Genetica **115:** 1–12, 2002.

Genome deterioration: loss of repeated sequences and accumulation of junk DNA

A. Carolin Frank, Haleh Amiri & Siv G.E. Andersson*
*Department of Molecular Evolution, University of Uppsala, Uppsala, S-751 36 Sweden; *Author for correspondence (Phone: +46-18-4714379; Fax: +46-18-471 64 04; E-mail: Siv.Andersson@ebc.uu.se)*

Key words: gene loss, genome degradation, microbial genomes, repeated sequences

Abstract

A global survey of microbial genomes reveals a correlation between genome size, repeat content and lifestyle. Free-living bacteria have large genomes with a high content of repeated sequences and self-propagating DNA, such as transposons and bacteriophages. In contrast, obligate intracellular bacteria have small genomes with a low content of repeated sequences and no or few genetic parasites. In extreme cases, such as in the 650 kb-genomes of aphid endosymbionts of the genus *Buchnera* all repeated sequences above 200 bp have been eliminated. We speculate that the initial downsizing of the genomes of obligate symbionts and parasites occurred by homologous recombination at repeated genes, leading to the loss of large blocks of DNA as well as to the consumption of repeated sequences. Further sequence elimination in these small genomes seems primarily to result from the accumulation of short deletions within genic sequences. This process may lead to temporary increases in the genomic content of pseudogenes and 'junk' DNA. We discuss causes and long-term consequences of extreme genome size reductions in obligate intracellular bacteria.

Introduction

During the last few years, we have witnessed an explosion in the field of microbial genomics. To date, over 60 microbial genome sequences have been published and another 100 genomes are in different stages of completion (http://www.tigr.org). These represent all three kingdoms and most bacterial and archaeal phyla. To interpret this flood of information from an evolutionary perspective, we have to develop models that describe causes and consequences of genome size variations and integrate these with our knowledge of microbial relationships, that is, rRNA phylogenetic trees (Olsen et al., 1994).

Microbial genomes vary more than tenfold in size, from 580 kb in *Mycoplasma genitalium* to 9.2 Mb in *Myxococcus xanthus* (Casjens, 1998). Even within bacterial subdivisions, the entire size span of genome sizes may be represented. For example, in the α-proteobacterial subdivision genomes vary in size from 1.1 Mb in *Rickettsia prowazekii* to 8.7 Mb in *Bradyrhizobium japonicum*. Here, there are also several examples of bacteria with complex genome structures, that is, more than one replicating circle, as observed for example in *Agrobacterium tumefaciens, Rhodobacter capsulatus, Brucella melitensis* (Casjens, 1998) and *Rhizobium* (Flores et al., 2000). The role of repetitive sequences as targets for homologous recombination and rearrangements has recently been demonstrated in *Rhizobium* (Flores et al., 2000).

Genome sizes can vary quite extensively even for closely related strains and species that show low degrees of divergence at the nucleotide sequence level. For example, a recent investigation of different isolates of *Bartonella henselae* revealed a size variation between 1.7 and 2.9 Mb (Maruyama et al., 2001). Likewise, a comprehensive study of natural isolates of *Escherichia coli* shows that these vary in genome size from 4.5 to 5.5 Mb (Bergthorsson & Ochman, 1998). The two completely sequenced *E. coli* genomes exemplify this size variation (Perna et al., 2001). Thus, the genome of strain O157:H7 is substantially larger (5.5 Mb) than the genome of strain K12 (4.6 Mb) and there is a 422 kb inversion around the replication

terminus. The two strains share a linear 'backbone' of 4.1 Mb, and the difference in size is accounted for by numerous strain-specific 'islands', that either represent imported DNA or retention of fragments that have been lost in the other strain (Perna et al., 2001).

In free-living organisms with large population sizes that frequently adapt to environmental changes, some of these differences may be due to the insertion or deletion of 'cassettes' of DNA that confer selective advantages in particular growth environments (Moxon et al., 1994; Ochman, Lawrence & Groisman, 2000). Pathogenicity islands (PAI), plasmids, bacteriophages, transposons and integrons are examples of mobile genomic islands (Hacker & Kaper, 2000; Hacker & Carniel, 2001). The loss or gain of such an island will quickly reach fixation in the population if it is associated with a selective advantage. Indeed, it has been shown that the gain and loss of mobile virulence elements have occurred several times in parallel in separate lineages of pathogenic *E. coli* (Reid et al., 2001). Mechanistically, this flexibility in structure is often caused by recombination at repetitive elements that flank the mobile element.

In contrast, the loss or gain of 'core' genes that code for proteins with functions in basic cellular functions (Hacker & Carniel, 2001) is likely to represent a (slightly) deleterious event under most environmental conditions. Consequently, we expect cells with nucleotide changes and/or deletions in these genes to be out-competed by more rapidly growing cells with no mutations. In effect, rearrangements and indels in the core genes will only rarely reach fixation in large, free-living microbial populations, irrespective of the selective demands imposed by the different growth habitats.

Obligate intracellular bacteria, however, are fundamentally different from free-living bacteria in that they form populations of small sizes that frequently undergo bottlenecks (Andersson & Kurland, 1998). The genomes of these organisms have evolved primarily by reductive evolutionary processes, during which even genes that are part of the highly conserved genomic core have already been eliminated or are currently in the process of being eliminated (Andersson & Andersson, 1999a). For example, more than 1000 pseudogenes have been identified in *M. leprae* (Cole et al., 2001). Likewise, *R. prowazekii* seem to have at least a few hundred genes in different stages of degradation (Andersson et al., 1998; Ogata et al., 2001). Thus, as much as 25–50% of the genome of an obligate intracellular parasite may represent fairly recent gene inactivation events (Andersson et al., 1998; Cole et al., 2001).

Studies of the rates and patterns of changes in bacterial pseudogenes have been done most extensively in *Rickettsia*. A detailed comparative analysis of the *metK* gene coding for S-adenosylmethionine synthetase (Andersson & Andersson, 1999a) and other pseudogenes (Andersson & Andersson, 2001) has shown short deletions predominate over short insertions in sequences that evolve by neutral processes in *Rickettsia* (Andersson & Andersson, 1999b, 2001). No signs of preexisting open reading frames could be identified in the noncoding DNA of *R. prowazekii*, probably due to extensive accumulation of frameshift mutations (Andersson et al., 1998). However, sequence similarity between a gene in one species and a stretch of noncoding DNA in a closely related species has confirmed that the 'junk' DNA represents highly degraded remnants of ancestral gene sequences (Andersson & Andersson, 2001; Ogata et al., 2001).

It may seem surprising that most other microbial genomes have so little junk DNA despite high rates of DNA influx in the form of genetic parasites, such as plamids, transposable elements and prophages (Andersson & Andersson, 1999b; Lawrence et al., 2001). To explain this paradox, it has been suggested that there is a selection for high deletion rates in the genomes of free-living bacteria to compensate for the attacks of invading phage DNA and other pieces of selfish DNA (Lawrence, 2001). Since obligate intracellular bacteria are no longer threatened by parasitic genetic elements, it is thought that they can tolerate a lower deletion rate, which would result in a gradual accumulation of pseudogenes (Lawrence et al., 2001).

Alternatively, the molecular mechanisms whereby pseudogenes are eliminated in obligate intracellular bacteria may be very different from the mechanisms whereby genetic parasites are eliminated from the genomes of free-living bacteria. Indeed, prophages and transposons are often distinguished from other types of DNA by being flanked by moderately long repeated sequences that serve as targets for the recombination/deletion machinery. No such repeats are flanking the identified pseudogenes in *Rickettsia* (Andersson et al., 1998; Andersson & Andersson, 1999a, 2001).

To understand the role of repeated sequences in promoting genome instability and deterioration, we need to quantify the distribution of repeated genes in microbes with different genome sizes and life styles. In this paper, we present an analysis of repeat densities in microbial genomes and we discuss the evolutionary

forces that drive genome degradation, with a particular focus on the loss of repeated genes in obligate intracellular bacteria.

Repeat density in microbial genomes

Despite their compactness, many prokaryotic genomes contain a variety of repetitive sequences, either as recently expanded gene families, or as arrays of non-coding repeated sequences or as transposable elements and insertion sequences. A number of studies on repeated sequences have focused on the occurrence and functionality of short repetitive elements, as for instance in *E. coli* (Bachellier et al., 1999). Repeat length, divergence and location are all factors that influence the rate of ectopic recombination (Bachellier et al., 1999). As pointed out earlier by Rocha, Danchin and Viari (1999a, b), large, identical repeats are particularly important because of their potential role in genome instability and evolution.

A graphical representation of repeated sequences longer than 200 bp gives a visual impression of how the content of large repeats vary in 37 completely sequenced microbial genomes (Figure 1). To enable a quantitative comparison, we have plotted genome size as a function of repeat density (defined as the number of nucleotides involved in a repetitive DNA segment per total number of nucleotides in the genome) (Figure 2). On the average, the density of identical repeats that are larger than 200 bp is 1.7% per genome, not including extra-chromosomal elements (Figure 2). Thus, if the correlation between genome size and repeat content is perfect, all microbial genomes are expected to have a repeat density close to 1.7%. However, the variation in the relative fraction of repeats seems to be larger than what would be expected solely by the genome size differences (Figure 2). For example, *Neisseria meningitidis* has a repeat density of 6.2% and the symbiotic plasmid of *Rhizobium* (pNGR) of 6.8%, which is considerably higher than expected from their genome sizes (Figure 2). *N. meningitidis* is known to contain a large number of repeats involved in recombination-mediated variation (Parkhill et al., 2000b) and an unusually high frequency of insertion sequences (Tettelin et al., 2000; Parkhill et al., 2000b).

On the other extreme, we find that there is a significantly lower density of repeated sequences in obligate intracellular bacteria with genomes in the 1 Mb range (repeat density is <0.6%) than in other bacteria ($p < 0.026$, Wicoxon signed rank test). For example, *R. prowazekii* contains only four repeated sequences longer than 200 bp and these accounts for only 0.23% of the genome (Andersson et al., 1998). Likewise, repeat densities in *R. conorii* (Ogata et al., 2001), *Chlamydia pneumoniae* and *Chlamydia muramidum* are lower than 0.6% (Figure 2). The most extreme case is the 650 kb genome of *Buchnera aphidicola* with no repeats longer than 200 bp (Shigenobou et al., 2000). We may wonder: Why and how have these repeated sequences been lost?

Loss of duplicated genes in obligate intracellular bacteria

Since the loss of redundant genes may not necessarily be a lethal event, we expect genes that are present in more than copy to be the first targets for sequence elimination during adaptations to intracellular growth environments. Such genes include those coding for ribosomal RNAs (*rrs, rrl* and *rrf*) and elongation factor Tu (*tuf*) (Ohama et al., 1987; Buttarelli et al., 1989; Meng, Shinozaki & Sugira, 1989; Tiboni et al., 1991; Vijgenboom et al., 1994). For example, *E. coli* has seven rRNA operons, *Vibrio cholera* has eight and *Bacillus subtilis* has as many as 10 rRNA operons. Since each rRNA operon is about 5 kb in size, as much 25–50 kb of the repeated sequences in these organisms may be accounted for by rRNA genes. A direct comparison of rRNA and total repeat contents show that the rRNA genes represent 40–60% of the repeated sequences in *E. coli* and *V. cholera* and more than 90% in *B. subtilis*. However, in many other organisms there is no such strong correlation between rRNA gene copy numbers and repeat densities. For example, *N. meningitidis*, has the highest density of repeated sequences, but its four rRNA operons represent only 13% of the total fraction of repeated sequences.

Multiple rRNA operons are found in bacteria with large genomes; whereas only one or two rRNA operons are present in genomes that are less than 2 Mb in size. For example, only one copy of the rRNA and the *tuf* genes are present in *R. prowazekii* (Andersson et al., 1995; Syvänen 1996; Andersson et al., 1998) and in other obligate intracellular bacteria (Stephen et al., 1998; Shigenobou et al., 2000). A comparative study of several *Rickettsia* species has shown that the loss of duplicate copies of the rRNA and the *tuf* genes occurred already prior to the divergence of species within the genus *Rickettsia* (Andersson et al., 1998; Amiri, Alsmark & Andersson, 2002).

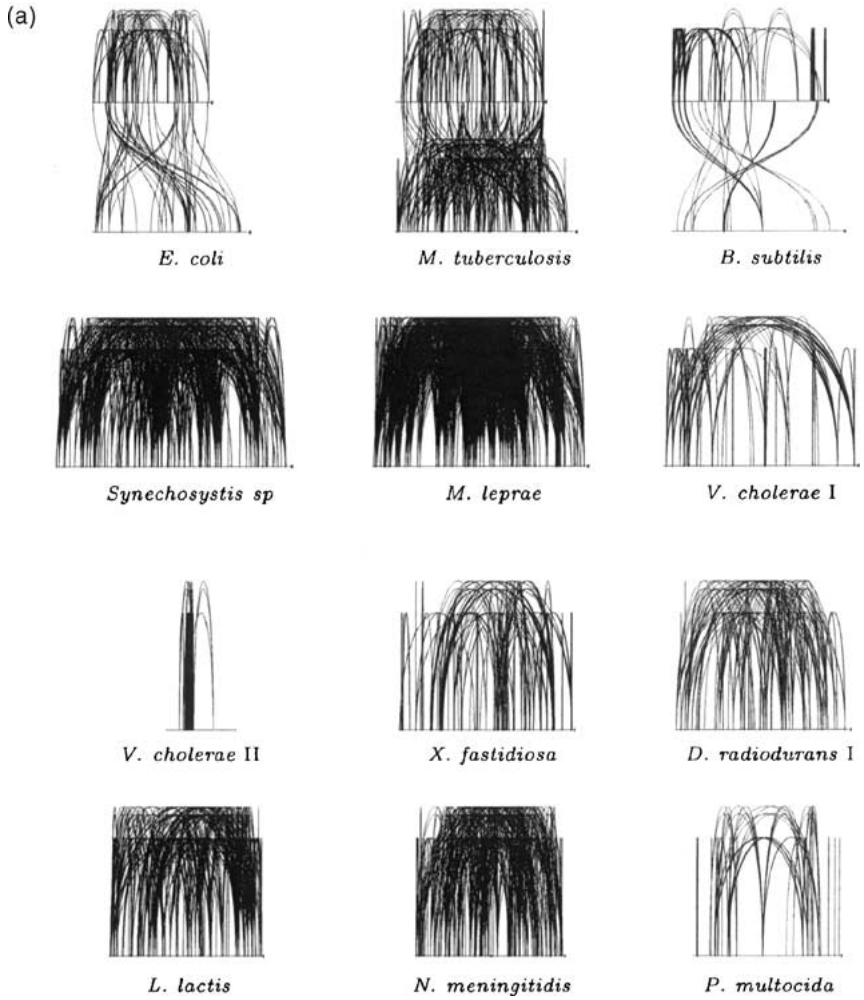

Figure 1. Content of repeated sequences larger than 200 bp in 37 microbial genomes. The images were produced using the program miropeats (Parsons, 1995) (http://www.ebi.ac.uk/~jparsons/packages/miropeats/miropeats.html), which detects regions of sequence similarity in a sequences and then presents this similarity information graphically. On the vertical line, which represents the genome, an arc is drawn between any two repeated regions. Genomes larger than 4 Mb in size are divided in two halves. (a) Repeat content in genomes of sizes 4.6 to 2.2 Mb. Genome sequence data were taken from Blattner et al. (1997) (U00096); Cole et al. (1998) (AE000666); Kunst et al. (1997) (AL009126); Kaneko et al. (1996) (AB001339); Cole et al. (2001) (AL450380); Heidelberg et al. (2000) (AE003852,AE003853); Simpson et al. (2000) (AE003849); White et al. (1999) (AE000513); Bolotin et al. (2001) (AE005176); Parkhill et al. (2000b) (AE002098); May et al. (2001) (AE004439). (b) Repeat content in genomes of size 2.1 to 1.1 Mb. Genome sequence data were taken from Klenk et al. (1997) (AE000782); Ng et al. (2000) (AE004437); Nelson et al. (1999) (AE000512); Ferretti et al. (2001) (AE004092); Fleishmann et al. (1995) (L42023); Kawarabayasi et al. (1998) (BA000001); Smith et al. (1997) (AE000666); Kawarabayasi et al. (1999) (BA00000); Tomb et al. (1997) (AE000511); Bult et al. (1996) (L77117); Parkhill et al. (2000a) (AL111168); Kawashima et al. (2000) (BA000011); Ruepp et al. (2000) (AL139299); Deckert et al. (1998) (AE000657); Kalman et al. (1998) (AE001363). (c) Repeat content in genomes of size 1.1 to 0.5 Mb. Genome sequence data were taken from Fraser et al. (1998) (AE000520); Ogota et al. (2001) (AE006914) Andersson et al. (1998) (AJ235269); Read et al. (2000) (AE002160); Stephens et al. (1998) (AE001273); Fraser et al. (1997) (AE000783); Himmelreich et al. (1996) (NC_000912); Glass et al. (2000) (AF222894); Fraser et al. (1995) (L43967).

Duplicated genes like the rRNA and the *tuf* genes are not only targets for sequence elimination but also for genome rearrangements (Anderson & Roth, 1977; Petes & Hill, 1988; Liu & Sanderson, 1996). For example, the 16S and 23S rRNA genes in *R. prowazekii* are not located in the normal rRNA operon structure (16S-23S-5S) (Andersson et al., 1995). Furthermore, it has been suggested that the atypical organization of the single *tuf* gene in *Rickettsia* is the result of an intra-chromosomal recombination event at the two ancestral *tuf* genes that were located in inverse orientation, leading to an inversion of the intervening

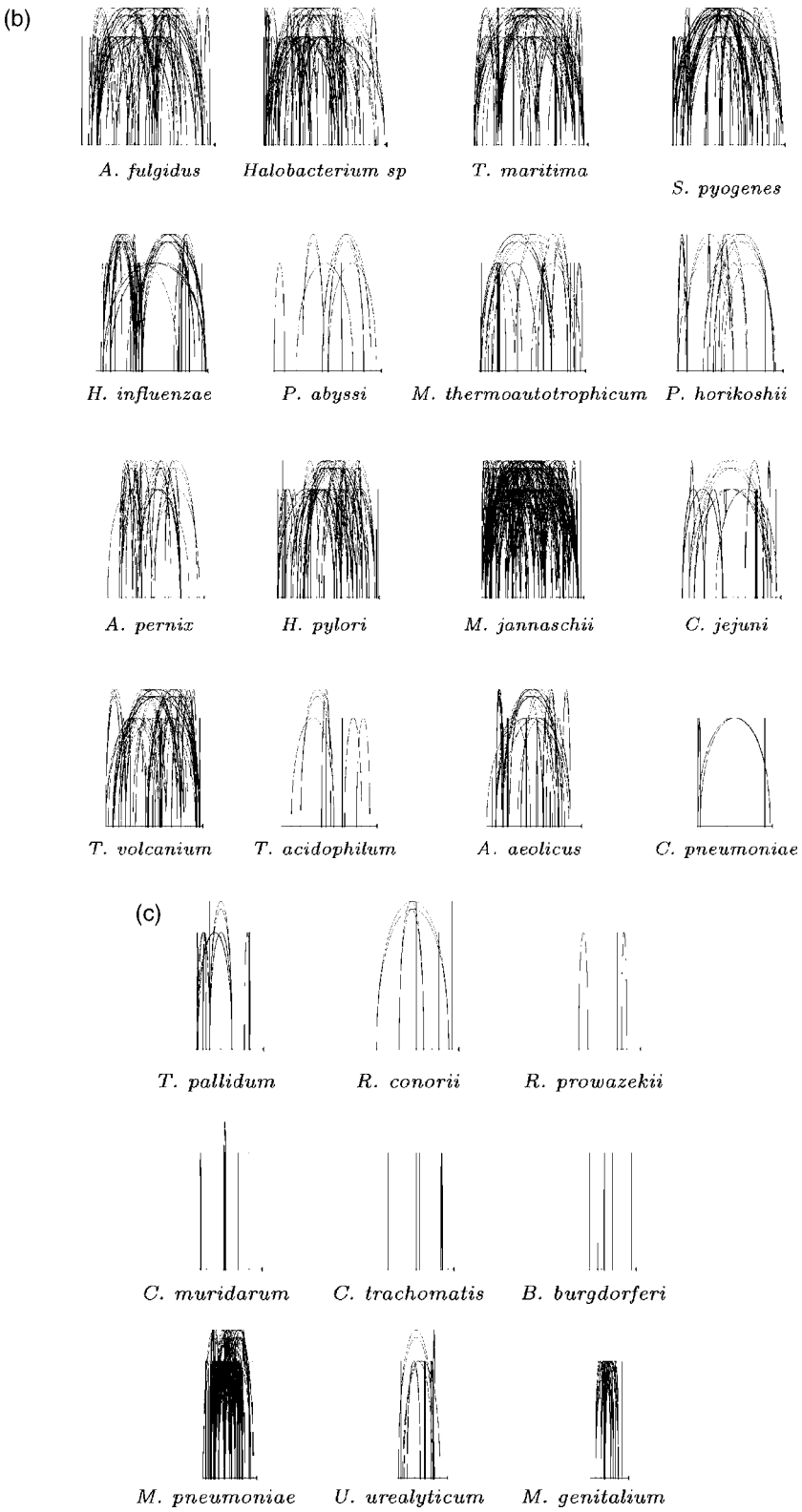

Figure 1. Continued.

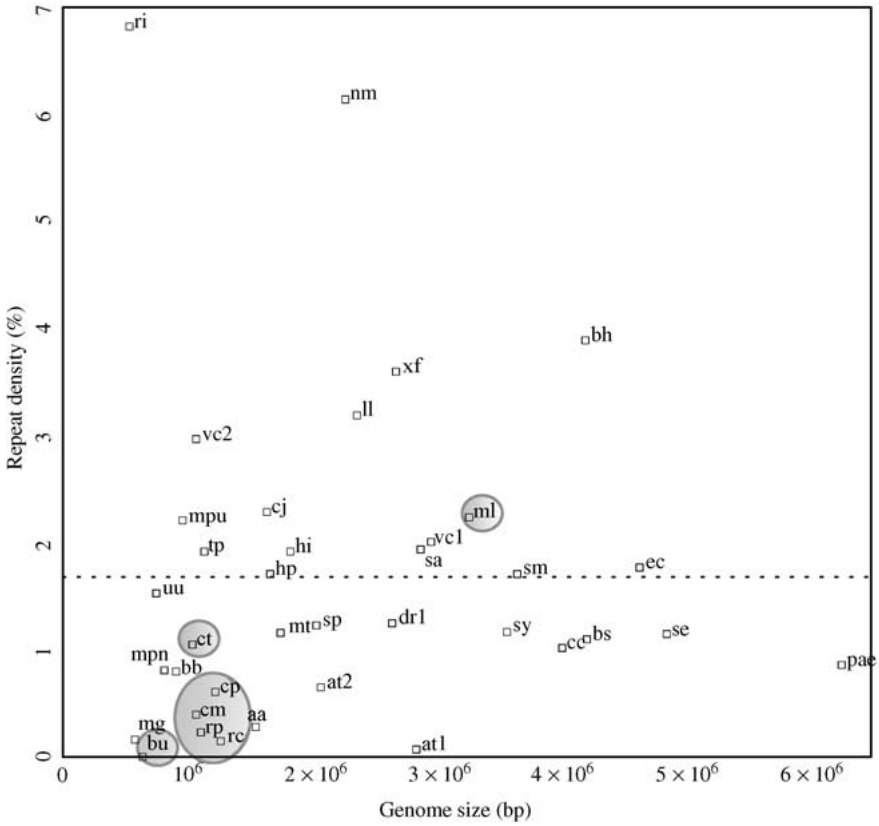

Figure 2. Comparisons of repeat density and genome size. Exact repeats larger than 200 bp were detected using the software Reputer (Kurtz & Schleiermacher, 1999) (http://bibiserv.techfak.uni-bielefeld.de/reputer/). The number of basepairs present in one or several repeats larger than 200 bp was calculated for each genome. The repeated fraction of the genome is plotted against genome size and the average repeat density (1.7%) is indicated with a dotted line. Obligate intracellular parasites are shown in shaded colour. Genomic data not present in Figure 1 were taken from: Bolotin et al. (2001) (AE005176), Ogata et al. (2001) (AE006914), McClelland et al. (2001) (AE006468), Takami et al. (2000) (BA000004), Chambaud et al. (2001) (AL445566), Hoskins et al. (2001) (AE007317). Freiberg et al. (1997) (AE000083), Capela et al. (2001) (AL591688). Abbreviations are used as follows: aa (*A. aeolicus*), at1 (*A. tumefaciens* circular chromosome), at2 (*A. tumefaciens* linear chromosome), bb (*B. burgdorferi*), bh (*B. halodurans*), bs (*B. subtilis*), bu (*Buchnera*), cc (*C. crescentus*), cj (*C. jejuni*), cm (*C. muridarum*), cp (*C. pneumoniae*), ct (*C. trachomatis*), dr1 (*D. radiodurans* chromosome1), ec (*E. coli*), hi (*H. influenzae*), hp *(H. pylori*), ll (*L. lactis*), mg (*M. genitalium*), mp (*M. pneumoniae*), mpu (*M. pulmonis*), mt (*M. tubercolosis*), nm (*N. meningitidis*), rc (*R. conorii*), rp (*R. prowazekii*), ri (*Rhizobium etli*, plasmidcpNGR234a), sa (*S. aureus*), se (*S. enterica*), sm (*S. meliloti*), sp (*S. pneumoniae*), sy (*Synechocystis*), tp (*T. pallidum*), uu (*U. urealytica*), vc1 (*V. cholerae* chromosome 1), vc2 (*V. cholerae* chromosome 2), xf (*X. fastidiosa*).

segment (Syvänen et al., 1996). In some species, genome rearrangements at repeated sequences occur at frequencies that are measurable in real time. For example, populations with all possible rearrangements matching those predicted from the location of directly repeated sequences in the symbiotic plasmid of *Rhizobium* has been identified (Flores et al., 2000). In this case, direct-repeated sequences ranging in size from 2.6 to 4.3 kb are flanking amplicons of 36–307 kb (Flores et al., 2000). A total of 14 potential rearrangements generated by homologous recombination at these sites could be identified in the *Rhizobium* population (Flores et al., 2000).

Short repetitive sequences have been identified near to the rRNA genes and the *tuf* genes in several *Rickettsia* species and these are flanking short deletions in other species (Andersson et al., 1995; Amiri, Alsmark & Andersson, 2002). This suggests that homologous recombination at repeated sites do not only cause major genome rearrangements but also contributes to the loss of short sequence segments (Andersson et al., 1995; Amiri, Alsmark & Andersson, 2002). A novel type of repetitive element was recently identified in the *R. conorii* genome, called the Rickettsia Palindromic Element (Ogata et al., 2001). We have shown that this element flourished already

prior to species divergence within the genus *Rickettsia*, and that the absence of this element in some species is most likely due to recent loss rather than to recent proliferation (Amiri, Alsmark & Andersson, unpublished).

All of this suggests that multiple genome configurations may be present in natural microbial populations, in particular in microbes with high degrees of DNA reiteration. However, this diversity in genomic structures may get lost in small populations that undergoes frequent bottlenecks since lost sequences can not easily be obtained by inter-chromosomal recombination in these isolated populations. In effect, repeats that have once been eliminated can normally not be recovered again. In the extreme, even short repeats may get consumed in this process, as observed in the tiny genome of the aphid endosymbiont *B. aphidicola* (Shigenobou et al., 2000).

Discussion

Transitions to intracellular environments have occurred frequently in the evolutionary past. These transitions impose a new set of challenges for the invading bacterium, which has to learn how to enter, survive and exit the new environment. Initially, the bacteria may explore the intracellular habitat as facultative parasites while maintaining their free-living status. Most of these facultative intracellular parasites have a broad host range and must, therefore, have evolved efficient mechanisms to evade attacks from the immune system in each of these hosts. The capacity for rapid variation in cell surface molecules is prevalent among pathogens (Peterson et al., 1995; Himmelreich et al., 1996; Fraser et al., 1997; Cole et al., 1998; Faguy et al., 2000). This is commonly obtained through the presence of repetitive gene families that encode variants of outer membrane proteins. These copies can be differentially expressed or their frequency can be altered through conversion and recombination. In these bacteria, a dynamic and modular genomic structure has enabled rapid responses to new challenges in the growth environments (Romero & Palacios, 1997). Changes between growth habitats allow bacteriophages, transposons and plasmids to flourish which contribute to some of the repetitive DNA in these genomes. Thus, high repeat contents, flexible genomes and rapid cycles of expansion and degradation are to be expected for free-living and facultative intracellular parasites (Figure 3).

When the relationship with the host becomes obligate the bacterium is restricted to growth in a few selected host cells that it has to be transmitted between without intervening stages of free living growth. Insects of various kinds often serve as hosts or vectors of transmission between hosts. Since bacteria growing under these conditions are isolated and do not come in contact with many of the mobile genetic elements, their small genomes lack much of the repetitive DNA found in the free-living organisms. Small population sizes, bottlenecks and integration with the host metabolic system may lead to additional genomic changes, such as the loss of large blocks of genes by homologous recombination at repeated sites (Andersson & Kurland, 1998). However, since the repeated sequences will gradually be consumed in this process, the possibilities for further recombinatorial events are reduced. At the very end, even genes involved in the recombination machinery may be destroyed by mutations, making any further deletions by homologous recombination impossible. Low repeat contents and static genomes are, therefore, to be expected for bacteria that have evolved very intimate host-microbe relationships (Figure 3).

Throughout this process, we expect substitutions and deletion mutations to accumulate in single-copy genes. Some of these may be neutral or even advantageous for growth in the intracellular environment. For example, the cytoplasm of an eukaryotic cell is nutritionally very rich which makes the biosynthesis of small molecules such as carbohydrates, amino acids and nucleotide monophosphates nonessential. Not surprisingly, obligate intracellular parasites have no or very few genes coding for the biosynthesis of small metabolites (Andersson et al., 1998; Stephens et al., 1998).

However, in small populations that undergo frequent bottlenecks, weakly deleterious mutations may also be fixed in the population (Andersson & Kurland, 1998). This means that not only the redundant gene function may be eliminated, but also genes in the essential genomic ‘core’ may be inactivated by mutations. These genes are normally not contained within cassette-like structures, and even if they would, the flanking repeats may already have been eliminated from the genome. Thus, the removal of DNA sequences subsequent to the loss of repeated genes is a very tedious process that occurs by numerous small deletion events, each of which only affects a few nucleotides. This slow process leads to a temporary accumulation of pseudogenes (Figure 3), in striking

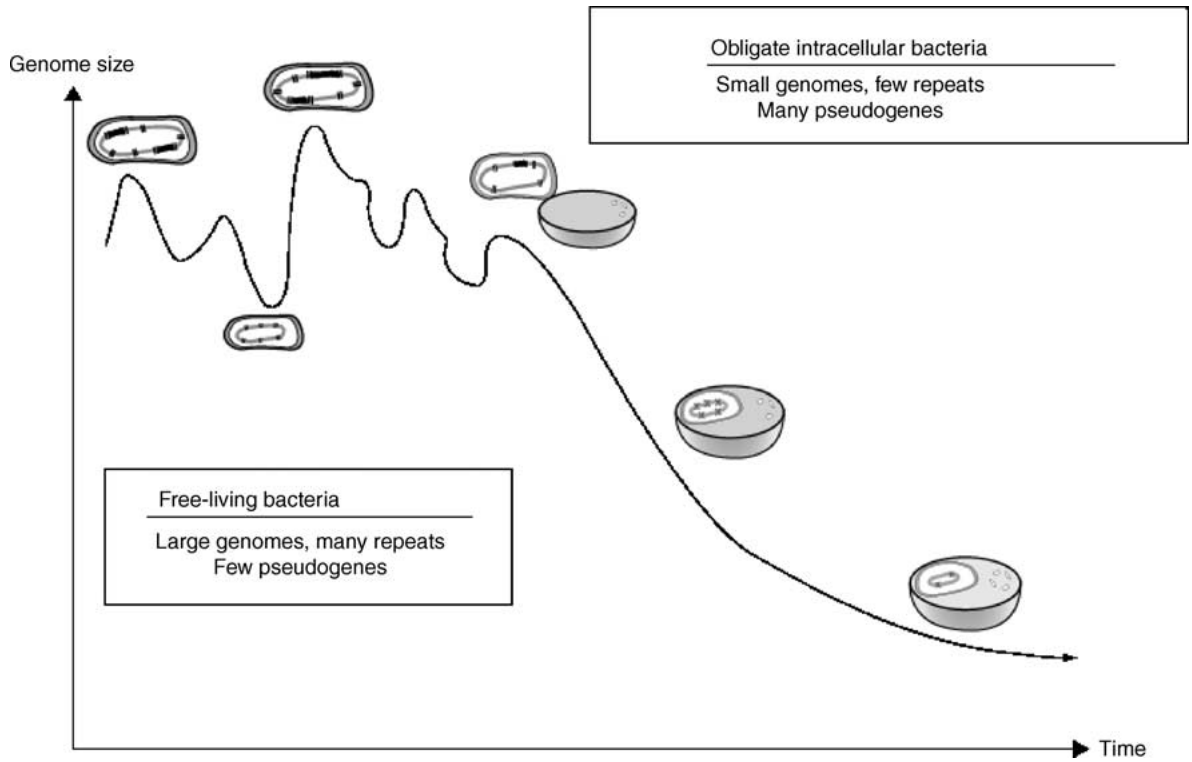

Figure 3. Schematic illustration of genome size variations as a function of time during transitions to intracellular growth habitats. Filled boxes represent mobile genetic elements. Genomes of obligate intracellular bacteria are smaller and have a lower content of repeated sequences (//) and a higher content of pseudogenes (x) than genomes of free-living bacteria and facultative intracellular parasites.

contrast to the rapid and clean excision of sequences that are flanked by repeated sequences.

The reductive evolutionary processes that we observe in the genomes of modern obligate intracellular parasites and symbionts may cast some light on the way in which the mitochondrial genomes were once reduced in size (Andersson & Kurland, 1999; Kurland & Andersson, 2000). The mitochondrion is thought to have originated from an endosymbiotic bacterium that evolved by a massive elimination of genes, some of which genes may have been transferred from the endosymbiont to the nuclear genome (Andersson & Kurland, 1999; Karlberg et al., 2000; Kurland & Andersson, 2000). Studies of the evolution of obligate intracellular parasites and symbionts have already provided extensive information about the first of these processes, that is, genome degradation. Whether we can also identify examples of the other two processes, that is, downloading of bacterial genes to the host nuclear genome and/or recruitment of host nuclear genes for service in the parasite/symbionts remain to be determined.

Acknowledgements

The authors work is supported by the National Science Research Council and the Foundation for Strategic Research.

References

Amiri, H., U.C.M. Alsmark & S.G.E. Andersson, 2002. Proliferation and deterioration of Rickettsia Palindromic Elements. Mol. Biol. Evol. In press.

Andersson, J.O. & S.G.E. Andersson, 2001. Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. Mol. Biol. Evol. 18: 829–839.

Andersson, J.O. & S.G.E. Andersson, 1999a. Insights into the evolutionary process of genome degradation. Curr. Opin. Genet. Dev. 9: 664–671.

Andersson, J.O. & S.G.E. Andersson, 1999b. Genome degradation is an ongoing process in *Rickettsia*. Mol. Biol. Evol. 16: 1178–1191.

Andersson, S.G.E., A. Zomorodipour, J.O. Andersson, T. Sicheritz-Ponten, U.C. Alsmark, R.M. Podowski, A.K. Naslund, A.S. Eriksson, H.H. Winkler & C.G. Kurland, 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396: 133–140.

Andersson, S.G.E. & C.G. Kurland, 1998. Reductive evolution of resident genomes. Trends Microbiol. 6: 263–268.

Andersson, S.G.E. & C.G. Kurland, 1999. Origins of mitochondria and hydrogenosomes. Curr. Op. Microbiol. 2: 535–541.

Andersson, S.G.E., A. Zomorodipour, H.H. Winkler & C.G. Kurland, 1995b. Unusual organization of the rRNA genes in *Rickettsia prowazekii*. J. Bacteriol. 177: 4171–4175.

Bachellier, S., E. Gilson, M. Hofnung & C.W. Hill, 1999. Repeated sequences, pp. in *Escherichia coli* and *Salmonella*; Cellular and Molecular Biology, edited by F.C. Neidhart. ASM Press, Washington DC.

Bergthorsson, U. & H. Ochman, 1998. Distribution of chromosome length variation in natural isolates of *Escherichia coli.* Mol. Biol. Evol. 15: 6–16.

Blattner, F.R., G. Plunkett, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew, J. Gregor, N.W. Davis, H.A. Kirkpatrick, M.A. Goeden, D.J. Rose, B. Mau & Y. Shao, 1997. The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453–1474.

Bolotin, A., P. Wincker, S. Mauger, O. Jaillon, K. Malarme, J. Weissenbach, S.D. Erlich & A. Sorokin, 2001. The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. lactis IL1403. Genome. Res. 11: 731–753.

Bult, C.J., O. White, G.J. Olsen, L. Zhou, R.D. Fleischmann, G.G. Sutton, J.A. Blake, L.M. FitzGerald, R.A. Clayton, J.D. Gocayne, A.R. Kerlavage, B.A. Dougherty, J.F. Tomb, M.D. Adams, C.I. Reich, R. Overbeek, E.F. Kirkness, K.G. Weinstock, J.M. Merrick, A. Glodek, J.L. Scott, N.S. Geoghagen & J.C. Venter, 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 273: 1058–1073.

Buttarelli, F.R., R.A. Calogero, O. Tiboni, C.O. Gualerzi & C.L. Pon, 1989. Characterization of the *str* operon genes from *Spirulina platensis* and their evolutionary relationship to those of other prokaryotes. Mol. Gen. Genet. 217: 97–104.

Capela, D., F. Barloy-Hubler, J. Gouzy, G. Bothe, F. Ampe, J. Batut, P. Boistard, A. Becker, M. Boutry, E. Cadieu, S. Dreano, S. Gloux, T. Godrie, A. Goffeau, D. Kahn, E. Kiss, V. Lelaure, D. Masuy, T. Pohl, D. Portetelle, A. Puhler, B. Purnelle, U. Ramsperger, C. Renard, P. Thebault, M. Vandenbol, S. Weidner & F. Galibert, 2001. Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021. Proc. Natl. Acad. Sci. USA 98: 9877–9882.

Casjens, S., 1998. The diverse and dynamic structure of bacterial genomes. Annu. Rev. Genet. 32: 339–377.

Chambaud, I., R. Heilig, S. Ferris, V. Barbe, D. Samson, F. Galisson, I. Moszer, K. Dybvig, H. Wroblewski, A. Viari, E.P. Rocha & A. Blanchard, 2001. The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. Nucl. Acids Res. 29: 2145–2153.

Cole, S.T., K. Eiglmeier, J. Parkhill, K.D. James, N.R. Thomson, P.R. Wheeler, N. Honore, T. Garnier, C. Churcher, D. Harris, K. Mungall, D. Basham, D. Brown, T. Chillingworth, R. Connor, R.M. Davies, K. Devlin, S. Duthoy, T. Feltwell, A. Fraser, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, C. Lacroix, J. Maclean, S. Moule, L. Murphy, K. Oliver, M.A. Quail, M.A. Rajandream, K.M. Rutherford, S. Rutter, K. Seeger, S. Simon, M. Simmonds, J. Skelton, R. Squares, S. Squares, K. Stevens, K. Taylor, S. Whitehead, J.R. Woodward & B.G. Barrell, 2001. Massive gene decay in the leprosy bacillus. Nature 409: 1007–1011.

Cole, S.T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S.V. Gordon, K. Eiglmeier, S. Gas, C.E. Barry, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels & B.G. Barrell, 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393: 537–544.

Deckert, G., P.V. Warren, T. Gaasterland, W.G. Young, A.L. Lenox, D.E. Graham, R. Overbeek, M.A. Snead, M. Keller, M. Aujay, R. Huber, R.A. Feldman, J.M. Short, G.J. Olsen & R.V. Swanson, 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature 392: 353–358.

Faguy, D.M., 2000. The controlled chaos of shifty pathogens. Curr. Biol. 10: R498–501.

Ferretti, J.J., W.M. McShan, D. Ajdic, D.J. Savic, G. Savic, K. Lyon, C. Primeaux, S. Sezate, A.N. Suvorov, S. Kenton, H.S. Lai, S.P. Lin, Y. Qian, H.G. Jia, F.Z. Najar, Q. Ren, H. Zhu, L. Song, J. White, X. Yuan, S.W. Clifton, B.A. Roe & R. McLaughlin, 2001. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. Proc. Natl. Acad. Sci. USA 98: 4658–4663.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty & J.M. Merrick, 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269: 496–512.

Flores, M., P. Mavingui, X. Perret, W.J. Broughton, D. Romero, G. Hernandez, G. Davila & R. Palacios, 2000. Prediction, identification, and artificial selection of DNA rearrangements in *Rhizobium*: towards a natural genomic design. Proc. Natl. Acad. Sci. USA 97: 9138–9143.

Fraser, C.M., S.J. Norris, G.M. Weinstock, O. White, G.G. Sutton, R. Dodson, M. Gwinn, E.K. Hickey, R. Clayton, K.A. Ketchum, E. Sodergren, J.M. Hardham, M.P. McLeod, S. Salzberg, J. Peterson, H. Khalak, D. Richardson, J.K. Howell, M. Chidambaram, T. Utterback, L. McDonald, P. Artiach, C. Bowman, M.D. Cotton & J.C. Venter, 1998. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science 281: 375–388.

Fraser, C.M., S. Casjens, W.M. Huang, G.G. Sutton, R. Clayton, R. Lathigra, O. White, K.A. Ketchum, R. Dodson, E.K. Hickey, M. Gwinn, B. Dougherty, J.F. Tomb, R.D. Fleischmann, D. Richardson, J. Peterson, A.R. Kerlavage, J. Quackenbush, S. Salzberg, M. Hanson, R. van Vugt, N. Palmer, M.D. Adams, J. Gocayne & J.C. Venter, 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 390: 580–586.

Fraser, C.M., J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann, C.J. Bult, A.R. Kerlavage, G. Sutton & J.M. Kelley, 1995. The minimal gene complement of *Mycoplasma genitalium*. Science 270: 397–403.

Freiberg, C., R. Fellay, A. Bairoch, W.J. Broughton, A. Rosenthal & X. Perret, 1997. Molecular basis of symbiosis between *Rhizobium* and legumes. Nature 387: 394–401.

Glass, J.I., E.J. Lefkowitz, J.S. Glass, C.R. Heiner, E.Y. Chen & G.H. Cassell, 2000. The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. Nature 407: 757–762.

Hacker, J. & J.B. Kaper, 2000. Pathogenicity islands and the evolution of microbes. Annu. Rev. Microbiol. 54: 641–679.

Hacker, J. & E. Carniel, 2001. Ecological fitness, genomic islands and bacterial pathogenicity. EMBO Rep. 2: 376–381.

Heidelberg, J.F., J.A. Eisen, W.C. Nelson, R.A. Clayton, M.L. Gwinn, R.J. Dodson, D.H. Haft, E.K. Hickey, J.D. Peterson, L. Umayam, S.R. Gill, K.E. Nelson, T.D. Read, H. Tettelin, D. Richardson, M.D. Ermolaeva, J. Vamathevan, S. Bass, H. Qin, I. Dragoi, P. Sellers, L. McDonald, T. Utterback, R.D. Fleishmann, W.C. Nierman & O. White, 2000. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. Nature 406: 477–483.

Himmelreich, R., H. Hilbert, H. Plagens, E. Pirkl, B.C. Li & R. Herrmann, 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucl. Acids Res. 24: 4420–4449.

Hoskins, J., W.E. Alborn Jr., J. Arnold, L.C. Blaszczak, S. Burgett, B.S. DeHoff, S.T. Estrem, L. Fritz, D.J. Fu, W. Fuller, C. Geringer, R. Gilmour, J.S. Glass, H. Khoja, A.R. Kraft, R.E. Lagace, D.J. LeBlanc, L.N. Lee, E.J. Lefkowitz, J. Lu, P. Matsushima. S.M. McAhren, M. McHenney, K. McLeaster, C.W. Mundy, T.I. Nicas, F.H. Norris, M. O'Gara, R.B. Peery, G.T. Robertson, P. Rockey, P.M. Sun, M.E. Winkler, Y. Yang, M. Young-Bellido, G. Zhao, C.A. Zook, R.H. Baltz, S.R. Jaskunas, P.R. Rosteck Jr., P.L. Skatrud & J.I. Glass, 2001. Genome of the bacterium *Streptococcus pneumoniae* strain R6. J. Bacteriol. 183: 5709–5717.

Kalman, S., W. Mitchell, R. Marathe, C. Lammel, J. Fan, R.W. Hyman, L. Olinger, J. Grimwood, R.W. Davis & R.S. Stephens, 1999. Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. Nat. Genet. 21: 385–389.

Kaneko, T., S. Sato, H. Kotani, A. Tanaka, E. Asamizu, Y. Nakamura, N. Miyajima, M. Hirosawa, M. Sugiura, S. Sasamoto, T. Kimura, T. Hosouchi, A. Matsuno, A. Muraki, N. Nakazaki, K. Naruo, S. Okumura, S. Shimpo, C. Takeuchi, T. Wada, A. Watanabe, M. Yamada, M. Yasuda & S. Tabata, 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire. DNA Res. 3: 109–136.

Karlberg, O., B. Canback, C.G. Kurland & S.G.E. Andersson, 2000. The dual origin of the yeast mitochondrial proteome. Yeast 17: 170–187.

Kawarabayasi, Y., Y. Hino, H. Horikawa, S. Yamazaki, Y. Haikawa, K. Jin-no, M. Takahashi, M. Sekine, S. Baba, A. Ankai, H. Kosugi, A. Hosoyama, S. Fukui, Y. Nagai, K. Nishijima, H. Nakazawa, M. Takamiya, S. Masuda, T. Funahashi, T. Tanaka, Y. Kudoh, J. Yamazaki, N. Kushida, A. Oguchi & H. Kikuchi 1999. Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. DNA Res. 6: 83–101.

Kawarabayasi, Y., M. Sawada, H. Horikawa, Y. Haikawa, Y. Hino, S. Yamamoto, M. Sekine, S. Baba, H. Kosugi, A. Hosoyama, Y. Nagai, M. Sakai, K. Ogura, R. Otsuka, H. Nakazawa, M. Takamiya, Y. Ohfuku, T. Funahashi, T. Tanaka, Y. Kudoh, J. Yamazaki, N. Kushida, A. Oguchi, K. Aoki & H. Kikuchi, 1998. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. DNA Res. 5: 55–76.

Kawashima, T., N. Amano, H. Koike, S. Makino, S. Higuchi, Y. Kawashima-Ohya, K. Watanabe, M. Yamazaki, K. Kanehori, T. Kawamoto, T. Nunoshiba, Y. Yamamoto, H. Aramaki, K. Makino & M. Suzuki, 2000. Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium*. Proc. Natl. Acad. Sci. USA 97: 14257–14262.

Klenk, H.P., R.A. Clayton, J.F. Tomb, O. White, K.E. Nelson, K.A. Ketchum, R.J. Dodson, M. Gwinn, E.K. Hickey, J.D. Peterson, D.L. Richardson, A.R. Kerlavage, D.E. Graham, N.C. Kyrpides, R.D. Fleischmann, J. Quackenbush, N.H. Lee, G.G. Sutton, S. Gill, E.F. Kirkness, B.A. Dougherty, K. McKenney, M.D. Adams, B. Loftus & J.C. Venter, 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature 390: 364–370.

Kunst, F., N. Ogasawara, I. Moszer, A.M. Albertini, G. Alloni, V. Azevedo, M.G. Bertero, P. Bessieres, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M. Braun, S.C. Brignell, S. Bron, S. Brouillet, C.V. Bruschi, B. Caldwell, V. Capuano, N.M. Carter, S.K. Choi, J.J. Codani, I.F. Connerton, A. Danchin et al. 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature 390: 249–256.

Kurland, C.G. & S.G.E. Andersson, 2000. Origin and evolution of the mitochondrial proteome. Microbiol. Mol. Biol. Rev. 64: 786–820.

Kurtz, S. & C. Schleiermacher, 1999. REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics 5: 426–427.

Lawrence, J.G., R.W. Hendrix & S. Casjens, 2001. Where are the pseudogenes in bacteiral genomes? Trends Microbiol. (in press).

Maruyama, S., R.W. Kasten, H. Boulouis, N.A. Gurfield, Y. Katsube & B.B. Chomel, 2001. Genomic diversity of *Bartonella henselae* isolates from domestic cats from Japan, the USA and France by pulsed-field gel electrophoresis. Vet. Microbiol. 79: 337–349.

McClelland, M., K.E. Sanderson, J. Spieth, S.W. Clifton, P. Latreille, L. Courtney, S. Porwollik, J. Ali, M. Dante, F. Du, S. Hou, D. Layman, S. Leonard, C. Nguyen, K. Scott, A. Holmes, N. Grewal, E. Mulvaney, E. Ryan, H. Sun, L. Florea, W. Miller, T. Stoneking, M. Nhan, R. Waterston & R.K. Wilson, 2001. Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. Nature 413: 852–856.

May, B.J., Q. Zhang, L.L. Li, M.L. Paustian, T.S. Whittam & V. Kapur, 2001. Complete genomic sequence of *Pasteurella multocida*, Pm70. Proc. Natl. Acad. Sci. USA 98: 3460–3465.

Meng, B.Y., K. Shinozaki & M. Sugiura, 1989. Genes for the ribosomal proteins S12 and S7 and elongation factors EF-G and EF-Tu of the cyanobacterium, *Anacystis nidulans*: structural homology. Mol. Gen. Genet. 216: 25–30.

Moxon, R., 1994. Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr. Biol. 1: 24–33.

Nelson, K.E., R.A. Clayton, S.R. Gill, M.L. Gwinn, R.J. Dodson, D.H. Haft, E.K. Hickey, J.D. Peterson, W.C. Nelson, K.A. Ketchum, L. McDonald, T.R. Utterback, J.A. Malek, K.D. Linher, M.M. Garrett, A.M. Stewart, M.D. Cotton, M.S. Pratt, C.A. Phillips, D. Richardson, J. Heidelberg, G.G. Sutton, R.D. Fleischmann, J.A. Eisen & C.M. Fraser, 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 399: 323–329.

Ng, W.V., S.P. Kennedy, G.G. Mahairas, B. Berquist, M. Pan, H.D. Shukla, S.R. Lasky, N.S. Baliga, V. Thorsson, J. Sbrogna, S. Swartzell, D. Weir, J. Hall, T.A. Dahl, R. Welti, Y.A. Goo, B. Leithauser, K. Keller, R. Cruz, M.J. Danson, D.W. Hough, D.G. Maddocks, P.E. Jablonski, M.P. Krebs, C.M. Angevine, H. Dale, T.A. Isenbarger, R.F. Peck, M. Pohlschroder, J.L. Spudich, K.W. Jung, M. Alam, T. Freitas, S. Hou, C.J. Daniels, P.P. Dennis, A.D. Omer, H. Ebhardt, T.M. Lowe, P. Liang, M. Riley, L. Hood & S. DasSarma, 2000. Genome sequence of *Halobacterium* species NRC-1. Proc. Natl. Acad. Sci. USA 97: 12176–12181.

Ochman, H., J.G. Lawrence & E.A. Groisman, 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405: 299–304.

Ogata, H., S. Audic, V. Barbe, F. Artiguenave, P.E. Fournier, D. Raoult & J.M. Claverie, 2000. Selfish DNA in protein coding genes. Science 290: 347–350.

Ogata, H., S. Audic, P. Renesto-Audiffren, P.E. Fournier, V. Barbe, D. Samson, V. Roux, P. Cossart, J. Weissenbach, J.M. Claverie & D. Raoult, 2001. Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. Science 293: 2093–2098.

Ohama, T., F. Yamao, A. Muto & S. Osawa, 1987. Organization and codon usage of the streptomycin operon in Micrococcus luteus, a

bacterium with a high genomic G + C content. J. Bacteriol. 169: 4770–4777.

Olsen, G.J., C.R. Woese & R. Overbeek, 1994. The winds of (evolutionary) change: breathing new life into microbiology. J. Bacteriol. 175: 3893–3896.

Parkhill, J., B.W. Wren, K. Mungall, J.M. Ketley, C. Churcher, D. Basham, T. Chillingworth, R.M. Davies, T. Feltwell, S. Holroyd, K. Jagels, A.V. Karlyshev, S. Moule, M.J. Pallen, C.W. Penn, M.A. Quail, M.A. Rajandream, K.M. Rutherford, V.l.i.e.t. van, S. Whitehead & B.G. Barrell, 2000a. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. Nature 403: 665–668.

Parkhill, J., M. Achtman, K.D. James, S.D. Bentley, C. Churcher, S.R. Klee, G. Morelli, D. Basham, D. Brown, T. Chillingworth, R.M. Davies, P. Davis, K. Devlin, T. Feltwell, N. Hamlin, S. Holroyd, K. Jagels, S. Leather, S. Moule, K. Mungall, M.A. Quail, M.A. Rajandream, K.M. Rutherford, M. Simmonds, J. Skelton, S. Whitehead, B.G. Spratt & B.G. Barrell, 2000b. Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. Nature 404: 502–506.

Parsons, J.D., 1995a. Improved tools for DNA comparison and clustering. Comput. Appl. Biosci. 11: 603–613.

Parsons, J.D., 1995b. Miropeats: graphical DNA sequence comparisons. Comput. Appl. Biosci. 11: 615–619.

Perna, N.T., G. Plunkett, V. Burland, B. Mau, J.D. Glasner, D.J. Rose, G.F. Mayhew, P.S. Evans, J. Gregor, H.A. Kirkpatrick, G. Posfai, J. Hackett, S. Klink, A. Boutin, Y. Shao, L. Miller, E.J. Grotbeck, N.W. Davis, A. Lim, E.T. Dimalanta, K.D. Potamousis, J. Apodaca, T.S. Anantharaman, J. Lin, G. Yen, D.C. Schwartz, R.A. Welch & F.R. Blattner, 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 409: 529–533.

Petes, T.D. & C.W. Hill, 1988. Recombination between repeated genes in microorganisms. Annu. Rev. Genet. 22: 147–168.

Peterson, S.N., C.C. Bailey, J.S. Jensen, M.B. Borre, E.S. King, K.F. Bott & C.A. Hutchison, 1995. Characterization of repetitive DNA in the *Mycoplasma genitalium* genome: possible role in the generation of antigenic variation. Proc. Natl. Acad. Sci. U S A 92: 11829–11833.

Reid, S.D., C.J. Herbelin, A.C. Bumbaugh, R.K. Selander & T.S. Whittam, 2000. Parallell evolution of virulence in pathogenic *Escherichia coli*. Nature 406: 64–67.

Rocha, E.P., A. Danchin & A. Viari, 1999a. Functional and evolutionary roles of long repeats in prokaryotes. Res. Microbiol. 150: 725–733.

Rocha, E.P., A. Danchin & A. Viari, 1999b. Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in *Bacillus subtilis* and other competent bacteria. Mol. Biol. Evol. 16: 1219–1230.

Romero, D., R. Palacios, 1997. Gene amplification and genomic plasticity in prokaryotes. Annu Rev. Genet. 31: 91–111.

Ruepp, A., W. Graml, M.L. Santos-Martinez, K.K. Koretke, C. Volker, H.W. Mewes, D. Frishman, S. Stocker, A.N. Lupas & W. Baumeister, 2000. The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. Nature 407: 508–513.

Sanderson, K.E. & S.L. Liu, 1998. Chromosomal rearrangements in enteric bacteria. Electrophoresis 19: 569–572.

Shigenobu, S., H. Watanabe, M. Hattori, Y. Sakaki & H. Ishikawa, 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. Nature 407: 81–86.

Shirai, M., H. Hirakawa, M. Kimoto, M. Tabuchi, F. Kishi, K. Ouchi, T. Shiba, K. Ishii, M. Hattori, S. Kuhara & T. Nakazawa, 2000. Comparison of whole genome sequences of *Chlamydia pneumoniae* J138 from Japan and CWL029 from USA. Nucl. Acids Res. 28: 2311–2314.

Simpson, A.J., F.C. Reinach, P. Arruda, F.A. Abreu, M. Acencio, R. Alvarenga, L.M. Alves, J.E. Araya, G.S. Baia, C.S. Baptista, M.H. Barros, E.D. Bonaccorsi, S. Bordin, J.M. Bove, M.R. Briones, M.R. Bueno, A.A. Camargo, L.E. Camargo, D.M. Carraro, H. Carrer, N.B. Colauto, C. Colombo, F.F. Costa, M.C. Costa, C.M. Costa-Neto, L.L. Coutinho, M. Cristofani, E. Dias-Neto, C. Docena, H. El-Dorry, A.P. Facincani, A.J. Ferreira, V.C. Ferreira, J.A. Ferro, J.S. Fraga, S.C. Franca, M.C. Franco, M. Frohme, L.R. Furlan, M. Garnier, G.H. Goldman, M.H. Goldman, S.L. Gomes, A. Gruber, P.L. Ho, J.D. Hoheisel, M.L. Junqueira, E.L. Kemper, J.P. Kitajima & C.L. Marino, 2000. The genome sequence of the plant pathogen *Xylella fastidiosa*. Nature 406: 151–157.

Smith, D.R., L.A. Doucette-Stamm, C. Deloughery, H. Lee, J. Dubois, T. Aldredge, R. Bashirzadeh, D. Blakely, R. Cook, K. Gilbert, D. Harrison, L. Hoang, P. Keagle, W. Lumm, B. Pothier, D. Qiu, R. Spadafora, R. Vicaire, Y. Wang, J. Wierzbowski, R. Gibson, N. Jiwani, A. Caruso, D. Bush & J.N. Reeve, 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. J. Bacteriol. 179: 7135–7155.

Stephens, R.S., S. Kalman, C. Lammel, J. Fan, R. Marathe, L. Aravind, W. Mitchell, L. Olinger, R.L. Tatusov, Q. Zhao, E.V. Koonin & R.W. Davis, 1998. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science 282: 754–759.

Syvanen, A.C., H. Amiri, A. Jamal, S.G.E. Andersson & C.G. Kurland, 1996. A chimeric disposition of the elongation factor genes in *Rickettsia prowazekii*. J. Bacteriol. 178: 6192–6199.

Takami, H., K. Nakasone, Y. Takaki, G. Maeno, R. Sasaki, N. Masui, F. Fuji, C. Hirama, Y. Nakamura, N. Ogasawara, S. Kuhara & K. Horikoshi, 2000. Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. Nucl. Acids Res. 28: 4317–4331.

Tettelin, H., N.J. Saunders, J. Heidelberg, A.C. Jeffries, K.E. Nelson, J.A. Eisen, K.A. Ketchum, D.W. Hood, J.F. Peden, R.J. Dodson, W.C. Nelson, M.L. Gwinn, R. DeBoy, J.D. Peterson, E.K. Hickey, D.H. Haft, S.L. Salzberg, O. White, R.D. Fleischmann, B.A. Dougherty, T. Mason, A. Ciecko, D.S. Parksey, E. Blair, H. Cittone, E.B. Clark, M.D. Cotton, T.R. Utterback, H. Khouri, H. Qin, J. Vamathevan, J. Gill, V. Scarlato, V. Masignani, M. Pizza, G. Grandi, L. Sun, H.O. Smith, C.M. Fraser, E.R. Moxon, R. Rappuoli & J.C. Venter, 2000. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. Science 287: 1809–1815.

Tiboni, O., R. Cantoni, R. Creti, P. Cammarano & A.M. Sanangelantoni, 1991. Phylogenetic depth of *Thermotoga maritima* inferred from analysis of the fus gene: amino acid sequence of elongation factor G and organization of. J. Mol. Evol. 33: 142–151.

Tomb, J.F., O. White, A.R. Kerlavage, R.A. Clayton, G.G. Sutton, R.D. Fleischmann, K.A. Ketchum, H.P. Klenk, S. Gill, B.A. Dougherty, K. Nelson, J. Quackenbush, L. Zhou, E.F. Kirkness, S. Peterson, B. Loftus, D. Richardson, R. Dodson, H.G. Khalak, A. Glodek, K. McKenney, L.M. Fitzegerald, N. Lee, M.D. Adams, J.C. Venter & J.C. Venter, 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 388: 539–547.

Vijgenboom, E., L.P. Woudt, P.W. Heinstra, K. Rietveld, Haarlem. van, Wezel. van, S. Shochat & L. Bosch, 1994. Three *tuf*-like genes in the kirromycin producer *Streptomyces ramocissimus*. Microbiology 140: 983–998.

White, O., J.A. Eisen, J.F. Heidelberg, E.K. Hickey, J.D. Peterson, R.J. Dodson, D.H. Haft, M.L. Gwinn, W.C. Nelson, D.L. Richardson, K.S. Moffat, H. Qin, L. Jiang, W. Pamphile, M. Crosby, M. Shen, J.J. Vamathevan, P. Lam, L. McDonald, T. Utterback, C. Zalewski, K.S. Makarova, L. Aravind, M.J. Daly & C.M. Fraser, 1999. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science 286: 1571–1577.

Woods, S.A. & S.T. Cole, 1990. A family of dispersed repeats in *Mycobacterium leprae*. Mol. Microbiol. 4: 1745–1751.