University of South Bohemia

Faculty of Science

Bachelor Thesis

Molecular phylogeny of intracellular symbiotic Gammaproteobacteria in insects

Filip Husník

Supervisor: Prof. RNDr. Václav Hypša, CSc.

České Budějovice 2010

Husník, F., 2010: Molecular Phylogeny of Intracellular Symbiotic Gammaproteobacteria in Insects. Bc Thesis, in English. - 45 p., Faculty of Science, The University of South Bohemia, České Budějovice, Czech Republic.

Annotation:

Many groups of insects harbor mutualistic intracellular bacteria. These bacteria originated mainly from two bacterial phyla: Bacteroidetes and Proteobacteria. The thesis is focused on phylogeny of Enterobacteriaceae - the most diverse group of endosymbiotic bacteria within Gammaproteobacteria. The study brings new phylogenetical data on "primary" symbionts and summarizes the current state of knowledge on their phylogeny, evolution and diversity.

Prohlašuji, že svoji bakalářskou práci jsem vypracoval samostatně, pouze s použitím pramenů a literatury uvedených v seznamu citované literatury.

Prohlašuji, že v souladu s § 47b zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své bakalářské práce, a to v nezkrácené podobě, Přírodovědeckou fakultou elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách.

V Českých Budějovicích, 30. dubna 2010

Filip Husník

Acknowledgements:

I would like to thank my supervisor Václav Hypša for great theoretical background and my technical advisor Tomáš Chrudimský for valuable help and constructive advice in our lab. The access to the MetaCentrum supercomputing facilities provided under the research intent MSM6383917201 is highly appreciated. This work was supported by Student Grant Agency of the Faculty of Science, University of South Bohemia.

Contents:

1. Introduction

1.1 Intracellular symbionts of insects

Insect symbionts are an astonishingly diverse group of bacteria, both phylogenetically and ecologically. They affect host ecology and evolution, influencing reproductive systems, defense mechanisms and nutrition. Some of the symbiotic bacterial lineages display a surprising flexibility; for example the genera *Arsenophonus*, *Sodalis* or *Wolbachia* can give rise to several different symbiotic lifestyles . In many lineages, their effect on the host, however, remains unknown (Moran et al., 2008; Wernegreen, 2002).

Until the boom of molecular methods in the last decade of the 20th century, the research of symbiotic bacteria was based mainly on morphology, histology and metabolic functions (Buchner, 1965). Evolutionary history was almost neglected. During this period, two main ecological categories were defined – primary (P) and secondary (S) symbionts. In spite of the fact that there is no firm boundary between P and S-symbionts, these terms are constantly used to simplify the whole concept.

1.1.1 *Primary vs. secondary symbionts*

P-symbionts (also called obligate) are intracellular bacteria inherited maternally by vertical transmission and cospeciating with their host. They are harbored directly in cytoplasm of the host cells called bacteriocytes that can form a special organ – bacteriom (mycetocytes, mycetom – older terms) (Baumann, 2005; Douglas, 1989). They provide their host with compounds missing in nutritionally-poor diet or metabolize waste products (Moran et al., 2003; Sabree et al., 2009). As a consequence of this relationship, they are required for the survival of their hosts. Extreme importance for the host fitness was shown e.g. in the louse *Pediculus humanus* by the removal of the bacteriom (Aschner, 1932, 1934; Baudisch, 1958), by antibiotic treatment (Hill et al., 1973; Pell and Southern, 1976; Schlein, 1977; Southwood et al., 1975), or by the attack of lysosyme and specific antibodies in tsetse flies (Nogge, 1978).

In comparison to P-symbionts, S-symbionts (also called facultative symbionts or guest microorganisms) is ecologically much more heterogeneous group of bacteria (Haynes et al., 2003; Moran et al., 2005a; Sandstrom et al., 2001; Tsuchida et al., 2006; Tsuchida et al., 2002). They are not necessary for the host reproduction (Douglas et al., 2006) and can invade various cell types and tissues, including haemolymph (Fukatsu et al., 2000). In contrast to P-symbionts, they are often horizontally transferred to other species (Russell et al., 2003; Sandstrom et al., 2001; Thao et al., 2000; Thao et al., 2002). They can contribute to host resistance to parasitoids (Hansen et al., 2007; Oliver et al., 2008; Oliver et al., 2005, 2006; Oliver et al., 2003), increase tolerance to heat stress

(Chen et al., 2000; Montllor et al., 2002; Russell and Moran, 2006), protect against pathogenic fungi (Scarborough et al., 2005), and influence host-plant specialization (Tsuchida et al., 2004). Some symbionts/parasites, such as *Wolbachia*, *Cardinium*, *Spiroplasma* and *Arsenophonus*, are known to manipulate reproductive systems of the host (Duron et al., 2008).

Only one percent of the known bacteria can be cultivated in laboratory conditions (Rappe and Giovannoni, 2003). All attempts to cultivate P-symbionts have so far been unsuccessful. Some of the S-symbionts can be maintained in insect cell lines (*Arsenophonus triatominarum*, *Hamiltonella defensa*, *Regiella insecticola*, *Wolbachia pipientis*) and only two species (*Arsenophonus arthropodicus* and *Sodalis glossinidius*) can grow in axenic culture (Pontes and Dale, 2006). Since, according to the current rules of bacterial nomenclature, description of a new genus/species requires *in vitro* culture (Murray and Schleifer, 1994), many of the insect symbionts have been named under provisional *Candidatus* status. To simplify the text, I have omited using this status hereafter; formal names can be found in the references.

1.1.2 *Co-symbiosis, transitions, losses and replacements*

Good example of the vague distinction between P and S symbionts is provided by occurance of multiple symbiotic bacteria within a single host. Complex symbioses involving multiple bacteria are quite common and their number increases with improved accuracy of the used methods (Haynes et al., 2003). Several bacterial lineages comprising both, primary and secondary symbionts (*Arsenophonus*, *Sodalis*) imply transitions between facultative and obligate symbiotic lifestyles.

Paul Baumann suggested term coprimary symbiont(s) for the occurrence of more than one (primary) obligate symbiont within a single host. This situation is well illustrated by the symbionts of Auchenorrhyncha (Takiya et al., 2006). Ancestor of the whole group was infected by the ancestor of the symbiont *Sulcia muelleri* (Bacteroidetes) more than 260 million years ago (Moran et al., 2005b). In addition, sharpshooters harbor *Baumannia cicadellinicola* (γ-Proteobacteria) and cicadas harbor *Hodgkinia cicadicola* (α-Proteobacteria). Both of these latter symbionts are obligate and cospeciating with their hosts for millions years. Despite being phylogenetically distant, they have converged to similar gene content that allows them to supplement *Sulcia* metabolically and to provide their host with the compounds it needs for survival (McCutcheon et al., 2009a; McCutcheon and Moran, 2007; Wu et al., 2006).

Because of their ecological/physiological flexibility, S-symbionts are able to supplement metabolic functions of P-symbionts or even replace them in some cases. Such situation was proved experimentally by replacement of *Buchnera* with *Serratia* in aphids (Koga et al., 2003). Extremely small genome (see Table 2) and the loss of metabolic functions of *Buchnera* from *Cinara cedri* aphid (Perez-Brocal et al., 2006) implies that *Buchnera* is already being complemented by *Serratia*

symbiotica and might be eventually replaced even in natural populations (Lamelas et al., 2008). Another example of replacement is occurrence of primary *Sodalis* sp. within weevils of the genus *Sitophilus*. In this case, *Sodalis* has replaced *Nardonella* symbiont found in other members of the family Dryopthoridae (Conord et al., 2008; Lefevre et al., 2004). Secondary *Sodalis* symbionts that could be source of such replacement were recently found in *Curculio* weevils (Toju et al., 2009).

Remarkable situation of co-symbiosis is reported also from whitefly *Bemisia tabaci* (Gottlieb et al., 2008), where S-symbionts are localized inside bacteriocytes together with P-symbiont (*Portiera aleyrodidarum*). Five different symbionts have been identified within this system (*Arsenophonus*, *Cardinium*, *Hamiltonella*, *Rickettsia*, *Wolbachia*) in various combinations (but maximally two S-symbionts) depending on the *Bemisia* biotype. This finding can indicate influence of symbiotic composition on origin of new species and potential competition between lineages with similar phenotype (e.g. *Hamiltonella* and *Arsenophonus* were never found in the same individual). The possibility of horizontal transfer of genes inside this intracellular ecosystem should also be addressed.

1.2 Diversity and distribution

Two "phylogenetic hotspots" of the insect intracellular symbionts origin are generally recognized, Proteobacteria and Bacteroidetes. Symbiotic lineages originated several times independently within these phyla and there is no doubt about their polyphyly on broader phylogenetic level (Moran et al., 2008). It is uncertain which, if any, features predispose members of these bacterial phyla for the intracellular symbiotic lifestyle. In addition to these hotspots, one symbiotic lineage (*Spiroplasma*) originated within Mollicutes (Gasparich, 2002) and one lineage (*Fritschea*) within Chlamydiae (Everett et al., 2005; Thao et al., 2003).

1.2.1 *Bacteroidetes*

Mutualistic Bacteroidetes symbionts can be found in cockroaches and *Mastotermes* termites – *Blattabacterium* (Bandi et al., 1995); armored scale insects – *Uzinura diaspidicola* (Gruwell et al., 2007); cicadas, leafhoppers, treehoppers, planthoppers and spittlebugs – *Sulcia muelleri* (Moran et al., 2005b). Genus *Cardinium* contains reproductive manipulators invading various arthropods (Zchori-Fein and Perlman, 2004). Male killing in ladybugs is caused by *Flavobacterium* (Hurst et al., 1999).

1.2.2 *Proteobacteria*

Proteobacteria are Gram-negative bacteria comprising the largest and phenotypically the most diverse group among prokaryotes. According to molecular data, they are divided into six subgroups named by the Greek letters alpha through zeta (Gupta, 2000). Insect symbionts originated within alpha (α), beta (β), and gamma (γ) subdivisions.

The most common bacterium found in arthropods and even in nematods is α-proteobacterial genus *Wolbachia*. Other known α-proteobacterial symbionts are of the genus *Rickettsia* found either as primary intranuclear symbiont of booklice (Perotti et al., 2006) or as secondary symbionts of various insects (Sakurai et al., 2005). *Hodgkinia cicadicola* found in cicadas has currently the smallest known genome among insect symbionts (144 kb), but with unusually high 58.4% guanine-cytosine (GC) content (McCutcheon et al., 2009a, b).

Betaproteobacteria are represented by *Tremblaya princeps* - primary symbiont of mealybugs with also atypically high (57.1%) GC content (Baumann et al., 2002). Inside this symbiont, another γ-proteobacterial symbiont occurs. It is the only known bacterial example of extraordinary intracellular symbiosis inside intracellular symbiont (von Dohlen et al., 2001).

Gammaproteobacteria gave rise to numerous lineages of insect symbionts (Table 1) of which almost all originated within Enterobacteriaceae (see chapter 1.5). The first known exception is the P-symbiont of the louse genus *Polyplax* that originated within *Legionella* clade (Hypsa and Krizek, 2007). P-symbionts of psyllids (*Carsonella rudii*) and whiteflies (*Portiera aleyrodidarum*) appear to be related to *Zymobacter palmae* (Spaulding and von Dohlen, 1998; Thao and Baumann, 2004b), a bacterium isolated from palm sap clustering with *Pseudomonas*. This position of *Carsonella* was also observed in phylogenomic analysis of Gammaproteobacteria (Williams et al., 2010) after the removal of AT rich taxa.

Table 1. The diversity of symbiotic Gammaproteobacteria in insects. Unnamed taxa are listed at the bottom of the table. S-symbionts of various psyllids and mealybugs are not included.

Symbiotic lineage	**Hosts**	**Phenotype**
Arsenophonus spp. (Incl. *Riesia* spp. and *Phlomobacter* spp.)	various insects, ticks, spiders, plants	various
Baumannia cicadellinicola	sharpshooters	co-primary
Blochmannia spp.	carpenter ants	primary
Buchnera aphidicola	aphids	primary
Carsonella ruddii	psyllids	primary
Curculioniphilus buchneri	*Curculio* weevils	primary
Hamiltonella defensa *(Consessoris aphidicola, T - type, PABS)*	aphids, whiteflies, psyllids	secondary
Macropleicola spp.	*Macroplea* reed beetles	primary
Nardonella spp.	weevils	primary
Portiera aleyrodidarum	whiteflies	primary
Puchtella pedicinophila	*Pedicinus* lice	primary
Regiella insecticola *(Adiaceo aphidicola, U - type, PAUS)*	aphids	secondary
Serratia symbiotica (R - type, PASS)	aphids	secondary, coprimary
Sodalis spp. (Incl. SOPE, SZPE)	various insects	secondary, primary
Stammerula tephritidis	tephritid flies	primary
Wigglesworthia glossinidia	tsetse flies	primary
Endosymbiont of *Cicadella*	*Cicadella viridis* leafhopper	unknown
Endosymbiont of *Cimex*	bedbugs	secondary
Endosymbiont of Donaciinae	Donaciinae reed beetles (incl. *Macroplea*)	primary
Endosymbionts of *Formica fusca*	*Formica fusca* ant	primary
Endosymbiont of *Haematomyzus*	*Haematomyzus* lice	primary
Endosymbiont of *Haematopinus*	*Haematopinus* lice	primary
Endosymbiont of *Polyplax*	*Polyplax* lice	primary
Endosymbiont of *Pseudolynchia*	*Pseudolynchia canariensis* pigeon louse fly	unknown
Endosymbiont of *Solenopotes*	*Solenopotes* lice	primary
Endosymbiont of *Euscelidius*	*Euscelidius variegatus* leafhopper	parasite

Table 2. Chromosome size and GC content of completely sequenced symbiotic Gammaproteobacteria of insects. If more than one strain has been sequenced from the same host (*Buchnera* strains from *Acyrthosiphon pisum*), only the first published is presented here.

Symbionts and hosts (in brackets)	**GC content**	**Chromosome size**	**Reference**
Arsenophonus nasoniae (*Nasonia vitripennis*)	37.7 %	c. 3.3 Mbp (draft sequence)	(Darby et al., 2010; Wilkes et al., 2010)
Baumannia cicadellinicola (*Homalodisca coagulata*)	33.2 %	686,192 bp	(Wu et al., 2006)
Blochmannia pensylvanicus (*Camponotus pensylvanicus*)	29.6 %	791,654 bp	(Degnan et al., 2005)
Blochmannia floridanus (*Camponotus floridanus*)	27.4 %	705,557 bp	(Gil et al., 2003)
Buchnera aphidicola str. APS (*Acyrthosiphon pisum*)	26.2 %	640,681 bp	(Shigenobu et al., 2000)
Buchnera aphidicola str. Bp (*Baizongia pistaciae*)	25.3 %	615,980 bp	(van Ham et al., 2003)
Buchnera aphidicola str. Cc (*Cinara cedri*)	20 %	416,380 bp	(Perez-Brocal et al., 2006)
Buchnera aphidicola str. Sg (*Schizaphis graminum*)	26.3 %	616,380 bp	(Tamas et al., 2002)
Carsonella ruddii (*Pachypsylla venusta*)	16.5 %	159,662 bp	(Nakabachi et al., 2006)
Hamiltonella defensa str. 5AT (*Acyrthosiphon pisum*)	40.1 %	2,110,331 bp	(Degnan et al., 2009b)
Regiella insecticola str. LSR1 (*Acyrthosiphon pisum*)	42.4 %	2,035,106 bp (slightly larger)	(Degnan et al., 2009a)
Riesia pediculicola	unpublished	574,390 bp	unpublished
Sodalis glossinidius (*Glossina morsitans*)	54.7 %	4,171,146 bp	(Toh et al., 2006)
Wigglesworthia glossinidia (*Glossina brevipalpis*)	22 %	697,724 bp	(Akman et al., 2002)

1.3 Phylogenetics

1.3.1 *Evolutionary forces shaping symbiotic genomes*

Molecular evolution of DNA sequences is strongly dependent on population genetics formed by selection, mutation, random genetic drift and gene flow. For symbiotic bacteria, as well as for other host-restricted vertically transferred bacteria, extremely small effective population size (N_e) accompanied by strong bottlenecks at the stage of inoculation of new hosts and reduced purifying selection result in elevated effect of genetic drift (Mira and Moran, 2002; Wernegreen and Moran, 1999). Accordingly, rate of fixation of mildly deleterious and deleterious mutations is increased (Lambert and Moran, 1998); unessential genes become inactive pseudogenes and are subsequently lost. Mutation rate is consequently further raised by the loss of DNA repair and recombination genes (Dale et al., 2003). These mutations also cause low thermal stability of proteins, which have to be buffered by elevated expression of heat shock proteins (Fares et al., 2002). Irreversibility of these processes is caused by Muller's ratchet (Moran, 1996) working within small populations with effectively no recombination. The ratchet-like one-way degradation can be only slightly slowed down by selection (Allen et al., 2009).

1.3.2 *Long-branch attraction (LBA)*

As described above, phylogenetically independent (and some related) bacterial lineages have experienced parallel changes during their associations and these convergences between unrelated lineages produce the main source of artifacts introduced into phylogenetic analyses. The most important (for phylogenetic analyses) shared features are extremely small genomes with biased nucleotide composition favoring adenine-thymine bases (see Table 2) and rapid sequence evolution resulting in accelerated rates of amino acid substitutions. The logical consequence of these traits is that primary symbionts essentially always appear as long branches in phylogenetic trees.

Long-branch attraction (LBA) is a situation in which taxa with rapid evolution rates are inferred as related due to their convergent changes (Bergsten, 2005). Since the number of different states of each character in the sequence is finite and considerably low (4 for nucleotides, 20 for amino acids) and mutations are not entirely random (because of physical and chemical constraints of coding substances), occurrence of the same independently acquired character is quite common (Felsenstein, 2003).

LBA affects all phylogenetic methods. Maximum Parsimony (MP) is the most frequently influenced (Aguinaldo et al., 1997; Anderson and Swofford, 2004; Kennedy et al., 1999) as it interprets homoplasy as a synapomorphy, but Maximum likelihood (ML) and Bayesian inference (BI) with various models used (from JC69 to GTR+I+Γ) are also not completely immune (Schwarz et al.,

2004). There have been a lot of methods suggested and used to deal with LBA artifacts. The necessary procedure for any phylogenetic analysis is usage of various methods (MP, ML, and BI) and configuration. Especially, rate heterogeneity modeled by gamma distribution in ML has been proven to reduce LBA (Nikaido et al., 2003; Sullivan and Swofford, 1997). For distance methods, paralinear or log-determinant (LogDet) distances (Lake, 1994; Lockhart et al., 1994) were designed to eliminate the effect of nucleotide composition bias and even of amino acid bias (Thollesson, 2004). Standard phylogenetic methods assume homogeneity of nucleotide/amino acid composition and substitution rates among taxa. In real data, e.g. in phylogenomic study of γ-Proteobacteria, 73% of used taxa violate this assumption (Williams et al., 2010). Galtier and Gouy distance (1995) was designed to deal with heterogeneous base composition and this approach was quickly modified and introduced into the ML field as a nonhomogeneous model of DNA sequence evolution (Galtier and Gouy, 1998). Study by Herbeck et al. (2005) used this model and showed polyphyly of P-symbionts on 16S rRNA and *GroEL* genes.

Long-branch taxa exclusion and multiple analyses with only one of these taxa included are also powerful tool (Williams et al., 2010), as well as adding more taxa to break the long branches (Hendy and Penny, 1989). Characters can be weighted according to their evolutionary rate or even excluded as third positions in codons (Sullivan and Swofford, 1997) or AT biased amino acids (FYMINK) (Comas et al., 2007a). Transitions/transversions can be also weighted in MP approach (Reyes et al., 1999). Slow-fast analysis (Bapteste et al., 2008; Brinkmann and Philippe, 1999; Kostka et al., 2008) is very strong method for inferring deeper phylogeny affected by LBA. This method omits in several steps the fastest evolving positions and thus increases phylogenetic information/noise ratio. Similar approach is based on removing saturated amino acid positions by program called ASaturA (Van de Peer et al., 2002). Adding and combining more data of various kinds and analyzing with several outgroups (Xiang et al., 2002) also helps to overcome LBA artifacts.

1.3.3 *Horizontal (lateral) gene transfer (HGT)*

Horizontal gene transfer (HGT) is one of the most serious problems in phylogenetic analyses, especially in bacteria, in which gene flow is maintained by conjugation, plasmids, bacteriophages and even by incorporating genes from other dead organisms (Pennisi, 2004). Many studies have estimated the frequency of HGT (Cortez et al., 2009; Dagan and Martin, 2007; Ge et al., 2005; Lerat et al., 2003; Susko et al., 2006): while under the current state of knowledge it is difficult to draw any clear conclusion, HGT seems to be extensive among free living Enterobacteriaceae (ancestors of symbiotic bacteria).

For symbiotic bacteria, the proportion of horizontally transferred genes is still unclear, but higher numbers of horizontally transferred genes are found in dynamic genomes of S-symbionts than in the genomes of P-symbionts (Degnan et al., 2009a; Degnan et al., 2009b; Toh et al., 2006;

Wilkes et al., 2010). If the theory that some P-symbionts originated from S-symbionts (Wernegreen et al., 2009) is correct, horizontally transferred genes in P-symbionts are now either lost or strongly overwritten by rapid mutational rate. Moreover, detection of HGT will be complicated if these genes are transferred from another Enterobacteriaceae species such as *Photorhabus*, *Salmonella* or *Yersinia* as is quite usual in the above mentioned examples.

Horizontal transfer of symbiont genes to the host chromosomes is difficult to identify (Nikoh and Nakabachi, 2009). However, completely sequenced genomes of insects provide opportunity to find these genes as was shown in study of Nikoh et al. (2010) on only 12 bacterial genes transferred to its aphid host. These genes were transferred from *Buchnera* and unidentified α-Proteobacteria and deny hypothesis that the genes missing in the reduced *Buchnera* genome were transferred to the host nuclear chromosome.

1.3.4 *Ribosomal genes*

Until recently, small subunit ribosomal RNA (SSU rRNA) had been the most frequently used gene for inferring organismal phylogeny. Several properties make it a good phylogenetic marker, such as a universal presence, evolutionary conservation and rare horizontal transfer (Jain et al., 1999; Woese, 1987). However, it has some significant disadvantages in respect to bacterial phylogeny, e.g. low number of informative sites causing different topologies when the evolutionary model or method is changed (Comas et al., 2007a; Herbeck et al., 2005), occurrence of several copies of ribosomal operon in the genome, problematic alignment and even rare horizontal transfer (Ueda et al., 1999; Yap et al., 1999).

In Proteobacterial phylogeny, information provided by 16S rRNA gene is incapable of resolving deeper phylogenetic relationships (Williams et al., 2010; Williams et al., 2007), so that it is currently used as a first-choice tool to characterize undescribed taxa (Kolsch et al., 2009; Toju et al., 2009) or in combination with 23S rRNA gene (Downie and Gullan, 2005; Gruwell et al., 2007; Thao and Baumann, 2004a, b; Thao et al., 2002) or with protein coding genes in multi gene analyses.

1.3.5 *Protein coding genes*

Enterobacteriaceae, as a group containing many medically important species, is extensively studied bacterial family. Phylogenetic information obtained from SSU rRNA gene is insufficient for genus-level phylogenetic analyses (Naum et al., 2008) and new genes (mostly housekeeping) are thus tested and used for inferring Enterobacteriaceae phylogeny: *gapDH* (Brown et al., 2000), *gyrB* (Dauga, 2002), *gapA*, *groEL*, *gyrA*, *ompA*, and *pg* (Wertz et al., 2003), *tuf* and *atpD* (Paradis et al., 2005), *dnaJ* (Pham et al., 2007), non coding *oriC* (Roggenkamp, 2007), *gyrB*, *rpoB*, *atpD* and *infB* (Brady et al., 2008), *gapA*, *phoP*, *mdh* and *recA* (McQuiston et al., 2008), *recN*, *rpoA* and *thdF* (Kuhnert et al., 2009).

For insect symbiotic bacteria phylogeny, various protein coding genes have been used as an alternative and complementing (to ribosomal genes) source of information: *atpAGD* and *rpoBC* (Thao et al., 2001), *atpD* and *aroQ/pheA* (Moya et al., 2002), *groEL*, *gidA*, and *rpsB* (Degnan et al., 2004), *gyrB* and *recA* (Moran et al., 2005a), *fusA* (Fukatsu et al., 2007), *groEL*, *ftsA*, *secA*, *gapA*, *nagB*, *rplB* (Sorfova et al., 2008). Multi-gene approach allows increasing information (Wertz et al., 2003), but also leads to possible problems with horizontal transfer (McQuiston et al., 2008), missing genes (Mignard and Flandrois, 2008) or paralogous genes (Horiike et al., 2009).

1.4 Phylogenomics

With an increasing number of completely sequenced genomes, phylogenomic approach (Delsuc et al., 2005) has allowed to infer species trees not affected by the most frequent problems associated with single gene phylogenies. For dynamic bacterial genomes with common horizontal gene transfer, possibility to work with all the present genes is appreciated. It has been proved that not all γ-proteobacterial core genes have a similar phylogenetic signal and that they rarely favor a single tree (Susko et al., 2006).

1.4.1 *Genome structure data*

The sequencing of complete genomes of many bacterial species provided source of phylogenetic information coded in genome structure. There are various types of these idiosyncratic markers (Hypsa and Novakova, 2008), but the most often used marker in bacterial phylogeny is arrangement of the genes within the genome. Gene order may be affected by inversions and translocations, gene removals and insertions, by xenologous genes or by partial duplications of a genome (Belda et al., 2005)

In symbiotic bacteria, gene order was originally used during pregenomics era on *dnaA* region by Hassan et al. (1996) to confirm phylogenetic position of *Buchnera* within Enterobacteriaceae. In the published gene order studies, symbiotic bacteria were either not included at all (Comas et al., 2006), or only *Buchnera* was present in the dataset (Korbel et al., 2002). The only study dealing specifically with P-symbionts (Belda et al., 2005) included *Buchnera* (APS, Sg, Bp), *Blochmannia floridanus* and *Wigglesworthia*. Inferred phylogenetic tree presents *Buchnera* in paraphyly with respect to *Wigglesworthia* and places *Blochmannia* as a closer relative to *Escherichia*/*Shigella*/*Salmonella* (ESS) clade than *Yersinia pestis*. These methods are not biased by GC content, but may be strongly affected by the number of shared orthologous genes, for example *Wigglesworthia* shares only 69% of its coding sequences with *Buchnera* (Akman et al., 2002).

1.4.2 *Genome sequence data*

Although amino acid sequences are not free from the base composition artifacts (Foster and Hickey, 1999; Singer and Hickey, 2000), they are much more frequently used for phylogenomic studies than nucleotide data. It should be noted that symbiotic bacteria phylogeny inferred from amino acid information may be strongly influenced by other artifacts, particularly by amino acid composition convergences between unrelated symbiotic bacteria within hosts feeding on similar diet (Lopez-Sanchez et al., 2009).

There are two ways of dealing with multiple gene alignments. First, alignments can be concatenated into a supermatrix and analyzed together to increase phylogenetic signal. Second, all single-gene alignments can be analyzed independently and the acquired topologies are then integrated into a single consensus tree. If the initial dataset does not include the same taxa in all the gene trees, supertree can be constructed on the basis of overlapping topologies (Daubin et al., 2002).

Monophyly/polyphyly of *Wigglesworthia* - *Buchnera* clade is the most widely and exhaustively discussed question in symbiotic bacteria research. Phylogenomic studies frequently imply monophyly of P-symbionts (Canback et al., 2004; Ciccarelli et al., 2006; Comas et al., 2007a, b; Gao et al., 2009; Lerat et al., 2003; Wu and Eisen, 2008), but it should be taken into account that they are usually based on largely incomplete taxon sampling. Secondary symbionts and plant pathogens that break the P-symbionts monophyly in the analysis using nonhomogeneous model (Herbeck et al., 2005) cannot be included because their genomes have not yet been available. Nevertheless, the most rigorous analysis by Comas et al. (2007a) compared supermatrix, supertree and consensus approaches and demonstrated monophyly of P-symbionts (*Buchnera*/*Blochmannia*/*Wigglesworthia*). It should be mentioned that monophyly of these taxa does not necessarily mean unique origin from free-living bacterium, but rather origin of P-symbionts within closely related enterobacterial lineages (Hypsa and Novakova, 2008).

Recently, Williams et al. (2010) performed a "telescoping" multi-protein phylogenomic analysis of 104 γ-Proteobacterial genomes. Phylogeny of Enterobacteriaceae has been difficult to resolve. LBA was dealt with removing the AT rich genomes and performing separate analyses with only one P-symbiont retained. *Photorhabdus* was identified as disruptor of the root and *Sodalis* was identified as attractor of P-symbionts (*Baumannia*, *Blochmannia* and *Wigglesworthia*). The P-symbionts sensitive to the presence of *Sodalis* were therefore excluded and the final Enterobacteriaceae tree is presented with *Buchnera* as root and *Sodalis* within free-living bacteria (in paraphyly with *Pectobacterium*). This study shows that appropriate taxon sampling is essential for reconstructing phylogeny of Enterobacteriaceae; taxa that were scarcely used before (*Pectobacterium, Photorhabdus, Serratia, and Sodalis*) help to resolve deeper phylogeny, but also introduce new questions.

1.5 Enterobacteriaceae

Enterobacteriaceae comprises species that have been studied extensively for decades and became model for research of intracellular symbiosis (*Buchnera*, *Blochmannia*, *Wiglesworthia*), species that have been broadly studied only for last few years (e.g. *Baumannia*, *Hamiltonella*, *Reggiella*) and lot of almost neglected symbiotic lineages (secondary symbionts of mealybugs and psyllids or primary symbionts of lice). Currently, it is well established that insect symbionts originated many times independently within Enterobacteriaceae: *Serratia symbiotica* originated within *Serratia* genus (maybe more than once, see Burke et al., 2009), *Stammerula tephritidis* originated within *Erwinia* genus (Mazzon et al., 2008), *Arsenophonus* clade originated within *Photorhabdus*/*Proteus* clade (Gherna et al., 1991), *Hamiltonella* and *Regiella* are sister species probably clustering with *Yersinia* (Degnan et al., 2009b). It is no surprise that symbiotic species with well established phylogenetic position are usually S-symbionts, for P-symbionts affected by genome degeneration, phylogenetic studies are often incongruent.

Enterobacteriaceae do not contain only intracellular, but also much less studied extracellular symbionts. The only extracellular enterobacterial symbionts formally described are the gut symbionts of plataspid stinkbugs (*Ishikawaella capsulata*; Hosokawa et al., 2006) and acanthosomatid stinkbugs (*Rosenkranzia clausaccus*) (Kikuchi et al., 2009). Their phylogenetic position close to *Buchnera* clade can imply potential origin of intracellular symbionts from gut symbionts, and their genetic traits (high AT content, small genomes, accelerated molecular evolution) demonstrate that the intracellular habitat is not responsible for these characteristics. Candidate mechanism is attenuated purifying selection caused by small effective population size and strong bottleneck effect (Kikuchi, 2009).

This thesis is focused on phylogeny of insect symbionts within Enterobacteriaceae. Due to the problems discussed above, a reliable phylogeny of these highly modified organisms can only be obtained by phylogenomic or at least multi-gene approach. However, while the genome sequencing evolves rapidly, it is impossible to sequence genomes of all symbiotic bacteria for a complex phylogenomic analyses. Paradoxically, the problem is not caused by inefficiency of sequencing methods but by the difficulties with cultivation/separation of the symbionts. Multi-gene phylogenetic approach is thus a trade-off to be used until the sequencing of insects together with their symbionts becomes more frequent. The main purpose of this study is to extend the multi-gene dataset of Enterobacteriaceae with additional insect symbionts with two main goals: to investigate behaviour of the P-symbiont topologies under these new conditions and to explore phylogenetic position of the newly sequenced lineages. This approach is based on the conviction that more extensive sampling could help to resolve the widely discussed question of the monophyly/polyphyly of *Buchnera*/*Ishikawaella* + *Baumannia*/*Blochmannia*/*Wigglesworthia* clade.

As the additional lineages of the P-symbionts, I chose symbiotic bacteria from four lice species. Generally, the lice symbionts are undoubtedly polyphyletic with at least four known independent origins: *Pediculus* lice acquired *Riesia* clustering within *Arsenophonus* clade (Allen et al., 2007), *Columbicola columbae* lice acquired *Sodalis* sp. symbiont (Fukatsu et al., 2007) and *Polyplax* lice harbor *Legionella* sp. (Hypsa and Krizek, 2007). Genera *Haematomyzus*, *Haematopinus*, *Solenopotes* and *Pedicinus* harbor symbionts clustering within *Blochmannia*/*Wigglesworthia* clade (Fukatsu et al., 2009; Hypsa and Krizek, 2007) and have been selected in this work for the dataset extension and multi-locus phylogenetic analysis.

2. Aims

- To identify genes that will be used for multi-locus phylogenetic analyses of symbiotic bacteria within Enterobacteriaceae.
- To design primers and optimize PCR conditions.
- To sequence identified genes from primary symbionts of louse (genera *Haematopinus* and *Haematomyzus*) and all the other available symbionts clustering with enteric bacteria.
- To screen for new unidentified primary symbionts and to distinguish multiple symbionts occurrence in mixed bacterial DNA by using TGGE method.
- To analyze the phylogeny of enterobacteria with particular focus on primary symbionts

3. Materials and methods

3.1 Samples and DNA extraction

The insect species used for DNA extraction were obtained from various sources: *Haematopinus suis and H. eurysternus* (laboratory collection), *Trachys minuta, Polydrusus formosus, Rhampus pubicarius* (collected by M. Volf), *Haematomyzus elephantis* (collected by the author at the Prague ZOO), *Planococcus citri* (collected by the author in České Budějovice). The insect samples with their taxonomy are summarized in Table 3. Insects were homogenized in liquid nitrogen using pestle and the total genomic DNA was isolated by DNEasy Tissue Kit (QIAGEN) according to manufactures' instructions. Previously isolated DNA of *Solenopotes capillatus* (Hypsa and Krizek, 2007) was acquired from laboratory sources.

Table 3. List of insect samples with their taxonomy.

Species	**Taxonomy (Order: Suborder: Family)**
Haematomyzus elephantis	Pthiraptera: Rhynchopthirina: Haematomyzidae
Haematopinus eurysternus	Pthiraptera: Anoplura: Haematopinidae
Haematopinus suis	Pthiraptera: Anoplura: Haematopinidae
Polydrusus formosus	Coleoptera: Polyphaga: Curculionidae
Planococcus citri	Hemiptera: Sternorrhyncha: Pseudococcidae
Rhampus pulicarius	Coleoptera: Polyphaga: Curculionidae
Solenopotes capillatus	Pthiraptera: Anoplura: Linognathidae
Trachys minuta	Coleoptera: Polyphaga: Buprestidae

3.2 Polymerase chain reaction (PCR)

Degenerated primers were designed manually in BioEdit program (Hall, 1999) on the basis of Enterobacteriaceae dataset, checked for occurrence of hairpins and dimers in OligoAnalyzer 3.1 (http://eu.idtdna.com/analyzer/Applications/OligoAnalyzer/) and *in silico* PCR amplification (http://insilico.ehu.es/PCR/) was carried out against available symbionts. Primers for 16S rRNA gene with different specifity were adopted from previous publications (see Table 4 for all used primers).

Despite of using different sets of primers and cycling conditions, reaction mixtures were always based on the same composition. The mixture contained 19 µl of ddH_2 , 2.5 µl of 10X High Fidelity PCR Buffer supplemented with 15mM $MgCl_2$ (Fermentas), 0.5 µl of 10mM dNTP mix (Invitrogen), 0.5 µl of each primer (25 µmol), 0.25 µl of High Fidelity PCR enzyme mix (Fermentas) and 1.75 µl of total genomic DNA. Annealing temperatures of newly designed primers were determined by gradient PCR (temperature range 40 – 55 °C). Amplification was carried out using

thermocycler Xp (Bioer). Program consisted of 2 minutes of initial denaturation (94 °C) followed by 30 cycles of denaturation (50 sec, 94 °C), primers annealing (50 sec, temperature determined by gradient) and extension (72 °C, 1 min/1 kb of product); 10 minutes of final extension (72 °C) terminated the program.

Fifty microliters of PCR samples for TGGE (see 3.2.2) were prepared using forward primer F968 with GC clamp and reverse primer R1060 (Table 4), these primers are designed to amplify hypervariable region V6-V8 of 16S rRNA gene. Samples were concentrated by heated vacuum centrifugation (Jouan RC 1009, 37 °C, 1 hour).

Table 4. List of primers used for polymerase chain reactions displayed with their sequences, *E.coli* (NC_010473) ortologous genes positions, lengths of amplified products and references to original publications.

Designation (*E. coli* position)	**Sequence (5´- 3´)**	**Product**	**Reference**
16S forward (45-61)	GCTTAACACATGCAAG	16S rDNA c. 1180 bp	(O'Neill et al., 1992)
16S reverse (1242-1227)	CCATTGTAGCACGTG T		
F40 forward (40-57)	GCGGCAAGCCTAACACAT	16S rDNA c. 1045 bp	(Hypsa and Krizek, 2007)
R1060 reverse (1060-1085)	CTTAACCCAACATTTCTCAAC ACGAG		
F968 forward (968-985) (GC clamp in italics)	*CGCCCGGGGCGCGCCCCGGG CGGGGCGGGGGCACGGGGGG* AACGCGAAGAACCTTAC	16S rDNA c. 449 bp	(Nubel et al., 1996)
R1401 reverse (1401-1417)	CGGTGTGTACAAGACCC		
GroEL forward (7-26)	GCTAAAGAIGTIAARTTYGG	GroEL 1643 bp	this study
GroEL reverse (1630 -1650)	TTACATCATICCICCCATICC		
LepA forward (247 – 266)	GAYTTYTCATAYGAAGTWTC	LepA 1427 bp	this study
LepA reverse (1654 – 1673)	ACRTCNCCRCCRTARCAYTT		
AceE forward (564-584)	GCAATTTCCIACIGTWTCWAT	AceE 1932 bp	this study
AceE reverse (2473-2495)	CGACCAAAICCITCIGTICCIAA		

3.2.1 *Agarose gel electrophoresis*

Agarose gel electrophoresis was used to separate fragments acquired by PCR and to confirm the presence of insert in isolated plasmids digested by EcoRI. PCR samples (25 µl) were mixed with 5 µl of loading dye (see Appendix 1); plasmid samples (10 µl) were mixed with 2 µl of loading dye. Mixed samples were loaded on 0.75-1.5 % agarose gel. DNA Ladder (GeneRuler^{TM} 1kb, Fermentas)

was used to determine the size of DNA fragments. Gels were examined under UV transillluminator TI3 (Biometra). PCR products were excised and isolated from agarose by PureLink™ Quick Gel Extraction Kit (Invitrogen) according to manufacturer instructions.

3.2.2 *Temperature gradient gel electrophoresis (TGGE)*

Conventional nucleic acid electrophoresis separates molecules according to their size; temperature gradient electrophoresis adds new parameter for separation, namely the melting behavior of a molecule. The melting behavior is determined by primary sequence and secondary/tertiary structure of the molecule, thus DNA fragments of identical size but different sequence can be separated. During the electrophoresis, samples migrate along a temperature gradient. At a certain temperature, DNA fragments (AT richer) start to melt and their migration is slowed down compared to GC richer fragments. GC rich forward primer is used to clamp partial single strands of DNA and avoid complete denaturation. Currently, this method is used especially for detection of mutations (Balogh et al., 2004; Ke et al., 1993) and diversity analysis of complex samples (Beier et al., 2008; Brummer et al., 2003; Wei et al., 2007).

Glass sandwich with hydrophilic TGGE MAXI Polybond film (Biometra) for parallel gels was assembled according to TGGE MAXI System Manual. Gel solution was prepared for 50 ml according to Appendix 2. Solution was stirred at 50 °C to completely dissolve urea and cooled down to room temperature. Polymerization was initiated with 80 µl of 10% APS (ammonium persulfate) and 110 µl of TEMED (tetramethylethylendiamine) and gel solution was then poured into the glass sandwich. Gel was polymerized for approximately 3 hours at room temperature. TGGE MAXI System (Biometra) electrophoresis unit was prepared – buffer tanks were filled with 500 ml of running buffer (1x TAE), buffer wicks were soaked, 1.5 ml of thermal solution (0.1% Triton, Sigma) was applied on the thermoblock, gel was placed on the thermoblock, applicator strip was applied, gel was covered with TGGE MAXI cover film (Biometra) and pre-soaked buffer wicks were attached on top and bottom of the gel. Five microliters of each previously prepared sample (see PCR) were mixed with 5 µl of loading buffer (see Appendix 1) and loaded. Gel cover plate was attached, safety lid was closed and electrophoretic run was started. TGGE was set to 35-55 °C temperature gradient, 100 V, 18 hours. Gel was stained with ethidium bromide solution (0.5 µg/ml ethidium bromide in 1 X TAE) for 1 hour and analyzed under UV transilluminator TI3 (Biometra). Separated fragments were extracted from the gel using scalpel, incubated in 20 µl of ddH_2O overnight at 37 °C and used for PCR re-amplification with forward primer F968 without GC clamp and reverse primer R1401 (Table 4).

3.3 Cloning and sequencing

PCR products were cloned into pGEM®–T Easy vector (Promega). Ligation mixture was mixed as follows: 0.5 µl vector, 2.5 µl 2x rapid ligation buffer (Promega), 1.5 µl PCR product, 0.5 µl T4 ligase (Promega). Ligation mixture was incubated for 12 hours in 4 °C or 2 hours in 25 °C. Chemically competent cells *Escherichia coli* DH5α (25 µl) were thawed on ice; ligation reaction was added and mixture was incubated for 30 minutes on ice. Transformation was done by heat shock (42 °C) in Digital Dry Bath (Labnet) for 50 seconds. Cells were cooled on ice for 2 minutes, then 125 µl of tempered (37 °C) SOC medium was added and bacteria were cultivated 1.5 hour in shaker incubator (37 °C, 210 RPM). Transformed cells were spread on preheated 1% agar plates with ampicilin (see Appendix 3) coated with 40 µl of X-gal (40 mg/ml) (5–bromo–4–chloro–3–inodyl–beta–D–galactosidase, Serva) and 4 µl of IPTG (isopropyl thiogalactoside, Fermentas). Petri dishes were incubated at 37 °C overnight. Blue-white screening detection method was used for selection of successful ligations in vector. The basic principle of this method is that the vector encodes the α-subunit of Lac Z protein; PCR product is inserted into the multiple cloning site (MCS) within this protein and makes it nonfunctional. Lac Z protein forms β-galactosidase - enzyme which hydrolyzes X-gal and causes characteristic bright blue color of the colonies. Blue colonies thus do not possess vector containing insert. White colonies were inoculated into tubes with LB medium (5 ml) and ampicilin (100 mg/l). Tubes were incubated in a shaker for 12 hours (37 °C, 210 RPM). Plasmids were isolated by QIAprep®Spin Miniprep Kit (QIAGEN). Presence of insert was confirmed by plasmid hydrolysis using restriction endonuclease. One microliter of 10X REact® 3 buffer (Invitrogen), 0.3 µl of EcoRI (Invitrogen), 2 µl of isolated plasmid and 6.7 µl of ddH_2O were mixed and incubated for 1.5 hour (37 °C). After restriction digest, samples were loaded on 1% agarose gel (see electrophoresis) where length of particular inserts was confirmed. DNA Sanger sequencing was carried out on ABI Automatic Sequencer 3730XL by Macrogen Inc. (Geumchun-gu-Seoul, Korea) using universal SP6 and T7 primers and designed inner primers (for sequences see Appendix 4).

3.4 Sequence identification and assembly

Quality of the raw sequences was checked in Sequence Scanner 1.0 (Aplied Biosystems). Vector contamination identified by vecscreen (BLAST, NCBI) was removed in EditSeq (DNASTAR). SeqMan (DNASTAR) was used for contig assembly. Forward and reverse primers were removed from the sequences in BioEdit (Hall, 1999) to prevent introduction of an artificial signal in phylogenetic analyses. Reading frame was confirmed by ORF Finder (http://www.ncbi.nlm.nih.gov/projects/gorf) and GC content of sequences was determined in BioEdit (Hall, 1999). Initial sequence identification was performed by NCBI BlastSearch (http://www.ncbi.nlm.nih.gov/Blast.cgi) using either nucleotide or protein query.

3.5 Matrices and multiple alignments

Housekeeping genes were chosen as a source of phylogenetic information. Rules for selection were as follows: one copy of the gene in genomes of Enterobacteriaceae, universal presence in both free-living and symbiotic bacteria, known function assuring correct annotation, unpublished/unknown horizontal transfer. Gene for translation elongation factor (*EF-Tu*) was available in GenBank (http://www.ncbi.nlm.nih.gov/) from previous study (Hypsa and Krizek, 2007). This gene has usually two copies in enterobacterial species, but these copies are coelvolving via gene conversion and clustering together in phylogenetic trees. Nevertheless, the phylogeny based on *EF-Tu* was computed with both paralogs present in the dataset and the copy with shorter branch was chosen for further analyses. *Yersinia* species are known to possess very divergent copies of this gene (Isabel et al., 2008). I excluded this taxon because of different phylogenetic positions of paralogs were observed also in my analyses.

All matrices of protein-coding sequences were based on Enterobacteriaceae taxa with available genomes in GenBank (see Appendix 5). One dataset (*GroEL* max) was complemented with sequences of symbionts available only for this gene. For 16S rRNA gene phylogeny, extensive taxon sampling was done on the basis of published studies. Heuristic Blast search is incapable to find all related sequences, which can be shown e.g. on total omission of symbiotic enterobacteria of leeches in all phylogenetic analyses of symbionts, although they cluster within P-symbionts (Kikuchi and Fukatsu, 2002; Perkins et al., 2005). A ML guide tree (GTR+I+ Γ) using Phyml v3.0 (Guindon and Gascuel, 2003) was used to filter only symbionts that cluster within Sodalis/P-symbionts clade.

The reduced 16S rDNA datatset was aligned in the Mafft program based on iterative refinement (Katoh et al., 2009; Katoh et al., 2002) (http://align.bmr.kyushu-u.ac.jp/mafft/online/server/index.html), using the E-INS-i strategy with default parameters. Poorly aligned positions were eliminated in Gblocks program (Castresana, 2000) with tree different levels of stringency. First, less strict flanking positions, smaller final blocks and gap positions within the final blocks were allowed. Second, only gaps were allowed within final blocks, and third, alignment was the same as previous, but with manually removed long gaps.

All protein coding sequences were aligned after translating into amino acids in Clustal W (Thompson et al., 1994) implemented in BioEdit (Hall, 1999) using default parameters. Ambiguously aligned positions were excluded from codon sequences in Gblocks (Castresana, 2000; Talavera and Castresana, 2007) with gaps allowed within the final blocks and alignments were checked manually in BioEdit (Hall, 1999). If there were long gaps at the beginning and end of my partial sequences (*AceE, LepA*), alignments were also trimmed and effect on the tree topology was tested. Since this trimming did not affect the gene tree topology, longer alignments were further used and all gaps were treated as missing characters. Matrices (nucleotides and amino acids) were saved in interleaved nexus

and phylip formats. Lengths of alignments and number of conserved/parsimony-informative sites were determined in Mega 4.1 (Tamura et al., 2007). Protein coding genes were concatenated in SeaView v4 (Gouy et al., 2010).

3.6 Phylogenetic analyses

Four different methods were used to infer phylogenetic trees (for workflow see Table 5) from single-gene and concatenated nucleotide and amino acid matrices: maximum parsimony (MP) and distance methods (LogDet) implemented in PAUP* 4.0 (Swofford, 2002), maximum likelihood (ML) using PhyML v3.0 (Guindon et al., 2009; Guindon and Gascuel, 2003) and Phylo_win 2.0 (Galtier et al., 1996) and Bayesian inference (BI) using MrBayes 3.1.2 (Huelsenbeck and Ronquist, 2001). If possible, parallel versions of these programs were used on MetaCentrum computational clusters (http://meta.cesnet.cz/cms/opencms/en/index.html) to reduce computational time. Incongruence-length difference test (ILD) (Farris et al., 1995) implemented in PAUP* 4.0 (Swofford, 2002) was used to test homogeneity (nreps=10000) of concatenated genes.

For likelihood-based methods, evolutionary models of DNA and protein evolution were selected in jModelTest (Posada, 2009) and ProtTest (Abascal et al., 2005). For nucleotide sequences, General Time Reversible (GTR) model (Rodriguez et al., 1990) was used with estimated proportion of invariable sites (I) and heterogeneity of evolutionary rates modeled by 6 substitution rate categories of gamma distribution (Γ) with gamma shape parameter (alpha) estimated from the dataset. In addition, nonhomogeneous model of sequence evolution (T92+Γ+varGC) implemented in Phylo_win (Galtier et al., 1996) was used to infer ML trees. For amino acid data, LG model (Le and Gascuel, 2008) in ML analyses and WAG model (Whelan and Goldman, 2001) in BI analyses were used with estimated proportion of invariable sites (I) and variability of evolutionary rates across sites modeled by 6 substitution rate categories of gamma distribution (Γ) with gamma shape parameter (alpha) estimated from the dataset. A codon-based model of nucleotide substitution (Goldman and Yang, 1994) implemented in MrBayes (Huelsenbeck and Ronquist, 2001) was also used for protein-coding sequences with GTR+I+Γ modeling nucleotide changes. Subtree pruning and regrafting (SPR) tree searching algorithm was used in ML analyses to decrease the chance of getting stuck in a local maximum of the likelihood function. Bootstrap resamplings (100) were performed to test reliability of inferred trees.

In MrBayes program, following parameters were set for the MCMC sampler: nruns= 2, nchains=4 (three heated, one cold), ngen = 3000000, samplefreq= 100 and printfreq = 100. Markov chain Monte Carlo (MCMC) convergence and number of trees to discard was explored by AWTY (Nylander et al., 2008) web based program (http://ceb.csit.fsu.edu/awty). Trees were summarized by sumt command and burnin trees were discarded.

The MP analyses were carried out by heuristic search with TBR (tree-bisection and reconnection) branch swapping algorithm and 100 replicates of random sequence addition. LogDet distances were computed with invariable positions excluded.

Slow-fast analysis was performed in SlowFaster (Kostka et al., 2008), monophyletic phyla were specified on rooted ML tree. Produced datasets with increased phylogenetic signal/noise ratio were analyzed by different methods.

Computed trees were viewed in TreeView (Page, 1996) and rooted by outgroups.

Table 5. Phylogenetic analyses workflow used for single-gene and concatenated datasets.

Nucleotides			**Amino-acids**	
All positions	**2nd/3rd positions excluded**	**Slow-fasted**	**All**	**Slow-fasted**
ML (GTR+I+ Γ)	ML (GTR+I+Γ)	ML (GTR+I+Γ)	ML (LG+I+Γ)	ML (LG+I+Γ)
ML (T92+Γ+varGC)	BI (GTR+I+Γ)	BI (GTR+I+Γ)	BI (Wag+I+Γ)	BI (Wag+I+Γ)
BI (GTR+I+Γ)	LogDet	LogDet	MP	MP
BI (Codon model)	MP	MP		
LogDet				
MP				

4. Results

4.1 Sequences

I have acquired 6 partial sequences encoding small ribosomal subunit gene and 8 partial sequences of housekeeping genes from intracellular symbiotic bacteria of 8 insects. Sequenced protein coding genes were as follows: chaperonine *GroEL*, GTP-binding protein *LepA* and pyruvate dehydrogenase subunit E *AceE*. All the sequences with their lengths and GC contents are summarized in Table 6. TGGE analyses did not confirm presence of S-symbionts in tested lice, except for diverse lineages of *Wolbachia* previously identified in lice (Covacin and Barker, 2007; Kyei-Poku et al., 2005). First weevil species (*Polydrusus formosus*) was negative for symbiotic bacteria and two sequences of symbiotic bacteria were acquired from the second tested species (*Rhampus pulicarius*). From *Solenopotes capillatus*, only 16S rRNA gene was acquired.

Table 6. Sequenced genes of symbiotic bacteria of insects with gene fragment length and GC content.

Insect species	**Symbiont**	**Gene**	**Length (bp)**	**GC content (%)**
Haematomyzus elephantis	primary	*GroEL*	1587	35.35
		LepA	1386	24.53
Haematopinus eurysternus	primary	*GroEL*	1602	36.20
		LepA	1386	24.17
Haematopinus suis	primary	*GroEL*	1602	36.77
		AceE	1886	27.31
Planococcus citri	S-cosymbiont	*GroEL*	1584	47.16
		16S rRNA	1328	51.36
	primary *Tremblaya princeps*	16S rRNA	396	54.8
Rhampus pulicarius	secondary	*GroEL*	1590	56.23
		16S rRNA	1013	56.37
	Wolbachia	16S rRNA	973	45.63
Solenopotes capillatus	primary	16S rRNA	1030	49.13
Trachys minuta	*Wolbachia*	16S rRNA	1291	47.05

4.2 Multiple alignments

Lengths of edited multiple alignments and number of parsimony-informative/conserved sites for nucleotide and amino-acid data are presented in Table 7. In fact, lower number of parsimony informative sites is available for inferring origin of symbiotic bacteria from free-living ancestors (see Figure 1). Because of house-keeping nature of chosen genes, there was only small number of gaps distributed within the alignments. Amino-acid deletions were generally located on terminal parts of symbiotic bacteria genes that are under the lowest selection constraints. Contrary to protein-coding genes, 16S rRNA gene alignments were gapped by numerous indels.

Incongruence length difference (ILD) test rejected congruence between the single-gene datasets. I tried to exclude individual species to determine the source of noise. Exclusion of a single species did not significantly raise the P-value. Deleting multiple taxa resulted in an acceptance of the null hypothesis only if the retained taxa were closely related species. E.g. for six species from *Escherichia*/*Salmonella*/*Citrobacter* clade and concatenated *AceE*/*GroEL* datasets, P-value was only 0.08. However, problematic nature and limited power of this test have been proven many times in literature (e.g. Darlu and Lecointre, 2002; Dolphin et al., 2000). Since here the results of the ILD test are likely to be due to the unusual character of the data (see discussion), I used the joined datasets

despite of the negative results of ILD tests. Two concatenates were prepared. The first consisted of genes *AceE*/*EF-Tu*/*GroEL*/*LepA* with available lice symbionts sequences. The second dataset consisted of all 12 house-keeping genes. This dataset was used to test, if adding data affect the inferred tree topology.

Table 7. Lengths of edited multiple alignments and computed number of parsimony-informative/conserved sites.

Alignments	**Edited length (bp)**	**Number of sites (nucleotides/amino acids)**	
		Parsimony-informative	**Conserved**
AceE	2639	1683/495	718/267
AtpD	1365	685/141	572/251
EF-Tu	1173	589/124	502/217
FusA	2079	1190/322	678/260
GapA	990	652/203	239/75
GidA	1833	1219/341	502/193
GroEL	1581	790/186	667/271
GroEL max	1581	823/203	656/266
GyrB	2403	1571/460	635/237
LepA	1779	1148/342	468/162
PurB	1356	893/265	339/120
PyrG	1620	1004/271	484/185
RpoB	4005	2227/525	1434/598
16S rDNA 1	1422	597	637
16S rDNA 2	1220	494	576
16S rDNA 3	898	362	423

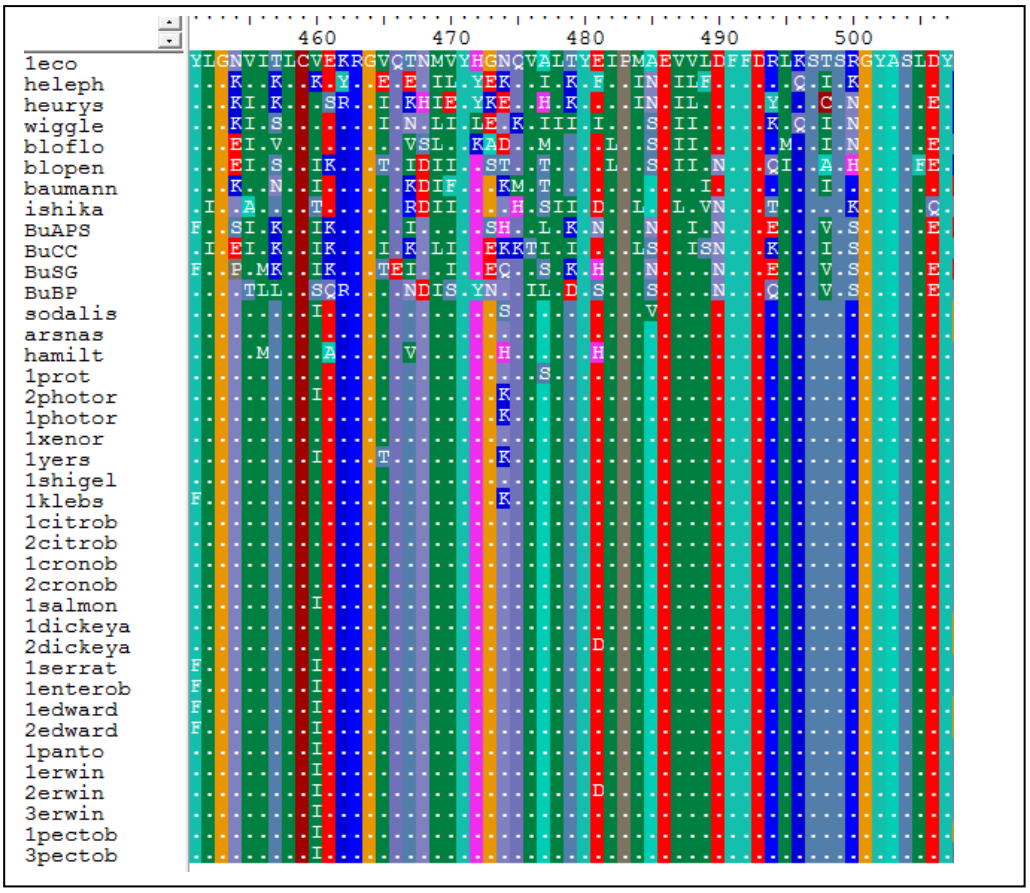


Figure 1. Section of translated *LepA* alignment demonstrates rapid evolution of insect symbionts in comparison with free-living bacteria. Dots represent amino-acids identical with *Escherichia coli*. Bacteria abbreviations are listed in Appendix 5, heleph-*Haematomyzus elephantis* endosymbiont, heurys-*Haematopinus eurysternus* endosymbiont.

4.3 Phylogenetic trees

Considerable number of phylogenetic trees was computed from nucleotide and amino-acid sequences using different weighting (excluded positions, slow-fast) and methods (MP, LogDet, ML, BI) with several complex models of evolution (codon model, nonhomogeneous model).

Tree topology derived from 16S rRNA gene varied considerably in dependence on the used method, alignment length and taxon sampling. Only a few clades were stable in all derived topologies. Therefore, the phylogenetic tree (Figure 2) computed by BI (GTR+I+Γ) from the longest alignment is presented only as an illustration of symbiotic Enterobacteriaceae diversity and should be taken with caution. Endosymbiont of *Haematomyzus* clustered almost exclusively with *Trioza magnoliae* psyllid endosymbiont, the placement congruent with the previous study (Hypsa and Krizek, 2007). The newly described taxa added into analysis (*Nardonella* spp.) form a sister clade. Endosymbiont of *Solenopotes*

clustered either with endosymbionts of *Haematopinus* or with endosymbionts of psyllids *Calophya schini* and *Heteropsylla cubana*. The last endosymbiont of lice in dataset is *Puchtella pedicinophila* clustering with *Wigglesworthia*. Gammaproteobacterial symbiont of weevil *Rhampus pulicarius* is a member of a broad *Sodalis* clade.

Analyses (MP, ML, BI, LogDet) of nucleotide data (single genes/concatenates) sorted symbionts according to their branch lengths (Figure 3) into one cluster joined to the longest branch within tree (e.g. *Arsenophonus nasoniae* or *Proteus mirabilis*). This topology was apparently incorrect as it even split up monophyly of *Buchnera* species and clustered *Wigglesworthia*/lice symbionts among them. Exclusion of third/third+second codon positions reduced the evident erorrs within symbionts, but did not help to eliminate the affinity of symbionts to the longest branch available. Likelihood analyses using nonhomogeneous model of DNA evolution, Bayesian approach with codon model and LogDet distances also did not convincingly avoid clustering of symbionts clade with the longest branch in tree. If the long-branched taxa were excluded, symbionts formed sister clade with remaining free-living bacteria. However, two clades of symbionts were generally resolved: *Buchnera*/*Ishikawaella* and *Baumannia*/*Blochmannia*/*Wigglesworthia*/lice symbionts. *Sodalis glossinidius* as the less degenerated symbiont was not usually attracted by long branches and clustered with pathogenic bacteria of plants (*Dickeya and Pectobacteiurm*) or animals (*Edwardsiella*). In contrary, *Hamiltonella defensa* was strongly attracted by the rest of symbionts and had to be excluded from most of the analyses.

Phylogenetic analyses using amino-acid data (single genes/concatenates) implied almost exclusively monophyly of P-symbionts (Figure 5), only analysis of GroEL with extensive taxon sampling (Figure 4) have split up monophyly and placed *Buchnera*/*Ishikawaella* as a sister clade to *Pantoea*/*Erwinia*.

Slow-fast method produced several datasets with reduced number of saturated positions. When analyzed one by one, the most reduced datasets (0-4) showed repeatedly independent position of *Buchnera* clade, which was clustered with *Erwinia*/*Pantoea* or *Edwardsiella* species. Phylogenetic tree inferred from the third matrix produced by slow-fast analysis of concatenated *AceE*, *EF-Tu*, *GroEL*, and *LepA* amino-acid data is presented here (Figure 6).

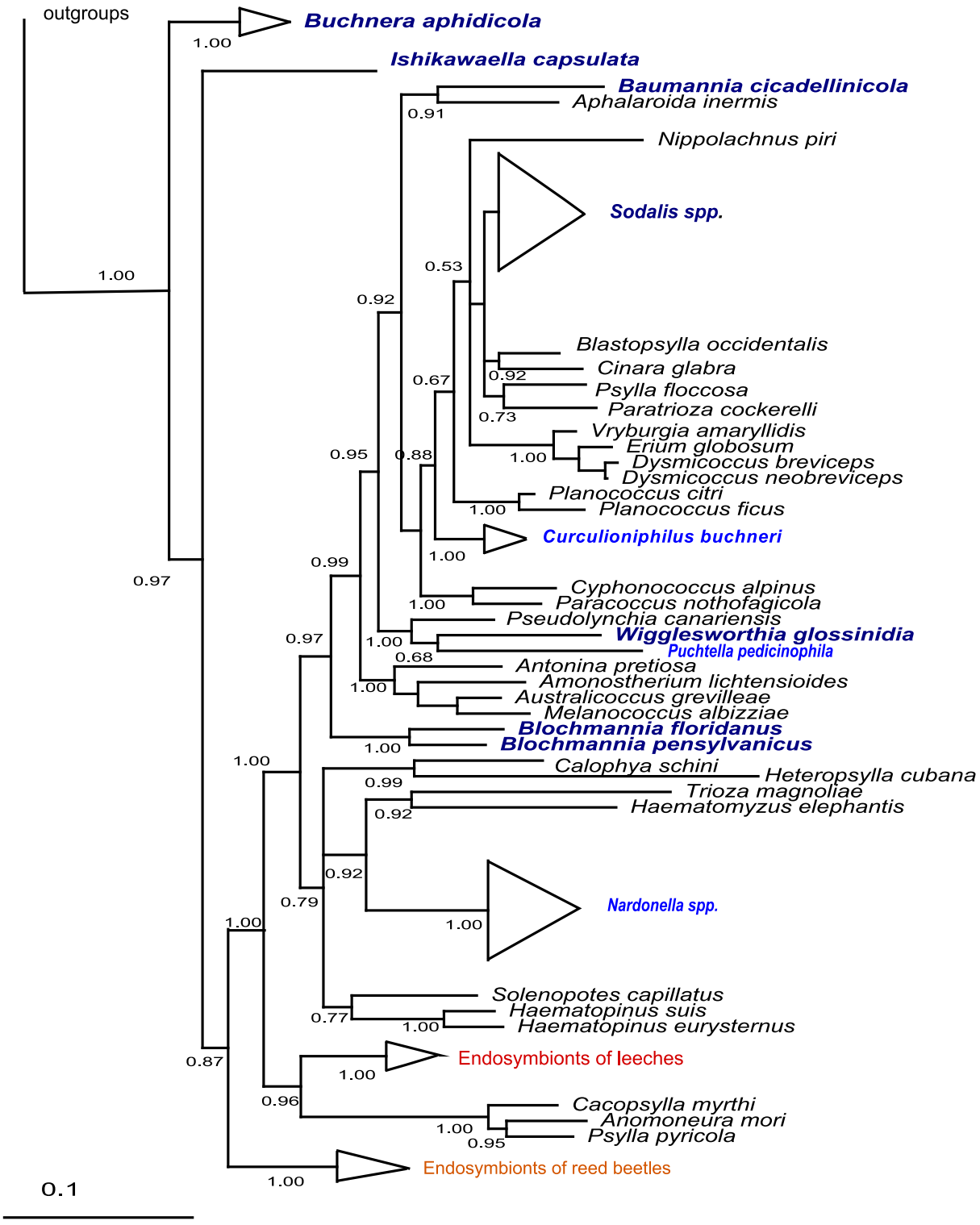


Figure 2. Phylogeny of insect symbionts derived from 16S rRNA gene using Bayesian inference (GTR+I+Γ). If possible, bacterial names are presented (symbionts are printed in color), otherwise host species are used. Triangles represent monophyletic clades with numerous species.

Figure 3. Nucleotide data analyses were typically affected by the long-branch attraction artifacts. Maximum likelihood tree (GTR+I+ Γ) inferred from *LepA* nucleotide data with excluded third codon positions. Insect symbionts are displayed in blue. Undescribed bacteria are presented by their host taxa and marked with an asterisk.

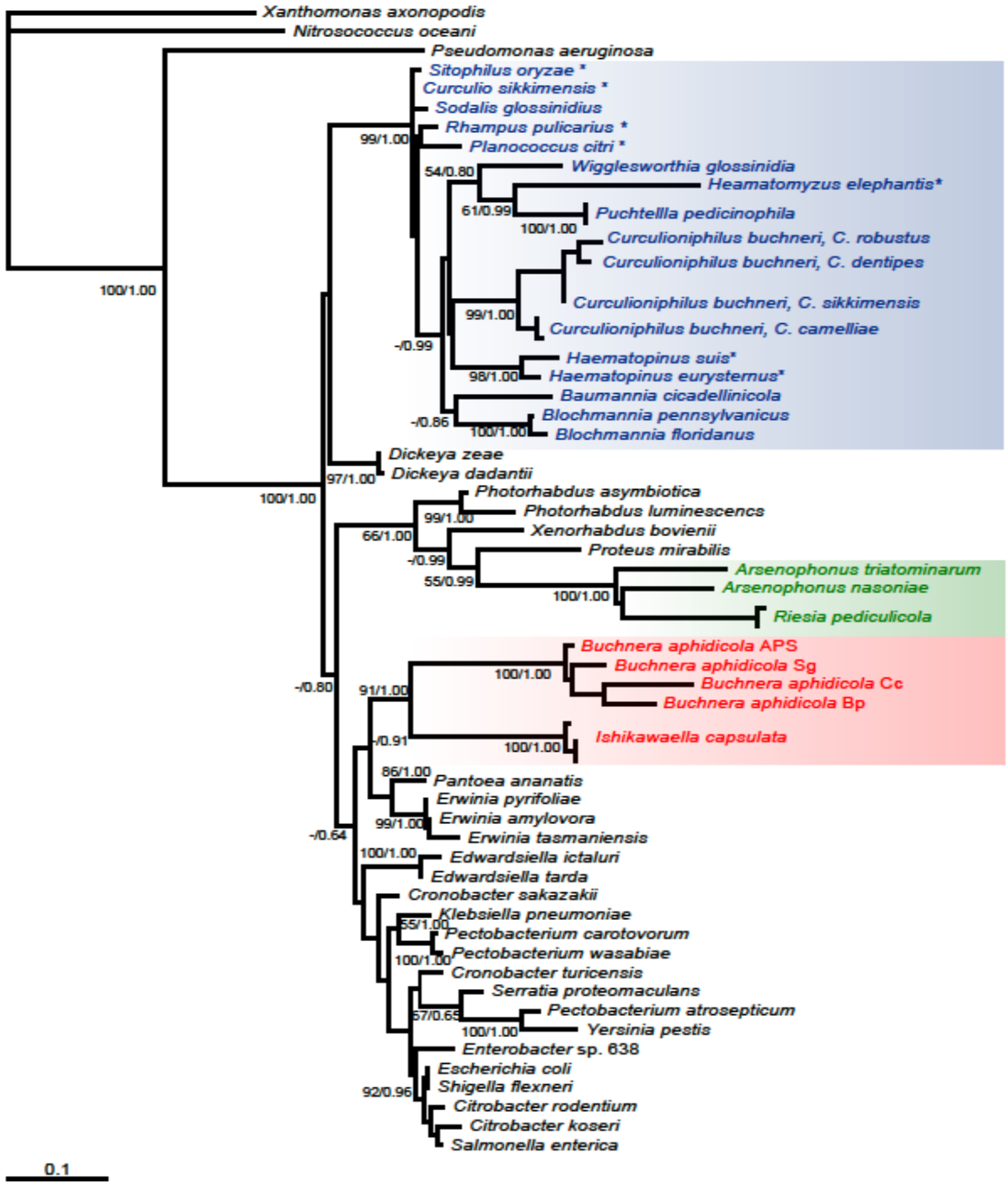


Figure 4. Phylogeny of Enterobacteriaceae inferred from *GroEL* amino-acid data by ML (LG+I+Γ) and BI (Wag+I+Γ). Three clusters of symbiotic bacteria of insects are printed in color. Undescribed bacteria are presented by their host taxa and marked with an asterisk.

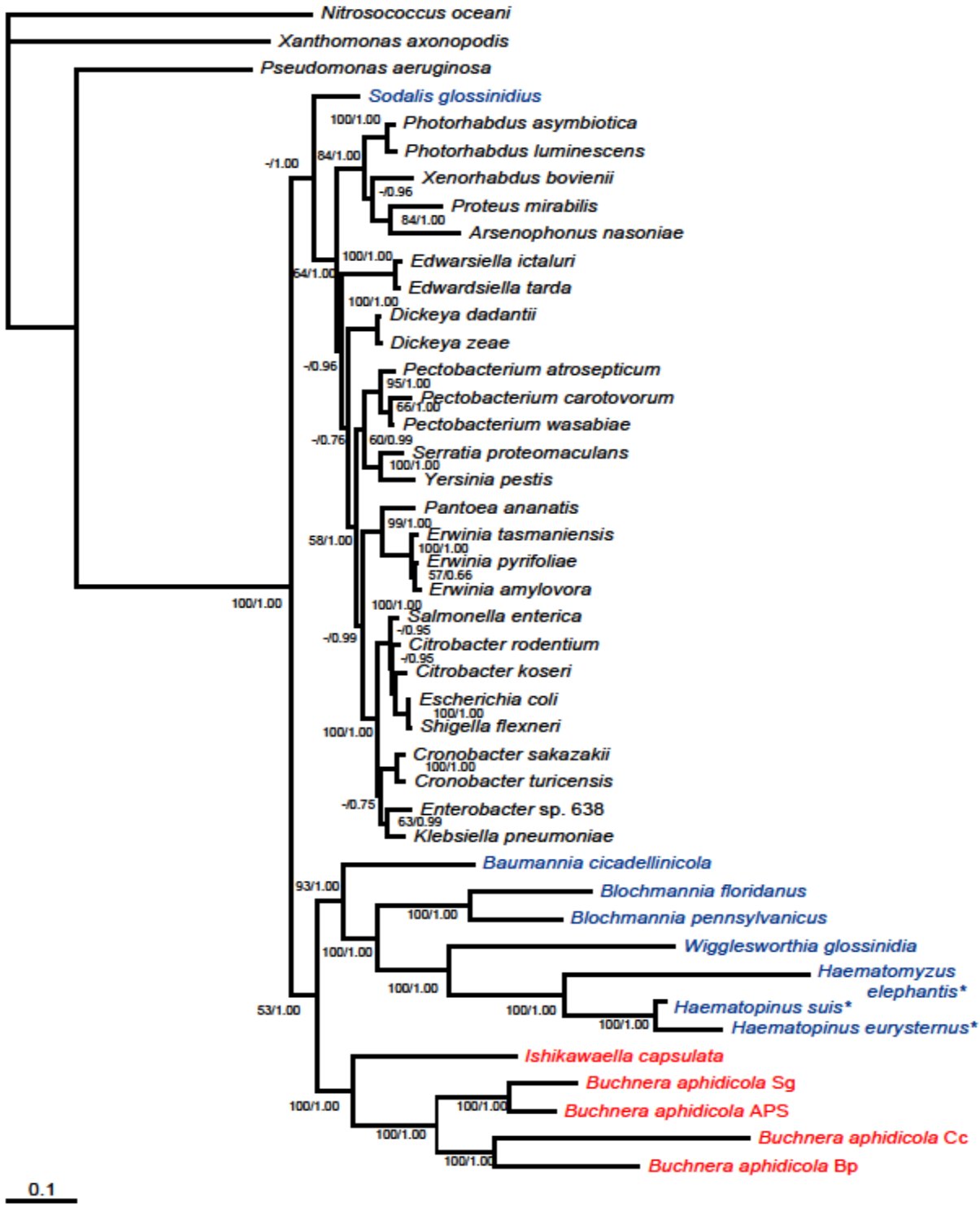


Figure 5. Phylogeny of Enterobacteriaceae inferred from concatenated *AceE*, *EF-Tu*, *GroEL* and *LepA* amino-acid data by ML (LG+I+Γ) and BI (Wag+I+Γ). Undescribed bacteria are presented by their host taxa and marked with an asterisk.

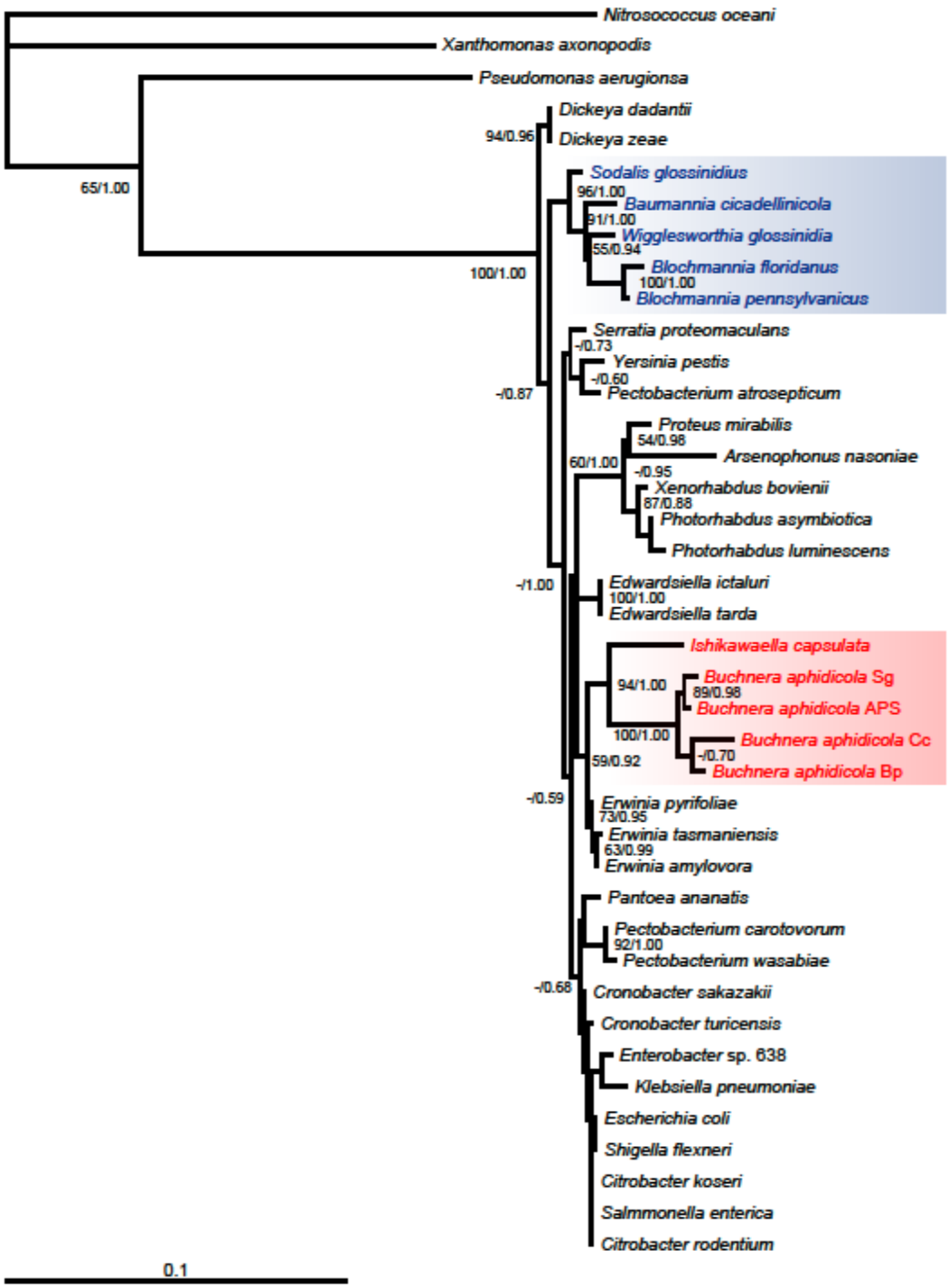


Figure 6. Phylogenetic tree inferred by ML (LG+I+Γ) and BI (Wag+I+Γ) from the third matrix produced by slow-fast analysis of concatenated *AceE*, *EF-Tu*, *GroEL* and *LepA* amino-acid data.

5. Discussion

5.1 PCR results, matrices, alignments

Negative PCR results for one of two closely related species of the louse symbionts (*LepA* and *AceE* from *Haematopinus* species) were probably caused by mismatches between primer and gene sequence. This problem is often encountered in symbiotic bacteria research due to rapid DNA substitution rate and thus higher sequence divergence between species. Degenerated *GroEL* primers were the only pair that worked properly with all of the tested species (except for the *Solenopotes* long-time stored and fragmented DNA). Good taxon sampling is available for *GroEL* in GeneBank as this gene often supplements/replaces ribosomal markers in insect symbionts phylogenies (Degnan et al., 2004; Fukatsu et al., 2009; Toju et al., 2009; Wertz et al., 2003). However, this heat-shock protein is under strong positive selection (Fares et al., 2004) and it has been shown that symbionts from two distantly related bacterial phyla have fixed identical amino-acids substitutions in functionally important regions of *GroEL* (see Appendix 6, Fares et al., 2005).

Alignments of the house-keeping genes are unequivocal and do not introduce any noise or false alignment-mediated phylogenetic signal into the analyses. In contrast, the16S rDNA alignment(s) contained high number of ambiguously aligned positions. Different stringency of alignment trimming demonstrated that characters affecting topology are within these regions.

The concatanated matrix of the housekeeping genes was used in the phylogenetic analyses despite the rejection by ILD tests. Strict rules of this test were many times discussed and it has been proven by simulations that ILD test fails if there are not numerous characters present and if there is not homogeneous substitution rate among sites (Barker and Lutzoni, 2002; Cunningham, 1997; Darlu and Lecointre, 2002; Dolphin et al., 2000; Graham et al., 1998; Yoder et al., 2001). Considering problematic character of the symbionts sequences (i.e. the strong compositional bias and the rate nonhomogeneity) and the fact that the housekeeping genes belong to the "core" genes with low probability of HGT, it is likely that the rejection by ILD test is an artifact and does not reflect a real incompatibility of the phylogenetic signals.

5.2 Phylogenetic analyses

Phylogenetic analyses were accompanied by easily detectable but hard to eliminate long-branch attraction artifacts. As expected, analyses based on 16S rDNA did not lead to a clear conclusion. Low informative value of this marker is being discussed since the first phylogenetic studies of symbiotic bacteria (O'Neill et al., 1992). At least, inferred topology indicates some essential information needed for further multi-gene analyses. Crucial role of various S-symbionts for P-symbionts phylogeny has

been noticed and my results support the theory that (some) primary symbionts originated from secondary-symbiont ancestors (Wernegreen et al., 2009).

Only two studies have been published that present *Buchnera* and *Wigglesworthia/Blochmannia* clades in polyphyly (Belda et al., 2005; Herbeck et al., 2005). It should be mentioned, that phylogenomic studies imply monophyly of this clade on the basis of very inappropriate taxon sampling: presence of only few free-living/pathogenic enterobacteria in these datasets is not rare (Canback et al., 2004; Comas et al., 2007a). Species which cause polyphyly of insect symbionts in the study of Herbeck et al. (2005) are only distantly related to *Escherichia*/*Salmonella*/*Yersinia* clade. Monophyly of P-symbionts derived from phylogenomic dataset, which includes only these three genera is logically inevitable, but does not contain any information.

My results suggest two independent origins of the studied insect symbionts (Figures 4 and 6). The topology is partially congruent with the tree previously derived by the nonhomogeneous model of sequence evolution (Herbeck et al., 2005). *Buchnera* + *Ishikawaella* form one clade while the rest of symbionts, including *Sodalis*, the other clade. Both these groups cluster with plant pathogenic bacteria, but it would be premature to define ancestors of the symbionts because taxon sampling of plant pathogenic bacteria is extremely poor. It is generally assumed that insect symbionts originated from bacteria common in the environment. For example, cicadas spend most of their lives underground and feed primarily on plant roots. It is no surprise, that their α-Proteobacterial symbiont *Hodgkinia cicadicola* originated within Rhizobiales (McCutcheon et al., 2009b), a group that is known especially because of symbiosis with plant roots. Insects that harbor symbiotic Enterobacteriaceae are predominantly feeding on plants. Pathogenic bacteria of plants fit well in the above mentioned theory as bacteria that infected ancestors of the given insect group and become symbionts. It can be only speculated that they first become "S-symbiont type" and were horizontally transferred to various other insect species. Within some of the infected species, facultative symbionts became obligate primary symbionts. Identical situation can be observed in clades with numerous species such as *Wolbachia* (e.g. Hosokawa et al., 2010), *Sodalis* (Fukatsu et al., 2007; Heddi et al., 1998; Novakova and Hypsa, 2007) or *Arsenophonus* (Novakova et al., 2009).

To determine a particular bacterial lineage as an ancestor of endosymbionts is very complicated due to the rapid evolution of symbionts. In my phylogenetic analyses of Enterobacteriaceae, lack of informative positions needed to infer deeper phylogeny and detect ancestor(s) of insect symbionts was the main problem. If there is not sufficient amount of information matching symbionts with free-living bacteria, they cluster together. Long-branch attraction artifacts are only the consequence of overwritten ancient information and can be eliminated by various methods. One of the proposed methods is the nonhomogeneous model of sequence evolution (Galtier and Gouy, 1998). However, this model did not bring convincing results when applied to my datasets. Two different explanations

can account for this fact. First, there is no phylogenetic signal that can be increased by implementing heterogeneous base composition assumption. Second, this model works differently if used with different datasets. In my dataset, there is substantial number of species with homogeneous base composition which could cause an insufficient capability of the nonhomogeneous model to recognize the correct topology. LogDet distances were also incapable to resolve origin(s) of P-symbionts. I attribute the failure of these methods to the low number of informative positions that can be used to infer deeper phylogeny. Ancient phylogenetic information is overwritten by rapid evolution in the insect host, e.g. by approximately 180 millions of years for *Buchnera* species (Moran et al., 1993).

Slow-fast analysis increased the ratio of the ancient phylogenetic information and I consider the inferred tree as an alternative theory to the monophyly of insect symbionts. I have to admit that plant pathogenic bacteria are also plausible species which could exchange genes with symbiotic bacteria of plant sap-sucking insects. Phylogenomic study will have to be done to test possible horizontal gene transfers and confirm/reject my preliminary results.

Phylogeny of lice endosymbionts inferred from 16S rRNA gene implies their independent origins. However, all the endosymbionts included in my datasets of house-keeping genes (*Haematomyzus*, *Haematopinus*, *Solenopotes*, *Pedicinus*), cluster with *Wigglesworthia* in genome-wide taxon sampling. This problem is caused by missing taxa. Therefore, it was impossible to convincingly prove the independent origins indicated by 16S rRNA gene phylogeny.

In future, I will extend my dataset both with additional taxa and data for other genes. Aspirant taxa are mainly the secondary symbionts clustering with P-symbionts. If genomes are available, various pathogenic Enterobacteriaceae will be added to the dataset. These bacteria are not as strongly affected by loss of ancient phylogenetic information and the resulting LBA artifacts. For data extension, I will try to overcome negative PCR results for some symbionts by newly designed primers. I also intend to prepare Enterobacteriaceae phylogenomic dataset based on core of all orthologous genes in symbiotic genomes (approximately 200 genes) and automate the process of sequence downloading, filtering and aligning e.g. by workflow of Perl scripts. With phylogenomic dataset finished, it will be possible to evaluate amount of genome-wide phylogenetic signal favoring my current results.

6. Conclusion

Phylogenetic analyses of Enterobacteriaceae did not unambiguously resolve phylogenetic position of symbiotic bacteria in insects. However, there is a certain level of phylogenetic signal favoring independent origin of *Buchnera*/*Ishikawaella* and *Baumannia*/*Blochmannia*/*Wigglesworthia* clades. Unfortunately, this signal is strongly overwritten by rapid evolution of symbionts and by convergently acquired characters.

Endosymbiotic bacteria of lice were observed to cluster within the second clade mentioned, but analysis of 16S rRNA gene implies that the "genome" taxon sampling is very poor for precise placement and 16S rRNA gene alone is insufficient marker. Importance of endosymbionts of psyllids and weevils should not be overlooked because these symbionts form clades related to primary symbionts of lice.

Ancestors of insect symbionts were probably bacteria that somehow come in contact with insects. My results suggest pathogenic bacteria of plants and animals as a possible source of bacterial species that infected ancestors of insects and become domesticated.

7. Appendices

Appendix 1. Loading buffers composition.

Agarose gel electrophoresis loading buffer	**TGGE loading buffer TAE**
0.01% bromophenol blue (Sigma)	0.01% bromophenol blue (Sigma)
0.01% xylene cyanol FF (Sigma)	0.01% xylene cyanol (Sigma)
500 x SYBR® Green I (Sigma)	2 mM EDTA (Sigma)
30% glycerol (Sigma)	0.1% Triton-X 100 (Sigma)
ddH_2O to 10 ml	

Appendix 2. Acryl amid gel composition.

Stock solution	**for 50 ml**	**Final concentrations**
40 % (37,5 :1) acrylamid (Merck)	10 ml	[8%]
urea - solid (AppliChem)	21 g	[7M]
50 x TAE (Duchefa)	1ml	[1x]
40% glycerol (Sigma)	2,5 ml	[2%]
dH_2O	to 50 ml	adjusted with aqua bidest

Appendix 3. Composition of 1% agar plates.

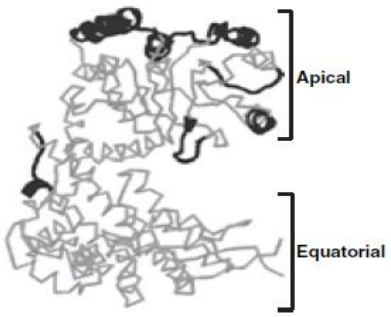
2.5 g LB medium (Amresco)
1g bacteriological agar (Amresco)
100 ml dH_20
100 µg/ml ampicilin (Invitrogen)

Appendix 4. Forward and reverse primers designed to obtain contiguous gene sequence.

Primer - species	**Sequence**	**Position**
GroEL f *Haematopinus*	GGTACTATATCTGCAAATGC	1048-1067
GroEL r *Haematopinus*	GCTTCATCTCTTTGTTGATT	442-462
GroEL f *Haematomyzus*	GAAGCTATGGATAGAGTAGG	490-509
GroEL r *Haematomyzus*	TTTATTACCGCTACACCTCC	1120-1140
GroEL f *Rhampus*	GAAGCCATGGCGAAAGTGGG	490-509
GroEL r *Rhampus*	TTGACAACCGCAACGCCGCC	1120-1140
AceE f *Haematopinus*	AAGGGTTATGGTATGGGGTC	593-612
AceE r *Haematopinus*	AGTATCTGATCATTATCTTG	1064-1083

Appendix 5. Bacterial species with genomes available in GenBank used as a core of all datasets.

Accession number	Bacterium	Abbreviation
NC_002528	*Buchnera aphidicola*	BuAPS
NC_004545	*Buchnera aphidicola*	BuBP
NC_008513	*Buchnera aphidicola*	BuCC
NC_004061	*Buchnera aphidicola*	BuSG
NC_007984	*Baumannia cicadellinicola*	baumann
NC_005061	*Blochmannia floridanus*	bloflo
NC_007292	*Blochmannia pennsylvanicus*	blopen
NC_012752	*Hamiltonella defensa*	hamilt
NC_007712	*Sodalis glossinidius*	sodalis
NC_004344	*Wigglesworthia glossinidia*	wiggle
NC_009792	*Citrobacter koseri ATCC BAA-895*	1citrob
NC_013716	*Citrobacter rodentium ICC168*	2citrob
NC_009778	*Enterobacter sakazakii ATCC BAA-894*	1cronob
NC_013282	*Cronobacter turicensis*	2cronob
NC_013592	*Dickeya dadantii Ech586*	1dickeya
NC_012912	*Dickeya zeae Ech1591*	2dickeya
NC_012779	*Edwardsiella ictaluri 93-146*	1edward
NC_013508	*Edwardsiella tarda EIB202*	2edward
NC_009436	*Enterobacter sp. 638*	1enterob
NC_012214	*Erwinia pyrifoliae Ep1/96*	1erwin
NC_010694	*Erwinia tasmaniensis Et1/99*	2erwin
NC_013961	*Erwinia amylovora*	3erwin
NC_013956	*Pantoea ananatis LMG 20103*	1panto
NC_010473	*Escherichia coli str. K-12 substr. DH10B*	1eco
NC_012731	*Klebsiella pneumoniae NTUH-K2044*	1klebs
NC_004547	*Pectobacterium atrosepticum*	1pectob
NC_012917	*Pectobacterium carotovorum subsp. carotovorum PC1*	2pectob
NC_013421	*Pectobacterium wasabiae WPP163*	3pectob
NC_012962	*Photorhabdus asymbiotica*	1photor
NC_005126	*Photorhabdus luminescens subsp. laumondii TTO1*	2photor
NC_013892	*Xenorhabdus bovienii SS-2004*	1xenor
NC_010554	*Proteus mirabilis HI4320*	1prot
NC_003197	*Salmonella enterica subsp. enterica serovar Typhimurium str. LT2*	1salmon
NC_009832	*Serratia proteamaculans 568*	1serrat
NC_008258	*Shigella flexneri 5 str. 8401*	1shigel
NC_003143	*Yersinia pestis CO92*	1yers
project	*Ishikawaella capsulata Mpkobe*	ishika
scaffolds	*Arsenophonus nasoniae*	arsnas
NC_003919	*Xanthomonas axonopodis pv. citri str. 306*	xanthaxo
NC_007484	*Nitrosococcus oceani ATCC 19707*	nitroso
NC_002516	*Pseudomonas aeruginosa PAO1*	1pseudo



Appendix 6. *GroEL* monomer from *Escherichia coli* with locations of amino acid residues (in dark) detected to be under positive selection in both Bacteroidetes and γ-Proteobacteria endosymbionts. Adapted from Fares et al. (2005).

8. References

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21, 2104-2105.

Aguinaldo, A.M., Turbeville, J.M., Linford, L.S., Rivera, M.C., Garey, J.R., Raff, R.A., Lake, J.A., 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387, 489-493.

Akman, L., Yamashita, A., Watanabe, H., Oshima, K., Shiba, T., Hattori, M., Aksoy, S., 2002. Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. Nat Genet 32, 402-407.

Allen, J.M., Light, J.E., Perotti, M.A., Braig, H.R., Reed, D.L., 2009. Mutational meltdown in primary endosymbionts: selection limits Muller's ratchet. PLoS One 4, e4969.

Allen, J.M., Reed, D.L., Perotti, M.A., Braig, H.R., 2007. Evolutionary relationships of "*Candidatus* Riesia spp.," endosymbiotic *Enterobacteriaceae* living within hematophagous primate lice. Appl Environ Microbiol 73, 1659-1664.

Anderson, F.F., Swofford, D.L., 2004. Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA. Mol Phylogenet Evol 33, 440-451.

Aschner, M., 1932. Experimentelle unter suchungen über die symbiose der kleiderlaus. Naturwissenschaften 20, 501-505.

Aschner, M., 1934. Studies on the symbiosis of the body louse. I. Elimination of the symbionts by centrifugation of the eggs. Parasitology 26, 309-314.

Balogh, K., Patocs, A., Majnik, J., Racz, K., Hunyady, L., 2004. Genetic screening methods for the detection of mutations responsible for multiple endocrine neoplasia type 1. Mol Genet Metab 83, 74-81.

Bandi, C., Sironi, M., Damiani, G., Magrassi, L., Nalepa, C.A., Laudani, U., Sacchi, L., 1995. The establishment of intracellular symbiosis in an ancestor of cockroaches and termites. Proc Biol Sci 259, 293-299.

Bapteste, E., Susko, E., Leigh, J., Ruiz-Trillo, I., Bucknam, J., Doolittle, W.F., 2008. Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the prokaryotic phylogeny. Mol Biol Evol 25, 83-91.

Barker, F.K., Lutzoni, F.M., 2002. The utility of the incongruence length difference test. Syst Biol 51, 625-637.

Baudisch, K., 1958. Beiträge zur zytologie und embryologie einiger insektensymbiosen. Zeit. Morph. Okol. Tiere 47, 436-488.

Baumann, L., Thao, M.L., Hess, J.M., Johnson, M.W., Baumann, P., 2002. The genetic properties of the primary endosymbionts of mealybugs differ from those of other endosymbionts of plant sap-sucking insects. Appl Environ Microbiol 68, 3198-3205.

Baumann, P., 2005. Biology of bacteriocyte-associated endosymbionts of plant sap-sucking insects. Annu Rev Microbiol 59, 155-189.

Beier, S., Witzel, K.P., Marxsen, J., 2008. Bacterial community composition in Central European running waters examined by temperature gradient gel electrophoresis and sequence analysis of 16S rRNA genes. Appl Environ Microbiol 74, 188-199.

Belda, E., Moya, A., Silva, F.J., 2005. Genome rearrangement distances and gene order phylogeny in γ-Proteobacteria. Mol Biol Evol 22, 1456-1467.

Bergsten, J., 2005. A review of long-branch attraction. Cladistics 21, 163-193.

Brady, C., Cleenwerck, I., Venter, S., Vancanneyt, M., Swings, J., Coutinho, T., 2008. Phylogeny and identification of *Pantoea* species associated with plants, humans and the natural environment based on multilocus sequence analysis (MLSA). Syst Appl Microbiol 31, 447-460.

Brinkmann, H., Philippe, H., 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol 16, 817-825.

Brown, E.W., Davis, R.M., Gouk, C., van der Zwet, T., 2000. Phylogenetic relationships of necrogenic *Erwinia* and *Brenneria* species as revealed by glyceraldehyde-3-phosphate dehydrogenase gene sequences. Int J Syst Evol Micr 50, 2057-2068.

Brummer, I.H., Felske, A., Wagner-Dobler, I., 2003. Diversity and seasonal variability of β-Proteobacteria in biofilms of polluted rivers: analysis by temperature gradient gel electrophoresis and cloning. Appl Environ Microbiol 69, 4463-4473.

Buchner, P., 1965. Endosymbiosis of animals with plant microorganisms. Interscience Publishers, New York.

Burke, G.R., Normark, B.B., Favret, C., Moran, N.A., 2009. Evolution and diversity of facultative symbionts from the aphid subfamily Lachninae. Appl Environ Microbiol 75, 5328–5335.

Canback, B., Tamas, I., Andersson, S.G., 2004. A phylogenomic study of endosymbiotic bacteria. Mol Biol Evol 21, 1110-1122.

Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17, 540-552.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., Bork, P., 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311, 1283-1287.

Comas, I., Moya, A., Azad, R.K., Lawrence, J.G., Gonzalez-Candelas, F., 2006. The evolutionary origin of Xanthomonadales genomes and the nature of the horizontal gene transfer process. Mol Biol Evol 23, 2049-2057.

Comas, I., Moya, A., Gonzalez-Candelas, F., 2007a. From phylogenetics to phylogenomics: the evolutionary relationships of insect endosymbiotic γ-Proteobacteria as a test case. Syst Biol 56, 1-16.

Comas, I., Moya, A., Gonzalez-Candelas, F., 2007b. Phylogenetic signal and functional categories in Proteobacteria genomes. BMC Evol Biol 7 Suppl 1, doi:10.1186/1471-2148-1187-S1181-S1187.

Conord, C., Despres, L., Vallier, A., Balmand, S., Miquel, C., Zundel, S., Lemperiere, G., Heddi, A., 2008. Long-term evolutionary stability of bacterial endosymbiosis in Curculionoidea: additional evidence of symbiont replacement in the Dryophthoridae family. Mol Biol Evol 25, 859-868.

Cortez, D., Delaye, L., Lazcano, A., Becerra, A., 2009. Composition-based methods to identify horizontal gene transfer. Methods Mol Biol 532, 215-225.

Covacin, C., Barker, S.C., 2007. Supergroup F *Wolbachia* bacteria parasitise lice (Insecta: Phthiraptera). Parasitol Res 100, 479-485.

Cunningham, C.W., 1997. Can three incongruence tests predict when data should be combined? Mol Biol Evol 14, 733-740.

Dagan, T., 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.

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

Darby, A.C., Choi, J.H., Wilkes, T., Hughes, M.A., Werren, J.H., Hurst, G.D., Colbourne, J.K., 2010. Characteristics of the genome of *Arsenophonus nasoniae*, son-killer bacterium of the wasp *Nasonia*. Insect Mol Biol 19 Suppl 1, 75-89.

Darlu, P., Lecointre, G., 2002. When does the incongruence length difference test fail? Mol Biol Evol 19, 432-437.

Daubin, V., Gouy, M., Perriere, G., 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Res 12, 1080-1090.

Dauga, C., 2002. Evolution of the *gyrB* gene and the molecular phylogeny of *Enterobacteriaceae*: a model molecule for molecular systematic studies. Int J Syst Evol Microbiol 52, 531-547.

Degnan, P.H., Lazarus, A.B., Brock, C.D., Wernegreen, J.J., 2004. Host-symbiont stability and fast evolutionary rates in an ant-bacterium association: cospeciation of *Camponotus* species and their endosymbionts, *Candidatus* Blochmannia. Syst Biol 53, 95-110.

Degnan, P.H., Lazarus, A.B., Wernegreen, J.J., 2005. Genome sequence of *Blochmannia pennsylvanicus* indicates parallel evolutionary trends among bacterial mutualists of insects. Genome Res 15, 1023-1033.

Degnan, P.H., Leonardo, T.E., Cass, B.N., Hurwitz, B., Stern, D., Gibbs, R.A., Richards, S., Moran, N.A., 2009a. Dynamics of genome evolution in facultative symbionts of aphids. Environ Microbiol, doi:10.1111/j.1462-2920.2009.02085.x.

Degnan, P.H., Yu, Y., Sisneros, N., Wing, R.A., Moran, N.A., 2009b. *Hamiltonella defensa*, genome evolution of protective bacterial endosymbiont from pathogenic ancestors. Proc Natl Acad Sci U S A 106, 9063-9068.

Delsuc, F., Brinkmann, H., Philippe, H., 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6, 361-375.

Dolphin, K., Belshaw, R., Orme, C.D.L., Quicke, D.L.J., 2000. Noise and incongruence: Interpreting results of the incongruence length difference test. Mol Phylogenet Evol 17, 401-406.

Douglas, A.E., 1989. Mycetocyte symbiosis in insects. Biol Rev Camb Philos Soc 64, 409-434.

Douglas, A.E., Francois, C.L.M.J., Minto, L.B., 2006. Facultative 'secondary' bacterial symbionts and the nutrition of the pea aphid, *Acyrthosiphon pisum*. Physiol Entomol 31, 262-269.

Downie, D.A., Gullan, P.J., 2005. Phylogenetic congruence of mealybugs and their primary endosymbionts. J Evol Biol 18, 315-324.

Duron, O., Bouchon, D., Boutin, S., Bellamy, L., Zhou, L.Q., Engelstadter, J., Hurst, G.D., 2008. The diversity of reproductive parasites among arthropods: *Wolbachia* do not walk alone. BMC Biol 6, doi:10.1186/1741-7007-1186-1127.

Everett, K.D., Thao, M., Horn, M., Dyszynski, G.E., Baumann, P., 2005. Novel chlamydiae in whiteflies and scale insects: endosymbionts '*Candidatus* Fritschea bemisiae' strain Falk and '*Candidatus* Fritschea eriococci' strain Elm. Int J Syst Evol Microbiol 55, 1581-1587.

Fares, M.A., Moya, A., Barrio, E., 2004. GroEL and the maintenance of bacterial endosymbiosis. Trends Genet 20, 413-416.

Fares, M.A., Moya, A., Barrio, E., 2005. Adaptive evolution in GroEL from distantly related endosymbiotic bacteria of insects. J Evol Biol 18, 651-660.

Fares, M.A., Ruiz-Gonzalez, M.X., Moya, A., Elena, S.F., Barrio, E., 2002. Endosymbiotic bacteria - GroEL buffers against deleterious mutations. Nature 417, 398-398.

Farris, J.S., Kallersjo, M., Kluge, A.G., Bult, C., 1995. Testing significance of incogruence. Cladistics 10, 315-319.

Felsenstein, J., 2003. Inferring phylogenies. Sinauer Associates, Inc., Sunderland, Massachusetts.

Foster, P.G., Hickey, D.A., 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J Mol Evol 48, 284-290.

Fukatsu, T., Hosokawa, T., Koga, R., Nikoh, N., Kato, T., Hayama, S., Takefushi, H., Tanaka, I., 2009. Intestinal endocellular symbiotic bacterium of the macaque louse *Pedicinus obtusus*: Distinct endosymbiont origins in anthropoid primate lice and the old world monkey louse. Appl Environ Microbiol 75, 3796-3799.

Fukatsu, T., Koga, R., Smith, W.A., Tanaka, K., Nikoh, N., Sasaki-Fukatsu, K., Yoshizawa, K., Dale, C., Clayton, D.H., 2007. Bacterial endosymbiont of the slender pigeon louse, *Columbicola columbae*, allied to endosymbionts of grain weevils and tsetse flies. Appl Environ Microbiol 73, 6660-6668.

Fukatsu, T., Nikoh, N., Kawai, R., Koga, R., 2000. The secondary endosymbiotic bacterium of the pea aphid *Acyrthosiphon pisum* (Insecta: Homoptera). Appl Environ Microbiol 66, 2748-2758.

Galtier, N., 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., 1998. Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol Biol Evol 15, 871-879.

Galtier, N., Gouy, M., Gautier, C., 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci 12, 543-548.

Gao, B., Mohan, R., Gupta, R.S., 2009. Phylogenomics and protein signatures elucidating the evolutionary relationships among the Gammaproteobacteria. Int J Syst Evol Micr 59, 234-247.

Gasparich, G.E., 2002. Spiroplasmas: evolution, adaptation and diversity. Front Biosci 7, 619-640.

Ge, F., Wang, L.S., Kim, J., 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biol 3, e316.

Gherna, R.L., Werren, J.H., Weisburg, W., Cote, R., Woese, C.R., Mandelco, L., Brenner, D.J., 1991. *Arsenophonus nasoniae* gen. nov, sp. nov, the causative agent of the son-killer trait in the parasitic wasp *Nasonia vitripennis*. Int J Syst Bacteriol 41, 563-565.

Gil, R., Silva, F.J., Zientz, E., Delmotte, F., Gonzalez-Candelas, F., Latorre, A., Rausell, C., Kamerbeek, J., Gadau, J., Holldobler, B., van Ham, R.C., Gross, R., Moya, A., 2003. The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. Proc Natl Acad Sci U S A 100, 9388-9393.

Goldman, N., Yang, Z., 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11, 725-736.

Gottlieb, Y., Ghanim, M., Gueguen, G., Kontsedalov, S., Vavre, F., Fleury, F., Zchori-Fein, E., 2008. Inherited intracellular ecosystem: symbiotic bacteria share bacteriocytes in whiteflies. FASEB J 22, 2591-2599.

Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27, 221-224.

Graham, S.W., Kohn, J.R., Morton, B.R., Eckenwalder, J.E., Barrett, S.C., 1998. Phylogenetic congruence and discordance among one morphological and three molecular data sets from Pontederiaceae. Syst Biol 47, 545-567.

Gruwell, M.E., Morse, G.E., Normark, B.B., 2007. Phylogenetic congruence of armored scale insects (Hemiptera: Diaspididae) and their primary endosymbionts from the phylum Bacteroidetes. Mol Phylogenet Evol 44, 267-280.

Guindon, S., Delsuc, F., Dufayard, J.F., Gascuel, O., 2009. Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol 537, 113-137.

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52, 696-704.

Gupta, R.S., 2000. The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. FEMS Microbiol Rev 24, 367-402.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41, 95-98.

Hansen, A.K., Jeong, G., Paine, T.D., Stouthamer, R., 2007. Frequency of secondary symbiont infection in an invasive psyllid relates to parasitism pressure on a geographic scale in California. Appl Environ Microbiol 73, 7531-7535.

Hassan, A.K., Moriya, S., Baumann, P., Yoshikawa, H., Ogasawara, N., 1996. Structure of the *dnaA* region of the endosymbiont, *Buchnera aphidicola*, of aphid *Schizaphis graminum*. DNA Res 3, 415-419.

Haynes, S., Darby, A.C., Daniell, T.J., Webster, G., Van Veen, F.J., Godfray, H.C., Prosser, J.I., Douglas, A.E., 2003. Diversity of bacteria associated with natural aphid populations. Appl Environ Microbiol 69, 7216-7223.

Heddi, A., Charles, H., Khatchadourian, C., Bonnot, G., 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.

Hendy, M.D., Penny, D., 1989. A framework for the quantitative study of evolutionary trees. Syst Zool 38, 297-309.

Herbeck, J.T., Degnan, P.H., Wernegreen, J.J., 2005. Nonhomogeneous model of sequence evolution indicates independent origins of primary endosymbionts within the enterobacteriales (γ-Proteobacteria). Mol Biol Evol 22, 520-532.

Hill, P., Saunders, D.S., Campbell, J.A., 1973. The production of "symbiont-free" *Glossina morsitans* and an associated loss of female fertility. Trans R Soc Trop Med Hyg 67, 727-728.

Horiike, T., Miyata, D., Hamada, K., Saruhashi, S., Shinozawa, T., Kumar, S., Chakraborty, R., Komiyama, T., Tateno, Y., 2009. Phylogenetic construction of 17 bacterial phyla by new method and carefully selected orthologs. Gene 429, 59-64.

Hosokawa, T., Kikuchi, Y., Nikoh, N., Shimada, M., Fukatsu, T., 2006. Strict host-symbiont cospeciation and reductive genome evolution in insect gut bacteria. PLoS Biol 4, e337.

Hosokawa, T., Koga, R., Kikuchi, Y., Meng, X.Y., Fukatsu, T., 2010. *Wolbachia* as a bacteriocyte-associated nutritional mutualist. Proc Natl Acad Sci U S A 107, 769-774.

Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754-755.

Hurst, G.D., Bandi, C., Sacchi, L., Cochrane, A.G., Bertrand, D., Karaca, I., Majerus, M.E., 1999. *Adonia variegata* (Coleoptera: Coccinellidae) bears maternally inherited flavobacteria that kill males only. Parasitology 118 (Pt 2), 125-134.

Hypsa, V., Krizek, J., 2007. Molecular evidence for polyphyletic origin of the primary symbionts of sucking lice (Phthiraptera, Anoplura). Microb Ecol 54, 242-251.

Hypsa, V., Novakova, E., 2008. Insect symbionts and molecular phylogenetics. In: Bourtzis, K., Miller, T. (Eds.), Insect symbiosis. CRC Press, pp. 1-32.

Chen, D.Q., B., M.C., H., P.A., 2000. Fitness effects of two facultative endosymbiotic bacteria on the pea aphid, *Acyrthosiphon pisum*, and the blue alfalfa aphid, *A. kondoi*. Entomol Exp Appl 95, 315-323.

Isabel, S., Leblanc, E., Boissinot, M., Boudreau, D.K., Grondin, M., Picard, F.J., Martel, E.A., Parham, N.J., Chain, P.S.G., Bader, D.E., Mulvey, M.R., Bryden, L., Roy, P.H., Ouellette, M., Bergeron, M.G., 2008. Divergence among genes encoding the elongation factor Tu of *Yersinia* species. J Bacteriol 190, 7548-7558.

Jain, R., Rivera, M.C., Lake, J.A., 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A 96, 3801-3806.

Katoh, K., Asimenos, G., Toh, H., 2009. Multiple alignment of DNA sequences with MAFFT. Methods Mol Biol 537, 39-64.

Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30, 3059-3066.

Ke, S.H., Kelly, P.J., Wartell, R.M., Hunter, S., Varma, V.A., 1993. Selecting DNA fragments for mutation detection by temperature gradient gel electrophoresis: application to the p53 gene cDNA. Electrophoresis 14, 561-565.

Kennedy, M., Paterson, A.M., Morales, J.C., Parsons, S., Winnington, A.P., Spencer, H.G., 1999. The long and short of it: branch lengths and the problem of placing the New Zealand short-tailed bat, *Mystacina*. Mol Phylogenet Evol 13, 405-416.

Kikuchi, Y., 2009. Endosymbiotic bacteria in insects: their diversity and culturability. Microbes Environ 24, 195-204.

Kikuchi, Y., Fukatsu, T., 2002. Endosymbiotic bacteria in the esophageal organ of glossiphoniid leeches. Appl Environ Microbiol 68, 4637-4641.

Kikuchi, Y., Hosokawa, T., Nikoh, N., Meng, X.Y., Kamagata, Y., Fukatsu, T., 2009. Host-symbiont co-speciation and reductive genome evolution in gut symbiotic bacteria of acanthosomatid stinkbugs. BMC Biol 7, doi:10.1186/1741-7007-1187-1182.

Koga, R., Tsuchida, T., Fukatsu, T., 2003. Changing partners in an obligate symbiosis: a facultative endosymbiont can compensate for loss of the essential endosymbiont *Buchnera* in an aphid. Proc Biol Sci 270, 2543-2550.

Kolsch, G., Matz-Grund, C., Pedersen, B.V., 2009. Ultrastructural and molecular characterization of endosymbionts of the reed beetle genus *Macroplea* (Chrysomelidae, Donaciinae), and proposal of "*Candidatus* Macropleicola appendiculatae" and "*Candidatus* Macropleicola muticae". Can J Microbiol 55, 1250-1260.

Korbel, J.O., Snel, B., Huynen, M.A., Bork, P., 2002. SHOT: a web server for the construction of genome phylogenies. Trends Genet 18, 158-162.

Kostka, M., Uzlikova, M., Cepicka, I., Flegr, J., 2008. SlowFaster, a user-friendly program for slow-fast analysis and its application on phylogeny of *Blastocystis*. BMC Bioinformatics 9, 341.

Kuhnert, P., Korczak, B.M., Stephan, R., Joosten, H., Iversen, C., 2009. Phylogeny and prediction of genetic similarity of *Cronobacter* and related taxa by multilocus sequence analysis (MLSA). Int J Food Microbiol 136, 152-158.

Kyei-Poku, G.K., Colwell, D.D., Coghlin, P., Benkel, B., Floate, K.D., 2005. On the ubiquity and phylogeny of *Wolbachia* in lice. Mol Ecol 14, 285-294.

Lake, J.A., 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc Natl Acad Sci U S A 91, 1455-1459.

Lambert, J.D., Moran, N.A., 1998. Deleterious mutations destabilize ribosomal RNA in endosymbiotic bacteria. Proc Natl Acad Sci U S A 95, 4458-4462.

Lamelas, A., Perez-Brocal, V., Gomez-Valero, L., Gosalbes, M.J., Moya, A., Latorre, A., 2008. Evolution of the secondary symbiont "*Candidatus* Serratia symbiotica" in aphid species of the subfamily Lachninae. Appl Environ Microbiol 74, 4236-4240.

Le, S.Q., Gascuel, O., 2008. An improved general amino acid replacement matrix. Mol Biol Evol 25, 1307-1320.

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

Lerat, E., Daubin, V., Moran, N.A., 2003. From gene trees to organismal phylogeny in prokaryotes: the case of the γ-Proteobacteria. PLoS Biol 1, e19.

Lockhart, P.J., Steel, M.A., Hendy, M.D., Penny, D., 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol 11, 605-612.

Lopez-Sanchez, M.J., Neef, A., Pereto, J., Patino-Navarrete, R., Pignatelli, M., Latorre, A., Moya, A., 2009. Evolutionary convergence and nitrogen metabolism in *Blattabacterium* strain Bge, primary endosymbiont of the cockroach *Blattella germanica*. PLoS Genet 5, e1000721.

Mazzon, L., Piscedda, A., Simonato, M., Martinez-Sanudo, I., Squartini, A., Girolami, V., 2008. Presence of specific symbiotic bacteria in flies of the subfamily Tephritinae (Diptera Tephritidae) and their phylogenetic relationships: proposal of '*Candidatus* Stammerula tephritidis'. Int J Syst Evol Microbiol 58, 1277-1287.

McCutcheon, J.P., McDonald, B.R., Moran, N.A., 2009a. Convergent evolution of metabolic roles in bacterial co-symbionts of insects. Proc Natl Acad Sci U S A 106, 15394–15399.

McCutcheon, J.P., McDonald, B.R., Moran, N.A., 2009b. Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. PLoS Genet 5, e1000565.

McCutcheon, J.P., Moran, N.A., 2007. Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. Proc Natl Acad Sci U S A 104, 19392-19397.

McQuiston, J.R., Herrera-Leon, S., Wertheim, B.C., Doyle, J., Fields, P.I., Tauxe, R.V., Logsdon, J.M., Jr., 2008. Molecular phylogeny of the Salmonellae: relationships among *Salmonella* species and subspecies determined from four housekeeping genes and evidence of lateral gene transfer events. J Bacteriol 190, 7060-7067.

Mignard, S., Flandrois, J.P., 2008. A seven-gene, multilocus, genus-wide approach to the phylogeny of mycobacteria using supertrees. Int J Syst Evol Microbiol 58, 1432-1441.

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

Montllor, C.B., Maxmen, A., Purcell, A.H., 2002. Facultative bacterial endosymbionts benefit pea aphids *Acyrthosiphon pisum* under heat stress. Ecol Entomol 27, 189-195.

Moran, N., Munson, M., Baumann, P., Ishikawa, H., 1993. A molecular clock in endosymbiotic bacteria is calibrated using the insect hosts. Proc R Soc Lond [Biol] 253, 167-171.

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

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

Moran, N.A., Plague, G.R., Sandström, J.P., Wilcox, J.L., 2003. A genomic perspective on nutrient provisioning by bacterial symbionts of insects. Proc Natl Acad Sci U S A 100, 14543-14548.

Moran, N.A., Russell, J.A., Koga, R., Fukatsu, T., 2005a. Evolutionary relationships of three new species of *Enterobacteriaceae* living as symbionts of aphids and other insects. Appl Environ Microbiol 71, 3302-3310.

Moran, N.A., Tran, P., Gerardo, N.M., 2005b. Symbiosis and insect diversification: an ancient symbiont of sap-feeding insects from the bacterial phylum Bacteroidetes. Appl Environ Microbiol 71, 8802-8810.

Moya, A., Latorre, A., Sabater-Munoz, B., Silva, F.J., 2002. Comparative molecular evolution of primary (*Buchnera*) and secondary symbionts of aphids based on two protein-coding genes. J Mol Evol 55, 127-137.

Murray, R.G., Schleifer, K.H., 1994. Taxonomic notes: a proposal for recording the properties of putative taxa of procaryotes. Int J Syst Bacteriol 44, 174-176.

Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H.E., Moran, N.A., Hattori, M., 2006. The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. Science 314, 267.

Naum, M., Brown, E.W., Mason-Gamer, R.J., 2008. Is 16S rDNA a reliable phylogenetic marker to characterize relationships below the family level in the Enterobacteriaceae? J Mol Evol 66, 630-642.

Nikaido, M., Cao, Y., Harada, M., Okada, N., Hasegawa, M., 2003. Mitochondrial phylogeny of hedgehogs and monophyly of Eulipotyphla. Mol Phylogenet Evol 28, 276-284.

Nikoh, N., McCutcheon, J.P., Kudo, T., Miyagishima, S.Y., Moran, N.A., Nakabachi, A., 2010. Bacterial genes in the aphid genome: absence of functional gene transfer from *Buchnera* to its host. PLoS Genet 6, e1000827.

Nikoh, N., Nakabachi, A., 2009. Aphids acquired symbiotic genes via lateral gene transfer. BMC Biol 7, 12.

Nogge, G., 1978. Aposymbiotic tsetse flies, *Glossina morsitans morsitans* obtained by feeding on rabbits immunized specifically with symbionts. J Insect Physiol 24, 299-304.

Novakova, E., Hypsa, V., 2007. A new Sodalis lineage from bloodsucking fly Craterina melbae (Diptera, Hippoboscoidea) originated independently of the tsetse flies symbiont Sodalis glossinidius. FEMS Microbiol Lett 269, 131-135.

Novakova, E., Hypsa, V., Moran, N.A., 2009. *Arsenophonus*, an emerging clade of intracellular symbionts with a broad host distribution. BMC Microbiol 9, doi:10.1186/1471-2180-1189-1143.

Nubel, U., Engelen, B., Felske, A., Snaidr, J., Wieshuber, A., Amann, R.I., Ludwig, W., Backhaus, H., 1996. Sequence heterogeneities of genes encoding 16S rRNAs in *Paenibacillus polymyxa* detected by temperature gradient gel electrophoresis. J Bacteriol 178, 5636-5643.

Nylander, J.A., Wilgenbusch, J.C., Warren, D.L., Swofford, D.L., 2008. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24, 581-583.

O'Neill, S.L., Giordano, R., Colbert, A.M., Karr, T.L., Robertson, H.M., 1992. 16S rRNA phylogenetic analysis of the bacterial endosymbionts associated with cytoplasmic incompatibility in insects. Proc Natl Acad Sci U S A 89, 2699-2702.

Oliver, K.M., Campos, J., Moran, N.A., Hunter, M.S., 2008. Population dynamics of defensive symbionts in aphids. Proc Biol Sci 275, 293-299.

Oliver, K.M., Moran, N.A., Hunter, M.S., 2005. Variation in resistance to parasitism in aphids is due to symbionts not host genotype. Proc Natl Acad Sci U S A 102, 12795-12800.

Oliver, K.M., Moran, N.A., Hunter, M.S., 2006. Costs and benefits of a superinfection of facultative symbionts in aphids. Proc Biol Sci 273, 1273-1280.

Oliver, K.M., Russell, J.A., Moran, N.A., Hunter, M.S., 2003. Facultative bacterial symbionts in aphids confer resistance to parasitic wasps. Proc Natl Acad Sci U S A 100, 1803-1807.

Page, R.D., 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. Comput Appl Biosci 12, 357-358.

Paradis, S., Boissinot, M., Paquette, N., Belanger, S.D., Martel, E.A., Boudreau, D.K., Picard, F.J., Ouellette, M., Roy, P.H., Bergeron, M.G., 2005. Phylogeny of the *Enterobacteriaceae* based on genes encoding elongation factor Tu and F-ATPase beta-subunit. Int J Syst Evol Microbiol 55, 2013-2025.

Pell, P.E., Southern, D.I., 1976. Effect of the coccidiostat, sulphaquinoxline, on symbiosis in the tsetse fly, *Glossina* species. Microbios Letters 2, 203-211.

Pennisi, E., 2004. Microbiology - Researchers trade insights about gene swapping. Science 305, 334-335.

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

Perkins, S.L., Budinoff, R.B., Siddall, M.E., 2005. New Gammaproteobacteria associated with blood-feeding leeches and a broad phylogenetic analysis of leech endosymbionts. Appl Environ Microbiol 71, 5219-5224.

Perotti, M.A., Clarke, H.K., Turner, B.D., Braig, H.R., 2006. *Rickettsia* as obligate and mycetomic bacteria. FASEB J 20, 1646-1656.

Pham, H.N., Ohkusu, K., Mishima, N., Noda, M., Monir Shah, M., Sun, X., Hayashi, M., Ezaki, T., 2007. Phylogeny and species identification of the family *Enterobacteriaceae* based on *dnaJ* sequences. Diagn Microbiol Infect Dis 58, 153-161.

Pontes, M.H., Dale, C., 2006. Culture and manipulation of insect facultative symbionts. Trends Microbiol 14, 406-412.

Posada, D., 2009. Selection of models of DNA evolution with jModelTest. Methods Mol Biol 537, 93-112.

Rappe, M.S., Giovannoni, S.J., 2003. The uncultured microbial majority. Annu Rev Microbiol 57, 369-394.

Reyes, S.G., Cooper, S.J.B., Schwarz, M.P., 1999. Species phylogeny of the bee genus *Exoneurella michener* (Hymenoptera : Apidae : Allodapini): Evidence from molecular and morphological data sets. Ann Entomol Soc Am 92, 20-29.

Rodriguez, F., Oliver, J.L., Marin, A., Medina, J.R., 1990. The general stochastic model of nucleotide substitution. J Theor Biol 142, 485-501.

Roggenkamp, A., 2007. Phylogenetic analysis of enteric species of the family Enterobacteriaceae using the *oriC*-locus. Syst Appl Microbiol 30, 180-188.

Russell, J.A., Latorre, A., Sabater-Munoz, B., Moya, A., Moran, N.A., 2003. Side-stepping secondary symbionts: widespread horizontal transfer across and beyond the Aphidoidea. Mol Ecol 12, 1061-1075.

Russell, J.A., Moran, N.A., 2006. Costs and benefits of symbiont infection in aphids: variation among symbionts and across temperatures. Proc Biol Sci 273, 603-610.

Sabree, Z.L., Kambhampati, S., Moran, N.A., 2009. Nitrogen recycling and nutritional provisioning by *Blattabacterium*, the cockroach endosymbiont. Proc Natl Acad Sci U S A 106, 19521-19526.

Sakurai, M., Koga, R., Tsuchida, T., Meng, X.Y., Fukatsu, T., 2005. *Rickettsia* symbiont in the pea aphid *Acyrthosiphon pisum*: novel cellular tropism, effect on host fitness, and interaction with the essential symbiont *Buchnera*. Appl Environ Microbiol 71, 4069-4075.

Sandstrom, J.P., Russell, J.A., White, J.P., Moran, N.A., 2001. Independent origins and horizontal transfer of bacterial symbionts of aphids. Mol Ecol 10, 217-228.

Scarborough, C.L., Ferrari, J., Godfray, H.C., 2005. Aphid protected from pathogen by endosymbiont. Science 310, 1781.

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

Schlein, Y., 1977. Lethal effect of tetracycline on tsetse flies following damage to bacteroid symbionts. Experimentia 33, 450-451.

Schwarz, M.P., Tierney, S.M., Cooper, S.J., Bull, N.J., 2004. Molecular phylogenetics of the allodapine bee genus *Braunsapis*: A-T bias and heterogeneous substitution parameters. Mol Phylogenet Evol 32, 110-122.

Singer, G.A.C., Hickey, D.A., 2000. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol Biol Evol 17, 1581-1588.

Sorfova, P., Skerikova, A., Hypsa, V., 2008. An effect of 16S rRNA intercistronic variability on coevolutionary analysis in symbiotic bacteria: molecular phylogeny of *Arsenophonus triatominarum*. Syst Appl Microbiol 31, 88-100.

Southwood, T.R., Khalaf, S., Sinden, R.E., 1975. The micro-organisms of tsetse flies. Acta Trop 32, 259-266.

Spaulding, A.W., von Dohlen, C.D., 1998. Phylogenetic characterization and molecular evolution of bacterial endosymbionts in psyllids (Hemiptera: Sternorrhyncha). Mol Biol Evol 15, 1506-1513.

Sullivan, J., Swofford, D.L., 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J Mamm Evol 4, 77-86.

Susko, E., Leigh, J., Doolittle, W.F., Bapteste, E., 2006. Visualizing and assessing phylogenetic congruence of core gene sets: A case study of the γ-proteobacteria. Mol Biol Evol 23, 1019-1030.

Swofford, D.L., 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Takiya, D.M., Tran, P.L., Dietrich, C.H., Moran, N.A., 2006. Co-cladogenesis spanning three phyla: leafhoppers (Insecta: Hemiptera: Cicadellidae) and their dual bacterial symbionts. Mol Ecol 15, 4175-4191.

Talavera, G., Castresana, J., 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56, 564-577.

Tamas, I., Klasson, L., Canback, B., Naslund, A.K., Eriksson, A.S., Wernegreen, J.J., Sandstrom, J.P., Moran, N.A., Andersson, S.G., 2002. 50 million years of genomic stasis in endosymbiotic bacteria. Science 296, 2376-2379.

Tamura, K., Dudley, J., Nei, M., Kumar, S., 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24, 1596-1599.

Thao, M.L., Baumann, L., Hess, J.M., Falk, B.W., Ng, J.C., Gullan, P.J., Baumann, P., 2003. Phylogenetic evidence for two new insect-associated Chlamydia of the family *Simkaniaceae*. Curr Microbiol 47, 46-50.

Thao, M.L., Baumann, P., 2004a. Evidence for multiple acquisition of *Arsenophonus* by whitefly species (Sternorrhyncha: Aleyrodidae). Curr Microbiol 48, 140-144.

Thao, M.L., Baumann, P., 2004b. Evolutionary relationships of primary prokaryotic endosymbionts of whiteflies and their hosts. Appl Environ Microbiol 70, 3401-3406.

Thao, M.L., Clark, M.A., Baumann, L., Brennan, E.B., Moran, N.A., Baumann, P., 2000. Secondary endosymbionts of psyllids have been acquired multiple times. Curr Microbiol 41, 300-304.

Thao, M.L., Clark, M.A., Burckhardt, D.H., Moran, N.A., Baumann, P., 2001. Phylogenetic analysis of vertically transmitted psyllid endosymbionts (*Candidatus* Carsonella ruddii) based on *atpAGD* and *rpoC*: comparisons with 16S-23S rDNA-derived phylogeny. Curr Microbiol 42, 419-421.

Thao, M.L., Gullan, P.J., Baumann, P., 2002. Secondary (γ-Proteobacteria) endosymbionts infect the primary (β-Proteobacteria) endosymbionts of mealybugs multiple times and coevolve with their hosts. Appl Environ Microbiol 68, 3190-3197.

Thollesson, M., 2004. LDDist: a Perl module for calculating LogDet pair-wise distances for protein and nucleotide sequences. Bioinformatics 20, 416-418.

Thompson, J.D., Higgins, D.G., 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. Nucleic Acids Res 22, 4673-4680.

Toh, H., Weiss, B.L., Perkin, S.A., Yamashita, A., Oshima, K., Hattori, M., 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.

Toju, H., Hosokawa, T., Koga, R., Nikoh, N., Meng, X.Y., Kimura, N., Fukatsu, T., 2009. "*Candidatus* Curculioniphilus buchneri", a novel clade of bacterial endocellular symbionts from weevils of the genus *Curculio*. Appl Environ Microbiol 76, 275-282.

Tsuchida, T., Koga, R., Fukatsu, T., 2004. Host plant specialization governed by facultative symbiont. Science 303, 1989-1989.

Tsuchida, T., Koga, R., Sakurai, M., Fukatsu, T., 2006. Facultative bacterial endosymbionts of three aphid species, *Aphis craccivora*, *Megoura crassicauda* and *Acyrthosiphon pisum*, sympatrically found on the same host plants. Appl Entomol Zool 41, 129-137.

Tsuchida, T., Koga, R., Shibao, H., Matsumoto, T., Fukatsu, T., 2002. Diversity and geographic distribution of secondary endosymbiotic bacteria in natural populations of the pea aphid, *Acyrthosiphon pisum*. Mol Ecol 11, 2123-2135.

Ueda, K., Seki, T., Kudo, T., Yoshida, T., Kataoka, M., 1999. Two distinct mechanisms cause heterogeneity of 16S rRNA. J Bacteriol 181, 78-82.

Van de Peer, Y., Frickey, T., Taylor, J.S., Meyer, A., 2002. Dealing with saturation at the amino acid level: a case study based on anciently duplicated zebrafish genes. Gene 295, 205-211.

van Ham, R.C., Kamerbeek, J., Palacios, C., Rausell, C., Abascal, F., Bastolla, U., Fernandez, J.M., Jimenez, L., Postigo, M., Silva, F.J., Tamames, J., Viguera, E., Latorre, A., Valencia, A., Moran, F., Moya, A., 2003. Reductive genome evolution in *Buchnera aphidicola*. Proc Natl Acad Sci U S A 100, 581-586.

von Dohlen, C.D., Kohler, S., Alsop, S.T., McManus, W.R., 2001. Mealybug β-proteobacterial endosymbionts contain γ-proteobacterial symbionts. Nature 412, 433-436.

Wei, G., Lu, H., Zhou, Z., Xie, H., Wang, A., Nelson, K., Zhao, L., 2007. The microbial community in the feces of the giant panda (*Ailuropoda melanoleuca*) as determined by PCR-TGGE profiling and clone library analysis. Microb Ecol 54, 194-202.

Wernegreen, J.J., 2002. Genome evolution in bacterial endosymbionts of insects. Nat Rev Genet 3, 850-861.

Wernegreen, J.J., Kauppinen, S.N., Brady, S.G., Ward, P.S., 2009. One nutritional symbiosis begat another: Phylogenetic evidence that the ant tribe Camponotini acquired *Blochmannia* by tending sap-feeding insects. BMC Evol Biol 9, 292.

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

Wertz, J.E., Goldstone, C., Gordon, D.M., Riley, M.A., 2003. A molecular phylogeny of enteric bacteria and implications for a bacterial species concept. J Evolution Biol 16, 1236-1248.

Whelan, S., Goldman, N., 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18, 691-699.

Wilkes, T.E., Darby, A.C., Choi, J.H., Colbourne, J.K., Werren, J.H., Hurst, G.D., 2010. The draft genome sequence of *Arsenophonus nasoniae*, son-killer bacterium of *Nasonia vitripennis*, reveals genes associated with virulence and symbiosis. Insect Mol Biol 19 Suppl 1, 59-73.

Williams, K.P., Gillespie, J.J., Sobral, B.W., Nordberg, E.K., Snyder, E.E., Shallom, J.M., Dickerman, A.W., 2010. Phylogeny of Gammaproteobacteria. J Bacteriol 192, 2305-2314.

Williams, K.P., Sobral, B.W., Dickerman, A.W., 2007. A robust species tree for the *Alphaproteobacteria*. J Bacteriol 189, 4578-4586.

Woese, C.R., 1987. Bacterial evolution. Microbiol Rev 51, 221-271.

Wu, D., Daugherty, S.C., Van Aken, S.E., Pai, G.H., Watkins, K.L., Khouri, H., Tallon, L.J., Zaborsky, J.M., Dunbar, H.E., Tran, P.L., Moran, N.A., Eisen, J.A., 2006. Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. PLoS Biol 4, e188.

Wu, M., Eisen, J.A., 2008. A simple, fast, and accurate method of phylogenomic inference. Genome Biol 9, doi:10.1186/gb-2008-1189-1110-r1151.

Xiang, Q.Y., Moody, M.L., Soltis, D.E., Fan, C.Z., Soltis, P.S., 2002. Relationships within Cornales and circumscription of Cornaceae - *matK* and *rbcL* sequence data and effects of outgroups and long branches. Mol Phylogenet Evol 24, 35-57.

Yap, W.H., Zhang, Z., 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.

Yoder, A.D., Irwin, J.A., Payseur, B.A., 2001. Failure of the ILD to determine data combinability for slow loris phylogeny. Syst Biol 50, 408-424.

Zchori-Fein, E., Perlman, S.J., 2004. Distribution of the bacterial symbiont *Cardinium* in arthropods. Mol Ecol 13, 2009-2016.