino acid sequence.

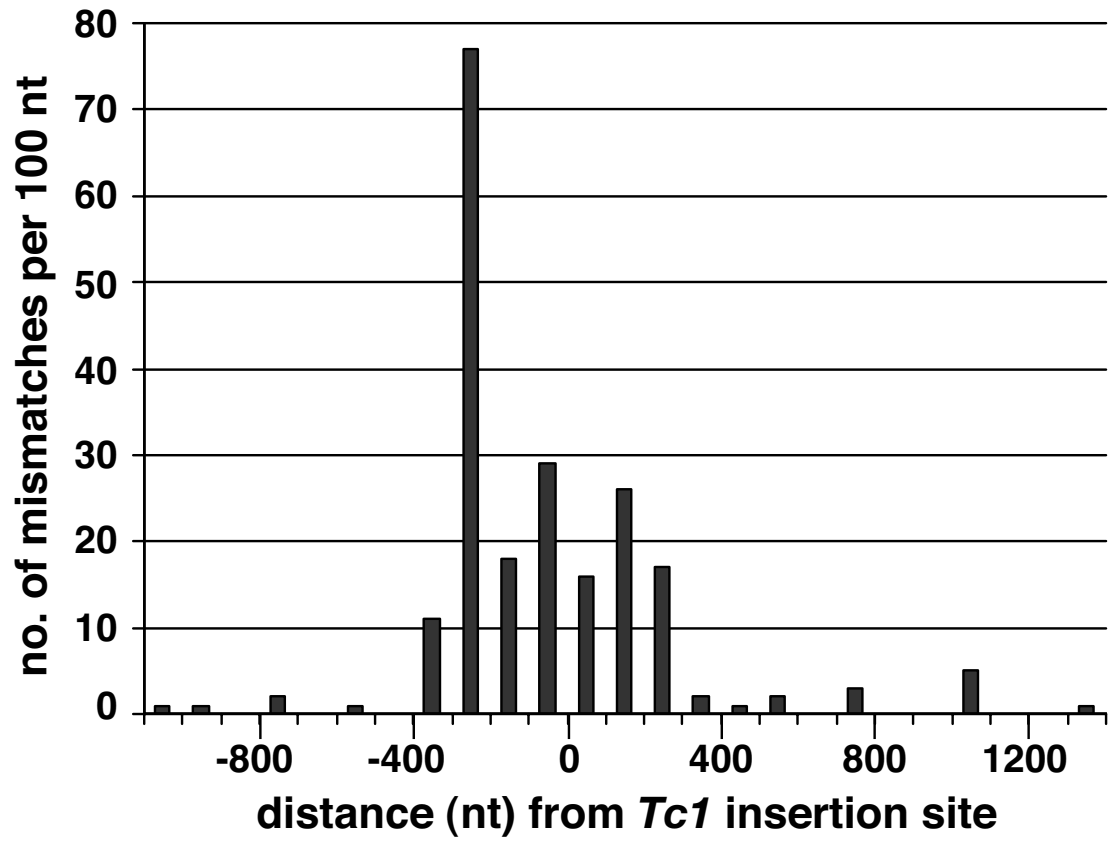


Fig. S55. *M^I*-associated sequence variation flanking the *Tc1* insertion site. Homologous *M^I* and non-*M^I* (GA2) sequences were aligned by using AlignX. The x axis is the distance in nucleotides from the *Tc1* insertion point. The y axis represents the number of nonidentical residues in each contiguous 100-nt segment. The large strain difference in the segment 200–300 nt centromeric (to the left) of the insertion point reflects the copy number difference in the tandem repeat (five copies in non-*M^I*, approximately two copies in *M^I*).