

- 13 Gene Ontology Consortium, (2001) Creating the gene ontology resource: design and implementation. *Genome Res.* 11, 1425–1433
- 14 Hamalainen, H.K. *et al.* (2001) Identification and validation of endogenous reference genes for expression profiling of T helper cell differentiation by quantitative real-time RT-PCR. *Anal. Biochem.* 299, 63–70
- 15 Lee, P.D. (2002) Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. *Genome Res.* 12, 292–297
- 16 Karolchik, D. *et al.* (2003) The UCSC genome browser database. *Nucleic Acids Res.* 31, 51–54
- 17 Akashi, H. (2001) Gene expression and molecular evolution. *Curr. Opin. Genet. Dev.* 11, 660–666
- 18 Duret, L. and Mouchiroud, D. (1999) Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4482–4487
- 19 Moriyama, E.N. and Powell, J.R. (1998) Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. *Nucleic Acids Res.* 26, 3188–3193
- 20 Urrutia, A.O. and Hurst, L.D. (2001) Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics* 159, 1191–1199

0168-9525/03/\$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved.  
doi:10.1016/S0168-9525(03)00140-9

## Strand misalignments lead to quasipalindrome correction

Vera van Noort<sup>1</sup>, Peder Worning<sup>2</sup>, David W. Ussery<sup>2</sup>, William A. Rosche<sup>3</sup> and Richard R. Sinden<sup>4</sup>

<sup>1</sup>Nijmegen Center for Molecular Life Sciences, P/A Center for Molecular and Biomolecular Informatics, Nijmegen, The Netherlands

<sup>2</sup>Center for Biological Sequence Analysis, The Technical University of Denmark, DK-2800 Lyngby, Denmark

<sup>3</sup>Department of Biological Science, The University of Tulsa, Tulsa, Oklahoma 74104–3126, USA

<sup>4</sup>Laboratory of DNA Structure and Mutagenesis, Center for Genome Research, Institute of Biosciences and Technology, Texas A and M University System Health Sciences Center, Houston, TX 77030, USA

**Quasipalindromes, or imperfect inverted repeats, undergo spontaneous mutation to more-perfect inverted repeats. These mutations have been observed in many organisms, ranging from bacteria to humans, where they are associated with mutations leading to disease. We determined the relative frequency of quasipalindromes and perfect palindromes in more than 100 sequenced prokaryotic genomes. In nearly all cases, perfect palindromes were relatively more frequent than quasipalindromes, suggesting that quasipalindrome correction is a general mechanism for mutation in prokaryotes.**

Apart from simple misincorporation mutations, primer-template misalignments are probably the predominant cause of spontaneous mutations, and can lead to different types of mutation, such as frameshifts, deletions, duplications, inversions and complex mutations [1,2]. Misalignment requires sequence complementarity, such as direct or inverted repeats. Simple misalignment can occur along a linear template, and complex misalignment can be directed by DNA secondary structure. Imperfect inverted repeats, or quasipalindromes, can undergo spontaneous mutation to form a perfect inverted repeat (Fig. 1). Such mutations have been observed in bacteriophage T4, yeast and prokaryotes [3]. In addition, they have been associated with several human genetic diseases, including hereditary angioneurotic oedema, Duchenne muscular dystrophy, osteogenesis imperfecta, Lesch–Nyhan syndrome, and familial hypertension [4]. Here we provide evidence that a complex mutation, the correction of a quasipalindrome

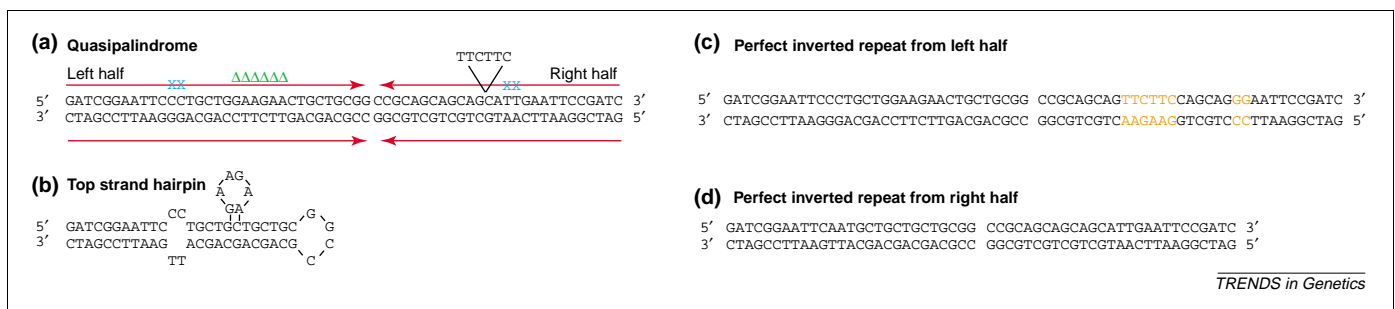
(an imperfect inverted repeat) to a palindrome (a perfect inverted repeat), occurs frequently in prokaryotes.

In 1982, Ripley [5] proposed two models for the correction of quasipalindromes to perfect inverted repeats: the intramolecular strand-switch model (also known as the hairpin-correction model), and the intermolecular strand-switch model (Fig. 2). In the latter model, the unpaired 3' end of the nascent strand pairs with the quasipalindrome in the opposite template strand; that is, hybridization from the leading to lagging template strand. We have demonstrated that quasipalindrome correction in *Escherichia coli* occurs from a misalignment that is caused by an intermolecular strand switch, preferentially during leading strand replication [6]. Other quasipalindrome correction mutations might occur by an intramolecular (hairpin) replication mechanism [7].

### Frequency of quasipalindromes within complete genomes

One would expect that correction of quasipalindromes over a period of time should result in an increase in the frequency of perfect palindromes within the bacterial genome. The repetitiveness of genomes has been investigated before [8], as well as frequencies of specific repeats in single genomes [9,10]. Here we test sequenced bacterial genomes for the frequencies of quasipalindromes and perfect palindromes, and compare them with expected values. In Fig. 3, we show that the relative frequencies of perfect palindromes are generally higher than the relative frequencies of quasipalindromes. In particular, the relative frequency of perfect palindromes in *E. coli* (orange triangle) is much higher than that of quasipalindromes.

Corresponding author: David W. Ussery (dave@CBS.dtu.dk).



**Fig. 1.** Quasipalindromes and quasipalindrome correction mutations. (a) A model quasipalindrome (i.e. an imperfect inverted repeat). The left half contains a CC dinucleotide, whereas the right half contains a TT dinucleotide (denoted by XX above the sequences). In addition, the left half contains the sequence GAAGAA ( $\Delta\Delta\Delta\Delta$  above the sequence). This sequence is missing in the right half, and its position is denoted by the sequence above the brackets pointing to the site where from which the sequence is missing. (b) The top strand of the quasipalindrome is shown in its hairpin conformation in which the 2-bp mismatch and the GAAGAA loop are evident. (c,d) Left–left (c) and right–right (d) quasipalindrome correction, in which perfect inverted repeats have been created. In (c) the bases that are inserted are shown in orange.

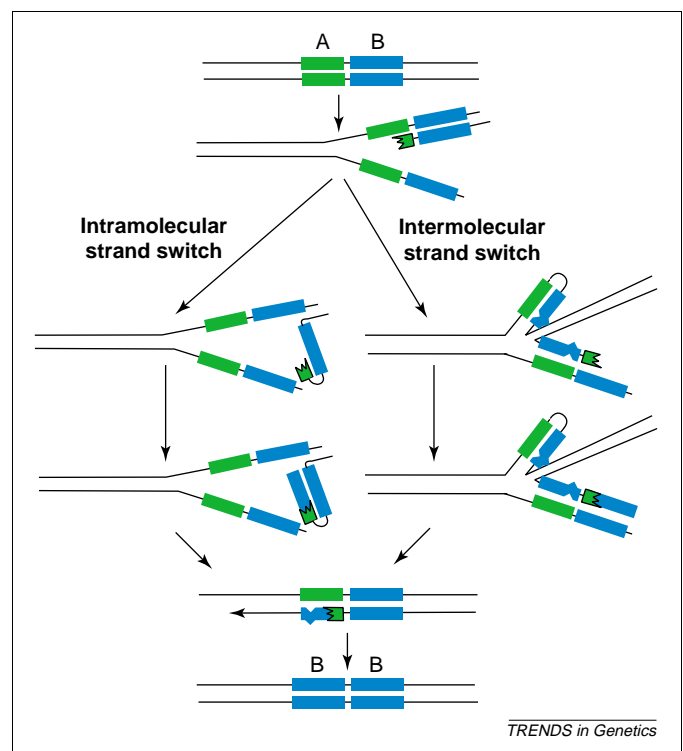
Relative frequencies are calculated by dividing ‘observed’ by ‘expected’ frequencies. This supports quasipalindrome correction being a general mechanism for mutation.

We define a quasipalindrome as an inverted repeat within a 30-bp window where six out of seven bases are complementary, and a perfect palindrome as a perfect inverted repeat of seven bases within a 30-bp window. These values are based on experiments on the size and stability of cruciforms [11]. A complete table with results of analysis of 106 prokaryotic genomes (16 archaeal and 90 bacterial) can be found following the ‘Microbial Database Tables’ link from the main atlas web page: <http://www.cbs.dtu.dk/services/GenomeAtlas/>. The ‘observed’ frequencies are compared with what would be expected given the length and dinucleotide composition of the genome. The ‘expected’ quasipalindrome and perfect palindrome frequencies were measured as the average frequencies in 20 DNA sequences, of the length of the original genome, generated by a first order (dinucleotide) Markov model trained on the original genome [12]. The relative standard deviations are typically  $\sim 0.5\%$  of the average, meaning that an observed:expected ratio of more than 1.01 or less than 0.99 differs significantly from the expected value. Expected frequencies were also calculated based on zeroth order (mononucleotide) Markov models, in which case the differences between expected and observed frequencies were larger. The expected frequencies did not change significantly when using second-order (trinucleotide) instead of first-order Markov models.

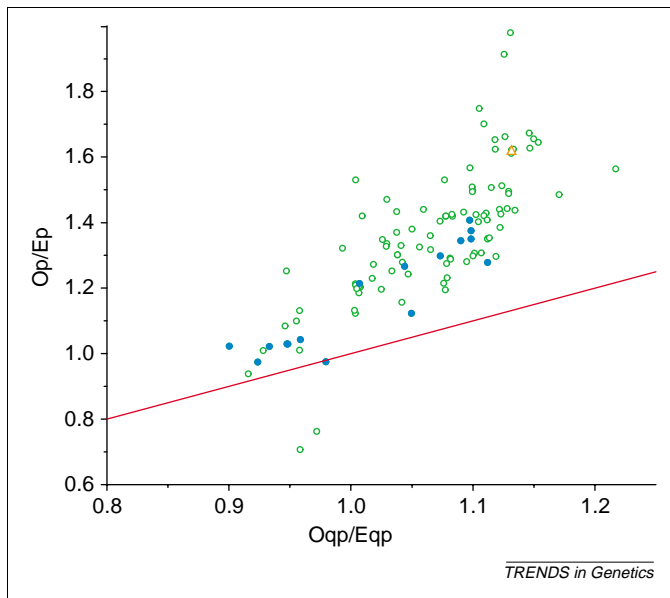
The highest number of quasipalindromes is found in *Ureaplasma urealyticum*, where they make up 50.4% of the chromosome; this is to be expected, as this bacterium has the most extreme A + T content of the genomes in the list. When the A + T content has an extreme value (i.e. far away from 50%), the genomic sequence is restricted, corresponding to a lower number of different nucleotides (e.g. only two rather than four nucleotides in the extreme case). This results in a higher probability of inverted (and direct) repeats. The pattern of the expectation numbers, as a function of A + T content in Fig. 4 shows this trend quite clearly.

The number of quasipalindromes observed in sequenced genomes is consistently higher than would be expected from the dinucleotide composition. For quasipalindromes, only seventeen genomes show a value that is lower than the expected, whereas most of the remaining

genomes have values much higher than the expected. The same is true for the perfect palindromes, where only five genomes have a lower percentage of palindromes than expected. Significantly, the ratio of the observed to the expected values for perfect palindromes (Op/Ep in Fig. 3) is almost always higher than the ratio for quasipalindromes (Oqp/Eqp), whereas random mutagenesis of a quasipalindrome would predict a decrease in palindromic symmetry. The observation of larger observed:expected values for perfect palindromes than for quasipalindromes for 103 out



**Fig. 2.** Models for quasipalindrome correction. The quasipalindrome is represented as the shaded rectangles: A is a short arm, and B is a long arm. In the intramolecular strand switch model, replication occurs through the centre of the quasipalindrome. Intrastrand misalignment of cDNA sequences creates a hairpin. By copying the long arm B the A arm is converted to a B arm. Realignment of the hairpin sequences to the leading template strand creates a heteroduplex in the left half of the quasipalindrome. Following the next round of replication, the plasmid derived from the bottom strand contains a perfect B–B inverted repeat. In the intermolecular strand switch model, only the misalignment is different, occurring between the long B arm of the quasipalindrome in the progeny leading strand and the short A arm of the quasipalindrome in the lagging template strand. Continued synthesis of the quasipalindrome produces the same mutational event as the intramolecular strand switch; that is, conversion of the A arm to a B arm.



**Fig. 3.** The frequencies of quasipalindromes and perfect inverted repeats were calculated, as described in the text. Oqp, observed quasipalindromes, as a percentage of the whole genome; Eqp, expected quasipalindromes, based on the dinucleotide base composition; Op, observed palindromes, as a percentage of the whole genome; Ep, expected palindromes, based on the dinucleotide base composition. The red line is where Oqp/Eqp is equal to Op/Ep. Open green circles indicate Bacteria, filled blue circles indicate Archaea. The orange triangle is *Escherichia coli*.

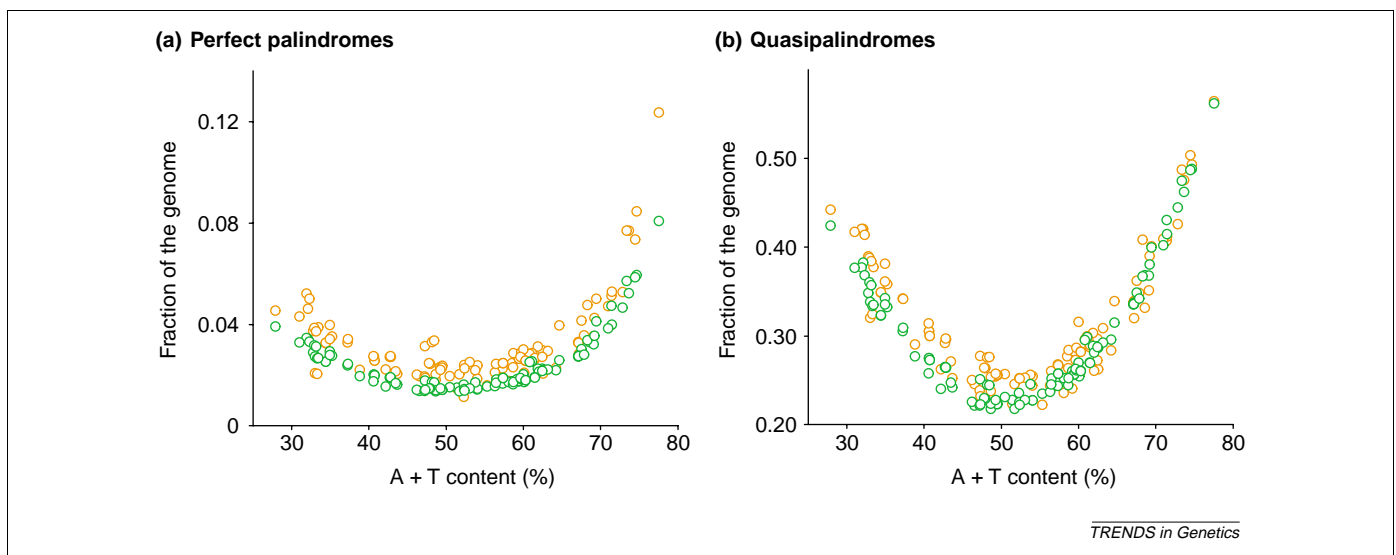
of the 106 genomes examined could reflect the occurrence of spontaneous strand switch mutations creating perfect palindromes from quasipalindromes. Another possibility is that several specific palindromes are genetically selected for a function like protein binding sites.

The archaeal genomes (blue filled circles in Fig. 3) tend to have quasipalindrome and perfect palindrome frequencies closer to the expected values than the bacterial genomes. Furthermore, within the bacterial genomes, the *Neisseria* species have the largest deviation from

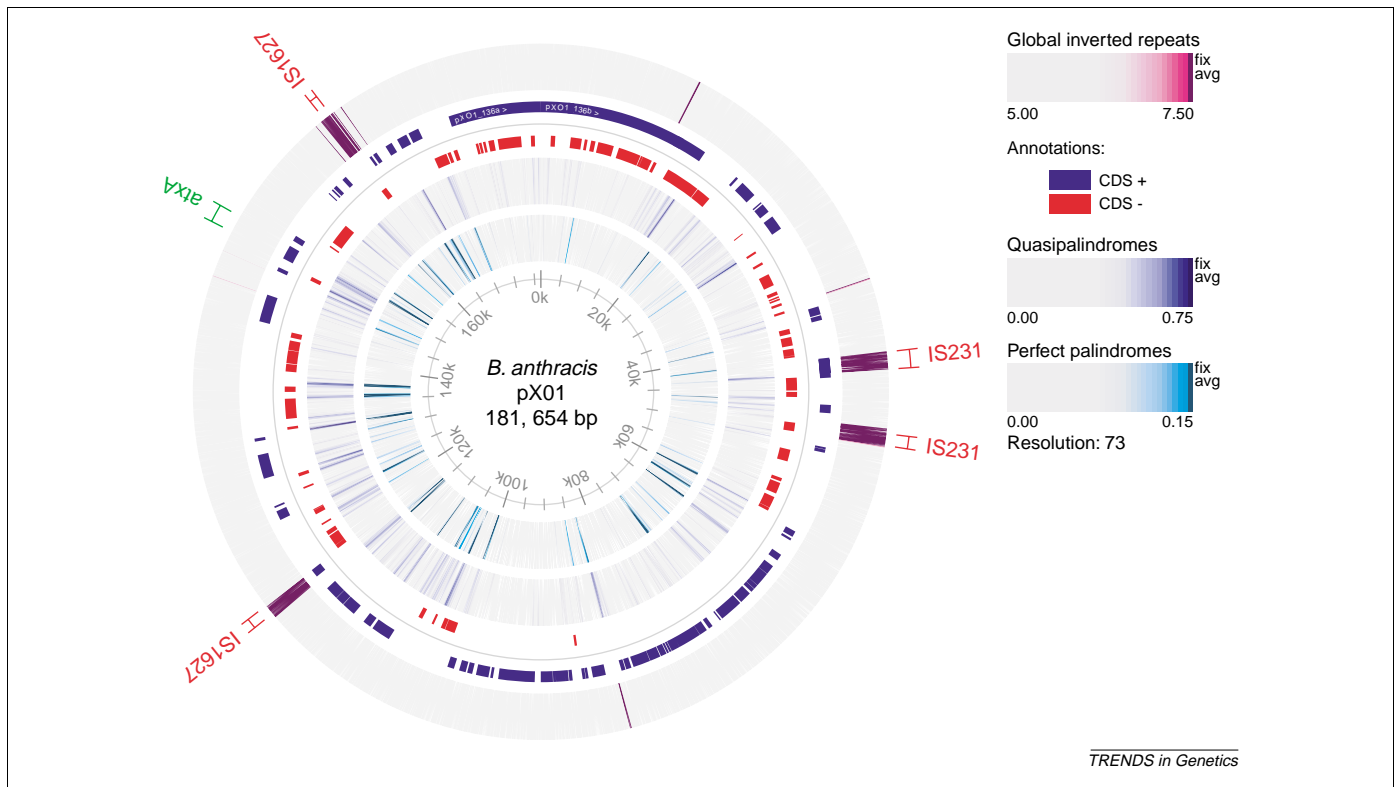
expectation; that is, twice as many perfect palindromes as expected, whereas *Archaeoglobus*, *Deinococcus* and *Synechocystis* seem to have consistently fewer palindromes than expected. The gamma subdivision of proteobacteria (of which *E. coli* is a member) also has higher than expected frequencies of quasipalindromes and perfect palindromes. This difference might reflect the inherent propensity for the occurrence of this type of mutation by the replication apparatus of the individual bacteria.

### Cruciform atlases for localization of palindromes with genomes

Palindromes are not homogeneously distributed throughout the chromosome, and the localization of quasipalindromes and perfect palindromes can be plotted in a 'DNA atlas' format, as described previously [13–15]. Figure 5 shows a cruciform atlas for *Bacillus anthracis* plasmid pX01 [16]. The left-hand side of the figure has a higher fraction of palindromes than the rest of the genome. This region contains a 44 800-bp pathogenicity island, located between the two IS1627 insertion sequence (IS) elements marked in the outer 'global inverted repeats' circle. These sites represent known regions of inversion for different isolates of pX01 plasmids. Although in general many of the quasipalindromes and perfect palindromes are in intergenic regions (often corresponding to stem-loop structures in mRNA at the 5' ends of genes), a few genes contain a high fraction of palindromes within the coding region. For example, in Fig. 5 it can be seen that the anthrax toxin activator gene *atxA* (located at ~150 kbp) has a high number of quasipalindromes. This could result in a higher mutational frequency for this gene, and variability in the amino acid sequence of this protein might aid in the bacteria's ability to evade the immune system. Cruciform Atlases are available on our web pages, for archaeal



**Fig. 4.** Fraction of the genome containing local quasipalindromes. (a) The observed (orange) and the predicted (green) fraction of the genome that is part of a perfect inverted repeat, plotted against the A + T content of the genome. A 'perfect palindrome' for the purposes of this analysis is an exact match of any 7-bp piece of DNA within a 3-bp window. (Thus, these repeats can have asymmetric centers of up to 16 bp.) (b) The observed (orange) and the predicted (green) fraction of the genome that is part of a quasipalindrome, plotted against the A + T content of the genome. Quasipalindromes were calculated as the best match of a 7-bp window in a 30-bp length of DNA. The cut-off value for scoring a quasipalindrome was 80%, which allows one mismatch in the 7-bp window. To calculate the expected value for quasipalindromes and perfect inverted repeats (in (a) and (b)), 20 random DNA sequences with the same nucleotide composition and length were generated and analysed in the same fashion as the entire genome. The values are mean values with a relative standard deviation of ~0.5%.



**Fig. 5.** Cruciform atlas for *Bacillus anthracis* pX01. The cruciform atlas was constructed from the pX01 GenBank file (AF065404). The quasipalindromes and perfect palindromes are calculated as described in the text, and other repeats are calculated as before [19]. The genes containing insertion sequence (IS) elements are shown in red. Abbreviations: CDS, coding sequence; avg, average.

([http://www.cbs.dtu.dk/services/GenomeAtlas/cruciform/index\\_Archaea\\_Organism.html](http://www.cbs.dtu.dk/services/GenomeAtlas/cruciform/index_Archaea_Organism.html)) and bacterial ([http://www.cbs.dtu.dk/services/GenomeAtlas/cruciform/index\\_Bacteria\\_Organism.html](http://www.cbs.dtu.dk/services/GenomeAtlas/cruciform/index_Bacteria_Organism.html)) genomes; the web tables also contain further information and references to the different prokaryotic genomes used.

#### Potential for DNA directed mutational change at quasipalindrome

Quasipalindrome correction mutations constitute one class of mutation hotspot. The DNA symmetry elements allow the formation of DNA secondary structures that promote mutation. In addition, the inverted repeat nature of the sequence provides the opportunity that during leading strand replication of the quasipalindrome, a second complementary copy of the template exists in single stranded form in the lagging strand template, which might also contribute to the high frequency of mutation at quasipalindromes. The outcome of this spontaneous mutational event is the formation of more-perfect inverted repeats. Long perfect inverted repeats (> 50 bp) are also genetically unstable, especially those that include direct repeats at their ends. Presumably, instability occurs by primer template misalignment during replication, with the misalignment stabilized by hairpin formation [11,17–21]. The outcome for spontaneous mutations involving long perfect inverted repeats is deletion. Thus, on an evolutionary basis, it might seem that there is a driving force for weak quasipalindromes to become longer, more-perfect inverted repeats and then for these sequences to be deleted from the genome. The

remarkably high frequency of perfect palindromes in bacterial genomes argues that spontaneous mutation involving quasipalindromes could be a frequent event. This hypothesis is supported by demonstration of quasipalindrome-associated mutational hotspots in *E. coli* [7,22]. The observation that the ratio of observed to expected short perfect inverted repeats (7/7 bp stems, which as defined in this analysis can have asymmetric centers) is higher than the ratio for quasipalindromes (6/7 bp stems) suggests that these repeats are stable and that they do not accumulate by chance.

#### Acknowledgements

Support for research on quasipalindrome correction mutations was provided by grant ES 05508 from the National Institute of Environmental Health Sciences, National Institutes of Health to R.R.S. D.W.U. and P.W. are supported by a grant from the Danish Research Foundation. Work in the W.A.R. laboratory is supported by a grant from the Bovaird Center for Studies in Molecular Biology and Biotechnology. The authors thank Hans Henrik Stærfeldt, Lars Juhl Jensen and Jacob L. Reimers for their help.

#### References

- 1 Streisinger, G. *et al.* (1966) Frameshift mutations and the genetic code. *Cold Spring Harb. Symp. Quant. Biol.* 31, 77–84
- 2 Ripley, L.S. (1990) Frameshift mutation: determinants of specificity. *Annu. Rev. Genet.* 24, 189–213
- 3 Ripley, L.S. and Shoemaker, N.B. (1982) Polymerase infidelity and frameshift mutation. *Basic Life Sci.* 20, 161–178
- 4 Bissler, J.J. (1998) DNA inverted repeats and human disease. *Front. Biosci.* 3, d408–d418
- 5 Ripley, L.S. (1982) Model for the participation of quasi-palindromic DNA sequences in frameshift mutation. *Proc. Natl. Acad. Sci. U. S. A.* 79, 4128–4132

- 6 Rosche, W.A. *et al.* (1997) Leading strand specific spontaneous mutation corrects a quasipalindrome by an intermolecular strand switch mechanism. *J. Mol. Biol.* 269, 176–187
- 7 Viswanathan, M. *et al.* (2000) A novel mutational hotspot in a natural quasipalindrome in *Escherichia coli*. *J. Mol. Biol.* 302, 553–564
- 8 Hancock, J.M. (2002) Genome size and the accumulation of simple sequence repeats: implications of new data from genome sequencing projects. *Genetica* 115, 93–103
- 9 Cox, R. and Mirkin, S.M. (1997) Characteristic enrichment of DNA repeats in different genomes. *Proc. Natl. Acad. Sci. U. S. A.* 94, 5237–5242
- 10 Saunders, N.J. *et al.* (1998) Simple sequence repeats in the *Helicobacter pylori* genome. *Mol. Microbiol.* 27, 1091–1098
- 11 Sinden, R.R. *et al.* (1991) On the deletion of inverted repeated DNA in *Escherichia coli*: effects of length, thermal stability, and cruciform formation *in vivo*. *Genetics* 129, 991–1005
- 12 Baldi, P. *et al.* (1994) Hidden Markov models of biological primary sequence information. *Proc. Natl. Acad. Sci. U. S. A.* 91, 1059–1063
- 13 Jensen, L.J. *et al.* (1999) Three views of microbial genomes. *Res. Microbiol.* 150, 773–777
- 14 Pedersen, A.G. *et al.* (2000) A DNA structural atlas for *Escherichia coli*. *J. Mol. Biol.* 299, 907–930
- 15 Petersen, L. *et al.* (2002) Visualization and significance of DNA structural motifs in the *Campylobacter jejuni* genome. *Genome Lett.* 1, 16–25
- 16 Okinaka, R.T. *et al.* (1999) Sequence and organization of pXO1, the large *Bacillus anthracis* plasmid harboring the anthrax toxin genes. *J. Bacteriol.* 181, 6509–6515
- 17 Weston-Hafer, K. and Berg, D.E. (1989) Palindromy and the location of deletion endpoints in *Escherichia coli*. *Genetics* 121, 651–658
- 18 Weston-Hafer, K. and Berg, D.E. (1991) Deletions in plasmid pBR322: replication slippage involving leading and lagging strands. *Genetics* 127, 649–655
- 19 Bissler, J.J. *et al.* (1994) Contiguous deletion and duplication mutations resulting in type 1 hereditary angioneurotic edema. *Hum. Genet.* 93, 265–269
- 20 Ketterling, R.P. and Sommer, S.S. (1994) Microdeletions in the factor IX gene: three recurrences associated with a quasipalindromic sequence. *Hum. Mol. Genet.* 3, 191–192
- 21 Pomponio, R.J. *et al.* (1996) Deletion/insertion mutation that causes biotinidase deficiency may result from the formation of a quasipalindromic structure. *Hum. Mol. Genet.* 5, 1657–1661
- 22 Rosche, W.A. *et al.* (1998) 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

0168-9525/03/\$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved.  
doi:10.1016/S0168-9525(03)00136-7

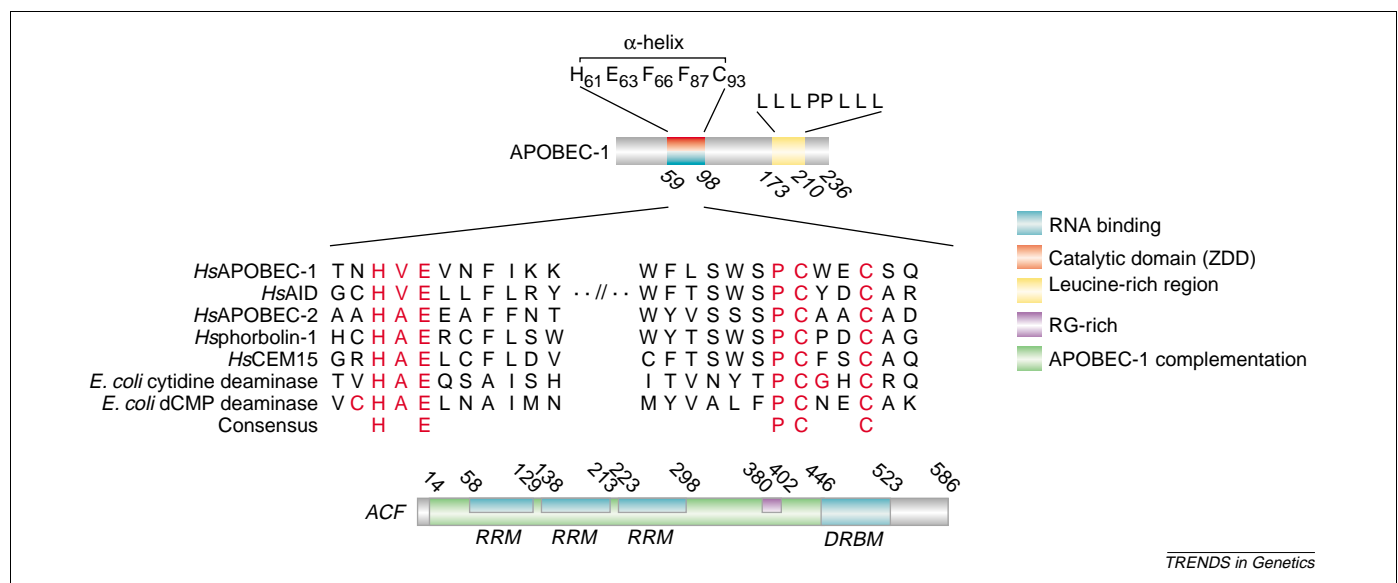
## Erratum

## Erratum: Messenger RNA editing in mammals: new members of the APOBEC family seeking roles in the family business

Wedekind, J.E. *et al.* (2003) *Trends in Genetics* 19, 207–216

There was a small mistake in the alignments in Fig. 1, for which the authors apologize. The correct figure should be:

0168-9525/03/\$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved.  
doi:10.1016/S0168-9525(03)00145-8



**Fig. 1.** Functional domains of APOBEC-1 and ACF. Conserved residues within the zinc-dependent deaminase domain (ZDD) are shown for APOBEC-1 and homologous cytidine deaminases. The catalytic domain of APOBEC-1 is characterized by a ZDD with three zinc ligands (either His or Cys), a glutamic acid, a proline residue and a conserved primary sequence spacing [17]. The ZDD of other deaminases and APOBEC-1 related proteins are shown for comparison along with a consensus ZDD. The indicated residues in the catalytic site of APOBEC-1 bind AU-rich RNA with weak affinity. The leucine rich region (LRR) of APOBEC-1 has been implicated in APOBEC-1 dimerization and shown to be required for editing [19,65] although structural analysis suggests that LRR forms the hydrophobic core of the protein monomer [67]. ACF complements APOBEC-1 through its APOBEC-1 and RNA-binding activities. The RNA recognition motifs (RRMs) are required for mooring-sequence-specific RNA binding, and these domains plus sequences flanking them are required for APOBEC-1 interaction and complementation [21,28]. APOBEC-1 complementation activity minimally depends on ACF binding to both APOBEC-1 and mooring sequence RNA. A broad APOBEC-1 complementation region is indicated that is inclusive of all regions implicated in this activity [21,28].