Facultat de Ciències Biològiques

Institut Cavanilles de Biodiversitat i Biologia Evolutiva

Evolutionary genomics and functional studies on the metabolic role of *Blattabacterium,* primary endosymbiont of cockroaches

Memòria presentada per Rafael Patiño Navarrete per optar al grau de Doctor en Biotecnologia per la Universitat de València.

Directors:

Dra. Amparo Latorre Castillo. Catedràtica de Genètica de la Universitat de València

Dr. Juli Peretó Magraner. Professor Titular de Bioquímica i Biologia Molecular de la Universitat de València

València, 2013

AMPARO LATORRE CASTILLO Catedràtica del Departament de Genètica de la Universitat de València i JULI PERETÓ MAGRANER, Professor Titular del Departament de Bioquímica i Biologia Molecular de la Universitat de València,

CERTIFIQUEN que la memòria titulada “Evolutionary genomics and functional studies on the metabolic role of *Blattabacterium* primary endosymbiont of cockroaches” ha estat realitzada sota la seua direcció al Institut Cavanilles de Biodiversitat i Biologia Evolutiva per RAFAEL PATIÑO NAVARRETE.

I perquè així consti, signen el següent certificat.

València, a de de 2013

Amparo Latorre Castillo Juli Peretó Magraner

« Il y a des millions d’années que les fleurs fabriquent des épines. Il y a des millions d’années que les moutons mangent quand même les fleurs. Et ce n’est pas sérieux de chercher à comprendre pourquoi elles se donnent tant de mal pour se fabriquer des épines qui ne servent jamais à rien?...”

Antoine de Saint-Exupéry

Le Petit Prince

Aquest llibre marca el final d'un cicle ben important a la meva vida, tant pel temps i l'esforç que ha suposat com pel període vital en el que ha ocorregut. Una volta superat un repte com aquest, al llarg del qual es passen grans períodes de temps sol, es reconfortant aturar-se, reflexionar i adonar-se que al voltant teu hi ha hagut una gran quantitat de gent que d'una forma o d' altra t'han ajudat a dur a terme aquesta aventura. Així que seria imperdonable no aprofitar aquest espai per donar les gràcies a tots aquells que al llarg de la meva vida m' han ajudat a assolir aquest desafiament.

En primer lloc vull donar les gracies a Juli i Amparo, per haver confiat amb mi, just quan vaig acabar la llicenciatura. Però sobretot, agraeixo els consells, ànims i empentes que m' han donat durant el període de la tesi, així com els moments on fora del àmbit mes acadèmic he pogut gaudir del seu bon humor i atenció. Per altra banda no oblide la resta de professors del departament de genètica evolutiva al Institut Cavanilles que directa o indirectament han contribuït al meu treball.

Al llarg de la tesi es fàcil passar per moments de desànim o frustració, que gràcies al suport dels companys de laboratori s' han fet molt més tolerables. Així en primer lloc voldria donar les gràcies a Maria José, ja que fou ella la primera que em va acollir al laboratori, i em va introduir al mon de la recerca. Amb ella hi havia un grup de becaris "grans" que em van acollir amb els braços oberts. Especialment vull agrair a Eugeni i Araceli tot el que em van ensenyar, especialment quan vaig començar a analitzar dades de seqüenciació massiva i em vaig introduir al mon del Linux, sense els seus consells m'haguera vist perdut. A part vaig gaudir de molt bons moments d' esbarjo amb ells i la resta de gent (Silvia, Manel, Teresa, Mireia i Vicente). Poc a poc aquesta gent va anar marxant, però foren substituïts per nous companys. Primer, els meus veïns de laboratori, Sergio, un orador divertit i agradable, i Manzano, tot un expert en l'art de la bioinformàtica, que amb la

seua “màgia” ens facilita el treball. No volgués deixar d’anomenar Diego, sempre disposat a donar un cop de mà, Quelo, el “valencianot” del grup i a tots els mexicans que s’han unit al grup (Mariana, Oceguera i Carlos) i a la meua veïna, aquesta del poble, Vanessa. Tampoc m’oblide de les companyes d’ insecte, Elisa i Carmen. A part de la gent de la línia de genètica voldria agrair al Professor Ximo Baixeras i la gent del laboratori de Verdugo per l’ ajuda que als últims moments m’estan oferint per obtenir fotos de microscòpia electrònica. Una menció a part es mereix Peris, al que vaig conèixer a l’autoescola, i d’ ençà hem seguit una mena de vides paral·leles, fent carrera, màster i tesi junts.

Durant aquest temps he tingut la oportunitat de fer dues estades a altres laboratoris. Primer vaig estar a Lió, amb el Professor Abdellaziz Heddi, on per desgracia vaig anar lesionat, no obstant tant ell com la gent del seu laboratori em van mimar moltíssim, i em van atendre com si fos a casa. Encara recorde les seues dots culinàries! Després vaig marxar a Barcelona, al laboratori del Professor Xavier Bellés a l’Institut de Biologia Evolutiva. D’allà em va impressionar la calidesa i entusiasme que mostraven tant el Professor Bellés com la Professora Maria Dolors Piulach. Durant l’estada a Barcelona, després d’una bona temporada cara al ordinador, vaig tornar a fer treball de bancada, aquesta transició fou molt més fàcil gràcies a l’ajuda de l’Erica.

Per finalitzar els agraïments de l’àmbit acadèmic, voldria donar les gracies a tots els mestres i professors que he tingut des de l’escola fins la universitat, passant per l’institut, sempre al sistema públic. Sense ells aquesta tesi no hagués estat possible, per tant cal reconèixer la seua tasca, especialment en aquest temps tant durs.

Fora de l’àmbit acadèmic vull donar les gracies a tots els amics que he anat fent al llarg de la meua vida. La gent de Paiporta (Ferran, Jaume, Toni, Amanda, Màrius, Àngela, Gabri, Maria, Borja, Teresa, Paco, Borja i Mario) que m’ha acompanyat des de que tenia quatre anys, i que sense cap dubte han contribuït al meu creixement. Durant la tesi els moments que he passat amb vosaltres m’ han ajudat moltíssim a refrescar la ment i deixar de banda les preocupacions del laboratori. En aquest últim aspecte, també m’ha sigut de molta ajuda tots “breaks” que fèiem molts dimecres a poqueta nit (per trencar la setmana), a La Esquinita, amb Borja, Amanda i Maria, Al Bar Murta, amb Ivan Ferran i Àlex o últimament als trivial del Racó amb Maria Jesus, Àngela i la resta del equip de les Pollastretes.

Finalment, com no donar les gracies als mon pare i ma mare, que tant s’han esforçat, per a que jo em traga el doctorat. Ells son els que sempre m’han animat a continuar treballant, i m’han donat les eines per a que tinga una vida independent.

Content

I Introduction

1.- Symbiosis

1.1 Definition and classification

Symbiosis (which derives from Greek *sym* “together” and *biosis* “living”) is defined in biology as long-term associations between individuals of two or more species that have established a relationship of interdependence at any biological level (behavioural, metabolic or genetic).

The term was proposed in 1879 by the botanist Anton de Bary in his work “*Die Erscheinungen des symbiose*”, based on his observations on the lichen biology. In the de Bary’s definition, it was not expressed any judgement about the effects of the interaction on the fitness of the participants. In the most extreme cases, both members can constitute a new living being, with his own physiologically properties. In this sense, short-term associations are not considered symbiotic. This definition of symbiosis is not universally accepted. Since the birth of the term, several authors restricted the symbiosis to those cases where both organism take profit of the interaction (Saffo 1992), or in the most extreme sense, only when new metabolic capabilities arise in a determined organism through the association with other organism of a different species (Douglas 1994). The debate between those that defend this more restricted definition of symbiosis and those which still defend the broad sense, i.e., closer to the original definition proposed by de Bary (Paracer and Vernon 2000), remains open.

Under the broad sense conception of symbiosis, we can distinguish among three subtypes of association depending on the fitness effect that the establishment of the symbiotic relationship has on both partners. Thereby, if one of the members benefits from the other without causing neither harm nor benefit to the other partner, we are talking of **commensalism**, while if one of the members is beneficed by causing a decrease in the fitness of the other

partner the relationship is known as **parasitism**. Finally, when both members are benefited from the relationship the term used is **mutualism**. The three terms were described, even before de Bary proposed the term of symbiosis, by Pierre-Joseph van Beneden in the book *Les commensaux et les parasites* (1876).

Depending on the localization of the symbiont respect the host it is called **ectosymbiont** when the symbiont lives on the host's body surface, including internal surfaces like the digestive tube lining and the ducts of glands, and **endosymbiont**, when the symbiont dwells within a cell of the host. Finally, according to the degree of the symbiotic association it is **facultative** or **obligate,** depending on the ability of the organism to survive outside of the symbiotic consortium or not.

1.2 Symbiosis as a major evolutionary force

Despite the controversy about the role of the symbiosis in evolution, there is a common agreement that symbiosis is essential on the origin and early evolution of the eukaryotic cell. The first eukaryote may have appeared through a symbiotic event between a bacterium and an archea (Duve 2007). Subsequently, some of the organelles of the eukaryotic cell were the result of the invasion of the primitive eukaryotic cell by several prokaryotes, a hypothesis known as Serial Endosymbiotic Theory (SET) proposed by Lynn Margulis (Margulis 1993 and references therein) (Figure 1). Thus, the mitochondrion may be originated from an ancestral α-proteobacterium in one unique event of symbiosis. On the other hand, the chloroplast evolved from an ancestral cyanobacterium that have been engulfed and retained by the host cell. Despite the initial controversy, the symbiotic origin of mitochondria and chloroplasts is nowadays well supported by physiological,

morphological and molecular data (Gray and Spencer 1996; Martin et al. 1998).

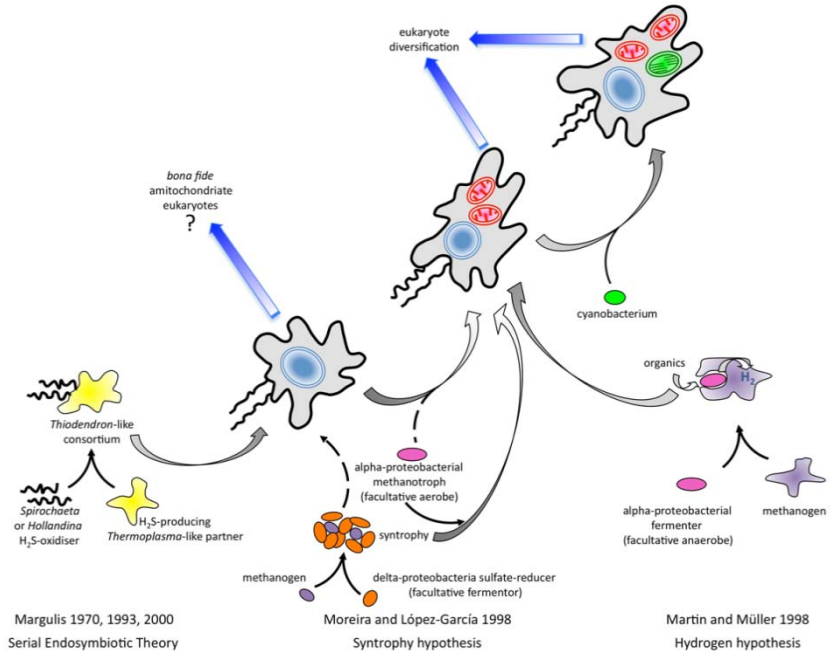

Figure 1. Origin of eukaryotic cell, and organelles through symbiogenesis (Moya and Peretó 2011 adapted from Latorre et al. 2011).

In addition to the role in the origin and early evolution of eukaryotic cells, symbiosis has shaped the evolution of life in many other ways. Thereby, the evolution of the immune system of complex eukaryotes has been deeply influenced by pathogenic interactions with other organisms. The mutualistic relationships that plants and animals established with prokaryotes and fungi would had furnished them with new metabolic capabilities, allowing the colonization of other niches that otherwise would be inaccessible.

An indicator to evaluate the importance of symbiosis in the evolution of life is the huge variety of symbiotic associations described in nature with examples in the three domains of life, with all type of combinations (Moya

et al. 2008; Mcfall-Ngai 2008; Moya and Peretó 2011). Particularly numerous are those symbiotic associations in which at least one of the members is a prokaryotic microorganism, but there are also examples for symbiosis where the two members are eukaryotes. Of these, probably the most renowned are lichens, which are associations between an algae and a fungus, but there are also associations of fungi with protists, animals, plants or other fungi. Paracer and Ahmandjian (2000) and Moya et al. (2008) have compiled multiple examples for these associations.

Prokaryotes possess an impressive set of metabolic capabilities that allowed them to colonize a broad range of ecosystems and environmental conditions, even the most extremes. This feature makes the prokaryotes good candidates to colonize the surfaces and inner spaces of other organism. Thus, many prokaryotes live physically attached to other living beings. Biofilm formation is a clear example of an association between different species of prokaryotes (Hall-Stoodley et al. 2004). But the symbiotic association between eukaryotes and prokaryotes are the most studied. Actually, there have been described symbiotic associations with prokaryotes in practically all branches of the eukaryotic tree of life (Figure 2). Eukaryotes, and especially animals, have limited metabolic capabilities and through the association with bacterial symbionts they acquired several metabolic pathways like nitrogen fixation (Kneip et al. 2007), methanogenesis (Schink 1997), chemolithoautotrophy (Stewart et al. 2005), nitrogen assimilation (Minic and Hervé 2004) and biosynthesis of several nutrients lacking in the diet (Zientz et al. 2004).

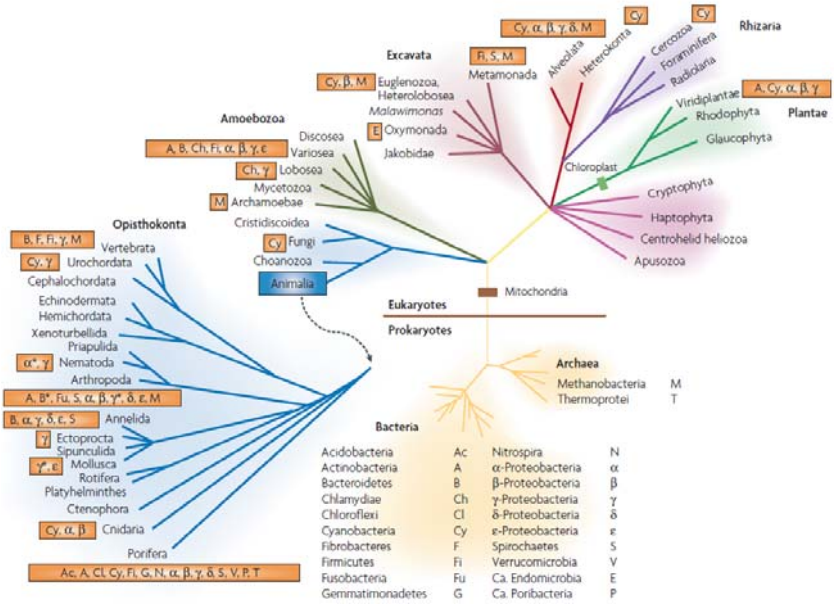

Figure 2. Phylogenetic distribution of the symbioses between eukaryotes and prokaryotes (bacteria or archea). Orange boxes indicate the phyla of the symbiont. Asterisks indicate that the genome sequence was available at the time of the review (Moya et al. 2008).

1.3 Bacterial endosymbionts in insects

Endosymbiosis is the most intimate relationship that a symbiont can establish with its host. In this case, the symbiont lives inside the host cell. Many bacteria live during part o through its whole life cycle inside the cells from other organisms. These organisms can be other bacteria, like in the case of *Bdellovibrio bacteriovorus* (Rendulic et al. 2004), or eukaryotic cells, like *Legionella pneumophila* (Cazalet et al. 2004). In both examples, the symbiotic bacteria only dwell within the host cell during a small part of their life cycle. In some other cases the relationship is even more intimate, like bacterial endosymbionts of animals, where the bacteria are transmitted

vertically from mother to offspring and usually reside in specialized cells , called bacteriocytes. Paul Buchner catalogued a large amount of these associations in 1965 in his book *Endosymbiosis of Animals with Plant Microorganisms*. Since these organisms are uncultivable, this work has been the main source of knowledge for this kind of associations until the emergence of modern genomic sciences. Thanks to new sequencing technologies it has been possible to deal with the study of these consortia at genomic level. Thus, during the last 15 years, and thanks to a wide range of studies on these systems, it has been possible to elucidate the role several of these bacteria play on the biology of their host. In his work, Buchner described endosymbionts in many animal phyla, including nematodes (Taylor et al. 2005), sponges (Schmitt et al. 2007), annelids (Graf et al. 2006), bryozoans (Sharp et al. 2007) and molluscs (Newton et al. 2007). But, by far, insects are the phyla that have received the most attention to the study of their intracellular symbionts.

Insects appeared in the Devonian (between 350-400 Mya "million years ago") (Engel and Grimaldi 2004), and diversified very quickly until they become the most diverse and successful lineage among the animals with approximately 1,000,000 described species around the world (according to diverse estimations other 5,000,000 remain to be discovered or described) (Mora et al. 2011). Insects have colonized a broad range of ecosystems and despite being quite uniform regarding nutritional requirements (Dadd 1985), they feed on a number of different diets. There are examples from generalist (omnivores or scavengers) to specialists that feed, for instance, on blood, sap, skin, fungi or nectar (Slansky and Rodriguez 1987). In many cases, the food source of insects does not satisfy their nutritional requirements, so they rely on the metabolic ability of microorganisms for the synthesis of diverse compounds absent in the diet. There are also cases where the symbionts participate in the digestion of nutrients, or in the detoxification of certain

foodstuff (Douglas 2009). In summary, symbiotic associations between insects and endosymbiotic bacteria are widespread in nature. Actually insects seem to be, among multicellular organisms, the most tolerant to live with other organisms, out or within their bodies. Thereby, it has been estimated that around 15 to 20% of insects have established symbiotic associations with intracellular bacteria. Particularly, three orders of insects are especially rich in species with endosymbiotic bacteria, Blattaria, Homoptera and Curculionidae (Dasch et al. 1984). A common feature for many endosymbionts is the diet of their host, being this usually very specialized and imbalanced for some nutrients, like the phloem or the xylem (lacking essential amino acids and other metabolites, Sandström and Pettersson 1994), blood (poor in vitamin B) or stored grains (poor in both vitamins and amino acids). In the abovementioned cases, it seems clear that the association has a nutritional value. Nevertheless, these associations may be involved in other functional roles since there are also present in insects with complex diets like cockroaches (Blochmann 1887) or ants (Blochmann 1884). The associations with mutualistic endosymbiotic bacteria may, in some way, explain the great success of insects that, thanks to the metabolic capabilities of their symbionts, can exploit food sources that otherwise would not be enough to fulfill their nutritional demands.

Depending on the grade of dependence and the age of the association, we can distinguish between two kinds of symbionts: primary endosymbionts (P-endosymbionts) and secondary or facultative endosymbionts (S-endosymbionts). P-endosymbionts are transmitted vertically from mother to offspring, live inside specialized host cells called bacteriocytes, and cannot survive out of their host. In these associations, that are generally very ancient, both members are necessary for the survival of their partner, i.e., it is a case of obligate mutualism. On the other hand, S-endosymbionts are facultative, and its presence is not universal among their probable host. Like

P-endosymbiont they are vertically transmitted; however, cases of horizontal infection have been documented (Sandstrom et al. 2001; Russell et al. 2003). S-endosymbionts are not necessarily confined in bacteriocytes or in a specific tissue, and can be found in other places like the haemolymph, glands or other non-specialized tissues, surrounding the bacteriocytes. In some cases, they can live inside their own bacteriocytes called in this case secondary bacteriocytes. S-endosymbionts cannot be considered simple commensals since it has been experimentally proved that their presence has several effects (both positive and negative) on the fitness of their host (McGraw and O'Neill 2004). The effects of the S-endosymbiont on the host are varied: some confer resistance to thermic stress (Chen et al. 2000; Russell and Moran 2006), and other protect the host against parasitoid attacks (Oliver et al. 2003, 2005) or fungal parasites (Ferrari et al. 2004). Finally, in many cases the effect of the S-endosymbionts depend on the environmental conditions in which the host grows (Tsuchida et al. 2004; Oliver et al. 2008). A third category of endosymbionts would appear when the metabolic capabilities of the P-endosymbiont are not enough to fulfill the requirements of their host. In these cases a second symbiont is needed to carry out the metabolic steps lacking in the P-endosymbiont, establishing a consortium of three members: the host and two co-primary endosymbionts with complementary metabolic pathways (Wu et al. 2006; Gosalbes et al. 2008; Lamelas et al. 2011a). Examples of symbiotic associations between bacteria and insects are summarized on Table 1.

1.3.1 Primary endosymbionts

One model system for studying symbiotic associations between insects and mutualistic intracellular bacteria is the one formed by aphids (Hemiptera:

Aphididae), specifically the aphid *Acyrtosiphon pisum*, and the γ-proteobacterium *Buchnera aphidicola*. Aphids feed on phloem, which is rich in carbohydrates but certain essential amino acids and vitamins are scarce (Douglas 1993; Sandström and Pettersson 1994; Dinant et al. 2010). At present, the genome sequence for seven strains of *B. aphidicola* are available: strains of the aphids *A. pisum* (Shigenobu et al. 2000), *Schizapis graminum* (Tamas et al. 2002), *Baizongia pistacea* (van Ham et al. 2003), *Cinara cedri* (Pérez-Brocal et al. 2006), *Cinara tujafilina* (Lamelas et al. 2011b), *Acyrthosiphon kondoi* and *Uroleucon ambrosiae* (Degnan et al. 2011). The analyses of these genomes show that the biosynthetic pathways of essential amino acids are complete, while those for the synthesis of several non-essential amino acids are absent. Additionally, in some lineages the genes involved in the synthesis of tryptophan and leucine are amplified through its translocation to plasmids (Latorre et al. 2005). Thus, the role of *B. aphidicola* as essential amino acid supplier, suggested by previous experimental works on aposymbiotic aphids, and metabolic inference (Prosser and Douglas 1992; Douglas 1998) has been confirmed after its genome sequencing.

Table 1. Endosymbiotic bacteria in insects, its classifcation and function

Host	Diet	Symbionts (Type)	Phyla	Metabolic capability
Aphids	Phloem	*Buchnera aphidicola* (primary)	γ-proteobacteria	Essential amino acids and vitamins
(*Acyrthosiphon pisum*)		*Hamiltonella defensa* (secondary)	γ-proteobacteria	Protection against parasitoids
		Regiella insecticola (secondary)	γ-proteobacteria	Protection against pathogenic fungi
		Serratia symbiotica (secondary)	γ-proteobacteria	Protection against heat stress
Aphids	Phloem	*Buchnera aphidicola* (coprimary)	γ-proteobacteria	Essential amino acids and vitamins
(*Cinara cedri*)		*Serratia symbiotica* (coprimary)	γ-proteobacteria	Tryptophan and vitamins
Psyllids (*Pachyosylla venusta*)	Phloem	*Carsonella rudii* (primary)	γ-proteobacteria	Essential amino acids?
Tsetse fly	Blood	*Wigglesworthia glossinidia* (primary)	γ-proteobacteria	B-family vitamins
(*Glossina brevipalpis*)		*Sodalis glossinidia* (secondary)	γ-proteobacteria	Inmunity
Louse (*Pediculus humanus*)	Blood	*Ca.* Riesia pediculicula (primary)	γ-proteobacteria	B-family vitamins
Sharpshooters	Xylem	*Sulcia muelleri* (coprimary)	Bacteroidete	Essential amino acids
(*Homalodisca vitripennis*)		*Baumannia cicadellinicola* (coprimary)	γ-proteobacteria	Histidine, methionine, cofactors and vitamins
Cicadas	Xylem	*Sulcia muelleri* (coprimary)	Bacteroidete	Essential amino acids
(*Diceroprocta semicincta*)		*Hodgkinia cicadicola* (coprimary)	α-proteobacteria	Histidine and methionine
Spittlebugs	Xylem	*Sulcia muelleri (*coprimary*)*	Bacteroidete	Essential amino acids
(*Clastoptera arizonana*)		*Zinderia insecticola* (coprimary)	β-proteobacteria	Histidine, methionine and tryptophan
Ants (*Camponotus* spp.)	Omnivorous	*Blochmannia* spp. (primary)	γ-proteobacteria	Nitrogen metabolism
Mealybug	Phloem	*Tremblaya princeps* (corpimary)	β-proteobacteria	Essential amino acids
(*Planococcus citri*)		*Moranella endobia* (corpimary)	γ-proteobacteria	Essential amino acids
Weevils (*Sythophilus oryzae*)	Grain	SOPE (*S. oryzae* primary endosymbiont)	γ-proteobacteria	Essential amino acids and vitamins
Arthropods		*Wolbachia* spp.	α-proteobacteria	Host reproductive parasite

The association between *Carsonella rudii* (γ-proteobacteria) and the psyllid *Pachyosylla venusta* (Hemiptera: Psylloidea) is one of the most striking associations of this kind described so far. The genome of *C. rudii* has only 160 kb and 182 coding sequences (Nakabachi et al. 2006). Psyllids feed also on phloem and would need an additional source for essential amino acids, so the function of *C. rudii* must be similar to that of *B. aphidicola* in aphids. Further analyses on this genome revealed that *Carsonella* has lost the capacity to provide the host with the essential amino acids histidine, phenylalanine and tryptophan. In addition, it has lost essential genes to carry out DNA replication, transcription and translation, thus it should not be considered anymore a living being but a new entity between a living cell and an organelle (Tamames et al. 2007). The most remarkable trait of this system is that many other psyllids harbours a S-endosymbiont, that is not the case of *P. venusta*, where *Carsonella* seems to be the only known endosymbiont (Fukatsu and Nikoh 1998; Thao et al. 2000, Nakabachi et al. 2006, Sloan and Moran 2012).

Tsetse flies (Diptera: Glossinidae) also harbour a mutualistic endosymbiotic γ-proteobacteria, *Wigglesworthia glossinidia*. Like aphids and psyllids, tsetse flies feed on a restricted diet, in this case vertebrate blood, which is deficient in vitamins. Antibiotic treated flies showed reduced growth and fecundity rates (Nogge 1976), which can be restored when B-complex vitamins are supplied (Nogge 1981), suggesting that *Wigglesworthia* is supplying these vitamins to the fly. The analysis of the genome sequence from *W. glossinidia* endosymbiont of *Glossina brevipalpis* reinforced these evidence, since it codes for a number of genes involved in the biosynthesis of cofactors and vitamins (Akman et al. 2002). The human body louse, *Pediculus humanus*, also feeds on blood and like tsetse flies needs a supply of vitamins B, in this case these vitamins are synthetized by

another γ-proteobacteria endosymbiont, *Riesia pediculicola* (Kirkness et al. 2010).

Ants from the genus *Camponotus,* in spite of being omnivorous, also harbour mutualistic intracellular bacteria. Their association with *Blochmannia* spp., is the first bacteryocite associate symbiosis described (Blochmann 1884). Currently, there are available three genomes of *Blochmannia* species: *B. floridanus*, symbiont from the ant *C. floridanus* (Gil et al. 2003), *B. pennsylvanicus*, symbiont from *C. pennsylvanicus* (Degnan et al. 2005) and *B. vafer* the symbiont of *C. vafer* (Williams and Wernegreen 2010). Given the diet and that aposymbiotic ants did not showed any deficiency, the role of *Blochmannia* was not clear (Sauer et al. 2002). However, the analysis of the genome of *B. floridanus* pointed out that *Blochmannia* was supplying their host with essential amino acids and reduced sulfur, as well as being involved in the metabolism of nitrogen (Gil et al. 2003). Further analyses showed that when ants were fed with a diet enriched with essential amino acids, there were no differences between aposymbiotic and ants carrying the endosymbiont. However, when the diet lacked essential amino acids, aposymbiotic ants raised less brood than xenic ones (Feldhaar et al. 2007). It has been observed that wild ants feed on plant or sap sucking insect exudates and could be considered as "secondary hervibores", having the same diet deficiencies than other sap sucking insects (Davidson et al. 2003). Thus, the presence of *Blochmannia* may allow *Camponotus* ants to exploit other sources of food that otherwise would be limited in several nutrients. Other works suggest that *Blochmannia* may not be necessary during the whole life cycle of ants, since the replication activity of *B. floridanus* varies greatly according to the developmental stage of the host, being higher in pupae close to the hatching and decreasing in older adults. Even in old queens there are no more symbionts (Wolschin et al. 2004). Other studies showed how the expression of several genes involved in

nitrogen recycling and metabolism of aromatic amino acids is regulated according to the developmental stage (Zientz et al. 2006). These results suggested that *Blochmannia* is important at definite points of the development of the host. Finally, works on *C. fellah* suggested that *Blochmannia* also improves the host immune system (De Souza et al. 2009).

In some cases, two or more endosymbionts cohabit in the same host, being all of them necessary for its survival, i.e. co-primary endosymbionts. That is the case of the consortium formed by *B. aphidicola* (BCc) and *Serratia symbiotica* (SCc) in the cedar aphid, *C. cedri*. The 416 kb genome of *B. aphidicola* BCc is the most reduced genome among sequenced *Buchnera* strains (Gil et al. 2002). *B. aphidicola* BCc cannot synthetize tryptophan and riboflavin, thus being unable to fulfil the nutritional requirements of its host (Pérez-Brocal et al. 2006). *S. symbiotica* is a facultative symbiont for a number of aphid species, however it has been observed that is present in all individuals of *C. cedri* with a similar density to that of *B. aphidicola* (Gómez-Valero et al. 2004). The sequence and metabolic reconstruction of *S. symbiotica* from *C. cedri* (SCc) genome suggested that its main role would be the biosynthesis of cofactors and vitamins, but the synthesis of essential amino acids would still be under the responsibility of *B. aphidicola* with the exception of tryptophan, since both symbionts are necessary for the synthesis of this amino acid (Gosalbes et al. 2008; Lamelas et al. 2011a).

Insects classified into the suborder Auchenorryncha from the order Hemiptera, which comprises, among others sharpshooters, cicadas or spittlebugs, possess one of the most striking symbiotic systems, and a number of species have two or more symbionts. The most widely distributed is *Sulcia muelleri,* a Bacteroidete that infected the common ancestor to all Auchenorryncha in the Permian between 260 to 280 Mya (Moran et al. 2005). Up to now, four *S. muelleri* genomes have been sequenced, those

from the sharpshooters *Homalodisca vitripennis* (GWSS) and *Draeculacephala minerva* (DMIN) (McCutcheon and Moran 2007; Woyke et al. 2010) the cicada *Diceroprocta semicincta* (DSEM) (McCutcheon et al. 2009a) and the spittlebug *Clastoptera arizonana* (CARI) (McCutcheon and Moran 2010). All these strains possess an extremely reduced genome ranging from 243 kb to 277 kb. The analysis of their metabolic networks indicate that all strains are able to synthetize most of the essential amino acids with the exception of histidine and methionine, and the strain CARI has also lost the entire tryptophan operon. As pointed above, most of Auchenorryncha have additional symbionts, and three of them have also been sequenced, namely, *Baumannia cicadellinicola* (γ-proteobacteria) from *H. vitripennis* (Wu et al. 2006), *Hodgkinia cicadicola* (α-proteobacteria) from *D. semicincta* (McCutcheon et al. 2009b) and *Zinderia insecticola* a β-proteobacteria symbiotic from *C. arizonana* (McCutcheon and Moran 2010). Like *S. muelleri,* the last two possess extremely reduced genomes (144 and 208 kb respectively); whereas the genome of *B. cicadellinicola* (686 kb) is not so reduced. These three co-symbionts have evolved to complement the genome of their respective *S. muelleri* partners, thus all three are able to synthetize histidine and methionine, *Zinderia* possess the machinery necessary for the synthesis of tryptophan (McCutcheon and Moran 2010), and *B. cicadellinicola* can also synthetize cofactors and B-family vitamins (Wu et al. 2006).

Probably, the most extreme case of dual symbiosis is the one described in the mealybug *Planococcus citri* (Hemiptera: Pseudococcidae) which harbours two endosymbionts, a β-proteobacteria (*Tremblaya princeps*) and a γ-proteobacteria (*Moranella endobia*) that dwells inside the first endosymbiont (Dohlen et al. 2001). *T. princeps* is distributed among all mealybugs and seems to be originated in a single event of infection in the common ancestor of the mealybugs. Thereby, the second partner should

have been acquired through subsequent infections of bacteria from multiple origins (Thao et al. 2002). The sequencing and analysis of the genome of both symbionts suggest that they are supplying their host with essential amino acids since, similarly to aphids, mealybugs feed on phloem sap (McCutcheon and Dohlen 2011; López-Madrigal et al. 2011).

1.3.2 Secondary endosymbionts

Several additional bacteria can infect aphids, besides the association that they have established with *Buchnera*. As stated before, these bacteria, known as secondary or facultative symbionts, are not essential for the survival and reproduction of their host and, in contrast to *Buchnera,* are not universally distributed. But in some occasions the infection by these symbionts provides the aphid with new adaptive traits. Thus, aphids infected with *Hamiltonella defensa* (γ-proteobacteria) show resistance to the attack of the parasitoid wasps *Aphidius ervi* and *Aphidius eadyi* (Oliver et al. 2003). The genome of *H. defensa* possess several strains of a lysogenic bacteriophage called APSE (*A. pisum* Secondary Endosymbiont) that code for several eukaryotic toxins (Degnan et al. 2009), which are related to the resistance *A. pisum* phenotype associated to *H. defensa* infection (Degnan and Moran 2008). *Regiella insecticola,* a γ-proteobacterium, confers protection against the infection by the fungi *Pandora neoaphidis* (Ferrari et al. 2004). *H. defensa* and *R. insecticola* share a relatively recent common ancestor, and both are themselves obligate symbionts depending on the essential amino acids synthetized by *Buchnera*, since they are only capable to synthetize lysine and threonine, but possess transporters for the other essential amino acids (Degnan et al. 2010). Other secondary symbiont, *S. symbiotica*, protect the host against environmental heat stress (Russell and Moran 2006). This symbiont, just like *Hamiltonell*a and *Regiella*, has lost the genes for the

synthesis of several essential amino acids (Burke and Moran 2011). It is worth to mention, as pointed above, that in some lineages of the subfamily Lachninae, *S. symbiotica* has become a co-obligate endosymbiont together with *B. aphidicola*, because both endosymbionts are needed to fulfil the metabolic requirements of the whole system (Lamelas et al. 2008; 2011b).

Tsetse flies, besides *W. glossinidia* possess a facultative S-endosymbiont, *Sodalis glossinidia*. There is evidence suggesting that the infection by this symbiont favours the infection by trypanosomes in the midgut of the fly (Aksoy 2000) and the selective elimination of this symbiont reduces de lifespan of the host (Dale and Welburn 2001). The genome and metabolic network of *S. glossinidius* is closer to a free living bacterium than to a symbiotic one (Toh et al. 2006; Belda et al. 2010; Belda et al. 2012).

The most widespread intracellular bacteria are *Wolbachia* (α-proteobacteria), because they infect up to 40% of terrestrial arthropods (Zug and Hammerstein 2012). This high incidence is probably due to the fact that these bacteria are able to manipulate the host reproduction in their own benefit through mechanisms such as cytoplasmic incompatibility, parthenogenesis, male-killing and feminization (Stouthamer et al. 1999).

1.4 Genomic changes in endosymbiotic bacteria

There are several drastic changes that occur during the transition from free-living bacteria to intracellular endosymbionts. The most prominent are reduced genome size encoding for a small number of genes and high AT content (McCutcheon and Moran 2012). Two main factors can drive genome shrinkage in these bacteria: first, the particular environment where the bacteria reside and, second, the particular population dynamics to which endosymbionts are subjected (Moya et al. 2008). Living inside a eukaryotic cell renders several genes unnecessary or redundant, since the coded

functions can be achieved by the host. Thus, the loss of these genes has no effect on the bacterial fitness and the pressure of natural selection over them is relaxed. The strict vertical transmission from the mother to the offspring reduces the effective population size, increasing the effects of random genetic drift (Moran 1996). These two factors facilitate the fixation of slightly deleterious mutations in non-essential genes producing their inactivation and posterior loss. Among the genes affected by these processes are those involved in DNA repair, recombination and DNA uptake. Losing genes in such functions brings to further increases in the mutation rate and prevent the genetic exchange by means of homolog recombination. In addition, the isolation to which these bacterial populations are subjected eliminates the possibility of gaining new genetic material through horizontal gene transfer events that could compensate the losses.

The genome reduction processes can be divided in two main phases. First, soon after the establishment of the symbiosis, there is a huge proliferation of mobile elements, like insertion sequences (IS), favoured by the relaxation of purifying selection (Moran and Plague 2004; Gil et al. 2008). The accumulation of IS enhances the reduction process by increasing the events of intrachromosomal homolog recombination, which induces genome rearrangements and loss of large genomic fragments (Parkhill et al. 2003). Further transposition can inactivate individual gene that would be degraded afterwards (Gil et al. 2008). In a second stage the genome reduction continues through the pseudogenization and loss of individual genes scattered throughout the genome (Silva et al. 2003). At this point genomes are completely stabilized and there are not more rearrangements (Tamas et al. 2002), since mobile elements are absent in the genomes of bacteria that have reached this stage (Moya et al. 2008) (Figure 3).

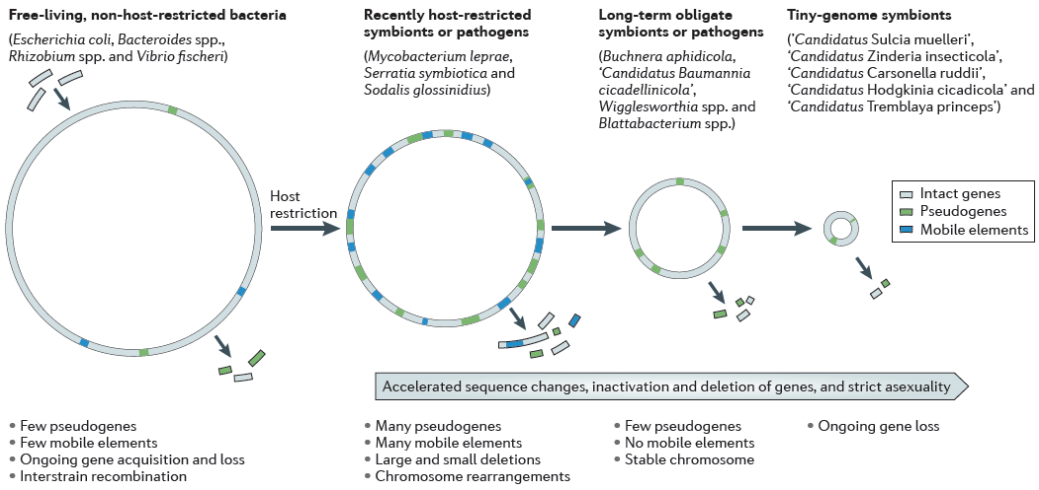

Figure 3. Graphic representation of the different stages of the genome reduction process in obligate endosymbiotic bacteria (McCutcheon and Moran 2012).

Two possible explanations have been proposed to explain the strong bias towards AT that show these genomes. The first one suggest that this bias is a consequence of the loss of DNA repair systems and the mutational pressure from GC to AT (Wernegreen 2005; Lind and Andersson 2008). The higher energetic cost of the biosynthesis of GTP and CTP nucleotides with respect to the cost of producing ATP or TTP has also been suggested as a possible reason of the bias towards AT (Rocha and Danchin 2002). The existence of this bias has strong consequences, such as the loss of codon-usage bias observed in endosymbionts (Rispe et al. 2004). Additionally, the amino acid composition of the proteome of these bacteria has been dramatically altered decreasing the thermal stability of their proteins. The overexpression of a chaperonin like GroEL could compensate the altered structure of many of these proteins (Fares et al. 2004).

2.- The cockroaches

Cockroaches (Blattodea) together with termites (Isoptera) and mantids (Mantodea) form the superorder of Dictyoptera, mainly characterized by having a perforation in the tentorium (the internal skeleton of the head) and enclosing their eggs in an ootheca (Kristensen 1981). The relationships within this group have been recently revised demoting the order Isoptera to a family (Termitidae) nested in the order Blattodea, rendering the cockroaches as a paraphyletic group (Inward et al. 2007).

Cockroaches are one of the most ancient groups of insects described so far, as they are the oldest exemplars reported in the fossil record coming from deposits from the Middle Carboniferous. Later, in the same period, cockroaches were the dominant insect but it was not until the transition between the Jurassic and the Cretaceous (140 Mya) that the extant families appeared (Table 2). Nowadays, about 5,000 living species are described with 1,000 more being extinct (Vrsansky et al. 2002). The families Blattidae, Blattellidae and Blaberidae comprise the majority of the 5,000 classified species. Cockroaches have three developmental stages: egg, nymph and adult, and like other hemimetabolous insects, early stage nymphs resemble adult specimens without wings (Bell et al. 2007).

Despite its negative reputation, only four species are adapted to human habitats, and therefore responsible of domestic pests. Most cockroaches inhabit forests and are much less adapted to other environments, like water, caves or nest than other insect species. However, the vast majority of studies are focused on those species which live in close contact to humans, mainly because the health problems linked to these species, like allergies, asthma or infections caused by pathogenic organisms associated with them (Robinson 2005).

Table 2. Classification of the order Dictyoptera based on the proposition from Inward et al. 2007.

Order	Superfamily	Family
Blattodea	Polyphagoidea	Nocticolidae
		Polyphagidae
	Blattoidea	Blattidae
		Cryptocercidae
		Termitidae
	Blaberoidea	Blattellidae
		Blaberidae
Mantodea		Mantoida

2.1 *Blatta orientalis* and *Blattella germanica*

Two species of cockroaches have been used during the development of this work, *Blatta orientalis* (Blattoidea: Blattidae) and *Blattella gemanica* (Blaberoidea: Blattellidae). Natural populations of *B. orientalis* occur in the litter leaf of forest from zones with warm summer and moderate winter on the Crimean Peninsula and the regions around the Black and Caspian seas (Robinson 2005). Despite no natural populations have been described for *B. germanica*, their natural habitats probably are, like for *B. orientalis*, moist litter leaf in the ground of forest in west Africa (Robinson 2005; Guthrie and Tindall 1968). Both species are now worldwide distributed, and always associated to human habitats, being two of the four cockroach species considered as domestic pests. The major biological properties of these two species are summarized in Table 3, notwithstanding most of these features are highly influenced by environmental conditions, mainly the temperature and the humidity.

Table 3. Biological features of the two cockroach species used during the development of this work (Guthrie and Tindall 1968; Robinson 2005; Short and Edwards 1991).

Cockroach species	***B. orientalis***	***B. germanica***
Classification	Blattoidea: Blattidae	Blaberoidea: Blattellidae
Habitat	Fresh (2–29 ºC) and humid places like basements, cellars crawls, and pipes. Fresh and moist environments.	Live only indoors, in warm (30 ºC), humid dark places, like behind refrigerators, stoves or water heaters in the kitchens.
Morphology	Black colour and wings not fully developed.	Pale brown, with two marked stripes in the pronotum. Wings fully developed.
Males	17–29 mm, reduced wings that do not reach the tip of the abdomen.	10–13 mm, elongated abdomen, larger than the wings.
Females	20–27 mm, wings do not reach the abdomen.	12–15 mm, darker than males, with a more round abdomen.
Embryonic development	45–56 days.	15–30 days.
Nymph period	185–215 days, 7-9 instars in males and 8–10 in females.	60–65 days and 6 instars.
Adult life span	143.6 ± 4.1 days for females and 87.2 ± 9.1 days for males.	180 ± 6 days for both sexes.
No. of eggs/ oothecae	12–18.	35–48, declining with the age.

2.2 Fat body anatomy

In cockroaches, like in other insects, the fat body has a central role in a number of physiological processes during the insect life. Beyond the role in the storage of nutrients, this organ has also endocrine functions, participates in the insect immune system and has a role in the detoxification of nitrogen (Arrese and Soulages 2010). Additionally, it is in the fat body where cockroaches harbour its endosymbiotic bacteria, *B. cuenoti* (Buchner 1965). Cockroach fat body is distributed among the abdominal haemocoel very close to the gut. Three cell types have been identified in the cockroach fat bodies (Figure 4): thropocytes, bacteriocytes and uricocytes. Thropocytes are located on the peripheral layers of the fat body lobes, whereas the other two cell types remain internal (Cochran 1985). The trophocytes, the most numerous, show different features according to the cockroach life stage. In nymphs and aged adult females these cells contain lipid droplets and a huge amount of glycogen. However, in young adult females, glycogen is severely reduced and the cell is completely filled with lipid droplets. The bacteriocytes, the cells containing *Blattabacterium* (Figure 4A and 4B), are rich in glycogen, with few organelles and most of the cytoplasm filled with bacteria (Piceis et al. 1986). The last cell type, uricocytes, are rich in glycogen, but the most remarkable trait of these cells is the presence of the urate structural units, which are vacuolar bodies with a dark centre surrounded by a greyish area. Urates are deposited around the dark area (Piceis et al. 1986; Cochran et al. 1979).

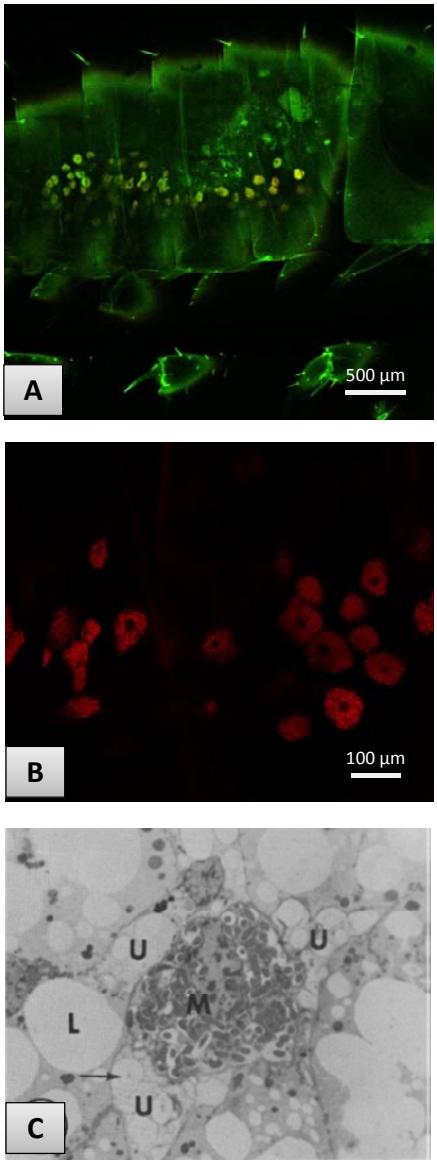

Figure 4. Visualising *Blattabacterium* cells**. A** and **B.** Bacteriocytes from a whole mount *B. germanica* hybridised with probes Eub338 for eubacteria and blb6 specific for *Blattabacterium* (López-Sánchez et al. 2008), labelled with FAM and Cy3 respectively using confocal microscopy (following the protocol described by Koga et al. 2009). **C.** The three cell types on the *P. americana* fat body. L, Trophocytes, U, uricocytes; arrow, urate spherule (amplification x850) (Cochran et al. 1979). Note how uricocytes are surrounding the bacteriocytes (M).

3.- *Blattabacterium cuenoti,* primary endosymbiont of cockroaches

3.1 Morphological characterization, localization and transmission

The endosymbionts of cockroaches were one of the first intracellular mutualistic bacteria described (Blochmann 1887). Initially they were known as Blochmann bodies, and later were named as *Bacillus cuenoti* (Mercier 1906), and is not until 1931 that a new genera was proposed to classify these bacteria that were renamed as *Blattabacterium cuenoti* (Hollande and Favre 1931). They are found inside specialized cells called bacteriocytes (formerly known as mycetocytes) in the visceral fat body of the abdomen of all cockroaches, and in the termite *Mastotermes darwiniensis* (Bandi et al. 1995). The only group of cockroaches that have lost *Blattabacterium* are cave dweller cockroaches from the genus *Nocticola* (Lo et al. 2007), whereas it has been lost in all the higher termites. The cells of *Blattabacterium* are rod shaped, with a diameter of 0.9 μm and between 1.5 to 8 μm of length that varies according the species: in *B. orientalis* are 2.5–5.3 μm long, while in *B. germanica* are about 3 μm (Kambhampati 2010). Endosymbionts are also localized in the follicular epithelium of the ovarioles and in this case they are shorter than those located in the fat body (Sacchi et al. 1996). The cytoplasm is uniformly granular and dense, with a low density area in the centre corresponding to the nuclear body (Piceis et al. 1986). The cell wall is very thin (from 5 to 10 nm) and there is a plasma membrane measuring between 1.7 and 3.5 nm inside it (Brooks 1970). Each bacterium is surrounded by a host derived vacuolar membrane, leaving a variable width vacuolar space between the bacterium and the membrane (Piceis et al. 1986; Kambhampati 2010).

Like other obligate mutualistic intracellular bacteria, *Blattabacterium* is transmitted vertically from mother to offspring. The infection of the ovaries

by the symbionts starts early in the development as, during embryonic life, bacteriocytes are in close contact with the ovary. However, the ovaries are sterile until the hatching and at this time, bacteriocytes invade the ovarioles and get adhered to the oocyte. Later, the bacteria are released from the bacteriocytes and migrate through the tunica propia and the follicular epithelium to the space between the latter and the oocyte (Sacchi and Grigolo 1989). In a further step, bacteria are engulfed by a pseudopod-like structure from the egg membrane and internalized (Sacchi et al. 1996).

3.2 Phylogenetic position

Initially, cockroach endosymbionts were classified into the genera *Bacillus* (Mercier 1906). However, due to these bacteria do not form spores, in 1931 Hollande and Favre (1931) re-described the symbiont and proposed a new genus, *Blattabacterium*. Until the availability of molecular data, the taxonomic assignation of *Blattabacterium* was quite difficult, mainly due to the fact that it is not possible to growth these bacteria outside of cockroaches. *Blattabacterium* were initially classified within the order Rickettsiales (Dasch et al. 1984). However, the first molecular phylogenetic analysis clearly clustered *Blattabacterium* into the Cytophaga-Flavobacterium-Bacteroides phylum (Bandi et al. 1994). This result was lately corroborated by our group (Figure 5) (López-Sánchez et al. 2008). In addition, the endosymbiont of the termite *M. darwiniensis* was also unambiguously placed into the clade formed by *Blattabacterium* (Bandi et al. 1995).

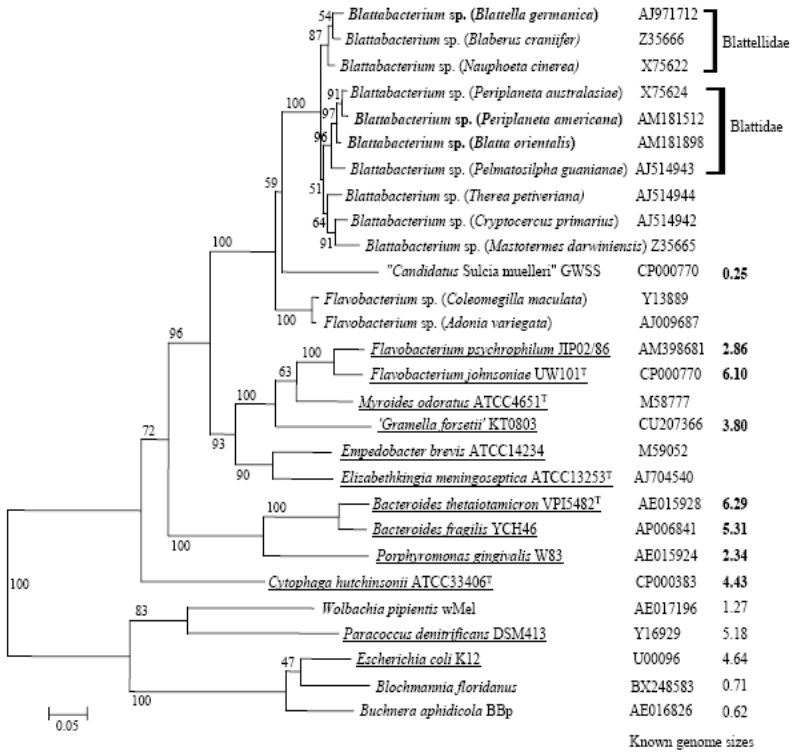

Figure 5. Phylogenetic position of *Blattabacterium* among the Bacteriodetes, according the 16S rDNA sequences. Tree obtained with the maximum likelihood method by López-Sánchez and coworkers (2008).

The tree representing the phylogeny of different strains of *Blattabacterium* is congruent with the tree of their hosts (Lo et al. 2003). These results point to a monophyletic origin of the primary endosymbionts of cockroaches and a co-cladogenesis between hosts and endosymbionts. Thus, a single infection event may have occurred in the common ancestor of cockroaches and termites more than 140 Mya, when the extant families of cockroaches appeared in the fossil record (Vrsansky et al. 2002), but

probably cockroach ancestors and *Blattabacterium* are coevolving much earlier.

3.3 Functional role of *Blattabacterium*

Traditionally, most of the knowledge about the biochemical and physiological functions of *Blattabacterium* came from studies with aposymbiotic roaches. The offspring of aureomycin-treated parents were free of *Blattabacterium*. These nymphs were weaker, smaller and lighter than normal nymphs. Additionally, when they were reared with commercial dog food, the usual laboratory diet, cockroaches did not moult even after thirty days, when normal roaches usually moult every ten days. Aposymbiotic nymphs only reached mature state if their diet was supplemented with dried brewer's yeast and, even in this case, they needed twice the time than normal roaches to become adults, and in no case these animals were fertile (Brooks and Richards 1955). Experimental data showed that *B. germanica* could incorporate sulfate into cysteine and methionine. However, aposymbiotic individuals when injected with labelled [^{35}S]-sulfate were not able to incorporate it to cysteine and methionine (Block and Henry 1962). Furthermore, aposymbiotic hosts are unable to synthetize the essential amino acids phenylalanine, isoleucine, valine, threonine, arginine and tyrosine (Henry 1962). It is also remarkable that the fat body of aposymbiotic *P. americana* individuals has lower concentrations of ascorbic, folic and pantothenic acids (Ludwig and Gallagher 1966).

Besides the synthesis of essential amino acids and vitamins, classical works also pointed to an involvement of *Blattabacterium* in the uric acid metabolism of their hosts, since it was observed that in aposymbiotic cockroaches there was a huge increase in the amounts of uric acid stored in the fat body (Malke and Schwartz 1966; Valovage and Brooks 1979). This

association between the endosymbionts and the uric acid production and storage was reinforced by the observation that bacteriocytes are usually surrounded by uricocytes (Cochran et al. 1979). Contrary to most insects, cockroaches do not excrete excess of nitrogen as uric acid. Instead, the major nitrogen waste product is ammonia (Mullins and Cochran 1972; Mullins and Cochran 1976; Cochran 1985), thus cockroaches seem ammonotelic and not uricotelic as expected for a terrestrial insect (Needham 1937). By contrast, it is well known that cockroaches are able to produce uric acid and store it in the fat body mainly in the uricocytes (McEnroe and Forgash 1957). The stores of uric acid became extremely increased when cockroaches were reared with diets with a high protein content, even reaching a toxic level (Haydak 1953; Mullins and Cochran 1974; Mullins and Cochran 1975a). In other experiments, cockroaches were allowed to increase their levels of stored uric acid with a protein-rich diet. Later on, these animals were fed with a diet poor in nitrogen. In these cases the amount of uric acid in the fat body drooped dramatically (Mullins and Cochran 1974; Mullins and Cochran 1975b). From these observations, it was assumed that stored urates in the fat body act as a metabolic reservoir of nitrogen that would be mobilized in periods of scarcity of this element (Cochran 1985). At cellular level, it became clear that as nitrogen dietary levels increase, uricocytes become enlarged and filled with urate spherules in *P. americana* (Figure 6), while individuals fed on a nitrogen free diet for 4 to 6 weeks showed a lower number of bacteriocytes and a decrease in uricocyte size (Cochran et al. 1979).

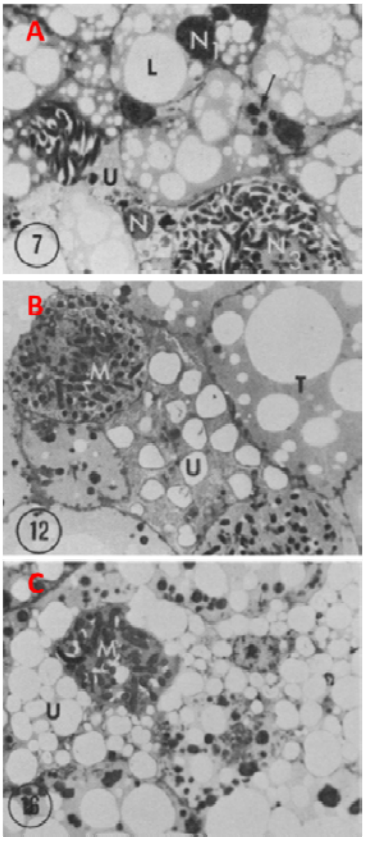

Figure 6. Cytological changes in the fat body of *P. americana* according the protein content of their diet (X800) (Cochran et al. 1979). U, uricocytes, M, bacteriocyte; T, trophocyte. White spheres within uricocytes are the urate spherules. **A,** Cockroach fed with dextrin, 0% protein. N_1, N_2 and N_3 are the nucleus of the trophocyte, the urocyte and the bacteriocytes respectively. **B**, Cockroach with a diet with 42% of protein content. **C**, Cockroach with a diet with 66% of protein content. Note how the urate spherules increase with the diet protein content, and compare with Figure 5 (animals fed with commercial dog food).

3.4 Genomic features of *Blattabacterium*

During the development of this thesis five genome sequences from different *Blattabacterium* strains have been released: BBge, endosymbiont of *B. germanica* (Blattellidae) (López-Sánchez et al. 2009) and BBgi, from

Blaberus giganteus (Blaberidae) (Huang et al. 2012) both members of the Blaberoidea superfamily, BPam the endosymbiont from *P. americana* (Blattidae) (Sabree et al. 2009), and BCpu and BMad, endosymbionts from the subsocial cockroach *Cryptocercus punctulatus* (Cryptocercidae) (Neef et al. 2011), and the termite *M. darwiniensis* (Termitidae) (Sabree et al. 2012), respectively. The last three species belong to the superfamily Blaberoidea, and in contrast to the other roaches that are considered omnivorous, both, *M. darwiniensis* and *C. punctulatus,* are xylophagous (See Table 2 for taxonomy).

The genome of *Blattabacterium* shows all the typical features of other insect primary endosymbionts: (i) a reduced genome (from 590 to the 640 kb) in comparison with free living Flavobacteria (genome sizes ranging from 2.2 to 6.1 Mb); (ii) all strains show a low GC content (from 23.8 to 28.2 %); (iii) despite the long-time divergence among all the *Blattabacterium* compared, synteny has been maintained mainly unaltered, with only two detected chromosomal rearrangements: one inversion in BPam (about 19 kb) and other in BMda of 242 kb (Sabree et al. 2010, 2012); (iv) finally, these bacteria exhibit elevated rates of nonsynonymous substitutions per nonsynonymous site (dN) (comparison between BPam and BBge) in comparison with free living relatives (Sabree et al. 2010).

Blattabacterium from omnivorous cockroaches (i.e., strains BPam, BBge and BBgi) possess the biosynthetic capabilities for the synthesis of ten essential and six non-essential amino acids, being auxotrophic for glutamine, glycine, proline and asparagine. Additionally, they must be able to supply their hosts with cofactors and vitamins (López-Sánchez et al. 2009; Sabree et al. 2009, Huang et al. 2012). These results are in concordance with observations obtained previously with aposymbiotic animals that suggested a possible role of *Blattabacterium* as an essential amino acid and cofactor supplier for the cockroaches (Henry 1962; Ludwig and Gallagher 1966). The

symbionts from the *C. punctulatus* and *M. darwiniensis*, have lost the capability to synthetize seven out of ten essential amino acids, retaining only the pathways for the synthesis of phenylalanine, histidine, tyrosine and arginine and BCpu has also lost the genes for the synthesis of cysteine. These functional losses could be compensated by gut microbiota (Neef et al. 2011; Sabree et al. 2012).

All *Blattabacterium* strains sequenced up today possess the urea cycle as well as the genes coding for urease. As suggested by López-Sánchez et al. (2009) (Figure 7), the presence of this enzyme is a key factor to explain the intriguing ammonotelism showed by the cockroaches (Mullins and Cochran 1972; Mullins and Cochran 1976), and the physiological use of uric acid as a nitrogen storage (Cochran et al. 1979; Cochran 1985). Nevertheless, urease is not enough to explain the process of nitrogen recycling from uric acid, since additional activities for the synthesis and catabolism of urates, as well as the presence of the activity glutamine synthase to incorporate the released ammonia by the urease into the metabolism, are required. None of these activities are encoded by the endosymbiont genome and thus must be carried out by the host, and/or by the gut microbiota. Since it has been demonstrated that some tissues of *P. americana* code for urate oxidase, it seems plausible that animal could produce urea from uric acid degradation (Figure 7).

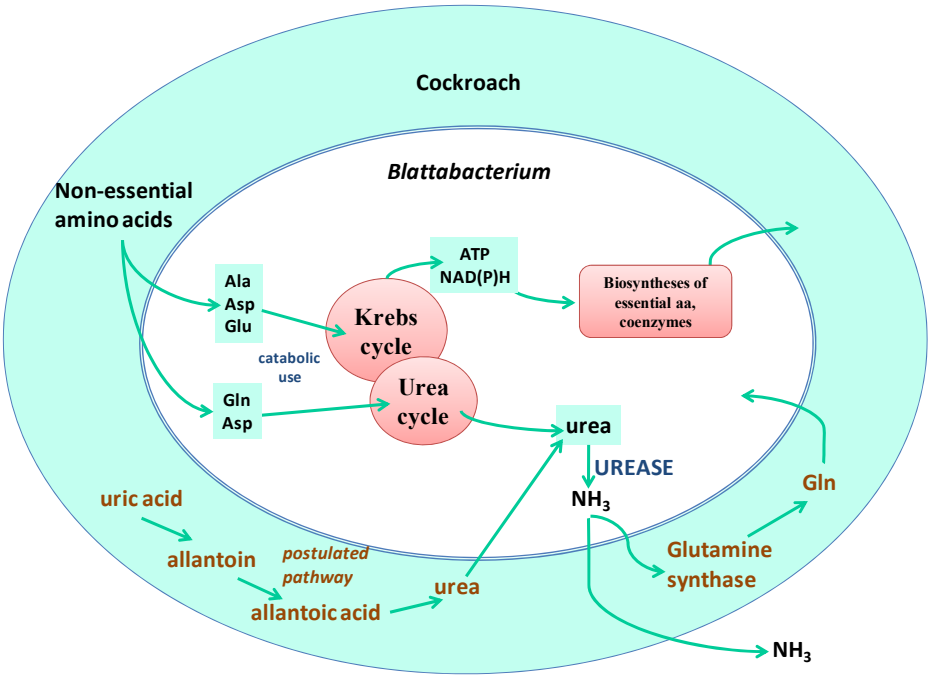

Figure 7. Proposed model for the nitrogen recycling in cockroaches by López-Sánchez et al. (2009).

The stoichiometric analysis of the central metabolic pathways of BBge showed the potential of the endosymbiont to produce ammonia (López-Sánchez et al. 2009).

II Objetives

This work is embedded in a research program devoted to the study of the reductive evolution in endosymbiotic bacteria from insects, as well as the adaptation of these bacteria to the intracellular life. Comparative genomic studies carried out on the genomes of the sequenced endosymbiotic bacteria in different evolutive stages, have shed light on how the transition from free-living bacteria to an obligate symbiont of insects might have occurred. Most of these studies have been performed with endosymbionts belonging to the γ-Proteobacteria, and there is less information from other phyla of endosymbiotic bacteria. This work is focused on the symbiotic consortium formed by cockroaches and its Bacteroidete endosymbiont *B. cuenoti*. The genome analysis of several strains of this bacterium suggests its role in nitrogen metabolism, particularly how the endosymbiont allows cockroaches to use uric acid as nitrogen storage.

The main goal of this work is depicting the evolutionary history of *Blattabacterium* through the genome sequencing of a new strain, the endosymbiont of the cockroach *Blatta orientalis* (BBor) and its comparison with other *Blattabacterium* genomes, as well as the function of the shared nitrogen metabolism between the host and the symbiont using *B. germanica* as model.

The specific objectives are:

1. To obtain the complete genome sequence and annotation of the genome of *Blattabacterium,* strain BBor.
2. To perform comparative evolutionary and metabolic analyses of the six strains of *Blattabacterium* with complete genome published so far.

3. To perform functional studies by sequencing transcriptomes from different host tissues.
4. To compare the expression of genes involved in uric acid metabolism in response to dietary protein levels.

The results of this thesis, and other related works, have been published and presented in scientific meetings.

Refeered journal papers

- López-Sánchez MJ, Neef A, Patiño-Navarrete R, Navarro L, Jimenéz R, Latorre A, Moya A. 2008. Blattabacteria, the endosymbionts of cockroaches, have small genome sizes and high genome copy numbers. *Environmental Microbiology* 10: 3417-3422.
- López-Sánchez MJ, Neef A, Peretó J, Patiño-Navarrete R, Pignatelli R, Latorre A, Moya A. 2009. Evolutionary convergence and nitrogen metabolism in *Blattabacterium* strain Bge, primary endosymbiont of the cockroach *Blattella germanica*. PLoS Genetics. 5:e1000721.
- González-Domenech C, Belda E, Patiño-Navarrete R, Moya, A, Peretó J, Latorre A, 2012. Metabolic stasis in an ancient symbiosis: genome scale metabolic networks from two *Blattabacterium cuenoti* strains, primary endosymbionts of cockroaches. BMC Microbiology 12 (Suppl 1):55.
- Patiño-Navarrete R, Moya A, Latorre A, Pereto, J. 2013. Comparative genomics of *Blattabacterium cuenoti*: the frozen legacy of an ancient endosymbiont genome. Genome Biology and Evolution 5:351-361.

- Patiño-Navarrete R, Piulachs MD, Moya A, Bellés X, Latorre A, Peretó J. Uric acid degrading enzymes genes response to different dietary protein levels in the fat body of *Blattella germanica*. In preparation.

Communications and conference papers

- Patiño-Navarrete R, Peretó J, López-Sánchez MJ, Neff A, Moya A, Latorre A. (2009) Secuenciación y análisis evolutivo de *Blattabacterium cuenotii*, endosimbionte primario de *Blatta orientalis*. Segundo Congreso de la Sociedad Española de Biologia Evolutiva.València. (Poster).
- Patiño-Navarrete R, Peretó J, López-Sánchez MJ, Neef A, Moya A, Latorre A. (2010). Genome sequence and evolutionary analysis of *Blattabacterium cuenotii*, endosymbiont of the cockroach *Blatta orientalis*. Workshop of Genomics and Metagenomics, COST action FAO0701 "Arthropod symbioses: from fundamental studies to pest ans disease management". Funchal, Portugal. (Poster).
- Patiño-Navarrete R, Peretó J, Latorre A. (2010). Comparison of the strains of the cockroach endosymbiotic bacteria *Blattabacterium cuenoti*, from the cockroaches *Blattella germanica*, *Periplaneta americana* and *Blatta orientalis*. GDRE of Comparative Genomics meeting. Barcelona. (Oral Communication)
- Patiño-Navarrete R, Peretó J, Piulachs MD, Bellés X, Moya A, Latorre A. (2011). Análisis del transcriptoma de tres tejidos de *Blattella germanica*: Metabolismo compartido entre *B. germanica* y *Blattabacterium* sp. XXXVIII Congreso de la Sociedad Española de Genética. Murcia. (Poster)
- Patiño-Navarrete R, Ponce de León M, Montero F, Moya A, Peretó J, Latorre A. (2012). Chromosomal and metabolic stasis in

Blattabacterium cuenoti, the ancient primary ensdosymbiont of cockroaches. Global Questions on Advanced Biology, Societat Catalana de Biologia. Barcelona. (Poster)

III Material and Methods

1.- Insects

Two species of cockroaches had been used; *B. orientalis* (Blattaria: Blattidae), to sequence its *Blattabacterium* endosymbiont, and *B. germanica* (Blattaria: Blattellidae) for transcriptomic and physiological studies.

1.1 *Blatta orientalis*

The strain of *B. orientalis* used during the present work was originated from a population sample collected in 2005 at the locality of Tarazona de la Mancha (Spain) and maintained in the laboratory at 26 ºC and 70% of relative humidity at the insect chambers of the "Institut Cavanilles de Biodiversitat i Biologia Evolutiva". The animals are fed with commercial dog food and water is supplied *ad libitum*.

1.2 *Blattella germanica*

The *B. germanica* individuals used for the whole transcriptome analysis were obtained from the population reared at the "Institut de Biologia Evolutiva (CSIC-UPF)" (IBE), Barcelona at 30 ºC and 70% of relative humidity and fed with commercial dog food.

The *B. germanica* individuals used to quantify the expression of genes involved in nitrogen metabolism under different dietary protein levels were also maintained at the facilities of our center under the conditions previously described for *B. orientalis*. This population was founded from a sample of the above-mentioned population from the IBE.

The individuals used in the gene expression experiments were reared in the abovementioned conditions during their whole life cycle until they moult into the adult form. As soon as they moult in the adult form, they were separated from the colony and reared with commercial dog food for two days; thereafter the animals were put on the experimental diets for two more days, until they are four old days' adults. Some animals were maintained on the common diet and were used as controls. The experimental diets are based on those described by Mullins and Cochran (1974). Diets are described in Table 4.

Table 4. Composition and protein content of the diets administered to the cockroaches.

	Dextrine	**Low protein**	**High protein**	**Control**
Protein content	**0%**	**5%**	**50%**	**~25%**
Composition	4g WSM[a]	4g WSM[a]	4g WSM[a]	Commercial dog food
	96g dextrin	10g YE[b]	10g YE[b]	
	20g cellulose	106g cellulose	45g caseine	
			41g Dextrine	

[a]Wesson salt mixture (Wesson 1932).
[b]Yeast Extract (Scharlau).

2.- Nucleic acids isolation and purification

2.1 Isolation and purification of genomic DNA from *Blattabacterium* strain BBor

2.1.1 Enrichment of *Blattabacterium* from fat body samples of *B. orientalis*

Adult *B. orientalis* females were sacrified and dissected while immersed in buffer A (35 mM Tris-HCL; 250 mM sucrose; 100 mM EDTA; 25 mM KCl; pH 7.5). Fat body, which is easily recognizable because its characteristic white color (Figure 8), is collected with the aid of entomological forceps and stored at 4 ºC.

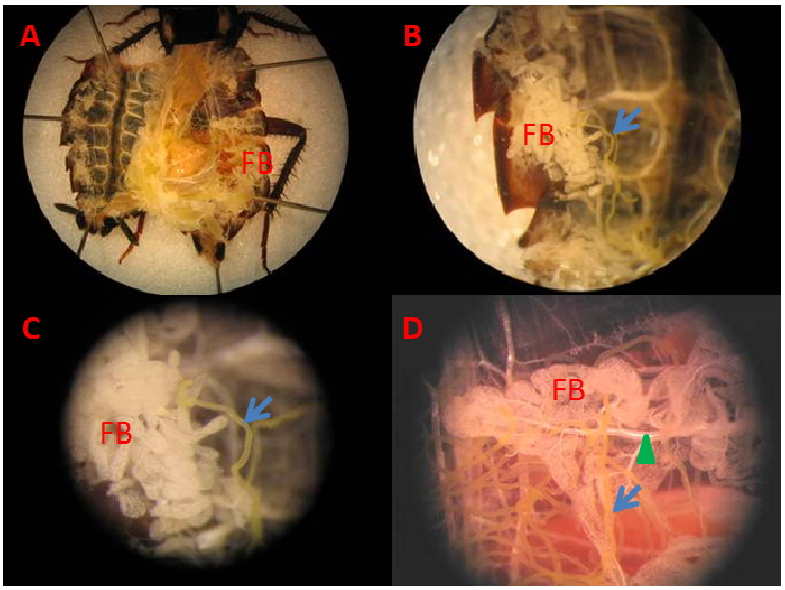

Figure 8. Internal anatomies of the cockroach *B. orientalis*. **A**, Panoramic view of a dorsally dissected animal. **B**, Detail of a fat body lobe. **C**, Zoom of the image B. **D**, Fat body enclosing trachea. FB, fat body; blue arrow, malpighian tubule; green arrowhead, trachea.

The sample was enriched for *Blattabacterium* following the method of Harrison (Harrison et al. 1989) with modifications as described in López-Sánchez et al. (2009). The whole procedure was carried out on ice to reduce as much as possible the histolysis and the degradation of genomic DNA from the symbionts. The fat body was homogenized in pre-cooled buffer A, and sieved in a series of filters with a decreasing pore size (80 µm, 20 µm and 11µm; Millipore) up to a final volume of 50 ml. The homogenate was later on centrifuged at 4000*g* for 15 min at 4 ºC, by this procedure the fat, which remained in the top of the Falcon, was separated from the other biological material, including the *Blattabacterium* cells, which were concentrated on the bottom. The *Blattabacterium* enriched pellet was washed twice with PBS (Phosphate buffer saline: 137 mM NaCl; 2.7 mM KCl; 10 mM Na_2HPO_4; 2 mM KH_2PO_4; pH 7.4), and finally resuspended in 1 ml of DNase I buffer (10 mM Tris-HCl pH 7.6; 2.5 mM $MgCl_2$; 0.5 mM $CaCl_2$). The sample was digested with DNase I (1mg/ml, Roche) for 10 min at room temperature. In this way, genomic DNA from the insect is degraded by the action of the DNase, whereas genomic DNA from *Blattabacterium* is protected by the bacterial cell wall. To inhibit the DNase, 500 µl of 0.5 M EDTA were added. Finally, the mixture is centrifuged (4000*g*, 15 min, 4 ºC), after removing the supernatant (containing the DNase); the pellet is washed twice with PBS for a proper elimination of the possible rests of DNase. Finally, the pellet was resuspended on TE buffer (10 mM Tris-HCl pH 7.5; 1 mM EDTA).

2.1.2 Extraction, purification and quantification of genomic DNA

Genomic DNA from *Blattabacterium* was extracted using the CTAB/Na (Cetyltrimethylammonium bromide) method (Ausubel, 1999; Murray and Thompson, 1980). This method was originally designed for the extraction and purification of good quality DNA from plants, since it is especially indicated for the elimination of polysaccharides and polyphenolic compounds that otherwise may alter the quality of DNA. These features make this protocol suitable to extract DNA from the fat body of cockroaches (López-Sánchez et al. 2009; Sabree et al. 2012). The protocol includes a phenolization step that extracts those contaminants that do no precipitate during the CTAB incubation like proteins, polysaccharides and other cellular components. The resultant DNA pellet is finally resuspended in LTE buffer (1 mM Tris-HCl pH 7.5; 0.1 mM EDTA) with RNase (20 µg/ml).

DNA was quantified spectrophotometrically with the system Nanodrop® ND-1000. To assess the purity of DNA, the 260/280 and 260/230 ratios are calculated, both of them must be between 1.8 and 2.0.

2.2 Total RNA extraction from several tissues of *B. germanica*

2.2.1 Insect dissection and storage

Insects were anesthetized with CO_2 and kept on ice until the dissection. The insects were fixed with entomological needles over its dorsal side, immersed in Krebs-Ringer bicarbonate buffer (Sigma-Aldrich) and opened ventrally in a way that the organs of interest remain exposed (ventral fat body lobes, ovaries (Figure 9) and the epithelium beneath the pronotum).

Tissues dissected were immediately frozen with liquid nitrogen and stored at -80 ºC until the RNA extraction.

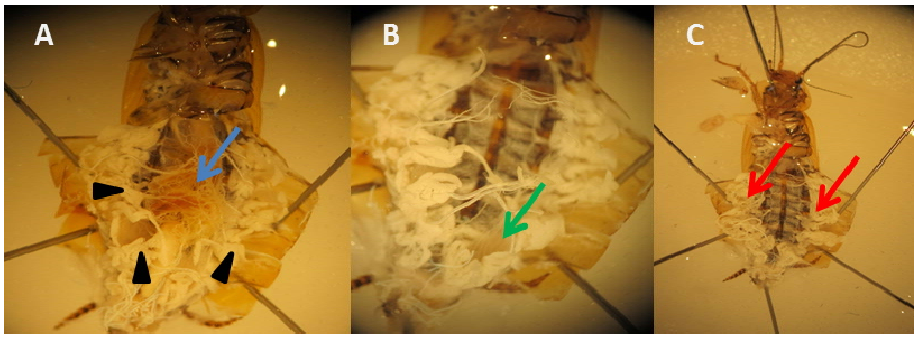

Figure 9. Dissection process for *B. germanica*. **A**. The animal was fixed with the ventral part up;after opening the abdomen the gut appears well defined (blue arrow); fat body lobes were in the laterals (black arrowheads). **B**. Once removed the gut, the ovaries become clearly visible (green arrow). **C**. After removing the ovaries the fat body (red arrows) was swiped.

2.2.2 Total RNA extraction and quantification

All the material and surfaces used during the RNA extraction process were chemically treated with RNase Zap Wipes (Ambion), or by heat 12 hours at 180 ºC, for inactivating the RNases. Total RNA was extracted with the GenElute™ Mammalian Total RNA Miniprep Kit (Sigma-Aldrich). This system avoids the use of organic compounds such as phenols or chloroform; instead, RNA binds to a silica membrane. For the RNase inactivation, in addition to guanidine thyocianate, uses a reducing agent like β-mercaptoethanol. After this protocol, the RNA integrity was evaluated by agarose gel electrophoresis. Finally, the quantity as well as the quality was measured with the spectrophotometer (Nanodrop® ND-1000; Nanodrop Technologies, Inc), in a similar way that described in the previous section for the DNA.

3.- Nucleic acid sequencing, assembly and quantification of selected genes

3.1 Genome sequencing of *Blattabacterium* strain BBor

3.1.1 Sanger approach

Shotgun libraries were constructed after random mechanical shearing of the genomic DNA with ultrasounds. Different sized DNA fragments were separated by pulse field gel electrophoresis (PFGE) and those fragments corresponding to the ranges 5-2 kb, 2-1 kb, and < 1kb were selected to construct the genomic libraries. The DNA was purified from the agarose slides with the system Ultrafree® (Millipore), and cleaned by phenolization.

DNA was prepared for ligation with the Single dA™ Kit (Novagen), which blunts DNA ends and adds a single deoxyadenine nucleotide on the 3' end, rendering the DNA fragment suitable to be cloned. Once the extremes were prepared, DNA fragments were ligated into the plasmid vector pCR®-XL-TOPO® (Invitrogen). Afterwards, these plasmids were used to transform *Escherichia coli* cells by electroporation applying 1.8 kV with the Electroporator 2510 (Eppendorf). Resistance genes to kanamycin and ampicilin harbored in the vector were used to select those cells that had incorporated the vector. The discrimination of those cells that incorporate non-recombinant vectors was possible because the insert interrupts the gene *lacZ*-α, making that colonies remain white in presence of X-Gal. Those transformed colonies with recombinant plasmids were grown (overnight at 37 ºC in agitation) and subjected to a miniprep procedure with the system MultiPROBE II with reactives of Millipore. Finally, the plasmidic DNA was sequenced with the universal primers for the XL-Topo, pUC18b (5'-

GTAAAACGACGGCCAGT) and pUC18g (5'-CAGGAAACAGCTATGAC), with the kit BigDye (Applied Biosystems). Sequencing was carried out in the sequencing facility of the SCSIE (Servei Central de Suport a la Investigació Experimental) at the Universitat de València.

3.1.2 Pyrosequencing

The emergence of the so-called next-generation sequencing technologies (NGS) has improved the ability to sequence new genomes. To complete the genome of the *Blattabacterium* strain BBor the system Genome Sequencer FLX (454 Life Sciences, Roche), based on the pyrosequencing technology (Ronaghi et al. 1998) was used. A total of 5 µg of DNA was sent to Life Sequencing (València) where they were processed and sequenced as described in Margulies et al. (2005).

3.1.3 Assembly

All the sequences obtained with the classical Sanger method, were edited with the program Trev 1.9 included in the software Staden Package, to trim out vector or bad quality sequences (Staden et al. 2000). Then, the reads were analyzed by homology search against a protein database with the BLASTX algorithm of the BLAST (Altschul et al. 1997). All sequences which matched with eukaryotic organisms or non-Bacteriodete bacteria were discarded. After the trimming process, selected sequences were assembled with Gap 4 (also included in the Staden Package), creating a database. The assembly of the sequences coming from the Genome Sequencer FLX was done with the Newbler software (Roche). Since the output format from

Newbler is not readable by Staden Package, it was necessary to convert it into a readable format with CAF2GAP. Both databases, the one with the Sanger sequences and the second with 454 sequences, were joined with the tool copy_db implemented in the Staden Package. The final assembly was performed manually on the Staden Package. The different contigs were properly orientated taking as reference the gene order on the other *Blattabacterium* strain genomes, given that in endosymbiotic bacteria the stasis is highly maintained along its evolution (Tamas et al. 2002; Latorre et al. 2005). The comparison of the contigs obtained for the endosymbiont of *B. orientalis* with the strains symbiotic from *P. americana* and *B. germanica* was done with the Artemis Comparative Tool (Carver et al. 2005).

3.2 Sequencing of RNA samples from different tissues of *B. germanica*

Whole transcriptome from three different tissues were analyzed with the aim of getting a deeper understanding about the relationships at gene expression level between *Blattabacterium* and its host *B. germanica*. Two tissues, fat body and ovaries harbor *Blattabacterium,* while the third one, the epithelium, was chosen as a control tissue, given that do not harbor *Blattabacterium*. Between the two tissues where *Blattabacterium* dwells, the main interest was to describe the differences of how *B. germanica* interacts whit its symbionts in the fat body, where they carry out the metabolic functions and in the ovaries where the symbionts are only to infect the progeny and thus few transcriptional activity is expected. Fat body and ovaries of adult females from 3 to 5 days old were dissected, while the control tissue was obtained from the epithelial tissue beneath the pronotum of stage 6 nymphs from days 5, 6 and 7. RNA was sequenced in GATC-Biotech (Konstanz, Germany), thereby 9.59, 9.63 and 9.88 µg of RNA from fat body, ovaries and epithelium respectively were sent to sequence.

Complementary DNA (cDNA) was synthetized using the Smart cDNA Construction Kit (Clontech). For the synthesis of the first strand an oligo (dT) was used obtaining by this way only mRNA from the host followed by a cap-primed second strand synthesis. Given that there were cDNAs coming from three different tissues it was necessary to mark each DNA with a barcode depending on its origin, the barcodes are defined in Table 5.

The cDNA was sequenced using the same procedure described in the section 3.1.2, but in this case the Titanium chemistry was used, which allows obtaining longer reads. As a result of the sequencing process, a sff file with the raw sequences was obtained, which was afterwards trimmed for bad quality regions, giving as output a Fasta and Fasta Quality files. Trimmed sequences were cleaned of endosymbiotic reads by mapping all the sequences against the genome sequence from the *Blattabacterium* strain BBge using the program MEGABLAST (Altschul et al. 1997) with an e-value cut-off of 10^{-5} and an identity percentage of 95. The sequences identified as belonging to the endosymbiont were removed from the Fasta and Fasta Quality files using a customized perl script. Once the sequences from the endosymbiont have been removed, the reads are assembled with MIRA 3.2 (Chevreux et al. 1999) in the EST working mode.

Table 5. Barcodes assigned to each tissue.

Tissue	Barcode
Fat body	AATGT
Ovaries	CTGCT
Epithelium	ACTCT

3.3 Quantitative real time PCR.

After two days on the experimental diets described in Table 4, *B. germanica* females were dissected. Then, total RNA was extracted from fat body and ovaries of these females. A total amount of 400 ng of this RNA in a final concentration of 50 ng/µl was used for synthetize the cDNA. Even when the GenElute^{TM} kit used during RNA extraction shears and remove most DNA contamination, prior cDNa synthesis the sample was treated for 30 min at 37 ºC with 1 U of DNaseI (Promega). The Transcriptor First Strand cDNA Synthesis Kit (Roche) was selected for cDNA synthesis. The primming method used in this case was the random hexamers that allows the synthesis of cDNA from bacterial mRNA. All the procedure was performed according the manufacturer instructions, but an optional step of denaturation (10 min at 65 ºC) to the RNA-random hexamers mixture was applied to denature secondary structures in RNA that may impair the union of the hexamers to the template. The synthesis reaction was done with 10 U of reverse transcriptase, and an incubation of 10 min at 25 ºC for extending the primers, 30 min at 55 ºC for the extension and finally 5 min at 85 ºC for inactivating the enzymes. In all procedures, one sample was used as a negative control (known as "RT-"), to prove the inexistence of genomic DNA, those controls were treated equally than the normal samples, but no reverse transcriptase was used. The result of the synthesis was tested by amplifying a 300 bp fragment of the actin gene with the primers 95 and 96 (Table 6), with a profile of 35 cycles of 95 ºC for 30 s, 58 ºC for 30 s and 72 ºC for 30 s. The response of genes involved in uric acid metabolism as well as the glutamine synthetase (Table 6) to the protein dietary levels was evaluated by real-time quantitative PCR by relative quantification method. Three individuals fed with each diet for all the analyses were used. The endogenous controls were the actin for those genes coded by the host and the

gene for the elongation factor EF-Tu (*tuf*) for the endosymbiotic urease gene. All qPCR were carried out in a Lightcycler 2.0 (Roche) with the kit FluoCycle II™ SYBR® (Euroclone). The genes measured and the primers selected for each gene are described in Table 6.

Each quantitative PCR was run with 1 U of Taq polymerase and 250 µM of each primer. The temperature profile was 95 ºC for 10 min followed by 40 cycles of 95 ºC for 5 s, 55 ºC for 8 s, and 72 ºC for 20 s. Finally a melting curve analysis was performed to ensure that there is only one product amplified. Each reaction was run per triplicate, and non-template negative controls were used.

Statistical analysis were done with the REST package tool (Pfaffl et al. 2002). This software estimates the changes of gene expression between two groups (in our case a control group, formed by those animals fed with dog food and the experimental comprising the animals fed with one of the artificial diets), taking the distribution of the threshold values (C_T) as input without any assumption about the distribution. It uses a pair wise fixed reallocation test to evaluate the significance of the results (Pfaffl et al. 2002). Results are given graphically as copies of RNA per 1000 copies of the reference gene, based on the method proposed by Schmittgen and Livak (2008).

Table 6. Genes and primers used in protein level experiments.

Gene	Primer	Sequence
Actin	95	5'-TCGTTCGTGACATCAAGGAGAAGC-3'
	96	5'-TGTCGGCAATTCCAGGGTACATGGT-3'
	512	5'-AGCTTCCTGATCGTCAGGTGA-3'
Urate oxidase	URTOf	5'-AAATCATTGGTGGGCTTCGTGGTC-3'
	URTOr	5'-TCACGTCTGGCAACGTTCTGTACT-3'
Allantoinase	ALNf	5'-TCTGACAGCAGAAACCTGTCACCA-3'
	ALNr	5'-GCTGCCCAAAGACGTTCCTTGTTT-3'
Allantoicase	ALTf	5'-GGAATTATGCACCTCGCTTCTCTC-3'
	ALTr	5'-CACTCCCTATTCTACTGTTCCGAC-3'
Glutamine synthetase	BgGSaf	5'-TACAAAGATCCATTCAGGCCA-3'
	BgGSar	5'-CACGTATGCCTTTGATTTGTGG-3'
Elongation factor EF-Tu	TU1f	5'-AAGGAAGAAGGAGGACGACACACT-3'
	TU1r	5'-TAGGCTGATGCAATTCCACCTCCA-3'
Urease	UC1f	5'-GTCCAGCAACTGGAACTATAGCCA-3'
	UC1r	5'-CCTCCTGCACCTGCTTCTATTTGT-3'

4.- Annotation and comparative analyses

4.1 Annotation and functional analysis of *Blattabacterium* strain BBor

The annotation process consisted on the identification and localization of genes, pseudogenes and intergenic regions (IGS) in the genome sequence. First of all, the putative Open Reading Frames (ORF) in the genome were predicted with the software GLIMMER v.3.02 (Delcher et al. 2007). Since this program needs to be trained, the complete set of nucleotide sequences for all protein coding genes from the *Blattabacterium* strains of *B.*

germanica and *P. americana,* were used as a training set. The minimum gene length was fixed in 100 bp. Once the open reading frames (ORF) of all probable protein coding genes were identified, their coordinates were charged on the genome sequence using the genome browser ARTEMIS (Rutherford et al. 2000), which allows us to manually identify and curate start and stop codons as well as annotate probable ORF not detected by GLIMMER. The protein sequences of all ORF identified was written in a multifasta file. Once the ORF were identified, it was necessary to assign a function to each gene. First, orthologous genes in the previously sequenced *Blattabacterium* strains as well as with the free living Bacteriodete *F. psychrophilum* were identified using the OrthoMCL algorithm (Chen et al. 2006). Then, the protein coding genes of all bacteria were compared all-against-all using BLASTP, minimum E-value was established at 1e-05, a per cent match cutoff of 70 and inflation value of 1.5. As a result, it was obtained a table were each gene was classified in groups of orthology. BLAST searches were also performed against the gene non-redundant KEGG database (www.genome.jp) with a cut-off e-value of 10^{-3} and an identity value of 50. Finally, searches with BLASTN and BLASTX were performed on the IGS to identify possible genes or, most probably, pseudogenes overlooked by GLIMMER.

Protein coding genes were classified according to the functional categories described in Cluster of Orthologous Groups of proteins (COG) (Tatusov et al. 2003). The COG database was downloaded from NCBI, and the proteomes of the six strains of *Blattabacterium* were compared with this database whit BLASTP searches (e-value 0.0001). Genes were also assigned, when it was possible, to KEGG Orthology (KO) numbers (www.genomenet.jp). Finally, EC numbers according to Brenda enzymes database (http://www.brenda-enzymes.info/) were assigned to each enzyme coding gene. Once identified all genes it became evident that, as in the

other*Blattabacterium* strains, there are neither dnaA boxes nor *dnaA* gene, so it was not possible to locate the origin of replication. In that case, the method used to predict the origin was through the GC skew using the software OriginX (Worning et al. 2006).

RNA genes were identified using INFERNAL (Nawrocki et al. 2009), an algorithm that searches structures and sequences against an RNA database. By this way ribosomal RNAs, ncRNAs and the tRNAs were detected. For the last ones an additional search with the tRNA-Scan (Lowe and Eddy 1997) was performed. All RNA genes have been also confirmed through searches on the Rfam database (Gardner et al. 2011).

Genes were named respecting, when possible, the name assigned in the previously sequenced strains. In case of conflict, the name assigned in *E. coli* was chosen. In addition, each gene was identified with a number, corresponding to its position respect to the origin of replication, preceded by the locus tag code BLBBOR.

Metabolic reconstruction was done with the KASS-KEGG server (Moriya et al. 2007, http://www.genome.jp/tools/kaas/), a web-tool able to generate maps with the metabolic paths of an organism from the amino acid sequences of their CDS. However, the metabolic map was also manually curated with the information obtained from the metabolic reconstructions of the strains BPam and BBge (González-Domenech et al. 2012) and the information found in the MetaCyc database (www.metacyc.org).

The GC content of the whole genome for all strains was calculated with the software GeeCee, included in the EMBOSS package (Rice et al. 2000). For the estimations of the average GC content in genes, IGR and CDS (in the whole gene and in the different codon position) a costumized perl script was used.

4.2 Comparative genomics and inferred metabolisms among the *Blattabacterium* sequenced strains

Graphic representation of genome-compared graphs between the six *Blattabacterium* strains were obtained with the genoPlotR package (Guy et al. 2010). Orthologous genes linked in the representations were obtained from the orthology table generated with orthoMCL (Annex 1).

For the comparative analyses among all six *Blattabacterium* strains, the first step was the construction of the pan-genome, thus it would be possible to determine which genes form the core (i.e those genes present in all six strains), and which genes form the dispensable genome. The graphic representation of an Euler diagram was done with the R package Vennerable (Swinton 2011).

The gene counts for each space were made from the gene orthology table obtained from the orthoMCL. The coverage of the genome was measured by means of rarefaction curves performed with the specaccum function of the Vegan library of R (Oksanen et al. 2012), which is used in ecological studies to calculate the species richness for a given number of samples. In our case, the species were the number of orthologous groups, while the number of samples was the number of sequenced genomes.

COG distribution profiles were compared among all *Blattabacterium* strains. To assess if there were different distributions, a χ^2 test was performed, using as a reference the COG profile of BBge, given that possess the most complete metabolism. With the heatmap.2 from the gplots library (Warnes 2011) a heat map was drawn using the COG distribution frequencies for all six *Blattabacterium* strains, with the pseudodata sets generated from the pan-genome and the core genome COG profiles, using the free living Bacteriodete *F. psychrophilum* as an outgroup. All strains

were clustered putting closer the most similar datasets; COG categories ordered in the rows were also clustered in a similar way.

To assess the effects of gene loss in the metabolism of BBor, a stoichiometric analysis was performed with METATOOL (Pfeiffer et al. 1999), centered on the reactions involving the urea and Krebs cycles.

4.3 Annotation and functional analysis of RNA samples

The annotation and functional analysis of the complete transcriptomes for the three chosen tissues of *B. germanica* was carried out with Blast2GO (www.blast2go.com) (Conesa et al. 2005). This program starts with BLAST searches to identify similar sequences to the input ones. In our case, the contigs obtained from the assembly process with MIRA have been used as input sequences. The nucleotide sequences were compared with BLASTX to the non-redundant protein database from the NCBI with an e-value cut-off fixed on 1e-06. The minimal length of the hit was established in 30 nucleotides, and low complexity regions were filtered. Once the BLASTX was finished, Gene Ontology terms (GO) were assigned to our sequences by searching the hits of each sequence on the mapping files provided by the NCBI, gene_info (ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/) and gene2accesion (www.ncbi.nlm.nih.gov/entrez/query/static/help/LL2G.htmlfiles), as well as in the GO database directly (www.geneontology.org). Finally, an EC number was assigned to the enzymes. Blast2GO also offers the possibility of generate the KEGG metabolic maps. In some occasions, these pathways were interrupted, and manual searches on the transcriptome were done in order to close those gaps in the pathways, always with the help of the MetaCyc database (www.metacyc.org).

4.4 Comparative analysis of the transcriptome

Blast2GO has integrated the Gossip (Blütghen et al. 2005) that allows us to perform the Fisher exact test, which is indicated to assess the differences between the annotation of two sets of sequences with respect to the distribution of GO terms. The data were corrected for multiple testing, and the two-tailed option was used.

5.- Evolutionary analysis

5.1 dN/dS test

The synonymous (dS) and non-synonymous (dN) substitution rate were calculated with the program yn00 included in the software package PAML version 4.6 (Yang 2007) using the approximated method proposed by Yang and Nielsen (2000), which takes into account the transition/transversion rate bias and the base frequency, the two most important features in the mutation dynamics, since it has been shown that adding further complications do not improve the results (Yang 1994). Once estimated these two rates, the value of the ratio among them (dN/dS, ω) will indicate the evolutionary force operating in each gene, positive when $\omega > 1$, neutral when $\omega = 1$ and purifying when $\omega < 1$. Previous to running YN00, each pair of proteins were aligned with MAFFT (Katoh et al. 2005) and a nucleotide alignment was obtained using the amino acid alignment as a template with Tranalign (Rice et al. 2000).

5.2 Phylogenetic analysis

Protein sequences for all genes present in the core genome with an orthologous gene in *F. psycrhophilum* were aligned with MAFFT, using accurate oriented method L-INS-i (Katoh et al. 2005). These alignments were posteriorly concatenated with a customized perl script catfasta2phyml.pl (http://www.abc.se/~nylander/). Thereafter, the best-fit models of amino acid replacement was estimated with ProtTest 3 with the AIC information criteria (Darriba et al. 2011). The maximum likelihood (ML) best tree was calculated with 100 bootstrap replicates with RAxML using the PROTGAMMA algorithm (Stamatakis et al. 2005).

5.3 Testing the molecular clock hypothesis

Nucleotide alignments of all genes present in the core genome of *Blattabacterium* were obtained with the program Tranalign from the EMBOSS package (Rice et al. 2000) using protein alignments obtained with MAFFT (Katoh et al. 2005) as a template. To minimize the problem of saturation of nucleotide site, third codon positions were removed from the analysis. Maximum likelihood models are known to be robust to violation of the model, including divergence times and saturation. Moreover, first and second positions account for non-synonymous sites, which are often subject to selection and are less prone to saturation. The best fit evolutionary model for each alignment was selected with the jModelTest 2.1, with three substitution schemes (Darriba et al. 2012). A likelihood ratio test (LRT) was performed running the Baseml program in the PAML package (Yang 2007), with the selected evolutionary models. Using the tree topology estimated by

ML in the previous section, but without *F. psychrophilum*, each gene was tested under two models, one assuming homogenous rates (rooted model where *n-1* branch lengths are estimated) were all branches evolve at the same rate, and the other which allows each branch to evolve under different rates (unrooted model, *2n-3* branch lengths are estimated). The LRT compares the differences in the log likelihood (*l*) values for each model ($2\Delta l$) with a χ^2 distribution with *n-2* degrees of freedom, where *n* was the number of *Blattabacterium* strains. Those genes that (i) do not reject the null hypothesis of homogeneous rates, (ii) possess an orthologous gene in the *F. psychrophilum* genome, (iii) do not accumulate more than 2.5 substitutions per site, and (iv) continue to accept the molecular clock after the addition of the free living Bacteroidete, where used to determine the date of the divergence between pairs of strains (i.e., BBor/BPam, BMda/BCpu and BBge/BBgi). We used as a calibration point the divergence between the Blaberoidea (BBge and BBgi) and the Blattoidea (BBor, BPam, BCpu and BMda) superfamilies of cockroaches, estimated by fossil data in 140 Mya (Sabree et al. 2010; Vrsansky et al. 2002). Nucleotide alignments were obtained following the above mentioned procedure, also removing third codon positions.

IV Results

1.- Genome sequencing and analysis of *Blattabacterium* from *Blatta orientalis* (BBor).

1.1 Genome characteristics

The genome of *Blattabacterium* strain BBor has a total size of 638,184 bp and is composed by a circular chromosome of 634,449 bp and a plasmid of 3,735 bp, with a GC content of 28.2 and 30.6%, respectively.

As pointed out in Material and Methods, the chromosome was sequenced by a hybrid approach, using the Sanger method and the Genome Sequencer FLX system of Roche (454). Using the Sanger approach, 727 sequences belonging to *Blattabacterium* were obtained, which were assembled in 351 contigs spanning 138 kb of the genome, with a low coverage.

The 133,562 good quality reads obtained with the 454 technology were assembled with the Newbler software obtaining 1,201 contigs with an average length of 846 nucleotides and coverage of 30X. Nonetheless, 97% of these reads (129,647 readings) were located into the 39 bigger contigs (those ≥500 bp), being twelve of them bigger than 5,000 kb. Both databases were combined using the copy_db tool implemented in the Staden Package, obtaining a hybrid database with contigs coming from both technologies. These contigs were assembled and finally, produced two circular contigs, one of 634,449 bp and 125,049 reads for the chromosome and other one of 3,735 bp and 2,275 reads for the plasmid. The final average consensus coverage was 42X with 41.5X for the chromosome and 118X for the plasmid, when taken separately.

The number of annotated genes as well as IGS present in the genome of *Blattabacterium* BBor is summarized in Table 7. In this genome there are several overlapping regions between different genes that have been considered as a category of IGS.

Table 7. Number of genes and intergenic regions in the genome of *Blattabacterium* strain BBor. Gene numbers in the plasmid are in parentheses.

1. Genes	627 (7)
1.1 CDS	579 (7)
1.2 RNAs	39
1.2.1 tRNA	33
1.2.2 rRNA	3
1.2.3 Other ncRNAs	3
1.3 Pseudogenes	9
2. Intergenic regions	628 (8)
2.1 non-overlapping	504 (7)
2.2 overlapping	124 (1)

The GC content was calculated for the whole genome as well as for the genes taken individually. Additionally, in protein coding genes, GC content has been measured separately for the 1^{st}+2^{nd} and for the 3^{rd} codon position. Finally, it was also calculated for the intergenic non-coding regions (Table 8). The detailed results for each gene can be consulted in the Annex 2.

Comparing the six sequenced strains, only BCpu shows lower values in GC content. The values for the other five *Blattabacterium* strains are quite similar, especially the symbionts from *P. americana* and *B. orientalis*, where the values are almost equal in several categories.

As usual in endosymbiotic bacteria, it was identified only one copy of each ribosomal gene. In BBor, as in the other *Blattabacterium*, all three ribosomal genes (23S, 16S and 5S) form an operon. There were identified 33 tRNA, with the anti-codon sequence for all 20 amino acids, and three non-coding RNA genes: tmRNA, RNA of the signal recognition particle, and the RNA fraction of the ribonuclease-P.

Table 8. Summary of GC content in the genome for all *Blattabacterium* strains analyzed.

G+C (%)	BBor	BPam	BCpu	BMda	BBgi	BBge
1. Genome	28.2	28.2	23.9	27.5	26.0	27.2
1.1 Chromosome	28.2	28.2	23.8	27.5	25.7	27.1
1.2 Plasmid	30.6	28.5	30.4	31.9	30.9	29.9
2. Genes	28.6	28.6	24.6	27.8	27.1	27.5
2.1 CDS	28.4	28.4	24.2	27.6	25.6	27.2
2.1.1 1^{st}+2^{nd} position	33.7	33.4	30.8	33.1	31.5	33.0
2.1.2 3^{rd} position	17.7	17.6	11.2	16.6	13.7	15.7
2.2 Pseudogenes	24.8	28.0	19.5	25.3	23.0	19.3
2.3 RNA genes	47.0	47.1	43.1	46.7	50.3	46.8
3. Intergenic region	20.7	20.1	14.6	18.8	15.7	18.9

Nine genes were finally annotated as pseudogenes. Two of them, *cysH* and *cysI,* encoding for products involved in the sulfate assimilatory pathway, 3'-phosphoadenosine 5'-phosphosulfate sulfotransferase and the hemoprotein subunit of the sulfite reductase, respectively. Three are involved in the synthesis of heme groups, *hemC*, *hemD* and *cysG,* coding for hydroxymethylbilane synthase, uroporphyrinogen III synthase and siroheme synthase, respectively, being the last one indirectly related to the sulfate assimilatory pathway, since the siroheme is the heme group present in sulfite reductase (Murphy and Siegel 1973). The other pseudogenes were *dut*, coding for deoxyuridine triphosphatase, which participates in the synthesis of timidine nucleotides; a thermonuclease family protein annotated as *lpxP* in BMda and BPam; an ATP-binding cassette transporter, with functional homologs in BCpu, BMda, BBgi and BBge clustered in the orthology group blb_595 (see Annex 1). The last one code for an hypothetical protein, with homologs in BPam, BCpu, BBge and BBgi, which are clustered in the orthology group blb_0578 (Annex 1).

Seven duplicated genes were found, those coding for 2-methylthioadenine synthetase (*miaB*), dihydrolipoyl dehydrogenase (*lpdA*), rod-shape determining protein RodA (*rodA*), phosphoserine transaminase (*serC*), peptidylprolyl isomerase (*ppiC*), acetylornithine transaminase (*argD*) and ATP-dependent DNA helicase (*uvrD*). Additionally, for two of the pseudogenes, *dut* and *hemD*, a functional copy was found, being remarkable that functional *dut* copy is the one located in the plasmid.

The calculus of gene length, summarized in Table 9, was performed at different levels.

Table 9. Total and mean length of the genome of *Blattabacterium* strain BBor.

	Length (nt)	**Mean (nt)**
1. Genome	638184	-
1.1 Chromosome	634449	-
1.2 Plasmid	3735	-
2. Genes	607213	968.4
2.1 CDS	594851	1027.4
2.2 RNA	7625	195.5
2.2.1 tRNA	2521	76.4
2.2.2 rRNA	4337	1445.7
2.2.3 Other ncRNAs	765	109.3
2.2 Pseudogenes	4746	527.3
3. Intergenic region	31635	64.6

Close to 95% of the genome encodes for any class of gene. This value is in concordance with the values observed for the other *Blattabacterium*, which ranged between 96.3% in BBge and 93.4% in BCpu. Due to the high density in gene content in this genome, the length of IGS is quite reduced, 41% of them have less than 20 nucleotides of length (Figure 10). Besides, in 124 cases the end of a gene and the beginning of next gene are overlapped (Table 7).

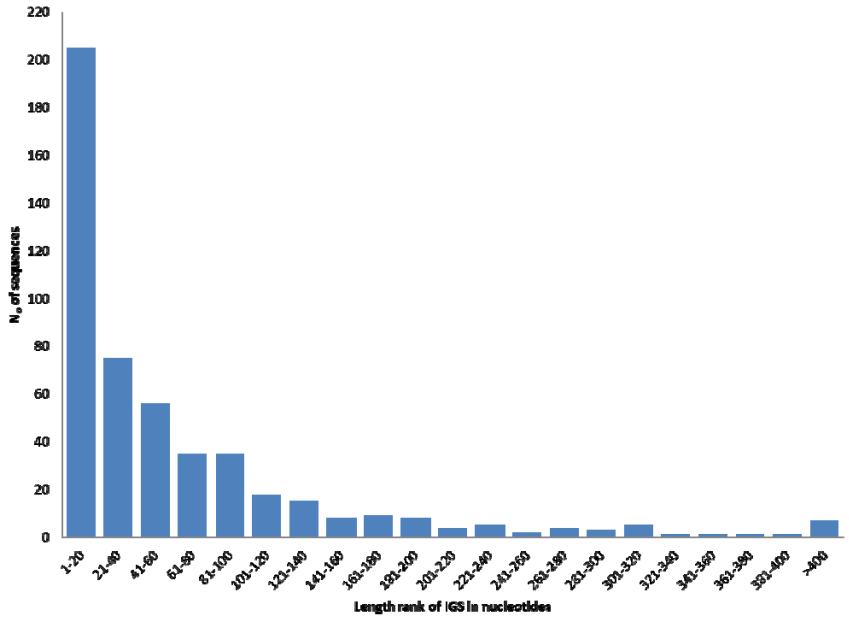

Figure 10. IGS length distribution in BBor genome.

Every coding gene was classified by its function according to the COG categories (Table 10). It was possible to assign a COG category to 511 of 579 coding genes; five of them were assigned to two different COG categories. All genes with two assigned COG functional categories code for proteins that possess two domains, each of them with a different orthologous gene in the COG database. Thus the carbohydrate kinase encoded by *yjeF* has been classified into the categories G and S, *lpxC* has homologs into the categories I and M, a *fhlA* into K and T categories, *topA* was classified in L and R, and finally *ccoN* was grouped into the categories C and O.

Table 10. COG classification

COG	Description
J	Translation
K	Transcription
L	Replication, recombination and repair
D	Cell cycle control
M	Cell/wall membrane biogenesis
N	Cell motility
O	Posttranslational modification protein turnover, chaperones
P	Inorganic ion transport and metabolism
T	Signal transduction mechanism
U	Intracellular trafficking and secretion
V	Defense mechanism
C	Energy production and conversion
E	Amino acid transport and metabolism
F	Nucleotide transport and metabolism
G	Carbohydrate transport and metabolism
H	Coenzyme transport and metabolism
I	Lipid transport and metabolism
Q	Secondary metabolites biosynthesis transport and catabolism
R	General function prediction only
S	Function unknown

The categories involved in metabolic processes (C, G, E, F, H, I and Q) account for 36.6% of the genes, 25.7% were placed into the informational processing and storage categories (J, K and L), 17.9% are part of the cellular processes categories (D, O, M, N, P, T, U and V), 8.8% were classified into the poorly characterized proteins (R and S), and finally for 68 genes it was not possible to assign any functional category. The most represented functional categories are J, E, C and H with 105, 68, 37 and 36 genes, respectively.

Genes also were classified according to the KEGG orthology system (KO) and 468 genes were assigned to their corresponding KO category. Finally, an EC number was assigned to 296 proteins encoded in this genome.

1.2 Comparative genomics of *Blattabacterium* strains

After the identification, annotation and classification of all genes present in the genome of *Blattabacterium* strain BBor, we performed a comparative genomic analysis of this strain with the five previously sequenced, those from the cockroaches *P. americana* (BPam), *B. germanica* (BBge), *B. gigantenus* (BBgi) and *C. punctulatus* (BCpu), and the strain from the termite *M. darwinensis* (BMda). The general genomic features for the six sequenced *Blattabacterium* strains are summarized in Table 11. Data for the previously sequenced strains were extracted from the GenBank files available for each genome (see Material and Methods). By this study we aimed to obtain a major comprehension of the evolutionary history of the symbiotic system established more than 140 Mya (Lo et al. 2003) in the common ancestor of cockroaches and termites. It was especially interesting the comparison of *Blattabacterium* strains from omnivore cockroaches (BBor, BPam, BBge and BBgi) with that of xylophagous (BCpu and BMda).

The genomes from BBor, BPam and BBge show a high degree of similarity in all analysed parameters. The genomes of BCpu and BMda are 5% and 8% smaller, respectively, when compared with the other strains. The average size of CDS and RNA genes varies scarcely among the different strains. The most remarkable difference lies on the longer IGS found in BCpu than in the other strains. However, this last feature has not a great influence in the genome structure, since IGSs represent a minor fraction within these genomes. Thus, the genomic reduction in BCpu and BMda is mainly due to gene loss events.

Table 11. General genomic features of the six sequenced *Blattabacterium* strains: BBor, *Blattabacterium* from *B. orientalis*; BPam, *Blattabacterium* from *P. americana*; BBge, *Blattabacterium* from *B. germanica*; BCpu, *Blattabacterium* from *C. punctulatus* and BMda for *Blattabacterium* from *M. darwiniensis*.

Strain	BBor	BPam	BCpu	BMda	BBge	BBgi
Genome size (bp)	638183	640442	609561	590554	640335	632588
Plasmids	1	1	1	1	1	1
Plasmid size (bp)	3735	3448	3816	3306	3485	3423
Chromosome size (bp)	634448	636994	605745	587248	636850	629165
GC content (%)	28.1	28.2	23.8	27.5	27.1	26.0
Total number of genes	618 (7)[a]	620 (4)[a]	586 (3)[a]	582 (4)[a]	630 (4)[a]	616 (4)[a]
CDSs	579	582	548	544	590	577
rRNAs	3	3	3	3	3	3
tRNAs	33	33	32	34	34	33
Other ncRNAs	3	3	3	3	3	3
Pseudogenes	9	6	3	9	1	1
CDS Coding region %	93.2	93.8	91.9	94.3	95.1	95
CDS average length	1027	1033	1023	1024	1032	1039
IGS average length	65	59	83	45	52	58

[a]In parentheses, number of genes coded in the plasmid.

The gene order in BBor was compared with the other sequenced strains. The synteny (i.e., gene order conservation) for those orthologous genes conserved among the different strains remains mainly unaltered (Figure 11). Actually, there are few re-arrangements; only three inversions were detected (in blue in Figure 11): one of ~242 kb in BMda, the breakpoints being localized first at the IGS between the genes *carB* and *argF*, and the second one at the IGS between the genes coding for cytochrome *c* and the probable peptidase M16. The second inversion is a fragment of ~19 kb in the Blattidae family symbionts (BBor and BPam), with the breakpoints localized at the IGS between the genes *mdh* and one coding for a thiamine pyrophosphokinase, the first and the second in the IGS of the genes *lnt* and *yjeF*. From these two previously described inversions, only the second one is

present in BBor. It has been identified a third inversion in BMda, with a size of 2.9 kb that contains the genes *rffH*, *dut* and *wzxC*.

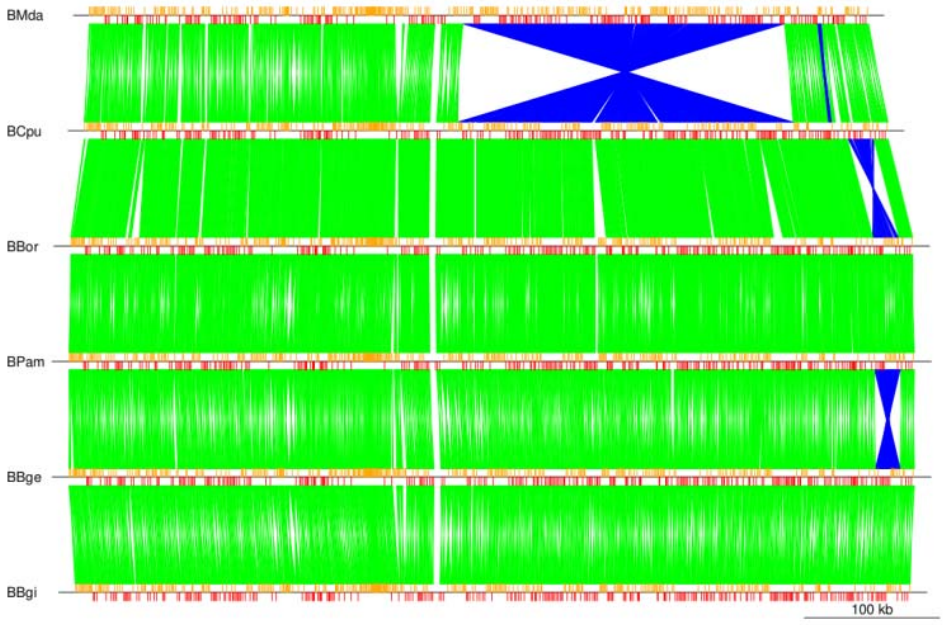

Figure 11. Whole-genome comparison between the six sequenced strains of *Blattabacterium*, yellow boxes represent the CDS coded in the leading strand, red boxes represent CDS coded in the lagging strand. Green lines connect orthologous genes with the same orientation. Blue lines connect orthologous genes with different orientation.

Despite that genomes from insect endosymbiotic bacteria are quite compact, nine duplicated genes were identified in BBge (López-Sánchez et al. 2009). All strains maintained at least one of the copies for all nine genes (Table 12) and only five of the nine genes maintain both copies in all strains, *rodA*, *uvrD*, *lpdA*, *miaB* and *argD*. The gene *hemD* is only duplicated in BBge, the other strains only keep one of the copies, but different paralogs are retained in different strains.

Table 12. Duplicated genes on *Blattabacterium* strain BBge, and their orthologs in other strains. It has been been taken as reference the position on the genome of BBge strain. +, presence; –, absence; ψ, pseudogene.

gene	Locus in BBge	BBor	BPam	BCpu	BMda	BBgi
rodA	BLBBGE_005	+	+	+	+	+
	BLBBGE_387	+	+	+	+	+
uvrD	BLBBGE_138	+	+	+	+	+
	BLBBGE_475	+	+	+	+	+
lpdA	BLBBGE_147	+	+	+	+	+
	BLBBGE_214	+	+	+	+	+
miaB	BLBBGE_207	+	+	+	+	+
	BLBBGE_350	+	+	+	+	+
argD	BLBBGE_315	+	+	+	+	+
	BLBBGE_623	+	+	+	+	+
ppiC	BLBBGE_093	+	+	+	+	+
	BLBBGE_620	+	+	+	-	+
serC	BLBBGE_144	+	+	-	+	+
	BLBBGE_428	+	+	+	+	+
dut	BLBBGE_606	ψ	+	+	+	+
	BLBBGE_p002	+	+	+	+	+
hemD	BLBBGE_290	+	+	+	-	+
	BLBBGE_576	ψ	ψ	-	+	-

1.3 Pan-genome reconstruction and functional profile

The full complement of genes within any bacterial species forms the pan-genome, which is composed by the core complement, which includes those genes present in all strains, and the dispensable or accessory genome, formed by those genes present in at least one, but not all strains of the species. A previous step to describe the pan-genome is the identification of all orthologous groups among the CDS coded in all strains of a species. In our case, from a data set of 3,417 CDS from all six *Blattabacterium* strains orthoMCL clustered 3,399 proteins (>99 % of total set of CDS among the six strains) in 599 clusters of proteins, leaving 18 proteins with any known orthologs in the other strains, which are the so-called strain-specific proteins. There were identified 502 groups containing sequences for all six strains,

which form the core complement of *Blattabacterium*, while the remaining 97 groups plus the 18 ungrouped proteins form the accessory genome, which is distributed as follows: 18 CDS are strain specific, 15 are shared by two strains, 8 by three strains, 35 by four strains and 39 by 5 strains.

It is worth to mention that two annotated CDS that were embedded in the 23S rRNA gene in BMda (Sabree et al. 2012) one of them also present in BBgi (Huang et al. 2012), namely MADAR_308 and his ortholog in BBgi (BGIGA_329) annotated as cell wall hydrolase, and MADAR_309 (hypothetical hydrolase), were removed from the accessory genome since they should be considered false positives as result of missannotations as shown by Tripp et al. (2011). Figure 12 shows, through an unweight Euler diagram, the localization of protein clusters in the different subspaces of the pan-genome.

Additionally, 37 RNA genes (31 tRNAs, 1 rRNA operon and three more ncRNA) should be added to the core genome. Three more tRNA genes are present in the accessory genome, one for proline (anti-codon GGG) lost in three independent events in BCpu, BPam and BBgi), other for arginine (anti-codon CCG), lost in BCpu and the last one for valine (anticodon TAC), lost in BBor.

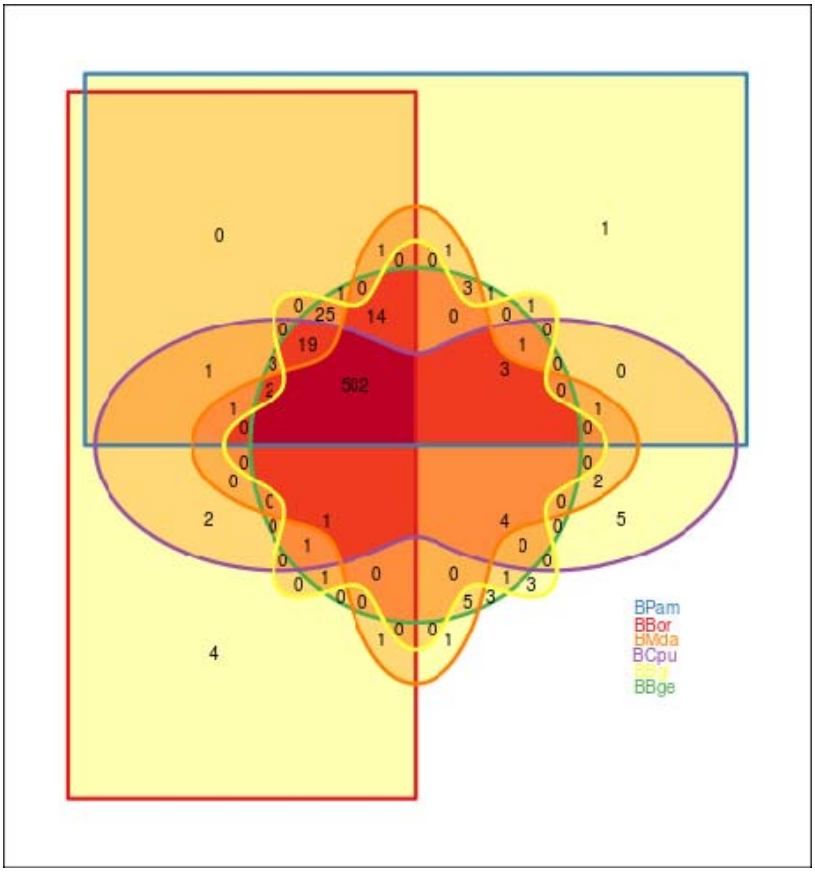

Figure 12. Non-proportional Euler diagram representing the pan-genome of *Blattabacterium* strains; numbers represent the total amount of clusters of orthologous proteins in each subspace.

In summary, the pan-genome of *Blattabacterium* would be composed by 655 genes (615 CDS, 34 tRNAs, 3rRNA and 3 ncRNA) while the core genome represents the 82.3 % of the pan-genome and comprises 539 genes (502 CDS, 31 tRNAs, 3 rRNA and 3 ncRNA). When BCpu and BMda, the two symbionts from xylophage hosts, were removed from the analysis, the differences between the pan-genome and the core (647 and 598 genes respectively) were reduced. In this case the core represented the 92.4 % of the pan-genome.

Finally, a rarefaction analysis was performed to assess the coverage of the *Blattabacterium* pan-genome (Figure 13). Rarefaction curves show when *Blattabacterium* pan-genome has reached a plateau, indicating that the addition of new strains would not increase the number of genes significantly (what it is referred as a closed pan-genome). These data are in concordance with other insect endosymbionts, like *B. aphidicola* or *S. muelleri* (this last not depicted in the graphic). By contrast free living bacteria possess an open pan-genome because they usually need of more genomes to reach a plateau, like in the case of *E. coli*, where the curve is still increasing despite eight genomes were added to the analysis. In facultative intracellular parasites, like in the case of the bacterial genus *Bartonella,* the pan-genome also remains open (Figure 13).

Genes for the six *Blattabacterium* strains were classified according COG functional categories. The total number of genes classified in the different COG categories among *Blattabacterium* strains show a similar functional profile (Figure 14) with the only exception of genes involved in the amino acid transport an biosynthesis (E category), which are reduced in BMda and BCpu. As usual, the most represented functional category in the endosymbiont genomes, are genes involved in translation (J) accounting for 20% of all genes, while the second most represented, especially on the omnivorous species, is the category of genes devoted to the amino acid transport and metabolism (E), and shows the main gap between xylophages species and the rest.

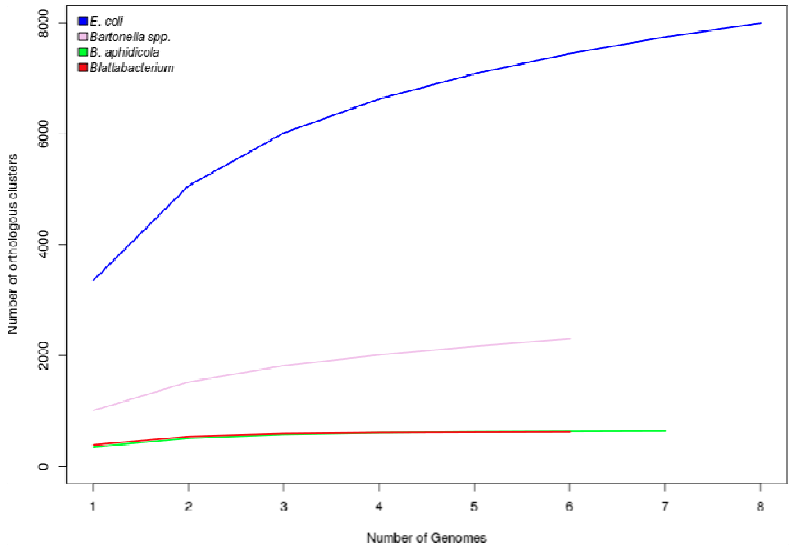

Figure 13. Rarefaction curves applied to different strains of *Blattabacterium* (6 genomes), *E. coli* (8 genomes), *Bartonella* (6 genomes) and *B. aphidicola* (7 genomes).

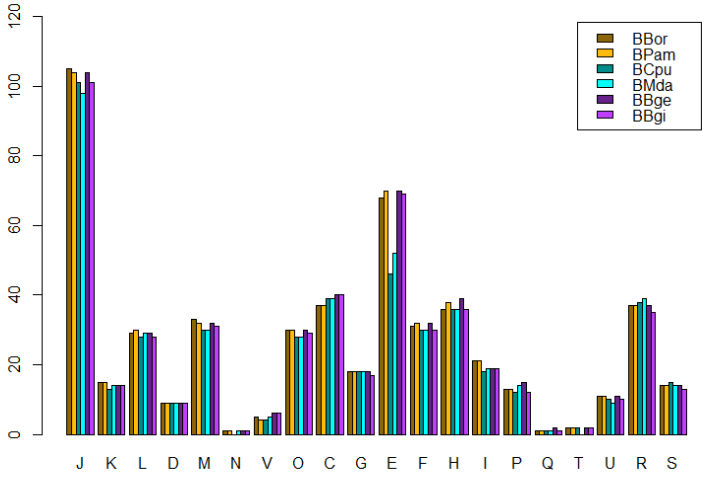

Figure 14. Absolute numbers of genes for each COG category in every *Blattabacterium* strain sequenced. COG categories as in Table 10.

A χ^2 test was applied to check if there are differences in COG distribution among the different strains of *Blattabacterium*. The observed distribution for COG categories in BBge was taken as a model for calculating the expected distribution in the other strains, since BBge is the strain with a most complete CDS dataset. Table 13 compiles the observed and expected values as well as the results of the analysis.

Table 13. Observed COG values in the six strains and values for the Chi square test performed over the COG distribution values in the different strains. For avoiding the multiple testing problems, the Bonferroni correction has been applied, so the significance level for each individual test was 0.05/5.

COG	BBge	BBor	BPam	BCpu	BMda	BBgi
J	104	105	104	101	98	101
K	14	15	15	13	14	14
L	29	29	30	28	29	28
D	9	9	9	9	9	9
M	32	33	32	30	30	31
N	1	1	1	0	1	1
V	6	5	4	4	5	6
O	30	30	30	28	28	29
C	40	37	37	39	39	40
G	18	18	18	18	18	17
E	70	68	70	46	52	69
F	32	31	32	30	30	30
H	39	36	38	36	36	36
I	19	21	21	18	19	19
P	15	13	13	12	14	12
Q	2	1	1	1	1	1
T	2	2	2	2	0	2
U	11	11	11	10	9	10
R	37	37	37	38	39	25
S	14	14	14	15	10	13
Chi^2 Statistic		1.7042	1.9715	8.8034	6.392	1.1877
Df		19	19	19	19	19
Chi^2 p-value		1	1	0.977	0.997	1
MC p-value		1	1	0.972	0.998	1
P<0.01		Not rejected	Not rejected	Not rejected	Not rejected	Not rejected

The results for the χ^2 showed that there are not differences in COG distribution between the different strains. Despite the results obtained on the statistical test, the gene loss events in the strains BCpu and BMda are focused in those genes involved in amino acid metabolism and transport, grouped in the COG category E (Figure 14 and Table 13), even so χ^2 test performed only in the E category yields no statistical significative differences (p-value = 0.105).

Finally a clustering diagram was done using the functional profile of *F. psychrophilum* as out-group. In this diagram, symbionts from omnivorous cockroaches show a functional profile closer to that observed in the pan-genome, being BBge the strain most closely related to this simulated dataset. On the other hand, the functional profiles of the strains BCpu and BMda cluster with the core genome dataset. The heat-map shows that the main gap between wood-feeding species and the rest are the loss of genes grouped in the E category in the BCpu and BMda (Figure 15).

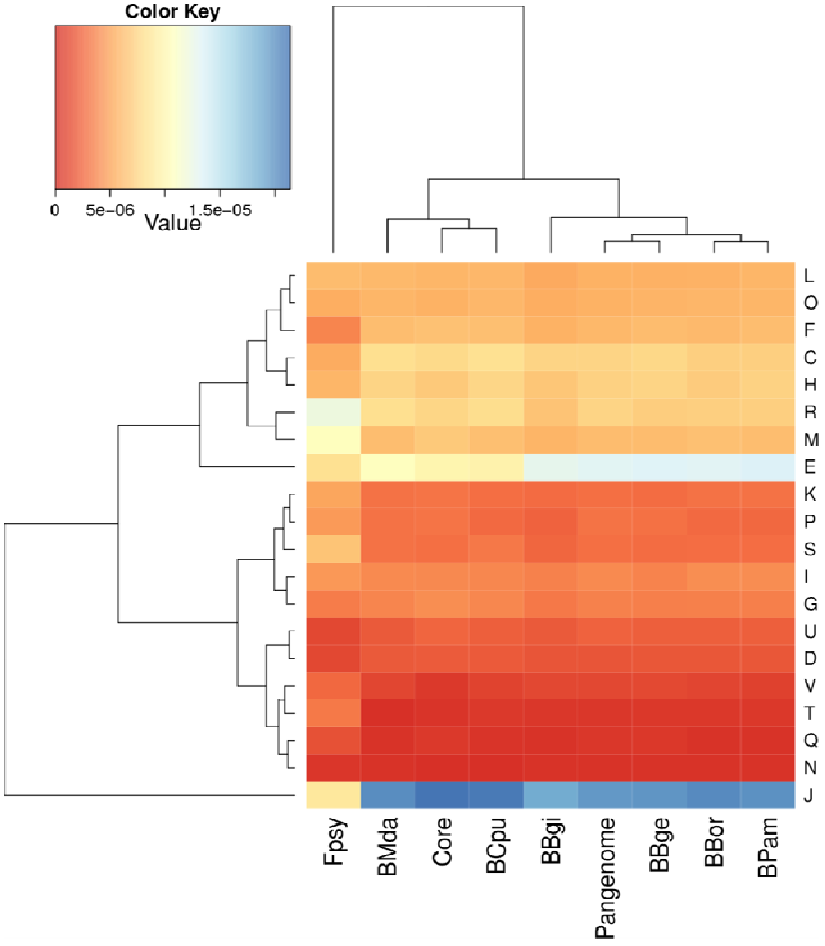

Figure 15. Heat-map comparisons of COG frequency profiles among different *Blattabacterium* strains with their pan-genome and core genome and the free-living Bacteroidete *F. psychrophilum*.

1.4 Differential gene losses and evolutionary history of *Blattabacterium* genomes

In order to obtain a reliable topology of the different *Blattabacterium* lineages and reconstruct the chronology of gene losses, a phylogenetic reconstruction was performed with the six *Blattabacterium* strains and the free-living *F. psychrophilum* as outgroup. The 465 protein coding genes

found to be homologs among *Blattabacterium* strains and possessing orthologs in *F. psychrophilum* were used for the analysis (see the list on Annex 1). They gave rise to a concatenated alignment of 173,523 sites. The best evolutionary model for this dataset, estimated with ProtTest 3 (Darriba et al. 2011) was CpREV+G+F. The Phylogeny obtained (Figure 16) places both Blattidae endosymbionts as a sister clade to the one formed by BCpu and BMda, while BBge and BBgi are clustered together thus corroborating the previous analyses with host genes (Inward et al. 2007).

The phylogenetic reconstruction of the six *Blattabacterium* lineages, the corresponding pan-genome, and the set of retained genes in each genome were used to establish the evolutionary history of gene losses, following the same strategy proposed by Lamelas et al. (2011b). The underlying assumption to our analyses is that the endosymbiotic genomes have experienced no gene gain, thus the retained genes are only vertically inherited. Despite this assumption is commonly accepted, since these bacteria are enclosed in a eukaryotic cell and lack of the complete set of recombination genes, the existence of events of horizontal gene transfer (HGT) was tested without finding any evidence for such events (Annex 3). Through these assumptions, it was possible to reconstruct the ancestral genome of *Blattabacterium* that would contain 655 genes, 648 in the chromosome and seven in the plasmid. Of the 655 genes, 615 were protein coding genes and the remaining 40 coding for RNA genes (34 tRNA, 3 rRNA, SRP RNA, tmRNA and the RNA component of the ribonuclease P) (Figure 17).

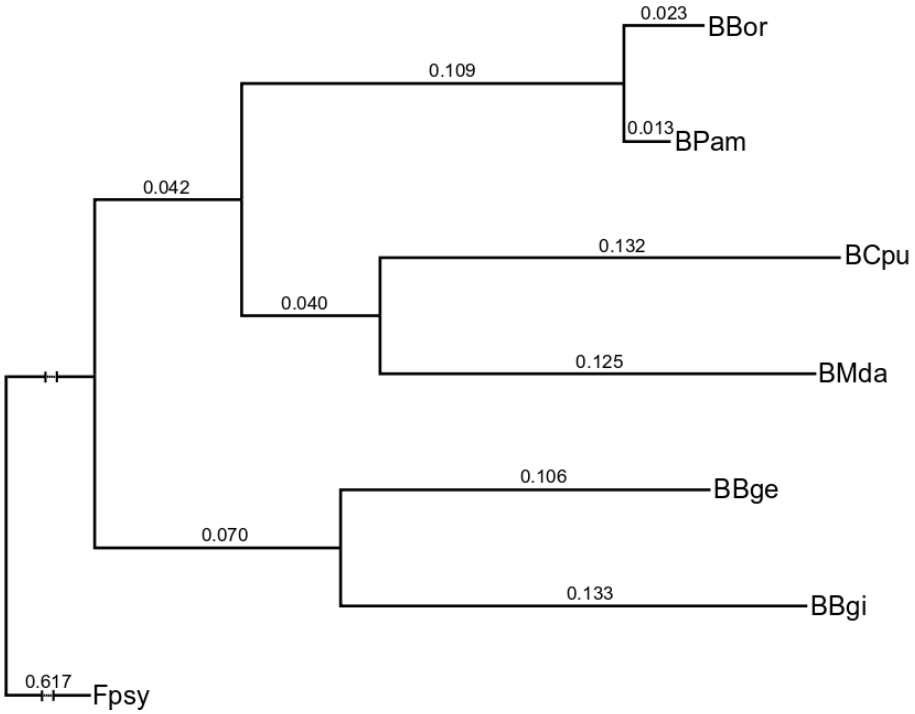

Figure 16. Phylogenetic reconstruction obtained for *Blattabacterium* strains sequenced up today. Tree obtained by maximum likelihood using RAxML on amino acid sequences of 465 concatenated proteins (173,523 aligned sites). Numbers represent the number of amino acid substitutions per site. *F. psychrophilum* was used as outgroup and its branch has been manually shortened.

Taking into account that no HGT events have been detected, the status of each ancestral gene in each extant *Blattabacterium* genome was evaluated: unique or convergent losses and active or pseudogenaized sequence (Table 14). Unique losses are those affecting one specific strain or occurring before divergence of related strains (i.e., the loss took place in their most recent common ancestor). On the other hand, when unrelated lineages show specific pseudogenaized sequences or the absence of a gene, a convergent gene loss event was assumed. During the evolutive history of *Blattabacterium*, since the LSCA (Last Common Symbiotic Ancestor) to the extant strains, 183 gene loss events were identified (Table 14). In these

events only 113 and 3 tRNAs have been involved, because several events of convergent gene loss have occurred. Of them, 70 genes have been lost in unique events. Convergent losses affect to 43 genes, 21 of them have been lost twice, while 22 have been lost three times, these convergent losses account for a total of 114 events. The gene loss events were placed over the topology obtained for the *Blattabacterium* strains in the phylogenetic analysis (removing *F. psycrophilum*) (Figure 17).

Since both superfamilies (Blattoidea and Blaberoidea) diverged, few gene loss events took place during the evolution of Blattoidea before split between Blattidae and Cryptocercidae plus Termitidae families, losses affected 6 genes coding for hypothetical proteins and the gene *sirBC* (bifunctional precorrin-2 dehydrogenase/sirohydrochlorin ferrochelatase).

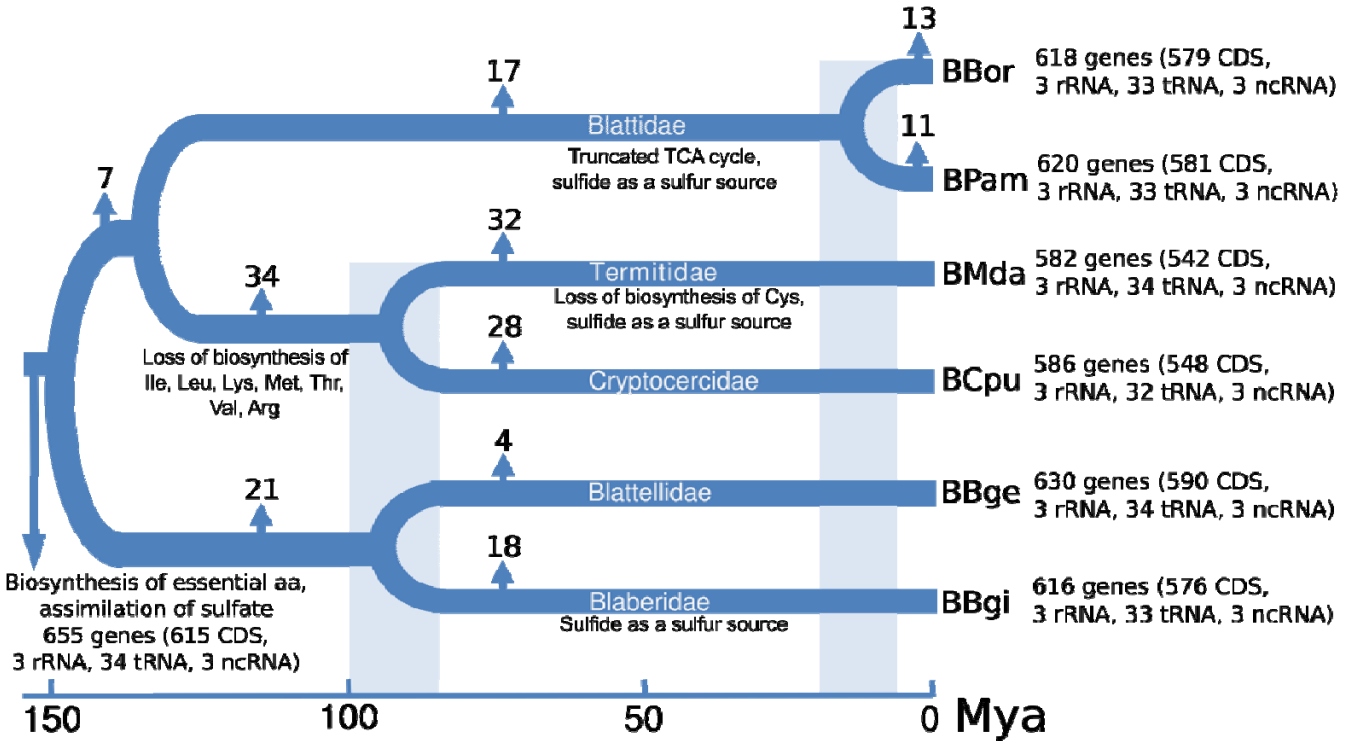

Figure 17. Gene losses during *Blattabacterium* diversification. Number of genes in each strain is indicated. Numbers above the branches indicate gene loss events. Host family names are indicated on each branch. Last common symbiotic ancestor would be situated in the root. Blue rectangles indicate approximated split times**.**

Table 14. Classification of gene losses in different *Blattabacterium* strains. +, present gene; -, absent gene, or pseudogene.

BBor	BPam	BMda	BCpu	BBge	BBgi	n	Genes
One gene lost event							
-	-	-	-	+	+	1	blb_613
+	+	+	+	-	-	1	blb_523
-	-	+	+	+	+	4	*gltA, acnA, icd, msbA*
+	+	-	-	+	+	25	*ilvA, ilvB, ilvC, ilvD, ilvH, leuA, leuC, leuD, leuB, trpE, trpG, trpD, trpC, trpB, trpA, thrB, thrC, lysA, argH, metC, ygfA, asnC, ung,* blb_0546, blb_0548
-	+	+	+	+	+	4	*dut*, blb_543, blb_544, tRNA-val
+	-	+	+	+	+	1	*secE*
+	+	-	+	+	+	19	*ubiE, ribD, nadD, mvaK1, mvaD, efp, trmH, nth, lolE, hemD, ppiC, clpX, dksA, secDF,* blb_0525, blb_0518, blb_0522, blb_0525, blb_0541
+	+	+	-	+	+	14	*metE, metF, cysE, cysK, serC, tgt, luxE, clpB, mdlA, msbA*, blb_0537, blb_0507, blb_0536, blb_0542
+	+	+	+	+	-	3	*ccoQ, rpmF,* tRNA-arg
Two gene lost events							
-	-	-	-	-	+	3	BGIGA_344, BGIGA_431, BGIGA_467
-	-	-	-	+	-	3	*sirBC*, BLBBGE_159, BLBBGE_195
-	-	+	+	-	-	2	*ywrO*, blb_0586
+	+	+	-	-	-	1	*desA*
+	+	-	+	-	-	1	blb_565
+	+	-	-	+	-	1	blb_035
+	+	-	+	+	-	3	*truB, era*, blb_529
+	-	-	+	+	+	1	*lolD*
+	-	-	-	+	+	1	blb_591
-	+	+	+	-	-	1	*ccoS*
-	+	+	-	+	-	3	*cysH, cysG, hemC*
-	+	-	+	+	+	1	blb_578
Three gene lost events							
+	-	-	-	-	-	4	BLBBOR_609, BLBBOR_p001, BLBBOR_p002, BLBBOR_p007
+	-	+	-	+	-	1	tRNA-pro
+	-	+	-	-	-	1	blb_585
+	-	-	+	-	-	2	*rnpA,* blb_587
-	+	-	-	-	-	1	BPLAN_099
-	+	+	-	-	-	1	*lpxP*
-	+	-	-	+	-	1	blb_596
-	+	-	-	-	+	1	*trpF*
-	-	+	-	-	-	1	MADAR_453
-	-	+	-	+	-	5	*cysD, cysN, cysI, hemD*, blb_595
-	-	-	+	-	-	5	BLBCPU_006, BLBCPU_149, BLBCPU_186, BLBCPU_463, BLBCPU_511

The common ancestor of BBor and BPam should have already suffered the deletion of three genes from the Krebs cycle: *gltA* (citrate synthase), *acnA* (aconitase), and *icd* (isocitrate dehydrogenase). The sulfate assimilatory pathway would be also impaired because the inactivation of the genes coding for ATP sulfurylase (*cysN* and *cysD*) and the loss of one of the subunits of the sulfite reductase (*cysI*) (Figure 18). The gene *hemD* and an ABC-type multidrug transporter and twelve genes coding for hypothetical proteins were also inactivated. During the evolution of BBor two more genes of the sulfate assimilatory pathway, *cysH* and *cysG* (phosphoadenosine phosphosulfate sulfotransferase and siroheme synthase, respectively), have been inactivated and the synthesis of uroporphyrinogen III should be affected by the inactivation of *hemC* (hydroxymethylbilane synthase). The losses of *dut* and *trpF* during the evolution of this lineage (Table 14) have no metabolic effect since a copy for *dut* remains active in the plasmid, and *trpF* is fused to *trpB* in the genome of BBor. The other genes that have been lost are *lpxP*, *ccoS* and five coding for hypothetical proteins. The endosymbiont of *P. americana* (BPam) has lost two genes involved in the mobilization and transport of proteins through membranes, *secE* and *lolD*, the first is the translocase of the Sec secretory system and the last is the ATPase of the Lol system involved in sorting lipoproteins to the outer membrane. The other genes lost or inactivated are *rnpA* (the protein component of the ribonuclease P) and 7 genes coding for hypothetical proteins.

The branches leading to BMda and BCpu account for the major number of losses, 34 in the branch leading to their common ancestor and 32 and 28 during the evolution of BMda and BCpu, respectively (Figure 17). The ability to synthetize several essential amino acids was already lost before the split of both strains, thus the common ancestor to BCpu and BMda would be unable to synthetize branched chain amino acids because of the loss of *ilvA*, *ilvH*, *ilvC*, *ilvD*, *leuA*, *leuC*, *leuD* and *leuB*. In the same way, the operon for

the synthesis of tryptophan was completely loss. The synthesis of amino acids of the aspartate family is impaired by the losses of *thrB*, *thrC*, *metC* and *lysA*. Finally, the loss of *argH* (argininosuccinate lyase) affects the biosynthesis of arginine and breaks the urea cycle. In addition the genes *ygfA* (5-formyl tetrahydrofolate cycloligase), which participates in the recycling of 5-formyltetrahydrofolate, *ung* (uracil-DNA glycosylase) a DNA repair gene and the gene coding for the transcription regulator AsnC had been lost.

Along the evolution of BCpu, it has been lost the complete pathway for sulfate assimilation and cysteine biosynthesis (Figure 18), as well as the genes for the synthesis of heme groups (*hemC* and *hemD*). Besides, additional genes that participate in amino acid synthesis are lost, like the genes *metF* and *metE*, involved in the synthesis of methionine, or the gene *serC* that codes for a transaminase that participates in the serine biosynthesis. Other gene losses in this lineage affect the maturation of tRNAs (*tgt*) or the synthesis of linoleic acid (*desA*). Other losses are summarized in Table 14. During the evolution of BMda, like in the case of BPam, the transport and sorting of proteins through membranes is affected by the losses of genes for the Sec secretory system (*secDF*) and genes for the Lol system (*lolD* and *lolE*). The synthesis of several cofactors could be affected by the inactivation of the genes *ubiE*, *ribD* and *nadD*. The loss of *mvaK* (mevalonate kinase) and *mvaD* (diphosphomevalonate decarboxylase) prevent the synthesis of isopentenyl diphosphate. A number of genes involved in the informational storage and processing are lost or inactivated in this genome, like *def* (peptide deformilase), *era* (involved in ribosome maturation), *efp* (elongation factor P), three genes involved in the maturation of tRNAs (*truB*, *trmH* and *rnpA*), a transcriptional regulator (*dksA*) and an endonuclease III involved in DNA repair.

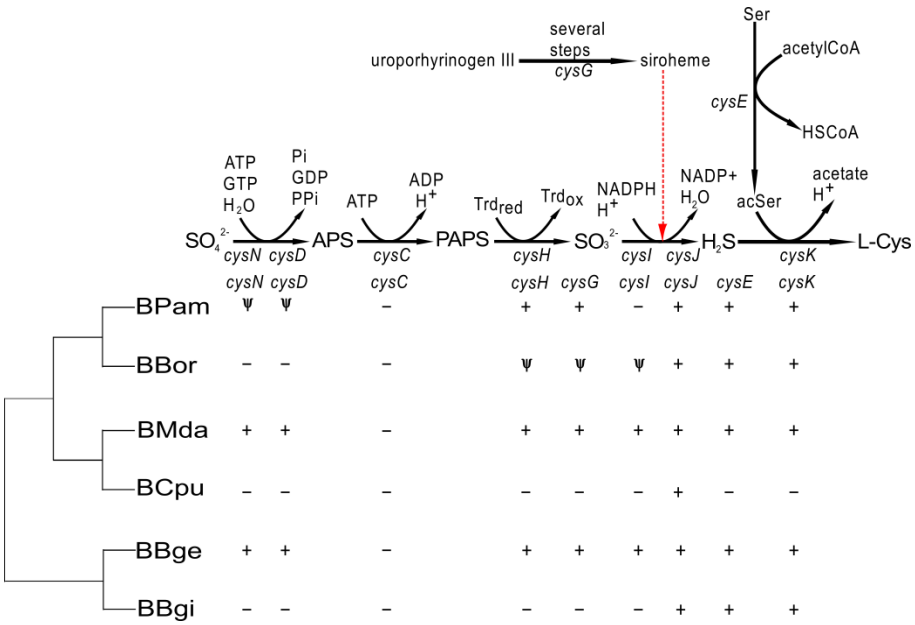

Figure 18. Sulfate assimilatory pathway genes in different *Blattabacterium* strains. +, present gene; –, absent gene ψ, pseudogene. Trd, thioredoxin; acSer, acetylserine.

Finally, the lineage leading to BBge and BBgi has undergone 21 losses before the split between both strains. Most of them are genes for hypothetical proteins, or with only general predicted functions (16 out of 21). All genes with an assigned function lost in this lineage have also been lost in some of the other strains (*ywrO, lpxP, ccoS, rnpA* and *desA*) and their losses do not affect any metabolic pathway. Along the evolution of the lineage BBgi, 17 additional CDS have been lost, among them, the sulfate assimilatory pathway, two genes involved in the synthesis of heme groups, *hemC* and one of the copies of *hemD* and two genes involved in informational storage and processing *era* and *truB*. Two genes are exclusively lost in this strain, the ones coding for the subunit IV of the cytochrome oxidase (*ccoQ*) and the ribosomal protein L32 (*rpmF*). Finally, only four genes losses occurred in the branch to BBge, three of them

affecting hypothetical proteins and the fourth is the gene *trpF*, which in this strain is fused to *trpB*.

1.5 Comparative metabolism within *Blattabacterium* strains

The metabolism of the symbiotic *Blattabacterium* from the four omnivore cockroaches (*B. germanica, P. americana, B. gigantenus* and *B. orientalis*), and probably also the metabolism of the LCSA, is remarkably conserved. Actually, the metabolic network of BBge and the LCSA are nearly equivalents, as far as the LCSA is an estimation inferred only from the sequenced strains. As it is pointed in the previous section, gene losses in BBge mainly involve genes coding for hypothetical proteins.

In the above section, it was stated that one of the main metabolic differences in *Blattabacterium* strains from the Blattidae family were the absence of the three first steps of Krebs cycle. To evaluate the effect of these losses, an stoichiometric analysis was performed using the program METATOOL on the enzymes involved in the Krebs and urea cycles, together with the enzymes that catalyse related reactions like urease and malate dehydrogenase, as well as those enzymes involved in the amino acid biosynthesis linked to these pathways (Figure 19). There were obtained five elementary modes (i.e., minimal metabolic routes feasible in steady state in a determined network), three of them forming the convex basis (i.e. those reactions that characterize the whole metabolism in the enzyme subset) (Table 15). The output of this analysis shows that both BBor and BPam, with the same metabolic structure, are able to produce ammonia through the catabolism of amino acids. The input and output files for METATOOL are included in the Annex 4.

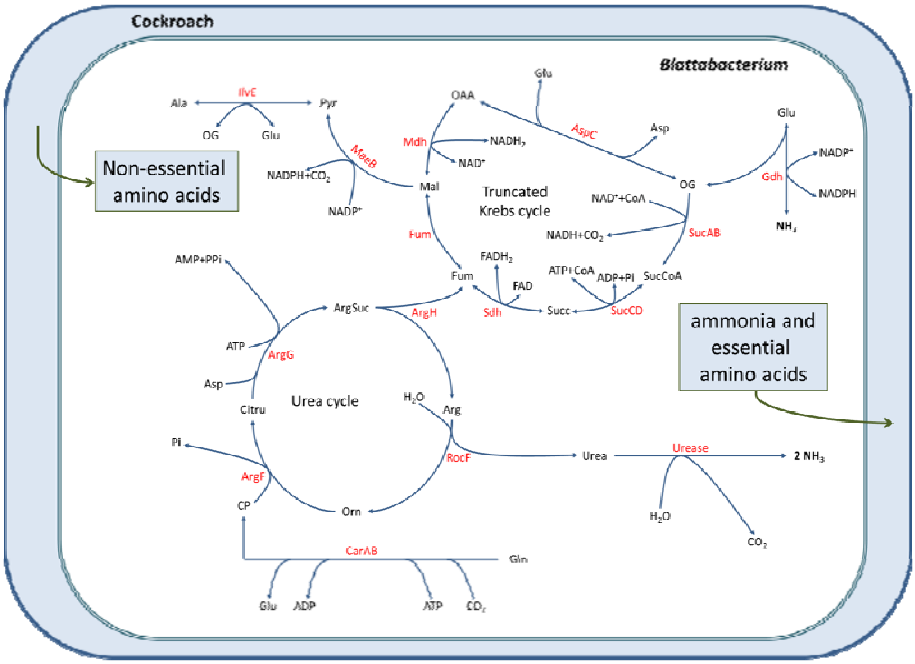

Figure 19. *Blattabacterium* metabolic pathways used for the stoichiometric analysis. Double-headed arrows indicate reversible reactions. Metabolites and enzymes are written as in the input file for METATOOL (Annex 4). Metabolites are written in black (Pyr, pyruvate; Mal, malate; OAA, oxaloacetate; OG, 2-oxoglutarate; SucCoA, succinyl-CoA; Succ, succinate; Fum, fumarate; CP, carbamoylphosphate; Orn, ornithine; Citru, citrulline; ArgSuc, argininosuccinate, Ala, alanine; Glu, glutamate; Arg, arginine; Asp, asparagine). Enzymes are shown in red (IlvE, branched chain aminotransferase; MaeB, malate dehydrogenase; Mdh, malate dehydrogenase NAD-dependent; AspC, aspartate aminotransferase; Gdh, glutamate dehydrogenase; SucAB, 2-oxoglutarate dehydrogenase; SucCD, succinyl-CoA synthetase; Sdh, succinate dehydrogenase; Fum, fumarase; ArgH, argininosuccinate lyase; RocF, arginase; CarAB, carbamoylphosphate synthase; ArgF, ornithine carbamoyltransferase; ArgG, argininosuccinate synthase).

The analysis in BBge resulted in 14 elementary modes, five of them corresponding to the convex basis, and 11 out of 14 elementary modes producing ammonia as a final product (López-Sánchez et al. 2009).

Table 15. Results of the stoichiometric analysis on the central networks for BPam and BBge.

	Overall reaction	Participating enzymes	Metabolic function
	Convex basis		
1	NADP + NADH + Asp = CO2 + NADPH + NAD + Ala	AspC -Mdh IlvE maeB irreversible	Transhidrogenase dependent of Asp decarboxylation to Ala
2	ADP + FAD + 2 NAD + Glu + P = CO2 + ATP + FADH2 + 2 NADH + Asp	-AspC SucCD Sdh Fum Mdh SucAB irreversible	Partial catabolism of Glu (second half of Krebs cycle)
3	2 ATP + NADPH + NAD + Gln + 3 H2O = ADP + NH3 + NADP + NADH + Glu + AMP + diPP + P	-AspC Fum Mdh Gdh carAB ArgF ArgH rocF ArgG Urease irreversible	Urea cycle coupled with urease.
	Elementary modes		
4	ADP + FAD + NADP + NAD + Glu + P = 2 CO2 + ATP + FADH2 + NADPH + NADH + Ala	SucCD Sdh Fum IlvE SucAB maeB irreversible	Partial catabolism of Glu
5	2 ATP + Asp + Gln + 3 H2O = ADP + CO2 + NH3 + Ala + Glu + AMP + diPP + P	Fum Gdh IlvE maeB carAB ArgF ArgH rocF ArgG Urease irreversible	Decarboxylation of Asp to Ala

One of the main metabolic differences in the different strains is the sulfur source (Figure 18). While the symbionts from *B. germanica* and *M. darwiniensis* possess the complete set of genes for assimilate inorganic sulfate, the other strains (BCpu, BPam, BBor and BBgi) must use sulfide directly. As mentioned above, the genes *cysDNHIJG* are lost in a convergent manner in three events, one in the branch leading to BCpu and other in the branch leading to BBgi. Finally, in the family Blattidae the genes that code for ATP-sulfurylase (*cysND*) and the hemoprotein component of the sulfide reductase (*cysI*) have been lost before the split between *P. americana* and *B. orientalis*. Furthermore, during the evolution of *B. orientalis* the genes for the PAPS sulfotransferase (*cysH*) and siroheme synthase (*cysG*) have been

inactivated. In the strain BPam active copies for *cysH* ans *cysG* remain, the last not directly implied in the pathway, since only participates in the synthesis of the heme group of the sulfite reductase, which do not seem to be active. Actually in the FBA analyses carried out for this strain, this step was removed (González-Domenech et al. 2012).

Metabolic differences among omnivorous roaches, both Blattidae (BPam and BBor) and Blaberoidea (BBge and BBgi), are basically restricted to both cases pointed above. Beyond the Krebs cycle and sulfate assimilation, the only remarkable difference is that BBor, BPam and BBgi are all three able to convert oleic acid into linoleic acid by means of the desaturase activity of the product of *desA*.

1.6 Evolutionary analysis on *Blattabacterium*

1.6.1 Dating the split times

Genes that follow a molecular clock were identified by performance of a LRT analysis, where each gene in the core was tested under two models, first assuming molecular clock, which would be the null hypothesis (H_0), and second allowing each branch to evolve at different rate, which would be the alternative hypothesis (H_1). LRT supported molecular clock hypothesis for 296 CDS out of a total of 502 (Annex 5). From this set of 296 core CDS, 275 have an ortholog in *F. psychrophylum* and 261 of them do not reject the molecular clock hypothesis. In addition, to deal with the problem that nucleotide substitution saturation can suppose, the 37 genes which show more than 2.5 substitution per site were discarded, resting, finally, 224 CDS. These genes allowed us to estimate the divergence time between the Blattidae lineages as well as between both Blaberoidea and the divergence

among xylophage lineages. Thus, the divergence time between BPam and BBor must have occurred 12.3 ± 7.6 Mya, while the split between the BCpu and BMda should have occurred 87.0 ± 18.8 Mya. In the branch leading to the Blaberoidea, it have been estimated that the split between the *B. germanica* and *B. gigantenus* must be happened parallel to the split of wood-feeding branch individuals, 89.5 ± 17.7 Mya (see blue rectangles in figure 17).

1.6.2 Synonymous and non-synonymous nucleotide substitutions.

The number of synonymous (dS) and non-synonymous (dN) substitutions per synonymous and non-synonymous site, respectively between each pair of orthologous genes was estimated using the approximated method proposed by Yang and Nielsen (2000), which takes into account the transition/transversion ratio bias and unequal base frequencies. The dN/dS changes have been calculated for every coding gene to determine whether they are under purifying, neutral or positive selection. The values for the dN, dS and the ratio dN/dS (ω) for each pair of orthologous among the different *Blattabacterium* strain can be found in the Annex 6. The results for each comparison are summarized in the Table 16.

Purifying selection is the main force operating in these organisms. Nevertheless, when we compare genes between the strains BBor and BPam, two genes seem to be under neutral selection (*ccoQ* and *atpH*, coding for a cbb_3-type cytochrome oxidase subunit and the subunit III of the ATPase, respectively), whereas other three (*greA*, *nadD* and *ribE*, a transcription elongation factor, the nicotinate-nucleotide adenylyltransferase and the riboflavin synthase alpha subunit, respectively) seem to be subject of positive selection. However, these results might be taken cautiously as they must be due to sampling errors. Additionally, with regards to dS values,

there are elevated in most comparisons above mentioned, with an average value close to 2. This high value indicates saturation for synonymous sites.

Table 16. Comparison of dN, dS and ω among the six *Blattabacterium* strains

Strains	No. of comparisons	Average ω	Average dS	Average dN
BBor-BPam	565	0.218 ± 0.744	0.104 ± 0.056	0.016 ± 0.012
BBor-BCPu	529	0.073 ± 0.063	1.850 ± 0.840	0.108 ± 0.059
BBor-BMda	515	0.078 ± 0.059	1.629 ± 0.705	0.106 ± 0.055
BBor-BBge	566	0.068 ± 0.055	2.292 ± 0.979	0.126 ± 0.065
BBor-BBgi	563	0.077 ± 0.064	2.355 ± 4.171	0.140 ± 0.069
BPam-BCpu	532	0.075 ± 0.066	1.782 ± 0.815	0.106 ± 0.059
BPam-BMda	529	0.081 ± 0.070	1.593 ± 0.751	0.104 ± 0.057
BPam-BBge	575	0.069 ± 0.069	2.264 ± 0.953	0.124 ± 0.068
BPam-BBgi	565	0.075 ± 0.059	2.351 ± 4.166	0.137 ± 0.068
BCpu-BMda	516	0.087 ± 0.071	1.406 ± 0.704	0.094 ± 0.042
BCpu-BBge	530	0.066 ± 0.066	2.497 ± 1.013	0.128 ± 0.060
BCpu-BBgi	531	0.074 ± 0.060	2.520 ± 4.297	0.138 ± 0.062
BMda-BBge	529	0.064 ± 0.055	2.301 ± 0.949	0.128 ± 0.061
BMda-BBgi	525	0.080 ± 0.073	2.181 ± 0.888	0.140 ± 0.067
BBge-BBgi	572	0.079 ± 0.070	1.685 ± 4.126	0.096 ± 0.050

2.- Transcriptome analysis of host tissues

The studies carried out on the reconstruction of the metabolic networks of the different *Blattabacterium* strains have been useful to suggest its key role in the physiology of cockroaches. However, the activities carried out by the host have only been inferred and functional studies should be done. As a first approach, the whole transcriptome of three tissues of *B. germanica* was characterized. Two of these tissues, fat body and ovaries, harbour *Blattabacterium*, while in the third one, the epithelium, endosymbionts are absent and was chosen as control. Moreover, it has been postulated that *Blattabacterium* activity is reduced if any, in the ovaries.

2.1 Assembly and annotation of the 454 sequences

From the sequencing process, 554,403 reads from the three analysed tissues were retrieved. These sequences were sorted according to their origin thanks to the barcode sequence attached to them during the cDNA synthesis process, and classified in three datasets, one for each tissue. The reads were also trimmed for low quality regions. The statistics for the trimmed sequences are summarized in Table 17.

Table 17. Main statistics of the sequencing project.

Sample	Fat Body	Ovary	Epithelium
No. of sequences	164,677	166,672	223,054
Bases (bp)	38,729,289	39,279,123	50,422,530
Reads range (bp)	15-785	15-823	15-842
Mean length (bp)	235	235	226
N50[a] (bp)	299	297	293
CG content (%)	41.01	38.71	37.91

[a]Length for which the all contigs of equall or more length contains the 50% of the sequences.

Prior to the assembly, all the reads from each dataset were mapped against the genome of *Blattabacterium* strain BBge to identify those sequences belonging to the endosymbiont. In the fat body, 10,578 reads were identified, representing the 6.42% of the sequences of this dataset. On the other hand, only 104 sequences matched with *Blattabacterium* genes in the ovaries, and just three reads from the epithelial tissue. *B. germanica* sequences were posteriorly assembled with Mira 3.2. In a second step, all the sequences showing similarity to *Blattabacterium* sequences, where removed from the analysis to deal only with host mRNAs. The results of the assembly process are summarized in Table 18.

Table 18. Statistics for the assembly process.

Sample	Fat Body	Ovary	Epithelium
Input reads	154,099	166,568	223,051
Assembred reads	83,802	110,628	136,070
Contigs	11,905	17,159	23,318
≥500 bp	3,500	6,393	6,393
Total Consensus	5,529,891	8,738,873	10,269,711
Coverage	3.89	3.29	3.25
N50 (bp)	512	572	491
Mean (bp)	466.9	512.4	442.5
Range (bp)	40-9,738	40-4,267	40-5,014

Most of the reads were incorporated to contigs bigger than 500 bp. Thus, 71.4, 73.5 and 66.1% of the assembled reads from the fat body, ovary and epithelium, respectively, were incorporated in the so-called big contigs.

After the assembly process, the contigs were annotated with Blast2GO software (Conesa et al. 2009) that perform BLAST searches to identify similar sequences to our query. Afterwards, GO terms are assigned by using the BLAST hits as a query against the GO database. This gene ontology initiative seeks to standardize the name and function of every gene product across all domains of life. Each gene was described in three main domains,

according its properties. These domains were: (i) cellular component, which refers to its localization, or the structure that the products form; (ii) biological process, which refers to a series of molecular events with defined starts and ends, and (iii) molecular function, which describes the activity carried out by the gene product. In last step, Blast2GO, uses the GO to EC mapping file in the GO web site to assign an EC annotation to the enzymes present in the query dataset. The annotation process for all three tissues is summarized in Table 19 and in the Figure 20.

Table 19. Annotation statistics.

Tissue	Fat body	Ovary	Epithelium
No. of Contigs	11,905	17,159	23,318
BLASTX[a]	5,848 (3,113)	8,052 (6,782)	6,396 (4,126)
Contigs with assigned GO	3,993	5,376	3,623
Average GOs per contig	5.3±1.9	5.7±1.9	5.7±1.9

[a] In parenthesis, number of different accession numbers

Close to a half of the sequences from fat body and ovaries have significantly positive BLAST hits, conversely only the 27.4% of the sequences from the epithelium have significant matches with sequences from the non-redundant NCBI database. However, since there are several contigs that code for the same product, the number of different accession numbers is smaller than the number of contigs showing similarity with sequences in the NCBI database.

The accession numbers of the BLAST hits were used to assign GO terms to each sequence through searches against the GO database. Thus, 33.5, 31.3 and 15.5% of the total amount of contigs of the fat body, ovaries and epithelium datasets, respectively, were properly annotated (Figure 20).

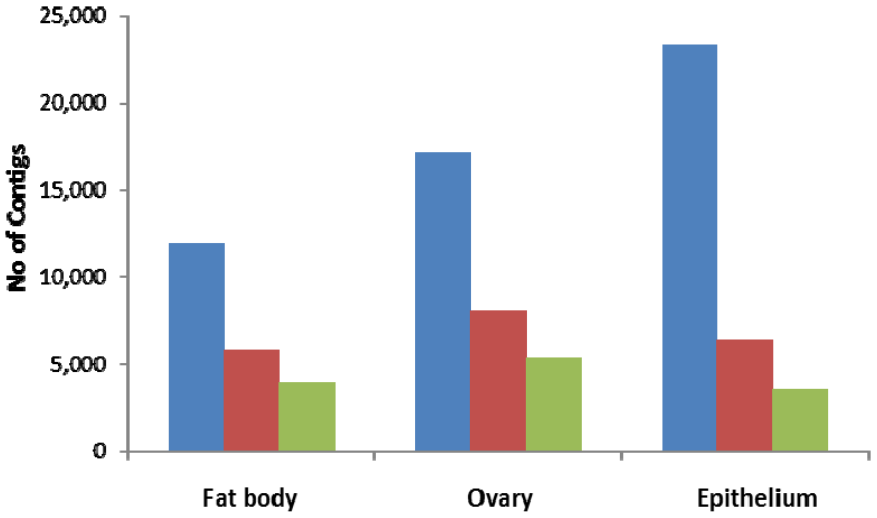

Figure 20. Overview of the annotation process in the three different analysed tissues. In blue, the initial number of contigs; in red, sequences matching with any sequence when searched against non-redundant GenBank protein database with BLASTX, and in green, number of sequences properly annotated.

2.2 Metabolic capabilities of three tissues in *B. germanica*

2.2.1 Fat body

A total of 2,708 unique GO terms in 17,109 occurrences were assigned to the fat body sequences. A total of 3,273 contigs were grouped under biological processes, 3,684 under molecular function and 2,060 under cellular components. To have a broad level overview, GO terms were mapped against the GO-slim vocabulary. Given that GO classification is organized as a Directed Acyclic Graphs (DAG) where each term is linked to one or more other terms, most of the times of the same domain, and the higher domains are too broad, pie charts of the distribution for the lowest

levels were obtained for each one of the three domain of the GO classification (Figure 21).

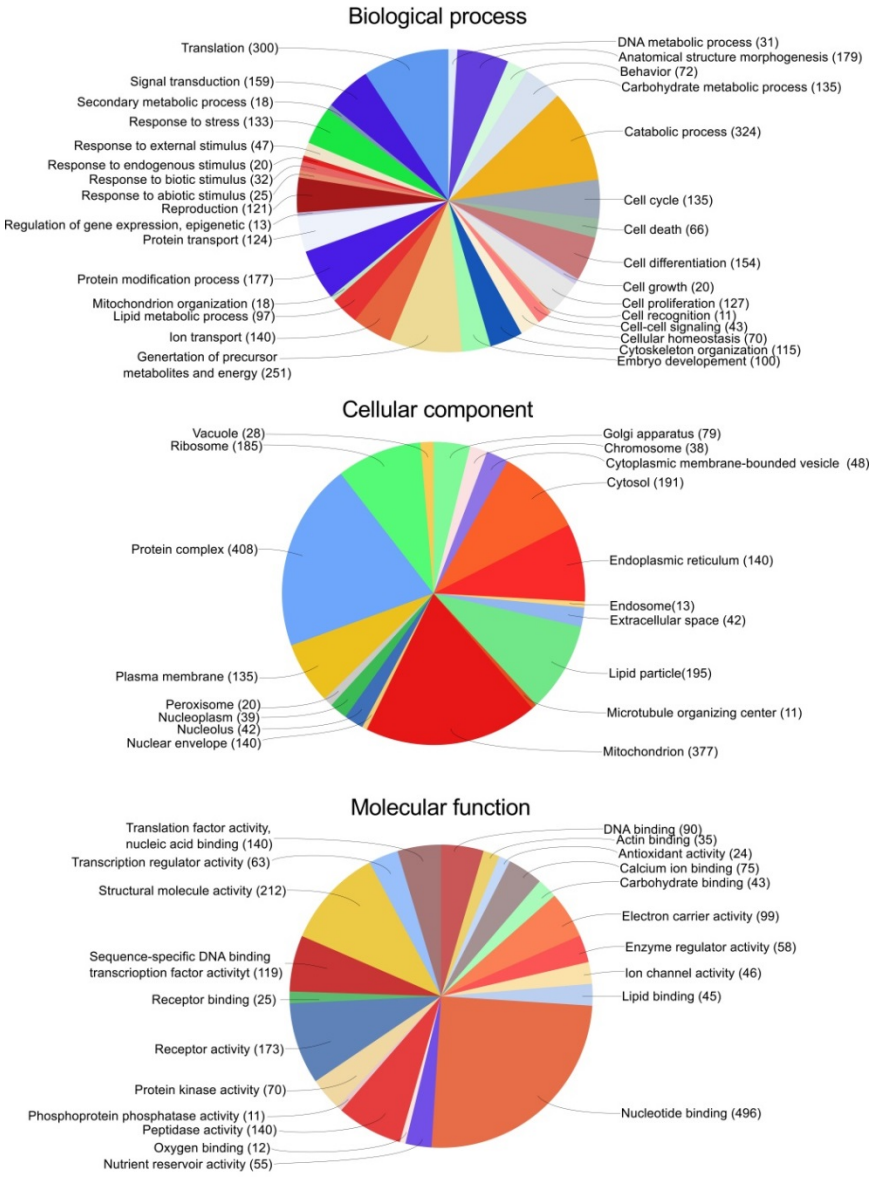

Figure 21. Fat body multilevel pie chart of sequence distribution according to GO classification for the three main GO categories.

Using the information of GO terms, 1,559 Enzyme Commission numbers (EC numbers) were assigned to 1,146 different sequences. These enzymes are localized in 116 out of 164 KEGG metabolic maps.

For evaluating the completeness of the library, four of the central metabolic pathways present in animal phyla were searched, namely glycolysis and gluconeogenesis, pentose phosphate pathway and the Krebs cycle.

The fat body dataset include annotated sequences for eight out of ten genes coding for enzymes involved in the glycolysis pathway (Figure 22), being the non-detected transcripts those genes coding for phosphofructokinase (PFK) and phosphoglycerate mutase (PGM). The gluconeogenesis process shares most of the enzymes with glycolysis, except for the three irreversible steps in the catabolic pathway, i.e., hexokinase (HK), PFK and pyruvate kinase (PK), which are replaced by glucose-6-phosphatase (G6Pase), fructose biphosphatase (FBPase) and the couple pyruvate carboxylase/PEP carboxykinase (PC/PEPCK) (Figure 22), respectively. In this second pathway only transcripts for two genes were absent, the corresponding to PGM and G6Pase. For the pentose phosphate pathway transcripts for six out of seven required genes were identified. Four were properly annotated, the ones coding for glucose-6-phosphate dehydrogenase (G6PDH), phosphogluconate dehydrogenase (PGDH), ribulose-5-phosphate 3-epimerase (RPE) and transaldolase (TA). Transcripts for two other enzymes, 6-phosphogluconolactonase (PGLS) and transketolase (TK) were identified with BLAST searches among the transcriptome sequences. Only for ribulose-5-phosphate isomerase (RPI) there was no signal in the whole dataset (Figure 23). Finally, there are annotated sequences for all genes encoding for Krebs cycle enzymes except for aconitase (Figure 24).

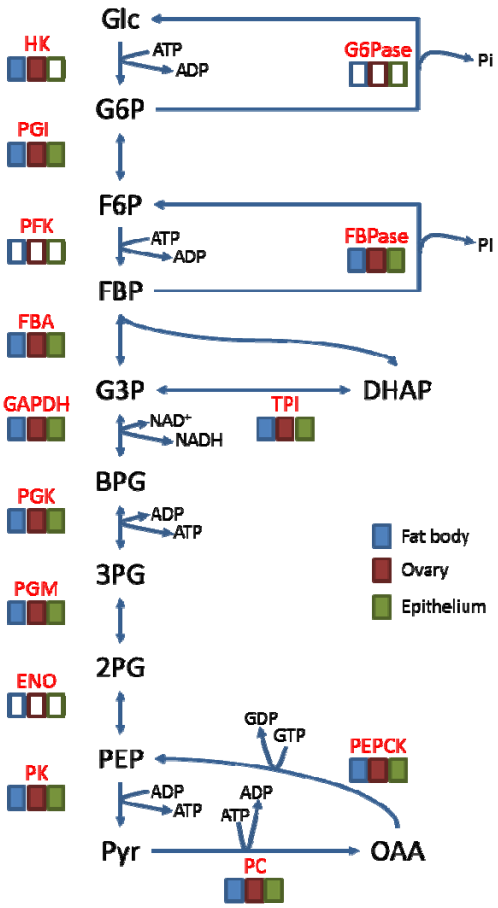

Figure 22. Glycolysis and gluconeogenesis. Empty squares indicate that no transcripts for the corresponding enzyme have been found in the corresponding tissue. Compounds: Glc, glucose; G6P, glucose-6-phosphate; F6P, fructose-6-phosphate; FBP, fructose-1,6-bisphosphate; G3P, glyceraldehyde-3-phosphate; DHAP, dihydroxyacetonephosphate; BPG, 1,3-bisphosphoglycerate; 3PG, 3-phosphoglycerate; 2PG, 2-phosphoglycerate; PEP, phosphoenolpyruvate; Pyr, pyruvate; OAA, oxaloacetate. Enzymes: HK, hexokinase; PGI, phosphoglucose isomerase; PFK, phosphofructokinase; FBA, fructose bisphosphate aldolase; GAPDH, glyceraldehyde phosphate dehydrogenase; TPI, triosephosphate isomerase; PGK, phosphoglycerate kinase; PGM, phosphoglycerate mutase; ENO, enolase; PK, pyruvate kinase; PC, pyruvate carboxylase; PEPCK, phosphoenolpyruvate carboxykinase.

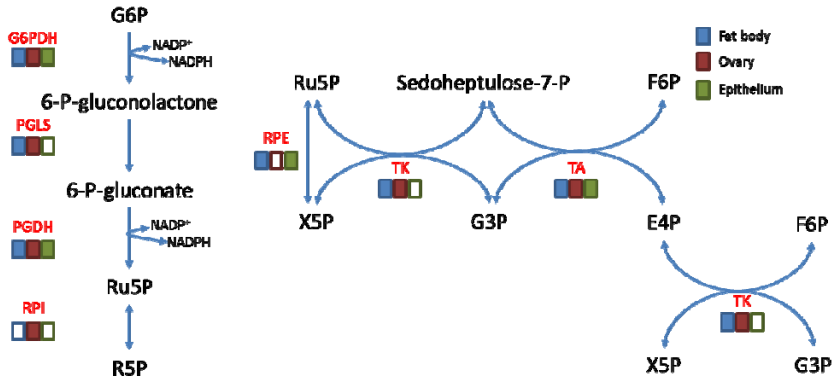

Figure 23. Pentose phosphate pathway. Empty squares indicate that no transcripts for the corresponding enzyme have been found in the corresponding tissue. Compounds: Ru5P, ribulose-5-phosphate; R5P, ribose-5-phosphate; X5P, xylulose-5-phosphate; E4P, erythrose-4-phosphate. Enzymes: G6PDH, glucose-6-phosphate dehydrogenase; PGLS, 6-phosphogluconolactonase; RPI, ribose-5-phosphate isomerase; RPE, ribulose-5-phosphate epimerase; TK, transketolase; TA, transladolase.

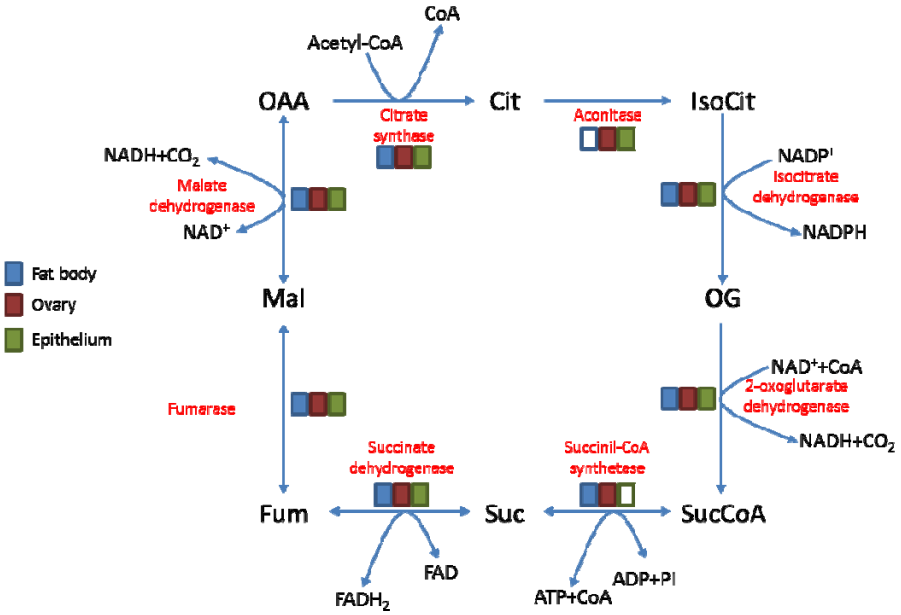

Figure 24. Krebs cycle. Empty squares indicate that no trasncripts for the corresponding enzyme have been found. Compounds: Cit, citrate; IsoCit, isocitrate; OG, α-ketoglutarate; SucCoA, succinyl-CoA; Suc, succinate; Fum, fumarate; Mal, malate; OAA, oxaloacetate.

For the metabolic reconstruction of *Blattabacterium* strain BBge, it was assumed that several non-essential amino acids like asparagine, glutamine, glycine and proline, as well as other compounds such as porphobilinogen and pantotheine-4-phosphate, should be supplied by the host (González-Domenech et al. 2012). The transcripts of the genes required for the synthesis of all four non-essential amino acids were annotated among the sequences of the library from the fat body. Thus, asparagine can be synthetized from oxaloacetate via aspartate, since transcripts for aspartate amino transferase and asparagine synthetase have been identified. There have been found transcripts for glutamate dehydrogenase and glutamine synthetase required for the synthesis of glutamine. There have been identified transcripts of genes coding for the required enzymes for the synthesis of glycine from 3-phospho-D-glycerate via serine, namely: phosphoglycerate dehydrogenase, phophoserine transaminase, phosphoserine phosphatase (which synthetize serine from 3-phospho-D-glycerate), and serine hydroxymethyl transferase that converts serine into glycine. Finally proline is synthetized from glutamate-δ-semialdehyde with pyrroline-5-carboxylate reductase. Glutamate-δ-semialdehyde can be synthetized from glutamate by glutamate demialdehyde dehydrogenase, or from the ornithine through the action of ornithine-δ-transaminase.

In addition to genes needed for the synthesis of non-essential amino acids, the expression of genes involved in the synthesis of porphobilinogen from glycine, aminolevulinic acid synthase and porphobilinogen synthase, were detected. Besides, there is the complete pathway for the synthesis of protoheme IX, with the exception of uroporphirinogen III synthase (this enzyme is encoded by the endosymbiont genome) and protoporphyrinogen oxidase. It was also detected the expression of the gene for pantothenate kinase gene. Finally, transcripts from genes coding for the necessary

enzymes for the synthesis (xanthine dehydrogenase) and degradation (urate oxidase, allantoinase and allantoicase) of uric acid were annotated.

2.2.2 Ovaries

A total of 28,889 GO's, corresponding to 4,236 different terms were assigned to the ovary dataset. A total of 4,003 sequences were assigned to GO terms from the biological process domain, 4,651 sequences were assigned with GO linked to molecular function domain and 3,633 in the cellular compartment domain. Like in the fat body, GO terms assigned in this tissue were mapped against the GO-slim library, and multilevel pies were generated for each one of the main GO categories (Figure 25).

As for the fat body, the principal metabolic pathways were analyzed at first. In this tissue, the annotated genes for the glycolysis/gluconeogenesis pathway showed the same profile than in the fat body transcriptome, where there were no annotated sequences for the PFK, PGM and G6Pase (figure 22). The same situation was observed for the pentose phosphate cycle, with the identification of six out of seven enzymes with the exception of the RPE (Figure 23). As in the fat body, TK was identified only by BLASTX searches against the non-redundant database. Finally, the complete set of enzymes for the Krebs cycle was annotated (Figure 24).

Finally, no transcripts from genes coding for the necessary enzymes for the synthesis (xanthine dehydrogenase) and one involved in the uric acid degradation (allantoinase) were found.

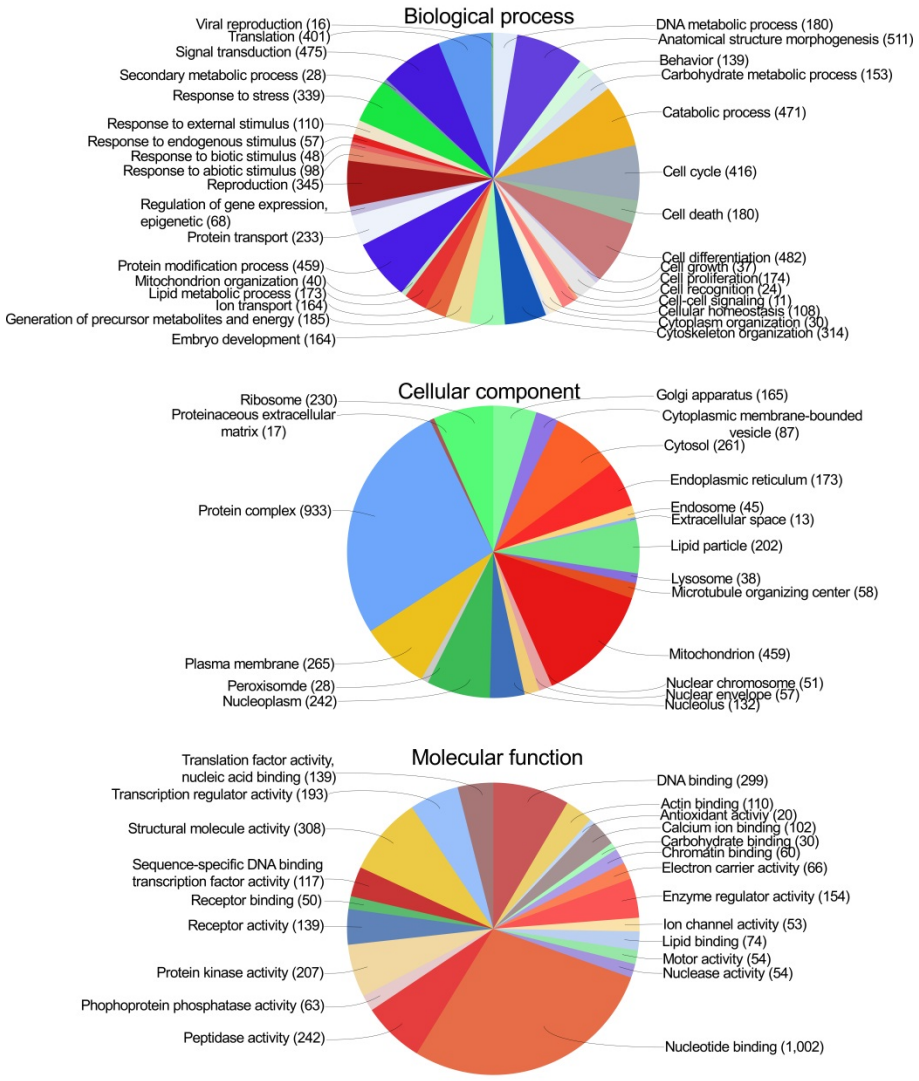

Figure 25. Ovary multilevel pie chart of sequence distribution according to GO classification for the three main GO categories

2.2.3 Epithelium

A total of 18,903 GO terms, which counted to 2,923 different terms, were assigned to the sequences obtained from this tissue. The three main domains accounted for 2,687 sequences in the biological process category, 3,227 on the molecular function and 2,334 on the cellular function category. As for the previous tissues, multilevel pies with the sequence/GO distributions were generated after mapping the GO terms obtained to the GO-slim library (Figure 26).

In the glycolysis/gluconeogenesis pathway, the expression of the four genes coding for PFK, PGM, G6Pase and HK were not detected (Figure 22). Only four genes (out of seven) were annotated for the phosphate pentose pathway being the lacking enzymes PGLS, RPI and TK (Figure 23). Sequences for all the Krebs cycle enzymes, with the sole exception of the succinyl-CoA synthetase were detected (Figure 24).

In this case transcripts from genes coding for the necessary enzymes for the synthesis (xanthine dehydrogenase) and degradation of uric acid (urate oxidase, allantoinase and allantoicase) were not found.

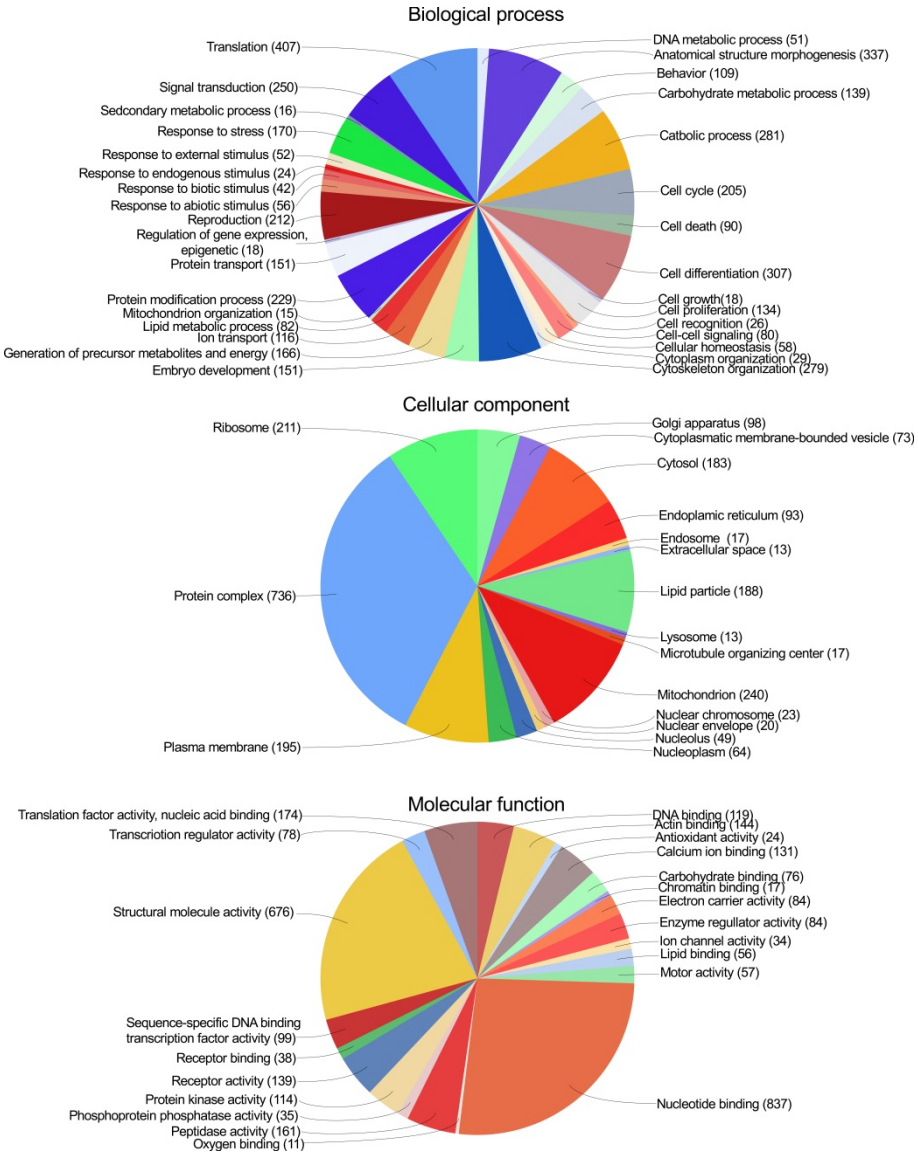

Figure 26. Epithelial multilevel pie chart of sequence distribution according to GO classification for the three main GO categories

2.3 Comparative tissue metabolism

2.3.1 Amino acid metabolism

As mentioned above, *Blattabacterium* possess the enzymes for synthetizing most of the essential amino acids, however needs an external supply for several non-essential amino acids, namely asparagine, glutamine, glycine and proline. The presence of gene expression for genes encoding the required enzymes for the synthesis of these amino acids was assessed in the three tissues. As shown in Table 20, the expression of the genes for all the enzymes needed for the synthesis of the above-mentioned amino acids has been detected only in the fat body.

Table 20. Presence (+) or absence (-) of several gene transcripts related to non-essential amino acid biosynthesis in the three analysed tissues.

Gene	EC	Fat body	Ovary	Epithelium
Asparagine biosynthesis				
Aspartate aminotransferase	2.6.1.1	+	+	+
Asparagine synthetase	6.3.5.4	+	+	-
Glutamine biosynthesis				
Glutamate dehydrogenase	1.4.1.3	+	+	+
Glutamine synthetase	6.3.1.2	+	+	+
Proline biosynthesis				
Glutamate-semialdehyde dh	2.7.2.11	+	-	-
Ornithine-δ-transaminase	2.6.1.3	+	+	-
Pyrroline-5-carboxylate reductase	1.5.1.2	+	+	-
Glycine biosynthesis				
Phophoglycerate dehydrogenase	1.1.1.95	+	+	+
Phosphoserine transaminase	2.6.1.52	+	+	-
Phosphoserine phosphatase	3.1.3.3	+	-	-
Serine hydroxymethyl transferase	2.1.2.1	+	+	+

2.3.2 Biosynthesis of heme groups

Blattabacterium metabolic network seems to require the input of porphobilinogen (González-Domenech et al. 2012). In the fat body there were present all but the sequences coding for two enzymes of the heme biosynthetic pathway, being the tissue were the pathway was more complete (Table 21).

Table 21. Presence (+) or absence (-) of transcripts for heme biosynthetic enzymes in the three analysed tissues.

Gene	EC	Fat body	Ovary	Epithelium
Aminolevulinic acid synthase	2.3.1.37	+	+	-
Porphobilinogen synthase	4.2.1.24	+	-	+
Porphobilinogen deaminase	2.5.1.61	+	+	+
Uroporphobilinogen III synthase	4.2.1.75	-	+	-
Uroporphobilinogen III decarboxylase	4.1.1.37	+	+	-
Coproporphyrinogen III oxidase	1.3.3.3	+	-	-
Protoporphyrinogen oxidase	1.3.3.4	-	-	-
Ferrochelatase	4.99.1.1	+	+	+

2.3.3 Uric acid metabolism

The process of nitrogen recycling in cockroaches involves the degradation of uric acid to urea, and later degradation of this metabolite by a *Blattabacterium* urease. We have previously postulated that the released ammonia by the endosymbiont would be used by the host encoded glutamine synthetase to produce glutamine (López-Sánchez et al. 2009; González-

Domenech et al. 2012). This could be the way to incorporate ammonia to the insect amino acid metabolism, probably also through the enzyme glutamate synthase. Transcripts for this enzyme have been identified in all three tissues. The genes involved in the synthesis and degradation of uric acid were annotated in the fat body transcriptome. The complete pathway was not found in the ovaries, while no transcripts were found in the epithelium (Table 22).

Table 22. Presence (+) or absence (-) of transcripts for uric acid metabolism enzymes, in the three analysed tissues.

Gene	EC	Fat body	Ovary	Epithelium
Xanthine dehydrogenase	1.17.1.4	+	-	-
Urate oxidase	1.7.3.3	+	+	-
Allantoinase	3.5.2.5	+	-	-
Allantoicase	3.5.3.4	+	+	-

2.4 Statistical analyses on the fat body and ovary transcriptomes

The libraries from fat body and ovaries were compared among them to find those functional terms enriched in the fat body, the tissues where *Blattabacterium* is supposed to carry out its metabolic function. A two-tailed Fisher test applying a multiple test correction was used. Only most specific terms were reviewed in this analysis. In Table 23, the over-represented ones at cellular process level are summarized. By contrast 31 terms were under-represented; most of them with nucleus or nucleic acid machinery related GO terms (Annex 7).

Table 23. Cellular component GO terms significatively enriched in the fat body respect to the ovaries (corrected p-value by False Discovery Rate [FDR] control)

GO identity	Term	FDR
GO:0005615	extracellular space	1.88E-005
GO:0070469	respiratory chain	9.02E-005
GO:0005788	endoplasmic reticulum lumen	0.00181
GO:0005753	mitochondrial proton-transporting ATP synthase complex	0.02774

Regarding biological processes GO category, fat body library was enriched for nine GO terms (Table 24). Most of them are related to energy production processes and the metabolism of energy sources. Terms related with the serine metabolism were also enriched in the fat body. By contrast 101 terms were overrepresented in the ovary (Annex 7).

Table 24. Biological processes GO terms enriched significatively in the fat body (corrected p-value by False Discovery Rate [FDR] control).

GO identity	Term	FDR
GO:0006869	lipid transport	1.85E-265
GO:0009303	rRNA transcription	5.40E-007
GO:0046113	nucleobase catabolic process	0.01373
GO:0006096	Glycolysis	0.01947
GO:0042775	mitochondrial ATP synthesis coupled electron transport	0.01952
GO:0009152	purine ribonucleotide biosynthetic process	0.02207
GO:0009070	serine family amino acid biosynthetic process	0.03867
GO:0006094	gluconeogenesis	0.04920
GO:0006563	L-serine metabolic process	0.04920

Finally, a total of 15 GO terms at molecular level were found over-represented in the fat body (Table 25). Among them those related to the lipid metabolism of the storage of nutrient reservoirs. Face the 15 terms in which the fat body is enriched respect the ovaries there are 35, which are over-represented, in the second tissue (Annex 7).

Table 25. Molecular processes GO terms enriched significatively in the fat body (corrected p-value by False Discovery Rate [FDR] control).

GO identity	Term	FDR
GO:0005319	lipid transporter activity	1.08E-270
GO:0005344	oxygen transporter activity	7.50E-035
GO:0045735	nutrient reservoir activity	1.19E-017
GO:0004497	monooxygenase activity	5.25E-013
GO:0003837	beta-ureidopropionase activity	2.97E-007
GO:0020037	heme binding	5.62E-005
GO:0009055	electron carrier activity	1.70E-004
GO:0004872	receptor activity	2.74E-004
GO:0004129	cytochrome-c oxidase activity	2.98E-004
GO:0008134	transcription factor binding	0.00296
GO:0009374	biotin binding	0.00626
GO:0000036	ACP phosphopantetheine attachment site binding involved in fatty acid biosynthetic process	0.03077
GO:0016885	ligase activity, forming carbon-carbon bonds	0.03453
GO:0005272	sodium channel activity	0.04214
GO:0008553	hydrogen-exporting ATPase activity, phosphorylative mechanism	0.04275

3.- Uric acid metabolism: the response to dietary nitrogen levels

To evaluate the response of *B. germanica* to the dietary protein levels, recent moult adult females (day 0) were separated from the colony and fed during two days with commercial dog food (approx. protein content 25%). Then, some animals were maintained fed in dog food as a control and others were changed to one of the three experimental diets, with different protein content: (i) dextrin (with 0% of protein); (ii) yeast extract with added 5% casein; (iii) 50% casein (see Material and Methods, Table 4, for diet composition). We, then, measured the differential expression of genes involved in the catabolism of uric acid (urate oxidase, allantoinase and allantoicase), as well as the *Blattabacterium* encoded urease, which degrades urea to CO_2 and ammonia. Besides, it was measured the expression of host glutamine synthetase since this enzyme would incorporate the released ammonia to metabolism. According to the model presented in the introduction (Figure 7), during periods of nitrogen scarcity, uric acid would be degraded to urea by the action of the three first enzymes, whereas the generated urea must enter into the endosymbiont cell where it would be degraded to CO_2 and ammonia. The released ammonia should be used by the host glutamine synthetase to produce glutamine from glutamate (Figure 27). Actine was used as a reference gene for all host genes and the *Blattabacterium* elongation factor EF-Tu for *Blattabacterium* urease. The gene expression was measured in ovary and fat body by qPCR.

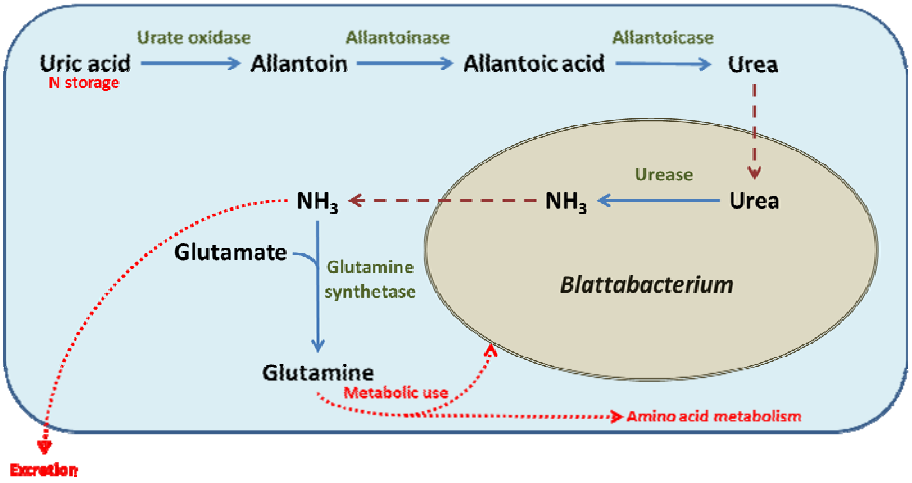

Figure 27. Proposed model for uric acid mobilisation. Uric acid is degraded to urea by the successive action of urate oxidase (an oxygen-dependent, peroxisomal enzyme), allantoinase and allantoicase. Then, urea is degraded to ammonia and carbon dioxide by the endosymbiotic urease. Finally, ammonia would be incorporated to glutamine through glutamine synthetase. The overflow of produced ammonia may be released by the system and explains the classic observation on the apparent ammonotelism in cockroaches (Mullins and Cochran 1972; 1976).

Among the five measured genes only statistical significative differential expression was detected in two of them. Table 26 shows the expression levels for each gene in those animals fed with the experimental diets respect to those fed with commercial dog food.

The first gene of the proposed pathway (urate oxidase) showed significant differential expression levels in both tissues when animals are fed with low level protein diets, respect to those kept on commercial dog food (Figure 28). In those animals fed with dextrin diet (without protein), urate oxidase was expressed a mean of 8.3 and 3.1 times more in fat body and ovary respect those animals fed with dog food. In animals fed with 5% protein diet urate oxidase expression was increased 6.9 and 2.5 times in fat body and ovaries respectively, when compared to control animals.

Table 26. Relative expression for uric acid metabolism genes. Relative expression mean value levels with its ranges in parenthesis are shown.

Genes	Fat body			Ovary		
	Dextrine	5%	50%	Dextrine	5%	50%
Urate oxidase	8.301 (3.66-17.51)	6.948 (3.37 -17.89)	1.745 (0.42 - 6.46)	3.124 (1.73 - 5.17)	2.549 (1.39 - 4.09)	1.630 (1.18 - 2.39)
Allantoinase	1.220 (0.70 – 3.00)	1.292 (0.28 - 5.60)	0.577 (0.04 - 4.56)	1.036 (0.49 - 3.14)	1.366 (0.50 - 2.76)	1.269 (0.49 - 3.33)
Allantoicase	1.033 (0.58 - 2.410)	1.223 (0.65 - 2.43)	1.879 (0.87 - 6.06)	1.256 (0.73 - 2.22)	1.200 (0.70 - 1.86)	1.256 (0.72 - 1.99)
Urease	0.691 (0.31 - 1.26)	0.839 (0.50 - 1.54)	0.725 (0.43 - 1.24)	1.562 (1.11 - 2.42)	1.028 (0.60 - 2.46)	1.950 (1.04 - 4.85)
Glutamine synthetase	2.271 (1.17 - 7.37)	1.794 (0.74 - 4.66)	1.069 (0.29 - 4.62)	0.961 (0.60 - 1.61)	0.364 (0.17 - 1.25)	0.448 (0.33 - 0.78)

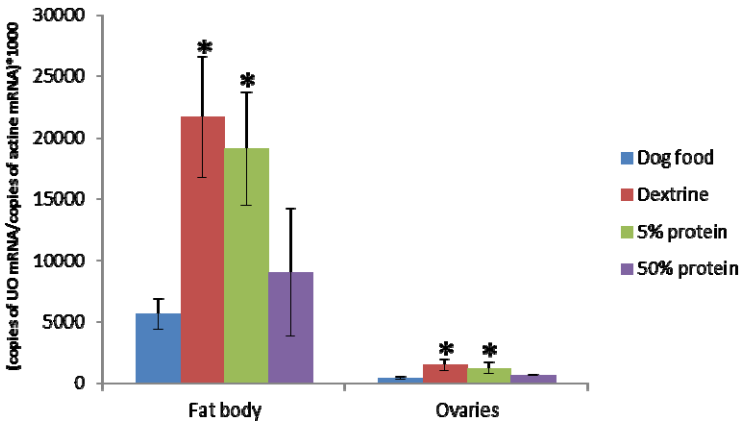

Figure 28. Expression pattern of urate oxidase (UO) gene in the fat body and ovary under the different diets expressed in copies of urate oxidase mRNA per copies of actin mRNA per 1000. Asterisk represents statistically significant differences respect to control ($p<0.05$).

For allantoinase and allantoicase, the other two necessary enzymes for degrading uric acid to urea, even if there were detected small increases in the expression in both low protein diets respect to control, these were not statistically significant (Figure 29). Neither was the case for the urease encoded in *Blattabacterium* (Figure 30).

Finally, it was found a significant differential expression of glutamine synthetase in fat body on dextrine fed animals, with expresion levels 2.3 times higher when compared with control-fed animals. Additionally, this enzyme seem to be significatively down-regulated in the ovaries of the animals fed with protein rich diets (Figure 31)

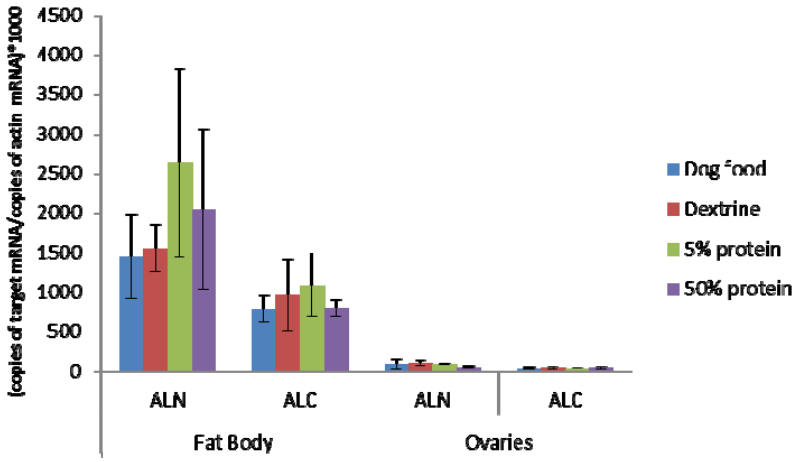

Figure 29. Expression pattern for allantoinase (ALN) and allantoicase (ALC) in the fat body and ovaries under the different diets expressed as in Figure 28.

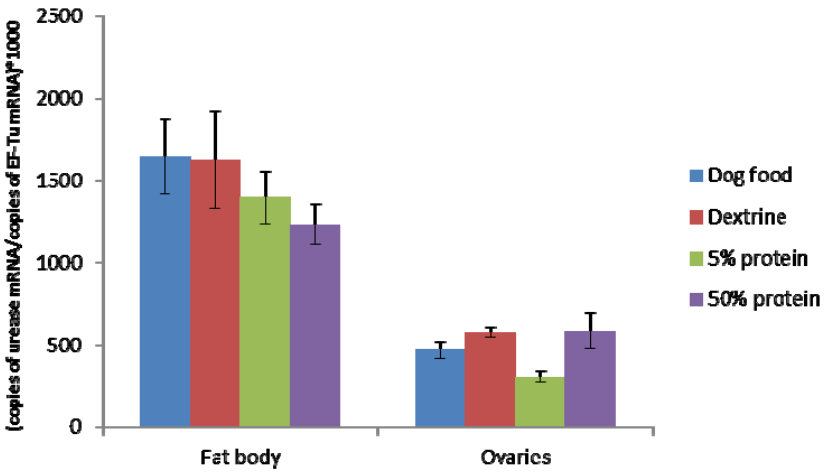

Figure 30. Expression pattern for urease in the fat body and ovaries under the different diets expressed as in Figure 28**.**

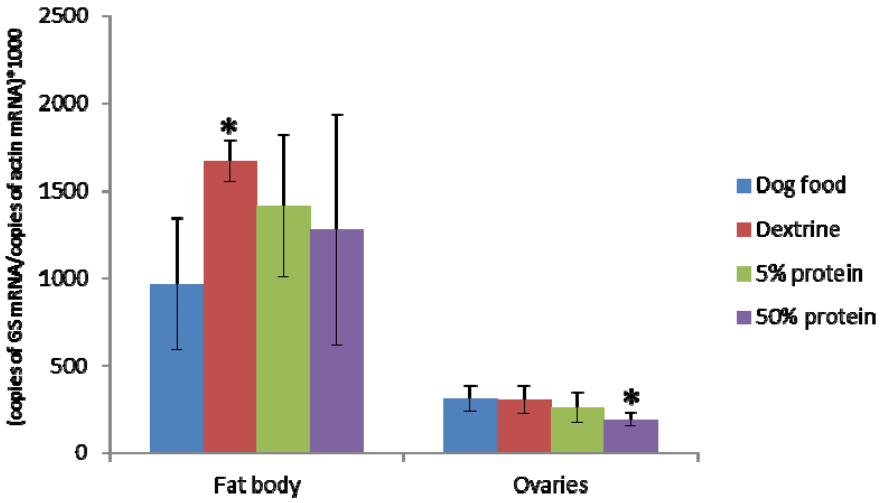

Figure 31. Expression pattern of the glutamine synthetase (GS) in the fat body and ovaries under the different diets expressed as in Figure 27. Asterisk represents statistically differences respect control ($p<0.05$).

V- Discussion

Blattabacterium ***comparative genomics: the frozen legacy of an ancient endosymbiont genome***

Cockroaches and *Blattabacterium* are evolving together for at least 140 Mya, since the common ancestor of cockroaches and termites was infected by an ancestral member of the Flavobacteriaceae family (Lo et al. 2003; Inward et al. 2007). Actually, all the extant cockroaches, with the sole exception of species from the genus *Nocticola*, live together with *Blattabacterium* (Lo et al. 2007). Conversely, *Blattabacterium* has been lost in the Termitidae, with the sole exception of the basal termite *M. darwiniensis* (Bandi et al. 1995). Many of the insects that harbour intracellular mutualistic bacterial symbionts are characterized by feeding on unbalanced food sources. In these cases the insect take advantage of the metabolic capabilities of the bacterium to obtain the nutrients lacking in the diet (Moya et al. 2008; Baumann 2005). However, in cockroaches, which feed on a complex diet, the role of endosymbionts was less clear. Studies on aposymbiotic cockroaches suggested that *Blattabacterium* may participate in the biosynthesis of essential amino acids (Henry 1962), and in the metabolism of nitrogen (Cochran 1985). Most recently, the complete genome sequence of the symbionts from *B. germanica* (López-Sánchez et al. 2009) and *P. americana* (Sabree et al. 2009) as well as the FBA performed over their inferred metabolic networks (González-Domenech et al. 2012), reinforced the thesis that this endosymbiont plays a role in the host nitrogen metabolism. Nowadays, the genome sequences for six different *Blattabacterium* strains are available, the two above mentioned plus the strains from *B. gigantenus* (Huang et al. 2012), *B. orientalis* (Patiño-Navarrete et al. 2013), the wood-feeding roach *C. punctulatus* (Neef et al. 2011), and the wood-feeding termite *M. darwiniensis* (Sabree et al. 2012). All these strains possess typical features showed by other insect

endosymbionts as a reduced genome size, ranging from 591 kb to 640 kb, low GC content (between 23.8% and 28.2%) and gene order maintenance.

The genome of BBor presented in this work has confirmed the extreme stability of the genome architecture in *Blattabacterium*, despite that all strains are evolving separately for more than 140 Mya. Only three chromosomal inversions have been identified but the synteny is highly maintained. Other remarkable feature is that all six strains maintain an extremely similar gene content with up to 82% of the genes from the pan-genome represented in the core. If xylophage roaches are removed from the analysis this proportion raises to 93.4%. These values are similar to those observed in *Blochmannia* spp., the primary endosymbionts from *Camponotus* ants where the core represents the 93.5% of the pan-genome (Williams and Wernegreen 2010). However the divergence time among the three *Blochmannia* strains was estimated in 20 Mya (Degnan et al. 2004; Gómez-Valero et al. 2008), while the divergence among the cockroaches is dated, as stated before, in at least 140 Mya. When the comparison is made with similar ancient symbiotic association like the one formed by aphids and *B. aphidicola*, established between 86 and 164 Mya (Dohlen and Moran 2000), and removing from the analysis the *Buchnera* symbiotic strains that need a complementation with other bacteria (like the case of the cedar aphids, which need to be complemented with *S.symbiotica*), the core spans for only the 74% of the pan-genome, a much lower proportion than in cockroaches. These facts strongly suggest that massive gene losses occurred in *Blattabacterium* soon after the establishment of its new intracellular way of life, and quickly reached a kind of optimal genome size capable of fulfilling the host requirements.

As pointed above, these requirements are likely related to the role of *Blattabacterium* on its host nitrogen metabolism (González-Domenech et al. 2012). Actually, most of gene losses among omnivorous *Blattabacterium*

lineages affect only to peripheral metabolic activities, affecting for instance the sulfur source, or the loss in BBor and BPam of the genes coding for the three first enzymes of the Krebs cycle. It was proposed that at least two of these steps may be carried out by the 3-isopropylmalate dehydratase (coded by *leuCD*) and the 3-isopropilmalate dehydrogenase (*leuD*), those catalysed by aconitate hydratase (*acnA*) and the isocitrate dehydrogenase (*icd*) respectively (Sabree et al. 2010), but none other gene was proposed to complement the function of the citrate synthase (*gltA*). Nevertheless, data from FBA of the reconstructed network pointed out that these steps could be dispensable, even if the complementation could not be ruled out. Even if these three steps are absent, the network of BPam is still functional in terms of biomass production, provided that the supply of glutamate to the network of BPam results in similar performances of the BBge network (González-Domenech et al. 2012).

The presence of the urease allows a metabolic network capable of catabolizing nitrogenous compounds yielding ammonia as a final waste product (López-Sánchez et al. 2009; González-Domenech et al. 2012). The presence of urease also can give a biochemical explanation to the classical hypothesis stating that urate deposits found in the fat body act as a nitrogen storage (Mullins 1974) as well as to the intriguing ammonotelism of cockroaches (Mullins and Cochran 1972; González-Domenech et al. 2012). Besides, the presence of urease genes in the core genome suggests that the role of *Blattabacterium* in the nitrogen recycling is ancestral and conserved among all *Blattabacterium* strains. The ammonia released by the endosymbionts that cannot be incorporated to the metabolism would be the one detected by Cochran and Mullins (1972) in their studies about waste nitrogen excretion in cockroaches

The strains from wood-feeding hosts accumulate the major number of gene loss events. These events affect the biosynthesis of both, essential and

non-essential amino acids, and according to the phylogenetic reconstruction, the loss of amino acid biosynthetic capacities took place very early in their evolutionary history, before the split between the families Termitidae and Cryptocercidae. These loses contrast with the situation of the omnivorous cockroaches, where the endosymbionts are able to supply its host with the whole set of essential amino acids. It has been postulated that in wood-feeding roaches the supply of amino acids must be compensated by the gut microbiota (Neef et al. 2011; Sabree et al. 2012). Additionally, in BMda and BCpu, the gene *argH* has been lost, which codes for the enzyme argininosuccinate lyase (ASL). ASL catalyses the last step of the arginine biosynthesis and seems probable that a host-encoded ASL may participate in the arginine biosynthesis. By BLAST analysis, we have identified genes coding for ASL in eight out of ten insect genomes present in the KEGG database. Actually, ASL gene was only absent in the aphid *A. pisum,* albeit it is present in its primary endosymbiont (*B. aphidicola*) genome, and in the human body louse, *P. humanus*, where the arginine could be acquired from the diet (blood). A recent work on metabolic modelling studies supported by transcriptomic and proteomic data (Macdonald et al. 2012) suggest that the action of host enzymes at the end of amino acid biosynthesis in the pea aphid may exhibit regulatory functions.

During the evolution of the diverse *Blattabacterium* strains there are a number of convergent metabolic traits. The most remarkable is related to sulfur metabolism. Among the six strains, only the symbionts BBge and BMda possess the genes coding for all the necessary enzymes for the assimilation of sulfate, with the sole exception of a 5'-phosphosulfate kinase (encoded by *cysC*). Despite the lack of *cysC*, this pathway must be operative, at least in BBge, since experimental data in aposymbiotic individuals of *B. germanica* indicated the inability to incorporate sulfate into cysteine and methionine (Block and Henry 1962). As this pathway is absent in the other

strains, and according to the phylogenetic reconstruction, it was most likely lost in at least three independent events, once in the lineage leading to *Cryptocercus*, other in the lineage leading to BBgi and the last one during the Blattidae species evolution. In BPam, *cysI* is absent and both *cysN* and *cysD* are found as pseudogenes, whereas in BBor, these two genes are completely lost while *cysI* is inactivated. Respect to the remaining genes in the pathway, *cysH* and *cysG* are found as pseudogenes in BBor, but in BPam they seem to be functional. Finally, the gene *cysJ* is present in all sequenced *Blattabacterium* strains, even in those that are not able to assimilate sulfate. The product of this gene is a flavoprotein that together with the product of *cysI* forms sulfite reductase. In this case, *CysJ* may have been recruited by other processes like the regeneration of reduced ribonucleotide reductase (Covès et al. 1993), or as an electron relay (Siegel and Davis 1974). Seven duplicated genes were found in the genomes of BBor, BCpu and BMda (Neef et al. 2011; Sabree et al. 2012), while BPam and BBgi possess eight duplicated genes (Sabree et al. 2009; Huang et al. 2012). Finally in BBge there are nine duplicated genes (López-Sánchez et al. 2009). The presence of such genes in a reduced genome like that of *Blattabacterium* is surprising. The fact that most of these genes are present in all six strains (for five out of nine genes both copies are intact in the six genomes) points to a possible functional role in the bacterium physiology and could be considered as ecoparalogs (Sanchez-Perez et al. 2008).

Similar to what is found in other long-term symbiotic association, hosts and their primary endosymbionts show co-cladogenesis when phylogenetic reconstruction of cockroaches and *Blattabacterium* are compared (Lo et al. 2003; lópez-Sanches et al. 2008). The phylogeny obtained during this thesis shares the same topology than the one obtained for the cockroaches by Inward and collaborators (2007). The presence of *Blattabacterium* genes evolving under a molecular clock has allowed the determination of split

times between the host species. The divergence between Blaberoidea cockroaches and the Termitidae and Crypotcercidae lineages was calculated close to 90 Mya, whereas the divergence between BPam and BBor was stablished around 12 Mya. Summarizing, even if the hosts harbouring *Blattabacterium* are evolving separately for such a long-time, the genomic and metabolic architectures of their endosymbionts have remained strikingly stable. Thus, the basic genetic and metabolic features of *Blattabacterium* were established in a short period of time since these bacteria infected the common ancestor of cockroaches and termites.

The shared role of endosymbiont and host in nitrogen metabolism as revealed by transcriptomic analyses

For most of the insect endosymbionts sequenced up today, the genome sequence of the host has not been still sequenced. Nowadays, there are only the sequences of the pea aphid, *A. pisum* host of *B. aphidicola* (Eisen et al. 2010), the human body louse, *P. humanus* host of *Riesia* sp. (Kirkness et al. 2010) and from *C. floridanus* host of *B. floridanus* (Bonasio et al. 2010). The availability of the host genome sequence has allowed the characterization of the metabolic interdependence network between *Buchnera* and its host in the aphid bacteriocytes (Hansen and Moran 2011). The lack of genomic information of the host severely impairs the understanding of other systems. The characterization of the transcriptome has been proved to be a good tool to overcome the lack of genomic data in a non-model organism (Vera et al. 2008). Given that the genomic information on *B. germanica* was limited, it was decided that the first step to gain insight into the relationships and shared metabolism between *Blattabacterium* and its host was the characterization of the cockroach transcriptome. We chose three tissues, two harbouring the endosymbiont, fat body and ovary, and a third one free of the

endosymbiont, the epithelial tissue localized underneath the pronotum of 6 stage nymphs.

The cDNA synthesis was done using an oligo-dT, avoiding the tRNA and rRNA and enriching the sample in mRNAs. By this method it should be expected that mRNAs from *Blattabacterium* should be also discarded, since bacterial mRNA lack of polyA queue. However, the genome of *Blattabacterium* is extremely AT rich, thus a good amount of messengers from the endosymbiont has been swept during the cDNA synthesis procedure. Up to 6.5% of the reads from the fat body library belong to *Blattabacterium*. Conversely, in the ovaries, where the endosymbiont is also present, this proportion is much lower, and hardly a hundred of reads had their origin in the bacteria, probably because in this stage *Blattabacterium* is metabolically less active and/or the bacterial population size per cell is lower.

Close to 50% of the sequences from the fat body and ovary libraries show significant matches in the BLASTX searches against the non-redundant NCBI peptide database, this proportion is reduced to 27.4% in the epithelium library. These results might be due to the fact that the closest relative to cockroaches with a sequenced genome is *Tribolium castaneum*. Actually most of the best BLAST hits match *T. castaneum* sequences. Further, the proportion of properly annotated sequences is also quite low: around 30% for the fat body and ovary libraries and 15% for the transcriptome coming from the epithelium are sequences with assigned GO terms. On average, each annotated sequence was assigned to 5.7 different GO terms. Albeit these proportions may seem to be low, they are in concordance with the observations in other transcriptomic analyses performed in non-model organism with no close relatives sequenced (Vera et al. 2008; Meyer et al. 2009; Coppe et al. 2010).

Although the precise coverage of the whole transcriptome is difficult to estimate without the full genomic sequence, it has been possible a broad estimation analysing the expression of genes related to several central metabolic pathways present in all animal phyla. The selected pathways were the glycolysis/gluconeogenesis, the pentose phosphate pathway and the Krebs cycle. In the fat body and ovaries, all pathways are well represented, and almost equally distributed between the two tissues. We failed to detect transcripts for three out of 11 genes in the glycolysis/gluconeogenesis pathways; however it should be taken into account that glycolytic and gluconeogenic enzymes may not be expressed at the same time. Also, there have not been identified transcripts for one out of seven enzymes involved in the pentose phosphate pathway. The Krebs cycle is complete in the ovary but in the fat body there is one gene below the detection level, the one coding for aconitase. In the epithelium library, these pathways are much less complete, lacking three genes from the glycolysis/gluconeogenesis, three more in the phosphate pentose pathway and one in the Krebs cycle. Overall, we can conclude that there is a good representation of the transcriptome at least in ovaries and fat body tissues.

Statistical analyses confirmed that the library coming from the fat body tissue is enriched in GO terms, in all three main categories, related mainly to energy production and nutrient storage and mobilization. As expected for metabolic active cells, the terms related to lipid and sugar catabolism are enriched in this tissue.

During both, the metabolic reconstruction (López-Sánchez et al. 2009) and the FBA of the genome-scale metabolic network (González-Domenech et al. 2012), it has been shown that *Blattabacterium* is auxotroph for some non-essential amino acids, namely: proline, asparagine, glutamine and glycine (Figure 32). In *Buchnera* the three first amino acids should also be supplied externally as well as glutamate, tyrosine, alanine and serine (Figure

32). *Blochmannia* strains, that like *Blattabacterium* are symbionts from insects with a complex diet such as ants, rely on a host supply of serine, proline, asparagine, alanine and arginine (Figure 32). In addition, *B. vafer* do not encode for a glutamine synthase, so like *Blattabacterium* is host dependent for glutamine. Among these instances, asparagine and proline need to be supplied in all three cases. It is worth mentioning that non-essential amino acids are especially abundant in arthropod haemolymph; for example in the aphid *A. pisum*, alanine, glutamine, proline and asparagine represent the 50% of the total free amino acid content in the haemolymph (Macdonald et al. 2012). Proline, glycine and tyrosine are the most abundant amino acids in the haemolymph of both *P. americana* and *Blaberus discoidalis* (Stevens 1961; Osborne and Neuhoff 1974; Sowa and Keeley 1996). In mosquito's haemolymph, the most represented amino acids are alanine, proline, glutamine and glutamate (Pennington et al. 2003). The abundance of proline seems a common feature among insect species. For example, this amino acid is used as energy source during flight in mosquitoes (Scaraffia et al. 2003) and beetles (Gäde and Auerswald 2002). In addition, in the mosquito *A. aegypti* proline has been also proposed as nitrogen sink (Pennigton et al. 2003). In arthropods other than insects, like spiders, glutamine is the most represented amino acid, followed by far by proline and glycine (Tillinghast and Townley 2008). Manual searches among the annotation and the KEGG maps retrieved for each tissue confirmed that the genes required for the metabolic synthesis of asparagine, glutamine, proline and glycine are expressed in the host fat body. These pathways seem incomplete in the ovaries and epithelial tissues. The host supply of non-essential amino acids has been postulated as a possible mechanism of host control on the metabolism of its symbionts. Thus, the FBA analysis performed on the metabolic network of *Blattabacterium* suggested that cockroaches could use the supply of glutamine to control their

endosymbionts metabolic behaviour or growth rate (González-Domenech et al. 2012). Similarly, the obligate aerobic metabolism of the endosymbiont could be a target for host control mechanisms considering the sensitivity of metabolic functionality to the oxygen availability (González-Domenech et al. 2012). Also, *A. pisum* can control the amount of essential amino acids produced by the endosymbiont by controlling the amount of metabolic precursors supplied to *Buchnera*. Some of these precursors are non-essential amino acids like aspartate, required for the synthesis of lysine and threonine or glutamate, a precursor in the synthesis of arginine (Macdonald et al. 2011). Interestingly, both, amino acid and oxygen supply have been described as control factors of plant hosts over the population of nitrogen fixing bacteria (Prell et al. 2009; Kiers et al. 2003). In *Rhizobium*-legume symbioses, a symbiotic auxotrophy for essential branched amino acids has been described (Prell et al. 2010).

Other host compound required by *Blattabacterium* metabolism is porphobilinogen. In the fat body, transcripts of the genes for all enzymes involved in the synthesis of this compound have been found, with the exception of uroporphorbirinogen III synthase, which is encoded in the endosymbiont, and protoporphyrinogen oxidase. This second one is not encoded in the *Blattabacterium* genome, however it has been empirically demonstrated that under aerobic conditions the reaction catalysed by this enzyme can occur spontaneously (Sasarman et al. 1979). Nevertheless, not detecting transcripts of the above-mentioned genes does not mean that the corresponding genes are not present in the *B. germanica* genome.

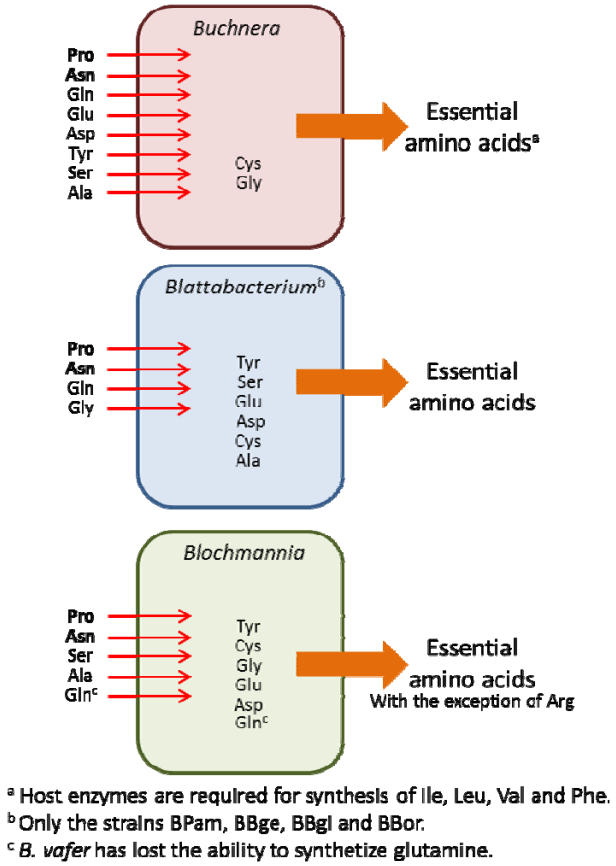

Figure 32. Non-essential amino acid auxotrophies shown by three different insect endosymbionts. Inside the area representing each symbiont, the non-essential amino acids that can be synthetized are listed.

One of the most interesting points of these analyses, was to establish the pathway of uric acid recycling in cockroaches (Cochran 1985). In the fat body there have been identified transcripts for the three enzymes necessary for degrading uric acid (uricase, allantoinase, allantoicase), as well as for the uric acid synthesis enzyme xanthine dehydrogenase. In the ovaries, we only have detected urate oxidase and allantoicase transcripts, while in the epithelium we were not able to find any transcript for these genes. This pathway is only present in other three of the insect genomes represented in

the KEGG database: the mosquitoes *Anopheles gambiae* and *Aedes aegypti* and the Jewell wasp, *Nasonia vitripennis*. The presence of these enzymes, together with *Blattabacterium* urease and a host encoded glutamine synthetase, complete a pathway for uric acid recycling in a similar way as the one described in the shield bug, *Parastrachia japonensis* (Kashima et al. 2006), but with one remarkable difference. While in the shield bug, the whole pathway is encoded by *Erwinia*-like bacteria, an ectosymbiont located in the midgut (Hosokawa et al. 2010), in *B. germanica* the pathway for the nitrogen recycling is shared between the endosymbiont and its host. It is worth mentioning that in higher termites, where the *Blattabacterium* endosymbiont has been lost, this metabolic process is carried out by the gut microbiota (Potrikus and Breznak 1981). It could be interesting to further study the transition from a metabolism shared between cockroach fat body cells and the bacterial endosymbiont to a gut localized nitrogen metabolism performed by ectosymbionts, especially in putative transitional species like the wood feeding cockroach *C. punctulatus* and the endosymbiont harbouring termite *M. darwiniensis*. Although these metabolic studies have not been yet performed, the gut microbiota composition in *M. darwiniense*, *C. punctulatus* and *P. americana* was studied using the V6-V9 region of the bacterial 16S rRNA gene (Sabree et al. 2011). They found that *M. darwiniensis* and *C. punctulatus* gut microbiome components were similar in both species and very constant between individuals. Moreover, they included specialized termite gut-associated bacteria that were previously postulated that collaborate in fixing nitrogen, degrading lignocellulose, and producing nutrients (Warnecke et al. 2007). On the contrary, *P. americana* gut microbiota was very variable among individuals and with a large proportion of sequences that were most closely related to environmental sequences and not to symbionts represented in current databases (Sabree et al. 2011).

Classical studies on the cockroach *P. americana* describe how the fat body deposits of urates raise when individuals of this species are fed with a protein rich diet. Conversely, when animals previously kept on the protein rich diet where changed to a diet with a low protein content, the amount of uric acid stored in the fat body extremely decreased (Cochran et al. 1979). These observations suggested that uric acid was actually a reservoir of nitrogen to be mobilized on periods of protein scarcity. The identification of sequences coding for the necessary enzymes for uric acid recycling prompted us to investigate how the expression of these genes are affected in response to the dietary protein levels. After feeding females of *B. germanica* with diets containing 0 or 5% of protein, an increase in the number of transcripts for urate oxidase gene (the first step of the catabolic pathway) was observed, as well as in the glutamine synthetase gene expression, the last step of the pathway of nitrogen mobilization. Actually, in several experimental systems, glutamine synthetase has been described as the main controlling step for nitrogen metabolism (Urich 1990), an observation coherent with the central role played by glutamine as nitrogen donor. Thus, when dietary nitrogen sources were scarce, the expression of genes encoding for urate oxidase and glutamine synthetase increased. However, in the present work it remains unexplained whether there are additional posttranslational regulatory mechanisms, especially in those steps where we were unable to detect changes in the expression levels. Further functional studies, e.g., both proteomic and metabolomic approaches, focusing on possible post-transcriptional regulatory mechanisms are needed to establish a model of how cockroaches cope with severe changes in nitrogen intake.

Besides the supply of essential amino acids and cofactors, the metabolism of nitrogenous waste products seems to be one of the main roles of insect mutualistic bacteria. Actually, while some strains of *Blattabacterium* had lost its ability to synthetize essential amino acids (Neef et al. 2011; Sabree et

al. 2012) (Figure 17), the genes for urease and for all the genes needed for the urea cycle (with the sole exception of the gene for the ASL, lost in BCpu and BMda) are maintained in all sequenced strains (Patiño-Navarrete et al. 2013). The pathway proposed for *Blattabacterium* and the cockroaches depicted in the Figure 33A, has also been proposed with some modifications for other insects. Thus the shield bug, *P. japonensis*, during its diapause period also recycles nitrogenous waste products via the uricotelic pathway. In this case, all three enzymes are encoded in an *Erwinia*-like bacteria located in the gastric cecum of the insect (Kashima et al. 2006). The same solution has also been adopted by the brown planthopper (*Nilaparvata lugens*), but in this case yeast-like endosymbionts are the responsible of degrading the uric acid, at least the step catalysed by uricase (Sasaki et al. 1996). However, the urease activity has not been found in any of the two examples. In a similar way than in cockroaches, *Blochmannia*, endosymbionts of *Camponotus* ants, also participate in the nitrogen recycling via urease and glutamine synthetase, but in this case, both enzymes are encoded in the bacterial genome (Figure 33B) (Gil et al. 2003; Degnan et al. 2005; Feldhaar et al. 2007). However, the *Blochmannia* strain from *C. vafer* has lost the gene coding for glutamine synthetase, and as a consequence ammonia should be incorporated to amino acid metabolism by a host encoded glutamine synthetase, as it is the case in the cockroaches (Williams and Wernegreen 2010). The silkworm nitrogen recycling system is extremely similar to the one we have proposed for the cockroach-*Blattabacterium* association. In this insect, genes for all three uricolytic enzymes (uricase, allantoinase, allantoicase) have been annotated but urease is not encoded by any microbial symbiont, rather it is obtained from the mulberry leaves that the caterpillar eats (Hirayama et al. 1999; Hirayama et al. 2000). *Buchnera* cells in *A. pisum* are also an obligate partner for nitrogen recycling. Initially it was proposed that ammonia may be incorporated to

amino acid metabolism through glutamine synthetase and glutamate synthase, and then, both glutamine and glutamate, would be used by *Buchnera* to synthetize essential amino acids (Hansen and Moran 2011). However, FBA and functional (i.e. transcriptomic and proteomic) analyses carried out on the integrated aphid bacteriocyte and *Buchnera,* suggest that it is the host cell the one that produces ammonia, and afterwards this compound is incorporated into glutamine and glutamate through the activities of glutamine synthetase and glutamate synthase by the so-called GS/GOGAT (GS, glutamine synthetase; GOGAT, glutamate synthase) cycle. Most of the ammonia assimilated through the GS/GOGAT cycle is later incorporated to essential amino acids, especially those whose synthesis is shared between *Buchnera* and *A. pisum* (Figure 33C) (Macdonald et al. 2012).

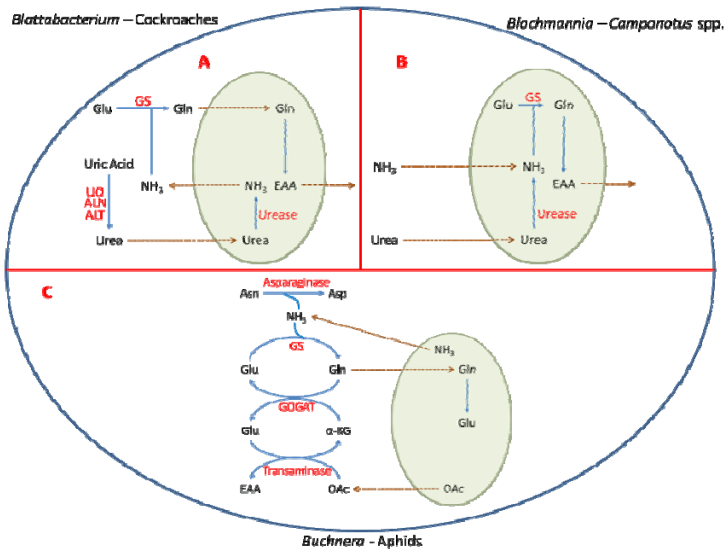

Figure 33. Nitrogen recycling models for different insect-endosymbiotic bacteria systems based on the different schemes (modified from Macdonald et al. 2012). **A** Cockroach-*Blattabacterium*. **B** Ant-*Blochmannia*, with the exception of the system in *C. vafer*. **C** Aphid-*Buchnera*. GS, glutamine synthase; UO, urate oxidase; ALN, allantoinase; ALT, allantoicase; GOGAT, glutamate synthase, EAA, essential amino acid; OAc, oxoacid.

Future studies on the cellular localization and quantification of enzymes involved in the metabolism nitrogen, as well as the measurement of activities could help to clarify how the system works. Alongside with these researches, metabolomic studies focused on the products related to nitrogen metabolism must be also valuable. Finally, no *Blattabacterium* strains from the basal family Polyphagidae are sequenced up today. Since the cockroaches of the genus *Nocticola* (Lo et al. 2008), the ones that have lost *Blattabacterium,* belong to this family, the availability of the genome sequence from a *Blattabacterium* strain from one species of this family can give us insights about the process underlying the loss of the endosymbiont in *Nocticola*.

The studies during this thesis have been useful to describe the evolutionary history of *Blattabacterium* since the divergence of the extant cockroach families. The observations obtained depict *Blattabacterium* as a buffered system with no significant variation in the metabolic capabilities of the different strains during more than 140 Mya of parallel evolution. In addition, the functional studies carried out help us to unravel the metabolic machinery behind the classic observations, suggesting that cockroaches use uric acid, normally a waste product in terrestrial animals, as a source of nitrogen.

VI- Conclusions

1- The genomic characteristics of *Blattabacterium*, primary endosymbiont from *B. orientalis*, are virtually equivalent to the main features of the other sequenced strains, especially those from symbiotic bacteria from omnivorous cockroaches.

2- Despite being evolving separately for more than 140 Mya, the genomic architecture of *Blattabacterium* shows an extreme stability, with only few rearrangements.

3- Gene content in *Blattabacterium* strains has also been maintained along its evolutionary history, especially among omnivorous strains. The proportion of core genes respect to the pan-genome is similar to species like *Blochmannia* spp., which are evolving separately for a shorter period than *Blattabacterium*.

4- The architecture of the central metabolic network has been maintained across the evolutionary history of *Blattabacterium*. Most of gene loss events affect poorly characterized genes, or genes involved in peripheral metabolic activities.

5- The gene content of the LCSA is not significantly different from its closest strain, in terms of gene content, the one from *B. germanica*. Thus it can be postulated that the reduction process occurred soon after the establishment of the symbiotic relationship.

6- The symbiotic strains from the wood-feeding hosts have lost the capacity of synthetizing essential amino acids. This loss may have occurred soon after the split among the branches leading to the Blattidae family members and the branch leading to the Crypotcercidae and Termitidae families, and before the split between these last two families.

7- The presence of urease genes in the core genome strengthens the hypothesis that relates *Blattabacterium* to the host urate and nitrogen metabolism.

8- Sequences encoding for the necessary enzymes involved in the synthesis of those non-essential amino acids that should be supplied to the bacterium have been identified in the host fat body.

9- Transcripts for urate oxidase, allantoinase and allantoicase, the required enzymatic repertoire for the catabolism of uric acid to urea, have been identified in the fat body, as well as the corresponding to xanthine dehydrogenase, involved in uric acid biosynthesis.

10- Genes for urate oxidase and glutamine synthetase are over-expressed when *B. germanica* is deprived of a protein source. This observation is consistent with our hypothesis that cockroaches are mobilizing in their fat body the urate reserves to be used as nitrogen source.

VII References

Akman L, Yamashita A, Watanabe H, Oshima K, Shiba T, Hattori M and Aksoy S. 2002. Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. Nat Genet. 32:402–7.

Aksoy S. 2000. Tsetse--A haven for microorganisms. Parasitol Today. 16:114–8.

Altschul S, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–402.

Arrese E, Soulages J. 2010. Insect fat body: energy, metabolism, and regulation. Annu Rev Entomol. 55:207–25.

Ausubel FM. 1999. Short Protocols in Molecular Biology: a Compendium of Methods from Current Protocols in Molecular biology. John Wiley & Sons: New York.

Bandi C, Damiani G, Magrassi L, Grigolo A. 1994. Flavobacteria as intracellular symbionts in cockroaches. Proc Biol Sci. 257:43–8.

Bandi C, Sironi M, Damiani G, Magrassi L, Nalepa CA, Laudani U, Sacchi L. 1995. The establishment of intracellular symbiosis in an ancestor of cockroaches and termites. Proc Biol Sci. 268:293–9.

Baumann P. 2005. Biology bacteriocyte-associated endosymbionts of plant sap-sucking insects. Annu Rev Microbiol. 59:155–89.

Belda E, Moya A, Bentley S, Silva F. 2010. Mobile genetic element proliferation and gene inactivation impact over the genome structure and metabolic capabilities of *Sodalis glossinidius*, the secondary endosymbiont of tsetse flies. BMC Genomics. 11:449.

Belda E, Silva FJ, Peretó J, Moya A. 2012. Metabolic networks of *Sodalis glossinidius*: a systems biology approach to reductive evolution. PLoS One. 7:e30652.

Bell W, Roth L, Nalepa CA. 2007. Cockroaches. Ecology, Behavior, and Natural History. The Johns Hopkins University Press: Baltimore.

Beneden P. 1876. Animal Parasites and Messmates. Appleton: New York.

Blochmann F. 1884. Über die metamorphose der kerne in den ovarialeiern und über den beginn der blastodermbildung bei den ameisen. l. naturhist. Med. 3.

Blochmann F. 1887. Über das regelmässige vorkommen von backterienähnlichen gebilden in den geweben und eiern versichiedener Insekten. Z Biol. 24:6.

Block R, Henry S. 1962. Metabolism of the sulphur amino-acids and of sulphate in *Blattella germanica*. Nature. 191:392–3.

Blütghen N, Brand K, Cajavec B, Swat M, Herzel H, Beule D. 2005. Biological profiling of gene groups utilizing gene ontology. Genome Inf. 16:106-15.

Bonasio R, Zhang G, Ye C, Mutti NS, Fang X, Qin N, Donahue G, Yang P, Li Q, Li C, Zhang P, Huang Z, Berger SL, Reinberg D, Wang J, Liebig J. 2010. Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. Science. 329:1068-71.

Brooks M, Richards A. 1955. Intracellular symbiosis in cockroaches. I. Production of aposymbiotic cockroaches. Biol Bull. 109:22–39.

Brooks M. 1970. Comments on the classification of intracellular symbiotes cockroaches and a description of the species. J Invertebr Pathol. 16:249–58.

Buchner P. 1965. Endosymbiosis of Animals with Plant Microorgansims. John Wiley & Sons: New York.

Burke G, Moran NA. 2011. Massive genomic decay in *Serratia symbiotica*, a recently evolved symbiont of aphids. Genome Biol Evol. 3:195–208.

Carver T, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhil J. 2005. ACT: the artemis comparison tool. Bioinformatics. 21:3422–3.

Cazalet C, Rusniok C, Brüggemann H, Zidane N, Magnier A, Ma L, Tichit M, Jarraud S, Bouchier C, Vandenesch F, Kunst F, Etienne J, Glaser P, Buchrieser C. 2004. Evidence in the *Legionella pneumophila* genome for exploitation of host cell functions and high genome plasticity. Nat Genet. 36:1165–73.

Chen DQ, Montllor CB, Purcell AH. 2000. Fitness effects of two facultative endosymbiotic bacteria on the pea aphid, *Acyrthosiphon pisum*, and the blue alfalfa aphid, *A. kondoi*. Entomol Exp Appl. 95:315–23.

Chen F, Mackey AJ, Stoeckert CJ Jr, Roos DS. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res. 34:D363–8.

Chevreux B, Wetter T, Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information. In: Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB). Heidelberg, Germany pp. 45–6.

Cochran D. 1985. Nitrogen excretion in cockroaches. Annu Rev Entomol. 30:29–49.

Cochran DG, Mullins DE, Mullins KJ. 1979. Cytological changes in the fat body of the american cockroach, *Periplaneta americana*, in relation to dietary nitrogen levels. Ann Entomol Soc Am. 72:197–205.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 21:3674–6.

Coppe A, Pujolar JM, Maes GE, Larsen PF, Hansen MM, Bernatchez L, Zane L, Bortoluzzi S. 2010. Sequencing, de novo annotation and analysis of the first *Anguilla anguilla* transcriptome: EelBase opens new perspectives for the study of the critically endangered European eel. BMC genomics. 11:635.

Covès J, Nivière V, Eschenbrenner M, Fontecave M. 1993. NADPH-sulfite reductase from *Escherichia coli*. A flavin reductase participating in the generation of the free radical of ribonucleotide reductase. J Biol Chem. 268:18604–9.

Dadd RH. 1985. Nutrition: organisms. In: Comprehensive Insect Physiology, Biochemistry and Pharmacology. Kerkut GA & Gilbert LI, editors Vol 4. Pergamon Press: Oxford pp. 313–390.

Dale C, Welburn SC. 2001. The endosymbionts of tsetse flies: manipulating host-parasite interactions. Int J Parasitol. 31:628–31.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. 27:1164–5.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 9:772.

Dasch GA, Weiss E, Chang K. 1984. Endosymbionts of insects. In: Bergeys Manual of Systematic Bacteriology. Krieg NR, editor. Vol. 1 Williams & Williams: Baltimore pp. 881–883.

Davidson DW, Cook SC, Snelling RR, Chua TH. 2003. Explaining the abundance of ants in lowland tropical rainforest canopies. Science. 300:969–72.

de Duve C. 2007. The origin of eukaryotes: a reapparisal. Nat Rev Genet. 8:395-405.

de Souza DJ, Bézier A, Depoix D, Drezen J-M, Lenoir A. 2009. *Blochmannia* endosymbionts improve colony growth and immune defence in the ant *Camponotus fellah*. BMC Microbiol. 9:29.

Degnan PH, Lazarus AB, Brock CD, Wernegreen JJ. 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 PH, Lazarus AB, Wernegreen JJ. 2005. Genome sequence of *Blochmannia pennsylvanicus* indicates parallel evolutionary trends among bacterial mutualists of insects. Genome Res. 15:1023–33.

Degnan PH, Moran NA. 2008. Diverse phage-encoded toxins in a protective insect endosymbiont. Appl Environ Microbiol. 74:6782–91.

Degnan PH, Yu Y, Sisneros N, Wing RA, Moran NA. 2009. *Hamiltonella defensa*, genome evolution of protective bacterial endosymbiont from pathogenic ancestors. Proc Natl Acad Sci USA. 106:9063–8.

Degnan PH, Leonardo TE, Cass BN, Hurwitz B, Stern D, Gibbs RA, Richards S, Moran NA. 2010. Dynamics of genome evolution in facultative symbionts of aphids. Environ Microbiol. 12:2060–9.

Degnan PH, Ochman H, Moran NA. 2011. Sequence conservation and functional constraint on intergenic spacers in reduced genomes of the obligate symbiont *Buchnera*. PLoS Genet. 7:e1002252.

Delcher A, Bratke K, Powers E, Salzberg S. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics. 23:673–9.

Dinant S, Bonnemain JL, Girousse C, Kehr J. 2010. Phloem sap intricacy and interplay with aphid feeding. C R Biol. 333:504–15.

Dohlen CD, Moran NA. 2000. Molecular data support a rapid radiation of aphids in the Cretaceous and multiple origins of host alternation. Biol J Linnean Soc. 71:689–717.

Dohlen CD, Kohler S, Alsop ST, McManus WR. 2001. Mealybug beta-proteobacterial endosymbionts contain gamma-proteobacterial symbionts. Nature. 412:433–6.

Douglas AE. 1993. The nutritional quality of phloem sap utilized by natural aphid populations. Ecol Entomol. 18:31–8.

Douglas AE. 1994. Symbiotic Interactions. Oxford University Press: Oxford.

Douglas AE. 1998. Insect-microbial symbioses: aphids and their symbiotic bacteria *Buchnera*. Annu Rev Entomol. 43:17–37.

Douglas AE. 2009. The microbial dimension in insect nutritional ecology. Funct Ecol. 23:38–47.

Eisen JA, plus 338 authors. 2010. Genome sequence of the pea aphid *Acyrthosiphon pisum*. PLoS Biol. 8:e1000313.

Engel MS, Grimaldi DA. 2004. New light shed on the oldest insect. Nature. 427:627–30.

Fares MA, Moya A, Barrio E. 2004. GroEL and the maintenance of bacterial endosymbiosis. Trends Genet. 20:413-6.

Feldhaar H, Straka J, Krischke M, Berthold K, Stoll S, Mueller MJ, Gross R. 2007. Nutritional upgrading for omnivorous carpenter ants by the endosymbiont *Blochmannia*. BMC Biol. 5:48.

Ferrari J, Darby AC, Daniell TJ, Godfray HCJ, Douglas AE. 2004. Linking the bacterial community in pea aphids with host-plant use and natural enemy resistance. Ecol Entomol. 29:60–5.

Fukatsu T, Nikoh N. 1998. Two intracellular symbiotic bacteria from the mulberry psyllid *Anomoneura mori* (Insecta, Homoptera). Appl Envir Microbiol. 64:3599–606.

Gäde G, Auerswald L. 2002. Beetles' choice--proline for energy output: control by AKHs. Comp Biochem Physiol B Biochem Mol Biol. 132:117-29.

Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe DL, Eddy SR, Bateman A. 2011. Rfam: Wikipedia, clans and the "decimal" release. Nucleic Acids Res. 39:D141–5.

Gil R, Sabater-Muñoz B, Latorre A, Silva FJ, Moya A. 2002. Extreme genome reduction in *Buchnera* spp.: toward the minimal genome needed for symbiotic life. Proc Natl Acad Sci USA. 99:4454–8.

Gil R, Silva FJ, Zientz E, Delmotte F, González-Candelas F, Latorre A, Rausell C, Kamerbeek J, Gadau J, Hölldobler B, van Ham RC, Gross R, Moya A. 2003. The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. Proc Natl Acad Sci USA. 100:9388–93.

Gil R, Belda E, Gosalbes MJ, Delaye L, Vallier A, Vincent-Monégat C, Heddi A, Silva FJ, Moya A, Latorre A. 2008. Massive presence of insertion sequences in the genome of SOPE, the primary endosymbiont of the rice weevil *Sitophilus oryzae*. Int Microbiol. 11:41-8.

Gómez-Valero L, Soriano-Navarro M, Pérez-Brocal V, Heddi A, Moya A, García-Verdugo JM, Latorre A. 2004. Coexistence of *Wolbachia* with *Buchnera aphidicola* and a secondary symbiont in the aphid *Cinara cedri*. J Bacteriol. 186:6626–33.

Gómez-Valero L, Latorre A, Gil R, Gadau J, Feldhaar H, Silva FJ. 2008. Patterns and rates of nucleotide substitution, insertion and deletion in the endosymbiont of ants *Blochmannia floridanus*. Mol Ecol. 17:4382–92.

González-Domenech CM, Belda E, Patiño-Navarrete R, Moya A, Peretó J, Latorre A. 2012. Metabolic stasis in an ancient symbiosis: genome-scale metabolic networks from two *Blattabacterium cuenoti* strains, primary endosymbionts of cockroaches. BMC Microbiol. 12:S5.

Gosalbes MJ, Lamelas A, Moya A, Latorre A. 2008. The striking case of tryptophan provision in the cedar aphid *Cinara cedri.* J Bacteriol. 190:6026–9.

Graf J, Kikuchi Y, Rio RV. 2006. Leeches and their microbiota: naturally simple symbiosis models. Trends Microbiol. 14:365–71.

Gray M, Spencer D. 1996. Organellar evolution. In: Evolution of Microbial Life. Roberts D, Sharp P, Altderson G, Collins M, editors. Cambridge University Press: Cambrige pp. 109–126.

Guthrie D, Tindall A. 1968. The Biology of the Cockroach. Edward Arnold (Publishers) Ltd.: London.

Guy L, Kultima JR, Andersson SG. 2010. genoPlotR: comparative gene and genome visualization in R. Bioinformatics. 26:2334–5.

Hall-Stoodley L, Costerton JW, Stoodley P. 2004. Bacterial biofilms: from the natural environment to infectious diseases. Nat Rev Microbiol. 2:95–108.

Hansen AK, Moran NA. 2011. Aphid genome expression reveals host-symbiont cooperation in the production of amino acids. Proc Nat Acad Sci U S A. 108:2849–54.

Harrison C, Douglas AE, Dixon F. 1989. A rapid method to isolate symbiotic bacteria from aphids. J Invertebr Pathol. 53:427–8.

Haydak M. 1953. Influence of the protein level of the diet on the longevity of cockroaches. Ann Entomol Soc Am. 46:547–60.

Henry S. 1962. The importance of microorganisms in the nutrition of insects. Trans N Y Acad Sci 24:676–83.

Hirayama C, Sugimura, Shinbo H. 1999. Recycling of urea associated with the host plant urease in the silkworm larvae, *Bombyx mori*. J Insect Physiol. 45:15-20.

Hirayama C, Sugimura, Saito H, Nakamura M. 2000. Host plant urease in the haemolymph of the silkworm, *Bombyx mori*. J Insect Physiol. 46:1415-21.

Hollande AC, Favre R. 1931. La structure cytologique de *Blattabacterium cuénoti* (Mercier N. G.), symbiote du tissu adipeux des Blattides. C R Soc Biol. 107:752–4.

Hosokawa T, Kikuchi Y, Nikoh N, Meng XY, Hironaka M, Fukatsu T. 2010. Phylogenetic position and peculiar genetic traits of a midgut bacterial symbiont of the stinkbug *Parastrachia japonensis*. Appl Environ Microbiol. 76:4130–5.

Huang CY, Sabree ZL, Moran NA. 2012. Genome sequence of *Blattabacterium* sp. strain BGIGA, endosymbiont of the *Blaberus giganteus* cockroach. J Bacteriol. 194:4450-1.

Inward D, Beccaloni G, Eggleton P. 2007. Death of an order: a comprehensive molecular phylogenetic study confirms that termites are eusocial cockroaches. Biol Lett. 3:331–5.

Kambhampati S. 2010. Blattabacteriaceae. In: Bergey's Manual of Systematic Bacteriology. Krieg NR, Staley JT, Brown DR, Hedlund BP, Paster BJ, Ward NL, Ludwig W and Whitman WB, editors Vol. 4: Baltimore pp. 315–321.

Kashima T, Nakamura T, Tojo S. 2006. Uric acid recycling in the shield bug, *Parastrachia japonensis* (Hemiptera: Parastrachiidae), during diapause. J Insect Physiol. 52:816–25.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33:511–8.

Kiers ET, Rousseau RA, West SA, Denison RF. 2003. Host sactions and the legume-*Rhizobium* mutualism. Nature. 425:78-81

Kirkness EF, Haas BJ, Sun W, Braig HR, Perotti MA, Clark JM, Lee SH, Robertson HM, Kennedy RC, Elhaik E, Gerlach D, Kriventseva EV, Elsik CG, Graur D, Hill CA, Veenstra JA, Walenz B, Tubío JM, Ribeiro JM, Rozas J, Johnston JS, Reese JT, Popadic A, Tojo M, Raoult D, Reed DL, Tomoyasu Y, Kraus E, Mittapalli O, Margam VM, Li HM, Meyer JM, Johnson RM, Romero-Severson J, Vanzee JP, Alvarez-Ponce D, Vieira FG, Aguadé M, Guirao-Rico S, Anzola JM, Yoon KS, Strycharz JP, Unger MF, Christley S, Lobo NF, Seufferheld MJ, Wang N, Dasch GA, Struchiner CJ, Madey G, Hannick LI, Bidwell S, Joardar V, Caler E, Shao R, Barker SC, Cameron S, Bruggner RV, Regier A, Johnson J, Viswanathan L, Utterback TR, Sutton GG, Lawson D, Waterhouse RM, Venter JC, Strausberg RL, Berenbaum MR, Collins FH, Zdobnov EM, Pittendrigh BR. 2010. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. Proc Natl Acad Sci USA. 107:12168–73.

Kneip C, Lockhart P, Voss C, Maier UG. 2007. Nitrogen fixation in eukaryotes--new models for symbiosis. BMC Evol Biol. 7:55.

Koga R, Tsuchida T, Fukatsu T. 2009. Quenching autofluorescence of insect tissues for in situ detection of endosymbionts. Appl Entomol Zool. 44:281-91.

Kristensen NP. 1981. Phylogeny of insect orders. Annu Rev Entomol. 26:135–57.

Lamelas A, Pérez-Brocal V, Gómez-Valero L, Gosalbes MJ, Moya A, Latorre A. 2008. Evolution of the secondari symbiont "*Candidatus* Serratia symbiotica" in aphid species of the subfamily Lachninae. Appl Environ Microbiol. 74:4236-40.

Lamelas A, Gosalbes MJ, Manzano-Marín A, Peretó J, Moya A, Latorre A. 2011a. Serratia symbiotica from the aphid *Cinara cedri*: a missing

link from facultative to obligate insect endosymbiont. PLoS Genet. 7:e1002357.

Lamelas A, Gosalbes MJ, Moya A, Latorre A. 2011b. New clues about the evolutionary history of metabolic losses in bacterial endosymbionts, provided by the genome of *Buchnera aphidicola* from the aphid *Cinara tujafilina*. Appl Environ Microbiol. 77:4446–54.

Latorre A, Gil R, Silva FJ, Moya A. 2005. Chromosomal stasis versus plasmid plasticity in aphid endosymbiont *Buchnera aphidicola*. Heredity. 95:339–47.

Latorre A, Durbán A, Moya A, Peretó J. 2011. The role of symbiosis in eukaryotic evolution. In: Origins and Evoution of Life- an astrobiological perspective. Gargaud P, López-Garcia P and Martinin H, editors. Cambridge University Press. Cambridge: pp. 326-43.

Lind PA, Andersson DI. 2008. Whole-genome mutational biases in bacteria. Proc Natl Acad Sci USA. 105:17878–83.

Lo N, Bandi C, Watanabe H, Nalepa C, Beninati T. 2003. Evidence for cocladogenesis between diverse dictyopteran lineages and their intracellular endosymbionts. Mol Biol Evol. 20:907–13.

Lo N, Beninati T, Stone F, Walker J, Sacchi L. 2007. Cockroaches that lack *Blattabacterium* endosymbionts: the phylogenetically divergent genus *Nocticola*. Biol Letters. 3:327–30.

López-Madrigal S, Latorre A, Porcar M, Moya A, Gil R. 2011. Complete genome sequence of "*Candidatus* Tremblaya princeps" strain PCVAL, an intriguing translational machine below the living-cell status. J Bacteriol. 193:5587–8.

López-Sánchez MJ, Neef A, Patiño-Navarrete R, Navarro L, Jiménez R, Latorre A, Moya A. 2008. Blattabacteria, the endosymbionts of cockroaches, have small genome sizes and high genome copy numbers. Environ Microbiol. 10:3417–22.

López-Sánchez MJ, Neef A, Peretó J, Patiño-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.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: A Program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25:0955–64.

Ludwig D, Gallagher MR. 1966. Vitamin synthesis by the symbionts in the fat body of the cockroach, *Periplaneta americana* (L.). J N Y Entomol soc. 74:134–9.

Macdonald SJ, Thomas GH, Douglas AE. 2011. Genetic and metabolic determinants of nutritional phenotype in an insect-bacterial symbioses. Mol Ecol. 20:2073-84.

Macdonald SJ, Lin GG, Russell CW, Thomas GH, Douglas AE. 2012. The central role of the host cell in symbiotic nitrogen metabolism. Proc Biol Sci. 279:2965–73.

Malke H, Schwartz W. 1966. Untersuchungen uber die symbiose von tieren mit pilsen und bakterein XII. Die bedeuntung der blattiden symbiose. Z Allg Mikribiol. 6:34–68.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 437:376–80.

Margulis L. 1993 Symbiosis in Cell Evolution (2nd ed.). W. H. Freeman & Co: New York.

Martin W, Müller M. 1998. The hydrogen hypothesis for the first eukaryote. Nature. 392:37–41.

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

McCutcheon JP, Moran NA. 2007. Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. Proc Natl Acad Sci USA. 104:19392–7.

McCutcheon JP, McDonald BR, Moran NA. 2009a. Convergent evolution of metabolic roles in bacterial co-symbionts of insects. Proc Natl Acad Sci USA. 106: 15394-9.

McCutcheon JP, McDonald BR, Moran NA. 2009b. Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. PLoS Genet. 5:e1000565.

McCutcheon JP, Moran NA. 2010. Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. Genome Biol Evol. 2:708–18.

McCutcheon JP, Dohlen CD. 2011. An interdependent metabolic patchwork in the nested symbiosis of mealybugs. Curr Biol. 21:1366–72.

McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol. 10:13–26.

McEnroe W, Forgash A. 1957. The in vivo incorporation of C^{14} formate in the groups ureide of uric acid by *Periplaneta americana* (L.). Ann Entomol Soc Am. 50:429–31.

Mcfall-Ngai M. 2008. Are biologist in "future shock?" Symbiosis integrates biology across domains. Nat Rev Microbiol. 6:789–92.

McGraw EA, O'Neill SL. 2004. *Wolbachia pipientis*: intracellular infection and pathogenesis in *Drosophila*. Curr Opin Microbiol. 7:67–70.

Mercier L. 1906. Les corps bactéroïdes de la Blatte (*Periplaneta orientalis*): *Bacillus cuenoti* (n. sp. L. Mercier). (Note préliminaire). C R Soc Biol. 61:682–4.

Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, Willis BL, Matz MV. 2009. Sequencing and de novo

analysis of a coral larval transcriptome using 454 GSFlx. BMC Genomics. 10:219.

Minic Z, Hervé G. 2004. Biochemical and enzymological aspects of the symbiosis between the deep-sea tubeworm *Riftia pachyptila* and its bacterial endosymbiont. Eur J Biochem. 271:3093–102.

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B. 2011. How many species are there on earth and in the ocean? PLoS Biol. 9: e1001127

Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci USA. 93:2873–8.

Moran NA, Plague GR. 2004. Genomic changes following host restriction in bacteria. Curr Opin Genet Dev. 14:627–33.

Moran NA, Tran P, Gerardo NM. 2005. Symbiosis and insect diversification : an ancient symbiont of sap-feeding insects from the bacterial phylum bacteroidetes. Appl Environ Microbiol. 71:8802–10.

Moreira D, Lopez-Garcia P. 1998. Symbiosis between methanogenic archaea and delta-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. J Mol Evol. 47:517–30.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35:W182–5.

Moya A, Peretó J, Gil R, Latorre A. 2008. Learning how to live together: genomic insights into prokaryote-animal symbioses. Nat Rev Genet. 9:218–29.

Moya A, Peretó J. 2011. Simbiosis, Seres que Evolucionan Juntos. Síntesis: Madrid

Mullins DE. 1974. Nitrogen metabolism in the American cockroach: an examination of whole body ammonium and other cations excreted in relation to water requirements. J Biol Exp. 61:541–56.

Mullins DE, Cochran DG. 1972. Nitrogen excretion in cockroaches: uric acid is not a major product. Science. 177:699–701.

Mullins DE, Cochran DG. 1974. Nitrogen metabolism in the American cockroach: an examination of whole body and fat body regulation of cations in response to nitrogen balance. J Exp Biol. 61:557–70.

Mullins DE, Cochran DG. 1975a. Nitrogen metabolism in the american cockroach—I. An examination of positive nitrogen balance with respect to uric acid stores. Comp Biochem Physiol A Comp Physiol. 50:489–500.

Mullins DE, Cochran DG. 1975b. Nitrogen metabolism in the american cockroach—II. An examination of negative nitrogen balance with respect to mobilization of uric acid stores. Comp Biochem Physiol A Comp Physiol. 50:501–10.

Mullins DE, Cochran DG. 1976. A comparative study of nitrogen excretion in twenty-three cockroach species. Comp Biochem Physiol A Comp Physiol. 53:393–9.

Murphy MJ, Siegel LM. 1973. Siroheme and sirohydrochlorin the basis for a new type of porphyrin-related prosthetic group common to both assimilatory and dissimilatory sulfite reductases. J Biol Chem. 248:6911–9.

Murray MG, Thompson WF. 1980. Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res. 8:4321–5.

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

Nawrocki EP, Kolbe DL, Eddy SR. 2009. Infernal 1.0: inference of RNA alignments. Bioinformatics. 25:1335–7.

Needham BJ. 1937. Contributions of chemical physiology to the problem of reversibility in evolution. Biol Rev. 13:225–51.

Neef A, Latorre A, Peretó J, Silva FJ, Pignatelli M, Moya A. 2011. Genome economization in the endosymbiont of the wood roach *Cryptocercus punctulatus* due to drastic loss of amino acid synthesis capabilities. Genome Biol Evol. 3:1437–48.

Newton IL, Woyke T, Auchtung TA, Dilly GF, Dutton RJ, Fisher MC, Fontanez KM, Lau E, Stewart FJ, Richardson PM, Barry KW, Saunders E, Detter JC, Wu D, Eisen JA, Cavanaugh CM. 2007. The *Calyptogena magnifica* chemoautotrophic symbiont genome. Science. 315:998–1000.

Nogge G. 1976. Sterility in tsetse flies (*Glossina morsitans* Westwood) caused by loss of symbionts. Experientia. 32:995–6.

Nogge G. 1981. Significances of symbionts for the maintenance of an optimal nutritional state for successful reproduction in hematophagous arthropods. Parasitology. 82:101–4.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MH, Wagner H 2012. vegan: Community Ecology Package. http://cran.r-project.org/package=vegan.

Oliver KM, Russell JA, Moran NA, Hunter MS. 2003. Facultative bacterial symbionts in aphids confer resistance to parasitic wasps. Proc Natl Acad Sci USA. 100:1803–7.

Oliver KM, Moran NA, Hunter MS. 2005. Variation in resistance to parasitism in aphids is due to symbionts not host genotype. Proc Natl Acad Sci USA. 102:12795–800.

Oliver KM, Campos J, Moran NA, Hunter MS. 2008. Population dynamics of defensive symbionts in aphids. Proc Biol Sci. 275:293–9.

Osborne NN, Neuhoff V. 1974. Amino acid and serotonin content in the nervous system, muscle and blood of the cockroach *Periplaneta americana*. Brain Res. 80:251-64.

Paracer S, Vernon A. 2000. Symbiosis. An Introduction to Biological Associations. 2nd ed. Oxford University Press: NY.

Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, Harris DE, Holden MT, Churcher CM, Bentley SD, Mungall KL, Cerdeño-Tárraga AM, Temple L, James K, Harris B, Quail MA, Achtman M, Atkin R, Baker S, Basham D, Bason N, Cherevach I, Chillingworth T, Collins M, Cronin A, Davis P, Doggett J, Feltwell T, Goble A, Hamlin N, Hauser H, Holroyd S, Jagels K, Leather S, Moule S,

Norberczak H, O'Neil S, Ormond D, Price C, Rabbinowitsch E, Rutter S, Sanders M, Saunders D, Seeger K, Sharp S, Simmonds M, Skelton J, Squares R, Squares S, Stevens K, Unwin L, Whitehead S, Barrell BG, Maskell DJ. 2003. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat Genet. 35:32–40.

Pennigton JE, Goldstrohm DA, Wells MA. 2003. The role of haemolymph proline as nitrogen sink during blood meal digestion by mosquito *Aedes aegypti*. J Insect Physiol. 49:115-21.

Pérez-Brocal V, Gil R, Ramos S, Lamelas A, Postigo M, Michelena JM, Silva FJ, Moya A, Latorre A. 2006. A small microbial genome: the end of a long symbiotic relationship? Science. 314:312–3.

Pfaffl MW, Horgan GW, Dempfle L. 2002. Relative expression software tool (REST) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. Nucleic Acids Res. 30:e36.

Pfeiffer T, Sánchez-Valdenebro I, Nuño JC, Montero F, Schuster S. 1999. METATOOL: for studying metabolic networks. Bioinformatics. 15:251–7.

Piceis PP, Sacchi L, Grigolo A, Laudani U. 1986. Fine structure of the fat body and its bacteroids in *Blattella germanica* (Blattodea). Acta Zoologica. 67:63–71.

Potrikus CJ, Breznak JA. 1981. Gut bacteria recycle uric acid nitrogen in termites: a strategy for nutrient conservation. Proc Natl Acad Sci USA. 78:4601–5.

Prell J, White JP, Bourdès A, Bunnewell S, Bomngaerts RJ, Poole PS. 2009. Legumes regulate *Rhizobium* bacteroid development and persistence by the supply of branched-chain amino acids. Proc Natl Acad Sci USA. 106:12477-82.

Prell J, Bourdès A, Kumar S, Lodwig E, Hosie A, Kinghorn S, White J, Poole P. 2010. Role of symbiotic auxotrophy in the *Rhizobium*-legume symbioses. PLoS One. 5:e13933.

Prosser WA, Douglas AE. 1992. A test of the hypotheses that nitrogen is upgraded and recycled in an aphid (*Acyrthosiphon pisum*) symbiosis. J Insect Physiol. 38:93–9.

Rendulic S, Jagtap P, Rosinus A, Eppinger M, Baar C, Lanz C, Keller H, Lambert C, Evans KJ, Goesmann A, Meyer F, Sockett RE, Schuster SC. 2004. A predator unmasked: life cycle of *Bdellovibrio bacteriovorus* from a genomic perspective. Science. 303:689–92.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16:276–7.

Rispe C, Delmotte F, Van Ham RC, Moya A. 2004. Mutational and selective pressures on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids. Genome Res. 14:44–53.

Robinson WH. 2005. Blattaria. In: Urban Insects and Arachnids. A Handbook of Urban Entomology. Cambridge University Press: New York pp. 35–64.

Rocha EP, Danchin A. 2002. Base composition bias might result from competition for metabolic resources. Trends Genet. 18:291–4.

Ronaghi M, Uhlén M, Nyrén P 1998. DNA sequencing: a sequencing method based on real-time pyrophosphate. Science. 281:363–5.

Russell JA, Latorre A, Sabater-Munoz B, Moya A, Moran NA. 2003. Side-stepping secondary symbionts: widespread horizontal transfer across and beyond the Aphidoidea. Mol Ecol. 12:1061–75.

Russell JA, Moran NA. 2006. Costs and benefits of symbiont infection in aphids: variation among symbionts and across temperatures. Proc Biol Sci. 273:603–10.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. 2000. Artemis: sequence visualization and annotation. Bioinformatics. 16:944–5.

Sabree ZL, Kambhampati S, Moran NA. 2009. Nitrogen recycling and nutritional provisioning by *Blattabacterium*, the cockroach endosymbiont. Proc Natl Acad Sci USA. 106:19521–6.

Sabree ZL, Degnan PH, Moran NA. 2010. Chromosome stability and gene loss in cockroach endosymbionts. Appl Environ Microbiol. 76:4076–9.

Sabree ZL, Huang CY, Arakawa G, Tokuda G, Lo N, Watanabe H, Moran NA. 2012. Genome shrinkage and loss of nutrient-providing potential in the obligate symbiont of the primitive termite *Mastotermes darwiniensis*. Appl Environ Microbiol. 78:204–10.

Sacchi L, Grigolo A. 1989. Endocytobiosis in *Blattella germanica* L. (Blattodea): recent acquisitions. Endocyt Cell Res 6:121–47.

Sacchi L, Corona S, Grigolo A, Laudani U, Selmi MG, Bigliardi E. 1996. The fate of the endocytobionts of *Blattella germanica* (Blattaria: Blattellidae) and *Periplaneta americana* (Blattaria: Blattidae) during embryo development. Ital J Zool. 63:1–11.

Saffo M. 1992. Coming to terms with a field: words and concepts in symbiosis. Symbiosis. 14:17–31.

Sanchez-Perez G, Mira A, Nyiro G, Pasić L, Rodriguez-Valera F. 2008. Adapting to environmental changes using specialized paralogs. Trends Genet. 24:154–8.

Sandström J, Pettersson J. 1994. Amino acid composition of phloem sap and the relation to intraspecific variation in pea aphid (*Acyrthosiphon pisum*) performance. J Insect Physiol. 40:947–55.

Sandstrom JP, Russell Jacob A, White JP, Moran NA. 2001. Independent origins and horizontal transfer of bacterial symbionts of aphids. Mol Ecol. 10:217–28.

Sasarman A, Chartrand P, Lavoie M, Tardif D, Proschek R, Lapointe C. 1979. Mapping of a new heme gene in *Escherichia coli* K12. J Gen Microbiol. 113:297-303.

Sasaki T, Kawamura M, Ishikawa H. 1996. Nitrogen recycling in the brown planthooper, *Nilaparvata lugens*: involvement of yeast-like endosymbionts in uric acid metabolism. J Insect Physiol. 42:125-29.

Sauer C, Dudaczek D, Holldobler B, Gross R. 2002. Tissue localization of the endosymbiotic bacterium "*Candidatus* Blochmannia floridanus"

in adults and larvae of the carpenter ant *Camponotus floridanus*. Appl Environ Microbiol. 68:4187–93.

Scaraffia PY, Wells MA. 2003. Proline can be utilized as an energy substrate during flight of *Aedes aegypti* females. J Insect Physiol. 49:591-601.

Schink B. 1997. Energetics of syntrophic cooperation in methanogenic degradation . Microbiol Mol Biol Rev. 61:262–80.

Schmitt S, Weisz JB, Lindquist N, Hentschel U. 2007. Vertical transmission of a phylogenetically complex microbial consortium in the viviparous sponge *Ircinia felix*. Appl Environ Microbiol. 73:2067–78.

Schmittgen TD, Livak KJ. 2008. Analyzing real-time PCR data by the comparative C(T) method. Nat Prot. 3:1101–8.

Sharp KH, Davidson SK, Haygood MG. 2007. Localization of "*Candidatus* Endobugula sertula" and the bryostatins throughout the life cycle of the bryozoan *Bugula neritina*. ISME J. 1:693–702.

Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H. 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. Nature. 407:81–6.

Short JE, Edwards JP. 1991. Reproductive and developmental biology of the oriental cockroach *Blatta orientalis* (dictyoptera). Med Vet Entomol. 5:385–94.

Siegel LM, Davis PS. 1974. Reduced nicotinamide adenine dinucleotide phosphate-sulfite reductase of enterobacteria. IV. The *Escherichia coli* hemoflavoprotein: subunit structure and dissociation into hemoprotein and flavoprotein components. J Biol Chem. 249:1587–98.

Silva FJ, Latorre A, Moya A. 2003. Why are the genomes of endosymbiotic bacteria so stable? Trends Genet. 19:176–80.

Slansky F, Rodriguez JG. 1987. Nutritional Ecology of Insects, Mites, Spiders, and Related Invertebrates. John Wiley & Sons: New York.

Sloan DB, Moran NA. 2012. Genome reduction and co-evolution between the primary and secondary bacterial symbionts of psyllids. Mol biol Evol. 29:3781-92.

Sowa SM, Keeley LL. 1996. Free amino acids in the haemolymph of the cockroach, *Blaberus discoidalis*. Comp Biochem Physiol. 113:131-4.

Staden R, Beal K, Bonfield J. 2000. The Staden package, 1998. Methods Mol Biol. 132:115–30.

Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics. 21:456–63.

Stewart FJ, Newton IL, Cavanaugh CM. 2005. Chemosynthetic endosymbioses: adaptations to oxic-anoxic interfaces. Trends Microbiol. 13:439–48.

Stevens TM. 1961. Free amino acids in the haemolymph of the american cockroach, *Periplaneta americana* L.. Comp Biochem Physiol. 3:304-9.

Stouthamer R, Breeuwer JA, Hurst GD. 1999. *Wolbachia pipientis*: microbial manipulator of arthropod reproduction. Annu Rev Microbiol. 53:71–102.

Swinton J. 2011. Vennerable. https://r-forge.r-project.org/projects/vennerable/

Tamames J, Gil R, Latorre A, Peretó J, Silva FJ, Moya A. 2007. The frontier between cell and organelle: genome analysis of *Candidatus* Carsonella ruddii. BMC Evol Biol. 7:181.

Tamas I, Klasson L, Canbäck B, Näslund AK, Eriksson AS, Wernegreen JJ, Sandström JP, Moran NA, Andersson SG. 2002. 50 Million years of genomic stasis in endosymbiotic bacteria. Science. 296:2376–9.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale

DA. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 4:41.

Taylor MJ, Bandi C, Hoerauf A. 2005. *Wolbachia* bacterial endosymbionts of filarial nematodes. Adv Parasitol. 60:245–84.

Thao ML, Moran NA, Abbot P, Brennan EB, Burckhardt DH, Baumann P. 2000. Cospeciation of psyllids and their primary prokaryotic endosymbionts. Appl Environ Microbiol 66:2898–905.

Thao ML, Gullan PJ, Baumann P. 2002. Secondary (gamma-Proteobacteria) endosymbionts infect the primary (beta-Proteobacteria) endosymbionts of mealybugs multiple times and coevolve with their hosts. Appl Environ Microbiol. 68:3190–7.

Tillinghast EK, Townley MA. 2008. Free amino acids in spider haemolymph. Comp Biochem Physiol B Biochem Mol Biol. 151:286-95

Toh H, Weiss BL, Perkin SA, 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. 16:149–56.

Tripp HJ, Hewson I, Boyarsky S, Stuart JM, Zehr JP. 2011. Misannotations of rRNA can now generate 90% false positive protein matches in metatranscriptomic studies. Nucleic Acids Res. 39:8792–802.

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

Urich K. 1994. Small nitrogenous compounds. In: Comparative Animal Biochemistry. Springer-Verlag: Berlin pp 403-54.

Valovage WD, Brooks MA. 1979. Uric acid quantities in the fat body of normal and aposymbiotic german cockroaches, *Blattella germanica*. Ann Entomol Soc Ann. 72:687–9.

van Ham RC, Kamerbeek J, Palacios C, Rausell C, Abascal F, Bastolla U, Fernández JM, Jiménez L, Postigo M, Silva FJ, Tamames J, Viguera E, Latorre A, Valencia A, Morán F, Moya A. 2003. Reductive

genome evolution in *Buchnera aphidicola*. Proc Nat Acad Sci U S A. 100:581–6.

Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH. 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol Ecol. 17:1636–47.

Vrsansky P, Vishniakova VN, Rasnitsyn AP. 2002. Order Blattida Latreille, 1810. The cockcraches. In: History of Insect Orders. Rasnitsyn AP, Quicke DLJ, editors. Kluwer Academic Publishers. pp. 263–270.

Warnecke F, Luginbühl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, Cayouette M, McHardy AC, Djordjevic G, Aboushadi N, Sorek R, Tringe SG, Podar M, Martin HG, Kunin V, Dalevi D, Madejska J, Kirton E, Platt D, Szeto E, Salamov A, Barry K, Mikhailova N, Kyrpides NC, Matson EG, Ottesen EA, Zhang X, Hernández M, Murillo C, Acosta LG, Rigoutsos I, Tamayo G, Green BD, Chang C, Rubin EM, Mathur EJ, Robertson DE, Hugenholtz P, Leadbetter JR. 2007. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature. 450:560-5.

Warnes G. 2011. gplots: Various R programming tools for plotting data. http://cran.r-project.org/web/packages/gplots/index.html

Wernegreen JJ. 2005. For better or worse: genomic consequences of intracellular mutualism and parasitism. Curr Opin Genet Dev. 15:572–83.

Wesson LG. 1932. A modification of the osbornemendel salt mixture containing only inorganic constituents. Science. 75:339–40.

Williams LE, Wernegreen JJ. 2010. Unprecedented loss of ammonia assimilation capability in a urease-encoding bacterial mutualist. BMC Genomics. 11:687.

Wolschin F, Ho B, Gross R, Zientz E. 2004. Replication of the endosymbiotic bacterium *Blochmannia floridanus* is correlated with the developmental and reproductive stages of its ant host. Appl Environ Microbiol. 70:4096–102.

Worning P, Jensen LJ, Hallin PF, Staerfeldt HH, Ussery DW. 2006. Origin of replication in circular prokaryotic chromosomes. Environ Microbiol. 8:353–61.

Woyke T, Tighe D, Mavromatis K, Clum A, Copeland A, Schackwitz W, Lapidus A, Wu D, McCutcheon JP, McDonald BR, Moran NA, Bristow J, Cheng JF. 2010. One bacterial cell, one complete genome. PLoS One. 5:e10314.

Wu D, Daugherty SC, Van Aken SE, Pai GH, Watkins KL, Khouri H, Tallon LJ, Zaborsky JM, Dunbar HE, Tran PL, Moran NA, Eisen JA. 2006. Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. PLoS Biol. 4:e188.

Yang Z. 1994. Estimating the pattern of nucleotide substitution. J Mol Evol. 39:105–11.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–91.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 17:32–43.

Zientz E, Dandekar T, Gross R. 2004. Metabolic interdependence of obligate intracellular bacteria and their insect hosts metabolic interdependence of obligate intracellular bacteria and their insect hosts. Microbiol Mol Biol Rev. 68:745–70.

Zientz E, Beyaert I, Gross R, Feldhaar H. 2006. Relevance of the endosymbiosis of *Blochmannia floridanus* and carpenter ants at different stages of the life cycle of the host. Appl Environ Microbiol. 72:6027–33.

Zug R, Hammerstein P. 2012. Still a host of hosts for *Wolbachia*: analysis of recent data suggests that 40% of terrestrial arthropod species are infected. PLoS One. 7:e38544.

VIII Resum

Introducció

La simbiosi, entesa com la relació de interdependència entre dos o més individus de diferents espècies, ha tingut un profunda influència al llarg de l'evolució de la vida. De fet, hi ha un ampli consens respecte a la importància de les associacions simbiòtiques, en l'aparició i evolució primerenca de la cèl·lula eucariota, així com en l'evolució del sistema immunitari en eucariotes complexos o en l'adquisició de noves capacitats metabòliques en plantes i animals, permetent-los així explorar noves fonts d'aliments o ocupar nous nínxols ecològics. Segons l'efecte de la interacció sobre els membres de l'associació, es poden distingir tres tipus: **comensalisme**, quan un dels membres es beneficia sense afectar l'aptitud de l'altre (o altres); **parasitisme**, quan un dels participants treu un avantatge disminuint l'aptitud de l'altre; i finalment **mutualisme**, quan ambdós socis es beneficien de la interacció. A més, també es poden classificar segons la localització del simbiont respecte de l'hoste entre **endosimbionts** (el simbiont habita a l'interior de cèl·lules de l'hoste) i **ectosimbionts** (quan el simbiont viu sobre la superfície de l'hospedador). Finalment, segons el grau de dependència, trobarem simbionts **facultatius** o **obligats**. S'han catalogat un gran nombre d'associacions simbiòtiques, de tot tipus, repartides per tot l'arbre de la vida. Aquesta gran diversitat d'interaccions és un clar indicador de la importància de les interaccions simbiòtiques per a l'evolució de la vida.

Un dels casos més estudiats és la simbiosi d'insectes amb un ample espectre de bacteris. La presència en molts insectes de bacteris intracel·lulars mutualistes, coneguts com endosimbionts, pot ser una de les claus per explicar l'enorme èxit evolutiu d'aquest grup d'animals. En general tots els insectes tenen unes necessitats nutricionals molt paregudes. No obstant mostren una gran diversitat respecte a la dieta que segueixen. Molts insectes depenen d'una única font d'aliment, que en molts casos és deficient per a

certs nutrients. Aquestes deficiències nutricionals de la dieta són compensades per les capacitats metabòliques de l'endosimbiont. Així, els pugons obtenen els aminoàcids essencials del seu bacteri endosimbiont, *Buchnera aphidicola*, ja que la saba de la que s'alimenten, si bé és rica en sucre és molt pobra en aminoàcids. D'una forma similar, insectes hematòfags com la mosca tse-tse o els polls dels cabells, obtenen vitamines del grup B a través de bacteris endosimbiotics. D'aquesta forma, l'establiment d'aquestes interaccions hauria permès als insectes explotar noves fonts d'aliment que d'altra forma serien insuficients per cobrir les seues necessitats nutricionals. No obstant, no tots els insectes que posseeixen bacteris endosimbionts depenen d'una única font d'aliment, per exemple, formigues o panderoles tot i ser generalistes també posseeixen bacteris endosimbionts de caràcter mutualista. La seqüenciació dels genomes de *Blochmannia* spp. i *Blattabacterium* sp., endosimbionts de formigues del gènere *Camponotus* i de panderoles respectivament, ha confirmat la contribució d'aquests bacteris al metabolisme del nitrogen dels seus hostes així com en la síntesi d'aminoàcids essencials.

Segons el grau de dependència i l'edat de l'associació, es poden distingir dos tipus de simbionts. Per una banda, tenim els **endosimbionts primaris** (P-endosymbionts), transmesos de forma estrictament vertical i localitzats a l'interior de cèl·lules especialitzades de l'hoste anomenades bacteriòcits. Aquestes associacions són molt antigues i les relacions de dependència són tan fortes que cap dels dos membres del consorci podria sobreviure sense l'altre. Molts dels endosimbionts que complementen la dieta de l'hoste són d'aquest tipus, com l'abans esmentada *B. aphidicola* als pugons, *Wigglesworthia glossinidia* a les mosques tse-tse, *Riesia pediculicula* als polls o *Carsonella rudii* als psil·lids. Per altra banda, hi ha els **endosimbionts secundaris** (S-endosymbionts), que al contrari que els primers, no són indispensables per la supervivència de l'hoste, així que no

estan universalment distribuïts. Al igual que els endosimbionts primaris, el seu mode de transmissió és vertical encara que s'han descrit casos de transferència horitzontal. A diferència dels endosimbionts primaris, els secundaris no estan necessàriament confinats a un únic teixit o a l'interior del citoplasma de cèl·lules especialitzades de l'hoste, podent-se trobar a l'hemolimfa, glàndules o altres teixits no especialitzats. Lluny d'ésser simples comensals, la presència d'aquest simbionts exerceix efectes tant positius com negatius sobre l'aptitud dels hostes. Així, alguns secundaris protegeixen els hostes front a l'estrès tèrmic (*Serratia symbiotica* als pugons), confereixen resistència front a l'atac de vespes parasitoides (*Hamiltonella defensa* als pugons també) o infeccions fúngiques (*Regiella insecticola*). Altres, en canvi, manipulen la sexualitat de l'hoste en benefici propi amb *Wolbachia* spp. es un exemple paradigmàtic ja que s'ha estimat que vora un 40% de les espècies d'artròpodes poden ser infectades per aquest gènere de bacteris. Finalment, hi ha casos on un sol bacteri no és capaç de satisfer les necessitats nutricionals dels seus hostes, i n'és necessari un segon, que complemente les capacitats metabòliques del primer. En aquest cas els simbionts es coneixen com co-primaris. Aquest tipus d'associació es troba al pugó del cedre (*Cinara cedri*). *Buchnera* BCc ha perdut la major part de gens necessaris per sintetitzar triptòfan, així els passos absents a *Buchnera* són duts a terme per *Serratia symbiotica*, que en aquesta espècie de pugó ha assolit el paper de simbiont co-primari.

Durant la transició de vida lliure a intracel·lular, es donen de sèrie de canvis que modifiquen molt l'estructura genòmica d'aquests bacteris. Les característiques més destacades són la reducció genòmica i l'increment en contingut en AT. Dos factors són decisius per explicar el procés de reducció genòmica d'aquests bacteris. El primer és que al viure a l'interior d'una cèl·lula eucariota moltes funcions del bacteri esdevenen innecessàries o redundants. Així, la pèrdua dels gens que codifiquen per a les proteïnes

encarregades d'aquestes funcions no tindria cap efecte en l'aptitud del bacteri, d'aquesta forma la pressió de la selecció natural sobre ells està relaxada. El segon factor es la dinàmica poblacional d'aquests bacteris. La transmissió vertical provoca continus colls d'ampolla reduint la mida efectiva de la població, incrementat així l'efecte de la deriva genètica. Els dos factors esmentats, faciliten l'acumulació de mutacions lleugerament deletèries a gens no essencials fins inactivar-los i posteriorment eliminar-los. El procés de reducció genòmica es dóna en dues fases ben diferenciades. Primer, just després l'establiment de la simbiosi, hi ha una proliferació de seqüències d'inserció (IS), afavorida per la relaxació de la selecció purificadora. L'increment d'IS afavoreix el procés de reducció genòmica i provoca reordenacions cromosòmiques mitjançant esdeveniments de recombinació homòloga. En una segona fase, les IS desapareixen i es dóna pas a un període d'estabilitat genòmica. No obstant, el procés de reducció genòmica continua a través de la pseudogeneització i eliminació individual de gens repartits per tot el genoma. La pèrdua dels sistemes de reparació del DNA durant el procés de reducció genètica així com la pressió mutacional de GC cap AT explicarien l'alt contingut en AT que tenen aquests genomes.

La simbiosi entre *Blattabacterium cuenoti* i les panderoles és una de les associacions d'aquest tipus més antigues descrites. S'estima que ambdós espècies estan evolucionant juntes des de fa més de 140 milions d'anys, quan un Flavobacteri de vida lliure va infectar l'avantpassat comú de panderoles i tèrmits. Els bacteriòcits a l'interior dels quals viu *Blattabacterium* es localitzen al cos gras envoltats per un segon tipus cel·lular, els uricòcits, cèl·lules especialitzades dintre de les quals s'acumulen cristalls d'àcid úric. La major part de la informació sobre el paper de *Blattabacterium* al metabolisme del seu hoste ve de estudis clàssics basats en l'obtenció de panderoles aposimbiòtiques. D'aquesta forma es va poder relacionar *Blattabacterium* amb la síntesi de aminoàcids essencials. Altres estudis

clàssics clarament relacionen la presència de *Blattabacterium* amb el metabolisme del nitrogen. Tot i ser amonotèliques, les panderoles tenen la capacitat de sintetitzar àcid úric, que en compte de ser eliminat és emmagatzemat als uricòcits en forma de sals d'urat. Aquests depòsits es veuen incrementats en individus aposimbiòtics. Paral·lelament, s'ha observat que a les panderoles que són alimentades amb una dieta rica en proteïnes, els dipòsits d'àcid úric s'incrementen notablement, així com el volum i nombre de uricòcits. No obstant, aquests dipòsits es redueixen ràpidament si seguidament les mateixes panderoles són alimentades amb una dieta pobra en proteïnes.

Durant els últims anys, s'han seqüenciat els genomes de cinc soques de *Blattabacterium*, tres d'elles simbionts de panderoles omnívores: *Blattella germànica* (BBge), *Periplaneta americana* (BPam) i *Blaberus gigantenus* (BBgi), membres de les famílies Blattellidade, Blattidae i Blaberidae respectivament. A més, també hi ha disponible el genoma dels simbionts de la panderola *Cryptocercus punctulatus* (Cryptocercidae) (BCpu) i del tèrmit *Mastotermes darwiniensis* (Termitidae) (BMda), ambdós xilòfags i l'últim a més, l'únic membre de la família Termitidae que manté l'endosimbiont. Els genomes d' aquests bacteris presenten les mateixes característiques que la resta d'endosimbionts: genomes reduïts (de 590 a 640 kb); baix contingut en GC (del 23.8 al 28.2%); manteniment de l'ordre gènic i una elevada taxa de substitucions nucleotídiques. Les soques simbionts de panderoles omnívores tenen la capacitat de sintetitzar tots els aminoàcids essencials. Per contra, tant BCpu com BMda han perdut l'habilitat de sintetitzar set d'aquests aminoàcids. En aquestes dos últimes soques, s'ha postulat que es la microbiota intestinal l'encarregada de proveir els aminoàcids essencials als hostes.

Finalment, totes les soques de *Blattabacterium* tenen gens que codifiquen per a l'enzim ureasa. A més, les soques de BBge, BPam i BBgi posseeixen

els gens codificants per a tots els enzims del cicle de la urea, mentre que a BCpu i BMda s'ha perdut el gen *argH*, que codifica per a la argininosuccinat liasa, interrompent d'aquesta forma el cicle de la urea. La presència d'ureasa ha permès postular un model que explica tant l'amonotelisme, com el reciclatge del nitrogen de l'àcid úric emmagatzemat als uricòcits. La ureasa degradaria urea formant amoníac i CO_2. L'amoníac generat deurà ésser incorporat al metabolisme per l'acció de l'enzim glutamina sintetasa. Ara bé, per a que el model funcione cal inferir la presència tant de la glutamina sintetasa i dels enzims encarregats de degradar l'àcid úric (urat oxidasa, allantoinasa i allantoicasa) a l'hoste, ja que no estan codificats a *Blattabacterium*. L' anàlisi de balanç de fluxos (FBA) realitzat sobre la xarxa metabòlica dels simbionts de *B. germanica* i *P. americana* assenyala que *Blattabacterium* té el potencial per a generar amoníac, quan s'optimitza la xarxa per a la producció de biomassa. S'ha proposat un model on es suggereix que el nitrogen excedent és emmagatzemat als uricòcits presents al cos gras en forma d'urat. En períodes de carència, aquests dipòsits seran mobilitzats per l'acció dels enzims urat oxidasa, allantoinasa i allantoicasa, que degradaran l'àcid úric a urea. Aquesta serà hidrolitzada a CO_2 i amoníac per l'acció de la ureasa codificada a *Blattabacterium*. Finalment, l'amoníac alliberat per l'endosimbiont o bé serà incorporat al metabolisme del insecte gràcies a l'acció de la glutamina sintetasa de l'hoste, o bé serà excretat.

Material i mètodes

Seqüenciació, anotació i genòmica comparada amb les diferents soques de Blattabacterium

Les mostres de cos gras obtingudes de femelles adultes de *B. orientalis* van ser enriquides amb *Blattabacterium* a través de diversos passos de centrifugació i filtració. Seguidament, abans de lisar les cèl·lules de

l'endosimbiont, la mostra va ser tractada amb DNasa I, per eliminar el DNA de l'hoste però no el de *Blattabacterium*, ja que aquest últim està protegit per la membrana bacteriana. Finalment el DNA genòmic de *Blattabacterium* es va obtenir utilitzant el mètode del CTAB (bromur de cetiltrimetilamoni). Una volta purificat el DNA, es va quantificar mitjançant espectrofotometria.

La seqüenciació es va fer mitjançant el sistema GS-FLX de Roche. Addicionalment, es van generar llibreries que es van seqüenciar per el mètode de Sanger. Les seqüències resultants del procés de piroseqüenciació van ser assemblades amb el programa Newbler. Després es va crear una base de dades compatible amb el programa GAP4, dintre del paquet *Staden Package*, per revisar l'assemblatge manualment. A aquestes base de dades van ser adjuntades les seqüències obtingudes pel mètode Sanger.

Quan el genoma va ser tancat, es va començar el procés d'anotació. Les pautes de lectura oberta (ORF), es van identificar mitjançant el programa GLIMMER, encara que posteriorment van ser revisades manualment per identificar tant els codons d'inici com de parada, a més d'identificar possibles ORF obviades pel GLIMMER. L'assignació funcional d'aquestes ORF es va fer mitjançant la identificació d'ortòlegs a les altres soques de *Blattabacterium* i amb el Bacteroidete de vida lliure *Flavobacterium psychrophilum* mitjançant l'algorisme OrthoMCL. A més, es van fer cerques mitjançant BLASTP contra la base de dades no redundant de la Kyoto Encyclopedia of Genes and Genomes (KEGG). Finalment, també es van realitzar cerques mitjançant BLASTX i BLASTN a les regions intergèniques per tal d'identificar possibles pseudogens no identificats pel GLIMMER. A cada gen codificant per proteïnes (CDS) va ser assignat, sempre que fos possible, una o diverses categories COG (Cluster of Orthologous Genes), un número de KO (KEGG orthology) així com codis EC (Enzyme Comission number). Per la identificació dels gens de RNA es va fer una cerca mitjançant el algorisme INFERNAL contra una base de dades de seqüències

i estructures de RNA. Davant la manca del gen *dnaA* o les caixes DnaA, l'origen de replicació es va determinar mitjançant la desviació en GC, utilitzant el programa OriginX. El contingut G+C per al total del genoma es va determinar mitjançant el programa GeeCee, dintre el paquet EMBOSS. En canvi per determinar el contingut G+C als diferents gens, així com a diferent posicions de codons es va utilitzar un script de Perl. Finalmentm, la representació gràfica de la comparativa genòmica entre les sis soques de *Blattabacterium* seqüenciades es van fer mitjançant el paquet de R genoPlotR.

Per a realitzar l'estudi comparatiu entre les sis soques de *Blattabacterium* el primer pas va ser l'obtenció del pan-genoma. La representació gràfica del pan-genoma es va fer amb R, utilitzant el paquet Vennerable. Els valors per a cadascun dels subespais es van extreure a partir de la taula d'ortologia obtinguda amb OrthoMCL. Es van realitzar corbes de rarefacció per veure la cobertura del pan-genoma amb sis soques, mitjançant la funció speccaccum del paquet Vegan per R.

Per veure si la distribució de les diferents categories COG a les sis soques era estadísticament diferent es va realitzar una proba de la χ^2, utilitzant com a referència la distribució a BBge. Les diferents soques també es van agrupar segons les distribucions COG utilitzant la funció heatmap.2 del paquet gplots per a R.

Els efectes de la pèrdua dels tres primers gens del cicle de Krebs, als endosimbionts de *B. orientalis* i *P. americana*, es van mesurar mitjançant anàlisis estoiquiomètriques amb el programa METATOOL.

Per determinar les relacions filogenètiques de les diferents soques de *Blattabacterium*, es va utilitzar un concatenat de tots els alineaments obtinguts amb Mafft de les proteïnes presents a les sis soques de *Blattabacterium*, que a més, tingueren un ortòleg a *F. psychrophilum*. El millor model evolutiu es va estimar amb ProtTest, i l'arbre es va obtenir per

màxima versemblança, amb 100 pseudorèpliques utilitzant el programa RAxML. Posteriorment, es van situar els esdeveniments de pèrdua gènica sobre la topologia resultant.

Finalment es van realitzar dos tipus d'anàlisis evolutives, a partir d'aliniaments de nucleòtids obtinguts utilitzant aliniaments de proteïnes com a motlle. Primer es van calcular les taxes de substitució sinònima (dS) i no sinònima (dN), així con la ràtio entre ambdues (dN/dS), per tal de determinar el tipus de pressió evolutiva que està actuant sobre cada gen, utilitzant el programa YN00 dintre del PAML. Seguidament es va determinar, mitjançant un test de raó de versemblança (LRT), quins dels gens presents a totes les soques, estaven evolucionant sota el rellotge molecular. Per evitar el problema que la saturació de substitucions nucleotídiques pot suposar per aquest tipus d'anàlisi, es van eliminar les terceres posicions de cada codó així com els gens que tenien una taxa de substitució igual o superior a 2.5 substitucions per lloc. El LRT es va realitzar tal com ve determinat al programa Baseml utilitzant el model evolutiu seleccionat per a cada aliniament mitjançant jModelTest. Es va determinar el valor de versemblança sota dues suposicions, primer assumint rellotge molecular i l'altra deixant cada rama evolucionar a una taxa diferent. Tots aquells gens que no rebutjaren el rellotge molecular, i que a més tingueren un ortòleg a *F. psychrophilum*, es van utilitzar per a situar els temps de divergència dels diferent nodes de la filogènia de *Blattabacterium*, utilitzant com a calibrador el temps de la divergència entre les soques BBge/BBgi i la resta, determinat pel registre fòssil fa 140 milions d'anys.

Anàlisi transcriptòmica i resposta al contingut en proteïna de la dieta

Es va obtenir el cos gras i ovaris (tots dos teixits on *Blattabacterium* s'allotja) de femelles adultes, de 3 a 5 dies. Com a control es va utilitzar el teixit epitelial situat baix el pronot de nimfes a l'estadi 6 de

desenvolupament. La extracció de l'RNA total per a cadascun dels teixits es va fer, en condicions lliures de RNases amb el kit GenElute™ Mammalian Total RNA Miniprep Kit de Sigma-Aldrich. Prop de 10 μg de RNA de cadascun dels teixits es van enviar a la empresa GATC-Biotech, on es va sintetitzar el cDNA utilitzant el kit Smart cDNA Construction Kit de clontech. Donat que hi havia mostres de tres teixits diferents, es va adjuntar una seqüencia marcadora a cadascuna de les llibreries per tal de diferenciar-les en posteriors processos. Les mostres es van seqüenciar mitjançant piroseqüenciació al 454 de Roche, utilitzant la química Titanium. Com a resultat d'aquest procés vam rebre tres conjunts de dades, un per cada teixit, on prèviament s'havien eliminat les seqüències de baixa qualitat. Seguidament es van identificar i eliminar les seqüencies provinents de *Blattabacterium*, mitjançant un mapeig amb MEGABLAST contra una base de dades formada pel genoma de BBge. Finalment, l'assemblatge es va fer amb el MIRA 3.2 treballant amb el mode EST.

Els *contigs* obtinguts de l'assemblatge es van anotar amb Blast2GO; aquest programa inicia el procés de anotació mitjançant cerques per BLASTX contra la base de dades no-redundant del NCBI, seguidament assigna termes de *Gene Onthology* (GO) i codis enzimàtics a cadascuna de les seqüències. A partir de les dades obtingudes de la assignació de números EC, Blast2GO ens ofereix la possibilitat de generar mapes de rutes metabòliques utilitzant els mapes de la base de dades KEGG. Aquest mapes van ser curats manualment amb l'ajuda de las base de dades MetaCyc per tancar buits a les diferents rutes. Finalment es va comparar la distribució de termes GO entre els transcriptomes mitjançant un test de Fischer de dos cues.

Durant l'anàlisi del transcriptoma, es van identificar seqüències gèniques per a tots tres gens implicats en la síntesi i degradació de l'àcid úric a urea, (xantina deshidrogenasa, urat oxidase, allantoinasa i allantoicasa), així com

per a la glutamina sintetasa. Per estudiar la resposta d'aquests gens a la quantitat de proteïna ingerida a la dieta, es van dissenyar tres preparats alimentaris per a les panderoles amb diferents continguts en proteïna: del 0, 5 i 50% respectivament. L'aliment per a gossos fou control (25%). Femelles just acabades de mudar a forma adulta foren separades de la població general i es van mantenir durant dos dies amb la dieta control. Al tercer dia es va substituir la dieta control per un dels preparats abans esmentats per dos dies més. Al quart dia d'ençà que van mudar a adultes, es van viviseccionar per obtenir el cos gras i els ovaris. Es va extreure l'RNA total d'ambdós teixits, i es va sintetitzar el cDNA utilitzant hexàmers al atzar. D'aquesta forma es sintetitza el cDNA a la vegada de teixits de l'hoste i del bacteri. L'expressió relativa dels gens involucrats en el metabolisme de l'àcid úric a *B. germanica* i ureasa a *Blattabacterium*, així com la glutamina sintetasa de l'hoste es va mesurar mitjançant PCR quantitativa. Com a control intern es van gastar els gens de l'actina i el factor d'elongació EF-TU segons el gen d'interès fos de l'eucariota o del simbiont, respectivament. L'anàlisi estadística es va fer amb el programa REST.

Resultats

Seqüenciació i anàlisi comparada de Blattabacterium *endosimbiont de* B. orientalis

El genoma de *Blattabacterium* endosimbiont de *B. orientalis*, està composat per un cromosoma de 634.449 bp i un plasmidi de 3735 bp. En total s'han anotat 627 gens (set d'ells al plasmidi) distribuïts de la següent forma: 579 codifiquen per a proteïnes, 39 per gens de RNA (33 rRNA, un operó ribosòmic, i tres ncRNA), finalment 9 són pseudogens. Igual que les altres soques de *Blattabacterium*, el contingut G+C es molt baix (28,2%). Com a la majoria dels endosimbionts en estats avançats, l'ordre gènic es manté a totes les soques, i tan sols s'han descrit tres reordenacions, dos a la

soca simbiòtica de *M. darwiniensis* (una d'elles de 242 kb i altra de 2.9 kb) i una tercera de unes 19 kb a les soques BPam i BBor. Tot i que els genomes dels endosimbionts de insectes solen ser reduïts, a la soca BBge es van trobar nou gens duplicats. D'aquests hi ha cinc que mantenen ambdues còpies a totes les soques (*rodA, uvrD, lpdA, miaB* i *argD)*. El fet que s'hagen mantés al llarg de la història evolutiva de *Blattabacterium*, ens indica que la seua presència podria tenir cert significat fisiològic, i podrien ser considerats com ecoparàlegs.

El pangenoma de *Blattabacterium* consta de 655 gens (615 CDS, 3 rRNA, 34 tRNA i 3 ncRNA). La corba de rarefacció mostra que ha arribat a la saturació. El *core* (tots aquells gens presents a les 6 soques) per contra consta de 539 gens (502 CDS, 3 rRNA, 31 tRNA i 3 ncRNA), i representa el 82.3% del contingut gènic del pan-genoma. Si treiem de l'anàlisi les soques simbionts de les panderoles xilòfagues, la proporció del pan-genoma coberta per el *core* puja fins al 93.4%. Aquestes dades són un clar indicador del grau d'estabilitat assolit per l'associació entre *Blattabacterium* i les paneroles, ja que tot i dur evolucionant per separat més de 140 milions d'anys el contingut gènic és pràcticament idèntic a les diferents soques. L'estabilitat a *Blattabacterium* es destaca quan comparem les dades abans esmentades amb associacions amb una edat similar, com la de *Buchnera* amb els pugons, on el *core* representa el 74% del pangenoma. Els valors de *Blattabacterium* són similars als de *Blochmannia*, endosimbionts de les formigues del gènere *Camponotus*, que tan sols porten evolucionant per separat uns 20 milions d'anys.

L'anàlisi estadística sobre el perfil funcional, indica que no hi ha diferències significatives respecte a la distribució de les categories COG als sis genomes, tot i que BCpu i BMda han perdut gran nombre de gens implicat en el metabolisme d'aminoàcids (Categoria E). L'anàlisi d'agrupació que es va fer, agrupa les soques BCpu i BMda amb el *core*,

mentre que les altres es van agrupar junt al pan-genoma. Al *heat-map* es veu clarament com les categories COG millor representades son la J (traducció) i la E (Metabolisme i transport de aminoàcids), tot i que aquesta última està en una proporció menor tant en BMda com en BCpu.

Amb 465 gens presents a les 6 soques que a més posseïen un ortòleg a *F. psychrophilum* es va fer una reconstrucció filogenètica per màxima versemblança. Es van aliniar les seqüències proteiques per a cadascun d'aquests gens. Després els aliniaments es van concatenar donant com a resultat un aliniament de 174.707 llocs. El millor model evolutiu estimat per ProtTest va ser el CpREV+G. La filogènia resultant va deixar BBge i BBgi com a grups basals corresponent a la superfamília Blaberoidea. Les altres quatre soques, de la superfamília Blattoidea, es van agrupar, formant dos subgrups, un on s'incloïen BMda i BCpu, mentre que a l'altre van anar a parar BBor i BPam. Aquestes dades coincidien amb la filogènia obtinguda per als hostes on es passaven les termites d'ordre a família. La topologia resultant d'aquesta anàlisi es va utilitzar per situar tots els esdeveniments de pèrdua gènica al llarg que la història evolutiva d'aquest simbionts. Així es van identificar un total de 183 pèrdues gèniques, 70 d'elles, esdeveniments únics. Per altra banda 43 gens es deuen d'haver perdut al menys en dues ocasions durant l'evolució de *Blattabacterium*. D'aquesta anàlisi es pot deduir que la pèrdua de la capacitat de sintetitzar aminoàcids essencials en BMda i BCpu deu haver ocorregut prompte al llarg la evolució de l'ancestre comú entre ambdues soques. La pèrdua dels tres primers gens del cicle de Krebs deu haver ocorregut abans la separació dels llinatges BBor i BPam. Pel que fa a la capacitat per assimilar sulfat inorgànic, s'ha perdut en almenys 3 ocasions, una durant al evolució de les panderoles de la família Blattidae, altra durant la evolució de BCpu i finalment una tercera volta al llarg de la evolució de BBgi.

Entre els 502 CDS presents al *core*, 296 no rebutgen la hipòtesi de rellotge molecular, d'aquests 275 tenen un homòleg a *F. psychrophilum.* Per la datació dels temps de divergència es van utilitzar finalment 224 gens, aquells que continuaven sense rebutjar la hipòtesi de rellotge molecular després de l'addició del grup extern a l'anàlisi i que, a més, tingueren una taxa de substitucions nucleotídiques per lloc menor de 2,5. Així es va determinar que la divergència entre els simbionts de *P. americana* i *B. orientalis* va tenir lloc fa uns 12,3 ± 7,6 Mya. Per contra les divergències entre BBge i BBgi, i BCpu i BMda fou estimada en 89,5 ± 17,7 Mya i 87,0 ± 18,8 Mya respectivament. Finalment les anàlisis de la raó dN/dS indiquen clarament com la selecció purificadora és la que està actuant sobre la gran majoria de gens.

Anàlisi del transcriptoma

Com a resultat de la seqüenciació es van obtenir 554.403 lectures per als tres teixits analitzats, a raó de 164.677 per al cos gras, 166.672 per als ovaris i 223.054 per a l'epiteli. Entre les seqüencies del cos gras un 6,42% eren de transcrits provinents de *Blattabacterium*, en canvi, tant als ovaris com a l'epiteli, tan sols van ser identificades 104 i tres lectures originaries de l'endosimbiont respectivament. Del procés d'assemblatge d'aquestes dades es van obtenir 11.905 contigs per al transcriptoma del cos gras, 17.159 *contigs* per a l'ovari i 23.318 per a l'epiteli.

Per 5.848 dels *contigs* obtinguts del cos gras, un 49,1% del total, hi va haver resultats positius a les cerques per BLASTX contra la base de dades no redundant del NCBI. Ara bé, tan sols 3.993 van ser correctament anotats, amb assignació de termes GO o codi enzimàtic. Als ovaris, 8.052 *contigs* (46,9%) van tenir resultats positius a la cerca mitjançant BLASTX, i com al cas del cos gras, tan sols 5.376 foren correctament anotades. Al

transcriptoma de l'epiteli, la proporció de *contigs* amb similitud significativa a les bases de dades és menor que als altes dos teixits, així tan sols 6.396 *contigs*, un 27,4%, van tenir resultats positius a las cerques per similitud, i d'elles tan sols 3.623 van ser correctament anotades.

Al cos gras, es va assignar al menys un codi enzimàtic a 1.146 seqüències diferents. Amb aquesta informació es van cercar els transcrits dels enzims que participen en les principals rutes metabòliques als animals, glucòlisi/gluconeogènesi, ruta dels fosfats de pentosa i el cicle de Krebs. Es van identificar huit dels deu gens implicats en la glucòlisi. La gluconeogènesi comparteix la major part dels enzims de la glucòlisi amb l'excepció de tres enzims irreversibles, la hexocinasa, la fosfofructocinasa i la piruvat cinasa que són reemplaçats en aquesta ruta per la glucosa-6-fosfatasa, fructosa-bifosfatasa i la parella formada per la piruvat-cinasa/PEP carboxicinasa, es van detectar transcrits per aquests tres últims enzims. Tot i que cal tenir en compte que aquestes dues rutes no sempre estaran activades al mateix temps. Pel que fa a la ruta dels fosfats de pentosa, es van identificar transcrits per sis dels set gens implicats. Finalment, del cicle de Krebs tan sols per un dels enzims, l'aconitasa, no es van trobar transcrits del seu gen.

A l'ovari, tant les rutes per a la glicòlisi/gluconeogènesi com per als fosfats de pentosa mostraven el mateix perfil que al cos gras, en canvi per al cicle de Krebs hi havia transcrits de tots els gens implicats. Finalment a l'epiteli, per a la glucòlisi/gluconeogènesi a més dels tres enzims que tampoc van ser detectats als anteriors teixits, cal afegir que no es va trobar cap transcrit del gen de l'hexocinasa. A la ruta dels fosfats de pentosa, a part de no detectar transcrits per a la ribulosa-5-fosfat-isomerasa, tampoc se'n van trobar per a la 6-fosfoglucolactonasa i la transcetolasa. Del cicle de Krebs hi havia transcrits de tots els gens menys de la succinil-CoA-sintetasa.

Durant la reconstrucció del metabolisme de *Blattabacterium*, i posteriorment a través de les anàlisis per FBA, es va veure que ella bacteri necessitaria un aportació externa d'aminoàcids no essencials com l'asparagina, glicina, glutamina i prolina, així com altres compostos com el porfobilinogen i la pantoteina-4-fosfat. Al cos gras es van trobar transcrits de tots els gens codificant de enzims implicats en la síntesi de tots quatre aminoàcids, així com gran part de la ruta per a síntesi de porfobilinogen. A mes, s'ha detectat transcrits per al gen de la pantotenat cinasa necessari per a la síntesi de la pantoteïna-4-fosfat. Als ovaris, hi ha transcrits necessaris per obtenir els enzims per la síntesi d'asparagina i glutamina. No obstant, no s'han detectat transcrits per la glutamat-semialdehid-deshidrogenasa, implicada en la síntesi de prolina, ni per a la fosfoserina-fosfatasa, que participa a la síntesi de glicina. Finalment a l'epiteli tan sols s'ha trobat completa la ruta per a la síntesi de glutamina.

Finalment, al cos gras es van detectar transcrits del gens codificant per a la síntesi (xantina-deshidrogenasa) i degradació (urat-oxidasa, allantoinasa i allantoicasa) de l'àcid úric. Aquesta ruta, essencial per poder explicar l'ús que les panderoles fan de l'àcid úric com a reservori de nitrogen, tan sols es troba completa al cos gras. De fet cap d'aquest transcrits ha estat detectat a l'epiteli, i tan sols transcrits per la urat oxidasa i la al·lantoicasa s'han detectat a l'ovari.

Les anàlisis estadístiques comparades fetes sobre l' anotació del cos gras i l'ovari, indiquen, com era d'esperar, que el cos gras està enriquit en termes GO associats a la producció i conversió de metabòlits energètics, així com l'emmagatzemament i mobilització de substàncies de reserva. En canvi, a l'ovari destaquen tots les processos relacionats amb la regulació de la expressió gènica, control del cicle cel·lular, o localització cel·lular.

Resposta a diferent nivells de nitrogen a la dieta

L'expressió del gen per a la urat-oxidasa es veu significativament incrementada als animals que són alimentats amb una dieta sense o amb baixa proporció (5%) de proteïnes. Aquest augment es detecta tant al cos gras com a l'ovari. Cap dels altres gens implicats en la degradació de l'àcid úric a urea, així com la ureasa de *Blattabacterium*, veu modificat el seu nivell d'expressió segons la quantitat de proteïna a la dieta. Finalment, la glutamina-sintetasa, encarregada de recuperar el amoni alliberat com a conseqüència del catabolisme de la urea, veu el seu nivell d'expressió incrementat, tan sols al cos gras, quan el animal està alimentat amb una dieta sense cap traça de proteïna. A més, quan el nivell de proteïnes ingerit es molt elevat, baixa la seua expressió a l'ovari.

Conclusions

Durant la present tesi, s'ha seqüenciat una nova soca de *Blattabacterium*, simbiont de la panderola *B. orientalis*, i s'ha comparat amb les soques prèviament seqüenciades. Aquests estudis han confirmat l'extrema estabilitat de l'arquitectura genòmica a *Blattabacterium* tot i que les diferents soques porten evolucionant per separat des de fa més de 140 milions d'anys. El contingut gènic a penes a variat al llarg de tot aquest temps, sobretot si tan sols ens fixem en aquelles soques simbiòtiques d'espècies omnívores. Així es pot assumir que el procés de reducció es va donar d'una forma ràpida just després de la infecció de l'avantpassat de totes les panderoles i tèrmits, arribant ràpidament a una espècie de genoma "òptim" capaç de satisfer les necessitats de l'hoste, molt probablement relacionades amb el metabolisme del nitrogen. De fet, la major part de les pèrdues gèniques afecten proteïnes del metabolisme més perifèric o pobrament caracteritzades. La presència d'un nombre destacable de gens que segueixen el rellotge molecular i

acumulen un nombre baix de substitucions ha permès establir els temps de divergència de les diferents soques: les soques BCpu i BMda haurien divergit fa uns 87 milions d'anys, mentre que les dues soques de la superfamília Blaberoidea ho haurien fet fa uns 90 milions d' anys. La divergència entre BBor i BPam seria molt més propera, fa uns 12 milions d'anys.

L'anàlisi del transcriptoma a revelat que al cos gras s'expressen els gens que codifiquen els enzims necessaris per produir els aminoàcids que *Blattabacterium* necessita adquirir de l'exterior. A més, s'han trobat els transcrits dels gens implicats en la degradació del àcid úric a urea, així s' han identificat les eines que fan possible que les panderoles utilitzen l'àcid úric com un reservori de nitrogen. Finalment, s'ha comprovat com davant d'un període d'escassetat en nitrogen, s'incrementa la transcripció per al gen urat-oxidasa. No obstant no s'han identificat canvis significatius als altres gens implicats. Per altra banda, el gen per a la glutamina-sintetasa, que s'encarregaria d'incorporar l'amoníac alliberat per *Blattabacterium* al metabolisme, també mostra un increment en la seua expressió.