Molecular Evolutionary Studies of Genome Degradation in Bacteria

BY

JAN O. ANDERSSON

Dissertation for the Degree of Doctor of Philosophy in Molecular Biology presented at Uppsala University in 1999

ABSTRACT

Andersson, J. O. 1999. Molecular Evolutionary Studies of Genome Degradation in Bacteria. Acta Universitatis Upsaliensis. *Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 499.50 pp Uppsala. ISBN 91-554-4621-3.

The genus *Rickettsia* belongs to the α-proteobacteria and consists of obligate intracellular bacteria, which often are pathogenic for humans. All *Rickettsia* have small genomes, highly adapted to an intracellular lifestyle. However, the ancestors of *Rickettsia* were most likely free-living organisms with substantially larger genomes. This thesis is a study of the reductive evolutionary processes by which *Rickettsia* has adapted to a life inside eukaryotic cells.

The *Rickettsia prowazekii* genome sequence confirmed the close phylogenetic relationship between the genus *Rickettsia* and the mitochondria. In addition, 12 putative pseudogenes and an unusually large fraction of non-coding DNA (24%) were identified. Analysis of the *metK* genomic region in different *Rickettsia* species identified *metK* as a pseudogene in all but two lineages. The pattern of mutations indicated that the pseudogenes are no longer under purifying selection, and that *metK* was inactivated several times independently in different lineages. Similar patterns were found in many other *Rickettsia* pseudogenes, revealing an ongoing genome degradation process in the *Rickettsia*.

Analysis of neutrally evolving pseudogenes showed that deletions dominate over insertions, and that there is a mutational bias towards A+T nucleotides, in the *Rickettsia* genomes. In agreement, the long intergenic regions in the *R. prowazekii* genome have a decreased G+C content. Several of these regions showed sequence homology to genes in orthologous positions in other *Rickettsia* genomes, which indicated that the long intergenic regions represent old genes that are disappearing from the genome.

The ancestor of the two major *Rickettsia* groups may have encoded 200-300 additional genes compared to *R. prowazekii*. Differential loss of mostly genus specific genes during the evolution resulted in the present-day *Rickettsia* genomes. Currently, *Rickettsia* inactivates genes at a higher than they are eliminated from the genome by fixation of deletions.

Jan O. Andersson, Evolutionary Biology Centre, Department of Molecular Evolution, Norbyvägen 18, SE-752 36 Uppsala, Sweden (Jan.Andersson@molbio.uu.se)

ISSN 1104-232X
ISBN 91-554-4621-3

Printed in Sweden by Lindbergs Grafiska HB, Uppsala 1999

till mina morföräldrar
Elis och Svea Mattsson

Main references

The thesis is based on the following papers, which will be referred to in the text by Roman numerals I-V.

I. **Andersson J. O.** & Andersson S. G. E. (1997). Genomic rearrangements during evolution of the obligate intracellular parasite *Rickettsia prowazekii* as inferred from an analysis of 52015 bp nucleotide sequence. *Microbiology*, **143**, 2783-2795.

II. **Andersson J. O.** & Andersson S. G. E. (1999). Genome degradation is an ongoing process in *Rickettsia*. *Mol Biol Evol*, **16**, 1178-1191.

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

IV. **Andersson J. O.** & Andersson S. G. E. (1999). Pseudogenes, junk DNA and the dynamics of *Rickettsia* genomes. *manuscript*

V. **Andersson J. O.** & Andersson S. G. E. (1999). Insights into the evolutionary process of genome degradation. *Curr Opin Genet Dev*, **9**, 664-671.

Paper I was reproduced by permission from the Society for General Microbiology, paper II was reproduced by permission from the Society for Molecular Biology and Evolution, paper III was reproduced by permission from Macmillan Magazines Ltd. and paper V was reproduced by permission from Elsevier Science Ltd.

Table of contents

Preface

In 1965 Zuckerkandl and Pauling pointed out that DNA molecules are not only the genetic blueprint for all the functions in the cell, DNA molecules are also "documents of evolutionary history" (1). The work presented in this thesis work explores these "historical documents" for a group of bacteria called *Rickettsia*. The reconstructed history turns out to be extremely fascinating, telling a story about the radical, and still ongoing, transition process for a bacterium from a free-living to an intracellular lifestyle. Ironically, the evolutionary history of these pathogenic bacteria, that have caused millions of human deaths throughout history, included lots of sudden, and sometimes maybe mysterious, deaths of genes.

However, as in all science, the ambition has been to create theories that do not only explain the observed data. My hope is that this work on the dynamics of the *Rickettsia* genomes has contributed to the understanding of bacterial genome sequences in general, and to the understanding of genome evolution during the transition processes that involve reduction in genome size and content in particular.

The first part of the thesis presents the biological background to the genus *Rickettsia* in order to give the reader the opportunity to become familiar with these strange bacteria. The rest of the thesis is devoted to the topic of genome degradation. A summery of the results from the papers is followed by a presentation of the known cases of genome degradation in bacterial and organellar genomes. Finally the evolutionary forces that may drive this process are discussed. The goal is to give the reader a coherent view of reductive evolution and genome degradation in bacteria.

Uppsala, November 1999

The genus *Rickettsia*

The genus *Rickettsia* was named after Howard Ricketts, the pathologist who together with Stanislaus Prowazek discovered the causative agent of epidemic typhus, *Rickettsia prowazekii* in 1909 (2). The same year Charles Nicolle was able to show that typhus was transmitted between humans by the louse (3). The search for a cure against the deadly disease could begin, and in 1930 Rudolph Weigl had developed a vaccine (4). However, early *Rickettsia* research was a dangerous business, several people in Weigl's staff developed typhus and died, which had also happened to both Ricketts and Prowazek before (5). On the other hand, the only survivor of the pioneers, Charles Nicolle, received the Nobel Prize in 1928 "for his work on typhus" (6).

Besides the obvious medical interest, studies of the genus *Rickettsia* can be motivated from an evolutionary point of view. The close phylogenetic affinity to the mitochondria makes the genus interesting for studies on the bacterial contribution to the eukaryotic cell. In addition, the genus has adopted an obligate intracellular lifestyle, an evolutionary process that is associated with extensive genomic degradation, which is the main topic of this thesis.

The phylogenetic position of the genus *Rickettsia*

The members of the genus *Rickettsia* are small obligate intracellular parasitic gram-negative coccobacilli measuring 0.7-1.0 µm in length and 0.3-0.5 µm in width, (7). Historically the name "rickettsia" has been used for any small rod that could not be cultivated. It was not until the introduction of molecular taxonomic methods that the phylogenetic relationships between these bacteria could be studied properly, which caused a lot of reorganisations of the earlier classification (8). Today the genera *Rickettsia* and *Orientia* make up the tribe *Rickettsieae*, which together with *Ehrlichieae* and *Wolbachieae* form the family *Rickettsiaceae*. The order *Rickettsiales*, in turn, is a member of the α subdivision of the proteobacteria and contains the families *Rickettsiaceae*, rickettsia group ciliate endosymbionts and some environmental samples (9).

This work focus on the genus *Rickettsia*, and every time the word *Rickettsia* is used it refers to the genus.

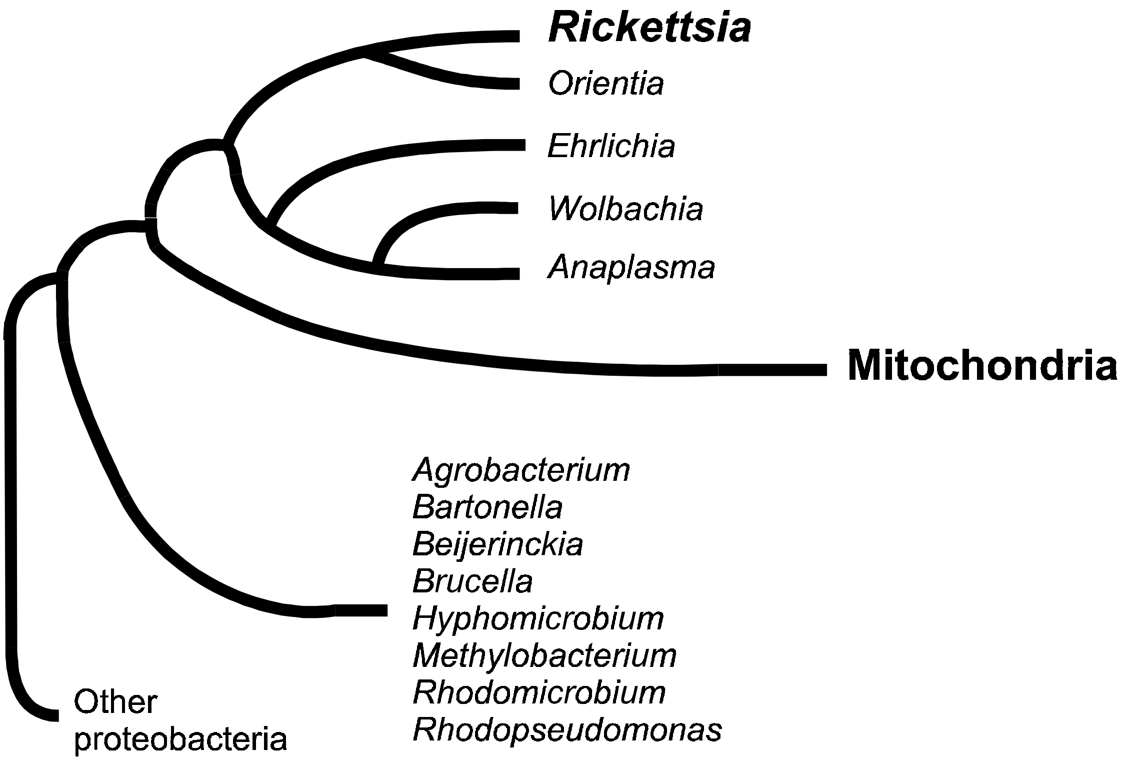

Figure 1. Schematic phylogenetic tree showing the specific phylogenetic relationship between mitochondria and the *Rickettsiaceae*. This tree is a subset of a tree based on 16S rRNA[1] sequences, containing only the α-proteobacteria showing the closest relationship with the mitochondria (10).

α-proteobacteria and the close phylogenetic relationship with the mitochondria

The mitochondrion is the eukaryotic organelle that carries out the aerobic respiration in the cell. Morphologically the organelle resembles bacteria, and the finding that it contained its own genome led to the hypothesis of an endosymbiotic origin of the mitochondrion (11). Studies of the mitochondrial genome established the α-proteobacteria as the closest modern relative to the mitochondria (12, 13). More dense sampling of the 16S rRNA gene both among mitochondria and bacteria, and sequencing of a couple of protein coding genes, showed a specific relationship between the family *Rickettsiaceae*, within the α-proteobacteria, and the mitochondria (Figure 1; 10, 14-16).

The complete *R. prowazekii* genome offered a more detailed picture of the evolutionary connection between *Rickettsia* and mitochondria (paper III). Phylogenetic analysis of an extended set of proteins confirmed the special relationship with the mitochondria (Figure 5b & 6b in paper III). Analysis of gene order structures in *R. prowazekii* and mitochondria showed similarities, as expected from the common ancestry, as well as differences that have occurred during the evolution of the two

[1] **rRNA** ribosomal ribonucleic acid.

distinct lineages (Figure 5a & 6a in paper III). Genes coding for proteins similar to more than half of 300 proteins encoded in the nucleus of *Saccharomyces cerevisiae*, but targeted to the mitochondria, could be found in the *R. prowazekii* genome (Figure 4 in paper III). This indicated that there, indeed, has been an extensive transfer of genes from the mitochondria to the nucleus, but also that many proteins utilised in the mitochondria are derived from other sources than the endosymbiont that gave rise to the organelle.

The two major groups of *Rickettsia*

Historically the genus *Rickettsia* has been divided into three groups: the TG[2], the SFG[3], and the STG[4] (17). The only member of the STG, *Rickettsia tsutsugamushi*, has recently been shown to be phylogenetically distinct from the other *Rickettsia*, and has therefore been transferred to its own genus, *Orientia*, within the tribe *Rickettsieae* (18). The TG has only two members, *R. prowazekii* and *Rickettsia typhi*, both pathogenic for humans. The SFG contains a continuously increasing number of members, both pathogenic and non-pathogenic. Currently there are 13 pathogenic SFG species described. In addition, about 20 species have been isolated from their arthropod vectors and identified on a molecular level, but have not yet been shown to be associated with any human disease (Table 1; 9). Some species previously classified as belonging to the TG or the SFG, i. e. *Rickettsia bellii* and *Rickettsia canada*, have now been shown to be phylogenetically close, but distinct to the two groups (Figure 2; 19).

Rickettsia - as pathogens

Throughout history the diseases caused by bacteria of the genus *Rickettsia* have been responsible for millions of deaths. The plague of Athens in 430-426 BC may have been caused by a typhus epidemic (20), and in the wake of the First World War at least 3 million people died of typhus in Eastern Europe and Russia (21). Even today epidemic typhus is a threat, and tends to show up whenever there are problems with social conditions leading to poor hygienic conditions. During the last few years there have been outbreaks among inhabitants of refugee camps in Burundi (22, 23), and in Russia, where the political transformation have led to poor living conditions in some parts of the country (24).

[2] **TG** typhus group *Rickettsia*.
[3] **SFG** spotted fever group *Rickettsia*.
[4] **STG** scrub typhus group *Rickettsia*.

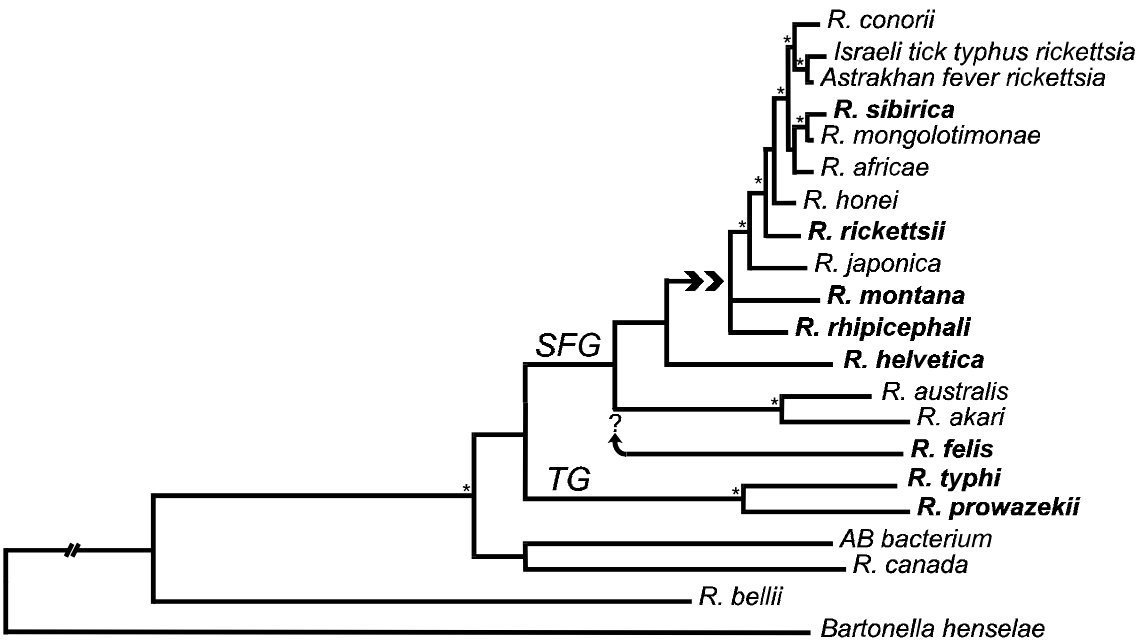

Figure 2. Phylogenetic relationship within the *Rickettsia* genus for the species listed in Table 1. A combination of phylogenetic trees based on *ompA* (upper part) and *gltA* (lower part), separated by the double arrow. Both parts are based on nucleotide sequences which were aligned using Clustalw (25) and analysed with both the Neighbour Joining and Maximum Parsimony methods with 1000 bootstrap replicates within the Phylo_win package (26). Both methods generated the same topology. The Neighbour Joining trees are presented. The *ompA* tree is rooted using the information of that part available from the *gltA* analysis. Note that the branch lengths are not comparable between the two parts of the tree. Nodes that were supported with bootstrap values above 90% in both phylogenetic methods are indicated with an asterisk (*). The two main *Rickettsia* groups are indicated with TG and SFG. The hypothetical placement of *Rickettsia felis* is indicated by an arrow and is based on earlier studies (paper II; 27, 28), which all placed *R. felis* as an early diverging SFG. The species included in this work (paper II, IV) are in boldface. Data taken from (19, 29, 30).

The typhus group

Humans are the main reservoir for the etiologic agent of epidemic typhus, *R. prowazekii*. The disease is transmitted between humans by the body louse, *Pediculus humanus corporis* (3). The lice, that take frequent large blood meals on humans, have a tendency to desert febrile hosts to seek new healthy hosts. This behavior effectively spreads the disease between human individuals harbouring lice. However, the louse also suffers from the rickettsial infection and is killed within 1-2 weeks (17). The best way to stop an epidemic is to try to keep the lice away, which is done by bathing and changing of clothes. The clinical symptoms at the onset of disease are high fever and headache and after 5 to 7 days the patients develop a rash. The disease is fatal in 10 to 30% of the patients (31). Fortunately, the disease is efficiently and easily treated by antibiotics, a single dose of 200 mg doxycycline usually cures a patient (32).

A milder form of typhus in humans, murine typhus, is caused by *R. typhi*, the other member of the TG. Rats are the main reservoirs and the disease is transferred via the flea to humans living in close contact with rats. The symptoms are similar to epidemic typhus, but milder (31).

The spotted fever group

The SFG contains a lot of species, some that have been known for a long time, and some that are newly described. The most severe disease is Rocky Mountain spotted fever, caused by transfer of *R. rickettsii* from a tick bite. The disease is characterised by malaise, chills, headache and fever (39). The mortality rate was high before the era of antibiotics, but today Rocky Mountain spotted fever is effectively cured if recognised and treated promptly with tetracycline (17). Among the other SFG, only *Rickettsia conorii* causes a disease as severe as Rocky Mountain spotted fever. The other pathogenic species cause similar, but milder, diseases (31).

Rickettsia helvetica has been detected at a prevalence of 22% in *Ixodes ricinus* ticks in Sweden (40, 41). Recently it was found that two young Swedish men that died of sudden cardiac failure during exercise were infection by *R. helvetica*. This was the first time the presence of *R. helvetica* in humans could be shown (38). *R. helvetica* is just one example of newly described *Rickettsia* diseases, and emphasises that no

Table 1. Summary of 20 *Rickettsia* species.

Species	Group	Disease[1]	Reservoir[1]	L_G[2]	GC_G[3]
R. conorii	SFG	Mediterranean spotted fever	ticks	1.24	33.2
Isreali tick typhus rickettsia	SFG	Israeli tick typhus	ticks	1.22	ND[4]
Astrakhan fever rickettsia	SFG	Astrakhan fever	ticks	ND[4]	ND[4]
R. sibirica	SFG	Siberian tick typhus	ticks	1.24	32.5
R. mongolotimonae	SFG	Unnamed spotted fever	ticks	ND[4]	ND[4]
R. africae	SFG	African tick bite fever	ticks	1.26	ND[4]
R. rickettsii	SFG	Rocky Mountain spotted fever	ticks	1.27	32.6
R. japonica	SFG	Japanese spotted fever	ticks	1.28	ND[4]
R. montana	SFG	unknown	ticks	1.24	ND[4]
R. rhipicephali	SFG	unknown	ticks	1.25	ND[4]
R. helvetica	SFG	unnamed disease[5]	ticks	1.40	ND[4]
R. australis	SFG	Queensland tick typhus	ticks	1.27	ND[4]
R. akari	SFG	Rickettsialpox	mite	1.23	33.2
R. felis	SFG	California flea rickettsiosis	fleas	ND[4]	ND[4]
R. honei	SFG	Flinders Island spotted fever	ND[4]	ND[4]	ND[4]
R. typhi	TG	Murine typhus	rats	1.13	29.0
R. prowazekii	TG	Epidemic typhus	humans	1.11	29.0
AB bacterium		unknown	beetle	ND[4]	ND[4]
R. canada		unknown	ticks	ND[4]	ND[4]
R. bellii		unknown	ticks	1.66	ND[4]

[1] (31, 33).
[2] L_G = estimated genome size in Mb (34, 35).
[3] GC_G = estimated genomic G+C content (36, 37).
[4] Data not available
[5] (38)

Rickettsia should be treated as non-pathogenic (31).

The lifecycles of *Rickettsia*

For all *Rickettsia* species except *R. prowazekii*, humans are only incidental hosts[5] (Table 1). The SFG and *R. typhi* have symbiotic relationships with their arthropod hosts and are maintained through transovarial transmission[6] (33). Individual *Rickettsia* species are usually not specific to a single arthropod species, and a single arthropod species may be infected by different *Rickettsia* species (31). For example, both *R. typhi* and *R. felis* are maintained transovarially in the same species of fleas. Although both *Rickettsia* species have been identified in the same geographical areas, and co-infection have been shown in the laboratory (42), natural infections of both species have not yet been observed in individual fleas, which is an indication of transovarial interference (33, 43). Similar observations have been done for ticks, where an individual infected by *Rickettsia peacockii* is resistant to infection and subsequent transovarially transmission of *R. rickettsii.* The questions about the molecular basis for the transovarial interference, and if the bacteria induce reproductive incompatibility in the ticks, remain to be answered (33). However, there are no indications of long-term interaction and co-evolution between ticks and *Rickettsia* from phylogenetic studies (31), as observed between *Buchnera* and their aphid hosts (44).

Since *R. prowazekii* kills the louse host and there is no evidence that the bacterium is maintained in lice vertically, the role of humans as the reservoir in nature is important (33). The bacteria can persist for a human lifetime in patients who have survived epidemic typhus. The bacteria are activated under stressful conditions and the patient develops recrudescent typhus. A single case of the disease, known as Brill-Zinsser disease, can initiate an outbreak of epidemic typhus if louse infestations are widespread in the population (22, 45).

The molecular biology of *Rickettsia*

The intracellular lifestyle and pathogenicity of many species have made studies of *Rickettsia* difficult, complicated and sometimes dangerous. The bacteria can only be cultivated inside the host cells, which has made the development of a genetic system problematic. However, the first genetic manipulation of a rickettsial gene could recently

[5] **incidental host** a host that is not part of the normal lifecycle, but may be infected.

[6] **transovarial transmission** transfer of the bacteria from mother to offspring via the egg.

be demonstrated by incorporation of rifampin resistance from a plasmid into the genome of *R. prowazekii* (46). This represent the initial step in the establishment of a genetic system for *Rickettsia*, which, together with the availability of the complete genome sequence of *R. prowazekii* (paper III), radically will change the possibilities to study and understand the molecular biology of these intracellular parasites.

The genome sizes and G+C content

The genus *Rickettsia* is characterised by small A+T[7] rich genomes (Table 1). *Rickettsia* once had a free-living ancestor with a much larger genome, which gradually decreased in size during the adaptation to an intracellular lifestyle (47). During this reductive process, which is the main topic of this thesis, the majority of the genes of the ancestor have been lost and the genome has been extensively rearranged (papers I-V; 47, 48).

In any genome with a biased G+C[8] composition different parts of the genome will respond differently to the bias, depending on the functional constraint[9] acting on the parts (49). Firstly, the G+C content of an A+T rich genome will be lowest in non-coding regions, higher in protein coding genes and highest in tRNA[10] and rRNA genes. Secondly, within protein coding genes the G+C content values are expected to vary with the highest value in the first codon position and the lowest value in the third codon position (49).

The genomic G+C content in *Rickettsia* is about 29% within the TG and 32-33% within the SFG (Table 1). The different parts of the *R. prowazekii* genome showed the expected pattern of G+C content values variations, both regarding the three codon positions and the difference between genic and intergenic regions (papers I, III; 50). The SFG species also showed the expected pattern within genes, while some of the intergenic regions seem to be influenced by other mutational and/or selectional forces (papers II, IV). However, the characteristic G+C patterns within genes have successfully been used to identify coding regions in both groups (papers I-IV).

Regulation on the transcription level as a response for changes in the environment has been shown to occur for the 16S rRNA gene and for the genes encoding ATP/ADP translocase and citrate synthase in *R. prowazekii* (51, 52). Analysis of the major macromolecular operon identified three transcripts that indicated the existence of operons and co-ordinated gene regulation in *R. prowazekii* (53). The

[7] **A+T** the amount of adenine and thymine in DNA.
[8] **G+C** the amount of guanine and cytosine in DNA.
[9] **functional constraint** the degree of intolerance towards nucleotide substitutions.
[10] **tRNA** transfer ribonucleic acid.

transcription initiation signals at positions -35 and -10 upstream of the start of translation of the genes were identified, and shown to have sequence similarities to the *Escherichia coli* consensus sequences (51, 53, 54). However, the elements were A+T rich and not perfectly conserved, which make predictions of transcriptional start sites based on sequences problematic, due to the biased nucleotide composition in the genome.

Interaction between *Rickettsia* and the host cell during infection

The endothelial cells are the major host cells for *Rickettsia* during human infection, although the bacteria are capable of invading almost any cell type *in vitro* (55). The mechanism by which *Rickettsia* binds to and enters into the host cell is not known in detail. One suggestion is induced phagocytosis in which metabolically active *Rickettsia* cells trigger a microfilament-dependent phagocytic process in the host cell (56). The rate-limiting step in the binding and entry processes seems to be the adherence of the bacteria to the host cell (57).

Once inside, *Rickettsia* grows within the cytoplasm. Despite the fact that the host cell offers a very rich media containing many metabolites in high concentrations, the generation time for *R. prowazekii* has been measured to be as long as nine to twelve hours (58, 59). The SFG and *R. typhi* exit the host cells during the growth process, resulting in a continuous spread of the parasites to new uninfected host cells (57, 59). On the contrary, *R. prowazekii* grows inside the host cell until the cell bursts simply due to the load of the large number of bacteria inside. Hundreds of bacteria are escaping and are able to infect new host cells (58).

Although *R. prowazekii* contains only a single copy of the 16S rRNA gene (60) RNA and ribosome concentrations are very similar to those in *E. coli*, which have seven copies of the 16S rRNA gene and a generation time of 40 min (61). It has been speculated that the slow growth is an adaptation to maximise the number of bacteria from each infected host cell and to minimise the rate of host cell destruction. This would increase the probability that a louse obtains at least one parasite in a blood meal from an infected human, which would aid transmission to another human host (62).

Genome degradation

All parts of any genome are constantly mutated over evolutionary time. However, any part of the genome that provides a positive function for the host organism will be conserved through purifying selection, which remove the variants that destroy the function. On the other hand, if a stretch of DNA codes for a function that become unnecessary for the organism, the purifying selection acting on that part of the genome will disappear, and the stretch of DNA will be free to fix any mutations.

It was recognised early on that many bacterial genomes are densely packed with genes separated by short intergenic regions containing mainly regulatory elements. Indeed, all 15 completely sequenced bacterial genomes to date, except *R. prowazekii*, have a coding content close to 90% (63-77). From that perspective, the findings of pseudogenes[11] and long, apparently non-coding, intergenic regions that corresponded to 24% of the total genome length in *R. prowazekii* were unique (paper I, III).

This section will present how these pseudogenes and non-coding regions have been used to study reductive evolution and genome degradation in *Rickettsia*. Additionally, data from other bacteria, as well as organelles, will be presented, that indicate that the observed processes within *Rickettsia* may be general for a wide range of organisms, both intracellular and free-living.

The genus *Rickettsia* as a model for reductive evolution

Intracellular lifestyles have developed many times independently in different bacterial lineages (47). All these intracellular bacteria have small genomes compared to their free-living relatives, indicating that the free-living ancestor to *Rickettsia* most likely had a much larger genome than the modern *Rickettsia* genomes (table 1; 47). The exact genome size of the free-living ancestor is impossible to know, but since it must have had a metabolic repertoire similar to modern free-living bacteria a genome size of 4-5 Mbp and approximately 4000 genes may be a reasonable guess (48).

The order *Rickettsiales* is dominated by intracellular bacteria with small reduced genomes (78-80), indicating that the common ancestor of the *Rickettsiales* had started to adapt to an intracellular lifestyle. Consequently, the common ancestor of the TG and the

[11] **pseudogene** a functionless segment of DNA exhibiting sequence homology to a functional gene.

SFG already had lost many genes only needed for a free-living lifestyle. Nevertheless, *Rickettsia* is still losing genes, which indicates that the adaptation to the intracellular environment is an ongoing process (papers II-V). Therefore the genus is an ideal model for reductive evolution. Similarities between the TG and the SFG facilitate reconstruction of the genome repertoire in the common ancestor of the genus, while differences between members of the genus indicate recent evolutionary changes.

R. prowazekii is highly adapted to an intracellular lifestyle

The eukaryotic host cell is a rich and constant environment compared to many other ecological niches where bacteria are found. This leads to a decrease of the functional constraint acting on genes for biosynthesis of small molecules and regulatory functions for intracellular bacteria compared to free-living bacteria. Within all small derived bacterial genomes sequenced, genes for biosynthesis of small molecules that are easily imported from the host cell are absent, while most of the genes coding for functions in translation, transcription, replication and energy metabolism are present

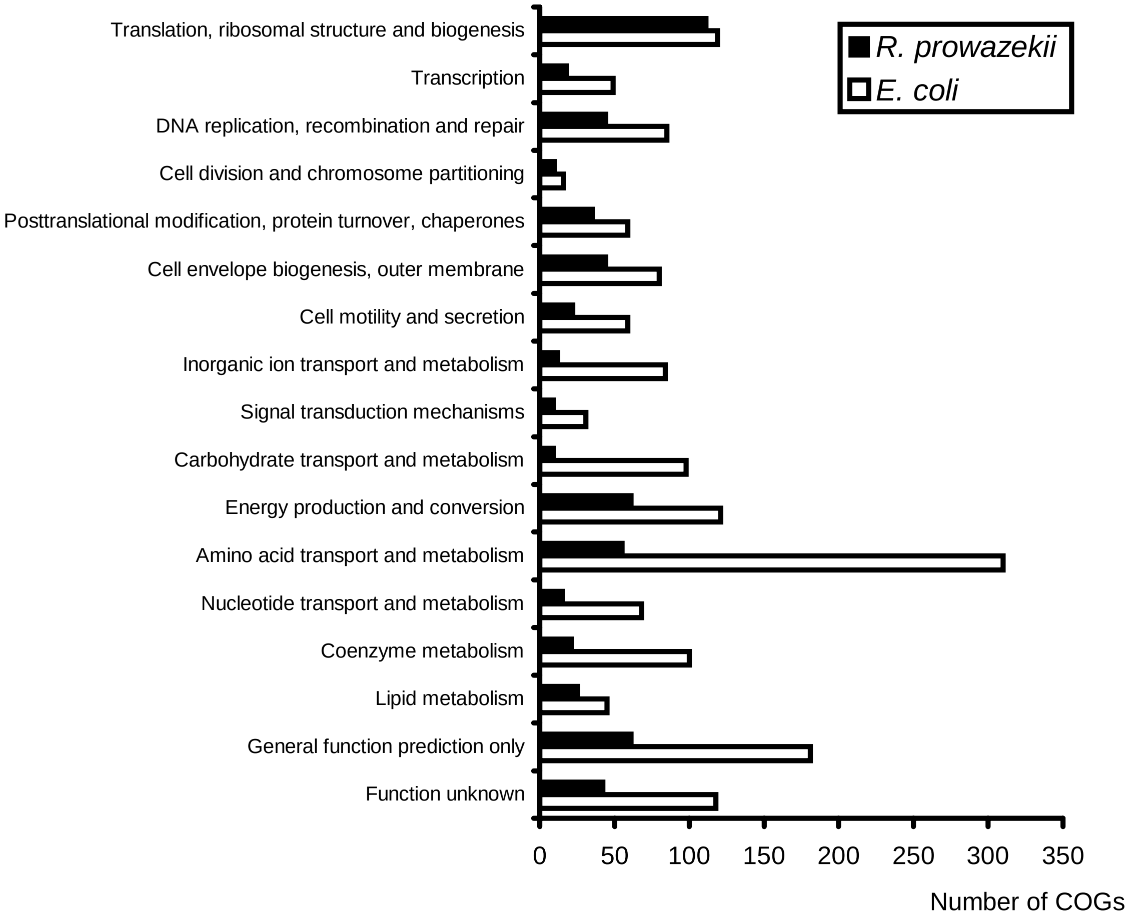

Figure 3. The distribution of COG members. The number of COGs with protein members in 17 different functional categories from *R. prowazekii* (filled bars) and *E. coli* (open bars). Data taken from (81-83).

(Figure 3; paper III; 64, 70, 72, 74, 76).

When two genomes from intracellular parasites, *R. prowazekii* and *Chlamydia trachomatis* were compared, striking similarities were identified (84). Firstly, both lineages have gone through a massive loss of genes for biosynthesis resulting in a very similar distribution of genes in the different functional categories, although the identities of the genes within each category are often different. Secondly, both lineages have recruited a transport system for ATP and ADP, which is unique to *Rickettsia*, *Chlamydia* and plastids (84, 85). Since these two parasites do not share an intracellular ancestor, the similarities observed are indicative of extensive, independent reductive co-evolution of the two unrelated lineages (84).

Genome rearrangements during reductive evolution

Comparisons of gene arrangements in bacteria reveal some operons that are conserved between distantly related species (86, 87). Several of these have been found in unique arrangements in *Rickettsia*. The single copies of the genes coding for rRNA are found in unusual arrangements in the *R. prowazekii* genome (60, 88), and the arrangements of the translational elongation factors genes, *tuf* and *fus*, suggest rearrangement between the two operons that normally harbour the genes (89). The *dnaA* gene is usually found close to the origin of replication in a conserved gene arrangement (90). Conversely, in both *R. prowazekii* (paper I) and *Wolbachia* (91) the gene is found in unique arrangements, which indicates that this genomic region has undergone multiple rearrangements within *Rickettsiales*. Taken together, all these rearrangements indicate that the transition from a free-living to an intracellular lifestyle is associated with the scrambling of genes in the genome of *Rickettsia*.

It has been found through comparative sequencing within the *Rickettsia* genus that all these rearrangement occurred before the formation of the genus (paper IV; H Amiri, UCM Alsmark, SGE Andersson, personal communication; 28). Together with the absence of gene order rearrangement in the 5% of the genomes that were sequenced for four different *Rickettsia* species (paper IV), this may indicate that the rate of gene shuffling decreases once the organism is adapted to an intracellular lifestyle.

Pseudogenes reveal different modes of ongoing gene inactivation and loss

Rickettsia may have lost 80% of the genes present in the free-living ancestor during the evolution. This massive gene loss should have left traces in the modern *Rickettsia* genome, especially if gene loss is an ongoing process. Early in our effort to sequence the *R. prowazekii* genome we could detect one such trace in the form of an internal termination codon in the *metK* gene, coding for AdoMet[12] synthetase (paper I).

The presence of a termination codon within an otherwise functionally conserved protein coding gene could be an intrinsic feature of the gene, and function, for example, as a regulation signal for programmed frameshifting (92). This has been observed within the *R. prowazekii* genome, where the *prfB* gene, coding for peptide chain release factor 2, contains a termination codon and frameshift in the same position as in the *E. coli* gene (paper III; 93). Another putative function for the termination codon is as an alternative code for incorporation of an unusual amino acid, which is known to occur for incorporation of selenocysteine into proteins in other organisms (94). A third mechanism by which the gene might encode a functional protein is by editing of the transcript at the mRNA[13] level prior to the translation, a phenomenon known to occur in a wide range of organisms (95). A fundamentally different plausible explanation of the internal termination codon in the *metK* gene is that it represents the first deleterious mutation, making the gene a pseudogene that is no longer under purifying selection.

The facts that the identity of the termination codon, TAG, is widely used in the *R. prowazekii* genes as a functional termination codon, and that the termination codon is located in a well conserved region of the protein, favoured the pseudogene explanation (paper I). That led us to initiate a comparative sequencing approach of the *metK* gene in different *Rickettsia* species, which showed that the *metK* had accumulated between 2 and 5 mutations destroying the reading frame in the *metK* gene in the SFG (Figure 2 in paper II). To determine if there was any purifying selection still acting on the hypothetical amino acid sequences, several different analysis of the non-synonymous[14] and synonymous[15] substitution rates were performed. All analyses indicated that the *metK* gene is not under purifying selection in the SFG (Figures 2-4 in paper II). Taken together, this strongly suggested that the *metK* genes are present as pseudogenes in the SFG. The most likely explanation for the termination codon in the *metK* sequence in *R. prowazekii* must be that it is the first sign of genic degradation, rather than an intrinsic feature of a functional gene (paper II). Another example of a pseudogene was found downstream of *metK* in the SFG. When the intergenic regions from the different SFG *Rickettsia* were aligned a number of indel[16] mutations could be inferred. ORFs[17] were

[12] **AdoMet** a methyl group donor in a wide range of reactrions in the cell, also known as S-Adenosyl-methionine (SAM).

[13] **mRNA** messenger ribonucleic acid.

[14] **non-synonymous substitution** a substitution that alters a codon to that for another amino acid.

[15] **synonymous substitution** a nucleotide substitution resulting in a codon specifying the same amino acid as before.

[16] **indel** insertion(s) and/or deletion(s).

[17] **ORF** open reading frame.

generated by removal of these indels, which showed all characteristics of functional *Rickettsia* ORFs (paper II).

A similar mode of transition from functional gene to pseudogene by fixation of random mutations throughout the genes was detected by comparative sequencing for three of the *R. prowazekii* pseudogenes (papers III, IV). The numerous unique pseudogenes found in the SFG in genomic regions that contained pseudogenes in *R. prowazekii* were also of this kind, with deletions, insertions and internal termination codons dispersed throughout their entire length (Figure 1 in paper IV). Similarly, the majority of the long intergenic regions in *R. prowazekii* showed sequence similarities to the genes or pseudogenes detected in paralogous genomic locations in the SFG, which indicated that they represented badly damaged pseudogenes (Table 2 and Figure 1 and 3 in paper IV). This indicated that the majority of the long non-coding regions of the *R. prowazekii* genome are remnants of inactivated genes in a late stage of degradation (Figure 3 in paper IV).

Four of the twelve putative pseudogenes were absent in the corresponding genomic location in the other *Rickettsia* species studied, in agreement with their pseudogene status in *R. prowazekii* (papers III, IV). The last four putative pseudogenes showed a more complex pattern. All are fragments of the *spoT* gene, encoding ppGpp[18] hydrolase (paper III). Combined two and two, these gene fragments roughly correspond to the synthetase and hydrolase activity of the protein, while the missing part corresponds to a part with regulatory function (paper IV; 96, 97). The *spoT* fragments seem to be under purifying selection, but are associated with length variation and loss (paper IV). This may represent a different mode of gene inactivation where the regulatory part of a multifunctional protein is lost first, giving a truncated unregulated protein that is able to perform the function at some rate. By rigorous studies of the *R. prowazekii* proteome[19] it might be possible to find more examples of this kind of intermediate between a functional gene and a pseudogene. However, the fragmentation and duplication of the gene and the absence of a proper initiation codon in some fragments (paper IV) are puzzling and make the *spoT* sequence data very difficult to interpret in the absence of functional studies.

[18] **ppGpp** guanosine 5'-diphosphate 3'-disphosphate, a compound that accumulates during starvation and inhibits a number of biosynthesis pathways.

[19] **proteome** the complete set of proteins encoded in the genome.

Parallel losses of *Rickettsia* specific genes

37% of the putative protein coding genes in the *R. prowazekii* genome could not be assigned to any function (paper III). Many of these are expected to encode proteins that perform *Rickettsia* specific functions, such as virulence. From this point of view, it is interesting that the majority of the proteins that have been inactivated and/or lost in at least one of the *Rickettsia* species cannot be assigned to a function, or have a weak functional prediction (Table 2 in paper IV). These genes may have been recruited by the ancestor of *Rickettsia* for functions in response to the host cell, with subsequent differential loss in different lineages in the adaptation process to specific hosts and vectors.

The rule seems to be that each gene is inactivated several times independently in different *Rickettsia* lineages rather than once in evolution (paper II, IV). The multiple inactivation events of the *metK* gene and its downstream ORF are good examples of this phenomenon (paper II). Similarly, for several of the inactivated genes in the four-species dataset multiple inactivation events have to be inferred to explain the observed patterns of deleterious mutations (Figure 1 in paper IV). A compensating event, for example the invention of an import system that happened in the common ancestor of the lineages is one explanation for the occurrence of parallel losses. Both systems were then working in parallel until the new system was good enough to fully compensate for the gene. Speciation occurred during this time and the gene was lost independently in the two lineages.

Mutational patterns and the fate of inactivated genes in the *Rickettsia* genome

Studies of neutrally evolving pseudogenes showed that there exists a mutational bias in the *Rickettsia* genome for transitions over transversion. Among the transitions there is a bias towards A and T nucleotides, similar to both *Drosophila* and mammals (Figure 4; paper II; 98-100), which will result in an enrichment of A+T in a sequence without functional constraint. Indeed, the non-coding part of the *R. prowazekii* genome has a G+C content of 23.7% compared to the overall G+C content of 29.1% (paper III). Similarly, in *Drosophila* the G+C contents of introns are decreased compared to the exons (100). Among the indel mutations a strong bias for deletions over insertions, both in size and number, were observed (papers II, IV, V). Again, this is comparable to the available data from *Drosophila* and mammals (Table 2; 101-103).

The *Rickettsia* genome is constantly inactivating genes, and the resulting pseudogenes will fix mutations in a random manner (papers II-V). In principle, it is possible to calculate the rate of disappearence of a pseudogene by fixation of random deletions according to a continuous decay formula (Table 2; 102, 104). It was calculated that a gene in the ancestor of the SFG and the TG that was inactivated at the time for the divergence between the groups has decreased to 11% of the length of the functional gene in the modern *Rickettsia* genome by fixation of random deletions. The G+C enrichment process of an inactivated gene can be calculated according to a formula for sequence amelioration derived by Lawrence and Ochman (paper IV; 105). Accordingly, the remaining fragment of a gene that had a G+C content of 30% at the time of the inactivation in the common ancestor of the two groups is calculated to have a G+C content of 28% in the modern *Rickettsia* genome (paper IV).

However, there are indications that the ratio between indel and point mutations are higher in the SFG than in the TG (Table 3 in paper IV; see discussion below), which would lead to maintenance of pseudogenes for a longer time in the TG. Additionally, large statistical fluctuations are expected since the datasets are small and the presence of large deletions will strongly dominate over frequent short deletions in the estimation of mean deletion size, even if they are very rare. In any case, these rough calculations do not disagree with the occurrence of pseudogenes and long non-coding regions in the different *Rickettsia* lineages (paper II, IV). Indeed, the observed differences of the rate of indel mutations between *Drosophila* and mammals have been used to explain the

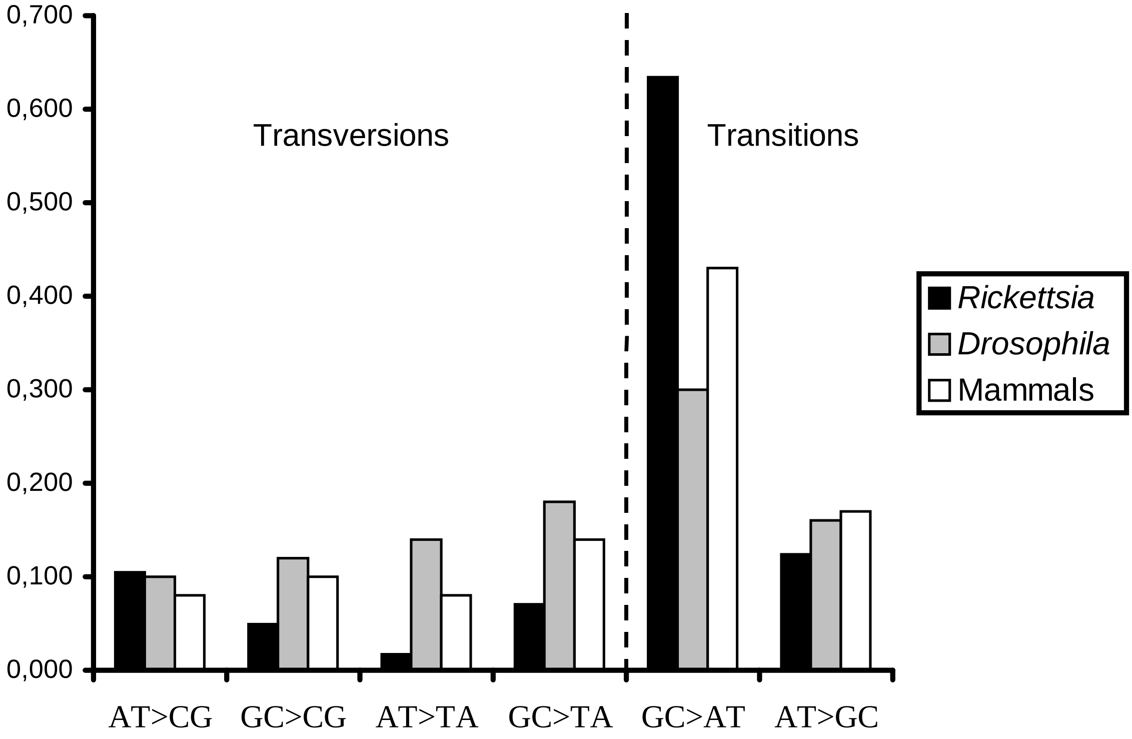

Figure 4. The relative rates of nucleotide pair substitutions. The complementary nucleotide substitutions are grouped. For example, AT>CG indicate the sum of the frequencies of A to C and T to G substitutions. Data taken from (paper II; 98-100).

Table 2. Indel evolution in *Rickettsia*, *Drosophila* and mammals.

	Rickettsia[1]	*Drosophila*[2]	Mammals[3]
Ratio of insertions to point mutations	0.055	0.015	0.010
Ratio of deletions to insertions	3.3	8.7	4.7
Ratio of deletions to point mutations	0.18	0.13	0.049
Mean deletion size (bp)	68.9	24.9	3.2
Mean insertion size (bp)	1.1	2.9	2.5
Half life of a pseudogene (point mutations per nucleotide)[4]	0.056	0.21	4.42
Half life of a pseudogene (Myr)	7-14[5]	14.3	884

[1] Data are from (paper II).
[2] Data are from (101, 102).
[3] Data are from (103).
[4] Calculated using a continuous decay formula: $L = L_0e^{-rt}$, where L is length, r is deletion rate (product of ratio of deletions to point mutations and mean deletion size), and t is time (in point mutations per nucleotide).
[5] Substitution rates of *Buchnera* used (108).

differences in pseudogene frequency, intron sizes and genome sizes in the two lineages (101, 102, 104, 106, 107).

Although the exact rates of the processes remain to be determined, it is clear that an inactivated gene in the *Rickettsia* genomes will decrease in G+C content and accumulate small deletions (Table 2 and Figure 4). Both processes will blur the traces of the gene. If the region escapes large deletions the blurring process may go as far as to removal of all traces of a functional gene, even if a closely related orthologous gene is known. For example, some of the unique genes found in the SFG could only be recognised in the non-coding regions in the TG through weak BLAST hits (Table 3 in Paper IV). From a genomic perspective, the high fraction of non-coding DNA in the *Rickettsia* genome is the result of the balance between the rate of which genes are inactivated and the ratio between insertions and deletions (papers II-V). Currently, *Rickettsia* inactivates genes at a much higher rate than they are eliminated from the genome by deletions (papers II, IV, V). A stronger selection on genome size, a higher intrinsic deletion mutation rate, and/or a slower rate of which genes are inactivated would make the genome richer in coding regions.

Gene loss and intergenic region differences between the two *Rickettsia* groups

As presented in the first section there are many differences within the *Rickettsia* genus regarding lifestyle, pathogenicity, genome size, and base composition. These differences are expected to have shaped the individual genomes somewhat differently.

The most obvious difference between the genomes from the different groups is the number of genes and pseudogenes. The corresponding regions to the regions harbouring the twelve pseudogenes detected in the complete *R. prowazekii* genome were sequenced in three other *Rickettsia* species. The fraction of the genomic DNA that corresponds to recent pseudogenes was highest in *R. rickettsii*, although only approximately 5% of the genome were sequenced (Table 1 in paper IV). This strongly suggests that recently inactivated pseudogenes are far more frequent in the SFG than in the TG, while intergenic regions corresponding to more anciently inactivated genes are more frequent in the TG (paper IV). A rough calculation indicated that the common ancestor of the two groups maybe contained 300 genes that are missing in the modern *R. prowazekii* (paper IV). The majority of these were inactivated early in the branch leading to the TG, and are now totally lost or present as intergenic regions with weak homologies to SFG genes. On the other hand, most of these 300 genes remained under purifying selection for a longer time in the SFG, and many are still present as functional genes (paper IV). Most likely the species in the SFG contain more protein coding genes than *R. prowazekii*, in agreement with the larger genome sizes (Table 1; paper IV). Nevertheless, some genes still present in the *R. prowazekii* genome are expected to have been lost in the SFG (papers II-IV).

The *R. prowazekii* genome contains unusually few repeats (paper III). In contrast, the intergenic regions of the SFG contain a high frequency of short G+C rich repeats, often in the form of palindromic sequences (papers II, IV). The repeats are most likely mobile within the SFG genomes, since one 6-bp repeat motif found in the SFG are identical to a described *Neisseria* repeat motif, and a long palindromic repeat containing this motif was detected within an insertion in the *dnaA* gene in *R. felis* (paper II).

There is a small but significant difference in the estimated G+C contents between the TG and the SFG (Table 1), which is confirmed by the partial genomic sequences (papers II, IV). Unfortunately, all the data on point substitutions comes from the SFG (paper II), which make direct comparisons of the mutational biases impossible. However, for 43 orthologous genes, the average G+C content values in the third codon positions within genes were 18.8% for the TG and 23.2% for the SFG (paper IV), which indicate that there must have been a change in mutational patterns after the divergence of the two groups. Thirteen indel mutations and 14 termination codons are found within the pseudogenes in the TG, while 36 indel mutations and 8 termination codons are found in the SFG (Table 3 in paper IV). Furthermore, there seems to be an equal number of insertions and deletions within the indel mutations in the TG, while deletions are twice as frequent compared to insertions in the SFG. Contrary, the average size of deletions is much larger than the average size of insertions in both groups (Table 3 in paper IV). Although the numbers are small and statistical fluctuations are expected to be large, there seems to be a difference in the mutational patterns between the two groups.

Anyway, if there is a significant difference, it will have implications on how long pseudogenes will remain in the different genomes, as discussed above.

Unfortunately, it is difficult to connect the observed differences between the individual species to the differences in lifestyle and pathogenicity. The problems are that many of the genes that differ are ORFs without predicted function (paper IV), and that too little is known about the different cell biology of the individual species to be able to make any hypothesis.

Pseudogenes as indicators of genome degradation in bacteria

The applicability of the study of pseudogenes for the study of genome degradation was described in the previous section. It should be possible to deduce how general the processes observed within *Rickettsia* is, by studying pseudogenes in other bacteria. An increasing amount of pseudogenes are being discovered in bacteria (Table 3), which will be presented in this section.

Pseudogenes can loosely be divided into two groups: (i) ancient genes in the genome that have become pseudogenes because the function of the gene products are no longer needed due to change of lifestyle, and therefore the purifying selection is relaxed, and (ii) genes that have recently been introduced into the genome, either by horizontal gene transfer[20] or duplication of ancient genes, and are no longer needed. *Rickettsia* pseudogenes described above are of the first category (papers II-V), while most of the eukaryotic pseudogenes described are of the second type (109-111).

Pseudogenes creating permanent loss of function

The most straightforward example of pseudogenes that create permanent loss of function comes from *Lactococcus lactis* (112, 113). It was found that *L. lactis* strains isolated from diary products were auxotrophs for branched-chain amino acids, while most strains isolated from non-dairy sources were prototrophs. Sequencing of the *leu* genes from an auxotroph revealed 99% sequence homology to the prototroph and the presence of two internal termination codons and two small deletions (113). Although the *leu* genes appear to be transcribed and regulated similarly in the auxotroph and prototroph the presence of deleterious mutations most likely explains the auxotrophy (113). In an accompanying report similar results were presented for the *his* genes in *L.*

[20] **horizontal gene transfer** transfer of genetic information between different species (same as lateral gene transfer).

Table 3. Pseudogenes in bacteria.

Species	Pseudogenes	Location	References
Lactococcus lactis	*leu* and *his* operons	chromosome	[1]
Mycobacterium leprae	39 pseudogenes	chromosome	[2]
Yersinia pestis	*gumO, ompF, bvgS, hha, malK, inv*	chromosome	[3]
Yersinia pestis	*yopA, ylpA, yadA*	plasmid	[4]
Neisseria	*porA, opcA, orfC, orfE, orfD, comEA*	chromosome	[5]
Buchnera sp.	*trpEG, repAC*	plasmid	[6]
Borrelia burgdorferi	167 pseudogenes	plasmid	[7]

References: [1] (112, 113); [2] (114, 115); [3] (116, 117); [4] (118, 119); [5] (120, 121); [6] (122-124); [7] (125).

lactis. However, mutations in the promoter region of this operon were detected, as well as a decreased level of expression (112). The observations from the *L. lactis* are best explained by the fact that these two amino acid biosynthesis pathways became redundant once the strains adapted to the rich environment of milk products. The amino acids became constantly available in the surroundings, and the purifying selection acting on amino acid biosynthesis operons was lost (112, 113).

As many as 39 pseudogenes were found within the 66% that were sequenced from the 2.8 Mbp genome of the obligate intracellular parasite *Mycobacterium leprae* (114). The pseudogenes corresponded to as much as 3.5% of the possible coding regions and each pseudogene contained at least three, and on average nine, frame-destroying mutations (Table 1; 114). In a preliminary report the 90% complete *M. leprae* genome was compared with the completely sequenced and 1.6 Mbp larger genome of the closely related *Mycobacterium tuberculosis*. Of the 13 genes present between *rif* and *str* operons in *M. tuberculosis* only two are found as intact genes in *M. leprae,* while the other genes were absent or found as pseudogenes in the *M. leprae* genome (71, 115). Overall, the patterns of genome degradation seem to be similar in the obligate intracellular parasites *M. leprae* and *R. prowazekii* (paper IV; 71, 115).

Several pseudogenes have been detected in *Yersinia pestis*, the causative agent of plague, which has been responsible for millions of deaths during history. Sequencing of a 119-kb part of the *Y. pestis* chromosome covering the *pgm* locus and its flanking regions identified three genes that contained frameshift mutations and two genes interrupted by insertion elements (116). The pseudogenes identified were non-functional in all examined *Y. pestis* strains, while they appeared functional in all strains of the much less virulent, but closely related species *Yersinia pseudotuberculosis*. Several of the putative functions of these *Y. pestis* pseudogenes as well as two earlier described *Y. pestis* pseudogenes, *inv* and *yopA* (117, 118) are related to virulence (116). Two more pseudogenes, both with unknown function, have been found on the Low-Ca^{2+}-response plasmid (pCD1) of *Y. pestis* (119). It has been speculated that the enhanced virulence of *Y. pestis* is a response to the inactivation of these genes, since many of the *Y. pestis* pseudogenes are present as functional genes in closely related and less virulent *Yersinia* species (116, 118). When a functional *yopA* gene was introduced into *Y. pestis* the virulence of the strain was reduced (126). Similar observations have been made in *Shigella* spp. where the introduction of the *E. coli* gene *cadA* attenuated virulence (127). These observations suggest that the transition process to a highly virulent pathogen not only requires the introduction of genes in the form of pathogenicity islands (128), but the inactivation of genes detrimental to a pathogenic lifestyle may also be necessary (116, 127).

Functionally redundant pseudogenes within the genome

The detection of a pseudogene within a genome is not necessarily the consequence of a change in the lifestyle of the organism. It may also be the result of the inactivation of a gene that was very recently fixed in the genome. The source of the gene was either duplication within the genome, or horizontal gene transfer of a gene from another species.

The finding of a *porA* pseudogene in the *Neisseria gonorrhoeae* genome was unexpected since the gene is absent in seven closely related *Neisseria* species (121). Studies of *opcA* and ψ*opcB* genomic regions in three *N. gonorrhoeae* and *Neisseria meningitidis* strains revealed a complex evolutionary pattern with insertion elements, insertions, deletions and termination codons within genes. In total five pseudogenes could be detected in different strains (Table 3) which all were located on a DNA island that probably had been imported from unrelated bacteria (120). Both these publications indicated that the pseudogenes in *Neisseria* are the results of the inactivation of recently acquired genes (120, 121). However, this is not surprising given the frequent transfer of genetic information that has been shown within individual *Neisseria* species (129-131), between species within the genus (132-134), and even from distantly related species, such as *Haemophilus* (135).

The role of *Buchnera*, an obligate intracellular symbiont in aphids, is to supply the host with essential amino acids (136, 137). The capacity for amino acid synthesis has been increased in order to strengthen the symbiotic relationship. Several genes involved in amino acid synthesis have been incorporated in multiple copies on plasmids (124, 138-140). The tandem repeats of the plasmid-borne *trpEG* genes have been found to consist of one functional copy and a pseudogene in two *Buchnera* species (122, 123). Recently it was found that the gene coding for the replication initiation protein, *repAC*, is also present as a pseudogene on the *trpEG* plasmid in *Buchnera aphidicola* from aphids of the family Pemphigidae (124). The congruency of the phylogenetic trees between the plasmid-associated *trpE* gene and the chromosomal *trpB* gene indicate an ancient origin of the *trpEG* containing plasmids with no subsequent exchange of plasmids between different *Buchnera* species (123, 124). The authors speculate that the plasmid gene inactivations may be due to changes in the exogenous amino acid supply in combination with a lack of a mechanism for reducing the number of gene copies (122, 123).

The Lyme disease spirochete, *Borrelia burgdorferi*, contains one linear chromosome, twelve linear plasmids and nine circular plasmids (70, 125). The 21 plasmids were predicted to encode 535 functional genes and 167 pseudogenes. Less than 10% of the functional plasmid genes show sequence homology to genes outside of *Borrelia*. There is a high sequence similarity between some linear plasmids (125), which is known to facilitate recombination between *Borrelia* plasmids (141). These recent DNA rearrangements have created regions of high similarity, which harbour genes of paralogous gene families (125). Many of the members of these families contain internal termination codons and/or frameshifts compared to the homologous proteins, which indicate that they are no longer under purifying selection. A typical example of a paralogous gene family contains the intact genes BBE02 and BBH09 from two different linear plasmids and four badly damaged paralogs on other linear plasmids (125). 87% of the 168 pseudogenes found in the complete *B. burgdorferi* genome are located on ten linear plasmids, which have an average coding content of only 41%. The other 11 plasmids and the chromosome have few pseudogenes and coding contents close to 90% (125). It was speculated that *B. burgdorferi* is currently undergoing a rapid evolution that is manifested by a high rate of apparently random duplications. After the duplication event the selection is lost for one copy of the gene, which then will be present as a pseudogene and be degraded (125, 142).

Genome sequences and genome degradation

A majority of the publications of complete genome sequences did not report any observations of ongoing genome degradation in the form of pseudogenes. On the other hand, a high frequency of horizontal gene transfers has been seen in several bacterial genomes. Since the bacterial genome sizes are fairly constant over evolutionary time, genome degradation should be at least as frequent as horizontal gene transfer. Comparisons of genome sequences from closely related strains and species offer unique insights into the gene turnover rates.

Genome evolution - horizontal gene transfer

4288 protein coding genes were annotated in the complete genome of *E. coli* K-12 (69). Based on the nucleotide frequencies and codon usage biases it has been suggested that as many as 755 of these have been introduced into the *E. coli* K-12 genome since the divergence from the *Salmonella* lineage 100 millions years ago, in an ongoing process that introduces 16 kb of foreign DNA per million years (105, 143, 144). The genome sizes among different isolates of *E. coli* range from 4.5 to 5.5 Mb. The heterogeneity has been generated by multiple changes throughout the genome rather than by a few large duplications, which supports a rapid exchange of DNA in the species (145, 146). Identification and sequencing of strain-specific accessory DNA showed that it was mostly of exogenous origin, presumably introduced by horizontal gene transfer (147).

Massive gene transfers were also inferred from the genome sequence of *Thermotoga maritima*, a thermophilic bacteria for which 24% of the genes were most similar to archaeal genes (77), and several analyses based on phylogenetic reconstructions have identified widespread gene transfers between distantly related microorganisms (148, 149). It is obvious that horizontal gene transfer is a widespread phenomenon, although the rate variations between different categories of genes and organisms remain to be worked out (144, 150).

Compared to the examples described above, the *R. prowazekii* genome showed strikingly few horizontal gene transfers, only a couple of tRNA synthetase genes and few other genes showed phylogenetic patterns that indicated transfer events (paper III; B Canbäck, personal communication; 151). Furthermore, no recent horizontal gene transfer could be detected among the 5% of the genome that were sequenced in three additional *Rickettsia* species (paper IV). This is not surprising since the process of horizontal gene transfer must involve contact with foreign DNA and the lifestyle of *Rickettsia* offers few opportunities to mix and mingle with other bacteria. On the other hand, host cell DNA should be available in the surroundings, and incorporation of host cell DNA into the *Rickettsia* genome should be possible. In fact, 16 rickettsial genes showed the strongest similarity to eukaryotic genes (152). In the absence of well-sampled phylogenetic trees based on the putative horizontally transferred genes it is difficult to estimate the contribution of host-cell to parasite transfer in the genome evolution of *Rickettsia*. However, a comparison of the available data for *Rickettsia* and *E. coli* suggests that horizontal gene transfer is far more frequent in free-living *E. coli* than in the obligate intracellular parasite *R. prowazekii*.

Comparison of the estimated rate of horizontal gene transfer and the genome size in *E. coli* indicates that the horizontal gene transfers do not lead to a significant genome size expansion over evolutionary time (143, 144, 146). As a consequence, genome degradation has to be as frequent as horizontal gene transfer. The process of genome degradation in organisms that experience extensive horizontal gene transfer should be possible to detect by comparative sequencing, as it has been for *Rickettsia* (papers II, IV). The amount of ongoing genome degradation that is possible to be detected will depend on the rate of gene inactivation, the intrinsic insertion and deletion mutation rates and the selection pressure on genome size. Probably, the main reason for the absence of pseudogenes in *E. coli* is a strong selection on genome size, forcing the inactivated *E. coli* genes to disappear from the genome much quicker than in *Rickettsia* (see discussion on page 30). The *Neisseria* genome, on the other hand, contains several pseudogenes that most likely represent inactived genes that were recently introduced into the genome (Table 3; 120, 121). *Neisseria* may be an example of a genome where horizontal gene transfer is frequent, but selection on genome size is less pronounced than for *E. coli*.

Comparison of closely related bacteria show fast gene exchange

It is possible to detect recent evolutionary events by comparison of genomes from closely related species, since all observed differences must have occurred in the lineages during the evolutionary time since the divergence of the species. Currently there are three bacterial genera from which genome sequences of two members are published; *Mycoplasma*, *Helicobacter* and *Chlamydia* (Table 4; 64, 66, 68, 74-76).

Comparison of the two *Mycoplasma* species showed that all genes present in the *Mycoplasma genitalium* genome were present in the larger *Mycoplasma pneumoniae* genome. The *M. pneumoniae* genome codes for an additional 209 putative genes, of which 110 are unique and the rest are duplications of genes present in the *M. genitalium* genome (Table 4; 153). Since the divergence of the two species, *M. genitalium* has lost many genes and *M. pneumoniae* has duplicated many genes, creating a genome size difference of 236 kb. During this process the gene orders of the orthologous gene pairs in the two genomes were well conserved, as indicated by the identification of six large segments of highly conserved gene orders between the two species (153).

Both completely sequenced *Chlamydia* species, *Chlamydia pneumoniae* and *C. trachomatis*, are obligate intracellular parasites that cause diseases in humans. The *C. pneumoniae* genome is 187 kb larger than the *C. trachomatis* genome and contains 214 unique coding sequences. Only 28 of these, and only 10 of the 70 genes unique to *C. trachomatis*, show similarity to genes in the public databases (Table 4; 76). Gene order

Table 4. Distribution of ORFs in three pairs of closely related bacteria.

Species and strains	Number of ORFs	Unique ORFs[a]	Without homology[b]	Reference[c]
Mycoplasma genitalium	470	-	-	1
Mycoplasma pneumoniae	677	110	77 (70%)	1
Chlamydia trachomatis	894	70	60 (86%)	2
Chlamydia pneumoniae	1.073	214	186 (87%)	2
Helicobacter pylori 26695	1.552	117	91 (78%)	3
Helicobacter pylori J99	1.495	89	56 (63%)	3

[a] Number of ORFs without a homologue in the closely related species or strain.
[b] Number of the unique ORFs without a homologue in the public databases, percentage in parenthesis.
[c] References: [1] (153); [2] (76); [3] (75)

analysis of the two *Chlamydia* species indicated that the overall gene order is conserved with unique genes interspersed between orthologous genes, with only a few large rearrangements (T. Sicheritz-Pontén, personal communication).

Another example of a comparison of two closely related complete genome sequences comes from *H. pylori*, for which the strains 26695 and J99 are sequenced (68, 75). Although both sequences are from the same species 89 genes unique to strain J99 and 117 genes unique to strain 26695 could be detected, compared to 1406 orthologous genes found in both strains. A significant homologue could be found for only 59 of the 206 strain specific genes (Table 4; 75). The overall arrangement of the orthologous genes was conserved, with only a few major rearrangements. Half of the unique genes clustered in a hypervariable region. The authors also briefly mentioned that some genes appeared to contain a frameshift in one of the two strains (75).

The *Mycoplasma*, *Chlamydia*, *Helicobacter* and *Rickettsia* show the striking similarities that they seem to undergo genomic evolution where losses of genus specific genes play an important role. The overall genome organisations, on the other hand, seem to be stable with only a few major rearrangements observed (paper V; T Sicheritz-Pontén, personal communication; 75, 76, 153). This may indicate that these species are in a late state of reductive evolution where most of the genes that can be lost easily are already gone. The differences in gene content observed within the genera are to a large part genes associated with the specific virulence of the strain or species.

Organellar genomes - window into the future?

The mitochondria descend from within α-proteobacteria (Figure 1) and have, similarly to *Rickettsia*, highly reduced genomes. However, current data indicate that the mitochondria and *Rickettsia* have undergone independent reductive evolution, and that the bacterial ancestor of mitochondria most likely had a much larger genome than modern *Rickettsia* (paper III; 154, 155). Extensive reductive evolution has also occurred and is currently occurring in the evolution of chloroplasts from their cyanobacterial ancestor (156). This section will present the current knowledge of genome degradation in organellar genomes.

Extensive parallel gene losses and gene transfers in organelles

All mitochondria are descendants of a single endosymbiotic event between a host cell and an α-proteobacteria, as indicated by both gene content (157), gene arrangement (paper III; 157, 158) and phylogenetic data (paper III; 15, 16). Similarly, all plastids are thought to be descendants of a single cyanobacterial ancestor which formed a symbiotic relationship with an eukaryotic cell, also based on gene content data (159) and phylogenetic reconstructions (156).

Both plastids and mitochondria seem to have undergone a fast extensive gene loss after the establishment as symbionts. A comparison of the completely sequenced chloroplasts could trace 205 genes to the common ancestor of these genomes, compared to at least 2000 genes in the cyanobacterial ancestor of the chloroplasts (159). In the case of mitochondria, the 97 genes found in the freshwater protozoon *Reclinomonas americana* mitochondrial genome may represent the gene complement of the common ancestor of all mitochondria, since all other sequenced mitochondria contain a subset of these genes (157, 158, 160, 161).

A pattern where parallel gene losses in different lineages are the rule rather than the exception emerged, when gene content of the mitochondria and chloroplasts were mapped onto the respective phylogenetic tree (157, 159, 161). For example, of the 205 genes in the common ancestor to the chloroplasts, 46 were present in all nine genomes analysed, 58 genes were lost once, 44 genes were lost twice, 43 genes were lost three times and 14 genes were lost as much as four times in independent lineages (159).

Ongoing genome degradation in organelles

The loss of a gene from an organelle does not necessarily mean that the information is lost altogether, it may also indicate that the gene has been transferred to the nucleus and that the gene product might be imported back to the organelle. Indeed, the close phylogenetic relationship between a subset of nucleus-encoded genes and eubacteria usually are interpreted as indications of transfers from organelles to the nucleus (162, 163). A successful transfer includes at least three distinct phases: (i) physical transfer of the gene to the nucleus, (ii) activation of the nuclear copy of the gene and inactivation of the mitochondrial copy, and (iii) physical loss of the mitochondrial gene.

The gene *coxII* was transferred to the nucleus before the divergence of five phylogenetic groups during the evolution of flowering plants (164). In four of these groups, the nuclear copy of *coxII* is expressed and the mitochondrial copy is silenced, while the fifth still uses the mitochondrial copy. Of the four groups that express the nuclear copy, two have lost the mitochondrial copy (164). RNA editing, a widespread phenomenon among land plant mitochondria (95), can be utilised for detection of the source of a transferred gene. Transfer via mRNA seems to be the most common mechanism for transfer of genetic information from the mitochondria to the nucleus, as inferred from the presence of edited copies of the mitochondrial genes in the nucleus (164-168). The nature of the ongoing process is indicated by the liverwort *Marchantia* and seed plant *Arabidopsis*, where genes shown to be present in the nucleus are present as pseudogenes in the mitochondria (167, 168). For the *rps19* gene in *Arabidopsis* it was also shown that the transferred copy was expressed in the nucleus, and that the protein actually was imported to the mitochondria (167).

An interesting example of genome degradation in organelles comes from the plastid genome of *Epifagus virginiana*, a nonphotosynthetic parasitic plant living on the roots of beech trees. Only 42 functional genes could be found in the 70-kb genome, of which 38 specified components of the gene-expression apparatus of the plastid (169). Compared to *Nicotiana tabacum*, *E. virginiana* has lost all photosynthetic genes, six ribosomal protein and 13 tRNA genes, of which some were present as pseudogenes. An increased rate of molecular evolution could be detected in the retained ribosomal protein and tRNA genes (169, 170). At least one gene in the genome must encode a protein with an essential non-genetic function. The maintenance of this function requires maintenance of functional replication, transcription and translation systems in the plastid (169). Sequencing of the tRNA genes of *Orobanche minor*, which share a nonphotosynthetic ancestor with *E. virginiana*, identified nine shared intact tRNA genes. Most of the tRNA species that were conserved in nonphotosynthetic plants have never been shown to be imported to plant mitochondria, which suggests that they are preserved in the chloroplast genome because the encoded tRNA cannot be imported into the chloroplast (171).

Photosynthesis has been lost at least five times independently within the monophylogenetic group of parasitic plants that *Epifagus* belongs to. In different nonphotosynthetic lineages, the photosynthetic gene *rbcL*, encoding RuBisCo[21], was present as either an intact gene, a pseudogene or was totally deleted (172). A detailed analysis of individual amino acid changes in the *rbcL* genes revealed that the pseudogenes no longer were under purifying selection. However, the putative amino acid sequences from intact *rbcL* genes from non-photosynthetic species seemed to be under purifying selection, indicating that RuBisCO may have a non-photosynthetic function in these lineages (173).

Highly derived organellar genomes

Genome degradation is an ongoing process in organelle genomes, and there does not seem to be any limit to how degenerate these genomes may become. All mitochondrial genomes can be divided into two groups, ancestral and derived (161). The *R. americana* genome is the best example of an ancestral genome with many genes preserved compared to the derived genomes, a standard genetic code, an eubacterial-like gene clusters in a tightly packed genome and a complete or nearly complete set of tRNA genes (158, 161). Derived genomes, on the other hand, may show extensive gene loss, accelerated evolutionary rate, non-standard genetic code and RNA editing (161). The primitive multicellular animal *Dicyma* represents an extremely derived genome where three mitochondrial genes are encoded on three distinct circular DNA molecules of 1.6 to 1.7 kb length (174). A similar organisation has been reported from the chloroplast of the dinoflagellate *Heterocapsa triquetra*, where two ribosomal RNA genes and seven protein coding genes were each shown to be present on a small minicircle (175, 176).

At least seven photosynthetic groups have derived their plastids by uptake of a whole plastid-containing cell, a phenomenon known as secondary endosymbiosis (177). Two of these groups have retained remnants of the nucleus of the secondary symbiont as a nucleomorph (178). Genetic analysis of the nucleomorph showed a stunning example of miniaturisation of an eukaryotic genome with very small introns, very short intergenic regions and co-transcription of genes (178-180).

However, the example of the most reduced genome may come from the hydrogenosome, a hydrogen producing organelle that might be a modified mitochondrion (181, 182). Until recently, it was believed that hydrogenosomes lacked a genome altogether, but a recently found hydrogenosome with a genome (183) may clarify the origin of this strange organelle.

[21] **RuBisCo** ribulose-1,5-biphosphate carboxylase/oxygenase, the carbon dioxide fixation protein in plants.

The order and driving forces of reductive evolution

The presence of a pseudogene in a genome indicates that the organism can survive without that gene in a functional state. However, all genes in a genome do not have the same probability to become redundant. In addition, the same gene does not have the same probability to be degraded in all organisms. This section discusses the order of gene inactivations during the evolutionary transition from free-living to intracellular, as well as the evolutionary driving forces behind the process.

The evolutionary order of gene inactivation

Each gene in an organism could, in principal, be categorised according to how easy it is made redundant by the effect of the environment. During the evolutionary transition from a free-living to an intracellular lifestyle different categories of genes will therefore be lost at different phases of the adaptation process.

Among the categories of genes that are easily made redundant are the genes coding for biosynthesis of small molecules, which are easily taken up by the bacterium. For example, a constant supply of an amino acid makes the genes coding for that amino acid redundant and subject to genic degradation, as described earlier for *Lactococcus* (112, 113). With a few exceptions, all genes coding for proteins involved in amino acid biosynthesis are lost from the *Rickettsia* genome (paper III), and no traces of such genes are found among the pseudogenes detected in the comparative sequencing projects (papers II, IV). This is reasonable since amino acid genes most likely were lost early in the adaptation process to the intracellular environment.

All metabolites are not easily imported into the bacterial cell. To utilise these metabolites, specific transport systems have to be invented. When these systems are in place they may make some genes in the bacterial genome redundant. *metK*, coding for AdoMet synthetase, may be an example of this scenario. *metK* is present in all sequenced bacterial genomes, except in the genomes of the intracellular parasites. The gene is lost from *Chlamydia* (74, 76), and is disappearing from the *Rickettsia* genome (paper II). Since there are many genes encoding enzymes that utilise AdoMet present in these genomes, the biological basis for the losses are most likely a consequence of a specific import system for AdoMet (papers II, V).

Genes coding for proteins that are involved in the transfer of information will remain for a long time in the genomes, since the import of functional enzymes from the host cell is needed to replace them. Most likely the imported enzymes have to be of the bacterial type, since the enzymes for these processes often are dissimilar in bacteria and eukaryotes. This would require a transfer of the bacterial gene to the host nucleus, in a similar way to the transfer of organellar genes to the cell nucleus in eukaryotes (164-168). However, for *Rickettsia* it seems unlikely to occur, since the same gene has to be transferred independently to the different hosts in the lifecycle.

Muller's ratchet acting on organelle genomes

Intracellular bacteria and organelles experience gene loss as an ongoing process. For many genes, the loss can be explained by the neutralising effect of the host genome that makes the genes redundant. However, both organelles and intracellular bacteria live inside host cells and experience frequent bottlenecks during transition from one host to the next, with little opportunity for recombination between variants. It has been predicted from theoretical work that small, asexual populations are expected to fix slightly deleterious mutations in an irreversible manner by genetic drift. The reason is that individuals cannot produce descendants with a reduced number of deleterious mutations in the absence of recombination. In addition, the smaller a population is the higher is the chance for fixation of a slightly deleterious mutation. Once a slightly deleterious mutation is fixed, the fitness of the organisms has decreased in an irreversible manner. Cumulative chance fixation of mildly deleterious mutations will gradually decrease the fitness of small and/or asexual populations. This phenomenon is known as Muller's ratchet and may drive the genome degradation of organelles and intracellular bacteria (48, 184, 185).

The reductive evolutionary forces acting on the organelle genomes are not only manifested by gene inactivation and gene transfer to the nucleus. In animal mitochondria the genetic code has been re-assigned in order to minimise the number of tRNAs needed to read all the codons (186, 187). Comparison of the tRNA set from animal mitochondria and nuclear genomes showed that the mitochondrial tRNA genes accumulate nucleotide substitutions much more rapidly. These substitutions are mildly deleterious as inferred from decreased binding stability in the stems, higher variability in the loop length, and fewer invariable sites in the mitochondrial tRNA compared to the nuclear encoded tRNA (188). In a larger study it was shown that the rate of evolution is generally higher in organelles than in prokaryotes and the nucleus of eukaryotes. This effect is due to a reduction in the efficiency of selection on new mutations in organelles, rather than an increased mutations rate (189). This supports the idea that genomes of small, asexual populations are subject to long-term gradual fitness loss. An extreme consequence of Muller's ratchet is that the degeneration of the organelle genomes may lead to the extinction of the host species (189).

Does Muller's ratchet shape the genome evolution of intracellular parasites?

The absence of fossil data for dating of speciation events complicates studies of Muller's ratchet for bacteria. Fortunately, the biology of the obligate endosymbionts in aphids, *Buchnera,* circumvents that problem. As described earlier, the role of *Buchnera* is to supply the aphid with amino acids. The bacteria are impossible to cultivate on artificial media, and the aphids are sterilised or killed by treatment with antibiotics, indicating a mutual obligate symbiotic relationship. The *Buchnera* cells are transmitted maternally from one aphid generation to the next through transovarial transmission (136, 137). *Buchnera* was acquired once by aphids, probably 160-280 million years ago (44). The congruency between phylogenies of the *Buchnera* genus and their hosts indicates that the hosts and the embosymbionts have diversified in parallel, and therefore the fossil record of the aphids can be used to date bacterial nodes (44, 137). By using these dates, *Buchnera* have been shown to fix substitutions that probably are slightly deleterious at a higher rate than the free-living bacteria, both in protein coding genes (108, 190, 191) and the 16S rRNA gene (192).

It is reasonable to believe that Muller's ratchet also affect genomes of other intracellular bacteria, such as *Rickettsia*, although direct evidence for an increased substitution rate on the amino acid level is difficult to obtain in the absence of fossil records for other genera (48). However, several indications that *Rickettsia* experience higher fixation rates for slightly deleterious mutations compared to free-living bacteria can be observed. Non-random use of synonymous codons, a phenomenon known as codon usage bias, is thought to be the result of interplay between mutational and selective forces. For example, free-living microorganisms usually have a tendency to use only a subset of the codons in the highly expressed genes in order to maximise rate and/or accuracy of translation, while lowly expressed genes have a more randomise codon usage (193, 194). In *R. prowazekii* codon usage is roughly similar among genes, with a general bias towards A+T in third codon position (50). This is expected from a genome of low G+C content, if no selection for translational efficiency on the synonymous codons occur, and only mutational forces shape the codon usage (49). Similarly, the unusual long intergenic regions in the *R. prowazekii* genome, which have been shown to represent inactivated genes to a large part (paper IV), show little functional constraint with a G+C value of 23-24%, significantly lower than the overall genomic value of 29% (papers I, II-V; 50). The maintenance of pseudogenes and non-coding regions in the *Rickettsia* genome indicate that the selection on genome size is weaker than in free-living bacteria, such as *E. coli*. The absence of selection for translational efficiency and the weak selection on genome size in the *R. prowazekii* genome may indicate that the effect of Muller's ratchet is acting on the *Rickettsia* genomes, which may drive the lineages on a one-way route towards extinction.

Concluding remarks

Conclusions

The sequencing of the complete *R. prowazekii* genome identified a dozen putative pseudogenes and a large fraction of non-coding DNA. The comparative sequencing of the genomic regions that contained pseudogenes in other species of the *Rickettsia* genus exposed extensive ongoing gene inactivation and loss. Several significant features of reductive evolution and genome degradation in *Rickettsia* could be identified (bold face), all of which have been shown to have parallels in other genera (normal text):

- **Pseudogenes are remnants of permanently inactivated genes without functional constraint and are useful in the study of genome degradation.** There are an increasing number of bacterial pseudogenes described in the literature (Table 3) and the majority of these most likely are without functional constraint.

- **The rate of gene inactivation is higher than the rate of removal of pseudogenes, resulting in a high fraction of non-coding DNA.** A similar conclusion can be drawn from observations that have been made in the partially sequenced *M. leprae* genome and in the chloroplast of non-photosynthetic plant *E. virginiana*, but this was not observed in the two *Chlamydia* genomes.

- **A gene present in the ancestral node seems to have been inactivated several times in parallel in different branches of the phylogenetic tree, rather than in one or a few rare events.** No extensive comparable bacterial dataset exists, but the phenomenon is in agreement with what is known about reductive evolution in chloroplasts and mitochondria.

- **A high fraction of the inactivated and lost genes are genus specific, and the gene losses are not associated with gene rearrangements.** Similar observations were done in the three other genus pairs of bacterial genomes sequenced (Table 4).

Future prospects

Although the comparative sequencing approach was successful in outlining which genes that were inactivated in the different lineages, it did not give detailed information on how and why these genes were inactivated. Expression studies of the *Rickettsia* pseudogene regions should clarify if gene expression must be turned off before a pseudogene is allowed to pick up frameshift mutations. If a pseudogene is expressed, identification of the protein is needed to truly understand the process. The biological background in the inactivation of individual genes is also possible to study experimentally. In the case of *metK*, the testable hypothesis is that *Rickettsia* is able to import AdoMet through a specific import system.

R. conorii is the subject of complete genome sequencing (195) and will further clarify the amount of genome degradation within the *Rickettsia* genus. The generality of conclusions drawn from the *Rickettsia* sequences for other bacteria will be clarified within the next few years when more bacterial genome sequences become available, such as *M. leprae* and *N. meningitidis* (196).

The question if Muller's ratchet is the driving force for reductive evolution in *Rickettsia* remains to be answered. One way to find out if it is would be to purify orthologous proteins from *Rickettsia* and a free-living bacterium and measure if the rickettsial proteins are less effective in performing the biological functions, which would indicate that they have fixed slightly deleterious mutations. Until these kinds of experiments have been performed the long term evolutionary fate of the *Rickettsia* lineage remains unresolved. However, for the human pathogen *R. prowazekii* the evolutionary mistake may have been to become dependent on humans for survival. What a threat is Muller's ratchet when you live inside an animal as self-destructive as humans?

Acknowledgements

I would like to thank **everybody** at the former Department of Molecular Biology, programme 1 & 4, for creating a friendly and stimulating atmosphere.

I am especially grateful to:

Thomas for all the help with the bioinformatic stuff and for being a great competitor, collaborator, colleague and friend; **Morten** for doing a lot of boring but necessary computer stuff for me and everybody else without being *snarstucken*; **Ali** for always taking time out for a discussion, and for showing that it is possible to graduate from the group!; **Ann-Sofie** for excellent care of the sequencing machines and for always being willing to solve practical problems so that I could use the machines during nights and weekends; **Kristina** for knowledge of all the practical things and for always being helpful and nice; **Haleh** for being a nice colleague throughout the years; **Ivica** for all the discussions about everything; **Cecilia** for joining our group at a time when your enthusiasm certainly was needed in the genome project; **Björn** for creating a relaxed atmosphere and being a good lunch friend**; Åsa** for being who you are ;-); **Daniel** for the revival of the Beer Club this autumn; **Olof** for being the new linux guru in the group; **Håkan** for being just the person that we needed in our group; **Linda** for all the nice decorations and drinks at the parties; **Sean** for all the crazy and entertaining discussions; **David** for all the help during the writing process of my thesis and for being the host of the Beer Club throughout the years; **Robert** for sharing your enormous knowledge about everything; **Otto** for creating a sharp contrast to the rest of the deparment with your calmness; **Marie** for always being nice; **Karin** for excellent care of all the economical stuff; **Chuck** for the initiation of this research topic in Uppsala and for creating an inspiring scientific atmosphere in the department with your never ending enthusiasm; **Siv** for always letting me work independently in the lab and at the computer, I really appreciated the enormous freedom you gave me the last year. Thank you for a fun and fruitful co-operation during the last four and a half years!

Men det har ju också funnits en värld utanför BMCs gråa korridorer, den s k *verkligheten*. Mina besök där har ofta gått via **Roger** och **Anette** som fixat en aldrig sinande ström av sammankomster i glada vänners lag. Tack till er och alla andra som vet vad jag pratar om!

Ett stort tack till släkten! **Walter** och **Margaretha** för att ni låtit mig vara 5% bonde - efter 15 års höskörd är organisationen av ett sekvensprojekt en barnlek!; Storebror **Pär** för att du delar de flesta av mina intressen och ointressen, och för att så många gånger fört bort mina tankar från forskning genom att diskutera viktiga saker som t ex vilka skivor med Barbara Acklin man borde ha. Till sist ett extra stort tack till mina föräldrar, **Nils-Olov** och **Margaretha,** för att ni alltid låtit mig vara den jag är och alltid litat på mitt omdöme, trots att jag tagit denna märkliga och något omständiga väg i livet!

References

1. Zuckerkandl E, Pauling L. Molecules as documents of evolutionary history. *J Theor Biol* 1965;**8**:357-66.
2. Ricketts HT. *J Am Med Assoc* 1909;**52**:379-80.
3. Nicolle C, Comte C, Conseil E. *C R Acad Sci* 1909;**149**:486-9.
4. Weigl R. *Pol Acad Sci Bull Med* 1931;**1**:1-23.
5. Gross L. How Charles Nicolle of the Pasteur Institute discovered that epidemic typhus is transmitted by lice: reminiscences from my years at the Pasteur Institute in Paris. *Proc Natl Acad Sci U S A* 1996;**93**:10539-40.
6. Raju TN. The Nobel chronicles. 1928: Charles Jules Henry Nicolle (1866-1936). *Lancet* 1998;**352**:1791.
7. Ormsbee RA. Rickettsiae as organisms. *Acta Virol (Praha)* 1985;**29**(5):432-48.
8. Woese CR. Bacterial evolution. *Microbiol Rev* 1987;**51**:221-71.
9. NCBI Taxonomy Database, http://www.ncbi.nlm.nih.gov/.
10. Olsen GJ, Woese CR, Overbeek R. The winds of (evolutionary) change: breathing new life into microbiology. *J Bacteriol* 1994;**176**:1-6.
11. Margulis L. Origin of eukaryotes. New Haven, CT: Yale Univ. Press; 1970.
12. Gray MW, Doolittle WF. Has the endosymbiont hypothesis been proven? *Microbiol Rev* 1982;**46**:1-42.
13. Yang D, Oyaizu Y, Oyaizu H, Olsen GJ, Woese CR. Mitochondrial origins. *Proc Natl Acad Sci U S A* 1985;**82**:4443-7.
14. Viale AM, Arakaki AK. The chaperone connection to the origins of the eukaryotic organelles. *FEBS Lett* 1994;**341**:146-51.
15. Gray MW. In Evolution of microbial life. Roberts DM, Sharp PM, Alderson G, Collins M. Cambridge: Cambridge University Press 1996, pp109-26.
16. Sicheritz-Ponten T, Kurland CG, Andersson SGE. A phylogenetic analysis of the cytochrome b and cytochrome c oxidase I genes supports an origin of mitochondria from within the Rickettsiaceae. *Biochim Biophys Acta* 1998;**1365**:545-51.
17. Weiss E, Moulder JW. Genus I. Rickettsia. In Bergey's manual of systemtic bacteriology, vol. 1. Krieg NR, Holt JG. Baltimore, Md.: The Williams & Wilkings Co. 1984, pp688-98.
18. Tamura A, Ohashi N, Urakami H, Miyamura S. Classification of *Rickettsia tsutsugamushi* in a new genus, *Orientia* gen. nov., as *Orientia tsutsugamushi* comb. nov. *Int J Syst Bacteriol* 1995;**45**:589-91.
19. Roux V, Rydkina E, Eremeeva M, Raoult D. Citrate synthase gene comparison, a new tool for phylogenetic analysis, and its application for the rickettsiae. *Int J Syst Bacteriol* 1997;**47**:252-61.
20. Retief FP, Cilliers L. The epidemic of Athens, 430-426 BC. *S Afr Med J* 1998;**88**:50-3.
21. Hambraeus L. Krig har sot i sitt följe - om kriget och epidemierna. In Epidemiernas historia och framtid. Evengård B. . Natur och Kultur 1992, pp85-111.
22. Raoult D, Roux V, Ndihokubwayo JB, Bise G, Baudon D, Marte G, Birtles R. Jail fever (epidemic typhus) outbreak in Burundi. *Emerg Infect Dis* 1997;**3**:357-60.

23. Raoult D, Ndihokubwayo JB, Tissot-Dupont H, Roux V, Faugere B, Abegbinni R, Birtles RJ. Outbreak of epidemic typhus associated with trench fever in Burundi. *Lancet* 1998;**352**:353-8.

24. Tarasevich I, Rydkina E, Raoult D. Outbreak of epidemic typhus in Russia. *Lancet* 1998;**352**:1151.

25. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;**22**:4673-80.

26. Galtier N, Gouy M, Gautier C. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* 1996;**12**:543-8.

27. Stothard DR. The evolutionary history of the genus *Rickettsia* as inferred from 16S and 23S rRNA genes and the 17 kDa cell surface antigen gene. Dissertion. Columbus, OH: The Ohio State University; 1995.

28. Andersson SGE, Stothard DR, Fuerst P, Kurland CG. Molecular phylogeny and rearrangement of rRNA genes in *Rickettsia* species. *Mol Biol Evol* 1999;**16**:987-95.

29. Fournier PE, Roux V, Raoult D. Phylogenetic analysis of spotted fever group rickettsiae by study of the outer surface protein rOmpA. *Int J Syst Bacteriol* 1998;**48**:839-49.

30. Stenos J, Roux V, Walker D, Raoult D. *Rickettsia honei* sp. nov., the aetiological agent of Flinders Island spotted fever in Australia. *Int J Syst Bacteriol* 1998;**48 Pt 4**:1399-404.

31. Raoult D, Roux V. Rickettsioses as paradigms of new or emerging infectious diseases. *Clin Microbiol Rev* 1997;**10**:694-719.

32. Perine PL, Krause DW, Awoke S, McDade JE. Single-dose doxycycline treatment of louse-borne relapsing fever and epidemic typhus. *Lancet* 1974;**2**:742-4.

33. Azad AF, Beard CB. Rickettsial pathogens and their arthropod vectors. *Emerg Infect Dis* 1998;**4**:179-86.

34. Eremeeva ME, Roux V, Raoult D. Determination of genome size and restriction pattern polymorphism of *Rickettsia prowazekii* and *Rickettsia typhi* by pulsed field gel electrophoresis. *FEMS Microbiol Lett* 1993;**112**:105-12.

35. Roux V, Raoult D. Genotypic identification and phylogenetic analysis of the spotted fever group rickettsiae by pulsed-field gel electrophoresis. *J Bacteriol* 1993;**175**:4895-904.

36. Tyeryar Jr FJ, Weiss E, Millar DB, Bozeman FM, Ormsbee RA. DNA base composition of rickettsiae. *Science* 1973;**180**:415-7.

37. Schramek S. Deooxyribonucleic acid base composition of Rickettsiae belonging to the Rocky Mountain spotted fever group isolated in Czechoslovakia. *Acta Virol (Praha)* 1974;**18**:173-4.

38. Nilsson K, Lindquist O, Påhlson C. Association of *Rickettsia helvetica* with chronic perimyocarditis in sudden cardiac death. *Lancet* 1999;**354**:1169-73.

39. Walker DH. Rocky Mountain spotted fever: a seasonal alert. *Clin Infect Dis* 1995;**20**:1111-7.

40. Nilsson K, Jaenson TGT, Uhnoo I, Lindquist O, Pettersson B, Uhlen M, Friman G, Påhlson C. Characterization of a spotted fever group *Rickettsia* from *Ixodes ricinus* ticks in Sweden. *J Clin Microbiol* 1997;**35**:243-7.

41. Nilsson K, Lindquist O, Liu AJ, Jaenson TGT, Friman G, Påhlson C. *Rickettsia helvetica* in *Ixodes ricinus* ticks in Sweden. *J Clin Microbiol* 1999;**37**:400-3.

42. Noden BH, Radulovic S, Higgins JA, Azad AF. Molecular identification of *Rickettsia typhi* and *R. felis* in co-infected *Ctenocephalides felis* (Siphonaptera: Pulicidae). *J Med Entomol* 1998;**35**:410-4.

43. Williams SG, Sacci Jr JB, Schriefer ME, Andersen EM, Fujioka KK, Sorvillo FJ, Barr AR, Azad AF. Typhus and typhuslike rickettsiae associated with opossums and their fleas in Los Angeles County, California. *J Clin Microbiol* 1992;**30**:1758-62.

44. Moran NA, Munson MA, Baumann P, Ishikawa H. A molecular clock in endosymbiotic bacteria is calibrated using the insect hosts. *Proc R Soc Lond B Biol Sci* 1993;**253**:167-71.

45. Baxter JD. The typhus group. *Clin Dermatol* 1996;**14**:271-8.

46. Rachek LI, Tucker AM, Winkler HH, Wood DO. Transformation of *Rickettsia prowazekii* to rifampin resistance. *J Bacteriol* 1998;**180**:2118-24.

47. Andersson SGE, Kurland CG. Genomic evolution drives the evolution of the translation system. *Biochem Cell Biol* 1995;**73**:775-87.

48. Andersson SGE, Kurland CG. Reductive evolution of resident genomes. *Trends Microbiol* 1998;**6**:263-78.

49. Muto A, Osawa S. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Natl Acad Sci U S A* 1987;**84**:166-9.

50. Andersson SGE, Sharp PM. Codon usage and base composition in *Rickettsia prowazekii*. *J Mol Evol* 1996;**42**:525-36.

51. Pang H, Winkler HH. Transcriptional analysis of the 16s rRNA gene in *Rickettsia prowazekii*. *J Bacteriol* 1996;**178**:1750-5.

52. Cai J, Winkler HH. Transcriptional regulation in the obligate intracytoplasmic bacterium *Rickettsia prowazekii*. *J Bacteriol* 1996;**178**:5543-5.

53. Shaw EI, Marks GL, Winkler HH, Wood DO. Transcriptional characterization of the *Rickettsia prowazekii* major macromolecular synthesis operon. *J Bacteriol* 1997;**179**:6448-52.

54. Cai J, Pang H, Wood DO, Winkler HH. The citrate synthase-encoding gene of *Rickettsia prowazekii* is controlled by two promoters. *Gene* 1995;**163**(1):115-9.

55. Wisseman Jr CL. Selected observations on rickettsiae and their host cells. *Acta Virol (Praha)* 1986;**30**:81-95.

56. Walker TS, Winkler HH. Penetration of cultured mouse fibroblasts (L cells) by *Rickettsia prowazekii*. *Infect Immun* 1978;**22**:200-8.

57. Winkler HH. *Rickettsia* species (as organisms). *Annu Rev Microbiol* 1990;**44**:131-5.

58. Wisseman Jr CL, Waddell AD. In vitro studies on rickettsia-host cell interactions: intracellular growth cycle of virulent and attenuated *Rickettsia prowazeki* in chicken embryo cells in slide chamber cultures. *Infect Immun* 1975;**11**:1391-404.

59. Wisseman Jr CL, Edlinger EA, Waddell AD, Jones MR. Infection cycle of *Rickettsia rickettsii* in chicken embryo and L-929 cells in culture. *Infect Immun* 1976;**14**:1052-64.

60. Pang H, Winkler HH. Copy number of the 16S rRNA gene in *Rickettsia prowazekii*. *J Bacteriol* 1993;**175**:3893-6.

61. Pang H, Winkler HH. The concentrations of stable RNA and ribosomes in *Rickettsia prowazekii*. *Mol Microbiol* 1994;**12**:115-20.

62. Winkler HH. *Rickettsia prowazekii*, ribosomes and slow growth. *Trends Microbiol* 1995;**3**:196-8.

63. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM *et al*. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995;**269**:496-512.

64. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM *et al*. The minimal gene complement of *Mycoplasma genitalium*. *Science* 1995;**270**:397-403.

65. Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S *et al*. Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res* 1996;**3**(3):109-36.

66. Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li B-C, Herrman R. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res* 1996;**24**:4420-49.

67. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S *et al*. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* 1997;**390**(6657):249-56.

68. Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA *et al*. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* 1997;**388**:539-47.

69. Blattner FR, Plunkett GI, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF *et al*. The complete genome sequence of *Escherichia coIi* K-12. *Science* 1997;**277**:1453-74.

70. Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK *et al*. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* 1997;**390**:580-6.

71. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3 *et al*. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 1998;**393**:537-44.

72. Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, Dodson R, Gwinn M, Hickey EK, Clayton R, Ketchum KA *et al*. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* 1998;**281**:375-88.

73. Deckert G, Warren PV, Gaasterland T, Young WG, Lenox AL, Graham DE, Overbeek R, Snead MA, Keller M, Aujay M *et al*. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* 1998;**392**(6674):353-8.

74. Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q *et al*. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* 1998;**282**:754-9.

75. Alm RA, Ling LS, Moir DT, King BL, Brown ED, Doig PC, Smith DR, Noonan B, Guild BC, deJonge BL *et al*. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* 1999;**397**:176-80.

76. Kalman S, Mitchell W, Marathe R, Lammel C, Fan J, Hyman RW, Olinger L, Grimwood J, Davis RW, Stephens RS. Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nat Genet* 1999;**21**:385-9.

77. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA *et al*. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* 1999;**399**:323-9.

78. Rydkina E, Roux V, Raoult D. Determination of the genome size of *Ehrlichia* spp., using pulsed field gel electrophoresis. *FEMS Microbiol Lett* 1999;**176**:73-8.

79. Alleman AR, Kamper SM, Viseshakul N, Barbet AF. Analysis of the *Anaplasma marginale* genome by pulsed-field electrophoresis. *J Gen Microbiol* 1993;**139**:2439-44.

80. McGraw EA, O'Neill SL. Evolution of *Wolbachia pipientis* transmission dynamics in insects. *Trends Microbiol* 1999;**7**:297-302.

81. COG Database, http://www.ncbi.nlm.nih.gov/.

82. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science* 1997;**278**:631-7.

83. Koonin EV, Tatusov RL, Galperin MY. Beyond complete genomes: from sequence to structure and function. *Current Opinion in Structural Biology* 1998;**8**:355-63.

84. Zomorodipour A, Andersson SGE. Obligate intracellular parasites: *Rickettsia prowazekii* and *Chlamydia trachomatis*. *FEBS Lett* 1999;**452**:11-5.

85. Winkler HH, Neuhaus HE. Non-mitochondrial ATP transport. *Trends Biochem Sci* 1999;**24**:64-8.

86. Siefert JL, Martin KA, Abdi F, Widger WR, Fox GE. Conserved gene clusters in bacterial genomes provide further support for the primacy of RNA. *J Mol Evol* 1997;**45**:467-72.

87. Itoh T, Takemoto K, Mori H, Gojobori T. Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol Biol Evol* 1999;**16**:332-46.

88. Andersson SGE, Zomorodipour A, Winkler HH, Kurland CG. Unusual organization of the rRNA genes in *Rickettsia prowazekii*. *J Bacteriol* 1995;**177**:4171-5.

89. Syvänen AC, Amiri H, Jamal A, Andersson SGE, Kurland CG. A chimeric disposition of the elongation factor genes in *Rickettsia prowazekii*. *J Bacteriol* 1996;**178**:6192-9.

90. Yoshikawa H, Ogasawara N. Structure and function of DnaA and the DnaA-box in eubacteria: evolutionary relationships of bacterial replication origins. *Mol Microbiol* 1991;**5**:2589-97.

91. Sun LV, Babaratsas A, Savakis C, O'Neill SL, Bourtzis K. Gene organization of the *dnaA* region of *Wolbachia*. *J Bacteriol* 1999;**181**:4708-10.

92. Farabaugh PJ. Programmed translational frameshifting. *Annu Rev Genet* 1996;**30**:507-28.

93. Craigen WJ, Cook RG, Tate WP, Caskey CT. Bacterial peptide chain release factors: conserved primary structure and possible frameshift regulation of release factor 2. *Proc Natl Acad Sci U S A* 1985;**82**:3616-20.

94. Böck A, Forchhammer K, Heider J, Baron C. Selenoprotein synthesis: an expansion of the genetic code. *Trends Biochem Sci* 1991;**16**:463-7.

95. Smith HC, Gott JM, Hanson MR. A guide to RNA editing. *RNA* 1997;**3**:1105-23.

96. Martinez-Costa OH, Fernandez-Moreno MA, Malpartida F. The *relA*/*spoT*-homologous gene in *Streptomyces coelicolor* encodes both ribosome-dependent (p)ppGpp-synthesizing and -degrading activities. *J Bacteriol* 1998;**180**:4123-32.

97. Gentry DR, Cashel M. Mutational analysis of the *Escherichia coli spoT* gene identifies distinct but overlapping regions involved in ppGpp synthesis and degradation. *Mol Microbiol* 1996;**19**:1373-84.

98. Gojobori T, Li WH, Graur D. Patterns of nucleotide substitution in pseudogenes and functional genes. *J Mol Evol* 1982;**18**:360-9.

99. Li WH, Wu CI, Luo CC. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J Mol Evol* 1984;**21**:58-71.

100. Petrov DA, Hartl DL. Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. *Proc Natl Acad Sci U S A* 1999;**96**:1475-9.

101. Petrov DA, Lozovskaya ER, Hartl DL. High intrinsic mutation rate of DNA loss in *Drosophila*. *Nature* 1996;**384**:346-9.

102. Petrov DA, Hartl DL. High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. *Mol Biol Evol* 1998;**15**:293-302.

103. Graur D, Shuali Y, Li WH. Deletions in processed pseudogenes accumulate faster in rodents than in humans. *J Mol Evol* 1989;**28**:279-85.

104. Petrov DA, Hartl DL. Trash DNA is what gets thrown away: high rate of DNA loss in *Drosophila*. *Gene* 1997;**205**:279-89.

105. Lawrence JG, Ochman H. Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* 1997;**44**:383-97.

106. Petrov DA, Chao Y-C, Stephenson EC, Hartl DL. Pseudogene evolution in *Drosophila* suggests a high rate of DNA loss. *Mol Biol Evol* 1998;**15**:1562-7.

107. Moriyama EN, Petrov DA, Hartl DL. Genome size and intron size in *Drosophila*. *Mol Biol Evol* 1998;**15**:770-3.

108. Brynnel EU, Kurland CG, Moran NA, Andersson SGE. Evolutionary rates for tuf genes in endosymbionts of aphids. *Mol Biol Evol* 1998;**15**:574-82.

109. Gottlieb LD, Ford VS. A recently silenced, duplicated *PgiC* locus in *Clarkia*. *Mol Biol Evol* 1997;**14**:125-32.

110. Garcia-Meunier P, Etienne-Julan M, Fort P, Piechaczyk M, Bonhomme F. Concerted evolution in the GAPDH family of retrotransposed pseudogenes. *Mamm Genome* 1993;**4**(12):695-703.

111. Kvarnheden A, Albert VA, Engstrom P. Molecular evolution of cdc2 pseudogenes in spruce (*Picea*). *Plant Mol Biol* 1998;**36**(5):767-74.

112. Delorme C, Godon JJ, Ehrlich SD, Renault P. Gene inactivation in *Lactococcus lactis*: histidine biosynthesis. *J Bacteriol* 1993;**175**:4391-9.

113. Godon JJ, Delorme C, Bardowski J, Chopin MC, Ehrlich SD, Renault P. Gene inactivation in *Lactococcus lactis*: branched-chain amino acid biosynthesis. *J Bacteriol* 1993;**175**:4383-90.

114. Smith DR, Richterich P, Rubenfield M, Rice PW, Butler C, Lee HM, Kirst S, Gundersen K, Abendschan K, Xu Q *et al*. Multiplex sequencing of 1.5 Mb of the *Mycobacterium leprae* genome. *Genome Res* 1997;**7**:802-19.

115. Cole ST. Comparative mycobacterial genomics. *Curr Opin Microbiol* 1998;**1**:567-71.

116. Buchrieser C, Rusniok C, Frangeul L, Couve E, Billault A, Kunst F, Carniel E, Glaser P. The 102-kilobase pgm locus of *Yersinia pestis*: sequence analysis and comparison of selected regions among different *Yersinia pestis* and *Yersinia pseudotuberculosis* strains. *Infect Immun* 1999;**67**:4851-61.

117. Simonet M, Riot B, Fortineau N, Berche P. Invasin production by *Yersinia pestis* is abolished by insertion of an IS200-like element within the *inv* gene. *Infect Immun* 1996;**64**:375-9.

118. Skurnik M, Wolf-Watz H. Analysis of the *yopA* gene encoding the Yop1 virulence determinants of *Yersinia* spp. *Mol Microbiol* 1989;**3**:517-29.

119. Perry RD, Straley SC, Fetherston JD, Rose DJ, Gregor J, Blattner FR. DNA sequencing and analysis of the low-Ca^{2+}-response plasmid pCD1 of *Yersinia pestis* KIM5. *Infect Immun* 1998;**66**:4611-23.

120. Zhu P, Morelli G, Achtman M. The *opcA* and Ψ*opcB* regions in *Neisseria*: genes, pseudogenes, deletions, insertion elements and DNA islands. *Mol Microbiol* 1999;**33**:635-50.

121. Feavers IM, Maiden MC. A gonococcal *porA* pseudogene: implications for understanding the evolution and pathogenicity of *Neisseria gonorrhoeae*. *Mol Microbiol* 1998;**30**:647-56.

122. Lai CY, Baumann P, Moran N. The endosymbiont (*Buchnera* sp.) of the aphid *Diuraphis noxia* contains plasmids consisting of *trpEG* and tandem repeats of *trpEG* pseudogenes. *Appl Environ Microbiol* 1996;**62**:332-9.

123. Baumann L, Clark MA, Rouhbakhsh D, Baumann P, Moran NA, Voegtlin DJ. Endosymbionts (*Buchnera*) of the aphid *Uroleucon sonchi* contain plasmids with *trpEG* and remnants of *trpE* pseudogenes. *Curr Microbiol* 1997;**35**:18-21.

124. Van Ham RCHJ, Martinez-Torres D, Moya A, Latorre A. Plasmid-encoded anthranilate synthase (TrpEG) in *Buchnera aphidicola* from aphids of the family pemphigidae. *Appl Environ Microbiol* 1999;**65**:117-25.

125. Casjens S, Palmer N, van Vugt R, Huang WM, Stevenson B, Rosa P, Lathigra R, Sutton G, Peterson J, Dodson RJ *et al*. A bacterial genome in flux: The twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*. *Mol Microbiol* 2000;**in press**.

126. Rosqvist R, Skurnik M, Wolf-Watz H. Increased virulence of *Yersinia pseudotuberculosis* by two independent mutations. *Nature* 1988;**334**:522-4.

127. Maurelli AT, Fernandez RE, Bloch CA, Rode CK, Fasano A. "Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. *Proc Natl Acad Sci U S A* 1998;**95**:3943-8.

128. Hacker J, Blum-Oehler G, Muhldorfer I, Tschape H. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol Microbiol* 1997;**23**:1089-97.

129. Zhou J, Spratt BG. Sequence diversity within the *argF*, *fbp* and *recA* genes of natural isolates of *Neisseria meningitidis*: interspecies recombination within the *argF* gene. *Mol Microbiol* 1992;**6**:2135-46.

130. Seiler A, Reinhardt R, Sarkari J, Caugant DA, Achtman M. Allelic polymorphism and site-specific recombination in the *opc* locus of *Neisseria meningitidis*. *Mol Microbiol* 1996;**19**:841-56.

131. Holmes EC, Urwin R, Maiden MC. The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*. *Mol Biol Evol* 1999;**16**:741-9.

132. Morelli G, Malorny B, Muller K, Seiler A, Wang JF, del Valle J, Achtman M. Clonal descent and microevolution of *Neisseria meningitidis* during 30 years of epidemic spread. *Mol Microbiol* 1997;**25**:1047-64.

133. Feil E, Zhou J, Maynard Smith J, Spratt BG. A comparison of the nucleotide sequences of the *adk* and *recA* genes of pathogenic and commensal *Neisseria* species: evidence for extensive interspecies recombination within *adk*. *J Mol Evol* 1996;**43**:631-40.

134. Zhou J, Bowler LD, Spratt BG. Interspecies recombination, and phylogenetic distortions, within the glutamine synthetase and shikimate dehydrogenase genes of *Neisseria meningitidis* and commensal *Neisseria* species. *Mol Microbiol* 1997;**23**:799-812.

135. Kroll JS, Wilks KE, Farrant JL, Langford PR. Natural genetic exchange between *Haemophilus* and *Neisseria*: intergeneric transfer of chromosomal genes between major human pathogens. *Proc Natl Acad Sci U S A* 1998;**95**:12381-5.

136. Moran NA, Telang A. The evolution of bacteriocyte-associated endosymbionts in insects. *Bioscience* 1998;**48**:295-304.

137. Douglas AE. Nutritional interactions in insect-microbial symbiosis: aphids and their symbiotic bacteria *Buchnera*. *Annu Rev Entomol* 1998;**43**:17-37.

138. Lai CY, Baumann L, Baumann P. Amplification of *trpEG*: adaptation of *Buchnera* aphidicola to an endosymbiotic association with aphids. *Proc Natl Acad Sci U S A* 1994;**91**:3819-23.

139. Baumann L, Baumann P, Moran NA, Sandström J, Thao ML. Genetic characterization of plasmids containing genes encoding enzymes of leucine biosynthesis in endosymbionts (*Buchnera*) of aphids. *J Mol Evol* 1999;**48**:77-85.

140. Rouhbakhsh D, Clark MA, Baumann L, Moran NA, Baumann P. Evolution of the tryptophan biosynthetic pathway in *Buchnera* (aphid endosymbionts): studies of plasmid-associated *trpEG* within the genus *Uroleucon*. *Mol Phylogenet Evol* 1997;**8**:167-76.

141. Stevenson B, Casjens S, Rosa P. Evidence of past recombination events among the genes encoding the Erp antigens of *Borrelia burgdorferi*. *Microbiology* 1998;**144**:1869-79.

142. Casjens S. Evolution of the linear DNA replicons of the *Borrelia* spirochetes. *Curr Opin Microbiol* 1999;**2**(5):529-34.

143. Lawrence JG, Ochman H. Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci U S A* 1998;**95**:9413-7.

144. Martin W. Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *Bioessays* 1999;**21**:99-104.

145. Bergthorsson U, Ochman H. Heterogeneity of genome sizes among natural isolates of *Escherichia coli*. *J Bacteriol* 1995;**177**:5784-9.

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

147. Hurtado A, Rodriguez-Valera F. Accessory DNA in the genomes of representatives of the *Escherichia coli* reference collection. *J Bacteriol* 1999;**181**:2548-54.

148. Jain R, Rivera MC, Lake JA. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A* 1999;**96**:3801-6.

149. Wolf YI, Aravind L, Grishin NV, Koonin EV. Evolution of aminoacyl-tRNA synthetases-analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res* 1999;**9**:689-710.

150. Doolittle WF. Phylogenetic classification and the universal tree. *Science* 1999;**284**:2124-9.

151. Andersson SGE, Kurland CG. Ancient and recent horizontal transfer events: the origins of mitochondria. *APMIS Suppl* 1998;**84**:5-14.

152. Wolf YI, Aravind L, Koonin EV. Rickettsiae and Chlamydiae: evidence of horizontal gene transfer and gene exchange. *Trends Genet* 1999;**15**:173-5.

153. Himmelreich R, Plagens H, Hilbert H, Reiner B, Herrmann R. Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. *Nucleic Acids Res* 1997;**25**:701-12.

154. Gray MW. *Rickettsia*, typhus and the mitochondrial connection. *Nature* 1998;**396**:109-10.

155. Müller M, Martin W. The genome of *Rickettsia prowazekii* and some thoughts on the origin of mitochondria and hydrogenosomes. *Bioessays* 1999;**21**:377-81.

156. Palmer JD, Delwiche CF. The origin and evolution of plastids and their genomes. In Molecular Systematics of Plants. II. DNA Sequencing. Soltis DE, Soltis PS, Doyle JJ. Norwell, MA: Kluwer Academic Publisher 1998, pp375-409.

157. Gray MW, Lang BF, Cedergren R, Golding GB, Lemieux C, Sankoff D, Turmel M, Brossard N, Delage E, Littlejohn TG *et al*. Genome structure and gene content in protist mitochondrial DNAs. *Nucleic Acids Res* 1998;**26**:865-78.

158. Lang BF, Burger G, O'Kelly CJ, Cedergren R, Golding GB, Lemieux C, Sankoff D, Turmel M, Gray MW. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature* 1997;**387**:493-7.

159. Martin W, Stoebe B, Goremykin V, Hansmann S, Hasegawa M, Kowallik KV. Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* 1998;**393**:162-5.

160. Palmer JD. The mitochondrion that time forgot. *Nature* 1997;**387**:454-5.

161. Gray MW, Burger G, Lang BF. Mitochondrial evolution. *Science* 1999;**283**:1476-81.

162. Martin W, Schnarrenberger C. The evolution of the Calvin cycle from prokaryotic to eukaryotic chromosomes: a case study of functional redundancy in ancient pathways through endosymbiosis. *Curr Genet* 1997;**32**:1-18.

163. Keeling PJ, Doolittle WF. Evidence that eukaryotic triosephosphate isomerase is of alpha-proteobacterial origin. *Proc Natl Acad Sci U S A* 1997;**94**:1270-5.

164. Nugent JM, Palmer JD. RNA-mediated transfer of the gene *coxII* from the mitochondrion to the nucleus during flowering plant evolution. *Cell* 1991;**66**:473-81.

165. Grohmann L, Brennicke A, Schuster W. The mitochondrial gene encoding ribosomal protein S12 has been translocated to the nuclear genome in Oenothera. *Nucleic Acids Res* 1992;**20**:5641-6.

166. Wischmann C, Schuster W. Transfer of *rps10* from the mitochondrion to the nucleus in *Arabidopsis thaliana*: evidence for RNA-mediated transfer and exon shuffling at the integration site. *FEBS Lett* 1995;**374**:152-6.

167. Sanchez H, Fester T, Kloska S, Schroder W, Schuster W. Transfer of *rps19* to the nucleus involves the gain of an RNP-binding motif which may functionally replace RPS13 in *Arabidopsis mitochondria*. *EMBO J* 1996;**15**:2138-49.

168. Kobayashi Y, Knoop V, Fukuzawa H, Brennicke A, Ohyama K. Interorganellar gene transfer in bryophytes: the functional *nad7* gene is nuclear encoded in *Marchantia polymorpha*. *Mol Gen Genet* 1997;**256**:589-92.

169. Wolfe KH, Morden CW, Palmer JD. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proc Natl Acad Sci U S A* 1992;**89**:10648-52.

170. Wolfe KH, Morden CW, Ems SC, Palmer JD. Rapid evolution of the plastid translational apparatus in a nonphotosynthetic plant: loss or accelerated sequence evolution of tRNA and ribosomal protein genes. *J Mol Evol* 1992;**35**:304-17.

171. Lohan AJ, Wolfe KH. A subset of conserved tRNA genes in plastid DNA of nongreen plants. *Genetics* 1998;**150**:425-33.

172. dePamphilis CW, Young ND, Wolfe AD. Evolution of plastid gene *rps2* in a lineage of hemiparasitic and holoparasitic plants: many losses of photosynthesis and complex patterns of rate variation. *Proc Natl Acad Sci U S A* 1997;**94**:7367-72.

173. Wolfe AD, dePamphilis CW. The effect of relaxed functional constraints on the photosynthetic gene *rbcL* in photosynthetic and nonphotosynthetic parasitic plants. *Mol Biol Evol* 1998;**15**:1243-58.

174. Watanabe KI, Bessho Y, Kawasaki M, Hori H. Mitochondrial genes are found on minicircle DNA molecules in the mesozoan animal *Dicyema*.. *J Mol Biol* 1999;**286**:645-50.

175. Zhang Z, Green BR, Cavalier-Smith T. Single gene circles in dinoflagellate chloroplast genomes. *Nature* 1999;**400**:155-9.

176. McFadden G. Ever decreasing circles. *Nature* 1999;**400**:119-20.

177. Douglas SE. Plastid evolution: origins, diversity, trends. *Curr Opin Genet Dev* 1998;**8**:655-61.

178. Gilson PR, Maier UG, McFadden GI. Size isn't everything: lessons in genetic miniaturisation from nucleomorphs. *Curr Opin Genet Dev* 1997;**7**:800-6.

179. Gilson PR, McFadden GI. The miniaturized nuclear genome of a eukaryotic endosymbiont contains genes that overlap, genes that are cotranscribed, and the smallest known spliceosomal introns. *Proc Natl Acad Sci U S A* 1996;**93**:7737-42.

180. Gilson PR, McFadden GI. Good things in small packages: the tiny genomes of chlorarachniophyte endosymbionts. *Bioessays* 1997;**19**:167-73.

181. Embley TM, Martin W. A hydrogen-producing mitochondrion. *Nature* 1998;**396**:517-9.

182. Andersson SGE, Kurland CG. Origins of mitochondria and hydrogenosomes. *Curr Opin Microbiol* 1999;**2**(5):535-41.

183. Akhmanova A, Voncken F, van Alen T, van Hoek A, Boxma B, Vogels G, Veenhuis M, Hackstein JH. A hydrogenosome with a genome. *Nature* 1998;**396**:527-8.

184. Muller JJ. The relation of recombination to mutational advance. *Mutat Res* 1964;**1**:2-9.

185. Felsenstein J. The evolutionary advantage of recombination. *Genetics* 1974;**78**:737-56.

186. Andersson SGE, Kurland CG. An extreme codon preference strategy: codon reassignment. *Mol Biol Evol* 1991;**8**:530-44.

187. Kurland CG. Evolution of mitochondrial genomes and the genetic code. *Bioessays* 1992;**14**:709-14.

188. Lynch M. Mutation accumulation in transfer RNAs: molecular evidence for Muller's ratchet in mitochondrial genomes. *Mol Biol Evol* 1996;**13**:209-20.

189. Lynch M. Mutation accumulation in nuclear, organelle, and prokaryotic transfer RNA genes. *Mol Biol Evol* 1997;**14**:914-25.

190. Moran NA. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc Natl Acad Sci U S A* 1996;**93**:2873-8.

191. Wernegreen JJ, Moran NA. Evidence for genetic drift in endosymbionts (*Buchnera*): analyses of protein-coding genes. *Mol Biol Evol* 1999;**16**:83-97.

192. Lambert JD, Moran NA. Deleterious mutations destabilize ribosomal RNA in endosymbiotic bacteria. *Proc Natl Acad Sci U S A* 1998;**95**:4458-62.

193. Andersson SGE, Kurland CG. Codon preferences in free-living microorganisms. *Microbiol Rev* 1990;**54**:198-210.

194. Akashi H, Eyre-Walker A. Translational selection and molecular evolution. *Curr Opin Genet Dev* 1998;**8**:688-93.

195. Genescope, http://www.genoscope.cns.fr/.

196. The Sanger Centre, http://www.sanger.ac.uk/.