The "Domino Theory" of Gene Death: Gradual and Mass Gene Extinction Events in Three Lineages of Obligate Symbiotic Bacterial Pathogens

Tal Dagan, Ran Blekhman,†[1] and Dan Graur‡*

*Institüt für Botanik III, Heinrich-Heine Universität Düsseldorf, Düsseldorf, Germany; †Bioinformatics Undergraduate Program, Tel Aviv University, Ramat Aviv, Israel; and ‡Department of Biology and Biochemistry, University of Houston

During the adaptation of an organism to a parasitic lifestyle, various gene functions may be rendered superfluous due to the fact that the host may supply these needs. As a consequence, obligate symbiotic bacterial pathogens tend to undergo reductive genomic evolution through gene death (nonfunctionalization or pseudogenization) and deletion. Here, we examine the evolutionary sequence of gene-death events during the process of genome miniaturization in three bacterial species that have experienced extensive genome reduction: *Mycobacterium leprae*, *Shigella flexneri*, and *Salmonella typhi*. We infer that in all three lineages, the distribution of functional categories is similar in pseudogenes and genes but different from that of absent genes. Based on an analysis of evolutionary distances, we propose a two-step "domino effect" model for reductive genome evolution. The process starts with a gradual gene-by-gene-death sequence of events. Eventually, a crucial gene within a complex pathway or network is rendered nonfunctional triggering a "mass gene extinction" of the dependent genes. In contrast to published reports according to which genes belonging to certain functional categories are prone to nonfunctionalization more frequently and earlier than genes belonging to other functional categories, we could discern no characteristic regularity in the temporal order of function loss.

Introduction

Reductive evolution is typical of organisms that, like Blanche Dubois, have become dependent on "the kindness of strangers." Extensive genome reduction was observed in obligate intracellular parasites and endosymbionts (e.g., Cole et al. 2001; Parkhill et al. 2003; van Ham et al. 2003; Wei et al. 2003). In free-living organisms, reductive genome evolution is a very rare phenomenon and so far was only documented in the oxyphototrophic marine prokaryote, *Prochlorococcus marinus* (Dufresne, Garczarek, and Partensky 2005). During the adaptation of the parasite to its new niche, various gene functions may be rendered superfluous due to the fact that the host may supply these needs more cheaply and effortlessly. The resulting relaxation of the selection pressures may sometimes be followed by complete gene inactivation.

The most extreme example of reductive evolution is most probably the case of *Mycobacterium leprae*, the causative agent of Hanson's disease (leprosy). In a comparison of the *M. leprae* genome to the genome of the very closely related species, *Mycobacterium tuberculosis*, it was found that *M. leprae* lost about a quarter of its genes (Cole et al. 2001). Remnants of these once functional genes are found as 1,114 pseudogenes within the *M. leprae* genome. Two additional examples of extensive reductive evolution are *Shigella flexneri*, a pathogen responsible for many bacillary dysentery cases, and *Salmonella typhi*, the etiological agent of typhoid fever. A comparison between *S. flexneri* and *Escherichia coli* genomes revealed that at least 254 genes have become nonfunctional in the *S. flexneri* lineage (Wei et al. 2003). A comparison between the genomes of *Salmonella typhimurium* with that of *S. typhi* revealed that at least 204 protein-coding genes have become pseudogenes in *S. typhi* (Parkhill et al. 2001).

A comparison of the gene content in the genomes of obligate symbionts to that in free-living bacteria showed that many eliminated genes encode for information-processing functions, such as DNA repair, replication, transcription, and translation (Moran and Mira 2001; Andersson et al. 2002; Parkhill et al. 2003). Moreover, pathways of intermediate metabolism and amino acid biosynthesis are almost entirely absent from the genomes of pathogens (Moran and Wernegreen 2000; Moran 2002).

Gene death is the end outcome of a process that starts with the relaxation of functional constraints. When a mutation that incapacitates gene function is fixed in the population, the gene becomes a pseudogene. Functionless pseudogenes may linger within the genome or they may mutate to such an extent that they are no longer recognizable as homologs of functional genes. Alternatively, they may be deleted altogether. Therefore, the search for "lost functions" accompanying reductive evolution should include two groups of sequences: pseudogenes and absent genes (deleted or unrecognizable). A pseudogene is identified as such through its sequence similarity to a functional gene, and its nonfunctionality is verified by the presence of molecular defects precluding proper expression as well as by the lack of telltale signs of purifying selection on its sequence (Graur and Li 2000). Absent genes may be found by comparison of the genome under study to the genome of a closely related species. If we assume that at the time of divergence, the gene content of the two closely related species was identical, then genes that are present in one genome but absent from the other were most probably deleted during evolution. A different explanation for the existence of certain genes in the genome of one species and their absence in the other is horizontal transfer. This option, however, may have a much lower a priori probability because horizontally transferred genes frequently lose their functionality upon transfer (Kurland, Canback, and Berg 2003). We note that the number of absent genes identified by the above

[1] Present address: Department of Human Genetics, University of Chicago, Chicago, Ill.

Key words: genome reduction, *Mycobacterium leprae*, obligate symbioints, *Salmonella typhi*, *Shigella flexneri*, unitary pseudogenes.

E-mail: dgraur@uh.edu.

Mol. Biol. Evol. 23(2):310–316. 2006
doi:10.1093/molbev/msj036
Advance Access publication October 19, 2005

© The Author 2005. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved.
For permissions, please e-mail: journals.permissions@oxfordjournals.org

rationale may be inflated due to the existence of strain- or species-specific genes (e.g., Welch et al. 2002).

Determining the rate of gene extinction is important in elucidating the evolutionary sequence of events leading to parasitism. A spatial comparison of the *Buchnera aphidicola* genome with that of *E. coli* revealed large fragments that are present in *E. coli* but are absent from *B. aphidicola*. This piece of evidence led to the conclusion that genome reduction proceeds mainly through large deletions accompanying chromosome rearrangements (Moran and Mira 2001). In contrast, Silva, Latorre, and Moya (2001) proposed a model by which the genome of *Buchnera* sp. experienced a continual and gradual gene-by-gene nonfunctionalization process. Subsequently, the resulting pseudogenes accumulated neutral substitutions until they ceased to resemble the functional gene. According to Silva, Latorre, and Moya (2001), several such relic pseudogenes that reside in proximity to one another resulted in the creation of "gene deserts" that are frequently misinterpreted as large deletions of several genes at once. A third interesting possibility for the process of genome degradation takes into account the fact that genes (and their products) interact with one another. Thus, the inactivation of one gene may result in the relaxation of selection of the genes with which it interacts (e.g., genes that are part of the same metabolic pathway as the inactivated gene). This relaxation of selection triggered a mass nonfunctionalization of "dependent" genes. An example for such a "domino effect" of gene loss was proposed for *M. leprae*. It has been suggested that the loss of two genes involved in hypoxia response (*Dev*R and *Dev*S) resulted in the inability of the bacterium to respond to oxygen deprivation. Subsequently, other genes that were formerly involved in the response to hypoxia became superfluous and ceased to be constrained by function. Indeed, almost 70% of all genes involved in this response became pseudogenes or were deleted from the genome (Tyagi and Saini 2004).

The purpose of this study is to document the temporal dynamics of gene extinction during the process of reductive evolution. Toward this aim, we analyzed pseudogenes and absent genes from three bacterial lineages that have experienced extensive genome reduction during evolution.

Methods

Data

Our study required bacterial genomes that experienced substantial genome reduction and which contain numerous pseudogenes. We therefore sorted completely sequenced bacterial genomes according to the number of pseudogenes in their genomes using the National Center for Biotechnology Information (NCBI) annotation (http://www.ncbi.nlm.nih.gov/). The genomes that came at the top of the list were chosen for analysis: *M. leprae* (NC_002677), *S. typhi* (NC_003198), and *S. flexneri* (NC_004337). The genomes of two additional species, *Bordetella pertussis* and *Bordetella parapertussis*, also contained large numbers of pseudogenes. However, they were not included in the analysis because we could not find sufficiently divergent controls for the purposes of our analyses. For calculation of genetic distances, *M. leprae*, *S. typhi*, and *S. flexneri* were paired with *M. tuberculosis* CDC1551 (NC_002755), *S. typhimurium* LT2 (NC_003197), and *E. coli* K12 (NC_000913), respectively. Pseudogenes were identified according to the annotations in the January 2004 version of the NCBI database. Additional pseudogenes that have been detected in the genome of *S. flexneri* by Lerat and Ochmann (2004) were added.

Genetic Distances

Orthologous gene-pseudogene pairs were identified using the two-step reciprocal Blast routine (Altschul et al. 1990; Tatusov, Koonin, and Lipman 1997). In the first step, each pseudogene was Blasted against the genome of the paired species. The best Blast hit from among the hits with $e < 0.001$ was used in a Blast search against the species from which the pseudogene was derived. If the best Blast hit with $e < 0.001$ turned out to be the original pseudogene, then the pair of sequences was used in subsequent analyses. All other pairs were discarded from further consideration. The orthologous gene-pseudogene pair was aligned with ClustalW (Thompson, Higgins, and Gibson 1994). Genetic distances were calculated for each of the pairwise alignments according to Kimura's two-parameter (2p) model as implemented in the ODEN package (Ina 1994). The 2p model may be inappropriate when significant differences exist in GC content. We note, however, that although the GC content differences between the genomes of *M. leprae* and *M. tuberculosis* are quite large, the differences in GC content between the gene-pseudogene pairs used in our study are approximately 2 ± 1.6%. Under the assumption that genes evolve much slower than pseudogenes, we use the genetic distances as proxies for the age of the pseudogenes.

Absent Genes

The clusters of orthologous genes (COG) database (Tatusov et al. 2001) was used to identify absent genes. COG contains clusters of orthologous genes from more than 60 bacterial genomes. COGs that included a representative gene from the control species (e.g., *M. tuberculosis*) but not from the test species (e.g., *M. leprae*) were marked. A lost function for which we could not identify a homologous pseudogene was deemed to be absent.

Functional Assignment

Each pseudogene was classified according to the function of its functional ortholog into one of four main functional categories in the COGs database: (1) information processing, (2) cellular processes, (3) metabolism, and (4) unknown. In COG, these main functional categories are further subdivided into 26 functions. In order not to deal with very small numbers, we decided to use only the primary division into the four functional categories.

The genomes of *S. typhi* and *S. flexneri* are not included in the COG version we have used but the genomes of their closely related species *S. typhimurium* and *E. coli* are. Therefore, we used a reciprocal Blast procedure (as described above) between the pairs of these closely related species in order to classify the genes of these missing

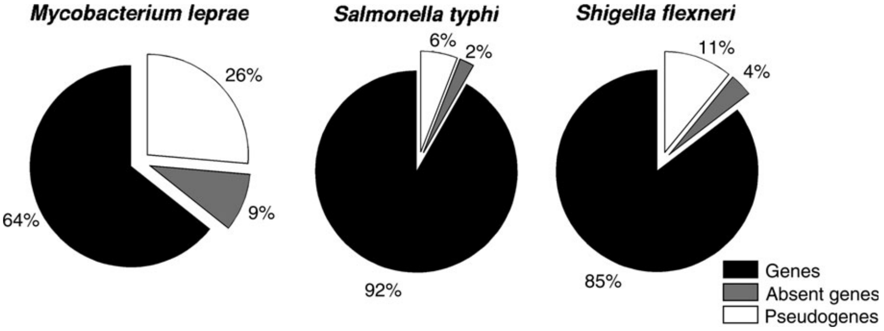

FIG. 1.—Proportions of genes, absent genes, and pseudogenes in three bacterial genomes.

genomes into existing COGs. The resulting COGs are listed in the COG*plus* database (http://cogplus.tau.ac.il/).

Statistical Tests

The contingency table χ^2 test with $\alpha = 0.05$ (Zar 1999) was used to compare among frequencies of genetic elements. Our null hypotheses were that the distributions of functional genes, absent genes, and pseudogenes are independent of lineage. Comparisons among assignments to functional categories were also performed with the χ^2 test, the null hypothesis being that the functional category is independent in genetic state (either present as a functional gene, present as a pseudogene, or absent from the genome).

To test whether the order of nonfunctionalization events that produced pseudogenes out of genes that performed a certain function is random, the pseudogenes were first ordered according to their distances from the functional ortholog. Then, all pseudogenes whose functional homologs belong to a certain functional category, for example, information processing, were assigned "1," and the rest were assigned "0." We then used the normalized runs test (Zar 1999) to determine whether the rank positions of the pseudogenes belonging to the functional category under study was random with respect to the other pseudogenes. We repeated this test for each of the four main functional categories. Because normalized runs tests involved multiple comparisons, we used the Bonferroni correction (Zar 1999). Thus, only results yielding α values lower than $0.05/4 = 0.0125$ were deemed significant. The comparison among the distances in the different functional groups was done by Kruskal-Wallis nonparametric test (Zar 1999). Post hoc comparisons were done using Mann-Whitney nonparametric test (Zar 1999).

Results

The *M. leprae* genome contains 1,605 genes, of which 1,168 are clustered into 886 COGs, and 1,114 pseudogenes, of which 556 pseudogenes could be matched to functional orthologs from *M. tuberculosis*. From these comparisons we deduced that 129 COGs are absent in *M. leprae*. The 2,085 genes from *S. typhi* were classified into 1,917 COGs, and the 2,131 genes from *S. flexneri* were classified into 1,658 COGs. Functional orthologs were found for 152 pseudogenes from *S. typhi* and 197 pseudogenes from *S. flexneri*. There were 50 and 72 absent COGs from the genomes of *S. typhi* and *S. flexneri*, respectively.

Genome Reduction

The proportions of functional genes, pseudogenes, and absent genes are significantly different between the three species ($P < 0.05$). Pseudogenes and absent genes comprise 35% of the genes in *M. leprae*, while *S. typhi* and *S. flexneri* have only lost 8% and 15% of their genes, respectively (fig. 1). The proportion of pseudogenes among the lost functions (pseudogenes and absent genes) is similar in the three species; in *M. leprae* the pseudogenes comprise 74% of the lost functions, while in *S. typhi* and *S. flexneri* there are pseudogenes for 75% and 78% of the lost functions, respectively (table 2).

The distribution of the genetic distances between the *M. leprae* pseudogenes and the *M. tuberculosis* genes

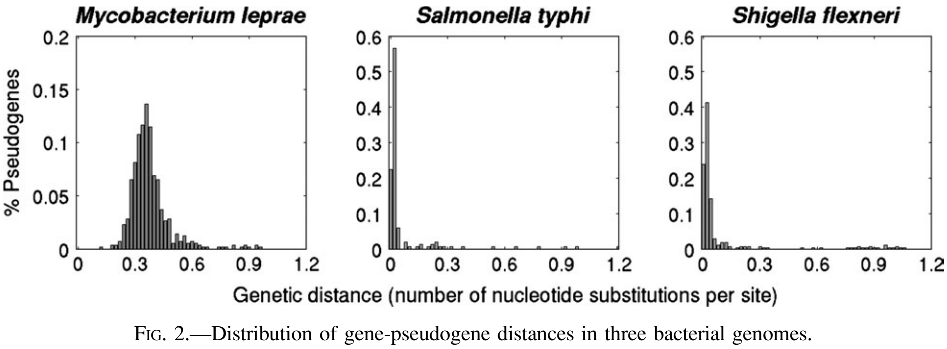

FIG. 2.—Distribution of gene-pseudogene distances in three bacterial genomes.

Table 1
Statistics of Genetic Distances of Pseudogenes in the Three Analyzed Genomes

Bacteria	Range	Mean ± SD	Median	Skew	Kurtosis
Mycobacterium leprae	0.13–0.95	0.37 ± 0.10	0.35	2.31	11.36
Shigella flexneri	0.00–1.05	0.10 ± 0.24	0.02	2.94	10.26
Salmonella typhi	0.00–0.98	0.06 ± 0.15	0.01	4.22	22.01

is leptokurtic (fig. 2 and table 1). About 90% of the distances are found within the very narrow range of 0.10–0.45. The rest 10% of the distances are dispersed over distances higher than 0.45. The distributions of distances in *S. typhi* and *S. flexneri* are leptokurtic as well, with a majority of distances very close to zero. The distribution of distances in *S. flexneri* is somewhat wider than that of *S. typhi*. About 90% of the distances are smaller than 0.4 in *S. flexneri*, while in *S. typhi* most distances are smaller than 0.2 (fig. 2).

Loss of Functions

In the genome of *M. leprae* we found significant differences in the functional compositions of genes, pseudogenes, and absent genes ($P < 0.05$; table 2). The fraction of information-processing sequences in the functional gene category (24%) was larger than their fraction in either the absent gene category (12%) or the pseudogene category (12%; fig. 3). The fraction of cellular processes genes is the smallest in the pseudogenes category. Genes involved in metabolism comprise about 50% of the *M. leprae* functional genome. A similar percentage is found as pseudogenes, while their share in the absent gene category was only 9%. No significant difference in the functional compositions of genes and pseudogenes was found in either *S. flexneri* or *S. typhi* ($P = 0.15$ and $P = 0.70$, respectively). However, the composition of absent genes is significantly different from that of genes and pseudogenes in these two species ($P < 0.05$; table 2).

We next tested whether or not the distribution of gene nonfunctionalization events is random with respect to functional category. In our three bacterial species, all main functions were found to be distributed randomly ($P > 0.05$). We tested for differences between the distributions of distances between pseudogenes and their respective functional homologs in the different functional groups in the three species. In the genome of *M. leprae*, we found no significant difference in the distribution of distances among the four main functional categories ($P = 0.26$). However, several pseudogenes from all categories exhibit extremely large distances. The extremes in the group of information-processing genes include two DNA-directed RNA polymerases (COGs 86 and 1595; Table S1, in the Supplementary Material online) and a transcriptional regulator from the *Lux*R family (COG 2771). In the cellular processes group, the extremes are two genes within the signal transduction function (COGs 589 and 2114). The vast majority of the extreme distances are seen in genes involved in metabolism, and most of these are involved in the biosynthesis of secondary metabolites (e.g., COGs 318, 3315, and 2124).

In the genome of *S. typhi*, we found statistically significant differences ($P < 0.05$) among the genetic distances from the different functional categories. Post hoc comparisons showed that pseudogenes belonging to the information-processing group show on average larger distances than the other functions. In contrast, in the genome of *S. flexneri*, we found no significant difference among distances from the different functional categories ($P = 0.12$), even though most of the pseudogenes that exhibit extremely large distances in *S. typhi* and *S. flexneri* are classified as genes involved in DNA replication, recombination, and repair. In *S. typhi* the extreme values are exhibited by two transposases (COGs 2963 and 2801) and a DNA polymerase (COG 389). In *S. flexneri* three transposases (COGs 3385, 2801, and 1662) exhibit extreme genetic distances. In the genome of *S. typhi*, pseudogenes, which are derived from genes involved in metabolism and cellular processes, are associated with exceptionally large genetic distances. It is, however, impossible to classify them into a single functional category. In the genome of *S. flexneri*, metabolism pseudogenes exhibiting extremely large distances are mostly derived from amino acid transport and metabolism genes (COGs 1897, 111, and 477), as well as from genes

Table 2
Number of COGs, Pseudogenic COGs, and Absent COGs in the Three Test Species

Test Species	Control Species	Function	Number of COGs in Control Species	Number of COGs in Test Species	Number of Pseudogenic COGs	Number of Absent COGs	Number of Lost Functions
Mycobacterium leprae	*Mycobacterium tuberculosis*	Information	273	214	44	15	59
		Cellular	193	137	38	18	56
		Metabolism	591	409	170	12	182
		Unknown	321	126	111	84	195
Salmonella typhi	*Salmonella typhimurium*	Information	326	297	18	11	29
		Cellular	392	365	19	8	27
		Metabolism	862	802	53	7	60
		Unknown	510	453	33	24	57
Shigella flexneri	*Escherichia coli*	Information	317	283	25	9	34
		Cellular	367	314	48	5	53
		Metabolism	829	710	88	31	119
		Unknown	429	351	51	27	78

NOTE.—The number of missing functions is calculated as the sum of pseudogenic COGs and absent COGs.

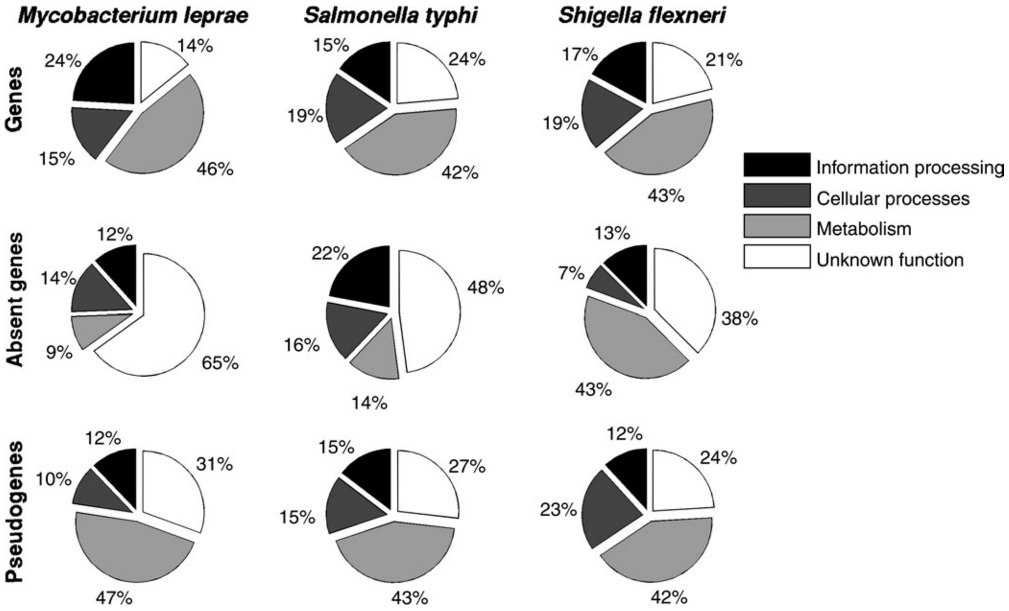

FIG. 3.—Division of genes, absent genes, and pseudogenes into main functional categories.

involved in cellular processes, mainly secretory pathways (COGs 3468 and 1989).

Only in *M. leprae* we had sufficient data to compare distances among subfunctions belonging to the same functional category. The genetic distances associated with pseudogenes derived from functions belonging to the information-processing category were significantly different from one another ($P = 0.04$). Post hoc comparisons showed that the distances associated with pseudogenes within the replication, recombination, and repair categories are significantly larger than those within the other subfunctions within the information-processes category. No significant differences were found among the functions within either the cellular processes or the metabolism category.

Discussion

In this study, we aimed to characterize the sequence of gene-death events during reductive evolution. We showed that there is no significant difference in the frequency of genes, absent genes, and pseudogenes in the three tested species. Absent genes comprise about a quarter of the lost functions in all the three genomes. Lawrence, Hendrix, and Casjens (2001) suggested that deletion of "junk" DNA is a defense mechanism in bacterial genomes against outside invasions. According to this suggestion, such mechanisms are weaker in intracellular parasites, which take advantage of the host defense, and as a result, there is an accumulation of pseudogenes within their genomes. According to this hypothesis, pseudogenes in extracellular parasites such as *S. flexneri* and *S. typhi* should be prone to deletion more than those in intracellular parasites, such as *M. leprae*. Our results do not support this hypothesis.

We assumed that the genetic distance between the pseudogene and its functional ortholog is proportional to the nonfunctionalization time. This allows us to use the distribution of distances to infer the temporal dynamics of gene death. All three distance distributions are highly leptokurtic (peaked). That is, many nonfunctionalization events are inferred to have occurred over very short periods of evolutionary time. The differences among the three distributions may be explained by either the pseudogenes in *M. leprae* being older than those in *S. typhi* and *S. flexneri* or by a higher deletion rate of pseudogenes in *S. typhi* and *S. flexneri*. The relative antiquity of the pseudogenes from *M. leprae* is supported by the fact that there were no statistically significant differences in the frequency of genes, absent genes, and pseudogenes among the three tested species.

The distance distributions in all three species are also skewed to the right, i.e., a significant minority of pseudogenes are quite old. This finding, in conjunction with the observation that the distance distribution exhibits a narrow peak, may be interpreted as evidence for a scenario involving nonfunctionalization events occurring continuously over relatively long periods of time and causing a gradual accumulating of pseudogenes, followed by a relatively brisk "mass extinction" of genes. We thus propose that the process of genome miniaturization begins slowly through gradual gene-by-gene nonfunctionalization, as suggested by Silva, Latorre, and Moya (2001). These "background" nonfunctionalization events damage the functionality of a certain biological pathway and trigger a domino effect of nonfunctionalization of other genes involved in the damaged pathway. The latter stages of the nonfunctionalization process are usually very rapid, as in the "mass extinction of genes" suggested by Moran and Mira (2001). Madan Babu (2003) proposed a two-step model for the accretion of pseudogenes in the *M. leprae* genome. According to this model, the mass nonfunctionalization of genes in *M. leprae* was triggered by two independent losses of several sigma factors.

Are certain functions more prone to be lost than others during reductive evolution? In *M. leprae*, gene decay has eliminated many metabolic functions, small molecule catabolism, energy metabolism, and synthesis and modification of macromolecules. Additional damaged systems include cell envelope constituents, transport and binding proteins, and proteins performing regulatory functions (Cole et al. 2001). Our results show that *M. leprae* lost about a third of the genes involved in metabolism and cellular processes and about a fifth of the information functions. In *S. typhi* all the main functional categories experienced about the same level of gene loss. Only 7% and 32% of the metabolism functions and cellular processes, respectively, were lost from the *M. leprae* genome. Information genes are the most conserved group in the *M. leprae* genome as only 21% of the information functions were lost and about quarter of the lost functions were deleted from the genome entirely. Most of the lost functions in *S. typhi* are related to metabolism. *Salmonella typhi* lost 7% of its metabolism functions, most of which are found as pseudogenes within its genome. Similarly, 9% of the information functions and 7% of the cellular processes functions were lost. Again, most of the lost functions can be still found within its genome. In *S. flexneri* most pseudogenes were derived from two functional subcategories: transport (belonging to the metabolism category) and cellular structure (within cellular processes). Approximately 14% of the cellular process functions and 15% of the metabolism functions were lost. The information category lost 11% of its genes.

To summarize the above, metabolism genes comprise the largest group among the lost functions. However, their fraction in the pseudogenes category is similar to their fraction in the functional genes in all three genomes. Moreover, the relative fraction of the main functional COG categories within pseudogenes and functional genes is similar in all three genomes. Therefore, we conclude that during the process of genome reduction genes belonging to all functional categories are equally prone to nonfunctionalization. We note, however, that information genes may be lost at slightly lower frequencies in *M. leprae* and *S. flexneri*. Our results agree with those for *Buchnera* (Gomez-Valero, Latorre, and Silva 2004).

In all three genomes, the functional composition of absent genes differs significantly from that of pseudogenes and functional genes. Because the functional composition of the genes category is highly similar between the test and control species, absent genes most likely resulted from deletion of genes or pseudogenes. In the first case, deletion events are expected to be functionally constrained. However, the significant difference between the proportions of absent genes and that of pseudogenes and genes suggests a more erratic evolutionary process. Therefore, the most probable option in our opinion is that the functional composition of absent genes results from a neutral process of deletion of nonfunctional DNA, i.e., pseudogenes, such as genome rearrangements.

In our study, we attempted to tackle the early stage of genome reduction. In particular, we were interested in whether or not certain functional categories of genes become superfluous earlier than others. Some tentative answers have been supplied from studies of organisms in early stages of adaptation to a symbiotic lifestyle (Dale et al. 2003; Moran 2003). The endosymbionts of maize and rice weevils have lost two DNA recombination-repair enzymes. It was suggested that this loss resulted in deleterious mutations and genome degradation. However, in a stable association, this loss of function may be compensated by a supply of metabolites from the host, in a process that will eventually give rise to a new holobiont (symbiont-host association). Our study yielded no significant differences in the distances of pseudogenes from their functional orthologs classified among the different functions. Moreover, the distribution of the nonfunctionalization events of the different functions was found to be significantly random temporally. A close examination of pseudogenes with relatively large genetic distances from their functional orthologs indicated that the oldest pseudogenes in *M. leprae* include RNA polymerases, signal transduction functions, and genes involved in the biosynthesis of secondary metabolites. In the genomes of *S. typhi* and *S. flexneri*, the oldest pseudogenes include DNA replication as well as recombination and repair functions, a few transposases, genes involved in secretory pathways, and amino acid transport and metabolism functions. Hence, we may conclude that there are no general rules pertaining to the genetic functions that should be lost to trigger a mass extinction of genes during reductive evolution.

Supplementary Material

Table S1 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Moshe Mevarech and Giddy Landan for their helpful feedback. This study was supported in part by the Israel Science Foundation. D.G. was supported in part by a Grant to Enhance and Advance Research from the University of Houston. T.D. was supported in part by a scholarship in Complexity Science from the Yeshaia Horvitz Association.

Literature Cited

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. **215**: 403–410.

Andersson, S. G. E., C. Alsmark, B. Canback, W. Davids, C. Frank, O. Karlberg, L. Klasson, B. Antoine-Legault, A. Mira, and I. Tamas. 2002. Comparative genomics of microbial pathogens and symbionts. Bioinformatics **18**:S17.

Cole, S. T., K. Eiglmeier, J. Parkhill, K. D. James, N. R. Thomson, P. R. Wheeler, N. Honore, T. Garnier, C. Churcher, D. Harris et al. 2001. Massive gene decay in the leprosy bacillus. Nature **409**:1007–1011.

Dale, C., B. Wang, N. Moran, and H. Ochman. 2003. Loss of DNA recombinational repair enzymes in the initial stages of genome degeneration. Mol. Biol. Evol. **20**:1188–1194.

Dufresne, A., L. Garczarek, and F. Partensky. 2005. Accelerated evolution associated with genome reduction in a free-living prokaryote. Genome Biol. **6**:R14.

Gomez-Valero, L., A. Latorre, and F. J. Silva. 2004. The evolutionary fate of nonfunctional DNA in the bacterial endosymbiont *Buchnera aphidicola*. Mol. Biol. Evol. **21**:2172–2181.

Graur, D., and W.-H. Li. 2000. Fundamentals of molecular evolution. Sinauer Associates, Sunderland, Mass.

Ina, Y. 1994. ODEN—a program package for molecular evolutionary analysis and database search of DNA and amino acid sequences. Comput. Appl. Biosci. **10**:11–12.

Kurland, C. G., B. Canback, and O. G. Berg. 2003. Horizontal gene transfer: a critical view. Proc. Natl. Acad. Sci. USA **100**: 9658–9662.

Lawrence, J. G., R. W. Hendrix, and S. Casjens. 2001. Where are the pseudogenes in bacterial genomes? Trends Microbiol. **9**:535–540.

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

Madan Babu, M. 2003. Did the loss of sigma factors initiate pseudogene accumulation in *M. leprae*? Trends Microbiol. **11**:59–61.

Moran, N. A. 2002. Microbial minimalism: genome reduction in bacterial pathogens. Cell **108**:583–586.

Moran, N. A. 2003. Tracing the evolution of gene loss in obligate bacterial symbionts. Curr. Opin. Microbiol. **6**: 512–518.

Moran, N. A., and A. Mira. 2001. The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. Genome Biol. **2**:R54.

Moran, N. A., and J. J. Wernegreen. 2000. Lifestyle evolution in symbiotic bacteria: insights from genomics. Trends Ecol. Evol. **15**:321–326.

Parkhill, J., G. Dougan, K. D. James, N. R. Thomson, D. Pickard, J. Wain, C. Churcher, K. L. Mungall, S. D. Bentley, M. T. G. Holden et al. (40 co-authors). 2001. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature **413**:848–852.

Parkhill, J., M. Sebaihia, A. Preston, L. D. Murphy, N. Thomson, D. E. Harris, M. T. G. Holden, C. M. Churcher, S. D. Bentley, K. L. Mungall et al. (52 co-authors). 2003. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat. Genet. **35**:32–40.

Silva, F. J., A. Latorre, and A. Moya. 2001. Genome size reduction through multiple events of gene disintegration in *Buchnera* APS. Trends Genet. **17**:615–618.

Tatusov, R. L., E. V. Koonin, and D. J. Lipman. 1997. A genomic perspective on protein families. Science **278**:631–637.

Tatusov, R. L., D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, and E. V. Koonin. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. **29**:22–28.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 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.

Tyagi, J. S., and D. K. Saini. 2004. Did the loss of two-component systems initiate pseudogene accumulation in *Mycobacterium leprae*? Microbiology **150**:4–7.

van Ham, R., J. Kamerbeek, C. Palacios, C. Rausell, F. Abascal, U. Bastolla, J. M. Fernandez, L. Jimenez, M. Postigo, F. J. Silva et al. (16 co-authors). 2003. Reductive genome evolution in *Buchnera aphidicola*. Proc. Natl. Acad. Sci. USA **100**:581–586.

Wei, J., M. B. Goldberg, V. Burland, M. M. Venkatesan, W. Deng, G. Fournier, G. F. Mayhew, G. Plunkett, D. J. Rose, A. Darling et al. (17 co-authors). 2003. Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. Infect. Immunol. **71**:2775–2786.

Welch, R. A., V. Burland, G. Plunkett III et al. (19 co-authors). 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc. Natl. Acad. Sci. USA **99**:17020–17024.

Zar, J. H. 1999. Biostatistical analysis. Prentice Hall, Upper Saddle River, N.J.

Pierre Capy, Associate Editor

Accepted October 10, 2005