a significant force on organelle gene-sequence evolution (*37*, *38*), but selective arguments for the architectural features of organelle genomes have remained elusive. Although it has been suggested that an intracellular "race to replication" is responsible for the streamlining of animal mitochondrial genomes (*38*), it is unclear whether broader phylogenetic patterns in organelle evolution can be explained by variation in intracellular competition. Perhaps differential metabolic demands and/or organelle turnover rates are involved, but this remains to be demonstrated. The arguments presented above help explain not just the phylogenetic variation in noncoding organelle DNA, but also the peculiar distribution of genetic code changes and mRNA editing. Thus, while serving as a useful null model, the hypothesis that genome evolution is strongly influenced by nonadaptive forces appears to have broad explanatory power, with variation in nuclear-genome architecture being primarily driven by variation in $N_e$ (*1*, *2*), and differences in μ making a major contribution to organelle evolution.

**References and Notes**


1. M. Lynch, *Mol. Biol. Evol.* **23**, 450 (2006).
2. M. Lynch, J. S. Conery, *Science* **302**, 1401 (2003).
3. Materials and methods are available as supporting material on *Science* Online.
4. M. W. Gray *et al.*, *Nucleic Acids Res.* **26**, 865 (1998).
5. W.-H. Li, *Molecular Evolution* (Sinauer Associates, Sunderland, MA, 1997).
6. D. R. Denver, K. Morris, M. Lynch, L. L. Vassilieva, W. K. Thomas, *Science* **289**, 2342 (2000).
7. N. Howell *et al.*, *Am. J. Hum. Genet.* **72**, 659 (2003).
8. R. S. Balaban, S. Nemoto, T. Finkel, *Cell* **120**, 483 (2005).
9. A. A. Johnson, K. A. Johnson, *J. Biol. Chem.* **276**, 38097 (2001).
10. P. A. Mason, R. N. Lightowlers, *FEBS Lett.* **554**, 6 (2003).
11. Y. Cho, J. P. Mower, Y. L. Qiu, J. D. Palmer, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 17741 (2004).
12. J. H. Gillespie, *Genetics* **155**, 909 (2000).
13. S. R. Palumbi, F. Cipriano, M. P. Hare, *Evol. Int. J. Org. Evol.* **55**, 859 (2001).
14. S. Berlin, H. Ellegren, *Nature* **413**, 37 (2001).
15. N. W. Gillham, *Organelle Genes and Genomes* (Oxford Univ. Press, Oxford, UK, 1994).
16. C. W. Birky Jr., T. Maruyama, P. M. Fuerst, *Genetics* **103**, 513 (1983).
17. M. Lynch, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 6118 (2002).
18. M. W. Hahn, J. E. Stajich, G. A. Wray, *Mol. Biol. Evol.* **20**, 901 (2003).
19. M. Lynch, D. G. Scofield, X. Hong, *Mol. Biol. Evol.* **22**, 1137 (2005).
20. M. J. H. van Oppen *et al.*, *J. Mol. Evol.* **55**, 1 (2002).
21. G. Pont-Kingdon *et al.*, *J. Mol. Evol.* **46**, 419 (1998).
22. T. L. Shearer, M. J. H. van Oppen, S. L. Romano, G. Worheide, *Mol. Ecol.* **11**, 2475 (2002).
23. R. D. Knight, S. J. Freeland, L. F. Landweber, *Nat. Rev. Genet.* **2**, 49 (2001).
24. T. H. Jukes, S. Osawa, *Comp. Biochem. Physiol. B* **106**, 489 (1993).
25. J. Swire, O. P. Judson, A. Burt, *J. Mol. Evol.* **60**, 128 (2005).
26. T. L. Horton, L. F. Landweber, *Curr. Opin. Microbiol.* **5**, 620 (2002).
27. D. V. Lavrov, W. M. Brown, J. L. Boore, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 13738 (2000).
28. P. Giegé, A. Brennicke, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 15324 (1999).
29. R. Hiesel, B. Combettes, A. Brennicke, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 629 (1994).
30. T. Tsudzuki, T. Wakasugi, M. Sugiura, *J. Mol. Evol.* **53**, 327 (2001).
31. T. Miyamoto, J. Obokata, M. Sugiura, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 48 (2004).
32. E. Kotera, M. Tasaka, T. Shikanai, *Nature* **433**, 326 (2005).
33. C. G. Phreaner, M. A. Williams, R. M. Mulligan, *Plant Cell* **8**, 107 (1996).
34. D. C. Shields, K. H. Wolfe, *Mol. Biol. Evol.* **14**, 344 (1997).
35. K. H. Wolfe, W. H. Li, P. M. Sharp, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 9054 (1987).
36. B. S. Gaut, B. R. Morton, B. C. McCaig, M. T. Clegg, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 10274 (1996).
37. M. Lynch, *Mol. Biol. Evol.* **14**, 914 (1997).
38. D. M. Rand, *Annu. Rev. Ecol. Syst.* **32**, 415 (2001).


39. Supported by grants from the NIH and NSF to M.L., an NSF predoctoral fellowship to B.K., and an NSF Integrative Graduate Education and Research Traineeship Program (IGERT) fellowship to S.S. Some key comments from J. Palmer led us to pursue this work. We are grateful to M. Neiman, J. Palmer, A. Richardson, and the reviewers for helpful comments.


**Supporting Online Material**

www.sciencemag.org/cgi/content/full/311/5768/1727/DC1
SOM Text
Fig. S1
Tables S1 to S5
References

10.1126/science.1118884

# The Nature and Dynamics of Bacterial Genomes

**Howard Ochman* and Liliana M. Davalos**

**Though generally small and gene rich, bacterial genomes are constantly subjected to both mutational and population-level processes that operate to increase amounts of functionless DNA. As a result, the coding potential of bacterial genomes can be substantially lower than originally predicted. Whereas only a single pseudogene was included in the original annotation of the bacterium *Escherichia coli*, we estimate that this genome harbors hundreds of inactivated and otherwise functionless genes. Such regions will never yield a detectable phenotype, but their identification is vital to efforts to elucidate the biological role of all the proteins within the cell.**

The organization of bacterial genomes is simple and elegant. These genomes are small, ranging from 500 to 10,000 kb, and are tightly packed with genes and other functional elements. The coding regions themselves are intronless and short, averaging a scant 1 kb, and are aligned almost contiguously along the chromosome. The common view is that the streamlining of bacterial genomes is the result of selection acting on replication efficiency and growth rates. Although this idea is warranted by the relatively low ceiling on bacterial genome size, there is no clear association between chromosome length and cell division rates either within or across bacterial species, implying that factors other than selection on overall chromosome size contribute to the compactness of bacterial genomes (*1*).

The elucidation of complete sequences has helped define the forces that shape bacterial genomes. Early research showed bacterial genomes to be tightly packed with functional elements, but unprecedented discoveries from genome analyses have shown that the genetic information encoded within bacterial genomes decays over evolutionary time scales (*2*–*4*). At first, this feature seems at odds with the high gene density observed in most bacterial genomes, but it is actually one of the primary determinants of their streamlined organization. All organisms accumulate mutations that can disrupt and degrade functional regions, but in bacteria (as well as in several eukaryotes) there is a mutational bias toward deletions over insertions (*1*, *5*–*7*). When disruptions occur in genes that are no longer required, the nonfunctional regions can be maintained in the genome for some time, but they gradually erode and are eventually eliminated, as is evident from comparisons of bacterial pseudogenes with their functional counterparts. In this manner, bacterial genomes maintain high densities of functional genes.

The primary force countering the erosion of genomes is natural selection, which serves to maintain the functional regions. The degree of selection varies along a continuum depending on the role of a gene in cell survival and replication: Genes with little contribution to fitness are more susceptible to inactivating and deletional mutations, whereas those that are critical will resist such mutation. Moreover, the degree of selection acting on any particular gene can change over time and according to a specific ecological context. For example, the inactivation or loss of one or more genes can increase the value of others, and changes in bacterial ecology or lifestyle might render some genes redundant (*8*).

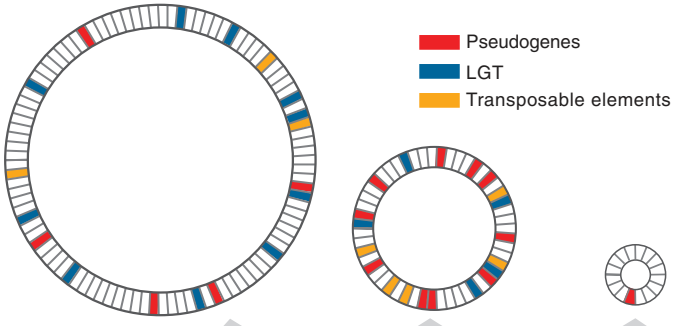
As important as the intensity of selection is the effectiveness of selection, which depends on population size and structure. In very large populations, deleterious mutations in beneficial genes are not likely to become fixed by chance; but in small populations, even useful genes can

Department of Biochemistry and Molecular Biophysics, University of Arizona, Tucson, AZ 85721, USA.

*To whom correspondence should be addressed. E-mail: hochman@email.arizona.edu

be inactivated and deleted. Population-level processes operate on all genomes, and their effects are even observed in the functional regions of the human genome (*9*). Thus, organisms with small population sizes are more prone to gene decay and loss, even for genes that are usually considered beneficial. This is the case for pathogens, which typically require only small inocula to infect each new host, resulting in a reduction in the effective size of their populations.

The impact of these factors is evident from the contents and organization of sequenced bacterial genomes. When compared with their free-living relatives, facultative bacterial pathogens are seen to harbor substantial numbers of pseudogenes (*5*, *10*) and a population structure that promotes the maintenance of deleterious mutations (*1*, *11*, *12*). Such pathogens represent an intermediate stage of genome erosion relative to the obligately host-associated lineages, whose genomes are highly reduced, having undergone massive inactivation and loss of nonessential genes (Fig. 1). When considering all bacterial genomes, there appears to be a set of only 50 to 100 genes that are universally maintained. The losses observed in degraded and reduced genomes include not only those genes that are clearly superfluous but also many that are beneficial but not essential, such as those involved in DNA repair (*2*, *13*, *14*).




| | Free-living | Recent or facultative pathogen | Obligate symbiont or pathogen |
|---|---|---|---|
| **Genome size** | Large (5-10 MB) | Intermediate (2-5 MB) | Small (0.5-1.5 MB) |
| **Number of pseudogenes** | Few | Many | Rare |
| **Incidence of LGT** | Frequent | Frequent to rare | Rare to none |
| **Selfish genetic elements** | Few | Common | Rare |
| **Genome organization** | Stable or unstable | Unstable | Stable |
| **Effective population size** | Large | Small | Small |

**Fig. 1.** Trends in the size and contents of bacterial genomes. Upon classifying bacteria according to their degree of association or dependence on a eukaryotic host, several general features emerge (*8*). Because of their relatively large population sizes, selection operates most effectively in free-living species—i.e., those that replicate in the environment independently of a host. Because selection is effective in removing deleterious sequences, these species usually possess large genomes containing relatively few pseudogenes (red) or mobile genetic elements (yellow). In recently derived pathogens, the availability of host-supplied nutrients combined with decreases in effective population sizes allows for the accumulation of pseudogenes and of transposable elements. In long-term host-dependent species, the ongoing mutational bias toward deletions has removed all superfluous sequences, resulting in a highly reduced genome containing few, if any, pseudogenes or transposable elements. LGT, lateral gene transfer.

### Redundant Regions in Bacterial Genomes

Some fraction of genes in each bacterial genome will not be functional because of ongoing mutational and population-level processes. Most genome annotations carry in them the implicit assumption that all predicted coding DNA sequences (CDSs) confer some (although often unspecified, hypothetical, or putative) function (*15*–*18*), but the critical task of identifying inactivated genes has been largely ignored.

To understand the functional status of a genome, we need to recognize the pseudogenes within it. Unfortunately, there are inconsistencies in the methods by which pseudogenes are defined (*19*, *20*). Hence, the assignment of pseudogenes must be based on some a priori assumptions about the spectrum of alterations in a gene that will abolish the function of its encoded protein. For example, genes in adenine + thymine (A + T)–rich genomes are often shorter than their homologs in other genomes, owing to their propensity toward mutations that form stop codons (*21*, *22*). Such truncations cannot automatically be taken to reflect a global inactivation of genes. Wherever possible, predicted pseudogenes should also be appraised in light of the physiological features of the cell. For example, the bacterial pathogen *Yersinia pestis*, unlike its sister taxon *Y. pseudotuberculosis*, requires exogenous methionine and phenylalanine for growth because it contains disabled versions of the *metB* and *pheA* genes (*4*, *20*).

### Finding Pseudogenes

The availability of genome sequences from closely related organisms has considerably altered the way that genes are recognized and annotated, but just how useful and accurate are genome annotations for the downstream analyses of microbial systems? Their reliability can be best assessed in *E. coli* K-12, which, owing to its long history as an experimental organism, has the highest proportion of genes for which functions have been determined directly (*23*–*25*). Furthermore, many genomic features first discovered in *E. coli* (such as gene and operon structure, occurrence of mobile elements, mutational patterns, and rates of gene exchange and acquisition) subsequently have been found to be characteristic of other bacterial groups.

Early molecular genetic analyses gave little indication that *E. coli*, or any other bacterial genome, might harbor substantial numbers of nonfunctional genes. The original annotation of the *E. coli* K-12 reported only a single pseudogene among its 4288 coding regions. Because this genome is expected to possess all of the hallmarks of free-living bacteria (Fig. 1), we set out to identify the inactivated and nonfunctional portion of the *E. coli* K-12 genome and to assess the number and authenticity of pseudogenes predicted by several analytical approaches.

### Genes and Proteins with Structural Alterations

We searched the annotated genome of *E. coli* K-12 (*26*) for genes that differed substantially in length from their homologs in the genome sequences of close relatives (i.e., enteropathogenic *E. coli* EDL933, uropathogenic *E. coli* CFT073, and *Shigella flexneri* 2a, which is more closely related to K-12 than either of the other two *E. coli* strains). As a consequence, more than 160 putative coding regions in *E. coli* K-12 were predicted to be pseudogenes (*27*, *28*). Contained within intergenic regions, there were 47 additional pseudogenes that were not recognized as such in the original annotation.

Although the criteria for distinguishing pseudogenes differ among studies, the overall rationale is the same: The predicted protein must be altered to a degree that abolishes its function. Such thresholds should derive from the known size and organization of functional domains within proteins, the observed length variation within individual gene families conserved across bacteria, and available information on experimentally disrupted proteins. In general, we classed as pseudogenes those cases

**Table 1.** Predicted pseudogenes in the *E. coli* K-12 genome. Functional categories, obtained from the 1 July 2004 annotation deposited in GenBank, are as follows: Hypothetical genes are CDSs of no known function; putative genes are CDSs with homology-derived functions; functional genes have an experimentally confirmed function. Essential genes are those determined experimentally to be required for aerobic growth in nutrient-rich conditions, as assayed in (*29*), with numbers of genes of annotated (known) and hypothetical (unknown) functions shown. Dashes denote the cases where CDSs predicted to be pseudogenes were not tested or yielded ambiguous results in (*29*). Ten genes of known function remain annotated as a "hypothetical" in GenBank. Conversely, there are 23 instances in which the GenBank annotation assigned named functions to CDSs based solely on sequence homology, whereas they should be classified as "putative." Gene names, identification numbers, genomic coordinates, strand orientations, and functional assignments of annotated CDSs in each category of predicted pseudogenes are available in tables S1 to S5.

| | Method of recognizing pseudogenes | | | | |
|---|---|---|---|---|---|
| | Disruption | | $K_a/K_s$ | SIFT‡ | Not expressed |
| | 2002§ | 2004¶ | | | |
| Total number detected* | 95 | 68 | 4 | 5 | 62 |
| Intergenic† | | 47 | | | |
| Hypothetical | 56 | 59‖ | 4 | 1 | 16 |
| Putative | 27 | 14‖ | 0 | 2 | 17 |
| Functional | 11 | 15 | 0 | 2 | 28 |
| Essential (known/hypothetical) | 0/5 | 2/4 | –/– | –/– | 0/6 |

*Numbers of CDSs in the original genome annotation of *E. coli* K-12 that are putative pseudogenes. Matches to intergenic regions not included. †Intergenic pseudogenes are disrupted copies of CDSs in other *E. coli* genomes but do not match any annotated CDS in the *E. coli* K-12 genome. Some entries are blank because the methods only apply to annotated CDSs. ‡SIFT (*35*) searches for amino acid replacements predicted to disrupt protein function. §Data from (*27*). ¶Data from (*28*). Pseudogenes considered in this column do not overlap with those reported in (*27*). ‖Includes matches of CDSs and intergenic regions to named genes in other *E. coli* genomes.

in which a stop codon or deletion has resulted in an encoded protein that is less than 80% of the length of its counterpart in the contrasted genome and those cases in which a frameshift or insertion has altered more than 20% of the amino acid sequence (*28*).

Because pseudogenes have no role in cellular processes, their status has to be supported by the lack of any direct evidence indicating a function. While such validations are conditional and do not prove that a region is inert, authentic pseudogenes are expected to be among those CDSs with functions that have not been (and can never be) resolved (Table 1). Despite the extensive biochemical, physiological, and molecular genetic characterization of *E. coli*, no functions have been uncovered for more than 90% of the 207 disrupted CDSs we found, bolstering their designation as pseudogenes. Most of these predicted pseudogenes are annotated as "hypothetical," although some (those designated "putative" in Table 1) have been assigned a provisional name or function based solely on their similarity to sequences already characterized in other organisms. Pseudogenes that are generated from genes of known function are among the most biologically interesting portions of a genome, because they disclose which specific functions are vulnerable to loss during the process of genome decay. Although it has been hypothesized that disrupted regions might take on a new role after their initial inactivation (*19*), there are as yet no verified cases in bacteria in which a pseudogene has assumed a function.

Extending this analysis to the other sequenced members of the *E. coli*/*Shigella* clade has resolved similarly high numbers of pseudogenes within each genome. Among these strains, the genome of the human pathogen *S. flexneri* displayed the highest level of genome erosion and degradation, containing more than 400 truncated or inactivated genes, as well as substantial numbers of translocatable insertion sequences (ISs) (*9*, *28*). Although the sequenced strains within the *E. coli*/*Shigella* clade are very closely related (averaging only 1% in sequence divergence), the set of disrupted genes harbored by each strain is distinct. The lack of pseudogenes shared among multiple strains indicates that pseudogenes are generated continually, but older pseudogenes are eliminated and only rarely persist in bacterial genomes.

Functions have been specified for 15 of the *E. coli* K-12 pseudogenes we recognized (Table 1), including two found to be essential for growth in nutrient-rich conditions (*29*). Sequencing errors resulted in incorrect assignment as pseudogenes, which have since been corrected; the other cases, which represent less than 3% of the total sample, are likely to be false positives.

### Unconstrained Substitutions

The most widely used method for identifying nonfunctional genes has been the $K_a/K_s$ test, which compares nucleotide substitution rates at synonymous sites ($K_s$) to those at nonsynonymous sites ($K_a$) (*30*). Regions that are unfettered by functional constraints, such as pseudogenes, are expected to have $K_a/K_s$ ratios that do not differ significantly from one. We searched for pseudogenes by estimating the $K_a/K_s$ ratio for each of nearly 1500 coding regions shared between *E. coli* K-12 and *Salmonella enterica* serovar Typhimurium strain LT2. We first excluded genes subject to extreme codon-usage bias as well as short leader peptides, which have been shown to give the appearance of neutral evolution (*30–32*), and then we applied a likelihood ratio test for detecting selection (*33*) to genes in the top 5% of the distribution ($K_a/K_s > 0.104$).

The $K_a/K_s$ ratio test exposed only four of the *E. coli/Salmonella* homologs as pseudogenes (Table 1). If most pseudogenes in the *E. coli* genome are newly derived (and strain specific), most of the evolutionary divergence between the *E. coli/Salmonella* homologs must have occurred while both copies were intact and undergoing purifying selection. There has simply not been sufficient time since the inactivation event for new mutations to equilibrate the levels of divergence at synonymous and nonsynonymous sites (*28*), resulting in $K_a/K_s$ ratios much less than one.

The $K_a/K_s$ test, although commonly used for detecting nonfunctional regions in eukaryotic genomes, is of limited value for uncovering pseudogenes in bacterial genomes, where most pseudogenes are of recent origin (*34*). Nonetheless, this method is robust and those few bacterial pseudogenes uncovered in this manner are likely to be authentic. As expected, no functions have been ascribed to any of the four high $K_a/K_s$ genes in the *E. coli* K-12 genome.

### Radical Amino Acid Replacements

Owing to the high gene density of bacterial genomes, the majority of point mutations result in amino acid replacements that can potentially alter protein function. To search for pseudogenes caused by missense mutations, we used a homology-based tool [termed Sorting Intolerant from Tolerant (SIFT)] (*35*) that sorts intolerant from tolerant substitutions. SIFT reveals differences in the primary structure of an *E. coli* protein that were sufficient to disrupt protein function when compared with orthologs present in at least three other sequenced members of the Gammaproteobacteria. The method examines all possible amino acid substitutions with a set of proteins and predicts those that are likely to be deleterious. SIFT identified only five genes with encoded proteins that contained potentially deleterious amino acid replacements (Table 1), suggesting that the proportion of single amino acid substitutions that completely abolishes protein function and generates a pseudogene is low.

### Transcriptional Pseudogenes

In addition to those encoding nonfunctional proteins, pseudogenes can arise when tran-

scription is permanently obstructed such that no protein is produced. To identify genes that are transcriptionally inactivated, we analyzed the results of studies of global gene expression in *E. coli* MG1655 (*36*–*38*), searching for genes that produce no detectable transcripts. There were 62 *E. coli* K-12 genes for which no transcripts were detected under any of the tested conditions, but for nearly half of these genes a function had been assigned. Because our analyses were limited to the three studies that used identical methods (and reflected a narrow range of growth conditions), it is likely that we overlooked the particular contexts under which many genes are expressed. By including an additional experiment that used similar but not identical methods (*39*), the occurrence of completely silenced genes was reduced to one (the putative acetyltransferase *ycdJ*). The low number of transcriptional pseudogenes reflects the evidence that the sequences regulating gene expression span much shorter regions than do the coding sequences within bacterial genomes, thereby providing a smaller target for inactivating mutations that silence genes.

### The Functional Component of Bacterial Genomes

By relying solely on comparisons of full genome sequences, we estimate that nearly 1 in 20 of the annotated coding regions in the *E. coli* K-12 genome are pseudogenes and that they tend to occur among genes of unknown function. Because *E. coli* K-12 has one of the smallest genomes of any *E. coli* strain (*40*), it is usually considered to have a relatively compact genome; however, the number of nonfunctional genes is likely to be higher than we predict.

There are also hundreds of genes in the *E. coli* K-12 genome whose functional status cannot be assessed by comparative approaches because they have no counterparts in closely related genomes. These include 104 known prophage- and IS element–associated genes, which may encode functional proteins but are generally assumed to be dispensable for host-cell survival and fitness, as well as numerous sequences acquired from distant sources.

Although it is not possible to specify which of the acquired genes are still functional on the basis of their sequence characteristics, the proportion that will be retained might be estimated by examining their incidence throughout the history of the *E. coli* lineage. Assuming that rates of gene transfer and loss have remained relatively constant over evolutionary time scales, we calculate that an additional 77 K-12 genes are expendable and could be removed from the genome (*41*). When combined with the 172 putative pseudogenes and the 104 phage- and IS-associated genes, this brings the total number of inactivated, nonfunctional, and expendable genes in the *E. coli* K-12 genome to 353, constituting 8% of the annotated coding regions.

Cumulatively, these analyses demonstrate that the coding potential of *E. coli* K-12 is less than that expected from the high density of predicted CDSs. However, the functional component of a genome is not confined to protein coding regions. There are not only transfer and ribosomal RNAs that operate in translation but also numerous small noncoding RNAs (sRNAs), with roles in regulation, as well as in RNA processing, stability, and degradation that have been detected in the intergenic regions of the *E. coli* genome. The numbers of predicted sRNAs in *E. coli* vary by an order of magnitude, depending on the particular procedure used to search the genome (*42*–*45*). However, more than a dozen sRNAs have been functionally characterized in *E. coli*, and there is strong support for at least 50 others (*46*, *47*).

The ultimate objective of genomics is to elucidate the biological role and phenotypic consequence of every nucleotide within a genome. Given that most bacterial lineages experience a massive turnover of genes through acquisition of foreign DNA and loss of existing sequences, any given genome will possess the "debris" generated by these dynamics. Delineating these sites and regions must be a central goal if we are to attain a comprehensive understanding of genomes.

*E. coli* K-12, and indeed most other bacteria, contains fewer functional protein-coding regions than anticipated, but many functional elements may lay hidden in the noncoding portion of genomes. Such insights alter our expectations about the contents and organization of bacterial genomes. We now recognize that many apparent coding regions might never surrender a meaningful phenotype, no matter how sensitive the assay. Alternatively, those genomic regions not regularly surveyed are likely to control many phenotypes.

### References and Notes


1. A. Mira, H. Ochman, N. A. Moran, *Trends Genet.* **17**, 589 (2001).
2. S. G. E. Andersson *et al.*, *Nature* **396**, 133 (1998).
3. S. T. Cole *et al.*, *Nature* **409**, 1007 (2001).
4. J. Parkhill *et al.*, *Nature* **413**, 523 (2001).
5. J. O. Andersson, S. G. E. Andersson, *Curr. Opin. Genet. Dev.* **9**, 664 (1999).
6. J. O. Andersson, S. G. E. Andersson, *Mol. Biol. Evol.* **18**, 829 (2001).
7. D. A. Petrov, D. L. Hartl, *J. Hered.* **91**, 221 (2000).
8. N. A. Moran, G. R. Plague, *Curr. Opin. Genet. Dev.* **14**, 627 (2004).
9. P. D. Keightley, M. J. Lercher, A. Eyre-Walker, *PLoS Biol.* **3**, e42 (2005).
10. Q. Jin *et al.*, *Nucleic Acids Res.* **30**, 4432 (2002).
11. D. I. Andersson, D. Hughes, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 906 (1996).
12. V. S. Cooper, R. E. Lenski, *Nature* **407**, 736 (2000).
13. C. M. Fraser *et al.*, *Science* **270**, 397 (1995).
14. N. A. Moran, J. J. Wernegreen, *Trends Ecol. Evol.* **15**, 321 (2000).
15. P. Bork, *Genome Res.* **10**, 398 (2000).
16. S. Bocs, A. Danchin, C. Medigue, *BMC Bioinformatics* **3**, 5 (2002).
17. D. W. Ussery, P. F. Hallin, *Microbiology* **150**, 2015 (2004).
18. E. Kolker *et al.*, *Nucleic Acids Res.* **32**, 2353 (2004).
19. H. Ogata *et al.*, *Science* **293**, 2093 (2001).
20. P. S. Chain *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 13826 (2004).
21. J. L. Oliver, A. Marin, *J. Mol. Evol.* **43**, 216 (1996).
22. H. Charles, D. Mouchiroud, J. Lobry, I. Goncalves, Y. Rahbe, *Mol. Biol. Evol.* **16**, 1820 (1999).
23. M. H. Serres, S. Goswami, M. Riley, *Nucleic Acids Res.* **32**, D300 (2004).
24. I. M. Keseler *et al.*, *Nucleic Acids Res.* **33**, D334 (2005).
25. R. V. Misra, R. S. Horler, W. Reindl, I. I. Goryanin, G. H. Thomas, *Nucleic Acids Res.* **33**, D329 (2005).
26. F. R. Blattner *et al.*, *Science* **277**, 1453 (1997).
27. K. Homma, S. Fukuchi, T. Kawabata, M. Ota, K. Nishikawa, *Gene* **294**, 25 (2002).
28. E. Lerat, H. Ochman, *Genome Res.* **14**, 2273 (2004).
29. S. Y. Gerdes *et al.*, *J. Bacteriol.* **185**, 5673 (2003).
30. M. Nei, S. Kumar, *Molecular Evolution and Phylogenetics* (Oxford Univ. Press, New York, 2000).
31. H. Ochman, *Trends Genet.* **18**, 335 (2002).
32. J. G. Lawrence, *Trends Genet.* **19**, 131 (2003).
33. Z. Yang, *Mol. Biol. Evol.* **15**, 568 (1998).
34. E. Lerat, H. Ochman, *Nucleic Acids Res.* **33**, 3125 (2005).
35. P. C. Ng, S. Henikoff, *Nucleic Acids Res.* **31**, 3812 (2003).
36. A. Lobner-Olesen, M. G. Marinus, F. G. Hansen, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 4672 (2003).
37. R. W. Corbin *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 9232 (2003).
38. S. J. Arends, D. S. Weiss, *J. Bacteriol.* **186**, 880 (2004).
39. D. W. Selinger *et al.*, *Nat. Biotechnol.* **18**, 1262 (2000).
40. U. Bergthorsson, H. Ochman, *Mol. Biol. Evol.* **15**, 6 (1998).
41. The fraction of acquired genes that becomes dispensable is estimated by comparing the absolute numbers of acquired genes that are unique to *E. coli* K-12 to those shared among *E. coli* strains. Since its split from *Salmonella*, the *E. coli* lineage leading to strain K-12 acquired 394 genes, of which 91 are restricted to the genome of the K-12 strain (*48*). These 91 genes accumulated over a period spanning just under 5% of the history of the lineage, with the disproportionate representation of recent acquisitions attributable to the loss of older arrivals. If the acquisition rate has been relatively constant, only 303 genes remain from a total of 2093 that were introduced up to the time when K-12 diverged from other *E. coli* strains. Because most pseudogenes are of recent origin, the attrition rate of 85% can be applied to the 91 genes unique to *E. coli* K-12, yielding 77 expendable genes. In this regard, it is noteworthy that M. Taoka *et al*. (*49*) found that only a small fraction (~10%) of genes mapping to *E. coli* regions specific to K-12 are translated into proteins.
42. L. Argaman *et al.*, *Curr. Biol.* **11**, 941 (2001).
43. R. J. Carter, I. Dubchak, S. R. Holbrook, *Nucleic Acids Res.* **29**, 3928 (2001).
44. E. Rivas, R. J. Klein, T. A. Jones, S. R. Eddy, *Curr. Biol.* **11**, 1369 (2001).
45. K. M. Wassarman, F. Repoila, C. Rosenow, G. Storz, S. Gottesman, *Genes Dev.* **15**, 1637 (2001).
46. R. Hershberg, S. Altuvia, H. Margalit, *Nucleic Acids Res.* **31**, 1813 (2003).
47. J. Vogel *et al.*, *Nucleic Acids Res.* **31**, 6435 (2003).
48. V. Daubin, H. Ochman, *Genome Res.* **14**, 1036 (2004).
49. M. Taoka *et al.*, *Mol. Cell. Proteomics* **3**, 780 (2004).


50. Funded by NIH grant GM56120 to H.O. We thank N. Moran, A. Corthals, E. Groisman, and the three anonymous reviewers for their comments on the manuscript; E. Lerat for providing information about *E. coli* pseudogenes; P. Ng for assistance with SIFT; and S. Miller and B. Nankivell for technical support. The GenBank accession number for the *E. coli* MG1655 genome is U00096. Gene designations in tables S1 to S5 follow those in the most recent release (obtained from www.genome.wisc.edu/sequencing/updating.htm).


### Supporting Online Material

www.sciencemag.org/cgi/content/full/311/5768/1730/DC1
Tables S1 to S5

10.1126/science.1119966