Genome reduction in prokaryotic obligatory intracellular parasites of humans: a comparative analysis

Kishore R. Sakharkar,¹ Pawan Kumar Dhar¹ and Vincent T. K. Chow²

Correspondence

Vincent T. K. Chow
micctk@nus.edu.sg

¹Bioinformatics Institute, Singapore

²Human Genome Laboratory, Department of Microbiology, Faculty of Medicine, National University of Singapore, Singapore

Obligatory intracellular parasites have undergone significant genome reduction by gene loss over time in the context of their obligate associations with the host. The flux, streamlining and elimination of genes in these genomes constitute a selective and ongoing process. Comparative analyses of five completely sequenced obligatory intracellular parasite genomes reveal that these genomes display marked similarities in patterns of protein length and frequency distribution, with substantial sharing of a 'backbone genome'. From category distribution based on the database of cluster of orthologous groups of proteins (COG), it is clear that habitat is a major factor contributing to genome reduction. It is also observed that, in all five obligatory intracellular parasites, the reduction in number of genes/proteins is greater for proteins with lengths of 200–600 amino acids. These comparative analyses highlight that gene loss is function-dependent, but is independent of protein length. These comparisons enhance our knowledge of the forces that drive the extreme specialization of the bacteria and their association with the host.

INTRODUCTION

Obligatory intracellular parasitism serves as an excellent model to study how bacteria exploit the functions of their host cells. Obligatory intracellular parasites possess small genomes and display a tendency towards further genome reduction (Andersson & Andersson, 1999) (Fig. A and Table A, available as supplementary material in IJSEM Online). Gene degradation appears to be a common feature of obligatory intracellular parasites, targeting overlapping subsets of potentially dispensable genes while adapting to the selective pressures of different niches (Andersson & Andersson, 2001). Therefore, genes found as multiple copies may outline their specific adaptations (Andersson & Kurland, 1998). The flux, streamlining and elimination of genes in genomes of obligate intracellular parasitic species constitute an ongoing process, and may represent a function of lifestyle and genome-coding capacity in terms of compactness. There have been many reports on the

evolution of these bacteria from larger genomes by genome deterioration (Andersson *et al.*, 1998). Zomorodipour & Andersson (1999) have provided examples of reductive convergent evolution in the genomes of *Rickettsia prowazekii* and *Chlamydia trachomatis*, and have associated this phenomenon with metabolic parasitism in response to intracellular habitat. The high fraction of non-coding DNA in many genomes is speculated to represent ancient genes in the process of elimination (Andersson *et al.*, 1998). Furthermore, pseudogenes are postulated to be genes in the process of becoming lost. Andersson *et al.* (1998) have shown in *Rickettsia* that deletions are far more common than insertions, and on average much larger in size. They speculated that once a rickettsial gene becomes non-functional, it will be eliminated from the genome solely by mutational events. The higher proportion of intergenic DNA may be scattered remnants of genes lost in a stepwise process (Lawrence *et al.*, 2001; Tamas *et al.*, 2001; Frank *et al.*, 2002).

C. trachomatis, *Chlamydia pneumoniae*, *Mycobacterium leprae*, *R. prowazekii* and *Rickettsia conorii* are completely sequenced eubacterial obligate intracellular parasites that are pathogenic for humans (Andersson *et al.*, 1998; Stephens *et al.*, 1998; Kalman *et al.*, 1999; Cole *et al.*, 2001; Ogata *et al.*, 2001). Despite their similarity in biology and reduced genome size, these species display extreme diversity in tissue tropism and disease expression that hitherto remains a major unanswered question in microbial behaviour. In

Published online ahead of print on 23 April 2004 as DOI 10.1099/ijs.0.63090-0.

Bar charts showing a representation of bacterial genome sizes and gene numbers (Fig. A), the distribution of COG categories and their percentage representation in the bacterial genomes (Fig. B) and the percentage change in protein length distributions in the five obligatory intracellular parasites compared to *E. coli* (Fig. C), and tables listing gene numbers and genome sizes of obligatory intracellular parasites (Table A) and the percentage distribution of different COG categories (Table B) are available as supplementary material in IJSEM Online.

efforts to characterize the evolutionary forces underlying genome reduction and to understand the dynamics of microbial genomes, a critical issue concerns the size, processes and content of deletions. This study is an attempt to find 'fingerprints' by comparing the genomes of the five completely sequenced obligatory intracellular parasites with that of the free-living bacterium *Escherichia coli* (Perna *et al.*, 2001). This may shed light on the differential loss of genes in response to bacterial lifestyle. Knowledge in this area may contribute to elucidating the fundamental mechanisms of host-pathogen interactions, with specific reference to the recognition of determinants responsible for host specificity, virulence and disease pathogenesis, and to the identification of new targets for vaccine and drug design.

METHODS

The genome sequences of the five completely sequenced prokaryotic obligatory intracellular parasites (*C. trachomatis*, *C. pneumoniae*, *M. leprae*, *R. prowazekii* and *R. conorii*) and of the free-living bacterium *E. coli* were downloaded from the National Center for Biotechnology Information (NCBI) website together with their GenBank annotations. The COG (database of cluster of orthologous groups of proteins) files (Natale *et al.*, 2000) describing the gene categories were also downloaded from GenBank (Benson *et al.*, 2003) for further analysis. The Proteome Profile Database (PPD) program (<http://www.bii.a-star.edu.sg/~kishore/PPD/PPD.html>) was employed for clustering and tabulation of protein sequences according to their length distributions, and percentage changes in genomes were calculated (Sakharkar & Chow, 2004).

RESULTS AND DISCUSSION

In general, prokaryotic obligate intracellular parasites of humans have reduced genomes compared to the free-living bacterium *E. coli*, indicating a continual selective pressure for a minimal genome. One explanation is the intracellular habitat preventing the uptake of foreign DNA by these organisms as suggested by Tamas *et al.* (2002). Other reasons include the differential loss of genes in these organisms with loss of gene function, or genome decay with increased adaptation to the host (Andersson *et al.*, 1998). Hence, reductive convergent evolution may be a consequence of prolonged intracellular life.

We present here the results of our comparative analyses of the genome sizes, gene distributions in various COG categories (Natale *et al.*, 2000) and protein lengths of five obligatory intracellular prokaryotes with respect to *E. coli*.

Genome size and number of genes

Comparison of these genome sequences revealed wide variation in size as well as in number of genes and proteins. Obligatory intracellular parasites have small genomes. Since most bacterial genomes primarily contain coding DNA, genome reduction in prokaryotes must involve the loss of metabolic functions and physiological capacities with important phenotypic implications (Andersson &

Kurland, 1998; Bergthorsson & Ochman, 1998; Ochman & Moran, 2001). Moreover, it can be inferred that the number of genes decreases with reduction in genome size. The loss of these biochemical capabilities may also account for the inability to culture these bacteria in cell-free systems. It is clear that there is massive genome reduction and convergent evolution in response to lifestyle in all the five obligatory intracellular parasites, with the maximum reduction in genome size noted for *C. trachomatis*. The least number of genes is observed in *R. prowazekii*. The distributions of genome size and number of genes based on NCBI annotation are illustrated in Fig. A and Table A (available as supplementary material in IJSEM Online). Convergent evolution is suggested to be a potent indicator of optimal design. Since the loss of redundant genes may not necessarily be lethal, this reduction targets potentially dispensable genes while adapting to the selective pressures of different niches.

Genome decay and host adaptation

The five obligatory intracellular genomes display marked similarities in COG category distributions (Fig. B and Table B, available as supplementary material in IJSEM Online). About half of the proteins in all the obligatory intracellular parasites as well as in *E. coli* (the free-living bacterium selected for comparison) are of unknown biological function. Many genes are present in one organism but absent from another. Identification of such genes is of particular importance for the mutually exclusive biological, virulent and pathogenic capabilities of each species.

Extensive gene loss is a general attribute of obligatory intracellular parasites. Genome comparisons of these organisms reveal numerous cases of orthologous pairs of open reading frames with assigned functions. These data were derived from the COG division of genome data from NCBI. From COG category distribution, it is clear that habitat is a major factor contributing to genome reduction. Supply of energy, nutrients and metabolites from the host supplements the bacterium's potential to synthesize them. Thus, genes involved in many metabolic pathways (e.g. enzymes) are partially or completely lost from these parasites, and the loss of genes from these pathways underlies the reduction in number of genes and corresponding genome size. However, this must be complemented by increase in transporter systems for uptake from their milieu. Thus, parasitic lifestyle gives rise to problems that must be solved by homologous or analogous systems from the host or environmental niche.

Our data show that genes for translation (J), co-enzyme transport and metabolism (H), lipid transport and metabolism (I), intracellular trafficking and secretion (U) increase their representation per genome size in all the five obligatory intracellular parasites compared to *E. coli*. However, genes belonging to the functional categories of signal transduction mechanisms (T), carbohydrate transport and metabolism (G), inorganic ion transport and metabolism (P) decrease in

percentage of genome representation. As the cytoplasm of an eukaryotic cell is nutritionally very rich, the biosynthesis of small molecules such as carbohydrates, amino acids and nucleotides in the prokaryotic parasite is rendered non-essential. Given that there is no mechanism for the gain of genes in the closed intracellular environment of the host-pathogen conglomerate, it is possible that the percentage increase in representation may be due to loss of genes from other categories. Some functional categories are lost altogether or show a drastic reduction in certain genomes, for example secondary metabolite biosynthesis, transport and catabolism (Q), cell motility (N), defence mechanisms (V), and this may be explained based on parasitic lifestyle. Thus, identification of shared genes supports the requirement for these capabilities in biological systems that have evolved over long-term associations with mammalian host cells, to reduce metabolic capacities while optimizing survival, growth and transmission of pathogens. In addition, identification of lost genes supports the organism's optimal requirements according to its niche and habitat.

Protein length distribution

The parasitic genomes display marked similarities in patterns of protein length and frequency distribution. The protein length distribution profiles for all five obligatory intracellular parasites and *E. coli* are depicted in Fig. 1. The length of a protein sequence is generally determined by its function, and the wide variance in the lengths of an

organism's proteins reflects the diversity of specific functional roles for these proteins. It is noteworthy that about 56% of the proteins in *R. conorii* are less than 200 amino acids in length. This is much more than its closest neighbour, *R. prowazekii*, about 34% of whose proteins are less than 200 amino acids. Interestingly, an increase in percentage genome representation of genes encoding proteins less than 200 amino acids in length is evident in *R. conorii* compared to *E. coli*. Genome reduction is a common phenomenon in obligatory intracellular parasites. A detailed analysis of proteins less than 200 amino acids in length in all five parasitic bacteria reveals that these proteins are mainly essential ones involved in translation (e.g. tRNA and ribosomal proteins), and are thus vital for survival.

R. conorii also exhibits maximum gain of genes (increase in percentage representation of the genome) for genes encoding proteins of less than 200 amino acids. This may result from a lack of loss of genes in this category compared to other categories, culminating in overall gain in percentage genome representation. Another possible explanation may be the gene-splitting phenomenon in *R. prowazekii* compared to *R. conorii*, as described by Ogata *et al.* (2001). Our results thus concur with previous findings, such as gene splitting or fission that may lead to smaller proteins.

The reduction in genome size of obligatory intracellular

Fig. 1. Protein length distribution profiles of the prokaryotic genomes under study.

parasites results in genome stability via loss of selected groups of proteins, e.g. loss of prophage component proteins. These proteins are mainly involved in genome dynamics as well as gene mobility, and loss of these proteins decreases the rate of gene rearrangement which is a requirement for an obligatory intracellular lifestyle. These prophage components represent more than 600 proteins in *E. coli*, but their number is reduced to less than three in all five obligatory intracellular parasites. These proteins thus contribute significantly to genome reduction in parasitic bacteria.

In most of the micro-organisms studied, there is maximum reduction in the number of genes encoding proteins of 200–600 amino acids in length (Fig. C, supplementary material in IJSEM Online, and Fig. 2). It was noted that most of the proteins in this range are metabolic enzymes (data available online). Despite being more than 800 amino acids long, enzymes categorized as housekeeping and that account for essential requirements including DNA replication, repair, transcription and translation (e.g. RNA polymerase, gyrase, exonuclease) are not lost in any of the obligatory intracellular parasites. Interestingly, the distributions of protein lengths are substantially similar in all the genomes under study, except *R. conorii* which has the maximum number of small proteins. This analysis clearly shows that the loss of genes from obligatory intracellular

parasites is not dependent on protein length. Both long and short proteins of all prokaryotes are functionally important in their own right, and bacteria selectively lose genes based on their environmental niche.

Conclusion

Reductive evolution is found to be common among obligatory intracellular parasites. During genome reduction, genes that once had important functions accumulate mutations and eventually disappear because they are no longer or less under strong selective pressures, and their functions are of negligible relevance to the organism's lifestyle. This phenomenon may be explained by a change in the niche of the organism or its inability to retain certain genes under weak selection pressure. Such comparisons increase our knowledge of the forces that drive the extreme specialization of bacteria to facilitate their association with their hosts. Analysis of other genomes should allow us to test the generality of the observed patterns, to identify the host-specific functions in obligatory intracellular parasites, and to better understand the strategies by which bacteria form intimate one-sided associations with eukaryotic cells. Our comparative analyses thus strongly support massive genome reduction in all five completely sequenced obligatory intracellular parasites of humans, and highlight that gene loss is function-dependent but is independent of

Fig. 2. Distribution of the number of proteins of specific lengths in the five obligatory intracellular parasites compared to *E. coli*.

protein length. Detailed analyses on obligatory intracellular parasites will provide insights into the reasons for observed reduction in genes of specific lengths, and the forces responsible for differential maintenance or loss of specific genes and for variations in genome size. Protein length profiles of these genomes reveal substantial sharing of a 'backbone genome' in all the obligatory intracellular parasites (results available online).

Caveats

While gene prediction algorithms work better for prokaryotic genomes than for eukaryotic genomes, it should be emphasized that the gene prediction is not perfect since absence of a gene from the annotation for bacterial genomes is not proof that the gene is missing from an organism's genome. Furthermore, there are limitations with respect to evolutionary diversity as our sample size of five genomes may not represent the complete set of obligatory intracellular parasites. Notwithstanding this, our data provide an overall picture of the genomes of obligatory intracellular parasites of humans against our limited understanding of the biology of micro-organisms in general.

REFERENCES

- Andersson, J. O. & Andersson, S. G. (1999). Insights into the evolutionary process of genome degradation. *Curr Opin Genet Dev* **9**, 664–671.
- Andersson, J. O. & Andersson, S. G. (2001). Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. *Mol Biol Evol* **18**, 829–839.
- Andersson, S. G. & Kurland, C. G. (1998). Reductive evolution of resident genomes. *Trends Microbiol* **6**, 263–268.
- Andersson, S. G., Zomorodipour, A., Andersson, J. O. & 7 other authors (1998). The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**, 133–140.
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. (2003). GenBank: update. *Nucleic Acids Res* **32**, D23–D26.
- Bergthorsson, U. & Ochman, H. (1998). Distribution of chromosome length variation in natural isolates of *Escherichia coli*. *Mol Biol Evol* **15**, 6–16.
- Cole, S. T., Eiglmeier, K., Parkhill, J. & 41 other authors (2001). Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007–1011.
- Frank, A. C., Amiri, H. & Andersson, S. G. (2002). Genome deterioration: loss of repeated sequences and accumulation of junk DNA. *Genetica* **115**, 1–12.
- Kalman, S., Mitchell, W., Marathe, R. & 7 other authors (1999). Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nat Genet* **21**, 385–389.
- Lawrence, J. G., Hendrix, R. W. & Casjens, S. (2001). Where are the pseudogenes in bacterial genomes? *Trends Microbiol* **9**, 535–540.
- Natale, D. A., Galperin, M. Y., Tatusov, R. L. & Koonin, E. V. (2000). Using the COG database to improve gene recognition in complete genomes. *Genetica* **108**, 9–17.
- Ochman, H. & Moran, N. (2001). Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* **292**, 1096–1099.
- Ogata, H., Audic, S., Renesto-Audiffren, P. & 8 other authors (2001). Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. *Science* **293**, 2093–2098.
- Perna, N. T., Plunkett, G., 3rd, Burland, V. & 25 other authors (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**, 529–533.
- Sakharkar, K. R. & Chow, V. T. K. (2004). PPD – Proteome Profile Database. *In Silico Biol* **4**, 0019.
- Stephens, R. S., Kalman, S., Lammel, C. J. & 9 other authors (1998). Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282**, 754–759.
- Tamas, I., Klasson, L. M., Sandstrom, J. P. & Andersson, S. G. (2001). Mutualists and parasites: how to paint yourself into a (metabolic) corner. *FEBS Lett* **498**, 135–139.
- Tamas, I., Klasson, L. M., Canback, B., Naslund, A. K., Eriksson, A. S., Wernegreen, J. J., Sandstorm, J. P., Moran, N. A. & Andersson, S. G. (2002). 50 million years of genomic stasis in endosymbiotic bacteria. *Science* **296**, 2376–2379.
- Zomorodipour, A. & Andersson, S. G. (1999). Obligate intracellular parasites: *Rickettsia prowazekii* and *Chlamydia trachomatis*. *FEBS Lett* **452**, 11–15.