

Microbial Minimalism: Genome Reduction in Bacterial Pathogens

Minireview

Nancy A. Moran¹

Department of Ecology and Evolutionary Biology
University of Arizona
Tucson, Arizona 85721

When bacterial lineages make the transition from free-living or facultatively parasitic life cycles to permanent associations with hosts, they undergo a major loss of genes and DNA. Complete genome sequences are providing an understanding of how extreme genome reduction affects evolutionary directions and metabolic capabilities of obligate pathogens and symbionts.

Introduction

Known genome sizes of bacteria range from under 0.6 to ~10 megabases (mb). At the lowest extreme of this range are the mycoplasmas and related bacteria, with genome sizes reported as low as 530 kilobases. Among the many early revelations from molecular phylogenetic studies of bacteria (Woese, 1987) was the recognition that the mycoplasmas represented an evolutionarily derived condition rather than a primitive one, as once believed. Now that phylogenetic relationships and genome sizes are determined for a broader array of organisms, it is clear that the mycoplasmas are just one example of genome shrinkage that has occurred in a variety of obligately host-associated bacteria. Other prominent examples are *Rickettsia* and related pathogens within the α -proteobacteria; insect symbionts within the γ -proteobacteria, as exemplified by *Buchnera aphidicola* in aphids; the chlamydiae; and the parasitic spirochetes, such as *Borrelia burgdorferi* (the agent of Lyme disease).

Small genome size in these organisms is associated with other distinctive genetic features, including rapid evolution of polypeptide sequences and low genomic G+C content (Figure 1). The repeated evolution of these features in unrelated bacteria indicates that an obligate association with host tissues somehow promotes genome reduction. Understanding the causes of these genome level changes will help to reveal the processes that are important in pathogen and symbiont evolution.

Over 50 eubacterial genomes are now fully sequenced and annotated, with many more near completion. These sequences have corroborated a link between obligate host-associated lifestyles and a distinctive set of genomic features that include small size. Furthermore, they are yielding detailed information on the evolutionary basis for DNA loss and the functional implications of this loss.

Which Genes Disappear

Bacterial genomes are comprised mostly of coding genes: in almost all of the fully sequenced genomes, over 80% of the sequence consists of intact ORFs. This, combined with the fact that gene length is effectively

constant (averaging ~1 kb per gene) across genomes, implies that small genomes have few genes and correspondingly limited metabolic capabilities. Whereas bacteria with free-living stages, such as *Escherichia coli*, *Salmonella* species, or *Bacillus* species, typically encode 1500 to 6000 proteins, obligately pathogenic bacteria often encode as few as 500 to 1000 proteins (Figure 1).

The simplest possibility would be that reduced genomes converge on a set of universal genes that underlie the core processes of cellular growth and replication, with each genome also containing some loci corresponding to that species' ecology or host-relationship. But this possibility is contradicted by the full genome sequences. The set of orthologs that are universal, or nearly so, among eubacteria constitutes only a small proportion (<15%) of each genome, totaling about 80 genes (Koonin, 2000). Thus, each lineage has taken a different evolutionary route to minimalism. Since universal cellular processes require many more than 80 genes, differences in gene inventories imply that the same functions can be achieved by retention of nonhomologous genes.

Use It or Lose It

Nevertheless, some intelligible patterns do emerge from comparing gene sets of fully sequenced genomes. One clear basis for genome reduction is that bacteria living continuously in hosts can obtain many compounds of intermediate metabolism from host cytoplasm or tissue; thus, they can discard the corresponding biosynthetic pathways and genes. Such elimination of unneeded pathways explains a substantial proportion of observed gene losses. For instance, many of the genes involved in energy metabolism are eliminated from *Rickettsia* species, *Mycoplasma* species, and *Buchnera*, which can rely on consistent availability of particular energy substrates from hosts (Figure 2).

Likewise, most small genomes have eliminated genes underlying biosynthesis of amino acids, which are taken up from host cells. A remarkable exception—of the type that proves the rule—occurs in *Buchnera*, an obligate maternally transmitted symbiont of aphids. A basis for the mutualism is the provisioning of essential amino acids to hosts, and *Buchnera* retains 54 genes (comprising ~10% of its genome) for biosynthesis of essential amino acids, but has lost pathways for amino acids that the host can produce itself (Shigenobu et al., 2000). Pathways for nucleotide biosynthesis, and vitamin biosynthesis, are also missing from many reduced genomes. Individual genomes retain unique combinations of anabolic pathways, probably relating to different environmental conditions.

Small genomes have lost many regulatory elements, including sigma factors. This aspect also may be partly attributable to a lack of need: living continuously within the host eliminates the extreme environmental fluctuations encountered by free-living bacteria.

Use It, but Lose It Anyway

The premise that useful genes are retained and useless ones eliminated oversimplifies the evolutionary processes that affect persistence of genes in genomes.

¹Correspondence: nmoran@u.arizona.edu

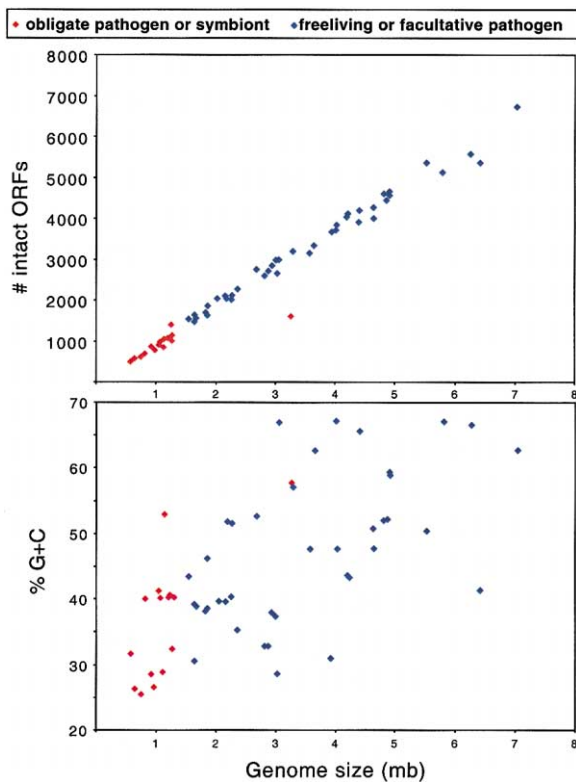


Figure 1. Size of Eubacterial Genomes in Relation to Number of intact ORFs (top) and % G+C Content in Genome (bottom)

The red spots correspond to obligate pathogens, belonging to a variety of unrelated lineages. *M. leprae* contains inactivated genes not included in the tally of intact ORFs. Included are all eubacterial genomes available January 2002 in NCBI Entrez Genomes (http://www.ncbi.nlm.nih.gov:80/PMGifs/Genomes/eub_g.html), with numbers of intact ORFS taken from this annotation.

And it cannot explain why many of the discarded genes encode products that would seem to be just as useful in obligate pathogens as in other organisms. Many eliminated genes encode accessory proteins or regulatory products involved in universal cellular processes, including replication, transcription, and translation (e.g., Andersson, et al. 1998; Moran and Wernegreen, 2000; Figure 2). Some genes underlying DNA recombination and repair pathways are eliminated from every small genome, although the precise set discarded varies. Also, small genomes contain fewer tRNAs, retaining only one for many amino acids. Thus, a single anticodon must pair with multiple codons, presumably resulting in less efficient translation machinery. It is not clear why obligate intracellular pathogens would benefit by retaining fewer tRNAs and fewer DNA repair enzymes.

Also important in the evolutionary processes determining patterns of gene persistence are the changes in population structure that accompany a shift to an obligately pathogenic lifestyle. Acquisition of an obligately host-associated lifestyle will often greatly diminish the genetic population size of a lineage, due to restricted habitat (hosts), and to bottlenecks in bacterial numbers at the time of infection (Andersson and Kurland, 1998; Moran and Wernegreen, 2000). The resulting

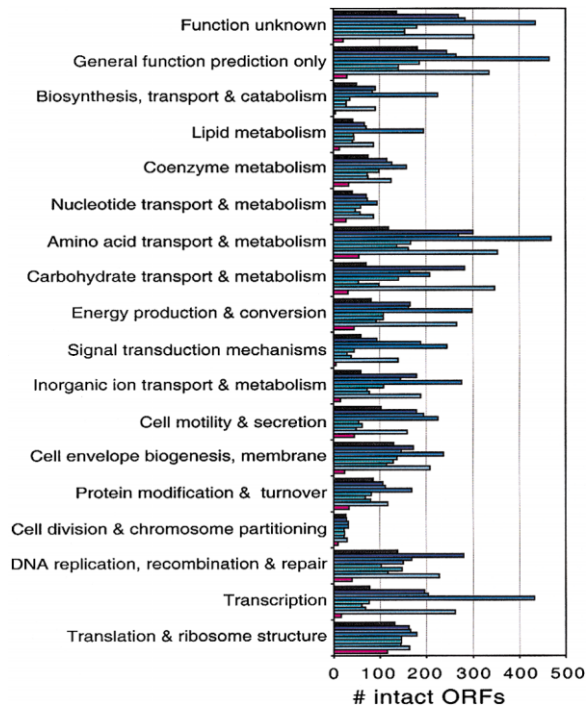


Figure 2. Numbers of ORFs Assigned to Different Functional Categories for Fully Sequenced Genomes of Members of the β - and γ -Proteobacteria

From top, they are *Xylella fastidiosa*, *Yersinia pestis*, *Vibrio cholerae*, *Pseudomonas aeruginosa*, *Pasteurella multocida*, *Neisseria meningitidis*, *Haemophilus influenzae*, *Escherichia coli* K12, and *Buchnera aphidicola*. *Buchnera*, the only organism in this group that is obligately associated with hosts and that has a highly reduced genome, shows reduced numbers of ORFs in all categories. Categories are those corresponding to the COGs database (Tatusov, R.L., Koonin, E.V., and Lipman, D.J., 1997, *Science* 278, 631-637, <http://www.ncbi.nlm.nih.gov/COG/xindex.html>).

genetic drift can lead to the fixation of mutations that inactivate genes that are useful (but obviously not essential), or they can decrease the efficiency of gene products. Thus, entirely useless genes, such as those not needed in a newly acquired niche, will lose functionality due to mutations that disrupt the coding region, but even beneficial genes may be lost or degraded if genetic drift precludes effective purifying selection.

How Many Genes Are Required?

Studies aiming to define the minimal gene inventory of a cellular organism received new focus with the publication of the complete sequence of *M. genitalium*, the smallest sequenced bacterial genome (Maniloff, 1996). As mentioned, the set of universally distributed genes is small and insufficient for independent cellular growth and replication, implying that small genome organisms accomplish the same set of cellular processes by retaining different sets of genes. This is achieved in part through nonorthologous gene displacement: the role of one gene is replaced by an unrelated gene accomplishing the same function (Koonin, 2000). Redundancy within the ancestral, large genome appears to be eliminated through different routes. The final gene set may depend on the gene content of chromosomal deletions that occur early in the course of genome reduction.

Even the tiny genome of *M. genitalium* harbors many genes that are dispensable, at least for growth in vitro. Based on a study in which single genes of *M. genitalium* were inactivated using transposon-mediated mutagenesis, at least 129 of that organism's 484 ORFs were unnecessary for growth. Thus, a substantially smaller genome is plausible. It must be remembered, however, that the bare minimum is not necessarily what we expect in naturally occurring organisms, in which selection will eliminate less competitive genotypes.

Every sequenced genome contains a set of ORFs with no assigned function, with the proportion varying among genomes according to the phylogenetic proximity to a laboratory model organism. One kind of insight to be gained from small genome sequences is the identification of unassigned genes that are good targets for further study. For example, almost every *Buchnera* gene has a clear ortholog in *E. coli*, indicating that *Buchnera* provides a good approximation of a minimal *E. coli* genome (in the context of an intracellular environment). For about 50 of these shared genes, no function is assigned. As more small genomes in the γ -proteobacteria are completed, the set of genes indispensable within this group will be defined; these can focus experimental studies directed at extending knowledge of gene function in this model clade of organisms.

Selective Advantage for Smallness?

A frequently proposed explanation for genome reduction is that selection has favored small genome size for the sake of growth efficiency or competitiveness within the host. The implication is that selection for efficient replication would suffice to eliminate DNA corresponding to some genes. But changes in DNA content, on the scale corresponding to individual genes, have not been shown to affect rate of bacterial cell replication. Also, genome sequence analyses contradict some predictions of the hypothesis that selection drives elimination of DNA. One line of evidence against significant selection for small genome size is the retention of nonfunctional DNA in the form of pseudogenes within small genomes, as in *Rickettsia* and *Buchnera*. If genome reduction were driven by a replication advantage of a minimal genome, gene inactivation would not be expected to precede loss of DNA. Finally, small genomes are no more tightly packed, based on overall length of spacer regions or on comparison of homologous chromosomal regions between the tiny *Buchnera* genome and the much larger genome of the related *E. coli* (Mira et al., 2001).

Mutation to Smallness (Deletional Bias)

In contrast to eukaryotes, in which nonfunctional DNA often persists, bacterial genomes are tightly packed with genes, implying elimination of nonfunctional DNA. Analyses of pseudogene sequences in *Rickettsia* and a wide range of other bacteria reveals widespread mutational bias toward DNA loss (deletional bias) (Andersson and Andersson, 2001; Mira et al., 2001). Thus, DNA is retained in the bacterial genome only if selection is acting effectively to preserve it. If gene functions are rendered useless, due to redundancy within the host environment, then mutations will inactivate the sequences, and the corresponding DNA will be eroded over time through mutational patterns favoring deletion.

Although small genomes present a clear progression

from fragmentation or disruption of an ORF to elimination of the corresponding DNA, it is not yet obvious when inactivated genes cease to be transcribed and translated. In the genome of *Rickettsia conorii*, numerous interrupted ORFs were found to be (at least) transcribed (Ogata et al., 2001). Transcription of nonfunctional or unneeded genes might impose a selective cost, possibly favoring elimination of the corresponding DNA. Thus, even if nontranscribed, inactivated genes impose no selective cost; deletional bias may confer a benefit through elimination of DNA that is transcribed but not useful.

Mutational Pressure for A+T Enrichment

Many obligate pathogenic bacteria display A/T-enriched sequences that reflect mutational bias. In each A/T-biased small genome, the bias toward A/T is evident in all types of genes and positions but is strongest in neutral DNA positions, such as noncoding spacers and third codon positions of ORFs. However, the tendency to greater genomic A+T content affects even those nucleotides that effect amino acid replacements. As a result, polypeptides of small genome bacteria are enriched in amino acids, such as lysine, that contain more A or T in the codon family. One consequence is higher predicted isoelectric point (pI) of polypeptides; for example, the average pIs for polypeptides of *Buchnera* and of its relative *E. coli* are 9.6 and 7.2, respectively (Shigenobu et al., 2000). Among currently sequenced genomes, the most extreme A/T bias occurs in *Ureaplasma urealyticum* (25.5% GC), in which the most A/T biased ORFs correspond to genes shown to be expendable in gene knockout studies on the related *M. genitalium*. Thus, the A+T content of a particular set of nucleotide positions reflects the opposing pressures of mutational bias toward A+T versus purifying selection for preservation of gene function.

The basis of mutational pressure toward increased A+T may reflect the elimination of genes encoding DNA repair enzymes, or the decreased efficiency of these enzymes. In particular, the incorporation of uracil into DNA, due either to replication error or to C \rightarrow U deamination, will result in mutational pressure toward A+T if not prevented or corrected, and the enzymes mediating these changes are sometimes missing or less efficient (Glass et al., 2000). Another possible explanation for the A/T mutational bias might involve nucleotide pools favoring A or T. However, this is unlikely to provide a general reason for the pattern as small genomes vary both in capability for nucleotide biosynthesis and in location within hosts.

Reconstructing Genome Reduction

In most cases, the small eubacterial genomes are only distantly related to any larger genome organisms, a situation that precludes a reliable reconstruction of how the genome reduction occurred. For example, *Rickettsia* species and relatives such as *Wolbachia pipientis* and *Ehrlichia* species comprise an α -proteobacterial clade characterized by consistently small genomes, and this clade is only distantly related to species with larger genomes. Likewise, the mycoplasmas and chlamydiae are embedded in ancient groups with uniformly small genome size. However, some of the symbiotic bacteria of insects and other arthropods fall quite closely related to larger genome species within the Enterobac-

teriaceae such as *E. coli*, *Yersinia pestis*, and *Salmonella* species. These symbiotic bacteria, which include *Buchnera* in aphids and *Wigglesworthia* in tsetse flies, provide the opportunity to reconstruct the process of genome reduction. Such an attempt to reconstruct the pattern of gene deletions during the evolution of *Buchnera* suggested that, in addition to gradual erosion of some individual genes through small deletions, some deletions were large and spanned dozens of ancestral genes (Moran and Mira, 2001). One plausible scenario is that the initial transition to the obligately symbiotic (or pathogenic) lifestyle is accompanied by massive genomic changes with some large deletions being fixed within the lineage. These deletions might establish themselves due to a combination of reduced competition and selection in the newly invaded niche and of increased genetic drift arising from population bottlenecks that occur at the time of infection (with many hosts invaded by a single genotype). It is clear as well that some *Buchnera* genes were lost individually through a process of inactivation followed by decay (Moran and Mira, 2001, Silva et al., 2001), as documented in *Rickettsia* (Andersson and Andersson, 2001).

A snapshot of genome degradation in progress is provided by the complete sequence of the genome of *Mycobacterium leprae*, the infectious agent for leprosy. *M. leprae* stands out among the mycobacteria in having a genome that is both reduced and rearranged (Cole et al., 2001). Comparing *M. leprae* to its more typical relative, *Mycobacterium tuberculosis*, indicates that the *M. leprae* lineage has discarded more than 2000 genes. DNA corresponding to more than 1000 of these genes is still present as partial copies or as nonfunctional pseudogenes. Proteome analyses confirm that *M. leprae* does express a much reduced complement of proteins as compared to *M. tuberculosis*, indicating that the apparent pseudogenes have indeed been silenced. This organism has the largest proportion of noncoding DNA of any fully sequenced bacterial genome; only about half of its sequence encodes proteins, as contrasted with 90% in *M. tuberculosis*. In addition, *M. leprae* exhibits some very large deletions that span multiple loci, and its genome size is considerably reduced (3.3 mb as compared to 4.4 mb in other mycobacteria). It has also undergone a shift in base composition toward lower G+C%. The most plausible interpretation of this unusual bacterial genome is that the reductive evolution of *M. leprae* is recent, perhaps linked to its becoming an obligate pathogen during the last few million years.

Lineages that have acquired pathogenic lifestyles recently appear to have embarked on some of the same processes of gene decay and deletion that have progressed to extensive genome shrinkage in more ancient pathogenic groups. For example, large numbers of pseudogenes have been identified in both *Yersinia pestis*, the agent for plague, and *Salmonella enterica* serovar Typhi (Parkhill et al., 2001). The acquisition of a pathogenic life cycle may impose a relatively constant environment, rendering many genes useless, as well as population bottlenecks, resulting in greater levels of genetic drift and resulting gene inactivation. Support for the latter comes from recent studies of the population genetics of human pathogens *Y. pestis* and *M. tuberculosis*, which show very low levels of polymorphism and

indicate high levels of genetic drift relative to related free-living bacteria (Achtman et al., 1999).

Outlook

From the extensive documentation of lateral transfer of pathogenicity islands, we know that gene acquisition often enables pathogenic life. Yet the evolutionarily ancient obligate pathogens possess genomes in which gene loss is far more extensive than gene acquisition. Analysis of the varying solutions to genome minimalism, as presented by different small genome organisms, promises to yield information about interdependencies of gene products. Such information is not evident from studies based on single gene knockouts. Understanding of the basis for the observed differences in gene inventories will depend in part on identifying the kinds of DNA deletions that occur at different evolutionary stages of genome reduction, and this will soon be possible, in view of the current rapid rate of publication of new genome sequences.

Selected Reading

- Achtman, M., Zurth, K., Morelli, C., Torrea, G., Guiryoule, A., and Carniel, E. (1999). Proc. Natl. Acad. Sci. USA 96, 14043–14048.
- Andersson, J.O., and Andersson, S.G.E. (2001). Mol. Biol. Evol. 18, 829–839.
- Andersson, S.G.E., and Kurland, C.G. (1998). Trends Microbiol. 6, 263–278.
- Andersson, S.G.E., Zomorodipour, A., Andersson, J.O., Sicheritz-Ponten, T., Alsmark, U.C.M., Podowski, R.M., Naslund, A.K., Eriksson, A.S., Winkler, H.H., and Kurland, C.G. (1998). Nature 396, 133–140.
- Cole, S.T., Eiglmeier, K., Parkhill, J., James, K.D., Thomson, K.D., Wheeler, P.R., Honore, N., Garnier, T., Churcher, C., Harris, D., et al. (2001). Nature 409, 1007–1011.
- Goebel, W., and Gross, R. (2001). Trends Microbiol. 9, 267–273.
- Glass, J.L., Lefkowitz, E.J., Glass, J.S., Hlener, C.R., Chen, E.Y., and Cassell, G.H. (2000). Nature 407, 757–762.
- Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B., and Herrmann, R. (1997). Nucleic Acids Res. 25, 701–712.
- 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). Science 286, 2165–2169.
- Koonin, E.V. (2000). Ann. Rev. Genom. Hum. Gen. 1, 99–116.
- Maniloff, J. (1996). Proc. Natl. Acad. Sci. USA 93, 10004–10006.
- Moran, N.A., and Mira, A. (2001). Genome Biol., in press.
- Mira, A., Ochman, H., and Moran, N.A. (2001). Trends Genet. 17, 589–596.
- Moran, N.A., and Wernegreen, J.J. (2000). Trends Ecol. Evol. 15, 321–326.
- Mushegian, A.R., and Koonin, E.V. (1996). Proc. Natl. Acad. Sci. USA 93, 10268–10273.
- 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). Science 293, 2093–2098.
- Parkhill, J., Dougan, G., James, K.D., Thomson, N.R., Pickard, D., Wain, J., Churcher, C., Mungall, K.L., Bentley, S.D., Holden, M.T.G., et al. (2001). Nature 413, 848–852.
- Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., and Ishikawa, H. (2000). Nature 407, 81–86.
- Silva, F.J., Latorre, A., and Moya, A. (2001). Trends Genet. 17, 615–618.
- Woese, C.R. (1987). Microbiol. Rev. 51, 221–271.