Molecular Characteristics of Spontaneous Deletions in the Hyperthermophilic Archaeon *Sulfolobus acidocaldarius*

Dennis W. Grogan* and Josh E. Hansen

Department of Biological Sciences, University of Cincinnati, Cincinnati, Ohio 45221

Received 20 September 2002/Accepted 18 November 2002

Prokaryotic genomes acquire and eliminate blocks of DNA sequence by lateral gene transfer and spontaneous deletion, respectively. The basic parameters of spontaneous deletion, which are expected to influence the course of genome evolution, have not been determined for any hyperthermophilic archaeon. We therefore screened a number of independent pyrimidine auxotrophs of *Sulfolobus acidocaldarius* for deletions and sequenced those detected. Deletions accounted for only 0.4% of spontaneous pyrE mutations, corresponding to a frequency of about $10^{-8}$ per cell. Nucleotide sequence analysis of five independent deletions showed no significant association of the endpoints with short direct repeats, despite the fact that several such repeats occur within the pyrE gene and that duplication mutations in pyrE reverted at high frequencies. Endpoints of the spontaneous deletions did not coincide with short inverted repeats or potential stem-loop structures. No consensus sequence common to all the deletions could be identified, although two deletions showed the potential of being stabilized by octanucleotide sequences elsewhere in pyrE, and another pair of deletions shared an octanucleotide at their 3′ ends. The unusually low frequency and low sequence dependence of spontaneous deletions in the *S. acidocaldarius pyrE* gene compared to other genetic systems could not be explained in terms of possible constraints imposed by the 5-fluoroorotate selection.

Sequence analysis of entire genomes has made it clear that despite the gradual accumulation of nucleotide substitutions, much of the divergence among prokaryotic species occurs through acquisition and loss of blocks of DNA sequences. Although bacterial genomes carry significant amounts of sequences originating in other lineages, related species exhibit similar genome sizes (21, 22), implying a general balance of losses against gains. Notable exceptions to this size constancy occur in species that have adopted a parasitic life strategy; the genomes of these species are either small or rapidly shrinking, reflecting the elimination of many genes whose functions have become superfluous (1, 6, 24). The mechanistic basis of genome reduction appears to be a short-term capacity to delete DNA that exceeds the capacity to acquire it through lateral gene transfer. This kinetic bias in favor of deletion implies that the maintenance of genome size seen in most bacterial lineages reflects selection for function of the vast majority of genes in the genome (24). It also explains the genome features observed in most free-living prokaryotes, including compactness (i.e., high open reading frame [ORF] density) and a dearth of full-length but nonfunctional ORFs (1, 24).

Data from genomes of hyperthermophilic archaea suggest a similar interplay among DNA acquisition, DNA loss, and selection for function. For example, *Pyrococcus* species can acquire large blocks of DNA by lateral transfer (9), yet they have also maintained relatively constant genome sizes during speciation (35). However, these and other archaea from geothermal biotopes have extreme evolutionary divergence from the organisms in which most DNA transactions have been analyzed at the molecular level, and they grow optimally at temperatures which destabilize the primary and secondary structure of DNA. Thus, the frequency with which deletions occur in nonessential sequences, the average size of individual deletions, and the influence of DNA sequences on the positioning of endpoints cannot be assumed to correspond to those of the bacterial or eukaryotic systems in which the process of spontaneous deletion has been analyzed. Since these parameters should influence the rate and course of DNA removal, determining them in hyperthermophilic archaea represents an important step in understanding the evolutionary dynamics of their genomes.

In this study, we combined genetic selection and sequence analysis to determine the frequency and molecular nature of spontaneous deletions in the chromosome of *Sulfolobus acidocaldarius*, a crenarchaeote that populates acidic hot springs and grows optimally at about 80°C and pH 3. To our knowledge, this is the first quantitative molecular analysis of deletion formation in any archaeon or hyperthermophile. In contrast to the bacterial and eukaryotic target systems that have been similarly examined, spontaneous deletion events in the *S. acidocaldarius pyrE* gene were infrequent and not directed by repeated sequences at their ends.

**MATERIALS AND METHODS**

**Recovery of deletions.** A total of 155 independent 5-fluoroorotic acid (FOA)-resistant (Foa r) mutants of wild-type *S. acidocaldarius* (strain DG185) were screened to identify those exhibiting reversion, the pyrE coding region was amplified from 84 mutants by PCR (28). After the strains were screened to eliminate those exhibiting reversion, the pyrE coding region was amplified from 84 mutants by PCR (28). The PCR products were first screened by electrophoresis in 1.5% agarose gels; any that appeared to be smaller than normal were then digested with I 1 digestion and resolved on a 20% sequencing gel. The sequencing gel was then dried and exposed to X-ray film. The bands of interest were excised and subjected to sequencing by the dideoxy chain-termination method.

* *Corresponding author. Mailing address: Department of Biological Sciences, University of Cincinnati, 614 Rieveschl Hall, ML0006, Cincinnati, OH 45221-0006. Phone: (513) 556-9748, Fax: (513) 556-5299, E-mail: grogandw@email.uc.edu.*
RESULTS

Frequency of spontaneous deletions in *S. acidocaldarius*. Having observed pyrE deletion mutations in the course of other studies on *S. acidocaldarius* Foar mutants, we conducted a systematic search to recover additional deletions in this gene and measure their frequency. Screening additional mutants by PCR yielded two strains, MR123 and MR311, in which pyrE deletions were confirmed by DNA sequencing. When these results are combined with the results of directly sequencing a smaller set of mutants (15), deletions account for about 0.4% of spontaneous pyrE mutations. On the basis of the overall frequency of Foar mutants (Materials and Methods), pyrE deletions occur at a frequency of about $10^{-8}$ in *S. acidocaldarius* cultures. For comparison, deletions account for about 13% of all *Escherichia coli lacI* mutants, representing an overall frequency of about $2.5 \times 10^{-7}$ per viable cell (30).

To our knowledge, *S. acidocaldarius* strains MR123 and MR311, together with strains MR31, MR103, and a strain which we provisionally designate JDS22 (Materials and Methods), comprise the only collection of mutants of a hyperthermophilic archaean having spontaneous deletions. For simplicity, the alleles are here designated $\Delta$123, $\Delta$31, $\Delta$31, $\Delta$103, and $\Delta$22, respectively. Figure 1 shows the relative sizes and positions of the deletions. Only two ($\Delta$31 and $\Delta$123) preserve the translational reading frame. Although overlaps occur among four of the deletions, the endpoints do not coincide, and the endpoints do not appear to be tightly clustered (Fig. 1).

The deletion endpoints do not correspond to sites prone to other forms of mutation. The pyrE gene contains four mutational hot spots (Fig. 1), each of which accounts for 5% or more of Foar mutants sequenced (15). One of these regions

![Diagram of deletions and other sequence features of the *S. acidocaldarius* pyrE gene.](Image)

**FIG. 1.** Deletions and other sequence features of the *S. acidocaldarius* pyrE gene. Horizontal open bars on the genetic map depict coding sequences of the pyrB, pyrE, and pyrF genes of *S. acidocaldarius*; arrows show direction of transcription. Open bars above the map show the size and location of spontaneous deletions (designated in italics). All other numbers identify nucleotide positions in the pyrE coding sequence (GenBank accession number Y12822). Positions of the first and last nucleotides removed are given above the genetic map on either side of the open bars. Regions of potential stem-loop structures, numbered in order of decreasing thermodynamic stability, are indicated below the map. The first and last nucleotides of these regions (in roman numerals) are as follows: I, 200 to 217; II, 133 to 153; III, 67 to 88; IV, 528 to 541; V, 481 to 508; VI, 340 to 350. Regions which have been duplicated by various spontaneous mutations are represented by the trapezoidal symbols below the genetic map. Sites of frequent mutation are shown as partial sequences in parentheses below the map; numbers indicate the position of the first nucleotide in each region. More-detailed information on the gene sequence and mutations are provided in reference 15.

The frequency of deletion formation in the wild-type pyrE gene was calculated from the isolations that yielded strains MR123, MR311, and JDS22. The calculations had to account for the fact that about 60% of all spontaneous Foar mutants have a leaky phenotype. This class of mutants was included in a set of 79 strains that yielded JDS22 (15) but had been purged during assembly of the remaining strains (284 strains). Therefore, deletion mutants MR123, MR311, and JDS22 represent the results of screening a total of 79 + (284/0.40), or approximately 790, independent Foar mutants initially isolated.

**Sequence analysis.** Deletion endpoints were defined by aligning their pyrE sequences with the wild-type pyrE (orotate phosphoribosyltransferase) coding sequence of 594 nucleotides (nt) (GenBank accession number Y12822). This sequence, and for some analyses, a sequence of 2,251 nt encompassing pyrE and the adjacent aspartate carbamyltransferase (pyrB) and orotidin 5'-monophosphate decarboxylase (pyrF) coding regions, was analyzed using programs of the GCG package supported by the SeqWeb interface (Accelrys Inc., Madison, Wis.). Direct repeats (DRs) and inverted repeats (IRs) were found by COM-ARE, and consensus sequences were found by PILEUP, using penalties of 5 for gap creation and 1 for gap extension. Sequences surrounding the deletion end-points were randomized by SHUFFLE, and six potential stem-loop structures stable at 65°C were identified by M-FOLD.

**Probabilities.** The expected numbers of short DRs or IRs near deletion endpoints were calculated from the number of N-mer repeats that fit within a 20-nt window and the probability of matching any given N-mer at both ends of the deletion. Thus, for $N = 3$, there are 18 overlapping trimers in each 20-nt region and a 1/64 probability that a given trimer in the 5'-region will define a DR or complementary IR compared with a given trimer in the 3'-region. When the trimers at the ends are not paired in phase (i.e., when position relative to the deletion endpoint is ignored), all unordered combinations of trimers, i.e., $1/2 (18)(19) = 171$, are considered. This leads to an expected value per deletion of 171/64 for $N = 3$. Additional calculations yield values of 153/256 for $N = 4$ and 136/1024 for $N = 5$. These expected yields were multiplied by the number of deletions available (five) to give the overall expected number of observed DRs or IRs.

The probability that $n$ of 10 deletion endpoints would fall within duplication-prone regions comprising 104 bp (17.5%) of the pyrE gene is slightly greater than the Poisson term $P_n = (e^{-n}n^n)/n!$ for 10 endpoints distributed randomly among six bins. In the latter case, $n = 10/6$, and the probability that three endpoints would fall into any one bin is $P_3 = (10/6)^3/3! = 0.146$.

The probability that any given 104-bp region in the pyrE gene would contain more than three of these endpoint mutations is very small. Therefore, assuming a Poisson distribution, the expected number of 104-bp regions that would contain at least three endpoints is only about 0.1. In practice, 8 of the 104-bp regions included in the pyrE coding region contained at least three endpoint mutations, for a value of about 8/104, or 0.077. More precisely, because the Poisson distribution is not very accurate for small values of $n$, the expected number of 104-bp regions that would contain at least three endpoints is only about 0.087.
includes a short trinucleotide repeat, whereas the others are short homopolymeric runs of purines in the nontranscribed (sense) strand (Fig. 1). None of the deletion endpoints occur within these four mutation-prone regions. The 3' endpoint of Δ22 could be considered to be located as close as 2 bp from the first region, but this result is not significant at the 95% confidence level. (A total of 16 bp, or 2.7%, of the pyrE sequence is within 2 bp of one of these four mutation-prone regions, so the probability that at least 1 of 10 deletion endpoints randomly distributed within pyrE would fall within 2 bp of a hot spot is about 0.27.) Similarly, deletion endpoints were not significantly associated with sites promoting duplication. Duplications of pyrE regions that total 104 bp have been found by sequencing spontaneous mutants, but none of the duplication endpoints coincide with deletion endpoints (Fig. 1). Deletion Δ22 lies entirely within a duplication-prone region of 43 bp at the 3' end of the coding region, and one endpoint of Δ103 falls within a region near the 3' end of pyrE. However, the probability that 3 of 10 randomly distributed endpoints would fall within regions of pyrE totaling 104 bp is greater than 0.146 (see Materials and Methods), so this result is not considered significant.

Sequence analysis of deletion endpoints. (i) DRs. In other genetic systems, most deletions occur between DRs of less than 15 nt and remove one of the repeats. Mechanisms proposed to explain this pattern include misalignment of a growing 3' end between DRs, such that a stretch of template is bypassed during replication, or resection of the 5' ends of a double-strand break to generate single-stranded 3' ends which then anneal at DRs and support gap filling (3). In a previous study, Δ22 was found to have occurred between two GCT triplets in the 5' region of the pyrE gene, leaving one triplet (15). This fits the pattern seen in other systems, but the DR is shorter than most of the other examples (8, 13, 25, 30). Only one other spontaneous deletion in the set, Δ123, has endpoints defined by a repeat, and this repeat is only 1 nt (Table 1).

We assessed the significance of the Δ22 DR in two ways. The first was based on the observation that Δ22 could have arisen by removing 20 consecutive bp starting at any of four positions in the nontranscribed strand (Table 1): the first G, the first C, the first T, or the next nucleotide (C). This reflects the general property that a deletion involving two DRs N nucleotides long has N + 1 possible pairs of internucleotide breakpoints. As a result, a given deletion event will meet the criterion of involving a DR of N nucleotides if any of N + 1 pairs of windows of N nucleotides wide coincides with a DR when scanned in phase across the actual breakpoints. Thus, any given deletion has a 4/64 probability of involving DRs of 3 bp, and the probability that at least one of the five deletions would exhibit this property by chance is 20/64, or 0.31. Our second assessment of significance was based on the proposed role of DRs as sites of strand annealing. Since the stability of annealing increases dramatically with longer complementary sequences, we searched the region for longer DRs that would anneal with correspondingly higher probability. The pyrE sequence itself contains one DR of 10 nt, nine DRs of 8 nt, and 13 DRs of 7 nt, and spacings between these DRs do not differ markedly from the size distribution of the pyrE deletions recovered (Table 2). Thus, the failure to use longer DRs for deletion does not reflect their scarcity in this region or a constraint on the size of deletions recovered in the selection. These considerations argue that the coincidence of deletion endpoints with the GCT repeat is largely fortuitous.

We also evaluated experimentally the potential of DRs to promote deletion, using a genetic selection provided by tandem duplications that inactivate the pyrE gene (Fig. 1). In these mutants, precise deletion of the duplication restores gene function, making this a selectable event. Five duplications were assayed for phenotypic reversion, and five pyrE frameshift mutations were assayed for amino acid change.

### Table 1. Sequence contexts of deletion endpoints

<table>
<thead>
<tr>
<th>Deletion</th>
<th>No. of bp deleted</th>
<th>5' Region†</th>
<th>3' Region†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δ31</td>
<td>18</td>
<td>TGCAGTCGACATAGTAAAAG</td>
<td>AGATTATAATTTGATAGGA</td>
</tr>
<tr>
<td>Δ22</td>
<td>20</td>
<td>TTTGCGGAAAGCTACTTGATA</td>
<td>ATAAAAATATTTTGATAGGA</td>
</tr>
<tr>
<td>Δ123</td>
<td>27</td>
<td>GCTGGAAGTTAAGTTGAACAT</td>
<td>GTTATAGTCGATAGAAGA</td>
</tr>
<tr>
<td>Δ311</td>
<td>322</td>
<td>CTTACCTCTGATTCAAGGA</td>
<td>GTTGAACTGCTCTAGTAT</td>
</tr>
<tr>
<td>Δ103</td>
<td>329</td>
<td>GACATATTTTTCTCTGTCG</td>
<td>TGAATAAAATTGAGATAG</td>
</tr>
</tbody>
</table>

† Each region shown encompasses 10 bp on either side of the deletion endpoints. The sequences that remain after the deletion event are in bold type; the GCT triplet a region near the 3' end of the coding region, and one endpoint of Δ103 is italic and underlined.

### Table 2. DRs in the wild-type pyrE sequence

<table>
<thead>
<tr>
<th>DR</th>
<th>Position (nt) of DR</th>
<th>5' end</th>
<th>3' end</th>
<th>No. of intervening bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>7-nt DRs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>387</td>
<td>367</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>549</td>
<td>528</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>452</td>
<td>411</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>223</td>
<td>182</td>
<td></td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>254</td>
<td>185</td>
<td></td>
<td></td>
</tr>
<tr>
<td>181</td>
<td>578</td>
<td>389</td>
<td></td>
<td></td>
</tr>
<tr>
<td>267</td>
<td>584</td>
<td>309</td>
<td></td>
<td></td>
</tr>
<tr>
<td>310</td>
<td>334</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>316</td>
<td>434</td>
<td>110</td>
<td></td>
<td></td>
</tr>
<tr>
<td>319</td>
<td>490</td>
<td>163</td>
<td></td>
<td></td>
</tr>
<tr>
<td>380</td>
<td>491</td>
<td>103</td>
<td></td>
<td></td>
</tr>
<tr>
<td>447</td>
<td>543</td>
<td>88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>543</td>
<td>572</td>
<td>21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8-nt DRs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>442</td>
<td>398</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>171</td>
<td>125</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95</td>
<td>383</td>
<td>280</td>
<td></td>
<td></td>
</tr>
<tr>
<td>232</td>
<td>401</td>
<td>161</td>
<td></td>
<td></td>
</tr>
<tr>
<td>262</td>
<td>535</td>
<td>265</td>
<td></td>
<td></td>
</tr>
<tr>
<td>378</td>
<td>394</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>447</td>
<td>572</td>
<td>116</td>
<td></td>
<td></td>
</tr>
<tr>
<td>468</td>
<td>487</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>547</td>
<td>585</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10-nt DR</td>
<td>66</td>
<td>396</td>
<td>320</td>
<td></td>
</tr>
</tbody>
</table>
one end of a deleted sequence, we identified the six most stable stem-loop structures predicted in single-stranded DNA of the pyrE region (Fig. 1). These structures encompass a total of 124 nt, or 21% of the coding sequence, but no deletion endpoints fall within any of them. One deletion endpoint is adjacent to one of the stem-loop structures (the 3’ end of Δ31 and stem-loop II) (Fig. 1). We also searched for sequences in pyrE and the two adjacent genes that could anneal across both ends of any of the five deletions. The two best examples of such potentially “templating” sequences are shown in Fig. 2. Each represents an octanucleotide, the first half of which matches the 5’ flank of a deletion and the last half of which matches the 3’ flank. Annealing of these segments would have minimal base pairing for stabilization (4 bp for each segment) and would entail formation of large single-stranded regions (Fig. 2). Nevertheless, three observations make us hesitant to exclude a mechanistic role for these sequences in the formation of Δ22 and Δ103. (i) The two deletions have different sizes and positions, yet their potential templates occur at nearly the same location. (ii) The two potentially templating octanucleotides do not occur in the flanking pyrB or pyrF genes and are predicted to occur by chance once in every 20 to 50 kb of S. acidocaldarius DNA. (iii) In E. coli, one lacI deletion, S24, is known whose proposed intermediate closely resembles the structures depicted in Fig. 2 (12).

(iii) Consensus sequences. Finally, we investigated the possibility that spontaneous deletion in S. acidocaldarius occurs predominantly at a consensus sequence which may not appear as a DR or IR due to interruptions of symmetry or lack of correspondence between the two ends of the deletion. For example, Δ31 and Δ22 share an octanucleotide sequence (TT TGATA) at their 3’ ends (Table 1). In order to test the generality of shared sequences, we aligned the 5’-end regions of the deletions with each other, the 3’-end regions with each other, and the 5’-end regions with the reverse complements of the 3’-end regions. Alignments involving all five deletions revealed no base conserved at a level greater than 60% at any position fixed with respect to the endpoint. To test for a sequence whose position may vary, we aligned the 5’-end regions with each other and the 3’-end regions with each other, without constraint with respect to the position of the deletion endpoint. This yielded 100% consensus sequences TNNT (where N is any nucleotide) (5’ ends) and A (3’ ends), with the deletion endpoints falling over a 10-nt range and a 9-nt range, respectively. To confirm that these consensus sequences are not significant, we separately randomized each of the end region sequences and repeated the alignments. This yielded 100% consensus sequences of ANT and ANA for the 5’- and 3’-end regions, respectively. Thus, the consensus sequences derived from randomized end regions were of equal or higher quality as those found near the actual deletion endpoints.

**DISCUSSION**

**Comparison to other target genes.** Our review of the literature revealed few examples of chromosomal target genes in which unconstrained, spontaneous deletion has been analyzed both quantitatively and in terms of nucleotide sequence. Fortunately, however, the genetic systems meeting these criteria are phylogenetically diverse, representing animals, fungi, and
gram-positive and gram-negative bacteria. Certain features of spontaneous deletion in the *S. acidocaldarius* pyrE gene become apparent when compared to these other systems (Table 4). One is the low frequency of spontaneous deletions in *S. acidocaldarius*. When the data are normalized with respect to the size of the target gene, only *Caenorhabditis elegans* exhibits a distinctly lower frequency, whereas the bacterial systems exhibit 13- and 65-fold-higher frequencies (Table 4). Alternatively, when deletions are measured as a proportion of all spontaneous mutants, *S. acidocaldarius* falls well outside the range of the bacterial, fungal, and animal systems. Another feature of *S. acidocaldarius* is the low correlation of deletion endpoints with DRs, even those as short at 3 nt. This property is not shared by any of the other systems (Table 4).

Properties of the FOA selection. Evaluating these results must account for the fact that analysis of spontaneous deletion generally requires a target gene whose inactivation can be selected, which has the potential drawback that properties of the selection can influence the types of deletions recovered. This study used FOA plus uracil to select mutations inactivating the pyrE or pyrF gene (14); this selection is widely used in yeast genetics and has been successfully applied to other hyperthermophilic archaea (33). More importantly, its possible influences on the types of deletions recovered can be evaluated on the basis of published data for *S. acidocaldarius*.

For example, the frequency of deletions in *S. acidocaldarius* pyrE would resemble that of *E. coli* lacI if about 90% of pyrE deletions went undetected (Table 4). Potential reasons for inefficient detection of a deletion include the following: (i) a leaky phenotype, (ii) inviability, or (iii) failure to be recognized as a deletion during screening. Interference from a leaky phenotype (possibility i) is excluded by the fact that FOA selects single-nucleotide frameshifts, small in-frame deletions, and small in-frame insertions throughout pyrE (15). Inviability of a

![FIG. 2. Octanucleotides corresponding to novel joints. As a convenient way to show the relative positions of the sequences, the upper strand of the wild-type pyrE gene is drawn annealed to the octanucleotide in the lower strand that represents the novel joint of the corresponding deletion. This method of depiction is that used in other deletion studies (11, 12, 16, 30) and is not a proposal of a specific mechanism. Sequences remaining after deletion are shown in bold type, as for Table 1. Numbers with vertical lines indicate the nucleotide position in each strand with respect to the pyrE coding sequence; the remaining numbers indicate the sizes of the loops shown.](http://jb.asm.org/)

<table>
<thead>
<tr>
<th>Organism</th>
<th>Target gene</th>
<th>Size (bp)</th>
<th>Deletion frequency</th>
<th>Proportion (%) of mutants</th>
<th>No. analyzed</th>
<th>% at DRs of &gt; 2 nt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Entire gene</td>
<td>Per kbp</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>S. acidocaldarius</em></td>
<td>pyrE</td>
<td>594</td>
<td>$1 \times 10^{-8}$</td>
<td>$1.7 \times 10^{-8}$</td>
<td>0.4</td>
<td>5</td>
</tr>
<tr>
<td><em>L. bulgaricus</em></td>
<td>lac</td>
<td>~3,100</td>
<td>$3.3 \times 10^{-6}$</td>
<td>$1.1 \times 10^{-6}$</td>
<td>9.3</td>
<td>8</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>lac&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1,080</td>
<td>$2.5 \times 10^{-7}$</td>
<td>$2.3 \times 10^{-7}$</td>
<td>12.6</td>
<td>22</td>
</tr>
<tr>
<td><em>N. crassa</em></td>
<td>mtr</td>
<td>1,472</td>
<td>$1.0 \times 10^{-7}$</td>
<td>$6.8 \times 10^{-8}$</td>
<td>35</td>
<td>11</td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td>UNC-54</td>
<td>~6,000</td>
<td>$4.5 \times 10^{-8}$</td>
<td>$7.5 \times 10^{-9}$</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>Hamster</td>
<td>APRT&lt;sup&gt;d&lt;/sup&gt;</td>
<td>670&lt;sup&gt;a&lt;/sup&gt;</td>
<td>$1 \times 10^{-8}$</td>
<td>$1.5 \times 10^{-8}$</td>
<td>10</td>
<td>6</td>
</tr>
</tbody>
</table>

<sup>a</sup> Data are for *S. acidocaldarius* (this work), *Lactobacillus bulgaricus* (25), *E. coli* (30), *Neurospora crassa* (8), *C. elegans* (27), and Chinese hamster ovary cells (26).

<sup>b</sup> Similar results have been reported in other studies on this gene (11, 16).

<sup>c</sup> Includes introns.

<sup>d</sup> Estimated from data of reference 10.
large proportion of pyrE deletions (possibility ii) is inconsistent
with recovery of deletions removing most of the gene and
deletions that sharply decrease expression of the cotranscribed
pyrF gene (15, 28). With regard to the failure to be recognized
as a deletion (possibility iii), our screening of PCR products on
agarose gels successfully identified an 18-bp deletion, and our
screening probably has a detection threshold of 12 to 15 bp.
Thus, a significant underestimation of the deletion frequency
under our conditions would require the majority of pyrE dele-
tions to be very short (i.e., less than about 15 bp) or to occur in
pyrF. Neither of these alternatives agrees with the spectrum of
spontaneous mutation in the pyrE and pyrF genes, in which
only one deletion was observed among 108 independent, ran-
domly chosen Foa' mutants. This deletion (Δ22) removes 20 bp
and lies in pyrE, which was the location of about 95% of all
spontaneous mutations conferring the Foa' phenotype (15).

Another property, the relatively low proportion of deletions
for the S. acidocaldarius pyrE gene compared to the other
genes in Table 4 could, in principle, result from the following:
(i) production of other classes of mutation at a much higher
rate than those in the other genetic systems or (ii) selection of
leaky mutants at a much higher rate than those in the other
systems. Possibility i is excluded by two sets of independent
mutation rate measurements, which show that pyrE has a for-
ward mutation rate very close to those of E. coli lact and other
bacterial genes (15, 18). Situation ii is more difficult to exclude,
as about 60% of mutants selected by 50 μg of FOA per ml exhib-
it a leaky phenotype, and it is not known whether the other
systems of Table 4 yield a significant proportion of leaky
mutants. Therefore, the frequency of deletions among non-
leaky S. acidocaldarius pyrE mutants (about 1%) may provide
a more conservative value for comparison to the other systems,
but this would still be only one-ninth of the closest value in
Table 4.

Our failure to recover a deletion extending beyond the
boundaries of the pyrE gene also raises the question of whether
the FOA selection can recover such deletions in S. acidocal-
darius. These would be precluded if DNA sequences flank-
ing pyrE were essential, for example. The transcript beginning
with pyrE includes pyrF and at least one other ORF, designated orf8
(32). However, two observations argue against essential rules
for either pyrF or orf8: (i) frameshift mutations in pyrE which
exert strong polar effects on pyrF expression are frequently
isolated (28), and (ii) several pyrF frameshifts and one non-
sense mutation, which inactivate pyrF and should exert similar
polarity on orf8 expression, have also been isolated (15). An
alternative explanation for the observed placement of dele-
tions is that some position-dependent property of pyrE makes
it more susceptible than pyrF to various forms of mutation,
including deletion. This is consistent with the following facts:
duplication mutations are relatively common in pyrE but have
not been found in pyrF (15); point mutations are 20-fold more
frequent in pyrE than pyrF (15); in S. solfataricus, IS elements
insert threefold more frequently into pyrE than into pyrF (23);
and diverse forms of DNA damage all induce mutation in S.
acidocaldarius (29, 34).

Potential deletion mechanisms. The low frequency of dele-
tions in S. acidocaldarius impeded our efforts to collect many
examples for analysis, yet sequences of the available alleles
distinguish the dominant mode (or modes) of their formation
from those commonly proposed for other organisms. No asso-
ciation of deletion endpoints with DRs, IRs, or stem-loop
structures was supported at the 95% confidence level, despite
the fact that these structures are about as abundant in the S.
acidocaldarius pyrE gene as in the various target genes where
they have been found to promote spontaneous deletions (3, 8,
12, 25, 26). However, genetic assays in S. acidocaldarius do
provide evidence of strand slippage mechanisms of deletion
under certain conditions. Tandem duplications reverted at
high frequencies, which increased with increasing length of the
repeated sequence (Table 3). Furthermore, homopolymeric
runs and short repeats have been shown to promote frameshift
mutations and triplet expansions, respectively, in the wild-type
pyrE gene (15). Our observations that tandem duplications
were deletion-prone, whereas natural DRs were not may re-
fect the fact that the natural DRs are shorter (on average) and
interrupted by nonrepeated sequences.

In some systems, consensus sequences occur at the end-
points of spontaneous deletions and have been attributed to
normal or aberrant processing by DNA breaking or joining
enzymes such as topoisomerases (3, 5). The available set of S.
acidocaldarius deletions included only one example of a com-
mon sequence found at the 3' ends of two deletions (Δ22 and
Δ31), and this remains difficult to interpret mechanistically.
The lack of AA or TT dinucleotides at deletion endpoints
argues that an activity resembling topoisomerase II of Sulfo-
lobus shibatae, which cleaves DNA predominantly at AA or TT,
yielding 2-nr, 5' extensions (4), is not a significant source of
small- and medium-sized deletions in S. acidocaldarius. How-
ever, it should be emphasized that DNA sequence specificity
has not been established for other potentially relevant enzymes
of S. acidocaldarius. For example, the reverse gyrase of S.
acidocaldarius (a topoisomerase I) exhibits the same minimal
sequence specificity of its counterpart in S. shibatae (19), so it
is difficult to rule it out as a potential source of deletions.
The possible role of nearby sequences in stabilizing nascent dele-
tions ("templating" [12]) could be inferred in two cases and
will require isolation of additional deletions to evaluate. Our
observations also seem consistent with repair of double-strand
breaks by some form of nonhomologous end joining (3), which,
in turn, is consistent with the associations of MRE11 and
RAD50 homologues in the genome of S. acidocaldarius and
other hyperthermophilic archaea (7).

Implications for genome evolution. According to mutational
analysis of the pyrE and pyrF genes, S. acidocaldarius has one
of the lowest genomic error rates yet measured, despite its
extremely harsh growth conditions (15). The present study
further shows that, within the low mutant frequency, deletions
make up an unusually low proportion of spontaneous muta-
tions. The relative frequencies of different classes of mutations
now documented in S. acidocaldarius predict a pattern for the
elimination of a gene following release of selective pressure.
Initially, frameshift mutations should accumulate in homopoly-
meric runs, then elsewhere in the sequence, followed by slower
accumulation of base pair substitutions and duplications, fol-
lowed in turn by yet slower accumulation of deletions. It also
seems significant that in S. acidocaldarius the more-frequent
classes of mutations are reversible. This would serve to provide
a window of time during which function of a mutated gene can
be restored should moderate selection be reapplied. This idea
finds experimental support in the observation that eight lineages of *S. acidocaldarius* subjected to multiple cycles of forward and reverse mutation failed to generate a functional *pyrE* gene with an altered sequence (2). Also, analysis of other genomes, such as that of *Saccharomyces cerevisiae*, has identified defective but recoverable genes which represent a reserve of potential functionality (17).

The slow, incremental nature of gene elimination predicted by the mutational processes in *S. acidocaldarius* seems to favor genome streamlining while minimizing concomitant inactivation of beneficial genes. It should be noted, however, that this situation may not typify all hyperthermophilic archaea. *S. acidocaldarius* has few, if any, active IS elements on the basis of genetic assays (15) and preliminary sequence analysis of the genome (R. Garrett, personal communication), but other *Sulfolobus* species have a number of them (20, 23, 31). In particular, transposition of several distinct families of IS elements causes a high frequency of spontaneous mutation at the *pyrE* and *pyrF* genes of *S. solfataricus* and may also generate frequent chromosomal rearrangements (23, 31). In such lineages, IS elements may assume the dominant role in inactivating nonessential genes (through insertion) and in removing them (through transposase-catalyzed imprecise excision).

ACKNOWLEDGMENTS

We thank M. Reilly for providing strains; New England Biolabs, Inc., for providing D.W.G. with computing facilities; and J. Drake for communicating unpublished results and for comments on the manuscript.

This work was supported in part by grant MCB9733303 from the National Science Foundation.

REFERENCES


