genes

Review

Factors Behind Junk DNA in Bacteria

Rosario Gil [1],* and Amparo Latorre [1,2]

[1.] Institut Cavanilles de Biodiversitat i Biologia Evolutiva, Universitat de València, Apartado Postal 22085, 46071 València, Spain; Departament de Genètica, Universitat de València, Dr. Moliner, 50, 46100 Burjassot (València), Spain

[2] Área de Genómica y Salud, Centro Superior de Investigación en Salud Pública (CSISP), Avenida de Cataluña 21, 46020 Valencia, Spain; E-Mail: amparo.latorre@uv.es

* Author to whom correspondence should be addressed; E-Mail: rosario.gil@uv.es;
Tel.: +34-96-35-43824; Fax: +34-96-35-43670.

Abstract: Although bacterial genomes have been traditionally viewed as being very compact, with relatively low amounts of repetitive and non-coding DNA, this view has dramatically changed in recent years. The increase of available complete bacterial genomes has revealed that many species present abundant repetitive DNA (*i.e.*, insertion sequences, prophages or paralogous genes) and that many of these sequences are not functional but can have evolutionary consequences as concerns the adaptation to specialized host-related ecological niches. Comparative genomics analyses with close relatives that live in non-specialized environments reveal the nature and fate of this bacterial junk DNA. In addition, the number of insertion sequences and pseudogenes, as well as the size of the intergenic regions, can be used as markers of the evolutionary stage of a genome.

Keywords: junk DNA; pseudogenes; intergenic regions (IGR); insertion sequences (IS); genome degradation

1. Introduction

1.1. Genome Size and Junk DNA

The availability of an increasing number of complete genome sequences has allowed researchers to extract general rules about their shape and dynamics, as well as to analyze the molecular and evolutionary forces involved. Eukaryotes have a wide range of genome sizes, from the 6.5 Mb of the pneumonia-causing fungi *Pneumocystis carinii* f. sp. *muris* to the 133 Gb of the marbled lungfish *Protopterus aethiopicus* [1]. However, it has been demonstrated that the increase of genome size is not correlated with an increase in complexity or gene number, as bigger eukaryotic genomes tend to increase the amount of non-coding DNA (ncDNA) and contain large amounts of transposable elements [2]. The high percentage of ncDNA present in higher eukaryotes genomes was already noticed in the 1970s, when it was considered as "junk DNA" [3]. However, it has become obvious that many of these non-coding sequences are fully functional. Therefore, it would be more appropriate to refer to junk DNA only when we are talking about DNA that has "little obvious function" and "contributes little or nothing to the fitness of the organism" [4]. This would include inter- or intragenic regions that have no known function, pseudogenes, repeated sequences and selfish DNA such as transposons and viral elements.

On the other side, prokaryotic genomes are in general smaller and less variable than eukaryotic ones, ranging from less than 200 kb to more than 11 Mb (http://www.genomesonline.org). The size distribution in the different bacterial taxonomic groups shows that species with large and small genomes coexist in the different lineages and important size variation can even be observed among strains of the same species. *Escherichia coli* genomes can range from 4.6 to 5.7 Mb; *Procholoroccus marinus*, from 1.6 to 2.7 Mb; *Pseudomonas fluorescens*, from 6.3 to 7.1 Mb, or *Buchnera aphidicola,* from 0.4 to 0.7 Mb. All these examples illustrate that genome size is highly variable in prokaryotes and that, contrary to eukaryotes, it may change drastically even within a short divergence time. Variations in genome size are mostly related to the bacteria lifestyle, although different mechanisms seem to be acting in free-living and host-dependent bacteria [5]. Regarding gene content, and also contrary to eukaryotes, bacteria and archaea genomes tend to be very compact, with genes occupying around 90% of the genome for most of the species [6]. Intergenic regions (IGRs) are usually short but variable in size, with mean values as small as 3 bp in *Pelagibacter ubique*, 85 bp in *E. coli*, or 151 bp in *Yersinia pestis* [7]. Therefore, in general, genome size shows a strong positive correlation with gene number, although exceptions to this compactness are accumulating in recent years. Many sequenced bacterial genomes present a lower gene density than average due to the accumulation of different types of sequences that, similarly to eukaryotic genomes, can be considered "junk DNA". Some organisms possess a large quantity of pseudogenes, as it is the case of *Mycobacterium leprae* [8], *Serratia symbiotica* SAp [9] or *Sodalis glossinidius* [10,11], while others, such as *S. symbiotica* SCs [12] or *Rickettsia prowazekii* [13], show long IGRs. There are also genomes where genetic parasites, such as transposable elements and bacteriophages, are highly present, as it is the case of SOPE [14]. In this review we will not take into consideration bacteriophages as junk DNA because it has been shown that genes of a phage origin have been co-opted by different bacteria for their own profit. Thus, bacteriophages are known to carry key virulence factors for pathogenic bacteria [15] and, although

their roles in symbiotic bacteria are less understood, Oliver *et al.* (2009) [16] suggested that a phage-encoded toxin is responsible for the protecting role of *Hamiltonella defensa*, a facultative symbiont of the pea aphid *Acyrtosiphon pisum,* against the attack of parasitoids. Furthermore, though several studies are trying to manage this problem [17–19], the present standards of annotation for genomes available in the databases make difficult to assess when a putative functional gene has a phage origin or is just a non-functional remnant of an ancestral phage gene [20].

1.2. Bacterial Endosymbionts as a Model

Our group has been involved for more than a decade in the study of the genomes of endosymbionts, bacteria that live inside specialized eukaryotic cells. Many bacteria live in obligate association with eukaryotes (see [21] and [22] for complete reviews on the subject). Based on the effect of the bacterium on its host, these relationships are classified as mutualistic, commensal or parasitic but, from the bacterial point of view, the basic requirements to successfully maintain this kind of association are the same. They need to overcome the physical, cellular and molecular barriers presented by the host to achieve survival, proliferation and the infection of a new host. Therefore, the evolutionary processes suffered by both pathogens and mutualists are very similar. The rich and stable environment provided by the host cells leads to dramatic changes in the bacterial genome composition due to a relaxation of the selective forces on genes that become non-essential in the new environment, alongside an increase in the effects of genetic drift caused by bottlenecking over bacterial generations. These changes include a gradual reduction in genome size as well as variation in the coding capacity and gene density, together with changes in the number of genes and pseudogenes, and in the presence of insertion sequences (IS). Although endosymbiotic bacteria that keep a long term relationship with their hosts have extremely reduced and (usually) compact genomes, without traces of phage remnants, sequences acquired through horizontal gene transfer (HGT) and ISs and with limited amounts of pseudogenes, newly acquired endosymbionts still retain many of these features in their genomes, probably reflecting an intermediate step before the massive genome reduction starts. To better understand the molecular and evolutionary bases of endosymbiosis, the genomes of bacteria in different stages of the symbiotic integration process have been analyzed. These genomes and their comparison with those from their free-living relatives can be also useful in understanding the fate of bacterial junk-DNA. Or, on the other hand, the analysis of junk DNA can be used as a marker of the integration process of endosymbiotic bacteria.

2. The Impact of Pseudogenes in Bacterial Genomes

Pseudogenes are inactivated copies of known genes, which present an erosion or disruption either of their reading frames or their regulatory regions. They are usually detected by comparative analysis with their functional orthologs in close relative species. The alignment of such orthologous genes allows the detection of mutations, insertions/deletions (indels), frameshifts, premature stop codons and insertion of transposable elements, all of which cause the appearing of a truncated protein, or the inactivation or loss of essential functional domains. However, it is not always easy to assess the functional status of each annotated coding region within a genome for several reasons. First of all, orthology cannot be detected above a threshold of divergence between sequences. Second, not all

frameshifts, indels, stop mutations and rearrangements will inactivate a gene and there is a wide length variation of homologous proteins within families. An additional problem is that there is no clear definition of what has to be considered a pseudogene and there are inconsistencies in the methods used in different studies, so that it is not always possible to compare the pseudogene content from different genomes. In some studies, any shortened open reading frame (ORF) is annotated as a pseudogene [23], while in others considerably shortened homologs are annotated as genes if they keep complete domains specifying some function [24]. Lerat and Ochman (2005) [25] considered a pseudogene only when more of 20% of the length of the primary sequence of the encoded protein was lost, while some other studies excluded from this category impairments in homology matches within a "cutoff" region at either end, considering that slightly shorter alignments can reflect functional protein [26]. The most widely used method to identify non-functional genes involves the comparison of nucleotide substitution rates at synonymous (K_s) and non-synonymous sites (K_a), using the K_a/K_s test [27]. Regions without functional constraints, such as pseudogenes, are expected to have K_a/K_s ratios not significantly different from one. However, comparative analyses can only detect pseudogenes in genes that have orthologs in other available genomes, and only if the inactivation is due to truncations or disruptions of the original ORF, which represents only a fraction of the potentially inactivated genes. Genes that have been inactivated by missense mutations or changes in regulatory regions will also remain undetected by these methods.

At the beginning of the Genomics era, pseudogenes were thought to be unusual in bacteria. This idea has completely changed now that it is possible to perform comparative genomics on different strains of the same species, on closely related species with different lifestyles or on those living in different environments. In fact, when the *E. coli* MG1665 genome was first reported, only one pseudogene was annotated among its 4288 coding regions, but recent studies revealed that this genome contains about 100 genes that retain less than 20% of the annotated sequence in other *E. coli* strains, some of which (at least) are likely to be non-functional [28]. Nevertheless, pseudogenes were early noticed in significant amounts in some bacterial pathogens such as *R. prowazekii* [13] or *M. leprae* [8], the latter being one of the most extreme cases known to date, with 1614 protein-coding genes and 1133 annotated pseudogenes, 40% of its 3.2-Mb genome. Since then, several comprehensive studies have been performed trying to extract general rules that explain the dynamics of pseudogenes within bacterial genomes.

Liu *et al.* (2004) [26] analyzed 64 prokaryote genomes and defined a conservative set of conditions to detect their proportion of pseudogenes. In this analysis, they do not include hypothetical or putative proteins, because a large proportion of them could be over-annotated. However, in order to maximize the efficiency of pseudogene finding, they searched for remnants of ancient genes in the regions that had been considered as IGRs in the original annotations. Using this approach, they found that pseudogenes are pervasive in prokaryotes, accounting for 1 to 5% of most analyzed genomes. They didn't find a clear correlation between percentage of pseudogenes and lifestyle: taking aside the extreme case of *M. leprae*, the pseudogene fraction of the analyzed archaea, non-pathogenic and pathogenic bacteria were fairly similar (3.6, 3.9 and 3.3% respectively). However, and as mentioned in the introduction, it has been shown later on that some bacteria that have recently adopted an obligate host-dependent lifestyle (either mutualistic or parasitic) present higher amounts of pseudogenes, probably related with a rapid change to the new environments in which some previously encoded

functions are no longer needed. The rapid adaptation to new environments also explains the situation of *M. leprae* and its differences with its close relative *M. tuberculosis* [29] in that, with 3959 protein-coding genes, has only six identified pseudogenes. A comparative analysis of both genomes suggested that the pseudogenes in *M. leprae* have degenerated by gene-by-gene inactivation mostly after the divergence of these two clades [30]. Additionally, genomes from host-dependent symbionts that are suffering a genome reductive syndrome are known to contain many pseudogenes in regions that were previously annotated as IGRs. Thus, when the genome of *R. prowazekii*, the α-proteobacteria that causes typhus was sequenced [13], it was found that a considerable portion of its 1.1-Mb genome (22.9%) is covered by ncDNA and pseudogenes. A comparative analysis of this genome with the genome of three representative species of the genus *Rickettsia* showed that most of the IGRs were, in fact, remnants of ancestral genes with a high degree of degradation [24,31]. Even in smaller genomes, such as those from four different strains of *B. aphidicola* (the bacterial endosymbiont of aphids), longer IGRs in one strain contain the remnants of lost genes in other lineages [32].

As it would be expected, the pseudogenes identified by Liu *et al.* (2004) [26] cluster mainly in specific families related with environmental responses (such as specific nutrient transporters or processing, and related with antigenic variation). Other disabled genes correspond to hypothetical and unknown proteins, belong to transposable elements and bacteriophages, or are remnants of horizontally-transferred genes. Later on, Lerat and Ochman (2005) [25] determined the pseudogene content of 11 available complete genomes from four bacterial genera (*Staphylococcus, Streptococcus, Yersinia* and *Vibrio*), each of which include at least one human pathogen, which were supposed to accumulate pseudogenes compared with their free-living relatives. In addition to the gene families identified in the previous study [26], in obligate host-associated bacteria several broadly distributed genes involved in nucleotide processing, repair or replication appear as pseudogenes. The loss of genes needed for DNA repair and recombination are also among the first losses detected in bacteria that have recently acquired an obligate endosymbiotic lifestyle [33], thus contributing to the accumulation of pseudogenes in these species [34].

3. IS Elements Shaping Bacterial Genomes

Mobile genetic elements, such as phages, plasmids and transposable elements are widespread in prokaryotes, where they can represent a significant proportion of their DNA and play important roles in shaping their genomes. The simplest and most abundant transposable elements in bacteria are ISs. Although several classification schemes have been proposed, the one defined by Mahillon and Chandler (1998) [35] has become the most widely used. It is the format used by ISfinder (www-is.biotoul.fr) [36], a repository of ISs isolated from bacteria and archaea that also provides extensive background information, as well as an updated and comprehensive classification in IS families and subfamilies, and a proposal for a coherent nomenclature. These small elements are very compact and consist on a short DNA sequence, usually between 0.6 and 2.5 kb in length, able to translocate within and among replicons. A typical IS element only codes for proteins involved in its transposition, flanked by short inverted repeats (IR) of around 10 to 40 bp. Usually, their transposition generates a small duplication (2 to 14 bp) of the target DNA flanking the insertion point. Based on similarities in their transposases and IRs, and the length of their target site sequence, ISs can be

grouped into 20 major families. IS elements are important factors involved in genetic variability, since they can promote genomic rearrangements, and generate mutations by their insertion within genes or regulatory sequences. This capacity to generate genome damage is probably one of the reasons why transposition is usually strongly regulated and maintained at a low level in free-living bacteria [37]. In fact, most bacterial genomes contain only a few copies of a limited number of IS types [38]. However, in bacteria that have recently adopted an intracellular lifestyle, there is a reduction in the selective pressure to keep the ISs regulated, and a massive expansion of those elements can take place (see below).

As it happens with the pseudogenes, the quality of IS annotation in sequenced genomes is highly heterogeneous due to the different annotation methods and nomenclature used by different researchers, and the diverse interests that drive the functional analysis of a given genome. In fact, it is not uncommon that annotation focuses only on the potential protein-coding sequences included in the elements, but ignores their IR boundaries and the directed repeats generated by their insertions, as well as the vestiges of ancestral ISs [36]. Nevertheless, several global surveys of available bacterial genomes have been performed in recent years, allowing conclusions to be obtained regarding the distribution, dynamics and evolution of these elements.

Many different hypothesis have been proposed to explain the abundance and variability of IS elements among prokaryote genomes. Wagner *et al.* (2007) [38] used IScan, a free open source package, to examine the more than 2000 IS elements present in 438 completely sequenced bacterial genomes in the most comprehensive analysis to date. They found a high homogeneity pattern of IS families across vast taxonomic scales, consistent with previous and more limited works [39–41]. Such high-sequence homogeneity could be explained by the rapid spreading of ISs within a genome, in addition to other genetic mechanisms such as gene conversion. In the evolutionary scenario proposed by Wagner *et al.*, after an IS enters a genome, its copy number expands rapidly through transposition. Consequently, there is a low degree of diversity among the different copies of an IS present in a given genome. Eventually, due to the deleterious effect of their accumulation, IS elements tend to disappear and may become extinct from the lineage. However, later on, it may be reintroduced through HGT. The impact of HGT in the spreading of these sequences is revealed by the presence of closely related IS elements of the same family in non-closely-related genomes. However, HGT appears to be necessary but not sufficient for the presence of ISs, since their abundance within a genome does not depend on the level at which genomes are invaded by the elements [42].

Touchon and Rocha (2007) [42] performed a comprehensive reannotation and analysis of the putative functional ISs present in 262 prokaryote genomes in order to test for IS family specificity, the influence of host genome size, pathogenicity, or human association in the IS abundance or density. They found that IS distribution in prokaryotic genomes strongly correlates only with genome size, probably due to a decrease on the density of highly deleterious insertion sites with genome size. They hypothesized that IS abundance is mostly determined by selection and, therefore, the effective population sizes of the microorganism would strongly influence IS abundance. Thus, the high increase in the amount of ISs that has been detected in some bacterial pathogens, even though they have smaller genomes, would be the consequence of a recent reduction of effective population sizes, not of their parasitic lifestyle. However, it must be taken into account that it is their host-dependent lifestyle that determines that many genes become redundant or superfluous, which also diminishes the effectiveness

of natural selection, because there are more sites that can be substrate of IS transposition without affecting the fitness of the microorganism.

On a more recent study, Newton and Bordenstein (2011) [43] analyzed 384 bacterial genomes on the light of their phylogenetic relationships, genome sizes and ecology, in order to test whether there is a correlation between any of these factors and the mobile elements density (which in this study included phages, plasmids and transposable elements). They found that the density of mobile DNA elements only correlates with bacterial ecology. ISs account for the majority of mobile DNA elements in nearly half of the analyzed genomes, and there is a significant variation on the amount of them depending on the bacterial lifestyles (free-living, facultative intracellular and obligate intracellular bacteria), and also between horizontally and vertically transmitted obligate intracellular bacteria. The increase in the amount of IS elements in bacteria that have recently evolved as specialized pathogens or acquired an intracellular lifestyle had already been noticed in previous studies [44]. Since these bacteria sequestered inside eukaryotic cells can not exchange material with other bacteria through HGT, the massive presence of ISs must be due to an increase in the replicative transposition of elements that were resident at the onset of the obligate symbiosis [45], when many genes that had become non-essential can accumulate ISs without a detrimental effect. In addition, the high abundance of IS elements is also an important source of chromosomal rearrangements suffered by these genomes at this point [46]. Some studies emphasize that the persistence of IS elements in bacterial genomes can not be only explained by the balance of their propagation as selfish elements and the control of the host genome to avoid deleterious effects, but also due to their ability to promote adaptive evolution of the host genomes by generating beneficial mutations that increase the fitness of the host [47]. However, such benefits do not seem to be acting on extremely reduced genomes with a large evolutionary history with their hosts, since IS elements have been completely lost in their genomes. Their absence in the latter stages of the endosymbiosis can mostly explain the extreme chromosomal stasis observed in the genomes of the different *B. aphidicola* strains that have been sequenced [48].

4. Junk DNA as a Marker of the Symbiotic Integration Process

In recent years, new information has been obtained by the study of bacterial endosymbionts in different stages of their relationship with their respective hosts. Several cases are especially useful to illustrate the dynamics of junk DNA in the genomes of bacteria that acquire a specialized lifestyle and will be analyzed in detail in this section and summarized in Table 1.

4.1. Serratia symbiotica, the Missing Link from Free-Living to Obligate Mutualism

S. symbiotica is a γ-proteobacterium that appears in symbiotic association with different aphid species. In most cases, as for the pea aphid *A. pisum,* it appears as a facultative symbiont (strain SAp) [9]. However, in the case of the cedar aphid *Cinara cedri*, *S. symbiotica* SCc has established a permanent and obligate consortium with *B. aphidicola* [12], the primary endosymbiont of aphids. Therefore, S. *symbiotica* SCc is present in all cedar aphid populations and both bacteria maintain a metabolic complementation for the provision of some essential nutrients, such as tryptophan, as revealed by functional analyses of their genomes. It is worth noticing that *B. aphidicola* BCc from *C. cedri* possess the smallest genome among all the sequenced *Buchnera* strains [52].

Table 1. Relevant genomic features of selected bacteria with different lifestyles

Species	Lifestyle	Genome Size (kb)	CDS	Pseudogenes	IGR Mean Size (bp)	% Coding Density	Presence of ISs (%)	Data Source
M. leprae TN	human parasite	3,268	1614	1293	ND	ND	Yes	[30,49]
M. tuberculosis H37Rv	human parasite	4,411	4006	6*	ND	ND	1.5	[29,44,50]
S. symbiotica SAp	pea aphid facultative symbiont	2,789	2098	550	204.3	60.9	Yes	[9,12]
S. symbiotica SCc	cedar aphid facultative symbiont	1,763	672	58	1672.01	38.7	No	[12]
S. proteamaculans	free-living	5,496	4942	12	165.67	87.1	Yes	[12]
B. aphidicola BAp	pea aphid obligate endosymbiont	656	574	1	126.9	86.7	No	[12,51]
B. aphidicola BCc	cedar aphid obligate endosymbiont	422	362	3	135.8	90.0	No	[52]
S. glossinidius str. "morsitans"	tse-tse fly facultative symbiont	4,293	2516	1501	ND	50.9	2.72	[10,11]
*W. pipientis w*Mel	*D. melanogaster* reproductive parasite	1,268	1270	94	ND	80	7.7	[44,53]
*W. pipientis w*Ri	*D. simulans* reproductive parasite	1,445	1150	114	ND	ND	10	[54]
*W. pipientis w*Pip	*C. pipiens* reproductive parasite	1,482	1386	97	ND	82	Yes	[55]
*W. pipientis w*Bm	*B. malayi* obligate endosymbiont	1,080	806	98	ND	67	5.4**	[56]
R. prowazekii str. Madrid E	human parasite	1,112	835	12	ND	76	0.3	[13]

* Excluding IS elements. ** Various repeats including ISs. ND: non-determined

The genome comparison of both obligate and facultative *S. symbiotica* strains and other free-living *Serratia* reveals the gradual changes in gene content at different stages in the transition from free-living to endosymbiosis. *S. symbiotica* SCc has a moderately reduced genome (36.8%), compared to *S. symbiotica* SAp, and a 67.7% reduction compared to free-living *Serratia* such as *S. proteamaculans*. However, the data indicating genome decay are more extreme when the coding capacity is compared: the obligate strain SCc presents only 672 protein-coding genes and 58 pseudogenes, whereas the facultative strain SAp has 2098 genes and 550 pseudogenes. Thus, the overall coding density of strain SCc is 38.7%, the lowest among insect endosymbionts described so far. As a consequence, the IGRs are very long in this strain, with an average length of 1,672 bp, the highest among all endosymbionts analyzed. No traces of homology have been found in these IGRs compared to coding regions of other sequenced bacterial genomes. This ncDNA must represent ancient pseudogenes that are being gradually eroded until their total (or almost total) disappearance, as it has happened in ancient obligate bacterial endosymbionts [57]. Altogether, these data indicate that this genome is in the last steps of genomic degradation but previous to that of long-term endosymbionts

such as *B. aphidicola*. In fact, if we substitute the size of the IGRs in *S. symbiotica* SCc for the size of these regions in *B. aphidicola* BCc (135.8 bp on average), the chromosomal length would be 771,075 bp, in the range of other obligate endosymbionts. In addition, no IS sequences remain in this strain, whereas they have been described in SAp as well as in free-living *Serratia*. An analysis of the genome synteny between both symbionts revealed that an important number of rearrangements have occurred in both *S. symbiotica* lineages: This fact is in accordance with the possibility of the presence of active mobile elements in their ancestor at the onset of the symbiosis, which are currently unidentifiable in the *S. symbiotica* SCc genome [12].

4.2. Wolbachia pipiensis, from Reproductive Parasite to Intracellular Mutualist

One of the cases most extensively studied correspond to different strains of *W. pipiensis*, an α-proteobacterium that has established parasitic and mutualistic symbiosis with different animal hosts [58]. Most *W. pipientis* strains are reproductive parasites of insects. The complete genome of the strains that parasite *Drosophila melanogaster* (*w*Mel) [53], *D. simulans* (*w*Ri) [54], and mosquitoes from the *Culex pipiens* group (*w*Pip) [55,59] are available. Additionally, *W. pipiensis w*Bm was identified as the obligate mutualistic symbiont of *Brugia malayi*, a human filarial parasitic nematode, and its genome was also sequenced [56]. All these bacteria have streamlined genomes compared with other free-living α-proteobacteria, as expected for host-associated microorganism, being more extreme in the case of the mutualistic strain *w*Bm. However, all of them also contain high levels of repetitive DNA and mobile elements. When the genome of the strain *w*Mel was determined [53], it was found that the mobile DNA identified didn't have homologues in other α-proteobacteria but have been found in other *Wolbachia* strains. Therefore, the authors proposed that it was acquired short after the separation of the *Wolbachia* and *Rickettsia* lineages but before the radiation of the *Wolbachia* group. Another study revealed that most of the differences in genome size between *w*Mel with *w*Pip come from repetitive and mobile elements (not only ISs, but also prophages), and there is almost no overlapping between both genomes, which might be due to a recent invasion after the separation of both lineages or to different losses after such event [55]. Although many of the mobile elements seem to be defective, they must have played a key role in shaping the evolution of this genus, and are probably responsible for most of the genome rearrangements found among strains and compared with free-living relatives, similar to what has happened in *S. symbiotica* [12]. The comparison with *W. pipiensis w*Bm revealed that it also shares some repetitive elements with *w*Mel, although they are considerably less abundant in the mutualistic symbiont (5.4%) [56]. Additionally, the *w*Bm genome has an extremely low density of predicted functional genes, similar to what has been found in *R. prowazekii* [13] and *M. leprae* [8]. As in these other genomes, the main reason for such low-coding density is the presence of a considerable number of pseudogenes (which, even so, were underestimated at that time, since many putative ORFs correspond to fragmented former genes).

4.3. The Sodalis-like Group of Symbionts, at the Transition Point from Facultative to Obligate Symbionts

Phylogenetic analysis demonstrated the close relationship between *S. glossinidius*, secondary symbiont of the tse-tse fly and the primary endosymbiont of grain weevils of the genus *Sitophilus*, as it

is the case of SOPE and SZPE, endosymbionts of rice and maize weevils, *S. oryzae* and *S. zeamays*, respectively. *Sitophilus* symbionts and *S. glossinidius* provide a good model for the study of the evolutionary transition from facultative to obligate-mutualism, because the divergence of these lineages from a common ancestor has produced two different symbiotic outcomes. The genome sequencing and reannotation of *S. glossinidius*, showed that about a third of it is composed of inactivated genes in different degrees of disintegration, and it also contains a limited amount of IS elements of five different types, representing 2,72% of the genome [10,11]. Among the 1501 identified pseudogenes, only 18 were originated by the insertion of an IS element. This big amount of pesudogenes confirms the very recent symbiotic association with their insect host [60,61], while it is clear that ISs have not been determinant in the degeneration of the genome.

Although the genome sequence of SOPE is not yet available, preliminary studies showed that this genome has been massively invaded by IS elements, which have been estimated to represent more than 20% of the genome [14]. This is the most extreme case of IS abundance in any known bacterial genome. The presence of a Sodalis-like γ-proteobacteria in cereal weevil lineages has been explained by a replacement of the ancestral endosymbiont *Nardonella* (present in the rest of the members of the family to which the rice weevil belongs) less than 25 million years ago [62,63], thus being a very young obligate mutualist. Therefore, it seems that after the establishment of the obligate symbiosis, the evolutionary path taken by SOPE differs from the observed in *S. glossinidius*, with a massive proliferation of at least four types of IS elements. Only two of these elements are shared with *S. glossinidius,* which as in the case of *Wolbachia*, indicates that they were present before the split of the two lineages, and the other two must have either been acquired by HGT in the ancestor of SOPE, or lost in *S. glossinidius*. The importance of the ISs in the process of pseudogenization can not be completely estimated until the full genome is available. Although it is known that many genes are interrupted by ISs, it is first necessary to determine if the interruption occurred on genes that were already pseudogenized in the common ancestor of *Sodalis* and SOPE, which seems to be the case. Preliminary studies performed in our group indicate that many genes inactivated by IS insertions encode hypothetical proteins or proteins of unknown function, as well as transposases and bacteriophage related proteins [14] and that a high proportion of them appear already as pseudogenes in *Sodalis* (unpublished results). In any case, it is clear that ISs must be involved in genomic recombination events leading to the loss of the region between two elements. Thus, similar to what is found in SOPE, ISs must be a key factor in the genome degradation that occurs in the first stages of integration to intracellular life even though the random pseudogenization of genes that have become non-essential started before their proliferation.

5. Concluding Remarks: Dynamics of Junk-DNA during the Evolutionary Reduction Process

Comparative studies have revealed that bacterial genomes are under selective pressures that have a deep impact on their shape and gene content, including a previously unexpected degree of diversity in gene repertoires within and between species. In addition to a set of genes that are shared by all members of a monophyletic group, there are a considerable number of genes and other functional features that can be highly specific depending on the environmental context. The evolution of these gene repertoires can be explained by processes of gene gains through HGT and duplications, and gene

loses, through pseudogenization and gene excision [28]. Since it is well known that HGT has a great impact on prokaryotes [64], reductive evolutionary processes due to gene degradation and elimination must also occur very frequently in order to maintain compact genomes [31,65]. Pseudogenes must be quickly removed from the genomes, as it can be deduced from the small proportion of them those that are present simultaneously in different strains. When mutations occur in genes that are no longer needed as an adaptation to particular living conditions, the inactivated genes can remain in the genome for some time and they gradually erode in a random manner until they are completely removed [57,66], because within bacteria (as well as within several eukaryotes) there is a mutational bias toward deletions over insertions [31,67–69].

The dynamics of degradation and elimination of junk sequences has been studied in intracellular specialists by comparative genomics with their free-living relatives (Figure 1). Free-living bacteria have larger genomes with a moderate content of repeated sequences and self-propagating DNA, such as transposons, bacteriophage, a moderate amount of pseudogenes (usually between 1% and 5%) and relatively short IGRs. In the transition from free-living to host-dependent lifestyles, junk DNA increases their content, leading some times to an expansion of genome size parallel to a decrease in gene-coding capacity. The accumulation of pseudogenes in different degrees of decay (sometimes leading to long IGRs in which no traces of former pseudogenes can be detected except based on genome synteny studies), as well as the accumulation and losses of IS elements in the genomes of these bacteria can be explained by genetic and population factors. At the beginning of the association, the new rich, protected and stable niche provided by the host makes unnecessary or redundant some gene functions. Due to a decreased efficiency of the purifying selection, they can rapidly accumulate slightly deleterious mutations on genes that have become non-essential. The inefficiency of the purifying selection also explains the increase in the amount of IS elements present in these genomes at the onset of symbiosis, which in some cases seems to be accompanied by an increase in the transposition rate and the concomitant production of pseudogenes created by IS insertions. At the same time, the random genetic drift increases due to a drastic reduction of the bacterial effective population size when they are transmitted from one host to another. The uncontrolled proliferation of IS elements can lead to the loss of large stretches of genomic DNA by unequal recombination when they appear in direct orientation. This was one of the predicted effects in the model for the reductive syndrome of endosymbiont genomes, which takes place in two stages: first, a massive reduction and a second stage of pseudogenization and progressive disappearing of non-functional sequences by gene-by-gene erosion and deletion, leading to the highly compact and streamlined genomes of symbionts with a long established obligatory intracellular relationship with their hosts. IS elements also suffer the accumulation of mutations that render them inactive for transposition, thus becoming real junk DNA. The process occurs in a random manner so that some genomes can accumulate dramatic amounts of ISs (as in the case of SOPE) or lose most of them at the end of the first stage. The possibility for rapid genome degradation by means of deletion events is unlikely in the latter genomes and therefore, as it is the case of *S. symbiotica* SCc, they enter the second step of gene-by-gene degeneration while still retaining great amounts of pseudogenes which, with evolutionary time, will appear as long IGRs.

Figure 1. Dynamics of gain and loss of junk DNA in bacteria that establishes a host-dependent lifestyle. Details are given in the text. Orange arrows: essential genes; pink arrows: non-essential genes; white arrows: pseudogenes; blue and green boxes ISs that are active (dark colour) or inactive (light colour); light purple boxes: IGRs.
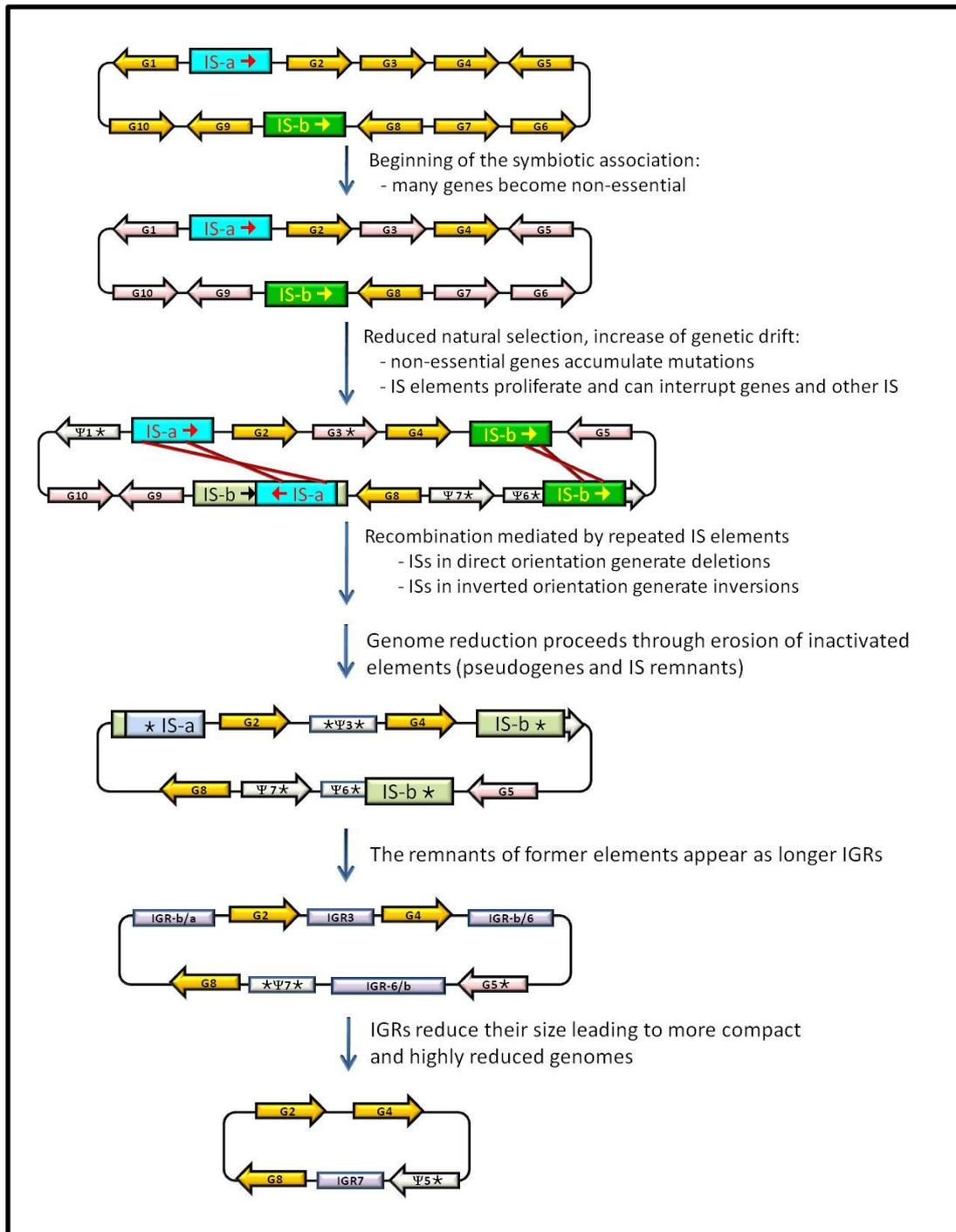
References and Notes

1. Gregory, T.R.; Nicol, J.A.; Tamm, H.; Kullman, B.; Kullman, K.; Leitch, I.J.; Murray, B.G.; Kapraun, D.F.; Greilhuber, J.; Bennett, M.D. Eukaryotic genome size databases. *Nucleic Acids Res.* **2007**, *35*, D332–D338.
2. Gregory, T.R. Synergy between sequence and size in large-scale genomics. *Nat. Rev. Genet.* **2005**, *6*, 699–708.
3. Ohno, S. So much "Junk" DNA in our genome. *Brookhaven Symp. Biol.* **1972**, *23*, 366–370.
4. Orgel, L.E.; Crick, F.H. Selfish DNA: The ultimate parasite. *Nature* **1980**, *284*, 604–607.
5. Delaye, L.; Gil, R.; Pereto, J.; Latorre, A.; Moya, A. Life with a few genes: A survey on naturally evolved reduced genomes. *Open Evol. J.* **2010**, *4*, 12–22.
6. Achaz, G.; Coissac, E.; Netter, P.; Rocha, E.P. Associations between inverted repeats and the structural evolution of bacterial genomes. *Genetics* **2003**, *164*, 1279–1289.
7. Giovannoni, S.J.; Tripp, H.J.; Givan, S.; Podar, M.; Vergin, K.L.; Baptista, D.; Bibbs, L.; Eads, J.; Richardson, T.H.; Noordewier, M.; *et al.* Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **2005**, *309*, 1242–1245.
8. Cole, S.T.; Eiglmeier, K.; Parkhill, J.; James, K.D.; Thomson, N.R.; Wheeler, P.R.; Honore, N.; Garnier, T.; Churcher, C.; Harris, D.; *et al.* Massive gene decay in the leprosy bacillus. *Nature* **2001**, *409*, 1007–1011.
9. Burke, G.R.; Moran, N.A. Massive genomic decay in *Serratia symbiotica*, a recently evolved symbiont of aphids. *Genome Biol. Evol.* **2011**, *3*, 195–208.
10. Toh, H.; Weiss, B.L.; Perkin, S.A.; Yamashita, A.; Oshima, K.; Hattori, M.; Aksoy, S. Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. *Genome Res.* **2006**, *16*, 149–156.
11. Belda, E.; Moya, A.; Bentley, S.; Silva, F.J. Mobile genetic element proliferation and gene inactivation impact over the genome structure and metabolic capabilities of *Sodalis glossinidius*, the secondary endosymbiont of tsetse flies. *BMC Genomics* **2010**, *11*, 449.
12. Lamelas, A.; Gosalbes, M.J.; Manzano-Marin, A.; Pereto, J.; Moya, A.; Latorre, A. *Serratia symbiotica* from the aphid *Cinara cedri*: A missing link from facultative to obligate insect endosymbiont. *PLoS Genet.* **2011**, *7*, e1002357.
13. Andersson, S.G.; Zomorodipour, A.; Andersson, J.O.; Sicheritz-Ponten, T.; Alsmark, U.C.; Podowski, R.M.; Naslund, A.K.; Eriksson, A.S.; Winkler, H.H.; Kurland, C.G., The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **1998**, *396*, 133–140.
14. Gil, R.; Belda, E.; Gosalbes, M.J.; Delaye, L.; Vallier, A.; Vincent-Monegat, C.; Heddi, A.; Silva, F.J.; Moya, A.; Latorre, A. Massive presence of insertion sequences in the genome of SOPE, the primary endosymbiont of the rice weevil *Sitophilus oryzae*. *Int. Microbiol.* **2008**, *11*, 41–48.
15. Brussow, H.; Canchaya, C.; Hardt, W.D. Phages and the evolution of bacterial pathogens: From genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev.* **2004**, *68*, 560–602.
16. Oliver, K.M.; Degnan, P.H.; Hunter, M.S.; Moran, N.A., Bacteriophages encode factors required for protection in a symbiotic mutualism. *Science* **2009**, *325*, 992–994.
17. Lima-Mendez, G.; Van Helden, J.; Toussaint, A.; Leplae, R. Prophinder: A computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics* **2008**, *24*, 863–865.

18. Zhou, Y.; Liang, Y.; Lynch, K.H.; Dennis, J.J.; Wishart, D.S. PHAST: A fast phage search tool. *Nucleic Acids Res.* **2011**, *39*, W347–W352.

19. Akhter, S.; Aziz, R.K.; Edwards, R.A. *PhiSpy*: A novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res.* **2012**, doi: 10.1093/nar/gks406.

20. Canchaya, C.; Proux, C.; Fournous, G.; Bruttin, A.; Brussow, H. Prophage genomics. *Microbiol. Mol. Biol. Rev.* **2003**, *67*, 238–276.

21. Moya, A.; Pereto, J.; Gil, R.; Latorre, A., Learning how to live together: Genomic insights into prokaryote-animal symbioses. *Nat. Rev. Genet.* **2008**, *9*, 218–229.

22. Moran, N.A.; McCutcheon, J.P.; Nakabachi, A. Genomics and evolution of heritable bacterial symbionts. *Annu. Rev. Genet.* **2008**, *42*, 165–190.

23. Chain, P.S.; Carniel, E.; Larimer, F.W.; Lamerdin, J.; Stoutland, P.O.; Regala, W.M.; Georgescu, A.M.; Vergez, L.M.; Land, M.L.; Motin, V.L.; *et al.* Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia. pseudotuberculosis*. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 13826–13831.

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

25. Lerat, E.; Ochman, H. Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Res.* **2005**, *33*, 3125–3132.

26. Liu, Y.; Harrison, P.M.; Kunin, V.; Gerstein, M. Comprehensive analysis of pseudogenes in prokaryotes: Widespread gene decay and failure of putative horizontally transferred genes. *Genome Biol.* **2004**, *5*, R64.

27. Nei, M.; Kumar, S. *Molecular Evolution and Phylogenetics*; Oxford University Press: New York, NY, USA, 2000.

28. Abby, S.; Daubin, V. Comparative genomics and the evolution of prokaryotes. *Trends Microbiol.* **2007**, *15*, 135–141.

29. Cole, S.T.; Brosch, R.; Parkhill, J.; Garnier, T.; Churcher, C.; Harris, D.; Gordon, S.V.; Eiglmeier, K.; Gas, S.; Barry, C.E., 3rd; *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **1998**, *393*, 537–544.

30. Gomez-Valero, L.; Rocha, E.P.; Latorre, A.; Silva, F.J. Reconstructing the ancestor of *Mycobacterium leprae*: The dynamics of gene loss and genome reduction. *Genome Res.* **2007**, *17*, 1178–1185.

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

32. Perez-Brocal, V.; Latorre, A.; Gil, R.; Moya, A. Comparative analysis of two genomic regions among four strains of *Buchnera aphidicola*, primary endosymbiont of aphids. *Gene* **2005**, *345*, 73–80.

33. Dale, C.; Wang, B.; Moran, N.; Ochman, H. Loss of DNA recombinational repair enzymes in the initial stages of genome degeneration. *Mol. Biol. Evol.* **2003**, *20*, 1188–1194.

34. Gil, R.; Latorre, A.; Moya, A. Evolution of prokaryote-animal symbiosis from a genomics perspective. In *(Endo) symbiotic Methanogenic Archaea*; Hackstein, J.H.P., Ed. Springer-Verlag: Berlin, Germany, 2010; pp. 207–233.

35. Mahillon, J.; Chandler, M. Insertion sequences. *Microbiol. Mol. Biol. Rev.* **1998**, *62*, 725–774.

36. Siguier, P.; Perochon, J.; Lestrade, L.; Mahillon, J.; Chandler, M. ISfinder: The reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **2006**, *34*, D32–36.

37. Nagy, Z.; Chandler, M. Regulation of transposition in bacteria. *Res. Microbiol.* **2004**, *155*, 387–398.

38. Wagner, A.; Lewis, C.; Bichsel, M. A survey of bacterial insertion sequences using IScan. *Nucleic Acids Res.* **2007**, *35*, 5284–5293.

39. Sawyer, S.A.; Dykhuizen, D.E.; DuBose, R.F.; Green, L.; Mutangadura-Mhlanga, T.; Wolczyk, D.F.; Hartl, D.L. Distribution and abundance of insertion sequences among natural isolates of *Escherichia coli*. *Genetics* **1987**, *115*, 51–63.

40. Lawrence, J.G.; Ochman, H.; Hartl, D.L. The evolution of insertion sequences within enteric bacteria. *Genetics* **1992**, *131*, 9–20.

41. Wagner, A. Periodic extinctions of transposable elements in bacterial lineages: Evidence from intragenomic variation in multiple genomes. *Mol. Biol. Evol.* **2006**, *23*, 723–733.

42. Touchon, M.; Rocha, E.P. Causes of insertion sequences abundance in prokaryotic genomes. *Mol. Biol. Evol.* **2007**, *24*, 969–981.

43. Newton, I.L.; Bordenstein, S.R. Correlations between bacterial ecology and mobile DNA. *Curr. Microbiol. 62* **2010**, 198–208.

44. Bordenstein, S.R.; Reznikoff, W.S. Mobile DNA in obligate intracellular bacteria. *Nat. Rev. Microbiol.* **2005**, *3*, 688–699.

45. Dale, C.; Moran, N.A. Molecular interactions between bacterial symbionts and their hosts. *Cell.* **2006**, *126*, 453–465.

46. Moran, N.A.; Plague, G.R. Genomic changes following host restriction in bacteria. *Curr. Opin. Genet. Dev.* **2004**, *14*, 627–633.

47. Schneider, D.; Lenski, R.E. Dynamics of insertion sequence elements during experimental evolution of bacteria. *Res. Microbiol.* **2004**, *155*, 319–327.

48. Latorre, A.; Gil, R.; Silva, F.J.; Moya, A. Chromosomal stasis *versus* plasmid plasticity in aphid endosymbiont *Buchnera aphidicola*. *Heredity* **2005**, *95*, 339–347.

49. Singh, P.; Cole, S.T. *Mycobacterium leprae*: Genes, pseudogenes and genetic diversity. *Future Microbiol. 6* **2010**, 57–71.

50. Camus, J.C.; Pryor, M.J.; Medigue, C.; Cole, S.T. Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiol.* **2002**, *148*, 2967–2973.

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

52. Perez-Brocal, V.; Gil, R.; Ramos, S.; Lamelas, A.; Postigo, M.; Michelena, J.M.; Silva, F.J.; Moya, A.; Latorre, A. A small microbial genome: The end of a long symbiotic relationship? *Science* **2006**, *314*, 312–313.

53. Wu, M.; Sun, L.V.; Vamathevan, J.; Riegler, M.; Deboy, R.; Brownlie, J.C.; McGraw, E.A.; Martin, W.; Esser, C.; Ahmadinejad, N.; *et al.* Phylogenomics of the reproductive parasite *Wolbachia. pipientis w*Mel: A streamlined genome overrun by mobile genetic elements. *PLoS Biol.* **2004**, *2*, E69.

54. Klasson, L.; Westberg, J.; Sapountzis, P.; Naslund, K.; Lutnaes, Y.; Darby, A.C.; Veneti, Z.; Chen, L.; Braig, H.R.; Garrett, R*., et al.* The mosaic genome structure of the *Wolbachia. w*Ri strain infecting *Drosophila simulans*. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 5725–5730.

55. Klasson, L.; Walker, T.; Sebaihia, M.; Sanders, M.J.; Quail, M.A.; Lord, A.; Sanders, S.; Earl, J.; O'Neill, S.L.; Thomson, N*., et al.* Genome evolution of *Wolbachia* strain *w*Pip from the *Culex pipiens* group. *Mol. Biol. Evol.* **2008**, *25*, 1877–1887.

56. Foster, J.; Ganatra, M.; Kamal, I.; Ware, J.; Makarova, K.; Ivanova, N.; Bhattacharyya, A.; Kapatral, V.; Kumar, S.; Posfai, J.; *et al.* The *Wolbachia* genome of *Brugia malayi*: Endosymbiont evolution within a human pathogenic nematode. *PLoS Biol.* **2005**, *3*, e121.

57. Silva, F.J.; Latorre, A.; Moya, A. Genome size reduction through multiple events of gene disintegration in *Buchnera* APS. *Trends Genet.* **2001**, *17*, 615–618.

58. Lo, N.; Paraskevopoulos, C.; Bourtzis, K.; O'Neill, S.L.; Werren, J.H.; Bordenstein, S.R.; Bandi, C. Taxonomic status of the intracellular bacterium *Wolbachia pipientis*. *Int. J. Syst. Evol. Microbiol.* **2007**, *57*, 654–657.

59. Salzberg, S.L.; Puiu, D.; Sommer, D.D.; Nene, V.; Lee, N.H., Genome sequence of the *Wolbachia* endosymbiont of *Culex quinquefasciatus* JHB. *J. Bacteriol.* **2009**, *191*, 1725.

60. Heddi, A.; Charles, H.; Khatchadourian, C.; Bonnot, G.; Nardon, P. Molecular characterization of the principal symbiotic bacteria of the weevil *Sitophilus oryzae*: A peculiar G+C content of an endocytobiotic DNA. *J. Mol. Evol.* **1998**, *47*, 52–61.

61. Dale, C.; Plague, G.R.; Wang, B.; Ochman, H.; Moran, N.A. Type III secretion systems and the evolution of mutualistic endosymbiosis. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12397–12402.

62. Lefevre, C.; Charles, H.; Vallier, A.; Delobel, B.; Farrell, B.; Heddi, A. Endosymbiont phylogenesis in the Dryophthoridae weevils: Evidence for bacterial replacement. *Mol. Biol. Evol.* **2004**, *21*, 965–973.

63. Conord, C.; Despres, L.; Vallier, A.; Balmand, S.; Miquel, C.; Zundel, S.; Lemperiere, G.; Heddi, A. Long-term evolutionary stability of bacterial endosymbiosis in Curculionoidea: Additional evidence of symbiont replacement in the Dryophthoridae family. *Mol. Biol. Evol.* **2008**, 859–868.

64. Koonin, E.V.; Makarova, K.S.; Aravind, L. Horizontal gene transfer in prokaryotes: Quantification and classification. *Annu. Rev. Microbiol.* **2001**, *55*, 709–742.

65. Lawrence, J.G.; Hendrix, R.W.; Casjens, S., Where are the pseudogenes in bacterial genomes? *Trends Microbiol.* **2001**, *9*, 535–540.

66. Ochman, H.; Davalos, L.M. The nature and dynamics of bacterial genomes. *Science* **2006**, *311*, 1730–1733.

67. Mira, A.; Ochman, H.; Moran, N.A. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **2001**, *17*, 589–596.

68. Andersson, J.O.; Andersson, S.G. Insights into the evolutionary process of genome degradation. *Curr. Opin. Genet. Dev.* **1999**, *9*, 664–671.

69. Petrov, D.A.; Hartl, D.L. Pseudogene evolution and natural selection for a compact genome. *J. Hered.* **2000**, *91*, 221–227.