10 Blanchard, J.L. and Schmidt, G.W. (1995) Pervasive migration of organellar DNA to the nucleus in plants. *J. Mol. Evol.* 41, 397–406
11 Richly, E. and Leister, D. (2004) NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs. *Mol. Biol. Evol.* 21, 1972–1980
12 Ricchetti, M. *et al.* (2004) Continued colonization of the human genome by mitochondrial DNA. *PLoS Biol.* 2, E273
13 Borensztajn, K. *et al.* (2002) Characterization of two novel splice site mutations in human factor VII gene causing severe plasma factor VII deficiency and bleeding diathesis. *Br. J. Haematol.* 117, 168–171
14 Turner, C. *et al.* (2003) Human genetic disease caused by *de novo* mitochondrial-nuclear DNA transfer. *Hum. Genet.* 112, 303–309
15 Long, M. *et al.* (2003) The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* 4, 865–875
16 Brosius, J. and Gould, S.J. (1992) On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". *Proc. Natl. Acad. Sci. U. S. A.* 89, 10706–10710
17 Krull, M. *et al.* (2005) Alu-SINE exonization: en route to protein-coding function. *Mol. Biol. Evol.* 22, 1702–1711
18 Decker-Walters, D.S. *et al.* (2004) Plastid sequence evolution: a new pattern of nucleotide substitutions in the Cucurbitaceae. *J. Mol. Evol.* 58, 606–614
19 Petrov, D.A. and Hartl, D.L. (1999) Patterns of nucleotide substitution in *Drosophila* and mammalian. genomes. *Proc. Natl. Acad. Sci. U. S. A.* 96, 1475–1479
20 Huang, C.Y. *et al.* (2005) Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. *Plant Physiol.* 138, 1723–1733
21 Noutsos, C. *et al.* (2005) Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants. *Genome Res.* 15, 616–628
22 Saxonov, S. *et al.* (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. U. S. A.* 103, 1412–1417
23 Bender, J. (2004) Chromatin-based silencing mechanisms. *Curr. Opin. Plant Biol.* 7, 521–526
24 Richly, E. and Leister, D. (2004) NUMTs in sequenced eukaryotic genomes. *Mol. Biol. Evol.* 21, 1081–1084
25 Sorek, R. *et al.* (2002) Alu-containing exons are alternatively spliced. *Genome Res.* 12, 1060–1067

0168-9525/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2007.08.016

Selection on the genic location of disruptive elements

M.W.J. van Passel and H. Ochman

Department of Biochemistry and Molecular Biophysics, University of Arizona, 1007 East Lowell Street, Tucson, AZ 85721, USA

Analyses of nucleotide patterns in coding regions of prokaryotes have revealed that selection acts on DNA and RNA stability and on translational accuracy. Here we examine the positions of mononucleotide repeats within microbial genes and detect a pervasive bias in the locations of these disruptive elements that becomes more pronounced with increases in repeat length. We argue that, because these repeats are mutagenic, this pattern arose to minimize the costs associated with transcribing and translating nonfunctional genes, supporting a view that pseudogenes need not be evolving in a strictly neutral manner.

Bacterial pseudogenes are ephemeral

Pseudogenes are often cited as the paradigm of neutral evolution, representing regions of the genome that are not under any selective constraints [1]. Pseudogenes have long been known to be present in eukaryotes [2], but until recently, these inactivated genes were considered to be rare in prokaryotes [3]. The elucidation of complete genome sequences revealed that pathogenic bacteria can contain hundreds of disrupted and eroded genes [4,5], and subsequent analyses divulged that pseudogenes are a relatively common feature of organisms in all domains of life [6–9]. In contrast to eukaryotes, the pseudogene repertoires of bacteria and archaea are largely genome-specific, suggesting that nonfunctional sequences are removed rapidly from the genome [6].

There are two potential reasons why pseudogenes do not persist in prokaryotic genomes. The first is that pseudogenes are eliminated by the pervasive mutational bias toward deletions in microbial genomes [10,11] — a mutation-based or 'passive removal' hypothesis. Alternatively, pseudogenes might be removed because of the costs associated with maintaining these disrupted genes — a selection-based or 'active removal' hypothesis. Although pseudogenes are usually considered to evolve neutrally (implying passive removal), they might actually be detrimental to the organism because of the energetic costs of their transcription or translation or through deleterious effects caused by nonfunctional proteins.

A cost to maintaining pseudogenes?

In prokaryotes, distinguishing between the active and passive removal processes cannot be achieved by simply comparing the rate at which pseudogenes are removed relative to noncoding sequences, because truly functionless DNA is limited in scale, and in most cases, derived from previously coding DNA. However, there is some evidence that transcription and translation might impose energetic costs on which selection might act. For example, Akashi and Gojobori [12] showed that highly expressed genes in *Escherichia coli* and *Bacillus subtilis* have a lower than average proportion of amino acids that are expensive to synthesize. Also, codon bias has been found to increase progressively along the length of genes, suggesting selection against nonsense mutations [13]. Finally, the genetic code seems to have been optimized such that stop codons follow frameshift mutations more quickly than would be the case in alternative genetic codes, which can serve to

Corresponding author: van Passel, M.W.J. (mvpassel@email.arizona.edu).
Available online 8 November 2007.

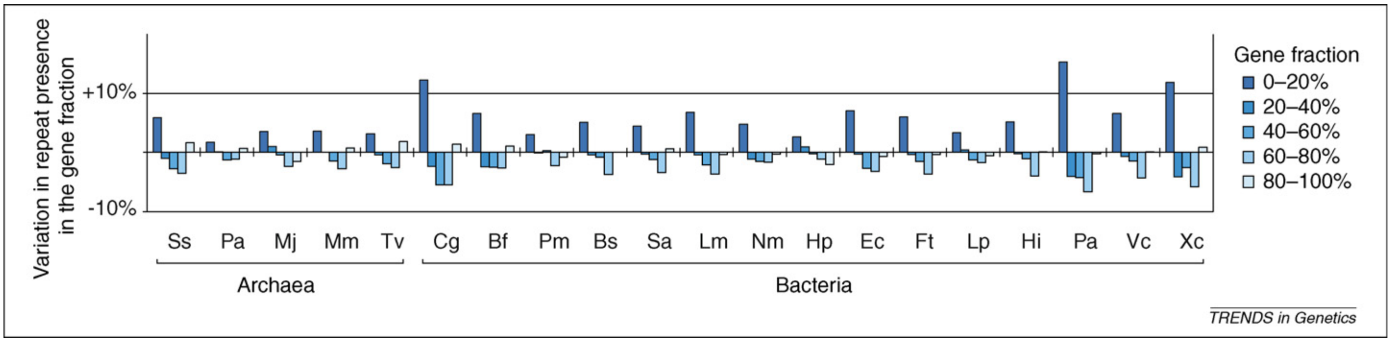


Figure 1. Positional bias of homopolymeric repeats (more than five nucleotides in length) within genes from 20 microbial genomes. In each genome (5 Archaea, 15 Bacteria), all genes were divided proportionally into five segments (with 0–20% denoting the initial one-fifth of the gene beginning at its 5′ end, and 80–100% denoting the 3′ segment), and numbers of simple sequence repeats (SSRs) were computed and summed for each segment. The two-letter strain designations are as follows: Ss, *Sulfolobus solfataricus* P2 (NC_002754); Pa, *Pyrobaculum aerophilum* str. IM2 (NC_003364); Mj, *Methanocaldococcus jannaschii* DSM 2661 (NC_000909); Mm, *Methanosarcina mazei* Go1 (NC_003901); Tv, *Thermoplasma volcanium* GSS1 (NC_002689); Cg, *Corynebacterium glutamicum* ATCC 13032 (NC_003450); Bf, *Bacteroides fragilis* NCTC 9343 (NC_003228); Pm, *Prochlorococcus marinus* str. MIT 9303 (NC_008820); Bs, *Bacillus subtilis* str. 168 (NC_000964); Sa, *Staphylococcus aureus* RF122 (NC_007622); Lm, *Leuconostoc mesenteroides* ATCC 8293 (NC_008531); Nm, *Neisseria meningitidis* MC58 (NC_003112); Hp, *Helicobacter pylori* 26695 (NC_0000915); Ec, *Escherichia coli* CFT073 (NC_004431); Ft, *Francisella tularensis* (NC_007880); Lp, *Legionella pneumophila* str. Lens (NC_006369); Hi, *Haemophilus influenzae* Rd KW20 (NC_000907); Pa, *Pseudomonas aeruginosa* PAO1 (NC_002516); Vc, *Vibrio cholerae* O1 str. N16961 (both chromosomes, NC_0002505 and NC_002506); Xc, *Xanthomonas campestris* str. 8004 (NC_007086). Gene lists of the selected genomes were downloaded as .ffn files from NCBI (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) and from TIGR (http://www.tigr.org/).

minimize the resources used to produce functionless transcripts and proteins [14].

If there are costs associated with synthesizing and translating useless sequences, one might expect that selection would favor mutations that diminish these costs by producing maximally truncated proteins. Therefore, we tested whether sites prone to disruptive mutations are biased toward certain locations within a gene. We focus on simple sequence repeats (SSRs), specifically mononucleotide runs, which are present in all genomes, although typically underrepresented in coding regions because of their highly mutagenic nature [15–18]. SSRs are known hotspots for small indels, the most common type of mutations that generate pseudogenes in bacteria and archaea [6,7,9]. Such mutations can, however, also be advantageous, as is the case in contingency loci in pathogenic bacteria, in which elevated mutation rates at SSRs are responsible for reversible gene inactivations, termed phase variation [19,20].

Intragenic arrangement of simple sequence repeats

We first examined the distribution of these mutation-prone repeats within the phase-variable contingency loci of *Neisseria meningitidis* MC58. Saunders *et al.* [19] identified 31 phase-variable genes whose coding sequences contained mononucleotide repeats greater than 5 bp in length. In 14 of 31 loci, the disruptive repeats occur in the first 20% of the gene, and there is a clear shift in SSR location toward the 5′-end of the gene with increased mononucleotide repeat length (data not shown).

Given the pattern observed in the neisserial contingency loci, we tracked the positions of mononucleotide SSRs in the annotated protein coding genes of 20 prokaryotic genomes. Both the 5′ bias in the intragenic location of SSRs (more than 5 bp; Figure 1) and the increase in bias with SSR length (Figure 2) are observed across genomes, indicating a general trend toward SSR avoidance in certain portions of the gene. Similarly, the 5′ location bias of mononucleotide repeats, as well as the 5′ shift with

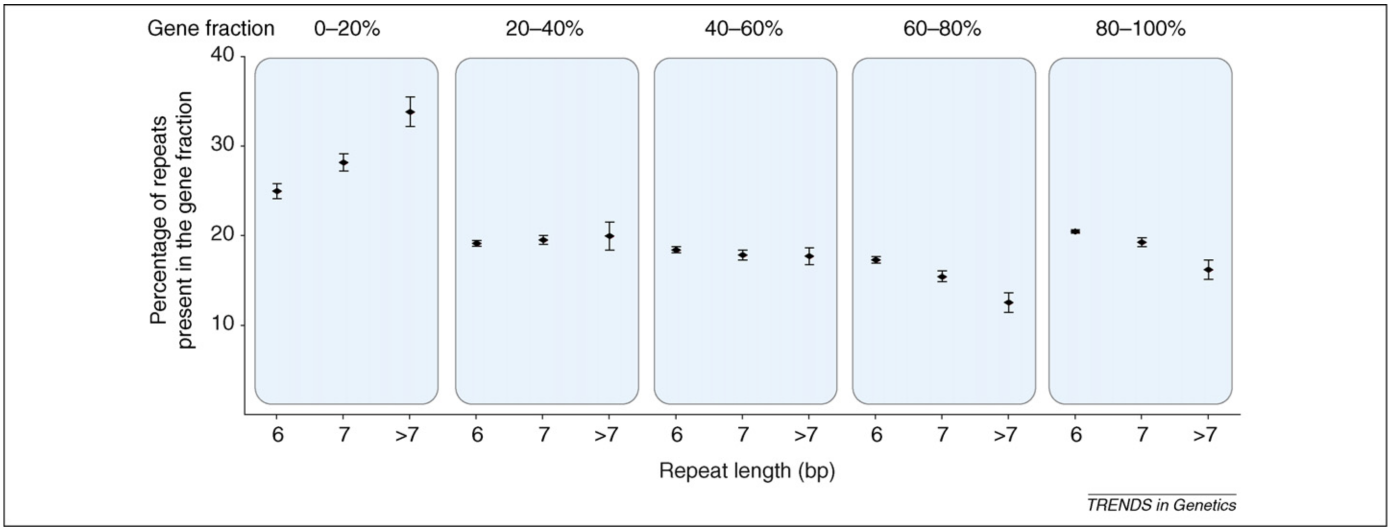
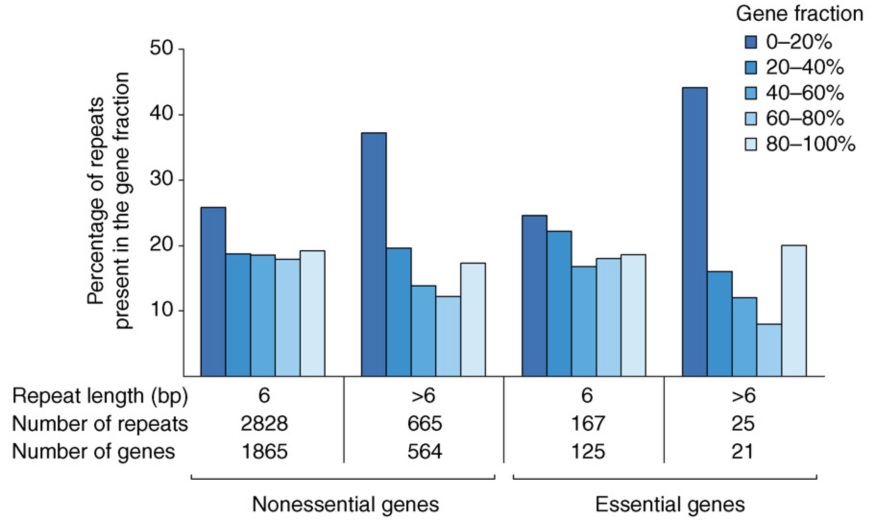


Figure 2. Intragenic shift in simple sequence repeat (SSR) location toward the 5′ end with increasing repeat length. Means and standard errors are based on data from the 20 microbial genomes presented in Figure 1. Repeat numbers are included only in cases where there were >50 SSRs in the coding portion of the genome.



TRENDS in Genetics

Figure 3. Distribution of homopolymeric repeats in essential and nonessential genes of *E. coli* K-12. Depicted are the repeat length (in bp), the number of repeats and the number of genes in which repeats of that repeat length are found.

increased repeat length, was also observed in *Saccharomyces cerevisiae* genes, albeit with longer repeat lengths (>7 bp). By contrast, the severely reduced genome of *Carsonella ruddii* [21] is extremely A + T-rich and shows neither the 5′ bias of SSRs nor the 5′ shift with increased repeat length (data not shown), even though a large number of SSRs are present in its coding sequences.

The N terminus of proteins frequently contain conserved peptide signal sequences [22], and these signal peptides can consist of amino acid motifs encoded by mononucleotide repeats. For example, the adjacent lysine residues in the MKK tripeptide at the N terminus of several signal sequences could generate homopolymeric adenine tracts. In a reanalysis of the *E. coli* K-12 and *N. meningitidis* MC58 genomes, we excluded all SSRs starting before the ninth nucleotide of each gene, and the 5′ location bias of SSRs, as well as the 5′ shift with increasing repeat size, is still observed. Furthermore, to test if variation in amino acid composition might be a source of mononucleotide repeat distribution, we compared the occurrence of amino acid residues across proteins in a genome. For the four amino acids that can be encoded by homogeneous codons [phenylalanine (TTY), glycine (CCN), lysine (AAR) and proline (GGN)], there is no genome-wide overrepresentation in the 5′ portion of genes, although lysine is somewhat more prevalent in the 3′ end (data not shown). Therefore, the tendency of SSRs to occur near the 5′ end of genes does not seem to be driven by amino acid use. In addition, Ackermann and Chao [15] have shown that selection acts to eliminate mononucleotide repeats in consecutive homogeneous codons with identical nucleotides.

The bias in the location of SSRs suggests that indels at different sites along a gene have differential effects on cell fitness as a result of the energetic and resource savings associated with investing less in functionless gene products. An alternative hypothesis for the 5′ bias in SSR location is that this pattern reflects selection on transcript properties, possibly affecting mRNA stability. On average, mononucleotide repeats are shorter [15] and less common in essential genes than in nonessential genes, suggesting that selection has acted to reduce the mutability of more highly required loci. However, some essential genes contain potentially disruptive repeats. Under the hypothesis that 5′ location bias of SSRs reflects selection to minimize investment in functionless gene products, we predict that essential genes would show little or no bias, because their products are essential to cell function. However, if SSR distributions are similar in essential and nonessential genes, it would suggest an additional role for these repeats.

Competing forces acting on SSR location

As in the study by Jordan *et al.* [23], we analyzed the set of essential protein-coding genes listed in the Profiling of the *E. coli* Genome database (http://www.shigen.nig.ac.jp/ecoli/pec/; 283 genes). We found that SSRs more than six nucleotides in length were underrepresented in the essential gene set (χ^2 test between the fraction of genes with repeats in the essential and nonessential gene sets, $P < 0.05$), supporting earlier results reporting shorter mononucleotide runs in essential genes [15]. Although there are significantly fewer and shorter disruptive repeats in essential genes, SSRs, when present, predominate at the 5′ end of the genes, and there is a 5′ shift with longer repeat lengths (Figure 3). This could be taken to mean that some subset of these genes might not be strictly essential or that additional factors are acting on the SSR distribution within genes.

Positional bias in homopolymers (particularly polyA tracts) could also result from selection against the formation of 5′ hairpin structures that might excessively stabilize transcripts and slow translation. We tested this possibility by examining the frequencies and locations of homopolymeric repeats in relation to transcription, using the codon adaptation index (CAI) as a proxy for expression levels (using CodonW, http://codonw.sourceforge.net). Although essential, highly expressed genes show a 5′ bias in the occurrence of homopolymeric repeats (suggesting that SSRs might serve to reduce hairpins), we also found that nonessential, low CAI genes display a 5′ bias, which becomes more pronounced as repeat lengths increase. Taken together, these results indicate that neither hairpin avoidance nor SSR mutability is the sole factor determining the intragenic location bias of homopolymeric repeats.

The placement of mononucleotide repeats toward the 5′ ends of genes supports the active removal hypothesis, whereby shorter polypeptides are preferred. Shorter pseudogenes might also be favored because their products are less disruptive to interacting proteins and to cellular processes in general. That selection has biased the intragenic location of disruptive elements is also supported by their distributions in other portions of the gene. As displayed in all three figures, there is less avoidance of homopolymeric tracts in the 3′ part of the gene (i.e. 80–100%). Hence, we think that avoidance of the SSR is lower in this portion of the gene than in the 40–80% portion of the gene. Also, mutagenic repeats at the end of the gene can contribute to fusions with downstream genes, as is the case in recoding events in mobile genetic elements [24].

Concluding remarks

Genome-wide analyses of sequence motifs have uncovered selective forces acting on DNA and RNA stability and on the translational accuracy of microbial genes [13,15,17]. Based on the sequence features of functional protein-coding genes, we were able to infer that pseudogenes are not evolving in a strictly neutral manner. If pseudogene evolution is affected by energetic costs associated with their transcription and translation, and their potential to disrupt cellular processes, we might also expect that pseudogenes will be transcriptionally inactivated or removed entirely from the genome at enhanced rates to ameliorate these costs.

Acknowledgements

The authors thank Becky Nankivell for assistance with the figures and Nancy Moran for comments on an earlier version of this manuscript. This work was supported by research grant GM56120 from the National Institutes of Health to H.O.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2007.08.017.

References

1 Li, W.H. *et al.* (1981) Pseudogenes as a paradigm of neutral evolution. *Nature* 292, 237–239

2 Jacq, C. *et al.* (1977) A pseudogene structure in 5S DNA of *Xenopus laevis*. *Cell* 12, 109–120

3 Lawrence, J.G. *et al.* (2001) Where are the pseudogenes in bacterial genomes? *Trends Microbiol.* 9, 535–540

4 Andersson, S.G. *et al.* (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* 396, 133–140

5 Cole, S.T. *et al.* (2001) Massive gene decay in the leprosy bacillus. *Nature* 409, 1007–1011

6 Lerat, E. and Ochman, H. (2004) Psi-Phi: exploring the outer limits of bacterial pseudogenes. *Genome Res.* 14, 2273–2278

7 Lerat, E. and Ochman, H. (2005) Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Res.* 33, 3125–3132

8 Liu, Y. *et al.* (2004) Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. *Genome Biol.* 5, R64

9 van Passel, M.W. *et al.* (2007) Gene decay in archaea. *Archaea* 2, 137–143

10 Mira, A. *et al.* (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet.* 17, 589–596

11 Nilsson, A.I. *et al.* (2005) Bacterial genome size reduction by experimental evolution. *Proc. Natl. Acad. Sci. U. S. A.* 102, 12112–12116

12 Akashi, H. and Gojobori, T. (2002) Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc. Natl. Acad. Sci. U. S. A.* 99, 3695–3700

13 Stoletzki, N. and Eyre-Walker, A. (2007) Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Mol. Biol. Evol.* 24, 374–381

14 Itzkovitz, S. and Alon, U. (2007) The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Res.* 17, 405–412

15 Ackermann, M. and Chao, L. (2006) DNA sequences shaped by selection for stability. *PLoS Genet.* 2, e22

16 Coenye, T. and Vandamme, P. (2005) Characterization of mononucleotide repeats in sequenced prokaryotic genomes. *DNA Res.* 12, 221–233

17 Mrazek, J. *et al.* (2007) Simple sequence repeats in prokaryotic genomes. *Proc. Natl. Acad. Sci. U. S. A.* 104, 8472–8477

18 van Belkum, A. *et al.* (1998) Short-sequence DNA repeats in prokaryotic genomes. *Microbiol. Mol. Biol. Rev.* 62, 275–293

19 Saunders, N.J. *et al.* (2000) Repeat-associated phase variable genes in the complete genome sequence of *Neisseria meningitidis* strain MC58. *Mol. Microbiol.* 37, 207–215

20 van der Woude, M.W. and Baumler, A.J. (2004) Phase and antigenic variation in bacteria. *Clin. Microbiol. Rev.* 17, 581–611

21 Nakabachi, A. *et al.* (2006) The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science* 314, 267

22 Sjostrom, M. *et al.* (1987) Signal peptide amino acid sequences in *Escherichia coli* contain information related to final protein localization. A multivariate data analysis. *EMBO J.* 6, 823–831

23 Jordan, I.K. *et al.* (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* 12, 962–968

24 Baranov, P.V. *et al.* (2006) Recoding in bacteriophages and bacterial IS elements. *Trends Genet.* 22, 174–181

0168-9525/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2007.08.017