Published online June 2, 2005

Nucleic Acids Research, 2005, Vol. 33, No. 10 ***3125–3132***
doi:10.1093/nar/gki631

Recognizing the pseudogenes in bacterial genomes

Emmanuelle Lerat and Howard Ochman[1,*]

Department of Ecology and Evolutionary Biology and [1]Department of Biochemistry and Molecular Biophysics, University of Arizona, Tucson, AZ 87521, USA

Received December 25, 2004; Revised March 1, 2005; Accepted May 15, 2005

ABSTRACT

Pseudogenes are now known to be a regular feature of bacterial genomes and are found in particularly high numbers within the genomes of recently emerged bacterial pathogens. As most pseudogenes are recognized by sequence alignments, we use newly available genomic sequences to identify the pseudogenes in 11 genomes from 4 bacterial genera, each of which contains at least 1 human pathogen. The numbers of pseudogenes range from 27 in *Staphylococcus aureus* MW2 to 337 in *Yersinia pestis* CO92 (e.g. 1–8% of the annotated genes in the genome). Most pseudogenes are formed by small frameshifting indels, but because stop codons are A + T-rich, the two low-G + C Gram-positive taxa (Streptococcus and Staphylococcus) have relatively high fractions of pseudogenes generated by nonsense mutations when compared with more G + C-rich genomes. Over half of the pseudogenes are produced from genes whose original functions were annotated as 'hypothetical' or 'unknown'; however, several broadly distributed genes involved in nucleotide processing, repair or replication have become pseudogenes in one of the sequenced *Vibrio vulnificus* genomes. Although many of our comparisons involved closely related strains with broadly overlapping gene inventories, each genome contains a largely unique set of pseudogenes, suggesting that pseudogenes are formed and eliminated relatively rapidly from most bacterial genomes.

INTRODUCTION

Although bacterial genomes are compact and contain relatively little non-coding DNA, pseudogenes are a regular feature and present even in the smallest of bacterial genomes (1,2). Because it is not feasible to directly assess the functional status of each annotated coding region within a genome, virtually all pseudogenes are recognized by a comparative approach, i.e. by aligning homologs and searching for truncated, or otherwise disrupted, CDSs. Such sequence comparisons become increasingly effective for identifying pseudogenes with the appearance of each additional genome sequence and are most valuable for groups of closely related genomes whose gene contents largely overlap and for which the extent of sequence divergence allows the recognition of individual point mutations.

Because the initial annotation guides much of the subsequent experimental and computational work on a genome, ranging from inclusion of particular sequences on arrays (3) to large-scale analyses of protein interactions (4), it would be most valuable if the functional status of predicted open reading frames (ORFs) (and the presence of pseudogenes) were known from the outset. Although bacterial genomes are gene-rich, and bacterial genes supply very strong signals on which to base gene recognition algorithms, pseudogenes cannot be discovered readily by *ab initio* methods of gene prediction. To this end, we developed a system to detect pseudogenes by comparing the genome sequences of closely related strains and species (5); and when applied to the four sequenced members of the *Escherichia coli* clade, we identified hundreds of new pseudogenes, even in those strains originally considered to have none. In this study, we found that each of the strains within this clade contained a largely unique set of pseudogenes, raising questions about the timescale by which pseudogenes originate and are removed, and about the particular types of genes that are inactivated in each bacterial genome.

The comparative approach for identifying pseudogenes has been applied with varying success as pairs of closely related bacterial genomes became available. Early on, it was noticed that bacterial pathogens, especially when compared with their free-living relatives, harbored large numbers of pseudogenes, presumably owing to the inactivation and degradation of genes that were no longer needed in the host environment (6–9). Unfortunately, the manner in which sequences are classified as pseudogenes is not consistent among studies, with some groups annotating any shortened ORF as a pseudogene (10) and others regarding considerably shortened homologs as still specifying some function (11)—making it problematic to directly compare the pseudogene contents, or the numbers and

*To whom correspondence should be addressed. Tel: +1 520 626 8355; Fax: +1 520 621 3709; Email: hochman@email.arizona.edu

types of mutational events that produce pseudogenes, across genomes.

To establish the criteria and thresholds by which sequences can be assigned as pseudogenes, we evaluated the complete gene repertoires from several densely sampled groups of sequenced genomes, each containing at least one human pathogen. These bacterial groups form clades of different phylogenetic depths, ranging from as little as 1% difference for the conspecific strains of *Staphylococcus aureus*, *Streptococcus pyogenes* and *Yersinia pestis* to ~10% difference for the *Vibrio* congeners, which allowed us to monitor the effects of strain divergence on pseudogene recognition. Because many of these genomes were sequenced and annotated well before there were suitable strains for comparative analysis, we provide the first glimpse into their pseudogene populations and the manner in which pseudogenes are formed in and eliminated from bacterial genomes.

MATERIALS AND METHODS

Genomes surveyed

To obtain the inventory of pseudogenes within the genomes of bacterial pathogens representing multiple taxonomic groups, we compared the full genome sequences of (i) three strains of *S.aureus* {Mu50 GenBank accession no. NC_002758 (12), N315 [BA000018 (12)] and MW2 [BA000033 (13)]}, (ii) four strains of *S.pyogenes* {SF370 serotype M1 [AE004092 (14)], MGAS315 serotype M3 [AE014074 (15)], MGAS8232 serotype M18 [AE009949 (16)] and SSI-1 serotype M3 [BA000034 (17)]}, (iii) three strains of *Y.pestis* {CO92 [AL5908842 (8)], KIM [AE009952 (9)] and 91001 [AE017042 (18)]} and *Yersinia pseudotuberculosis* IP32953 [BX936398 (10)] and (iv) both chromosomes of four *Vibrio* {*Vibrio cholerae* [AE003852 and AE003853 (19)], *Vibrio vulnificus* strains CMCP6 [AE016795 and AE016796 (20)] and YJ016 [BA000037 and BA000038 (21)] and *Vibrio parahaemolyticus* [BA000031 and BA000032 (22)]}.

Identifying pseudogenes

Pseudogenes within each genome were identified by comparing its genome contents with that of each of the other sequenced strains within the particular taxonomic group. For each genome, we first retrieved the set of the annotated proteins from GenBank, which were used to query the nucleotide sequence of the other genomes within the particular group using TBLASTN (23). For example, the proteins of *Y.pseudotuberculosis* were used to query the genome sequences of *Y.pestis* CO92, *Y.pestis* KIM and *Y.pestis* 91001. We then applied the program Ψ–Φ (5) on the BLAST outputs to recover candidate pseudogenes in each genome. This program allows the specification of any BLAST score and % identity cut-offs, and for our comparisons, proteins from two genomes were considered to be homologous if their BLAST score reached an *E*-value $<10^{-15}$ and their level of protein identity was >79%. In the case of *Vibrio*, the strains/species examined were not as closely related, so we applied different thresholds (*E*-values $<10^{-10}$ and a minimal percentage of protein identity of 49%) in order to identify homologous sequences. This program retrieves pseudogenes that result from nonsense mutations, frameshifts generated by small insertions or deletions, large insertions (such as those resulting from transposable elements) and truncations of any specified length as well as any incorrectly annotated spacers that resulted from degradation of a gene. Lists of candidate pseudogenes were curated manually, and the disrupting mutations were determined by aligning the nucleotide sequences of putative pseudogenes with their functional counterparts using CLUSTALW 1.8 (24).

RESULTS

Because the pseudogenes contents of each of the 11 bacterial genomes, which together constitute four divergent genera of widespread human pathogens, were recognized by the same criteria, we can begin to make some generalizations about the appearance and maintenance of pseudogenes as well as the specific types of genes that become inactivated in host-associated bacteria. Although many of our comparisons involve strains that diverged very recently—for example, the homologous genes of *Y.pestis* CO92 and *Y.pestis* 91001 differ, on average, by only 0.07% at the DNA level—each genome accommodates a large and unique set of pseudogenes. The numbers of pseudogenes (including those newly and previously recognized) range from 27 in *S.aureus* MW2 to 337 in *Y.pestis* CO92, representing from 1 to 8% of the annotated genes in the genomes. In all species analyzed, the vast majority of pseudogenes correspond to hypothetical and unknown proteins, but include some IS elements and prophage sequences. Pseudogenes are most commonly produced by small frameshifts, involving deletions of 1 or 2 nt. But because stop codons are A + T-rich, the pervasive mutational biases in the two low-G + C Gram-positive taxa (*Streptococcus* and *Staphylococcus*) have resulted in relatively high fractions of pseudogenes generated by nonsense mutations compared with what is observed in the more G + C-rich genomes (*Yersinia* and *Vibrio*) ($r^2 = 0.67$, Spearman's rank correlation test, $P < 0.001$). In addition to the initial inactivating mutation, older pseudogenes, as recognized by comparisons of more distantly related taxa, are often truncated at one end. The specific characteristics of the pseudogene contents within these genomes are as follows.

S.aureus

The three strains of *S.aureus* represent meticillin-resistant bacteria, which arose recently and are responsible for large numbers of clinical infections (13). In the original annotations, not a single pseudogene was identified; however, we ascertained the presence of 42, 27 and 28 pseudogenes in strains Mu50, MW2 and N315, respectively. Frameshifts, caused by small indels, and truncations are the most common disruptions and generate, on average, 37.44 and 30.05% of all pseudogenes. However, transversions to stop codons have inactivated several genes (22.17% of all pseudogenes, Table 1). Unlike *S.aureus* strains N315 and MW2, in which most of the disrupted genes were originally annotated as hypothetical (or conserved hypothetical), one quarter of the 42 pseudogenes in strain Mu50 have defined functions.

Table 1. Number of pseudogenes and mutation types in *S.aureus* strains

	F	D	Ins	S	T	Ins + Δ	Gene inserted	Ins + T	F + D	S + D	F + S	IS	Total (% of genes)
Mu50 2714 genes	7	2	2	8	13	0	0	7	1	1	1	0	42 (1.55)
MW2 2632 genes	12	2	1	4	7	1	0	0	0	0	0	0	27 (1.03)
N315 2593 genes	5	4	2	8	6	0	1	0	0	0	1	1	28 (1.08)

F, frameshift; D, internal deletions; Ins, insertion of >4 bp; S, nonsense mutation; T, truncation; IS, insertion of IS element.

Table 2. Number of pseudogenes and mutations types in *S.pyogenes* strains

	F	D	Ins	S	T	Total (% of genes)
MGAS315 1865 genes	17	3	1	10	11	42 (2.25)
MGAS8232 1845 genes	19	2	1	12	15	50 (2.71)
M1 SF370 1696 genes	22	4	3	14	17	60 (3.53)
SSI-1 1861 genes	18	5	1	9	18	51 (2.74)

F, frameshift; D, internal deletions; Ins, insertion of >4 bp; S, nonsense mutation; T, truncation.

Table 3. Number of pseudogenes in *Vibrio* species

	V.cholerae		*V.parahaemolyticus*		*V.vulnificus* CMCP6		*V.vulnificus* YJ016	
Chromosomes	Chr I	Chr II	Chr I	Chr II	Chr I	Chr II	Chr I	Chr II
Number of coding genes	2770	1115	3080	1752	2952	1562	3262	1697
Number of pseudogenes	80 + 33[a]	23 + 23[a]	63	38	146	69	52	28
% Pseudogenes	4.1	4.1	2	2.2	4.9	4.4	1.6	1.6

[a]Pseudogenes identified in the original annotations.

S.pyogenes

Of the taxa that we examined, the Group A *S.pyogenes*, at <2000 genes, have the smallest genomes; and again, no pseudogenes were reported in any of the original annotations (14–17). Applying Ψ–Φ, we found 42, 50, 60 and 51 pseudogenes in strains MGAS315, MGAS8232, SF370 and SSI-1, respectively, of which 32 are shared by two or more strains.

Similar to the staphylococci, most of the pseudogenes in *S.pyogenes* are formed by frameshifts caused by equal numbers of insertions and deletions (Table 2). In each sequenced strain, many pseudogenes correspond to disruptions in IS elements or in prophage proteins. Among pseudogenes whose counterparts have been assigned functions, in strain SF370, the 50S ribosomal protein L32 is disrupted; and in strain SSI-1, there are pseudogenes for *smeZ* encoding the mitogenic exotoxin Z, a major immunoactive agent (25), *sclB*, encoding collagen-like surface protein (26), and *dnaQ*, the epsilon subunit of the DNA polymerase III specifying the 3′–5′ exonuclease proofreading activity (27). As in *Staphylococcus*, the mutations that produce pseudogenes result in an A + T enrichment of these genomes.

Vibrio

As opposed to the analyses of *S.aureus* and *S.pyogenes*, which involved comparisons of genome contents among very recently diverged strains within a species, the four *Vibrio* genomes constitute three nominal species, one of which *V.parahaemolyticus* shows only 72% amino acid identity and shares only 2615 of its 4832 genes with *V.cholerae* (22). In the original annotation (19), several pseudogenes were annotated in *V.cholerae* (33 on chromosome I and 23 and chromosome II), but none was reported to occur in the other three sequenced *Vibrio* genomes. In addition to confirming 54% of the previously recognized pseudogenes in *V.cholerae*, we identified 80 and 23 new pseudogenes on chromosomes I and II, respectively. Within each of the *Vibrio* genomes, the ratios of pseudo-to-functional genes are virtually identical on each of the two chromosomes; however, the different strains vary more than 2-fold their total numbers of pseudogenes (Table 3), with extremes corresponding to one of the two strains of *V.vulnificus*.

Because many of *Vibrio* pseudogenes were identified through comparisons of divergent species, it is difficult, in some cases, to reconstruct the initial inactivating event because the pseudogene contains several disrupting mutations (Figure 1). The phylogenetic relationships among these *Vibrio* species are established (28), allowing us to determine when most of the pseudogenes appeared during the evolution of this group (Figure 2) and to relate the age of the pseudogenes to the numbers of disruptive mutations. Although the relatively deep divergence of these strains offers broad window on the accumulation of pseudogenes, the majority of pseudogenes has been acquired relatively recently, with only 18 pseudogenes ancestral to the two strains of *V.vulnificus* and four pseudogenes ancestral to *V.vulnificus* and *V.parahaemolyticus*.

With respect to the original functions of the pseudogenes now present in each genome, proteins whose functions have been labeled as 'hypothetical' or 'unknown' represent between 37 and 54% of the pseudogenes, depending on the particular species. However, it is most surprising that several broadly distributed genes involved in nucleotide processing, repair or replication are presently pseudogenes in one of the *V.vulnificus* genomes. Among these are elongation factors Tu and G, DNA

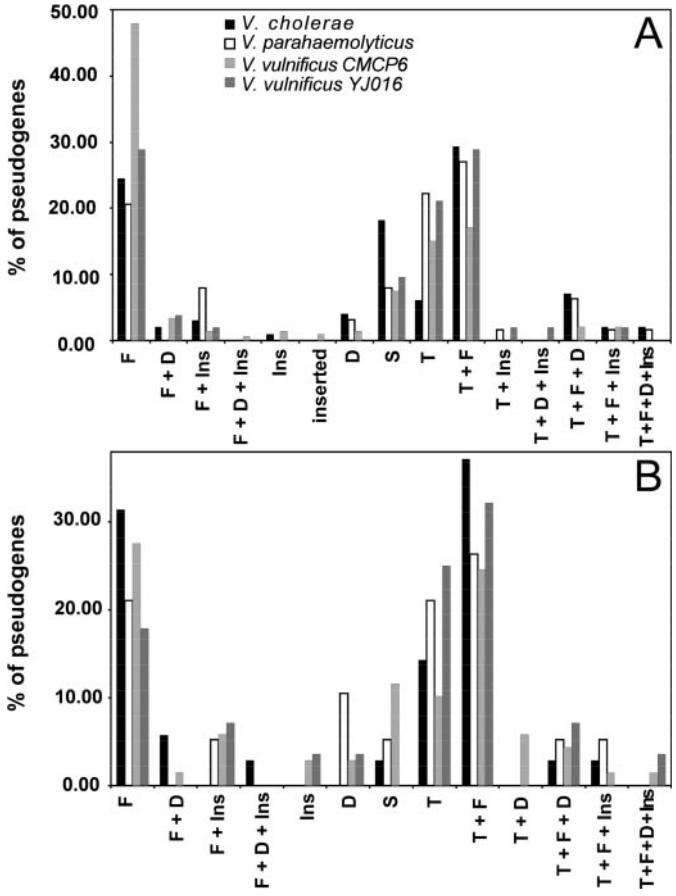

Figure 1. Vibrio pseudogenes on chromosome I (**A**) and chromosome II (**B**) classified according to the mutation type inactivating the gene. F, frameshift; D, internal deletions; Ins, insertion of >4 bp; S, nonsense mutation; T, truncation. Black boxes correspond to *V.cholerae*, white to *V.parahaemolyticus*, light gray to *V.vulnificus* CMCP6 and dark gray to *V.vulnificus* YJ016.

gyrase A, the helicases RuvB and RecG, and ribosomal proteins L5 and L35 (all of which are pseudogenes in strain CMCP6), and the DNA repair protein RecN (now a pseudogene in *V.vulnificus* YJ016).

Yersinia

Y.pestis is a recent human pathogen, which is thought to have emerged from *Y.pseudotuberculosis* <20 000 years ago (29). The three sequenced strains of *Y.pestis* represent two of the three known biovars: Mediaevalis (*Y.pestis* KIM and *Y.pestis* 91001) and Orientalis (*Y.pestis* CO92). Substantial numbers of pseudogenes were already identified in the previous annotations in these strains (149 pseudogenes in *Y.pestis* CO92, 54 in *Y.pestis* KIM, 140 in *Y.pestis* 91001 and 62 in *Y.pseudotuberculosis*), but our analysis uncovered a large population of new pseudogenes in each genome (188 new pseudogenes in *Y.pestis* CO92, 207 in *Y.pestis* KIM, 149 in *Y.pestis* 91001 and 124 in *Y.pseudotuberculosis*).

Unlike the other groups examined, a large proportion of the pseudogenes in the Yersiniae are truncated (Figure 3). Given that most new pseudogenes are created by frameshift mutations, the large most of these truncations probably occurred after the gene was already inactivated. The high numbers of pseudogenes in this group allows us to monitor the spectrum of

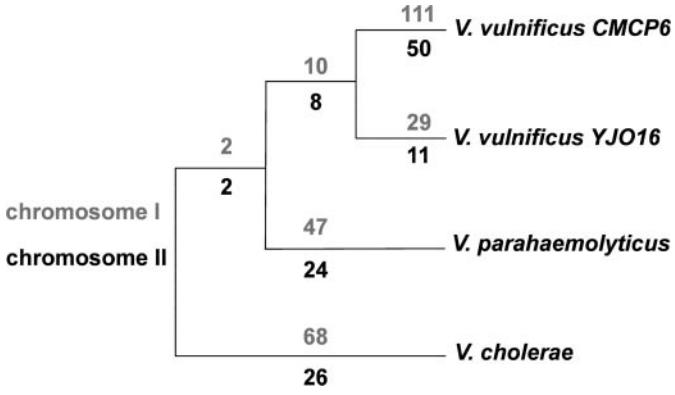

Figure 2. Phylogeny of Vibrio genomes showing numbers of pseudogenes that arose on each branch. Numbers above the lines (in gray) correspond to pseudogenes on chromosome I and those below the line (in black) correspond to pseudogenes on chromosome II. Note that despite the close relationship of the two sequenced strains of *V.vulnificus*, each harbors a largely unique set of pseudogenes.

mutational events that occur subsequent to the formation of a pseudogene. For example, among the pseudogenes caused by IS insertion, there is a case in *Y.pestis* CO92 where a recombination event between IS*100* elements linked two genes that were previously separated by 283 kb (Figure 4).

After producing the genome sequence of *Y.pseudotuberculosis*, Chain *et al*. (10) performed a comparative analysis to discover pseudogenes in *Y.pestis*. They designated 149 new pseudogenes in *Y.pestis* CO92 not recognized in the original annotation of this strain (8) where, coincidently, 149 pseudogenes were already annotated. Whereas the sequences of related genomes are expected to augment the ability to detect new pseudogenes, the difference in the numbers of new pseudogenes detected in *Y.pestis* CO92—the 149 by Chain *et al*. (10) versus the 100 that we detected when using this query sequence—is rather puzzling. Comparisons of the two sets of pseudogenes yielded matches for only 26 of the newly recognized pseudogenes. Eleven of the new pseudogenes reported by Chain *et al*. (10) are owing to IS insertions in non-coding regions (which are not considered by our analysis); however, the remaining 112 correspond to genes that are >90% of the length of their functional counterparts in *Y.pseudotuberculosis*, with some differing in length by only a single amino acid. Although this class of ORFs can certainly be considered truncated, many proteins retain function when shortened by <10% from either end and are not pseudogenes.

In contrast, among the set of 100 *Y.pestis* CO92 pseudogenes that we recognized using *Y.pseudotuberculosis* as query, 74 were not detected by Chain *et al*. (10): these include 13 that were originally annotated as functional CDSs and 61 that correspond to misannotations of non-coding DNA. Because the majority (66 of the 74) of these novel pseudogenes resulted from large truncations and additionally most approaches do not search non-coding regions for pseudogenes, these are generally difficult to recognize during genome annotation.

DISCUSSION

It was originally thought that bacteria would contain few, if any, pseudogenes because their genomes are small, gene dense

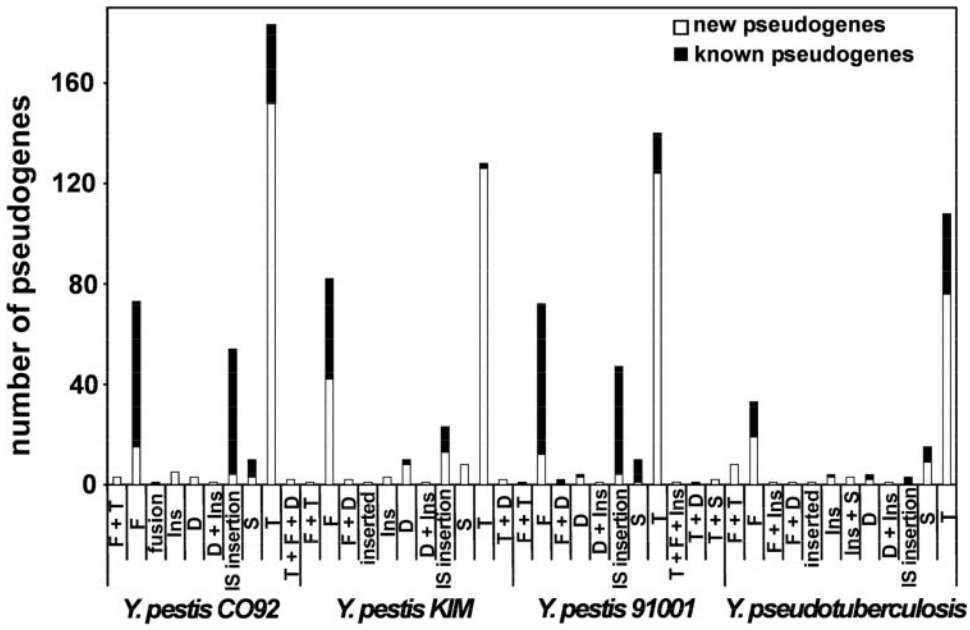

Figure 3. Percentage of types of mutations generating pseudogenes in *Yersinia*. Bars include previously annotated pseudogenes (black) and those newly recognized in the present study (white). F, frameshift; D, internal deletions; Ins, insertion of >4 bp; S, nonsense mutation; T, truncation.

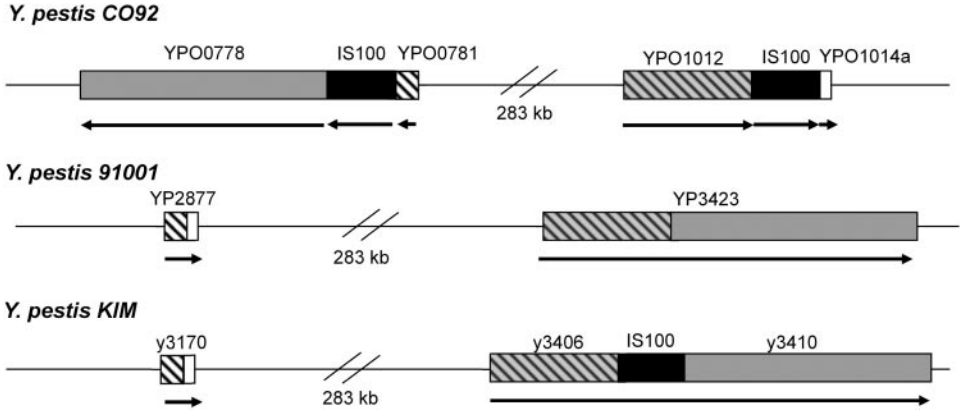

Figure 4. Example of rearrangements subsequent to the inactivation of two genes by IS elements in *Y.pestis* CO92. Arrows correspond to the gene orientation on the chromosome.

and contain very little non-coding DNA. However, virtually all bacterial genomes examined to date contain substantial numbers of pseudogenes, as evident from comparisons of the homologous sequences encoded by closely related genomes (5,30–34). Pseudogenes appear to be more common in the genomes of recent pathogens than in their benign or free-living relatives, a situation attributable to two factors: first, the reliance on the nutrients supplied within the host environment renders previously useful genes superfluous; and second, the reduction in the effective population size accompanying the infection of hosts results in a relaxation in the strength of selection and the accumulation of deleterious mutations.

On account of the potential for uncovering large numbers of inactivated genes and the availability of sequenced genomes whose relationships were suitable for comparative analysis, we focused on the pseudogenes present in four genera (*Staphylococcus*, *Streptococcus*, *Vibrio* and *Yersinia*), each of which contains human pathogens. The proportions of genes in these genomes that we recognize as pseudogenes range from ~1 to 8%, with the highest numbers detected in *Y.pestis* strain CO92. Although the original annotation of this strain reported 149 pseudogenes (8), our comparison of its genome to the newly available *Yersinia* sequences yielded a total of 337 pseudogenes.

The set of pseudogenes recognized by a comparative approach relies on two aspects of the analysis—the level of divergence between the genomes being compared and the thresholds applied to gene inactivation events. Although closely related strains might have broadly overlapping gene inventories, which foster the analysis of larger proportions of their genomes, there must be adequate time sequence or length variants to occur. At the other extreme, relatively little information about the functional status of genes can be gleaned from comparisons of very divergent homologs. Because the

comparisons performed in the present study involve genomes, whose pair-wise sequence divergences ranged from 1 to 10%, we can evaluate how the recognition of pseudogenes is linked to level of genetic relatedness. In *V.vulnificus* strain CMCP6, we recognized 96 pseudogenes on chromosome I when using the other *V.vulnificus* strain YJ016 as query, but only 64 pseudogenes (of which 42 were the same as those discovered with strain YJ016) from comparisons to *V.cholerae* and 66 (of which 50 overlapped with those discovered with strain YJ016) from comparisons to *V.parahaemolyticus*.

By performing comparative analyses among strains of known phylogenetic relationships, we also found that the majority of pseudogenes are of very recent origin and confined to a single genome. The paucity of older pseudogenes, i.e. those identical by descent in two present-day lineages—even in very closely related lineages—indicates that pseudogenes either degrade very quickly and are no longer recognizable or that they are eliminated rapidly from genomes. Although it is possible that there are pseudogenes ancestral to all constituent genomes and which are yet to be discovered due to the lack of a suitable outgroup, the rarity of pseudogenes originating at intermediate depths of these trees suggest that few, in any, such older pseudogenes exist in these genomes.

Although the mutational events that distinguish homologs are easy to recognize, it is rather more difficult to know whether these differences have disrupted gene function. Our analyses singled out those sequences that differed by frameshifts, nonsense mutations, transposon insertions or deletions from homologs in another genome and then inferred the events that inactivated gene function. Most genes disrupted by transposable elements are likely pseudogenes; however, the degree to which other forms of mutations affect the function depends upon their specific locations along a particular gene. Because not all frameshifts, deletions, rearrangements and stop mutations will inactivate a gene, we applied a fairly conservative criterion to distinguish pseudogenes from their functional counterparts, i.e. the disruption was required to alter >20% of the length or the primary sequence of the encoded protein.

This standard, based on both theoretical and experimental considerations, has been applied in other studies [e.g. (33,35)], and it certainly provides a more realistic view of the pseudogene contents of genomes than one considering all frameshifts, truncations and nonsense mutations as destroying gene function. For example, in their analysis of the genome sequence of *Y.pseudotuberculosis*, Chain *et al*. (10) have considered any ORFs shorter than its *Y.pestis* counterpart to be a pseudogene. However, the extent of length variation within the families of homologous proteins in the sequenced gamma-Proteobacteria (36) is, on average, 11.85% [and 10.71% when sequences from endosymbionts, which are known to possess smaller genes (37), are removed from the analysis]. Thus, genes differing by <10% in length, roughly 63% of the new pseudogenes in *Y.pestis* reported by Chain *et al*. (10), would best be reclassified as functional. In some genomes, it is possible that nonsense suppressors enable the production of functional proteins from genes that we deem as inactivated. In that genes disrupted by a single stop codon constitute a minor fraction of the pseudogenes within each of the strains that we considered, the presence of nonsense suppressors will not affect the functional status of pseudogenes in most cases.

Among the types of mutations forming pseudogenes, we find that frameshifts due to small indels are the most usual disruption, followed by truncations and nonsense mutations. To what extent might sequencing or assembly artifacts be the basis for the pseudogene-forming substitutions that we detected? These error rates are thought to be quite low for bacterial genome sequences; though admittedly, the sequences for some organisms have been altered and updated several times since their original release. Because most sequencing errors involve misinterpretations of individual nucleotides, their potential effect on our analyses can be evaluated in two ways. First, we note the number and spectrum of substitutions forming nonsense mutations differs among genomes and most closely resembles the inherent mutational biases of each organism. As mentioned, the more A + T-rich genomes were found to have more substitutions toward those nucleotides, whereas sequencing artifacts are not expected to be biased toward the particular G + C contents of the genome. Next, if small deletions were attributable to errors resulting from sequence compressions, frameshifts would occur preferentially within long tracts of identical nucleotides. Of 239 frameshifts detected in this study, 54 involve deletions occurring within runs of five or more repeated nucleotides. Of these, only 19 occur at G or C nucleotides, which are known to induce polymerase stuttering and the highest proportion of sequence compressions, hence it seems that very few pseudogenes are attributable to sequencing artifacts.

In *Y.pestis*, a large fraction of pseudogenes are caused by IS insertions, a situation also observed in *Shigella flexneri* (5). Both of these species harbor the highest numbers of pseudogenes among the sequenced members of their genera, and both are relatively recently emerged human pathogens: *Y.pestis* has emerged from *Y.pseudotuberculosis*, <20 000 years ago (29), and *Shigella* has emerged from *E.coli* <270 000 years ago (38). The accumulation of pseudogenes in recent pathogens reflects a process by which the effectiveness of selection has been reduced due to both the lower effective population sizes associated with host colonization coupled with the redundancy of many previously useful genes in the host environment. Moreover, the inefficiency of selection operating also accounts for the increases in the quantities of insertion sequences and the concomitant production of pseudogenes created by IS insertions. Despite these processes, *Staphylococcus* strains do not have very large numbers of pseudogenes (on average, the pseudogenes represent 1.30% of the genome); however, *S.aureus* genomes are relatively small and, some strong bias toward deletions, as observed in many bacteria (39), might prevent the accumulation of IS elements and pseudogenes.

In that, we recognize as pseudogenes only those ORFs that are truncated or otherwise disrupted over a sizeable portion of the coding sequence, only a fraction of the potential pseudogenes in the genome have been identified. Numerous other genes undoubtedly have been inactivated by missense mutations or by changes in their regulatory regions. Additionally, there are large sets of genes that are unique to each genome and whose functional status cannot be evaluated through sequence comparisons. Not only does the resolution of pseudogenes divulge the mutational and evolutionary processes occurring in bacterial genomes, such studies are relevant to current efforts to elucidate the functions of and interactions among every protein within a cell. Accurate

annotations are crucial to these investigations, because the scope of such large-scale analyses increases (often geometrically) with gene number.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by the National Institutes of Health, grant ROIGM-56120.

Conflict of interest statement. None declared.

REFERENCES

1. Andersson,J.O. and Andersson,S.G. (1999) Insights into the evolutionary process of genome degradation. *Curr. Opin. Genet. Dev.*, **9**, 664–671.
2. Cole,S.T., Eiglmeier,K., Parkhill,J., James,K.D., Thomson,N.R., Wheeler,P.R., Honore,N., Garnier,T., Churcher,C., Harris,D. *et al.* (2001) Massive gene decay in the leprosy bacillus. *Nature*, **409**, 1007–1011.
3. Rajashekara,G., Glasner,J.D., Glover,D.A. and Splitter,G.A. (2004) Comparative whole-genome hybridization reveals genomic islands in *Brucella* species. *J. Bacteriol.*, **186**, 5040–5051.
4. Rain,J.C., Selig,L., De Reuse,H., Battaglia,V., Reverdy,C., Simon,S., Lenzen,G., Petel,F., Wojcik,J., Schachter,V. *et al.* (2001) The protein–protein interaction map of *Helicobacter pylori*. *Nature*, **409**, 211–215.
5. Lerat,E. and Ochman,H. (2004) Psi–Phi: exploring the outer limits of bacterial pseudogenes. *Genome Res.*, **14**, 2273–2278.
6. Jin,Q., Yuan,Z., Xu,J., Wang,Y., Shen,Y., Lu,W., Wang,J., Liu,H., Yang,J., Yang,F. *et al.* (2002) Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res.*, **30**, 4432–4441.
7. Wei,J., Goldberg,M.B., Burland,V., Venkatesan,M.M., Deng,W., Fournier,G., Mayhew,G.F., Plunkett,G.,III, Rose,D.J., Darling,A. *et al.* (2003) Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. *Infect. Immun.*, **71**, 2775–2786.
8. Parkhill,J., Wren,B.W., Thomson,N.R., Titball,R.W., Holden,M.T., Prentice,M.B., Sebaihia,M., James,K.D., Churcher,C., Mungall,K.L. *et al.* (2001) Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature*, **413**, 523–527.
9. Deng,W., Burland,V., Plunkett,G.,III, Boutin,A., Mayhew,G.F., Liss,P., Perna,N.T., Rose,D.J., Mau,B., Zhou,S. *et al.* (2002) Genome sequence of *Yersinia pestis* KIM. *J. Bacteriol.*, **184**, 4601–4611.
10. Chain,P.S., Carniel,E., Larimer,F.W., Lamerdin,J., Stoutland,P.O., Regala,W.M., Georgescu,A.M., Vergez,L.M., Land,M.L., Motin,V.L. *et al.* (2004) Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. *Proc. Natl Acad. Sci. USA*, **101**, 13826–13831.
11. Ogata,H., Audic,S., Renesto-Audiffren,P., Fournier,P.E., Barbe,V., Samson,D., Roux,V., Cossart,P., Weissenbach,J., Claverie,J.M. and Raoult,D. (2001) Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. *Science*, **293**, 2093–2098.
12. Kuroda,M., Ohta,T., Uchiyama,I., Baba,T., Yuzawa,H., Kobayashi,I., Cui,L., Oguchi,A., Aoki,K., Nagai,Y. *et al.* (2001) Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. *Lancet*, **357**, 1225–1240.
13. Baba,T., Takeuchi,F., Kuroda,M., Yuzawa,H., Aoki,K., Oguchi,A., Nagai,Y., Iwama,N., Asano,K., Naimi,T. *et al.* (2002) Genome and virulence determinants of high virulence community-acquired MRSA. *Lancet*, **359**, 1819–1827.
14. Ferretti,J.J., McShan,W.M., Ajdic,D., Savic,D.J., Savic,G., Lyon,K., Primeaux,C., Sezate,S., Suvorov,A.N., Kenton,S. *et al.* (2001) Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proc. Natl Acad. Sci. USA*, **98**, 4658–4663.
15. Beres,S.B., Sylva,G.L., Barbian,K.D., Lei,B., Hoff,J.S., Mammarella,N.D., Liu,M.Y., Smoot,J.C., Porcella,S.F., Parkins,L.D. *et al.* (2002) Genome sequence of a serotype M3 strain of group A *Streptococcus*: phage-encoded toxins, the high-virulence phenotype, and clone emergence. *Proc. Natl Acad. Sci. USA*, **99**, 10078–10083.
16. Smoot,J.C., Barbian,K.D., Van Gompel,J.J., Smoot,L.M., Chaussee,M.S., Sylva,G.L., Sturdevant,D.E., Ricklefs,S.M., Porcella,S.F., Parkins,L.D. *et al.* (2002) Genome sequence and comparative microarray analysis of serotype M18 group A *Streptococcus* strains associated with acute rheumatic fever outbreaks. *Proc. Natl Acad. Sci. USA*, **99**, 4668–4673.
17. Nakagawa,I., Kurokawa,K., Yamashita,A., Nakata,M., Tomiyasu,Y., Okahashi,N., Kawabata,S., Yamazaki,K., Shiba,T., Yasunaga,R. *et al.* (2003) Genome sequence of an M3 strain of *Streptococcus pyogenes* reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. *Genome Res.*, **13**, 1042–1055.
18. Song,Y., Tong,Z., Wang,J., Wang,L., Guo,Z., Han,Y., Zhang,J., Pei,D., Zhou,D., Qin,H. *et al.* (2004) Complete genome sequence of *Yersinia pestis* strain 91001, an isolate avirulent to humans. *DNA Res.*, **11**, 179–197.
19. Heidelberg,J.F., Eisen,J.A., Nelson,W.C., Clayton,R.A., Gwinn,M.L., Dodson,R.J., Haft,D.H., Hickey,E.K., Peterson,J.D., Umayam,L. *et al.* (2000) DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature*, **406**, 477–483.
20. Kim,Y.R., Lee,S.E., Kim,C.M., Kim,S.Y., Shin,E.K., Shin,D.H., Chung,S.S., Choy,H.E., Progulske-Fox,A., Hillman,J.D. *et al.* (2003) Characterization and pathogenic significance of *Vibrio vulnificus* antigens preferentially expressed in septicemic patients. *Infect. Immun.*, **71**, 5461–5471.
21. Chen,C.Y., Wu,K.M., Chang,Y.C., Chang,C.H., Tsai,H.C., Liao,T.L., Liu,Y.M., Chen,H.J., Shen,A.B., Li,J.C. *et al.* (2003) Comparative genome analysis of *Vibrio vulnificus*, a marine pathogen. *Genome Res.*, **13**, 2577–2587.
22. Makino,K., Oshima,K., Kurokawa,K., Yokoyama,K., Uda,T., Tagomori,K., Iijima,Y., Najima,M., Nakano,M., Yamashita,A. *et al.* (2003) Genome sequence of *Vibrio parahaemolyticus*: a pathogenic mechanism distinct from that of *V. cholerae*. *Lancet*, **361**, 743–749.
23. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
24. Thompson,J.D., Higgins,D.G. and Gibson,T.L. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
25. Unnikrishnan,M., Altmann,D.M., Proft,T., Wahid,F., Cohen,J., Fraser,J.D. and Sriskandan,S. (2002) The bacterial superantigen streptococcal mitogenic exotoxin Z is the major immunoactive agent of *Streptococcus pyogenes*. *J. Immunol.*, **169**, 2561–2569.
26. Whatmore,A.M. (2001) *Streptococcus pyogenes sclB* encodes a putative hypervariable surface protein with a collagen-like repetitive structure. *Microbiology*, **147**, 419–429.
27. Scheuermann,R., Tam,S., Burgers,P.M., Lu,C. and Echols,H. (1983) Identification of the epsilon-subunit of *Escherichia coli* DNA polymerase III holoenzyme as the *dnaQ* gene product: a fidelity subunit for DNA replication. *Proc. Natl Acad. Sci. USA*, **80**, 7085–7089.
28. Thompson,F.L., Iida,T.M. and Swings,J. (2004) Biodiversity of vibrios. *Microbiol. Mol. Biol. Rev.*, **68**, 403–431.
29. Achtman,M., Zurth,K., Morelli,G., Torrea,G., Guiyoule,A. and Carniel,E. (1999) *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proc. Natl Acad. Sci. USA*, **96**, 14043–14048.
30. Coin,L. and Durbin,R. (2004) Improved techniques for the identification of pseudogenes. *Bioinformatics*, **20**, I94–I100.
31. Liu,Y., Harrison,P.M., Kunin,V. and Gerstein,M. (2004) Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. *Genome Biol.*, **5**, R64.
32. Harrison,P.M. and Gerstein,M. (2002) Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J. Mol. Biol.*, **318**, 1155–1174.
33. Homma,K., Fukuchi,S., Kawabata,T., Ota,M. and Nishikawa,K. (2002) A systematic investigation identifies a significant number of probable pseudogenes in the *Escherichia coli* genome. *Gene*, **294**, 25–33.
34. Andersson,J.O. and Andersson,S.G. (2001) Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. *Mol. Biol. Evol.*, **18**, 829–839.
35. Hutchison,C.A., Peterson,S.N., Gill,S.R., Cline,R.T., White,O., Fraser,C.M., Smith,H.O. and Venter,J.C. (1999) Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science*, **286**, 2165–2169.
36. Lerat,E., Daubin,V. and Moran,N.A. (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. *PLoS Biol.*, **1**, e19.

37. Charles,H., Mouchiroud,D., Lobry,J., Gonçalves,I. and Rahbe,Y. (1999) Gene size reduction in the bacterial aphid endosymbiont, *Buchnera*. *Mol. Biol. Evol.*, **16**, 1820–1822.
38. Pupo,G.M., Lan,R. and Reeves,P.R. (2000) Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc. Natl Acad. Sci. USA*, **97**, 10567–10572.
39. Mira,A., Ochman,H. and Moran,N.A. (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet.*, **17**, 589–596.