

FACULTAT DE CIÈNCIES BIOLÒGIQUES

DEPARTAMENT DE GENÈTICA

INSTITUTO CAVANILLES DE BIODIVERSITAT I

BIOLOGIA EVOLUTIVA



**PHYLOGENOMICS AND THE
EVOLUTION OF PROTEOBACTERIA**

Memoria presentada por Iñaki Comas Espadas para optar al
grado de doctor en ciencias Biológicas por la Universitat de
València

Directores:

Dr. Fernando González Candelas. Catedrático de la U.V.E.G

Dr. Andrés Moya Simarro. Catedrático de la U.V.E.G

Valencia 2007

COVER: Darwin's first sketch of an evolutionary tree from his *First Notebook on Transmutation of Species* (1837)

D. Andrés Moya Simarro, Doctor en Ciencias Biológicas y Catedrático del Departament de Genètica de la Universitat de València.

D. Fernando González Candelas, Doctor en Ciencias Biológicas y Catedrático del Departament de Genètica de la Universitat de València.

CERTIFICAN: Que Iñaki Comas Espadas, Licenciado en ciencias Biológicas por la Universitat de València, ha realizado bajo su dirección el trabajo que lleva por título: “Phylogenomics and the evolution of Proteobacteria”, para optar al Grado de doctor en Ciencias Biológicas por la Universitat de València.

Y para que conste, en el cumplimiento de la legislación vigente, firmo el presente certificado en Valencia, a 31 de Octubre de 2007.

Fdo.: Dr. Andrés Moya Simarro

Fdo.: Dr. Fernando González Candelas

AGRADECIMIENTOS

A la hora de hacer balance de un trabajo llevado a cabo durante cinco años lo único que uno puede hacer es acordarse de las personas que han transitado con él durante ese período. Pensar en todas ellas da alegría, pena y algo de nostalgia. Piensas en los que ya estaban cuando tu llegaste, los que llegaron contigo y ya no están, los que llegaron contigo y aún están, los que vinieron después y los recién llegados. Cómo os podéis honrar a mi fama de despistado) pero sí es así no lo hago con mala intención.

Empezaré pues por mis directores. Primero porque sin ellos no hubiera conocido a esa larga lista de personas, segundo porque de ellos es de los que más he aprendido. Andrés, te agradezco que te ofrecieras a ser mi director para así poder optar a una beca cuando todavía no me conocías de nada y te agradezco lo que te has preocupado por mí y mi futuro, por tu apoyo y tu interés. A Fernando porque fuiste tú quien me reclutó y me dio una oportunidad. Porque juntos hemos sufrido a revisores y editores, porque siempre encontrabas un momento para explicarme algo, porque siempre te preocupaste de buscar conmigo una solución cuando no la conocías, por corregir mi mal llamado inglés y por otras muchas cosas que aquí no caben y no creo que haga falta enumerar.

Gracias a toda la tropa de técnicos. Inma y Alma que estaban en mis comienzos como colaborador (Alma todavía debe temblar viéndome coger una pipeta). Pepa, Silvia y Nuria que siempre se han portado muy bien conmigo y que han contribuido a que todo fuera más fácil. A Pascual por toda la ayuda y consejos que me ha prestado.

Y por supuesto a los becarios que fueron, son y serán, a los que espero ver de doctores dentro de no mucho. A Alicia, pues sólo ella sabe lo que significa soportarme y cuya capacidad de trabajo y de aprender cosas nuevas es sorprendente. A Vicente Sentandreu con el que he compartido muchas cosas en estos cinco años, incluyendo la cama. No creo que haga falta decirte lo contento que estoy de haberte conocido. A Mireia, una de esas personas con tanta fuerza y que derrochan tanta personalidad que son capaces de conseguir todo lo que se proponen. A Yolima que todavía me recuerda que yo no le saludaba al principio y cuya amistad espero conservar estemos donde estemos los dos. A Laura, la persona más buena, incapaz de hacer daño a nadie que he conocido en mi vida. Gracias por tu apoyo, por escucharme y por ser como eres. A Vicente Pérez y a Eugeni, compañero de cafés, porque en estos años nos hemos reído mucho, y de muchos. A Vicky que empezó conmigo hace ya tanto tiempo. Y por supuesto a Teresa, Eugenia, Araceli, Pedro, Ana, Rafa, los Peris, Benja y Loreto, José, Lidia, Sergi...

To the Jeffreys Lawrence's lab in Pittsburgh, I had a great time there.
Thank you very much.

A Toni y David, compañeros de tantas cosas que sería imposible enumerar, ellos saben todo lo que les debo. A Cocheras y David por estar ahí y porque nadie sabe lo bien que nos lo hemos pasado juntos. A Elvira que ha sabido escucharme y con quien también me he reído de muchas cosas. Y también a Jeni, Juan y Diana, Use, Toni, Javi, Isa, Jorge...

A mi familia, porque todo lo que soy, para bien o para mal, es por ellos. Porque ellos me apoyaron y estuvieron siempre ahí. Espero poder devolverles algún día todo lo que les debo.

A mi gato Darwin, que nunca sabrá que lleva el nombre de un señor enterrado, mercedamente, al lado de Sir Isaac Newton.

TO THE HAPPY FEW

(The Red and the Black. Marie-Henri Beyle *Stendhal*, 1831)

Vivid, pues, y sed dichosos, hijos queridos de mi corazón, y no olvidéis nunca que hasta el día en que Dios se digne descifrar el porvenir al hombre, toda la sabiduría humana estará resumida en dos palabras: ¡Confiar y esperar!

*Vuestro amigo,
Edmundo Dantés,
Conde de Montecristo.*

(El Conde de Montecristo. Alejandro Dumas, 1844)

INDEX

PHYLOGENOMICS AND THE EVOLUTION OF PROTEOBACTERIA	3
1. GENERAL INTRODUCTION.....	17
1.1 NEW CONCEPTS IN MICROBIAL EVOLUTION AS REVEALED BY GENOME SEQUENCING PROJECTS	19
1.2 THE NATURE OF MICROBIAL EVOLUTION AND THE CONCEPT OF BACTERIAL SPECIES	26
1.3 GENE GAIN: PLURALITY OF PHYLOGENETIC SIGNALS IN MICROBIAL GENOMES	34
1.4 GENE LOSS: GENOME REDUCTION IN MICROBIAL EVOLUTION: GAMMA-PROTEOBACTERIA ENDOSYMBIONTS OF INSECTS AS A MODEL	37
1.5 PHYLOGENOMICS AND MICROBIAL EVOLUTION	40
1.6 A GLIMPSE ON THE GENOMES ANALYZED IN THIS THESIS	45
2. OBJECTIVES.....	49
3. FROM PHYLOGENETICS TO PHYLOGENOMICS: CURRENT TOOLS FOR THE EVOLUTIONARY ANALYSIS OF PROTEOBACTERIA GENOMES.....	53
3.1 INTRODUCTION.....	55
3.2 MATERIAL AND METHODS	61
3.2.1 Genomes and homologous genes selection.....	61
3.2.2 Obtaining the phylome of <i>Blochmannia floridanus</i>	66
3.2.3 Phylogenomic analyses	68
3.2.4 Phylogenomic cores analyses.....	72
3.3 RESULTS	75
3.3.1 Reference tree and search for putative orthologs	75
3.3.2 The phylome of <i>Blochmannia floridanus</i> and the 16S rDNA tree.....	77
3.3.3 Phylogenomic analyses	79
3.3.4 Phylogenomic cores.....	83
3.4 DISCUSSION.....	92
3.4.1 Comparison of methods: from phylogenetics to phylogenomics.....	92
3.4.2 Different phylogenomic data sets harbour different phylogenetic signals	97
3.5 CONCLUSIONS.....	102

4. THE PHYLOGENETIC LANDSCAPE OF GAMMA-PROTEOBACTERIA.....	105
4.1 INTRODUCTION	107
4.2 MATERIAL AND METHODS.....	109
4.2.1 <i>Data sets analyzed</i>	109
4.2.2 <i>Phylogenetic analyses under extreme conditions</i> ...	111
4.3 RESULTS	113
4.3.1 <i>The Gamma-Proteobacteria phylogenetic landscape</i>	113
4.3.2 <i>Phylogenomic analysis of Carsonella ruddii position</i>	115
4.4 DISCUSSION.....	120
4.4.1 <i>The Phylogenetic landscape of Proteobacteria: the placement of Xanthomonadales</i>	120
4.4.2 <i>The phylogenetic history of Gamma-Proteobacteria insect endosymbionts</i>	122
4.5 CONCLUSIONS.....	128
5. THE EVOLUTIONARY ORIGIN OF XANTHOMONADALES GENOMES AND THE NATURE OF THE HORIZONTAL GENE TRANSFER PROCESS	129
5.1 INTRODUCTION	131
5.2 METHODS	136
5.2.1 <i>Selection of homologs, gene alignments and gene trees</i> . 136	
5.2.2 <i>Congruence map analysis</i>	139
5.2.3 <i>Phylogenetic origin analysis</i>	139
5.2.4 <i>Testing for long-branch attraction artifacts</i>	140
5.2.5 <i>Testing for functional association and clustering along the genome</i>	140
5.2.6 <i>Detection of atypical genes</i>	142
5.3 RESULTS	143
5.3.1 <i>Data set analyzed and exploratory analyses</i>	143
5.3.2 <i>Congruence map</i>	145
5.3.3 <i>Atypical genes detection</i>	147
5.3.4 <i>Long-branch attraction artifact incidence</i>	148
5.3.5 <i>Phylogenetic origin test</i>	148

5.3.6	<i>Testing for functional association and clustering along the genomes</i>	150
5.4	DISCUSSION.....	153
5.4.1	<i>Xanthomonadales evolution illustrates the nature of the HGT process</i>	153
5.5	CONCLUSIONS.....	157
6.	FIGHTING FOR SURVIVAL: OPPOSING EVOLUTIONARY FORCES ACT IN THE LAST STAGES OF GENOME REDUCTION	159
6.1	INTRODUCTION.....	161
6.2	MATERIALS AND METHODS	164
6.2.1	<i>Putative orthologs analyzed and gene trees inference</i> 164	
6.2.2	<i>Evolutionary rates of endosymbiont genes</i>	165
6.2.3	<i>Whole genome A+T saturation measures</i>	166
6.2.4	<i>Positive selection (PS), relaxed constraints (RLC) or purifying selection (PUR)?</i>	168
6.2.5	<i>Testing for artefacts in the PS tests</i>	170
6.3	RESULTS	171
6.3.1	<i>Data set analyzed</i>	171
6.3.2	<i>Synonymous and nonsynonymous substitutions</i>	172
6.3.3	<i>Whole genome saturation measures</i>	175
6.3.4	<i>Positive selection, relaxed evolution and purifying selection</i>	178
6.3.5	<i>Testing for artefacts in PS detection</i>	181
6.4	DISCUSSION.....	185
6.5	CONCLUSIONS.....	192
7.	GENERAL DISCUSSION	195
7.1	PHYLOGENOMICS AND THE EVOLUTIONARY SIGNALS IN MICROBIAL GENOMES	200
7.1.1	<i>Influence of orthology assessment on phylogenomic analyses</i>	200
7.1.2	<i>Model-based methods of phylogenetic reconstruction</i> . 201	
7.1.3	<i>Supertrees, supermatrix and the signal detected</i>	202
7.1.4	<i>Incongruence and signal</i>	204

7.2	LESSONS ON MICROBIAL EVOLUTION FROM XANTHOMONADALES GENOMES.....	208
7.2.1	<i>Non-random genetic exchanges: internal and external factors</i> 210	
7.2.2	<i>A proposal for Bacteria and Archaea</i>	218
7.3	GENOME REDUCTION AND LIFESTYLE EVOLUTION IN MICROBIAL GENOMES	226
8.	GENERAL CONCLUSIONS	235
9.	REFERENCES	241
10.	BREVE RESUMEN EN CASTELLANO.....	273
11.	SUPPLEMENTARY INFORMATION.....	319

1. GENERAL INTRODUCTION

In this thesis we examine the nature of the evolutionary forces that shape the nucleotide sequence and contents of bacterial genomes and the best methodologies to achieve this objective. Hence, in this brief introduction we will introduce the most important features of bacterial genomes as revealed by sequencing projects and environmental genomics analyses. Next, we will deal with the two main evolutionary processes studied in this work: the mode of transmission of the genes and its influence on bacterial genome evolution and the dynamics of genome reduction in bacterial genomes. Finally, we will briefly review the methodologies that have allowed us to explore the incidence of these forces.

This introduction does not pretend to be an exhaustive description of the above topics as they will be analyzed in more depth in the corresponding chapters. Rather, our aim is to highlight the most significant advances in our knowledge of microbial genome evolution as derived from current studies based on comparative analyses of tens, even hundreds, of genomes. Although some aspects here referred will not be addressed directly in the following chapters, it is important to keep them in mind in order to achieve a comprehensive view of microbial evolution.

1.1 New concepts in microbial evolution as revealed by genome sequencing projects

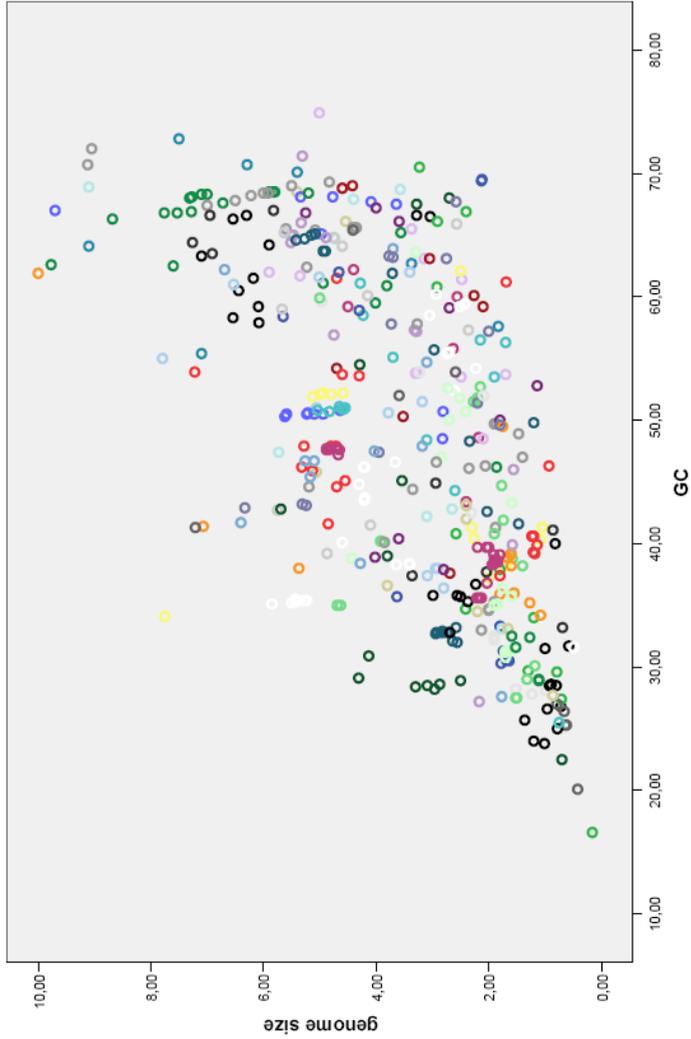
In 1995 (Fleischmann *et al.*, 1995) the first microbial genome was published. Since then, microbial genomics has shifted from the first stages of sequencing representative genomes of

representative phyletic groups or causative agents of important diseases to the massive sequencing of genomes even different strains from the same species. This transition has implied a change in the way in which genomes, even bacterial species, are considered, leading to the introduction of microbial studies into the population genomics era. Most efforts currently concentrate in two areas. One is the sequencing of a large number of genomes, from multiple strains of the same species to genomes of very distant taxonomic groups but which share common lifestyles (Figure 1). The results are allowing, for example, to quantify the roles of evolutionary processes among closely and distantly related genomes (Beiko *et al.*, 2005; Mau *et al.*, 2006; Zhaxybayeva *et al.*, 2006; Didelot *et al.*, 2007), challenging the definition of bacterial species (Doolittle and Papke, 2006; Tettelin *et al.*, 2005; Konstantinidis and Tiedje, 2005), the identification of genes with important biotechnological functions (Cowan *et al.*, 2005; Handelsman, 2004) or functions especially important for an specific environment (Venter *et al.*, 2004; Coleman *et al.*, 2006).

On the other hand, the sequencing of multiple markers of individuals from the same and different populations allows studying the evolution of microbial populations at a genome and worldwide-scale levels (Maiden, 2006; Perez-Losada *et al.*, 2006). This global survey of microbial population variation has revealed the evolutionary history, origin and transmission of common pathogens such as *Salmonella typhi* (Roumagnac *et al.*, 2006) or *Helicobacter pylori* (Linz *et al.*, 2007). Also, molecular epidemiology analyses and the study of evolutionary forces shaping microbial populations diversity are taking advantage of this multilocus

Figure 1.

The relationship between genome size and G+C content of the 471 microbial genomes sequenced up to March of 2007. Each color represents different species of the same genre following the taxonomy classification in the NCBI.



approach, allowing the comparison of clinical isolates with their environmental counterparts or reconsidering taxonomic ranks incorrectly assigned (Godoy *et al.*, 2003; Coscolla and Gonzalez-Candelas, 2007; McCombie *et al.*, 2006; Wirth *et al.*, 2006). As a result, a new consideration of the incidence of recombination and its role in shaping bacterial population variation is emerging for most bacterial taxa which contrasts with the traditional view of clonal complexes proposed in early years of molecular analyses of genetic variation in bacterial populations (Smith *et al.*, 1993; Feil *et al.*, 2001; Hanage *et al.*, 2006b).

But currently microbial genomes are considered and analyzed not only on the basis of the genes and associated functions but also considering the microbial diversity that surrounds them, in an approximation usually known as metagenomics. Metagenomic studies try to identify not only the bacterial diversity in a concrete niche, but also the global gene and genomic composition of these bacterial communities and their metabolic capabilities. Two main factors are relevant for these goals. On the one hand, laboratory techniques have been developed that allow the fast isolation and sequencing of uncultured, environmental samples and, on the other hand, as most of this metagenomic data represents unknown taxa it is necessary a sufficient taxon sampling in complete microbial genomes in order to assign sequences to their most closely sequenced taxa.

In the last few years, several studies have analyzed the gene and species composition of very different environments (Tringe *et al.*, 2005). Some have revealed a low bacterial diversity

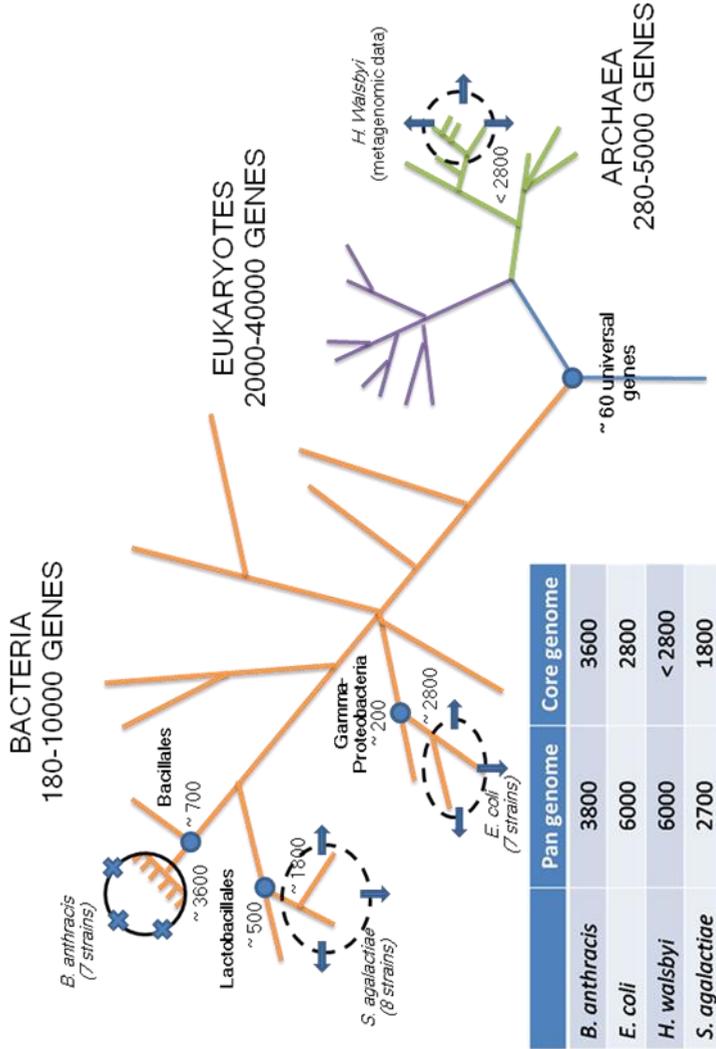
with communities dominated by only two or a few bacterial taxa (Tyson *et al.*, 2004), whereas other studies have identified niches where bacterial diversity is higher, with complex ecological networks among the species present therein (Venter *et al.*, 2004). The number of new genes, those with no known function or not identified in a any previously sequenced genome, is extraordinarily high even in those communities composed by species with close relatives in the databases. All these studies have highlighted the vast amount of unknown microbial diversity and the limitations that this ignorance is imposing on our current views of bacterial evolution (Doolittle and Baptiste, 2007).

For example, Tyson *et al.* (2004) analyzed the bacterial community structure of samples recovered from an acid mine drainage (AMD) microbial biofilm. Only a few lineages were shown to dominate the community, mostly *Leptospirillum* species. The low-complexity of the environment seems to explain the presence of this dominant lineage and other low-frequent Bacteria and Archaea (*Ferroplasma* type II genome was sequenced in this study). In such an environment, recombination seems to be the main evolutionary force that drives the speciation of the dominant *Leptospirillum* genome. Additionally, a few more cases of ancient horizontal gene transfer could be inferred, including cases of inter-domain exchange. At the other end, the analysis of samples from the Sargasso Sea revealed a rich ecological structure composed by a range of from 1800 to 48000 genomic species (Venter *et al.*, 2004). Although almost all bacterial phyla were represented, most of the species proved to be related with already sequenced Proteobacteria genomes. Other remarkable findings

from this study were the presence of spatial patchiness in the frequency and distribution of these species and the identification of around 1.2 millions of new genes.

The two examples outlined above highlight also the need of new conceptual frameworks for the evolutionary genomic analysis of bacteria. The term **pan-genome** has been proposed to describe all the genes contributed by the genomes from a species regardless of their presence in all of them (Tettelin *et al.*, 2005) see Figure 2 for some examples. Species like *Streptococcus agalactiae* have an open pan-genome and, as a result, it is expected that even the sequencing of one hundred more strains will add new genes to it (Tettelin *et al.*, 2005). The analysis of closely related *Shigella/Escherichia* species shows similar results (Chen *et al.*, 2006). Other species, like *Bacillus*, seem to have a closed pan-genome; almost all genes that are part of their genomes are known and new sequencing efforts will probably add not much more (Tettelin *et al.*, 2005). Analogously, a pan-genome for metagenomic studies can be outlined, the so-called **microbiome**. Focusing on the two metagenome examples given above, an environment such as the one dominated by *Leptospirillum* species is expected to have a closed microbiome while an environment such as that of the Sargasso Sea would have an open, difficult to assess microbiome.

Figure 2. Predicted pan-genomes and core-genomes of some taxonomic groups with a relevant number of strains already sequenced. Solid circles on nodes are the estimated size of the core genome of the corresponding monophyletic group. At the strain level solid line in *Bacillus anthracis* represents a closed pan-genome whereas a dashed line in the other species represent an open pan-genome as reflected in the table. Adapted from Abby and Daubin, 2007.



Complementary to the concept of pan-genome is that of **core genome**. The core is the set of genes shared by all the members of a taxonomic group. Usually, the core genome reduces with the addition of newly sequenced genomes especially for those cases of open pan-genomes, a fact already observed when trying to determine core gene sets at higher taxonomic ranks (Charlebois and Doolittle, 2004). Figure 2 presents a summary of some of the pan-genome and core genomes estimates derived from diverse sequenced species strains and taxonomic ranks.

1.2 The nature of microbial evolution and the concept of bacterial species

The Linnaean paradigm has been used to classify organisms based on their similarities since the XVIIIth century (Linnaeus, 1758). Based on shared features of organisms, it generates a hierarchical classification scheme that allows the formation of groups within groups in order to classify living organisms from species to domains. The theory of evolution by natural selection provided a good support to the Linnaean paradigm explaining the similarities of organisms as the result of shared adaptation due to common ancestry (Darwin, 1859). As a result, living organisms could be related to each other by means of a strictly bifurcating tree that shows the relationships between groups. The exchange of genetic information between these groups is not possible. Although the approach has been useful for most living organisms it is not clear whether bacteria, in the light of complete genome sequences, could fit this model of groups within groups. Underlying the Linnaean paradigm is the concept of species, the smallest group of organisms which are more related

between them than with other groups/species. A species definition, and therefore the taxonomic scheme derived from it, is useful only if it has a biological meaning (Hey, 2006). Although there are more than 20 proposed species concepts for living organisms, none of them is fully useful to the definition of bacterial species, because bacterial genomes and the forces acting on them follow very different routes than those acting on eukaryotes and sexually reproducing species.

In order to understand these forces, which are the clue to understand the nature of bacterial species, it is necessary to study them both at a population genetics level and at higher, interspecies taxonomic ranks. The population biology of bacteria has been studied from quite ago mainly because some of them are agents of important human and animal diseases. The first molecular analyses based on multilocus enzyme electrophoresis (MLEE) of bacterial populations revealed them as asexual clusters of individuals where only a few of all possible multilocus genotypes were recovered (see (Selander *et al.*, 1986) for a revision). Further studies in *Escherichia coli* and other bacteria corroborated this strong linkage disequilibrium in MLEE studies, thus indicating the secondary role of recombination in bacterial populations (Levin, 1981). From these results, the “periodic selection hypothesis” imposed as the paradigm (Levin, 1981) for the evolution and composition of bacterial populations. Under this hypothesis, the periodic selection of advantageous mutations present in the populations results in a bottleneck effect, reducing genetic variation in the population to a group of nearly identical individuals. Recombination is not as frequent as necessary to

break these lineages which were revealed as essentially clonal. The sampling of only a few of all possible lineages, even from populations separated by large geographic distances, was explained by the loss of alternative lineages due to limited geographic distribution, environmental changes or competitive exclusion, with some lineages being the most successful in the occupation of the corresponding niche.

The clonal model based on periodic selection was proposed for *Escherichia coli* populations; however, it was quickly adopted as the rule for all bacterial species. This model was not questioned until the early 90's, when analyses at the nucleotide level (Smith *et al.*, 1991) and the study of the evolution of antibiotic resistance related genes (Coffey *et al.*, 1991) showed that advantageous alleles could also spread by recombination. Smith *et al.* (1993) proposed a model in which recombination has an important role in bacterial evolution, thus offering alternative explanations to the MLEE observations of few, clonal lineages in bacterial species. This proposal, among other possibilities, explains clonal structures of some bacteria as transient states due to the "epidemic" structure of the population (Smith *et al.*, 1993), as exemplified in meningococcal populations (Holmes *et al.*, 1999) and proposed for other bacteria. However, the clonal model is not completely invalid since, in the same work, Smith *et al.* (1993) described a range of situations for different bacterial species, from those with a highly clonal structure at all levels (ex. the *Bacillus cereus* group) to those mainly shaped by recombination (ex. *Neisseria*) (see Figure 3).

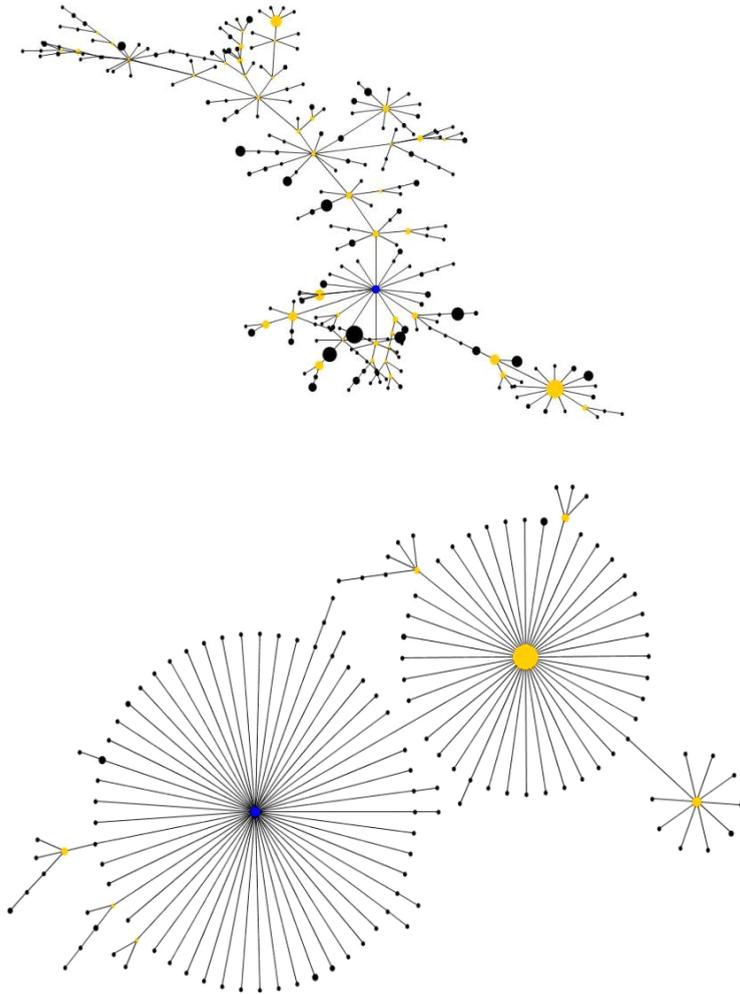


Figure 3. Different population structure of microbial pathogens as inferred by eBURST (Feil et al., 2004). The first network on the top corresponds to a *Burkholderia pseudomallei* survey. The second network refers to a clonal population of *Staphylococcus aureus*. The BURST algorithm calculates the distance between isolates on the basis of shared MLST alleles. Usually, in clonal populations like the one on the right, there is a highly frequent allelic profile and several alternatives which differ only in one or two of the loci. They are known as clonal complexes. When the diversity among isolates is very high, then there is not such a main profile.

With the advent of fast and cheap methods of sequencing, nucleotide-based approaches to microbial population genetics have blossomed. Multilocus sequence typing (MLST) was proposed in 1998 as a portable and universal method for characterizing bacterial isolates by unambiguous allelic profiles obtained from sequencing internal fragments of a variable number (usually 5-10) of genes (Maiden *et al.*, 1998). Although initially the selected loci were housekeeping genes, alternative markers have been developed to screen more variable regions and to extend these analyses to non-housekeeping, usually virulence related, genes (Coscolla and Gonzalez-Candelas, 2007; Castillo and Greenberg, 2007). Multilocus sequence typing has shed new light to the study of population structure of pathogens, revealing that recombination is more widespread than previously described with MLEE data for bacterial pathogens (Feil *et al.*, 2001; Feil *et al.*, 2000; Feil *et al.*, 1999; Scally *et al.*, 2005; Nubel *et al.*, 2006; Allen *et al.*, 2007; Perez-Losada *et al.*, 2006). Even an increasing number of intracellular bacteria, including maternally transmitted endosymbionts, has been reported as recombinogenic (Gomes *et al.*, 2007; Liu *et al.*, 2006; Baldo *et al.*, 2006).

In which way do these new ideas on the role of recombination affect our view of bacterial species? Recombination in bacterial populations is known to decrease with divergence. Distance-scaled recombination is a by-product of the mechanisms involved in the process. Bacterial mismatch correction systems are able to distinguish foreign DNA on the basis of the divergence with the recipient genome (Cohan, 2002; Majewski and Cohan, 1998). Therefore, this mechanism

represents a barrier at the population level to homologous recombination and therefore delineates a first boundary to the exchange among lineages. The clonality or sexuality of bacterial populations depends on the mutation/recombination rates ratio (Hanage *et al.*, 2006a). In clonal populations the emergence of distinct lineages and their subsequent differentiation is easier than under strong recombination, which acts as a cohesive force for the population. Once these lineages have started to diverge, distance-scaled recombination assures the cohesion of the nascent lineage and enhances divergence from other lineages, thus empowering the conditions for speciation. However, simulations taking into account these factors have revealed that recombination is useful to maintain the nascent lineages but it is not enough, under realistic parameter values, to originate them under a neutral drift model in sympatric populations (Fraser *et al.*, 2005; Hanage *et al.*, 2006a; Fraser *et al.*, 2007).

If recombination maintains the cohesion of lineages but it is not their cause, how can they arise? Different proposals have been formulated, although some of them lack experimental confirmation. The ecotype species model proposed by Cohan (2001) is based on the existence of strains sharing the ecological niche of mutants that out-compete all other strains and become fix in the population. Intra-lineage genetic variation is purged by periodic selection while inter-lineage genetic variation is maintained. Ecotypes are irreversibly separated from other ecotypes, therefore matching the basic principles of Mayr's biological species concept (Mayr, 1942). However, this definition works well only for cases in which each ecotype corresponds to

only one sequence cluster in MLST analyses, but there are several factors that could disrupt this one-to-one relation, including frequent recombination among ecotypes (Gevers *et al.*, 2005).

Horizontal gene transfer has also been proposed as a catalyst of bacterial speciation. Lawrence (2002) proposed a model in which members of a bacterial population gain genes by horizontal gene transfer. If these newly acquired genes have an ecological adaptive value for the strain, recombination around them will not be favoured and neutral variation will accumulate in regions of horizontally acquired genes. This accumulation results in a faster differentiation of these regions of the genome than others, therefore resulting in “fuzzy” species boundaries. This model is not incompatible with that proposed by Cohan, since regions of the genome will be subject to periodic selection, as they do not recombine, while other parts will remain freely recombinant (Cohan, 2004). In opposition, Doolittle *et al.* (Doolittle and Papke, 2006; Nesbo *et al.*, 2006) proposed that the acquisition of genes from unrelated sources enhances recombination between the donor and recipient genomes in this region, thus reinforcing the notion that the same organism could be a member of two distinct species. In any case, both proposals highlight the need for incorporating horizontal gene transfer as an important factor in bacterial speciation models, both because it allows the invasion of new niches (Ochman *et al.*, 2000) and because it promotes the genetic divergence among members of the same lineage (Doolittle and Papke, 2006; Lawrence, 2002).

The question about the existence of bacterial species as real entities remains open (Gevers *et al.*, 2005). It seems that only

completely clonal populations could fit a standard definition of biological species. In highly recombinogenic bacteria the distinction between clonal groups is much more difficult, with fuzzy boundaries which prevent the delimitation of natural taxonomic units (Hanage *et al.*, 2005). Analogously, horizontal gene transfer speeds up the speciation process but not necessarily in the whole genome, thus promoting also the creation of mosaic genomes with genes coming from multiple sources. This is currently observed in the study of pan-genomes and bacterial genome sizes (Tettelin *et al.*, 2005). Genome divergence and genome content are clearly uncoupled (see (Thompson *et al.*, 2005) for an example), therefore questioning the utility of Cohan's model (Cohan, 2001) based only on the former measure.

The nature of bacterial species does not affect only the realm of taxonomy, but also phylogenetic analyses. If most genes of a bacterial species come from illegitimate sources then the traditional bifurcating tree is clearly insufficient to grasp all the evolutionary history, since it will only reflect a part, and not necessarily the most important one, of the evolutionary processes that have shaped bacterial genomes (Doolittle, 1999b). Consequently, we have centred this study in two of the main, jointly with gene duplication, evolutionary forces that change the gene complement of a genome, gene gain through horizontal gene transfer and gene loss through reductive evolution, and their impact on bacterial phylogenetics.

1.3 Gene gain: plurality of phylogenetic signals in microbial genomes

The genomes of bacteria incorporate in their sequences different evolutionary signals as the result of the different evolutionary processes that act upon them. As a consequence, the phylogenetic information encoded in these genomes can be divided into three main categories: vertical signals, non-vertical signals and phylogenetic noise (Figure 4). The reconstruction of bacterial evolution and the appraisal of the different forces that have shaped their genomes depend on disentangling these signals.

The vertical signal is associated to the transmission of genetic information from ancestors to descendants. From a genomic perspective, this signal resides in the set of true orthologs shared by microbial genomes. However, establishing these sets of true orthologs is very difficult. In fact, the application of different strategies usually yields different sets of orthologous genes.

The non-vertical signal arises as the result of evolutionary processes that do not involve the immediate ancestors as donors of genetic material. The two most common processes at a genome level originating this signal are duplications and horizontal gene transfers. Paralogs are those genes resulting from a process of duplication. After their origin, paralogs may have different fates, from neo- or sub-functionalization to extinction through gene disintegration (Lynch and Conery, 2000). Xenologs are horizontally transmitted genes from a non-relative of the recipient genome obtain by non-homologous recombination (Gogarten and Olendzenski, 1999; Koonin, 2005).

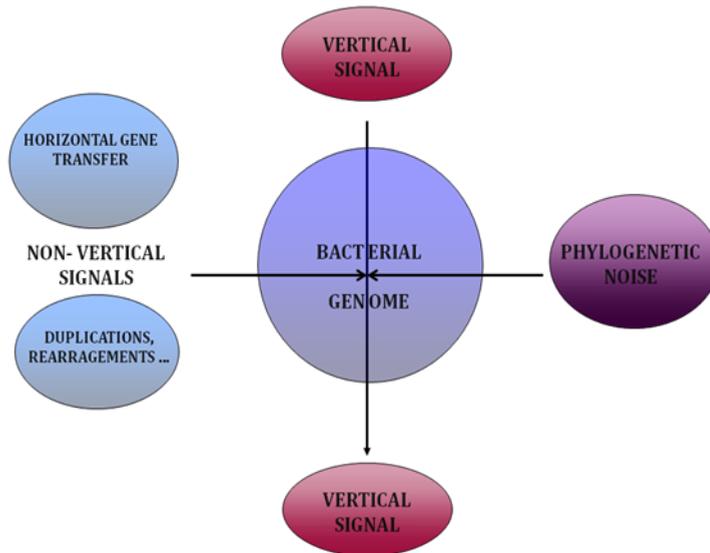


Figure 4. *Evolutionary forces acting on bacterial genomes and phylogenetic signals derived from them.*

The existence of horizontal gene transfer among microorganisms is known from quite ago (Avery *et al.*, 1944) and is currently recognized as one of the main processes influencing the evolution of bacteria (Lawrence, 2002; Gogarten and Townsend, 2005). The term synologs denotes the presence of more than one homolog within a genome regardless of the origin of the duplicate copies (paralogy or xenology) (Lerat *et al.*, 2005). Exhaustive analyses of the origin of the genes from 13 Gamma-Proteobacteria have shown that horizontal gene transfer is in fact the main factor introducing paralogy in microbial genomes (Lerat *et al.*, 2005).

In principle, it could be expected that most genes in bacteria would belong to the vertical category (Kunin and

Ouzounis, 2003; Beiko *et al.*, 2005) since genomes are vertically inherited every generation. However, the most important innovations in bacterial gene content, and consequently in possibilities of adaptation to new environments, seem to be from horizontal transfer events (Ochman *et al.*, 2000) and, to a lesser degree, of duplications (Gevers *et al.*, 2004). The exact fraction of genes belonging to each category is variable among different groups, even species, and difficult to assess. In fact, there is disagreement about the extent to which non-vertical processes, mainly horizontal gene transfer, influence the inference of genome phylogenies and the existence of a species tree for bacteria. If the rate of lateral gene transfer is high, then a phylogenetic reconstruction that relies on ancestor-descendant relationships will not be able to reflect the evolution of bacterial genomes that might be better described by means of networks (Doolittle, 1999a). However, if this rate is low enough then we will be able to represent bacterial evolution as a tree and not as a network (Kurland *et al.*, 2003). In their extreme versions, these two positions deny the importance of the vertical or the non-vertical signals, respectively.

A third factor influences the detection of these signals. Phylogenetic noise can be very difficult to distinguish from horizontal gene transfer. This noise is the result of processes non-related to the way of transmission and could be the cause of an important fraction of phylogenetic incongruence observed in microbial phylogenetics (Kurland *et al.*, 2003). Different factors introduce noise in phylogenetic analyses (Sanderson and Shaffer, 2002). It is known that different genes have different evolutionary

rates. Rapidly evolving genes on a lineage could make phylogenetic inference methods to err, overestimating the real distances to this lineage. Heterogeneity in nucleotide composition could also result in the wrong phylogenetic placement of the biased taxon. Therefore, before disentangling the horizontal/vertical signal it is important to evaluate the noise-to-signal ratio. Chapters 3 and 5 will deal with these difficulties. In chapter 3 the evolutionary features of endosymbiont genomes introduce noise in the phylogenetic analysis in the form of potential convergences. In chapter 5 we approach the distinction between noise and signal with a group of Gamma-Proteobacteria, the Xanthomonadales, which reveal their high past promiscuity for accepting external genetic material.

1.4 Gene loss: genome reduction in microbial evolution: Gamma-Proteobacteria endosymbionts of insects as a model

If gene gains through horizontal gene transfer are a major factor in bacterial evolution that have allowed bacteria to colonize multiple niches (Ochman *et al.*, 2000), gene losses have allowed them to refine their adaptation to some niches. This is usually the case when bacteria establish a symbiotic relationship with a host, either as commensalists (fitness advantage for one of the members but no reduction for the other), mutualists (all the members take advantage of the relationship) or parasites (one of the members increase its fitness decreasing its partner fitness) (Moran and Wernegreen, 2000). Possibly the best studied case of symbiotic relationship is that of mitochondria and chloroplasts as endosymbiotic genomes that lost most of their functional

capacities to the point of transforming into organelles within eukaryotic cells.

Among extant bacteria the best studied cases of endosymbiosis are those of the Gamma-Proteobacteria endosymbionts of insects (Wernegreen, 2002). All of them are characterized by the presence of an array of features which are very different to those of other bacteria. This set of features is known as the ‘resident syndrome’ and it is the hallmark of the transition from a pathogenic/free-living lifestyle to an endosymbiotic one (Andersson and Kurland, 1998). Residence inside a cell, usually in bacteriocytes, provides bacteria with a very stable environment. This means that most of the genes related to a non-endosymbiotic lifestyle or redundant with the functions provided by the host are superfluous; therefore they are suitable material for elimination, by becoming pseudogenes in the first stages of the relationship. Furthermore, this process is also characterized by a significant reduction or absence of horizontal gene transfer and recombination which results in no restoration of the DNA lost (Gil *et al.*, 2004a). However, the genome is still dynamic as these first stages of reductive evolution are correlated with a high number of rearrangements and the presence of repetitive DNA and insertion sequences (Wernegreen, 2005).

From a population genetics point of view, endosymbionts are maternally inherited, which imposes a bottleneck during their transmission. A reduced effective population size has been also proposed for endosymbiont populations which, coupled with the bottlenecks during transmission and the absence of recombination mechanisms, results in a process known as Muller’s ratchet

(Moran, 1996). The ratchet is the accumulation of slightly deleterious mutations that can be purged neither by recombination nor by purifying selection. Alternative proposals have questioned the importance of the ratchet and the reduction of the effective population size (Itoh *et al.*, 2002).

Although stasis has been proposed for the last stages of endosymbiotic relationships (Tamas *et al.*, 2002), in fact a different kind of genome dynamics is observed (Wernegreen, 2005). Pseudogenes have been completely lost in these stages and a perfectly conserved synteny among genomes of the same strains is observed. However, the reductive process continues as revealed by the genome sequence of *Buchnera aphidicola Cinara cedri* (Perez-Brocal *et al.*, 2006). It is also unclear which would be the evolutionary fate of these genomes, although replacement by a secondary endosymbiont has been proposed for the *Buchnera aphidicola Cinara cedri* (Perez-Brocal *et al.*, 2006) and transformation into a residing organelle for the most reduced bacterial genome known, *Carsonella ruddii* endosymbiont of psyllids (Nakabachi *et al.*, 2006).

It is paradoxical that the host, which initially provided a rich and stable environment for the endosymbiont ancestor, finally results to be the cause of the degeneration and reduction process. This process translates in the 'resident genome' syndrome mentioned above (Andersson and Kurland, 1998), which is characterized by high mutation rates due to the loss of DNA repair genes, the accumulation of deleterious mutations due to genetic drift, A+T bias in their nucleotide composition and the loss of codon usage optimization. But many questions remain

open. For instance, it is not clear which are the tempo and mode of the gene loss process. Initially, large deletions comprising several loci were proposed for the first stages of the genome reduction (Mira *et al.*, 2001). This proposal seems to be corroborated by the analysis of experimental populations evolving under severe bottlenecks, which demonstrated the occurrence of this kind of deletions although only a tiny fraction of them was viable (Nilsson *et al.*, 2005). Other authors have proposed that deletions are about the same size during the process, although large deletions are obviously more frequent at the initial stages. Analyses of the dynamics of deletions in intracellular genomes both in species in their first stages (*Mycobacterium tuberculosis*) and species with a long endosymbiotic relationship (*Buchnera*, *Blochmannia*) seem to corroborate this prediction (Gomez-Valero *et al.*, 2004a; Gomez-Valero *et al.*, 2007; Silva *et al.*, 2001).

1.5 Phylogenomics and microbial evolution

Our capacity to reveal and weight the importance of processes like horizontal versus vertical gene transfer, or the balance of evolutionary forces like mutation, selection and genetic drift depends on our ability to accurately reconstruct the evolutionary relationships of the taxa studied. This has been the aim of phylogenetics since the first reported phylogenetic trees (Zuckerandl and Pauling, 1965). Now that hundreds of microbial genomes have been sequenced, the question about which is the best way to derive these relationships taking advantage of genome sequences still revolves around the same two concepts: data sets and methods.

The selection of putative orthologs for phylogenomic analyses is a crucial step in the process (Snel *et al.*, 2005). Unfortunately, there is no agreement about the best way to derive those orthologs and very different approaches have been adopted in published studies (Dutilh *et al.*, 2007). The objective of the study is important in order to decide which is the best way to derive a set of putative orthologs. When very stringent criteria are used, only those homologs with very significant signal will remain in the data set, and this will not be useful to study the presence of xenology or paralogy. However, it will be useful if we want to derive the vertical signal of these genomes. For bacteria the choice is difficult because the number of xenologs could be very high and therefore filtering it as noise would eventually eliminate important biological and phylogenetic information from these genomes (Koonin, 2005). More relaxed criteria would allow for the presence of putative orthologs and putative xenologs but with the risk of retrieving misidentified sequences, therefore introducing noise in the analyses.

It is also in this context where a precision about nomenclature must be done. Throughout this thesis we will refer to the set of putative orthologs to those genes which will be used in the final analyses. By definition, a pair of orthologs is that set of two or more genes related with one another by vertical descent from a common ancestor, therefore as the result of a speciation event. However, to evaluate whether two genes are orthologs we need a reference tree of the ancestor-descendants relationships to compare with. Usually this reference tree is obtained from the set of orthologs by phylogenomic methods, so ultimately we enter in

a circular argument in while the assessment of orthology depends on the phylogenetic relationships of the genomes which, in turn, are obtained from the assesement of orthology. In consequence, we have abandoned the term orthologs in favor of the term **putative orthologs**: we presume that these groups of genes are *a priori* orthologs, only the *a posteriori* phylogenomic analyses will reveal them as true orthologs or as xenologs/paralogs.

Once a suitable set of genes has been selected, there is an array of available analytical methodologies based on genome sequences. Most properties that can be measured in a genome are reasonable good phylogenetic markers. This is expected since all the genomes share common ancestors with other genomes. As a consequence many different characters have been proposed to reconstruct phylogenetic relationships (Dutilh *et al.*, 2007; Delsuc *et al.*, 2005; Snel *et al.*, 2005). However, as in the case of the data sets, most of the times the use of different methods reveals different evolutionary properties. Some of them seem to mask all non-vertical signals (Gatesy and Baker, 2005), others allow exploring the possible presence of horizontal gene transfers (Gophna *et al.*, 2005), others have revealed systematic incongruence (Jeffroy *et al.*, 2006) and others reveal common themes in the evolution of genomes that are distantly related but which share some common features due to their niche, lifestyle or evolutionary pressures (Hughes *et al.*, 2005).

It is worth mentioning four methods proposed as good approaches for inferring phylogenomic relationships. The gene content of bacterial genomes has been shaped both by horizontal gene transfer, vertical transmission and paralogy. Usually, pairwise

distances of shared gene content are used in these analyses (Snel *et al.*, 1999). However, there are two main problems associated with this approach. Although it is expected that most of the shared genes are orthologs, the presence of horizontal gene transfer cannot be discarded mainly for the cases involving highly mosaic genomes. Despite this, the effect of horizontal gene transfer could be obscured by the expectedly high fraction of vertical events. This fraction is expected to be important because the number of genes analyzed in gene content trees is reduced with respect to the total number of genes available as the criterion of shared orthologs has to be met. On the other hand, using common genes reduces the data set to only a few hundred when large evolutionary distances are considered or very small genomes incorporated like in the case of endosymbionts. Although there have been attempts to correct this effect, known as the 'big genome attraction artifact' (Lake and Rivera, 2004), there is no reported case of gene content phylogeny able to correctly place several small genomes (Delsuc *et al.*, 2005) although some progress to avoid these problems has been made (Lake and Rivera, 2004; Huson and Steel, 2004).

Gene order trees are based on the conservation of order among the genomes analyzed. Inversions, duplications and translocations are the main phenomena that can alter gene order in a genome (Belda *et al.*, 2005). The advantage of this kind of events is that homoplasy is reduced because a change of state could result in numerous possible alternative states. One general observation on these measures is that usually gene order distances are larger than nucleotide distances and that there is variation in

the rate of change among taxa. This kind of data has been used successfully for phylogenetic reconstruction in eukaryotes and prokaryotes, and it is also interesting that these measures give hints about the evolution of genome architecture (Moret and Warnow, 2005). However the approach still does not use an explicit model of evolution incorporating all the above mentioned events and there is no way to correct for or take into account homoplasies.

Gene content and gene order phylogenies are products of the genome sequencing era, since both use characters beyond sequence data. Chapters 3 and 4 will treat in more depth the other two main phylogenomic approaches: supertrees and supermatrices. They are the heirs of the phylogenetic analyses as both are based on the evolutionary analysis of nucleotide or amino acid sequences. The two approaches are opposite representatives of the traditional distinction between the total evidence approach and the data partition approach (Kluge, 1989). Supermatrix approaches take advantage of the shared gene content of the genomes analyzed to construct a “supergene” through concatenation of individual alignments. This supermatrix is considered a total evidence approach as it analyzes simultaneously all the available characters. Although usually the putative orthologs used are those present in all the genomes, it is currently being studied which would be the impact of introducing non-universal genes and therefore increasing the amount of missing data (Philippe *et al.*, 2004). Therefore the limitation of shared gene content, which could reduce the characters analyzed drastically under certain circumstances, is being corrected. Supertrees are a

kind of data partition analysis based on the reconstruction of single gene trees and their reconciliation in the most plausible, compatible topology (Bininda-Emonds, 2004b). The advantage in this case resides in that there is no need for a highly overlapping data set. However, the accuracy of the methods depends on the quality of the gene trees which again could be affected by many processes (Dutilh *et al.*, 2007).

1.6 A glimpse on the genomes analyzed in this thesis

The different objectives outline above requires different experimental set ups. For example, endosymbiont genomes have not undergone horizontal gene transfer since their free-living ancestor with some exceptions. Therefore, they are not useful to detect instances of such events. In consequence different data sets were selected to achieve the different objectives of studying the performance of phylogenomic approaches when applied to endosymbionts data sets, analyzing the patterns of horizontal gene transfers and studying the forces behind endosymbiont sequences evolution.

Although phylogenomics approaches has been developed in numerous ways and applied to numerous data sets, the interest of our research group centered in the phylogenetic placement of Gamma-Proteobacteria endosymbiont sequences. As the most recent endosymbiont sequenced in our group as of year 2004 was *Blochmannia floridanus*, endosymbiont of carpenter ant, we selected it as our base genome to use for a further investigation of the relationships of these endosymbiont with three *Buchmaera aphidicola* genome sequences and the *Wigglesworthia brevipalpis* genome,

endosymbionts of aphids and tse-tse flies respectively. So this work was aimed to analyze the different phylogenomic tools and their performance in resolving an interesting evolutionary problem, the monophyly of the Gamma-Proteobacteria insect endosymbionts. Because of the resident syndrome explained in above sections, these sequences have characteristics high rates of evolution and low G+C content so were especially difficult to deal with them from a phylogenetic point of view.

Although previous studies were carried out with endosymbiont this work shows some distinctive features. 1) It carries out an analysis of the phylome of *Blochmannia floridanus* rather than analyzing only the common gene set. 2) It applies some corrections to the confounding effects of the extremely low G+C content of endosymbiont genomes. 3) Evaluates the performance of different phylogenomic approaches and the influence of factors like the number of genes analyzed or the influence of the functional role of these genes. 4) Uses two outgroups besides Gamma-Proteobacteria genomes, one sequence from Alpha-Proteobacteria and three from Beta-Proteobacteria.

The point around outgroup is not trivial. Its use revealed to us systematic incongruence around the phylogenetic position of the Xanthomonadales clade. These are plant pathogenic bacteria that use to be classified as Gamma-Proteobacteria and placed at the base of the group. Our previous phylogenomic study and a bibliographic revision revealed that Xanthomonadales use to be placed also as Beta-Proteobacteria or at the base of Alpha-Proteobacteria. Therefore this group was a suitable subject for the analysis of horizontal gene transfers as possible cause of

incongruence. It is important to note that for this study the number of genomes representing Alpha-, Beta- or Gamma-Proteobacteria must to be balanced, we selected representative genomes from these groups most of them also plant pathogens. It is worth mentioning that endosymbiont genomes were not used in this study because their little utility and their possible confounding effects on the interpretation of incongruence.

New sequenced endosymbiont genomes appeared in 2006. *Buchnera aphidicola* *Cianara cedri*, the smallest *Buchnera* genome, with only 0.42 Mb. and 20.1 % G+C content, and *Carsonella ruddii*, endosymbiont of psyllids, which is the smallest known bacterial genome with 0.16 Mb. and 16.6 % G+C content. These new genomes allowed us to study the role of positive, negative and neutral selection as well G+C content and rates of substitutions in genomes at different stages of symbiosis. We specially focused in the two new sequenced genomes and the action of these forces in the last stages of endosymbiosis. It is worth noting the increase in the number of available endosymbiont genomes with respect the first point mentioned above, one new *Blochmannia* specie and *Baumannia cicadellincola*, endosymbiont of sharpshooters, which is suspected in the first stages of endosymbiosis. This allowed us to expand the phylogenetic analysis carried out in the first point including new approaches to deal with G+C content of endosymbiont genomes and its influence on phylogenetic reconstructions.

2. OBJECTIVES

The present thesis deals with different microbial genome evolution aspects. Despite some of the objectives are pointed to answer concrete questions it is important to note that the methodologies used and their utility are also relevant to analyze large data sets, of gene or trees, like those derived from genome sequencing projects.

The general objectives of this thesis will center in three important evolutionary aspects in bacteria: 1) phylogenomics analyses, 2) horizontal gene transfer among bacterial genomes and 3) the balance of evolutionary forces acting on endosymbiont genomes. We have approached the accomplishment of these general objectives through the use of different genome data sets which allow us to formulate concrete questions about their evolution. These specific objectives are:

- 1) Which phylogenomic tools are most appropriate to analyze different aspects of microbial genome evolution and how they work when applied to endosymbiont genomes data sets?
- 2) Which are the evolutionary relationships among sequenced Gamma-Proteobacteria endosymbiont genomes?
- 3) Which was, and still is, the evolutionary impact of past and present horizontal gene transfer in the Gamma-Proteobacteria plant pathogens Xanthomonadales?
- 4) Which is the balance of evolutionary forces acting on endosymbiont genomes particularly in those that are

in their last stages of genome reduction like *Carsonella ruddii* and *Buchnera aphidicola* *Cinara cedri*?

**3. FROM PHYLOGENETICS TO
PHYLOGENOMICS: current tools for the
evolutionary analysis of Proteobacteria genomes**

3.1 INTRODUCTION

The evolution of bacteria is strongly influenced by the fluid nature of their genomes. Processes such as duplication, gene gain and loss limit the inference of their evolutionary relationships. However, in the last decade the genomic revolution has represented not only a change in scale for the analysis of sequences but also for phylogenetic inference. The phylogenomic approach, initially proposed as the application of phylogenetic analysis to help annotating complete genome sequences (Eisen, 1998), has transformed into the (possibly) most appropriate way to derive the phylogenetic history of organisms through genome-scale phylogenetic inference (O'Brien and Stanyon, 1999; Sicheritz-Potén and Andersson, 2001; Eisen and Fraser, 2003). This leap from phylogenetics to phylogenomics has allowed avoiding some of the inherent problems in the inference of species trees from single gene phylogenies. However, it has also brought new issues in the reconstruction of species trees and even questioned the existence of such a single tree for all the genes in the genome of a bacterial species (Doolittle, 1999b; Baptiste *et al.*, 2005; Susko *et al.*, 2006).

The single gene approach reached its peak with the use of ribosomal RNA sequences as phylogenetic markers for microorganisms (Woese, 1987). These genes have been, and still are, a powerful tool in bacterial phylogenetic analysis. Their properties as a good marker for phylogenetic inference, such as universal presence and evolutionary conservation, have enabled the proposal of a universal tree of life (Woese *et al.*, 1990) and the

classification and reconstruction of evolutionary relationships for the three domains in the absence of genomic data (Woese and Fox, 1977). However, even ribosomal markers have been shown to be laterally transferred in some cases (Asai *et al.*, 1999; Yap *et al.*, 1999).

The advantages of multiple gene approaches versus those based on single genes are *a priori* evident. In theory, we can evade the single gene evolutionary histories in favour of a common “true” phylogenetic signal; it is possible to avoid problems derived from insufficient sample size by addition of more sites from multiple genes or to compensate for biased base compositions. In practice, some of these problems will also affect phylogenies reconstructed from large data sets, but others will be substantially reduced and/or easily diagnosed. Several alternative methods to single gene tree phylogenies have been proposed recently (reviewed in Delsuc *et al.*, 2005). These methods are based on different kinds of sequence characters and genome structure, such as gene content (Snel *et al.*, 1999; Gu and Zhang, 2004; Huson and Steel, 2004), gene order (Wolf *et al.*, 2001; Korbelt *et al.*, 2002; Belda *et al.*, 2005), concatenated sequences (Brown *et al.*, 2001; Rokas *et al.*, 2003) or gene-tree-based techniques (Bininda-Emonds *et al.*, 2002; Bininda-Emonds, 2004b). In this chapter we focus on methods based on direct or indirect analyses of genomic sequence data.

The concatenation approach has proven useful in many cases and its use is increasingly common (Baptiste *et al.*, 2002; Rokas *et al.*, 2003). The method has been justified in cases where

single gene phylogenetic signals are insufficient (Herniou *et al.*, 2001), when there is heterogeneity in the evolutionary rates (Baptiste *et al.*, 2002; Gontcharov *et al.*, 2004) or when a high influence of non-standard evolutionary patterns in the shaping of gene trees, such as horizontal gene transfer, is to be expected (Brochier *et al.*, 2002). This method increases phylogenetic signal by joining the sequences from multiple genes thus creating a supermatrix of characters, and generally recovers accepted (and presumably correct) phylogenies with highly supported nodes thus evading many of the above pitfalls (for instance, insufficient amounts of informative sites or particular gene histories) as far as the correct model of evolution is used, although it may not overcome problems associated to systematic biases in the data (Phillips *et al.*, 2004).

Contrary to the concatenation approach, consensus and supertree approaches have an indirect association with the genome sequence. The consensus method is based on the integration of multiple source trees into a single topology. When the initial data set does not include the same taxa in all the gene trees, then a supertree is constructed combining the overlapping topologies (Bininda-Emonds, 2004b). All these methods have proven useful for phylogenetic inference (Daubin *et al.*, 2002; Rokas *et al.*, 2003), but they have different possible associated errors. For instance, the strength of the supermatrix approach decreases when the number of shared orthologous genes is low or when many of the genes concatenated are influenced by horizontal gene transfer events or hidden paralogies (Daubin *et al.*, 2002). On the other hand, a supertree approach has also its own

sources of problems. The use of wrong source trees, the indirect relationship with molecular data, taxon heterogeneity among the trees, or lack of a universal methodology for assessing the reliability of the nodes are among the most common problems encountered by consensus and supertree approaches (Gatesy *et al.*, 2002; Creevey, 2004).

On the other hand, not only the phylogenomic methodology is important but also the data set to which it is applied is of relevance. The nature of the genes that compose the data set to be analyzed can have a direct incidence on the phylogeny recovered and on the phylogenetic signals contained therein (Gophna *et al.*, 2005). From any genome, which is composed by a mixture of signals, different subsets can be derived. The term ‘minimal genome’ has been used to describe the set of genes that are supposed to be essential for a self-sustainable cell live (Mushegian and Koonin, 1996). There is no single, unique minimal genome and several proposals have been put forward (Mushegian and Koonin, 1996; Glass *et al.*, 2006). However, a review of different approaches has proposed a synthesis of 208 genes as the minimal genome needed for cellular life (Gil *et al.*, 2004b). It is expected that these genes, most of them characterized by their essentiality and their central role in the metabolic network, encode a good, vertical signal in agreement with the complexity hypothesis (Jain *et al.*, 1999; Jordan *et al.*, 2002).

Nevertheless, essentiality is not the only factor that could influence the presence of vertical signal in a set of genes. It is also important that these genes are shared by all the taxa

analyzed due to restrictions in the applicability of some phylogenomic methods (Lerat *et al.*, 2003). Consequently, a core of genes suitable for the phylogenomic analysis can be defined by the universality of their presence in all the genomes considered. The universality of this core is, in consequence, another factor to consider in the analysis of the evolutionary vertical signal of bacterial genomes.

Our work is based on the determination of the phylome (Sicheritz-Potén and Andersson, 2001) – the set of phylogenetic trees for each protein coding gene in the genome - of *Blochmannia floridanus* (Gil *et al.*, 2003) the primary endosymbiont of carpenter ants. We have used it as a starting point to explore the phylogenetic landscape of Gamma-Proteobacteria. In this work we have used an array of phylogenetic and phylogenomic techniques to infer the phylogenetic relationships among 21 Proteobacteria. In consequence, we have analyzed several phylogenetic hypotheses for these species through the examination of the phylogenies derived from the set of protein coding genes in *Blochmannia* and their comparison with topologies obtained from the 16S rDNA and different phylogenomic methods. Our aim in this chapter is mainly focused in analyzing the advantages and pitfalls associated to each class of approximation. Additionally we have centered on how to identify and extract the vertical signal from a real data set of bacterial genomes in the presence of incongruence by analyzing different subsets of the *Blochmannia* genome each one characterized by their own evolutionary and phylogenomic properties.

Table 1. Complete genome sequences of *Proteobacteria* used in this study.

Species (Strain)	Accession No.	Division	Order
<i>Rickettsia prowazekii</i>	NC_000963	ALPHA	Rickettsiales
<i>Neisseria meningitidis</i> MC58	NC_003112	BETA	Neisseriales
<i>Neisseria meningitidis</i> Z2491	NC_003116	BETA	Neisseriales
<i>Ralstonia solanacearum</i>	NC_003295	BETA	Bulkholderiales
<i>Blochmannia floridanus</i>	NC_005061	GAMMA	Enterobacteriales
<i>Buchnera aphidicola</i> BPI	NC_004545	GAMMA	Enterobacteriales
<i>Buchnera aphidicola</i> SGI	NC_004061	GAMMA	Enterobacteriales
<i>Buchnera</i> sp. APS	NC_002528	GAMMA	Enterobacteriales
<i>Escherichia coli</i> K12	NC_000913	GAMMA	Enterobacteriales
<i>Escherichia coli</i> O157:H7 EDL933	NC_002655	GAMMA	Enterobacteriales
<i>Haemophilus influenzae</i> Rd	NC_000907	GAMMA	Pasteurellales
<i>Pasteurella multocida</i>	NC_002663	GAMMA	Pasteurellales
<i>Pseudomonas aeruginosa</i>	NC_002516	GAMMA	Pseudomonadales
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar Typhi	NC_003198	GAMMA	Enterobacteriales
<i>Salmonella typhimurium</i> LT2	NC_003197	GAMMA	Enterobacteriales
<i>Vibrio cholerae</i>	NC_002505	GAMMA	Vibrionales
<i>Wigglesworthia glosinidia</i> <i>brevipalpis</i>	NC_004344	GAMMA	Enterobacteriales
<i>Xanthomonas axonopodis</i> pv. <i>citri</i> str. 306	NC_003919	GAMMA	Xanthomonadales
<i>Xanthomonas campestris</i> pv. <i>campestris</i> str. ATCC 33913	NC_003902	GAMMA	Xanthomonadales
<i>Xylella fastidiosa</i>	NC_002488	GAMMA	Xanthomonadales
<i>Yersinia pestis</i> KIM	NC_004088	GAMMA	Enterobacteriales

In this context, by incongruence we mean the presence of non-vertical signals or phylogenetic noise in the set of genes to be used in phylogenetic/phylogenomic analysis although how to address the source(s) of such incongruence is out of the scope of this chapter, and will be analyzed in detail in chapter 5. However, we study the effect of the presence of incongruence on the performance of the supermatrix and supertree methodologies mentioned above and address several points about the phylogenetic signal contained in the different functional categories and the role of essentiality and universality in the correct inference of vertical evolution. The impact of these phylogenomic analyses on endosymbiont genome studies will be treated in the chapter 4, where we center in the discussion on the phylogenetic relationships of Gamma-Proteobacteria endosymbionts, extending this work to the phylogenetic analysis of the two most reduced genomes up to now, *Carsonella ruddi* and *Buchnera aphidicola* *Cinara cedri*.

3.2 MATERIAL AND METHODS

3.2.1 Genomes and homologous genes selection

In this study we have used 21 complete genome sequences of Proteobacteria (Table 1) retrieved from GenBank (Benson et al., 2004). We included three Beta-Proteobacteria, *Neisseria meningitidis* MC58, *Neisseria meningitidis* Z2491 and *Ralstonia solanacearum*; one Alpha-Proteobacteria, *Rickettsia prowazekii*, and 17 Gamma-Proteobacteria, including five insect endosymbionts: the three *Buchnera* species sequenced so far, *Wigglesworthia glossinidia brevipalpis* and *Blochmannia floridanus*, which is our initial genome. A

complete and general scheme of the analyses carried out in this study and the relationships among them is presented in Figure 5.

The procedure to obtain the putative orthologs for each gene in the *Blochmannia floridanus* genome started from an initial reference tree. Since it is not possible to be certain about the truly orthologous nature of a gene until the phylogenetic analysis is completed, we considered the homologous genes found in the procedure described below as putative orthologs, which may eventually become true orthologs, paralogs, or xenologs, depending on their inferred evolutionary history. The reference tree was obtained as described in Gil *et al.* (2003). Briefly, we obtained the orthologs for 60 informational genes present in all the genomes considered. Protein sequences were aligned with CLUSTALW 1.8 (Thompson *et al.*, 1994) and processed with GBLOCKS (Castresana, 2000) with default parameters to eliminate areas of uncertain homology or low phylogenetic content before concatenation. The resulting concatenate was used to obtain a maximum likelihood tree using the quartet method implemented in TREEPUZZLE 5.1 (Schmidt *et al.*, 2002) with the following options: JTT model of substitution (Jones *et al.*, 1992), proportion of invariant sites (I) estimated from the data, eight discrete categories to approximate a gamma distribution accounting for evolutionary rate heterogeneity across sites (G), empirical amino acids frequencies (F) and 4000 puzzling steps.

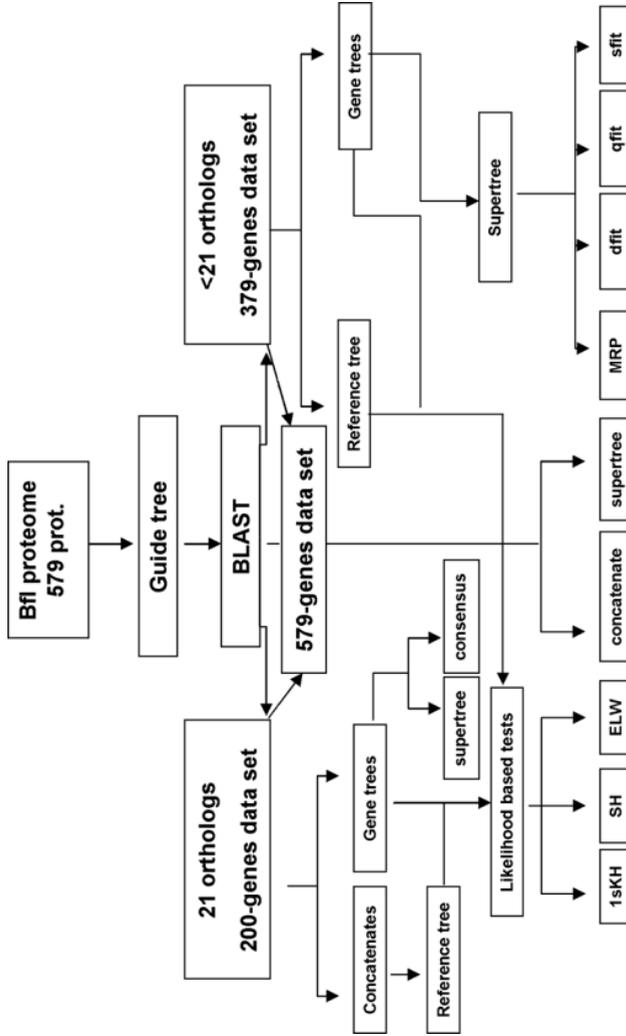


Figure 5. General diagram of the different methodologies used for assessing the phylogeny of *Blochmannia floridanus* as well as the phylogenomic analyses and testing. 1sKH: one-sided Kishino-Hasegawa test; SH: Shimodaira-Hasegawa test; ELW: Expected Likelihood Weight test; MRP: Matrix Representation using Parsimony; dfft: most similar supertree; qfit: maximum quartet fit; sfit: maximum splits fit.

We also obtained phylogenetic trees by Bayesian inference using MrBayes 3.0 (Ronquist and Huelsenbeck, 2003) (JTT + I + F + G and 100000 generations). This tree is an expanded version of the tree reported in Gil *et al.* (2003) with additional sequences from the non-Gamma-Proteobacteria genomes. With this reference tree, we assigned each genome to one of nine different groups (see Figure 6) in order to reduce the BLAST database and to speed up and refine the searches (see below).

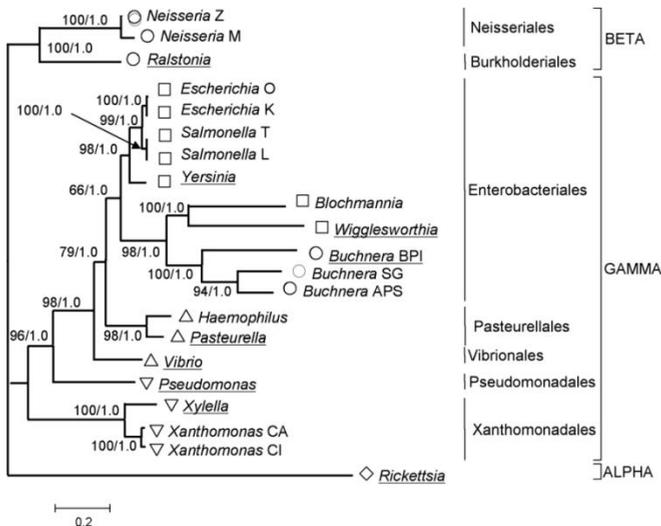


Figure 6. Reference tree (RT) obtained with a trimmed alignment of 60 concatenated proteins. For clarity, only genus names are represented except for cases with more than one species or strain (see Table 1). Numbers in nodes indicate support values in the form of proportion of quartets and Bayesian a posteriori probabilities for the corresponding inner branch. The symbol next to each species indicates the group of adscription for BLAST searches. Species with underlined names were used as initial target species for BLAST searches of the corresponding groups. The taxonomic classification of each species is also indicated according to the NCBI taxonomy database.

The *B. floridanus* genome contains 579 annotated protein coding genes (Gil *et al.*, 2003) and each of these was used as query for a BLASTP (Altschul *et al.*, 1997) search against the remaining genomes. To retrieve the homologous genes from the other genomes, we first performed a BLAST search using the representative genome from each of the nine previously described groups as target (Fig. 6). Next, we used the best hit from each representative species of the nine groups in the first BLAST search as query for a second BLAST against the remaining genomes in the corresponding group. This strategy was slightly modified for those *B. floridanus* genes for which no reliable homolog (see below a more detailed description of the criteria used) was found in the representative species of some group. In this case, a new BLAST search was performed against all the genomes in the group and the best hit was used as query for a second BLAST to retrieve the homologs from the remaining species in the group as described. In all cases, we used an E-value < 1E-3 as threshold for considering a matching sequence as a putative ortholog in the BLAST searches.

Once homologs for each of the 579 genes in *B. floridanus* were retrieved we performed a new filtering to ensure that these genes corresponded to putative orthologs. We first proceeded to obtain a multiple alignment and the corresponding gene tree as described below. With this information we considered five different criteria for each case: annotation (coincidence in functional annotation), multiple alignment (similarity extended over the complete sequence and not a short fragment), gene tree (strong conflict with the reference tree), BLAST significance, and

information from the Microbial Genome Database for Comparative Analysis (Uchiyama, 2003). Homologous genes were accepted as putative orthologs after simultaneous consideration of all these five criteria. Since these are independent, they allowed us to analyze orthology in the absence of a good annotation or when the position in the gene tree was unexpected. Should we had based strictly and only on annotation or position in the gene tree then we would had biased the data set towards genes with the most congruent phylogenies or best annotations. In most cases the identification of putative orthologs was easy. In the most difficult cases, which corresponded to non-annotated proteins or to very divergent gene trees from the usually accepted species phylogenies, we considered as putative orthologs those genes which were supported by at least three of the above mentioned criteria. This resulted in a set of 200 genes in *B. floridanus* with reliable orthologs in all of the other 20 genomes considered. We will refer to this subset as the '200-genes' data set. Whenever possible, and in order to test the effect of missing data, we performed the analyses described below with the remaining genes with reliable orthologs missing from at least one genome ('379-genes' data set) and also with the complete set of coding genes in the *B. floridanus* genome ('579-genes' data set).

3.2.2 Obtaining the phylome of *Blochmannia floridanus*

After obtaining putative orthologs for each protein of the *B. floridanus* genome we proceeded to obtain its phylome, the set of phylogenetic trees for each protein coding gene. The alignment

for each set of homologous proteins was obtained using CLUSTALW 1.8 (Thompson *et al.*, 1994) and the result was trimmed with GBLOCKS (Castresana, 2000) using default parameters. The maximum likelihood tree for each multiple alignment was inferred with the program PHYML 2.1b (Guindon and Gascuel, 2003) as it implements one of the fastest and most accurate algorithms for heuristic searches. For all the alignments we used JTT + I + G + F as model and parameters of evolution. Some of the initial alignments and trees were modified based on the *a posteriori* selection of homologs explained above. We decided not to obtain support values by bootstrap resampling to save computation time and not to bias the analysis towards the best supported topologies.

We used four phylogenetic approaches to obtain a 16S rDNA topology for the 21 species: Maximum parsimony (MP), maximum likelihood (ML), and two different distances with the neighbour joining (NJ) algorithm (Saitou and Nei, 1987). MP analysis was performed using PAUP*4b10 (Swofford, 1998) with heuristic search based in tree-bisection-reconnection (TBR) as branch-swapping algorithm and 1000 replicates to assess bootstrap support values. The best model for ML inference was selected using the Modeltest program (Posada and Crandall, 1998) following the Akaike information criterion (AIC) and implemented in PHYML 2.1b with 500 bootstrap replicates. For NJ we used two models that could handle heterogeneous base composition among lineages. Firstly we used MEGA 3 (Kumar *et al.*, 2004) to implement the LogDet distance modified by Tamura and Kumar (Tamura and Kumar, 2002) to account for

substitution pattern heterogeneity. Secondly we used PHYLO_WIN (Galtier *et al.*, 1996) implementing the Galtier and Gouy distance (Galtier and Gouy, 1995) for reducing nucleotide bias. A 1000 replicates bootstrap was carried out for both NJ analyses.

3.2.3 Phylogenomic analyses

We used three different approaches for phylogenomic analysis. Concatenated sequences, consensus trees and supertrees were used to explore the phylogenetic landscape of Gamma-Proteobacteria.

3.2.3.1 Supermatrix

The analysis of concatenates was based on the subset of putative orthologous genes shared by the 21 genomes, the ‘200-genes’ data set. We concatenated the GBLOCKS-trimmed alignments of these proteins and analyzed the resulting alignment using several methods. PAUP*4b10 (Swofford, 2002) was used for parsimony analysis with stepwise addition and tree bisection-reconnection for heuristic search and 500 bootstrap replicates. PHYML and TREEPUZZLE were used for maximum likelihood inference. Due to computational limitations, we had to reduce the number of parameters in the evolutionary models. In the analysis with TREEPUZZLE we used a JTT model with four gamma categories and 4000 puzzling steps. The JTT model of substitution with estimated proportion of invariant sites and 100 bootstrap replicates was used for inference with PHYML.

We also used this alignment of the ‘200-genes’ data set to explore the possible influence of heterogeneous base composition in the retrieval of the different identified clades, especially those involving endosymbiotic bacteria. We removed from the alignment those positions that included the amino acids more strongly affected by nucleotide bias (FYMINK) (Singer and Hickey, 2000). With this data set we carried out a Bayesian analysis using MrBayes 3.0 (four chains, 1000000 generations, 10% of burn-in) and a maximum-likelihood analysis with PHYML (500 bootstrap pseudo-replicates) with the same evolutionary model described above (JTT + I).

Finally, we obtained a concatenate tree using all available sequences (‘579-genes’ data set) by setting as “missing character” those amino acids that corresponded to a protein not present in any of the 21 genomes. It was analyzed with TREEPUZZLE using the JTT model of substitution with invariants estimated from the data set. Alternative, more complete methods were not feasible due to computational limitations.

3.2.3.2 Consensus trees

A majority rule consensus tree was obtained with the program CONSENSE from the PHYLIP package (Felsenstein, 2002) with the gene trees from the ‘200-genes’ data set.

3.2.3.3 Supertrees

For the ‘200-genes’ data set and for the set of gene trees with an incomplete number of genomes represented we used a

supertree approach (Sanderson et al., 1998). A supertree integrates all the topological information present in the source gene trees. In this case, we applied supertree reconstruction to three different data sets: (i) genes present in all the genomes ('200-genes' data set, the same set for which we obtained a consensus tree), (ii) genes missing in at least one genome ('379-genes' data set), and (iii) to the complete phylome ('579-genes' data set). We applied the different optimization algorithms implemented in program CLANN (Creevey, 2004; Creevey and McInerney, 2005) to our data set. The most commonly applied method in supertree reconstruction is Matrix Representation using Parsimony - MRP (Loomis and Smith, 1990; Baum, 1992; Ragan, 1992). This method considers each node of the source trees as a binary character, assigning a "1" to the taxa contained in the clade defined by the internal node, a "0" for the taxa not present in this clade and a "?" for the taxa not present in the tree. The resulting matrix is analyzed by parsimony. The other methods implemented and used in our work were Most Similar Supertree Method (dfit), Maximum Quartet Fit (qfit) and Maximum Split Fit (sfit) (Creevey, 2004). They are based in the optimization of distances between nodes, shared quartets or shared partitions between the supertrees proposed by the heuristic search and the source trees, respectively.

Non-parametric bootstrap by sampling with replacement over gene trees was used to obtain an indicative measure of the support of the derived clades in each supertree from the corresponding data. This approach seems to be the best suitable to evaluate the robustness of the supertree nodes to the gene tree

choice (Burleigh *et al.*, 2006). In our case it was evaluated from 100 pseudoreplicate samples for each data set used to derive the supertrees.

3.2.3.4 Analysis of incongruence

Likelihood-based tests of competing phylogenetic hypotheses were carried out once we obtained the complete phylogeny of *B. floridanus*. For each alignment we compared the topologies of the presumed species tree and the corresponding gene tree by means of three different tests implemented in TREEPUZZLE (Schmidt *et al.*, 2002): the one-side Kishino-Hasegawa test (1sKH) (Kishino and Hasegawa, 1989; Goldman *et al.*, 2000), Shimodaira-Hasegawa test (SH) (Shimodaira and Hasegawa, 1999) and the Expected Likelihoods Weights test (ELW) (Strimmer and Rambaut, 2002). For the two former tests, the null hypothesis assumes that the difference between the likelihoods associated to each topology is not significantly different from zero. The first test to be described was the KH test, but its validity when one of the competing hypotheses was derived from the alignment has been questioned (Goldman *et al.*, 2000). In this study we used the one-sided version of the test because the gene tree is the maximum likelihood tree and therefore we expected its likelihood to be always higher or equal than the likelihood associated to the species tree. In any case we have used it for comparative purposes and not as the reference test because of its problems with multiple comparisons corrections. The SH test is a multiple hypotheses contrast. Even though it is not free from errors (Goldman *et al.*, 2000) it overcomes some of the

problems associated with the KH test. We used it as our reference test for examining the rejection of the presumed species tree by each gene alignment. For both tests the resampling-estimated log likelihood (RELL) bootstrapping method with 1000 replicates was used. The ELW test is an alternative way of topology testing. It creates a confidence set of phylogenies that could include (acceptance) or not (rejection) our presumed species tree topology. The usual $\alpha=0.05$ level was used to delimit the acceptance/rejection region.

3.2.4 Phylogenomic cores analyses

3.2.4.1 Phylogenomic cores definition

We defined different phylogenomic cores from the 579 alignments with different genomic, evolutionary and phylogenetic meaning:

- '*Blochmannia*' core: composed by the 579 annotated protein coding genes of *Blochmannia floridanus* and their corresponding homologs in the other 20 genomes. In this set we deal with from genes present in the 21 genomes to genes present in only four.
- 'Universal' core: the 200 genes of *Blochmannia floridanus* that are also present in the remaining 20 genomes. This set represents those ubiquitous genes in this particular set of genomes but it does not mean that they are essential for bacterial cell life. In this set a fraction of true orthologs and xenologs/paralogs coexist.
- 'Essential' core: from the 200 genes of the 'universal' core we obtained those genes coincident with the proposal for the minimal genome by Gil et al. (2004b). This paper describes the 206 genes

needed by a cell for a self-sustainable life. From them, 133 genes were present in our 'universal' core and were selected for the 'essential' core and considered as a subset of genes with higher fraction of true orthologs and with essentiality as their common property.

Genes from each data set were assigned to different functional categories following their annotation in the *Blochmannia floridanus* genome. We used 18 specific functional categories and 4 general ones as defined in the COG database (Tatusov *et al.*, 2003).

3.2.4.2 Supermatrix analysis of phylogenomic cores

We first analyzed the performance of the concatenate analysis without taking into account the functional assignment of the genes. We carried out two different analyses, one for the 'essential' core and the other for the 'universal' core. One hundred concatenates of 10, 20, 30, 40, 50 and 60 genes were generated randomly from the pool of genes belonging to both core sets resulting in 600 concatenates for each data set. Each one of the 1200 concatenates was analyzed by maximum likelihood using PHYML under the JTT model of evolution and four gamma categories. The computational load prevented us from using more parameters in the evolutionary model. We compared the phylogeny derived from each concatenate with the reference tree (RT) shown in Figure 6 by using the Robinson-Foulds distance (Robinson and Foulds, 1981) and derived as explained in the above sections. This metric measure the number of partitions not shared between two phylogenies and is implemented in the program TREEDIST of the PHYLIP package (Felsenstein, 2002).

3.2.4.3 Supermatrix and supertree analyses of functional categories

We divided the genes in each core into 18 specific functional and four general categories. For the phylome set of genes, we screened the phylogenetic signal contained in each functional category by obtaining the supertrees derived from the gene trees of each alignment. Differences in the number of species represented in each gene alignment prevented us from performing a concatenate analysis of the whole phylome. However, for the ‘universal’ and ‘essential’ cores we were able to obtain the supertree and the concatenate alignments for each functional category.

All the supertrees were obtained with the CLANN software (Creevey and McInerney, 2005). We employed the commonly used Matrix Representation using Parsimony – MRP (Baum, 1992; Ragan, 1992) method. In this method each node of the source trees is coded as a character and a binomial code is assigned to the presence (1) or absence (2) of each taxon in the clade defined by the node. The resulting matrix is analyzed by parsimony. In some cases, the analyses resulted in more than one possible supertree in which case we took into account whether the RT topology was among the most parsimonious topologies found. With the concatenate alignments we obtained the maximum likelihood topology through PHYML (Guindon and Gascuel, 2003). For all the alignments, we used the JTT model of evolution, frequencies estimated from the data set, an estimated proportion of invariant sites and eight gamma rate categories.

Once a supertree and a concatenate phylogeny were obtained for each functional category and core set, we analyzed their phylogenetic signal through their comparison with the RT. The Robinson-Foulds distance as implemented in the program TREEDIST of the PHYLIP package was used to measure the similarity between the obtained topologies and the RT topology. The Shimodaira-Hasegawa test information obtained as explained above was also used taking into account the functional assignment of the genes.

3.3 RESULTS

3.3.1 Reference tree and search for putative orthologs

The construction of a phylogenetic tree for a set of species from their genome sequences needs, as starting point, a phylogeny on which to accept or reject which genes should be used for this kind of analysis on the basis of their true orthology. This is a circular question that we have approached by obtaining an initial or reference tree, testing its reliability, and then using it as benchmark for further decisions. For this, we started by obtaining the phylogenetic tree from a subset of conserved protein coding genes usually accepted as providing the most robust, reliable phylogenies when concatenated (Jain *et al.*, 1999).

The concatenation of 60 informational genes after their independent alignment and trimming of residues of uncertain homology resulted in an alignment with 8067 amino acid positions. Figure 6 shows the phylogenetic tree obtained by maximum likelihood and Bayesian inference methods. This

reference tree was identical to the one reported in Gil et al. (2003) except for the addition of 4 genomes from Proteobacteria in the Alpha- and Beta- divisions. The tree reflected monophyly not only for the three *Buchnera* species but also for a wider group including the other insect endosymbionts, *Blochmannia* and *Wigglesworthia*. This group is a sister clade to the YESS (*Yersinia-Escherichia-Shigella-Salmonella*) cluster. Due to its concordance with the accepted taxonomy, the high support values for the nodes and the fact that it recovers the monophyly of *Buchnera* species, we considered this as a good reference tree for the ensuing BLAST searches.

The similarity searches for homologs in the 21 genomes resulted in 215 protein coding genes present in all of them. After their initial multiple alignment and phylogenetic tree reconstruction, we proceeded to verify their reliability as putative orthologs according to the criteria indicated above. This resulted in the exclusion of 15 proteins with doubtful orthology for at least one Gamma-Proteobacteria species different from the insect endosymbionts, thus conforming a 200-genes data set. However, in order to analyze the complete *B. floridanus* phylome we included these 15 proteins after trimming the doubtful orthology into the data set of proteins absent from at least one taxon. Therefore this set was composed of 379 genes ('379-genes' data set) with a heterogeneous distribution of presence in the different genomes, summarized in Figure 7. Each protein was present on average in 16 of the 21 genomes. Nearly 60% of the proteins were present in at least 16 genomes and only 14% in less than 10.

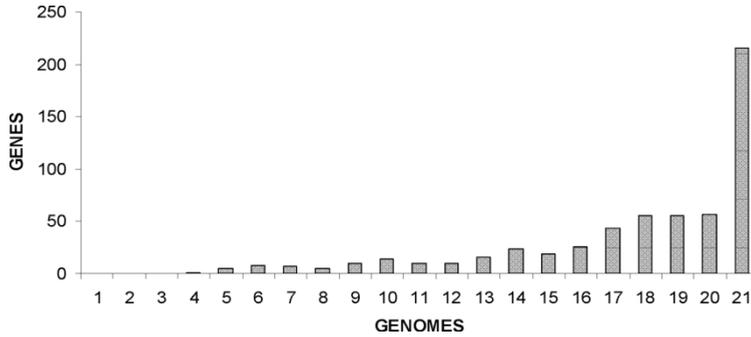


Figure 7. Histogram summarizing the distribution of putative orthologs for each protein coding gene (579 proteins of known function) in the *Blochmannia floridanus* genome among the 21 genomes considered in this study. There was 215 common genes (last column), with 15 of them considered of dubious orthology after the filtering process and removed and being analyzed in the 379-genes data set. Some genes (those present in only few genomes) were only identified in other endosymbiont genomes.

3.3.2 The phylome of *Blochmannia floridanus* and the 16S rDNA tree

We obtained all the phylogenetic trees derived from the set of protein coding genes of *B. floridanus* as explained in the Materials and Methods section. In the analysis of the ‘200-genes’ data set we identified only three genes (*rpoC*, *pheT*, and *mopA*) with the same topology (Robinson-Foulds distance = 0) than the reference tree (Table 2). In the ‘379-genes’ data set we found 40 genes whose topology was fully congruent with that of the reference tree, although this number reduced to 29 when only those present in at least 11 species were considered.

Apart from trees based on protein sequences, we obtained the tree based on the 16S rDNA. The four methods used (see Materials and Methods) resulted in placing Xanthomonadales in the Beta-Proteobacteria branch. For the endosymbiotic bacteria almost all methods retrieved a monophyletic group (see Figure 8 ML tree) although for the Galtier and Gouy distance this relationship was paraphyletic (see Figure 8 NJ tree). Other taxa like *Vibrio cholerae* or *Pseudomonas aeruginosa* also had a variable position depending on the model or method used (Figure 8 and data not shown).

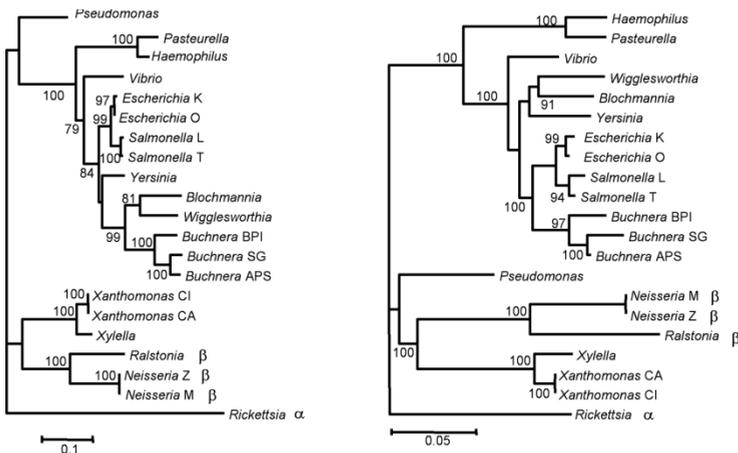


Figure 8. Phylogenies of the 16S rDNA gene inferred using maximum likelihood inference (left) and neighbour joining (right) with Galtier and Gouy distance. Only bootstrap values higher than 80% are shown. Divisions of Proteobacteria different from the Gamma division are indicated next to corresponding species.

3.3.3 Phylogenomic analyses

3.3.3.1 Supermatrix analysis

We obtained an alignment of 52029 amino acid positions, 40576 of which were variable and 32921 parsimony informative, from the concatenation of the ‘200-genes’ data set. The phylogenetic analysis resulted in a common topology for all the reconstruction methods used, with high support values for most of the nodes and methods (Figure 9). This topology agreed with the reference tree established by the preliminary analysis (see above) and the phylogenetic reconstruction in Gil *et al.* (2003) as well as the most commonly accepted phylogenetic history for these species. In consequence, we adopted this topology as the presumed species tree on which to base further comparisons. This topology included a monophyletic clade with the insect endosymbionts, which received a high support by the three different methods used (bootstrap analysis for maximum likelihood and parsimony, and quartet-puzzling maximum likelihood). This high support was common for most nodes in this tree and in fact there were only two cases in which one method provided a support lower than 70%. One corresponded to a large group of Gamma-Proteobacteria, including the Enterobacteriales and *H. influenzae* and *P. multocida*, which was recovered in only 69% by quartet puzzling analysis. The second case was the bootstrap support (59%) received by the three Gamma-Proteobacteria using parsimony analysis. In both cases, as well as for all other nodes, the bootstrap support of maximum likelihood analysis equaled 100%. It is also remarkable that the Xanthomonadales were placed at the base of Gamma-

Proteobacteria with very high support with the three methods used.

We also obtained the same topology in two other concatenate analyses. Firstly, we used the same concatenate alignment of the ‘200-genes’ data set but removing all the positions with those amino acids most influenced by a biased GC content (Singer and Hickey, 2000). After removal of positions with amino acids FYIMNK, there were 16772 positions left for analysis. We retrieved the same topology with even higher support for the nodes (data not shown). Secondly, we also carried out the analysis of the concatenated alignment of the whole data set (‘579–genes’ data set), incorporating “missing data” as necessary, and obtained again the same topology of the presumed species tree. The resulting concatenate included 137301 positions with an average of 14,59% sites encoded as ”missing” per genome (excluding *Blochmannia floridanus*).

3.3.3.2 Consensus and supertree analyses

Two kinds of methods dealing with the gene trees were used. Trees derived from the ‘200-genes’ data set were analyzed by traditional majority rule consensus and supertree approaches. Both analyses recovered the same topology (Figure 6) characterized by the deviation from the presumed species tree in one of the nodes. This divergence placed the Xanthomonadales group and *P. aeruginosa* with the Beta-Proteobacteria. These alternative groupings received relatively low bootstrap support and, in the most frequent alternative to the majority rule consensus tree, *P. aeruginosa* returned to its position in the

presumed species tree (data not shown). Despite the remaining clusters in the consensus were coincident with those in the presumed species tree, including the monophyly for insect endosymbionts, the frequency of each clade in the source trees was generally low. Even well defined groups such as free-living Enterobacteria received lower support than expected.

For the remaining gene trees ('379-genes' data set) from the *B. floridanus* genome we applied only the supertree approach due to the heterogeneity in the number of taxa represented in each tree. The four algorithms used resulted in the same topology (Figure 9). This topology was very close to the consensus and supertree topologies obtained from the '200-genes' data set (Figure 9). The Xanthomonadales group was also outside the Gamma-Proteobacteria but *P. aeruginosa* was recovered in its 'natural' group.

Finally, we also obtained a supertree from the complete phylome (579 gene trees) of *B. floridanus*. Interestingly, the topology obtained by heuristic search using the MRP algorithm agreed with the presumed species tree (Figure 6). Therefore, the complete phylome analysis recovered the position of Xanthomonadales as Gamma-Proteobacteria although with low bootstrap support. In fact, the bootstrap consensus tree (obtained after 100 pseudo-replicates of 579 gene trees analyzed by MRP) again showed Xanthomonadales grouping with Beta-Proteobacteria. This is a clear indication of the presence of conflicting phylogenetic signals in this data set.

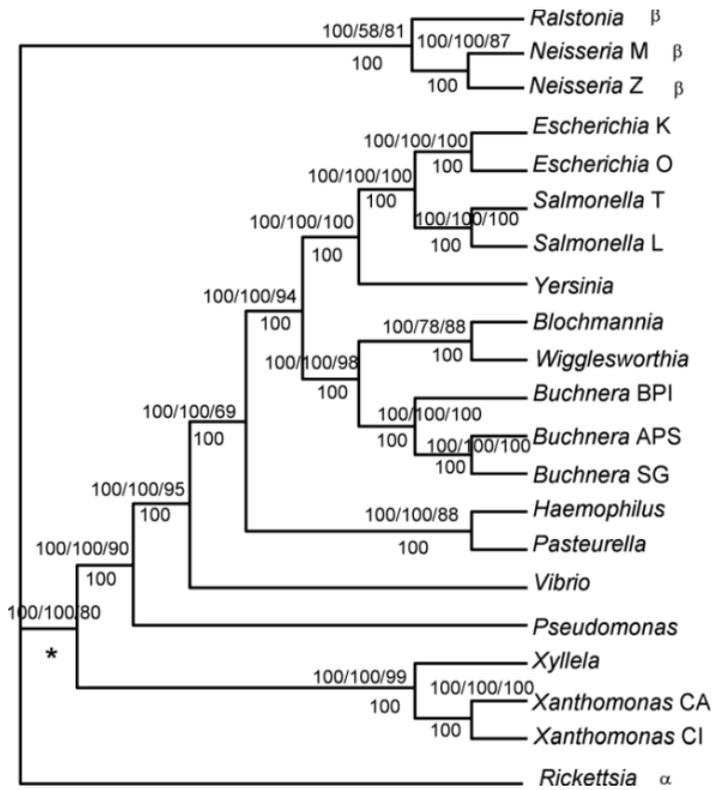


Figure 9. Common topology obtained from the analysis of the 200-genes concatenate and the 579-genes supertree. This topology was adopted as the presumed species tree. Values above the nodes indicate support values for the 200-genes concatenate obtained by maximum likelihood (bootstrapping), maximum parsimony (bootstrapping) and maximum likelihood (quartet-puzzling). Values below the nodes indicate bootstrap support for the 579-genes supertree. The asterisk indicates the node which appears in a different position in the 579-genes bootstrap supertree. Divisions of Proteobacteria different from the Gamma are indicated next to the corresponding species.

Table 2. Summary of tests for topologies. For each test (1sKH, one sided Kishino-Hasegawa; SH, Shimodaira-Hasegawa; ELW, expected-weighted likelihoods) percentage of cases in which the species tree topology was rejected is presented. The last column indicates the number of cases in which the species and gene tree topologies are identical.

GROUP	1sKH	SH	ELW	# genes
ALL (579 genes)	30.4 %	29.5 %	31.4 %	43
21 spp. (200 genes)	30.5 %	29 %	32 %	3
< 21 spp.(379 genes)	30.3 %	29.8 %	31.1 %	40

3.3.3.3 Incongruence analysis

We next tested whether the presumed species tree was significantly worse than each gene tree in the *B. floridanus* phylome. Table 2 summarizes the proportion of cases in which the former was rejected for the three likelihood based tests used. As expected, the SH test was more conservative than the other two although the results of the three tests were quite similar. Globally we observed a 30% rate of rejections with no noticeable difference between genes present in all ('200-genes' data set) or absent from some of the 21 genomes ('379-genes' data set). We also found slight differences in the rate of rejection among informational and operational genes (21.4% and 32.5% respectively).

3.3.4 Phylogenomic cores

The previously described search for putative orthologs identified 200 protein coding common genes which composed

what we called the ‘universal’ core, thus characterized by (quasi)universal genes. Of these, 133 genes were coincident with the proposal of a minimum number of genes for a self-sustainable cell by Gil et al. (2004) and composed what we called the ‘essential’ core, whose genes not only are universally distributed but also suspected to have an essential functional role. The distinction is important because minimal genome proposals take into account not only essential genes but also genes whose function could be replaced by other, alternative genes not included in the proposal. However, those genes included in ‘minimal genome’ proposals which have a universal distribution are probably essential genes.

3.3.4.1 Supermatrix analysis of phylogenomic cores

Our first approximation to the problem of analyzing the vertical signal of these genomes consisted in comparing the performance of the ‘universal’ and ‘essential’ cores in a supermatrix analysis. We generated 100 random concatenates of 10, 20, 30, 40, 50 and 60 genes for each core and analyzed their corresponding phylogenetic trees. Figure 10 summarizes the results of two metrics to evaluate the efficiency of each data set in recovering a reference tree (RT) congruent with current taxonomical classification of the species analyzed.

The ‘essential’ core performed better than the ‘universal’ core. The ‘essential’ core recovered the reference tree in all 60-genes concatenates generated, whereas the ‘universal’ core with 60 genes concatenated only yielded a null Robinson-Foulds (RF) distance to the reference tree in 41 of the 100 concatenates. In addition, the mean topological distance reflected the

differences between the two data sets. The average initial topological distances were 3.56 and 2.62 for the ‘universal’ and the ‘essential’ core concatenates, respectively. The behavior of the distance metric when the number of genes in the concatenates increased from 10 to 60 genes reflected very different dynamics for the two core sets. While the ‘essential’ core concatenates reduced the distance to the RT as more genes were added, the ‘universal’ core increased the gap as more genes were incorporated in the concatenates. The final value obtained for the 60-genes concatenates reflected this clear discrepancy: concatenates for the ‘essential’ core had RF distances of zero, since all of them recovered the reference tree, while the average distance of 60-genes concatenates from the ‘universal’ core was 5.78. The difference in the performance between these two data sets must reside, at least to a certain extent, in the 67 genes present in the ‘universal’ core and absent from the ‘essential’ core. In consequence, we included this subset of 67 genes in subsequent analyses and denoted it as ‘non-essential’ core.

When the complete sets of genes in the ‘universal’ and ‘essential’ cores were used to obtain the corresponding concatenates, the maximum likelihood trees showed identical topology than the reference tree (RF distance = 0). The same analysis with the ‘non-essential’ core resulted in a topology with RF distance = 4 to the reference tree, due to the unresolved position of Xanthomonadales at the base of the tree.

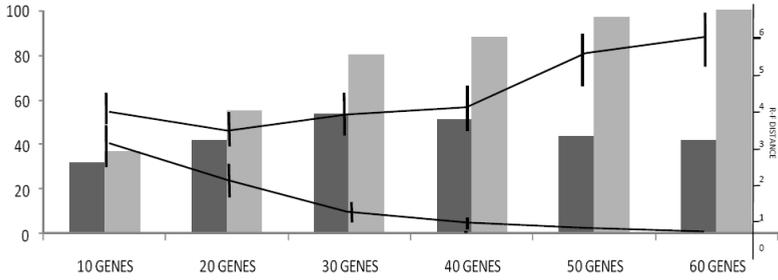


Figure 10. Number of concatenates out of 100 that recovered the RT (columns, left y-axis) for the 'essential' (dark-gray) and the 'universal' (light-gray) cores. The lines represent the average Robinson-Foulds distance (right y-axis) with standard errors from the 100 concatenates that compose each category for the 'essential' (down-line) and 'universal' (top-line) cores.

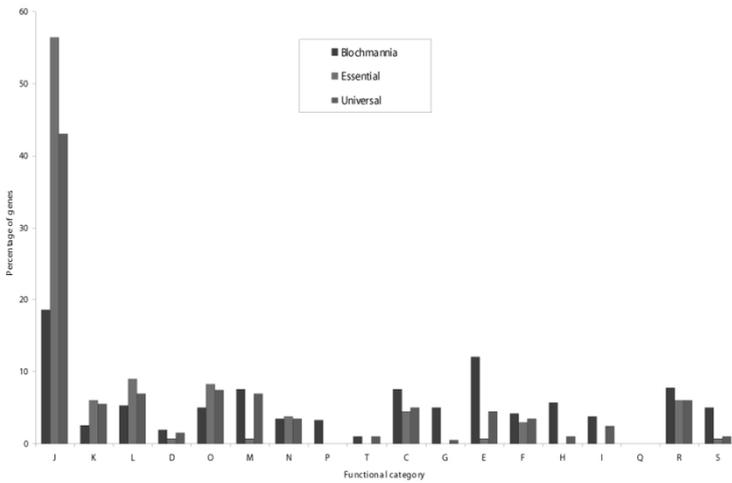


Figure 11. Percentage of genes in each functional category. Order of columns refers from left to right to the 'Blochmannia', the 'essential' and the 'universal' cores.

3.3.4.2 Functional analysis of the phylogenomic cores

Once the overall phylogenetic signal in the ‘universal’ and ‘essential’ cores had been evaluated, we proceeded to study the relationship between functional assignment of the genes and performance of the phylogenomics methods described. Table 3 shows the description of each functional category whereas Figure 11 shows the contribution in percentage of each category to each data set. As expected, both the ‘universal’ and ‘essential’ cores had an enriched fraction of the informational categories while other categories had almost disappeared. In this analysis we were interested in comparing the ‘universal’ and the ‘essential’ core and also the *Blochmannia* core, for which we had to introduce a supertree analysis, since in the latter the unequal number of sequences in the 579 multiple alignments prevented the application of a concatenate analysis. Also, due to the small number of genes present in the ‘non-essential’ core in the different functional categories considered, we did not include this subset in this analysis.

A summary of the supertree and concatenate analyses is shown in Figure 12. Overall, the K (‘Transcription’) and the J (‘Translation’) categories, both related to information processes, presented the best vertical signal. For the transcription category both supermatrix and supertree approaches recovered the RT of the ‘universal’ and ‘essential’ cores as did the supertree method when applied to the *Blochmannia* core subset. The reference tree was recovered from the subset of genes in the ‘Translation’ category only in the supermatrix analysis for the ‘universal’ and ‘essential’ cores, but neither in the supertree nor in the

Blochmannia' core. The other informational category, related to replication (L), did not recover the RT in any case. The supertree derived from all the individual trees of informational genes always recovered the RT as shown in Figure 9. In the remaining categories, the RT was obtained only in a few cases. For the general categories, only the *Blochmannia*' core subset of 'Cellular processes' recovered the RT in the supertree analysis. Among the additional specific functional categories, only genes related to posttranslational modification (category O), like chaperones, seemed to retain a good vertical signal. However, two cases grabbed our attention: on the one hand, the two concatenates derived from the 'Cell motility and secretion' (N) category recovered the RT; on the other hand, the general function (R) category also behaved well in the concatenate analysis.

For a more detailed quantitative analysis, we also analyzed the topological distance of the concatenate trees derived from each of these categories to the RT. Figure 13 shows the distances from the maximum likelihood-based phylogenies obtained with the concatenates derived from the 'universal' and 'essential' cores. The general category with the shortest distances to the RT was that of informational genes whereas the others had higher distances, above all the metabolism category. Surprisingly, the second category with shortest distance to the RT was that of 'poorly characterized' genes which comprises those of 'General function' (R) and 'Unknown function' (S).

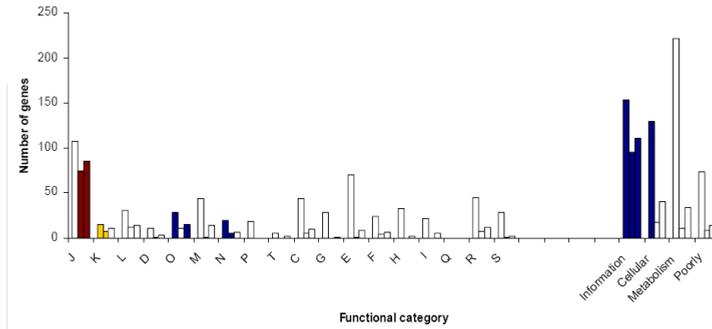


Figure 12. Supermatrix and supertree functional analyses. The categories recovering the reference tree through supermatrix (blue), supertree (red) or both methods (yellow) are shown as filled columns. For each category, the first column represents the results obtained with the 'Blochmannia' core, the second column corresponds to the 'essential' core and the third column to the 'universal' core. The height of each column represents the number of genes in each functional category for the three data sets. Note that for the four general categories only the supertree analysis was performed.

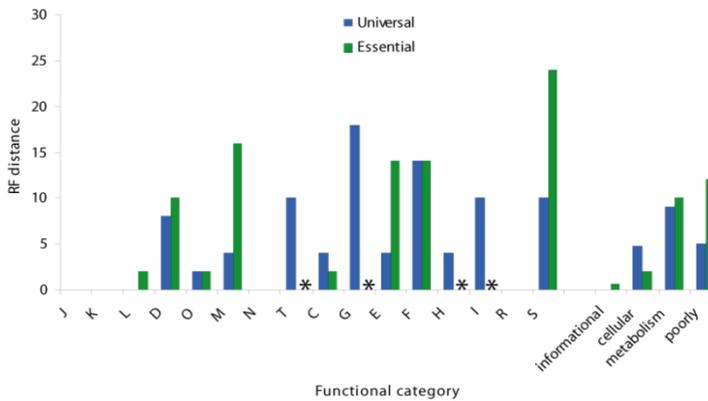


Figure 13. Comparisons (RF distances) between concatenate trees, by functional category, and the reference tree. Cases marked with an asterisk indicate that no genes were present in the corresponding category for the 'essential' core.

In fact, a detailed analysis of the more specific categories showed that the R category was the main contributor to the short distance of the general category, recovering the RT tree in both data sets. Meanwhile, categories G ('Carbohydrate transport and metabolism') and T ('Signal transduction mechanisms') presented the largest distances among specific categories. On the other side, categories O and N that were identified with good vertical signal were the two categories, apart from the informational, with shortest distances with respect to the RT.

Finally, we analyzed the performance of the individual gene trees in the different data sets for recovering the reference tree topology. The results were very similar for the 'universal', the 'essential' and the 'non-essential' cores, with average RF distance values of 12.19, 12.00 and 12.57, respectively. This statistic was not computable for the '*Blochmannia*' core as the number of sequences varies among the 579 individual gene trees considered. The results of the SH tests, at $\alpha = 0.05$, for each gene tree revealed a rejection rate of 29.5%, 29% , 27% and 34,3% for the '*Blochmannia*', the 'universal', the 'essential' and the 'non-essential' cores, respectively (Table 2). The same analyses were carried out taking into account the functional assignment of the genes. Only those genes of the K ('transcription') category present in the 'universal' and 'essential' core data sets showed a significantly lower rejection rate than the mean of their corresponding data sets. Conversely, genes from the 'non-essential' core in the E ('Amino acid transport and metabolism') and I ('Lipid metabolism')

categories had a significantly higher rejection rate of the RT using the SH test (Table 3).

Table 3. Percentage of gene trees that reject the reference tree. Each combination of core-functional category and the different cores proportions are shown. Each cell shows the percentage of cases in which the reference tree topology was rejected.

COG CATEGORY		'Blochmannia' core	'Universal' core	'Essential' core
Whole data set		29,5	29	27
INFORMATION				
Translation	J	31,5	26,7	26,7
Transcription	K	13,3	0	0
DNA replication	L	25,8	14,3	16,7
CELLULAR				
Cell division	D	18,2	33,3	NA
Posttranslational modification	O	31	33,3	36,4
Cell envelope biogenesis	M	34,1	35,7	NA
Cell motility and secretion	N	30	25	NA
Inorganic ion transport and metabolism	P	31,6	NA	NA
Signal transduction mechanisms	T	16,7	NA	NA
METABOLISM				
Energy production and conversion	C	22,7	20	16,7
Carbohydrate transport and metabolism	G	17,2	NA	NA
Amino acid transport and metabolism	E	28	77,8	0
Nucleotide transport and metabolism	F	33	28,6	50
Coenzyme metabolism	H	27,3	NA	NA
Lipid metabolism	I	36,4	NA	NA
Secondary metabolism	Q	NA	NA	NA
POORLY CHARACTERIZED				
General function prediction only	R	37,8	33,3	37,5
Function unknown	S	37,9	NA	NA

3.4 DISCUSSION

3.4.1 Comparison of methods: from phylogenetics to phylogenomics

We have used complete genome sequences of several Proteobacteria to analyze the phylogenetic relationships of a group of Bacteria characterized by their endosymbiotic relationship with different insects. Starting from the genome of one of these bacteria, *Blochmannia floridanus*, the endosymbiont of carpenter ants, we have obtained the phylogenetic trees for every protein coding gene in this species, by using single gene, standard phylogenetic methods, and we have also analyzed this data set with phylogenomic approaches.

Single gene phylogenies are representative of traditional phylogenetic analyses. They have several advantages over multi-gene approaches. Taxon sampling is widest and the acquisition of raw data in the laboratory and from public databases is easy and cheap. However, our analysis with 16S rDNA exemplifies the low robustness of this marker for phylogenetic reconstruction, since different topologies were obtained when the evolutionary model or reconstruction method were changed (see also Herbeck et al., 2004). As a case in point, only 3 of the genes present in all genomes had exactly the same topology than the presumed species tree obtained by the concatenate analysis (Figure 6). This high variability in reconstructed topologies is the result of limitations in the phylogenetic methods, sampling error due to the limited length of single gene alignments, the action of evolutionary forces and model misspecification that may result in single gene

phylogenies not coincident with the presumed species tree. Therefore, the topological heterogeneity obtained by using a single gene approach, both for the 16S rDNA and the phylome, does not justify in most cases their use in the inference of the evolutionary relationships among species, especially when whole genome data are available. However, still very often these are the only kind of data available for phylogenetic analysis. In this case, congruence with the known taxonomy and convergence with other phylogenetic markers available are reassuring but incongruence does not necessarily mean that a whole new evolutionary history has to be envisioned. Furthermore, the joint analysis of many single gene phylogenies provides a better, more accurate picture of the evolutionary history of the whole genome, in which different evolutionary trajectories may be revealed once noisy and/or unreliable signals are identified and considered appropriately.

On the other hand, phylogenomic methods, which can be divided in supermatrix, consensus and supertree approaches, have their own problems (Bininda-Emonds and Sanderson, 2001; Bininda-Emonds *et al.*, 2002; Bininda-Emonds, 2004a). While the concatenates are most useful for retrieving a main phylogenetic signal, techniques based on the analysis of underlying gene phylogenies may reveal the divergent signals encoded in the single genes that are usually ignored by concatenate analysis (see Gates and Baker, 2005).

In this study we have used the traditional consensus (with the '200-genes' data set) and supertree (with the '200-genes', '379-

genes', and '579-genes' data sets) approaches. Except for the complete phylome supertree, all the methods agreed in placing the Xanthomonadales with the Beta-Proteobacteria due to the high incidence of gene phylogenies with this placement instead of the expected Gamma placement. This case has been studied in detail chapter 5. Deviations in supertrees from the presumed species tree could arise from two methodological sources: errors due to heterogeneity in the taxa used for analysis or from differences in their most common topological position in the individual trees (Creevey, 2004). However, the former problem can be excluded in this case since we have worked with a very homogeneous data set and the conflicting result appears even in the two analyses based on the '200-genes' trees which are common to all the genomes considered. Thus, the three different supertree analyses ('200-genes', '379-genes' and '579-genes') have revealed a large amount of conflicting signals in these genes about the phylogenetic positioning of Xanthomonadales. This group is also the one responsible for most discrepancies between the consensus and supertree approaches, especially over the support values of the nodes. It has to be noted that consensus trees are summaries of the observed clades in the underlying gene trees, all obtained from the complete set of the same number of taxa, whereas support values in a supertree analysis are not a mere recount of observed clades but a measure of the support, derived from bootstrap resampling (Burleigh *et al.*, 2006), that the source gene trees provide to their most compatible topology.

Additionally, the use of supertrees is suitable for many situations where supermatrix or consensus may be limited or non

applicable. For example, in cases where the evolutionary distance among taxa is very high and the amount of shared genetic information is very low (Sanderson and Driskell, 2003) it is highly likely that sparse data matrices are obtained, leading to concatenate alignments with many missing characters and individual trees with incomplete sets of taxa. Furthermore, supertrees are useful in cases, such as this work, where more than one phylogenetic signal is present because they can reveal alternative signals appearing in single gene phylogenies. Finally, supertrees stand as the best choice method when phylogenetic data sources are heterogeneous, for instance when combining morphological and molecular data based phylogenies.

The other phylogenomic method commonly used is concatenation of shared sequences. With this approach we derived a phylogenetic tree which was identical to the reference tree and in agreement with the most commonly accepted phylogeny for the taxa considered. This lead us to adopt it as the ‘presumed species tree’, a hypothesis which received further support when alternative reconstruction methods were applied. But the concatenation approach cannot be considered free from errors and/or biases, the most important one being the assumption that a single, main phylogenetic signal exists and that it will be revealed by the simultaneous consideration of as many data as possible, regardless of their (evolutionary) congruence in terms of history or mode of evolution (Daubin *et al.*, 2002).

Further problems arise in the concatenation approach because the number of genes shared in all the genomes decreases

when the evolutionary distance between taxa increases (Charlebois and Doolittle, 2004). In this case we have used only a subset of 200 genes from the 579 present in the *B. floridanus* phylome. Although we have analyzed the concatenate of the complete genome, this approach also presents some problems, mainly the large number of missing characters in the supermatrix and the heavy computational load it imposes, with serious limitations on the methods and programs that can be used for its analysis. At least for the taxa we have studied, the subset of 200 common genes seems to be enough for recovering the main phylogenetic signal. In fact, even smaller subsets are still able to recover it, but in these cases there is more uncertainty in the inferred trees since these depends on the specific genes, and their evolutionary histories, chosen for analysis.

Contrary to the approaches based on gene trees, the concatenation methodology seems to favour mainly one signal, the most common or less incongruent with all the gene partitions in the alignment. The presence in the concatenates of genes that, when analyzed separately, reject the reference tree, implies that incongruent or conflicting phylogenetic signals are hidden at least in alignments corresponding to a large fraction of the genomes (Gatesy and Baker, 2005). Despite these potential problems, particularly important in bacterial phylogenetics, concatenation is the most popular phylogenomic method. Tools like the Microbial Genome Database for Comparative Analysis (Uchiyama, 2003) allow recovering the shared gene content from the growing number of completely sequenced bacterial genomes thus enabling an easier and faster analysis of concatenates. Similarly, recent

works propose algorithms for obtaining maximum concatenated sequences from databases (Driskell *et al.*, 2004) and the analysis of the required number of genes (Rokas *et al.*, 2003) or the impact of missing data to obtain a good phylogeny (Philippe *et al.*, 2004). These works are allowing the formalization of a concatenation methodology.

3.4.2 Different phylogenomic data sets harbour different phylogenetic signals

One of the main questions in phylogenomic analyses based on sequence information is the composition of the data set used. We have generated three different data sets derived from the genes present in the endosymbiont *Blochmannia floridanus* and other 20 genomes. These data sets, denoted 'Blochmannia' core, 'universal' core and 'essential' core, have allowed us to study the influence of different, presumably important factors on bacterial phylogenomics.

The main question we wanted to address was whether essentiality and universality were important factors influencing the efficiency of the commonly used concatenate methodology. Genes common to the 21 genomes, therefore expected to be quasi-universal at least at the Proteobacteria taxonomic level, were included in the 200-gene data set thus conforming the 'universal' core. On the other hand, the 133-genes common to the 21 genomes and simultaneously proposed to be minimal for a self-sustainable life conformed the 'essential' core, whose most relevant feature is essentiality. Their performance in the concatenate analyses was completely different: the 'essential' core

recovered the RT with fewer genes and with higher frequency than the 'universal' core. Clearly, essentiality seems to be an important factor. In fact, while the addition of genes had little effect over the 'universal' core, in the 'essential' core the mean distances to the RT reduced continuously until becoming null when 60 genes were concatenated. These results indicate that although the vertical signal is strong in the 'universal' core it still includes incongruent genes and therefore universality does not necessarily mean absence of factors like phylogenetic noise or lateral gene transfer (Susko *et al.*, 2006). Meanwhile, 'essential' genes seem to have an even stronger vertical signal, a result expected because of the increased proportion of informational genes in the 'essential' core data set (Nakamura *et al.*, 2004; Jordan *et al.*, 2002). The difficulties in recovering the RT mainly in the 10- and 20-genes concatenates revealed that some incongruence was still present in the 'essential' core. The analysis of the set of genes present in the 'universal' core and not included in the 'essential' core reveals that a substantial portion of the non-vertical signal that differentiates these two core sets is found in this 67-genes subset, which we have referred to as 'non-essential' core.

Therefore, we have shown that essentiality, defined as the intersection between universality and minimal gene set, is a more important factor than universality to recover the vertical signal of Proteobacterial genomes. However, we have also shown that the presence of incongruence is not always buffered even in cases where the number of concatenated genes is high. In consequence, we have analyzed the importance of a third factor, namely the function of the genes included in each data set. Due to the nature

of the three data sets we have been able to use both supertree and supermatrix approaches. Obviously, the composition of the core is clearly influenced by the special gene composition of the endosymbionts included in the study. These genomes have retained only those genes useful to their symbiotic association and to maintain the essential functions of the cell (Gil *et al.*, 2004b).

Many studies have shown a relationship between gene function and the evolutionary signal encoded therein, associating a higher frequency of lateral gene transfer to operational genes (Jain *et al.*, 1999; Nakamura *et al.*, 2004; Pal *et al.*, 2005). We have analyzed this signal in a phylogenomic context taking into account not only the functional category of the genes but also their assignment to each of the three data sets defined previously. In agreement with the results obtained in previous works, the informational categories seem to retain a better vertical signal than operational ones. The supertrees obtained for each of the three data sets with genes in the information category recovered the RT, whereas cellular, metabolism and poorly characterized genes showed a poor performance. In addition, the mean topological distance of each category to the RT confirms the high efficiency of the informational category with respect to the others, whose distance to the RT is significantly higher. However, a more detailed analysis reveals a more complex pattern.

Focusing in the three informational categories, the ‘transcription’ (K) category recovers the RT in all cases. Furthermore, this is the only category for which supertrees and concatenates perform equally well. Meanwhile the ‘Translation, ribosomal structure and biogenesis’ (J) category also presents a

good efficiency in the concatenate analysis. However, the ‘DNA replication, recombination and repair’ (L) category only recovers the RT in the ‘universal’ data set. Therefore, it seems that the ‘Transcription’ category is a good marker for phylogenomic exploration studies in which the vertical descent relationships of the species have to be assessed.

Metabolism genes usually represent the category with the highest frequency of horizontal gene transfer events (Pal *et al.*, 2005). Our analysis corroborates this result, as we have shown that the specialized categories encompassed by this general class have the largest distance to the RT. This result contrasts with the good performance of cellular categories, notably the ‘Posttranslational modification, protein turnover, chaperones’ (O) and ‘Cell motility and secretion’ (N) categories. In fact, the relative frequency of these categories is maintained or even increased over the three data sets analyzed. Even more interesting is the case of the ‘poorly characterized’ genes. Particularly, the ‘General function’ (R) behaves surprisingly well. Contrary to the ‘Function unknown’ (S) category, which practically disappears in the ‘universal’ and ‘essential’ cores, around 15 genes of the R category are present in these two data sets. The importance of these genes is being recognized now and their influence on bacterial evolution and adaptation is being studied (Galperin and Koonin, 2004; Glass *et al.*, 2006). Our results confirm the importance of some of these genes that seem to encompass a good vertical phylogenetic signal.

Finally, it is also remarkable the frequency of RT rejection through the SH test of genes belonging to each functional

category. Taking into account the whole genome, around 30% of the gene trees reject the RT and a similar fraction is maintained in the ‘universal’ and ‘essential’ cores. This incongruence could be due to the presence of non-vertical signals or to phylogenetic noise (for instance, insufficient signal in the corresponding multiple alignments). The same analysis but splitting the data set by functional category reveals that only the ‘Transcription’ (K) category has a significantly lower rate of rejection. This means that non-vertical processes and the presence of phylogenetic noise pervade all categories although, as we have shown, genes in some categories are better vertical markers than those in others.

We acknowledge the possible effects that including endosymbiont genomes could have in the recovered phylogenies. The evolution of endosymbiotic genomes is directly influenced by their lifestyle. Due to their relationship with the host, those genes that are not necessary for their survival are difficult to retain. This means that genes related to a free-living style or those related to motility are lost and most of the remaining ones are under weak selection or even in pseudogenization process (Wernegreen, 2002). This process of genome erosion translates most of the times into high A+T content and substitution rates that, from a phylogenomic point of view, imply possible convergences in the same clade of unrelated genomes, a phenomenon known as “long branch attraction” (Moran, 1996; Itoh *et al.*, 2002; Rispe *et al.*, 2004). These features have posed a challenge to traditional phylogenetic methods and are being revealed also as a conflicting point in genome phylogenies, mostly in those based on gene content. Our reference tree assumes the monophyly of the five

endosymbionts studied, a result derived in previous works although with some conflicting results (Herbeck *et al.*, 2004; Lerat *et al.*, 2003; Gil *et al.*, 2003; Charles *et al.*, 2001; Canback *et al.*, 2004). The inclusion in the data set of these genomes has two opposing effects. On the one hand, it reduces the number of genes shared among the species and thus affects the concatenate analyses. However, the number of genes shared by these Proteobacteria excluding these genomes is around 290, not much higher than the 200 genes found here (Charlebois and Doolittle, 2004). On the other hand, testing phylogenomic methods with these special conditionings also allows for testing their robustness and more general applicability.

3.5 CONCLUSIONS

As we have shown for this particular data set, supertree and supermatrix methods allow exploring all the phylogenetic signals contained in a bacterial genome. Both analyses are valid, complementary starting points in the evaluation of the incidence of events like vertical transmission, horizontal gene transfers, duplications or phylogenetic noise. Supermatrix methods are a double-edged sword, since they allow recovering the strongest phylogenetic signal even if the supermatrix is composed by alignments where this signal is hidden, i.e. it is not the strongest one (Gatesy and Baker, 2005; Gatesy *et al.*, 1999), but, at the same time, this could result in the masking of other alternative signals. The presence of these alternative signals is more easily revealed by analyses based on the underlying gene topologies. As shown in this work, these signals usually arise in a supertree or consensus framework in the form of incongruence around some taxa. For

instance, the supermatrix approach failed to reveal the phylogenetic incongruence around Xanthomonadales position in the base of the Gamma-Proteobacteria tree (Fig. 2) whereas both the '200-gene' trees consensus and the supertree of the '379- and 579-gene' trees placed them as Beta-Proteobacteria or as a low-supported Gamma-Proteobacteria clade. On the other hand, the selection of the data set it is also important. As we have shown, different data sets harbour different phylogenetic signals, the successful search for vertical, horizontal and noise signal may depend on the initial gene set selection.

Many factors have to be considered in selecting a methodology for whole genome phylogenetic analyses. We have shown that for intermediate phylogenetic depths it is possible to recover the primary phylogenetic signal of the genomes through the concatenation of shared genetic content although caution must be taken because usually a single model of evolution is applied to all the alignment, only recently the use of mixed models is being evaluated (Mark and Andrew, 2004). Meanwhile the supertree approach seems to be better suited to reveal conflicting gene trees signals (i.e. the positioning of Xanthomonadales) and when data are sparse (Sanderson and Driskell, 2003), which is usually the case for large evolutionary distances. One such example is the supertree obtained from 730 gene trees of 45 organisms from the three domains of life (Daubin *et al.*, 2002). The gene tree approach is suitable when working with species without complete genome sequences or for fast, exploratory phylogenetic analyses. These techniques are not incompatible but complementary, allowing the detection of possible sources of

error and pointing towards interesting evolutionary problems. As we have shown, the use of different methodologies and phylogenomic data sets allows the identification of the different phylogenetic signals encoded in bacterial genomes.

**4. THE PHYLOGENETIC LANDSCAPE OF
GAMMA-PROTEOBACTERIA**

4.1 INTRODUCTION

In chapter 3 we have presented some of the current approaches to bacterial phylogenetics taking advantage of whole genome data. However, the application of these new approximations does not guarantee success when applied to real data sets. Phylogenomic methods have inherited most of the problems of traditional single-gene approaches although others have been more easily overcome. For this reason we applied phylogenomic analyses in chapter 3 to a particular set of Proteobacteria which allowed us to test them under real conditions and also to address an interesting evolutionary question.

In particular, there is an ongoing debate on whether bacterial endosymbionts of insects from the Gamma subdivision of Proteobacteria conform a monophyletic group (Gil *et al.*, 2003; Lerat *et al.*, 2003; Canback *et al.*, 2004) or they are paraphyletic and their grouping results from artifacts in the phylogeny reconstruction process (Charles *et al.*, 2001; Herbeck *et al.*, 2004; Belda *et al.*, 2005). In chapter 3 we investigated the evolutionary relationships of five Gamma-Proteobacteria endosymbionts including the three *Buchnera* strains sequenced so far. Several evolutionary features of these genomes such as their high evolutionary rates, low G+C content and genome disintegration, lead to different methodological problems (Moreira and Philippe, 2000; Sanderson and Shaffer, 2002). New endosymbiont genome sequences were reported after these studies were carried out. Particularly interesting are a new *Buchnera* genome, *Buchnera*

aphidicola Cinara cedri, which is much more reduced than related genomes of the genus and *Carsonella ruddii*, endosymbiont of a psyllid which is characterized as the smallest bacterial genome with also the lowest G+C content of all endosymbiont genomes. This genome, as we will show in chapter 6, is particularly pervaded by A/T bias which translates from a phylogenetic point of view, into a special case of phylogenetic analysis under extreme conditions.

This chapter represents a blend of two phylogenomic analyses: an initial study of Gamma-Proteobacteria relationships with special emphasis in endosymbiont sequences carried out as explained in the chapter 3, and an extension of it as a result of the recent availability of new endosymbiont genome sequences, particularly that of *Carsonella ruddii*. However, although somewhat overlapping, these studies have very different objectives. The analyses in chapter 3 were aimed not only at exploring endosymbiont evolution but also the relationships among different Gamma-Proteobacteria genomes and at exploring phylogenetic incongruence which could eventually translate in the hallmark of horizontal gene transfer events as detailed in chapter 5. The second study is centred in the genome of *Carsonella*, whose composition and evolutionary rates force us to use new approaches in order to assess its most likely phylogenetic position.

4.2 MATERIAL AND METHODS

4.2.1 Data sets analyzed

In chapter 3 we have explained all the analyses carried out in order to establish the relationships among the genomes analyzed there, therefore this section and the Results section will be centred in the *Carsonella* genome analysis. For the *Carsonella* study we selected all the available Gamma-Proteobacteria endosymbiont genomes and also those from 19 additional species of this bacterial division in order to search for putative orthologs of each protein coding gene of *Carsonella*.

Genomes were downloaded from the NCBI repository (Benson *et al.*, 2002), see Table 4. For each protein sequence the reciprocal best hit was obtained (Evalue = 10^{-3}) and the annotation revised. However, the *Carsonella* genome has a large number of uncharacterized ORFs due to their much reduced similarity with other proteins, and its low GC content might also lead to some misidentifications. For these reasons, a few possible putative orthologs might have been missed in our search.

Once the putative orthologs were identified, we obtained alignments for the corresponding amino acid and nucleotide sequences with ClustalW (Thompson *et al.*, 1994). The resulting alignments were trimmed with Gblocks (Castresana, 2000) in order to eliminate positions of uncertain homology, phylogenetic noise mainly introduced by the highly divergent *Carsonella* sequences. In fact, the Gblocks procedure removed completely 37 amino acids alignments because of the many unconserved positions introduced by *Carsonella*. In order to assess how many genes of each alignment presented heterogeneity in the

composition at the amino acid level we performed the chi-square test implemented in TreePuzzle 5.2 (Schmidt *et al.*, 2002).

Table 4. Complete genome sequences of Gamma-Proteobacteria used in the *Carsonella* study. The abbreviations used for some figures are shown in the last column.

Organism	Size (Mb)	GC	RefSeq	Abbr.
<i>Acinetobacter</i> sp. ADP1	3.59	40.4	NC_005966.1	aci
<i>Baumannia cicadellinicola</i> str. Hc (<i>Homalodisca coagulata</i>)	0.69	33.2	NC_007984.1	bci
<i>Buchnera aphidicola</i> str. APS (<i>Acyrtosiphon pisum</i>)	0.66	26.4	NC_002528.1	buc
<i>Buchnera aphidicola</i> str. Bp (<i>Baiuzongia pistaciae</i>)	0.62	25.3	NC_004545.1	hap
<i>Buchnera aphidicola</i> str. Cc (<i>Cinara cedri</i>)	0.42	20.1	NC_008513.1	bac
<i>Buchnera aphidicola</i> str. Sg (<i>Schizaphis graminum</i>)	0.64	25.3	NC_004061.1	bas
<i>Candidatus Blochmannia floridanus</i>	0.71	27.4	NC_005061.1	bfl
<i>Candidatus Blochmannia pennsylvanicus</i> str. BPEN	0.79	29.6	NC_007292.1	bpen
<i>Candidatus Carsonella ruddii</i> PV	0.16	16.6	NC_008512.1	crs
<i>Coxiella burnetii</i> RSA 493	2.03	42.6	NC_002971.3	cbu
<i>Erwinia carotovora</i> subsp. <i>atroseptica</i> SCRI1043	5.06	51.0	NC_004547.2	eca
<i>Escherichia coli</i> K12	4.64	50.8	NC_000913.2	eco
<i>Escherichia coli</i> O157:H7 EDL933	5.62	50.3	NC_002655.2	ece
<i>Haemophilus influenzae</i> Rd KW20	1.83	38.1	NC_000907.1	hin
<i>Legionella pneumophila</i> subsp. <i>pneumophila</i> str. Philadelphia 1	3.39	38.3	NC_002942.5	lpn
<i>Mannheimia succiniciproducens</i> MBEL55E	2.31	42.5	NC_006300.1	msu
<i>Pasteurella multocida</i> a subsp. <i>multocida</i> str. Pm 70	2.26	40.4	NC_002663.1	pmu
<i>Photobacterium luminescens</i> subsp. <i>laumondii</i> TTO1	5.69	42.8	NC_005126.1	plu
<i>Pseudomonas syringae</i> pv. <i>tomato</i> str. DC3000	6.54	58.3	NC_004578.1	psb
<i>Salmonella typhimurium</i> LT2	4.95	52.2	NC_003197.1	stm
<i>Shewanella oneidensis</i> MR -1	5.13	45.9	NC_004347.1	son
<i>Shigella flexneri</i> 2a str. 2457T	4.59	50.9	NC_004741.1	sfx
<i>Sodalis glossinidius</i> str. 'morsitans'	4.29	54.5	NC_007712.1	sgl
<i>Vibrio cholerae</i> O1 biovar <i>eltor</i> str. N16961	4.03	47.5	NC_002505.1	vch
<i>Wigglesworthia glossinidia</i> endosymbiont of <i>Glossina brevipalpis</i>	0.69	22.5	NC_004344.2	wbr
<i>Yersinia pestis</i> CO92	4.88	47.6	NC_003143.1	ype

4.2.2 Phylogenetic analyses under extreme conditions

Endosymbiont sequences are characterized by a very low GC content. This is extremely exemplified in *Carsonella*, with a genomic GC content of 16%. As a result, it is difficult to evaluate in a phylogeny whether the relationships between low GC sequences are the result of an ancestor-descendant inheritance or they arise from evolutionary convergence. In consequence, we have implemented several phylogenetic approximations to correct the bias introduced by the extreme nucleotide compositions and fast evolving rates of endosymbionts.

For each nucleotide and amino acid alignment trimmed with Gblocks we obtained the corresponding gene tree by maximum likelihood using PHYML (Guindon and Gascuel, 2003). The general time reversible (GTR) model of nucleotide substitution and the Jones-Taylor-Thornton (JTT) model of amino acid evolution were used, respectively. In both cases, rate heterogeneity was approximated through a discrete gamma distribution with eight categories and shape parameter estimated from the data set, and a proportion of invariant sites was also incorporated in the likelihood model. Also the structural ribosomal RNAs, 16S, 5S and 23S were aligned and analyzed with the GTR + 8G + I model and, to correct to some extent for GC bias, the LogDet distance (Lockhart *et al.*, 1994) as implemented in the MEGA package (Kumar *et al.*, 2004).

In addition to this gene tree approach we also carried out a phylogenomic analysis. Supermatrices of amino acids and nucleotides were constructed using the 82 common genes to the

26 species. This resulted in two alignments of 67305 nucleotides and 20865 amino acids, respectively. The same models described previously for the gene trees were implemented. Different approaches were taken to account for the heterogeneity in GC content in the concatenated sequences. For the amino acids alignment, those peptides mainly affected by the bias introduced by enriched AT content (amino acids F, I, N, K, Y) were removed. The same procedure was adopted for the corresponding codons in the nucleotide alignment. Also, we implemented the RY-coding procedure that has been previously shown to reduce the effect of bias on phylogenies (Phillips *et al.*, 2004; Phillips and Penny, 2003). We codified as R (purine) or Y (pyrimidine) the third nucleotide position of each codon and analyzed the resulting alignment with a partitioned model in MrBayes (Ronquist and Huelsenbeck, 2003), one for the nucleotide partition (GTR + 8G+I) and a two state model (specified by the *nst=1* option) for the RY partition.

Supertrees of the gene trees were also retrieved by the MRP algorithm implemented in Clann software (Creevey and McInerney, 2005). We also used gene trees obtained after removal of FINKY amino acids in the corresponding gene alignments. Consensus trees were obtained for both the amino acid and nucleotide versions of the 82 common gene trees.

Finally, a congruence map analysis was carried out in order to evaluate the number of genes that rejected each gene tree topology. We selected those 82 genes common to the 26 genomes and their corresponding gene trees and carried out a maximum likelihood topology test known as expected likelihood weights (ELW) (Strimmer and Rambaut, 2002). Each gene alignment was

tested against every other gene tree and two other additional topologies whose main difference was in the position of *Carsonella* within or outside the endosymbiont clade (see below).

4.3 RESULTS

4.3.1 The Gamma-Proteobacteria phylogenetic landscape

As detail in chapter 3, the search for putative orthologs of the *Blochmannia floridanus* genome resulted in two sets of 200 common and 379 non-common genes. This allowed us to analyze the relationships among *Buchnera aphidicola*, *Blochmannia floridanus* and *Wigglesworthia brevialpilis* endosymbiont using several phylogenomic approaches.

To obtain a first notion of the relationships among our taxa, a first, reference tree was derived from the concatenation of 60 informational genes although similar results are obtained when 60 random genes chosen from the common gene pool are used. This reference tree pointed towards the monophyly of endosymbiont in a single clade near the YESS cluster. However, as indicated in chapter 3, this clade could be a phylogenetic artefact due to composition and high rate of change convergence.

Consequently, further phylogenomic analyses were aimed to establish if the reference tree topology was the most likely for the taxa studied. We increased on the one hand the number of characters of the matrix by concatenating the 200 common gene set (52029 positions, 40576 variable positions, 32921 parsimony informative positions) and on the other hand removed from this “200-genes” concatenate those aminoacids mainly affected by the

A+T bias (FINMKY) (16772 positions). Finally, an incomplete data matrix was obtained by concatenating the “579-genes” data set resulting in an alignment of 137301 aminoacid positions. The tree topology inferred from these analyses corroborated the reference tree topology with high support values and therefore supported the monophyly of the five endosymbiont genomes.

To evaluate the influence of the concatenate approach on the recovered relationships we also applied a supertree approach. We obtained the gene tree of each gene from the *Blochmannia* genome, the so called phylome, and analyzed their topologies first by a consensus tree of the 200 common gene trees and secondly by the comparative analyses of the supertrees derived from the 200-, 379-, and 579-gene tree data sets. The consensus tree retrieved the same topology than the reference tree for the endosymbiont clade but Xanthomonadales shifted to the Beta-Proteobacteria group. This possible signal of incongruence around Xanthomonadales position was corroborated in the 200- and the 379-gene supertree. The phylome supertree however recovered again Xanthomonadales in the Gamma-Proteobacteria branch although the bootstrap analyses supported their position as Beta-Proteobacteria.

4.3.2 Phylogenomic analysis of *Carsonella ruddii* position

The search for putative orthologs for the *Carsonella* phylogeny analyses resulted in a highly asymmetric distribution with 82 genes present in the 26 genomes and a group of genes only present in a range of 5-8 genomes (Figure 14). Most of these scarcely represented genes have orthologs in other endosymbiont genomes. In fact an analysis of the number of times that each genome is present in the data set revealed that the highest numbers of putative orthologs of *Carsonella* genes are found in the other endosymbiont genomes, thus possibly indicating the existence of a group of endosymbiont-specific genes. However, the influence of the high AT content and divergence of *Carsonella* and *Buchnera aphidicola* BCc sequences, which could lead to problems of convergence or loss of signal outside endosymbiont genomes, cannot be ruled out.

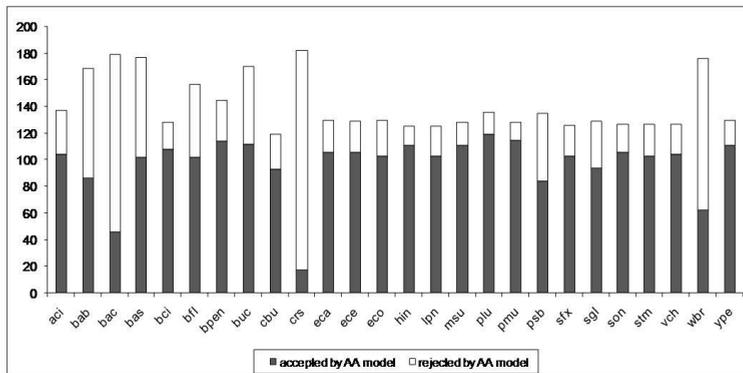
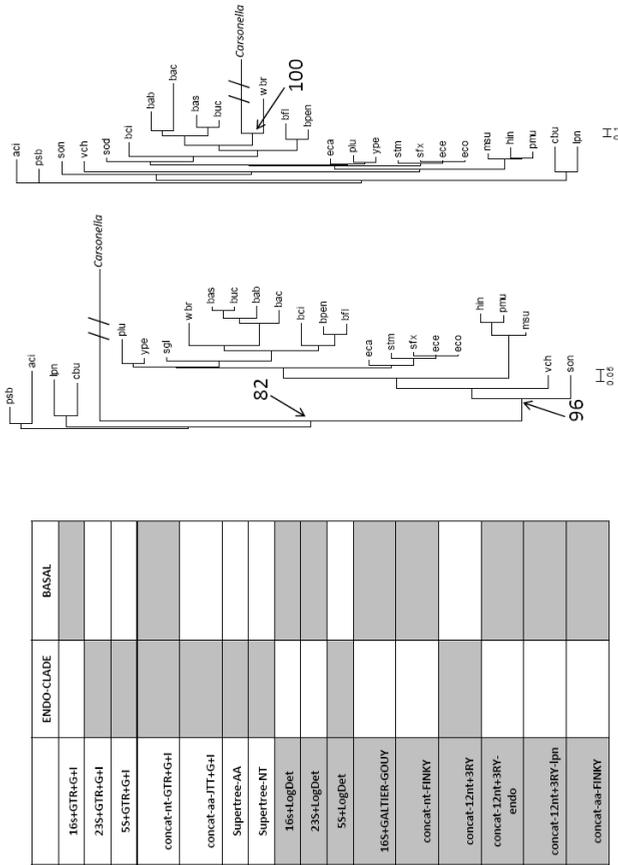


Figure 14. Number of genes per genome used in the *Carsonella* phylogenomic analysis. The black portion of the bars indicates the number of genes that rejected the assumed model of aminoacid evolution (JTT).

After alignment and trimming with Gblocks a number of genes were removed both at the amino acid (37) and the nucleotide level (1) due to the low quality of the resulting alignments, mainly due to *Carsonella*. In fact we analyzed the amino acid alignments without Gblocks trimming to search for sequences that might violate the homogeneity in composition assumption. Globally, 31.22 % of the total number of sequences (3671) rejected this assumption (Figure 14). The species-by-species analysis revealed, as expected, that it is the endosymbiont genomes the ones with a higher rejection rate. For *Carsonella*, 90.66 % of the sequences differed significantly in composition from the rest whereas other remarkable cases were those of *B. aphidicola* BCc (74.30%) and *Wigglesworthia* (64.77%). Among endosymbionts it is remarkable that *Blochmannia pennsylvanicus* genes do not have the same rejection level with values around those present in *Coxiella* or *E. coli* species. However, the most interesting case is that of *Baumannia*, whose rejection rate was among the lowest ones despite being an AT-rich endosymbiont genome. *Pasteurellales* were the group whose gene sequences violated the homogeneity assumption in a lesser degree. For phylogenetic methods, the presence of long branches, caused both by a presumably high rate of mutation (Itoh *et al.*, 2002) and an underlying AT bias, are very difficult to accommodate. We have approached this problem with multiple procedures, summarized in Figure 15. We have detected two main alternative phylogenetic positions for *Carsonella*, either at the base of Gamma-Proteobacteria near *Legionellaceae* or within the endosymbiont group, exemplified in the two trees shown in Figure 15. These placements were supported by different approaches.

Figure 15. Summary of the phylogenomic analyses applied to the Carsonella data set. Those aimed to correct potential A+T artefacts are shaded. Endo-clade refers to those which place Carsonella inside the endosymbiont monophyletic clade. Basal column refers to those analyses which place Carsonella out of the enterobacteria group, near Legionellaceae. Two representative alternative topologies. The one on the right correspond to the analysis of a nucleotide supermatrix from 82 common genes. The support value is a posteriori probabilities. The one on the left correspond to the 16s rDNA phylogenetic analyses using GTR, bootstrap values relevant for Carsonella placement are shown (see text for details).



In fact, different approaches aimed at correcting for the AT bias were also contradictory, as most of them placed *Carsonella* at the base of Gamma-Proteobacteria but with the remarkable exception of the RY coding approach.

Although some incongruence introduced in the phylogenetic reconstructions by other taxa could not be ruled out, we tested through a congruence map analysis which individual gene phylogenies were the most supported and whether these were congruent in the positioning of *Carsonella*. We tested each gene tree versus each gene alignment and added two plausible hypotheses, those derived from the RY coding and the one derived from the 16S rRNA (Figure 15). Figure 16 represents the congruence map derived from this analysis. On average each gene supported only 5.7 topologies which is revealing of the extreme phylogenetic signal diversity in this data set. The best supported phylogeny was that of the RY-coding which was accepted by 29 individual gene alignments (35% of the data set).

Traditional phylogenetic analyses using rDNA sequences supported both alternatives despite using Galtier-Gouy (1995) approach. So, the ML tree derived from 16S rRNA placed *Carsonella* at the base of the phylogeny with a relatively high bootstrap support (82%) while the same method applied to 23S rRNA sequences supported (BS = 91%) the *Carsonella* grouping with the remaining endosymbionts. Phylogenomic approaches also provided heterogeneous results. The concatenate of common amino acid sequences (20865 positions) supported the endosymbiont position whereas removal of amino acids most influenced by A+T bias (FINKY) (9968 positions) placed it at the

base of the tree, near the *Legionellaceae* clade (Figure 15). The nucleotide supermatrix (67305 positions) again placed *Carsonella* within the endosymbiont clade. In order to correct for the distorting effects introduced by the AT bias we also produced a concatenate of nucleotides with RY-encoding in third positions. A partitioned model was implemented in MrBayes and the best supported phylogeny retrieved *Carsonella* as a sister lineage of *Wigglesworthia* (Figure 15). Long-branch attraction artefacts cannot be ruled out completely in this reconstruction although the two extreme cases of high AT content (*B. aphidicola* BCc and *Carsonella*) were in different, monophyletic clades. Additionally, we removed the endosymbiont sequences except *Carsonella* and repeated the analysis which resulted in a basal position for the *Carsonella* genome, within the *Legionellaceae* group.

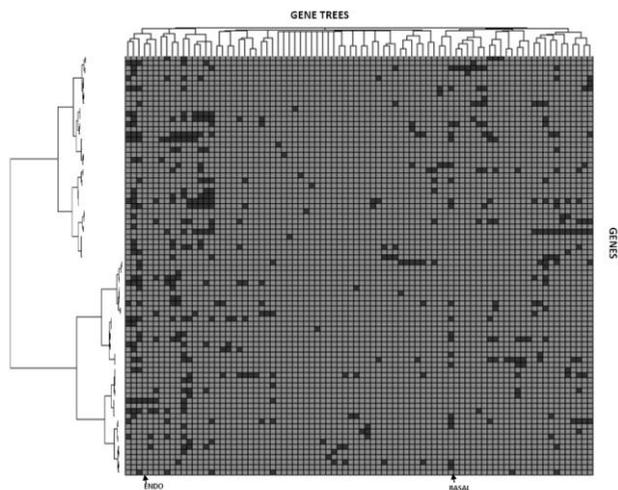


Figure 16. *Carsonella* congruence map. The figure represents the result of the phylogenetic congruence test of each gene versus each gene tree. Black squares correspond to cases where the gene tree is not rejected whereas gray squares mean rejection.

4.4 DISCUSSION

4.4.1 The Phylogenetic landscape of Proteobacteria: the placement of Xanthomonadales

We have identified two main points in the phylogeny of the Gamma-Proteobacteria bacteria from chapter 3. On one hand, the monophyletic or polyphyletic origin of insect endosymbionts has generated a debate. Most studies using a multi-gene approach have placed this group as a monophyletic clade sister to the YESS cluster (Lerat *et al.*, 2003; Canback *et al.*, 2004). However, studies with single gene phylogenies have pointed out that the group is not monophyletic (Charles *et al.*, 2001; Herbeck *et al.*, 2004), with the *Buchnera* clade as sister of the cluster YESS and suggesting that the other two endosymbionts (*B. floridanus* and *Wigglesworthia brevipalpis*) could have an independent origin related with secondary endosymbionts. On the other hand, the concatenate analysis, although presenting a strong primary phylogenetic signal, revealed some secondary topologies corroborated by the supertree/consensus analyses with incongruence in the placement of Xanthomonadales. In consequence, the topologies obtained in this study have been analyzed and discussed based on the position and evolutionary relatedness of the Xanthomonadales and insect endosymbionts groups.

The most frequently used phylogenetic marker in bacterial systematics is 16S rDNA (Woese, 1987). In this particular case, the 16S rDNA topology was not robust to different methods of phylogenetic reconstruction since the results obtained varied

depending on the method and model of evolution. However, despite these discrepancies all 16S rDNA topologies clustered the Xanthomonadales with the Beta-Proteobacteria. A screening of gene trees revealed that the three Xanthomonadales usually conform a monophyletic group whose placement in the Proteobacteria tree is unclear (see chapter 5).

As we have described, all the concatenate analyses revealed a main phylogenetic signal in which the Xanthomonadales were placed at the base of Gamma-Proteobacteria. However, a secondary phylogenetic signal was also observed. This signal also appeared in some single gene phylogenies in which the Xanthomonadales were excluded from the Gamma clade. Therefore, although the use of concatenates of different numbers of genes (60, 200 and 579) allowed us to establish the main phylogenetic signal, this approach also indicated the existence of alternative signals through different analyses evaluating the support for the main topology. Obviously, the larger the number of genes incorporating the alternative signal, the higher the support for conflicting phylogenies will be obtained.

Moreover, this secondary signal appeared more strongly in the consensus and supertree analyses. The majority rule consensus tree and the supertree of the 200 common genes as well as the supertree obtained from the other 379 gene trees placed again the Xanthomonadales with the Beta-Proteobacteria. These methods are only indirectly related to sequence data since they are based on gene trees derived from them. Because of this, they reflect what was already hinted in the single gene analysis, that an important

fraction of genes exclude the Xanthomonadales from the Gamma subtree, thus revealing the existence of either non-vertical evolutionary events, or methodological problems, or both, in the placement of Xanthomonadales in the source trees. The distinction between phylogenetic signal and noise around the position of the Xanthomonadales will be deeply analyzed in chapter 5.

4.4.2 The phylogenetic history of Gamma-Proteobacteria insect endosymbionts

The evolutionary relationship of Gamma-Proteobacteria endosymbionts remains a controversial issue. The first attempts to address this problem using single gene phylogenies produced conflicting results. For example, Heddi *et al.* (1998) and Schroder *et al.* (1996) supported the common ancestor hypothesis while Charles *et al.* (2001) defended a paraphyletic origin for insect endosymbionts. More recently, even with the availability of genomic data, the same conflicting results have been reproduced. Three other phylogenomic analyses support the common ancestor hypothesis: Gil *et al.* (2003) obtained this result after maximum likelihood and Bayesian analysis of 61 conserved proteins involved in translation; Lerat *et al.* (2003) obtained it from the common core of proteins, those present in the common ancestor and not involved in lateral transfers, for 13 Gamma-Proteobacteria species including one *Buchnera* and one *Wigglesworthia* species; and Canbäck *et al.* (2004) also examined the relationship of Proteobacteria including two *Buchnera* and *W. brevipalpis* using three different methods. These three studies agree with ours in

postulating a common ancestor for the endosymbionts and placing them as sister group to the YESS cluster.

However, two other studies have reached different conclusions: Herbeck *et al.* (2004) analyzed the 16S rDNA of endosymbionts and concluded that GC bias has influenced most published phylogenies. Using Galtier and Gouy's (1995) maximum likelihood method, they evaluated an array of possible phylogenies obtained by standard methods and models as well as other alternatives derived from the permutation of the *Buchnera* position in these source trees. The topologies preferred by Galtier and Gouy's method were those showing a paraphyletic origin of endosymbionts. This same strategy was used with the *groEL* gene and they reached the same conclusion. Although the two single gene analyses supported the paraphyly of Gamma-Proteobacteria insect endosymbionts, the corresponding topologies were incongruent and some taxa appeared in unlikely positions. A similar result has been obtained by Belda *et al.* (2005) using an alternative method for phylogeny reconstruction based on whole genome analysis. These authors have analyzed the number of pairwise genome rearrangements and gene order from 244 genes common to 30 Gamma-Proteobacterial genomes. Their analysis concluded that insect endosymbionts do not conform a monophyletic cluster, with *B. floridanus* being closer to *Escherichia coli* and separate from *Wigglesworthia* and *Buchnera*, genera that do not show a common origin themselves. Nevertheless, the lack of monophyly for insect endosymbionts is the only common result from these two studies, since the retrieved topologies are not compatible.

All our phylogenetic and phylogenomic analyses recover a single clade for endosymbionts with the notable exception of the 16S rDNA analysis with the Galtier and Gouy model which corrects for GC content biases. In consequence, we have addressed the possible influence of the GC content in the topologies recovered with the aim of discarding a convergence artifact in the rest of the analyses. Although we have based our phylome analysis on amino acid sequences, these are not free from base composition artifacts (Foster and Hickey, 2002). From the concatenate alignment of the 200 proteins common to all the genomes we removed those positions most likely to be influenced by GC bias. The trees obtained with two different methods were coincident and identical to the reference tree shown in Figure 6. Furthermore, the support values obtained for each node were even higher. Consequently, our results are robust to GC bias effects and they support the existence of a common ancestor for this group of insect endosymbionts.

The availability of *Carsonella ruddi* genome has allowed us in the second part of this chapter to determine whether the conclusions above exposed remains unchanged despite the addition of new endosymbiont genomes and to analyze the phylogenetic position of this extreme genome. There are few cases of *Carsonella* phylogenetics in the literature, and they are mostly based on single genes such as 16-23S rRNA (Thao *et al.*, 2000a; Thao *et al.*, 2000b; Thao and Baumann, 2004). We have approached this study with a variety of methods (Figure 15) aimed at correcting for the distorting effects on phylogenetic reconstructions introduced by the accelerated rate of evolution and increased AT contents of endosymbiont genomes and,

particularly, those of *Carsonella*. These features are usually responsible of long-branch attraction problems due to convergence caused by similar nucleotide or amino acid composition or by faster evolutionary rates as we have exposed previously. These two features have their extreme values in the genome of *Carsonella* and are most likely responsible for the difficulties in determining its correct evolutionary position with respect other bacteria.

As shown in Figure 14, 90% of the *Carsonella* sequences rejected the homogeneity in amino acid composition assumed by model of evolution. This means that all inferences derived from this data set cannot be considered reliable until the distorting effects possibly introduced by AT bias and evolutionary rates are corrected. Congruence maps are a mean to analyze alternative phylogenetic signals in gene tree sets. The one derived from the phylome of *Carsonella ruddii* left no room for such signals; almost every gene sequence is only compatible with its own gene tree. This result is a hallmark of phylogenetic noise. In opposition to the Xanthomonadales congruence map (which will be presented in chapter 5), there is no pattern of organization of the genes, a result expected when insufficient phylogenetic signal is present in the complete data set.

The phylogenomic analyses summarized in Figure 15 show that from the original data set and when no special methods are used to correct for the above mentioned effects there was a main agreement in considering *Carsonella* a member of the endosymbiont clade usually with *Wigglesworthia* as a sister taxon. The only exception was the 16S rRNA tree, which placed it at the

base of the tree, near *Legionellaceae*. This result was reversed when different corrections for the heterogeneity in composition were applied. Only two phylogenetic reconstructions retained *Carsonella* in the endosymbionts clade. One of these, the one derived from 5S rRNA, is not informative due to the short size of the sequence. The other exception is that of the RY-coding of third codon positions, which is supposed to be the best correction for AT bias. Again we found a major agreement, with a basal placement for *Carsonella*, with a very important exception.

In order to test the effect of the genome composition of the remaining endosymbiont genomes in the retrieved RY-coding phylogeny, we eliminated from the analysis the non-*Carsonella* endosymbiont taxa. This resulted in a basal position within *Legionellaceae* therefore indicating that the strong bias in composition is likely responsible for the positioning of *Carsonella* with the other endosymbionts in this phylogeny. However, our derived compositional tree showed that apart from endosymbionts, the genome with the most similar composition to *Carsonella* is that of *Legionella*. We tested again the effect of composition by eliminating *Legionella*, and in this case *Carsonella* remained as basal near *Coxiella*. Therefore, we consider that the most likely position of *Carsonella* in the Gamma-Proteobacteria tree is as a basal taxon, external to Enterobacteria. The failure of the RY coding to retrieve the basal position could be explained by the extreme AT bias of *Carsonella* and *Buchnera BCc* which pervade not only the third positions of the codons but also the other two, especially the first.

There are also some additional considerations to take into account. Firstly, none of the approaches breaks the grouping of the remaining endosymbionts in a single clade. That is, the correction for composition does not affect the monophyly of the *Buchnera-Blochmannia-Wigglesworthia* group except for the 16S rRNA with Galtier and Gouy's distance and therefore supports the main conclusion of the *Blochmannia* phylome analysis. Secondly, from a methodological point of view, the different approaches show both the strengths and the weaknesses of phylogenomics. Particularly, supermatrix approaches allow expanding the number of sites analyzed. However, as also reported elsewhere, the supermatrix does not solve bias problems and, on the contrary, they are apparently amplified by it (Phillips *et al.*, 2004). Examples are the phylogenies obtained with the RY-coding, although aimed at correcting the compositional bias, the extreme composition of *Carsonella*, and to a lesser degree of *B. aphidicola* BCc and *Wigglesworthia*, makes this correction insufficient as revealed by subsequent analyses. The Bayesian posterior probabilities of all the nodes in the RY-coding topology have a value of 1. The basal topology obtained after the removal of the endosymbiont taxa also retrieved BPP values of 1. Thus, support values in supermatrices, as argued in past decades for single-gene analyses, does not necessarily reflect the support for the correct phylogeny, but the extent in which the data and the model used to analyze them support the resulting, in this case probably incorrect, phylogeny.

4.5 CONCLUSIONS

In chapters 3 and 4 we have addressed a difficult evolutionary problem, the phylogenetic relationships of Gamma-Proteobacteria insect endosymbiont. We have presented evidence of the monophyletic relationships of most of these endosymbionts but the question about the evolutionary origin of *Carsonella* remains open. Although we consider that the most likely position of *Carsonella* is basal, out of the remaining endosymbionts clade, we acknowledge that the current analyses are not fully conclusive and that a position in the endosymbiont clade is still likely. Therefore this analysis shows the limitations of phylogenetic methodologies. Currently there are many methods and evolutionary models available for the phylogenetic analysis but they are not useful in extreme cases like this, where sequence degeneration and compositional bias pervade all the phylogenetic informative pieces and, therefore, make very unlikely the resolution of these particular cases even if more complex or adequate models are developed in the future.

**5. THE EVOLUTIONARY ORIGIN OF
XANTHOMONADALES GENOMES AND
THE NATURE OF THE HORIZONTAL
GENE TRANSFER PROCESS**

5.1 INTRODUCTION

On chapters 3 and 5 we have shown how it is possible to explore the different phylogenetic signals harbored by a genome. However, both chapters have been mainly devoted to, on the one hand, comparing current phylogenomic approaches and, on the other hand, answering questions on the relationships among endosymbiont genomes. Hence, both were centered in discriminating the vertical phylogenetic signals on these genomes which are also an exception among bacterial genomes, since they do not recombine and do not accept external genetic material. Therefore, the horizontal signal in endosymbiont genomes must be, if it still exists, very old and with few phylogenetic consequences. However, bacterial genomes are characterized by a continuous flux of gene losses through diverse mechanisms and gene gains mainly through horizontal gene transfer (Lerat *et al.*, 2005). Among the Gamma-Proteobacteria genomes analyzed in chapter 3, the Xanthomonadales were a group particularly difficult to place in the tree. Incongruence surrounding Xanthomonadales placement could be the hallmark of past horizontal gene transfer events or phylogenetic noise due to diverse causes. In this chapter will explore the phylogenetic signals in Xanthomonadales genomes, trying to differentiate not only among vertical and incongruent signals, but also among horizontal and phylogenetic noise signals.

Xanthomonadales is the most basal group of the Gamma-Proteobacteria clade and it is composed by phytopathogens ranging from obligate associations, like *Xylella* species, to non-obligate, like those belonging to the *Xanthomonas* genus (Van Sluys

et al., 2002). Previous works have revealed an unstable position of Xanthomonadales in the Proteobacteria tree (Van Sluys *et al.*, 2002; Beiko *et al.*, 2005). Both individual gene phylogenies and new genome phylogenies have placed them with the same frequency as Beta-, Gamma- or Alpha-Proteobacteria or as an external clade to the three groups (Van Sluys *et al.*, 2002; Omelchenko *et al.*, 2003; Martins-Pinheiro *et al.*, 2004; Bern and Goldberg, 2005; Dutilh *et al.*, 2005). Indeed a non-Gamma-Proteobacteria position for the Xanthomonadales is increasingly common in recently published phylogenies (Van Sluys *et al.*, 2002; Omelchenko *et al.*, 2003; Creevey *et al.*, 2004; Martins-Pinheiro *et al.*, 2004; Dutilh *et al.*, 2005; Studholme *et al.*, 2005). These reports point towards two most probable explanations: phylogenetic noise or horizontal gene transfer (HGT). Sometimes noise and HGT could not be discriminated but most of the times the hallmark of both processes is very different. In order to detect these differences is important to pay attention to the nature of the horizontal gene transfer process in bacteria.

It is known that the evolution of gene content of bacterial species is strongly influenced by their ability to incorporate DNA from other species in a process known as horizontal gene transfer (HGT) (Koonin *et al.*, 2001; Boucher *et al.*, 2003). The study of HGT events has shifted from reports of individual cases to genome-scale analyses taking advantage of the growing number of microbial genomes sequenced (Koonin and Galperin, 1997; Koonin *et al.*, 2001).

Although the importance of HGT as generator of evolutionary novelty is widely recognized (Ochman *et al.*, 2000), its

impact on the inference of organismal phylogenies is still hotly debated. The availability of a large number of microbial genome sequences has allowed the construction of genome phylogenies using different types of information (Snel *et al.*, 2005), with raw sequences, gene trees, shared gene content and shared gene order being most widely used. Typically, the impact of HGT on these genome phylogenies has been neglected and considered as mere phylogenetic noise in favor of a vertical signal resulting from the transmission of information from ancestors to descendants. Yet several authors (Doolittle, 1999b; Gogarten *et al.*, 2002; Kunin *et al.*, 2005) have claimed that retrieving a tree of life for bacteria is impossible, noting that 1) every gene has been transferred at least once during its evolutionary history (Dagan and Martin, 2007), and therefore 2) the phylogenetic signal associated to HGTs opposes, and often overcomes, the vertical signal, hence obscuring the deep phylogenetic relationships among current genomes.

The analysis of whole genomes has shown that incongruence between gene trees and organismal phylogenies is a pervasive feature of a significant fraction of genes from almost every bacterial genome except cases of obligate intracellular associations (Tamas *et al.*, 2002). This incongruence could arise from two main sources: phylogenetic noise and horizontal gene transfer. Phylogenetic noise primarily results from sequences with poor phylogenetic signal, high evolutionary rates for certain genes or lineages, or from long-branch attraction problems. On the contrary, the signals derived from HGT differ from noise because they usually reflect a robust and systematic incongruence towards

the donors. Therefore, conflicting phylogenetic signals coexist in bacterial genomes due to the vertical and horizontal histories of genes as well as phylogenetic noise. Divergent signals appear in core gene-sets (Baptiste *et al.*, 2005; Susko *et al.*, 2006) indicating that incongruence, often interpreted as HGT, could affect any gene and cellular function.

Gogarten *et al.* (2002) proposed a model that assumes that the likelihood of transfers is higher among related organisms thus generating a phylogenetic signal indistinguishable from the vertical one. Therefore, the tree-like evolution in bacteria is merely showing preferred paths for transfer events. In a recent analysis (Beiko *et al.*, 2005) 144 genomes were screened looking for the phylogenetic origin of all possible HGT events. This analysis suggested the preference of gene sharing among relatively closely related taxa, thus reaffirming the hypothesis that there are some constraints on HGT related to the compatibility of genome architectures and/or their phylogenetic distance (Gogarten *et al.*, 2002; Hendrickson and Lawrence, 2006).

Applying these concepts to Xanthomonadales evolution and the noise-HGT distinction may allow us to determine the source of phylogenetic incongruence in these genomes. On the one hand, phylogenetic noise is expected to affect up to certain degree gene phylogenies particularly for basal groups whose position may change due to limitations of phylogenetic reconstruction methods. On the other hand, following the Gogarten and coworkers' (2002) hypothesis, it is expected that transfers between Proteobacteria and Xanthomonadales genomes were more likely in the past when the divergence among the major

groups was relatively recent whereas recent transfers are expected to be among more closely related Xanthomonadales species. From a phylogenomic perspective, a high amount of ancient transfers from different donors to the Xanthomonadales ancestor might result in an unstable or unresolved position of these species in the Proteobacterial tree.

We have assessed whether the origin of the observed conflicting reports for Xanthomonadales in the Proteobacteria tree is the result of phylogenetic noise due to convergence and/or loss of signal or to recent or ancient HGT events. We have considered as phylogenetic noise the results of those processes unrelated to the form of transmission of a gene (horizontal or vertical) which violate the assumptions of phylogenetic reconstruction methods. Alternatively, we have considered as phylogenetic signal that derived from the vertical or horizontal transmission of genes. To separate these two components in the genomes of Xanthomonadales, we first identified all possible phylogenetic signals encoded therein. Next, we investigated the affinity of the genes for Gamma-, Beta- or Alpha-Proteobacteria clades. Our results indicate the existence of different, robust phylogenetic signals on the genomes of Xanthomonadales with origins in the three groups considered. We show that, unlike phylogenetic noise, these signals are not randomly distributed among genes; adjacent genes with the same phylogenetic signal appear more often than expected by chance in Xanthomonadales genomes, indicating that the signal detected is not due to selecting a conservative significance threshold for noise. These results are

analyzed in light of proposed models for the impact of HGT and preferred gene sharing paths in bacteria.

5.2 METHODS

5.2.1 Selection of homologs, gene alignments and gene trees.

Initially 18 Proteobacteria genomes were selected for analysis (see Table 5). The data set included a balanced number of representatives from the three major Proteobacteria groups and three Xanthomonadales genomes. We used *X. citri* as the base genome for selecting putative orthologs from the 17 additional genomes analyzed. For each protein coding gene we used a Reciprocal Best Hit BLAST strategy (Altschul *et al.*, 1997). We accepted as possible orthologs those genes that were reciprocal best hit between two genomes. BLAST searches were performed at the NGIBWS server (Charlebois *et al.*, 2003) with a very stringent criteria (E-value = 1E-10) to minimize problems associated to BLAST identifications. The annotation of each sequence and the corresponding multiple alignments were revised individually to discard wrongly identified putative orthologs.

As a first step, we analyzed possible incongruence not due to the evolutionary position of Xanthomonadales. We selected those proteins common to the 18 genomes and obtained their gene tree as explained below. With these gene trees we obtained a consensus topology. As the consensus reflects the most frequent position of each taxon in the tree, we could identify non-Xanthomonadales species with an unresolved phylogenetic

position and which therefore might affect future phylogenetic congruence analyses. Once these problematic taxa, *N. europaea* and *L. pneumophila*, were removed we repeated a BLAST search for the remaining genomes as explained above and finally only those sequences present in at least 10 species were considered for further analysis. This resulted in a set of 1051 genes of which 207 were present in the 16 genomes finally considered.

Table 5. *Genomes analyzed in this study.*

Organism	Taxonomy	Ref. Seq.	ORFs
<i>Xylella fastidiosa</i> 9a5c	GAMMA	NC_002488	2832
<i>Xanthomonas campestris</i> pv. <i>campestris</i> str. ATCC 33913	GAMMA	NC_003902	4181
<i>Xanthomonas axonopodis</i> pv. <i>citri</i> str. 306	GAMMA	NC_003919	4427
<i>Salmonella enterica</i> serovar Typhi strain Ty2	GAMMA	NC_004631	4324
<i>Rickettsia prowazekii</i> Madrid E	ALPHA	NC_000963	835
<i>Ralstonia solanacearum</i> GMI1000	BETA	NC_003295	5116
<i>Pasteurella multocida</i> PM70	GAMMA	NC_002663	2015
<i>Neisseria meningitidis</i> serogroup A strain Z2491	BETA	NC_003116	2065
<i>Mesorhizobium loti</i> MAFF303099	ALPHA	NC_002678	7275
<i>Haemophilus influenzae</i> Rd KW20	GAMMA	NC_000907	1657
<i>Escherichia coli</i> O157:H7	GAMMA	NC_002695	5361
<i>Caulobacter crescentus</i> CB15	ALPHA	NC_002696	3737
<i>Burkholderia mallei</i> ATCC 23344	BETA	NC_006349	4764
<i>Bradyrhizobium japonicum</i> USDA110	ALPHA	NC_004463	8317
<i>Bordetella bronchiseptica</i> RB50	BETA	NC_002927	4994
<i>Agrobacterium tumefaciens</i> C58	ALPHA	NC_003304	5299

Each selected protein from the set of putative orthologs was aligned with ClustalW (Thompson *et al.*, 1994) using default parameters. Phylogenetic trees were inferred by maximum likelihood with PHYML (Guindon and Gascuel, 2003), using JTT (Jones *et al.*, 1994) as the model of amino acid evolution with a Gamma distribution with 8 categories for modeling substitution rate heterogeneity among sites and an additional category of invariant sites estimated from the data set.

In order to explore the evolution of the genes identified in the previous step, we performed two different phylogenetic analyses. For assessing the different phylogenetic signals embedded in the Xanthomonadales genomes, we carried out firstly a “congruence map” analysis. Additionally, we addressed directly the question of the Proteobacterial origin of each gene by analyzing the support for different plausible evolutionary scenarios for each gene.

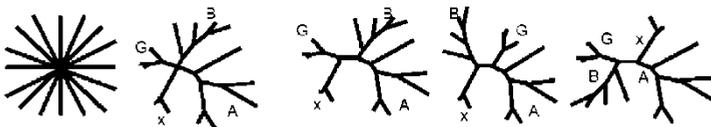


Figure 17. The five topologies used for the Phylogenetic preference test. The first two are named STAR1 and STAR2 and are used as phylogenetic signal control phylogenies. The other three topologies assume a placement of Xanthomonadales nearest to Gamma (G-tree), Beta (B-tree) or Alpha (A-tree) species respectively.

5.2.2 Congruence map analysis

We applied a previously described procedure (Baptiste *et al.*, 2005) to construct a congruence map for the 207-genes set. Each gene alignment was tested against every gene tree (207X207 comparisons) by means of the ELW test (Strimmer and Rambaut, 2002). For the congruence map, we only considered “acceptance” or “rejection” of the corresponding topology. The topologies (columns) were identified as Alpha-, Beta- or Gamma- by visual inspection.

5.2.3 Phylogenetic origin analysis

For the 1051 genes found in the search for putative orthologs, we tested their congruence to five possible phylogenetic hypotheses (Figure 17) using the ELW test of topologies. The five topologies were ‘artificially’ generated and can be divided into two groups. Firstly, we tested two control phylogenies: a star phylogeny (STAR1) with no resolved nodes, and a star-phylogeny (STAR2) in which the tips of the main phylogenetic groups (Alpha-, Beta-, Gamma-Proteobacteria and Xanthomonadales) were resolved. Secondly, we generated three phylogenies for which the only difference was the placement of Xanthomonadales at the base of Gamma- (G-tree) or Beta-Proteobacteria (B-tree) and between Alpha and the Beta-Gamma-split (A-tree).

For most genes the ELW test could not reject all but one phylogeny. In consequence, we eliminated from further analyses those genes that could not reject both star phylogenies, hence ensuring the presence of some phylogenetic signal. The

preferential phylogenetic origins of the genes were mapped into the three Xanthomonadales genomes considered.

5.2.4 Testing for long-branch attraction artifacts

Once the most plausible phylogenetic origin of each gene had been assessed, we tested for possible convergences due to shared high rates of substitutions and not to common origins. We used the program RRTree 1.0 (Robinson-Rechavi and Huchon, 2000) with the 207 common genes data set. We divided the species into four groups according to their taxonomic assignment (Alpha-, Beta-, Gamma-Proteobacteria and Xanthomonadales) and performed all possible pairwise comparisons of substitution rates between the four groups. As a common outgroup we chose the corresponding gene of *Rickettsia prowazekii*. The significance of the difference in number of substitutions for each comparison was assessed at the 0.0083 level (Bonferroni corrected significance level $[\alpha]$, 0.05) to take into account the six non-independent comparisons performed for each gene.

5.2.5 Testing for functional association and clustering along the genome

We have studied the relationship between phylogenetic origin and functional assignment of the genes. We used the functional categories described in the COG database (Tatusov *et al.*, 2000) to classify the genes in 4 general or 21 more detailed categories. We have also compared our results with a list of virulence-associated genes of *Xanthomonas citri*.

We used a subset of the previous data set to analyze whether genes with the same phylogenetic origin tended to cluster

in the three Xanthomonadales genomes. We selected those genes with at least one adjacent gene with phylogenetic information. We computed the number of the six possible combinations of Alpha-, Beta- and Gamma- assignments for all pairs of adjacent genes in each genome. We calculated the expected number of pairs in each category under the assumption of independent origin for each gene in a pair. Let AA, BB, GG, AB, AG and BG denote the possible observed pairs. We obtained the observed frequencies of each phylogenetic assignment (p_A , p_B , p_G) and calculated the expected frequency of each pair as the product of the individual frequencies of its components. For example, the expected number of Alpha-Beta-pairs is given by $2 \cdot p_A \cdot p_B \cdot N$, where N is the total number of pairs considered. The observed and the expected number of pairs were compared by means of a chi-square test with 3 degrees of freedom (df). A whole genome alignment of the three Xanthomonadales genomes was carried out with MAUVE (Darling *et al.*, 2004) in order to study the possible influence of rearrangements in the results of the clustering test.

Lastly, we tried to determine whether adjacent pairs of genes with the same phylogenetic origin were also in a potential operon. Pairs of genes used in the clustering analysis were assigned to two groups, adjacent and non-adjacent pairs. For each group we counted the number of times the members were in the same direction of transcription, the number of divergently transcribed gene pairs and the number of convergently transcribed gene pairs. A chi-square test was used for assessing the significance of the possible differences between adjacent and non-adjacent pairs with 2 df.

5.2.6 Detection of atypical genes

We have used a technique (Azad and Lawrence, 2005) which takes advantage of common parametric measures for HGT detection and AIC as a criterion for clustering. This technique allows the identification not only of clusters of native genes in the *X. citri* genome but also of different clusters of atypical genes. The two parametric measures used were nucleotide composition and codon bias. Briefly, these measures were computed for all the genes in the *X. citri* genome and the AIC criterion was used for deciding when the addition of more genes to a cluster was not significant. Frequently, this procedure retrieves a large gene cluster that usually corresponds to the native or typical genes of the genome analyzed and smaller clusters of genes with atypical features. Genes shorter than 300 bp were excluded because their low information content could result in GC content or codon usage biases (Lawrence and Ochman, 2002). A summary of the analysis pipeline followed in this work is shown in Figure 18.

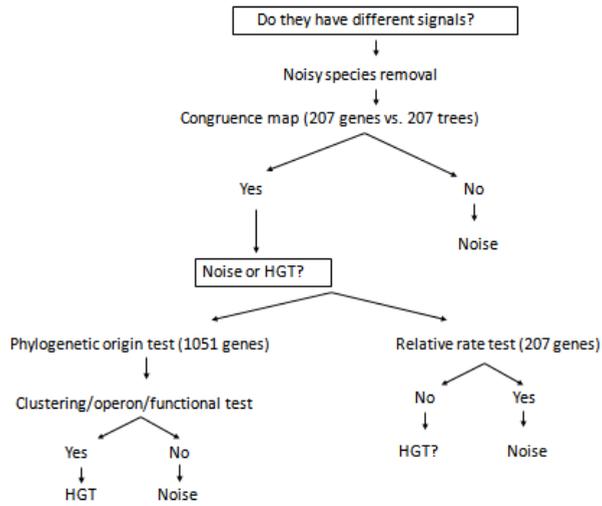


Figure 18. General pipeline of the methodology followed to analyze phylogenetic signal in the genomes of 3 *Xanthomonadales* species when compared to other *Proteobacteria* genomes.

5.3 RESULTS

5.3.1 Data set analyzed and exploratory analyses

We have analyzed a set of 18 *Proteobacteria* genomes in order to study the phylogenetic origin of *Xanthomonadales* genes. The dataset included a balanced number of representatives from the three major *Proteobacteria* groups and three *Xanthomonadales* genomes, *Xanthomonas axonopodis* pv. citri str. 306 (*X. citri*, Xci), *X. campestris* pv. campestris str. ATCC 33913 (*Xca*) and *Xylella fastidiosa* 9a5c (*Xy. fastidiosa*, Xy). Our goal was to analyze the different phylogenetic signals present in their genomes as well as to determine the phylogenetic origin of their genes from Gamma-, Beta- or Alpha-*Proteobacteria* ancestors.

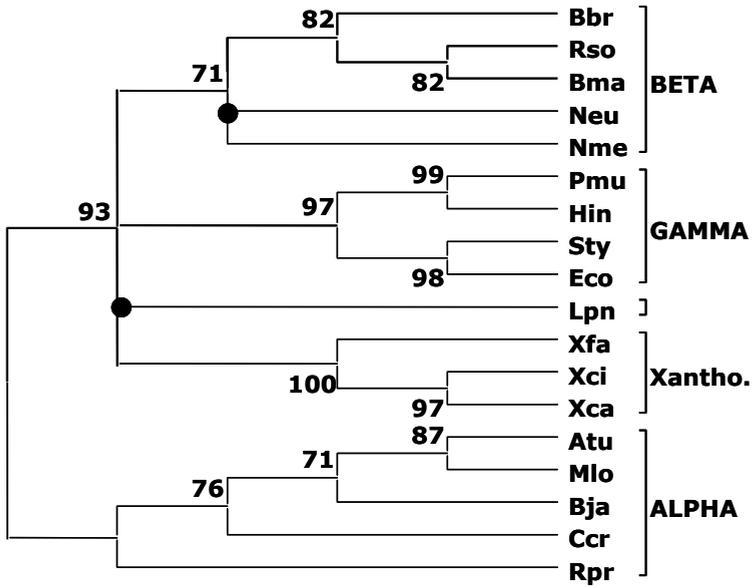


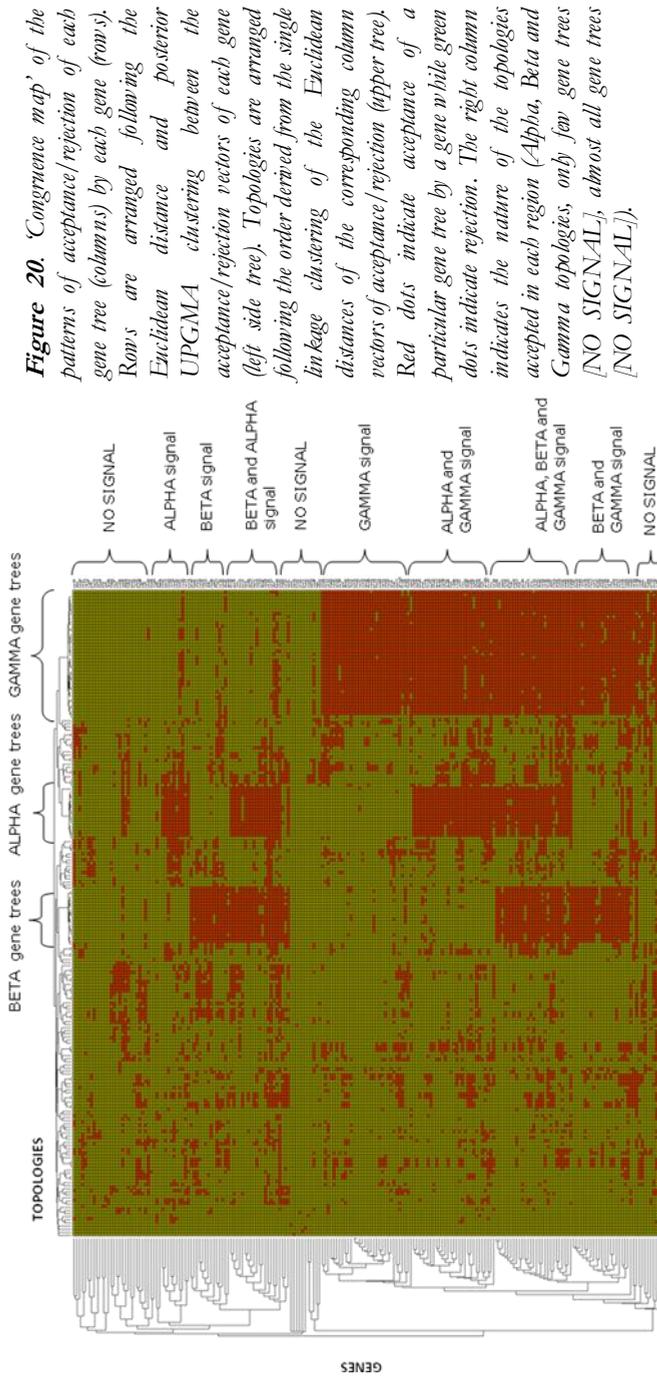
Figure 19. Majority rule consensus for the initial set of 18 genomes of the 207 gene trees used in the ‘Congruence Map’ analysis. The tree is arbitrarily rooted with the ALPHA branch. The nodes show the frequency of appearance of the corresponding group before removal of *Legionella pneumophila* and *Nitrosomonas europaea*. Species names and taxonomy groups according to NCBI.

Our initial search for putative orthologs retrieved 207 genes common to all the genomes. Our first goal was to identify “noisy” phylogenetic signals. As our analysis was intended to detect only incongruence related with Xanthomonadales we looked to eliminate other species that could introduce noise in the global phylogenetic analysis. We constructed the majority rule consensus tree of the trees obtained from these 207 common genes set as a way of summarizing the degree of incongruence present in each species (Figure 19). The tree identified three nodes with low resolution, those corresponding to *Nitrosomonas europaea*, *Legionella pneumophila* and the Xanthomonadales group. Since we

were only interested in the later, we discarded *Nitrosomonas europaea* and *Legionella pneumophila* from the ensuing analyses.

5.3.2 Congruence map

With these 207 common genes of the remaining 16 genomes we tested for the presence of different phylogenetic signals. We performed a ‘congruence map’ analysis in which each gene was tested for congruence against all the other gene trees (Figure 20); here, each row corresponds to a gene and each column to a gene tree. The analysis identified numerous genes whose phylogenetic reconstructions provide clear and robust support for other Alpha-, Beta- and Gamma- gene trees topologies. These were defined on the basis of the monophyletic grouping of Xanthomonadales sequences with the remaining sequences of each group. In addition, many tests were able to reject some, but not all, alternative phylogenies. The cases ranged from genes that were compatible only with their own gene tree to those that could not distinguish between Alpha-, Beta- or Gamma- topologies. Overall, the Gamma-topologies were the most frequently accepted, considering both cases in which this was the only topology selected and those with other topologies being also accepted. This analysis showed a mixture of noisy phylogenetic signals, which corresponded to genes only congruent with their own gene tree, and genes that were congruent with almost any topology. But, in addition to noisy signal, robust but divergent phylogenetic signals were detected in terms of acceptance or rejection of groups of topologies corresponding to different positions of Xanthomonadales with respect to other Proteobacteria.



The 207-gene analysis strongly suggested that HGT might have played an important role in the evolution of Xanthomonadales, but other alternatives could also be considered. The HGT hypothesis was tested by selecting a larger gene data set that would result in more robust statistics. This allowed us to analyze whether the cause that some genes were unable to reject alternative, incompatible hypothesis in the congruence map was the result of phylogenetic noise or the hallmark of past HGT events. Consequently, we extended the initial data set to incorporate genes from *X. citri* with orthologs in at least 10 genomes. The extended set of 1051 genes was composed of quasi-universal genes in Proteobacteria with functions not directly related to the virulence of *X. citri*. A comparison with a list of known virulence-related genes identified only 19 candidates.

5.3.3 Atypical genes detection

To determine if the genes contributing conflicting phylogenetic signals were recently introduced in the Xanthomonadales genomes, we identified atypical genes by a clustering methodology based on the AIC criterion using both codon usage bias and nucleotide composition as discriminating criteria (Azad and Lawrence, 2005). We found a main cluster of typical genes and several clusters of atypical ones. The analysis revealed that only 0.2% for codon usage and 13.61% for nucleotide composition of the genes used in this study were atypical. Meanwhile, the frequencies of atypical genes in the whole genome were 2.21% and 22.23% respectively. Therefore, the incongruence observed in the ‘congruence map’ analysis cannot be attributed to the confounding influence of recent HGT events.

Although some of these cannot be ruled out, this is an unlikely scenario for the whole set of conflicting genes, since these phylogenetically incongruent genes were found in all three Xanthomonadales genomes, with appropriate branching orders. These results suggest that the transfer events were old.

5.3.4 Long-branch attraction artifact incidence

To evaluate the possible incidence of long-branch attraction artifacts in our data sets we also carried out relative rate tests for the 207 genes common to all the genomes. The analyses revealed only one case of possible convergence due to shared high rates of substitutions (corresponding to the *hisS* gene). The remaining genes showed no evidence of grouping due to shared accelerated evolutionary rates and, in consequence, we excluded this phylogenetic artifact as responsible for apparently incongruent groupings.

5.3.5 Phylogenetic origin test

To verify that ancient transfers to Xanthomonadales genomes resulted in phylogenetic incongruence among quasi-universal genes, we examined the compatibility of each gene with five phylogenetic hypotheses (Figure 17). The STAR1 and STAR2 topologies are unresolved topologies with the difference that in the latter the tips of the major Proteobacteria groups are resolved; genes with strong phylogenetic signal should reject these topologies. The other three topologies placed the Xanthomonadales clade as the most basal group of the Gamma- (G-tree), Beta- (B-tree) or Alpha- (A-tree) groups. All of the 1051 genes analyzed rejected the STAR1 topology but 51 genes could not reject the STAR2 topology; these were removed from the

ensuing analyses. The distribution of the most likely phylogenetic origin of the remaining genes is shown in Figure 21.

A majority of genes preferred the A-tree. However, most of these genes were unable to reject some or all the other topologies. In consequence, posterior analyses were based on the most likely assignment regardless of their compatibility with other alternatives. In any case, an analysis of those genes selecting only one of the topologies revealed the same pattern, with Alpha topologies as the most preferred and Beta- topologies as the least.

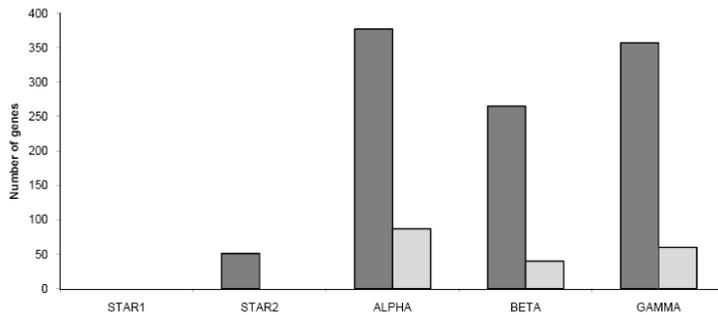


Figure 21. Histogram of the preferred phylogenetic assignment of the 1051 genes analyzed. As most genes are compatible with more than one topology the figure shows the results for the complete dataset (dark color bars) and that for the 187 genes that are congruent with only one of the five hypotheses (light color bars).

5.3.6 Testing for functional association and clustering along the genomes

As the phylogenetic origin test was not enough to distinguish between phylogenetic noise and phylogenetic signal, we tested the noise threshold by analyzing the distribution of the genes and their possible origins in the Xanthomonadales genomes. An adjacency analysis was carried out with a reduced subset of the 1051-genes set. We selected only those genes that were adjacent at least to another gene from this subset in the genomes of the Xanthomonadales. The number of pairs analyzed was 430 in *X. citri*, 438 in *X. campestris* and 377 in *Xylella fastidiosa*. This allowed us to test two alternative predictions. If incongruence was merely due to noise, then pairs of adjacent genes would show no association with respect to their phylogenetic origins. On the other hand, at least under certain models for HGT (Lawrence and Roth, 1996), pairs of genes adjacent in the recipient genome should tend to share the same phylogenetic origin.

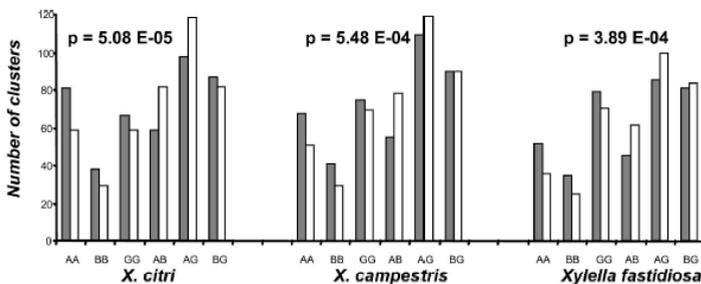


Figure 22. Number of expected (white) and observed (grey) pairs of consecutive genes with different combinations of Alpha- (A), Beta- (B) and Gamma- (G) origins. The *p*-value of the chi-square test is shown for the *Xanthomonas axonopodis* pv. *citri* (*X. citri*), *Xanthomonas campestris* (*X. campestris*) and *Xylella fastidiosa* (*Xy. fastidiosa*) genomes.

Our statistical tests revealed that the number of adjacent pairs of genes with the same phylogenetic origin in the Xanthomonadales genomes was higher than expected (Figure 22). Furthermore, such clustering was evident and significant for all three Xanthomonadales genomes and the three possible phylogenetic origins considered. This evidence of clustering highlighted two aspects. On the one hand, it rejected the possibility that most of these results were simply the product of phylogenetic noise although this was evidently present in some degree. On the other hand, although the adjacency test reduced the analysis to the observed and expected number of pairs, the actual size of the clusters identified tended to be larger than two genes, with examples including as many as eight genes. The mean size of the clusters was 2.35 genes, pointing towards horizontal gene transfer of operons as a possible mechanism of evolution.

If operons were being transferred between Gamma-, Alpha- and Beta-Proteobacteria lineages, resulting in the phylogenetic incongruence seen in Xanthomonadales genomes, then the overabundance of adjacent genes with common phylogenetic signals should be biased towards genes transcribed in the same direction. We studied the transcription direction of the genes present in the adjacent pairs analyzed above; as expected under the hypothesis of horizontal transfer of complete or partial operons, most of the genes identified were present in the same strand. Furthermore, the frequency in which adjacent genes were in the same direction of transcription was higher than that of non-adjacent cases, with significant statistical support for pairs of Alpha- ($p < 0.0002$) and Gamma- origin ($p < 0.002$), but not for

those of Beta- origin ($p = 0.3079$). Thus, it is likely that most of the clusters identified in our study, mainly those involving Gamma- and Alpha- origins, were operons.

We also investigated the relationship between functional assignment of the genes and their established phylogenetic origin to test whether informational genes were less prone to be transferred than non-informational ones (Jain *et al.*, 1999). We did not detect any association between functional classes and putative phylogenetic origin, but some patterns could be distinguished. For example, seven of the eight flagellar genes analyzed showed the same phylogenetic origin, B-tree. Alternatively, the informational category was richer in G-tree topologies while the predominant topology for metabolic genes was the A-tree. As a consequence, the different composition in functional categories of the 207-genes set (richer in informational genes) and the 1051-genes set might explain the differences in the most frequent origin of the genes in each set (G-trees and A-trees, respectively). Nevertheless, most of categories presented a mixture of topologies (Figure 23).

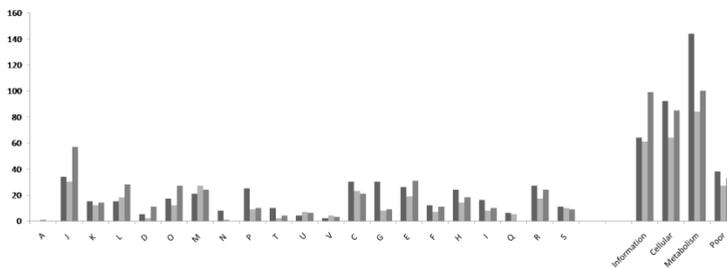


Figure 23. Functional assignment of the genes to a most plausible phylogenetic origin (from left to right: Alpha-, Beta-, Gamma-)

5.4 DISCUSSION

5.4.1 Xanthomonadales evolution illustrates the nature of the HGT process

The influence of horizontal gene transfer on the reconstruction of bacterial phylogenetic relationships and also on its relevance to shape their genomes has been a hotly debated issue. Different models have been proposed to explain patterns of transfers derived from complete microbial genomes (Jain *et al.*, 1999; Gogarten *et al.*, 2002; Kunin *et al.*, 2005). Gogarten *et al.* (2002) proposed that if there is a negative correlation between the likelihood of transfers and the evolutionary distance separating two taxa, then horizontal gene transfers are more likely among closely related taxa, thus generating a cohesive signal for the clade that agrees with that of vertical transmission.

One of the clues for testing the different proposals could be the barely studied subject of constraints to HGT. It is clear that there are some restrictions to transfers but most studies have focused on functional analyses (Jain *et al.*, 1999; Nakamura *et al.*, 2004; Pal *et al.*, 2005). However, two recent works have shed light on the probabilities of successful transfers from a phylogenetic point of view. The results of Beiko *et al.* (2005) are consistent with an uneven phyletic distribution of the transfers while Lawrence and coworkers (Lawrence and Hendrickson, 2003; Lawrence and Hendrickson, 2004; Hendrickson and Lawrence, 2006) have pointed out a possible molecular mechanism that limits the equally probable sharing of genes among all bacteria.

The analysis of 220,240 proteins from 144 genomes (Beiko *et al.*, 2005) has revealed a consistent vertical signal but also the relevance of horizontal transfer in shaping bacterial genomes. The gene tree for each protein was reconstructed and a reference supertree that is supposed to represent the vertical history of the species was derived. In this analysis, the mean number of steps for reconciling the trees was surprisingly low for a phylogeny in which all the taxonomic groups of Bacteria and some Archaea were represented. From a biological point of view, these translate into preferential sharing of genes among bacterial species from the same group or from close divisions such as among Proteobacteria clades. These preferences for gene sharing also reveal the presence of limitations to random transfers.

These results suggest that there are constraints in the genomic architecture that make gene acquisition from distantly related taxa unlikely. A possible mechanism for such constraints has been reported (Lawrence and Hendrickson, 2004; Hendrickson and Lawrence, 2006). These authors analyzed the distribution of octomers along the bacterial chromosome. Some octomers are preferentially found on leading strands, and increase in abundance towards the replication terminus; here, selection would be maximal for their role in efficient chromosome segregation (Hendrickson and Lawrence, 2006). This selection at the level of chromosome structure could have important implications for bacterial chromosome dynamics. For instance, DNA compatible with the octomer distribution of a recipient genome would have a higher chance of being transferred successfully. This implies that successful transfers will be more

likely between closely related, and therefore compatible, genomes although they are most difficult to identify with current methodologies (Lawrence and Hendrickson, 2003; Lawrence and Hendrickson, 2004).

Our results are compatible with the predictions of Gogarten and coworkers (2002) as reflected in Figure 24. We have found HGT events to the ancestor of Xanthomonadales, therefore prior to their diversification and, most likely, to the diversification of nascent Proteobacteria lineages. The low number of atypical genes detected among them also reveals the old age of these transfers. Ongoing, or recent, transfers to Xanthomonadales genomes are likely, as reflected by the proportion of atypical genes in the *X. citri* genome, but are not those detected by our phylogenetic origin test, since a much lower proportion of atypical genes was identified among the genes causing conflicting phylogenetic signal in this genome. Since these recent transfers do not result in an incongruent position of the involved Xanthomonadales taxa out from this group, it seems reasonable to assume that most of these recent transfers have occurred among members of the Xanthomonadales clade and not with other Proteobacteria or more external groups. Therefore, the age of transfers reveals the more likely partner(s) throughout the evolutionary history of the group: other Proteobacteria lineages in the past, when they had not diverged much yet, or other Xanthomonadales lineages in recent times, when the divergence between Xanthomonadales and other Proteobacteria groups is significant but not so within the Xanthomonadales group.

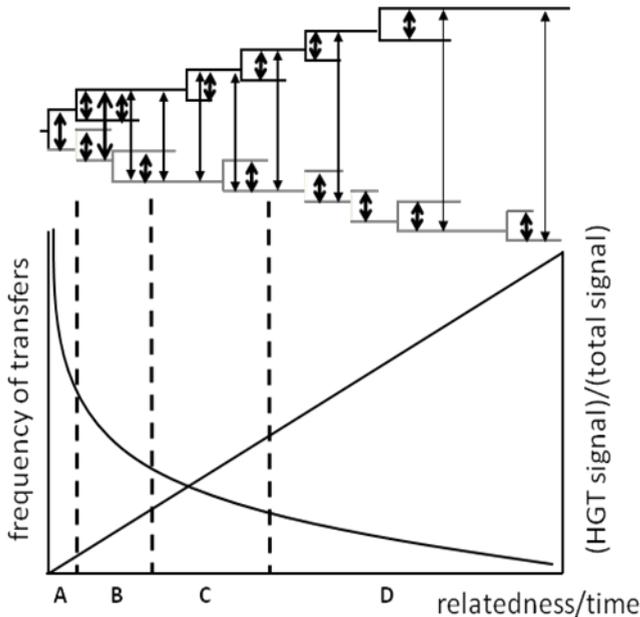


Figure 24. The figure reflects changes in the frequency of HGT events (curve line) and amount of accumulated HGT signal along time (straight line) in *Xanthomonadales* evolution. The frequency of transfers is also represented by the thickness of the arrows inside the genealogy. The figure can be interpreted in two alternative ways. Firstly, the x-axis could reflect the evolutionary relatedness of extant lineages. Then, regions A and B reflect exchanges within the *Xanthomonadales* clade, therefore between very close taxa. These transfers correspond to recombination within populations (A) and to relatively recent transfers detected in our atypical gene analysis (B). Regions C and D are those corresponding to current transfers from more distant groups, for example from other Proteobacteria (C) or even more distant taxa (non-Proteobacteria and Archaea) (D). Secondly, the x-axis could be interpreted as time, therefore reflecting the evolutionary history of a lineage from its origin to the present. In this case, the grey lineage represents the genealogy of current *Xanthomonadales* and the black lineage the genealogy of a hypothetical Proteobacterial lineage. In region A, since these lineages had not diverged much, the frequency of transfers between them was very high. This and region B reflect ancient transfers to the *Xanthomonadales* ancestor as the ones detected in our test of phylogenetic origin. As nascent Proteobacteria lineages started to diverge, the frequency of exchange between *Xanthomonadales* and other Proteobacteria began to decrease (C). Finally, the number of transfer events between Proteobacteria and *Xanthomonadales* have disappeared almost completely and only within group events, such as those revealed by our atypical gene analyses and the consistent monophyly of the *Xanthomonadales* clade, remain (D).

Additionally, this result reveals the different effects of HGT events on phylogenetic reconstructions depending on the time since the transfers. Those methods based on gene trees, such as supertrees and consensus (Figure 19), are appropriate to pinpoint (in the form of unresolved nodes) high amounts of transfers in the distant past. The higher the number of past transfer events on the ancestor of a clade the higher the likelihood of retrieving its corresponding branch as an unresolved position in bacterial genome phylogenies. On the other hand, all gene trees derived from the 207 common genes set support the monophyly of the Xanthomonadales clade. This strong signal may be powered not only by the vertical transmission of its members but also by current horizontal gene transfers inside the group (Tettelin *et al.*, 2005). Recombination between closely related strains and horizontal gene transfer between close, intra-genera species could result in the observation of a clear, vertical signal in genome phylogenies because there is not enough divergence to be detected at a genome-scale analysis.

5.5 CONCLUSIONS

To sum up, two conclusions have been outlined. On one hand, Xanthomonadales genomes have approximately the same number of genes with Beta-, Gamma- or Alpha-Proteobacteria affinity, making them an extreme case of mosaicism and preventing us from conclusively assigning them to one of the major Proteobacteria clades. On the other hand, we have shown that it is possible to disentangle noise from signal through

exhaustive and careful analysis even in the most complex cases like Xanthomonadales evolution. We have shown the existence of ancient and recent transfers despite possible phylogenetic artifacts. The effect of these transfers in phylogenomics and the resolution of ancestral nodes will depend on the vertical/horizontal signal ratio in the branches leading to a node. These ratios will determine which parts of the genome trees are tree-like and which are not. Obviously, Xanthomonadales seem to fit this model since they appear as a monophyletic group recovered in almost all gene phylogenies. Meanwhile their correct position in the Proteobacteria tree is obscured by the presence of a high amount of ancient transfer events to their common ancestor as shown in this analysis. Other groups such as Pseudomonadales and some Cyanobacteria also seem to present high ancient HGT rates (Beiko *et al.*, 2005) and may follow the same pattern. Alternatively, the evolutionary scenario of resolved tips and poorly-resolved deep nodes should not apply to genomes with less promiscuity or susceptibility to HGT in ancient times, in which case their vertical phylogenetic signal in the past had more weight than the horizontal one, thus allowing a better resolution of their deep phylogenetic relationships.

**6. FIGHTING FOR SURVIVAL: OPPOSING
EVOLUTIONARY FORCES ACT IN THE
LAST STAGES OF GENOME REDUCTION**

6.1 INTRODUCTION

Gene gain through horizontal gene transfer has been shown to be a major mechanism of evolutionary novelty for bacterial genomes as detailed in chapter 5. However, bacterial genomes sizes remain equal or lower than the estimates obtained from ancestral reconstruction of common ancestor genomes (Dagan and Martin, 2007). New genetic material introduced by HGT is compensated by a similar loss rate either because it results in the substitution of the native gene or because there is a process of sequence loss, mainly by the streamlining of the genomes through gene inactivation and loss or through reduction of non-coding DNA regions. Endosymbiont genomes are an extreme case of gene decay. Most of them have lost the capacity for genetic exchange as well as genes related to repair pathways (Wernegreen, 2005). The tempo and mode of these genome reductions are being studied. The most likely scenario seems to be a first stage of large deletions, possibly including more than one gene, and a second stage of a more specific gene inactivation, pseudogenization and loss through mutation and gradual deletions (Mira *et al.*, 2001; Silva *et al.*, 2001).

In this context, the extremely reduced genome sequences of the bacterial endosymbionts *Carsonella ruddii* (Nakabachi *et al.*, 2006) endosymbiont of the psyllid *Pachypsylla venusta*, and *Buchnera aphidicola* BCc (Perez-Brocal *et al.*, 2006), primary endosymbiont of the aphid *Cinara cedri*, have been recently reported and raised the question on where the limits

between an endosymbiotic way of life and the nature of organellar genomes should be drawn.

Most endosymbiotic genomes sequenced up to date belong to the Gamma-Proteobacteria clade. Their phylogenetic relationships have been established in chapter 4. They include four *Buchnera aphidicola* strains, primary endosymbionts of aphids to whom they supply essential amino acids missing from their phloem diets (Gil *et al.*, 2004a). Among these, *B. aphidicola* BCc has undergone the largest genome reduction with only 416 Kb (Perez-Brocal *et al.*, 2006). The two species of *Blochmannia*, *B. floridanus* and *B. pennsylvanicus*, primary endosymbionts of carpenter ants (Gil *et al.*, 2003; Degnan *et al.*, 2005); *Wigglesworthia brevipalpis* (Akman *et al.*, 2002), primary endosymbiont of the tsetse flies (*Glossinia* spp.); and *Baummania cicadellinicolla* (Wu *et al.*, 2006), primary endosymbiont of the sharpshooter (*Homalodisca coagulata*), have also established nutrition-based associations with their respective hosts, whereas *Sodalis glossidimus* (Toh *et al.*, 2006), secondary endosymbiont of tsetse flies and whose exact role in this association is not yet clear, has been proposed to be at a transition stage between a free-living and a mutualistic organism (Toh *et al.*, 2006). Finally, the smallest bacterial genome known corresponds to *Carsonella ruddii* (Nakabachi *et al.*, 2006), whose A+T content (84%), genome size (0.16 Mb) and coding capacity (182 ORFs) make it an extreme outlier, even among the other sequenced endosymbiotic genomes.

Bacterial endosymbionts share some common features due to their particular intracellular life-style. This ‘resident-genomes’ syndrome (Andersson and Kurland, 1998; Wernegreen,

2002) is characterized by high A+T content, small genome size, and accelerated rate of molecular evolution, along with low population sizes resulting from bottlenecks due to their host-to-host vertical transmission. The action of Muller's ratchet and the fixation of slightly deleterious mutations have been proposed to be responsible for the high number of both synonymous and nonsynonymous substitutions accumulated in their genomes and, therefore, for their accelerated rate of evolution in comparison with their free-living counterparts (Moran, 1996; Lynch, 1996; Lynch, 1997; Brynne *et al.*, 1998; Clark *et al.*, 1999). However, other authors have sustained that these features are the result of an enhanced mutation rate in endosymbiotic genomes (Itoh *et al.*, 2002).

It is difficult to anticipate which will be the evolutionary fate of these reduced genomes. Although stasis has been proposed for the genome size of *Buchnera* (Tamas *et al.*, 2002) there are still remarkable differences between strains of this species as revealed by the *Buchnera aphidicola* BCc genome (Perez-Brocal *et al.*, 2006). For this particular genome, a free-diffusion cell model has been proposed to explain how the cell is still able to obtain some metabolites despite lacking the necessary transporters. The authors also propose a replacement scenario in which *Buchnera* is being replaced by the secondary endosymbiont, *Serratia symbiotica*, which coexists with *B. aphidicola* BCc in the cells of the aphid (Gomez-Valero *et al.*, 2004b). The case of *Carsonella* is different; the authors (Nakabachi *et al.*, 2006) propose that it has experienced an organellar-like evolution, with functions transferred to the nucleus of the eukaryotic cells, although no

experimental evidence has been provided. Our study is aimed at revealing eventual differences in the evolutionary patterns of these extremely reduced genomes that allow us to ascertain what their evolutionary fate will be.

In this study we have focused on the evolutionary dynamics of *Carsonella* and *B. aphidicola* BCc genomes in comparison with those of the remaining Gamma-Proteobacteria endosymbionts and related non-endosymbiotic taxa. We have analyzed the action of different evolutionary forces acting on the, up to now, most advanced stages of reduction of endosymbiont genomes. We have obtained evidence for the action of positive selection in a significant portion of genes in both genomes, despite the pervading enrichment in A+T observed in them. Surprisingly, the action of positive selection is more pronounced in the most reduced genomes. Independently of their ultimate fate, both genomes seem to be actively opposing the degrading effects characterizing the “residents’ syndrome” in what seems to be their last fight for survival.

6.2 MATERIALS AND METHODS

6.2.1 Putative orthologs analyzed and gene trees inference

We selected all the available Gamma-Proteobacteria endosymbiont genomes and also those from 19 additional species of this bacterial division (Table 4) in order to search for putative orthologs of each protein coding gene of *Carsonella*. Genomes were downloaded from the NCBI repository (Benson *et al.*, 2002). For each protein sequence the reciprocal best hit was obtained (E-

value = 10^{-3}) and the annotation revised. However, the *Carsonella* genome has a large number of uncharacterized ORFs due to their very reduced similarity with other proteins, and its low GC content might also lead to some misidentifications. For these reasons, a few possible putative orthologs might have been missed in our search.

Once the putative orthologs were identified, we obtained alignments for the corresponding amino acid and nucleotide sequences with ClustalW (Thompson *et al.*, 1994). The resulting alignments were trimmed with Gblocks (Castresana, 2000) in order to eliminate positions of uncertain homology, likely representing phylogenetic noise mainly introduced by the highly divergent *Carsonella* sequences. The maximum likelihood approach implemented in PHYML v. 2.4.4 (Guindon and Gascuel, 2003) was applied for the gene tree phylogenetic inference. A general time reversible model and the JTT model of amino acid substitution was applied for the nucleotide and amino acid alignments respectively. Substitution rate heterogeneity was accounted by estimating a proportion of invariant sites and accommodating a gamma distribution defined by the estimated alpha parameter and eight rate categories.

6.2.2 Evolutionary rates of endosymbiont genes

For each gene present in all the endosymbiont genomes with an ortholog in the *E. coli* genome we carried out a relative rate test using *Pseudomonas* as outgroup based in the phylogenetic analyses carried out in the section 4 of this thesis. The test, implemented in the program RRTree (Robinson-Rechavi and Huchon, 2000), was applied to the amino acid sequences after

Gblocks treatment. A simple Jukes-Cantor model was applied for multiple-hit correction.

For each endosymbiotic sequence present in the alignments we measured the pairwise synonymous and nonsynonymous substitutions rates taking the genome sequence of *E.coli* K12 as reference for the comparison using the program SNAP (Korber, 2000). Due to the saturation of substitutions ($pS > 0.75$) in most endosymbiont sequences, the observed rate of synonymous and nonsynonymous substitutions could not be corrected for multiple hits. Therefore we used pS and pN values instead of the usual dS and dN parameters.

6.2.3 Whole genome A+T saturation measures

One common feature of endosymbiont genomes is their high content in A+T. We have designed two complementary measures to quantify this trend which can be used on any protein coding gene. We counted the number of A/T sites and G/C sites in each gene, classified them by their position in the corresponding codon and, lastly, classified each position as synonymous or nonsynonymous. We measured genome saturation by the number of positions that potentially could be G/C, because they are synonymous, but instead were occupied by A/T. The saturation measure for each genome (SMg) was calculated by averaging the individual gene values of saturation:

$$SMg = \frac{1}{n} \sum_{i=1}^n \left(1 - \frac{GC_{syn}}{Total_{syn}} \right)$$

where n is the number of genes in the genome, GC_{syn} is the number of sites with G/C that could be occupied synonymously by A/T, and $Total_{syn}$ is the total number of synonymous positions for each gene.

We also measured the resistance to A/T change of a genome as the average proportion of nonsynonymous G/C sites in its genes:

$$RMg = \frac{1}{n} \sum_{i=1}^n \frac{GC_{non}}{Total_{non}}$$

where n is the number of genes in the genome, GC_{non} is the number of sites with G/C that could be replaced nonsynonymously by A/T, and $Total_{non}$ is the total number of nonsynonymous positions for each gene.

Note that these measures do not rely on comparative analyses such as pS and pN calculations. In consequence, each genome can be characterized for saturation or resistance to A/T change regardless the number of its genes being shared with other genomes. This allowed us to include for comparative purposes the genome sequences of all available *Hexapoda* mitochondria, among them those of three insect species in which some of the endosymbionts studied reside: the sharpshooter *Homalodisca coagulata* (*Baumannia cicadellinicolla*), the psyllid *Pachypsylla venusta* (*Carsonella rudi*) and the aphid *Schizaphis graminum* (*Buchnera aphidicola* Sg).

6.2.4 Positive selection (PS), relaxed constraints (RLC) or purifying selection (PUR)?

To examine the pattern of selection acting in endosymbiont lineages we performed the branch-site positive selection test described in Zhang *et al.* (2005) and implemented in the PAML package (Yang *et al.*, 2000). We implemented the modified branch site model A in which a foreground lineage (endosymbiont) and several background lineages (remaining sequences) were defined. This model only allows for positive selection to act in the foreground lineage. It considers four site classes which appear in different proportions. Two of these proportions (p_0 and p_1) are common to all lineages and correspond to $w = dN/dS$ between 0 and 1 and those evolving neutrally with $w = 1$, which is known as model A. The other two classes of sites, denoted 2a and 2b, are different for the background and foreground lineages. These classes correspond to codons conserved or evolving neutrally in the background lineages but allowed to be under positive selection ($w > 1$) in the foreground lineage.

This model A is tested against two null hypotheses in order to, on the one hand, detect a significant increase in the number of nonsynonymous substitutions and, on the other hand, to test whether this significant fraction is the result of positive selection or relaxed constraints in the evolution of the sequence. The first null hypothesis model is based on the site-based model M1a in the PAML package and it is characterized by the presence of two site-classes, those evolving with $0 < w < 1$ and those evolving

with $w = 1$, therefore without lineage specific parameters. The second null model, denoted model A1, keeps the same proportions of sites than the modified branch-site model A but fixing in the foreground lineage the proportions of positively evolving codons (2a and 2b) to $0 < w < 1$ and $w = 1$, respectively. Likelihood ratio tests were used to compare each null hypothesis with the alternative model A; the distribution of the test was approximated using a χ^2 with 2 and 1 degrees of freedom, respectively. While the first null hypothesis is unable to distinguish between PS and RLC, although it does reveal the action of one or the other, the second null hypothesis allowed us to test directly the presence of PS in the gene of interest. For those cases in which the positive selection tests were significant, the codons under selection were identified in the corresponding sequences using a Bayes empirical Bayes method (Yang *et al.*, 2005).

Since we assume that part, maybe the most important one, of the sites detected above are the result of the A+T bias in the sequence, we only considered a codon to have evolved under positive selection if it met these 3 conditions: 1) it can be detected by the BEB analysis with an *a posteriori* probability of 0.995; 2) P-values were adjusted for multiple testing using the false discovery rate method and 3) that some of the sites of the codon change to G/C or maintain them with respect the corresponding codon in *Escherichia coli* K12., thus minimizing the chances that its detection in the BEB analysis is due to A+T bias. We compared the nucleotide in each codon position of the endosymbiont with the corresponding homolog in *Escherichia coli*. We counted the number of changes towards G/C and towards A/T as a measure of the

influence of the A+T bias in the detection of positively selected sites.

6.2.5 Testing for artefacts in the PS tests

We performed simulations of codon sequences in the absence of selection to detect incidence of false positives under A+T bias conditions. The program EVOLVER from the PAML package was used to generate sequences evolving under the branch-site model described in Zhang *et al.* (2005). The simulations required the specification of the four site categories (p_0 , p_1 , p_{2a} , p_{2b}) explained in the above section and the corresponding omegas for these categories in the foreground and background branches. We took for these values the real ones estimated in the analysis of the *rpoC* gene including its topology and branch lengths. This gene was chosen because it is large enough to overcome problems of limited number of characters and it presents a large number of positive selected sites detected (see below). However, the program EVOLVER is constrained in not allowing variation in the G+C content of the sequence. To overcome this limitation, we generated three sets of sequences under two different codon matrices, one derived from the *Escherichia coli* K12 *rpoC* sequence composition and the other derived from the *Carsonella* sequence. The (A+T)-unbiased *Carsonella* sequence derived from the first simulation was substituted by the (A+T)-biased sequence obtained in the second simulation. To ensure that we were comparing homologous sites, both simulations started from the same ancestral sequence composed by 54% of G+C and 3000 nucleotides. This biased set was composed by sequences with less A+T bias when compared

with the real *rpoC* gene. Consequently, a third set of sequences was generated taking as ancestors the *Carsonella* sequences from the second set in order to reach as much a similar (A+T)-bias as in the original sequence.

Following this procedure we generated 100 alignments for each of the two conditions tested. On the one hand, (A+T)-unbiased alignments were obtained, therefore composed by sequences with the same base composition although different branch lengths. On the other hand, 200 alignments were generated under the same conditions but with two (A+T)-biased composition. Once the new alignments had been generated, branch lengths for the *rpoC* topology were reestimated using the BASEML program. Both data sets were analyzed under the branch site model implemented in the codeml program of the PAML package as explained for the real data set. We also computed the pS, pN, dS and dN values for the comparison between each *Carsonella* sequence and the corresponding *Escherichia coli* K12 sequence.

6.3 RESULTS

6.3.1 Data set analyzed

The search for putative orthologs of the 182 genes in the *Carsonella* genome resulted in a highly asymmetric distribution, with 82 genes present in the 26 genomes and a relevant number of genes only present in from 5 to 8 genomes (Figure 15, section 4, page 107). Most of these genes had orthologs in the other endosymbiont genomes but not in other free living Gamma-Proteobacteria. In fact an analysis of the number of genes from each genome in the data set revealed that the highest numbers of

putative orthologs of *Carsonella* genes were found in the other endosymbiont genomes, possibly indicating the existence of a group of endosymbiont-specific genes. However, the influence of the high A+T content and divergence of *Carsonella* and *Buchnera aphidicola* BCc sequences, which could lead to a problem of convergence or loss of signal in non-endosymbiont genomes, cannot be ruled out. After alignment and trimming with Gblocks a number of genes were removed both at the amino acid (15 genes) and nucleotide levels (1 gene) due to the low quality of the resulting alignments, mainly due to the extreme divergence of *Carsonella*.

6.3.2 Synonymous and nonsynonymous substitutions

As previously noted (Moran, 1996), the relative rate tests of each endosymbiont gene compared with the corresponding *Escherichia coli* homolog showed an acceleration of evolutionary rates in the endosymbiont lineages, particularly for the cases of the two smallest genomes. *Buchnera aphidicola* BCc and *Carsonella ruddii* presented higher average values of evolutionary rates relative to *E. coli* than the remaining endosymbiont species analyzed (Figure 25). The estimated relative rate values, 2.11 ± 3.12 for *B. aphidicola* BCc and 2.8 ± 1.9 for *C. ruddii*, represent lower bounds of the actual acceleration, since we used the Gblocks-trimmed multiple alignments for these tests and, in consequence, rapidly evolving positions were eliminated prior to the analyses. In any case, only *Wigglesworthia glossinidia* (2.34 ± 3.1) presented rates similar to those of the two target genomes.

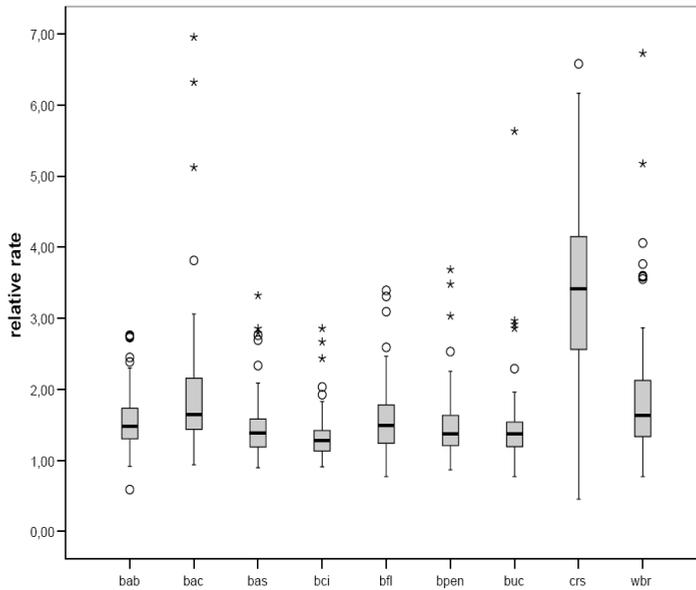


Figure 25. Box-plot of relative rate tests for each endosymbiont genome. Boxes represent the interquartile range, the black line corresponds to the median value of the relative evolutionary rates of each of the 101 common genes with respect to the corresponding homolog in *E. coli* taking *Pseudomonas* as the outgroup. Some extreme values were excluded for better viewing purposes.

To explore the source of this acceleration we obtained pairwise comparisons of the number of synonymous (pS) and nonsynonymous substitutions (pN) between each endosymbiont sequence and the corresponding *E. coli* K12 orthologs. The mean values of pS and pN (averaged for all the individual genes comparisons) per genome vs. average G+C content of the corresponding sequences are shown in Figure 26. The values obtained for pS were very high as expected (Moran, 1996). Most endosymbiont genomes were saturated in terms of synonymous substitutions due to their high rates of mutation. Only *Baumannia* (pS = 0.8509) was slightly less saturated, which could be due to being a more recent endosymbiont with the most complete

repertoire of genes involved in repair functions (Wu, Daugherty et al., 2006). In any case, all the genomes fell in the zone that prevents the use of Jukes-Cantor correction for multiple hits.

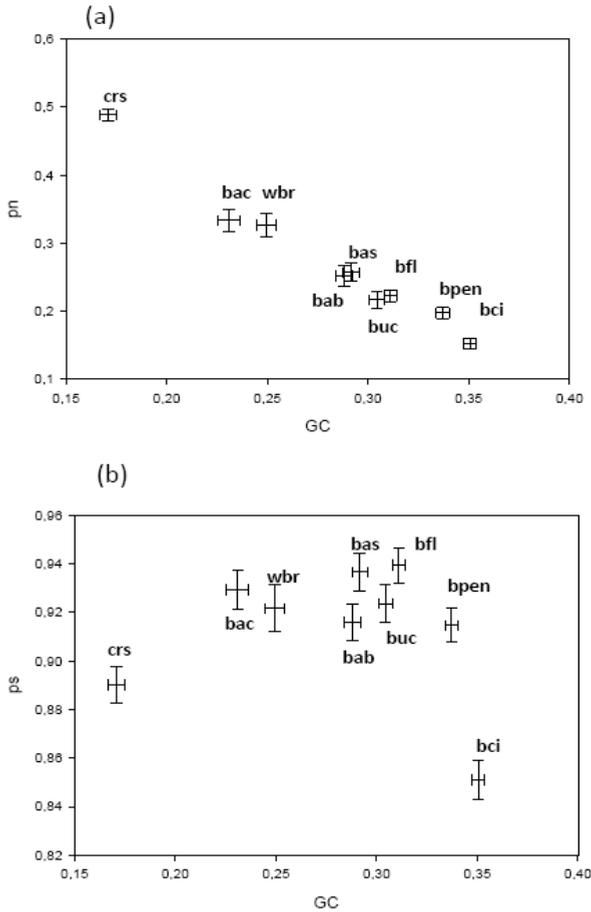


Figure 26. Number of non-synonymous (Fig. 26a) and synonymous substitutions (Fig. 26b) in pairwise comparisons between endosymbiont gene sequences and the corresponding *Escherichia coli* K12 ortholog plotted against the mean GC content of the analyzed gene sequences.

Obviously the values of pN were considerably lower than those of pS. In this case a clear correlation between the G+C content of the sequences and the mean pN parameter value was observed ($R^2 = 0.974$, $P < 0.001$). Note that although the plot suggests clear differences among some of the endosymbionts, this might be due to an incomplete sampling due to the low number of taxa available. In fact, we would have expected these values to fall in a continuum range determined by the G+C content of each genome, as observed in the following section for the genomic saturation measures when mitochondria were included in the analysis. In any case, *Carsonella* (0.4884) had the highest pN value whereas *Baumannia* (0.1521) and *B. pennsylvanicus* (0.1978) presented the lowest ones.

6.3.3 Whole genome saturation measures

The degree in which A+T bias pervades endosymbiont genomes was evaluated using two different measures (see Material and Methods). A saturation measure (SM), based on the potentially synonymous sites occupied by an A or a T, was obtained for the whole set of protein coding genes in each genome and plotted against average G+C content (Figure 27). The analysis revealed a negative correlation between SM and G+C content ($R^2 = 0.966$). As expected, the saturation in A+T increased with A+T contents, reaching a limit at around 94-95% represented by *Carsonella* and *B. aphidicola* BCc sequences. A clear discontinuity in SM values was found between the last endosymbiont (*Baumannia*, SM = 0.83) and the first non-endosymbiont genomes (*Haemophilus*, SM = 0.71), from which SM

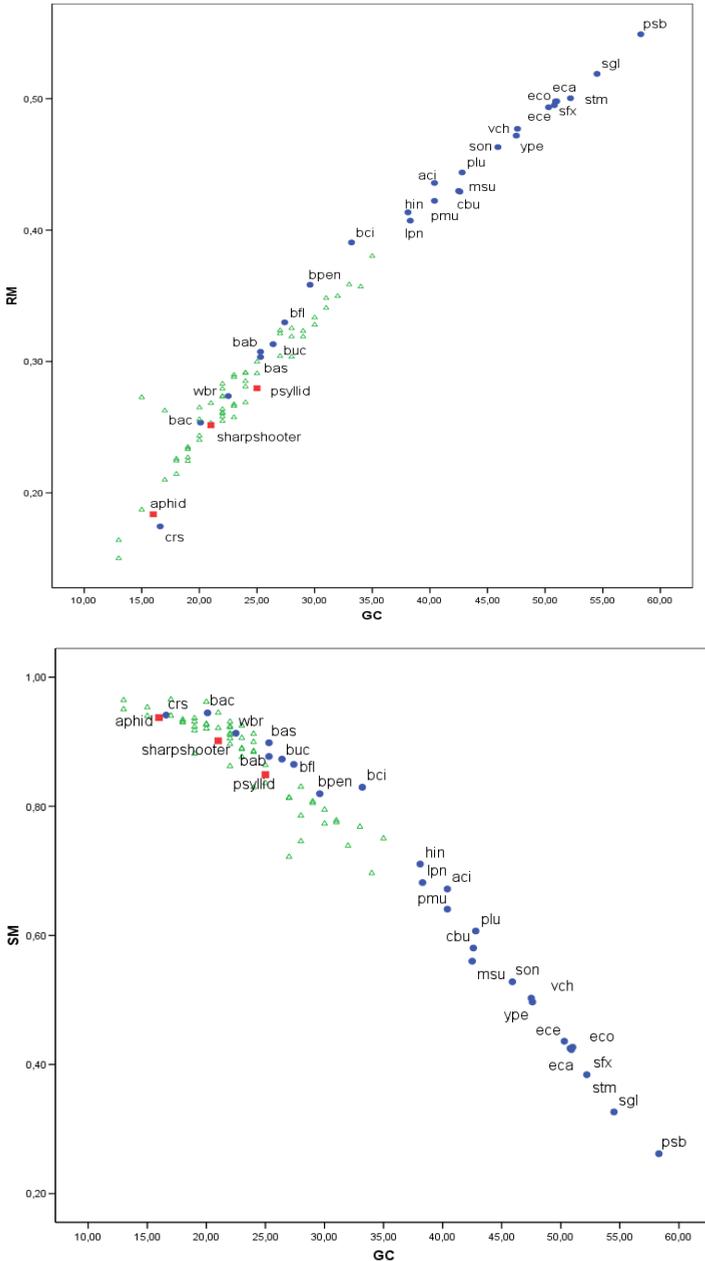


Figure 27. Saturation due to A+T content (SM) and proportion of nonsynonymous sites resistant to change towards A+T (RM) in complete endosymbiotic and non-endosymbiotic Gamma-Proteobacteria genomes (filled circles), and 61 mitochondrial genomes (triangles). Mitochondrial genomes from species also harboring a bacterial endosymbiont included in this study are indicated with a square.

decreased until the lowest values represented by *Sodalis* and *Pseudomonas* (SM = 0.33 and 0.26, respectively).

Similarly, we measured the capacity of the genomes to resist to changes towards A or T in nonsynonymous sites with what we denoted resistance measure (RM). This measure revealed a positive correlation with G+C content ($R^2 = 0.975$) with a remarkable difference between *Carsonella*, which had the lowest G+C and lowest and significantly different value of resistance (RM = 0.1745), and the next bacterial genome, *B. aphidicola* BCc (RM = 0.2535). For this measure, there was no such a clear discontinuity between endosymbiont and non-endosymbiont genomes, although their corresponding RM values did not overlap (Figure 27).

In order to evaluate the influence of the cellular environment on the evolution of nucleotide composition of intracellular (organelar and endosymbiont) genomes, we also analyzed the available mitochondrial genomes of *Hexapoda* species, including those of three species which also harbour Gamma-Proteobacteria endosymbionts (psyllids, sharpshooters and aphids). These genomes, with only 13 protein coding genes, are considerable smaller than those of bacterial endosymbionts. As shown in Figure 27, endosymbiont genomes had similar levels of saturation than mitochondria which established their symbiotic association far longer ago (around 2 Gyr) and whose genomes have been drastically reduced, with most functions having been transferred to the nucleus. The SM plot (Figure 27) revealed that *B. aphidicola* BCc and *Carsonella* have probably reached a maximum level of saturation since they had similar SM values to the most

extreme mitochondrial genomes. It is remarkable that some mitochondrial genomes had values even lower than those of the endosymbionts with the lowest values, such as *Baumannia* and *B. pennsylvanicus*. Similarly, RM values for bacterial endosymbionts spanned about the same range than for insect mitochondrial genomes, again with *Carsonella* presenting one of the lowest values among the whole data set, which again suggests that it has reached, or it is very close, to a limiting value. No correlation between the degree of saturation of the three mitochondrial genomes and the corresponding endosymbionts present in the same insect species (aphids, psyllids and sharpshooters) was observed.

6.3.4 Positive selection, relaxed evolution and purifying selection

Lineage specific tests for positive selection were carried out following a two-stage procedure. Firstly, the common alternative hypothesis, modified branch-site model A, was that there is a fraction of sites in the endosymbiont branch (foreground lineage) evolving under positive selection ($w > 1$) whereas the corresponding sites in the remaining genomes evolve neutrally or are selectively constrained. Hence, the first test (site-based null model 2a vs. alternative model A) was addressed at detecting genes with a significant fraction of sites with $w > 1$ in the endosymbiont branch. The second test (null model A1 vs. alternative model A) allowed us to differentiate between cases where those sites were the result of positive selection and cases where they resulted from the relaxation of natural selection in the foreground lineage, thus resulting in a higher rate of fixation of non-synonymous substitutions. Table 6 summarizes the results for

the two tests obtained with the set of orthologous genes to those present in the *Carsonella* genome (167 genes, after filtering with Gblocks). There was an important number of genes with evidence of positively selected codons in all the endosymbiont lineages compared with not a single case for *E. coli*, being those endosymbionts with the lowest G+C content the ones with the largest proportion of genes with codons under positive selection.

Codons potentially under positive selection in the endosymbiont sequences were compared with the homologous codon in *E. coli* K12. This allowed us to classify the observed changes for each position. We were particularly interested in those changes which were not in the direction of the A+T bias, i.e. in changes towards G/C and in those positions maintaining the G/C composition, due to the distorting effects introduced by the increasing A+T bias in the most reduced genome. Table 6 shows that there is an important fraction of positively selected codons changing to or maintaining G/C sites despite extreme A+T biased genomes, even in the case of *Carsonella* (51 (43 after FDR correction) of 78 putative selected genes). Taking into account the full protein length these sites represented 0.35-21.1% in *Carsonella* and 0-33.3% in *B. aphidicola* *Cc* genomes. Furthermore, the average G+C content in those genes in which a number of codons were detected to be under positive selection with changes opposing the A+T bias was significantly higher than the remaining genes included in the comparison (one-tailed $t = 2.679$, $P < 0.05$) (Figure 28).

Genome	genes analyzed	PUR	RLC	False PS genes	PS genes with G/C codons	Putative PS codons	PS codons with G/C	% of PS-G/C codons in the protein
<i>Escherichia coli</i> K12	121	119 (98,35)	2 (1,65)	0	0	0	0	0
<i>Baumannia cicadellinicolla</i>	118	93 (78,81)	16 (13,56)	9	0	0	0	0
<i>Blechnanmia floridanus</i>	146	94 (64,38)	28 (19,18)	19	5 (3,42)	3,8 ± 1,92	2,8 ± 1,09	0,69 ± 0,37
<i>Buchnera aphidivola</i> str. APS (<i>Acyrtosiphon pisum</i>)	166	121 (72,89)	18 (10,84)	22	5 (3,01)	1,8 ± 1,30	2,00 ± 1,22	0,65 ± 0,52
<i>Wigglesworthia glosinidia endosymbiont of <i>Closina brevipalpis</i></i>	164	69 (42,07)	34 (20,73)	30	31 (18,90)	21,03 ± 34,52	9,64 ± 13,00	3,31 ± 5,05
<i>Buchnera aphidivola</i> <i>Canara cedri</i>	166	54 (32,53)	27 (16,26)	34	51 (30,72)	17,74 ± 43,18	9,70 ± 25,18	3,24 ± 7,36
<i>Carsonella raddii</i>	167	40 (23,95)	49 (29,34)	33	43 (25,75)	94,81 ± 96,45	34,98 ± 36,93	9,86 ± 4,96

Table 6. Summary of positive selection tests. The two-step procedure used for testing positive selection allowed us to classify each gene as under purifying selection (PUR), relaxed evolution (RLC) or putatively under positive selection (PS). In brackets of the respective columns there is the percentage of the genes analyzed under each category. False putative PS genes category include those discarded by the FDR multiple test correction and those that do not present changes towards G/C in the codons identified or with no ortholog in *E. coli* K12. The mean number of codons identified for each gene before and after non G/C codons removal and the mean percentage of codons in the protein under PS are also indicated.

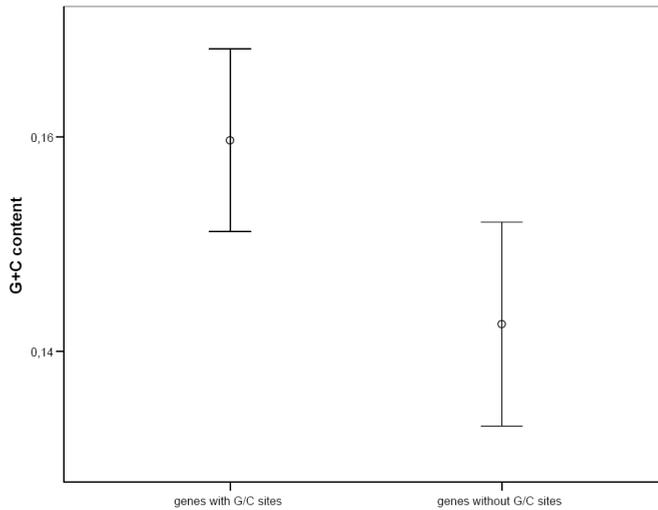


Figure 28. Mean G+C content for the genes with codons in which changes toward G/C are favoured by positive selection and mean G+C content for the remaining genes of the *Carsonella* genome.

6.3.5 Testing for artefacts in PS detection

We simulated 300 codon alignments of 26 sequences and 3000 nucleotides under the modified branch-site model A for positive selection detection but not allowing for positively selected sites ($w \geq 1$ for all the sites categories). A first set was generated with an homogeneous and (A+T)-unbiased sequence composition. A second set was generated substituting the non-biased *Carsonella* sequence of the first one by an (A+T)-biased evolved *Carsonella* sequence. The third set was generated taking as ancestor the *Carsonella* sequence from the second set and substituting it by the new one. The three simulations generated *Carsonella* sequences with different G+C content (0.54, 0.26 and 0.22 respectively).

Table 7 summarizes some of the parameters analyzed in these simulated data sets. The first column corresponds to the empirical estimates obtained from the real *rpoC* alignment. This is one of the genes with strongest evidence of positively selected codons found in our study. We selected it because it has a large number of codons (1224), the corresponding multiple alignment is very reliable with a high number of conserved positions despite the presence of *Carsonella*, and with a high number of positively selected codons detected. The original 328 codons detected to evolve under positive selection reduced to 117 when the second criterion, explained in above sections, was applied.

The same parameters were measured in the simulated data sets under biased and unbiased evolution conditions. Very different patterns were detected. The non-biased data set showed significant differences between the alternative model and the null model in 13 of the 100 alignments. However, these false positive cases did not allow almost any site to be identified as evolving under positive selection by subsequent application of the BEB procedure. The first biased data set showed no evidence of positive selection since the likelihoods of the null and the alternative models were almost the same in all cases (data not shown). Therefore no positively selected sites could be detected. The third biased data set showed significant differences in 16 of the 100 alignments. The BEB procedure identified no codons in three of the 16 positive cases. A high number of codons, between 16 and 778 codons, were detected for cases in which branch lengths were around 30-40 substitutions per codon. It is worth mentioning that some simulations in this third data set retrieved

branch lengths around 400 substitutions per codon but none of them were significant in the PS test.

To further explore whether the simulated sequences have had similar features than the real *rpoC* gene we also measured some of the parameters applied along this work. Although somewhat different, levels of synonymous and non-synonymous substitutions as well G+C content of the simulated sequences were quite similar to those observed for endosymbionts in the real alignment except for the third data set. In this case, the main differences were found in the large branch lengths and in the saturation not only of the synonymous sites but also of the non-synonymous site.

data set	rpoC (real)	no bias	bias	bias
length (codons)	1224	1000	1000	1000
p0	0.4275	0.9479 ± 0.023 (0.50-0.990)	0.6456 ± 0.160 (0.481-0.991)	0.7545 ± 0.3773 (0-0.99)
p1	0.03633	0.0156 ± 0.002 (0.009-0.022)	0.0113 ± 0.002 (0.007-0.018)	0.0134 ± 0.007 (0-0.024)
p2a	0.49417	0.0359 ± 0.021 (0-0.082)	0.3367 ± 0.158 (0-0.501)	0.2278 ± 0.377 (0-0.985)
p2b	0.042	0.0006 ± 0.000 (0-0.002)	0.0064 ± 0.003 (0-0.013)	0.0042 ± 0.007 (0-0.022)
% PS test	-	13	0	16
% PS codons	26.61 (338 codons)	0.0001 ± 0.0004 (0-2)	NA	24 ± 29.6 (0-778)
% PS codons with G/C sites	12.01 (123 codons)	-	-	-
G+C (crs)	0.1912	0.5417 ± 0.077 (0.53-0.56)	0.2615 ± 0.008 (0.24-0.27)	0.224 ± 0.006 (0.21-0.24)
branch length (per codon)	5.4405	2.96497 ± 0.167 (2.55-3.31)	7.56835 ± 0.44108 (6.56-8.66)	234,2541 ± 186.072 (29.90-508.83)
ps (eco-crs)	0.91	0.7984 ± 0.020 (0.75-0.86)	0.8633 ± 0.020 (0.81-0.98)	0.7988 ± 0.017 (0.75-0.84)
pni (eco-crs)	0.48	0.3289 ± 0.011 (0.30-0.35)	0.467 ± 0.011 (0.44-0.5)	0.7357 ± 0.008 (0.71-0.75)
ds (eco-crs)	NA	NA	NA	NA
dni (eco-crs)	0.76	0.43 ± 0.019 (0.39-0.47)	0.7323 ± 0.029 (0.67-0.82)	3.067 ± 0.511 (2.23-4.71)

Table 7. Summary of the simulations carried out in order to explore the influence of A+T bias and long branch lengths in the detection of positive selection. The first column shows some of the features of the rpoC sequence of the *Carsonella* genome. The proportions of sites under each omega category (see text for details) and the branch length were obtained directly from the rpoC analysis. The omega values for each category were not larger than one. Three data sets were generated, the first one with no compositional bias in any sequence and the other two with different A+T content for the *Carsonella* sequence. Means and standard deviation for each measure are given with the ranges of the values in brackets.

6.4 DISCUSSION

In this work we have analyzed with an in great detail the evolutionary forces acting on the last stages of endosymbiont genome sequence evolution. The extremely reduced genomes of *B. aphidicola* BCc and *C. ruddii* have allowed us to study the evolutionary dynamics of endosymbionts in the presumably last stages of their relationships with the respective hosts and to compare them with more recent and/or less degenerate endosymbiotic genomes. The introduction of two new measurements of saturation in nucleotide composition allowed us to further extend these comparisons to the most reduced genomes of bacterial origin, the mitochondria.

All Gamma-Proteobacteria endosymbiont genomes analyzed present a significant acceleration of their evolutionary rates and increased A+T content, in accordance with the previously described “resident genome” syndrome (Andersson and Kurland, 1998; Wernegreen, 2002). This syndrome is clearly exposed in our whole genome saturation measures analysis. The occupation of synonymous position by A/T provides a quantitative measure on the level of degeneration that each genome has reached and its capacity to buffer future A/T changes. The comparison with mitochondrial genomes, of only 13 genes, reveals that endosymbiont genomes have attained equivalent saturation levels and, as in the case of *Carsonella*, they can have more extreme values than most organelle genomes. Therefore, sequence degeneration seems to proceed much faster than genome disintegration given that for example *Buchnera* established

its relationship with the aphids around 200 Myr ago, whereas organellar genomes established it 2 Gyr ago.

Saturation in gene sequence can only be reached by an enhanced nucleotide substitution rate in endosymbionts, as revealed by relative-rate tests, jointly with an increased bias towards A/T change. This explains their high relative evolutionary rate with respect to *Escherichia coli*. However, since there are no differences in the number of synonymous substitutions between endosymbionts, since all of them are saturated even in the case of *Baumannia*, differences in acceleration rates among them must reside in differences in nonsynonymous rates. The original observation (Moran, 1996) based on a very small data set is corroborated in our analyses, which have revealed a strong correlation between three factors: A+T bias, genome size and number of nonsynonymous substitutions.

Different evolutionary forces could drive the fixation at such a high rate of nonsynonymous substitutions among which positive selection, relaxed evolution, genetic drift in the form of Muller's ratchet, and an increased rate of mutation have been discussed previously. Most arguments for and against each alternative are based on population genetics theory (Lynch, 1996; Moran, 1996; Itoh *et al.*, 2002; Lynch, 1997; Brynne *et al.*, 1998; Clark *et al.*, 1999). However, most of these proposals are mainly based on indirect evidence given the scarcity of reliable data on mutation rates and effective sizes of bacterial populations, either endosymbionts, pathogens or free-living.

Endosymbionts are vertically inherited and undergo regular bottlenecks during the mother-to-offspring transmission (Buchner, 1965). Furthermore, their mutation rates are also elevated because many of them completely lack of or have incomplete repair systems, thus favouring the fixation of both neutral and non-neutral substitutions. However, it could be argued that there is some room left for the action of positive selection; they are degraded genomes which are supposed to have an essential role in the survival of their hosts. Otherwise, they would have disappeared completely, either by displacement by a secondary endosymbiont (Perez-Brocal *et al.*, 2006) or by drift should they had become completely useless for their hosts once the relevant genes had been transferred to its nucleus as proposed for *Carsonella* (Nakabachi, Yamashita *et al.*, 2006). Therefore selection must act to maintain certain functions, mainly as purifying selection, preventing the loss of essential functional capabilities, and/or as positive selection, enabling adaptation to new conditions in the host-symbiont environment and evading its own extinction.

However, there is usually confusion when the forces behind this enhanced nonsynonymous rate are considered. Traditional explanations (Itoh *et al.*, 2002; Moran, 1996; Brynne *et al.*, 1998) assume that nonsynonymous substitutions are mainly due to Muller's ratchet effect, relaxed selection or high mutation rates. On the contrary, we propose here that different and opposing forces are enhancing the nonsynonymous rate of substitution. We will discuss each possibility in turn and try to

show that our data support the simultaneous action of different forces.

The accumulation of nonsynonymous substitutions due to continuous bottlenecks and the absence of recombination is known as Muller's ratchet effect (Muller, 1964). Its action results in the increasing accumulation of deleterious mutations which cannot be purged from the genome and may lead to the extinction of the lineage. The ratchet depends critically on the intensity of genetic drift, but we lack empirical data about the different demographic components determining the effective size of bacterial endosymbiont populations. These are usually assumed to be very low (Moran, 1996), due to the severe bottleneck during mother-to-offspring transmission (Mira and Moran, 2002), although some measurements point towards a less dramatic reduction of the effective population size (Moya and Latorre, 2007) than previously postulated. As argued in Itoh et al. (2002), the continuous degeneration of endosymbionts genomes is incompatible with the maintenance of their function 200 Myr later.

Itoh et al. (2002) favoured the high mutation rate hypothesis mainly due to the loss of repair machinery from most endosymbiont genomes. However, at the time their work was published an important number of endosymbiont sequences used in our study were not available yet. The analysis of new endosymbiont genomes has revealed that 1) the acceleration of *Baumannia*, which still maintains most of its repair machinery, with respect *Escherichia coli* is not very different to that of some *Buchnera* and *Blochmannia* species, which lack most genes for repair, and 2) if the enhanced mutation rate were the solely responsible of the

nonsynonymous rate, then no correlation between A+T bias and similar pN/pS rates for all of endosymbionts would be expected, which is not supported by our results. Relaxation of selection in the form of accumulation of slightly deleterious mutations due to small population sizes is also unlikely to fully explain differences in nonsynonymous rates, for similar reasons as indicated above for Muller's ratchet.

On the contrary, positive selection is a force whose role in the evolution of endosymbiont genomes has been scarcely explored with only some exceptions reported (Fry and Wernegreen, 2005; Fares *et al.*, 2002). However, the detection of positive selection has recently experienced a significant change (Zhang *et al.*, 2005; Arbiza *et al.*, 2006; Mes *et al.*, 2006) since the recognition that branch- or sites- only based methods are conservative in the detection of positive selection because the number of codons with $w > 1$ needed for its detection at the whole gene level must be very large. It is known that most of the sites in a protein are under purifying selection and that changes are permitted or favoured in only a few positions (Zhang *et al.*, 2005). As a consequence, most previous studies have failed to detect positive selection in endosymbiont genomes (Moran, 1996; Herbeck *et al.*, 2003; Fry and Wernegreen, 2005). Our branch-site test approximation has allowed us to detect cases with an enhanced frequency of sites in the $w > 1$ category along specific lineages and also to identify the involved codons. Accordingly, we have considered genes to be under positive selection when two conditions were met: the two-step procedure allowed us to detect a significant fraction of codons falling in the $w > 1$ category and

the BEB procedure was able to identify at least some of these with a posterior probability larger than 0.995. It is important to note that the number of codons under positive selection is only a tiny fraction of the whole gene. Although the vast majority of codons are evolving under purifying or relaxed selection there are traces of positive selection in some other codons.

However, there is a positive correlation between the numbers of genes with sites detected to evolve under positive selection and the A+T bias of the corresponding genome. Hence, it could be argued that the positive selection test might have been misled by the extreme compositions of endosymbiont genomes and the saturation of the synonymous substitutions, particularly of *Carsonella ruddii*. In this context, two hypothesis can be formulated: 1) positive selection in the detected codons results only from the A+T bias and the nonsynonymous substitutions result from the incapacity of natural selection to purge them, and 2) at least a fraction of these codons are the result of the true action of positive selection. We have taken two ways to discriminate between both possibilities: first, we have established an additional condition to the usual identification by BEB analyses with a higher than 0.995 posterior probability of a site belonging to the $w > 1$ class, i.e.; that the potential positively selected codon present changes against to the A + T compositional bias. Therefore, we have looked for the presence of sites among positively selected codons that either change towards G/C or maintain them (Table 6). Secondly, we have simulated sequence evolution under as similar conditions to those experienced by the actual sequences in order to evaluate the robustness of our testing

procedure for detecting sites evolving under positive selection to these factors.

Our analysis has revealed the existence of a significant fraction of these sites, even more markedly for *Carsonella* than for any other genome (Table 6). This kind of changes can only be explained by the action of positive selection. Therefore, selection is introducing changes that oppose the degenerative effects caused by the A/T bias pressure. Consequently, we propose that opposing evolutionary forces are simultaneously acting on these extremely reduced genomes, with positive selection counterbalancing the degenerative consequences of enhanced mutation rates, A/T bias and ratchet effects that surely keep pushing endosymbiont genomes down the whirl of extinction. These compensatory mutations might slow down the process of sequence degeneration, eventually to the point of attaining a transitory or permanent steady state. In fact, experimental approaches simulating Muller's ratchet for populations of repair deficient and wild type strains of yeast (Zeyl *et al.*, 2001) have shown that the probability of mutational meltdown by only this process in this particular experimental set up is very low and requires extremely reduced populations. The fitness of the mutator genotypes in these experiments did not change substantially and only a tiny fractions of the assays resulted in population extinctions, which is an evidence for the presence of compensatory mutations that allow fitness recovery. Furthermore, recent analyses have also identified the action of positive selection in mitochondrial genomes (Bazin *et al.*, 2006; Bazin *et al.*, 2006), most notably in insects and other invertebrates, despite the

traditional view of mitochondrial DNA as an essentially neutral marker.

6.5 CONCLUSIONS

In endosymbionts populations, the action of genetic drift is surely present and the accumulation of deleterious mutations is significant as derived from the number of genes evolving under relaxed and purifying selection and also from the proportion of changes towards A/T in a fraction of apparently positively selected codons. However, we have presented evidence for the existence of a significant fraction of advantageous mutations, marked by their change towards G/C in positively selected codons that somehow counter the degeneration process in these sequences and that might represent the last resource in the fight for survival of the most reduced bacterial genomes.

Three main evolutionary fates can be postulated for these genomes. If the genome reduction process is maintained despite the action of selection then the genome could eventually disappear. The most likely reason for this scenario comes from the presence of a secondary symbiont which takes over the role(s) that the primary symbiont was carrying out (Perez-Brocal *et al.*, 2006). Alternatively, it is possible that the endosymbiont retains only some genes whereas the most important ones for the host-symbiont association are transferred to the host genome in order to attain a better control of their transcription (Nakabachi *et al.*, 2006). This would end in the transformation of the endosymbiont genome into a new organellar structure. Finally, if no new selective pressure is present, the endosymbiont genome could reach a steady state where it continues providing the host with

benefits and therefore selection would favor, mainly (but not exclusively) by purifying selection, the maintenance of the integrity of the endosymbiont genome. Since the lapses of time necessary to test experimentally these alternatives are completely out of our reach, it is only through a more intensive comparative analysis and sampling of endosymbiont and organellar genomes in different genome size reduction stages that we will eventually be able to discern among them.

7. GENERAL DISCUSSION

Evolutionary analysis has undergone a revolution as a consequence of the exponential increase of sequenced genomes. This is particularly obvious in the case of prokaryotic genomes which are easier to obtain because of their reduced genome sizes. However, it would be a mistake to consider the evolutionary genomics of microbes simpler and less important than that of eukaryotic genomes. In fact according to the main measures used to describe a microbial genome (Bentley and Parkhill, 2004) the diversity among them even at very short evolutionary distances is surprising large and reveals that the incidence of evolutionary forces is very different in prokaryotic than in eukaryotic genomes.

Some of these forces are common to eukaryotes. For example selection, mutation, migration or genetic drift are common population processes with importance on bacterial genetic diversity (Perez-Losada *et al.*, 2006). In fact, despite it could be claimed that, as a general trend, bacteria have large population sizes, and therefore some of these processes are less important, there are many exceptions derived from the particular lifestyle of the different species due to their ability to occupy new ecological niches. At a genome level, other forces like genome rearrangements, that also have acted in eukaryotic genomes, mediated by insertion sequences elements (IS) have been proven to be a major factors in bacterial genome dynamics and usually reflect the lifestyle of the genomes analyzed (Belda *et al.*, 2005). For example, endosymbiotic bacteria use to have a stable genome once the relationship with the host is stable but are full of IS during the first stages of this relationship (Wernegreen, 2005). In pathogenic bacterial there are also examples of the importance of

IS mediated rearrangement in the arising of new pathogens (Nierman *et al.*, 2004).

However, the main force studied so far in bacterial genomes is the influence of genetic exchange among them. Horizontal gene transfers among microbial genomes have grabbed the attention of scientists both because it is a distinctive evolutionary feature with respect eukaryotic genomes and because it represents a challenge to the traditional view of evolution by vertical descent. Although horizontal gene transfer is not absent from eukaryotic genomes (see (Andersson, 2005) for a review) its importance as generator of evolutionary novelty seems to be more restricted to particular groups and to exceptions in some other specific lineages. In fact the main contribution of bacterial genomes comes from transfers from the bacterial origin organelles to the nuclei of the eukaryote (Martin and Herrmann, 1998). A very different picture is present in bacterial genomes where the exchange of genetic material seems to have shaped most of them and is an ongoing force of evolutionary novelty (Ochman *et al.*, 2000). Coupled with gene losses, HGT has contributed to the occupation of almost all ecological niches and is directly related with the harmful action of most bacterial pathogens.

Since Darwin's proposal of vertical inheritance with modification, different pictures of the evolutionary relationships among living organisms has been proposed. The advent of molecular techniques represented a Copernican revolution in the systematics field giving to the proposal of the three Domains of life: Eubacteria, Eukarya and Archaea (Woese *et al.*, 1990). Despite being initially well accepted, this proposal has suffered criticisms

related with the evolutionary and phylogenetic relationships among the three domains (Brinkmann and Philippe, 1999). Moreover, the sequencing of bacterial genomes also revealed that HGT was a widespread phenomenon in Eubacteria and Archaea, not only restricted to some particular traits like the transfer of resistance to antibiotics genes (Gogarten and Townsend, 2005). The mosaic nature of the sequenced genomes arose the question about the existence of a single tree of life in bacteria, still a hotly debated issue nowadays (Doolittle, 1999b). Our capacity to derive correctly, if it exists, a Tree of Life and to identify the existence of other phylogenetic signals is dependent on the current phylogenomic approaches and the cautionary steps that we have to adopt to distinguish between phylogenetic noise and signal.

This thesis has tried to address some of the aspects concerning bacterial evolution mentioned above. It started with the application of phylogenomics approaches to a particular issue, the phylogenetic position of Gamma-Proteoabacteria endosymbionts. Consequently, this discussion will start with some considerations about phylogenetic methodologies derived from the different analyses throughout this thesis which are important in order to accomplish the objective of differentiate noise and signal. Then, we will try to generalized what Xanthomonadales genomes and the analysis of the last stages of endosymbiont genomes tell us about two of the main bacterial evolution mechanisms, gene gain and gene loss respectively.

7.1 Phylogenomics and the evolutionary signals in microbial genomes

The discussions of chapter 3 and, partially, chapter 4 deeply explore the advantages and disadvantages of the phylogenetic approaches applied to the endosymbionts data sets. In this general discussion we would like to highlight some aspects derived from the different phylogenetic analyses in the four chapters of this thesis, although in some of them they were not the main goal of the chapter but only the basis on which some of the applied methods relied.

7.1.1 Influence of orthology assessment on phylogenomic analyses.

As mentioned in the Introduction, orthology assessment is not only a crucial step but also a problem because of the many different approaches to perform it. For example, Lerat *et al.* (2003) and chapter 3 of this thesis used a similar set of genomes of Gamma-Proteobacteria, although in the latter we used Beta- and Alpha-Proteobacteria as outgroups. In analyzing the presence of incongruence in a similar set of around 200 genes, in the first work we found that around only 1% of them rejected the reference tree whereas in the second case we found around 30% of rejections. Apart from some methodological differences in the reconstruction of gene trees, the main difference between both studies was the number of putative orthologs in each gene family. Lerat *et al.* (2003) used a very stringent criterion, filtering most of the potential incongruent genes as noise. Our work used a more relaxed selection based on several evidences, not only the BLAST report. Both studies are congruent/coincident in the sense that

both reached similar conclusions about Gamma-Proteobacteria topology but the gene set used in each case allowed answering different questions. The approach of Lerat *et al.* (2003) neglected the presence of possible horizontal gene transfer, filtering it as noise; this approach allowed them to address the question of the relationships among taxa, i.e. highlighting the vertical signal of these genomes. We were less restrictive so we allowed the presence of horizontal gene transfer and therefore were able not only to reveal the vertical signal but also to reveal potential horizontal gene transfer events in some of the taxa. Obviously the noise signal also increased but we worked under the supposition that most noise would be overcome by the genome-scale approach. This example is not a criticism to Lerat *et al.* (2003) analyses, but a reminder that the methodology used to select putative orthologs must keep in mind which are the objectives of the ensuing analyses.

7.1.2 Model-based methods of phylogenetic reconstruction.

Once a data set for the phylogenomic analyses has been assembled, the next crucial step is the selection of an appropriate phylogenetic reconstruction method. In this sense it is important to distinguish among those methods based on a model for the evolution of the characters studied and those with no model assumed. A model in statistics tries to grab the most important information of a process in order to facilitate the estimation of the parameters governing it. In phylogenetics they have been used to analyze sequence information in order to estimate the expected number of changes in a molecule during its evolution (Kelchner

and Thomas, 2007). Each parameter in a phylogenetic model represents a different property of the process (for example the shape of the gamma distribution used to model rate heterogeneity among sites or the proportion of invariants sites). Supermatrix, and indirectly supertrees, approaches have taken advantage of the development of explicit models for the evolution of sequences on single gene alignments. We have used them in chapters 3 and 5. On the contrary, gene content and gene order phylogenies have been possible only with the advent of genome sequences and therefore the modelizations of both processes have been developed only recently (Snel *et al.*, 2005; Delsuc *et al.*, 2005). Despite they have been successfully used to address several important issues in eukaryote (International Chicken Genome Sequencing Consortium, 2004) and prokaryote (Belda *et al.*, 2005) genome evolution, the insufficient development of these methods and the statistical framework for the evaluation of their performance led us not to use them. The use of model-based analyses has allowed us to explore the incidence of systematic biases in sequence data, a problem inherited from traditional phylogenetics, and the study of the accuracy of traditional phylogenetic procedures when transferred to phylogenomic analyses.

7.1.3 Supertrees, supermatrix and the signal detected.

The comparative analysis of supertrees and supermatrix performance has revealed its utility at different levels. The detection of the main phylogenetic signal, understanding main as the most frequent not as the most important, seems to be easier

for supermatrix analyses. It is true that using only common genes might reduce the data sets in some cases but for medium taxonomic ranges this loss is not significant. Our analyses of the 579-gene set versus the common 200-gene set of *Blochmannia* revealed no important effect of the addition of more, patchily distributed genes. Furthermore, if the genes selected come from the essential data set used in the second part of chapter 3, it is possible to retrieve this signal with only 60 genes. However, we want to challenge the view that using a supermatrix of common genes is enough to understand genome evolution (Dutilh *et al.*, 2007). Recently, a tree of life has been proposed using only 31 common genes to extant taxa (Ciccarelli *et al.*, 2006). This approach has been criticized by calling it the “tree of 1%” (Dagan and Martin, 2006). What is the utility of such a tree to study bacterial evolution when the most reduced microbial genome has 182 genes (Nakabachi *et al.*, 2006) and it is an outlier in the microbial world?

Likewise, for *Blochmannia floridanus* we have used a tree of 100% but of little utility for non-endosymbiont bacteria harboring thousand of genes. In this sense, the supertree analysis of the *Blochmannia* phylome revealed as a way of pointing towards interesting, because of being problematic, taxa. The use of consensus/supertrees in chapters 3 and 4 led us to explore which is the source of incongruence in Xanthomonadales genomes. Unresolved nodes, in this case the one that groups Xanthomonadales with the remaining genomes, revealed inconsistent signals in the gene trees that support the supertree and, therefore, point towards cases of incongruence or horizontal

gene transfer. The analyses in chapter 5 corroborated this initial supposition. In this work we also used the supertree as part of the protocol for removal of phylogenetic noise. Again, if our aim is to study incongruence around Xanthomonadales, the incongruence due to non-related taxa must be removed. The consensus tree obtained for the 18 selected genomes showed a strong incongruent signal around *Legionella pneumophila* and *Nitrosomonas europaea* which were removed from the ensuing analyses. Therefore, the summary of gene trees is best suited to point toward problematic clades which in turn could be further explored to reveal the source of incongruence or removed to avoid noise that could affect the objectives of the study.

7.1.4 Incongruence and signal.

There are several ways to explore the incongruence present in a set of putative orthologous groups. In chapters 3, 4 and 5 we used two approaches that, although relying on the same topology comparison test, gave us very different phylogenetic information. In the *Blochmannia* phylome data set we have compared each gene tree of the phylome with our presumed species tree topology. This kind of approach allowed us to explore how much phylogenetic incongruence is present in the data set by comparing with the expected relationships. However, there are two main problems associated with this approach: 1) it says nothing about the degree of incongruence among the genes, that is, we do not know whether the genes that reject the presumed species tree also reject each other; 2) it is possible that an alternative topology is globally less rejected than the presumed species tree topology, therefore questioning whether this

presumed topology is the most likely species tree or even whether such a tree exists (Baptiste *et al.*, 2004).

Which is the best way of using topology test information? Both in the *Carsonella* and in the Xanthomonadales data sets we have applied a congruence map analysis. That is, we have tested each gene tree versus each gene alignment and, in the *Carsonella* case, we also added two plausible presumed species trees. This way of visualizing incongruence has been applied to various data sets to demonstrate lack of a central phylogenetic tendency (species tree) in bacterial genomes due to the presence of noise or horizontal gene transfer (Baptiste *et al.*, 2005; Susko *et al.*, 2006). Our approach is not exactly the same: we have used it to test specific hypothesis of placement in the *Carsonella* case or transfers in the Xanthomonadales analysis. Therefore, we were not looking for incongruence in all the taxa but only in the groups of interest. The outcomes of such congruence maps can be basically three: 1) all the genes have been vertically transmitted, therefore most genes would be congruent among each other and only a tiny fraction would reject the others because of phylogenetic noise; 2) the genes have a different origin, most likely due to horizontal gene transfer and, if the signal is good, different groups will be detected; and 3) there is no pattern, each gene tree is rejected by most of the remaining genes because of phylogenetic noise due to systematic biases, model violation, insufficient phylogenetic signal or other reasons.

Carsonella and Xanthomonadales congruence maps are clear examples of the second and third alternatives. The Xanthomonadales data set allows differentiating clearly among

several groups as shown in Figure 29. The most plausible origin of the genes, Beta-, Alpha- or Gamma-Proteobacteria, is detected in some cases whereas in others a mixture of these possibilities is present. Noise is detected but it is also clearly different from the groups outlined before because it comes from genes that reject almost all topologies or from genes unable to reject any of them. This is also the case for the majority of genes from the *Carsonella* data set in which the most supported phylogeny is only compatible with about 30% of the genes, with a large array of heterogeneous phylogenies supported by only one or few genes. As this congruence map is only constructed with 82 genes and gene trees, present in all genomes and mainly from informative categories which tend to harbor a very good signal, most of the incongruence seems to be due to the endosymbiont sequences which in a high proportion violate the assumptions of the evolutionary model. In turn, the Xanthomonadales map is very different (Figure 29.b): instead of noise it reflects a systematic incongruence towards the donors, thus generating a clear clustering of genes. Therefore the presence or absence of patterns in a congruence map indicate us the source of phylogenetic incongruence in genomic data sets.

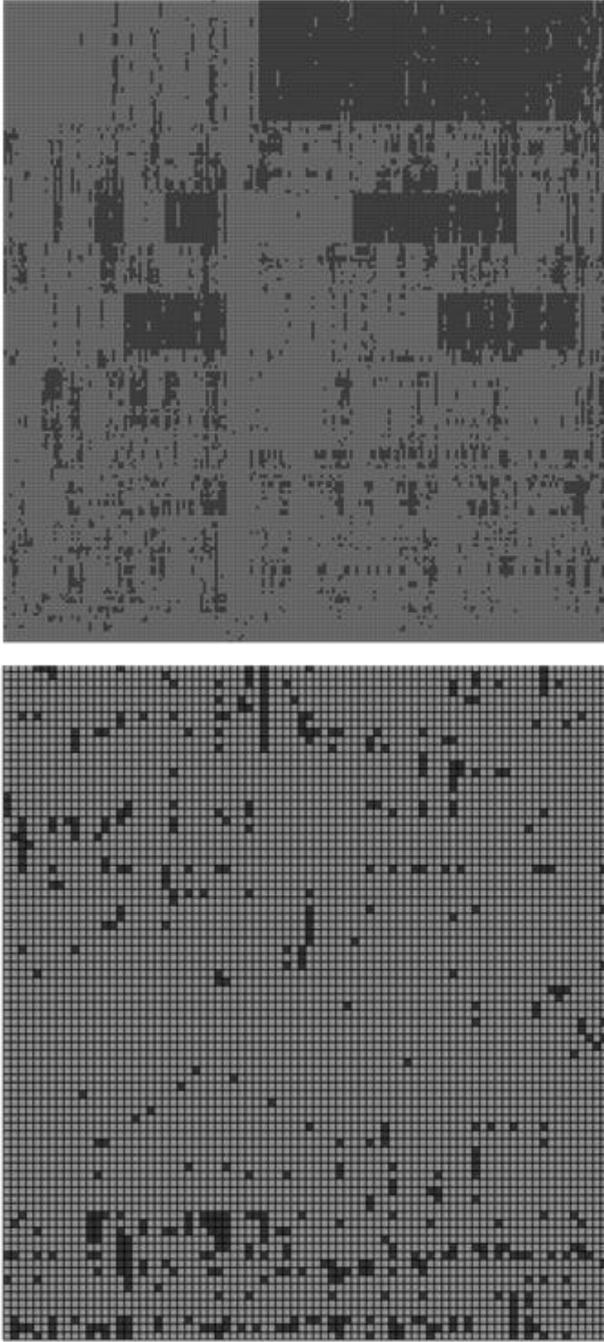


Figure 29. Noise versus signal in a congruence map analysis. The figure on the left correspond to the one obtained for the *Carsonella phylogenomic* analysis. The one on the right correspond to the obtained for the *Xanthomonadales* analysis.

7.2 Lessons on microbial evolution from Xanthomonadales genomes

Xanthomonadales are a group of plant-pathogens classified as Gamma-Proteobacteria. Most of their genomes are the result of multiple events of HGT to the ancestor of the genome from other Proteobacteria. The evolution of these genomes allowed us to delineate in chapter 5 a model for the effect of transfers along time on the content of the genomes as well on the analyses of genome phylogeny involving them. This model is expected to hold for highly promiscuous bacteria for which the main constraints to transfers are those related to genomic architecture incompatibilities and therefore from the time since divergence of the two lineages considered. This situation results in taxa like Xanthomonadales with a clear web-like behaviour in phylogenies due to the apparent absence of selection pressure against the majority of transfers and represents the most extreme case of genetic exchange among bacteria. This free-exchange scenario is represented in Figure 24. This model for Xanthomonadales assumed two principles:

- 1) genetic exchange in bacteria is widespread but not random, being more likely between more related bacteria (Gogarten *et al.*, 2002).
- 2) there are different effects of transfers on genome phylogenies depending on the time elapsed since the transfer and from the divergence between the donor and the recipient.

If both statements are true, then it is expected that the currently preferential partners of Xanthomonadales for

exchanging genetic material are other genomes of the same clade. In the past, when Proteobacteria and Xanthomonadales were nascent lineages, the preferential partners might have been found among other Proteobacteria. Our analysis of Xanthomonadales genomes corroborates this prediction. Genome phylogenies show two facts: on the one hand, the Xanthomonadales clade is monophyletic for all the gene trees; on the other hand the consensus tree is unable to show a single majoritary placement for the whole group. The second observation clearly points to ancient HGT to the ancestor of Xanthomonadales from other Proteobacteria, as our posterior analyses corroborated. The first observation is indicating either that vertical inheritance was the dominant force during the diversification of the clade or that there are HGT events not detected in genome-level phylogenetic analyses, whose signal is also vertical-like. An analysis of atypical composition of the *X. citri* genome corroborated the presence of a significant fraction of atypical genes (22%). Most of them were not detected in our phylogenetic analyses thus corroborating that recent intra-clade exchanges exist. Therefore the cohesion of the clade observed in the summaries of gene trees is powered not only by vertical descent but also by horizontal gene transfer as predicted elsewhere (Gogarten *et al.*, 2002).

This kind of preferential sharing between genomes is also pointing to the existence of mechanisms that limit HGT, making more likely those between closely related species. These mechanisms could be of different kinds and range from factors like genome compatibility to ecological opportunity. Our proposal for Xanthomonadales is that transfers to them are only limited by

genome architecture compatibility factors because they have diversified in a very rich environment (plants) and have had opportunities for interaction with many other plant-associated bacteria belonging to the most diverse taxonomic groups (Van Sluys *et al.*, 2002). Therefore, the success of transfers depends mainly on the phylogenetic distance between the donor and the Xanthomonadales ancestor. This free-exchange environment is reflected in Figure 24 in the straight line representing the amount of DNA acquired along time. As we will discuss later, other bacterial groups subject to other possible limiting mechanisms will change the shape of this curve for these genomes.

The extent to which this proposal could be applied to other microbial genomes is the main subject of this section. First, because it has to be demonstrated that distance-scaled horizontal gene transfers are possible. Second, because, unlike Xanthomonadales, many groups are limited in their capacity to accept external genetic material usually because they do not have the opportunity. As we have mentioned above there are different limiting factors to a widespread, random exchange of genetic material among bacteria. Most of these factors are known, but those related to genome architecture or compatibility have been reported only recently (see below for references).

7.2.1 Non-random genetic exchanges: internal and external factors

The first of these factors is the process of recombination, the homologous exchange of genetic material between closely related genomes, usually falling under the same species designation. As mentioned in the introduction there are two

important evolutionary features related to recombination in bacteria. There is decay in the rate of recombination with increasing phylogenetic distance due to the mismatch repair system (Majewski and Cohan, 1998). The tempo and mode of this decay are variable and different among bacterial groups. Distance-scaled recombination ensures the cohesion of nascent, divergent lineages and therefore contributes to their genetic isolation although different regions of the genomes seem to have different recombination rates and isolation in a region does not necessarily imply complete speciation. The acquisition of foreign material by non-homologous recombination seems to empower the speciation process by allowing regional isolation of the genomes, even though the exact process is still under discussion (Lawrence, 2002; Nesbo *et al.*, 2006; Cohan, 2004). Apart from homologous recombination, other mechanisms could help to the exchange of genetic material among distant bacteria (Thomas and Nielsen, 2005). However, they are partly out of the scope of this discussion since we are interested in (1) reviewing recently discovered mechanisms that prevent HGT between two lineages and (2) demonstrating that the by-product of these mechanisms is the preferential sharing of DNA between more closely related taxa.

Some recent reports of “in silico” and “wet” studies are revealing mechanisms by which bacteria allow or avoid the acquisition of foreign material. These mechanisms could be divided in intrinsic and extrinsic factors depending on whether they are derived from genomic features or are due to non-genomic properties. Genomic architecture features revealed by whole genome sequences, new molecular mechanisms besides

those that prevent recombination and the study of the functional roles of the genes from a genome should be remarked among internal factors. Among external factors the new insights about bacterial populations and their ecological niches derived from metagenomic analyses and the study of the diversity of phages are the most important.

Genome architecture. Recent progress in the analysis of the structure of bacterial genomes, aided by the continuous sequencing of new genomes, is allowing to describe and characterized them beyond the genes which compose them (Rocha, 2004). The preferential distribution of genes in the leading strand of genomes has been usually explained as a selective pressure for more efficient replication and expression (Brewer, 1988) but essentiality seems to be a more important factor, even constraining the number of possible rearrangements (Rocha and Danchin, 2003b; Rocha and Danchin, 2003a). Recent studies have also identified codon usage domains in bacterial chromosomes that are important for the control of the expression of the genes (Bailly-Bechet *et al.*, 2006). These results support the idea of domains in bacterial chromosomes of a higher level than genes, operons or even über-operons (Boccard *et al.*, 2005). The alteration of these macrodomains due to rearrangements or the insertion of foreign sequences could have a detrimental effect over the fitness of the organism and therefore a “compatibility filter” must exist in order to maintain the chromosome properties. A filter like those described above has been recently described (Hendrickson and Lawrence, 2006; Lawrence and Hendrickson,

2004). This implies that successful transfers will be more likely between closely, and therefore compatible, genomes.

Molecular mechanisms. The best known molecular mechanism controlling DNA uptake in bacteria is the methylation of genetic material. DNA not carrying the specific pattern of methylation of a species is digested (Jeltsch, 2003). Recently one of these systems has been proposed as responsible for foreign DNA control in *Staphylococcus aureus* (Waldron and Lindsay, 2006). This type I system requires the action of three genes for the methylation process (*hsdR*, *hsdM* and *hsdS*). Variations in these genes throughout sequenced *S. aureus* genomes and in multi-strain microarray analyses showed a significant correlation with the 10 known dominant lineages. Furthermore, the recognition of a sequence as alien by a lineage is based on different sequence profiles of these genes. The effect of this biased methylation is that (1) the intra-lineage exchange is more likely than the inter-lineage exchange and (2) there is preferential sharing between different lineages. Obviously, this mechanism also acts upon sequences obtained from other species and therefore the result of the presence of such a mechanism is the more likely exchange between *S. aureus* strains to the point that it seems to be the responsible for the identification of different lineages.

Other largely unresolved question about bacterial genome architecture is their characteristic GC/AT ratio (Bentley and Parkhill, 2004). Sometimes, lower GC contents are clearly associated with certain lifestyles, mainly in obligate intracellular pathogens and endosymbionts (Moran and Wernegreen, 2000). However, among the remaining genomes differences in GC

content are remarkable and genus-specific. Recently a possible role for these biases has been proposed (Lucchini *et al.*, 2006; Navarre *et al.*, 2006). The protein H-NS has the ability to bind DNA acting as a transcriptional regulator or even as a multimerization agent. In two independent experiments the genes mainly influenced by H-NS in *Salmonella* have been screened. Both approaches are based on microarray analysis of expression of *Salmonella* genes in mutant and non-mutated *hns* strains. Both experiments reached the same conclusions. H-NS affects mainly the expression of AT-rich genes in a process that has been called “xenogenic silencing” and it is expected that similar mechanisms are present in most bacteria genera. Most of the affected genes have a GC value lower than the average for the corresponding genome value. Moreover, most of them are not universally present in *Salmonella* genomes, which also points towards HGT as the most probable source. In the case of *Salmonella*, AT-rich sequences usually come from non-enterobacterial species. The silencing of these genes avoids the possible negative effect on the fitness of the genome. Which is the fate of these silenced genes? It seems that bacteria have evolved mechanisms in order to selectively activate some of these genes. The fate of the remaining genes could be their pseudogenization. Transfers from the same GC content donor maybe more likely to remain on the genome because the H-NS mechanism does not operate.

Functional association. Since the publication of the complexity hypothesis by Jain *et al.* (Jain *et al.*, 1999) it has been accepted that those genes belonging to informational categories are less prone to HGT than those of operational ones. The

hypothesis is based on two observations: (1) HGT has been continuous in the evolutionary history of bacteria, and (2) there are important differences between the degree of phylogenetic incongruence found in operational genes and metabolism genes. Jain *et al.* (1999) proposed that informational genes tend to be part of essential complexes interacting with many products whereas operational genes used to form part of simpler metabolic networks. Supporting this complexity hypothesis Nakamura *et al.* (2004) identified metabolism related genes (cell surface, DNA binding and pathogenicity-related functions) as the more transferred and found very few cases of informational genes. In addition, a more recent study with a very different approximation corroborates these observations (Pal *et al.*, 2005). The authors study the influence of different factors on the stability of metabolic networks of bacteria. They identified several instances of horizontal gene transfer events and mapped them in the metabolic network of *Escherichia coli*. The analyses revealed two facts: (1) that most genes are from metabolism categories, and (2) most of the genes occupied external nodes of the metabolic networks.

However, other studies analyzing the phenomena from a phylogenetic point of view only partly agree with these observations. Phylogenetic analysis of incongruence in core data sets has shown the presence of possible transfer events in an important proportion (Susko *et al.*, 2006; Baptiste *et al.*, 2005). Most of the genes in these cores are from informational categories. Recombination events in the elongation factor complex even at large phylogenetic distances have also been reported (Inagaki *et al.*,

2006). A new phylogenomic approach technique called embedded quartets has allowed to infer the number of transfer events between cyanobacteria genomes and among cyanobacteria and external species (Zhaxybayeva *et al.*, 2006). The analysis revealed these genomes as mosaics and the analyses of functional categories showed two distinct patterns. Those transfers among cyanobacteria genomes showed no clear pattern, implying that transfers of informational genes are more plausible between more related genomes whereas those transfers that implied a non-cyanobacteria donor seem to be more biased towards operational genes. The same pattern is observed for the Xanthomonadales genomes as outlined in chapter 5.

Ecological opportunity. The analysis of environmental metagenomes is important in the context of HGT because one of the most important extrinsic factors affecting the process is ecological opportunity. Two genomes may be compatible in terms of genome architecture but if they occupy very different niches with no opportunity for contact then a successful HGT event would be very difficult. The opposing cases of the acid mine metagenome (Tyson *et al.*, 2004) and the Sargasso Sea metagenome (Venter *et al.*, 2004) explained in the Introduction are very illustrative. Only few lineages are present in the former study, where some of them emerged by recombination of the ancestor thus generating a chimaeric organism. This low diversity contrasts with that found in the Sargasso Sea which is composed by thousands of phylotypes (expected species) harbouring millions of genes, some of them totally unknown before. Opportunities for genetic exchange in such an environment are much more likely

than in the acid mine. An open microbiome points clearly to HGT as the main agent of evolutionary novelty with a likelihood of transfers much higher in the second case. It is also important to note that transfers between very large phylogenetic distances are not impossible as demonstrated by Tyson *et al.* (2004). However, as illustrated in the study, transfers from Bacteria to Archaea are usually related with adaptation to a very specific niche, in which case the transfer of functions allowing the survival in this niche has a high associated fitness.

Phage genomics. Phages are clearly mosaic genomes resulting from multiple homologous and non-homologous recombination events. This mosaicism has been described both for double and single strand DNA phages (Lawrence *et al.*, 2002). However, the analysis of different T4-type phages has revealed limitations to this rampant gene transfer (Filee *et al.*, 2006). The authors localized two regions that seem not to be affected by HGT and whose main feature is the conservation of synteny across the analyzed genomes. The genes composing these regions have coupled metabolic functions, characterized by many interactions. A disruption of these regions could result in important losses of fitness. As we have described, genomic architecture conservation seems to be an important factor for successful transfers in bacterial genomes: the evidence from phage genomics corroborates its relevance.

On the other hand, our current view of phages has changed from early reports as agents of antibiotic-resistance carriers to treat them as agents catalyzing bacterial evolution (Canchaya *et al.*, 2003). Their lysogenic action causes the

conversion of a non-pathogenic organism to a pathogenic strain as shown by the phage encoded toxins of causative agents of diphtheria or shigellosis (Brussow *et al.*, 2004). However, the ecology of phages has to be taken into account. They have limited host ranges, most of the time at the species level, although cases of broad ranges covering large phylogenetic distances have been described (Jensen *et al.*, 1998; Beumer and Robinson, 2005). It is known that strains usually share a common phage pool and therefore phage-mediated transfers are more likely between more closely related strains. The same example of the mine drainage used in the above section is useful to illustrate this point. The authors demonstrate that recombination between *Leptospirillum* and between *Ferroplasma* species are phage-dependent as the analysis of the genomes has revealed (Tyson *et al.*, 2004). In another study, the analysis of *Prochlorococcus* genomes through a wide range of ocean light and depth conditions has revealed that there are different ecotypes depending on these variations (Coleman *et al.*, 2006). The distinguishing feature of these ecotypes is the composition of different genomic islands located in their genomes and that have been most likely introduced by phages. These studies are pointing to phages as the main way for HGT in closely related microbial genomes as multistrain analyses of sequenced genomes are revealing (Brussow *et al.*, 2004; Aziz *et al.*, 2005).

7.2.2 A proposal for Bacteria and Archaea

As stated at the beginning, our proposal for the nature and impact of the gene transfer process on microbial genomes is based on the concept that the likelihood of transfers decays with

the phylogenetic distance between the two lineages involved, in analogy to the lower recombination rate between more divergent sequences. We have explained this effect as a by-product of different mechanisms that allow bacteria to prevent the incorporation, silence or eliminate foreign material from more or less distant sources. Some of these mechanisms help to define intra-species lineages (Waldron and Lindsay, 2006) whereas others prevent inter-species exchange (Navarre *et al.*, 2006; Lucchini *et al.*, 2006) and, presumably, other mechanisms that surely act at other phylogenetic depths have still to be identified. However, horizontal gene transfer events between distant taxa are still detectable. Among them only those transfers from distant sources with a high fitness effect will be retained, which in turns are usually linked to niche conditions like the exchange of niche-specific genes between Bacteria and Archaea that share the same extreme environments (Mongodin *et al.*, 2005).

Although the decay of recombination rate with divergence has been demonstrated experimentally (Majewski and Cohan, 1998), only now the growing number of sequenced genomes is allowing to study the age and number of transfers between species. Different studies centring on specific taxonomic groups support our assumptions. The evolution of Xanthomonadales has been driven by ancient and recent HGT events from other Proteobacteria and from other Xanthomonadales genomes respectively. The Cyanobacteria genomes analyzed by Zhaxybayeva *et al.* (2006) revealed that around 50% of their genes have suffered at least an HGT event. What is more relevant for our proposal, intra-Cyanobacteria

transfers were more frequent than between Cyanobacteria and outer genomes. Other studies have shown similar patterns in *Corynebacteria* (Marri *et al.*, 2007). In addition, analyses of multiple genomes covering large phylogenetic distances point towards the hypothesis of phylogenetic preferential sharing. In the most complete analysis published up to now Beiko *et al.* (2005) screened 144 sequenced genomes ranging from Bacteria to Archaea. They derived a supertree from the single gene phylogenies obtained from all gene families retrieved from these genomes and then tried to reconcile the gene tree and the supertree. As the reconciliation of a gene tree equals to a putative HGT event, their conclusion was that most of the inferred transfers were between genomes of the same group, for example between genomes of Gamma-Proteobacteria. Therefore the analysis showed as preferential partners in HGT events those of more closely related genomes as predicted by Gogarten *et al.* (2002).

To simplify our argument we will consider a genome whose capability to accept new genetic material is constant through time. In this case, a linear accumulation of foreign material during the evolutionary history of the genome is expected, as shown in Figure 30a. Therefore, the current genome derived from this lineage will present ancient, recent and ongoing HGTs. From a methodological point of view the detection of these events requires very different approximations.

Ongoing transfers are composed mainly by homologous recombination within the lineage. This kind of transfers is only detectable in a population genetics framework that allows

inferring events and rates of recombination. Many examples exist for the use of these techniques in bacterial populations (Perez-Losada *et al.*, 2006). From a phylogenomic point of view there is no effect on genome phylogenies not only because there is not enough phylogenetic sampling, but also because at this scale recombination acts synergistically with the vertical signal, as a cohesive force of the group (Fraser *et al.*, 2007).

However, homologous recombination is not the only mechanism for acquiring foreign material. Recent transfers from other genomes could also be detected through surrogate methods (Lawrence and Ochman, 2002). These methods identify atypical regions in the genomes based on a significant difference in some compositional measure. Each genome is under different mutational pressures (Sueoka, 1988; Sueoka, 1992; Sueoka, 1993) which results in specific patterns of, for example, nucleotide composition (Lawrence and Ochman, 1997; Ochman *et al.*, 2000; Lawrence and Ochman, 1998), codon usage bias (Medigue *et al.*, 1991), dinucleotide frequencies (Karlin, 2001; Karlin and Burge, 1995) and sequence patterns detected by Markov models (Hayes and Borodovsky, 1998). In consequence, the introduction of new DNA from an HGT event results in the integration in the genome of a sequence with different features than the recipient genome (Ochman *et al.*, 2000). Usually each of these measures tends to generate different sets of atypical genes although it has been argued that the reason is that each one tests a different hypothesis, each one is best suitable to detect transfers from different genomes (Azad and Lawrence, 2005). Although not all the detected atypical genes results from transfers, they are a good

measure of the ongoing transfers in a genome. However, the main limitation of these methods is the elimination of atypical signals with time (Lawrence and Ochman, 1997). This process, called ameriolation, starts when the DNA integrates on the genome and therefore when the sequence starts to suffer the same biases than the rest of the genome. The process of ameriolation is relatively fast and therefore the power to detect atypical genes decreases with time, being valid only for recent transfers. The main advantage of this approach is that it does not rely on a comparative analysis and therefore those genes excluded from a phylogenetic analysis due to their presence in only one or a few lineages can be studied. Phylogenetic methods will detect these transfers if the divergence between the donor and receptor is large enough. For example, a recent transfer between bacteria and archaea is easy to detect whereas a recent transfer between two sequenced strains of *Escherichia coli* would only be detected if the number of strains used in the study is enough to reveal it.

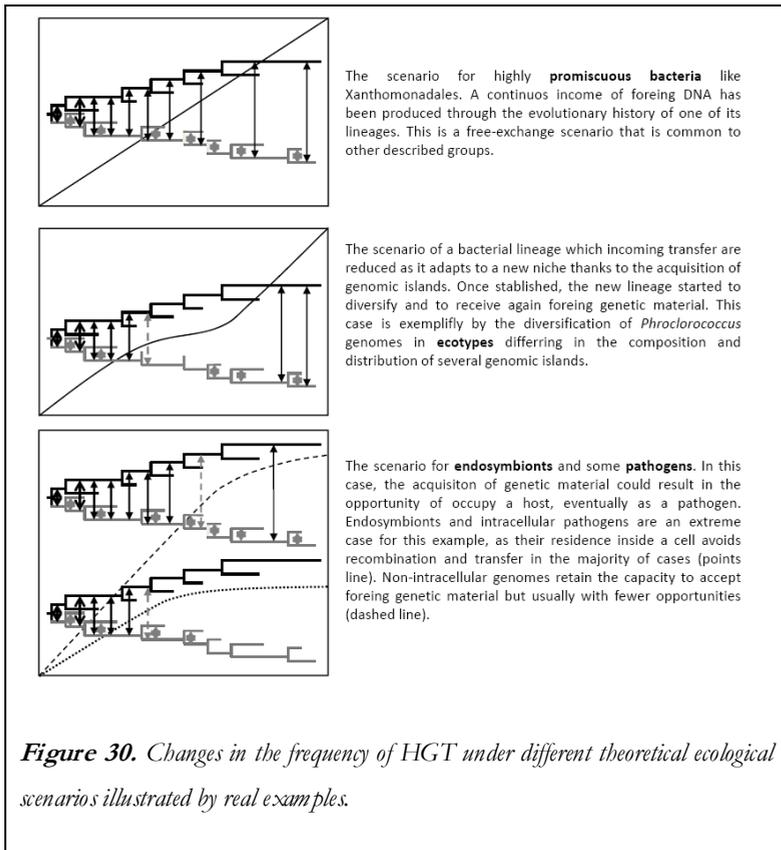
Older transfers can only be detectable by phylogenetic methods and their detection depends to a large extent on the time elapsed since the transfer. If the gene was transferred relatively recently, then it would show a clear horizontal signal, grouping the receptor in the donor's clade. However, if this transfer was older then the gene would carry two contradicting signals. The signal derived from the donor, originated before the transfer, and the signal from the receptor originated after the transfer. The first signal would suffer a process of phylogenetic ameriolation, which partly is due to the compositional ameriolation and to sharing selective pressures with related genomes. For example positive

selection could favour the same non-synonymous changes in the genes of a clade of genomes. Usually this translates into a low-bootstrap supported placement for these genes. When the transfer is very ancient, in many cases it would be impossible to differentiate between a transferred gene and the receptor genome genes because most of the phylogenetic history that could be inferred is shared with the rest of the related genomes. The deeper the transfer the more difficult would be to detect it. This also explains the presence of noise in terms of horizontal signal. Those genes not too old will show support both for the donor phylogeny and for the receptor phylogeny. Therefore, some noise, that is unable to differentiate between the assumed species tree and the transfer hypotheses, is the hallmark of a transfer event and is revealing a mixture of donor and receptor signals.

According to this model, the lineage studied will be receiving foreign DNA at a more or less constant rate, depending on its environment, as shown in Figure 30a. Then the interplay between selection and effective population sizes has the opportunity of filtering each transfer event. However, this is an ideal situation that could fit highly promiscuous bacteria like Xanthomonadales which seem to lack most mechanisms that prevent gene exchange. Other bacteria could have different trajectories.

For example, a free-living bacterial lineage could receive a gene island that eventually allows it to adapt to a new niche. This adaptation usually implies reduced population sizes, recombination rates and possibilities of gene transfers, thus lowering the transfer rate. However, as the lineage adapts to the

new environment, diversification by periodic selection events and acquisition of niche-specific genes could lead to an increase of bacterial lineages or ecotypes that could differentiate to the point that recombination is avoided between them although non-homologous transfers continue (Figure 30b).



This seems to be the case for the distribution of *Prochlorococcus* ecotypes in the ocean (Johnson *et al.*, 2006). Despite sharing a core genome among ecotypes, there is a large variability

in the remaining gene contents, which resides in genomic islands. A high correlation between genomic island composition and structure is found with the niche each ecotype occupies. These islands have been acquired by phages which therefore were the vehicle for the *Procholorococcus* diversification. The establishment of the lineage in a new niche allows new opportunities for recombination and gene sharing.

But the acquisition of genes that promote the invasion of a new niche could imply that new genetic material is not necessary because the genome has specialized in exploiting the new niche. This is the case for most pathogens derived from a free-living ancestor. The free-living genome receives transfers from other genomes. Eventually some of these transfers, presumably pathogenicity islands, allow it to infect a host as a pathogen. As the transition to a pathogenic lifestyle continues the capacity of receiving external genetic material decreases both because it is less important and because the environment offers fewer opportunities. Most bacterial pathogens show an intermediate pattern of reduced HGT capacity and genome decay signals due to constraints both in population sizes and in the selective advantages of new genetic material (Figure 30c). The most extreme cases are those of intracellular pathogens and endosymbionts which couple specialization in the new host with loss of non-essential genes for its lifestyle. Examples of intracellular pathogens are *Mycobacterium leprae* (Vissa and Brennan, 2001) or *Rickettsia prowazekii* (Müller and Martin, 1999) (Figure 30c). Endosymbiont genomes show the same processes than obligatory intracellular pathogens but much more accentuated,

with perfect gene order conservation among strains and with no opportunity for gene exchange (Wernegreen, 2002). For these cases, horizontal gene transfers to the non-intracellular ancestors of these genomes might be detected but only if they have not been lost along the genome reduction process.

7.3 Genome reduction and lifestyle evolution in microbial genomes

As mentioned earlier in this thesis, bacteria evolve through gene modification, gene gain, gene loss and gene duplication. Although the loss of DNA has been specially studied in intracellular bacteria, mainly endosymbionts, it is a common feature in pathogens and also has been documented in free-living taxa (Wernegreen, 2005). However, genome reduction is not always due to the same factors but responds to different environmental stimuli and evolutionary forces as we will detail in this section.

It is not surprising that two of the most reduced free-living bacterial genomes have been found in oceanic surveys. The extent of bacterial diversity on oceanic ecosystems has been one of the main targets of environmental genomics studies (Venter *et al.*, 2004; Rusch *et al.*, 2007). *Pelagibacter ubique* is the free-living bacteria with the smallest genome known (Giovannoni *et al.*, 2005). It is a heterotrophic marine Alpha-Proteobacteria. Its 1.308.759 base pairs retain most of the functions known for its taxonomic group but some features convert it in an exceptional case of

streamlining, the process by which natural selection favours minimalization in order to reduce the metabolic costs of cellular replication. Despite possessing the genes necessary to allow horizontal gene transfer, no such recent event has been detected. This is in agreement with the small number of paralogous genes that points towards a selection pressure to keep small gene family sizes. Finally, non-functional and redundant DNA is absent which is reflected in very small intergenic regions (Giovannoni *et al.*, 2005). This genome is one of the dominant clades in oceanic samples (Rusch *et al.*, 2007) revealing that it is a derived state of a larger ancestor and that it represents one of the clearest examples of adaptation by DNA loss, at least as efficient as other equivalent heterotrophic bacterium with much larger genomes (4-5 Mbp.) (Giovannoni *et al.*, 2005).

Another aspect of the genome reduction process is reflected in the marine unicellular cyanobacterium *Prochlorococcus marinus* (Dufresne *et al.*, 2003; Rocap *et al.*, 2003). It is also a free-living bacterium which is responsible for an important fraction of the global photosynthesis. The comparison of three strains of this species revealed specific adaptations to different minimum, maximum and optimal wave-lengths intensities. These adaptations are achieved by differential loss of genes from the common ancestor coupled with the gain of new functions by horizontal gene transfer, which seems to have driven the colonization of different niches by the differential transfer of phage-mediated genomic islands (Coleman *et al.*, 2006; Johnson *et al.*, 2006). This translates into important differences both in gene content (1.716-2.375 genes) and genome size (1.657.990-2.420.873 base pairs) and

also in G+C composition, being more (A+T)-rich the most reduced strain (30.8-50.74%). The two examples reported here show the divergent ways in which genome reduction can act even in free-living bacterium. The streamlining of *Pelagibacter* resides mainly in non-functional DNA and gene loss coupled with the absence of DNA gain events while it still retains most metabolic routes whereas in *Prochlorococcus* ecotypes differential gene losses and gains have allowed them to diversify and occupy different niches.

Gene loss has played a fundamental role in the evolution of most pathogens, both by fine-tuning the genomes of new pathogen lineages or by massive gene decay as a result of host specialization (Lawrence, 2005). Usually a pathogen strain arises due to the horizontal acquisition of genes, most of the times pathogenicity islands, by a commensalist ancestor. The term **pathoadaptive evolution** is used to refer to those genomic changes needed in recently arisen pathogens to remove those adaptations useful in the commensal ancestors but with no advantage for a pathogenic lifestyle (Maurelli, 2007). In bacterial genomes the most exemplar case of pathoadaptive changes is the evolution of the pathogenic *Shigella* species from non-pathogenic *Escherichia coli* genomes. Phylogenetic and population analyses demonstrate that *Shigella* and *Escherichia* species are not different species (Wirth *et al.*, 2006; Pupo *et al.*, 1997). *Shigella* pathogenicity has evolved multiple times after the acquisition by different *Escherichia coli* genomes of a virulence plasmid (Pupo *et al.*, 2000). This convergent evolution is also reflected in the independent loss of *cadA* and *ompT* genes in the different *Shigella* strains. Those

genes are a clear case of pathoadaptive evolution and, because they decrease fitness in the new pathogens, are considered antivirulence genes. They are useful for the non-pathogenic *Escherichia* counterparts but reduce the virulence of the *Shigella* strains (Maurelli *et al.*, 1998). Similar examples have been found in *Burkholderia pseudomallei* and its pathogenic counterpart *Burkholderia mallei*, with the difference that in this case the pathogenic lifestyle seems to be attained only by gene loss (antivirulence genes) with no evidence of gene gain (Moore *et al.*, 2004). A similar scenario is found in the relationship between the non-pathogenic *Bacillus cereus* and *Bacillus anthracis*, the causative agent of anthrax, which instead of living in soil is able to access mammalian blood and tissues. Some of the traits for the exploitation of a soil niche in *B. cereus* are clearly antipathogenic in *B. anthracis* and therefore they have been lost or modified during its evolution (Mignot *et al.*, 2001).

It is important to differentiate between pathoadaptive changes like those reported above, which sometimes require loss of antivirulence genes, and the nature of massive gene decay observed in some pathogens. The first case is linked to the inhibition of virulence that some traits of the non-pathogenic ancestor exert in the new pathogenic strain. Therefore, selection favours the loss of those genes. In the case of massive gene decay it is linked with an increasing niche specialization of the pathogen usually coupled with an intracellular lifestyle in the host. This is the case for pathogens like *Mycobacterium leprae* (Cole *et al.*, 2001) or *Bordetella pertussis* (Parkhill *et al.*, 2003) characterized by a high number of pseudogenes, around 1000 and 200 respectively,

because of the loss of selection pressures to maintain free-living or redundant functions in the genome. Other examples are found in the recently intracellular pathogen *Coxiella burnetii* (Beare *et al.*, 2006), a close relative of *Legionella pneumophila*. Therefore they are usually regarded as examples of the first stages of what is known as **reductive evolution** (Wernegreen, 2005). These genomes are usually characterized not only by a large number of pseudogenes but also by the presence of insertion sequences and mobile DNA that cause their instability both in gene order and gene content (Moran and Plague, 2004). More ancient associations like those of some *Rickettsia* species (c.a 400 mya) (Müller and Martin, 1999) or some Chlamydiales (c.a 700 mya) (Horn *et al.*, 2004) seem to be more stable, mainly because most of these elements have been lost and horizontal gene transfer is prevented or at least reduced but pseudogenes are still present in most of them. However, an increasing number of works are reporting cases of intracellular genomes with a fluid nature. For example, cases of recombination in intracellular pathogens (Baldo *et al.*, 2006) and non-homologous gene transfers are known (Ogata *et al.*, 2006; Ogata *et al.*, 2005; Van Ham *et al.*, 2000).

The extensive gene decay in intracellular pathogens mentioned above is common with the other extreme of intracellular lifestyle, the mutualists that increase the fitness of the host. Despite endosymbionts have been found in a wide variety of hosts, only insect endosymbionts, mainly of the Gamma-Proteobacteria group, have been sequenced up to date. It is worth mentioning that some genome sequences of closely relative strains are available (eg. *Blochmannia* and *Buchnera* strains) which has

allowed the development of the comparative genomics of endosymbionts. Consequently, the tempo and mode of gene loss has been studied (Gomez-Valero *et al.*, 2007; Silva *et al.*, 2001; Mira *et al.*, 2001), endosymbiont lifestyle genes have been identified (Gil *et al.*, 2003), a minimal endosymbiont and minimal bacterial genomes have been proposed (Gil *et al.*, 2004b) and inferences about the dynamics of the reductive evolution process have been made (Khachane *et al.*, 2007). However, the sequencing of the most reduced genomes known until today (Nakabachi *et al.*, 2006; Perez-Brocal *et al.*, 2006) has posed new questions about the final stages of the process and the ultimate fate of endosymbionts.

Like in the case of intracellular pathogens, the first stages of genome reduction in endosymbionts are characterized by a massive pseudogeneization process because of the redundant and useless information of most of their genes (Wernegreen, 2005). The irreversible association of the endosymbiont with the host and its strictly vertical transmission impose severe bottlenecks to its populations; the absence of documented recombination contributes to the fixation of slightly deleterious mutations by Muller's ratchet (Moran, 1996). The ratchet is thought to be the cause of the pseudogeneization of important but not essential genes in the next stages of gene decay. It is coupled with an increasing bias toward A+T content (Clark *et al.*, 1999), loss of effective codon usage (Rispe *et al.*, 2004) and low levels of intraspecific polymorphisms (Funk *et al.*, 2001). Furthermore, the loss of genes implied in repair pathways also seems to be related with high mutation rates (Wu *et al.*, 2006), which results in an

acceleration of evolutionary rates from free-living relatives and favours the fixation of deleterious mutations.

These features are also present in the most ancient bacterial endosymbionts, the organelles of eukaryotic cells. However, it is worth questioning the stability of the host-endosymbiont association. It could be argued that the process, as presented in the above paragraph, is irreversible and drives inexorably to the extinction of the genome. However, organellar genomes are clear representatives of a stable association. We can also discard that the endosymbiont genomes sequenced have reached an optimum, since different strains of *Buchnera* and *Blochmannia* have different genome sizes, therefore revealing that further genome reduction is possible. Maybe the key is in those genomes that are in the last stages of genome degradation: *Buchnera aphidicola* *Cinara cedri* and *Carsonella ruddi*.

Is reductive evolution an irreversible process? The analysis of the evolutionary forces acting on these two genomes seems to shed light to this question. Although usually a single evolutionary force has been proposed in order to account for the high evolutionary rates of these genomes and subsequent sequence degeneration and loss (Itoh *et al.*, 2002; Moran, 1996), we think that both genetic drift and an enhanced mutation rate are responsible for the mutation accumulation in endosymbiont genes. Increased mutation rates provide a higher supply of mildly and slightly deleterious mutations which coupled with the bottleneck transmission result in a higher rate of fixation due to genetic drift. This process is somewhat counterbalanced by the action of positive selection on specific sites which introduce non-

synonymous changes towards G/C and that seem to be more frequent in the most reduced genomes. This positive selection could be explained not only because the need to maintain certain functions but also the replicon structure of the whole genome opposing to changes towards A/T.

An inverse relationship between purifying and positive selection is observed across the endosymbiont genomes. It is possible that when the reductive process is so strong that purifying selection is unable to keep the integrity of the genes because A/T changes can not be removed fast enough, then positive selection appears favouring G/C changes. In fact, under this hypothesis it is not surprising that those genes with evidence for positive selection towards G/C have a significantly higher G+C composition. Is positive selection enough to stop the process? It is possible that a steady state could be reached, but we think that this equilibrium point, if exists, is unstable and that positive selection mainly slows down the degeneration of this extremely reduced genome.

8. GENERAL CONCLUSIONS

- Phylogenomic methodologies are the best suited methods to assess the relationships among bacterial taxa. They take advantage of the vast genome information generated nowadays although they are still limited by insufficient taxon sampling due to the costly time and economic efforts of sequencing complete or almost complete genomes.
- Usually the selection of a phylogenomic methodology is a trade-off between the number of characters available and the number of taxa available. Those based in common features for all the taxa such as gene content comparisons or sequence concatenation, are limited by the low number of shared genes when phylogenetic distances are high.
- Supermatrices of concatenated genes seem to reveal the most frequent phylogenetic signal in the alignment although tend to ignore alternative, equally important, phylogenetic signals.
- Supertrees and consensus trees are a good approximation to detect possible instances of rampant horizontal gene transfer in a genome. However, low resolved nodes do not necessarily imply this phenomenon, since it could result from a systematic phylogenetic artefact in the corresponding gene trees.
- Congruence maps are the best way to detect the presence of different phylogenetic signals in microbial genomes, although in some situations the differentiation between noise and signal might require further analyses.

- Gamma-Proteobacteria endosymbionts of the groups *Buchnera*, *Blochmannia* and *Wigglesworthia* seems to form a monophyletic clade near of the *Enterobacteriaceae* group, therefore indicating a common bacterium ancestor or a close relationship among the bacterial ancestor that established the endosymbiotic relationship.
- Endosymbiont genomes are under the influence of different evolutionary forces that enhance their non-synonymous substitution rates coupled with an increasing content in A+T positions.
- Enhanced mutation rates supply a large number of substitutions that are more likely to be fixed, even if they are detrimental, due to the intense genetic drift resulting from population bottlenecks due to their strict vertical transmission.
- Selection acts in the form of purifying and relaxed selection. However, the action of positive selection is also remarkable especially for the cases of the most reduced genomes, *Carsonella ruddii* and *Buchnera aphidicola* *Cinara Cedri*.
- Positive selection seems to introduce changes towards G/C in order to counterbalance the A+T accumulation typical of intracellular genomes. Evidence for the existence of positive selection is stronger in the most reduced genomes. Simulations have shown that this is not an artefact introduced by their high A+T content and their saturated synonymous substitution rate.

- *Carsonella ruddii*, the most reduced bacterial genome known, is not near the position of the remaining endosymbionts clade in the Proteobacteria tree. It's most likely position after correcting for the possible A+T bias artefact is as a basal clade near the *Legionellaceae* group.
- Xanthomonadales genomes are a clear example of mosaic origin of their genes, up to the point that it is very difficult to determine their most likely phylogenetic origin
- Past horizontal gene transfers to the Xanthomonadales genomes came from other Proteobacteria groups, when the divergence between them was still low.
- Recent horizontal gene transfers to the Xanthomonadales genomes primarily arrive from other Xanthomonadales and therefore do not alter their phylogenetic relationships with other Proteobacteria.

9. REFERENCES

- Abby,S., and Daubin,V. (2007) Comparative genomics and the evolution of prokaryotes. *Trends Microbiol* **15**: 135-141.
- Akman,L., Yamashita,A., Watanabe,H., Oshima,K., Shiba,T., Hattori,M., and Aksoy,S. (2002) Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. *Nat Genet* **32**: 402-407.
- Allen,E.E., Tyson,G.W., Whitaker,R.J., Detter,J.C., Richardson,P.M., and Banfield,J.F. (2007) Genome dynamics in a natural archaeal population. *Proc Natl Acad Sci U.S. A* **104**: 1883-1888.
- Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-3402.
- Andersson,J.O. (2005) Lateral gene transfer in eukaryotes. *Cell Mol Life Sci* **62**: 1182-1197.
- Andersson,S.G., and Kurland,C.G. (1998) Reductive evolution of resident genomes. *Trends Microbiol* **6**: 263-268.
- Arbiza,L., Dopazo,J., and Dopazo,H. (2006) Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS Computational Biology* **2**: e38.
- Asai,T., Zaporozhets,D., Squires,C., and Squires,C.L. (1999) An *Escherichia coli* strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria. *Proc Natl Acad Sci U.S. A* **96**: 1971-1976.
- Avery,O.T., MacLeod,C.M., and McCarty,M. (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Inductions of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J Exp Med* **79**: 137-158.
- Azad,R.K., and Lawrence,J.G. (2005) Use of artificial genomes in assessing methods for atypical gene detection. *PLoS Computational Biology* **1**: e56.

- Aziz,R.K., Edwards,R.A., Taylor,W.W., Low,D.E., McGeer,A., and Kotb,M. (2005) Mosaic prophages with horizontally acquired genes account for the emergence and diversification of the globally disseminated M1T1 clone of *Streptococcus pyogenes*. *J Bacteriol* **187**: 3311-3318.
- Bailly-Bechet,M., Danchin,A., Iqbal,M., Marsili,M., and Vergassola,M. (2006) Codon usage domains over bacterial chromosomes. *PLoS Computational Biology* **2**: e37.
- Baldo,L., Bordenstein,S., Wernegreen,J.J., and Werren,J.H. (2006) Widespread recombination throughout Wolbachia genomes. *Mol Biol Evol* **23**: 437-449.
- Baptiste,E., Boucher,Y., Leigh,J., and Doolittle,W.F. (2004) Phylogenetic reconstruction and lateral gene transfer. *Trends Microbiol* **12**: 406-411.
- Baptiste,E., Susko,E., Leigh,J., MacLeod,D., Charlebois,R.L., and Doolittle,W.F. (2005) Do orthologous gene phylogenies really support tree-thinking? *BMC Evolutionary Biology* **5**: 33.
- Baptiste,E., Brinkmann,H., Lee,J.A., Moore,D.V., Sensen,C.W., Gordon,P. *et al.* (2002) The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. *Proc Natl Acad Sci U.S. A.* **99**: 1414-1419.
- Baum,B.R. (1992) Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* **41**: 3-10.
- Bazin,E., Glemin,S., and Galtier,N. (2006) Population size does not influence mitochondrial genetic diversity in animals. *Science* **312**: 570-572.
- Beare,P.A., Samuel,J.E., Howe,D., Virtaneva,K., Porcella,S.F., and Heinzen,R.A. (2006) Genetic diversity of the Q fever agent, *Coxiella burnetii*, assessed by microarray-based whole-genome comparisons. *J Bacteriol* **188**: 2309-2324.

- Beiko,R.G., Harlow,T.J., and Ragan,M.A. (2005) Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci U.S. A* **102**: 14332-14337.
- Belda,E., Moya,A., and Silva,F.J. (2005) Genome rearrangement distances and gene order phylogeny in Gamma-Proteobacteria. *Mol Biol Evol* **22**: 1456-1467.
- Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A., and Wheeler,D.L. (2002) GenBank. *Nucleic Acids Res* **30**: 17-20.
- Bentley,S.D., and Parkhill,J. (2004) COMPARATIVE GENOMIC STRUCTURE OF PROKARYOTES. *Annu Rev Genet* **38**: 771-791.
- Bern,M., and Goldberg,D. (2005) Automatic selection of representative proteins for bacterial phylogeny. *BMC Evolutionary Biology* **5**: 34.
- Beumer,A., and Robinson,J.B. (2005) A broad-host-range, generalized transducing phage (SN-T) acquires 16S rRNA genes from different genera of bacteria. *Appl Environ Microbiol* **71**: 8301-8304.
- Bininda-Emonds,O.R. (2004a) Trees versus characters and the supertree/supermatrix "paradox". *Syst Biol* **53**: 356-359.
- Bininda-Emonds,O.R.P., Gittleman,J.L., and Steel,M.A. (2002) The (Super)tree of life: Procedures, problems, and prospects. *Annu Rev Ecol Syst* **33**: 265-289.
- Bininda-Emonds,O.R.P., and Sanderson,M.J. (2001) Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. *Syst Biol* **50**: 565-579.
- Bininda-Emonds,O.R.P. (2004b) The evolution of supertrees. *Trends Ecol Evol* **19**: 315-322.
- Boccard,F., Esnault,E., and Valens,M. (2005) Spatial arrangement and macrodomain organization of bacterial chromosomes. *Mol Microbiol* **57**: 9-16.

- Boucher, Y., Douady, C.J., Papke, R.T., Walsh, D.A., Boudreau, M.E.R., Nesbo, C.L. *et al.* (2003) Lateral gene transfer and the origins of prokaryotic groups. *Annu Rev Genet* **37**: 283-328.
- Brewer, B.J. (1988) When polymerases collide: Replication and the transcriptional organization of the *E. coli* chromosome. *Cell* **53**: 679-686.
- Brinkmann, H., and Philippe, H. (1999) Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Mol Biol Evol* **16**: 817-825.
- Brochier, C., Baptiste, E., Moreira, D., and Philippe, H. (2002) Eubacterial phylogeny based on translational apparatus proteins. *Trends Genet* **18**: 1-5.
- Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E., and Stanhope, M.J. (2001) Universal trees based on large combined protein sequence data sets. *Nat Genet* **28**: 281-285.
- Brussow, H., Canchaya, C., and Hardt, W.D. (2004) Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol Rev* **68**: 560-602.
- Brynnel, E.U., Kurland, C.G., Moran, N.A., and Andersson, S.G. (1998) Evolutionary rates for *tuf* genes in endosymbionts of aphids. *Mol Biol Evol* **15**: 574-582.
- Buchner, P. (1965) Endosymbiosis of animals with plant microorganisms. In Interscience. New York.
- Burleigh, J., Amy, D., and Michael, S. (2006) Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome-scale data sets. *Syst Biol* **55**: 426-440.
- Canback, B., Tamas, I., and Andersson, S.G.E. (2004) A phylogenomic study of endosymbiotic bacteria. *Mol Biol Evol* **21**: 1110-1122.
- Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M.L., and Brussow, H. (2003) Phage as agents of lateral gene transfer. *Current Opinion in Microbiology* **6**: 417-424.

- Castillo,J.A., and Greenberg,J.T. (2007) Evolutionary Dynamics of *Ralstonia solanacearum*. *Appl Environ Microbiol* **73**: 1225-1238.
- Castresana,J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**: 540-552.
- Charlebois,R.L., Clarke,G.D.P., Beiko,R.G., and Jean,A. (2003) Characterization of species-specific genes using a flexible, web-based querying system. *FEMS Microbiology Letters* **225**: 213-220.
- Charlebois,R.L., and Doolittle,W.F. (2004) Computing prokaryotic gene ubiquity: Rescuing the core from extinction. *Genome Res* **14**: 2469-2477.
- Charles,H., Heddi,A., and Rahbe,Y. (2001) A putative insect intracellular endosymbiont stem clade, within the Enterobacteriaceae, inferred from phylogenetic analysis based on a heterogeneous model of DNA evolution. *C R Acad Sci III* **324**: 489-494.
- Chen,S.L., Hung,C.S., Xu,J., Reigstad,C.S., Magrini,V., Sabo,A. *et al.* (2006) Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: A comparative genomics approach. *Proc Natl Acad Sci U.S. A* **103**: 5977-5982.
- Ciccarelli,F.D., Doerks,T., von Mering,C., Creevey,C.J., Snel,B., and Bork,P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**: 1283-1287.
- Clark,M.A., Moran,N.A., and Baumann,P. (1999) Sequence evolution in bacterial endosymbionts having extreme base compositions. *Mol Biol Evol* **16**: 1586-1598.
- Coffey,T.J., Dowson,C.G., Daniels,M., Zhou,J., Martin,C., Spratt,B.G., and Musser,J.M. (1991) Horizontal transfer of multiple penicillin-binding protein genes, and capsular biosynthetic genes, in natural populations of *Streptococcus pneumoniae*. *Mol Microbiol* **5**: 2255-2260.
- Cohan,F.M. (2001) Bacterial species and speciation. *Syst Biol* **50**: 513-524.

- Cohan,F.M. (2002) Sexual isolation and speciation in bacteria. *Genetica* **116**: 359-370.
- Cohan,F.M. (2004) Concepts of bacterial biodiversity for the age of genomics. In *Micronial genomes*. Fraser,C.M., Read,T., and Nelson,K.E. (eds). Humana Press, pp. 175-194.
- Cole,S.T., Eiglmeier,K., Parkhill,J., James,K.D., Thomson,N.R., Wheeler,P.R. *et al.* (2001) Massive gene decay in the leprosy bacillus. *Nature* **409**: 1007-1011.
- Coleman,M.L., Sullivan,M.B., Martiny,A.C., Steglich,C., Barry,K., DeLong,E.F., and Chisholm,S.W. (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* **311**: 1768-1770.
- Coscolla,M., and Gonzalez-Candelas,F. (2007) Population structure and recombination in environmental isolates of *Legionella pneumophila*. *Environ microbiol* **9**: 643-656.
- Cowan,D., Meyer,Q., Stafford,W., Muyanga,S., Cameron,R., and Wittwer,P. (2005) Metagenomic gene discovery: past, present and future. *Trends Biotechnol* **23**: 321-329.
- Creevey,C. (2004) Clann: Construction of Supertrees and exploration of phylogenomic information from partially overlapping datasets. <http://bioinf. May. ie/software/clann/>.
- Creevey,C.J., Fitzpatrick,D.A., Philip,G.K., Kinsella,R.J., O'Connell,M.J., Pentony,M.M. *et al.* (2004) Does a tree-like phylogeny only exist at the tips in the prokaryotes? *Proc R Soc Lond B Biol Sci* **271**: 2551-2558.
- Creevey,C.J., and McInerney,J.O. (2005) Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics* **21**: 390-392.
- Dagan,T., and Martin,W. (2007) Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci U.S. A.* **104**: 870-875.
- Dagan,T., and Martin,W. (2006) The tree of one percent. *Genome Biology* **7**: 118.

- Darling,A.C.E., Mau,B., Blattner,F.R., and Perna,N.T. (2004) Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14**: 1394-1403.
- Darwin,C. (1859) On the origin of species by means of natural selection. Murray, London.
- Daubin,V., Gouy,M., and Perriere,G. (2002) A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res* **12**: 1080-1090.
- Degnan,P.H., Lazarus,A.B., and Wernegreen,J.J. (2005) Genome sequence of *Blochmannia pennsylvanicus* indicates parallel evolutionary trends among bacterial mutualists of insects. *Genome Res* **15**: 1023-1033.
- Delsuc,F., Brinkmann,H., and Philippe,H. (2005) Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* **6**: 361-375.
- Didelot,X., Achtman,M., Parkhill,J., Thomson,N.R., and Falush,D. (2007) A bimodal pattern of relatedness between the *Salmonella Paratyphi* A and *Typhi* genomes: Convergence or divergence by homologous recombination? *Genome Res* **17**: 61-68.
- Doolittle,W.F. (1999a) Lateral genomics. *Trends Cell Biol* **9**: M5-M8.
- Doolittle,W.F. (1999b) Phylogenetic classification and the universal tree. *Science* **284**: 2124-2128.
- Doolittle,W.F., and Baptiste,E. (2007) Pattern pluralism and the Tree of Life hypothesis. *Proc Natl Acad Sci U.S. A.* **104**: 2043-2049.
- Doolittle,W.F., and Papke,R.T. (2006) Genomics and the bacterial species problem. *Genome Biol* **7**: 116.
- Driskell,A.C., Ane,C., Burleigh,J.G., McMahon,M.M., O'Meara,B.C., and Sanderson,M.J. (2004) Prospects for building the tree of life from large sequence databases. *Science* **306**: 1172-1174.
- Dufresne,A., Salanoubat,M., Partensky,F., Artiguenave,F., Axmann,I.M., Barbe,V. *et al.* (2003) From the Cover: Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a

nearly minimal oxyphototrophic genome. *Proc Natl Acad Sci U.S.A.* **100**: 10020-10025.

Dutilh,B.E., Huynen,M.A., Bruno,W.J., and Snel,B. (2005) The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. *J Mol Evol* **58**: 527-539.

Dutilh,B.E., van,N., V, van der Heijden,R.T., Boekhout,T., Snel,B., and Huynen,M.A. (2007) Assessment of phylogenomic and orthology approaches for phylogenetic inference. *Bioinformatics* **23**: 815-824.

Eisen,J.A. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res* **8**: 163-167.

Eisen,J.A., and Fraser,C.M. (2003) Phylogenomics: intersection of evolution and genomics. *Science* **300**: 1706-1707.

Fares,M.A., Barrio,E., Sabater-Munoz,B., and Moya,A. (2002) The evolution of the heat-shock protein GroEL from *Buchnera*, the primary endosymbiont of aphids, is governed by positive selection. *Mol Biol Evol* **19**: 1162-1170.

Feil,E.J., Maiden,M.C., Achtman,M., and Spratt,B.G. (1999) The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. *Mol Biol Evol* **16**: 1496-1502.

Feil,E.J., Holmes,E.C., Bessen,D.E., Chan,M.S., Day,N.P.J., Enright,M.C. *et al.* (2001) Recombination within natural populations of pathogenic bacteria: Short-term empirical estimates and long-term phylogenetic consequences. *Proc Natl Acad Sci U.S.A.* **98**: 182-187.

Feil,E.J., Li,B.C., Aanensen,D.M., Hanage,W.P., and Spratt,B.G. (2004) eBURST: Inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J Bacteriol* **186**: 1518-1530.

Feil,E.J., Smith,J.M., Enright,M.C., and Spratt,B.G. (2000) Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data. *Genetics* **154**: 1439-1450.

- Felsenstein, J. (2002) PHYLIP (phylogeny inference package). Version 3.6. Distributed by the author. Department of Genetics, University of Washington, Seattle.
- Filee, J., Baptiste, E., Susko, E., and Krisch, H.M. (2006) A selective barrier to horizontal gene transfer in the T4-Type bacteriophages that has preserved a core genome with the viral replication and structural genes. *Mol Biol Evol* **23**: 1688-1696.
- Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**: 496-512.
- Foster, P.G., and Hickey, D.A. (2002) Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J Mol Evol* **48**: 284-290.
- Fraser, C., Hanage, W.P., and Spratt, B.G. (2007) Recombination and the nature of bacterial speciation. *Science* **315**: 476-480.
- Fraser, C., Hanage, W.P., and Spratt, B.G. (2005) Neutral microepidemic evolution of bacterial pathogens. *Proc Natl Acad Sci U.S. A.* **102**: 1968-1973.
- Fry, A.J., and Wernegreen, J.J. (2005) The roles of positive and negative selection in the molecular evolution of insect endosymbionts. *Gene* **355**: 1-10.
- Funk, D.J., Wernegreen, J.J., and Moran, N.A. (2001) Intraspecific variation in symbiont genomes: bottlenecks and the aphid-*Buchnera* association. *Genetics* **157**: 477-489.
- Galperin, M.Y., and Koonin, E.V. (2004) 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucl Acids Res* **32**: 5452-5463.
- Galtier, N., and Gouy, M. (1995) Inferring phylogenies from DNA sequences of unequal base compositions. *Proc Natl Acad Sci U.S. A.* **92**: 11317-11321.

- Galtier,N., Gouy,M., and Gautier,C. (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* **12**: 543-548.
- Gatesy,J., and Baker,R.H. (2005) Hidden likelihood support in genomic data: can forty-five wrongs make a right? *Syst Biol* **54**: 483-492.
- Gatesy,J., Matthee,C., DeSalle,R., and Hayashi,C. (2002) Resolution of a supertree/supermatrix paradox. *Syst Biol* **51**: 652-664.
- Gatesy,J., Milinkovitch,M., Waddell,V., and Stanhope,M. (1999) Stability of cladistic relationships between Cetacea and higher-level artiodactyl taxa. *Syst Biol* **48**: 6-20.
- Gevers,D., Cohan,F.M., Lawrence,J.G., Spratt,B.G., Coenye,T., Feil,E.J. *et al.* (2005) Opinion: Re-evaluating prokaryotic species. *Nat Rev Microbiol* **3**: 733-739.
- Gevers,D., Vandepoele,K., Simillion,C., and Van de Peer,Y. (2004) Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol* **12**: 148-154.
- Gil,R., Silva,F.J., Zientz,E., Delmotte,F., Gonzalez-Candelas,F., Latorre,A. *et al.* (2003) The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. *Proc Natl Acad Sci U.S. A* **100**: 9388-9393.
- Gil,R., Latorre,A., and Moya,A. (2004a) Bacterial endosymbionts of insects: insights from comparative genomics. *Environ Microbiol* **6**: 1109-1122.
- Gil,R., Silva,F.J., Pereto,J., and Moya,A. (2004b) Determination of the Core of a Minimal Bacterial Gene Set. *Microbiol Mol Biol Rev* **68**: 518-537.
- Giovannoni,S.J., Tripp,H.J., Givan,S., Podar,M., Vergin,K.L., Baptista,D. *et al.* (2005) Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**: 1242-1245.

- Glass,J.I., Assad-Garcia,N., Alperovich,N., Yooseph,S., Lewis,M.R., Maruf,M. *et al.* (2006) Essential genes of a minimal bacterium. *Proc Natl Acad Sci U.S. A* **103**: 425-430.
- Godoy,D., Randle,G., Simpson,A.J., Aanensen,D.M., Pitt,T.L., Kinoshita,R., and Spratt,B.G. (2003) Multilocus sequence typing and evolutionary relationships among the causative agents of melioidosis and glanders, *Burkholderia pseudomallei* and *Burkholderia mallei*. *J Clin Microbiol* **41**: 2068-2079.
- Gogarten,J.P., and Olendzenski,L. (1999) Orthologs, paralogs and genome comparisons. *Curr Opin Genet Dev* **9**: 630-636.
- Gogarten,J.P., Doolittle,W.F., and Lawrence,J.G. (2002) Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* **19**: 2226-2238.
- Gogarten,J.P., and Townsend,J.P. (2005) HORIZONTAL GENE TRANSFER, GENOME INNOVATION AND EVOLUTION. *Nat Rev Micro* **3**: 679-687.
- Goldman,N., Anderson,J.P., and Rodrigo,A.G. (2000) Likelihood-based tests of topologies in phylogenetics. *Syst Biol* **49**: 652-670.
- Gomes,J.P., Bruno,W.J., Nunes,A., Santos,N., Florindo,C., Borrego,M.J., and Dean,D. (2007) Evolution of *Chlamydia trachomatis* diversity occurs by widespread interstrain recombination involving hotspots. *Genome Res* **17**: 50-60.
- Gomez-Valero,L., Latorre,A., and Silva,F.J. (2004a) The evolutionary fate of nonfunctional DNA in the bacterial endosymbiont *Buchnera aphidicola*. *Mol Biol Evol* **21**: 2172-2181.
- Gomez-Valero,L., Silva,F.J., Christophe Simon,J., and Latorre,A. (2007) Genome reduction of the aphid endosymbiont *Buchnera aphidicola* in a recent evolutionary time scale. *Gene* **389**: 87-95.
- Gomez-Valero,L., Soriano-Navarro,M., Perez-Brocac,V., Heddi,A., Moya,A., Garcia-Verdugo,J.M., and Latorre,A. (2004b) Coexistence of *Wolbachia* with *Buchnera aphidicola* and a Secondary symbiont in the aphid *Cinara cedri*. *J Bacteriol* **186**: 6626-6633.

- Gontcharov,A.A., Marin,B., and Melkonian,M. (2004) Are combined analyses better than single gene phylogenies? A case study using SSU rDNA and rbcL sequence comparisons in the *Zygnematophyceae* (*Streptophyta*). *Mol Biol Evol* **21**: 612-624.
- Gophna,U., Doolittle,W.F., and Charlebois,R.L. (2005) Weighted genome trees: refinements and applications. *J Bacteriol* **187**: 1305-1316.
- Gu,X., and Zhang,H. (2004) Genome phylogenetic analysis based on extended gene contents. *Mol Biol Evol* **21**: 1401-1408.
- Guindon,S., and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: 696-704.
- Hanage,W., Spratt,B., Turner,K., and Fraser,C. (2006a) Modelling bacterial speciation. *Phil Trans R Soc B* **361**: 2039-2044.
- Hanage,W., Fraser,C., and Spratt,B. (2005) Fuzzy species among recombinogenic bacteria. *BMC Biology* **3**: 6.
- Hanage,W.P., Fraser,C., and Spratt,B.G. (2006b) The impact of homologous recombination on the generation of diversity in bacteria. *J Theor Biol* **239**: 210-219.
- Handelsman,J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* **68**: 669-685.
- Hayes,W.S., and Borodovsky,M. (1998) How to interpret an anonymous bacterial genome: machine learning approach to gene identification. *Genome Res* **8**: 1154-1171.
- Heddi,A., Charles,H., Khatchadourian,C., Bonnot,G., and Nardon,P. (1998) Molecular characterization of the principal symbiotic bacteria of the weevil *Sitophilus oryzae*: a peculiar G + C content of an endocytobiotic DNA. *J Mol Evol* **47**: 52-61.
- Hendrickson,H., and Lawrence,J.G. (2006) Selection for chromosome architecture in bacteria. *J Mol Evol* **62**: 615-629.
- Herbeck,J.T., Degnan,P.H., and Wernegreen,J.J. (2005) Nonhomogeneous model of sequence evolution indicates

independent origins of primary endosymbionts within the enterobacteriales (Gamma-Proteobacteria). *Mol Biol Evol* **22**: 520-532.

Herbeck,J.T., Funk,D.J., Degnan,P.H., and Wernegreen,J.J. (2003) A conservative test of genetic drift in the endosymbiotic bacterium *Buchnera*: slightly deleterious mutations in the chaperonin *groEL*. *Genetics* **165**: 1651-1660.

Herniou,E.A., Luque,T., Chen,X., Vlak,J.M., Winstanley,D., Cory,J.S., and O'Reilly,D.R. (2001) Use of whole genome sequence data to infer baculovirus phylogeny. *J Virol* **75**: 8117-8126.

Hey,J. (2006) On the failure of modern species concepts. *Trends Ecol Evol* **21**: 447-450.

Holmes,E.C., Urwin,R., and Maiden,M.C. (1999) The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*. *Mol Biol Evol* **16**: 741-749.

Horn,M., Collingro,A., Schmitz-Esser,S., Beier,C.L., Purkhold,U., Fartmann,B. *et al.* (2004) Illuminating the evolutionary history of *Chlamydiae*. *Science* **304**: 728-730.

Hughes,A.L., Ekollu,V., Friedman,R., and Rose,J.R. (2005) Gene family content-based phylogeny of prokaryotes: the effect of criteria for inferring homology. *Syst Biol* **54**: 268-276.

Huson,D.H., and Steel,M. (2004) Phylogenetic trees based on gene content. *Bioinformatics*: bth198.

Inagaki,Y., Susko,E., and Roger,A.J. (2006) Recombination between elongation factor 1 α genes from distantly related archaeal lineages. *Proc Natl Acad Sci U.S. A* **03**: 4528-4533.

International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**: 695-716.

- Itoh,T., Martin,W., and Nei,M. (2002) Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. *Proc Natl Acad Sci U.S. A.* **99**: 12944-12948.
- Jain,R., Rivera,M.C., and Lake,J.A. (1999) Horizontal gene transfer among genomes: The complexity hypothesis. *Proc Natl Acad Sci U.S. A.* **96**: 3801-3806.
- Jeffroy,O., Brinkmann,H., Delsuc,F., and Philippe,H. (2006) Phylogenomics: the beginning of incongruence? *Trends Genet* **22**: 225-231.
- Jeltsch,A. (2003) Maintenance of species identity and controlling speciation of bacteria: a new function for restriction/modification systems? *Gene* **317**: 13-16.
- Jensen,E.C., Schrader,H.S., Rieland,B., Thompson,T.L., Lee,K.W., Nickerson,K.W., and Kokjohn,T.A. (1998) Prevalence of broad-host-range lytic bacteriophages of *Sphaerotilus natans*, *Escherichia coli*, and *Pseudomonas aeruginosa*. *Appl Environ Microbiol* **64**: 575-580.
- Johnson,Z.I., Zinser,E.R., Coe,A., McNulty,N.P., Woodward,E.M., and Chisholm,S.W. (2006) Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311**: 1737-1740.
- Jones,D.T., Taylor,W.R., and Thornton,J.M. (1994) A mutation data matrix for transmembrane proteins. *Febs Letters* **339**: 269-275.
- Jordan,I.K., Rogozin,I.B., Wolf,Y.I., and Koonin,E.V. (2002) Essential genes are more evolutionarily conserved than are nonessential genes in Bacteria. *Genome Res* **12**: 962-968.
- Karlin,S. (2001) Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol* **9**: 335-343.
- Karlin,S., and Burge,C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* **11**: 283-290.
- Kelchner,S.A., and Thomas,M.A. (2007) Model use in phylogenetics: nine key questions. *Trends Ecol Evol* **22**: 87-94.

- Khachane,A.N., Timmis,K.N., and Martins dos Santos,V.A.P. (2007) Dynamics of Reductive Genome Evolution in Mitochondria and Obligate Intracellular Microbes. *Mol Biol Evol* **24**: 449-456.
- Kishino,H., and Hasegawa,M. (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J Mol Evol* **29**: 170-179.
- Kluge,A.G. (1989) A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Syst Zool* **38**: 7-25.
- Konstantinidis,K.T., and Tiedje,J.M. (2005) Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U.S. A.* **102**: 2567-2572.
- Koonin,E.V., and Galperin,M.Y. (1997) Prokaryotic genomes: the emerging paradigm of genome-based microbiology. *Curr Opin Genet Dev* **7**: 757-763.
- Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* **39**: 309-338.
- Koonin,E.V., Makarova,K.S., and Aravind,L. (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol* **55**: 709-742.
- Korbel,J.O., Snel,B., Huynen,M.A., and Bork,P. (2002) SHOT: a web server for the construction of genome phylogenies. *Trends Genet* **18**: 158-162.
- Korber,B. (2000) HIV signature and sequence variation analysis. In Computational analysis of HIV molecular sequences. Rodrigo,A.G., and Learn,G.H. (eds). Dordrecht, Netherlands: Kluwer Academic Publishers, pp. 55-72.
- Kumar,S., Tamura,K., and Nei,M. (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* **5**: 150-163.

Kunin,V., and Ouzounis,C.A. (2003) The balance of driving forces during genome evolution in prokaryotes. *Genome Res* **13**: 1589-1594.

Kunin,V., Goldovsky,L., Darzentas,N., and Ouzounis,C.A. (2005) The net of life: Reconstructing the microbial phylogenetic network. *Genome Res* **15**: 954-959.

Kurland,C.G., Canback,B., and Berg,O.G. (2003) Horizontal gene transfer: a critical view. *Proc Natl Acad Sci U.S. A* **100**: 9658-9662.

Lake,J.A., and Rivera,M.C. (2004) Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. *Mol Biol Evol* **21**: 681-690.

Lawrence,J.G. (2002) Gene transfer in bacteria: speciation without species? *Theor Popul Biol* **61**: 449-460.

Lawrence,J.G. (2005) Common themes in the genome strategies of pathogens. *Curr Opin Genet Dev* **15**: 584-588.

Lawrence,J.G., Hatfull,G.F., and Hendrix,R.W. (2002) Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. *J Bacteriol* **184**: 4891-4905.

Lawrence,J.G., and Hendrickson,H. (2003) Lateral gene transfer: when will adolescence end? *Mol Microbiol* **50**: 739-749.

Lawrence,J.G., and Hendrickson,H. (2004) Chromosome structure and constraints on lateral gene transfer. *Dynamical Genet* 319-336.

Lawrence,J.G., and Ochman,H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* **44**: 383-397.

Lawrence,J.G., and Roth,J.R. (1996) Selfish operons: Horizontal transfer may drive the evolution of gene clusters. *Genetics* **143**: 1843-1860.

Lawrence,J.G., and Ochman,H. (1998) Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci U.S. A* **95**: 9413-9417.

- Lawrence, J.G., and Ochman, H. (2002) Reconciling the many faces of lateral gene transfer. *Trends in Microbiol* **10**: 1-4.
- Lerat, E., Daubin, V., and Moran, N.A. (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the Gamma-Proteobacteria. *PLoS Biol* **1**: E19.
- Lerat, E., Daubin, V., Ochman, H., and Moran, N.A. (2005) Evolutionary origins of genomic repertoires in Bacteria. *PLoS Biology* **3**: e130.
- Levin, B.R. (1981) Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. *Genetics* **99**: 1-23.
- Linnaeus, C. (1758) *Systema naturae per regna tria naturae, secundum classes, ordines, genera species, cum characteribus, differentiis, synonymis, locis. Holmiae.*
- Linz, B., Balloux, F., Moodley, Y., Manica, A., Liu, H., Roumagnac, P. *et al.* (2007) An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* **445**: 915-918.
- Liu, X., Gutacker, M.M., Musser, J.M., and Fu, Y.X. (2006) Evidence for recombination in *Mycobacterium tuberculosis*. *J Bacteriol* **188**: 8169-8177.
- Lockhart, P.J., Steel, M.A., Hendy, M.D., and Penny, D. (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol Biol Evol* **11**: 605-612.
- Loomis, W.F., and Smith, D.W. (1990) Molecular phylogeny of *Dictyostelium discoideum* by protein sequence comparison. *Proc Natl Acad Sci U.S. A.* **87**: 9093-9097.
- Lucchini, S., Rowley, G., Goldberg, M.D., Hurd, D., Harrison, M., and Hinton, J.C.D. (2006) H-NS Mediates the silencing of laterally acquired genes in Bacteria. *PLoS Pathogens* **2**: e81.
- Lynch, M. (1996) Mutation accumulation in transfer RNAs: molecular evidence for Muller's ratchet in mitochondrial genomes. *Mol Biol Evol* **13**: 209-220.

- Lynch,M. (1997) Mutation accumulation in nuclear, organelle, and prokaryotic transfer RNA genes. *Mol Biol Evol* **14**: 914-925.
- Lynch,M., and Conery,J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151-1155.
- Maiden,M.C., Bygraves,J.A., Feil,E., Morelli,G., Russell,J.E., Urwin,R. *et al.* (1998) Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U.S. A.* **95**: 3140-3145.
- Maiden,M.C. (2006) Multilocus sequence typing of bacteria. *Annu Rev Microbiol* **60**: 561-588.
- Majewski,J., and Cohan,F.M. (1998) The effect of mismatch repair and heteroduplex formation on sexual isolation in *Bacillus*. *Genetics* **148**: 13-18.
- Mark,P., and Andrew,M. (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst Biolo* **53**: 571-581.
- Marri,P., Hao,W., and Golding,G.B. (2007) The role of laterally transferred genes in adaptive evolution. *BMC Evolutionary Biology* **7**: S8.
- Martin,W., and Herrmann,R.G. (1998) Gene transfer from organelles to the nucleus: how much, what happens, and why? *Plant Physiol* **118**: 9-17.
- Martins-Pinheiro,M., Galhardo,R.S., Aires,K.A., Lima-Bessa,K.M., and Menck,C.F.M. (2004) Different patterns of evolution for duplicated DNA repair genes in bacteria of the *Xanthomonadales* group. *BMC Evol Biol* **4**.
- Mau,B., Glasner,J., Darling,A., and Perna,N. (2006) Genome-wide detection and analysis of homologous recombination among sequenced strains of *Escherichia coli*. *Genome Biol* **7**: R44.
- Maurelli,A.T. (2007) Black holes, antivirulence genes, and gene inactivation in the evolution of bacterial pathogens. *FEMS Microbiol Lett* **267**: 1-8.

Maurelli,A.T., Fernandez,R.E., Bloch,C.A., Rode,C.K., and Fasano,A. (1998) "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.* **95**: 3943-3948.

Mayr,E. (1942) Systematics and the Origin of Species. Columbia Univ. Press, New York.

McCombie,R.L., Finkelstein,R.A., and Woods,D.E. (2006) Multilocus sequence typing of historical *Burkholderia pseudomallei* isolates collected in Southeast Asia from 1964 to 1967 provides insight into the epidemiology of melioidosis. *J Clin Microbiol* **44**: 2951-2962.

Medigue,C., Rouxel,T., Vigier,P., Henaut,A., and Danchin,A. (1991) Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J Mol Biol* **222**: 851-856.

Mes,T.H.M., Doeleman,M., Lodders,N., Nubel,U., and Stal,L.J. (2006) Selection on protein-coding genes of natural cyanobacterial populations. *Environ Microbiol* **8**: 1534-1543.

Mignot,T., Mock,M., Robichon,D., Landier,A., Lereclus,D., and Fouet,A. (2001) The incompatibility between the PlcR- and AtxA-controlled regulons may have selected a nonsense mutation in *Bacillus anthracis*. *Mol Microbiol* **42**: 1189-1198.

Mira,A., and Moran,N.A. (2002) Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic bacteria. *Microb Ecol* **44**: 137-143.

Mira,A., Ochman,H., and Moran,N.A. (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet* **17**: 589-596.

Mongodin,E.F., Nelson,K.E., Daugherty,S., DeBoy,R.T., Wister,J., Khouri,H. *et al.* (2005) The genome of *Salinibacter ruber*: Convergence and gene exchange among hyperhalophilic bacteria and archaea. *Proc Natl Acad Sci U.S. A.* **102**: 18147-18152.

Moore,R.A., Reckseidler-Zenteno,S., Kim,H., Nierman,W., Yu,Y., Tuanyok,A. *et al.* (2004) Contribution of gene loss to the

pathogenic evolution of *Burkholderia pseudomallei* and *Burkholderia mallei*. *Infect Immun* **72**: 4172-4187.

Moran,N.A., and Wernegreen,J.J. (2000) Lifestyle evolution in symbiotic bacteria: insights from genomics. *Trends Ecol Evol* **15**: 321-326.

Moran,N.A. (1996) Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. *Proc Natl Acad Sci U.S. A* **93**: 2873-2878.

Moran,N.A., and Plague,G.R. (2004) Genomic changes following host restriction in bacteria. *Curr Opin Genet Dev* **14**: 627-633.

Moreira,D., and Philippe,H. (2000) Molecular phylogeny: pitfalls and progress. *Int Microbiol* **3**: 9-16.

Moret,B.M., and Warnow,T. (2005) Advances in phylogeny reconstruction from gene order and content data. *Methods Enzymol* **395**: 673-700.

Moya,A., and Latorre,A. (2007) Lessons in evolution from genome reduction in endosymbionts. In Horizontal gene transfer in the evolution of pathogenesis. Hensel,N., and Schmidt,H. (eds). Cambridge Univ. Press.

Muller,H.J. (1964) The relation of recombination to mutational advance. *Mutat Res* **106**: 2-9.

Müller,M., and Martin,W. (1999) The genome of *Rickettsia prowazekii* and some thoughts on the origin of mitochondria and hydrogenosomes. *Bioessays* **21**: 377-381.

Mushegian,A.R., and Koonin,E.V. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U.S. A* **93**: 10268-10273.

Nakabachi,A., Yamashita,A., Toh,H., Ishikawa,H., Dunbar,H.E., Moran,N.A., and Hattori,M. (2006) The 160-Kilobase Genome of the Bacterial Endosymbiont *Carsonella*. *Science* **314**: 267.

Nakamura,Y., Itoh,T., Matsuda,H., and Gojobori,T. (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet* **36**: 760-766.

- Navarre,W.W., Porwollik,S., Wang,Y., McClelland,M., Rosen,H., Libby,S.J., and Fang,F.C. (2006) Selective silencing of foreign DNA with low GC content by the H-NS protein in *Salmonella*. *Science* **313**: 236-238.
- Nesbo,C.L., Dlutek,M., and Doolittle,W.F. (2006) Recombination in *Thermotoga*: Implications for species concepts and biogeography. *Genetics* **172**: 759-769.
- Nierman,W.C., DeShazer,D., Kim,H.S., Tettelin,H., Nelson,K.E., Feldblyum,T. *et al.* (2004) From the Cover: Structural flexibility in the *Burkholderia mallei* genome. *Proc Natl Acad Sci U.S. A.* **101**: 14246-14251.
- Nilsson,A.I., Koskiniemi,S., Eriksson,S., Kugelberg,E., Hinton,J.C.D., and Andersson,D.I. (2005) From The Cover: Bacterial genome size reduction by experimental evolution. *Proc Natl Acad Sci U.S. A.* **102**: 12112-12116.
- Nubel,U., Reissbrodt,R., Weller,A., Grunow,R., Porsch-Ozcurumez,M., Tomaso,H. *et al.* (2006) Population structure of *Francisella tularensis*. *J Bacteriol* **188**: 5319-5324.
- O'Brien,S.J., and Stanyon,R. (1999) Phylogenomics. Ancestral primate viewed. *Nature* **402**: 365-366.
- Ochman,H., Lawrence,J.G., and Groisman,E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**: 299-304.
- Ogata,H., La Scola,B., Audic,S., Renesto,P., Blanc,G., Robert,C. *et al.* (2006) Genome sequence of *Rickettsia bellii* illuminates the role of amoebae in gene exchanges between intracellular pathogens. *PLoS Genet* **2**: e76.
- Ogata,H., Renesto,P., Audic,S., Robert,C., Blanc,G., Fournier,P.E. *et al.* (2005) The genome sequence of *Rickettsia felis* identifies the first putative conjugative plasmid in an obligate intracellular Parasite. *PLoS Biol* **3**: e248.
- Omelchenko,M., Makarova,K., Wolf,Y., Rogozin,I., and Koonin,E. (2003) Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. *Genome Biol* **4**: R55.

- Pal,C., Papp,B., and Lercher,M.J. (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* **37**: 1372-1375.
- Parkhill,J., Sebaihia,M., Preston,A., Murphy,L.D., Thomson,N., Harris,D.E. *et al.* (2003) Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat Genet* **35**: 32-40.
- Perez-Brocal,V., Gil,R., Ramos,S., Lamelas,A., Postigo,M., Michelena,J.M. *et al.* (2006) A small microbial genome: The end of a long symbiotic relationship? *Science* **314**: 312-313.
- Perez-Losada,M., Browne,E.B., Madsen,A., Wirth,T., Viscidi,R.P., and Crandall,K.A. (2006) Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. *Infect genet evol* **6**: 97-112.
- Philippe,H., Snell,E.A., Baptiste,E., Lopez,P., Holland,P.W.H., and Casane,D. (2004) Phylogenomics of eukaryotes: Impact of missing data on large alignments. *Mol Biol Evol* **21**: 1740-1752.
- Phillips,M.J., Delsuc,F., and Penny,D. (2004) Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol* **21**: 1455-1458.
- Phillips,M.J., and Penny,D. (2003) The root of the mammalian tree inferred from whole mitochondrial genomes. *Mol Phylogenet Evol* **28**: 171-185.
- Posada,D., and Crandall,K.A. (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**: 817-818.
- Pupo,G.M., Karaolis,D.K., Lan,R., and Reeves,P.R. (1997) Evolutionary relationships among pathogenic and nonpathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and *mdh* sequence studies. *Infect Immun* **65**: 2685-2692.
- Pupo,G.M., Lan,R., and Reeves,P.R. (2000) Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci U.S.A.* **97**: 10567-10572.

- Ragan,M.A. (1992) Matrix representation in reconstructing phylogenetic-relationships among the eukaryotes. *Bio.systems* **28**: 47-55.
- Rispe,C., Delmotte,F., Van Ham,R.C.H.J., and Moya,A. (2004) Mutational and selective pressures on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids. *Genome Res* **14**: 44-53.
- Robinson,D.F., and Foulds,L.R. (1981) Comparison of Phylogenetic Trees. *Math Biosci* **53**: 131-147.
- Robinson-Rechavi,M., and Huchon,D. (2000) RRTree: Relative-Rate Tests between groups of sequences on a phylogenetic tree. *Bioinformatics* **16**: 296-297.
- Rocap,G., Larimer,F.W., Lamerdin,J., Malfatti,S., Chain,P., Ahlgren,N.A. *et al.* (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424**: 1042-1047.
- Rocha,E.P.C. (2004) The replication-related organization of bacterial genomes. *Microbiology* **150**: 1609-1627.
- Rocha,E.P.C., and Danchin,A. (2003b) Gene essentiality determines chromosome organisation in bacteria. *Nucl Acids Res* **31**: 6570-6577.
- Rocha,E.P.C., and Danchin,A. (2003a) Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat Genet* **34**: 377-378.
- Rokas,A., Williams,B.L., King,N., and Carroll,S.B. (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*: 798-804.
- Ronquist,F., and Huelsenbeck,J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572-1574.
- Roumagnac,P., Weill,F.X., Dolecek,C., Baker,S., Brisse,S., Chinh,N.T. *et al.* (2006) Evolutionary history of *Salmonella Typhi*. *Science* **314**: 1301-1304.

Rusch,D.B., Halpern,A.L., Sutton,G., Heidelberg,K.B., Williamson,S., Yooseph,S. *et al.* (2007) The Sorcerer II global ocean sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biology* **5**: e77.

Sanderson,M.J., and Shaffer,H.B. (2002) Troubleshooting molecular phylogenetic analyses. *Annu Rev Ecol Syst* **33**: 49-72.

Sanderson,M.J., and Driskell,A.C. (2003) The challenge of constructing large phylogenetic trees. *Trends in Plant Science* **8**: 374-379.

Scally,M., Schuenzel,E.L., Stouthamer,R., and Nunney,L. (2005) Multilocus sequence type system for the plant pathogen *Xylella fastidiosa* and relative contributions of recombination and point mutation to clonal diversity. *Appl Environ Microbiol* **71**: 8491-8499.

Schmidt,H.A., Strimmer,K., Vingron,M., and von Haeseler,A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**: 502-504.

Schroder,D., Deppisch,H., Obermayer,M., Krohne,G., Stackebrandt,E., Holldobler,B. *et al.* (1996) Intracellular endosymbiotic bacteria of *Camponotus* species (carpenter ants): systematics, evolution and ultrastructural characterization. *Mol Microbiol* **21**: 479-489.

Selander,R.K., Caugant,D.A., Ochman,H., Musser,J.M., Gilmour,M.N., and Whittam,T.S. (1986) Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Appl Environ Microbiol* **51**: 873-884.

Shimodaira,H., and Hasegawa,M. (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol* **16**: 1114-1116.

Sicheritz-Potén,T., and Andersson,S.G.E. (2001) A phylogenomic approach to microbial evolution. *Nucleic Acids Res* **29**: 545-552.

Silva,F.J., Latorre,A., and Moya,A. (2001) Genome size reduction through multiple events of gene disintegration in *Buchnera* APS. *Trends Genet* **17**: 615-618.

- Singer, G.A.C., and Hickey, D.A. (2000) Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol Biol Evol* **17**: 1581-1588.
- Smith, J.M., Dowson, C.G., and Spratt, B.G. (1991) Localized sex in bacteria. *Nature* **349**: 29-31.
- Smith, J.M., Smith, N.H., O'Rourke, M., and Spratt, B.G. (1993) How clonal are Bacteria? *Proc Natl Acad Sci U.S. A.* **90**: 4384-4388.
- Snel, B., Bork, P., and Huynen, M.A. (1999) Genome phylogeny based on gene content. *Nat Genet* **21**: 108-110.
- Snel, B., Huynen, M.A., and Dutilh, B.E. (2005) Genome trees and the nature of genome evolution. *Annu Rev Microbiol* **59**: 191-209.
- Strimmer, K., and Rambaut, A. (2002) Inferring confidence sets of possibly misspecified gene trees. *Proceedings of the Royal Society of London Series B-Biological Sciences* **269**: 137-142.
- Studholme, D.J., Downie, J.A., and Preston, G.M. (2005) Protein domains and architectural innovation in plant-associated Proteobacteria. *BMC Genomics* **6**: 17.
- Sueoka, N. (1988) Directional mutation pressure and neutral molecular evolution. *Proc Natl Acad Sci U.S. A.* **85**: 2653-2657.
- Sueoka, N. (1992) Directional mutation pressure, selective constraints, and genetic equilibria. *J Mol Evol* **34**: 95-114.
- Sueoka, N. (1993) Directional mutation pressure, mutator mutations, and dynamics of molecular evolution. *J Mol Evol* **37**: 137-153.
- Susko, E., Leigh, J., Doolittle, W.F., and Baptiste, E. (2006) Visualizing and assessing phylogenetic congruence of core gene sets: a case study of the Gamma-Proteobacteria. *Mol Biol Evol* **23**: 1019-1030.
- Swofford, D.L. (1998) PAUP: Phylogenetic analysis using parsimony (and other methods), Version 4. *Sinauer Associates, Sunderland MA.*

- Tamas,I, Klasson,L., Canback,B., Naslund,A.K., Eriksson,A.S., Wernegreen,J.J. *et al.* (2002) 50 million years of genomic stasis in endosymbiotic bacteria. *Science* **296**: 2376-2379.
- Tamura,K., and Kumar,S. (2002) Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Mol Biol Evol***19**: 1727-1736.
- Tatusov,R.L., Galperin,M.Y., Natale,D.A., and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**: 33-36.
- Tatusov,R., Fedorova,N., Jackson,J., Jacobs,A., Kiryutin,B., Koonin,E. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinfo* **4**: 41.
- Tettelin,H., Massignani,V., Cieslewicz,M.J., Donati,C., Medini,D., Ward,N.L. *et al.* (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome". *PNAS* **102**: 13950-13955.
- Thao,M.L., Clark,M.A., Baumann,L., Brennan,E.B., Moran,N.A., and Baumann,P. (2000a) Secondary endosymbionts of psyllids have been acquired multiple times. *Curr Microbiol* **41**: 300-304.
- Thao,M.L., Moran,N.A., Abbot,P., Brennan,E.B., Burckhardt,D.H., and Baumann,P. (2000b) Cospeciation of psyllids and their primary prokaryotic endosymbionts. *Appl Environ Microbiol* **66**: 2898-2905.
- Thao,M.L., and Baumann,P. (2004) Evolutionary relationships of primary prokaryotic endosymbionts of whiteflies and their hosts. *Appl Environ Microbiol* **70**: 3401-3406.
- Thomas,C.M., and Nielsen,K.M. (2005) MECHANISMS OF, AND BARRIERS TO, HORIZONTAL GENE TRANSFER BETWEEN BACTERIA. *Nat Rev Micro* **3**: 711-721.
- Thompson,J.D., Higgins,D.G., and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl Acids Res* **22**: 4673-4680.

- Thompson,J.R., Pacocha,S., Pharino,C., Klepac-Ceraj,V., Hunt,D.E., Benoit,J. *et al.* (2005) Genotypic diversity within a natural coastal bacterioplankton population. *Science* **307**: 1311-1313.
- Toh,H., Weiss,B.L., Perkin,S.A.H., Yamashita,A., Oshima,K., Hattori,M., and Aksoy,S. (2006) Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. *Genome Res* **16**: 149-156.
- Tringe,S.G., von Mering,C., Kobayashi,A., Salamov,A.A., Chen,K., Chang,H.W. *et al.* (2005) Comparative metagenomics of microbial communities. *Science* **308**: 554-557.
- Tyson,G.W., Chapman,J., Hugenholtz,P., Allen,E.E., Ram,R.J., Richardson,P.M. *et al.* (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37-43.
- Uchiyama,I. (2003) MBGD: microbial genome database for comparative analysis. *Nucl Acids Res* **31**: 58-62.
- Van Ham,R.C.H.J., Gonzalez-Candelas,F., Silva,F.J., Sabater,B., Moya,A., and Latorre,A. (2000) Postsymbiotic plasmid acquisition and evolution of the repA1-replicon in *Buchnera aphidicola*. *Proc Natl Acad Sci U.S. A* **97**: 10855-10860.
- Van Sluys,M.A., Monteiro-Vitorello,C.B., Camargo,L.E.A., Menck,C.F.M., da Silva,A.C.R., Ferro,J.A. *et al.* (2002) COMPARATIVE GENOMIC ANALYSIS OF PLANT-ASSOCIATED BACTERIA. *Ann Rev Phytop* **40**: 169-189.
- Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso sea. *Science* **304**: 66-74.
- Vissa,V.D., and Brennan,P.J. (2001) The genome of *Mycobacterium leprae*: a minimal mycobacterial gene set. *Genome Biol* **2**: REVIEWS1023.
- Waldron,D.E., and Lindsay,J.A. (2006) *SauI*: a novel lineage-specific Type I restriction-modification system that blocks horizontal gene transfer into *Staphylococcus aureus* and between *S. aureus* isolates of different lineages. *J Bacteriol* **188**: 5578-5585.

- Wernegreen,J.J. (2002) Genome evolution in bacterial endosymbionts of insects. *Nat Rev Genet* **3**: 850-861.
- Wernegreen,J.J. (2005) For better or worse: genomic consequences of intracellular mutualism and parasitism. *Curr Opin Genet Dev* **15**: 572-583.
- Wirth,T., Falush,D., Lan,R., Colles,F., Mensa,P., Wieler,L.H. *et al.* (2006) Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Molecular Microbiology* **60**: 1136-1151.
- Woese,C.R. (1987) Bacterial evolution. *Microbiological Reviews* **51**: 221-271.
- Woese,C.R., and Fox,G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U.S.A.* **74**: 5088-5090.
- Woese,C.R., Kandler,O., and Wheelis,M.L. (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci U.S.A.* **87**: 4576-4579.
- Wolf,Y.I., Rogozin,I.B., Kondrashov,A.S., and Koonin,E.V. (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res* **11**: 356-372.
- Wu,D., Daugherty,S.C., Van Aken,S.E., Pai,G.H., Watkins,K.L., Khouri,H. *et al.* (2006) Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. *PLoS Biology* **4**.
- Yang,Z., Nielsen,R., Goldman,N., and Pedersen,A.M.K. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**: 431-449.
- Yang,Z., Wong,W.S.W., and Nielsen,R. (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol* **22**: 1107-1118.
- Yap,W.H., Zhang,Z., and Wang,Y. (1999) Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. *J Bacteriol* **181**: 5201-5209.

Zeyl,C., Mizesko,M., and de Visser,J.A. (2001) Mutational meltdown in laboratory yeast populations. *Evolution Int J Org Evolution* **55**: 909-917.

Zhang,J., Nielsen,R., and Yang,Z. (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* **22**: 2472-2479.

Zhaxybayeva,O., Gogarten,J.P., Charlebois,R.L., Doolittle,W.F., and Papke,R.T. (2006) Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events. *Genome Res* **16**: 1099-1108.

Zuckerandl,E., and Pauling,L. (1965) Molecules as documents of evolutionary history. *J Theor Biol* **8**: 357-366.

10. BREVE RESUMEN EN CASTELLANO

CAPÍTULO 1: Introducción general

En esta tesis vamos analizar varios aspectos relacionados con la evolución de los genomas bacterianos. Si hay un grupo de organismos que se ha beneficiado especialmente de las técnicas de secuenciación genómica cada vez más baratas y rápidas, ese ha sido el de los procariontes. En las primeras etapas de secuenciación iniciadas a partir de la publicación del primer genoma bacteriano completo, *Haemophilus influenzae* (Fleischmann et al., 1995), la selección del genoma a secuenciar debía ser muy cuidadosa no solo por el coste sino también por el esfuerzo necesario. Esto hizo que lo que podríamos llamar genomas carismáticos fueran los primeros en ser secuenciados. Entre dichos genomas se encuentran los de organismos modelo como *Escherichia coli* K12 o *Bacillus subtilis* o cepas patogénicas de agentes causantes de enfermedades de preocupación mundial tales como *Mycoplasma pneumoniae* M129, *Helicobacter pylori* o *Borrelia burgdorferi*. Sin olvidar, la secuenciación de los primeros genomas de Arquea que representaron un hito para el estudio de su biología pues no habían podido ser cultivados en laboratorio y sólo se habían identificado indirectamente a través de la secuenciación y análisis estructural de su 16s rDNA.

Ahora que las etapas iniciales de la revolución genómica se han consumido, una nueva etapa comienza en el estudio de los genomas bacterianos (Fraser-Liggett, 2005). Ya no sólo es importante caracterizar los genes que los componen o su organización, nuevos conceptos se están abriendo paso. Con la más fácil secuenciación de genomas se inicia la era de la genómica poblacional, la que es capaz de caracterizar poblaciones y sus

individuos en base a características genómicas como la presencia/ausencia de genes o las variaciones de parámetros de diversidad nucleotídica a lo largo del genoma. Nuevos conceptos se han introducido para esta nueva etapa. El pangenoma de una especie, el conjunto de genes que conforman todos los genomas de esa especie (Tettelin *et al.*, 2005), ha revelado la presencia de genes comunes compartidos por todas las cepas y conjuntos de genes dispensables, presentes en solo alguna/s de ellas. El análisis de múltiples cepas de microorganismos de diferentes grupos taxonómicos ha permitido determinar o predecir su pangenoma con casos que van desde aquellos que son cerrados, en los que la adición de nuevos genomas no es probable que dé lugar a un aumento de genes nuevos a aquellos que son abiertos en los que dicha adición podría continuar casi hasta el infinito. Los mismos avances permiten por otra parte contextualizar los genomas estudiados en un ambiente concreto, pudiéndose estudiar lo que se conoce como el metagenoma ambiental (Tringe *et al.*, 2005). La metagenómica, también conocida como la genómica ambiental o de comunidades, permite revelar la biodiversidad tanto a nivel a taxonómico como de funciones génicas que se encuentran en un determinado ambiente. Su utilización ha sido particularmente importante para sondear la gran cantidad de microorganismos no cultivables en laboratorio y que representan de largo la mayor fracción de especies microbianas.

Sin embargo, a pesar de la gran cantidad de temas que todos estos estudios permiten tratar, desde la bioquímica, a las consecuencias biotecnológicas de dichos descubrimientos, es la evolución microbiana y su naturaleza la que parece atraer el mayor

foco de atención. Los genes son “documentos de la historia evolutiva” (Zuckerlandl and Pauling, 1965) pero en bacterias más que en ningún otro grupo, son documentos de su propia historia evolutiva. La evolución en bacterias va más allá de las clásicas migración, mutación y deriva o incluso recombinación. La transferencia horizontal de genes entre genomas que pueden o no estar cercanos evolutivamente parece ser uno de los principales motores del cambio en bacterias (Ochman *et al.*, 2000). Las consecuencias de estos procesos dejan huella en los genomas procariontes. Dicha huella se puede rastrear en su mayor parte a través de las técnicas filogenéticas. De tal manera que en los genomas procariontes conviven diferentes señales filogenéticas (Figura 4) con el ruido filogenético. La diferenciación entre ruido señal es lo que se encuentra en el centro de la polémica.

Particularmente, tres son los campos que se van a tratar en esta introducción: el análisis filogenómico, la ganancia de genes a través de transferencia horizontal y la pérdida génica a través de la reducción genómica.

Análisis filogenómico

La secuenciación completa de genomas ha llevado a dar un salto cualitativo en la manera en que las relaciones filogenéticas entre especies, o más generalmente entre grupos de secuencias, son inferidas. Concretamente en el campo de la microbiología el marcador filogenético universal ha sido la subunidad pequeña del 16s ribosomal (16s rDNA). La facilidad en que las secuencias son obtenidas y su presencia universal en los genomas bacterianos junto con sus lentas tasas de evolución le convierten en un buen

marcador filogenético. Sin embargo, la secuenciación masiva de genomas y el desarrollo de técnicas de análisis de múltiples loci (MLST) están desplazándolo a favor de la inferencia filogenética a nivel genómico.

La aproximación filogenómica, inicialmente propuesta como el uso de técnicas filogenéticas para ayudar en las anotaciones genómicas (Eisen, 1998), se ha transformado en el que posiblemente sea el modo más apropiado de descifrar la historia filogenética de los organismos a través del análisis filogenético a escala genómica (O'Brien and Stanyon, 1999; Sicheritz-Potén and Andersson, 2001; Eisen and Fraser, 2003). Este salto de la filogenética a la filogenómica ha permitido evitar algunos de los problemas relacionados con la inferencia de árboles de especies a partir de árboles génicos. Sin embargo, a su vez ha generado nuevos problemas en la reconstrucción de árboles de especies e incluso cuestionado la existencia de un árbol único capaz de reconciliar todas las historias evolutivas individuales presentes en los genomas bacterianos (Doolittle, 1999; Baptiste *et al.*, 2005; Susko *et al.*, 2006).

Las aproximaciones basadas en árboles únicos alcanzaron un máximo de popularidad con el uso del RNA ribosomal como marcador molecular (Woese, 1987). Estos genes han sido, y todavía lo son, una herramienta útil en el análisis filogenético bacteriano. Sus propiedades como buenos marcadores para la inferencia filogenética, tales como su presencia universal y su conservación evolutiva, ha permitido la propuesta de un árbol de la vida universal (Woese and Fox, 1977; Woese *et al.*, 1990) y la clasificación y reconstrucción de las relaciones evolutivas dentro y

entre los tres dominios en ausencia de datos genómicos (Woese and Fox, 1977). Sin embargo, incluso los marcadores ribosomales no son inmunes a la transferencia horizontal (Asai *et al.*, 1999; Yap *et al.*, 1999).

Las ventajas de una aproximación basada en múltiples genes frente a aquellas basadas en genes únicos son evidentes *a priori*. En teoría, permiten evadir las historias evolutivas individuales en favor de la historia común; permiten por otra parte evitar problemas tales como un insuficiente tamaño de la secuencia gracias a la adición de nuevos sitios pertenecientes a múltiples genes o compensar casos particulares de composición de bases sesgada. En la práctica, algunos de estos problemas también pueden afectar a filogenias obtenidas a partir de grandes conjuntos de datos, aunque en otros casos sí pueden ser reducidos considerablemente o fácilmente diagnosticados. Diferentes alternativas basadas en datos extraídos de genomas completos han sido propuestas para sustituir el análisis tradicional basado en un solo gen (revisado en Delsuc *et al.*, 2005) (Delsuc *et al.*, 2005). Los caracteres genómicos más comúnmente usados por estos métodos son el contenido génico (Snel *et al.*, 1999; Gu and Zhang, 2004; Huson and Steel, 2004), el orden génico (Wolf *et al.*, 2001; Korbelt *et al.*, 2002; Belda *et al.*, 2005), secuencias concatenadas también llamadas supermatrices (Brown *et al.*, 2001; Rokas *et al.*, 2003) o técnicas basadas en múltiples árboles génicos (Bininda-Emonds *et al.*, 2002; Bininda-Emonds, 2004). Los primeros dos tipos de caracteres han sido tradicionalmente analizados por aproximaciones. Los últimos dos se basan directamente o

indirectamente en la secuencia genómica y son los métodos filogenómicos que hemos seleccionado para este capítulo.

Transferencia génica horizontal

La evolución del contenido génico de los genomas bacterianos está fuertemente influenciada por su habilidad para incorporar DNA de otras especies en un proceso conocido como transferencia génica horizontal (HGT) (Koonin *et al.*, 2001). El estudio de los eventos de transferencia horizontal ha cambiado del análisis de casos individuales a análisis de escala genómica gracias al número creciente de genomas microbianos secuenciados (Koonin and Galperin, 1997; Koonin *et al.*, 2001).

En principio, es de esperar que la gran mayoría de genes en bacterias pertenezcan a la categoría vertical (Kunin and Ouzounis, 2003; Beiko *et al.*, 2005). El genoma completo es heredado de manera vertical, sin embargo una parte de esos genes verticales como consecuencia de los procesos evolutivos pueden ser perdidos mientras que otra parte puede ser transferida horizontalmente a especies no emparentadas. De hecho, parece que las innovaciones más importantes se suelen adquirir como resultado de eventos de transferencia horizontal (Ochman *et al.*, 2000) y, en menor grado, de duplicaciones (Gevers *et al.*, 2004). La fracción exacta de genes pertenecientes a cada categoría es variable según grupos, o incluso especies, y difícil de determinar. Hay desacuerdo sobre hasta qué punto los procesos no verticales, sobretodo la transferencia horizontal, influyen en la inferencia de

filogenias genómicas así como en la existencia de un único árbol de especies para bacterias.

La disponibilidad de un gran número de genomas microbianos ha permitido la construcción de filogenias genómicas usando diferente tipos de información (Snel *et al.*, 2005), con datos de secuencias, árboles génicos, contenido génico compartido, y orden génico compartido como los más ampliamente utilizados. Típicamente, el impacto de la transferencia horizontal en estas filogenias genómicas ha sido obviado y considerado como puro ruido filogenético a favor de una señal vertical resultante de la transmisión de información de ancestros a descendientes. Sin embargo, diferentes autores (Doolittle, 1999; Gogarten *et al.*, 2002; Kunin *et al.*, 2005) han propuesto que la obtención de un único árbol de la vida para bacterias es imposible, puesto que 1) todo gen ha sido transferido al menos una vez durante su historia evolutiva, y por lo tanto 2) la señal filogenética asociada a las transferencias horizontales se opone, y muchas veces sobrepasa, la señal vertical, por tanto oscureciendo las relaciones filogenéticas más profundas que existen entre los actuales genomas. Si la tasa de transferencia horizontal es alta, entonces una filogenia que se basa en las relaciones ancestro-descendiente no será capaz de reflejar la evolución de los genomas bacterianos que podría ser descrita mejor como una red filogenética (Doolittle, 1999). Sin embargo, si esa tasa es lo suficientemente baja entonces deberíamos ser capaces de reflejar la evolución bacteriana como un árbol y no como una red (Kurland *et al.*, 2003).

Reducción genómica

Si la ganancia génica a través de la transferencia horizontal es un factor principal en la evolución de los genomas bacterianos permitiéndoles colonizar múltiples nichos, las pérdidas génicas les han permitido refinar su adaptación a esos nuevos nichos. Entre las bacterias actuales los casos mejor estudiados de pérdidas génicas masivas son los genomas de los endosimbiontes de insectos de la subdivisión Gamma-Proteobacteria (Wernegreen, 2002). Todos ellos se caracterizan por presentar una serie de características evolutivas muy diferentes a las del resto de bacterias. Este conjunto de propiedades es conocido como el “síndrome del residente” y es el resultado de la transición de un estilo de vida patogénico/libre a uno endosimbiótico (Andersson and Kurland, 1998). La residencia en el interior de la célula, generalmente en un bacteriocito, les provee de una ambiente estable que hace que muchos de los genes no necesario para la endosimbiosis o redundantes con las funciones provistas por el hospedador se encuentren bajo débil selección y se convierten en candidatos a pseudogenes y posterior eliminación en las primeras etapas de la endosimbiosis. Este hecho, junto con la pérdida en la capacidad de aceptar material génico externo, lleva a una continua reducción del tamaño genómico.

Desde el punto de vista de la genética de poblaciones, los endosimbiontes se heredan maternalmente, lo que impone un cuello de botella durante su transmisión. Este hecho junto con sus presumibles, pequeños tamaños poblacionales y la ausencia de recombinación tiene como resultado la acción de un proceso conocido como trinquete de Muller (Moran, 1996). La acumulación de mutaciones ligeramente desventajosas que no se

pueden purgar ni por recombinación ni por selección purificadora. Altas tasas de mutación debido a la pérdida de genes relacionados con los sistemas de reparación también han sido propuestas como mecanismo responsable de la acumulación de dichas mutaciones desventajosas (Itoh *et al.*, 2002).

Lo que es cierto es que el “síndrome del residente” se caracteriza por una alta tasa de mutación, la acumulación de mutaciones desventajosas por deriva genética pero también por un creciente sesgo hacia mayores contenidos en A+T y la pérdida de la optimización en el uso de codones. Desde un punto de vista del análisis filogenómico y evolutivo en general esto representa un problema porque aumenta la probabilidad de convergencias debido a la composición o a las altas tasas de mutación independientemente de que tengan o no un origen común.

CAPÍTULO 2: Objetivos

En esta tesis vamos a tratar:

¿Cómo afecta la revolución genómica a los métodos de reconstrucción filogenómica?

¿Cuáles son las relaciones evolutivas entre los endosimbiontes de insectos del la división Gamma-Proteobacteria?

Caracterizar el proceso de transferencia horizontal en un grupo concreto de genomas, las Xanthomonadales y cómo éste ha variado a lo largo del tiempo.

Caracterizar las fuerzas evolutivas que actúan en las últimas etapas del proceso de reducción genómica en endosimbiontes bacterianos mediante el análisis de los genomas de *Carsonella ruddii* y *Buchnera aphidicola* *Cinara cedri*.

CAPÍTULOS 3 Y 4: Filogenómica y la relación filogenética de los endosimbiontes de insectos

Los genomas bacterianos mantienen diferentes señales evolutivas como resultado de diferentes procesos evolutivos que actúan sobre ellos. Como consecuencia, la información codificada en sus genomas puede ser dividida en tres categorías principales: señales verticales, señales no-verticales y ruido filogenético. La reconstrucción de la historia evolutiva de las bacterias así como el análisis de las diferentes fuerzas evolutivas que sobre ellas actúan y que han definido sus genomas depende de nuestra capacidad de desentrañar estas señales.

La señal vertical está asociada a la transmisión de información genética de ancestros a descendientes. Desde una perspectiva de la genómica comparada, esta señal reside en el conjunto de ortólogos verdaderos compartidos por los genomas microbianos. La señal no-vertical aparece como resultado de un proceso que no implica al ancestro inmediatamente anterior como proveedor del material genético. Los dos procesos más comunes a escala genómica que originan este tipo de señal son las duplicaciones y la transferencia génica horizontal. Parálogos son aquellos genes provenientes de un proceso de duplicación. Después de su origen, los parálogos pueden tener diferentes

destinos desde la neo o sub-funcionalización a la extinción a través de la desintegración génica (Lynch and Conery, 2000). Xenólogos son genes transmitidos horizontalmente desde un genoma distinto del ancestro del genoma recipiente (Gogarten and Olenzenski, 1999; Koonin, 2005). La existencia de transferencia horizontal entre los microorganismos es un fenómeno conocido hace tiempo (Davies, 1996) y actualmente es reconocida como uno de los procesos principales que influyen la evolución de las bacterias (Lawrence, 2002; Gogarten and Townsend, 2005). Ortólogo, xenólogo y parálogo son términos usados comúnmente, sin embargo otros términos han sido propuestos con el objetivo de delimitar mejor la clasificación de los genes según su origen evolutivo, sobre todo cuando aparecen casos en los que los límites que define la terminología tradicional son difusos. Así, un sinólogo define un gen que existe como más de una copia en el genoma, a diferencia de un parálogo, las copias pueden ser debidas tanto a duplicaciones como a transferencias horizontales (Lerat *et al.*, 2005). Por último, el ruido filogenético puede tener diferentes fuentes y se corresponde con casos en los que la señal filogenética es insuficiente o los patrones evolutivos son tan complejos que limitan las posibilidades de una buena inferencia filogenética (Gribaldo and Philippe, 2002).

Del conjunto de aproximaciones filogenómicas descritas en la introducción general, en este capítulo nos vamos a centrar principalmente en dos: supermatrices (o concatenados) y superárboles. Como ha sido previamente mencionado, ambas se basan en el análisis filogenético de la secuencia genómica aunque

la relación de los superárboles con ésta es indirecta puesto que éstos se derivan de un conjunto de árboles génicos.

El uso de concatenados cada vez es más común, incluso usando secuencias genómicas parciales o datos provenientes de análisis de tipado basado en secuenciación multilocus (Baptiste *et al.*, 2002; Rokas *et al.*, 2003). El método ha sido elegido en casos donde la señal filogenética individual era muy pobre (Herniou *et al.*, 2001), cuando hay heterogeneidad en las tasas de evolución (Baptiste *et al.*, 2002; Gontcharov *et al.*, 2004) o cuando es de esperar la influencia de procesos no verticales tales como duplicaciones escondidas o transferencia génica horizontal (Brochier *et al.*, 2002). El método incrementa la señal filogenética mediante la unión de secuencias provenientes de múltiples genes, por lo tanto creando una supermatrix de caracteres, y generalmente recupera aceptables (y presumiblemente correctas) filogenias con valores de soporte de los nodos muy altos pudiendo evitar algunas de los fallos comentados en el párrafo anterior aunque no siempre es así (Phillips *et al.*, 2004).

Al contrario que en el uso de concatenados, las aproximaciones basadas en árboles consensos y superárboles tienen una asociación indirecta con la secuencia genómica. Los árboles consenso se basan en la integración de múltiples árboles iniciales en una única topología. Cuando el conjunto de datos inicial incluye árboles con diferente número de especies representadas, entonces es posible construir un superárbol combinando las regiones topológicas solapantes. A pesar de que todos estos métodos han sido utilizados con éxito para resolver diferentes problemas filogenéticos, también tienen diferentes tipos

de errores asociados. Por ejemplo, la potencia del uso de supermatrices decrece conforme la distancia entre los taxones considerados aumenta pues el número de secuencias compartidas decrece, también es posible que gran parte de los genes usados en el concatenado estén afectados por fenómenos como el de la transferencia horizontal o paralógias escondidas (Daubin *et al.*, 2002). Por otra parte, los superárboles tienen sus propios problemas asociados. Por ejemplo, el uso de árboles erróneos o afectados por un mismo sesgo sistemático, la relación indirecta con las secuencias moleculares, una marcada heterogeneidad en el número de taxones entre los árboles génicos o la falta de una metodología universal para evaluar el soporte de los nodos del superárbol se encuentran entre los problemas más comunes (Gatesy *et al.*, 2002; Creevey, 2004; 2006).

Por otra parte, no solo la metodología de inferencia filogenética es importante en filogenómica si no también el conjunto de datos a la que es aplicada. La naturaleza de los genes que componen el conjunto de datos a ser analizado puede tener un incidencia directa en la filogenia recuperada así como en las señales filogenéticas que se encontrarán (Gophna *et al.*, 2005). Diferentes subconjuntos génicos pueden ser derivados de un genoma, dependiendo de las características de los genes que ese subconjunto contenga se pueden derivar conclusiones muy distintas. Así, el término “genoma mínimo” ha sido usado para describir el número mínimo de genes necesarios para que una célula pueda subsistir (Mushegian and Koonin, 1996). No hay un único genoma mínimo por lo que ha habido diferentes propuestas a lo largo de los años (Mushegian and Koonin, 1996; Glass *et al.*,

2006). Sin embargo, una de las revisiones más recientes que intenta sintetizar el resultado de diferentes aproximaciones ha propuesto una síntesis de 206 genes como el genoma mínimo necesario para la vida celular (Gil *et al.*, 2004). Es de esperar que estos genes, muchos de ellos caracterizados por su esencialidad y su papel central en las redes metabólicas, codifiquen una buena señal filogenética de acuerdo con la hipótesis de la complejidad (Jain *et al.*, 1999; Jordan *et al.*, 2002).

En cualquier caso, la esencialidad no es el único factor que puede influir en la presencia de una mayor o menor cantidad de señal vertical en los genes analizados. Es también importante que esos genes se distribuyan por todos los taxones analizados debido a restricciones en la aplicación de algunos métodos filogenómicos. En consecuencia, es importante distinguir entre aquellos conjuntos de genes universales, presentes en todas las bacterias analizadas de aquellos que no lo son. Por lo que la universalidad de los conjuntos es otro factor a tener en cuenta en el análisis de la señal vertical en los genomas bacterianos.

En este capítulo hemos aplicado las técnicas filogenómicas con dos objetivos. El análisis de las relaciones evolutivas de los genomas analizados y el análisis del comportamiento de los métodos filogenómicos bajo la influencia de diferentes factores. Sobre la primera, existe un debate sobre si los endosimbiontes bacterianos de insectos pertenecientes al las Gamma-Proteobacteria forman un clado monofilético (Gil *et al.*, 2003; Lerat *et al.*, 2003; Canback *et al.*, 2004) o bien son parafiléticos y su agrupamiento es resultado de artefactos en la reconstrucción filogenética (Charles *et al.*, 2001; Herbeck *et al.*,

2004; Belda *et al.*, 2005). El presente estudio analiza las relaciones evolutivas de cinco endosimbiontes de Gamma-Proteobacteria incluyendo tres cepas de *Buchnera aphidicola*. Varias de las características evolutivas de estos genomas, como por ejemplo sus altas tasas de evolución, bajo contenido en G+C así como el procesos de desintegración genómica, acarrear diferentes problemas metodológicos (Moreira and Philippe, 2000; Sanderson and Shaffer, 2002). Sobre la segunda, nos hemos interesado por la relación entre supermatrices/superárboles, la reconstrucción filogenética en relación con la función de los genes, las señales filogenéticas contenidas en diferentes subconjuntos de genes de un genoma y la aplicación de diferentes metodologías para corregir el efecto sobre las filogenias del fuerte sesgo en A+T en endosimbiontes.

Resultados y discusión

La aproximación filogenómica y la exploración del paisaje adaptativo de las Gamma-Proteobacteria

Esta parte del capítulo toma como base la determinación del filoma (Sicheritz-Potén and Andersson, 2001) – el conjunto de árboles filogenéticos de cada gen codificante de proteínas en un genoma – de *Blochmannia floridanus* (Gil *et al.*, 2003) el endosimbionte primario de las hormigas carpinteras. Hemos usado su genoma como punto de inicio para explorar el paisaje filogenético de las Gamma-Proteobacteria. Para ello, hemos usado un conjunto de técnicas filogenéticas y filogenómicas con el objetivo de intentar inferir las relaciones evolutivas entre los 21 taxones estudiados que incluyen como endosimbiontes, además de

a *Blochmannia*, a tres cepas de *Buchnera aphidicola* (pulgones) y el genoma de *Wigglesworthia brevialpis* (mosca tse-tse). En consecuencia, hemos analizado varias hipótesis filogenéticas para estas especies a través del examen hecho a las diferentes filogenias derivadas del conjunto de genes codificantes de proteínas de *Blochmannia floridanus* y su comparación con topologías obtenidas del 16s rDNA así como de diferentes métodos filogenómicos.

El genoma de *Blochmannia floridanus* se compone de 579 genes codificantes, el primer paso dado en nuestro estudio fue determinar la presencia de genes ortólogos en el resto de genomas analizados. La determinación de un conjunto real de genes es difícil puesto que depende en la mayoría de ocasiones de determinar previamente las relaciones evolutivas entre los taxones considerados. Nos encontramos por lo tanto con la paradoja de que para obtener un conjunto de ortólogos fiables con los que obtener una filogenia viable es necesario previamente obtener esa filogenia. Para resolver dicho problema, nosotros hemos llevado a cabo una aproximación distinta. Hemos derivado un árbol guía a partir de homólogos fácilmente identificables (genes ribosomales) que nos ha permitido hacer una búsqueda de ortólogos putativos específica para cada clado identificado. Consideramos estos ortólogos como putativos pues sólo sabremos si son verdaderos ortólogos después del análisis filogenómico realizado. Usamos BLAST como primer criterio de búsqueda del conjunto de ortólogos putativos. Sin embargo, para aumentar la señal dividimos los genomas según el árbol guía en nueve grupos de BLAST, el primer hit contra cualquier miembro del grupo fue el utilizado para buscar en el resto de miembros del grupo. La

información de BLAST, junto con los alineamientos y el árbol génico derivados de ellos así como la anotación de los genes nos permitió clasificar cada uno de los genes encontrados como ortólogo putativo o bien descartarlo como ortólogo mal identificado. Este procedimiento identificó un conjunto de genes comunes a los 21 genomas analizados (200-genes) y otro de 379 con desigual distribución entre dichos genomas (379-genes).

La primera aproximación llevada a cabo con el objetivo de evaluar los diferentes métodos tanto filogenéticos como filogenómicos consistió en analizar las topologías derivadas de los alineamientos individuales (árboles génicos) así como la reconstrucción obtenida a partir del uso del marcador filogenético más común en bacterias, el 16s rDNA. El análisis del 16s rDNA reveló que todas las aproximaciones usadas coincidían en agrupar los endosimbiontes estudiados en un único clado. La única excepción fue el uso de la distancia de Galtier y Gouy (Galtier and Gouy, 1995), una excepción notable pues trata de corregir el posible efecto que la rica composición en A+T compartida por estos genomas puede tener. Por su parte el análisis de las topologías génicas reveló una gran heterogeneidad cuya causa podría ser tanto problemas de ruido filogenético introducido por las secuencias de endosimbiontes principalmente como a la presencia de transferencias horizontales o paralógicas escondidas. Como se detallará, los análisis filogenómicos nos permiten el desentrañar la presencia de estas diferentes señales, explicando la gran heterogeneidad en las topologías encontradas.

El método filogenómico usado está en gran parte condicionado por la naturaleza de los conjuntos de datos que

tenemos. Así por ejemplo, partiendo de un conjunto de 200 genes comunes podremos aplicar tanto el análisis de supermatrices como el análisis de superárboles. Sin embargo, para conjuntos de genes de distribución desigual es preferible el segundo pues la generación de una supermatriz requeriría la inserción de un gran número de indeterminaciones. En este trabajo hemos aplicado la supermatriz a dos conjuntos de datos: el de 200-genes y el de 579-genes introduciendo huecos. Además, hemos generado una supermatriz de 200-genes en la que las posiciones aminoacídicas más afectadas por el sesgo (FYMINK) fueron eliminadas con el objetivo de explorar la influencia de la composición sesgada de los endosimbiontes en la filogenia resultante. Las tres topologías resultantes corroboraron la previamente obtenida con el árbol guía: los endosimbiontes de los géneros *Buchnera*, *Blochmannia*, *Wigglesworthia* forman un clado monofilético hermano del de las enterobacterias. El uso de una supermatriz con genes de distribución desigual (579-genes) no alteró el resultado revelando que la de 200-genes es suficiente para obtener la misma topología con valores de soporte parecidos.

Por otra parte, la aplicación de superárboles a los tres conjuntos disponibles (200-, 379-, 579-genes) reveló información que las supermatrices no habían detectado. En los tres análisis los endosimbiontes siguieron formando un clado monofilético pero uno de los grupos basales, las Xanthomonadales, resultaron tener una posición inestable. De hecho, en los dos primeros conjuntos agrupan con Beta-Proteobacteria y sólo recuperan su supuesta posición Gamma en el de 579 genes pero sin ningún valor de soporte. Puesto que los superárboles se basan en los árboles

génicos subyacentes, lo que este resultado está poniendo de manifiesto es que existe una gran incongruencia en los árboles génicos alrededor de la posición de este grupo. Es esta incongruencia de la que trata el capítulo 5 y es la causante de la heterogeneidad en las topologías génicas observada previamente.

Existen otros dos factores que *a priori* pueden afectar el análisis filogenómico. Uno es la influencia del conjunto de datos elegidos en los diferentes métodos. El otro es la influencia de la función de los genes que componen dicho conjunto de datos. Para estudiar dichos factores generamos a partir del conjunto de 579-genes tres conjuntos de datos con muy diferentes características desde el punto de vista evolutivo:

- Conjunto *Blochmannia*: compuesto por los 579 genes codificantes de *Blochmannia* y que por lo tanto son una mezcla de genes con muy diferentes señales filogenéticas.
- Conjunto *universal*: compuesto por los 200 genes comunes a las 21 especies estudiadas y que son prácticamente comunes a todas las Proteobacteria. En este conjunto coexisten una fracción importante de ortólogos con otra de xenólogos/parálogos.
- Conjunto *esencial*: compuesta por los genes del conjunto de 200 comunes pero cuya función además ha sido descrita como esencial para el mantenimiento de la vida celular. Por lo tanto se espera que la

mayoría de estos genes estén libres de transferencias horizontales.

Lo primero que hicimos fue estudiar el comportamiento de las supermatrices cuando usábamos diferente número de genes de uno u otro conjunto. En este caso el conjunto de 579 genes fue excluido del análisis. Concatenamos genes en proporciones crecientes de 10 en 10 hasta los 60 genes elegidos al azar y analizamos las topologías resultantes. De esta manera para cada categoría se generaron 100 concatenados diferentes. El resultado fue muy diferente cuando usábamos cada uno de los conjuntos. Mientras que usando el conjunto *esencial* en más de un 50% de los casos con sólo 20 genes, para el conjunto *universal* dicha proporción sólo se alcanzó para las topologías de 30 genes concatenados permaneciendo en el resto por debajo. Es más la categoría compuesta por concatenados de 60 genes recupera el árbol de referencia en el 100% de los casos para el *esencial* y solo en 40% para el *universal*. Por lo tanto se demuestra que en el conjunto *universal* existe una mayor cantidad de señal conflictiva que impide la recuperación de la topología de referencia en muchas ocasiones.

La aplicación de supermatrices y superárboles para el análisis funcional de los conjuntos descritos reveló la presencia de incongruencia en prácticamente todas las categorías aunque con algunas salvedades a tener en cuenta. Por una parte, las categorías de “Transcripción” y “Traducción” fueron las que en mayor proporción de genes recuperaron el árbol de referencia. Mientras que entre el resto destaca la poca distancia topológica media con respecto al árbol de referencia de la categoría de “Función

general” cuyos genes sólo se sabe de manera genérica qué función tienen. En general se puede argumentar que las categorías informacionales poseen una mejor señal vertical pero que igualmente la incongruencia filogenética está presente prácticamente en cualquiera de las categorías.

Las relaciones filogenéticas entre endosimbiontes ante las nuevas secuencias genómicas

Por último y retomando las relaciones filogenéticas entre endosimbiontes, la aparición de nuevas e importantes secuencias genómicas de endosimbiontes nos llevó a preguntarnos cómo afectarían estos nuevos datos a la principal conclusión de este capítulo, la monofilia de los endosimbiontes. *Carsonella ruddii* (Nakabachi *et al.*, 2006) es el genoma bacteriano más pequeño conocido, de solo 182 genes, se ha propuesto que está en el camino de convertirse en un orgánulo. Adicionalmente una nueva cepa de *Buchnera* procedente del pulgón *Cimara cedri* (Perez-Brocal *et al.*, 2006) ha sido secuenciada con el menor tamaño descrito hasta ahora. En este nuevo análisis filogenómico aplicamos toda una serie de métodos no sólo a nivel aminoacídico sino también nucleotídico intentando corregir el gran sesgo hacia A+T presente en ambos genomas.

A nivel aminoacídico levamos a cabo un concatenado de los 82 genes comunes a los genomas analizados así como un segundo concatenado sin los aminoácidos afectados por el sesgo como ya hemos explicado en un apartado anterior. Adicionalmente, sobre el alineamiento de codones a parte de los

análisis con modelos de evolución tradicionales, generamos una supermatriz en la que la tercera posición estaba codificada como purina o pirimidina (R o Y). Las diferentes aproximaciones se encuentran resumidas en la figura 15. Básicamente, aquellas que no corregían por el sesgo coincidían en presentar a *Carsonella* con el resto de endosimbiontes en un mismo grupo, mientras que aquellas que sí corregían por el sesgo la presentaban como basal a las Gamma-Proteobacteria cerca del grupo de *Legionella*. *Carsonella* tiene tal sesgo hacia A+T (84%) que todas sus posiciones sinónimas han sido ocupadas por alguno de estos nucleótidos. Hasta tal punto es así que la corrección de la tercera posición introducida por la codificación R-Y parece insuficiente. A pesar que no podemos obtener una solución rotunda al emplazamiento de *Carsonella*, los análisis nos invitan a creer que no pertenece al mismo grupo que el resto de endosimbiontes analizados donde sí aparece *Buchnera aphidicola* *Cinara cedri* a pesar de su también fuerte sesgo hacia A+T.

CAPÍTULO 5: El origen evolutivo de las Xanthomonadales

El análisis de genomas completos ha evidenciado que la incongruencia entre los árboles génicos y las filogenias de especies es un hecho constante que afecta a una fracción más o menos importante de los genes que componen un genoma excepto en casos de asociaciones intracelulares (Tamas *et al.*, 2002). Esta incongruencia puede tener dos fuentes principales en bacterias: ruido filogenético o HGT. El ruido filogenético suele estar

asociado a casos donde las secuencias analizadas tienen una pobre señal filogenética, altas tasas de evolución para ciertos genes o linajes, o problemas de atracción de ramas largas (Sanderson and Shaffer, 2002). Por el contrario, la señal derivada de los procesos de transferencia horizontal difiere del ruido en que generalmente refleja el sesgo de transferencia desde unos donadores preferentes. Por lo tanto, señales filogenéticas conflictivas coexisten en los genomas bacterianos debido a la transmisión horizontal y vertical de los genes así como al ruido filogenético. Estas señales divergentes aparecen incluso en los “cores” (Baptiste *et al.*, 2005; Susko *et al.*, 2006) indicando que el ruido filogenético y la transferencia horizontal, detectados en forma de incongruencia afecta prácticamente a cualquier gen y función celular.

Gogarten y colaboradores (Gogarten *et al.*, 2002) propusieron un modelo que asume que la probabilidad de transferencia horizontal entre dos genomas decrece con la distancia evolutiva que los separa. Además, puesto que es más probable entre los genomas más cercanos filogenéticamente, ésta puede confundirse a escala filogenómica con la señal vertical, actuando por tanto como fuerza cohesionadora de un clado. Recientemente, 144 genomas de bacterias y arqueas (Beiko *et al.*, 2005) fueron analizados con la intención de detectar el mayor número posible de eventos de transferencia. El análisis sugirió preferencias en la transferencia de genes entre taxones relativamente cercanos, por tanto reafirmando la hipótesis de que existen restricciones al intercambio libre de material génico entre bacterias que deben estar relacionadas con la compatibilidad entre

las arquitecturas genómicas y/o su distancia filogenética (Gogarten *et al.*, 2002; Hendrickson and Lawrence, 2006).

En este capítulo, nos hemos centrado en un grupo particular de bacterias, las Xanthomonadales, que parecen haber sido especialmente afectadas por eventos de HGT. Es el grupo más basal del clado de Gamma-Proteobacteria, y está compuesto por fitopatógenos cuya relación con el hospedador va desde las asociaciones obligadas, como las especies de *Xylella*, a asociaciones no obligadas como el caso del género *Xanthomonas* (Van Sluys *et al.*, 2002). Trabajos anteriores han revelado una posición inestable del clado dentro del árbol de Proteobacteria (Van Sluys *et al.*, 2002; Beiko *et al.*, 2005). Tanto las filogenias de genes individuales como las filogenias genómicas las han emplazado con la misma frecuencia como Beta-, Gamma-, o Alfa-Proteobacteria o como un clado externo a los tres grupos (Van Sluys *et al.*, 2002; Omelchenko *et al.*, 2003; Martins-Pinheiro *et al.*, 2004; Bern and Goldberg, 2005; Dutilh *et al.*, 2005). De hecho, en las filogenias publicadas recientemente es común que aparezcan fuera del clado de las Gamma-Proteobacteria (Van Sluys *et al.*, 2002; Omelchenko *et al.*, 2003; Creevey *et al.*, 2004; Martins-Pinheiro *et al.*, 2004; Studholme *et al.*, 2005). Teniendo en cuenta estos trabajos dos son las explicaciones más probables: ruido filogenético o transferencia génica horizontal.

Por una parte, es de esperar que el ruido filogenético afecte un cierto número de los genes analizados sobretodo en los casos de grupos basales, cuya posición puede cambiar debido a limitaciones en la cantidad de información o de los métodos de reconstrucción filogenética. Por otra parte, siguiendo la hipótesis

de Gogarten y colaboradores (2002), es también de esperar que las transferencias entre Proteobacteria y Xanthomonadales fueran más probables en el pasado cuando la divergencia entre los principales linajes era menor, mientras que las transferencias recientes vendrían principalmente de otras Xanthomonadales. Desde un punto de vista filogenómico, transferencias antiguas al ancestro del grupo pueden resultar en su posición inestable o no resuelta en el árbol de Proteobacteria.

Resultados y discusión

En este capítulo pretendemos determinar el origen de los resultados conflictivos referenciados más arriba, y por tanto si éstos se deben a ruido filogenético debido a convergencias y/o pérdida de señal o bien a transferencias antiguas o recientes. Consideramos ruido filogenético aquella señal conflictiva resultado de procesos no relaciones con la forma de transmisión (vertical u horizontal) del gen, y que violan las asunciones de métodos de reconstrucción filogenética. Alternativamente, consideramos señal filogenética aquella derivada de la transmisión vertical u horizontal del gen. Para separar los dos componentes, transmisión o ruido, en los genomas de Xanthomonadales primero identificamos todas las posibles señales filogenéticas presentes. Seguidamente, investigamos la afinidad de cada uno de los genes a los clados de Gamma-, Beta, o Alfa-Proteobacteria. Nuestros resultados indican, la existencia de diferentes señales filogenéticas cuyo origen son los tres grupos de Proteobacteria considerados. Demostramos que, al contrario que el ruido filogenético, estas señales no se encuentran distribuidas al azar en el genoma; genes adyacentes tienden a tener la misma señal

filogenética más frecuentemente que lo esperado por azar, indicando que las señales detectadas no son el resultado de ruido sino de incongruencia sistemática hacia donadores de los tres grupos principales de Proteobacteria.

Hemos analizado un conjunto de 18 genomas de Proteobacteria con el objetivo de estudiar el origen filogenético de las Xanthomonadales. El conjunto de datos incluye un número balanceado de representantes de los tres principales grupos de Proteobacteria y tres genomas de Xanthomonadales, *Xanthomonas axonopodis* pv. *citri* str. 306 (*X. citri*, Xci), *Xanthomonas campestris* pv. *campestris* str. ATCC 33913 (*Xca*), y *Xylella fastidiosa* 9a5c (*Xy. Fastidiosa*, Xy).

Nuestra búsqueda inicial de ortólogos putativos identificó 207 genes comunes a todos los genomas. Nuestro primer objetivo se centró en identificar aquellas especies no pertenecientes a Xanthomonadales que introducen ruido en forma de incongruencia para los futuros análisis. Para ello reconstruimos el árbol consenso de las 207 topologías de tal manera que pudimos evaluar el grado de incongruencia introducido por cada especie. En el árbol identificamos tres nodos con baja resolución, aquellos correspondientes a *Nitrosomonas europaea*, *Legionella pneumophila*, y el grupo de Xanthomonadales. Como sólo estábamos interesados en este último, descartamos las otras dos especies para siguientes análisis.

Con estos 207 genes comunes a los restantes 16 genomas, buscamos la presencia de diferentes señales filogenéticas. Para ello, llevamos a cabo un mapa de congruencia en el que se prueba la

congruencia de cada uno de los genes por las 207 topologías génicas obtenidas; de esa manera cada fila se corresponde a un gen y cada columna a un árbol génico. El análisis identificó numerosos genes cuyas reconstrucciones filogenéticas eran claramente compatibles con otras topologías con orígenes en Alfa-, Beta- o Gamma-Proteobacteria que fueron identificadas en base al agrupamiento monofilético de las secuencias de Xanthomonadales con el resto de secuencias del grupo correspondiente. Sin embargo, muchas pruebas fueron incapaces de rechazar algunas topologías incongruentes entre sí. Los casos van desde genes que son compatibles con prácticamente todas las topologías, no pudiendo distinguir entre los diferentes orígenes posibles, a genes que solamente son compatibles con su propia topología. En conjunto, las topologías Gamma son las más aceptadas en el mapa de congruencia, considerando tanto casos en los que son las únicas topologías aceptadas como casos en los que aparte de topologías Gamma también aparecen topologías alternativas. El análisis por lo tanto revela una mezcla de ruido filogenético, que se corresponde con los genes congruentes sólo consigo mismos y aquellos compatibles con casi cualquier topología. Pero además de ruido, señales robustas y diferentes entre sí fueron detectadas en términos de aceptación o rechazo de grupos de topologías correspondientes a diferentes posiciones del clado de Xanthomonadales con respecto a otras Proteobacteria.

Para evaluar la posible influencia de artefactos de atracción de ramas largas en las filogenias estudiadas, llevamos a cabo pruebas de tasas relativas para los 207 genes comunes. En dichas pruebas nos interesamos por los casos en los que las

Xanthomonadales compartían ramas largas con un grupo (por lo tanto ambos grupos rechazaban por separado los otros dos) y además ese grupo era el mismo que el detectado como posible donador de la transferencia. El análisis reveló solo un caso de posible convergencia debida a compartir altas tasas de sustitución (correspondiente al gen *hisS*). El resto de genes no mostraron evidencia de artefactos de ramas largas y, en consecuencia, excluimos a este artefacto filogenético como responsable de agrupamientos aparentemente incongruentes.

Este análisis de 207 genes sugiere la presencia de una gran cantidad de eventos de transferencia horizontal que podrían haber jugado un papel central en la evolución del grupo aunque otras alternativas pueden ser posibles. Por ello, diseñamos una prueba específica que nos permitiera cuestionarnos directamente la presencia de HGT en los genomas de Xanthomonadales. Seleccionamos un conjunto de datos mayor que nos permitiera obtener estadísticas más robustas. Esto nos permitió analizar si la causa de que algunos genes fueran incapaces de rechazar topologías alternativas e incompatibles en el mapa de congruencia era el resultado de ruido filogenético o bien del sello de eventos de transferencia horizontal. Para ello elegimos aquellos ortólogos putativos identificados en al menos 10 genomas de los 16 estudiados. El conjunto extendido de 1051 genes estaba compuesto por genes casi-universales con funciones no relacionadas directamente con la virulencia de *Xanthomonas citri*. Una comparación con una lista conocida de genes relacionados con virulencia en dicho genoma sólo identificó 19 candidatos.

Para determinar si los genes que contribuyen a la señal conflictiva habían sido transferidos recientemente en los genomas de Xanthomonadales, identificamos genes atípicos mediante una metodología de agrupamiento basado en el criterio de AIC y usando el sesgo en el uso de codones y la composición nucleotídica como criterio discriminante (Azad and Lawrence, 2005). Encontramos un grupo principal de genes típicos así como diferentes grupos de genes atípicos. El análisis reveló que sólo un 0,2% para el uso de codones, y un 13,61% para composición nucleotídica de los genes usados en este estudio tienen una composición atípica. Mientras que, la frecuencia de genes atípicos en el genoma completo era de 2,21% y 22,23% respectivamente. Por lo tanto, la incongruencia observada en el mapa de congruencia no puede ser atribuida a la influencia confusa de eventos de transferencia horizontal reciente. Aunque algunos de éstos no se pueden descartar, una transferencia reciente es un escenario bastante improbable para gran parte del conjunto de genes conflictivos. La incongruencia filogenética aparece para los tres genomas estudiados de Xanthomonadales y por lo tanto la transferencia que generó dicha incongruencia debió darse en el ancestro más reciente de los genomas. Todos estos resultados sugieren que las posibles transferencias estudiadas deben ser antiguas.

Para verificar que las transferencias antiguas a los genomas de Xanthomonadales dan lugar a la incongruencia filogenética observada en los genes casi-universales analizados, examinamos también la compatibilidad de cada gen con cinco hipótesis filogenéticas posibles. Las topologías ESTRELLA1 y

ESTRELLA2 son topologías nos resueltas con la única diferencia de que en la última, los cladogramas terminales están resueltos; genes con una fuerte señal filogenética deberían rechazar ambas topologías. Las otras tres topologías probadas emplazan a las Xanthomonadales como el grupo más basal de los grupos de Gamma-, Beta, o Alfa-Proteobacteria. Los 1051 genes analizados rechazaron la topología ESTRELLA1, pero 51 genes fueron incapaces de rechazar la topología ESTRELLA2 por lo que fueron eliminados para los siguientes análisis. La distribución de los orígenes filogenéticos más probables del resto de genes muestra una cierta preferencia por Alfa. Sin embargo, muchos de esos genes son incapaces de rechazar alguna, o incluso ninguna, de las dos posibilidades alternativas. En consecuencia, los siguientes análisis se basaron en analizar la asignación más probable más allá de que el gen pueda ser compatible con otras topologías. En cualquier caso, un análisis de los genes que sólo seleccionaban una de las tres hipótesis reveló el mismo patrón, con las topologías Alfa como las más preferidas y las Beta como las que menos.

Puesto que el experimento sobre el origen filogenético de los genes no fue suficiente para distinguir entre ruido filogenético y señal filogenética, probamos el umbral de ruido analizando la distribución de los genes y su posible origen en los genomas de Xanthomonadales. Un análisis de adyacencia, agrupamiento, se llevó a cabo con un conjunto reducido de los 1051 genes usados en la prueba de origen filogenético. Seleccionamos solamente aquellos genes que estaban adyacentes por lo menos a uno más en los genomas de Xanthomonadales. El número de pares analizado fue de 430 en *X. citri*, 438 en *X. campestris*, y 377 en *X. fastidiosa*.

Esto nos permitió comprobar dos predicciones alternativas. Si la incongruencia fuera sólo debida al ruido filogenético, entonces los pares de genes adyacentes no mostrarían ninguna asociación con respecto al origen filogenético detectado para cada uno de ellos. Por otra parte, por lo menos bajo ciertos modelos de HGT (Lawrence and Roth, 1996), pares de genes adyacentes en el genoma recipiente deberían tender a compartir el mismo origen filogenético. Nuestra prueba estadística reveló que el número de genes adyacentes con el mismo origen filogenético en los genomas de Xanthomonadales era mayor de lo esperado. Lo que es más, dicho agrupamiento era evidente y significativo para los tres genomas de Xanthomonadales y los tres posibles orígenes filogenéticos estudiados. Esta evidencia de agrupamiento remarca dos aspectos. Por un parte, rechaza la posibilidad de que la mayor parte de los resultados sean simplemente el resultado de ruido filogenético aunque evidentemente, éste está presente de alguna manera. Por otra parte, a pesar de que la prueba de adyacencia reduce el análisis a pares de genes, en realidad los grupos de genes con el mismo origen solían ser mayores de dos genes. De hecho existen ejemplos de grupos de ocho genes, siendo la media de 2,35 genes por grupo, valor que concuerda con las estimas del número de genes que componen un operón medio y por tanto apuntando a la transferencia horizontal de operones como el mecanismo de evolución más probable.

En el caso de que fueran operones la unidad de transferencia entre linajes de Gamma-, Beta-, Alfa-Proteobacteria, dando lugar a la incongruencia detectada en los genomas de Xanthomonadales, entonces la mayoría de genes adyacentes con la

misma señal filogenética deberían también ser transcritos en la misma dirección. Estudiamos la dirección de transcripción de los genes presentes en el análisis de adyacencia; como esperábamos bajo la hipótesis de la transferencia horizontal de operones completos o parciales, la mayoría de los genes identificados se presentaban en la misma hebra. Para los casos de Alfa ($P < 0.0002$) y Gamma ($P < 0.02$) la frecuencia con la que los genes que componían un par y tenían la misma señal filogenética eran transcritos en la misma dirección era significativamente mayor que la frecuencia con la que genes no adyacentes se presentaban en la misma dirección. No así para los casos de origen de Beta-Proteobacteria ($P = 0.3079$). Por lo tanto, es probable que muchos de los grupos identificados en nuestro estudio, sobre todo aquellos que tienen que ver con un origen Gamma- o Alfa-, sean operones.

También investigamos la posible relación entre la asignación funcional de los genes estudiados y su origen filogenético para determinar si los genes pertenecientes a categorías funcionales eran menos susceptibles a ser transferidos que los no informacionales (Jain *et al.*, 1999). No detectamos ninguna asociación entre las clases funcionales y el supuesto origen filogenético, pero algunos patrones se podían distinguir. Por ejemplo, siete de los ocho genes flagelares comparten el mismo, Beta-, origen. Por otra parte, las categorías informacionales son más ricas en topologías Gamma, mientras que la más dominante entre los genes metabólicos es la Alfa. Como consecuencia, la diferente composición en categorías funcionales del conjunto de 207 genes (más rico en

informativos) y el conjunto de 1051 genes (más rico en metabólicos) podría explicar las diferencias en el origen más frecuente entre los dos conjuntos (Gamma y Alfa respectivamente). En cualquier caso, la mayoría de las categorías presentan una mezcla de topologías incluyendo algunos excelentes marcadores moleculares como los genes relacionados con la transcripción.

En resumen, nuestros resultados son compatibles con la hipótesis de Gogarten y colaboradores (2002) como se describe en la Figura 23. Presentamos evidencias de transferencias horizontales al ancestro de las Xanthomonadales, antes de su diversificación y seguramente antes de la diversificación de los linajes de Proteobacteria. El bajo número de genes atípicos detectados entre estas transferencias indican su antigüedad y el cambio en las preferencias de intercambio con el tiempo. La mayor parte de transferencias recientes se han dado entre miembros de las Xanthomonadales por lo que no afectan a los resultados filogenómicos del grupo completo. Mientras que aquellas transferencias más antiguas, al incidir sobre el ancestro provocan un cambio en la posición del todo el clado. En este caso además dichas transferencias son lo bastante importantes como para que las probabilidades de origen de los genes se distribuyan por igual entre Gamma-, Beta- y Alfa-Proteobacteria.

Además, estos resultados reflejan los diferentes efectos en la reconstrucción filogenética dependiendo del tiempo pasado desde la transferencia. Aquellos métodos basados en árboles génicos, superárboles y consensos, son apropiados para indicar, en la forma de nodos no resueltos, un importante número de

transferencias en el pasado. Cuanto mayor sea ese número, mayor será la probabilidad de no resolver la posición de dicho clado en las filogenias genómicas de bacterias. Por otra parte, todos los árboles genéticos derivados de los 207 genes comunes soportan la monofilia para el clado de Xanthomonadales lo que no solo puede ser debido a la transmisión vertical de los genes desde su diversificación sino también a continuas transferencias intra-Xanthomonadales que difícilmente se pueden ver reflejadas a escala genómica debido a que no hay suficiente divergencia entre ellas como para ser detectada.

CAPÍTULO 6: Fuerzas evolutivas opuestas actúan en las últimas etapas de la reducción genómica

El análisis de los tamaños genómicos en bacterias indica que la ganancia de material genético por transferencia horizontal suele venir contrarrestada por una tasa similar de pérdida génica (Dagan and Martin, 2007). Los genomas bacterianos de endosimbiontes son un caso extremo de pérdida génica en masa durante su evolución asociada a un hospedador. En este contexto, el extremadamente reducido genoma del endosimbionte de psílidos, *Carsonella ruddii* (Nakabachi *et al.*, 2006), y el genoma más pequeño conocido entre los endosimbiontes primarios de pulgones, *Buchnera aphidicola* *Cianra cedri* (Perez-Brocal *et al.*, 2006), han llevado a cuestionarse cuál es en último término el destino de estos genomas y cuáles son las fuerzas evolutivas que actúan en las últimas etapas de la reducción genómica.

Es difícil anticipar cuál será el destino de estos genomas. A pesar de que el estasis ha sido propuesto para el tamaño genómico de *Buchnera*, todavía existen diferencias remarcables entre cepas como ha revelado la secuenciación del genoma del endosimbionte del pulgón *Cianra cedri*. Para este genoma en particular, un modelo de libre-difusión celular ha sido propuesto (Perez-Brocal *et al.*, 2006) con el objetivo de explicar cómo es posible que la célula todavía sea capaz de obtener metabolitos ante la falta de muchos de los transportadores necesarios. Los autores también proponen un escenario en el que *Buchnera* está siendo reemplazada por el endosimbionte secundario *Serratia symbiotica*. El caso de *Carsonella* es diferente, los autores proponen que está sufriendo un proceso de reducción al estilo de los orgánulos eucariotas (Nakabachi *et al.*, 2006) y por tanto su destino final sería su incorporación en la célula eucariota como un orgánulo más.

Los endosimbiontes son heredados verticalmente sufriendo regulares cuellos de botella durante la transmisión. Además, sus tasas de mutación están incrementadas porque muchos de ellos carecen de los genes relacionados con los sistemas de reparación, por lo tanto favoreciendo la fijación de una gran cantidad de mutaciones tanto neutrales como no-neutrales. Sin embargo, puesto que se han mantenido en una relación más o menos estable con sus hospedadores es de esperar que todavía exista un cierto papel de la selección. Dicho papel bien puede ser en forma de selección purificadora para prevenir la pérdida e ciertas funciones esenciales o bien como selección positiva para permitir la adaptación a condiciones cambiantes en el ambiente hospedador-simbionte y por tanto evadir su propia

extinción. Este capítulo no se centra tanto en discutir cuál es el destino final de estos genomas como en intentar caracterizar patrones evolutivos comunes y diferentes entre los genomas de endosimbiontes, principalmente los dos mencionados más arriba.

Resultados y discusión

En este estudio hemos analizado 26 genomas de Gamma-Proteobacteria entre los que se encuentran 10 genomas de endosimbiontes descritos en la Tabla 4 y en el primer capítulo de este resumen. El análisis de ortología partió del genoma de *Carsonella ruddii* de solo 182 genes codificantes, identificando 82 genes comunes a los 26 genomas. Se obtuvieron los alineamientos tanto para aminoácidos como para los codones basándose en los primeros.

Lo primero que analizamos fue el grado en el que sesgo hacia A+T que presentan los endosimbiontes afecta a todas las posiciones nucleotídicas. Para ello desarrollamos dos medidas complementarias que se pueden aplicar a genomas completos:

$$SMg = \frac{1}{n} \sum_{i=1}^n \left(1 - \frac{GC_{syn}}{Total_{syn}} \right)$$

$$RMg = \frac{1}{n} \sum_{i=1}^n \frac{GC_{non}}{Total_{non}}$$

La medida de saturación (SM) se basa en los potenciales sitios sinónimos ocupados por A/T. Como es de esperar la saturación en A/T se incrementa con el aumento en contenido en A+T llegando a un límite alrededor de un 94% o 95% de

posiciones representadas por *Carsonella* y *Buchnera aphidicola* *Cinara cedri*. Además una clara discontinuidad se encuentra entre el endosimbionte menos saturado y el primer no-endosimbionte del conjunto analizado (ver Figura 27).

Para comprobar la relación entre la tasas de evolución acelerada de los endosimbiontes y la saturación en A+T calculamos la tasa relativa de sustitución entre pares de genes de cada endosimbionte con el correspondiente en *Escherichia coli* K12. Como era de esperar la aceleración en las tasas de evolución es evidente como queda reflejada en la Figura 25. Los dos genomas más acelerados vuelven a ser aquellos con un mayor contenido en A+T junto con *Wigglesworthia brevialpilis* que presenta valores similares. Sin embargo, las diferencias en las tasas de evolución entre endosimbiontes pueden ser debidas tanto a las sustituciones sinónimas como las no-sinónimas. Nos preguntamos por lo tanto dónde reside esta tasa de evolución acelerada.

La comparación dos a dos con secuencias de *Escherichia coli* reveló lo que ya se sabía, la tasa de sustitución sinónima está saturada. Por lo tanto, las diferentes tasas de evolución entre endosimbiontes debe residir en la tasa de sustitución no-sinónima como así se corrobora (Figura 26). La siguiente pregunta es por lo tanto qué fuerzas están detrás de esa tasa no-sinónima acelerada.

Previamente se han publicado estudios demostrando la intervención de dos factores en dicha tasa: la deriva genética (Moran, 1996) y altas tasas de mutación (Itoh *et al.*, 2002) debido a la pérdida de sistemas de corrección de la secuencia. A pesar de que tradicionalmente se le ha dado más peso a unas u otras parece

claro que ambas actúan sinérgicamente. Las altas tasas de mutación aumentan el número de mutaciones disponibles para ser fijadas por azar. Menos atención se le ha prestado sin embargo al papel de la selección en la tasa de sustitución no-sinónima aunque existe algún estudio centrado en genes particulares (Fares *et al.*, 2002; Fry and Wernegreen, 2005). La selección positiva es un fenómeno difícil de demostrar bajo ciertas condiciones. La falta de aproximaciones metodológicas fiables o que permitieran analizar en base no a toda la secuencia, sino codón a codón, no ha permitido un estudio riguroso de su acción en los genomas de endosimbiontes. Nosotros hemos aplicado un test (Zhang *et al.*, 2005) que permite por una parte detectar la presencia de selección en una rama concreta del árbol mientras por otra detectar que codones específicos están bajo selección.

Los resultados del test de selección positiva apuntan a dos conclusiones (Tabla 6): la selección positiva está más presente cuanto más reducción ha sufrido el genoma y el número de sitios detectados también aumenta con los genomas más reducidos. Aquí se plantea por lo tanto el problema de dirimir entre ruido y señal a la hora de hablar de que un codón este bajo selección. Puesto que los genomas más reducidos son aquellos que tienen un mayor contenido en A+T y unas tasas de sustitución más saturadas es lógica la prevención de que sean estos factores y no la acción de la selección lo que provoquen falsos positivos en el test. Para diferenciar entre ambas posibilidades lo primero que hicimos fue considerar un codon bajo selección solo en el caso de que pasara estos tres filtros: 1) que el test de selección fuera significativo, 2) que sea identificado con una probabilidad α

posteriori de 0,995 a través de una aproximación bayesiana y 3) que tenga un homólogo en *Escherichia coli* y que mantenga los sitios en G/C o cambie hacia G/C con respecto a ese homólogo.

Estos filtros revelan que sí existe un efecto del sesgo en A+T sobre las posiciones detectadas bajo selección pero sin embargo, las conclusiones cualitativamente no cambian. Existe selección positiva en los genomas de endosimbiontes más marcada en aquellos más reducidos.

DISCUSIÓN GENERAL

Filogenómica: metodologías y conjuntos de datos

Como se demuestra en la primera parte de este resumen, los superárboles y las supermatrices permiten explorar las señales filogenéticas en los genomas bacterianos. Ambos análisis son puntos de inicios válidos y complementarios para el estudio de dichas señales generadas por procesos como la transmisión vertical u horizontal de los genes o la presencia de ruido filogenético. Las supermatrices por ejemplo, son un espada de doble filo, permiten recuperar la señal filogenética más fuerte incluso cuando la supermatriz está compuesta por alineamientos en los que dicha señal está escondida. Sin embargo, esto puede resultar en el enmascaramiento de señales alternativas. La presencia de estas topologías alternativas es más fácil de revelar con los análisis que se basan en los árboles génicos como son los superárboles o los árboles consensos. Dicha incongruencia entre los árboles génicos suele reflejarse en forma de nodos no resueltos.

Por otra parte, como hemos demostrado, diferentes conjuntos de datos presentan diferentes señales filogenéticas, la identificación de dichas señales depende por lo tanto mucho del conjunto de genes iniciales usados.

Filogenómica: relaciones filogenéticas entre los endosimbiontes de insectos de Gamma-Proteobacteria

Como se ha detallado en el primer capítulo de esta tesis, tratar de determinar las relaciones filogenéticas entre endosimbiontes es una tarea ardua debido a la dificultad en distinguir entre convergencia y características derivadas de un ancestro común. A pesar de que hemos presentado evidencias de la relación monofilética entre muchos de esos endosimbiontes, la cuestión sobre el origen evolutivo de *Carsonella* se mantiene abierta. Nosotros consideramos que su posición más probable es basal, fuera del clado del resto de endosimbiontes, pero también reconocemos que los análisis no son totalmente concluyentes y no permiten descartar una posición dentro de dicho clado. A pesar de que actualmente hay muchos métodos y datos disponibles para el análisis filogenético, éstos pueden no ser suficientes en casos extremos como este en el que la degeneración de las secuencias, con un fuerte sesgo en A+T, afecta a todas las entidades con información filogenética haciendo muy difícil la resolución de dichos casos aunque se desarrollen nuevos métodos o modelos.

La evolución de los genomas de Xanthomonadales y la naturaleza del proceso de transferencia horizontal

Del capítulo que explora el origen filogenético de los genes de Xanthomonadales podemos extraer dos conclusiones. Por una parte, los genomas de Xanthomonadales tienen aproximadamente el mismo número de genes con un origen Beta-, Gamma- o Alfa-Proteobacteria haciendo de ellas un caso extremo de mosaicismo e impidiendo que se le pueda asignar al grupo completo un origen más probable en alguno de los clados principales de Proteobacteria.

Por otra parte, demostramos que es posible discernir entre ruido y señal a través de un análisis exhaustivo y cuidadoso incluso para casos tan complejos como el de Xanthomonadales. Hemos demostrado la existencia de transferencias recientes y más antiguas a pesar de posibles artefactos filogenéticos. El efecto de estas transferencias sobre la reconstrucción filogenómica y la resolución de nodos ancestrales dependerá de la relación entre la cantidad de señal vertical y horizontal asignada a la rama. Estas relaciones determinarán qué parte de los árboles genómicos siguen una pauta vertical de ancestro-descendiente y cuáles no. Obviamente, las Xanthomonadales se ajustan a este modelo puesto que por una parte aparecen como un grupo monofilético en los árboles genómicos mientras que su posición con respecto a otras Proteobacteria es indefinida debido a la incidencia de transferencia de horizontal desde diferentes Proteobacteria en la rama que lleva a su ancestro común. Otros grupos como Pseudomonadales o Cianobacterias parecen seguir patrones parecidos. Sin embargo, este escenario evolutivo en el que los

nodos internos no están resueltos no tiene porqué aplicar a otros genomas con menor promiscuidad o susceptibilidad a la transferencia en tiempos pasados, en cuyo caso la señal vertical tiene mayor peso que la horizontal y permite una mejor resolución de las relaciones filogenéticas más profundas.

La relación entre fuerzas evolutivas opuestas actuando en los genomas más reducidos

En las poblaciones de endosimbiontes tanto la deriva genética como las altas tasas de mutación son las principales fuerzas que aumentan tanto el número de sustituciones sinónimas como no-sinónimas. Éstas últimas se reflejan en el gran número de genes bajo selección relajada así como de codones falsamente detectados bajo selección positiva. Sin embargo, también hemos presentado evidencias de la presencia de una fracción importante de mutaciones ventajosas, detectadas como cambios hacia G/C en los codones detectados bajo selección, que hasta un cierto punto contrarrestan el proceso de degeneración continua que sufren estos genomas, principalmente en aquellos que están en las últimas etapas.

Tres principales escenarios evolutivos para estos genomas pueden ser descritos. En primer lugar, el proceso de reducción genómica se mantiene más allá de la acción de la selección positiva y por tanto en último término el genoma desaparecerá. La presión hacia la extinción más lógica en este escenario sería la presencia de

un endosimbionte secundario que supliera las funciones del primario. Alternativamente, es posible que el endosimbionte retenga sólo unos pocos genes, perdiendo la poca capacidad de autonomía que le quedaba y transfiriendo funciones importantes al genoma nuclear del hospedador. Finalmente, si no hay una presión selectiva que le lleve a su extinción, el genoma del endosimbionte puede llegar a un estado en equilibrio seguramente inestable en el que continúa proveyendo al hospedador con algún beneficio y por lo tanto la selección, sobretodo purificadora aunque también la positiva, favorece el mantenimiento de la integridad del genomas.

11. SUPPLEMENTARY INFORMATION

Supplementary information Table 1. Number of genes by phylogenomic core and functional category. Information about the core concatenates analyzed in chapters 3 and 4 is also shown.

	Number of genes	Positions analyzed	Number of variable sites	Number of parsimony informative sites
Phylome	579			
INFORMATION STORAGE AND PROCESS				
J	108	NA	NA	NA
K	15	NA	NA	NA
L	31	NA	NA	NA
CELLULAR PROCESSES				
D	11	NA	NA	NA
O	29	NA	NA	NA
M	44	NA	NA	NA
N	20	NA	NA	NA
P	19	NA	NA	NA
T	6	NA	NA	NA
METABOLISM				
C	44	NA	NA	NA
G	29	NA	NA	NA
E	70	NA	NA	NA
F	24	NA	NA	NA
H	33	NA	NA	NA
I	22	NA	NA	NA
POORLY CHARACTERIZED				
R	45	NA	NA	NA
S	29	NA	NA	NA
Universal	200	52029	40576	32921
INFORMATION STORAGE AND PROCESS				
J	86	19465	15280	12318
K	11	5047	3262	2589
L	14	4740	3839	3139
CELLULAR PROCESSES				
D	3	891	730	626
O	15	4328	2968	2339
M	14	3625	3228	2658
N	8	2281	1688	1401
P	NA	NA		
T	2	90	74	62
METABOLISM				
C	10	3366	2553	2060
G	1	151	147	124
E	9	2264	1906	1626
F	7	1617	1339	1042
H	2	524	474	367
I	4	837	706	580
POORLY CHARACTERIZED				
R	12	2317	1981	1677
S	2	147	120	106
Essential	133	36810	27873	22586
INFORMATION STORAGE AND PROCESS				
J	75	17177	13474	10946
K	8	4394	2864	2272
L	12	4532	3638	2997
CELLULAR PROCESSES				
D	1	296	207	173
O	11	3580	2428	1888
M	1	217	162	130
N	5	2064	1519	1258
P	NA	NA		
T	NA	NA		
METABOLISM				
C	6	1610	1148	891
G	NA	NA		
E	1	192	182	126
F	4	925	698	589
H	0	NA		
I	0	NA		
POORLY CHARACTERIZED				
R	8	1746	1490	1265
S	1	79	63	12

