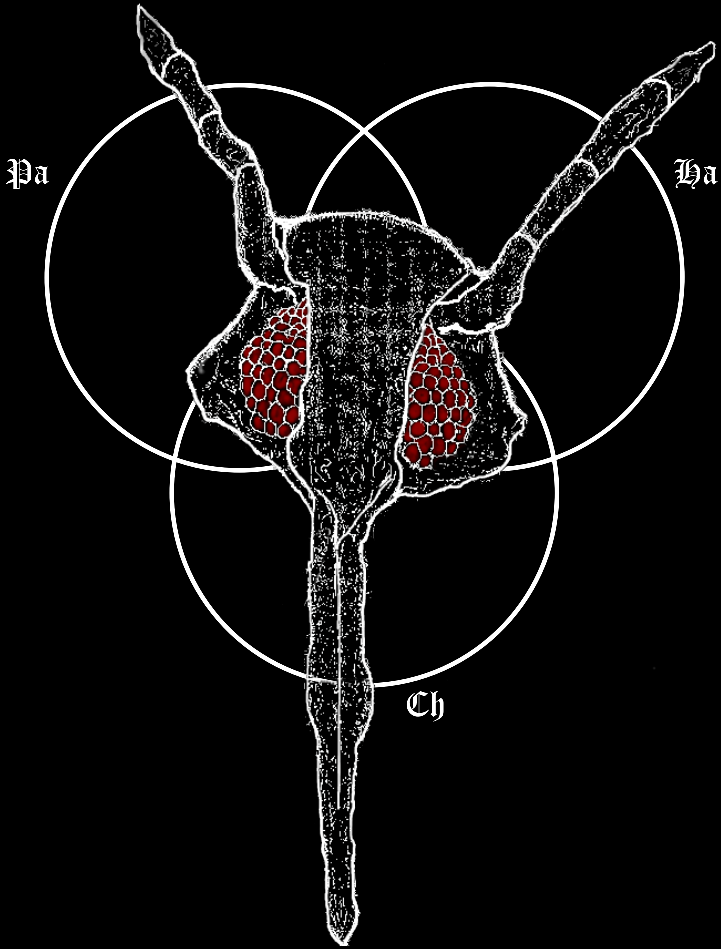
Endosymbiont communities in *Bemisia tabaci*: a metagenomic approach.



Diego Santos García
Tesis doctoral 2014

VNIVERSITAT DE VALÈNCIA

Facultad de Ciencias Biológicas

Instituto Cavanilles de Biodiversidad y Biología Evolutiva

Endosymbiont communities in *Bemisia tabaci*: a metagenomic approach.

Memoria presentada por Diego Santos García para optar al grado de Doctor en Biotecnología por la Universidad de Valencia.

Directores:
Dr. Francisco José Silva Moreno. Catedrático de Genética de la Universidad de Valencia.
Dra. Amparo Latorre Castillo. Catedrática de Genética de la Universidad de Valencia.
Dr. Andrés Moya Simarro. Catedrático de Genética de la Universidad de Valencia.

Valencia, 2014

FRANCISCO JOSÉ SILVA MORENO, Catedrático del Departamento de Genética de la Universidad de Valencia, AMPARO LATORRE CASTILLO, Catedrática del Departamento de Genética de la Universidad de Valencia y ANDRÉS MOYA SIMARRO, Catedrático del Departamento de Genética de la Universidad de Valencia,

CERTIFICAN que el trabajo para optar al grado de Doctor en Biotecnología, y que lleva por título "Endosymbiont communities in *Bemisia tabaci*: a metagenomic approach", ha estado realizado bajo su dirección en el Instituto Cavanilles de Biodiversidad y Biología Evolutiva por DIEGO SANTOS GARCÍA.

Y para que así conste, en el cumplimiento de la legislación vigente, firmamos el presente certificado en Valencia, a de de 2014.

Francisco José Silva Moreno Amparo Latorre Castillo

Andrés Moya Simarro

“Pero lo que no había aprendido era una cosa: a estar satisfecho de sí mismo y de su vida. Esto no pudo conseguirlo. Acaso ello proviniera de que en el fondo de su corazón sabía (o creía saber) en todo momento que no era realmente un ser humano, sino un lobo de la estepa.”

El lobo estepario - Hermann Hesse

“This thesis has been written in LaTeX. The electronic version contain hyperlinks to the different sections, figures, tables, bibliography and annex files. All references are linked to their on-line versions.”

Acknowledgements

La verdad es que ha sido un largo camino hasta llegar aquí. Este trabajo es la culminación de mis cinco años pasados en el grupo de Genética Evolutiva de la Universidad de Valencia, pero no habría llegado hasta aquí sin las enseñanzas, ayuda, comprensión y apoyo de un gran número de gente. Siento si me dejo a alguien, pero es difícil hacer memoria y recordar a toda la gente que ha influenciado en tu vida.

Lo primero es agradecer a todos los profesores y profesoras que se dejan la vida en la enseñanza pública. De mi paso por el Colegio Público Pablo Neruda recuerdo con especial cariño a Don Leandro ("Con cien cañones por banda...") y Benito (menudo puente colgante nos marcamos en su clase...). Del I.E.S. Jordi de Sant Jordi (uno de los institutos con más "alfaguays" por metro cuadrado de Valencia), quiero agradecerle a Aurelia (ya hay otro Biólogo salido del Jordi) y Fernando (me alegró verte por el Cavanilles) sus magníficas clases. Gracias a ellas no aprendí nada nuevo en la Universidad hasta segundo de carrera. ¡Qué paciencia y amor por la profesión la vuestra!

Agradecer los años pasados en el colegio con Miguel (Pilin), Toni Romero, Antonio, Adrián, Saceo, Julián, Jaume, Omar Badia, Omar Bakali,...Ha algunos os sigo viendo, otros habéis quedado ya en la negra bruma del olvido. A los compañeros del Shi-Ner-Khan: Dani (te deseo lo mejor en Chile), Luisa (que ha retornado a España con nosotros), Javi "Azahag" (hicimos mil locuras juntos, pero al final nuestros caminos se separaron) y a mi sensei Pancho (gracias por inculcarnos la filosofía del Judo). Agradecer también los años pasados en el cámping de Cheste (el Mono Amelio, el Gitano, Zaira, Zulema,...) y de Mareny de Barraquetes (David, Alberto, Samu, Eloy, Patas...). ¡Que buenos momentos tirando piedras al bus de Chocolate!

A la gente de Camporrobles, son muchísimos años juntos y los que quedan aún de autobús (aunque nos hacemos mayores y el botellón lo hemos ido cambiado por Gin&Tonics de tranqui): Josemi (mi primo, compañero de fechorías pero sobre todo gran amigo), Hector, Tere, Raúl, Oscar (otro gran amigo con el que nos quedan muchas birras por tomar, aunque cada vez sea más jipi), Didac (mi otro primo y amigo, también gamberro donde los haya y al que le debo el nombre), Hugo, Esther (aunque todos terminamos bailando, aquí estamos para lo que necesites), Laura y al resto de Minoría (o corto o necesitaré una segunda tesis).

Durante la Universidad conocí a muchos amigos biólogos, algunos ya emigrados otros aún pululando por aquí. Un abrazo a la Secta de Biólogos: José (el instigador y piedra angular del grupo), Marc, Juancho, Yago, Joan (¿alguien sabe salgo de él?), Juanan, Silvia, Patri, Pepe, Miguel, Susana, Laura, Isa,...A los que se apuntaron más tarde: Coca (otro paisano del Jordi), Raquel, Luz, Carmen (que también acabó en el Cavanilles), Chiara, Majo, Gema, Esther...

A la gente de Metal Bats, por mantener el espíritu del Heavy Metal en Valencia. En especial a Raúl “Pelilas” (ahora nos tendremos que ver en conciertos intencionales), Diego (Loof), Joseto, Monfort y Xavillin. Tanto concierto disfrutando juntos...Que sí, que el Heavy es una forma de vida.

A la banda que se forjó en el instituto y con la cual descubrí el Carmen: Pani (que haríamos sin ti...nunca cambies), Edu (tío ecléctico y “chistero” donde los haya), Santi (nuestro Quevedo particular), Saúl (el simpático del grupo...hasta que le hinchan las p*****s). Gracias amigos, porque se que por mucho que pasen los años puedo contar con vosotros. Nunca os podré agradecer suficientemente el detallazo con Naty y conmigo de la despedida...Por supuesto, mil gracias también a Jesica, Arantxa y Lidia.

También quisiera agradecer a la gente que me enseño mi oficio. Pedro e Isabel de Fisiología Vegetal, ellos me enseñaron todo lo que se sobre un laboratorio de biología molecular (incluso me “pelaba” algunas clases de la carrera para ir al lab). Al la gente de Diferenciación de Células Troncales del difunto CIFP y en especial a Arantxa, que con paciencia me enseño el desarrollo del *medaka* así como las técnicas para trabajar con estos pececillos. Nunca olvidaré que un 0.001% de sedante no es un gramo en 100 ml de agua, ni a los pececillos panza arriba...A la gente de Genética del Desarrollo: Núria, Rubén, Juanma, Vero, Favrice, Use, Amparo, Bea, Yaiza, Inma y Use. Use, siempre recordare el “A ti te mola Barón Rojo, lo veo en tu camiseta”. Espero que todo te vaya bien porque te lo mereces, altibajos los tenemos todos. A Carmén (por el buen humor y la amabilidad) y Raquel (una de las mejores científicas que conozco, gracias por enseñarme a hacer bio-ensayos y que al papel-paraquat hay que añadirle agua) de Valentia Biopharma.

Aunque siempre hemos dicho que esto no es una guardería y que aquí se viene a trabajar y no a hacer amigos, creo que al final me he llevado unos cuantos:

Gracias a la gente del lab de enfrente, porque con ellos he compartido muchas comidas, risas y buenos momentos: Silvia “Magic” PCR (voy a hecharte de menos allá donde vaya), Quelo “Això ho pague jo” (creo que es a tí a quien habría que estudiarle los ritmos circadianos...), Mariano “Maravillossso” (psstt, pásame el número de tu camello), Adrián y Teresa. A los del lab más “alejado”, tanto los que están como los que se fueron: David Peris “Pegatina” (como hecho de menos nuestras frikadas), Pilar, Maria Noel, Marin y Ron (eres muy grande, gracias por tu buen humor y nos vemos en tu tierra).

A los “locos” que formamos, o han formado, la infantería de Genética Evolutiva y del CSISP: Araceli (gracias por todos tus buenos consejos), Eugeni (gran profe de bioinformática), Rafa (la WikiPedia decidió que es el estereotipo de hypster), Sergio y Ana Elena (gracias por recomendarme a Amparo, esta aventura no habría sido lo mismo sin vosotros), Vanesa (la última Paiportina), Francesc Peris (mi mentor en el citómetro), Maria Dzunkova, Anouk y Laura. Quiero agradecer su amistad en especial a Manzano (gracias por picarme con el gusanillo de la programación), Mariana (espero que llegues muy lejos) y Carlos (otra gran adquisición para el lab). Gracias a los tres por ayudarme en esto de la bioinformática y los ánimos en épocas difíciles.

A la caballería de Genética Evolutiva y del CSISP: Ana (otra genómica más y una intensa post-doc), Giuseppe (un hombre renacentista, igual le da la bioinfomática que al wet lab), Sari, Juli y todo el profesorado de GenEvol.

A los estrategas que han dirigido esta tesis. Gracias Andrés por permitirme formar parte de este grupo, se que sin el Prometeo me habría sido muy difícil conseguir otra beca con mi nota. Muchas gracias Amparo por tus buenos consejos, tu paciencia (se que muchas veces me voy por las ramas y necesito que me recuerden mi objetivo), tu experiencia, la amabilidad y sobre todo por el apoyo que me has ofrecido a lo largo de los años. Como no, gracias a Paco del que he aprendido (pero solo algo) a sintetizar y no fiarme del todo de los métodos automáticos. Gracias por considerarme como un igual (echaré de menos las discusiones sobre los distintos proyectos que hemos llevado a cabo y los excels con colorines), pero sobre todo gracias por ayudarme a convertirme en un buen científico porque no he podido tener mejor maestro.

Gracias a Paco Beitia en el IVIA por los buenos momentos y las lecciones sobre la mosca blanca y sus parasitoides. Espero que podamos seguir colaborando en el futuro. Thanks to all the people I met at the First International Whitefly Symposium in Kolymbari and at the Internation Symbiosis Society congress in Krakow. My best wishes to the people at the Entomology Lab in Newe Ya'ar and all the people I met in Israel for their hospitality: Netta (I loved your pragmatism), Hanoch (we still owe you a visit to your kibbutz), Ayelet (thanks for the FISH lessons), Dr. Shai Morin (thanks for the post-doc opportunity) and all the people that I'm not able to remember their names. I want to thanks especially to Dr. Einat Zchori-Fein because

I consider her my fourth Ph. D. advisor. Thanks for all your kindly words.

Aunque ya casi al final, pero no menos importante, quiero dar las gracias a mi gran familia:

A mi tío José Lorenzo (mi padrino en muchos ámbitos) por inculcarme tu amor a la biología, los paseos por el campo, por el sensato intento de que no entrará en biología, por enseñarme que cuando se hace algo se llega hasta el final (incluso cunado robas botellas de vino con tus primos en el pueblo...jeje) y por tus consejos. A mi tía Ani y mis primos Ana Lorena y Loren.

A toda la familia Santos, ya que ahora llevó doble responsabilidad por ser el primer Doctor y el primer Doctor en emigrar de esta familia (siento robarte la frase tía Merche). A mis tías Mari Carmen (mi madrina), Fina (eres como una tía más), Tere, Milu, Merche, Charo y Ana (el superglue de la familia). A mis tíos Zésar, Felipe, Juán y José Luis (se te hecha mucho en falta). A mis primos Eva, Raúl, Víctor, Ana, José Luis, Dani y Alberto. A mis cuñados Carlos, Luis y Ángel con los que he compartido buenos y malos momentos. Aquí estoy para lo que os haga falta. A mi suegra Nati, porque también te voy a echar de menos y lo sabes. A Odín (ley universal, perro pequeño nombre grandilocuente) y Lalo (gato cabroncete donde los haya), la compañía que ofrecen no tiene precio (bueno sí, a cambio de sustento y techo...jeje).

Para el final me dejo lo más importante. Gracias a mi hermano Ernesto (mi antítesis), por las grande peleas y todas las historias que nos han hecho ser como somos. Puede que no seamos uña y carne, pero somos hermanos y por mi parte intentaré no perder nunca el contacto. A mis abuelos Heraclio y Josefina, por todo el amor que nos han dado durante estos años y los buenos momentos en Camporrobles. Siempre recordare con añoranza el primer día que fui con mi abuelo a las colmenas con el traje de apicultor o el olor a empanadillas de mi abuela. Es duro verlos envejecer. A mis abuelos paternos Juanita y Marcelino, a los que apenas conocí. A mis padres Marce e Isabel, por todo su amor, esfuerzo, alegrías y penas, grandes vacaciones y grandes castigos (papá, aún guardo la coleta de tela que me cortaste después de la feria Alternativa...), vuestro apoyo incondicional a este (a veces gruñón) biólogo. Por ser como sois. Porque yo soy quien soy gracias a vosotros.

A mi mujer, gracias por compartir durante tantos años (y los que nos quedan) conmigo el camino de la vida. Porque pese a que seamos como perro y gato, eres lo mejor que me ha pasado. Gracias por tu ayuda, tu paciencia (esta tesis se le hace cuesta arriba a cualquiera), tu amor, tu amistad, tu comprensión (porque se que la vida del científico no es fácil de aguantar), tu compañía, tanto buen momento juntos, porque para lo malo ya habrá otras ocasiones para hablar. Porque verte sonreír es un regalo. Gracias por venirte conmigo de "aventuras" dejándolo todo. Simplemente TE QUIERO.

Dedicado a mi mujer Naty y a mis
padres Marce e Isabel

Table of Contents

Acknowledgements I

Table of Contents IX

1 Introduction 1

1.1 Symbiosis 3

- 1.1.1 History and definitions 3
- 1.1.2 Symbiosis as an evolutionary driving force 5
- 1.1.3 Genome reduction in endosymbiosis 7

1.2 Symbiosis in insects 10

- 1.2.1 Primary endosymbionts 12
- 1.2.2 Secondary endosymbionts 15

1.3 Whiteflies 18

- 1.3.1 Whiteflies' biology 18
- 1.3.2 Endosymbionts of whiteflies 21
- 1.3.3 Endosymbiont transmission in whiteflies 23

1.4 Whiteflies used in this work 24

- 1.4.1 Aleyrodinae 24
  - 1.4.1.1 *Bemisia tabaci* 24
  - 1.4.1.2 *Trialeurodes vaporariorum* 27
- 1.4.2 Aleurodicinae 27
  - 1.4.2.1 *Aleurodicus dispersus* 27
  - 1.4.2.2 *Aleurodicus floccissimus* 28

2 Objectives 29

3 Material and Methods 33

3.1 Whiteflies samples 35

3.2 Microscopy techniques 35

- 3.2.1 Transmission Electron Microscopy 35
- 3.2.2 Fluorescent *in situ* hybridization 36

3.3 Endosymbiont enriched samples 37

- 3.3.1 Bacterial enriched samples 37

3.3.2 Bacteriome extraction and DNA amplification 38
3.4 DNA extractions, PCR reactions and quantification 39
3.4.1 Genomic DNA extraction 39
3.4.2 Chelex DNA extraction 40
3.4.3 PCR amplification 40
3.4.4 DNA purification and quantification 41
3.5 Genome Sequencing 42
3.5.1 Sanger sequencing 42
3.5.2 Next Generation Sequencing (NGS) 42
3.6 Genome assembly and annotation 43
3.6.1 Assembly 43
3.6.2 Annotation 48
3.7 Comparative Genomics 49
3.7.1 Orthologous proteins and synteny 49
3.7.2 Genome aligners 50
3.7.3 Last Common Ancestor (LCA) Reconstruction 50
3.7.4 Metabolic competition 51
3.8 Phylogenetic Methods 51
3.8.1 Alignments 51
3.8.2 Phylogenetic tree inference 52
3.8.3 Divergence dating 52
3.9 Evolutionary analyses 54
3.9.1 dN/dS sites analyses 54
3.9.2 Positive selection test 55
4 Results and Discussion 57
4.1 *Portiera* and its partner *Hamiltonella* 59
4.1.1 Background 59
4.1.2 *B. tabaci* QHC-VLC endosymbionts 60
4.1.3 *Portiera* BT-QVLC 68
4.1.3.1 *Portiera* BT-QVLC genomic features 68
4.1.3.2 Comparative genomics 70
4.1.3.2.1 *Portiera* strains from *B. tabaci* 76

4.1.4 *Hamiltonella* BT-QVLC 78
4.1.5 Metabolic integration 82
4.1.5.1 *Portiera* biosynthetic capabilities 83
4.1.5.2 *Hamiltonella* biosynthetic capabilities 86
4.1.5.3 Shared pathways 87
4.2 The third passenger: *Cardinium* cBtQ1 89
4.2.1 Background 89
4.2.2 General features of the genome of *Cardinium* cBtQ1 . . . 90
4.2.3 Taxonomic status of *Cardinium* cBtQ1 92
4.2.4 Comparative genomics 96
4.2.4.1 Mobile elements and genomic redundancy 96
4.2.4.2 Comparative genomics of *Cardinium* strains and *A. asiaticus* 99
4.2.4.3 Evolution of gene repertoires in the lineages of *A. asiaticus* and *Cardinium* 101
4.2.5 Biosynthetic capabilities in *Cardinium* cBtQ1 106
4.2.6 Gliding genes in *Cardinium* cBtQ1 109
4.2.6.1 Gliding machinery organization 114
4.2.6.2 Rhapidosomes in *Cardinium* 118
4.2.6.3 Possible gliding implications 120
4.3 Genome evolution of the genus *Portiera* 123
4.3.1 Background 123
4.3.2 Genomic features of *Portiera* strains 126
4.3.3 Comparative genomics and genome stasis in the genus *Portiera* 128
4.3.4 Metabolic blueprints of *Portiera* strains 134
4.3.5 Divergence times of *Portiera* lineages 138
4.3.6 Rates of nucleotide substitution in *Portiera* lineages . . . 145
4.3.7 Selective pressure in *Portiera* from *B. tabaci* 150
5 Conclusions 155
6 Bibliography 161

7 Appendix 183
7.1 List of Figures, Tables and Acronyms 185
7.2 Scientific production 192
8 Resumen 195
8.1 Introducción 197
  8.1.1 Simbiosis 197
  8.1.2 Simbiosis en insectos 197
  8.1.3 Moscas blancas 199
  8.1.4 Moscas blancas usadas en este trabajo 200
8.2 Objetivos 201
8.3 Material y Métodos 201
  8.3.1 Moscas blancas usadas 201
  8.3.2 Técnicas microscópicas 202
  8.3.3 Enriquecimiento de muestras en endosymbiontes 202
  8.3.4 Extracciones de ADN, PCR y cuantificación 203
  8.3.5 Secuenciación de genomas 203
  8.3.6 Ensamblaje y anotación de genomas 204
  8.3.7 Genómica comparativa 205
  8.3.8 Métodos filogenéticos 206
  8.3.9 Análisis Evolutivo 207
8.4 Resultados y discusión 207
  8.4.1 *Portiera* y su socia *Hamiltonella* 207
    8.4.1.1 Endosimbiontes en *B. tabaci* QHC-VLC 207
    8.4.1.2 Portiera BT-QVLC 208
      8.4.1.2.1 Genómica comparada 208
      8.4.1.2.2 *Hamiltonella* BT-QVLC 210
      8.4.1.2.3 Integración metabólica 211
  8.4.2 El tercer pasajero: *Cardinium* cBtQ1 212
    8.4.2.1 Características generales del genoma de *Cardinium* cBtQ1 212
    8.4.2.2 Estatus taxonómico de *Cardinium* cBtQ1 213

8.4.2.3 Genomica comparada 213
8.4.2.3.1 Elementos móviles y redundancia genómica 213
8.4.2.3.2 Genómica comparada entre las cepas de *Cardinium* y *A. asiaticus* 214
8.4.2.3.3 Evolución del repertorio génico en los linajes de *A. asiaticus* y *Cardinium* . . 214
8.4.2.4 El metabolismo de *Cardinium* cBtQ1 215
8.4.2.5 Genes de “deslizamiento” (*gliding*) en *Cardinium* cBtQ1 215
8.4.2.5.1 Organización de la maquinaria de *gliding* 216
8.4.2.5.2 Rhapidosomas en *Cardinium* 217
8.4.2.5.3 Posibles implicaciones del *gliding* . . . 217
8.4.3 Evolución genómica en el género *Portiera* 218
8.4.3.1 Características genómicas de las cepas de *Portiera* 218
8.4.3.2 Genómica comparada y estasis en el género *Portiera* 219
8.4.3.3 Los “planos” metabólicos de las cepas Portiera . 219
8.4.3.4 Tiempos de divergencia en los linajes de *Portiera* 220
8.4.3.5 Tasas de substitución nucleotídica en los linajes de *Portiera* 221
8.4.3.6 Selección positiva en *Portiera* de *B. tabaci* 223
8.5 Conclusiones 223

Part 1

Introduction

“Come down with fire
Lift my spirit higher
Someone’s screaming my name
Come and make me holy again
I’m the man on the silver mountain”

Rainbow

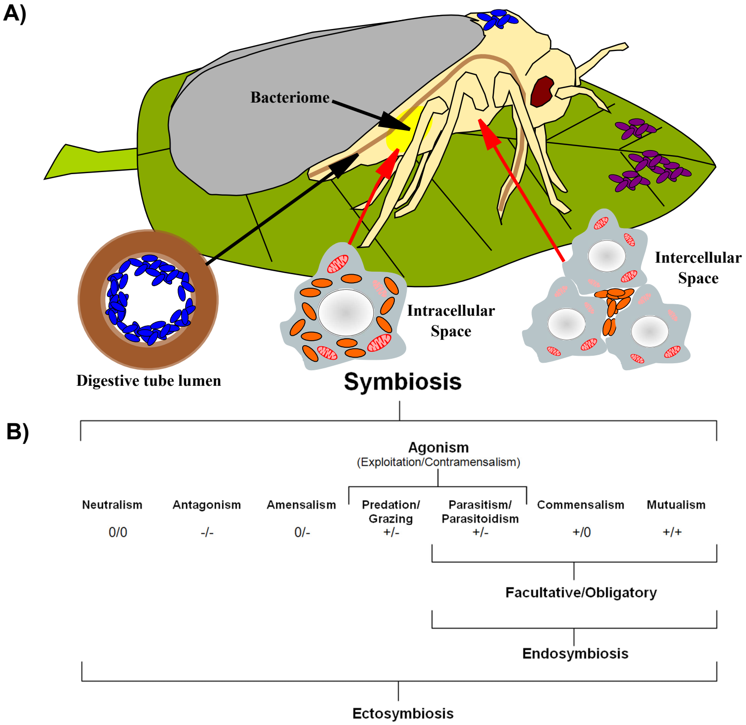
Symbiosis

1.1.1. History and definitions

The first biological meaning of the word “symbiosis” (derived from the Ancient Greek σύν “together” and βίωσις “living”) was introduced by Heinrich Anton de Bary. He coined the term as “the living together of unlike organisms” in his work “Die Erscheinungen des symbiose” without any restricted reference about the effects of the interaction between the symbiotic organism (DeBary, 1879). Since them, different semantic definitions have arisen with more or less astringency and acceptance. For example, a more restricted term considers only cases where the symbiotic organism takes profit of the interaction (Saffo, 1992) or when a new metabolic function arises as consequence of the interaction (Douglas, 1994). Despite problems associated with the use of the different meanings of symbiosis, it seems that a broad de Bary definition (that included mutualism, commensalism and parasitism) is the most widely accepted term (Martin and Schwab, 2012).

Regarding the aforementioned broad concept of symbiotic associations of host and symbiont, symbiosis can be also defined depending on the localization of the symbiont, the effects of the symbiotic relationship and the interdependence of the organisms (see Figure 1.1.1). If the symbiont is located in the external surface of the host, it is referred as **ectosymbiosis** (such as digestive tube symbionts, which are located in its inner surface). When the symbionts are localized inside the host, this association is called **endosymbiosis**. Also, **endosymbiosis** can be **extracellular** (the symbiont is in internal cavities or in intercellular spaces) or **intracellular** (inside the host’s cells isolated in host-derived vacuoles or freely in the cytoplasm).

Taking into account the effect of the relationships between host and symbiont, there can be different types of interactions: neutralism, antagonism, amensalism, predation, parasitism, commensalism and mutualism (Figure 1.1.1). The last three are the most widely known types of symbiosis.

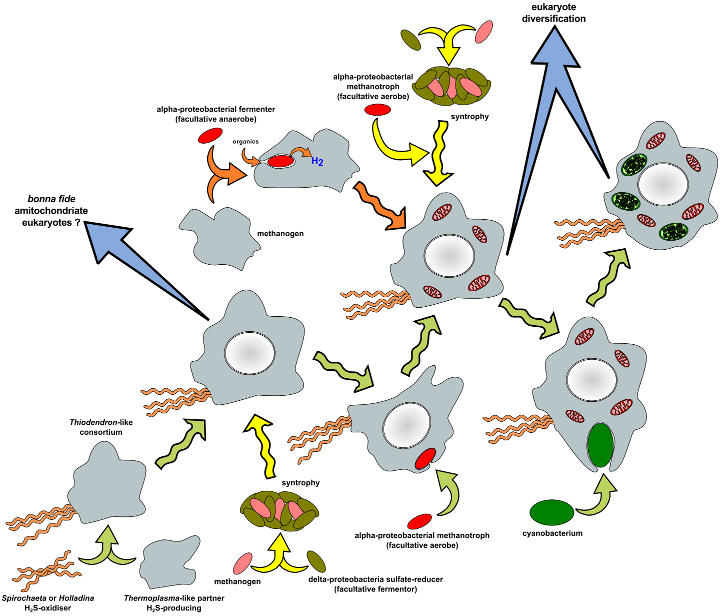

A)
Bacteriome
Intercellular Space
Intracellular Space
Digestive tube lumen
Symbiosis
B)
Agonism
(Exploitation/Contramensalism)
Neutralism
0/0
Antagonism
-/-
Amensalism
0/-
Predation/ Grazing
+/-
Parasitism/ Parasitoidism
+/-
Commensalism
+/0
Mutualism
+/+
Facultative/Obligatory
Endosymbiosis
Ectosymbiosis

Figures 1.1.1 A) Types of symbiosis attending to the localization of the symbiont relatively to the host. A whitefly is used as an example of host. The bacteriome will be explained in detail at Section Symbiosis in insects. Rod shapes represent: environmental bacteria (purple), ectosymbionts (blue) and endosymbionts (orange). **B)** Scheme of symbiosis interactions reproduced from Martin and Schwab (2012).

In the case of **parasitism**, the symbiont is benefited while the host has a detriment in its fitness. **Commensalism** defines a relationship where the fitness of one of the organism (usually the symbiont) is benefited but the other one (usually the host) does not suffer a detrimental fitness effect. Lastly, **mutualism** is considered when both, the host and the symbiont, have beneficial effects on their fitness. **Facultative** symbionts are those that do not require the symbiotic relationship for their survival. On the other hand, **obligate** symbionts cannot survive outside the symbiotic relationship (Figure 1.1.1).

1.1.2. Symbiosis as an evolutionary driving force

That prokaryotic symbiosis is a major evolutionary force is nowadays commonly accepted, but the prevailing idea before the 1960's was exclusively a "pathogenic" view of the microorganism, instead any of the other kinds of symbiotic interactions (Sapp, 1994). The aforementioned idea remained until Lynn Margulis in 1967 reintroduced the concept of the endosymbiosis as a primary evolutive force, based on the works of Konstantin Mereschkowski (1910) and Ivan Wallin (1920), and updated it with new scientific evidences (Sagan, 1967). This idea was revealed as the Serial Endosymbiotic Theory (SET) that considers an endosymbiotic origin of the eukaryotic cell, and it proposes that the eukaryotic cell is the product of series of endosymbiotic events (green arrows in Figure 1.1.2).


eukaryote diversification
alpha-proteobacterial methanotroph (facultative aerobe)
syntrophy
alpha-proteobacterial fermenter (facultative anaerobe)
organics
H_2
bonna fide amitochondriate eukaryotes ?
methanogen
Thiodendron-like consortium
syntrophy
alpha-proteobacterial methanotroph (facultative aerobe)
cyanobacterium
methanogen
delta-proteobacteria sulfate-reducer (facultative fermentor)
Spirochaeta or *Holladina* H_2S-oxidiser
Thermoplasma-like partner H_2S-producing

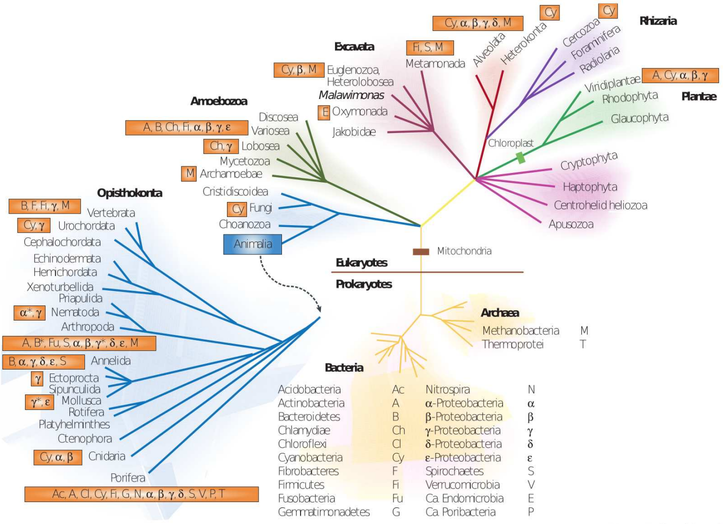
Figures 1.1.2 Different theories about the endosymbiotic origin of the eukaryotic cell. Modified from Latorre *et al.* (2011)

The initial step would be an association between an archaea and a spirochaeta followed by the evolution of the nucleus. After this event, the acquisition of an α-proteobacterium resulted in the mitochondria that are present in almost all eukaryotic cells. Lastly, secondary or tertiary endosymbiotic events of different cyanobacterial cells derived in the known chloroplast lineages (green plants, red algae, brown algae, etc.). Nevertheless, there are other valid hypothesis that explain the origin of the ancestral eukaryote, such as the Syntrophy (yellow arrows in Figure 1.1.2) (Moreira and Lopez-Garcia, 1998) or the Hydrogen (orange arrows in Figure 1.1.2) hypothesis (Martin and Müller, 1998).

The importance of symbiosis in evolution is not only because it shaped the origin of eukaryotes but also because symbiotic interactions are detected through the three domains of life (both in intra and inter domains). Indeed, the huge metabolic capabilities of prokaryotes and their adaptability to different environments have allowed the establishment of a wide range of symbiotic relationships with most of the eukaryotic lineages (McFall-Ngai, 2008; Moya *et al.*, 2008) (see Figure 1.1.3).

These associations allow the acquisition of new capabilities by the host, such as nitrogen fixation (Kneip *et al.*, 2007), sulphur and nitrogen assimilation (Nakagawa *et al.*, 2014), chemolithoautrotophy (Stewart *et al.*, 2005), toxin degradation (Adams *et al.*, 2013), nutrition (Hansen and Moran, 2014), etc.

Among symbiosis, one interesting case is the endosymbiosis between prokaryotes (mostly bacteria) and eukaryotes. The specific type of endosymbiosis between bacteria and animals can be one of the most intimate cases, where the endosymbiont is transmitted vertically from the mother to the offspring. The first compendia of bacterial endosymbiosis in animals was written by Buchner (1965) and even though the larger number of cases reported were from insects, other cases in nematodes, sponges, annelids, bryozoans and molluscs were included. This is not surprising, because it has been estimated that between 10 % to 20 % of insects have established endosymbiotic relationships with bacteria.


Cy, α, β, γ, δ, M
Cy
Cy
Rhizaria
Excavata
Fi, S, M
Metamonada
Alveolata
Heterokonta
Cercozoa
Foraminifera
Radiolaria
Cy, β, M
Euglenozoa,
Heterolobosea
Malawimonas
E
Oxymonada
Jakobidae
Viridiplantae
A, Cy, α, β, γ
Rhodophyta
Plantae
Glaucophyta
Amoebozoa
Discosea
A, B, Ch, Fi, α, β, γ, ε
Variosea
Ch, γ
Lobosea
Mycetozoa
M
Archamoebae
Chloroplast
Cryptophyta
Haptophyta
Centrohelid heliozoa
Apusozoa
Opisthokonta
Cristidiscoidea
B, F, Fi, γ, M
Vertebrata
Cy
Fungi
Cy, γ
Urochordata
Choanozoa
Animalia
Cephalochordata
Echinodermata
Hemichordata
Mitochondria
Eukaryotes
Prokaryotes
Xenoturbellida
Priapulida
α*, γ
Nematoda
Archaea
Arthropoda
Methanobacteria M
Thermoprotei T
A, B*, Fu, S, α, β, γ*, δ, ε, M
B, α, γ, δ, ε, S
Annelida
Bacteria
γ
Ectoprocta
Sipunculida
γ*, ε
Mollusca
Rotifera
Platyhelminthes
Ctenophora
Cy, α, β
Cnidaria
Porifera
Ac, A, Cl, Cy, Fi, G, N, α, β, γ, δ, S, V, P, T
Acidobacteria Ac
Actinobacteria A
Bacteroidetes B
Chlamydiae Ch
Chloroflexi Cl
Cyanobacteria Cy
Fibrobacteres F
Firmicutes Fi
Fusobacteria Fu
Gemmatimonadetes G
Nitrospira N
α-Proteobacteria α
β-Proteobacteria β
γ-Proteobacteria γ
δ-Proteobacteria δ
ε-Proteobacteria ε
Spirochaetes S
Verrucomicrobia V
Ca. Endomicrobia E
Ca. Poribacteria P

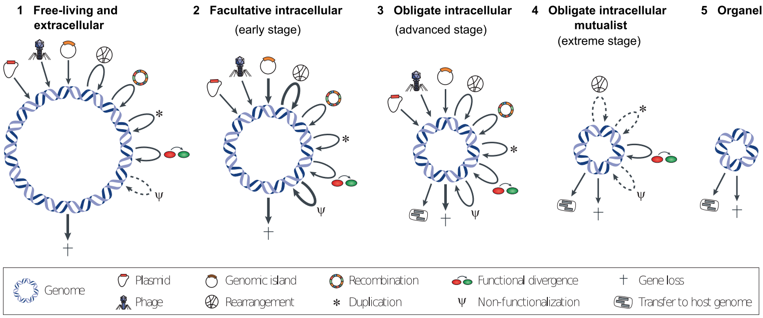
Figures 1.1.3 Phylogenetic tree showing eukaryotic hosts and their prokaryotic symbionts. Orange boxes indicate prokaryotic phyla. Asterisks denote genomes available at the time the review was published by Moya *et al.* (2008).

1.1.3. Genome reduction in endosymbiosis

Genome reduction is a common feature associated to the transition from free-living to endosymbiosis life-style. Two main causes are involved in this process. The relaxation of natural selection after endosymbiont arrival to a new stable environment, where an important amount of gene functions are not needed any more, and also, gene redundancy between the endosymbiont and the host, allow an extensive gene loss (Moya *et al.*, 2008). In addition, vertical transmission produces continuous bottlenecks in the endosymbiont's population that result in an increment of the effect of genetic drift and the associated Muller's ratchet effect[1] (Moran, 1996). These effects are also favoured by the progressive loss of genes involved in DNA repair, recombination and DNA uptake, avoiding the possibility of gaining exogenous genetic material (Silva *et al.*, 2003). This uptake is very limited in an endosymbiont because they are isolated inside the host and the only source are other endosymbionts (including themselves). The extreme cases of this process are the eukaryotic organelles (mitochondria and chloroplasts). Genome reduction can be divided in two phases: the first one occurs fast (in an evolutionary point of view) after the endosymbiosis takes place (step 1 to step 2 in Figure 1.1.4), and the second one starts after the engulfed bacteria reaches an obligate state of endosymbiosis (steps 3 and 4 in Figure 1.1.4) (Toft and Andersson, 2010).

---

[1] Endosymbionts become asexual populations by the loss of their recombination/reparation machinery and the bottlenecks result in the irreversible accumulation of deleterious mutations


1 Free-living and extracellular
2 Facultative intracellular (early stage)
3 Obligate intracellular (advanced stage)
4 Obligate intracellular mutualist (extreme stage)
5 Organelle
*
*
*
ψ
ψ
ψ
ψ
+
+
+
+
+
Genome
Plasmid
Phage
Genomic island
Rearrangement
Recombination
Duplication
Functional divergence
Non-functionalization
Gene loss
Transfer to host genome

Figures 1.1.4 Genome reduction process in endosymbionts during host adaptation and consequent co-evolution. Inner straight arrows indicate genetic material acquisitions, outer straight arrows indicate genetic material loss and looped arrows indicate internal genomic changes. Arrow thickness represents the importance of the process at each step (Toft and Andersson, 2010)

.

The first phase begins after the colonization of the new niche, in this case the eukaryotic cell, and involves a drastic process that can comprise as little as some millions of years (for an example of mitochondria evolution see Timmis *et al.* (2004)). It is characterized by the expansion and activation of different mobile elements, mainly Insertion Sequences (IS), that produce gene pseudogenizations (by IS insertions) and extensive amounts of genome rearrangements that can produce the loss of large genome regions (step 2 and in less manner step 3 in Figure 1.1.4) (Belda *et al.*, 2010; Gil *et al.*, 2008; Gillespie *et al.*, 2012; Parkhill *et al.*, 2003).

This reductive process plus the inability to acquire new genetic material lead to an irreversible genome size reduction. After this first step, the inactivation of the mobile elements plus the loss of the recombinatory machinery, reduces the rearrangement events to a minimum (step 3 in Figure 1.1.4) (Gil *et al.*, 2008; Penz *et al.*, 2012).

The second phase mechanisms are active during the first phase also, but are more evident when the processes of the last decrease their activity. At this point pseudogenization occurs mainly by accumulation of small insertion and/or deletion (indel) events, mutations or frameshifts and finally by a subsequent genome erosion[2] (Gómez-Valero *et al.*, 2004; Moran *et al.*, 2009; Silva *et al.*, 2001, 2003). Finally, while it seems that DNA from organelles are frequently transferred to the nucleus of the host by a mechanism known as Horizontal Gene Transfer (HGT), the integration of this DNA needs a long time of host-endosymbiont co-evolution to ensure the correct functionality of the transferred DNA (Timmis *et al.*, 2004). Moreover, although it seems that some barriers could be acting to avoid HGT from bacterial symbionts to the host nucleus, some cases have been reported (Husnik *et al.*, 2013; Sloan *et al.*, 2014). However, these barriers need to be explored because some reproduction manipulative endosymbionts, like *Wolbachia*, seem to bypass them probably due to their access to the germinal line (Brelsfoard *et al.*, 2014).

Genome reduction is associated in most cases with a higher Adenine (A) and Thymine (T) composition of the genomic sequence of the endosymbiont undergoing such reduction, with some exceptions like *Candidatus (Ca.)* Hodgkinia cicadicola and *Ca.* Tremblaya princeps (McCutcheon and Moran, 2012). Some possible non-exclusive, explanations have been proposed:

- Mutational bias: the higher spontaneous transversion rate of Guanine (G)/Cytosine (C) to A/T and the inactivation of a part of the DNA repair machinery lead to the increase in the A/T genome composition (Lind and Andersson, 2008).

---

[2] The pseudogene is no longer recognizable by homology and finally it is deleted

- Cost: the chemical structure of Guanosine triphosphate (GTP) and Cytosine triphosphate (CTP) are more costly to synthesize than the Adenosine triphosphate (ATP) or Thymidine triphosphate (TTP), so endosymbionts have a greater source of the last ones (Rocha and Danchin, 2002).
- Selection: maintains the G/C content at codon synonymous sites but the loss of selection in endosymbiotic bacteria could be the reason to the A/T bias (Hildebrand *et al.*, 2010).

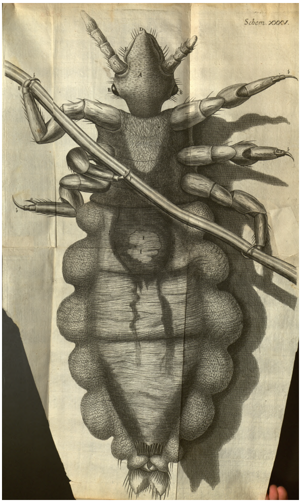
This A/T enrichment leads to the loss of a codon-usage bias in endosymbionts, which is common in free-living bacteria (Ermolaeva, 2001; Rispe *et al.*, 2004). This change in codon-usage seems to have altered and decreased the thermal stability of endosymbiotic proteins, a problem that could be ameliorated by the overexpression of GroEL (and maybe other chaperonin proteins). In fact, positive selection events have been detected in GroEL as an adaptation to the proteome instability problem in endosymbionts (Fares *et al.*, 2002, 2005)

Symbiosis in insects

The oldest fossil registry for insects are two Collembola fossils dated on the Devonian (*circa (ca.)* 370-400 million years (Myr)) and then radiated extensively during the Carboniferous (*ca.* 325 Myr) (Engel and Grimaldi, 2004; Wootton, 1981). Although a molecular study dated their origin a little earlier, during the Silurian (*ca.* 434 Myr), what is clear is that after Carboniferous' radiation insects became the most diverse animal taxon (actually, it is estimated that only 20 % of the species from *ca.* 5,000,000, are catalogued). Regardless of this broad biodiversity, it is interesting to denote that all insects have similar nutrient requirements for amino acids, vitamins, cofactors, and minerals. In fact, insects need a source of ten essential amino acids (the "rat essentials"[3] plus arginine) to maintain a 1:1 ratio of essential:non-essential amino acids (Chapman, 2013). During the Carboniferous' radiation, insects colonized a wide range of niches with a *ca.* 10 % of them based of unbalanced diets. Although it is known that insects can regulate their food uptake to fulfil their nutritional requirements, another way to achieve this goal is to establish different symbiotic associations with microorganisms, including endosymbiosis (Chapman, 2013). In fact, symbiotic association has been proposed as a key factor for insects' radiation succeed (Douglas, 1998).

---

[3] Phenylalanine, valine, threonine, tryptophan, methionine, leucine, isoleucine, lysine, and histidine



Figures 1.2.1 The first record of a bacteriome, reported by Hooke (1665), was from the human louse (*Pediculus humanus*). Bacteriomes, denoted by letter I, were confused with the louse kidney or pancreas.

Endosymbiosis in insects has been studied for many years and, although there are other orders that present many species with endosymbiotic bacteria, like the Blattodea or the Coleoptera orders (Bourtzis and Miller, 2003), Hemiptera seem to have received most of the attention (Baumann, 2005).

Endosymbionts from insects have been classified according to their mutual interdependency in obligate or Primary endosymbionts (P-endosymbionts) and facultative or Secondary endosymbionts (S-endosymbionts). While the first type is necessary for the survival of the insect and can not live outside its host, the second one is not needed for the survival of the insect and could not be present in all specimens.

Insects have developed specialized cells called bacteriocytes for harbouring P-endosymbionts that can form an organized tissue, the bacteriome (Figure 1.2.1). Usually, P-endosymbionts are isolated in the cytosol inside vacuoles and present a three-membranes system: the host's derived membrane and the two membranes from the bacterial gram-negative cell wall[4]. In contrast to P-endosymbionts, S-endosymbionts could present different localizations: sharing the same bacteriocyte with the P-endosymbionts, being in different bacteriocytes[5], distributed in different kind of tissues and cells, or freely in the haemolymph[6]. P-endosymbionts and S-endosymbionts are vertically transmitted from the mother to the offspring but S-endosymbionts can also present horizontal transmission between different host species (Chiel *et al.*, 2009; Koga *et al.*, 2012; Russell *et al.*, 2003).

Ectosymbionts, especially from the gut microbiota, seem to play different roles in food processing, nitrogen fixation, plant's secondary metabolites or toxins, but they are less studied (Engel and Moran, 2013; Fukatsu, 2012).

1.2.1. Primary endosymbionts

Most of the work on endosymbiosis has been done in the Sterrnorrhyncha suborder (Hemiptera), a group of phytophagous insects that feed on the plants' phloem or xylem. Phloem contains mainly sugars and non-essential amino acids (Douglas, 2006; Sandström and Pettersson, 1994), while xylem contains mainly inorganic compounds and minerals with small amounts of non-essential amino acids (Andersen *et al.*, 1989). Since the first sequenced P-endosymbiont, *Buchnera aphidicola* APS from *Acyrthosiphon pisum* (Shigenobu *et al.*, 2000), all P-endosymbionts sequenced from Sternorryncha (and from other insects that feed on unbalanced diets) are in charge of the insect's diet complementation supplying the lacking nutrients (amino acids, vitamins or cofactors) (Figure 1.2.2). There are omnivorous insects like cockroaches or carpenter ants that also harbour P-endosymbionts (*Blattabacterium cuenoti* and *Ca.* Blochmania spp. respectively). The sequence of their genomes have revealed that in addition to their role in diet complementation, similar to other endosimbionts, they have also a role in nitrogen recycling (Feldhaar *et al.*, 2007; Gil *et al.*, 2003; López-Sánchez *et al.*, 2009). In addition, P-endosymbionts that have evolved for a long time with their hosts usually present some characteristic features:

---

[4] Only one bacterial membrane is present in gram-positive P-endosymbiont

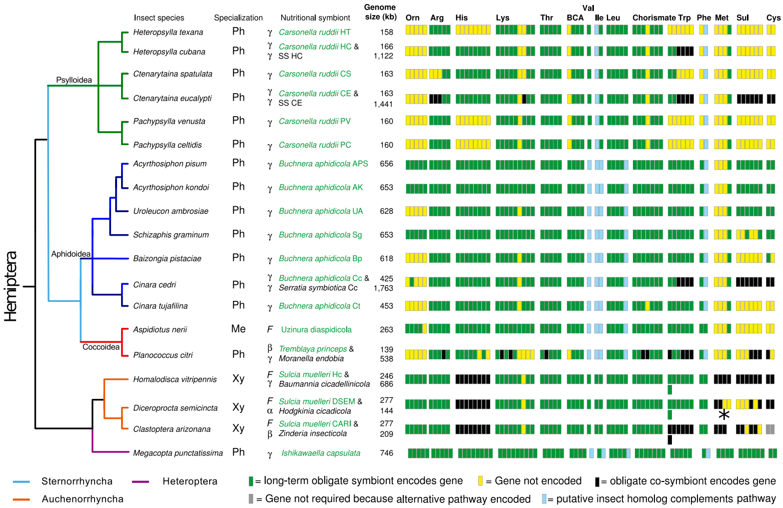
[5] Secondary bacteriocytes

[6] Some authors call them S-symbionts

- A reduced genome enriched in genes necessary to maintain the basic cellular functions and accomplish its symbiotic role within the host (Moya *et al.*, 2008; Shigenobu *et al.*, 2000).
- An A/T enriched genome (Moya *et al.*, 2008).
- No mobile elements (Shigenobu *et al.*, 2000).
- Minimal or no rearrangements (genome stasis) as a combination of mobile element loss and a minimal (or absent) recombination machinery (Latorre *et al.*, 2005; Patiño-Navarrete *et al.*, 2013; Silva *et al.*, 2003; Sloan and Moran, 2012b; Tamas *et al.*, 2002).
- Concordant phylogenies with its host (Moya *et al.*, 2008).

In some cases, these losses and the inability to complement the unbalanced diet of their hosts leads to the endosymbiont replacement by another endosymbiont with a less eroded genome. This is the case of some weevils where the long-term associated P-endosymbiont, *Ca.* Nardonella, has been replaced by a more recent one, *Ca.* Sodalis spp., or the several replacements that has suffered the P-endosymbiont *Ca.* Zinderia insecticola in spittlebugs (Conord *et al.*, 2008; Koga and Moran, 2014; Koga *et al.*, 2013; Lefèvre *et al.*, 2004; Oakeson *et al.*, 2014).

An intriguing result of genome reduction is present in some P-endosymbionts that have lost some essential amino acids, vitamins or cofactors biosynthetic pathways (see Psylloidea's endosymbiont *Ca.* Carsonella ruddii in Figure 1.2.2 as an example), or even part of their basic cell machinery (DNA replication, transcription and translation) (Gil *et al.*, 2004; Lamelas *et al.*, 2011a; Pérez-Brocal *et al.*, 2006). It is known that most of the cases where a metabolic complementation is needed, an endosymbiotic consortium has been established between a P-endosymbiont and a S-endosymbiont (Lamelas *et al.*, 2011b; Sloan and Moran, 2012b). Metabolic complementation can be at a complete



Figures 1.2.2 Sequenced long-term co-evolving P-endosymbionts and their metabolic capabilities. Plant tissue: Ph, phloem; Me, mesophyll and Xy, xylem. Endosymbiont bacterial taxa are represented by: ς, Gammaproteobacteria; β, Betaproteobacteria; α, Alphaproteobacteria; F, Flavobacteriales. Orn, ornithine; Arg, Arginine; His, histidine; Lys, lysine; Thr, threonine; BCA, shared branched-chain amino acid pathway (Val, Ile, Leu); Val, valine; Ile, isoleucine; Leu, leucine; chorismate, intermediate of the aromatic amino acids pathways (His, Trp, Phe); Trp, tryptophan; Phe, phenylalanine; Met, cobalamin-independent methionine pathway (starting from the homoserine intermediate); Sul, sulphur reduction and Cys, cysteine. ∗methionine cobalamin-dependet pathway requires the biosynthesis of vitamin B12 (genes not shown). Hosts phylogenetic relationships are based on Cryan and Urban (2012). Modified from Hansen and Moran (2014).

.

pathway level (each co-primary has the whole pathway for a certain amino acid or cofactor) or the pathway can be shared between both endosymbionts (Bennett and Moran, 2013; Gosalbes *et al.*, 2008; Lamelas *et al.*, 2011b; López-Madrigal *et al.*, 2013; Manzano-Marín and Latorre, 2014; McCutcheon *et al.*, 2009; Wu *et al.*, 2006). This special case of S-endosymbiont is considered a co-primary endosymbiont, because has established an obligatory mutualism with both the host and the P-endosymbiont, showing signatures common to the P-endosymbionts evolution. Also, there are cases where these relationships have been co-evolved for a long-time and both primary and co-primary presents an extreme reduced genome (see Auchenorrhyncha's endosymbionts in Figure1.2.2) (Bennett and Moran, 2013; Nakabachi *et al.*, 2013). However, these relationships not always are based on the amino acid, or co-factors, metabolic complementation and *Candidatus* Profftella armatura, a co-primary of *Ca.* Carsonella ruddii (hereafter *Carsonella*) in the psyllid *Diaphorina citri*, is an example of protective co-primary endosymbiont (Nakabachi *et al.*, 2013).

Recent works have revealed that some genes able to complement these losses are encoded in the host by self host's genes or by bacterial HGT genes (from the present endosymbiont or from other symbionts) (Husnik *et al.*, 2013; Sloan *et al.*, 2014). However, it is still unclear how the basic cell machinery losses are compensated. Other possibility is that some conserved proteins acquire new functionalities without losing their original ones (Kelkar and Ochman, 2013). The loss of the aforementioned functions in some P-endosymbionts have launched the question if these P-endosymbionts are closer to an autonomous cell or to an organelle (like *Carsonella* in Figure1.2.2) (Tamames *et al.*, 2007). In these context, the new term "symbionelle" was coined for these P-endosymbionts because their convergence with organelle evolution[7] (Reyes-Prieto *et al.*, 2014).

1.2.2. Secondary endosymbionts

Facultative or S-endosymbionts, in contrast to the P-endosymbionts, are not necessary for the survival or reproduction of the host. Usually, S-endosymbionts do not follow a strict vertical transmission and have horizontal transmission episodes producing not fully concordant phylogenies between the endosymbiont and the host. Also, genome reduction is not so marked as P-endosymbionts and the presence of mobile elements, phages, HGT events and rearrangements are common in S-endosymbionts genomes, maybe favouring their adaptation to new hosts (Duron, 2013; Ellegaard *et al.*, 2013; Gillespie *et al.*, 2012;

---

[7] With the difference that organelles evolved in a unicellular context while P-endosymbionts are evolving in a multicellular context

Moran *et al.*, 2008). Because vertical transmission is not ensured for S-endosymbionts, they can follow different strategies to guarantee their maintenance (Feldhaar, 2011), for instance:

- To increase the rate of horizontal transmission, so if the endosymbiont is lost in some host lineage they can recover it “jumping” from other host lineage.
- The host reproduction manipulation (e.g. male-killing, feminization, parthenogenesis, etc.) can ensure more females carrying the endosymbiont.
- Cytoplasmic Incompatibility (CI) can increase the fitness of infected females because they can produce viable offspring with infected and uninfected males while uninfected females can only do it with uninfected males.
- Direct increase of the host’s fitness can select the maintenance of the endosymbiont.

These strategies are not mutually exclusive. One reason is that fitness benefits are usually linked to the environment so when the selective pressure is not present, the S-endosymbionts, can be lost because the detrimental effects of their maintenance (Feldhaar, 2011; Ferrari and Vavre, 2011). A second reason is that reproductive manipulation can only be maintained for short periods of time. This is an effect of the effective population size reduction in the host. This reduction leads to a decrease of the host genetic diversity and finally the selection favours host’s alleles that counteract the endosymbiont manipulation (Ferrari and Vavre, 2011).

Although S-endosymbionts could be related to the complementation of the insect diet they can also affect the host’s fitness in other ways (reviewed in Oliver *et al.* (2010) and Ferrari and Vavre (2011)). Different experiments conducted in aphids have related some S-endosymbionts with different stress resistances:

- *Ca.* Hamiltonella defensa (hereafter *Hamiltonella*) seems to have an anti-parasitoid effect due to the toxins encoded in a phage (named as APSE) (Oliver *et al.*, 2008). Different *Hamiltonella* encode different APSE strains that are related to the level of protection against endoparasitoid wasps (Degnan and Moran, 2008).

- *Ca.* Regiella insecticola confers the aphid some protection against the parasitic fungus *Pandora neoaphidis* (Scarborough *et al.*, 2005). Although there are *Ca.* Regiella strains that can confer resistance to endoparasitoid wasps, this resistance is non phage dependent in contrast to *Hamiltonella* (Hansen *et al.*, 2012). Also, it seems that some *Ca.* Regiella strains in *A. pisum* can facilitate the shift to a new plant host (Tsuchida *et al.*, 2011).
- *Ca.* Serratia symbiotica (hereafter *S. symbiotica*) has two different lineages: one is present in the aphid *Cinara cedri* and *Cinara tujafilina* and is a co-primary endosymbiont (Lamelas *et al.*, 2011b; Manzano-Marín and Latorre, 2014), the second one is from *A. pisum* and confers certain grade of heat resistance (Montllor *et al.*, 2002).

While it is clear that the above mentioned S-endosymbionts provide some advantages to their hosts in a specific environment, it is not clear the effect of the reproductive manipulators endosymbionts. It is interesting that two of these manipulators are the most widespread arthropod endosymbionts: *Wolbachia* with a 40 % of prevalence while *Ca.* Cardinium hertigii (hereafter *Cardinium*) with a 16 %. Considering all the explained above, it is possible that manipulative endosymbiont (like *Wolbachia*, *Ca.* Arsenophonus or *Cardinium*), could produce some beneficial effects on the host to ensure their transmission and counteract the decrease in the host's genetic diversity they produce (Ferrari and Vavre, 2011).

Whiteflies

1.3.1. Whiteflies' biology

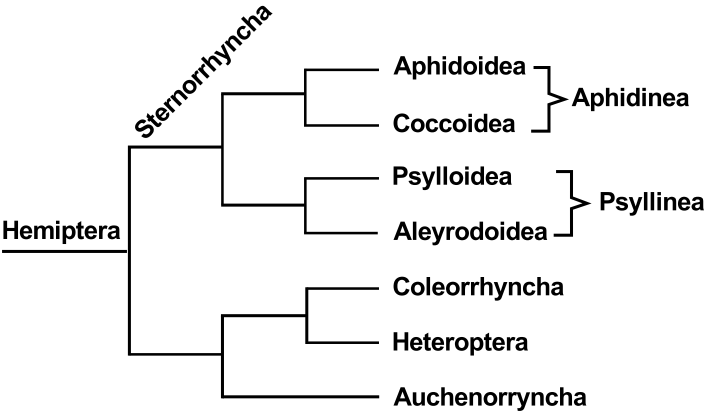
Hemiptera is an order of the class Insecta considered to be the largest group of hemimetabolous insects[8]. Their diversity seem to be related to angiosperm radiation. Hemiptera are classified in four suborders: Sternorrhyncha, Auchenorryncha, Heteroptera and Coleorrhyncha (Figure 1.3.1) (Cryan and Urban, 2012). They are characterized by a piercing mouthparts known as *rostrum*. The *rostrum* is composed by two stylets (formed by the mandibles and the maxillae) protected by a ribbed labium. All Hemipterans have a fluid diet: Sternorryncha, Auchenorryncha and Coleorrhyncha feed on plants' sap while some cases of predation are found in Heteroptera. In phytophagous hemipterans, the digestive system is adapted to this kind of diet and allows retaining nitrogenous compounds and other nutrients but quickly excreting the excess of sugar and water from the plant sap as "honeydew" (Grimaldi and Engel, 2005).

Molecular phylogenies grouped Sternorrhyncha suborder in four superfamilies: Aphidoidea, Coccoidea, Psylloidea and Aleyrodoidea (Cryan and Urban, 2012). Although molecular data usually give different topologies for the Sternorryncha (Campbell *et al.*, 1994), palaeontological studies supports two lineages, the Aphidinea and the Psyllinea (Figure 1.3.1) (Shcherbakov, 2000). The Aleyrodoidea superfamily (or whiteflies) is, with 1556 species, the less diverge among all the Sternorrhyncha (Martin and Mound, 2007). However, the later estimation could be incorrect because whiteflies research has taxonomical problems[9] and also most of the work has focused on crop pest species (Byrne and Bellows, 1991; Martin and Mound, 2007).

---

[8] Incomplete metamorphosis without pupal stage

[9] In contrast to other insects is only based on pupal stages

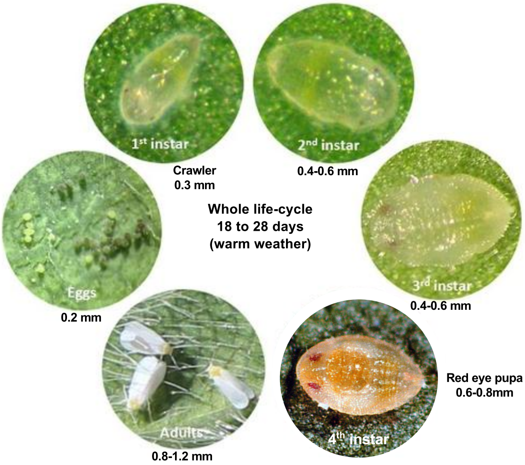


Figures 1.3.1 Hemiptera simplified phylogeny showing the four hemipteran suborders and the subdivision of Sternorrhyncha superfamily. Based on Campbell *et al.* (1994); Cryan and Urban (2012); Shcherbakov (2000).

Whiteflies have a paleotropical origin and are considered "the tropical equivalent of aphids" (Byrne and Bellows, 1991). It is possible that ancestral whiteflies feed on gymnosperms but it seems that their diversification was associated with the angiosperm radiation (Drohojowska and Szwedo, 2014). Adults are usually covered by a wax secreted by two pair of glands on the ventro-lateral part of the abdomen. The wax is distributed over the body by the whitefly using a set of combs placed on the hind legs. In nymphs, wax may appear as a gelatinous mass, or as different kind of projections (spike, cotton-like structures, etc.). Whiteflies present a unique structure among Sternorrhyncha, the vasiform orifice that is a dorsal depression with an operculum and a lingula where the anus finish. When this structure is filled by honeydew, the lingula catapult it away. This avoid fungal colonization of the nymphs and other problems related to the honeydew.

Whiteflies reproduction is mainly by arrhenotokous parthenogenesis with an X0 sex determination. Non-fertilized eggs produce males (X0), while females (XX) develop from fertilized eggs. Females attach the eggs to the leaf by a pedicel and a glue-like substance (see Figure 1.3.2).

When the egg cracks, the first-instar nymph (the only mobile instar) moves searching a minor vein and introduces its stylet into the phloematic tissue and continues its development. Second and third-instar are sessile and only increase in size. Fourth instar is known as "red eye pupa" because is a quiescent[10] nymph from which the adult emerges (Figure 1.3.2). Whiteflies' biology has been revised in Byrne and Bellows (1991), Grimaldi and Engel (2005) and Stansly and Naranjo (2010).


1st instar
Crawler
0.3 mm
2nd instar
0.4-0.6 mm
Whole life-cycle
18 to 28 days
(warm weather)
Eggs
0.2 mm
3rd instar
0.4-0.6 mm
Red eye pupa
0.6-0.8mm
Adults
0.8-1.2 mm
4th instar

Figures 1.3.2 Whiteflies' life cycle. Life cycle span can differ depending on whitefly species and climatic conditions. Modified from Surendra Dara blog at `http://ucanr.edu/blogs`. 4th instar photo by Paul de Barro.

Aleyrodoidea is composed by one family (Aleyrodidae). Whiteflies are formed by two extant subfamilies [11] that follow a West Gondwana-like distribution and an extinct one[12](Byrne and Bellows, 1991; Campbell *et al.*, 1994; Drohojowska and Szwedo, 2014; Martin and Mound, 2007). The two extant subfamilies are:

- **Aleyrodinae** subfamily. It groups most of the whiteflies in 91 non-synonymous genera, including the pest species *Bemisia tabaci* and *Trialeurodes vaporariorum*. Their body size is usually smaller (less than 2 mm) than Aleurodicinae and they have a worldwide distribution.

---

[10] Non-feeding nymph. Although it is not a true pupa because no metamorphosis occurs, this process is very different to the other hemimetabolous insects

[11] A third subfamily, the Udamoselinae, has been proposed but its phylogenetic position is still under discussion

[12] The Bernaeinae

- **Aleurodicinae**[13] subfamily. It is composed of 14 non-synonymous genera. Aleurodicinae usually have a bigger body size (greater than 2 mm) than Aleyrodinae and a Neotropical/Australasian distribution.

Molecular dating suggested that whiteflies origin could be in the Middle Cretaceous but the oldest fossil registry of a whitefly can be traced until the Upper Jurassic (Campbell *et al.*, 1994; Drohojowska and Szwedo, 2014). The first fossils of the present extant families were dated in the Early Cretaceous but it seems that whiteflies diversification started earlier, during the Late Jurassic probably associated to gymnosperm's forests (or with pro-angiosperms[14]). Angiosperms appeared during the Lower Cretaceous and radiated during Middle-Upper Cretaceous. Because most of the present whiteflies fed on angiosperms, it is possible that they changed from gymnosperm to angiosperms as host plants and this change triggered their diversification and originated the modern whiteflies (Drohojowska and Szwedo, 2014).

1.3.2. Endosymbionts of whiteflies

Although endosymbionts in whiteflies were described by Buchner (1965), their ultrastructure was firstly discussed in Costa *et al.* (1993) and extended in Costa *et al.* (1995). Two kind of endosymbionts were found in whiteflies' bacteriome, a predominant pleomorphic one shared by all whiteflies, and different coccoid-like ones that differed depending on the whitefly species. Coccoid-like endosymbionts were also detected in different tissues outside the bacteriome. This pleomorphic endosymbiont seemed to lack the cell wall, present in other endosymbionts, and in clear contrast to the coccoid-like symbionts (Costa *et al.*, 1993).

---

[13] Also referred as giant whiteflies

[14] Ancestors of angiosperms

The pleomorphic bacteria was designed as the P-endosymbiont of whiteflies and named as *Ca.* Portiera aleyrodidarum (hereafter *Portiera*) by Thao and Baumann (2004a) (Table 1.3.1). *Portiera* (Oceanospirillales:Halomonadaceae), together with *Carsonella* and *Ca.* Evansia muelleri (hereafter *Evansia*), forms an endosymbiotic group with *Halomonas elongata* and *Chromohalobacter salexigens* as their close free living relatives completely sequenced[15] (Santos-Garcia *et al.*, 2014b). This endosymbiotic group is related to *Pseudomonas*, in contrast to other insect P-endosymbionts that seem more related to the Enterobacteriaceae (Clark *et al.*, 1992; Thao and Baumann, 2004a). Although its origin, a nutritional role was suggested for *Portiera* due to phloem diet of whiteflies (Baumann, 2005).

Table 1.3.1 Endosymbionts identified in whiteflies. Adapted from (Stansly and Naranjo, 2010)

Type	Genus	Classification	Distribution	References
Primary	*Portiera*	γ	B*	48; 104; 288
	Hamiltonella	γ	B	44; 104; 190
	Arsenophonus	γ	B	48; 104; 190; 289
	Cardinium	Bacteroidetes	B/Sˆ	44; 104; 322
Secondary	*Wolbachia.*	α	B/S	48; 104; 289; 319
	Rickettsia	α	B/S	44; 103; 104
	Hemipteriphilus	α	B	20
	Fritschea	Chlamydiales	B	82; 291; 319

* Bacteriocyte ˆ Scattered through different tissues

Different coccoid-like organisms found in whiteflies were identified as S-endosymbionts[16] (Table 1.3.1). These S-endosymbionts could present two phenotypes regarding their distribution pattern in the host: they can share the same bacteriocytes as the P-endosymbiont *Portiera* (bacteriome-confined), or they can be found scattered across different tissues (including the hemolymph) (Gottlieb *et al.*, 2008). While *Hamiltonella*, *Arsenophonus*, *Hemipteriphilus* and *Fritschea* endosymbionts only present a bacteriome-confined phenotype, *Wolbachia*, *Rickettsia* and *Cardinium* present both phenotypes (references for each endosymbiont are in Table 1.3.1).

---

[15] Although the closest relative seems to be *Zymobacter palmae*, its genome is in at scaffolds level

[16] Most of the work is done on *B. tabaci* and *T. vaporariorum* but could be applied to other whiteflies

Scattered phenotype is usually associated with an early stage of facultative endosymbiosis, sometimes under a non-mutualistic relationship. In fact, *Hemipteriphilus*, *Chlamydia* and *Rickettsia* genera have pathogenic strains with this scattered phenotype. Also, *Cardinium*, *Wolbachia* and *Arsenophonus* genera are known as reproductive manipulators with scattered phenotypes in other hosts. Also, it is interesting to notice that, in *B. tabaci*, two *Rickettsia* patterns have been found and it could be possible that these patterns are due to two different *Rickettsia* strains[17] (Caspi-Fluger *et al.*, 2011).

However, not only parasitic endosymbionts present a scattered phenotype, as an example, *Hamiltonella* from *A. pisum* is recognized as a beneficial S-endosymbiont and present a scattered phenotype. The phenotypic transition from scattered to bacteriome-confined (e.g. *Hamiltonella*, *Arsenophonus*, etc.) could be related to the establishment of an obligate mutualistic relationship with the host and/or the P-endosymbiont environment in whiteflies. Moreover, the switch between scattered and bacteriome-confined phenotype could be and adaptation of the S-endosymbionts to the especial endosymbiont transmission mechanism in whiteflies.

1.3.3. Endosymbiont transmission in whiteflies

Whiteflies usually present a pair of orange/yellow roundish bacteriomes (Baumann, 2005; Buchner, 1965). However, there are some species that lack this specialized tissue and only present isolated bacteriocytes, usually in close relationship with the reproductive system (Coombs *et al.*, 2007; Szklarzewicz and Moskal, 2001). These bacteriocytes always harbour the P-endosymbiont *Portiera* and different S-endosymbionts (Costa *et al.*, 1993, 1995). Recent studies in *A. pisum* suggested that endosymbiont transmission is due to an exo/endocytosis mechanism between the bacteriocyte and the oocyte (Koga *et al.*, 2012). In contrast, whiteflies present a specialized mechanism for endosymbiont transmission different from other Sternorrhyncha[18]: some mother's bacteriocytes (number seems to depend on the species) migrate to the oocyte through the pedicele, ensuring endosymbiont transmission (Coombs *et al.*, 2007; Costa *et al.*, 1996; Szklarzewicz and Moskal, 2001). During this process the bacteriocyte and *Portiera* enlarge their shape. The apparently lack of cell wall in *Portiera* has been related to this process because a more flexible membrane is needed (Coombs *et al.*, 2007; Costa *et al.*, 1996; Szklarzewicz and Moskal, 2001). The bacteriocytes remain enclosed by the oocyte plasma membrane and do not enter in the ooplasm until the end of the oogenesis. At this point, it is unclear how the bacteriocyte, that has a maternal genome, integrates in the offspring development.

---

[17] One strain presents the scattered phenotype while the other the bacteriome-confined

Whiteflies used in this work

1.4.1. Aleyrodinae

1.4.1.1. *Bemisia tabaci*

B. tabaci, or the sweet potato whitefly, has a body length around 1 mm and can be identified by the more horizontal position of its wings (tent-like) compared to *T. vaporariorum*. It is distributed worldwide from tropical to subtropical temperatures and less expanded in temperate habitats. *B. tabaci* is one of the worst agricultural pests, being included in the 14*th* position of the 100 World's Worst Invasive Alien Species (`http://www.issg.org/`). Although it was considered that *B. tabaci* complex was composed of different biotypes[19], nowadays is accepted as a complex of species morphologically indistinguishable (cryptic). This classification is based on Mitochondrial Cytochrome C Oxidase subunit I (mtCOI) gene divergence and, until now, 31 lower groups, or species, has been described and grouped into 11 major groups (see Figure 1.4.2) (De Barro *et al.*, 2011; Lee *et al.*, 2013). In addition, species are composed by different haplotypes[20] that can be associated to a geographical origin. Also, haplotypes are divided into cytotypes, defined as an identical mtCOI haplotype associated with an endosymbiotic community (Gueguen *et al.*, 2010; Terraz *et al.*, 2014; Zchori-Fein *et al.*, 2014) (Figure 1.4.3). Finally, some endosymbionts can be involved in species formation due to their ability to manipulate their host's reproduction which produces a reproductive isolation (De Barro *et al.*, 2011). In conclusion, it is clear that more detailed phylogenetic analyses are needed in *B. tabaci* in order to solve the above mentioned problems.

---

[18] Cockroaches also show this transmission system

Figures 1.4.1 *B. tabaci*, courtesy of F. Beitia (IVIA).

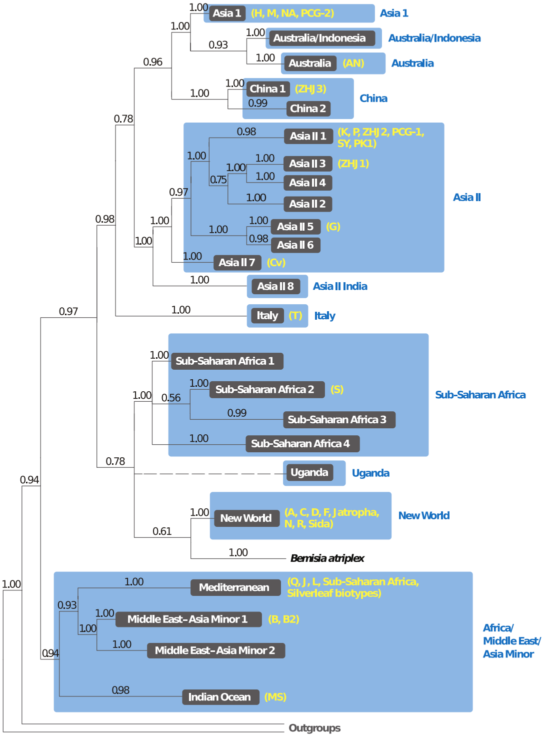
Among the 31 described species, two are the most invasive: the B biotype[21] or Middle East-Asia Minor 1 (MEAM1) and the Q biotype or Mediterranean (MED). While the B biotype (MEAM1) was the most widespread, the Q biotype (MED) has raised the invasive status due to its higher insecticide resistance (Stansly and Naranjo, 2010). Both species are able to produce important agricultural problems due to their polyphagy (can feed on more than 600 plant species), the direct physical damage due to the feeding action and the fungal infestations associated to the honeydew excreted by whiteflies. Another agricultural problem

---

[19] The same specie with different biological (phenotype) traits

[20] "Group of genes within an organism that was inherited together from a single parent" http://www.nature.com/scitable

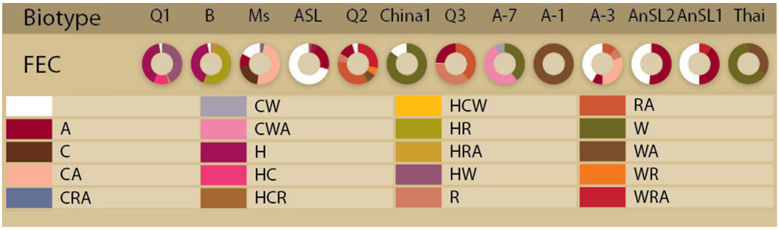
[21] In the present work the biotype nomenclature followed by the species definition in parentheses is used


1.00
Asia 1
(H, M, NA, PCG-2)
Asia 1
1.00
1.00
Australia/Indonesia
Australia/Indonesia
0.93
1.00
Australia
(AN)
Australia
0.96
1.00
1.00
China 1
(ZHJ 3)
0.99
China 2
China
0.78
0.98
Asia II 1
(K, P, ZHJ 2, PCG-1, SY, PK1)
1.00
1.00
Asia II 3
(ZHJ 1)
1.00
0.75
1.00
Asia II 4
0.97
1.00
Asia II 2
Asia II
0.98
1.00
1.00
1.00
Asia II 5
(G)
1.00
0.98
Asia II 6
1.00
Asia II 7
(Cv)
1.00
Asia II 8
Asia II India
0.97
1.00
Italy
(T)
Italy
1.00
Sub-Saharan Africa 1
1.00
Sub-Saharan Africa 2
(S)
1.00
0.56
Sub-Saharan Africa
0.99
Sub-Saharan Africa 3
1.00
Sub-Saharan Africa 4
0.78
0.94
Uganda
Uganda
1.00
New World
(A, C, D, F, Jatropha, N, R, Sida)
New World
0.61
1.00
Bemisia atriplex
1.00
1.00
Mediterranean
(Q, J, L, Sub-Saharan Africa, Silverleaf biotypes)
0.93
1.00
Middle East-Asia Minor 1
(B, B2)
1.00
1.00
Middle East-Asia Minor 2
Africa/ Middle East/ Asia Minor
0.94
0.98
Indian Ocean
(MS)
Outgroups

Figures 1.4.2 *B. tabaci* phylogenetic tree based on a mtCOI Bayesian analysis, with posterior probabilities displayed on the branches. *B. tabaci* species can be grouped in 11 high-level (blue boxes) and in 24 low-level (black boxes) groups. Biotype based nomenclature are listed in yellow inside parenthesis. Some biotypes could not be assigned at the time of this study. Reproduced from De Barro *et al.* (2011).

associated to *B. tabaci* is the large number of fitoviruses it can transmit (111 fitoviruses, in special *Begomovirus*) (reviewed in Navas-Castillo *et al.* (2011)).

Biotype Q (MED) is divided in four haplotypes, being the Q1 and Q2 the most widespread (Gueguen *et al.*, 2010; Terraz *et al.*, 2014; Zchori-Fein *et al.*, 2014) (Figure 1.4.3). While the Q1 cytotype is characterized for harbouring *Hamiltonella* usually combined with *Cardinium* or *Wolbachia*, the Q2 presents *Arsenophonus* and *Rickettsia*. In Q2 the presence of *Wolbachia* is sometimes detected.



Figures 1.4.3 Distribution of facultative endosymbiont combinations (FEC) representing natural assemblages of co-occurring S-endosymbionts in *B. tabaci* species. Abbreviations: A (*Arsenophonus*), H (*Hamiltonella*), C (*Cardinium*, W (*Wolbachia*) and R (*Rickettsia*). Modified from Zchori-Fein *et al.* (2014).

1.4.1.2. *Trialeurodes vaporariorum*

Figures 1.4.4 *T. vaporariorum*, courtesy of F. Beitia (IVIA).

T. vaporariorum, or greenhouse whitely, has a body length between 1-1.5 mm and its wings rest in a tent-manner making it distinguishable from *B. tabaci*. It is predominant in temperate regions and in greenhouses. Although few studies have been conducted on these whiteflies, detected endosymbiotic communities seem similar to *B. tabaci* and include most of the endosymbionts described in Table 1.3.1 (Skaljac *et al.*, 2010, 2012).

1.4.2. Aleurodicinae

1.4.2.1. *Aleurodicus dispersus*

Figures 1.4.5 *A. dispersus* (L. Buss, University of Florida).

A. dispersus body size range from 2-3 mm and due to the spiral forms of laying eggs with bits of wax interspersed, it is also known as spiralling whitefly (Russell, 1965). Nymph also produce complex wax structures and rods that can arise 8 mm and are a protection against natural enemies and pesticides. It is a polyphagous whitefly originated in the Caribbean and Central America region, but now is a serious pest in tropical and neotropical regions. Its distribution is more restricted than *B. tabaci* due to their less resistance to colder temperatures. As other whiteflies, its worldwide distribution seems an effect of human action. Its biology has been revised in Banjo (2010).

1.4.2.2. *Aleurodicus floccissimus*

A. floccissimus (formerly *Lecanoideus*) was firstly described in the Canary Island by Martin *et al.* (1997). The morphology of this whitefly is similar to the *A. dispersus* with which it shares even the same host plants. At the present time, these two whiteflies are an important pest in Canary Islands. However it seems possible to distinguish them based on molecular techniques like RAPD-PCR (Callejas *et al.*, 2005). Because it is a recent discovery, little is known about this whitely.

Figures 1.4.6 *A. floccisimus*, courtesy of F. Beitia (IVIA).

Part 2

Objectives

"Before the Gods of Hell sentence you to die
Remember well my friend a warlord never cries
These are the words that I've heard inside my mind
When Ragnarok comes down we'll all run out of time"

Manilla Road

This work is part of a research program with the aim to elucidate the evolution of endosymbiotic bacteria using the insect-bacteria consortia as a model. The different studies carried out on insects and their endosymbiotic bacteria have shed light on intracellular live-style changes and how different endosymbionts can interact forming endosymbiotic communities.

Two are the main goals of this work. The first (Chapters *Portiera* and its partner *Hamiltonella* and The third passenger: *Cardinium* cBtQ1) tries to analyse and describe the endosymbiotic community relationships in *B. tabaci* using a laboratory strain as a model. The second (Chapter Genome evolution of the genus *Portiera*) describes the evolution of the P-endosymbiont *Portiera* in the whiteflies and their use for molecular dating. These two goals can be divided according to:

1. ***Portiera* and its partner *Hamiltonella*.**
    - To confirm the distribution patterns of the *B. tabaci* QHC-VLC strain endosymbionts.
    - To revisit the ultraestructure of *Portiera*.
    - To sequence, annotate and analyse the *Portiera* BT-QVLC genome and infer its relationship with *B. tabaci*.
    - To sequence, annotate and analyse the *Hamiltonella* BT-QVLC genome and infer its relationship with *B. tabaci*.
    - To analyse the possible metabolic integration between both endosymbionts.
2. **The third passenger: *Cardinium* cBtQ1.**
    - To sequence, annotate and analyse the *Cardinium* cBtQ1 genome and infer its relationship with *B. tabaci*.
    - To establish the phylogenetic relationships of *Cardinium* cBtQ1 with other *Cardinium* strains and Bacteroidetes.
    - To compare the genomes of *Cardinium* cBtQ1 and cEper1, *Amoebophilus asiaticus*, and other sequenced Bacteriodetes.
    - To infer and analyse the gene content and its evolution among *Cardinium*, *Amoebophilus asiaticus*, and some related

Bacteroidetes.

- To infer the metabolism of *Cardinium* cBtQ1 and the relationship with *Portiera* and *Hamiltonella* BT-QVLC.
- To discuss the possible function of the gliding genes present in *Cardinium* cBtQ1.

3. **Genome evolution of the genus *Portiera***.
    - To sequence, annotate and analyse the *Portiera* strains genomes from *T. vaporariorum*, *A. dispersus*, and *A. floccissimus* whiteflies.
    - To compare the genomes and the metabolism of the *Portiera* strains sequenced in this work.
    - To estimate the divergence time of *Portiera* strains and their hosts.
    - To analyse the molecular evolution in *Portiera* genus.

Part 3

Material and Methods

“I don’t like people working all day
Only working just to see next tomorrow
But they are happy while they’re living that way
In that world I gotta beg, steel or borrow”

Accept

Whiteflies samples

Four whiteflies species have been analysed in this work: *B. tabaci*, *T. vaporariorum*, *A. dispersus* and *A. floccissimus*. *B. tabaci* is a homogeneous laboratory strain of Q1 haplotype that has been maintained for more than 5 years in laboratory conditions. In addition to *Portiera*, the P-endosymbiont, this strain harbours the S-endosymbionts *Hamiltonella* and *Cardinium*. The strain was named as QHC-VLC attending to the *B. tabaci* biotype, its S-endosymbionts and the localization of the laboratory (Valencia, Spain). A climatic chamber adjusted to 26°C, 60 % humidity and 12 hour (h) of light photoperiod was used for insect maintenance. Cotton (*Gossypium hirsutum*) was selected as host plant and it was grown in separated cages at the same chamber to avoid insect contamination from field populations.

T. vaporariorum was collected from a field population in Catalonia (Spain). The captured population harboured *Arsenophonus sp.* and *Wolbachia sp* as S-endosymbionts. This population was named as TVAW-BCN following the above reasoning.

Samples from *A. dispersus* and *A. floccissimus* were collected from crop fields in the Canary Islands. Both samples harboured the S-endosymbionts *Arsenophonus sp.* and *Wolbachia sp.* and were named as ADAW-CAI and AFAW-CAI respectively.

Microscopy techniques

3.2.1. Transmission Electron Microscopy

Whole *B. tabaci* eggs and nymphs were collected with a water-floss device in a mesh and briefly cleaned with 70 % ethanol and distilled water for taking out the whiteflies wax. Eggs and nymphs were fixed separately in Karnowsky's fixative (2 % paraformaldehyde and a 2.5 %

glutaraldehyde in 0.1 molar (M) cacodylate buffer at 7.2 pH) with 5 steps of 1 minute (min) in a vacuum pump and left for overnight (O/N) fixation at 4°C. Samples were washed and postfixed in 2 % OsO_4 for 2 h. After postfixation, samples were washed, dehydrated through ethanol series (30, 50, 70, 90 and 100), passed to propylene oxide and embedded in LR White resin. Resin blocks were cut in a Leica Ultracut EM UC6 (60-90 nm sections) and grids were contrasted with 2 % uranyl acetate and Reynolds' lead citrate. Pictures were taken with a JEOL JEM-1010 Transmission Electron Microscope (TEM) at 80kV.

For membrane measurements, two samples of *B. tabaci* QHC-VLC nymphs were collected on different days and fixed. Three different images from each sample clearly showing the *Portiera* cell wall were used to measure the membrane components. Five measurements were taken for each membrane component from each picture with **Fiji** (Schindelin *et al.*, 2012).

3.2.2. Fluorescent *in situ* hybridization

B. tabaci nymphs were collected with a water-floss device in a mess. Fluorescent In Situ Hybridization (FISH) procedure was followed as described by Gottlieb *et al.* 2006. Nymphs were directly transferred into modified Carnoy's fixative (6 chloroform:3 absolute ethanol:1 glacial acetic acid) and left O/N. Fixed nymphs were washed with ethanol and transferred to a 6 % solution of H_2O_2 (in ethanol) for at least two hours. Hybridization was performed O/N at Room Temperature (RT) in standard hybridization buffer (20 mM Tris(hydroxymethyl)aminomethane (Tris)-HCl [pH 8.0], 0.9 M NaCl, 0.01 % Sodium Dodecyl Sulfate (SDS), 30 % formamide) and washed (20 mM Tris-HCl [pH 8.0], 5mM Edetic acid (EDTA), 0.1 M NaCl, 0.01 % SDS) before slide preparation. Whole nymphs were viewed under an Olympus FV1000 confocal microscope. FISH specific probes for *Portiera*, *Hamiltonella* and *Cardinium* are listed in Table 3.2.1.

Table 3.2.1 16S rRNA FISH probes used in this work

Endosymbiont	Probe Name	Sequence	Dyes	Reference
Portiera	BTP1	TGTCAGTGTCAGCCCAGAAG	FAM	103; 104
Hamiltonella	BTH	CCAGATTCCCAGACTTTACTCA	Cy3	103; 104
Cardinium	Card	TATCAATTGCAGTTCTAGCG	Cy5	171

B . *tabaci* nymphs treated with RNase, non-probe controls and nymphs without *Cardinium* (B biotype with *Hamiltonella* and *Rickettsia*) were used as specificity probe controls. **Icy** software was used for FISH image channels analyses and composition of final images (de Chaumont *et al.*, 2012).

Endosymbiont enriched samples

3.3.1. Bacterial enriched samples

A modified protocol based on the method of Harrison and co-workers (Harrison *et al.*, 1989) was used for obtaining bacterial-enriched samples. This protocol allows the elimination of insect tissues and cell debris but maintaining an intact endosymbiont cell envelope. This is accomplished by the filtration steps after grinding the insect. The Dounce tissue grinder ensures the disruption of the insect cells without damaging the endosymbionts[22]. A DNase digestion step reduces the insect genomic DNA that could remain in the sample but maintaining the endosymbiont intact because is not able to degrade the DNA protected by the cell envelope. A relatively enriched sample in endosymbionts is finally obtained.

Around 40,000 adults from the *B. tabaci* strain QHC-VLC were collected. Whiteflies were briefly washed with 70 % ethanol and washed three times with Phosphate Buffer Saline (PBS) (137 mM NaCl; 2.7 mM

---

[22] This grinder has a space of 50 µm between the mortar and the pestle

KCl; 10 mM Na_2HPO_4; 2 mM KH_2PO_4; pH 7.4). Then, whiteflies were disrupted in 1 ml of pre-cooled Ringer-Krebs buffer (Sigma-Aldrich) with a Dounce homogenizer. The homogenate was filtered through decreasing pore size nylon membranes (twice each step): 1mm, 80 µm, 60 µm, 20 µm, 11 µm and 5 µm. The filtered homogenate was centrifuged at 8000 revolutions per minute (rpm) for 15 min in a pre-cooled centrifuge. The supernatant was discarded and the pellet re-suspended in 1 ml of Ringer and centrifuged three times at same conditions. The final pellet was re-suspended in 250 µm of Ringer plus 50 µm of 10x TURBO® DNase I Buffer and 5 µl (2 Units/µl) TURBO® DNase I (Ambion). The mix was stopped (30 µ of Inactivation Reagent) after a 30 min at 37°C incubation step. A final centrifugation (same conditions as above) was made for precipitate bacterial cells. This pellet was used directly for DNA extraction. All steps were made on ice or at 4°C if no temperature is indicated.

3.3.2. Bacteriome extraction and DNA amplification

Microneedles were made pulling glass capillaries (Drumond®) with a PC-100 Puller (Narishige®). The microneedles were adjusted to a microcap. Single bacteriomes were extracted from 4°th instar larvae (red eyes) by pricking in the bacteriome and taking up it with the microcap. The bacteriome was transferred to a 0.2 ml Polymerase Chain Reaction (PCR) tubes containing 10 µl of fresh made lysis solution (400 mM KOH, 10 mM EDTA, 100 mM Dithiothreitol (DTT)) and left on ice 10 min to obtain the genomic DNA (gDNA). Lysis solution was neutralized with an equal volume of fresh made neutralization buffer (400 mM Hcl, 600 mM Tris-HCl pH 7.5). The manufacturer's protocol was followed for the Whole Genome Amplification (WGA) reaction (GenomiPhi V2®, GE Healthcare) and the reaction mix was added to the PCR tube (7 µl Sample Buffer, 9 µl Reaction Buffer and 1 µl Enzyme Mix). Amplification reaction profile was 30 °C for 90 min and 65°C for 10 min.

As a general procedure, 10 reactions (10 bacteriomes from different individuals) were used for each species, which were pooled to diminish the impact of the possible chimeras formed during WGA. Pooled samples were sequenced by different Next Generation Sequencing (NGS) platforms.

DNA extractions, PCR reactions and quantification

3.4.1. Genomic DNA extraction

gDNA from endosymbiont enriched samples was extracted with the JetFlex® Genomic DNA purification kit following the manufacturer's instructions (Genomed). This kit is based on a salt precipitation procedure. It starts with the disruption of the cells by a combination of SDS for cell's membranes breakage and Proteinase K for tissue/potein digestion[23]. SDS is a surfactant that binds to the proteins/lipids and is precipitated by the addition of acetate (CH_3COOK or CH_3COONa) at high concentrations (3-5 M). The mixture is left for 10-30 min preferable at -20°C and centrifuged at maximum speed for pelleting the SDS-acetate complex. The supernatant is transferred to a new tube and precipitated by adding an equal volume of isopropanol[24], left for 10 min at RT and centrifuged at maximum speed for 15 min. The isopropanol is discarded, cold 70 % ethanol is added for extract salt excess and is centrifuged at maximum speed per 15 min at 4°C (two times). Finally, the gDNA pellet is left to air-dry the ethanol rests and resuspended in Tris-EDTA (TE) (10 mM pH 8 Tris, 1 mM EDTA) or Low TE (LTE) (10 mM pH 8 Tris, 0.1 mM EDTA) buffer or miliQ water depending on the downstream analyses[25].

---

[23] It is probable that the extraction buffer contains EDTA because it inhibits DNases action by chelating Mg^{2+}

[24] 2 volumes of cold absolute ethanol can be used. In this case, the procedure needs to be followed on ice and cooled centrifuge

[25] EDTA is a cation chelator and can inhibit PCR based methods because polymerases uses Mg^{2+}

3.4.2. Chelex DNA extraction

Chelex® DNA extraction method is based on the ability of this resin to bind cellular components and chelate cations like Mg^{2+}. This method is easy and fast and can be used for general PCR applications, but is not suitable for more accurate analysis or NGS sequencing. The sample is placed in a tube and grinded in presence of 5 %-10 % Chelex® (in miliQ water), incubated 20 min at 65°C and 20 min at 99°C (Walsh *et al.*, 1991). Proteinase K can be added for ensuring a better sample digestion. The extraction is then centrifuged at 5000 rpm for 5 min and can be stored until use at 4°C for 3-4 months or at -20°C for longer periods. For PCR is necessary to avoid taking Chelex® particles because their inhibitory effect.

3.4.3. PCR amplification

Standard PCR amplifications were performed on different parts of the work. A general PCR profile[26] was used most of the time only adjusting the Melting Temperature (Tm) according to each set of primers. Primers were designed with PrimerQuest tool from IDT (`http://eu.idtdna.com/PrimerQuest/Home/Index`). Wherever it was possible, all primers were designed with an optimal TM of 60°C. Primers used in this work can be found in the Annex Table A3.1. When a highest sensitivity was required, the LightCycler 2.0 (Roche) was used with the following PCR profile: 15 min of denaturalization step (95°C), 40x[95°C for 10 s, 58°C for 20 s and 72°C for 20 s] and a melting curve step (68°C to 95°C with a ramp rate of 0.2°C each second). LigthCycler FastStart DNA MasterPLUS SYBR Green I (Roche) mix was used as manufacturer recommendation. Melting curves were inspected to detect false positives amplifications (e.g. primer-dimer amplifications). All PCR amplicons were visualized by gel electrophoresis stained with ethidium bromide. When standard PCR amplification gave more than one product, a colony PCR was used for obtaining single PCR amplicons suitable for sequencing. An *Escherichia coli* DH5α strain was used in combination with the pGEM®-T Easy Vector System (Promega) for cloning the PCR product by electroporation following manufacturer recommendations. Transformants *E. coli* colonies were used for a direct colony PCR: a colony is picked by a sterilized toothpick and is gently stirred in a PCR tube with the PCR mix[27]. Finally, a standard PCR profile[28] is followed and the PCR product is purified and used for sequencing. Different colonies were amplified at the same time for having a representation of all the PCR amplicons cloned.

---

[26] 95°C for 2 min, 30x[95°C for 30 second (s), XX°C for 30 s, 72°C for 1min/1kb] and 72°C for 5 min

3.4.4. DNA purification and quantification

PCR amplicons were examined by gel electrophoresis. When only one single band was obtained the PCR product was directly purified with NucleoFast®96 PCR (Macherey-Nagel) following the manufacturer instructions. When more than one band was detected the wider band was cut and purified with the SpinPrep® Gel DNA Kit (Millipore). The purified products can be used for further analyses. Two methods for nucleic acid quantification were used:

- Spectrophotometry: Nanodrop® ND-1000 measures the sample's absorbance using an ultraviolet light. While nucleic acids absorb at 260 nm, proteins absorb at 280 nm. There are contaminants (phenols, co-purified salts) that can absorb at 230 nm and 270 nm. A 260/280 ratio of $\approx$ 1.8[29] indicates an almost pure nucleic acids sample. The contamination grade of the sample (organic solvents, co-purified salts, etc.) is given by the 230/260 ratio that must relay between $\approx$ 1.8-2.0. This method usually gives an overestimation of the nucleic acids amounts. While it can be used as routinely measurement protocols, for a more sensitive task a fluorimetric approach is recommended.

---

[27] It is dependent on commercial supplier but contains the DNA polymerase and its buffer, dNTPs, milliQ water, and SP6 and T7 promoter primers

[28] 95°C for 6 min, 30x[95°C for 30 s, 55°C for 30 s, 72°C for 3 min] and 72°C for 5 min

[29] A ratio of $\approx$ 2.0 for RNA

- Fluorometry: fluorescence emitted by nucleic acids intercalating dyes is measured by a fluorometer (in our case a Qubit® 2.0). Nowadays, Picogreen® is the most used fluorescent dye in DNA quantification and has become a standard prior to NGS sequencing[30]. Because only binds specifically to one conformational nucleic acid (double-stranded DNA, single-strand DNA, or RNA) the accuracy is better than spectrophotometry methods that measure all kinds of nucleic acids in the sample. Also the measurements are less impacted by contaminants.

Genome Sequencing

3.5.1. Sanger sequencing

PCR amplicons were marked with the BigDye® Terminator v3.1 kit (Applied Biosystems) following manufacturer instructions. The marking reaction was performed O/N in a PCR thermocycler[31]. Electroforetic capillary sequencing was carried on an ABI 3730, at a facility of the University of Valencia (SCSIE). Trace files generated by the ABI sequencer were used as input for the Staden Package: **Trev** was used to check the quality of the trace file, **Pregap4** for pre-process the trace files prior to build a **Gap4** database, and **Gap4** for read/contig editing and to build the final consensus sequence (Staden *et al.*, 2000).

3.5.2. Next Generation Sequencing (NGS)

Endosymbiotic genomes were sequenced using two NGS platforms: Genome Sequencer FLX+ (454 Life Sciences, Roche) and HiSeq2000 (Illumina). For a detailed review in NGS see Shendure and Ji (2008) and Metzker (2010). The libraries used were:

- *B. tabaci*: 1/2 single-end plate (shotgun) and a full paired-end plate

---

[30] Actually, different Picogreen® dyes can bind to RNA and proteins

[31] The PCR profile was: 95°C for 1 min and 99x[95°C for 10 s, 50°C for 10 s, 60°C for 4 min]

(3 kb of insert size) of GS FLX+ Titanium chemistry and a single lane of HiSeq200 mate-pair (5 kb of insert size). The GS FLX+ libraries were ordered to the Sequencing Facilities at the FISABIO (Valencia, Spain). The HiSeq library was ordered to Macrogen Inc. (Seoul, Republic of Korea).

- *T. vaporariorum*, *A. dispersus* and *A. floccissimus*: for each species a 1/4 single-end plate (shotgun) of GS FLX+ Titanium chemistry were ordered to the FISABIO. In addition, each species was tagged prior to library construction and a single multiplexed lane of HiSeq200 mate-pair library (5 kb of insert size) was ordered to Macrogen Inc.

Non processed sequences were received from the FISABIO and Macrogen. GS FLX+ platform generates a proprietary format called Standard flowgram format (SFF) while HiSeq results are delivered in FASTQ[32].

Genome assembly and annotation

3.6.1. Assembly

De-novo genome assembly is a complex field that is still in continuous improving. Basically, a chromosomal DNA molecule is fragmented and sequenced using different sequencing platforms. Assemblers take these DNA fragments, or reads, and try to reconstruct the original chromosome. For this issue, two kind of data are used: single reads or shotgun libraries, or paired reads or pair-ended libraries [33]. In the pair reads, the edges of a longer DNA fragment are sequenced, but not the central region, that is called the “insert”, giving positional information. When different reads overlap, their consensus sequence forms a “contig”. Contigs can be linked by the pair reads information to produce a gapped supercontig or “scaffold” (Baker, 2012). Assemblers are usually programmed to deal with isolated genomes and rely basically on three kind of algorithms (Pop, 2009):

---

[32] An example of these formats can be viewed here: `http://bioinf.comav.upv.es/courses/sequence_analysis/sequence_file_formats.html`

[33] Also called mate-pair depending on the NGS platform and the preparation technique

- Greedy: it is the more basic but the most expensive in computational terms. It starts with a read and adds more reads until no more can be added.
- Overlap-Layout-Consensus (OLC): basically computes all pairwise alignments between the input reads and search for overlaps. It uses large amount of computer resources, but less than the greedy algorithm, and needs reads larger than 100 base pairs (bp) to compute reliable overlaps. It is not suitable for short reads due to the large amount of data and the length of the reads. **Newbler** and part of **MIRA**[34] assembler code are based in this algorithm (Chevreux *et al.*, 1999).
- de Brujin graph: reads are decomposed on oligomers of k length (K-mer) that are used to construct the edges of a de Brujin graph. Finally the assembler tries to find a path that pass through this graph using every edge in the graph (Eulerian path)[35]. Because this algorithm relies in perfect k-mer matches, only technologies with low sequencing errors (e.g. Illumina) can be used. **Velvet**, **Celera** and **SOAPdenovo** are included in de Brujin graph assemblers.

De-novo genome assemblers deal with problems that interfere in the assembly process: sequencing errors, repeats (i.e. mobile elements), erroneous joins (chimeras), polymorphisms, etc.... For this reason, *de-novo* genome assembly of genomes that come from metagenomes are even less straightforward. Assemblage of metagenomes needs to confront two new problems: the diversity and the different abundance of each organism in the sample.

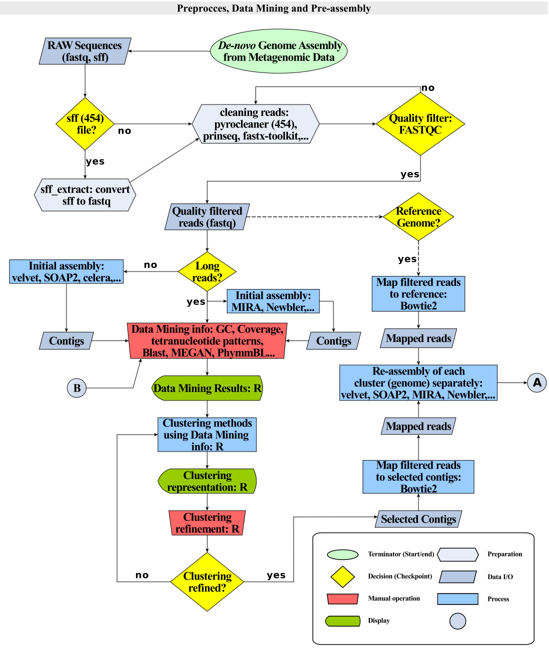
To alleviate the last problem a “divide and conquer” pipeline was implemented to assembly the different whiteflies endosymbiotic genomes. This pipeline is divided in three main steps and the software used is listed in Figure 3.6.1, 3.6.2 and 3.6.3:

---

[34] **MIRA** is one of the few assemblers that can perform hybrid assemblies with most of the current NGS platforms

[35] For an extensive review of the different algorithms see Miller *et al.* (2010)

1. Pre-process, Data mining and Pre-assembly (Figure 3.6.1): in this part the raw reads are trimmed and quality filtered to avoid several sequencing problems (linkers, adapters, bad quality regions, etc.). An initial assembly is performed and the contigs/reads are used for a data mining step that uses different sources of information (**BLAST** similarities, GC content, **PhymMBL** (Brady and Salzberg, 2009), contig coverage, etc...) to group the contigs/reads[36]. Each group of contigs could belong to the genome of the same organism and be used for a posterior re-assembly by mapping and selecting cleaned reads. If a reference genome is available, it can be used as another source for read selection.

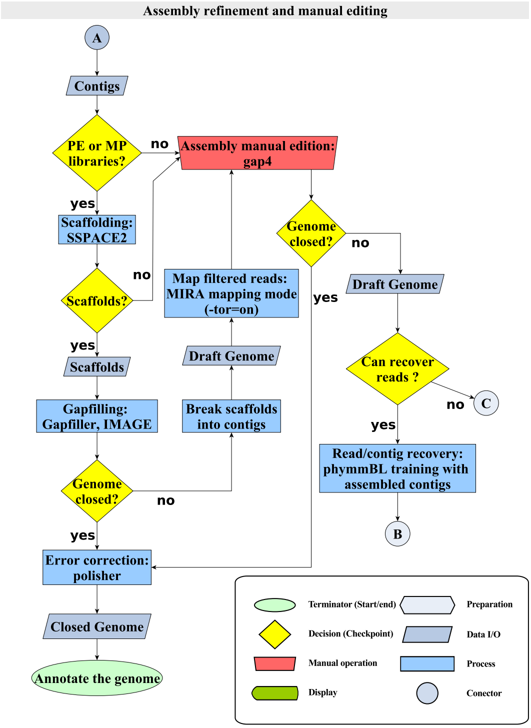

Preprocces, Data Mining and Pre-assembly
RAW Sequences (fastq, sff)
De-novo Genome Assembly from Metagenomic Data
no
sff (454) file?
no
cleaning reads: pyrocleaner (454), prinseq, fastx-toolkit,...
Quality filter: FASTQC
yes
sff_extract: convert sff to fastq
yes
Quality filtered reads (fastq)
Reference Genome?
yes
Initial assembly: velvet, SOAP2, celera,...
no
Long reads?
yes
Initial assembly: MIRA, Newbler,...
Map filtered reads to reference: Bowtie2
Contigs
Data Mining info: GC, Coverage, tetranucleotide patterns, Blast, MEGAN, PhymmBL...
Contigs
Mapped reads
B
Data Mining Results: R
Re-assembly of each cluster (genome) separately: velvet, SOAP2, MIRA, Newbler,...
A
Clustering methods using Data Mining info: R
Mapped reads
Clustering representation: R
Map filtered reads to selected contigs: Bowtie2
Clustering refinement: R
Selected Contigs
no
Clustering refined?
yes
Terminator (Start/end)
Preparation
Decision (Checkpoint)
Data I/O
Manual operation
Process
Display

Figures 3.6.1 Pre-process, Data mining and Pre-assembly steps of the *de-novo* general assembly pipeline used with the whiteflies metagenomic samples to isolate the endosymbiotic genomes.

---

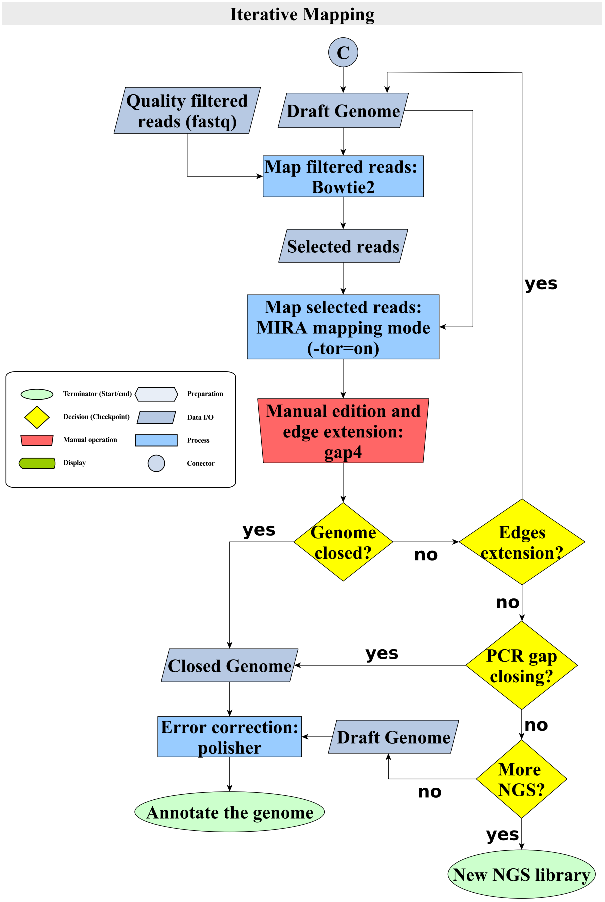
[36] An example of data mining can also be found in Albertsen *et al.* (2013)

2. Assembly refinement and manual editing (Figure 3.6.2): if paired end or mate-pair libraries were available, a scaffolding and gapfilling step was performed. If no pair information existed or the genome was not closed, then a manual joining step was performed. The manual edition relies on a mapping step of the draft genome with **MIRA** and the cleaned reads and its posterior edition on **Gap4** (Staden *et al.*, 2000). **MIRA** introduces information on the assembly about region repetitiveness and pair reads, which can be used for the manual joining step in **Gap4**. Also an additional step for trying to recover more reads that were not detected on the initial step can be performed.


Assembly refinement and manual editing
A
Contigs
PE or MP libraries?
no
Assembly manual edition: gap4
yes
Scaffolding: SSPACE2
no
Scaffolds?
yes
Scaffolds
Gapfilling: Gapfiller, IMAGE
Genome closed?
no
yes
Error correction: polisher
Closed Genome
Annotate the genome
Map filtered reads: MIRA mapping mode (-tor=on)
Draft Genome
Break scaffolds into contigs
Genome closed?
no
yes
Draft Genome
Can recover reads ?
no
C
yes
Read/contig recovery: phymmBL training with assembled contigs
B
Terminator (Start/end)
Preparation
Decision (Checkpoint)
Data I/O
Manual operation
Process
Display
Conector

Figures 3.6.2 Assembly refinement and manual editing steps of the *de-novo* general assembly pipeline used with the whiteflies metagenomic samples to isolate the endosymbiotic genomes.

3. Iterative mapping (Figure 3.6.3): is a looping process designed for manual scaffolding and gapfilling and to recover new reads. **MIRA** is used to extend the edges and **Gap4** for manual edition of the assembly. The mapping step is repeated until the genome is closed or no more reads are recovered.


Iterative Mapping
C
Quality filtered reads (fastq)
Draft Genome
Map filtered reads: Bowtie2
Selected reads
Map selected reads: MIRA mapping mode (-tor=on)
yes
Terminator (Start/end)
Preparation
Decision (Checkpoint)
Data I/O
Manual operation
Process
Display
Conector
Manual edition and edge extension: gap4
yes
Genome closed?
no
Edges extension?
no
Closed Genome
yes
PCR gap closing?
Error correction: polisher
Draft Genome
no
More NGS?
no
Annotate the genome
yes
New NGS library

Figures 3.6.3 Iterative mapping steps of the *de-novo* general assembly pipeline used with the whiteflies metagenomic samples to isolate the endosymbiotic genomes.

3.6.2. Annotation

As a general outline, the annotation pipeline for each genome was as follows[37]:

Initial Open Reading Frames (ORFs) predictions were performed with **Prodigal** (Hyatt *et al.*, 2010) and uploaded to the annotation servers **BASys** (Van Domselaar *et al.*, 2005) and **RAST** (Aziz *et al.*, 2008). Manual refinement of the annotation of the Coding DNA Sequence (CDS) was made using **Artemis** (Rutherford *et al.*, 2000) to integrate the information from several databases: **Pfam** (Punta *et al.*, 2012), **Uniprot** (The UniProt Consortium, 2012), **Interpro** (Hunter *et al.*, 2012), **BLAST** and **CDD** (Marchler-Bauer *et al.*, 2011) and **PHAST** (Zhou *et al.*, 2011).

Specific protein domains searches were conducted with **HMMER** using the **Pfam** Markov models (Eddy, 2011). **InterProScan** (Jones *et al.*, 2014) was used for **Gene Ontology (GO)** (Ashburner *et al.*, 2000), **TIGRFAM** (Haft *et al.*, 2003) and **Pfam** terms assignation to the annotated CDS. Transmembrane domains were predicted with **TMHMM2.0** (Käll *et al.*, 2007). **Cluster of Orthologous Categories (COG)** were assigned with a set of custom Perl scripts (**BLASTP** e-value cutoff of 1e-03) (Tatusov *et al.*, 2003).

Signal peptides were detected using **SignalP 4.0** Server (Petersen *et al.*, 2011) with signal P3.0 sensitivity selected. The transfer RNA (tRNA) genes were confirmed with **tRNAScan-SE** (Schattner *et al.*, 2005) and refined with **TFAM** (Ardell and Andersson, 2006). **Rfam** (Burge *et al.*, 2013) was used to predict non coding RNA genes.

Initial metabolic inferences were made using **KEGG** (Kanehisa *et al.*, 2012) and **KAAS** (Moriya *et al.*, 2007). 2007). Metabolic models for each genome were reconstructed using **pathway-tools** (Karp *et al.*, 2002) and the **EcoCyc** (Keseler *et al.*, 2013), **BioCyc** and **MetaCyc** databases (Caspi *et al.*, 2014).

When IS were detected (Gil *et al.*, 2008), they were annotated using the web server **ISsaga** and deposited in **ISfinder** database (Varani *et al.*, 2011). Reference copies for each IS were used to search with **BLASTX** against the non-redundant NCBI database (1e-3 e-value cutoff) and used as **MEGAN4** input for taxonomical assignments with default LCA parameters (Huson *et al.*, 2011). **Ori-Finder** was used to predict the replication origin of plasmids (Gao and Zhang, 2008). **Circos** was used for genome plotting (Krzywinski *et al.*, 2009).

---

[37] Software used can be slightly different between annotation because it is a constantly developing field

Comparative Genomics

3.7.1. Orthologous proteins and synteny

Translated CDS of the desired organism to compare were used as input for **OrthoMCL** as described previously (1.5 inflation value, 70 % match cut-off, 1e-5 e-value cut-off) (Li *et al.*, 2003; Manzano-Marín *et al.*, 2012). **COG** categories were assigned as explained above. Gene clusters may contain zero, one, two, or more CDS in each genome. Some CDS clusters were manually refined because **OrthoMCL** failed to recognize some orthologous CDS in endosymbionts due to the accelerated evolution rate. Clusters of Orthologous CDS were classified as core genome (CDS shared by all the genomes), CDS shared by two or more organisms and strain/organism specific CDS. Euler diagrams were plotted with **gplots** package (Warnes *et al.*, 2013) from **R** software R Core Team 2014 (2014).

Synteny among organisms was plotted using **genoPlotR** package (Guy *et al.*, 2010) from **R** software. Also, **BLAST** results and their positions between two organisms were plotted with **genoPlotR**. **MGR** (Bourque and Pevzner, 2002) was used to calculate the minimum number of rearrangements needed to explain the differences in the genomic architecture between desired genomes.

3.7.2. Genome aligners

Mauve aligner and **Sibelia** were used to compare nucleotide syntenic blocks between different genomes (Darling *et al.*, 2010, 2011; Minkin *et al.*, 2013). For plotting **Mauve** comparison, **genoPlotR** package was used. **Circos** was used for plotting **Sibelia** inferred syntenic blocks (Krzywinski *et al.*, 2009).

NUCmer from **MUMmer 3** was used to plot repetitive regions using the selected genome as query and subject (Kurtz *et al.*, 2004). Results were filtered and only sequences with at least 95 % identity and 500 bp length were used. **NUCmer** output files were used for assessing the level of genome redundancy using a custom python script.

3.7.3. Last Common Ancestor (LCA) Reconstruction

OrthoMCL results were used to reconstruct the putative Last Common Ancestors (LCAs) gene contents. The Most Parsimonious Reconstruction (MPR) function in **ape** package (Paradis *et al.*, 2004) from **R** was used to infer the ancestral state for each character (CDS clusters) in each LCA. Pseudogenes were manually selected and a **TBLASTX** was performed (e-value of 1e-5,80 % overlap) against the proteins present in the orthologous clusters. Pseudogenes that did not modify the LCA reconstruction (strain-specific CDS) were not considered. Pseudogenes that were mobile elements were also excluded. Parsimony reconstruction for orthologous groups that included the previously selected pseudogenes were checked using parsimony reconstruction of discrete characters in **Mesquite** (Maddison and Maddison, 2011).

For each reconstructed LCA and genome, **COG** categories were assigned. For each orthologous cluster, **COG** categories with less than a 10 % of a cluster, as well as the unassigned category, were removed. The LCA indeterminations (the presence/absence of the CDS in the LCA node could not be determined) were counted as half (0.5), instead of presence (1) or absence (0). Relative percentages of each **COG** were computed using one of the precedents LCA as reference and plotted using the **gplots** heatmap2 function without hierarchical clustering. Euler diagram was plotted using **gplots**. **COG** profiles, stated as the absolute number of **COG** categories divided by the total number of **COG** for each genome or LCA, were plotted as a heatmap with **gplots** allowing hierarchical clustering[38].

3.7.4. Metabolic competition

Competition between endosymbionts was checked with **NetCmpt** that reconstruct metabolic environments and checks the potential competition for different metabolites (Kreimer *et al.*, 2012). **NetCmp** calculates an index called Effective Metabolic Overlap (EMO) score for each pair of species that ranges from 0 (a pair of species does not compete) to 1 (a pair of species show strong competition and are mutually exclusive) (Freilich *et al.*, 2010). The European Community number (EC number) for bacterial species was extracted with a custom python script combining available genbank annotation files (NCBI) and in house pathway-tools ocelot database files.

Phylogenetic Methods

3.8.1. Alignments

If no other indication is given, genes or proteins (by itself or concatenated) were aligned with **MAFFT** using the L-INS-i algorithm (Katoh *et al.*, 2002). For 16S rRNAs genes, **ssu-aligner** was employed for the alignment[39] with predefined masking to ensure reproducibility in future alignments (Nawrocki, 2009). All alignments were refined with **Gblocks**, adjusting in each case the percentage of conserved gaps (half or none) (Castresana, 2000). Wherever it was possible, an outgroup sequence was incorporated to the alignment.

---

[38] Dendograms groups the most similar rows or columns together

[39] ssu-aligner takes into account the secondary structure, based on covariance models, of the 16S rRNA genes

For nucleotide alignments, **jModeltest2** was used for selecting the best evolutionary model (Darriba *et al.*, 2012) while **ProtTest3** was used in protein alignments (Darriba *et al.*, 2011).

Codon-based alignments were obtained using a protein alignment together with its nucleotide sequences as input for **PAL2NAL** (Suyama et al. 2006). These alignments were the datasets used for the molecular evolution and divergence analyses.

3.8.2. Phylogenetic tree inference

RaxML was used to calculate the Maximum Likelihood (ML) phylogenetic trees for all the alignments, using optimizations for branch lengths and model parameters, and 1,000 rapid bootstrap replicates (Stamatakis 2006). The evolutionary model was adjusted for each case depending on **jModeltest2** or **ProtTest3** results.

PhyloBayes3.3 was used to perform Bayesian analysis of the ML tree under the specified model (Lartillot *et al.*, 2009). In each case, the evolutionary model was adjusted to the model selected (described above), and three independent chains were run for each alignment. Following Lartillot *et al.* (2009) recommendations, each chain was left until maximum discrepancy between chains was less than 0.1 and all effective sizes were greater than 200 (**bpcomp** and **tracecomp** scripts). Finally, a majority rule posterior consensus tree was calculated for each alignment with **readpb** script. **Archaeopterix** was used for tree visualization and editing (Han and Zmasek, 2009).

3.8.3. Divergence dating

Divergence estimation was firstly computed with **BEAST2** (Bouckaert *et al.*, 2014) using three different dataset. Two datasets were a set of codon-aligned endosymbiotic genes while the third dataset was a codon-aligned mtCOI from different whiteflies. For each gene in the datasets, the evolutionary model was selected according to the **jModeltest2** results and used as priors in **BEAUti** (Bouckaert *et al.*, 2014). **BEAUti** was used to process the alignments, select the partitioning schema, the speciation models and the calibrations points. A lognormal relaxed clock with a Yule speciation process was selected for all datasets based on the results of the model comparison plugin (harmonic mean of the posterior probabilities with 100 bootstrap) implemented in **Tracer v1.6** (`http://tree.bio.ed.ac.uk/software/tracer/`). Two calibration points were inferred from previous works and set set to a uniform distribution: the emergence of the Sternorrhyncha suborder (250-278 My) and the divergence between the subfamilies Aleyrodinae and Aleurodicinae (125-135 My) (Wootton 1981, Shcherbakov 2000, Shi et al 2012, Drohojowska and Szwedo 2011a, 2011b, 2014). Finally, **BEAUti** produced the xml files used by **BEAST2**. Each dataset was firstly run with **BEAST2** under the prior to ensure that divergence dates are only estimated from the data and are not produced by the selected priors. Finally, eight independent runs were performed allowing 500 million generations and sampling every 50000th generation. Convergence, ESS suitability (larger than 200) and burn-in of the runs were checked and calculated with **Tracer v1.6**. Log files of the convergent runs were trimmed, reduced and combined with **Logcombiner** and used for obtaining the descriptive statistics with **Tracer v1.6**. Majority rule posterior consensus method implemented in **TreeAnnotator** was used for obtaining the consensus tree. **FigTree v1.3.1** was used for displaying the tree topologies (`http://tree.bio.ed.ac.uk/software/figtree/`).

To ensure the robustness of the obtained dates, **PhyloBayes3.3** was used for dating the divergences with the same datasets (Lartillot *et al.*, 2009). Because **PhyloBayes3.3** does not accept gene or codon partition, datasets that contained more than one gene alignment were concatenated in a single alignment. Also, fixed tree topologies are required for **Phylobayes3.3**, so the tree topology obtained from **BEAST2** analyses were used as input. Evolutionary models were selected as explained above and a chain under the prior was run for each dataset. Finally, three independent chains were run for each dataset until fitted Lartillot *et al.*

(2009) recommendations (see section 3.8.2). Descriptive statistics were obtained with the **readdiv** script from **PhyloBayes3.3**.

Evolutionary analyses

3.9.1. dN/dS sites analyses

Codon-based alignments of orthologous CDS clusters were used as input for **CodeML** from **PAML** (Yang, 2007). **CodeML** was used to estimate the number of synonymous substitutions per synonymous site (dS), the number of non-synonymous substitutions per non-synonymous site (dN), and their ratio (ω) under a ML approach. Analyses with three branch models were performed: m0 (one ω ratio for all the branches), m1 (free ω ratios for branches) and m2 (2 ω ratios, one for the background branches and one for the foreground branch). The best model for each orthologous cluster was selected using the Likelihood Ratio Test (LTR) values and the **chi2** tool from **PAML**.

Statistical analyses were performed on dN and dS values with **R**. Substitution rates per site and year were calculated based on the results from estimations of the divergence dates (i. e. dN/time of divergence). Exploratory analyses (descriptive statistics, histograms and density plots, boxplots, etc...) were used for cleaning the data of outliers and zero values (probably produced by decimals limits in **codeML**). Levene's test (homocedasticity) and Shaphiro's test (normality) were used as a previous step to select the appropriate statistical test. After logarithmic transformation (base 10) most of the distributions fitted a normal distribution, but some of them presented unequal variances. Two kind of tests were used to check the putative statistical differences between dN, dS or ω distributions among the organisms tested. The Student's T-test for equal and unequal (Welch's procedure) variances was used when data fitted a normal distribution. Kruskal-Wallis test, with its corresponding *post-hoc* tests with p-values corrected by Bonferroni's procedure, was used when the data was not normal distributed but presented equal variances. Finally, genomic dN and dS rates were calculated as a weighted arithmetic mean.

3.9.2. Positive selection test

For positive selection analysis, the pipeline described in Petersen *et al.* (2007) was followed with slightly modifications. Orthologous translated CDS clusters with less than 80 % identity were discarded from the analysis and codons were aligned as explained above. The branch-site model A implemented in **codeML** was used to infer if a selected branch (the foreground with a different ω ratio) has a different ω than the other branches (the background with same ω ratios for all the branches) (Zhang *et al.*, 2005). The model A allows two hypotheses: a null hypothesis (H0), where sites in the foreground and the background branches are under neutral or purifying selection, and an alternative hypothesis (HA) that considers that some sites in the foreground branch are under positive selection while in the background branch they are under neutral or purifying selection. For each orthologous codon-aligned protein LTR p-value was calculated and adjusted using Bonferroni's correction with **R**. If H0 was rejected, Bayes Empirical Bayes (BEB) was inspected for identifying putative sites under positive selection (Yang *et al.*, 2005). For genes that showed sites under positive selection, all alignments were manually inspected. Wherever was possible, the ancestral amino acid state was inferred by maximum parsimony. If the ancestral state could be inferred, genes under selection were only placed in the organism that showed the amino acid change that was different from the ancestral state.

Part 4

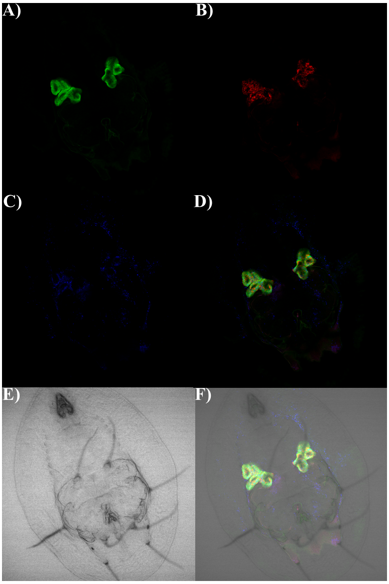
Results and Discussion

“Fly our flag, we teach them fear
Capture them, the end is near
Firing guns they shell burn
Surrender or fight there’s no return
Under Jolly Roger”

Running Wild

Portiera and its partner *Hamiltonella*

4.1.1. Background



Figures 4.1.1 FISH endosymbiont localization in a *B. tabaci* QHC-VLC nymph. **A)** *Portiera* probe (green), **B)** *Hamiltonella* probe (red), **C)** *Cardinium* probe (blue), **D)** merged endosymbiont channels under black field, **E)** bright field channel showing the nymph cuticle, and **F)** merged channels under bright field.

Whiteflies can harbour complex "intracellular ecosystems" usually composed by the P-endosymbiont *Portiera* and, at least, one S-endosymbiont (Costa *et al.*, 1993, 1995; Gottlieb *et al.*, 2006, 2008) that are sharing the same bacteriocyte (see Section Endosymbionts of whiteflies). Also, other kind of S-endosymbionts could be found displaying a scattered phenotype but, also inside the bacteriocytes (Gottlieb *et al.*, 2008). While S-endosymbionts always show a complete cell wall, *Portiera* was firstly described as a pleomorphic bacterium without a clear cell wall (Costa *et al.*, 1993). It has been postulated that the lack of a cell wall may be related to the endosymbiont transmission mechanism in whiteflies (see Section Endosymbiont transmission in whiteflies). Moreover, its closest relatives *Carsonella* (P-endosymbiont of psyllids) and *Evansia* (P-endosymbiont of moss bugs), present a clear cell wall (Santos-Garcia *et al.*, 2014b; Waku and Endo, 1987).

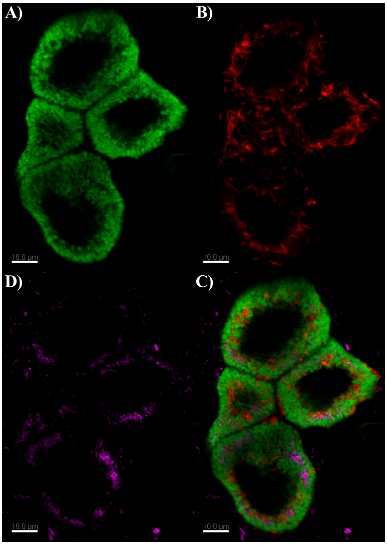
While it has been proposed that *Portiera* is probably involved in host diet complementation (as other Sternorrhyncha P-endosymbionts (Baumann, 2005)), not a clear function has been proposed for the S-endosymbionts that usually accompany it. It is possible that the S-endosymbionts that share the bacteriocytes with *Portiera* during the whole whitefly life cycle, could be involved in some metabolic complementation like *S. symbiotica*, *Ca.* Sulcia muelleri (*Sulcia*) or *Ca.* Baumannia cicadellinicola (Lamelas *et al.*, 2011b; Moya *et al.*, 2008).

The *Hamiltonella* strain found in the *B. tabaci* laboratory strain QHC-VLC presents a bacteriome-confined phenotype (Figure 4.1.1) and could be implied in complementing some metabolic pathways not present or degraded in *Portiera*, although a protective role cannot be discarded (Degnan *et al.*, 2009; Oliver *et al.*, 2010). The *Cardinium* putative role on the whitefly-endosymbiont system will be explored in Section The third passenger: *Cardinium* cBtQ1.

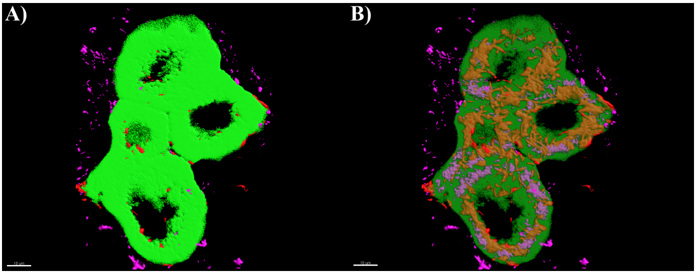
4.1.2. *B. tabaci* QHC-VLC endosymbionts

Whole mount FISH on *B. tabaci* QHC-VLC confirmed that *Hamiltonella* is always found inside the bacteriocyte while *Cardinium* could present a bacteriome-confined and a scattered phenotype (Gottlieb *et al.*, 2008; Skaljac *et al.*, 2010, 2012) (Figure 4.1.1 and 4.1.2). *Portiera* are the biggest cells and presented a non well defined, or pleomorphic, shape (Figure 4.1.2 A, Annex Movie A4.1.1). *Hamiltonella* cells seem to present two forms. The most common is an elongated shape, which sometimes is larger than 9 µm (Figure 4.1.2 B, Annex Movie A4.1.1). Finally, *Cardinium* cells are the smallest rod shape cells, sometimes forming dense aggregates (Figure 4.1.2 C, Annex Movie A4.1.1).

Three different subcellular distributions were detected: *Portiera* is occupying most of the bacteriocyte's cytosol and seems to surround the other endosymbionts, *Hamiltonella* cells seem to have a belt-like distribution and occupy the part of the cytosol that is closer to the bacteriocyte's nucleus, and *Cardinium* seems to occupy mainly the part of the cytosol close to the bacteriocyte's membrane and it could appear forming dense cell aggregates (Figure 4.1.3, Annex Movies A4.1.2 and A4.1.3). This distribution pattern was previously reported by TEM analyses in Costa *et al.* (1995). The conserved localization of *Hamiltonella* during all life-stages and in different species, could point to a conserved function that needs to be maintained in whiteflies (Gottlieb *et al.*, 2008; Skaljac *et al.*, 2010, 2012). In contrast, it seems that *Cardinium* localization is dependent on the host life-stage: while the bacteriocyte is migrating towards the egg, *Cardinium* presents a bacteriocyte-confined phenotype, but once the egg is attached to the leaf and starts its development, *Cardinium* shows a scattered phenotype in addition to the confined one (Gottlieb *et al.*, 2008). This is clear during nymphal stages where only a small amount of *Cardinium* cells remain in the bacteriome while the rest spreads through the whitefly body (Figure 4.1.1 and 4.1.2). This distribution pattern could indicate that *Cardinium* has a different role than the host diet complementation.

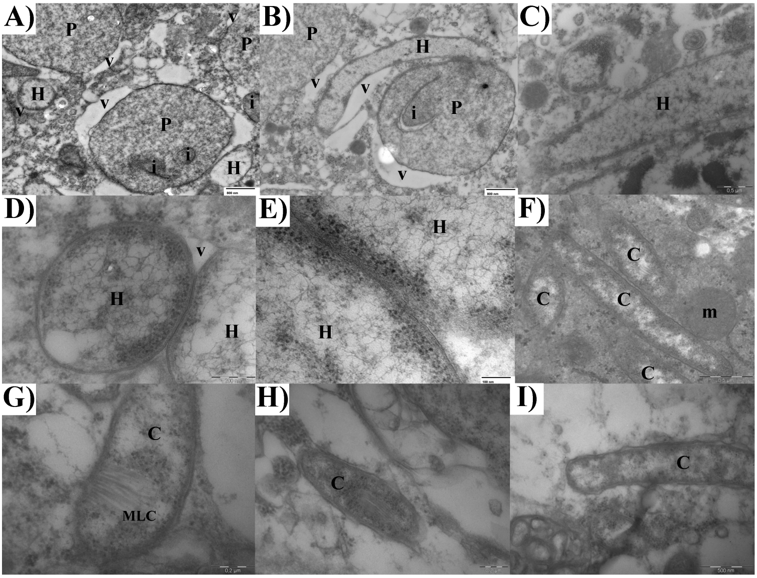


Figures 4.1.2 FISH endosymbiont localization in a *B. tabaci* QHC-VLC nymph bacteriome magnification. Four bacteriocytes can be seen with a blank space in the middle occupied by the cell nuclei. **A)** *Portiera* probe (green), **B)** *Hamiltonella* probe (red), **C)** *Cardinium* probe (blue), **D)** merged endosymbiont channels under black field.



Figures 4.1.3 Renderization from the same Z-stack as Figure 4.1.2. **A)** *Portiera* channel without transparency. **B)** *Portiera* channel with transparency allows to visualize the distribution of *Hamiltonella* and *Cardinium* inside the bacteriome. *Portiera* is displayed in green, *Hamiltonella* in red and *Cardinium* probe in purple. Six slices (3.24 µm) from a Z-stack were used for obtaining the rendered image (0.1 µm. detail level).

Attending to their ultrastructure, *Portiera*, *Hamiltonella* and *Cardinium* are morphologically distinguishable (Figure 4.1.4). *Portiera* are large



Figures 4.1.4 A-G) Bacteriocytes from *Bemisia tabaci* nymphs showing *Portiera* (P) and its membrane infoldings (i), *Hamiltonella* (H), *Cardinium* (C) and its Microtubule-Like Complexes (MLCs), a mitochondrion (m) and host vacuoles (v). **H-I)** *Cardinium* in two unidentified tissues outside the bacteriomes.

pleomorphic cells, harboured inside host vacuoles, with membrane infoldings and electron-dense aggregates that seem to be related spatially with the infoldings (Figure 4.1.4 A and B). *Hamiltonella* are large rod shape cells[40] with a clear cell wall and close to *Portiera* cells (Figure 4.1.4 B-E). As *Portiera*, they can be found inside host derived vacuoles (Figure 4.1.4 A, B, and D). *Cardinium* are rod shaped cells smaller than 3 µm and sometimes present a characteristic structure called MLC (Costa *et al.*, 1995; Zchori-Fein *et al.*, 2004) (Figure 4.1.4 F-I). *Cardinium* cells present a clear cell wall structure, are not usually harboured inside vacuoles and are predominantly distributed at the edge of the bacteriocyte (data not shown) (Costa *et al.*, 1995). In fact, they seem to "move" freely across the cytoplasm and outside the bacteriome through different tissues (Figure 4.1.4 H and I) (Costa *et al.*, 1995). The scattered phenotype of *Cardinium*, its apparent motility and the MLCs are developed in Section Gliding genes in *Cardinium* cBtQ1. For many years *Portiera* was proposed to be an exception for the three-membrane system, the bacterial cell wall plus the host's vacuolar membrane, but no differences were encountered when the putative metabolic capabilities of *Portiera*[41] were compared to those of other P-endosymbionts with three membranes (Table 4.1.1).

Table 4.1.1 Simplified membrane biosynthesis capabilities.

Species	Peptidoglycan	Cardiolipin	Other Fatty Acids / Lipids
Buchnera BCc	-	-	+
Buchnera BAp5A	+	+	+
Buchnera BBp	+	+	-
Carsonella HC	-	+	-
***Portiera* BT-QVLC**	-	+	-
Portiera TV	-	+	-
Evansia Xc1	-	-	-

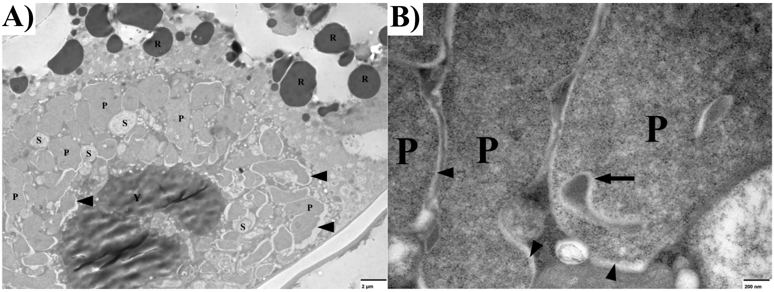
However, two types of *Portiera* membrane structures were found. The most common ultrastructure obtained did not present either a distinctive cell wall or the outer membrane, and was always separated from the host's vacuolar membrane (Baumann, 2005; Coombs *et al.*, 2007; Costa *et al.*, 1993; Szklarzewicz and Moskal, 2001) (Figure 4.1.5). In contrast, a less common structure composed of the host's vacuole membrane and a cell wall-like structure was detected (Figure 4.1.6a and 4.1.6c).

---

[40] Usually from 1 µm to 10 µm, as previously reported (Moran *et al.*, 2005)

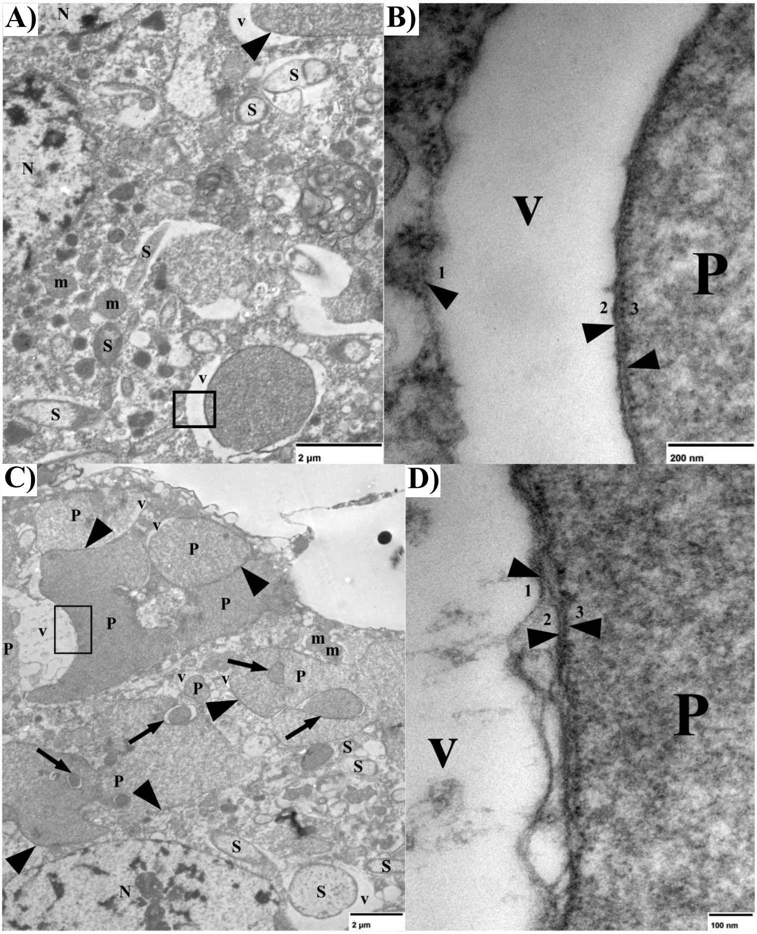
[41] At the time of writing this part of the work, five *Portiera* genomes were publicly available. Although is the most recent part, Sections are ordered for readability

Cells with the less common structure can be found attached to the vacuolar membrane or separated from it, but still with parts of the cell envelope in contact with the vacuolar membrane. In some images the cell envelope presented the typical structure of a gram-negative bacterium (Figure 4.1.6b and 4.1.6d).



Figures 4.1.5 A) Bacteriocytes from a *Bemisia tabaci* egg. Bacteriocyte is surrounded by a cell with reserve substances (R). Vitellogenic reserve (Y) is surrounded by the bacteriocytes. Primary (P) and Secondary endosymbionts (S) can be seen. **B)** Magnification from a nymph bacteriocyte showing different *Portiera* (P) cells without a clear cell wall. Arrowheads denote vacuolar spaces as results of *Portiera* degradation. Arrows denote *Portiera*'s membrane infoldings.

Even though the periplasmic space and both membranes were not always completely separated, the cell envelope showed a variable width depending on whether the periplasmic region was detectable or not (Table 4.1.2). The average widths for *Portiera* outer and inner membranes were 9.52 nanometre (nm) and 7.72 nm, respectively. The host's vacuolar membrane was a little bit wider than *Portiera* membranes (Table 4.1.2). indent When transmission electron pictures from *Carsonella*, *Evansia* and *Portiera* were compared, the former did not show the big vacuolar space usually reported in the latter (Coombs *et al.*, 2007; Costa *et al.*, 1993, 1995; Kuechler *et al.*, 2013; Szklarzewicz and Moskal, 2001; Thao *et al.*,


A)
N
v
S
S
N
S
m
m
S
v
S
2 µm
B)
V
1
P
2
3
200 nm
C)
v
v
P
P
P
P
v
P
P
m
m
v
P
v
P
S
S
P
S
S
N
S
v
2 µm
D)
1
2
3
P
V
100 nm

Figures 4.1.6 Bacteriocytes from *B. tabaci* nymphs. **A)** and **C)** General view of nymphal bacteriocytes. Some secondary endosymbionts (S) , mitochondria (m) and nuclei (N) are observed. Vacuolar spaces (V) can be seen but *Portiera* (P) cells still conserve a clear cell envelope (Arrowheads). Black boxes denote magnified area. **B)** and **D)** Magnified areas showing the three-membranes system of *Portiera*. Arrowheads point to the different membranes: 1) vacuolar membrane, 2) outer membrane, 3) inner membrane.

2000) (Figure 4.1.5). Although this could be an indication of a more degraded stage of the bacteriocyte and a fragile cell envelope in the latter, it is possible that the absence of the three-membrane system in *Portiera* is a technical artefact, produced by the difficulty to obtain a good fixation.

Table 4.1.2 *Portiera* membrane measurements in nm.

	Mean	Geometric Mean	Standard deviation
Outer membrane	9.52	9.26	2.20
Inner membrane	7.72	7.39	2.34
Periplasmatic space	6.88	6.36	2.59
Outer + Inner + Periplasm	30.18	29.52	6.69
Outer + Inner	21.12	20.47	5.40
Vacuole membrane	12.00	11.67	2.81

In fact, in well-conserved samples, with *Portiera* still in contact with the vacuolar membrane, the cell envelope is still observed (Figure 4.1.6). However, the components of the cell envelope could only be observed in some of the specimens with membranes in an initial state of degradation (Figure 4.1.6b and 4.1.6d). Also, it seems that no peptidoglycan (or only very small amounts from an unknown source) is deposited in the periplasmic space, because it is not clearly defined in non-degraded cell envelopes, and makes more difficult to distinguish the cell wall typical structure.

It is true that larger (in genome size terms) P-endosymbionts, like *B. aphidicola* BAp5A, retain the ability to synthesize a minimal cell envelope with all its parts clearly distinguishable, but also the reduced *B. aphidicola* BCc, possesses the three clearly visible membranes (Charles *et al.*, 2011) (Table 4.1.1). It has been postulated that this small *B. aphidicola* might be using the metabolites from the co-obligate endosymbiont *S. symbiotica* to produce its cell envelope (Lamelas *et al.*, 2011b). This suggests, that *Portiera* is using compounds from the secondary symbionts that share the bacteriocytes, or that a complementation or regulation with the host (Husnik *et al.*, 2013; Santos-Garcia *et al.*, 2012, 2014c) cannot be ruled out. In addition, *Carsonella*, that could derive from the same ancestral symbiotic infection event as *Portiera*, has an even more reduced genome and maintains the three-membrane structure (Baumann, 2005) with the same cell wall biogenesis capabilities than the latter (Santos-Garcia *et al.*, 2014b). Also, it needs to be mentioned the ability of *Evansia* to assemble a cell wall in absence of a cardiolipin pathway (Santos-Garcia *et al.*, 2014b). Lastly, neither *Portiera*, *Carsonella* or *Evansia* can synthesize peptidoglycan and it is expected a reduced or absent periplasmic space, as can be seen for *Portiera* in Figure 4.1.6 or for *Carsonella* and *Evansia* (Santos-Garcia *et al.*, 2014b; Waku and Endo, 1987). Because peptidoglycan is responsible for supplying mechanical force (and resistance to different environmental stresses) the cell envelopes of *Portiera*, *Carsonella* and *Evansia* must be extremely fragile and their integrity probably depends on the maintenance of an intact host's vacuole. However, it seems that *Portiera* is the most fragile and it could be related to the endosymbiont transmission route in whiteflies (Coombs *et al.*, 2007; Szklarzewicz and Moskal, 2001). It is known that small changes in membrane phospholipids composition can modify its homoeostatic and stability features (Cronan, 2003; Parsons and Rock, 2013). Without experimental procedures to determine endosymbionts' membrane composition, it is plausible that small phospholipids changes in the endosymbiont membrane are responsible for the fragility differences reported. However, it remains unclear the source of these phospholipids, being the host one of them.

Although it is still unclear how extremely reduced P-endosymbionts lacking most of the cell envelope biosynthetic genes produce their membranes, there are suggestions that it could be through a host's control mechanism or the use of host-derived membranes metabolites (Husnik *et al.*, 2013). Additionally, as stated before for *B. aphidicola* BCc, a second endosymbiont could provide the lacking cell envelope biogenesis functions in *Portiera*.

At present, the only other reported case of a two-membrane system was that of *B. aphidicola* BBp (Charles *et al.*, 2011). However, taking into account that according to its putative metabolic capabilities it is able to synthesize the two gram-negative membranes (Table 4.1.1), and considering the above mentioned, it cannot be discarded that this is also an artefact. Thus, similarly to *Portiera*, its membrane ultrastructure should be revisited to confirm its membrane organization.

4.1.3. *Portiera* BT-QVLC

Portiera from the *B. tabaci* QHC-VLC laboratory strain was named as BT-QVLC strain according to: its host (BT refers to *B. tabaci*), the biotype of the host (Q biotype or MED) and the region where the laboratory strain was obtained (VLC refers to Valencia).

4.1.3.1. *Portiera* BT-QVLC genomic features

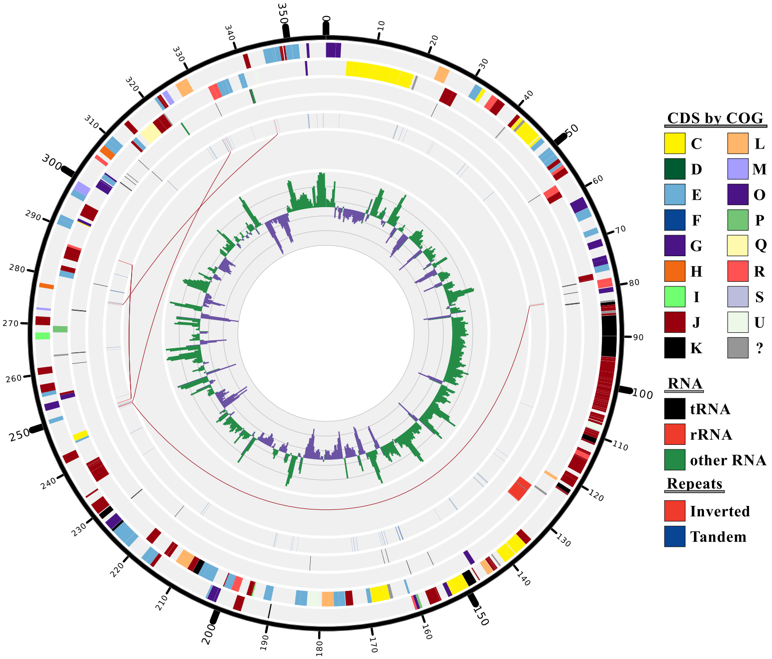
Portiera BT-QVLC is an extreme reduced P-endosymbiont with a circular chromosome of 357.472 bp. The hybrid *de-novo* assembly, single-end and 3-kb pair-end libraries 454 GS-FLX Titanium and a 5-kb mate-pair HiSeq2000 libraries, gave a combined coverage of 41X. *Portiera* BT-QVLC presented 246 CDS, eight pseudogenes (*argH*, *miaA*, *ruvC*, *dapB*, *clpX*, *clpP*, *galP*, and the ABC transporter *PAQ_201*), and 38 non-coding RNA genes, including the three ribosomal RNAs (rRNAs) forming a cluster (16, 23 and 5S), 33 tRNAs able to decode all messenger RNAs (mRNAs), one transfer-messenger RNA (tmRNA) and the *rnpB* (the RNA subunit of RNaseP) (Table 4.1.3 and Figure 4.1.7).

Although the genomic features of *Portiera* BT-QVLC are in general similar to other reduced P-endosymbionts, like *B. aphidicola* BCc or *Evansia* Xc1 (Pérez-Brocal *et al.*, 2006; Santos-Garcia *et al.*, 2014b), its number of CDS is unusually low according to its genome size, even when it is compared to very extreme reduced P-endosymbionts like *Carsonella* (Table 4.1.3 and Figure 4.1.7). The low coding density (68 %)

Table 4.1.3 Genomic features of *Portiera* BT-QVLC compared to other P-endosymbionts.

Symbiont	Genome size (bp)	GC (%)	Genes	CDS	Coding density (%)	rRNA	tRNA	Other RNA	Pseudo
Carsonella PV	159,662	17	213	182	97	3	28	0	0
Carsonella HC	166,163	14	223	192	98	3	28	0	0
***Portiera* BT-QVLC**	357,472	26	284	246	68	3	33	2	8
Evansia Xc1	357,498	25	369	330	94	3	33	3	0
Buchnera Cc*	422,434	20	403	365	87	3	31	4	3
Buchnera 5A	642,122	27	592	555	87	3	32	2	7

*Plasmid pLeu-BCc is included in the summary statistics

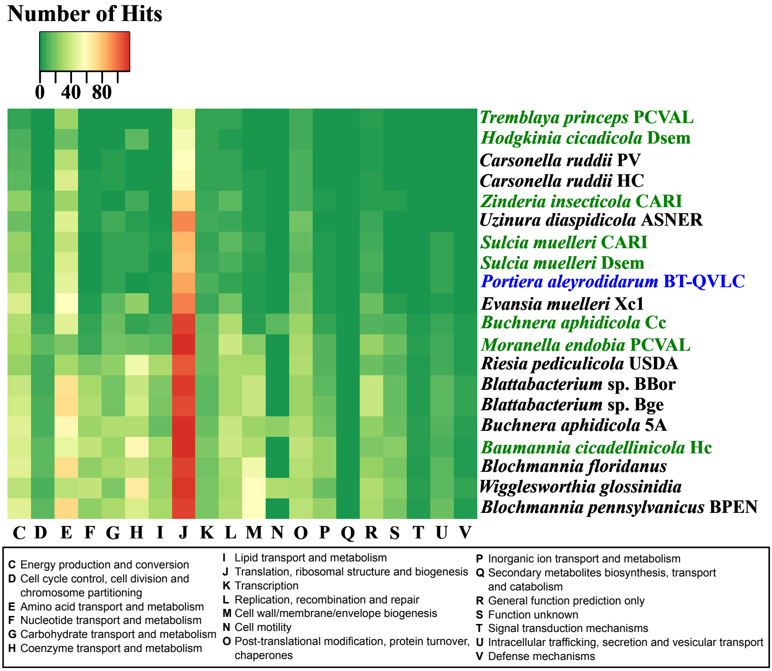

0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
260
270
280
290
300
310
320
330
340
350
CDS by COG
C
L
D
M
E
O
F
P
G
Q
H
R
I
S
J
U
K
?
RNA
tRNA
rRNA
other RNA
Repeats
Inverted
Tandem

Figures 4.1.7 Genome overview of *Portiera* strain BT-QVLC. From inner to outer tracks: (I) Positive (green) and negative (purple) GC skew across the genome. (II) Inverted repeats (red lines and links) and Tandem repeats (blue). (III) Complementary strand noncoding RNAgenes: rRNAgenes (red), transferRNA genes (black), other RNA genes (green). (IV) Direct strand noncoding RNA genes: rRNAgenes (red), transferRNAgenes (black), other RNAgenes (green). (V) Complementary strand CDS. (VI) Direct strand CDS. CDS were coloured according to their **COG** classification.

is due to the large Intergenic Regions (IGRs) present in *Portiera*. Another unusual feature not shown by other extreme reduced P-endosymbionts is the presence of an important number of repetitive regions, 112 tandem repeats and four inverted repeats, dispersed across the genome. *Portiera* BT-QVLC lacks an evident GC skew as can be found in other P-endosymbionts, a possible evidence of genomic rearrangements (low coding density and large IGRs are discussed in detail in Section Genome evolution of the genus *Portiera*).

4.1.3.2. Comparative genomics

Portiera BT-QVLC general metabolic capabilities were compared to other P-endosymbionts based on **COG** classification (Figure 4.1.8). According to its **COG** profile, *Portiera* BT-QVLC is more similar to *Sulcia* strains CARI and Dsem, P-endosymbionts with genomes around 270 kilobase pairs (kb) that present extremely reduced co-primary endosymbionts (*Zinderia insecticola* CARI and *Hodgkinia cicadicola* Dsem, respectively). When *Portiera* BT-QVLC is compared to its relative *Evansia* Xc1, that has the same genome size but its host does not present other endosymbionts in its bacteriomes, the profiles showed a general small number of **COG** hits in *Portiera* including the **C** (energy production), **E** (amino acid biosynthesis) and **H** (coenzyme metabolism) (Figure 4.1.8). This means that with an equal genome size and based on their functional categories, *Evansia* Xc1 has a greater metabolic repertoire than *Portiera* (Santos-Garcia *et al.*, 2014b). It is also surprising that **L** (DNA replication and repair) category is more reduced than in *Sulcia* and closer to its relative *Carsonella* (ca. 160 kb genome size) and to *Ca.* Tremblaya princeps (ca. 138 kb genome size), the P-endosymbiont of mealybugs (Figure 4.1.8). In addition, it seems that **O** (post-translational modification and protein turnover) category is diminished when it is compared to *Evansia* Xc1 but it is closer to *Sulcia*. Despite its genome size, it seems than *Portiera* BT-QVLC is closer to smaller P-endosymbionts that are usually found in community

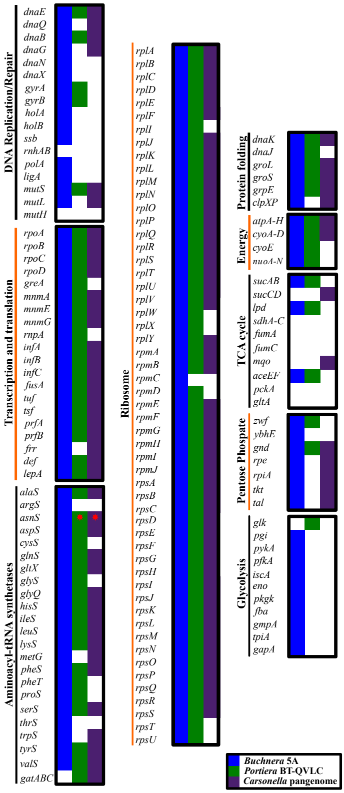


Figures 4.1.8 Heatmap of selected **COG** categories from different endosymbionts. Endosymbionts are sorted by genome size from the smallest (*Tremblaya princeps PCVAL*) to the biggest (*Blochmannia pennsylvanicus* BPEN). For each genome, the numbers of hits in each **COG** category are shown. The names of P-endosymbionts living with another co-primary endosymbiont are displayed in green. *Portiera* BT-QVLC, which shares the bacteriocytes with *Hamiltonella* and *Cardinium*, is displayed in blue. **COG** descriptions are showed in the bottom.

with another co-primary endosymbionts. This could be an effect of the low coding density of *Portiera* and seems to point that a mutualistic relationship with its partner *Hamiltonella* has started.

Regarding the reduction in **C**, **G**, **J**, **K**, **L**, and **O** **COG** categories, the basic cell machinery and the central metabolism of *Portiera* BT-QVLC was compared against *B. aphidicola* 5A (as a representative of a P-endosymbiont without a co-primary endosymbiont partner) and *Carsonella*, the closest relative of *Portiera* (a extreme reduced P-endosymbiont without a co-primary endosymbiont partner) (Figure

4.1.9). *B. aphidicola* 5A presented a complete set of DNA replication and repair machinery (**L**) while *Portiera* and *Carsonella* encoded a reduced one (Gil *et al.*, 2004; Tamames *et al.*, 2007) (Figure 4.1.9). In addition, *Portiera* presented even a more reduced set than *Carsonella* or *Ca.* Nasuia deltoce-phalinicola[42] (Bennett and Moran, 2013; Moran and Bennett, 2014). Only the *dnaE* (which encodes the polymerase activity) and *dnaB* (required for opening the replica-tion fork) polymerase subunits were present in *Portiera* but no signal of the *dnaQ* (proofreading activity), *dnaG* (primase activity), *dnaN* (the polymerase clamp), and *dnaX* (the dimerization unit) were detected. Only another extreme reduced polymerase has been reported, the case of *Ca.* Uzinura diaspidicola (hereafter *Uzinura*) from armoured scale insects (Sabree *et al.*, 2013) but, it is unknown how this organism deal with the apparent lack of proofreading activity and the increase in the polymerase instability (due to the absence of the clamp subunit).



Figures 4.1.9 Basic cell machinery and central metabolism comparison between *B. aphidicola* 5A, *Portiera* BT-QVLC and *Carsonella*'s pangenome from strains DC, HC, and PV. Red star denotes the alternative L-asparaginyl-tRNA pathway by the combination of the non-discriminating *aspS* and *gatABC*. Each gene was plotted only once.

The transcription, translation (**J**)

---

[42] It presents the smallest genome sequenced until the date and it is a co-primary endosymbiont of the leafhopper *Macrosteles quadrilineatus*

and ribosome biogenesis (**K**) categories has suffered some losses in *Portiera* and *Carsonella* (Figure 4.1.9). The loss of *frr* in some P-endosymbionts (including *Portiera*, *Uzinura*, and *Sulcia*), necessary for releasing the mRNA from the ribosome, suggest the possibility that they need to import it from the host cytosol. However, due to the small size of the different translation factors (and other small proteins) and the accelerated evolution reported in P-endosymbionts, it is possible that they cannot be identified by homology searches. *Portiera* BT-QVLC presents an almost complete ribosome, with the exception of *rpmC* that is lost also in other P-endosymbionts (Moran and Bennett, 2014). Moreover, *Portiera* BT-QVLC has lost four aminoacyl tRNA synthetases including *metG*, *trpS*, *argS*, and *thrS* (the latter two also lost in *Carsonella*) (Figure 4.1.9). In addition, it seems that *Portiera* BT-QVLC, *Carsonella* and *Evansia* Xc1 may produce Asn-tRNA (in many species produced by the aminoacyl tRNA synthetase *asnS*) through the action of a non-discriminating aspartyl-tRNA synthetase (encoded by *aspS*) followed by the action of glutamyl-tRNA(Gln) amidotransferase (encoded by *gatABC*). In support of the non-discriminating action is the presence of a histidine in position 30, a typical feature of non-discriminating enzymes, while discriminating enzymes possess a leucine (Bernard *et al.*, 2006). The differential loss of aminoacyl tRNA synthetases in P-endosymbiont has been explained by two different processes: the acquisition of new functions by the remaining aminoacyl tRNA synthetases (Moran and Bennett, 2014) or the import of nuclear encoded proteins, as suggested specifically for *Evansia* Xc1 ArgS (Santos-Garcia *et al.*, 2014b) and, in general, as one of the possible mechanisms able to compensate the loss of many important genes in *Tremblaya* (Husnik *et al.*, 2013). Some of these mechanisms involved HGT events into the nuclear insect genome. Although the authors detected examples of HGT events of bacterial origins into the nuclear insect genome, none of them compensated the lost aminoacyl tRNA synthetases (Husnik *et al.*, 2013). Recently, it has been confirmed that a nuclear encoded protein of bacterial HGT origin is specifically transported to the *B. aphidicola* cell (Nakabachi *et al.*, 2014).

From post-translational modification and protein turnover (**O** category) the most relevant is the absence of the ClpXP complex, in charge of recycling the misfolded proteins, in *Portiera* but present in *Carsonella* (Figure 4.1.9). In P-endosymbionts the GroEL-GroES chaperonin complex is in charge to help proteins to fold in a correct way. Also, in case of misfolded proteins this complex unfold the protein and aids to accomplish the correct tertiary structure of the protein, probably in a similar way to mitochondria (Tatsuta, 2009). When GroEL-GroES fails to re-fold the protein, the unfolded protein is degraded by the ClpXP serine-protease complex. ClpXP is able to unfold very stable misfolded proteins and degrade them into peptides as a recycling step (reviewed in Baker and Sauer (2012)). Also it recycles the proteins that stuck and fails to be released from the ribosome. Usually, extreme reduced P-endosymbionts only present this protease complex to recycle the misfolded proteins and avoid their accumulation, that usually has negative effects for the cell (see Annex Table A4.1.1). It is intriguing why *Portiera* from *B. tabaci* has lost the ClpXP although the closely related HlsUV complex could replaced it (Tatsuta, 2009).

Lastly, regarding the energy production and the central metabolism (**C** and **G** categories, respectively), *Portiera* encodes most of the electron transport chain (ATPsynthase, NADH dehydrogenase and the cytochrome bo oxidase) as bigger P-endosymbionts like *B. aphidicola* 5A (Figure 4.1.9). The presence of the electron transport chain components is variable among P-endosymbionts and indicates its dependence of an ATP source supplied by the host. When compared to *Portiera*, *Carsonella* has lost the NADH dehydrogenase. *Portiera*, as *B. aphidicola* 5A, maintains part of the tricarboxylic acid cycle (TCA) needed for maintaining the electron chain (it uses pyruvate to produce NADH). In contrast, some *Carsonella* strains have maintained a different set of TCA. These genes, as the *Portiera* ones, are in charge of supplying reductive power. Finally, while *Portiera* conserves the first step of the glycolysis, *Carsonella* has lost the whole pathway. In contrast, *Carsonella* presented a more complete pentose phosphate pathway while only two genes are maintained in *Portiera*. This indicates that while *Portiera* needs to import from the host all the intermediate metabolites produced by the pentose phosphate pathway, *Carsonella* is able to produce them (Figure 4.1.9).

The loss of essential genes related to informational processes in *Portiera* raises the question if it can be considered as a P-endosymbiont or it has crossed a biological borderline to be considered a subcellular entity from its host. Many years of discussion about where is the limit to consider an endosymbiont as a independent entity have produced several proposals. The work done in *Carsonella* by Tamames *et al.* (2007) suggested that the loss of essential genes related to the replication/repair and translation machinery in combination with some amino acid biosynthetic pathways points that *Carsonella* cannot be considered longer as a P-endosymbiont.

A recent work has been demonstrated that some of the lost amino acid pathways has been transferred to the host genome (Sloan *et al.*, 2014). Also, a similar case has been reporter for the tandem *Tremblaya-Moranella* where some metabolic functions has been assumed by the host after different HGT events from different bacteria (Husnik *et al.*, 2013). While it seems that transferring some metabolic function to the host is not so uncommon as previously thought, it remains unclear what happens in the case of informational genes (like the aminoacyl tRNA synthetases or some genes from the DNA polymerase).

This threshold is evident in the Halomonadaceae endosymbiont lineage (*Portiera*, *Evansia* Xc1 and *Carsonella*) where despite of their biosynthetic capabilities or genome size, they have lost their informational autonomy (Santos-Garcia *et al.*, 2014b). If these essential proteins are acquired from the host from a HGT nuclear-encoded gene targeted to the endosymbiont or if they are using the same proteins as the mitochondria (sharing their signalling pathway) it is a mystery.

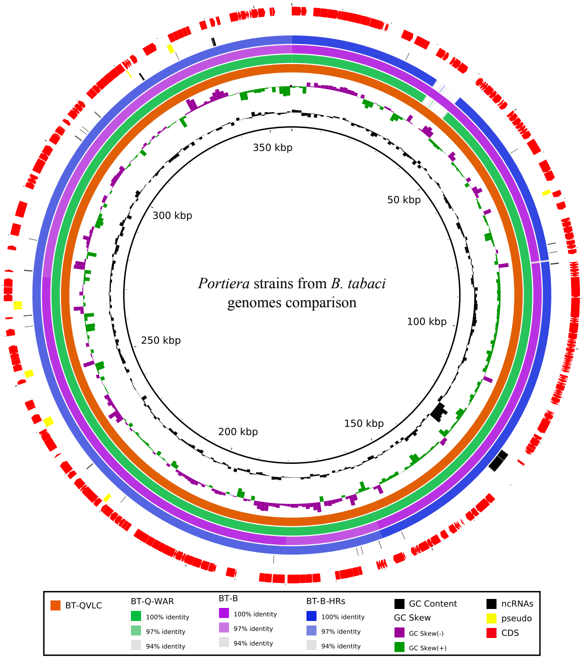
Although the “symbionelle” term seems to reinforce the idea that the evolutionary history of organelles and endosymbionts has been occurred in different context (at unicellular and multicellular organism, respectively), it should be revisited taking into account the above mentioned threshold rather than total gene content or biosynthetic capabilities (Reyes-Prieto *et al.*, 2014). Finally, with the experimental evidence of nuclear-encoded proteins targeted specifically to the endosymbiont reported by Nakabachi *et al.* (2014), the difference between symbionts and organelles becomes more blurred (McCutcheon and Keeling, 2014). In this context, the "symbionelle" term could help to categorize endosymbionts that are no longer autonomous at informational level but has not raised the organelle status.

4.1.3.2.1. *Portiera* strains from *B. tabaci*

Almost simultaneously, four *Portiera* strains from *B. tabaci*, including BT-QVLC, were released to the public domain: two strains from *B. tabaci* Q biotype (MED speciea), BT-QVLC (Santos-Garcia *et al.*, 2012) and BT-Q-WAR (Jiang *et al.*, 2013), and two more from the B biotype (MEAM1), BT-B (Sloan and Moran, 2012a) and BT-B-HRs (Jiang *et al.*, 2013). The average nucleotide identities by pairwise comparison were: 99.6 % BT-QVLC *vs* BT-B, 99.6 % BT-QVLC *vs* BT-B-HRs, and 99.9 % BT-QVLC *vs* BT-Q-WAR (Figure 4.1.10).

Because a mix of 454 and Illumina technology was used for BT-QVLC sequencing, this strain accumulated more homopolymers errors than the other strains and in consequence the real nucleotide identity between this strain and the others is even higher (closer to 100 %). Nucleotide differences between strains are mainly located at the IGRs and the tandem repeats.

Porteira BT-B-HRs and BT-Q-WAR strains (Jiang *et al.*, 2013) were released as incomplete genomes due to the presence of a gap as a result of polymorphic structural variants (Figure 4.1.10). This structural polymorphism was detected in *Portiera* BT-B (Sloan and Moran, 2013) although it is present in all the other strains. This polymorphism contains


350 kbp
50 kbp
300 kbp
100 kbp
250 kbp
150 kbp
200 kbp
Portiera strains from *B. tabaci*
genomes comparison
BT-QVLC
BT-Q-WAR
100% identity
97% identity
94% identity
BT-B
100% identity
97% identity
94% identity
BT-B-HRs
100% identity
97% identity
94% identity
GC Content
GC Skew
GC Skew(-)
GC Skew(+)
ncRNAs
pseudo
CDS

Figures 4.1.10 Comparison between the four *Portiera* strains sequenced from *B. tabaci*. Coloured circles represent the genomic comparison at nucleotide level (**BLASTN**) of the different strains against *Portiera* BT-QVLC. Blank region in the BT-Q-WAR strain (turquoise) and BT-B-HRs (blues) represents the 6.1 kb structural polymorphism present in all *Portiera* strains (Sloan and Moran, 2013).

three genes (*yidC*, *mnmE*, and *mnmG*) with a lenght of 6.1 kb and flanked by two identical tandem repeats. This region could be in two different structural conformations: integrated in the chromosome or as a separate subgenomic circle. Also, it could be present a variable amount of copies (from zero to three) in a tandem organization (Sloan and Moran, 2013).

Nevertheless, all these *Portiera* strains are identically, with some discrepancies due to annotation procedure, and all the conclusions made for *Portiera* BT-QVLC in this work can be extrapolated to the other three *Portiera* strains from *B. tabaci*.

4.1.4. *Hamiltonella* BT-QVLC

Bacterial enriched samples presented low amounts of *Hamitonella* cells compared to *Cardinium* or *Portiera* cells. After sequencing, reads belonging to *Hamiltonella* were less than 1 % of the library. The low amount of reads recovered (6X of 454 and 25X of Illumina coverage), the presence of repetitive elements (mobile elements and phage sequences) plus the chimeras formed during the WGA increased the complexity of the assemblage process. Finally, a draft assembly of *Hamiltonella* BT-QVLC (named following the same criteria as *Portiera*) was generated (Table 4.1.4). At the moment of writing this work, two more *Hamiltonella* genomes were available, the complete genome of *Hamiltonella* 5AT from *A. pisum* (Degnan *et al.*, 2009) and the draft genome of *Hamiltonella* MED from *B. tabaci* (Rao *et al.*, 2012).

Table 4.1.4 *Hamiltonella* strains assemblies statistics.

Strain	Genome size (Mb)	Scaffolds	N50 Scaffold (kb)	Contigs	N50 Contigs (kb)
5AT*	2.17	2	-	4	-
MED	1.84	404	14	372	12
BT-QVLC	1.61	85	26	101	43

*Plasmid pHD5AT (59 kb) is included in the summary statistics

Table 4.1.5 *Hamiltonella* strains general genomic features.

Strain	GC (%)	Genes	CDS	Coding density (%)	rRNA	tRNA	Other RNA	Pseudo
5AT*	40	2,200	2,148	81	9	43	-	1
MED	40	1,970	1,916	84	1†	38	15	-
BT-QVLC	40	1,897	1,839	81	?‡	33	24	-

*Plasmid pHD5AT (59 kb) is included in the summary statistics

† One 23S rRNA copy ‡ No rRNA genes were found

A total number of 101 contigs ordered in 85 scaffolds were obtained (Table 4.1.4). Although *Hamiltonella* 5AT presented a plasmid, it is not possible to know if it is conserved with this topology in the *Hamiltonella* BT-QVLC and MED assemblies. Approximately, the scaffolds spanned 1.61 megabase pair (Mb) that is 230 kb less than a published draft genome of *Hamiltonella* MED strain (from another *B. tabaci* Q biotype (MED)) (Rao *et al.*, 2012) (Table 4.1.4). Both genomes were more reduced than the *Hamiltonella* strain 5AT from *A. pisum* (Degnan *et al.*, 2009) with a difference of 560 kb when is compared to *Hamiltonella* BT-QVLC. When genomic features of *Hamiltonella* strains are compared, it seems that CDS number in BT-QVLC and MED did not correspond with the genome size and the coding density (Table 4.1.5). These increased number in CDS seems to be due to the presence of fragmented genes that are recognized as different CDS by annotation pipelines. Even, it is possible that most of these fragmented genes could be real pseudogenes, missassemblies problems can not be discarded. In *Hamitonella* MED, only one rRNA gene was detect probably due to the difficulty to assembly these regions. rRNA genes are under concerted evolution, which produces that these genes are almost identical, suggesting that at least two copies of the rRNA cluster should be present in *Hamitonella* MED and BT-QVLC strains. An almost complete set of tRNAs was detected in *Hamitonella* BT-QVLC with the exception of those charging histidine and isoleucine. Finally, a bigger number of other RNA genes were detected in *Hamitonella* BT-QVLC but this differences could be an effect of annotation pipelines.

Hamiltonella BT-QVLC and MED strains diverged recently because they present an average genome nucleotide identity value of 99.6 %[43] and differences in gene content and gene status (pseudogenes) are more likely an effect of genome assembly (loss of different contigs, homopolymers, repeats collapse, etc.) than real gene differences due to accommodation to the environment, that in fact is virtually the same (equal *B. tabaci* biotype and all *Portiera* strains have the same gene content).

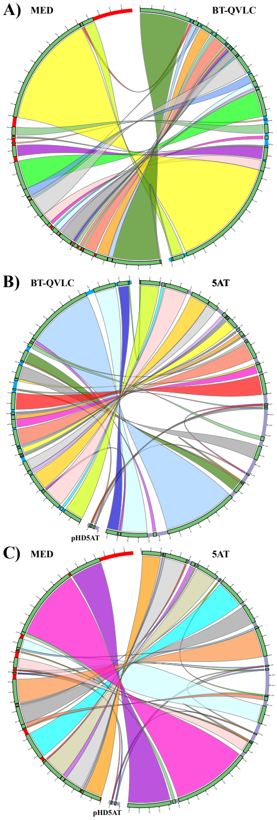
Differences due to genome assembly are also observed when nucleotide syntenic blocks larger than 5 kb are displayed (Figure 4.1.11).

---

[43] Genome nucleotide identity between *Hamitonella* BT-QVLC and 5AT is 96.6 %

Although scaffolds were joined and ordered according to *Hamitonella* 5AT genome for plotting reasons[44], syntenic blocks between *Hamitonella* BT-QVLC and MED strains showed that both genomes share most of their contents (Figure 4.1.11 **A**). As explained above, these difference could be due to the genome assembly process. In spite of the draft status of both genomes, it is clear that microsynteny (regional gene order), and probably macrosynteny (genomic architecture), is conserved between these two strains (Figure 4.1.11 **A**).

When *Hamitonella* BT-QVLC and MED syntenic blocks are compared against 5AT (Figure 4.1.11 **B** and **C**, respectively), it seems that both strains are a subset of 5AT despite some gains in the formers strains. However, it seems that *Hamiltonella* MED presented more syntenic blocks shared with 5AT than BT-QVLC, suggesting that this genome is more complete (Figure 4.1.11 **B** and **C**). Despite of the completeness genome of *Hamitonella* MED strain, a region of *ca.* 180 kb in this genome had not a counterpart in *Hamitonella* BT-QVLC or 5AT, indicative of a possible chimeric region not belonging to *Hamiltonella* (Figure 4.1.11 **A** and **C**). In fact, during this



Figures 4.1.11 Nucleotide syntenic blocks between the three *Hamiltonella* strains. Ideograms represent the compared genomes. Green boxes represent syntenic blocks connected by different colours. Strain specific regions are displayed in blue (BT-QVLC), red (MED) and purple (5AT).

---

[44] The number of scaffolds in *Hamitonella* MED and BT-QVLC becomes unintelligible if it is not reduced previously

work six contigs, *ca.* 50 kb, belonging to *Portiera* were found. Regarding the plasmid present in *Hamitonella* 5AT strain, only some regions of the pHD5AT plasmid have their counterparts in *Hamitonella* BT-QVLC or MED genomes, suggesting the absence of this plasmid in the latter strains. The APSE phage, present in *Hamitonella* 5AT, encodes different toxins and it has been related to the resistance against parasitoids in aphid (Degnan and Moran, 2008; Oliver *et al.*, 2003). Although it is not shown in Figure 4.1.11, APSE phages with some related toxins were detected in *Hamitonella* BT-QVLC and MED strains. This suggests that a protective role of *Hamiltonella* BT-QVLC and MED cannot be discarded.

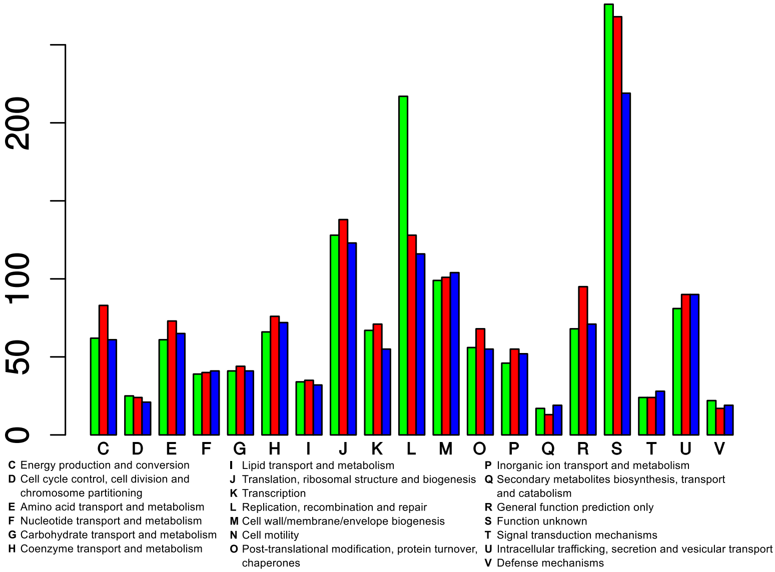
A possible explanation for region gains and losses in *Hamitonella* BT-QVLC and MED strains compared to 5AT could be due to HGT events after the divergence of both *Hamiltonella* lineages combined with gene losses in BT-QVLC and MED after the arrival to a new host.

Genome reduction usually produces a differentially shrinkage in the **COG** categories[45]. This process is a consequence of the accommodation to an intracellular life style and the new stable environment (Manzano-Marín *et al.*, 2012). In this context, **COG** categories distribution for each *Hamiltonella* strain were compared (Figure 4.1.12).

Hamiltonella 5AT, MED and BT-QVLC had 1430, 1444, and 1285 **COG** hits respectively. However, it seems that the higher number of **COG** hits in MED could be linked to the presence of fragmented genes or the big chimeric region (Figure 4.1.11). The greater differences were found in energy production (**C**), replication/recombination/repair (**L**), general (**R**) and unknown (**S**) function (Figure 4.1.12). While **C**, **R** and **S** seem artefactual results, the equal **L** reduction in *Hamitonella* BT-QVLC and MED seems to support this result. **L** reduction in both strains points to the idea that *Hamiltonella* BT-QVLC and MED are progressively loosing their autonomy, like other co-primary endosymbionts such as *S. symbiotica* (Lamelas *et al.*, 2011b; Manzano-Marín and Latorre, 2014; Manzano-Marín *et al.*, 2012). Despite of the great differences in **L**

---

[45] Specially in **D**, **J**, **K**, **O**, and **L** categories



Figures 4.1.12 COG categories distribution for *Hamiltonella* 5AT (green), MED (red) and BT-QVLC (blue).

category, it seems that all strains have a similar central metabolism (**G**), membrane (**M**) and nucleotides (**F**) biosynthesis capabilities, lipid (**I**) metabolism and response to the environment (**T**). Moreover, it is interesting to notice that categories regarding the biosynthesis of amino acids (**E**) and vitamins/cofactors (**H**) showed a slightly increase in *Hamitonella* BT-QVLC and MED compared to 5AP. This could point to increased amino acids and vitamins biosynthetic capabilities in the *Hamiltonellla* from *B. tabaci* that could be related to a possible role in metabolic complementation of *Portiera* (see next Section Metabolic integration).

4.1.5. Metabolic integration

Due to the draft state of *Hamiltonella* BT-QVLC and assuming that the MED strain is more complete, the metabolic models of the two strains were reconstructed with pathway-tools. All the reactions/pathways present in *Hamiltonella* BT-QVLC were compared to MED strains. If a hole (absence of an enzyme) was detected in a pathway from *Hamiltonella* BT-QVLC but it was present in MED, the MED gene was used for mapping the Illumina library and check for the presence and the state of the gene in BT-QVLC (e.g. *nadB*, *bioH*, *serA* were recovered in this way). Finally, each fragmented gene present in the pathways analysed in *Hamiltonella* BT-QVLC was reassembled alone and checked for sequencing errors that could produce an artefactual pseudogene (e.g homopolymeric stretches). *Portiera* BT-QVLC metabolism was also reconstructed using pathway-tools[46]. Insect metabolic capabilities were inferred using *A. pisum* AcypiCyc database (Vellozo *et al.*, 2011), KEGG database (Kanehisa *et al.*, 2012), the work made by Xie *et al.* (2012), and searching the corresponding enzymes by **TBLASTN** against all publicly available *B. tabaci* transcriptomes at NCBI.

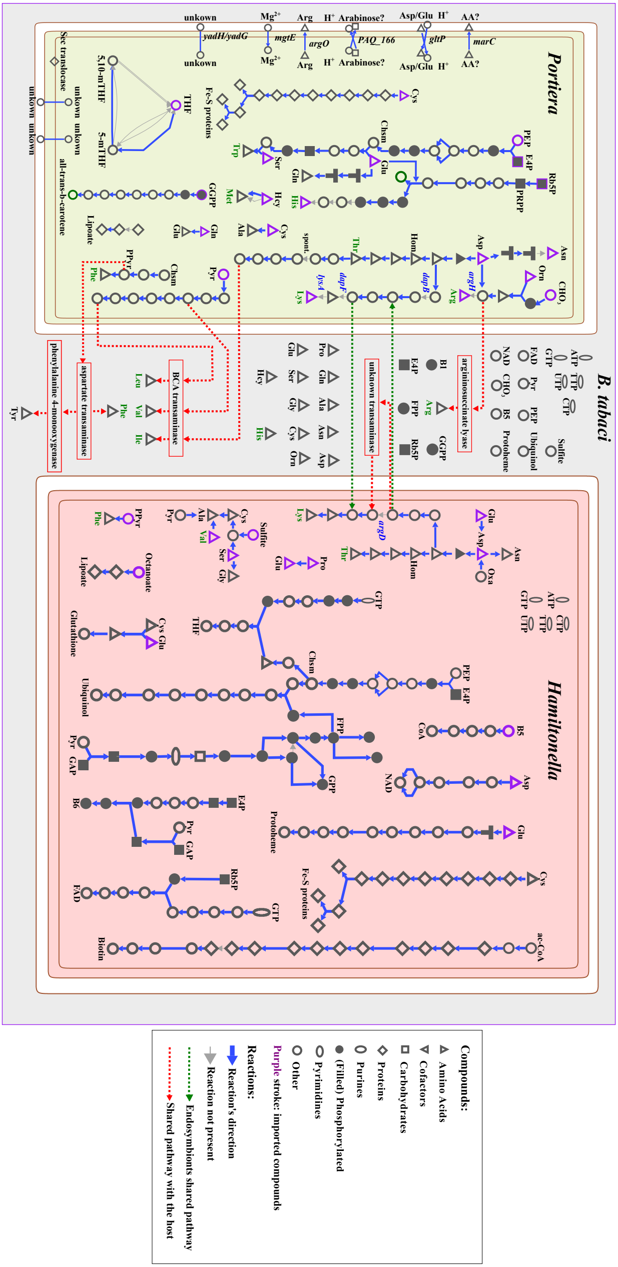
Portiera seems to be an “essential amino acids production factory” that conserves only the parts of the central and energy producing metabolism required for amino acid biosynthesis (production of reducing power, the electron transport chain for regenerate ATP and some intermediate metabolites). In contrast, *B. tabaci* is in charge to supply non-essential amino acids, intermediate and secondary metabolites, and some vitamins/cofactors (Xie *et al.*, 2012). *Hamiltonella* is mainly in charge of supplying the vitamins/cofactors not produced by the host but is able to produce most of the intermediate metabolites it needs. These relationships build a metabolic network able to produce all amino acids and most of vitamins/cofactors required by the three organisms (Figure 4.1.13).

4.1.5.1. *Portiera* biosynthetic capabilities

Portiera presents the complete biosynthetic pathways to produce Tryptophan (Trp) and Threonine (Thr). For Trp it needs to import from the host Phosphoenolpyruvate (PEP) (glycolysis), D-erythrose 4-phosphate

---

[46] Publicly available from BioCyc registry `http://biocyc.org/registry.html`


Portiera
B. tabaci
Hamiltonella
yadH/yadG
mgtE
argO
PAQ_166
gltP
marC
argH
dapB
dapF
lysA
argD
argininosuccinate lyase
unknown transaminase
BCA transaminase
aspartate transaminase
phenylalanine 4-monooxygenase
Sec translocase
Compounds:
Amino Acids
Cofactors
Carbohydrates
Proteins
Purines
(Filled) Phosphorylated
Pyrimidines
Other
Purple stroke: imported compounds
Reactions:
Reaction's direction
Reaction not present
Endosymbionts shared pathway
Shared pathway with the host

Figures 4.1.13 Metabolism integration between *Portiera*, *Hamiltonella* and *B. tabaci* at amino acid and vitamins/cofactors level. Only *Portiera* transporters are displayed. Only *B. tabaci* products imported by the endosymbionts are displayed. Abbreviations not in the main text: Geranyl diphosphate (GPP), (2e,6E)-farnesyl diphosphate (FPP), D-glyceraldehyde 3-phosphate (GAP), Oxaloacetate (Oxa), 5-phospho-a-D-ribose-1-diphosphate (PRPP).

(E4P) (pentose phosphate pathway), serine and glutamate while for Thr only needs Aspartate (Asp). *Portiera* presents an almost complete route for producing the Branched-chain amino acids (BCA): Isoleucine (Ile) from Thr, and Valine (Val) and Leucine (Leu) from host Pyruvate (Pyr). The last step of the BCA biosynthesis is supposed to be complemented by the host, replacing the lack of *ilvE* in *Portiera*, as a putative way of controlling the endosymbiont population (Poliakov *et al.*, 2011; Russell *et al.*, 2013; Shigenobu and Wilson, 2011; Wilson *et al.*, 2010; Xie *et al.*, 2012). Although the last step of Arginine (Arg)[47] biosynthesis is not present in *Portiera*, it is performed by the insect in the *Carsonella*-psyllid system (Sloan and Moran, 2012b; Sloan *et al.*, 2014; Xie *et al.*, 2012) so a similar case cannot be discarded. *Portiera* conserves the last step of Methionine (Met) biosynthesis (*metE*) that use the host Homocysteine (Hcy) to produce Met (Xie *et al.*, 2012).

Finally, although *Portiera* presented most of the Histidine (His) biosynthetic pathway (from the insect D-ribulose-5-phosphate (Rb5P)) it seems that it is not able to produce it. In this way, it has been proposed that His can be found freely in the plant phloem (Douglas, 2006) explaining why other endosymbiotic communities have lost this pathway, such as *Carsonella* or *Tremblaya-Moranella* (Hansen and Moran, 2014). Also, it is also possible that the gut microbiota are synthesizing it, because in Xie *et al.* (2012) a complete bacterial *his* operon were found but it was not detected in *Hamiltonella* genome and *Portiera* lacks the two last steps. However, it is not usually that extremely reduced P-endosymbionts conserve non-functional routes, suggesting that this route could be working in *Portiera* but it is unknown how it is finished because the last two enzymes seem not be present in the host. While it is possible that Phenylalanine (Phe) could be done by *Portiera* (see Section Shared pathways), it presents an incomplete Lysine (Lys) biosynthetic and it seems that this amino acid is synthesized by *Hamiltonella* (see

---

[47] In most of the reported cases of insect endosymbionts, the P-endosymbiont conserves the complete Arg biosynthetic pathway (Hansen and Moran, 2014)

Section *Hamiltonella* biosynthetic capabilities). In contrast to the amino acid biosynthetic machinery present in *Portiera*, its capabilities regarding vitamins/cofactors are scarce. It is only able to produce different carotenes conformations using the Geranylgeranyl diphosphate (GGPP) produced by the host. Although the canonical antioxidant function of carotenes is well known, it is possible that they are also related to an alternative source of reductive power for the endosymbiont and the host (Valmalette *et al.*, 2012). Also, *Portiera* is able to perform some Tetrahydrofolate (THF) transformations to obtain some cofactors, but it is not able to produce it by itself.

4.1.5.2. *Hamiltonella* biosynthetic capabilities

Hamiltonella, that still maintains an almost complete central metabolism (glycolysis, pentose phosphate pathways, etc.), is able to produce the essential amino acids Thr (from imported Asp) and Phe (from imported 2-oxo-3-phenylpropanoate (PPyr)) by itself, and a wide range of vitamins/cofactors without importing any compound from the host cytosol: THF (B9), ubiquinol, Pyridoxal 5'-phosphate (B6), and Riboflavin (B2) and its derivatives Flavin mononucleotide (FMN) and Flavin adenine dinucleotide (FAD). There are other vitamins/cofactors that require the import of some intermediate metabolites: (R)-pantothenate (B5) (probably captured from the diet) for the production of Co-enzyme A (CoA) and Biotin (B7), Asp producing Nicotinamide adenine dinucleotide (NAD), Glutamate (Glu) for protoheme IX and octanoate for the synthesis of lipoate. Thiamin diphosphate (B1) could be acquired from the diet. In the other hand, it is probable that a precursor could be transformed to its active form by the gut microbiota because an *panC* mRNA from *Pseudomonas*, that is usually found in different insets guts, was found in *B. tabaci* transcriptomes.

4.1.5.3. Shared pathways

Phe can be produced by *Portiera* from Chorismate (Chsm) and three non mutually exclusive options are possible for the last step of this pathway:

- *hisC* from *Portiera* has been replaced the transaminase activity of *aspC*.
- The insect is performing this step from the PPyr exported by *Portiera* and coupling it with the synthesis of Tyrosine (Tyr). Some authors consider Tyr as a essential amino acid because it is derived from Phe, which is not produced by the host Chapman (2013).
- The *aspC* encoded by *Hamiltonella* is finishing the pathway importing the PPyr produced by *Portiera*.

Lys biosynthesis is a special case and can be a case of within-pathway complementation with the first steps made by *Hamiltonella* until the N-succynil-2-amino-6-ketopimelate is reached. After that, it is passed to *Portiera* that can transform it to L,L-diaminopimelate and return it to *Hamiltonella* that finish it. Also, according to (Xie *et al.*, 2012) the host can also complement the *argD* absence in *Hamiltonella*, but no transcripts of this gene were detected in *B. tabaci* transcriptome.

In summary it seems that, as other P-endosymbionts, *Portiera* is maintaining only the amino acid biosynthetic capabilities while different S-endosymbionts seems to specialize in supplying all the vitamins/cofactors not synthesized by the host or the P-endosymbiont. Also, it is supposed that the products of this metabolic network are interchanged between the consortia either by osmosis or mediated by transporters. Although the small set of transporters encoded in the *Portiera* genome, it seems that all the required metabolites are covered by at least one transporter: *mgtE* for cations like Mg^{+2}; *argO*, *gltP* and *marC* for amino acids; three transporter with unknown function that could work as more general transporters; *PAQ_166* seems to be also a general transporter related to DitE from *Pseudomonas* (Santos-Garcia *et al.*, 2014b), the Sec translocase for exporting proteins, and a putative ompA-like domain-containing protein (PAQ_222) that could be located in the outer membrane present in *Portiera* (Santos-Garcia *et al.*, 2014a). In contrast, *Hamiltonella* encodes a large set of transporters able to import/export amino acids, vitamins/cofactors, cations, secondary metabolites, etc. Finally, it is possible that *B. tabaci* also encodes a different set of transporters for facilitate the metabolic interchange with the endosymbionts (Poliakov *et al.*, 2011; Price *et al.*, 2011).

The third passenger: *Cardinium* cBtQ1

4.2.1. Background

"*Ca*. Cardinium hertigii" (hereafter *C. hertigii* refers to the holotype) was first characterized in *Encarsia* wasps, which are parasitoids of *B. tabaci*, and it was proposed as the species type (Zchori-Fein *et al.*, 2004). However, in recent years, infections with bacteria belonging to the genus *Cardinium* have been detected not only in whiteflies but also in other insects (armored scale, sharpshooters, and *Culicoides* spp.) and other arthropods (mites, ticks, spiders, and copepods). Nowadays, the infection rate in arthropods has been estimated close to 7 % (Nakamura *et al.*, 2009). Based on molecular data (*16S rRNA* and *gyrB* genes) and the presence of Microtubule-Like Complexes (MLCs), a morphological feature shared by all known *Cardinium*, the genus has been divided into supergroups and strains, following a nomenclature similar to *Wolbachia* endosymbionts, with four described supergroups (A, B, C, and D) (Edlund *et al.*, 2012; Lo *et al.*, 2002; Nakamura *et al.*, 2009).

In several arthropod taxa, *Cardinium* has been described as a reproductive manipulator through diverse effects such as feminization, cytoplasmic incompatibility, and induction of parthenogenesis (White *et al.*, 2011). However, these effects have not been found in other species (e.g., *B. tabaci*), suggesting that *Cardinium* might also be a mutualistic endosymbiont. This aforesaid claim has been supported by the recently released genome of *Cardinium* cEper1 (endosymbiont of the wasp *Encarsia pergandiella*), which encodes a complete biotin biosynthetic pathway, suggesting a potential role in wasp nutrition (Penz *et al.*, 2012).

The laboratory strain *B. tabaci* QHC-VLC harbours *Cardinium* cBtQ1, which belongs to the C1 strain according to its *16S rRNA* gene. This strain coexists within bacteriocytes harbouring *Portiera* and *Hamiltonella* and can also be found scattered in different tissues of the whitefly.

4.2.2. General features of the genome of *Cardinium* cBtQ1

Cardinium cBtQ1's genome size is relatively small (1.065 Mb) and it is composed of a chromosome (1.013 Mb) and a large circular plasmid (52 kb) (Table 4.2.1). The chromosomal sequence is distributed in 11 contigs (ranging from 661.9 to 4.1 kb) with an average of 90X and 547X coverages for 454 and Illumina platforms respectively. The plasmid, named as pcBtQ1, is a single contig with 595X (454) and 4046X (Illumina) coverages. The higher coverage of the plasmid compared to the chromosomal contigs is an indicative of a multicopy plasmid, probably between 5 and 7 copies.

Cardinium cBtQ1 presents 709 and 30 coding genes on the chromosome and the plasmid, respectively (Table 4.2.1). Many of them were annotated as hypothetical or conserved proteins. Moreover, 156 pseudogenes were annotated in the chromosome: 132 derived from transposase genes, 24 from non-transposase (Annex Table A.4.2.1) and 4 in the plasmid (3 transposases and one resolvase). The genome contains one set of rRNA genes distributed in two segments, one including the 16S rRNA and the other the 23S plus the 5S rRNA genes. In addition, a set of 35 tRNA genes, which are able to completely decode the mRNA sequences and two other noncoding RNA genes (*rnpB* and *tmRNA*) were annotated (Table 4.2.1).

Also, *Cardinium* cEper1 genome, an endosymbiont of *Encarsia pergandiella* (a parasitoid wasp from *B. tabaci*) are publicly available (Penz *et al.*, 2012). Its genome is smaller than *Cardinium* cBtQ1's and *A. asiaticus*' ones. The number of genes in *Cardinium* cBtQ1 is smaller than in *Cardinium* cEper1, in spite of the former having a larger genome (Table 4.2.1). The number of pseudogenes in *Cardinium* cBtQ1 was closer to those in *A. asiaticus* than to *Cardinium* cEper1. Nonetheless, most of these differences in gene number and pseudogenes could be due to the gene annotation criteria followed[48].

Table 4.2.1 General Genomic Features of *Cardinium* Strains and *Amoebophilus asiaticus*

Bacterial genome	***Cardinium* cBtQ1**[a]		***Cardinium* cEper1**[b]		***A. asiaticus* 5a2**
Host	*Bemisia tabaci*		*Encarsia pergandiella*		*Acanthamoeba* spp.
	Chromosome	Plasmid	Chromosome	Plasmid	Chromosome
Contigs	11	1	1	1	1
Size (kb)	1,013	52	887	58	1,884
GC (%)	35	32	36	31	35
CDS	709	30	841	65	1,557
Average CDS length (bp)	1,033	1,389	911	733	990
Coding density (%)	79.7	80.1	85.5	82.1	81.8
rRNAs	3	-	3	-	3
tRNAs	35	-	37	-	35
Other RNA genes	2	-	-	-	-
Pseudogenes (total)	156	4	3	-	222
Pseudogenes (transposase)	132	3	3	-	-
Pseudogene (other CDS)	24	1	-	-	-

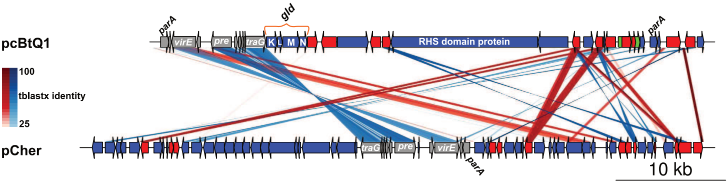
[a] High quality draft genome [b] Contains a single gap not closed due to repetitive elements

The average gene identity between *Cardinium* cEper1 and cBtQ1 was 92.9 % at nucleotide level and 91.8 % at amino acid level (with a standard deviation of 2.3 % and 4.7 % respectively). The genome fraction assigned to coding genes (designed as coding density in Table 4.2.1) was approximately 6 % smaller in *Cardinium* cBtQ1 than in *Cardinium* cEper1. Based on genomic features, the extent of the process of genome reduction has been higher in *Cardinium* cEper1 than in *Cardinium* cBtQ1.

Both *Cardinium* contain a plasmid of similar size (Table 4.2.1) but only a few genes are shared. These shared genes form a syntenic segment with a high level of nucleotide identity (Figure 4.2.1). For example, *Cardinium* cBtQ1's genes *CHV_p006* (*pre*, plasmid recombination enzyme), *CHV_p008* (*CHV_p008*, hypothetical protein) and *CHV_p011* (*traG*, putative conjugal transfer protein TraG) display a nucleotide identity that range from 83 to 91 % with their corresponding orthologous genes in *Cardinium* cEper1's plasmid. The degree of gene conservation and the conserved synteny in those genes may be an indication that both plasmids derive from the same ancestral plasmid and differences are due to the insertion of mobile elements, which can sometimes carry accessory genes. In addition, phylogenetic analysis corroborated that most of the shared genes form a monophyletic clade (e.g. *traG* gene at Annex Figure A.4.2.1)

---

[48] As an exmaple, some partial transposase domains are considered as CDS in automatic annotations instead pseudogenes



Figures 4.2.1 TBLASTX comparison of the plasmids from *Cardinium* cBtQ1 (pcBtQ1) and cEper1 (pCher). Gray arrows are genes included in the syntenic block, blue arrows non-transposase genes, red arrows transposase genes, and the green arrow is a resolvase pseudogene. Red and blue lines show genes in the same and in reverse orientation, respectively. Some gene names are shown in the plot.

4.2.3. Taxonomic status of *Cardinium* cBtQ1

To establish the relationship of *Cardinium* cBtQ1 to other *Cardinium* endosymbionts based on 16S rRNA sequences, a covariance model aligner was employed. This phylogeny showed that almost all *Cardinium* endosymbionts of *B. tabaci* (including cBtQ1) are present in a clade with other *Cardinium* endosymbionts of several *Encarsia* species (Figure 4.2.2, left). A phylogeny with *gyrB* coding genes was also performed, which corroborated the close phylogenetic relationship with *Cardinium* from *Encarsia* spp., but also showed that *Cardinium* cBtQ1 was embedded in the *Cardinium*-*Encarsia* clade and very close to *Encarsia inaron*[49] and *Encarsia pergandiella*[50] (Figure 4.2.2, right). Because 16S rRNA sequences of *Cardinium* cBtQ1 and the species type *C. hertigii* (symbiont of *Encarsia hispida*, (Zchori-Fein *et al.*, 2004)) show only 1.2 % of differences at a nucleotide level, it means that *Cardinium* cBtQ1 is a strain of the latter (Stackebrandt and Ebers, 2006), in agreement with previous authors (Zchori-Fein and Perlman, 2004). Finally, this *Bemisia/Encarsia* clade belongs to the *Cardinium* group A, that is well differentiated from the other two groups included in the analysis: the group C, specific of the genus Culicoides (Nakamura *et al.*, 2009) and the group D, present in some *Copepoda* spp. (Edlund *et al.*, 2012) (Figure 4.2.2, see Annex Table A.4.2.2 for the genbank ID of each 16S rRNA gene).

---

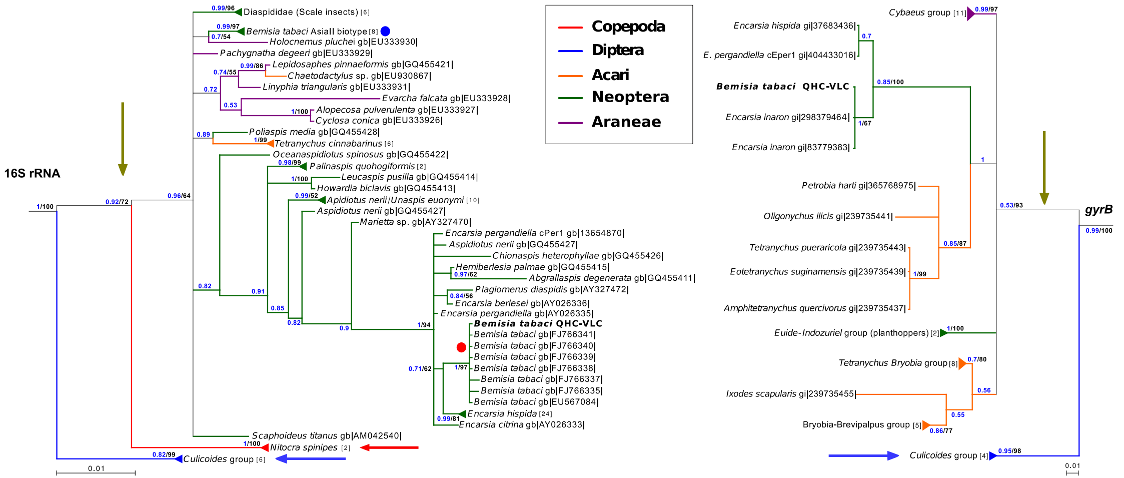
[49] >99 % identity for a *gyrB* 838 bp gene fragment

[50] 99.14 % identity for the whole 16S rRNA gene

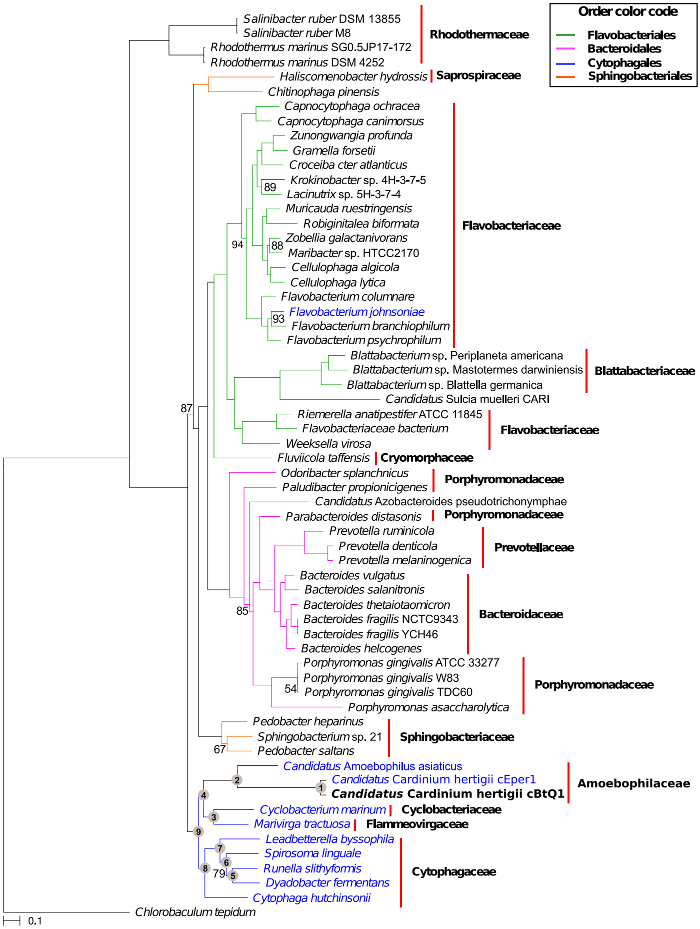
As previously described in Zchori-Fein *et al.* (2004), *C. hertigii* is closely related to the amoeba parasitic endosymbiont *A. asiaticus*. With the genome of *Cardinium* cBtQ1 sequenced and annotated, a Bacteroidetes phylogenomic reconstruction was performed, (see Annex Table A.4.2.2 for the locus tags used from each genome) finding that *Cardinium* and *A. asiaticus* formed a well differentiated clade related to the families Cyclobacteriaceae and Flammeovirgaceae, with the family Cytophagaceae slightly more distant. Because this phylogenomic reconstruction is consistent to other reported studies (Gupta and Lorenzini, 2007; Karlsson *et al.*, 2011) and due to the high bootstrap values obtained in the phylogeny, the *Cardinium/Amoebophilus* clade was proposed to form a new family and to be assigned to the order Cytophagales, instead of remaining in the non-classified Bacteroidetes. Moreover, the proposed name for the family was Amoebophilaceae[51], identified with the 1501348 Taxon ID at NCBI, related to the Cyclobacteriaceae and Flammeovirgaceae families (Figure 4.2.3). Finally, this phylogenomic reconstruction was used to select the genomes to compare in subsequent analyses and in the LCA reconstruction (denoted as grey numbered dots in 4.2.3).

---

[51] Naming as it is because *A. asiaticus* was the species with the first genome sequenced


16S rRNA
1/100
0.92/72
0.96/64
0.99/96 Diaspididae (Scale insects) [6]
0.99/97 Bemisia tabaci AsiaII biotype [8]
0.7/54 Holocnemus pluchei gb|EU333930|
Pachygnatha degeeri gb|EU333929|
0.99/86 Lepidosaphes pinnaeformis gb|GQ455421|
0.74/55 Chaetodactylus sp. gb|EU930867|
Linyphia triangularis gb|EU333931|
0.72
Evarcha falcata gb|EU333928|
0.53
1/100 Alopecosa pulverulenta gb|EU333927|
Cyclosa conica gb|EU333926|
0.89 Poliaspis media gb|GQ455428|
1/99 Tetranychus cinnabarinus [6]
Oceanaspidiotus spinosus gb|GQ455422|
0.98/99 Palinaspis quohogiformis [2]
1/100 Leucaspis pusilla gb|GQ455414|
Howardia biclavis gb|GQ455413|
0.99/52 Apidiotus nerii/Unaspis euonymi [10]
Aspidiotus nerii gb|GQ455427|
Marietta sp. gb|AY327470|
Encarsia pergandiella cPer1 gb|13654870|
Aspidiotus nerii gb|GQ455427|
Chionaspis heterophyllae gb|GQ455426|
Hemiberlesia palmae gb|GQ455415|
0.97/62 Abgrallaspis degenerata gb|GQ455411|
Plagiomerus diaspidis gb|AY327472|
0.84/56 Encarsia berlesei gb|AY026336|
Encarsia pergandiella gb|AY026335|
Bemisia tabaci QHC-VLC
Bemisia tabaci gb|FJ766341|
Bemisia tabaci gb|FJ766340|
Bemisia tabaci gb|FJ766339|
1/97 Bemisia tabaci gb|FJ766338|
Bemisia tabaci gb|FJ766337|
Bemisia tabaci gb|FJ766335|
Bemisia tabaci gb|EU567084|
Encarsia hispida [24]
0.99/81 Encarsia citrina gb|AY026333|
0.82
0.91
0.85
0.82
0.9
1/94
0.71/62
Scaphoideus titanus gb|AM042540|
1/100 Nitocra spinipes [2]
0.82/99 Culicoides group [6]
0.01
Copepoda
Diptera
Acari
Neoptera
Araneae
Cybaeus group [11] 0.99/97
Encarsia hispida gi|37683436|
0.7
E. pergandiella cEper1 gi|404433016|
0.85/100
Bemisia tabaci QHC-VLC
Encarsia inaron gi|298379464|
1/67
Encarsia inaron gi|83779383|
1
Petrobia harti gi|365768975|
Oligonychus ilicis gi|239735441|
0.53/93
gyrB
0.99/100
Tetranychus pueraricola gi|239735443|
0.85/87
Eotetranychus suginamensis gi|239735439|
1/99
Amphitetranychus quercivorus gi|239735437|
Euide-Indozuriel group (planthoppers) [2] 1/100
Tetranychus Bryobia group [8] 0.7/80
0.56
Ixodes scapularis gi|239735455|
0.55
Bryobia-Brevipalpus group [5] 0.86/77
Culicoides group [4] 0.95/98
0.01

Figures 4.2.2 Phylogenetic relationships among different *Cardinium* supergroups are shown based on the 16S rRNA and *gyrB* genes. The *Cardinium* group A strains (green vertical arrow) seems the most widespread group, infecting different arthropod orders including the subclass *Neoptera* (which includes whiteflies). Groups C (blue arrow) and D (red arrow) showed a more restricted pattern of hosts. Group C is restricted to the order *Diptera* while Group D is detected on subclass *Copepoda*. On the 16S rRNA phylogeny, *Cardinium* endosymbionts of *B. tabaci* are grouped in two different strains, the C1 (red dot) that belongs to the Mediterranean species (Q biotype) and the CI1 (blue dot) that belongs to the Asia II species. The GTR+I+G model was selected for 16S rRNA and the LG+G model for gyrB. Cardinium strain cBtQ1, endosymbiont of *B. tabaci* QHC-VLC, is displayed in bold type. In both cases, *A. asiaticus* was used as the outgroup, but was excluded from the figure for plotting reasons. Triangles represent collapsed branches of the same species or genus with the number of collapsed sequences between square brackets. *Arthropoda* taxonomic names and their respective colours are shown in the upper box. Accession numbers for non collapsed branches are displayed. Only Bayesian posterior values above 0.5 are displayed (blue) and branches under this score were condensed. Bootstrap values from Maximum Likelihood phylogenetic reconstruction above 50 % are also displayed (black).

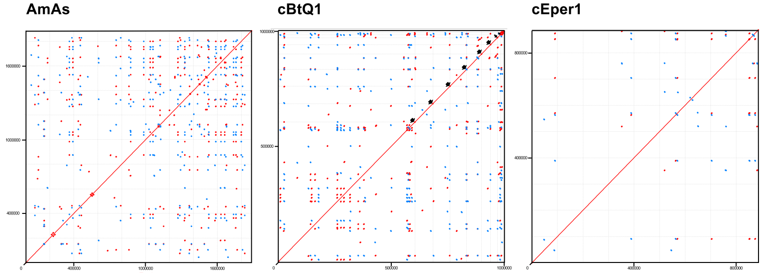

Order color code
Flavobacteriales
Bacteroidales
Cytophagales
Sphingobacteriales
Salinibacter ruber DSM 13855
Salinibacter ruber M8
Rhodothermus marinus SG0.5JP17-172
Rhodothermus marinus DSM 4252
Rhodothermaceae
Haliscomenobacter hydrossis
Saprospiraceae
Chitinophaga pinensis
Capnocytophaga ochracea
Capnocytophaga canimorsus
Zunongwangia profunda
Gramella forsetii
Croceiba cter atlanticus
Krokinobacter sp. 4H-3-7-5
89
Lacinutrix sp. 5H-3-7-4
Muricauda ruestringensis
Robiginitalea biformata
Flavobacteriaceae
94
Zobellia galactanivorans
88
Maribacter sp. HTCC2170
Cellulophaga algicola
Cellulophaga lytica
Flavobacterium columnare
Flavobacterium johnsoniae
93
Flavobacterium branchiophilum
Flavobacterium psychrophilum
Blattabacterium sp. Periplaneta americana
Blattabacterium sp. Mastotermes darwiniensis
Blattabacteriaceae
Blattabacterium sp. Blattella germanica
Candidatus Sulcia muelleri CARI
87
Riemerella anatipestifer ATCC 11845
Flavobacteriaceae bacterium
Flavobacteriaceae
Weeksella virosa
Fluviicola taffensis
Cryomorphaceae
Odoribacter splanchnicus
Porphyromonadaceae
Paludibacter propionicigenes
Candidatus Azobacteroides pseudotrichonymphae
Parabacteroides distasonis
Porphyromonadaceae
Prevotella ruminicola
Prevotella denticola
Prevotellaceae
Prevotella melaninogenica
Bacteroides vulgatus
Bacteroides salanitronis
Bacteroides thetaiotaomicron
85
Bacteroides fragilis NCTC9343
Bacteroidaceae
Bacteroides fragilis YCH46
Bacteroides helcogenes
Porphyromonas gingivalis ATCC 33277
Porphyromonas gingivalis W83
Porphyromonadaceae
54
Porphyromonas gingivalis TDC60
Porphyromonas asaccharolytica
Pedobacter heparinus
Sphingobacterium sp. 21
Sphingobacteriaceae
67
Pedobacter saltans
Candidatus Amoebophilus asiaticus
2
Candidatus Cardinium hertigii cEper1
Amoebophilaceae
1
Candidatus Cardinium hertigii cBtQ1
4
Cyclobacterium marinum
Cyclobacteriaceae
3
Marivirga tractuosa
Flammeovirgaceae
9
Leadbetterella byssophila
7
Spirosoma linguale
6
Cytophagaceae
Runella slithyformis
8
79
5
Dyadobacter fermentans
Cytophaga hutchinsonii
Chlorobaculum tepidum
0.1

Figures 4.2.3 Phylogenomic maximum likelihood reconstruction was done under the LG+G+F model on a concatenated alignment of 37 proteins. *Cardinium* genomes fall in the Cytophagales clade, with *Marivirga tractuosa* and *Cyclobacterium marinum* as the closest free-living relatives. *Cardinium* cBtQ1 is displayed in bold. Family names are displayed on the right delimited by a horizontal red line. The genomes used for the LCA reconstruction are shown in blue. Numbers inside grey dots show the LCAs reconstructed in each node. Only maximum likelihood bootstrap values below 95 % are displayed. Bayesian posterior probabilities for each node were above 0.95 and are also not displayed. *Chlorobaculum tepidum* was used as outgroup.

4.2.4. Comparative genomics

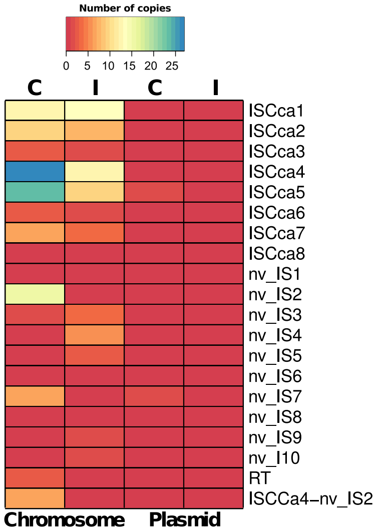
4.2.4.1. Mobile elements and genomic redundancy

The level of redundancy in the genome of *Cardinium* cBtQ1 (≈14 %) was twice as high as the level found in *Cardinium* cEper1 and *A. asiaticus* (≈7 % in both cases), with most of it associated to mobile elements (Figure 4.2.4). These mobile elements, are a typical feature of endosymbionts that have established a recent relationship with their hosts, such as *Sodalis pierantonius* from *Sitophilus oryzae* (formerly SOPE) (Gil *et al.*, 2008), *Sodalis glossinidius* from *Glossina morsitans* (Belda *et al.*, 2010), *S. symbiotica* from *C. cedri* (Manzano-Marín and Latorre, 2014). Also, other facultative endosymbionts with an unclear symbiotic relationship show an enrichment mobile elements: *Rickettsia* endosymbiont of *Ixodes scapularis* (Gillespie *et al.*, 2012), *Wolbachia wMel* endosymbiont of *Drosophila melanogaster* (Wu *et al.*, 2004) or *A. asiaticus* endosymbiont of *Acanthamoeba* sp. (Schmitz-Esser *et al.*, 2011).



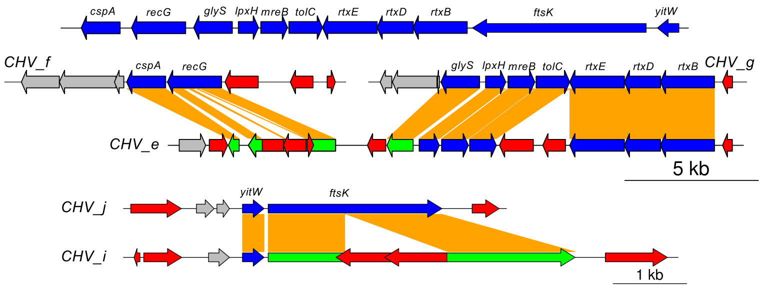
Figures 4.2.4 Mummer plot showing direct (red) and inverted (blue) genomic repeats with at least 500 base pair lengths and 95 % similarity. For *A. asiaticus* (AmAs) and *Cardinium* cEper1 (cEper1), inner plot lines denote the division of the chromosome in base pairs sections. Black arrows point contig ends for the largest contigs in *Cardinium* cBtQ1. These contigs were placed in order of decreasing length. Because plots are not scaled to genome size due to limitations of the software, it is noteworthy that the *A. asiaticus* genome is less repetitive than *Cardinium* cBtQ1 although the more compact plot in the former may alter that impression.

Cardinium cBtQ1's mobile elements, and their inactive derivatives, account for approximately 166 kb of the chromosome (196 copies) and 12.5 kb of the plasmid (12 copies) (Figure 4.2.5 and Annex Table A.4.2.3). From this number of mobile element copies, only 48 contained a functional transposase gene (eight in the plasmid) while 132 were transposase pseudogenes (three in the plasmid). These transposase proteins were classified in 20 different IS families, with only eight being complete IS elements (containing intact transposase genes and inverted repeats at their ends) and were named according to the ISfinder recommendations and deposited under the names ISCca1-8 (Figure 4.2.5 and Annex Table A.4.2.3). Only three mobile element types were specific of the *Cardinium* cBtQ1 (ISCca6, nv_IS3 and the Retron type one), while the rest of transposases were shared with *A. asiaticus*, *Cardinium* cEper1 or both. Some transposases are closely related to α-proteobacteria, probably from to the genera *Rickettsia* or *Wolbachia*, which are also secondary endosymbionts of *B. tabaci* and other arthropods. This supports the idea of HGT events between S-endosymbionts present in the same host (Duron, 2013; Penz *et al.*, 2012; Schmitz-Esser *et al.*, 2011; Toft and Andersson, 2010).



Figures 4.2.5 Insertion elements present in *Cardinium* cBtQ1 grouped as validated by the ISfinder (ISCca) and non-validated (nv_IS). IS are classified in functional copies (C) or inactive derivatives (I).

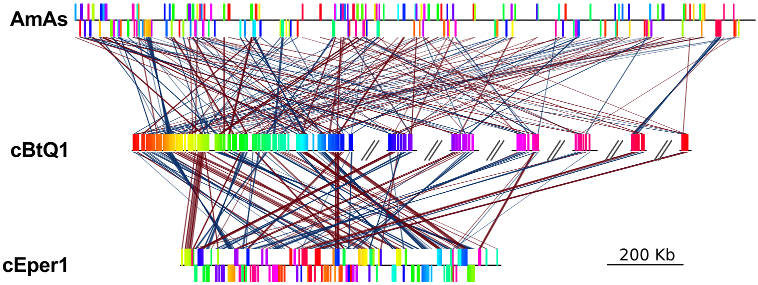
A possible signal that at least some IS are still active (e.g. ISCca4 and 5, see Figure 4.2.5), in contrast with *A. asiaticus*' case (Schmitz-Esser *et al.*, 2011), is that there are cases of very recent gene duplications (based on >99.9 % nucleotide identity) with one of the copies being later inactivated by the insertion of an IS (pseudogenes *recG* (*CHV_e0046*) or *ftsK* (*CHV_i0005*)) (Figure 4.2.6). Another important feature is the presence of a repetitive element composed by a copy of ISCca4 and a copy of nv_IS2, resulting in a composed IS that apparently can jump by itself. The inactivation of *ftsK* (*CHV_i0005*) was produced by the insertion of this composed IS. It is interesting that *lpxH*, *mreB*, *tolC*, *rtxBDE* and *yitW* conserve two intact copies while the other duplicated genes only maintain one, suggesting that these genes could provide an important function for *Cardinium* cBtQ1 (Figure 4.2.6).



Figures 4.2.6 Putative linear representation of the ancestral genomic region before duplication (on top) and the present state of the two duplications, which are distributed in 5 contigs (on bottom). Red arrows are mobile elements, blue arrows genes in the duplicated region, green arrows pseudogenized genes and grey arrows adjacent genes outside the duplication. Orange bars connect the two duplicated copies of each gene. Contig names are plotted at the beginning or the end of the contig (CH_) and only regions that contain the duplications are shown. The right ends of contigs CHV_g and CHV_e are connected through paired-end information with the right ends of either contig CHV_j or CHV_i. In both cases a complete ISCca1 copy, whose fragments are detected at the end of the contigs, is required for joining.

The presence of active IS elements and the high number of transposase copies throughout the genome, in combination with a complete replication and repair machinery that can produce recombination, is probably the cause of the massive number of rearrangements in the genome of *Cardinium* cBtQ1. While some microsynteny is still observed, the aforementioned statement explains the loss of macrosynteny when synteny blocks are compared between the genome of *Cardinium* cBtQ1, *Cardinium* cEper1 and *A. asiaticus* (Figure 4.2.7).

Taken all the data together, *Cardinium* cBtQ1 could be a facultative endosymbiont, like *Wolbachia* or *Rickettsia* endosymbionts, able to adapt to different niches due to the genome plasticity given by the active IS (Toft and Andersson, 2010). In contrast, *Cardinium* cEper1 is in an advanced genome reduction process with most of the IS elements, if not all, inactivated and in degradation process.

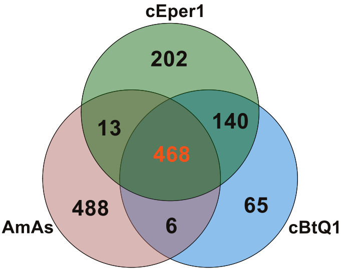


Figures 4.2.7 Common pairwise syntenic blocks of more than 1 kb for *A. asiaticus* (AmAs), *Cardinium* cBtQ1 (cBtQ1) and cEper1 (cEper1). The chromosome of cBtQ1 was taken as reference. Contigs in cBtQ1 are ordered in order of decreasing length and denoted by double backslashes. For plotting reasons, only the seven largest cBtQ1 contigs are shown. Red and blue lines show blocks in direct and inverted orientation. The stronger the line, the more nucleotide identity between synteny blocks.

4.2.4.2. Comparative genomics of *Cardinium* strains and *A. asiaticus*

Both *Cardinium* strains and *A. asiaticus* share a core genome of 468 CDS clusters, including 6 CDS clusters encoding putative host-interacting proteins. There are 140 unique CDS clusters present in both *Cardinium* but not in *A. asiaticus*, with an important part of them encoding hypothetical proteins (46), some membrane transport related proteins (15) and some putative host-interacting proteins (13). Among the remaining shared genes between both *Cardinium*, it is possible to find transposases (6), phage-derived proteins (Anti-feeding Prophage (Afp)-like proteins) (2), and some genes encoding vitamin biosynthetic proteins (5). *Cardinium* cEper1 has 202 strain specific gene clusters, which include, among others, CDS encoding hypothetical proteins (145), transposases (30), host-interacting proteins (6), and biosynthetic enzymes related to biotin (2) and pyridoxal (1) biosynthesis. *Cardinium* cEper1 and *A. asiaticus* share 13 gene clusters with most of them defined as hypothetical proteins (6), mobile elements (3), a cell-wall related protein, a membrane protein and a host-manipulation protein (Figure 4.2.8, Annex Table A.4.2.4).

Cardinium cBtQ1 contains 71 gene clusters (65 strain specific and 6 shared with *A. asiaticus*) that are not present in *Cardinium* cEper1. They include ankyrin-domains containing proteins (14), hypothetical proteins (35), and mobile elements (4). Because proteins with ankyrin domains can interact with the host's machinery, these proteins could yield some clues about the relationship, and the settlement, of *Cardinium* in the whitefly, but further studies are needed in this direction. The most interesting strain specific genes of *Cardinium* cBtQ1 are located in the multicopy plasmid. They include four gliding genes (*gldK*, *gldL*, *gldM* and *gldN*, see Figure 4.2.1) related to mobility in members of the phylum Bacteroidetes (also present in *A. asiaticus*) and the strain specific gene *CHV_p021* (ca. 14 kb). The fact that the chromosome contains four duplicated genes (*rtxB*, *rtxD*, *rtxE* and *tolC*) related to



Figures 4.2.8 Euler diagram representing the pan-genome, the core genome, the strain specific orthologous CDS clusters and the clusters shared by only two organisms. Numbers inside each subspace represent the number of orthologous CDS clusters assigned to its corresponding subspace. Core genome set is displayed in orange. Abbreviations: *Cardinium* cBtQ1 (cBtQ1), *Cardinium* cEper1 (cEper1), *A. asiaticus* (AmAs).

Type 1 Secretion System (T1SS)[52] is also remarkable because only a few sequenced Bacteroidetes harbour secretion systems type I, III, IV or VI (McBride and Zhu, 2013). The *rtxBDE* cluster seems to be another event of HGT, with RTX toxin transport system of *Vibrio* as **BLAST** best hits. The chromosomal segment involving these genes is duplicated in *Cardinium* cBtQ1 (Figure 4.2.6). It is important to state that genes related to motility were present in the *Cardinium* ancestor but were lost in *Cardinium* cEper1[53]. The *CHV_p021* gene encodes a RHS-repeat associated-core domain protein with C-terminus ankyrin repeats that seems a recent acquisition from an Alphaproteobacteria (probably a *Wolbachia* according to **BLAST** similarities). These kind of large proteins with RHS domains have been related with bacterial insecticidal toxins and intercellular signalling proteins (TIGR03696). The presence of ankyrins in the C-terminus domain in combination with a signal peptide has been attributed to protein secreted by T1SS (Kaur *et al.*, 2012). Although no clear signal peptide was bioinformatically detected in *CHV_p021*, it can not be rule out the possibility that this protein could be secreted. Because the best Blastx hits, with a 63 % query coverage and 36 % identity on average, belong to *Daphnia*, *Wolbachia* and different mosquitoes, this protein could be related to some conserved proteins in arthropods that are also exploited by *Wolbachia*. Whether this protein is a toxin or a host-interacting protein still remains unclear, but the fact that the gene is located in a multicopy plasmid and maybe could be secreted leads to consider that it is important for *Cardinium* cBtQ1 and its settlement in *B. tabaci*.

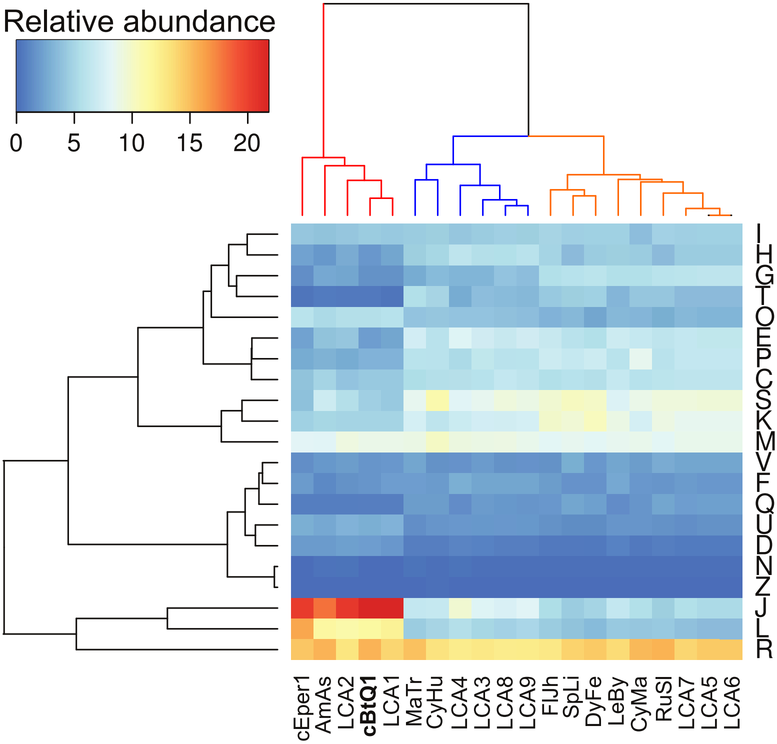
4.2.4.3. Evolution of gene repertoires in the lineages of *A. asiaticus* and *Cardinium*

Hierarchical clustering based on the relative abundance (percentage) of each **COG** category in each genome and LCAs was performed

---

[52] They form only 3 gene clusters because OrthoMCL placed the duplicated genes *rtxB* and *rtxE* in the same cluster

[53] See Evolution of gene repertories 4.2.4.3



C Energy production and conversion
D Cell cycle control, cell division and chromosome partitioning
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
J Translation, ribosomal structure and biogenesis
K Transcription
L Replication, recombination and repair
M Cell wall/membrane/envelope biogenesis
N Cell motility
O Post-translational modification, protein turnover, chaperones
P Inorganic ion transport and metabolism
Q Secondary metabolites biosynthesis, transport and catabolism
R General function prediction only
S Function unknown
T Signal transduction mechanisms
U Intracellular trafficking, secretion and vesicular transport
V Defense mechanisms

Figures 4.2.9 Hierarchical clustering heatmap representing the relative abundance (percentage) of each **COG** category in relation to the total number of gene clusters in each genome. Three main **COG** clusters (left) are observed: highly retained categories (**J**, **L**, **R**), medium retained categories (**I**, **H**, **G**, **T**, **O**, **E**, **P**, **C**, **S**, **K**, **M**) and low retained categories (**V**, **F**, **Q**, **U**, **D**, **N**, **Z**). Three main species/LCA cluster (up) are: cEper1, AmAs, cBtQ1, LCA1, LCA2 (only symbionts, left cluster); MaTr, CyHu, LCA3, LCA4, LCA8, LCA9 (middle cluster) and FlJh, SpLi, DyFe, LeBy, CyMa, RuSl, LCA5, LCA6, LCA7. Species clustering together by **COG** categories could have similar metabolic features and consequently, a similar ecological niche. *Cardinium* cEper1 (cEper1), *A. asiaticus* (AmAs), *Cardinium* cBtQ1 (cBtQ1), *M. tractuosa* (MaTr), *Cytophaga hutchinsonii* (CyHu), *Flavobacterium johnsoniae* (FlJh), *Spirosoma linguale* (SpLi), *Dyadobacter fermentans* (DyFe), *Leadbetterella byssophila* (LeBy), *C. marinum* (CyMa), *Runella slithyformis* (RuSl).

(Figure 4.2.9). Three main clusters were observed: one that contained the endosymbiotic genomes and the LCA1 and 2; a second that grouped *M. tractuosa* and *C. marinum* with the LCA3, 4, 8,and 9; and a third that contained the rest of the genomes and LCAs. The second cluster (Figure 4.2.9 blue) showed a clear reduction in some **COG** groups as **G** (Carbohydrates transport and metabolism) and **K** (transcription) but an enrichment in the **H** (coenzyme metabolism) and **J** (translation, ribosomal structures and biogenesis) groups when it was compared with the third cluster (Figure 4.2.9 orange). Hierarchical clustering indicates that LCA3 to 9 were, similar to free-living Bacteroidetes, able to occupy different niches. For example, the differences between the abundance of **G** category in the middle and right clusters could be related to a more restricted source of carbohydrates (niche specialization). It also seems that the increase of the **H** category in the middle cluster could be advantageous for the establishment of symbiotic relationships (*Cyclobacterium* was found in the celomic fluid of a sand dollar, Annex Table A.4.2.5). The symbiotic cluster (Figure 4.2.9 red) showed a stronger retention of genes in **J** (translation, ribosomal structure and biogenesis), **L** (replication, recombination and repair) and **O** (post-translational modification, protein turnover and chaperones) **COG** categories when were compared to the free-living Bacteroidetes genomes, a signal also observed in other symbiotic reduced genomes (Karlsson *et al.*, 2011). Attending to that the **E** category (amino acid transport and metabolism) was reduced in this cluster, it is clear that the common ancestor of both *Cardinium* strains and *A. asiaticus*, the LCA2, was also an endosymbiont with few biosynthetic capabilities (Figure 4.2.9, Annex Table A.4.2.5).

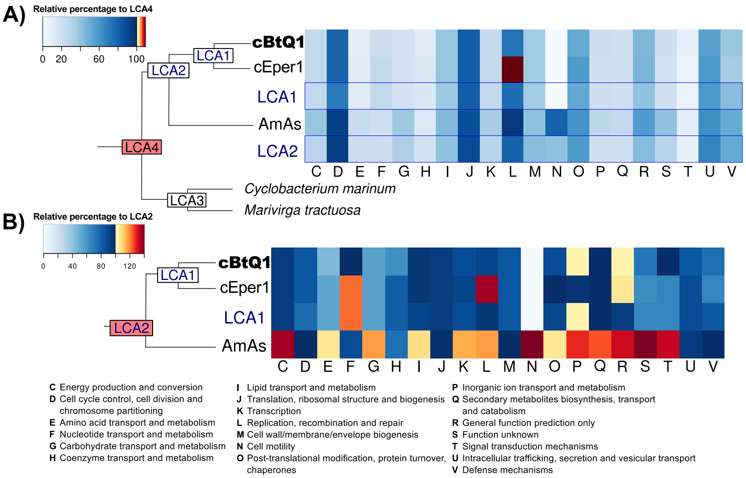
Because clustering of functional categories, like **COG** and **KEGG**, are in some manner correlated to the habitat, **COG** profiles were compared and and it was found that LCA4, the ancestor of the *Cardinium/A. asiaticus* lineage and family Cyclobacteriaceae, was close to the free-living Cyclobacteriaceae (Karlsson *et al.*, 2011). Because Cyclobacteriaceae seems to be predominantly a marine-related family that can stablish symbiotic relationships with different hosts, the **COG** profile of LCA4[54] result suggests that it was presumably a marine free living-bacterium with maybe the ability to stablish symbiotic relationships (Figure 4.2.9, Annex Table A.4.2.5). Also, it is likely that LCA4 was able to glide because it contained the whole set of gliding genes essential for gliding, including the *sprATE* genes (McBride and Zhu, 2013). This seems in concordance with a recently proposed evolutionary hypothesis where the ancestor of *Cardinium* changed its lifestyle from aquatic amoeba to arthropods (Penz *et al.*, 2012).

The transition from LCA4 to LCA2 had a strong impact in the number of gene clusters with more than half of them being lost (LCA2, 655 gene clusters plus 36 present/absent). The decrease was high for all **COG** categories except for some housekeeping categories such as **J**, **L** and **D** (cell cycle control, cell division and chromosome partitioning) (Figure 4.2.10 A, Annex Table A.4.2.5).

The transition from LCA4 to LCA2 was clearly a reductive process that affected almost all **COG** categories (Figure 4.2.10 A) producing an ancestral endosymbiont with few biosynthetic capabilities. Considering that the species derived from LCA2 were endosymbionts of amoebas (Horn *et al.*, 2001; Schmitz-Esser *et al.*, 2011) or insects (Penz *et al.*, 2012; Zchori-Fein *et al.*, 2004), the most probable reason for this reduction was the transition from a free living to intracellular life style, to start a symbiotic (either mutualistic or parasitic) relationship with a eukaryotic host. During this transition, the number of gene clusters and associated functions was reduced, although LCA2 maintained the ability to acquire new genes by HGT. In contrast, the transition from LCA2 to LCA1 (649 gene clusters) produced the loss of 160 gene clusters, although 118 new genes were acquired (Figure 4.2.10 B). Comparing the number of gene clusters of LCA2 to LCA1, and to both *Cardinium* and *A. asiaticus*, several differences were observed among **COG** categories (Figure 4.2.10

---

[54] The parsimony reconstruction assigned 1301 gene clusters to LCA4, and the equally parsimonious presence/absence of other 684 gene clusters



Figures 4.2.10 **A)** Heatmap showing the percentage of genes in each **COG** category, compared to the number of the same category in LC4 (100 %). In left, reduced phylogenomic reconstruction with the name of each Last Common Ancestor reconstructed. **B)** The same heatmap type comparing to LCA2 (100 %). **L** category in *Cardinium* cEper1 is an artefact produced by an incorrect annotation of inactivated transposases as CDS instead of pseudogenes. **COG** definitions are the same as Figure 4.2.9. Abbreviations: *Cardinium* cBtQ1 (cBtQ1), *Cardinium* cEper1 (cEper1), *A. asiaticus* (AmAs).

B , Annex Table A.4.2.5). First of all, *A. asiaticus* showed 331 strain specific gene clusters, distributed in several categories, not present in LCA2. This difference could be due to specific gene acquisitions in *A. asiaticus* but the possibility of a biased sample of genomes[55] and different annotation problems leaded to an overestimation of strain specific clusters. Secondly, the reductive evolution of the *Cardinium* lineage was more clearly observed in several **COG** categories, such as **E**, **G**, **H**, **S** (function unknown), **T** (signal transduction mechanisms) and **V** (defense mechanisms). The absence of gene clusters in *Cardinium* for the **N** (Cell motility) category was probably due to the fact that some genes related with motility have not been yet annotated in the **COG** database, especially those involved in gliding motility (discussed later) that, in fact, are present in *Cardinium* cBtQ1, LCA1, LCA2 and LCA4 (Figure 4.2.10 B).

---

[55] Maybe more genomes are needed for this kind of inferences

These results suggest that LCA1 contained the core genome of *Cardinium* and that the different *Cardinium* strains that compose supergroups seem to vary only in a few categories. These differences may provide the different strains the ability to exploit new niches (such as a new host) in a similar way of *Wolbachia* (Ellegaard *et al.*, 2013).

4.2.5. Biosynthetic capabilities in *Cardinium* cBtQ1

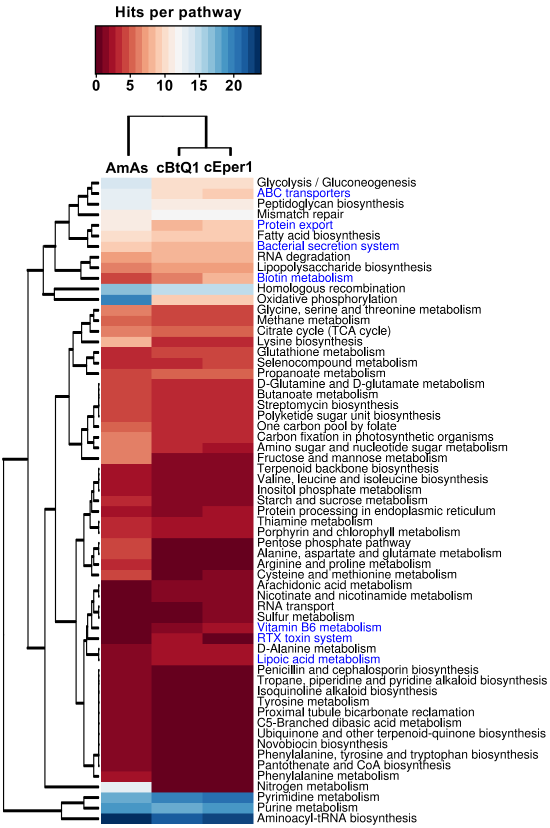
Cardinium cBtQ1, according to **KEEG** classification pathways, presents low biosynthetic capabilities (Figure 4.2.11), similar to those observed in *Cardinium* cEper1 and *A. asiaticus* (Karlsson *et al.*, 2011; Penz *et al.*, 2012). This was also confirmed after reconstructing *Cardinium* cBtQ1 metabolism with pathway-tools[56].

The main differences between the biosynthetic capabilities of both *Cardinium* strains are the biosynthesis of vitamins and cofactors. Both bacteria are able to produce lipoate, a key cofactor for intermediate metabolism and an important antioxidant molecule (Spalding and Prigge, 2010). While *Cardinium* cEper1 has the genes *pdxS* and *pdxT* and it can synthetize pyridoxal 5-phosphate (precursor of vitamin B6), the gene *pdxT* was pseudogenized by an IS transposition in cBtQ1. This event seems to have happened recently, because the *pdxT* pseudogene is 93.5 % identical to the cEper1 gene, a percentage higher than the average gene identity between these two strains. In addition, it is noteworthy that *Cardinium* cEper1 has maintained a complete biotin operon, a co-enzyme belonging to vitamin B class, that is a case of HGT from Alphaproteobacteria in the genus *Cardinium* (Figure 4.2.11).

Biotin can contribute with some benefits to the *E. pergandiella* host, although it is not experimentally demonstrated (White *et al.*, 2009, 2011). The loss of the ability to synthesize biotin in *Cardinium* cBtQ1 seems to have taken place by the combined effect of the insertion of a IS and

---

[56] Publicly available from BioCyc registry `http://biocyc.org/registry.html`



Figures 4.2.11 Comparative genomic analysis between *Cardinium* strains and *A. asiaticus* genomes. Heatmap showing the number of hits per KEGG pathway. Only pathways that showed differences between the three genomes were plotted. Interesting pathways from a comparative point of view between *Cardinium* strains are denoted in blue. Ribosomal and miss-sense pathways (like Cancer) were deleted. Abbreviations: *Cardinium* cBtQ1 (cBtQ1), *Cardinium* cEper1 (cEper1), *A. asiaticus* (AmAs).

a later deletion event, removing the complete *bioB* gene and almost the complete sequence of the adjacent *bioF* gene (92.5 % identical to cEper1 in the remnant segment). Another recent signal of the loss of a nutritional contribution is the pyridoxal-dependent enzyme cystathionine gamma-lyase (involved in the synthesis of cysteine) whose CDS contains an internal stop codon mutation that produces the pseudogene *CHV_c0068* in cBtQ1 (94.9 % identical to cEper1 gene). A phylogenetic analysis showed that the functional gene, present in this state in cEper1, was acquired by an ancestor through HGT from a unicellular eukaryote, perhaps an amoeba (Annex Figure A.4.2.2). The phylogenetic analysis, including the three in silico identified *Leishmania major*'s cystathionine metabolizing enzymes (Williams *et al.*, 2009), showed its closer relation with *L. major* cystathionine gamma-lyase rather than with *L. major cystathionine* beta-lyase, as previously annotated in *Cardinium* cEper1 (Penz *et al.*, 2012).

In *Cardinium* cBtQ1, the inability to synthetize pyridoxal and biotin suggests that these vitamins are obtained from the host. In the case of *B. tabaci* strain QHC-VLC, these vitamins are synthesized by *Hamiltonella* BT-QVLC and exported to the host cytosol (described in Section 4.1.5). Due to the reduced metabolic capabilities of *Cardinium* cBtQ1, the possible pairwise competence between the endosymbiotic community in *B. tabaci* QHC-VLC was checked with **NetCmpt** (Table 4.2.2).

Table 4.2.2 *B. tabaci* QHC-VLC endosymbionts' pairwise EMO scores.

	Cardinium cBtQ1	*Portiera* BT-QVLC	*Hamiltonella* BT-QVLC
Cardinium cBtQ1	-	0	0.66
Portiera BT-QVLC	0	-	0.2
Hamiltonella BT-QVLC	0.09	0.04	-

As mentioned in Section 3.7.4, **NetCmpt** reconstruct the metabolic environments for each endosymbiont and returns the Effective Metabolic Overlap (EMO) index, that reflects the level of competition between a pair of species (Kreimer *et al.*, 2012). *Portiera* and *Hamiltonella* seem unaffected by the presence of the other endosymbiont and present a normal EMO score for a P-endosymbiont[57] and a S-endosymbiont, respectively (Kreimer *et al.*, 2012). The 0.2 EMO score for the pairwise *Portiera-Hamiltonella* points that some of the metabolites required by *Portiera* are also used by *Hamiltonella* without producing a strong competence. In contrast, the high EMO score for *Cardinium* cBtQ1 against *Hamiltonella* BT-QVLC (0.66) confirms that *Cardinium* cBtQ1 needs to compete for the environmental metabolites (from the host) with *Hamiltonella* (Table 4.2.2).

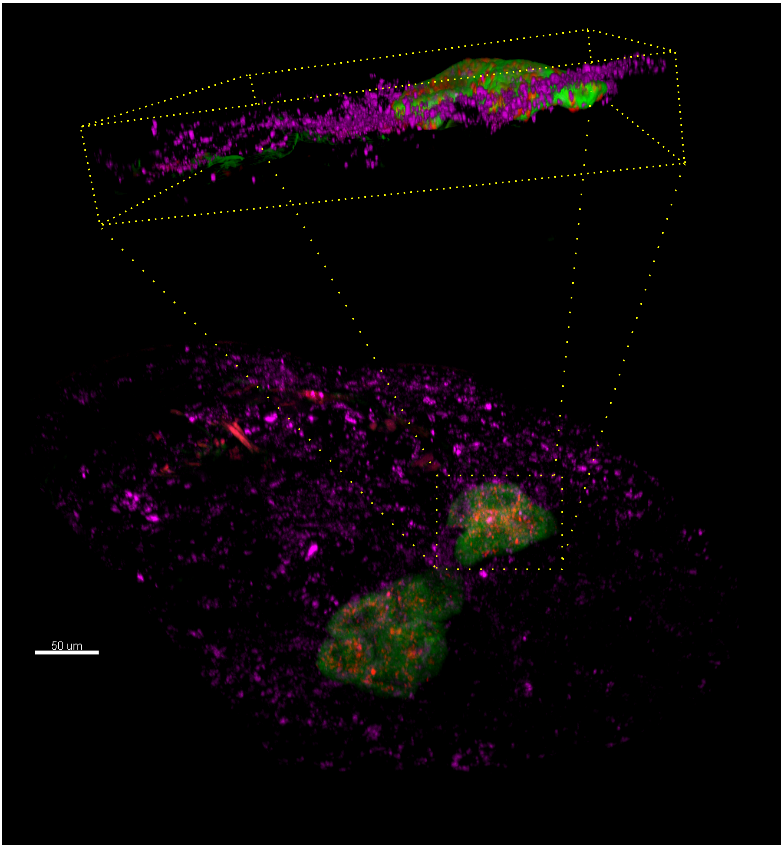
---

[57] Usually P-endosymbionts present a higher EMO score without competing with other endosymbionts

The loss of the biotin and pyridoxal pathways and the cysthationine gamma-lyase, are a clue that points to an accommodation of *Cardinium* cBtQ1 to a new environment where these metabolites or activities are supplied by other endosymbionts and the host, and available for *Cardinium* cBtQ1 from the host environment. These results lead to the hypothesis that, if *Cardinium* cBtQ1 has beneficial effects towards the host, they are not involved in nutrition. Moreover, it could be possible that the scattered phenotype (although other functions cannot be discarded) is a response to avoid the competition with *Hamiltonella* BT-QVLC for the resources in the bacteriocyte.

4.2.6. Gliding genes in *Cardinium* cBtQ1

Cardinium cBtQ1 (like other *Cardinium* from a broad range of hosts) could present different distribution patterns (displayed in purple on Figure 4.2.12 and in blue on Annex Movie A.4.2.1). They seem to have the ability to move inside and outside of the bacteriome (or the ovaries) of their host, and spread along the body of the insect, invading different tissues and cells (Bigliardi *et al.*, 2006; Gottlieb *et al.*, 2008; Kitajima *et al.*, 2007; Kurtti *et al.*, 1996; Nakamura *et al.*, 2009). In contrast, *Cardinium* cEper1 (as well as other strains, such as the *Cardinium* endosymbiont of Culicoides) is restricted to the ovaries of its wasp host (Morag *et al.*, 2012; Zchori-Fein *et al.*, 2001, 2004). Different Bacteroidetes possess the ability to move by a gliding mechanism, which is related to the ability to degrade some components present in the environment like chitin and cellulose (Braun *et al.*, 2005; McBride, 2004; Spormann, 1999) and may be related with a predatory behaviour (Furusawa *et al.*, 2003). Several examples of gliding have been reported in species of the class Cytophagia where *C. hertigii* was included (McBride and Zhu, 2013; Xie *et al.*, 2007) and it is possible that the scattered pattern detected in *Cardinium* cBtQ1 could be caused by a similar mechanism .



Figures 4.2.12 Whole mount FISH of a *B. tabaci* nymph and a Z-stack was used for reconstruct a three-dimensional picture of the nymph (bottom) and a bacteriome (upper). *Cardinium* cBtQ1 presents two kind of distribution patterns: one scattered and another confined into the bacteriome. Three different probes were used: *Poritera* in FAM (green), *Hamiltonella* in Cy3 (Red) and *Cardinium* in Cy5 (purple). For probe description see Material and Methods 3.2.2

The genome of *Cardinium* cBtQ1 reveals that the gliding genes detected (*gldK*, *gldL*, *gldM* and *gldN*) seem to be crucial for this organism because they are located in the multicopy plasmid, which denotes the possibility of overexpression required at specific points of development. Because these genes were lost in *Cardinium* cEper1, *Cardinium* from three *Encarsia* species were PCR screened for these genes. Two populations of *E. pergandiella* (one showing cytoplasmic incompatibility and the other parthenogenesis) one of *E. hispida* (parthenogenetic), as well as one of *E. inaron* (without phenotype) were checked for the presence of the four gliding genes in *Cardinium* to ensure that the non-motile phenotype could be related to the absence of these genes. None of the *Encarsia* species gave a positive result, suggesting, with caution, the absence of these genes in these *Cardinium* populations (Annex Table A.4.2.6). Based on these results, it is feasible that the gliding genes could be the cause of the motile phenotype (widespread pattern in different host tissues) in *Cardinium* cBtQ1 and other strains.

On the basis of the ancestor reconstruction analysis, the four gliding genes were present in LCA4, LCA2, and LCA1 and were lost in *Cardinium* cEper1. Although LCA4 conserved full gliding machinery, *sprATE*[58] was lost in LCA2 possibly due to its accommodation to an intracellular environment. Because LCA1 conserved the *gldKLMN* operon, this suggests that *gldKLMN* was lost in the *Cardinium* cEper1 lineage. Also, as in the closest Bacteroidetes genomes, such as *A. asiaticus*, *M. tractuosa*, or *C. marinum*, these genes are located in the chromosome, and we can postulate that in *Cardinium* cBtQ1, they have been translocated to a multicopy plasmid conserving the operon order. This supports the importance of these genes for *Cardinium* cBtQ1 and suggests that they may explain why the strain is not confined to a single tissue in *B. tabaci* in opposition to *Cardinium* cEper1 that is restricted to the ovaries of*Encarsia* (Penz *et al.*, 2012; Zchori-Fein *et al.*, 2004). Moreover, the gene amplification in *Cardinium* cBtQ1 not only of the four gliding genes but also of *mreB* and of the T1SS (RTX system) cluster *rtxBDE/tolC* (Figure 4.2.6) suggests that they may play an important role in this organism, as genome reduction is an ongoing process in this strain. There are two possible hypotheses:

- Those genes are involved in gliding as in other genomes (McBride and Zhu, 2013).

---

[58] Explained later

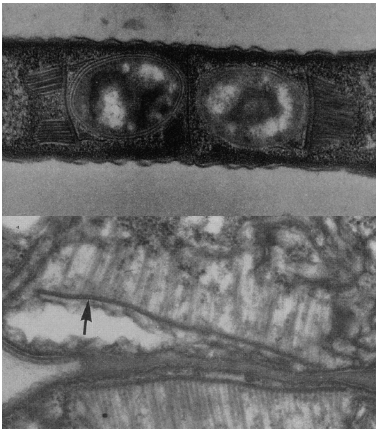
- They are involved in the novel Type 9 Secretion System (T9SS) (PorSS), which is also associated with the secretion of proteins involved in motility and toxins (McBride and Zhu, 2013; Sato *et al.*, 2010).

The first hypothesis considers that *Cardinium* cBtQ1 is able to glide and the *gldKLMN* operon is involved on this function. There are two main models for gliding proposed in Myxobacteria that have been shown some convergence in Bacteroidetes: the "slime" extrusion model and the motor based model (Braun *et al.*, 2005; Mauriello *et al.*, 2010b; Nakane *et al.*, 2013; Nan and Zusman, 2011; Spormann, 1999). The motor based model (or focal adhesion) is the most experimentally supported. It considers molecular motors that are associated with cytoskeletal filaments and use Proton Motive Force (PMF) to transmit force through the cell wall to attached dynamic focal adhesion complexes (adhesins) to the substrate, causing the cell to move forward (Mignot *et al.*, 2007; Sun *et al.*, 2011). The eukaryotic actin homolog MreB has been proposed as the cytoskeletal part of the gliding machinery (Kearns, 2007; Mauriello *et al.*, 2010a). Also, it is possible an association with FtsZ, a protein that is part of the bacterium cytoskeleton and can produce force by itself (Erickson *et al.*, 2010). Linkage between the cytoskeleton and gliding is supported by experimental data where the use of the compound A22, which is able to affect the MreB structure, inhibits the gliding motility in Myxobacteria (Mauriello *et al.*, 2010a; Nan and Zusman, 2011). In addition, although colchicine (a microtubule polymerization inhibitor) and derived drugs have little or no effects on bacterial FtsZ polymerization (Yu, 1998), the treatment with podophyllotoxin (other microtubule formation inhibitor not derived from colchicine) in *Saprospira* sp. suppress gliding motility (Furusawa *et al.*, 2003, 2005). The gliding proteins, detected in most Bacteroidetes, are the other part of the molecular machinery and 11 genes have been defined as essential (McBride and Zhu, 2013). Four of these genes (*gldB*, *gldD*, *gldH* and *gldJ*) have unknown function, while the remaining seven genes (*gldK*, *gldL*, *gldM*, *gldN*, *sprA*, *sprE* and *sprT*) encode the proposed PorSS system.

Cardinium cBtQ1 does not show the complete gliding machinery as it only contains four gliding core genes (*gldKLMN*) (McBride and Zhu, 2013). Neither homologous nor potential analogous genes of *gldBDHJ* have been detected. The *sprAET* genes are also absent, but their function would potentially be substituted by the cluster *rtxBDE/tolC*. RTX secretion system belongs to the T1SS and is able to transport proteins from the cytosol to the extracellular space in a SecYEG independent manner. Also, T1SS is able to secrete many different RTX family proteins and proteins without the C-terminal RTX nonapeptide (Kaur *et al.*, 2012; Linhartová *et al.*, 2010). The RTX system would secrete the adhesins (or other proteins that could interact with the host) across the bacterial membrane. Although no orthologs to known adhesin proteins were detected in *Cardinium* cBtQ1, its proteins with eukaryotic domains (such as ankyrins, TPR or WH2 domains) may function as adhesins in a multicellular eukaryotic organism. Moreover, *Cardinium* cBtQ1 could able to manipulate the host cytoskeleton to form a "scaffold", which could be used by the gliding machinery (Haglund *et al.*, 2010).

The second hypothesis would consider that the *gldKLMN* operon is not involved in gliding, but it is just required for secretion in the PorSS system, which was initially described for *Porphyromonas gingivalis* as a novel secretion system with eight proteins involved (PorK, PorL, PorM, PorN, PorT, PorW, Sov, and PorP) (Sato *et al.*, 2010). Putatively orthologous genes in the gliding system for the first seven are: *gldK*, *gldL*, *gldM*, *gldN*, *sprT*, *sprE*,and *sprA*. The proposed orthologous gene for *porP* in *Flavobacterium johnsoniae* was *Fjoh_3477*. A similar gene was not detected in either *A. asiaticus* or *Cardinium*. Proteins secreted by the PorSS systems are adhesins, as well as some enzymes such as chitinases, and gingipains in *F. johnsoniae* and *P. gingivalis*, respectively. Also, proteins secreted by the PorSS secretion system may contain a conserved C-terminal domain (TIGR4131 and 4183) (McBride and Zhu, 2013; Sato *et al.*, 2010). However, there were not proteins of *Cardinium* cBtQ1 with this domain. In the PorSS system, the presence of the protein complex GldKLMN is associated with the generation of the energy required for protein secretion by SprTEA. However, these proteins are not encoded in the genome of *Cardinium* cBtQ1 (they are also absent in *A. asiaticus*), and their substitution by the T1SS (RTX system) seems unlikely because T1SS has its own ATP-binding cassette, making the energy production function of GldKLM unnecessary. This suggests that the PorSS system does not work in *Cardinium* cBtQ1.

4.2.6.1. Gliding machinery organization



Figures 4.2.13 *Arthromitus* endospores with spore appendages by David G. Chase (upper). Plate filament array (arrow) in *Nostoc pruniforne*. Modified from Bermudes *et al.* (1994).

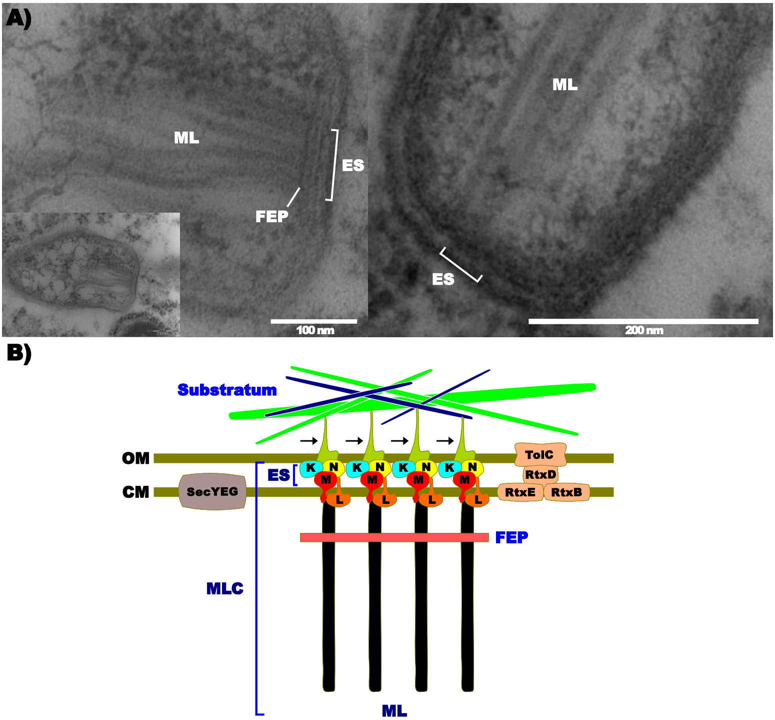
An important feature of all *Cardinium* strains is the presence of MLCs that were also described in different species of *Saprospira* (Bacteroidetes:Sphingobacteriia). MLCs-like structures are reported in other bacteria and it points to the idea that the proteins that form the MLCs need to be conserved (Figure 4.2.13). In consequence, they should be widespread among bacteria and probably involved in different cellular processes (Bermudes *et al.*, 1994; Bisalputra *et al.*, 1975; Burchard *et al.*, 1977). It seems that treatment with podophyllotoxin suppresses the gliding motility in *Saprospira*, but also the formation of MLCs structures (Furusawa *et al.*, 2003, 2005) linking these two processes in some manner. As suggested by Bigliardi *et al.* (2006), MLCs in *Cardinium* are divided in three components: the Microtubule-Like Structures (MLS), the Fibrous Electron-dense

Plaque (FEP) and the periplasmic Electron-dense Structure (ES) (Figure 4.2.14A). In contrast to *Cardinium* cEper1, ES in *Cardinium* cBtQ1 is visible in most of the images (Penz *et al.*, 2012; Zchori-Fein *et al.*, 2004)[59].

The genomic data from *Cardinium* cBtQ1 (duplicated genes and plasmid) and its differences with *Cardinium* cEper1 together with the Bacteroidetes proposed gliding machinery (McBride and Zhu, 2013; Sato *et al.*, 2010) and the previously proposed models for gliding (McBride and Zhu, 2013; Nakane *et al.*, 2013; Nan *et al.*, 2013), leads to speculate how the gliding motility machinery could be assembled in *Cardinium* cBtQ1. The proposed gliding apparatus is an adaptation of the previously proposed by Bigliardi *et al.* (2006). In this model, the tubulin homolog FtsZ, which can form straight tubules and can generate force by itself, and the actin homolog MreB (duplicated in *Cardinium* cBtQ1) which binds to cytoplasmic membrane, could be interacting in a complex that forms the ML and the FEP respectively (Erickson *et al.*, 2010; Fenton and Gerdes, 2013; Michie and Löwe, 2006; Salje *et al.*, 2011; Varma and Young, 2009) (Figure 4.2.14). It seems that both proteins are able to form a distribution pattern through the cell that can be congruent to the MLC distribution found in *Cardinium* (Chiu *et al.*, 2008; Thanedar and Margolin, 2004). It could be possible that ES is composed of a complex comprising the gliding proteins encoded by the genes *gldK*, *gldL*, *gldM* and *gldN*, which are present in an operon in *Cardinium* pcBtQ1's plasmid (Figure 4.2.14B). These proteins are part of the core proteins required for gliding in Bacteroidetes, such as *Flavobacterium johnsoniae*, as well as the PorSS secretion system detected in the non-motile pathogen *P. gingivalis* (Braun *et al.*, 2005; McBride and Zhu, 2013; Sato *et al.*, 2010; Shrivastava *et al.*, 2013). GldM and GldL are the only known gliding proteins that span the cytoplasmic membrane, and that have the required features to being the gliding motor (McBride

---

[59] Unfortunately, the TEM images were not as well defined as the images reported by Bigliardi *et al.* (2006), but ES were present in the periplasmic space between the inner and the outer membrane (Figure 4.2.14A)



Figures 4.2.14 A) Transmission Electron Microscopy images showing the Microtubule-Like Complex (MLC) of *Cardinium* cBtQ1. On the right image, Microtubule-like (ML) structures seem to span from the nucleoid of *Cardinium*, passing through the Fibrous Electron-dense Plaque (FEP) and contacting directly to the inner membrane. Electron-dense (ES) formations can be poorly detected between the inner and the outer membrane (both images). **B)** Schematic representation of the proposed gliding model for *Cardinium* cBtQ1. MLC has three components: the Microtubule-Like (ML), the Fibrous Electron-dense Plaque (FEP) and the Electron-dense structure (ES). ML and FEP could be composed of FtsZ and MreB, respectively. ES could contain the GldKLMN proteins. GldML proteins span across the inner membrane and are the putative molecular motor that can be in contact with the cytoskeletal part of the gliding machinery, the ML and FEP. GldKN are probably connecting the molecular motor to the outer membrane and the putative adhesins or secreted proteins (in green, marked with arrows) required for gliding. SecYEG can be the transporter of the GldKN proteins from the cytosol to the periplasmic space. RTX system (T1SS) would replace the SprAET system and would secrete the adhesins (or other proteins) across the bacterium membrane. The substratum is the putative target of adhesins in the host cell cytoskeleton or in the extracellular matrix.

and Zhu, 2013; Sato *et al.*, 2010; Sun *et al.*, 2011). These proteins have the same transmembrane domain profiles in *Cardinium* cBtQ1 as the ones observed in their orthologs from *F. johnsoniae* (Sonnhammer *et al.*, 1998). In addition, GldK (a lipoprotein) and GldN are located in the outer membrane (Sato *et al.*, 2010). *Cardinium* cBtQ1's GldN protein has a clear signal peptide (recognized by the SecYEG system), while the GldK protein possesses a putative site, but it is below the threshold level required to be considered as a signal peptide by SignalP4.1 (Petersen *et al.*, 2011). In contrast, in *F. johnsoniae* both proteins have a clear signal peptide that is required for their translocation to the periplasmic space. The conservation of transmembrane regions (that seems to be necessary for generating the PMF needed for gliding) and signal peptides (necessary for protein export to the periplasmic space where gliding motor is supposed to be assembled), suggest that these proteins are maintaining their functions in *Cardinium* cBtQ1.

Although *Cardinium* cBtQ1 has no orthologs for the rest of the proteins involved in gliding (such as ABC transporters), it has been suggested that they could be replaced by non-orthologous proteins (McBride and Zhu, 2013). The translocation of the GldK and GldN proteins to the periplasmic space requires the action of the GldAFG complex, an ABC transporter that translocates proteins from the cytoplasm to the periplasmic space. It may have been replaced by the functionally equivalent SecYEG transport system (Figure 4.2.14B). None of the other genes involved in the PorSS system (*sprT*, *sprE* and *sprA*) and the adhesin *sprB* were detected in *Cardinium* or *A. asiaticus*, although they were present in the LCA4. This suggests that a subsequent loss in these organisms probably occurred because these proteins lost their functions after the acquired intracellular lifestyle and the subsequent genome reduction process. All together points to the hypothesis that the RTX T1SS (duplicated in *Cardinium* cBtQ1), that is a more general secretion system, could be replacing the PorSS secretion system (Figure 4.2.14B), that seems to be a more specific secretion system, and *Cardinium* cBtQ1 is only retaining the necessary genes for the gliding movement (*gldKLMN*). It is possible that the MLS and the FEP conform the connection between the gliding machinery and the cytoplasm. As previously suggested, the FEP could be the area where the MLS assembly and insertion occurs (Bigliardi *et al.*, 2006). Moreover, MLS and FEP could have other primary cellular functions[60] but the gliding complex GldKLMN could be able to recruit, or stabilize, the MLS and the FEP for being used by the gliding machinery. GldLM was proposed as the gliding motor, but it is likewise possible that this motor is formed by the connection between the force generated by the MLS (FtsZ) and GldKN, which in turn, is supposed to be in contact with the RTX system secreted proteins that act like an anchor to the host cell cytoskeleton or to the extracellular environment (Jarrell and McBride, 2008; McBride *et al.*, 2009) (Figure 4.2.14B).

4.2.6.2. Rhapidosomes in *Cardinium*

Rhapidosomes[61] are rod shaped structures similar to defective phage tails that resemble microtubules (Bönemann *et al.*, 2010; Yamamoto, 1967). They are found in a wide range of bacterial lineages (including the Bacteroidetes) and seem to have diverged early and widespread through HGT events in Archaea. Rhapidosome proteins seem more related to the Type 6 Secretion System (T6SS) proteins than another phage-derived proteins (e.g. pyocins) and could have a similar function (Sarris *et al.*, 2014).

These structures were firstly described in *Sapropira* sp. by Delk and Dekker (1972) and renamed as SCFP by Furusawa *et al.* (2005). Recently, the *Saprospira grandis* genome was released and rhapidosomes presence was confirmed by genomic and proteomic approaches (Saw *et al.*, 2012).

Raphidosome proteins were named as Afp-like proteins in *Cardinium* cEper1 and were proposed as the components of the MLCs. In Penz *et al.*

---

[60] e.g. MreB is related to the insertion of pepditoglycan in the cell wall (Carballido-López, 2006) and this could be the reason why *Cardinium* cEper1 presents the MLCs without a scattered patter

[61] Also named as *Saprospira* cytoplasmic fibril proteins (SCFP), Afp-like proteins or more recently as Phage-Like-Protein-Translocation Structures (PLTS)

(2012), the MLCs were described as a T6SS based on the findings in the SCFP described by Furusawa *et al.* (2005) (Annex Table A.4.2.7). In this paper (Furusawa *et al.*, 2005), purified whole cell extracts of gliding *Saprospira* cells showed mainly structures that resemble rhapidosomes and were named as SCFP. Also, an over-expressed band of 61 KiloDalton (kDa) (SDS-PAGE gel) was detected in these extracts. This band was used to generate a polyclonal antibody against rhapidosomal proteins (or SCFP proteins) (Furusawa *et al.*, 2005). Whole cell extracts of induced gliding *Saprospira* cells (cultivated with 0.05 % polypeptone as amino acids source) were compared to non-gliding cells (0.5 % polypeptone) and subjected to immunoblot analyses, concluding that SCFP proteins are more abundant in gliding cells. Unfortunately, the polyclonal antibody seemed to have some unspecificity and more than one protein was detected (the previous band of 61 kDa and two more ranging from 20 to 30 kDa). It is interesting to notice tha some of the gliding proteins, FtsZ and MreB fall in this range and could be included in the detected bands.

From the methods presented in Furusawa *et al.* (2005), it is not possible to ensure that the higher amount of SCFP detected by immunoblot are only due to SCFP and it could be possible that other over-expressed proteins during the gliding state, like the above mentioned, were detected. In a more recent paper of the same group, expression of SCFP were the same in gliding as in non-gliding cells of *Saprospira* adding more incongruence to the previous results (Yoshikawa *et al.*, 2008)[62]. In the other hand, rhapidosomes of *Saprospira grandis* were analysed through proteomics approach using cell extracts of *Saprospira* cells cultivated under non-gliding conditions (0.5 % tryptone, a similar source of amino acids to polypeptone) indicating at least a great amount of rhapidosomes in non-gliding cells (Saw *et al.*, 2012).

If the aforementioned is taken into account, it seems that the association of SCFP to the MLC complex and gliding motility in *Saprospira* is not sufficiently congruent. A recent study of the T6SS using electron cryotomography (ECT) demonstrated a different, and less organized, structure than the one proposed for *Cardinium* cEper1 (Basler and Mekalanos, 2012). In the other hand, a structure of Afp-like proteins in *Pseudoalteromas luteoviolacea* that resembles to the MLCs from *Cardinium* triggers the metamorphosis of a marine tubeworm (Shikuma *et al.*, 2014). Without a clear experimental clue, is not possible to discard that the MLCs are composed by the rhapidosome proteins. Although it could be possible that these rhapidosomes are the MLS and FEP discussed above or that *Cardinium* could present a T6SS, the link between the MLC and the T6SS proposed for *Cardinum* cEper1 is not so clear Figure 4.2.14 (Penz *et al.*, 2012).

---

[62] The answer for this incongruence was that SCFP are regulated at translational level but not at transcriptional level

4.2.6.3. Possible gliding implications

Gliding seems to be a widespread direct invasion mechanism for different kinds of cells (Furusawa *et al.*, 2003; Sibley, 2004; Sibley *et al.*, 1998) and it is possible that *Cardinium* and other Bacteroidetes use this system to invade eukaryotic hosts. In fact, *Cardinium* endosymbiont of *I. scapularis* has been cultivated on insect cell lines, and is capable to invade new cells, even cell lines from different insect species, when they are added to the culture (Morimoto *et al.*, 2006; Nakamura *et al.*, 2011). It is possible that gliding could permit *Cardinium* to colonize new hosts and this could explain the horizontal transmission patterns detected in *Cardinium* from *B. tabaci* (Ahmed *et al.*, 2013). Eventually, new host-bacterial interactions would lead to the adaptation of the *Cardinium* strain to the new niche, sometimes with the loss of mobility and other genes (e.g. the Culicoides group). In *Cardinium* cBtQ1, the loss of the ability to synthetize biotin and pyridoxal seems to be related to the acquisition of these products from the host (including its other endosymbionts), which in *B. tabaci* QHC-VLC include *Portiera* and *Hamiltonella*. These losses restrict the new niches (hosts) to be invaded by *Cardinium* cBtQ1 to those with all these products available (e.g other whiteflies), reducing the putative targets for a new infection. This can explain the partial congruence shown between the phylogenies of several *B. tabaci* biotypes and the *Cardinium* strains that they harbour, where it seems that only some horizontal transmission has been occurred (Ahmed *et al.*, 2013).

In order to determine the presence of the gliding genes in *Cardinium* endosymbionts from wildtype *B. tabaci* populations, adult whiteflies were sampled in four different municipalities of Valencia (Spain). Twelve sample points were selected and four females per point were analysed (Annex Table A.4.2.8). The biotype was determined for each of them and, as expected for *B. tabaci* populations in Spain, most of them belonged to the biotype Q (Mediterranean species) but a few of them were from the biotype S (Sub-Saharan Africa species), an uncommon biotype in Spain (EMBL accession numbers HG421085-HG421096) (see Annex Figure A.4.2.3) . All analysed individuals harboured *Cardinium*, including biotype S (Annex Table A.4.2.8).

The *16S* rRNA genes from biotype S (sample F) and biotype Q (sample B) (1100 bp) were sequenced and resulted 100 % identical to *Cardinium* cBtQ1 (EMBL accession numbers HG421077-HG421084). The presence of gliding and the *CHV_p021* genes was also confirmed by PCR in all these individuals (Annex Table A.4.2.8). The biotype S individuals also harboured *Arsenophonus*, a symbiont that share the same distribution pattern of *Hamiltonella* and are fixed in some populations. Also, only a few cases of *Cardinium* and *Arsenophonus* sharing the same host have been reported suggesting a recent infection of biotype S by *Cardinium* cBtQ1 (Gnankiné *et al.*, 2013; Singh *et al.*, 2012; Thierry *et al.*, 2011; Zchori-Fein *et al.*, 2014) and the idea that this endosymbiont is still able to invade new similar hosts.

The maintenance of secondary endosymbiont infections is considered a trade-off between the costs of harbouring the symbiont, and the putative beneficial effects produced by it, *Cardinium* cBtQ1 could be able to confer some beneficial effects (Feldhaar, 2011; Ferrari and Vavre, 2011; Oliver *et al.*, 2008) to its host *B. tabaci* because of its abundant presence in the Valencia province and in the Q1 biotype in general (Zchori-Fein *et al.*, 2014). These effects could be related to its mobility feature, and the presence of some putative toxin-related genes in the plasmid like *CHV_p018* (low e-value blast hit againts RTX toxins of *Hamiltonella defensa* from *A. pisum*) and *CHV_p021*. *CHV_p021* could have a role in intercellular competition, intercellular signaling and insecticidal activity based on the presence of the RHS domain (Koskiniemi *et al.*, 2013). As a mobile endosymbiont, *Cardinium* cBtQ1 may contact directly the parasitoid and secrete insecticidal toxins near the parasitoid, could invade the parasitoid tissue and kill it by an unknown process, or by the cytotoxic effect of lipid A in a non-acclimated host (Ferrari and Vavre, 2011; Furusawa *et al.*, 2003; Hansen *et al.*, 2012; Oliver *et al.*, 2008; Rader *et al.*, 2012). Some other possibilities takes into account that MLCs can be related to a physically protective role as in some endosymbionts (Petroni *et al.*, 2000; Preer *et al.*, 1974). Furthermore, other effects that could increase the fitness of the host cannot be excluded, like heat-stress resistance or maybe some advantages that are given by the lipoate supplementation (Moran *et al.*, 2008; Moya *et al.*, 2008) but it is possible that fitness increase is only produced in some environments or climate conditions.

Recently, a beneficial effect of secreted Afp-like proteins (rhapidosomes) from *Pseudoalteromonas luteoviolacea* has been reported in *Hydroides elegans* (a marine tubeworm), where they seem to be beneficial for the latter. In this context, an advantageous effect of rhapidosomes from *Cardinium* can not be excluded (Shikuma *et al.*, 2014). However, *Cardinium* cBtQ1 could be only a manipulative endosymbiont and not confer any fitness advantage, although reproduction manipulation by *Cardinium* is rarely documented in *B. tabaci* (Stansly and Mckenzie, 2007; Thierry *et al.*, 2011).

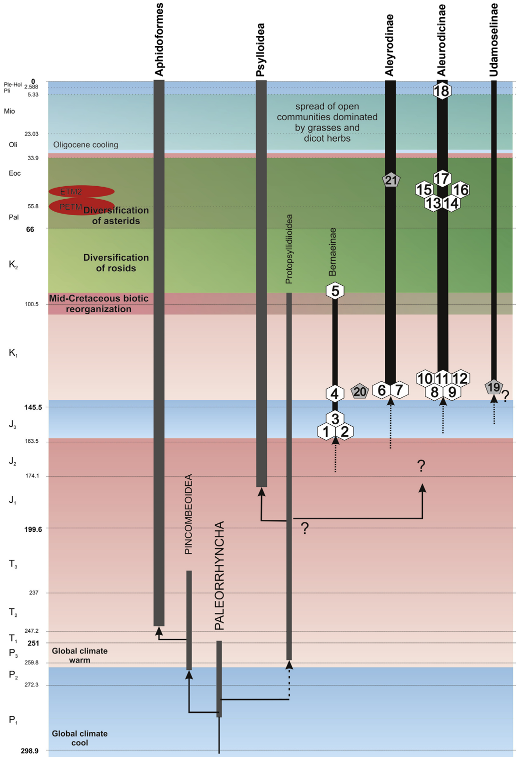
Genome evolution of the genus *Portiera*

4.3.1. Background

Whiteflies are proposed to be divided in four subfamilies, three extant ones and one extinct. The oldest whitefly subfamily, the extinct Bernaeinae, was the first subfamily to diverge and can be traced until the Upper Jurassic (fossils 1 and 2 from Figure 4.3.1) (Byrne and Bellows, 1991; Campbell *et al.*, 1994; Drohojowska and Szwedo, 2014; Shcherbakov, 2000). Although three extant subfamilies are proposed, the relationship of the Udamoseliane subfamily with the Aleurodicinae and Aleyrodinae families is still under discussion. Some authors proposed that Udamoseliane is closer to, and should be included in, the Aleurodicinae (Martin, 2007; Martin and Mound, 2007; Shcherbakov, 2000).

Fossils that can be assigned to the actual families are from the end of the Upper Jurassic and the beginning of the Lower Cretaceous (Drohojowska and Szwedo, 2014). The two oldest fossils from the subfamily Aleyrodinae, *Heidea cretacica* and *Baetylus kahramanus* (fossils 6 and 7 from Figure 4.3.1), and from the subfamily Aleurodicinae, *Gapenus rhinariatus* and *Aretsaya therina* (fossils 8 and 9 from Figure 4.3.1), were found in the Lebanese amber(125-135 Myr old) from the Lower Cretaceous (Drohojowska and Szwedo, 2011, 2013, 2014; Schlee, 1970).

The first work on whiteflies molecular dating was done by Campbell *et al.* (1994) and dated the divergence of Aleurodicinae and Aleyrodinae around 92 million years ago (mya). This result was approximately in agreement with the fossil data reported in Schlee (1970) and the paleotropical origin of whiteflies. Recently, another molecular dating work tried to date the divergence of the *Bemisia* genus and the *B. tabaci* complex (Boykin *et al.*, 2013). However, the calibration point was incorrectly set up with *'Aleurodicus' burmiticus*[63].



Figures 4.3.1 Whiteflies diversification and their stratigraphic distribution. Number denotes the assigned taxon to the fossil. Only taxa of interest for molecular dating are maintained (see): 1: *Juleyrodes visnyai* - Upper Jurassic, 2: *Juleyrodes sp.* - Upper Jurassic, 6: *Heidea cretacica* - Lower Cretaceous, 7: *Baetylus kahramanus* - Lower Cretaceous, 8: *Gapenus rhinariatus* - Lower Cretaceous, 9: *Aretsaya therina* - Lower Cretaceous, 21: *'Aleurodicus' burmiticus* - earliest Upper Cretaceous (Cenomanian). Other abbreviations: PETM - Palaeocene-Eocene Thermal maximum, ETM2 - Middle Eocene Climatic Optimum. Reproduced from Drohojowska and Szwedo (2014).

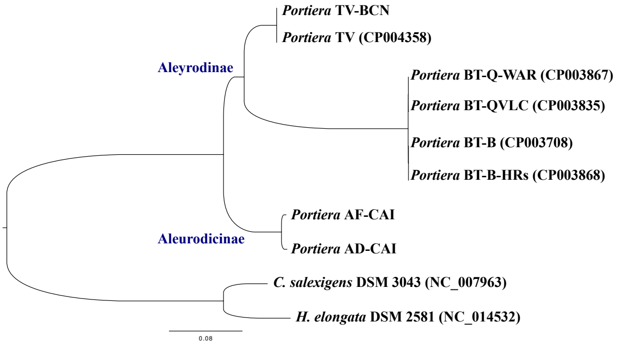
Up to date, five *Portiera* genomes from the subfamily Aleyrodinae have been sequenced. Four belongs to the *B. tabaci* complex, two from

---

[63] This fossil is younger than the values used in Boykin *et al.* (2013) with an estimated age of ca. 100 mya (Shi *et al.*, 2012)

Q biotype (MED) and two from B biotype (MEAM1) (Figure 4.3.2) (Jiang *et al.*, 2013; Santos-Garcia *et al.*, 2012; Sloan and Moran, 2012a). The remaining one is a strain from *Trialeurodes vaporariorum*, which highlighted that *Portiera* from *B. tabaci* has suffered a high number of rearrangements (Sloan and Moran, 2013). All these *Portiera* strains supply their hosts with some essential amino acids and carotenes (Santos-Garcia *et al.*, 2012; Sloan and Moran, 2012a).

In the present work, three additional *Portiera* strains are reported, one from a *Trialeurodes vaporariorum* host (Aleyrodinae) an named as TV-BCN, and two from the *Aleurodicus* genus (Aleurodicinae) (see Figure 4.3.2 and Annex Figures A4.3.1 and A4.3.2 for host phylogenies). From *Aleurodicus* genus, *Aleurodicus dispersus* and *Aleurodicus floccissimus* (formerly *Lecanoideus*) species were the ones selected and their *Portiera* strains were named as AD-CAI and AF-CAI, respectively.


Portiera TV-BCN
Portiera TV (CP004358)
Aleyrodinae
Portiera BT-Q-WAR (CP003867)
Portiera BT-QVLC (CP003835)
Portiera BT-B (CP003708)
Portiera BT-B-HRs (CP003868)
Portiera AF-CAI
Aleurodicinae
Portiera AD-CAI
C. salexigens DSM 3043 (NC_007963)
H. elongata DSM 2581 (NC_014532)
0.08

Figures 4.3.2 ML tree for a concatenated protein alignment (GroL, RpoB, RpoC, GyrA, GyrB, and DnaE summing up 5522 amino acids selected positions) of all the *Portiera* strains sequenced at the moment of writing this work. Whiteflies subfamilies are displayed in blue. ML tree was run under the cpREV with gamma distribution and empirical base frequencies model. All bootstrap values were 100. *C. salexigens* and *H. elongata* were used as outgroup.

These new genomes allowed to study the evolution of *Portiera* in the two subfamilies but also to test the possibility to use them to unravel the divergence history of whiteflies.

4.3.2. Genomic features of *Portiera* strains

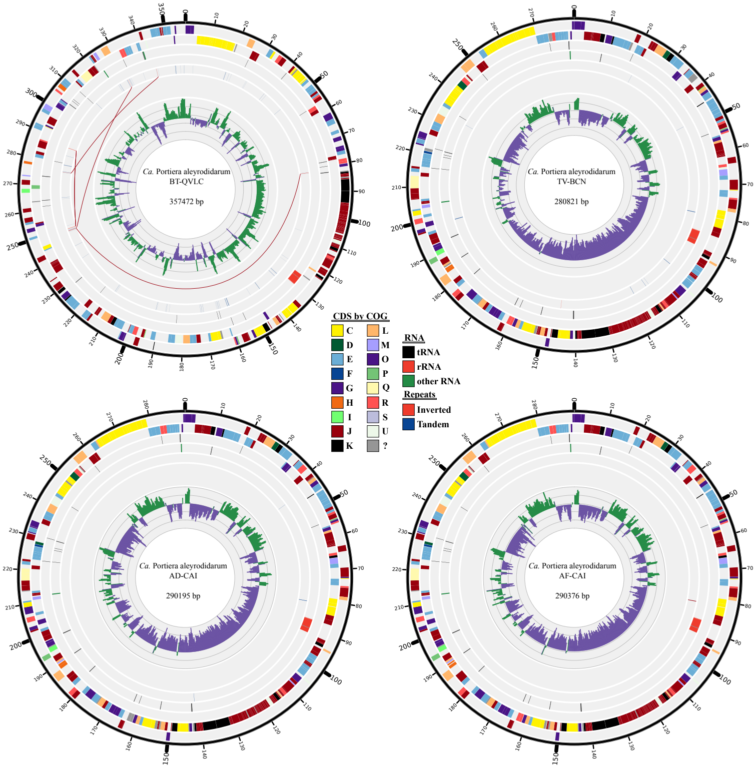
The genomes of *Portiera* strains TV-BCN, AD-CAI and AF-CAI were composed of a single circular closed contig with an approximate average coverage for each genome of 90X and 1500X for 454 and Illumina libraries, respectively. The general genomic features of the three new *Portiera* strains (TV-BCN, AD-CAI and AF-CAI) were partially similar to those of the previously sequenced *Portiera* genomes and to their sister lineage *Carsonella*. They have an extremely reduced genome (between 280 and 290 kb) with a low GC content and a high coding density without the large IGRs showed in *Portiera* from *B. tabaci* (only *Portiera* BT-QVLC is shown in Table 4.3.1 and Figure 4.3.3) (Santos-Garcia *et al.*, 2012; Sloan and Moran, 2012a, 2013).

Table 4.3.1 General Genomic Features of *Portiera* strains and *Carsonella* HC.

Symbiont	***Carsonella* HC**	***Portiera* TV**	***Portiera* TV-BCN**[b]	***Portiera* AD-CAI**[b]	***Portiera* AF-CAI**[b]	***Portiera* BT-QVLC**[a]
Host	*H. cubana*	*T. vaporariorum*	*T. vaporariorum*	*A. dispersus*	*A. floccissimus*	*B. tabaci*
Size (bp)	166,163	280,663	280,822	290,195	290,376	357,472
GC %	14	25	25	24	24	26
Genes	223	307	307	317	317	285
CDS	192	269	268	278	278	247
CDS %	98	94	94	95	95	68
rRNA	3	3	3	3	3	3
tRNA	28	34	34	34	34	33
Other RNA	0	1	2	2	2	2
Pseudo	0	0	1	1	0	7

[a] Re-annotated for this work [b] This work

The three new *Portiera* strains contain 39 non-coding RNA genes, which specify 34 tRNAs able to decode all mRNAs, the three rRNAs (16, 23 and 5S), one tmRNA and the RNA subunit of RNase P (*rnpB*).The differences in genome size between the three new genomes account for approximately 10 kb that correspond to the 10 CDS in which they differ. While the three new genomes maintained a clear GC skew pattern, it was not appreciable in none of the *Portiera* strains from *B. tabaci* (all strains sequenced from *B. tabaci* are virtually identical, see Section *Portiera* strains from *B. tabaci*). The loss of this GC skew pattern in *Portiera* from *B. tabaci* Q and B biotypes (MED and MEAM1, respectively), and the large IGRs detected in a DNA fragment from a *Portiera* from the New World 1 (AY268081) *B. tabaci* species (Baumann *et al.*, 2004), is an indication that this lineage has suffered recent rearrangements (in evolutionary terms), at least since its divergence from *Trialeurodes* (Table 4.3.1 and Figure 4.3.3) (Sloan and Moran, 2013).



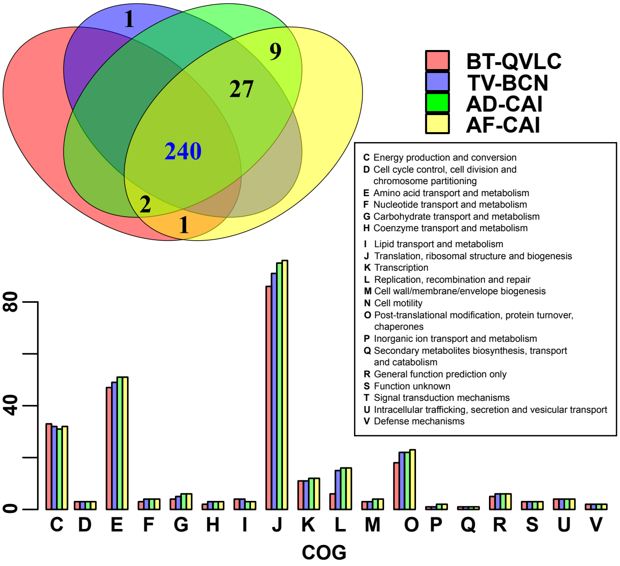
Figures 4.3.3 Genome overview of *Portiera* strains BT-QVLC, TV-BCN, AD-CAI and AF-CAI. From inner to outer tracks: (I) Positive (green) and negative (purple) GC skew across the genome. (II) Inverted repeats (red lines and links) and Tandem repeats (blue). (III) Complementary strand noncoding RNAgenes: rRNAgenes (red), transferRNA genes (black), other RNA genes (green). (IV) Direct strand noncoding RNA genes: rRNAgenes (red), transferRNAgenes (black), other RNAgenes (green). (V) Complementary strand CDS. (VI) Direct strand CDS. CDS were coloured according to their **COG** classification.

Although all *Portiera* strains have tandem repeats, it seems that they were accumulated in the Aleyrodinae subfamily, mainly in the *Bemisia* branch. The largest number of tandem repeats is accumulated in *Portiera* strains of *B. tabaci* BT-QVLC (112 tandems repeats) and BT-B (Sloan and Moran, 2013). In contrast, the *Portiera* strains of *T. vaporariorum* TV-BCN (10 tandem repeats) and TV (Sloan and Moran, 2013) presented a small amount of them. Finally, in the Aleurodicinae branch only a few tandem repeats were detected (AD-CAI presented 3 tandem repeats and AF-CAI only 1) (Figure 4.3.3, see phylogeny in Figure 4.3.2). Sloan and Moran (2013) proposed that the increase of these repeats in *Portiera* from *B. tabaci* seems an effect of the loss of essential genes from the replication and repair machinery (e.g. *dnaQ*) and that the repeats could explain the rearrangements and the large IGRs found in these genomes. Their proposal was that dispersed repeats are hotspots for recombinations while tandem repeats are a source of replication errors/DNA breaks.

Finally, compared to the recent published *Portiera* TV from *T. vaporariorum* (Sloan and Moran, 2013), *Portiera* TV-BCN is almost identical. The differences in the genome size increment could be due to assembly algorithms (collapsed or miss-assembled repeats), while the differences in the genome annotation are due to the pseudogene *miaA* and the *tmRNA* present in both strains, but only annotated in TV-BCN. It is interesting to notice that although TV is from North America (New Haven, Connecticut) and TV-BCN from Europe (Catalonia, Spain), both strains are close to 100 % identical at nucleotide level. This lack of nucleotide variation suggests that both strains are from the same population and were introduced in the different countries by human plant trade.

4.3.3. Comparative genomics and genome stasis in the genus *Portiera*

CDS from the three *Portiera* strains sequenced in this work, plus the BT-QVLC strain, were used to infer the pangenome and the core genome of the *Portiera* genus. While the core genome was composed of 240 clusters of orthologous CDS, the pangenome was only 40 clusters more than the core (Figure 4.3.4).



Figures 4.3.4 On the left an Euler diagram is displayed with each colour corresponding to a *Portiera* strain. The number of genes of the core genome is highlighted in blue. On the right, a bar plot represents the number of **COG** hits for each *Portiera* strain (in the same colours).

Most of these differences are due to the inclusion of *Portiera* BT-QVLC that lacks 37 CDS compared to the other strains, and thus decreasing the CDS in the pangenome. If *Portiera* BT-QVLC would be not included, the pangenome and the core would be mostly the same, with only 12 strain specific CDS: *lepB* in the *Portiera* TV-BCN, *ahpC* that is shared by AF-CAI and BT-QVLC and 11 shared by AD-CAI and AF-CAI (two of them shared also with BT-QVLC). This suggests that the LCA of all the *Portiera* strains possessed already an extreme reduced genome with 280 CDS, considering *alaS* (explained later in detail) one single gene and the ortholog of pseudogene *PAQ_201*, present in all the *B. tabaci* strains, as an active gene (279 pangenome CDS plus the pseudogene

PAQ_201). A few gene losses took place in most of the lineages, except in that of *Bemisia* where the important number of gene losses produced an unstable genome, as can be deduced also from the absence of GC Skew. Accordingly to the Euler Diagram (Figure 4.3.4), *Portiera* AF-CAI could be considered, in gene content terms, the closest to the LCA, because it includes all orthologous CDS clusters with the exception of *lepB* and *PAQ_201*. It is important to notice the initial impossibility to annotate some *Portiera* BT-QVLC genes (e.g. *rnpA*, *gatC*, etc.). This is an effect of the accelerated evolution shown by *Portiera* from *B. tabaci* that did not allow the recognition of some ORFs by homology analysis against free-living bacterial species or other endosymbionts. Only the new strains, that evolve at a lower rate and still maintain some homology to free-living relatives allowed to detect and annotate correctly these genes. This effect is also a problem detected in *Carsonella* strains where a large amount of genes remains as hypothetical proteins (Tamames *et al.*, 2007).

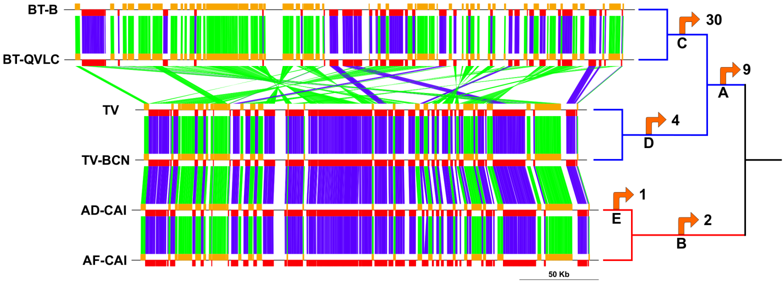
When CDS were assigned to a **COG** category, all strains with the exception of *Portiera* BT-QVLC shared a similar profile (see Figure 4.3.4 for **COG** category description). Also, as explained above, the inference of the pangenome and core genome was highly impacted by *Portiera* BT-QVLC as it can be seen in **COG** distribution, where the core genome bar was almost identical to the former (Figure 4.3.4). Although some **COG** categories presented small differences between strains[64] a similar metabolism should be expected for all four strains.

The majority of gene losses have been produced in the *B. tabaci* strains (branch C Figure 4.3.5 and Table 4.3.2). These losses included, among others, a great number of genes involved in DNA replication and repair machinery (8), the transcription/translation machinery (3), some genes from the amino acid biosynthetic pathways, the chaperone ClpB, and the almost universal protein recycling system ClpXP (explained in Section Comparative genomics).

---

[64] *Portiera* AF-CAI and AD-CAI showed the higher amount of hits on **G**, **E**, **K**, and **L** **COG** categories while in **J** and **O** was AF-CAI

The Aleyrodinae lineage (branch A) also accumulates an important number of gene losses (nine), most of them from the transcription/translation machinery and the *tktA* (the link between the glycolysis and the pentose phosphate pathway). In TV and TV-BCN strains (branch D) only four genes were lost, while in the Aleurodicinae lineage (branch B) only three were lost, two in their common LCA and one in AD-CAI (Table 4.3.2 and Figure 4.3.5).



Figures 4.3.5 Genomic synteny in *Portiera* strains sequenced, denoting the rearrangements produced in the *B. tabaci* lineage. Orange boxes represent genes in the direct strand, red boxes genes in the complementary strand, green lines connect genes with at least one of them in the direct strand while blue lines connect genes when both are in the complementary strand. The cladogram on the right represents the different host lineages (Aleyrodinae in blue and Aleurodicinae in red) and the gene losses in each branch represented by a letter (listed in Table 4.3.2).

Table 4.3.2 Gene losses during *Portiera* evolution.

	Branch				
	A	B	C	D	E
Gene losses	*miaA*^, *rnc*^, *rpmD*^, *glyA*⋆, *alaS*^, *hupB*^, *tktA*, *metG*^, *yqgF*^	*lepB*†, *PAQ_201*	*dnaQ**, *dnaX**, *dnaN**, *holA**, *holB**, *ruvC**, *ssb**, *mutL**, *upp*, *clpP*†, *clpX*†, *clpB*†, *lspA*†, *sohB*†, *lepB*†, *mucD*, *dapB*⋆, *lysA*⋆, *argH*⋆, *dapF*⋆, *trpS*^, *rsmA*^, *frr*^, *deaD*^, *tRNA-Ala*^, *era*, *lipB*, *galP*, *PAQ_201*	*hisE*⋆, *ahpC*, *rplA*^, *PAQ_201*	*ahpC*

*Replication, recombination and repair ^Transcription, translation and ribosome biogenesis

†Post-translational modification, protein turnover, and chaperones ⋆Amino acid biosynthesis

Genome rearrangement analysis using 235 genes shared between the two lineages of *Portiera* (Aleyrodinae and Aleurodicinae) showed that *Portiera* strains BT-B and BT-QVLC (*B. tabaci*) have accumulated all the 19 rearrangements needed to explain the actual genome architecture of these strains. In contrast *Portiera* TV and TV-BCN (*T. vaporariorum*), AF-CAI (*A. floccissimus*) and *Portiera* AD-CAI (*A. dispersus*) showed no rearrangements. When orthologous CDS clusters were plotted, the singular evolution of *Portiera* strains from *B. tabaci* (large IGRs and rearrangements) is clearly in contrast to the genome stasis in the other *Portiera* strains (Figure 4.3.5). A possible scenario for the special genome shape of the *Portiera* strains from *B. tabaci*, which undergoes an increase in genome size in contrast to the rest of P-endosymbionts that suffers a progressive genome reduction process, could be deduced taking into account that:

- Genome rearrangements usually occur early and in a short period of time during the genome reduction process (Belda *et al.*, 2005; Latorre *et al.*, 2005).
- Genome stability in reduced genomes seems a combination of recombination/repair gene and repetitive elements losses (Silva *et al.*, 2003; Tamas *et al.*, 2002).
- Illegitimate recombination is only dependent on repetitive sequences and does not require the recombination machinery[65] (Rocha and Danchin, 2002).
- Rearrangements, gene losses and repeat elements are accumulated in *Portiera* strains from *B. tabaci*.
- Few repeat elements were also detected in the *Portiera* strains TV-BCN, AD-CAI, and AF-CAI that present a more complete replication/repair machinery.

In this scenario, the loss of replication and repair genes in the *Portiera* from *B. tabaci* lineage produced the expansion of tandem repeats from the few dispersed repeat elements present in the *Portiea* LCA by mechanism like polymerase slippage, DNA breakage and linkage, recombination, etc. (Rocha and Danchin, 2002). Because tandem repeats have a high illegitimate recombination index, rearrangement events also increased producing more copies of the tandem repeats and triggering a loop where the tandem repeats were amplified and dispersed through the genome (Rocha and Danchin, 2002). Because IGRs can tolerate more rearrangements than coding regions, these tandem repeats (and also inverted repeats) were accumulated on them. The large IGRs present in *Portiera* strains from *B. tabaci* and their increase in genome size are a result of this loop[66]. A similar scenario was also proposed by Sloan and Moran (2013).

---

[65] In fact, the absence of a recombination and repair machinery can trigger these kind of recombination events

However, comparisons against other P-endosymbionts suggest that not only the loss of replication and repair genes are the responsible for the genome instability of *Portiera* from *B. tabaci*. *Uzinura*, that also presents the same extremely reduced polymerase machinery and few repetitive regions (11 tandem repeats and three inverted repeats), does not present the large IGRs and the low coding density of *Portiera* from *B. tabaci*. Although no more *Uzinura* genomes are available to understand the evolution of genome size in this P-endosymbiont, it seems that *Uzinura* is not suffering the same genomic instability as *Portiera* from *B. tabaci*, arisen some doubts to the central role of the reduced polymerase (the loss of *dnaQ*) in this process. In addition, *Tremblaya* presents large IGRs with a more complete polymerase (but still very basic) but it presents the usual genome reduction process[67]. It has been proposed that IGRs in *Tremblaya* are due to a process of recombination and pseudogenization (López-Madrigal *et al.*, 2013), pointing to the idea that the process behind its large IGRs are different than *Portiera* from *B. tabaci*. In conclusion, genomic instability in *Portiera* from *B. tabaci* is a complex scenario where loss of *dnaQ* and illegal recombination are only the tip of the iceberg.

---

[66] Maybe by the addition of intergenic region or gene fragments that suffered gene erosion and are no longer recognisable (Silva *et al.*, 2001) due to the rearrangement events, or by the accumulation of replication errors

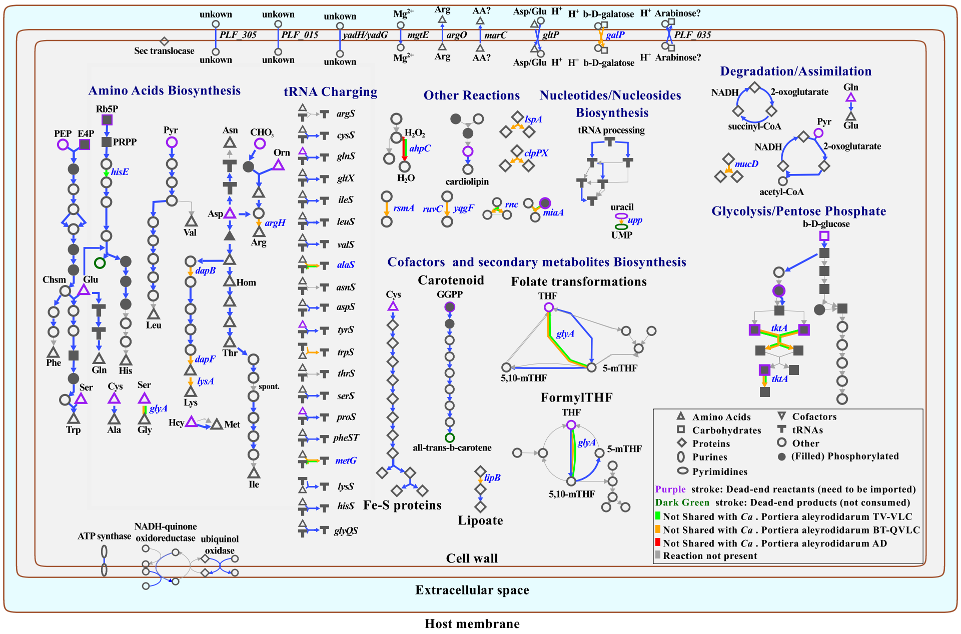
[67] *T. phenacola* genome size is ca. 0.17 Mb while *T. princeps* is ca. 0.14 Mb

4.3.4. Metabolic blueprints of *Portiera* strains

Although all *Portiera* strains share most of the metabolic reactions, different gene losses in the different strains have impacted their ability to synthesize amino acids, some cofactors and other reactions (Figure 4.3.6). *Portiera* AF-CAI, which had the most complete metabolism, was used as a reference for comparing the metabolism of the different *Portiera* strains (blue lines/arrows in Figure 4.3.6).

All the strains can produce carotenes and the Fe-S cluster proteins, decarboxilate the pyruvate for producing some intermediate metabolites and reducing power (nicotinamide adenine dinucleotide (NADH)), maintain most of the aerobic electronic transporter chain (*nuo* operon and ubiquinol oxidase) and the ATP synthase. In contrast, BT-QVLC and TV-BCN strains have lost *tktA*, one of the last remaining genes from the pentose phosphate pathway. Also, they have lost the ability to synthesize glycine and some folate transformations (*glyA*). This progressive loss of ability to synthesize intermediate metabolites and cofactors points to a still active genome reduction process and an increase in the dependency of the host environment. As an example, even that all *Portiera* strains encode the ubiquinol oxidase, they need to import ubiquinol (from the host or from an S-endosymbiont, explained in Section Metabolic integration); or the case of NADH, where they can be reduced/oxidized but it needs to be imported from the host. Another option is that *Portiera* strains, can get the lacking molecules from the S-endosymbionts that usually share the same bacteriocytes, as it occurs in the case of *Buchnera*/*S. symbiotica* consortium (Lamelas *et al.*, 2011b; Manzano-Marín and Latorre, 2014) and is the case of the tandem *Portiera*/*Hamiltonella* BT-QVLC (explained in Section Metabolic integration).

Portiera LCA genome encodes many enzymes involved in amino acid biosynthesis. *Portiera* strains AD-CAI and AF-CAI have retained all these


unkown
PLF_305
unkown
PLF_015
yadH/yadG
Mg2+
mgtE
Arg
argO
AA?
marC
Asp/Glu
H+
gltP
b-D-galatose
galP
Arabinose?
PLF_035
Sec translocase
Amino Acids Biosynthesis
tRNA Charging
Other Reactions
Nucleotides/Nucleosides Biosynthesis
Degradation/Assimilation
Rb5P
PEP
E4P
PRPP
hisE
Pyr
Asn
CHO3
Orn
Asp
argH
Arg
Val
Leu
Chsm
Glu
dapB
Hom
Phe
Gln
His
dapF
lysA
Lys
Thr
spont.
Ser
Cys
Ser
glyA
Trp
Ala
Gly
Hcy
Met
Ile
argS
cysS
glnS
gltX
ileS
leuS
valS
alaS
asnS
aspS
tyrS
trpS
thrS
serS
proS
pheST
metG
lysS
hisS
glyQS
H2O2
ahpC
H2O
cardiolipin
lspA
clpPX
rsmA
ruvC
yqgF
rnc
miaA
tRNA processing
uracil
upp
UMP
NADH
2-oxoglutarate
succinyl-CoA
Gln
Glu
Pyr
mucD
acetyl-CoA
Glycolysis/Pentose Phosphate
b-D-glucose
tktA
Cofactors and secondary metabolites Biosynthesis
Carotenoid
GGPP
all-trans-b-carotene
Folate transformations
THF
glyA
5-mTHF
5,10-mTHF
FormylTHF
Fe-S proteins
lipB
Lipoate
ATP synthase
NADH-quinone oxidoreductase
ubiquinol oxidase
Cell wall
Extracellular space
Host membrane
Amino Acids
Carbohydrates
Proteins
Purines
Pyrimidines
Cofactors
tRNAs
Other
(Filled) Phosphorylated
Purple stroke: Dead-end reactants (need to be imported)
Dark Green stroke: Dead-end products (not consumed)
Not Shared with Ca . Portiera aleyrodidarum TV-VLC
Not Shared with Ca . Portiera aleyrodidarum BT-QVLC
Not Shared with Ca . Portiera aleyrodidarum AD
Reaction not present

Figures 4.3.6 Metabolic comparison of *Portiera* strains. Four strains were analysed in this work. Metabolism from *Portiea* AF-CAI was assumed as the basal one (shared between all strains) and are represented by blue lines and arrows. Arrow indicate the direction of the reaction. Gene losses from different strains are displayed in a representative colour. Gene names in blue denote pseudogenes at least in one strain. Abbreviations: Chorismate (Chsm) Geranylgeranyl diphosphate (GGPP), Tetrahydrofolate (THF), Pyruvate (Pyr), Phosphoenolpyruvate (PEP), 5-phospho-a-D-ribose-1-diphosphate (PRPP), D-erythrose 4-phosphate (E4P), D-ribulose-5-phosphate (Rb5P), Homocysteine (Hcy).

enzymes. They encode complete biosynthetic pathways for Lys, Arg, Thr, and Trp. They also encode almost complete pathways for Phe, Ile, Leu and Val (the last step of these reactions can be complemented by the host) and for His (explained in detail in Section Metabolic integration). In addition, although they do not encode a Met complete pathway, they have retained *metE*, the gene encoding the last step of the pathway. The substrate of this reaction, Hcy, must be obtained from the host. They are also able to synthesize the non-essential amino acid glycine. As expected from the comparative genomic analysis, the most degraded metabolism was the one from *Portiera* BT-QVLC. It has lost the genes encoding three steps from the Lys (*dapBF* and *lysA*), one from the Arg (*argH*) biosynthetic pathway, and the ability to synthesize lipoate and UMP. These losses could be due to the presence of *Hamiltonella*, that has an almost complete Lys pathway and can produce lipoate and UMP. In contrast, *argH* seems to be complemented by the host (Sloan and Moran, 2012b; Sloan *et al.*, 2014; Xie *et al.*, 2012). This metabolic redundancy could have favoured a relaxation in purifying selection, that combined with the high substitution rate detected in this *Portiera* (see Section Rates of nucleotide substitution in *Portiera* lineages), has allowed the loss of these pathways in *Portiera* BT-QVLC. In addition, *Portiera* TV-BCN has lost the second step of the His biosynthetic pathway (*hisE*), but probably it is not required if the source of His is present in the diet of the host, or could be complemented in some manner by the host.

From the ten transporters probably present in the *Portiera* LCA, the galactose transporter (*galP*) has been pseudogenized in BT-QVLC strain and could be that different sugar molecules pass through diffusion across the membranes. Also, it is possible that due to the degradation of the glycolysis/pentose phospate pathway, at least BT-QVLC, does not require to import sugar molecules and only needs to import the intermediate metabolites required. Although few of these transporters have a known ligand, most of them should have a wide range of targets because all *Portiera* strains need to import mostly the same compounds/amino acids (see purple strokes in Figure 4.3.6) and not all of them can pass freely across the membranes. Finally, *Portiera* BT-QVLC and AF-CAI have maintained part of the superoxide detoxification pathway (*ahpC*), while TV-BCN and AD-CAI have lost it. It is possible that carotenes, well known antioxidants, supersede in some manner the superoxide protection of *ahpC*.

Although the genomes of all *Portiera* strains contain a set of tRNA genes for all amino acids, two of the genes encoding the aminoacyl tRNA synthetases responsible for charging each amino acid to its specific tRNA are absent (*argS* and *thrS*). Although the gene (*asnS*) is also absent, the synthesis of Asn-tRNA may be produced by the combination of *aspS* and *gatABC*, as explained in Section Comparative genomics. (Bernard *et al.*, 2006). Three more genes encoding aminoacyl tRNA synthetases have been lost in *Portiera* BT-QVLC (*alaS*, *metG* and *trpS*). The two former were also lost in *Portiera* TV-BCN. The *alaXp*[68] gene, that conforms the editing domain of *alaS*, is in charge of correcting the miss-charged tRNA^{Ala} avoiding its lethal effects (Chong *et al.*, 2008; Guo *et al.*, 2009). It is maintained in BT-QVLC and TV-BCN even than the rest of *alaS* has been lost. *Portiera* AD-CAI maintains both domains as separate CDS (*alaS* and *alaXp*), while AF-CAI has the whole *alaS* gene. In this context, it is possible that BT-QVLC and TV-BCN need to cover this lost function, maybe by importing the AlaS proteins or an already charged Ala-tRNA, but it is interesting that the protective function of *alaXp* is still needed and seems to be very important in all the *Portiera* strains. The lack of some aminoacyl tRNA synthetases, but the presence of their respective tRNAs would suggest that *Portiera* is acquiring these proteins from other source, maybe from a S-endosymbiont (Lamelas *et al.*, 2011b; Manzano-Marín and Latorre, 2014) (Section Metabolic integration) or from the host (Nakabachi *et al.*, 2014).

It is reasonable to think that ancestral whiteflies had similar nutritional requirements that the extant ones, because they were also sap-feeders probably related to gymnosperms (Drohojowska and Szwedo, 2014). The *Portiera* LCA, under a maximum parsimony scenario, lacked the ability to synthesize most of the vitamins and cofactors. Because the aforementioned lacks in *Portiera* LCA, it would be possible that ancestral whiteflies also had required a S-endosymbiont as seems to occur with the extant whiteflies. Although it is possible that no S-endosymbionts are required by whiteflies because some *B. tabaci* biotypes populations seem to lack them (Zchori-Fein *et al.*, 2014), it is clear that they confer some fitness advantages (Himler *et al.*, 2011). It could be also possible that in some environments, where different plants are available, or there is a rotation in plant species (e.g. seasonal plants), the quality (number of essential amino acids and vitamins/cofactors) and quantity (amount of these compounds) of the available sap suffers changes. When a good quality sap is available to the whiteflies, maybe S-endosymbionts are not needed and can be lost (Feldhaar, 2011; Ferrari and Vavre, 2011). A switch from good to bad sap, could enforce whiteflies to acquire again an S-endosymbiont to supply the amino acids and cofactors not produced by *Portiera* and fulfil their diet (Su *et al.*, 2014). These could explain why S-endosymbiont are not obligate endosymbionts in whiteflies and do not show the co-evolution pattern showed by *Portiera*.

---

[68] The gene *alaXP* was wrongly annotated as *alaS* in the first annotation version of *Portiera* BT-QVLC but correctly identified during the comparative genomics analyses of *Portiera* strains

In any case, the aforementioned could had produced an evolutionary constrain in whiteflies and the development of their special endosymbiont transmission route could be a solution to this problem (see Section Endosymbiont transmission in whiteflies). This transmission route ensures that a set of endosymbionts that can complement the unbalanced diet of the insect are transferred to the offspring.

4.3.5. Divergence times of *Portiera* lineages

Portiera strains divergence was estimated using the fossil record from their hosts and using *H. elongata* and *C. salexigens* as outgroups

(Figure 4.3.7). The first fossil of a whitefly (from the extinct subfamily Bernaeinae) was dated at the Upper Jurassic (163.5 mya-145 mya) in Shcherbakov (2000), while the oldest Aleyrodinae (*Baetylus kahramanus*) and Aleurodicinae (*Gapenus rhinariatus*) fossils were dated at the Lower Cretaceous (ca. 135-125 mya) in Drohojowska and Szwedo (2011) and Drohojowska and Szwedo (2013) respectively (Figure 4.3.1 in page 124). Although whiteflies were present since the Upper Jurassic, it was not until at some point in the Lower Cretaceous, when they diverged in the present subfamilies, so, the LCA or calibration point was set as a uniform distribution with an upper bound of 135 mya and a lower bound of 125 mya.

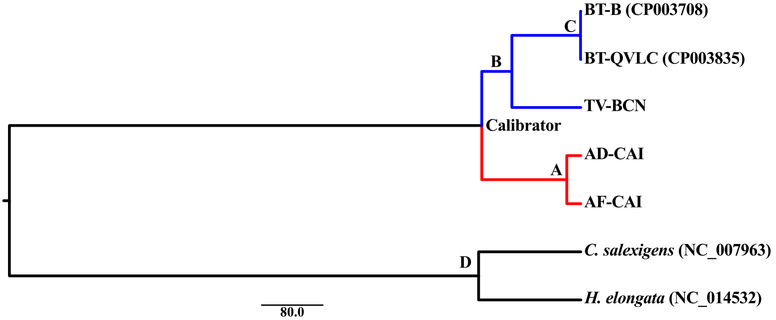
Two datasets were used for dating *Portiera* strains divergence: dataset A composed of *rpoB*, *rpoC*, *carB*, and *dnaE* genes (14280 bp) and dataset B of *sucA*, *aceE*, *valS* and *leuS* genes (13317 bp). They were selected for two reasons: they were the longest genes in *Portiera* genomes, and there were no significance differences between the branches leading to *Trialeurodes* and *Aleurodicus* (Relative rate test, data not shown). **BEAST2** Highest Posterior Density (HPD)[69] obtained with the two datasets for each estimated node overlapped, meaning that they were from the same distribution, and allowed the combination of both dataset to estimate the average parameters (Run AB from Table 4.3.3). Moreover, **PhyloBayes3** HPD also overlapped with **BEAST2** HPD, indicating that despite some differences in the wide of these HPD, all estimates came from the same distribution. The results for the divergence time estimations from **BEAST2** and **PhyloBayes3** are summarized in Table 4.3.3.

The estimated divergence of the two *Portiera* strains from *Aleurodicus*, *A. dispersus* and *A. floccissimus*, was 18.35 mya (node A in Figure 4.3.7 and Table 4.3.3) while the separation between *Portiera* strains from *T. vaporariorum* and *B. tabaci* was 90.1 mya (node B in Figure 4.3.7 and

---

[69] HPD - The x% highest posterior density interval is the shortest interval in parameter space that contains x% of the posterior probability (`http://www.beast2.org/wiki/index.php/Glossary`). For a detailed explanation about credible intervals, see `http://www.bayesian-inference.com/credible`

Table 4.3.3). *Aleurodicus* diverged during the last part of the Oligocene and the second-to-last period of the Miocene (Chattian-Tortonian, 28.1-11.62 mya). During this time, it took place the major evolution of the present flowering plants and the domination of open fields composed by grasses and dicotyledonous plants (Drohojowska and Szwedo, 2014). The split of the lineages conducting to *T. vaporariorum* and *B. tabaci* was during the Upper Cretaceous (100.5-66 mya). During this period flowering plants lineages (angiosperm), and probably herbivorous insects that were able to feed on them started to diverge (Drohojowska and Szwedo, 2014). The divergence between *Portiera* strains form *B. tabaci* B (MEAM1) and Q (MED) biotypes is more recent: 380,000 years ago (node C in Figure 4.3.7 and Table 4.3.3). If **PhyloBayes3** results are taken into account, it is possible that divergence between B (MEAM1) and Q (MED) biotypes occurred even in the late Pleistocene (Ionian-Tarantian, 0.781-0.0117 mya), before the actual geological period (Holocene, 0.0117-0 mya) 4.3.3. The divergence date between B (MEAM1) and Q (MED) biotypes was clearly in contrast to the 13 Myr (8-25 Myr) reported for a ca. 600 bp alignment of the mtCOI (Boykin *et al.*, 2013).

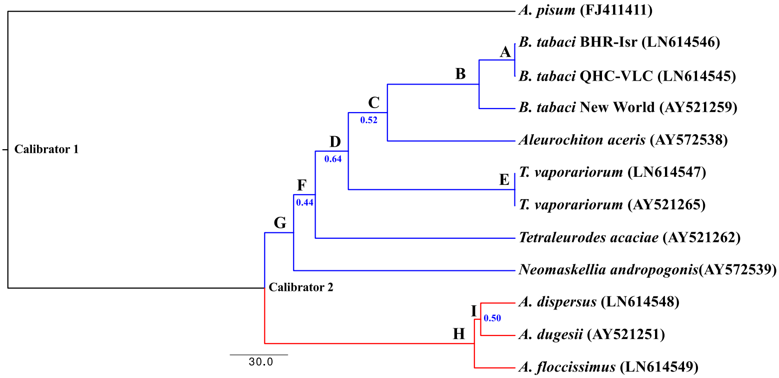


Figures 4.3.7 BEAST2 Bayesian inferred tree of *Portiera* strains. Each node whose divergence time was estimated is denoted by a bold uppercase letter (see Table 4.3.3). Each strain is displayed with its accession number in brackets. All posterior probabilities were 1. Branch lengths are displayed in Myr. *C. salexigens* and *H. elongata* were used as outgroup. Branches were colored according to the host subfamily: Aleyrodinae in blue and Aleurodicinae in red.

Table 4.3.3 Divergence dates, in Myr, for the different nodes of *Portiera* lineages.

Node	Description	Software	Run	Mean Age	G.M. Age	Median	Inf. 95 % HPD	Sup. 95 % HPD
Calibrator	Aleyrodidae Aleyrodinae - Aleurodicinae	BEAST2	A	129.67	125.00	134.39	129.64	129.50
			B	129.67	129.64	129.50	125.00 4	134.404
			AB	**129.47**	**129.44**	**129.22**	**125.00**	**134.31**
		PhyloBayes3	A	108.87			73.54	124.60
			B	109.41			76.07	124.51
A	Aleurodicinae *A. dispersus - A. floccissimus*	BEAST2	A	20.30	19.57	19.67	10.43	31.52
			B	17.68	17.14	17.16	9.62	26.71
			AB	**18.35**	**18.07**	**18.10**	**12.30**	**24.88**
		PhyloBayes3	A	30.97			14.83	55.19
			B	28.80			14.19	50.31
B	Aleyrodinae *T. vaporariorum – B. tabaci*	BEAST2	A	84.58	83.81	84.90	62.52	106.18
			B	93.54	92.86	94.02	71.81	114.44
			AB	**90.10**	**89.73**	**90.19**	**74.20**	**105.72**
		PhyloBayes3	A	63.80			40.91	84.91
			B	71.84			46.71	92.90
C	*B. tabaci* B(MEAM1) - Q(MED)	BEAST2	A	0.49	0.14	0.91	0.44	0.45
			B	0.35	0.31	0.32	0.07	0.69
			AB	**0.38**	**0.36**	**0.36**	**0.16**	**0.63**
		PhyloBayes3	A	0.10			0.04	0.19
			B	0.07			0.02	0.15
D	*H. elongata - C. salexigens*	BEAST2	A	114.81	110.73	111.36	58.99	177.02
			B	93.54	92.86	94.02	71.81	114.44
			AB	**133.71**	**131.54**	**132.01**	**88.18**	**181.17**
		PhyloBayes3	A	76.55			27.25	213.38
			B	130.88			38.94	396.41

To corroborate the *Portiera* dating results, the divergence of different whiteflies was estimated using a 1341 bp alignment of the mtCOI. Usually, only a fragment of the mtCOI is sequenced for phylogenetic analyses, and this could introduce some biases in the divergence dating. For avoiding this problem, only whiteflies whose mitogenomes were available were used. This cut-off diminished the number of species used in the mtCOI but ensures a good sequence data dating. The species included and their phylogenetic relationships are shown in the fixed tree[70] from Figure 4.3.8. Again, **BEAST2** and **PhyloBayes3** HPDs overlapped indicating the robustness of the estimates obtained (Table 4.3.4).


A. pisum (FJ411411)
B. tabaci BHR-Isr (LN614546)
A
B. tabaci QHC-VLC (LN614545)
B
B. tabaci New World (AY521259)
C
0.52
Aleurochiton aceris (AY572538)
Calibrator 1
D
0.64
T. vaporariorum (LN614547)
E
F
0.44
T. vaporariorum (AY521265)
G
Tetraleurodes acaciae (AY521262)
Neomaskellia andropogonis(AY572539)
Calibrator 2
A. dispersus (LN614548)
I
0.50
H
A. dugesii (AY521251)
30.0
A. floccissimus (LN614549)

Figures 4.3.8 BEAST2 Bayesian inferred tree of different whiteflies. Each node which divergence time was estimated are denoted by a bold uppercase letter (see Table 4.3.4). Each species are displayed with its accesion number in brackets. In blue are displayed the posterior probabilities below 1. Branch lengths are displayed in Myr. *A. pisum* was used as outgroup. Branches were colored according to the subfamily: Aleyrodinae in blue and Aleurodicinae in red.

In this case, mtCOI from *A. pisum* was selected as the outgroup for rooting the tree. Calibration points were set to an uniform distribution using different estimations of the emergence of the Sternorrhyncha suborder (250-278 mya) and the emergence of the Aleyrodinae and Aleurodicinae subfamilies (135-125 mya) (Drohojowska and Szwedo,

---

[70] One condition for some divergence dating programs is that phylogenetic trees can not include paraphyletic groups. In consequence, they force the trees to be monophyletic

2011, 2013, 2014; Shcherbakov, 2000; Shi *et al.*, 2012; Wootton, 1981). The divergence between *A. floccissimus* and the clade of *A. dispersus*-*A. dugesii* took place 20.25 mya (node H in Figure 4.3.8 and Table 4.3.4), very similar to the estimates using *Portiera* datasets. The separation of the *Trialeurodes* lineage and the lineage leading to *Bemisia* was estimated 86.07 mya (node D in Figure 4.3.8 and Table 4.3.4), also very similar to the *Portiera* estimates. Due to the incomplete *Bemisia* taxon sampling in this work, it could be roughly estimate the emergence of this genus between 66.05 mya and 18.43 mya (node C and B, respectively, in Figure 4.3.8 and Table 4.3.4). According to De Barro *et al.* (2011) (Figure 1.4.2 in page 26), *B. tabaci* B (MEAM1) and Q (MED) biotypes and New World biotype are part from two different clades that join in the basal branch of the *B. tabaci* complex. So, the time estimated for the divergence between B(MEAM1)/Q(MED) and New World (node B in Figure 4.3.8 and Table 4.3.4) should be the divergence of the *B. tabaci* complex, around 18.43 mya. Finally, the divergence between the B (MEAM1) and Q (MED) biotypes was 0.21 mya (0.03-0.55 mya), in the range of the estimates using *Portiera*. Although the *Bemisia* genus divergence time includes the Paleocene-Eocene Thermal Maximum (55 mya)[71] and it is possible that it diversified during this period, the *B. tabaci* complex diverged later in the Oligocene-Miocene (33.9-7.246 mya), probably linked to expansion of dicotyledonous plants and grasses and to the decrease in the global temperature. Finally, even though the *B. tabaci* complex diverged before the raise of the agriculture, it is not possible to rule out with a complete confidence (HPDs are so close to the Holocene epoch) that this was the reason underlying the divergence between B (MEAM1) and Q (MED) biotypes, that are the two most invasive biotypes from *B. tabaci* complex.

---

[71] Or PETM. This period of time showed an increase in the global temperature that was related to an increase of herbivore insects activity (Currano *et al.*, 2008)

Table 4.3.4 Divergence dates, in Myr, for different whiteflies based on a 1341 pb mtCOI fragment

Node	Description	Software	Mean Age	G.M. Age	Median	Inf. 95 % HPD	Sup. 95 % HPD
Calibrator 1	Sternorrhyncha	BEAST2	263.24	263.10	262.40	250.00	277.66
		PhyloBayes3	207.66			147.12	283.65
Calibrator 2	Aleyrodidae Aleyrodinae - Aleurodicinae	BEAST2	129.74	129.71	129.60	125.00	134.42
		PhyloBayes3	130.50			125.34	134.83
A	*B. tabaci* B(MEAM1) - Q(MED)	BEAST2	0.21	0.16	0.14	0.03	0.55
		PhyloBayes3	1.17			0.44	2.87
B	*B. tabaci* B(MEAM1)/Q(MED) - New World	BEAST2	18.43	17.80	17.73	9.85	28.50
		PhyloBayes3	19.87			11.16	32.44
C	*A. aceris* - *Bemisia*	BEAST2	66.05	65.15	65.63	45.15	87.41
		PhyloBayes3	61.39			41.44	83.16
D	*Trialeurodes* - *Bemisia*/*A. aceris*	BEAST2	86.07	85.28	85.95	63.80	108.73
		PhyloBayes3	81.94			59.94	103.43
E	*T. vaporariorum*	BEAST2	0.02	0.01	0.01	0.00	0.06
		PhyloBayes3	0.12			0.01	0.41
F	*T. acaciae* - *Trialeurodes*/*Bemisia*/*A. aceris*	BEAST2	103.09	102.46	103.53	81.23	125.08
		PhyloBayes3	95.38			73.06	116.14
G	*N. andropogonis* - other Aleyrodinae	BEAST2	114.39	113.93	115.54	94.94	132.21
		PhyloBayes3	113.17			91.49	130.08
H	Aleurodicus	BEAST2	20.25	18.52	17.26	8.27	37.31
		PhyloBayes3	47.94			26.14	78.68
I	*A. dispersus* – *A. dugesii*	BEAST2	17.11	15.60	14.71	6.67	32.09
		PhyloBayes3	38.80			24.48	65.82

It is worth to mentioning that the phylogenetic trees topologies for the *B. tabaci* complex were slightly different in Boykin *et al.* (2013) and De Barro *et al.* (2011). The former estimated the origin of this genus in 86 Myr (70-102 mya), the origin of the *B. tabaci* complex in 57 Myr (45-66 mya) and the divergence of *B. tabaci* B(MEAM1)/Q(MED) from *B. tabaci* New World (node B Figure 4.3.8 and Table 4.3.4) in 48 Myr (34-60 Myr). Because the dates obtained in this work for the *B. tabaci* B(MEAM1)/Q(MED) biotypes divergence and their separation from New World biotype (node A and B in Figure 4.3.8 and Table 4.3.4) are smaller than those obtained in Boykin *et al.* (2013), it seems that the origin of the *B. tabaci* complex and, especially, the divergence between *B. tabaci* B(MEAM1) and Q(MED) biotypes are more recent. It is possible that the differences between these results concerning the divergence of *Bemisia* genus and the *B. tabaci* complex could be due to the length of the mtCOI used. In (Boykin *et al.*, 2013), all the available mtCOI whiteflies sequences were used, producing a final alignment of less than 600 bp. Also it is possible that the use of a short gene fragment, the saturation of the phylogenetic signal, the presence of paraphyletic groups in the hosts inferred phylogenetic tree, and the use of a speciation model not recommended with intraspecific data (more than one individuals per species) have impacted their estimations, resulting in observed differences (Drummond *et al.*, 2006; Heled and Drummond, 2012; Ho *et al.*, 2005).

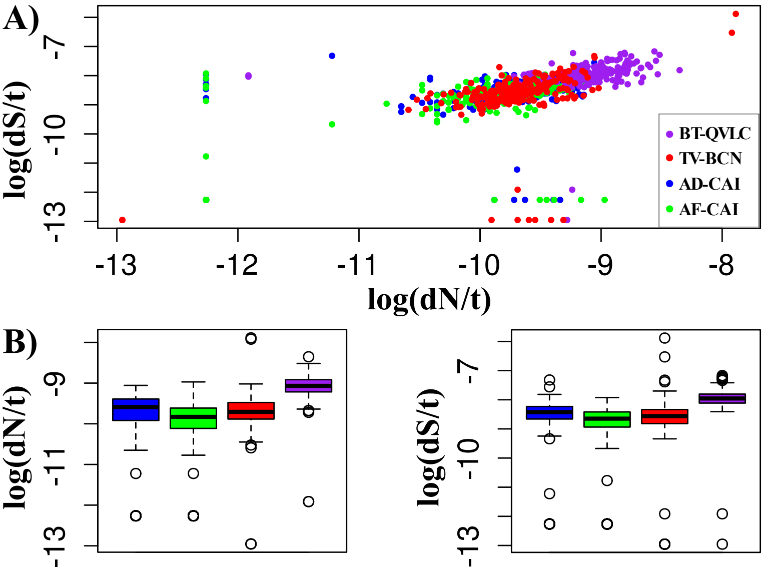
4.3.6. Rates of nucleotide substitution in *Portiera* lineages

The number of dS and dN were estimated in the lineages leading to *Portiera* BT-QVLC, TV-BCN, AD-CAI, and AF-CAI[72]. These values were divided by the mean age of the divergence times obtained in the run AB, to obtain the rates of substitutions/year (Table 4.3.3). dS and dN from *Portiera* BT-QVLC and TV-BCN were divided by 90.1 Myr while the ones from AD-CAI and AF-CAI by 18.35 Myr.

---

[72] The data used for the statistical analysis and the data can be found in Annex file dN_dS_data.tab

Firstly, a logarithmic transformation was performed[73]. When the initial raw data were plotted (240 genes), two main clusters were observed for most of the core genes: *Portiera* BT-QVLC was the one with the highest rate of dN/year and dS/year and TV-BCN, AD-CAI and AF-CAI formed a second cluster with a lower rate (Figure 4.3.9 A). During the exploratory analysis, a quality trimming of the data was performed: all outliers and values outside the 25 % and 75 % quartiles were removed, keeping a 60 % of the original data (146 genes out of 240) (Figure 4.3.9 B).

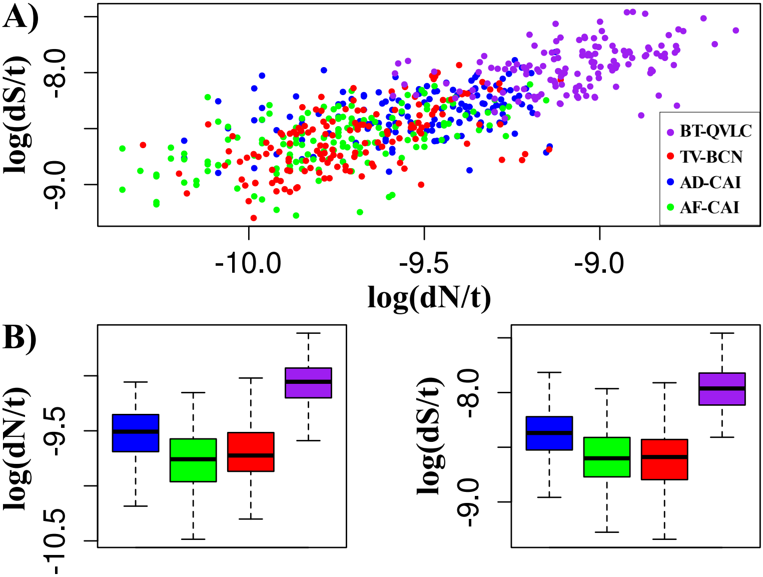


Figures 4.3.9 **A**) Scatter plot of the raw data output from codeML. Each dot represents the dN/year against dS/year logarithmic values of a single orthologous CDS. **B**) Box plot of the raw date before cleaning. Whiskers represents the 0 % and 100 % quartile. Colours representing each strain are the same as panel **A**. Notice that some dN/year and dS/year tendencies are masked by the outliers and extreme values.

---

[73] 0.0001 was added to all zero values of dN dS before the logarithmic transformation. It is important to notice that **codeML** have a maximum of 4 decimals, so it is probable that zero values are in fact very low dN or dS values

After trimming, three clusters were observed (from higher to lower dN/year against dS/year): *Portiera* BT-QVLC, AD-CAI, and TV-BCN/AF-CAI (Figure 4.3.10 A). This distribution was confirmed when the values for each strain were plotted (Figure 4.3.10 A). From the Figure 4.3.10 B, it is clear that *Portiera* BT-QVLC presented the highest distribution of dS/year and dN/year as it is expected for the large branches in the phylogenetic tree (Figure 4.3.2). This high substitution rate in *Portiera* BT-QVLC (and at least in *Portiera* strains from *B. tabaci* Q (MED) and B (MEAM1) biotypes) could be directly related with the replication and repair machinery losses produced in this lineage, which have probably increased the mutation rate. Moreover, it is probable that this tendency could be observed in all *Portiera* from the *B. tabaci* lineage, because the loss of these genes seems to have occurred early in the



Figures 4.3.10 A) Scatter plot of the cleaned data output from codeML. Each dot represents the dN/year against dS/year logarithmic values of a single orthologous CDS. **B**) Box plot of the cleaned data, Whiskers represents the 0 % and 100 % quartile. Colours representing each strain are the same as panel **A**. Notice that dN/year and dS/year tendencies are now clearly distinguishable.

divergence of the *B. tabaci* complex.

To determine if the rates of dN/year were significantly different among lineages, several test were performed. Because *Portiera* BT-QVLC failed to pass Levene's test when was compared to the other *Portiera* strains, and based on exploratory analysis, the dN/year distribution of this *Portiera* is clearly different from the other strains. To determine if the rates of dN/year were significantly different between the remaining *Portiera* strains, a Kruskal-Wallis test was performed[74]. The test gave a significant result (p-value $9e^{-12}$), supporting that not all the dN/year distributions were equal (Figure 4.3.10 B). *Post-hoc* Kruskal-Wallis test confirmed that there is statistical significance to assume that AD-CAI presents a different dN/year distribution compared to AF-CAI or TV-BCN (p-values, $4.4e^{-14}$ and $4.0e^{-08}$, respectively) and non significant differences between AF-CAI/TV-BCN (p-value = 0.039).

At dS/year level, *Portiera* BT-QVLC failed also to pass Levene's test, so it was not necessary to use another test to check if its dS/year distribution is statistically different to the other strains. Similar results to the dN/year distributions in the dS/year distributions were found when the remaining *Portiera* strains were compared. AD-CAI to AF-CAI or TV-BCN comparisons (T-test or Welch's procedure for unequal variances p-values: $2.213e^{-13}$ and $9.294e^{-12}$, respectively)[75] supported that AD-CAI had a statistically different dS/year mean. In contrast, AF-CAI and TV-BCN showed no differences at dS/year means (T-test with equal variance p-value = 0.859)[76]. In conclusion, it seems that AF-CAI and TV-BCN have similar substitution rates, lower than the ones reported for AD-CAI and BT-QVLC with the last being the extreme case.

In addition, the ω (dN/dS) was calculated for the orthologous CDS. CDS with dS values equal to zero or with a ω greater than ten were trimmed (53 and 1 out of 240, respectively). ω values of each population followed a non-normal distribution with equal variances. The median ω values for BT-QVLC, TV-BCN, AD-CAI and AF-BCN were 0.0743, 0.0735, 0.0643, and 0.0656, respectively[77]. When a Kruskal-Wallis test was applied, no significant differences were found between the ω distribution of the *Portiera* strains (p-value 0.2167). This implies that the core genes, on average, are evolving under purifying selection ($\omega < 1$). Also, the high increase in the rates of nucleotide substitutions in the BT-QVLC lineage affects in the same proportion to dN and dS. This means that natural selection seems not to be responsible for the evolutionary pattern observed in *Portiera* strains from *B. tabaci* and it could be due to other parameters such as population size, number of generations, mutation rate, etc. In fact, gene losses related to the replication and repair machinery (e.g. *dnaQ*) in *Portiera* from *B. tabaci* could be the principal reason of its accelerated evolution, as proposed by Sloan and Moran (2013).

---

[74] H_0: all populations have identical distribution functions; H_A: not all populations have identical distribution functions

[75] H_0: there are no differences between the means of the samples ; H_A: there is a difference between the means of the two samples

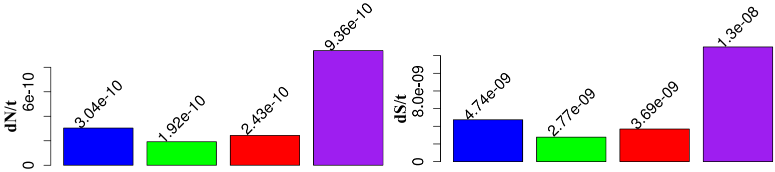
[76] H_0: there are no differences between the means of the samples ; H_A: there is a difference between the means of the two samples

Finally, a dN/year and dS/year genomic ratio was calculated for each *Portiera* strain (Figure 4.3.11). In general, AD-CAI showed approximately twice the ratio of AF-CAI in both dN/year and dS/year. TV-BCN dN/year and dS/year were between AD-CAI and AF-CAI but closer to the latter, as expected from the statistical analysis reported above. *Portiera* BT-QVLC presented a genomic dN/year and dS/year three times greater than AD-CAI and more than four times than the ones reported for AF-CAI and TV-BCN. Despite the accelerated evolution of *Portiera* BT-QVLC, it is still in the range of dS/year values given for other P-endosymbionts like *Buchnera* or *Blochmania* ($4.3e^{-09}$ and $1.5e^{-08}$ dS/year respectively)[78] (Gómez-Valero *et al.*, 2007, 2008) as well as the other *Portiera* strains. However, these values are far away from the values reported for *Buchnera* in Moran *et al.* (2009) (2.2e^{-07} dS/year)[79].

---

[77] The median is used because the data was non-normal distributed

[78] In this case the dS/year was estimated in non-functional regions. Intergenic regions were used for *Blochmania* while some pseudogenes were used for *Buchnera*. Because dS rate is neutral, or quasi-neutral, is considered that it does not depend on the function of the DNA region as the dN. This allows the comparison between non-functional and functional DNA region



Figures 4.3.11 Genomic dN/year and dS/year values using 240 orthologous CDS. *Portiera* strain colours are the same as Figure 4.3.9.

4.3.7. Selective pressure in *Portiera* from *B. tabaci*

The idea that IGRs in *Portiera* from *B. tabaci* could be involved in transcription differences between the Q (MED) and B (MEAM1) biotypes was proposed in Jiang *et al.* (2013), but later discarded in Sloan and Moran (2013). In this context, an analysis to explore the possibility that some CDS codons (or sites) have suffered positive selection events in *Portiera* strains of *B. tabaci* was conducted. For positive selection detection, a set of “background” branches (or branch) are used as reference for searching putative codons under selection in the “foreground” branches. Because this test relies in *a priory* hypothesis, strains (also populations or species) that show an outbreak (or an apparent fitness increase) compared to the other strains are set as the “foreground” branches. Because both *B. tabaci* seems to be the species with a higher pest impact compared to *T. vaporariorum*[80], these strains were the target of the positive selection test. Three tests for sites under positive selection were made: *Portiera* strains of *B. tabaci* compared to *T. vaporariorum* strains (TV and TV-BCN), Portiera of *B. tabaci* Q (MED) biotype (BT-QVLC and BT-Q-AWRs) compared to B (MEAM1) biotype (BT-B and BT-B-HRs) and *vice versa* (Table 4.3.5, Figure 4.3.2). Because *Portiera* strains from B (MEAM1) and Q (MED) biotypes diverged very recently and have more than 99 %

---

[79] This estimation was made between *Buchnera* strains from hosts that diverged less than 200 years ago. At this short period of time, mutation rate approximately equals the substitution rate and the latter is overestimated

[80] The worldwide widespread of *B. tabaci* compared to *T. vaporariorum* could be taken as a some kind of fitness indicator

nucleotide identity, it was necessary a third *Portiera* strain from a different *B. tabaci* biotype to be used in the “background” branches to infer the correct ancestral codon for a positive selection test. For this reason, the last two tests that compare the *Portiera* from *B. tabaci* strains could only be used as an information source of what proteins are accumulating significance changes between the two *Portiera* from *B. tabaci* strains, but not as an indicator of positive selection. Also, these two tests have less statistical power than the first one, as can be seen in the large increment of proteins at the 70 % Bayes Empirical Bayes (BEB) confidence threshold and the fact that some sites detected were changes from one amino acid to a similar one (e.g. change at position 376 of *trpB* were from valine to leucine in Q biotype and valine to isoleucine in B biotype, data not shown).

Table 4.3.5 Genes with codons that were positive for **codeML** selection analysis

BEB confidence	*B. tabaci*^{f}, *T vaporariorum*	*B. tabaci* Q^{f}, *B. tabaci* B	*B. tabaci* B^{f}, *B. tabaci* Q
95 %	trpG*, trpB*, thrB*, arcB*, asd*, aroA*, aroB*, leuC*, gltX$^{\wedge}$, rimM, rpoB, tufA, rplV, rpsQ, nuoJ, hlsV, hslU, yggt	lysS$^{\wedge}$, rplE, rpsF, rplI, cyoA, secA	trpA*, lysS$^{\wedge}$, cysS$^{\wedge}$, nuoJ, glyQ, hslV
70 %	trpG*, trpB*, thrB*, ilvI*, arcB*, asd*, aroA*, aroB*, leuC*, gltX$^{\wedge}$, rimM, rpoB, tufA, rplV, rpsQ, cyoA, nuoJ, hslV, hslU, yggt	trpE*, trpG*, trpC*, trpB*, lysC*, dapE*, ilvI*, ilvC*, arcB*, aroK*, aroA*, lysS$^{\wedge}$, proS$^{\wedge}$, leuS$^{\wedge}$, PAQ_222, lepA, lpd, hslV, tufA, rplA, gyrB, rplD, rplV, rplE, secA, secY, sucA, rplI, pnp, secA, hslU, nusA	trpC*, trpA*, trpB*, dapE*, hom*, thrB*, leuC*, ilvI*, aroB*, aroK*, glyQ$^{\wedge}$, gltX$^{\wedge}$, lysS$^{\wedge}$, tyrS$^{\wedge}$, proS$^{\wedge}$, cysS$^{\wedge}$, leuS$^{\wedge}$, nuoJ, sucA, rplV, der, rpoA, rpoB, rplD, rpsQ, cyoA, secA, secY, rplE, pnp, dnaE

$^{\wedge}$Transcription, translation and ribosome biogenesis *Aminoacid biosynthesis ^{f} Foreground branch

When *Portiera* strains of *B. tabaci* were compared to *T. vaporariorum* strains, 18 genes showed sites under positive selection after Bonferroni’s correction at a 95 % BEB confidence. This result needs to be interpreted carefully because, for example, some genes (e.g. *rimM*, *rplV*, *rpsQ*, *nuoJ*, *hslV*, and *yggt*) could be false positives due to their shortness. Although, for longer genes it seemed to be an enrichment in amino acid biosynthesis category at 95 % BEB confidence (one more gene is added to this category if a 70 % BEB confidence threshold is considered) (Table 4.3.5).

In comparison, *Portiera* strains of *B. tabaci* Q biotype against B biotype showed six genes with a positive selection signal. In this case, at 95 % BEB confidence, most of them were short genes with the exception of *lysS* and *secA*. At 70 % BEB confidence, 12 genes from 32 genes detected were related to the amino acid biosynthesis and three genes were aminoacyl tRNA synthetases. Finally, the list includes *secA* and *secY*, a kind of genes that are under positive selection in free-living bacteria such *E. coli* (Chen *et al.*, 2006; Petersen *et al.*, 2007) and could be considered a validation of the results presented.

When *B. tabaci* B biotype was compared against Q biotype, five genes showed sites under positive selection at 95 % BEB confidence and 31 genes at 70 % BEB confidence. Also *secA* and *secY* were detected, so it is possible that genes that were detected in both test are in fact false positives. At the higher confidence, one amino acid biosynthetic gene (*trpA*) and three aminoacyl tRNA synthetases (*lysS* and *cysS*) presented signs of positive selection. At the lower confidence level, nine genes were related to amino acid biosynthesis and seven were aminoacyl tRNA synthetases. Even if the genes only reported for one of the test are taken as valid ones, these results need to be interpreted cautiously.

It is possible that one way how selection can modify the fitness of the symbiont-host system is to increase the availability of essential amino acids. Taking into account that the *B. tabaci* strains is the most invasive whitefly, while *T. vaporariorum* is only a real problem in the greenhouses, it can be considered that, despite other factors, this amino acid increase could favour the fitness of the invasive species. This seems to be supported by the fact that an important part of genes that showed sites under positive selection are related (directly or indirectly) to the amino acids biosynthetic capabilities of *Portiera*. How these changes could favour the selection of the proteins remains unclear. One option is that some residues could increase the stability of the protein in certain environments. Also, it is possible that these changes are compensatory mutations due to the degradation that suffered the *Portiera* from *B. tabaci* lineage. However, it is important to mention that the changes detected by the tests could be an effect of the increased dN and dS detected in the *B. tabaci* lineages instead of positive selection signature. Also, the large divergence time between *B. tabaci* and *T. vaporariorum* could allow the accumulation of more changes in *B. tabaci* than in *T. vaporariorum* due to their different mutation rates.

Part 5

Conclusions

“ A tout le monde (To all the world)
A tout mes amis (To all my friends)
Je vous aime (I love you)
Je dois partir (I have to leave)
These are the last words
I’ll ever speak...
And they’ll set me free”

Megadeth

The conclusions obtained in this work can be summarized and grouped according to:

Portiera BT-QVLC and its partner *Hamiltonella* BT-QVLC

1. *Portiera* presents a canonical three-membrane system (host-derived vacuolar membrane and the gram-negative bacterial cell wall) as other P-endosymbionts.
2. *Portiera* from *B. tabaci* presents the same genomic features shared with other P-endosymbionts: low G/C content, reduced genome compared to free-living bacteria and absence of mobile elements. In contrast, it presents some special features: large intergenic regions, low coding density and abundance of tandem repeats.
3. *Portiera*, as its relatives *Carsonella* and *Evansia*, has lost its cellular autonomy and crossed the border-line between an endosymbiont and an organelle.
4. *Portiera* from *B. tabaci* synthesizes, or participates, in the synthesis of all essential amino acids except lysine. They serve to complement the deficient diet of the insect hosts. *Hamiltonella* contributes with many vitamins, cofactors and several amino acids, including lysine.
5. *Portiera* is able to synthesize carotenes that could act as an antioxidant but also as an alternative source of reductive power. It remains unclear if *Portiera* exports carotenes to the insect.
6. All *Portiera* from *B. tabaci* sequenced to date are almost identical. In consequence, the lysine pathway and the vitamins and cofactors should be supplied to the host by a bacteriome-confined S-endosymbiont (e.g. *Hamiltonella*, *Arsenophonus*, *Hemipteriphilus*, etc...).

The third passenger: *Cardinium* cBtQ1

7. *Cardinium* cBtQ1 endosymbiont of *B. tabaci* forms the new family Amoebophilaceae, together with *Cardinium* cEper1 and *A. asiaticus*.
8. *Cardinium* cBtQ1 presents a genome highly impacted by mobile

elements (17 % of its genome). Some of these mobile elements are still active and could play an important role in the plasticity and adaptation of *Cardinium* cBtQ1 to the *B. tabaci* environment.

9. The gene contents in *Cardinium* cBtQ1 and cEper1 are quite similar, with most of the differences due to hypothetical proteins, most of them probably artefacts. Most of the gene repertoire evolution in these genomes are due to a reductive process but maintaining the ability to acquire new genetic material though HGT events.
10. *Cardinium* cBtQ1 is not related to the host diet complementation and presents a strong competition with *Hamiltonella* for the host resources. It could be possible that the scattered phenotype of *Cardinium* cBtQ1 may be a strategy for avoiding this competence.
11. Gliding genes could be responsible of the scattered phenotype in *Cardinium* cBtQ1 because *Cardinium* cEper1 has lost these genes and not presents this phenotype. The gliding genes could form a minimal gliding machinery that allows *Cardinium* to move outside the bacteriome and invade new tissues. However, it is possible that the gliding genes form part of a T6SS secretory system.
12. It could be possible that *Cardinium* cBtQ1 confers some advantages to *B. tabaci* Q1 biotype (MED) because it is almost fixed in Valencia province. The scattered phenotype and a putative toxin in *Cardinium* cBtQ1 plasmid could be related to a defensive role.

Genome evolution of the genus *Portiera*

13. The species from the genus *Portiera* have maintained an almost perfect genome stasis for the last 125-135 Myr with only few gene losses in different lineages.
14. *Portiera* from *B. tabaci* presents an especial evolutionary pattern not shown by other *Portiera* strains from different whiteflies. This pattern is characterized by large intergenic regions, genome size increase, high number of recombinations events, tandem repeats, a large amount of gene losses and a high substitution rates.

15. A combination of replication, recombination and repair gene losses and the illegal recombination mechanism seems to be responsible of the genomic instability in *Portiera* from *B. tabaci*.
16. The *Portiera* LCA gene repertoire, composed of 319 genes and 280 CDS, was almost identical to *Portiera* AF-CAI. This strain is able to produce alone, or with some support of the host, the ten essential amino acids plus the non-essential glycine.
17. Metabolic capabilities of all *Portiera* strains, including the ability to synthesize carotenes, are quite similar and should need a S-endosymbiont for synthesize vitamins and cofactors.
18. The absence of essential aminoacyl tRNA synthetases, and other informational related genes, in all *Portiera* strains suggest that all of them have lost the cell autonomy for genetic information transfer systems.
19. Different CDS of *Portiera* and mitochondrial *COI* gene were used to estimate the divergence times of several whiteflies and *Portiera* lineages. Some divergences are remarkable such as those of *B. tabaci* B (MEAM1) and Q (MED) biotypes, estimated between 30,000 to 630,000 years, and that of the *B. tabaci* complex between 9.9 to 28.5 Myr. These dates are younger than previous published estimates.
20. The rates of gene nucleotide substitutions were estimated in four *Portiera* lineages. The faster evolving lineage was that of *B. tabaci*, followed by *A. dispersus* and *A. floccissimus*/*T. vaporariorum*.
21. No differences were obtained for gene average ω (dN/dS) values, reflecting that the acceleration of the substitution rate in the lineage of *Portiera* from *B. tabaci* was not due to a change in the pressure of natural selection.
22. The average rate of synonymous substitution in the genome (dS/year) was $3.7e^{-09}$ for all *Portiera* lineages except the fast evolving of *Portiera* from *B. tabaci* ($1.3e^{-08}$).

Part 6

Bibliography

“Quizá los hombres seamos a un tiempo Abel y Caín
quizá un día destruya lo oscuro que hay en mi
el destino no está marcado en la fe
yo he elegido ser lo que siempre seré
...HIJO DE CAIN”

Barón Rojo

Adams AS, *et al.* 2013. Mountain pine beetles colonizing historical and naive host trees are associated with a bacterial community highly enriched in genes contributing to terpene metabolism. *Appl. Environ. Microbiol.*, **79**(11):3468–3475. Link.

Ahmed MZ, De Barro PJ, Ren SX, Greeff JM, and Qiu BL. 2013. Evidence for horizontal transmission of secondary endosymbionts in the *Bemisia tabaci* cryptic species complex. *PLoS One*, **8**(1):e53084. Link.

Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, and Nielsen PH. 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.*, **31**(6):533–538. Link.

Andersen PC, Brodbeck BV, and Mizell RF. 1989. Metabolism of amino acids, organic acids and sugars extracted from the xylem fluid of four host plants by adult *Homalodisca coagulata*. *Entomol. Exp. Appl.*, **50**(2):149–159. Link.

Ardell DH and Andersson SGE. 2006. TFAM detects co-evolution of tRNA identity rules with lateral transfer of histidyl-tRNA synthetase. *Nucleic Acids Res.*, **34**(3):893–904. Link.

Ashburner M, *et al.* 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**(1):25–29. Link.

Aziz RK, *et al.* 2008. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, **9**:75. Link.

Baker M. 2012. De novo genome assembly: what every biologist should know. *Nat. Methods*, **9**(4):333–337. Link.

Baker TA and Sauer RT. 2012. ClpXP, an ATP-powered unfolding and protein-degradation machine. *Biochim. Biophys. Acta*, **1823**(1):15–28. Link.

Banjo AD. 2010. A review of on *Aleurodicus dispersus* Russel. (spiralling whitefly) [Hemiptera: Aleyrodidae] in Nigeria. *J. Entomol. Nematol.*, **2**(1):1–6. Link.

Basler M and Mekalanos JJ. 2012. Type 6 secretion dynamics within and between bacterial cells. *Science*, **337**(6096):815. Link.

Baumann L, Thao ML, Funk CJ, Ng JC, Baumann P, and Falk BW. 2004. Sequence Analysis of DNA Fragments from the Genome of the Primary Endosymbiont of the Whitefly *Bemisia tabaci*. *Curr. Microbiol.*, **48**(1):77–81. Link.

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

Belda E, Moya A, and Silva FJ. 2005. Genome rearrangement distances and gene order phylogeny in gamma-Proteobacteria. *Mol. Biol. Evol.*, **22**(6):1456–1467. Link.

Belda E, Moya A, Bentley S, and Silva FJ. 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. Link.

Bennett GM and Moran NA. 2013. Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a Phloem-feeding insect. *Genome Biol. Evol.*, **5**(9):1675–1688. Link.

Bermudes D, Hinkle G, and Margulis L. 1994. Do prokaryotes contain microtubules? *Microbiol. Rev.*, **58**(3):387–400. Link.

Bernard D, Akochy PM, Beaulieu D, Lapointe J, and Roy PH. 2006. Two residues in the anticodon recognition domain of the aspartyl-tRNA synthetase from *Pseudomonas aeruginosa* are individually implicated in the recognition of tRNAAsn. *J. Bacteriol.*, **188**(1):269–274. Link.

Bigliardi E, *et al.* 2006. Ultrastructure of a novel *Cardinium* sp. symbiont in *Scaphoideus titanus* (Hemiptera: Cicadellidae). *Tissue Cell*, **38**(4):257–261. Link.

Bing XL, Yang J, Zchori-Fein E, Wang XW, and Liu SS. 2013. Characterization of a newly discovered symbiont of the whitefly *Bemisia tabaci* (Hemiptera: Aleyrodidae). *Appl. Environ. Microbiol.*, **79**(2):569–575. Link.

Bisalputra T, Oakley BR, Walker DC, and Shields CM. 1975. Microtubular complexes in blue-green algae. *Protoplasma*, **86**(1-3):19–28. Link.

Bönemann G, Pietrosiuk A, and Mogk A. 2010. Tubules and donuts: a type VI secretion story. *Mol. Microbiol.*, **76**(4):815–821. Link.

Bouckaert R, *et al.* 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS computational biology*, **10**(4):e1003537. Link.

Bourque G and Pevzner PA. 2002. Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res.*, **12**(1):26–36. Link.

Bourtzis K and Miller TA, editors. 2003. *Insect Symbiosis, Volume 1*. CRC Press.

Boykin LM, Bell CD, Evans G, Small I, and De Barro PJ. 2013. Is agriculture driving the diversification of the *Bemisia tabaci* species complex (Hemiptera: Sternorrhyncha: Aleyrodidae)?: Dating, diversification and biogeographic evidence revealed. *BMC Evol. Biol.*, **13**(1):228. Link.

Brady A and Salzberg SL. 2009. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature methods*, **6**(9):673–6. Link.

Braun TF, Khubbar MK, Saffarini DA, and McBride MJ. 2005. *Flavobacterium johnsoniae* gliding motility genes identified by mariner mutagenesis. *J. Bacteriol.*, **187**(20):6943–52. Link.

Brelsfoard C, *et al.* 2014. Presence of Extensive *Wolbachia* Symbiont Insertions Discovered in the Genome of Its Host *Glossina morsitans morsitans*. *PLoS Negl. Trop. Dis.*, **8**(4):e2728. Link.

Buchner P. 1965. *Endosymbiosis of Animals with Plant Microorganisms*. 2. John Wiley & Sons, Inc.: Interscience Publ, New York.

Burchard A, Burchard R, and Kloetzel J. 1977. Intracellular, periodic structures in the gliding bacterium *Myxococcus xanthus*. *J. Bacteriol.*, **132**(2):666–672. Link.

Burge SW, *et al.* 2013. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, **41**(Database issue):D226–232. Link.

Byrne DN and Bellows TS. 1991. Whitefly Biology. *Annu. Rev. Entomol.*, **36**(1):431–457. Link.

Callejas C, Velasco A, Gobbi A, Beitia F, and Ochando MD. 2005. Fast discrimination (RAPD-PCR) of the species forming the pest complex *Aleurodicus dispersus-Lecanoideus floccissimus* (Hom: Aleyrodidae). *J. Appl. Entomol.*, **129**(7):382–385. Link.

Campbell BC, Steffen-Campbell JD, and Gill RJ. 1994. Evolutionary origin of whiteflies (Hemiptera: Sternorrhyncha: Aleyrodidae) inferred from 18S rDNA sequences. *Insect Mol. Biol.*, **3**(2):73–88. Link.

Carballido-López R. 2006. The bacterial actin-like cytoskeleton. *Microbiol. Mol. Biol. Rev.*, **70**(4):888–909. Link.

Caspi R, *et al.* 2014. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, **42**(Database issue):D459–471. Link.

Caspi-Fluger A, Inbar M, Mozes-Daube N, Mouton L, Hunter MS, and Zchori-Fein E. 2011. *Rickettsia* 'in' and 'out': two different localization patterns of a bacterial symbiont in the same insect species. *PLoS One*, **6**(6):e21096. Link.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**(4):540–552. Link.

Chapman RF. 2013. *The insects : structure and function*. Cambridge University Press, 5th editio edition.

Charles H, *et al.* 2011. A Genomic Reappraisal of Symbiotic Function in the Aphid/*Buchnera* Symbiosis: Reduced Transporter Sets and Variable Membrane Organisations. *PLoS ONE*, **6**:e29096. Link.

Chen SL, *et al.* 2006. Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. *Proc. Natl. Acad. Sci. U. S. A.*, **103**(15):5977–5982. Link.

Chevreux B, Wetter T, and Suhai S. 1999. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. *Comput. Sci. Biol. Proc. Ger. Conf. Bioinforma.*, **99**:45–56. Link.

Chiel E, Gottlieb Y, Zchori-Fein E, Mozes-Daube N, Katzir N, Inbar M, and Ghanim M. 2007. Biotype-dependent secondary symbiont communities in sympatric populations of *Bemisia tabaci*. *Bull. Entomol. Res.*, **97**(4):407–413. Link.

Chiel E, *et al.* 2009. Almost there: transmission routes of bacterial symbionts between trophic levels. *PLoS One*, **4**(3):e4767. Link.

Chiu SW, Chen SY, and Wong Hc. 2008. Dynamic localization of MreB in *Vibrio parahaemolyticus* and in the ectopic host bacterium *Escherichia coli*. *Appl. Environ. Microbiol.*, **74**(21):6739–6745. Link.

Chong YE, Yang XL, and Schimmel P. 2008. Natural homolog of tRNA synthetase editing domain rescues conditional lethality caused by mistranslation. *J. Biol. Chem.*, **283**(44):30073–30078. Link.

Clark MA, *et al.* 1992. The eubacterial endosymbionts of whiteflies (homoptera: Aleyrodoidea) constitute a lineage distinct from the endosymbionts of aphids and mealybugs. *Curr. Microbiol.*, **25**(2):119–123. Link.

Conord C, *et al.* 2008. Long-term Evolutionary Stability of Bacterial Endosymbiosis in Curculionoidea: Additional Evidence of Symbiont Replacement in the Dryophthoridae Family. *Mol. Biol. Evol.*, **25**(5):859–868. Link.

Coombs MT, Costa HS, De Barro P, and Rosell RC. 2007. Pre-Imaginal Egg Maturation and Bacteriocyte Inclusion in *Bemisia aff. gigantea* (Hemiptera: Aleyrodidae). *Ann. Entomol. Soc. Am.*, **100**(5):736–744. Link.

Costa HS, Westcot DM, Ullman DE, and Johnson MW. 1993. Ultrastructure of the endosymbionts of the whitefly, *Bemisia tabaci* and *Trialeurodes vaporariorum*. *Protoplasma*, **176**(3-4):106–115. Link.

Costa HS, Westcot DM, Ullman DE, Rosell R, Brown JK, and Johnson MW. 1995. Morphological variation in *Bemisia* endosymbionts. *Protoplasma*, **189**(3-4):194–202. Link.

Costa HS, Toscano NC, and Henneberry TJ. 1996. Mycetocyte Inclusion in the Oocytes of *Bemisia argentifolii* (Homoptera: Aleyrodidae). *Ann. Entomol. Soc. Am.*, **89**(5). Link.

Cronan JE. 2003. Bacterial membrane lipids: where do we stand? *Annu. Rev. Microbiol.*, **57**(17):203–224. Link.

Cryan JR and Urban JM. 2012. Higher-level phylogeny of the insect order Hemiptera: is Auchenorrhyncha really paraphyletic? *Syst. Entomol.*, **37**(1):7–21. Link.

Currano ED, Wilf P, Wing SL, Labandeira CC, Lovelock EC, and Royer DL. 2008. Sharply increased insect herbivory during the Paleocene-Eocene Thermal Maximum. *Proc. Natl. Acad. Sci. U. S. A.*, **105**(6):1960–1964. Link.

Darling AE, Mau B, and Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*, **5**(6):e11147. Link.

Darling AE, Tritt A, Eisen Ja, and Facciotti MT. 2011. Mauve assembly metrics. *Bioinformatics*, **27**(19):2756–7. Link.

Darriba D, Taboada GL, Doallo R, and Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, **27**(8):1164–1165. Link.

Darriba D, Taboada GL, Doallo R, and Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods*, **9**(8):772. Link.

De Barro PJ, Liu SS, Boykin LM, and Dinsdale AB. 2011. *Bemisia tabaci*: a statement of species status. *Annu. Rev. Entomol.*, **56**:1–19. Link.

de Chaumont F, *et al.* 2012. Icy: an open bioimage informatics platform for extended reproducible research. *Nat. Methods*, **9**(7):690–696. Link.

DeBary A. 1879. *Die Erscheinung der Symbiose*. Strassburg: Verlag von Karl J. Trubner.

Degnan PH and Moran NA. 2008. Diverse phage-encoded toxins in a protective insect endosymbiont. *Appl. Environ. Microbiol.*, **74**(21):6782–6791. Link.

Degnan PH, Yu Y, Sisneros N, Wing Ra, and Moran NA. 2009. *Hamiltonella defensa*, genome evolution of protective bacterial endosymbiont from pathogenic ancestors. *Proc. Natl. Acad. Sci. U. S. A.*, **106**(22):9063–9068. Link.

Delk AS and Dekker CA. 1972. Characterization of rhapidosomes of *Saprospira grandis*. *J. Mol. Biol.*, **64**(1):287–295. Link.

Douglas AE. 1994. *Symbiosis Interactions*. Oxford University Press: Oxford.

Douglas AE. 1998. Nutritional interactions in insect-microbial symbioses: aphids and their symbiotic bacteria Buchnera. *Annu. Rev. Entomol.*, **43**:17–37. Link.

Douglas AE. 2006. Phloem-sap feeding by animals: problems and solutions. *J. Exp. Bot.*, **57**(4):747–754. Link.

Drohojowska J and Szwedo J. 2011. A new whitefly from Lower Cretaceous Lebanese amber (Hemiptera: Sternorrhyncha: Aleyrodidae). *Insect Syst. Evol.*, **42**(2):179–196. Link.

Drohojowska J and Szwedo J. 2013. *Gapenus rhinariatus gen. sp. n. from the Lower Cretaceous amber of Lebanon (Hemiptera: Sternorrhyncha: Aleyrodidae)*. Brill. ISBN 9789004210714. Link.

Drohojowska J and Szwedo J. 2014. Early Cretaceous Aleyrodidae (Hemiptera: Sternorrhyncha) from the Lebanese amber. *Cretac. Res.* Link.

Drummond AJ, Ho SYW, Phillips MJ, and Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol.*, **4**(5):e88. Link.

Duron O. 2013. Lateral transfers of insertion sequences between *Wolbachia*, *Cardinium* and *Rickettsia* bacterial endosymbionts. *Heredity (Edinb).*, **2**(May):1–8. Link.

Eddy SR. 2011. Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**(10):e1002195. Link.

Edlund A, Ek K, Breitholtz M, and Gorokhova E. 2012. Antibiotic-induced change of bacterial communities associated with the copepod *Nitocra spinipes*. *PLoS One*, **7**(3):e33107. Link.

Ellegaard KM, Klasson L, Näslund K, Bourtzis K, and Andersson SGE. 2013. Comparative genomics of *Wolbachia* and the bacterial species concept. *PLoS Genet.*, **9**(4):e1003381. Link.

Engel M and Grimaldi D. 2004. New light shed on the oldest insect. *Nature*, **427**:627–630. Link.

Engel P and Moran NA. 2013. The gut microbiota of insects - diversity in structure and function. *FEMS Microbiol. Rev.*, **37**(5):699–735. Link.

Erickson HP, Anderson DE, and Osawa M. 2010. FtsZ in bacterial cytokinesis: cytoskeleton and force generator all in one. *Microbiol. Mol. Biol. Rev.*, **74**(4):504–528. Link.

Ermolaeva MD. 2001. Synonymous codon usage in bacteria. *Curr. Issues Mol. Biol.*, **3**(4):91–97. Link.

Everett KDE, Thao M, Horn M, Dyszynski GE, and Baumann P. 2005. Novel chlamydiae in whiteflies and scale insects: endosymbionts '*Candidatus* Fritschea bemisiae' strain Falk and '*Candidatus* Fritschea eriococci' strain Elm. *Int. J. Syst. Evol. Microbiol.*, **55**(Pt 4):1581–1587. Link.

Fares MA, Barrio E, Sabater-Muñoz B, and Moya A. 2002. The evolution of the heat-shock protein GroEL from Buchnera, the primary endosymbiont of aphids, is governed by positive selection. *Mol. Biol. Evol.*, **19**(7):1162–1170. Link.

Fares Ma, Moya a, and Barrio E. 2005. Adaptive evolution in GroEL from distantly related endosymbiotic bacteria of insects. *J. Evol. Biol.*, **18**(3):651–560. Link.

Feldhaar H. 2011. Bacterial symbionts as mediators of ecologically important traits of insect hosts. *Ecol. Entomol.*, **36**(5):533–543. Link.

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

Fenton AK and Gerdes K. 2013. Direct interaction of FtsZ and MreB is required for septum synthesis and cell division in *Escherichia coli*. *EMBO J.*, **32**(13):1953–1965. Link.

Ferrari J and Vavre F. 2011. Bacterial symbionts in insects or the story of communities affecting communities. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **366**(1569):1389–1400. Link.

Freilich S, Kreimer A, Meilijson I, Gophna U, Sharan R, and Ruppin E. 2010. The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic Acids Res.*, **38**(12):3857–3868. Link.

Fukatsu T. 2012. Next-generation sequencing sheds light on intricate regulation of insect gut microbiota. *Mol. Ecol.*, **21**(24):5908–5910. Link.

Furusawa G, Yoshikawa T, Yasuda A, and Sakata T. 2003. Algicidal activity and gliding motility of *Saprospira* sp. SS98-5. *Can. J. Microbiol.*, **49**(2):92–100. Link.

Furusawa G, Yoshikawa T, Takano Y, Mise K, Furusawa I, Okuno T, and Sakata T. 2005. Characterization of cytoplasmic fibril structures found in gliding cells of *Saprospira* sp. *Can. J. Microbiol.*, **51**(10):875–880. Link.

Gao F and Zhang CT. 2008. Ori-Finder: a web-based system for finding oriCs in unannotated bacterial genomes. *BMC Bioinformatics*, **9**(1):79. Link.

Gil R, *et al.* 2003. The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. *Proc. Natl. Acad. Sci. U. S. A.*, **100**(16):9388–9393. Link.

Gil R, Silva FJ, Peretó J, and Moya A. 2004. Determination of the core of a minimal bacterial gene set. *Microbiol. Mol. Biol. Rev.*, **68**(3):518–537. Link.

Gil R, *et al.* 2008. Massive presence of insertion sequences in the genome of SOPE, the primary endosymbiont of the rice weevil *Sitophilus oryzae*. *Int Microbiol*, **11**(1):41–48. Link.

Gillespie JJ, *et al.* 2012. A *Rickettsia* genome overrun by mobile genetic elements provides insight into the acquisition of genes characteristic of an obligate intracellular lifestyle. *J. Bacteriol.*, **194**(2):376–394. Link.

Gnankiné O, *et al.* 2013. Distribution of *Bemisia tabaci* (Homoptera: Aleyrodidae) biotypes and their associated symbiotic bacteria on host plants in West Africa. *Insect Conserv. Divers.*, **6**(3):411–421. Link.

Gómez-Valero L, Latorre A, and Silva FJ. 2004. The evolutionary fate of nonfunctional DNA in the bacterial endosymbiont *Buchnera aphidicola*. *Mol. Biol. Evol.*, **21**(11):2172–2181. Link.

Gómez-Valero L, Silva FJ, Christophe Simon J, and Latorre A. 2007. Genome reduction of the aphid endosymbiont *Buchnera aphidicola* in a recent evolutionary time scale. *Gene*, **389**(1):87–95. Link.

Gómez-Valero L, Latorre a, Gil R, Gadau J, Feldhaar H, and Silva FJ. 2008. Patterns and rates of nucleotide substitution, insertion and deletion in the endosymbiont of ants *Blochmannia floridanus*. *Mol. Ecol.*, **17**(19):4382–4392. Link.

Gosalbes MJ, Lamelas A, Moya A, and Latorre A. 2008. The striking case of tryptophan provision in the cedar aphid Cinara cedri. *J. Bacteriol.*, **190**(17):6026–6029. Link.

Gottlieb Y, *et al.* 2006. Identification and localization of a *Rickettsia* sp. in *Bemisia tabaci* (Homoptera: Aleyrodidae). *Appl. Environ. Microbiol.*, **72**(5):3646–3652. Link.

Gottlieb Y, Ghanim M, Gueguen G, Kontsedalov S, Vavre F, Fleury F, and Zchori-Fein E. 2008. Inherited intracellular ecosystem: symbiotic bacteria share bacteriocytes in whiteflies. *FASEB J.*, **22**(7):2591–2599. Link.

Grimaldi D and Engel MS. 2005. *Evolution of the Insects*. Cambridge University Press, New York. Link.

Gueguen G, *et al.* 2010. Endosymbiont metacommunities, mtDNA diversity and the evolution of the *Bemisia tabaci* (Hemiptera: Aleyrodidae) species complex. *Mol. Ecol.*, **19**:4365–4378. Link.

Guo M, Chong YE, Shapiro R, Beebe K, Yang XL, and Schimmel P. 2009. Paradox of mistranslation of serine for alanine caused by AlaRS recognition dilemma. *Nature*, **462**(7274):808–812. Link.

Gupta RS and Lorenzini E. 2007. Phylogeny and molecular signatures (conserved proteins and indels) that are specific for the Bacteroidetes and Chlorobi species. *BMC Evol. Biol.*, **7**:71. Link.

Guy L, Kultima JR, and Andersson SGE. 2010. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics*, **26**(18):2334–2335. Link.

Haft DH, Selengut JD, and White O. 2003. The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**(1):371–373. Link.

Haglund CM, Choe JE, Skau CT, Kovar DR, and Welch MD. 2010. *Rickettsia* Sca2 is a bacterial formin-like mediator of actin-based motility. *Nat. Cell Biol.*, **12**(11):1057–1063. Link.

Han MV and Zmasek CM. 2009. phyloXML: XML for evolutionary biology and comparative genomics. *BMC bioinformatics*, **10**:356. Link.

Hansen AK and Moran NA. 2014. The impact of microbial symbionts on host plant utilization by herbivorous insects. *Mol. Ecol.*, **23**(6):1473–1496.

Hansen AK, Vorburger C, and Moran NA. 2012. Genomic basis of endosymbiont-conferred protection against an insect parasitoid. *Genome Res.*, **22**(1):106–114. Link.

Harrison CP, Douglas AE, and Dixon. 1989. A rapid method to isolate symbiotic bacteria from aphids. *J. Invertebr. Pathol.*, **53**(3):427–428. Link.

Heled J and Drummond AJ. 2012. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Syst. Biol.*, **61**(1):138–149. Link.

Hildebrand F, Meyer A, and Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. *PLoS genetics*, **6**(9):e1001107. Link.

Himler AG, *et al.* 2011. Rapid spread of a bacterial symbiont in an invasive whitefly is driven by fitness benefits and female bias. *Science.*, **332**(6026):254–256. Link.

Ho SYW, Phillips MJ, Cooper A, and Drummond AJ. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol. Biol. Evol.*, **22**(7):1561–1568. Link.

Hooke R. 1665. *Micrographia*. London: The Royal Society.

Horn M, Harzenetter MD, Linner T, Schmid EN, Müller KD, Michel R, and Wagner M. 2001. Members of the Cytophaga-Flavobacterium-Bacteroides phylum as intracellular bacteria of acanthamoebae: proposal of '*Candidatus* Amoebophilus asiaticus'. *Environ. Microbiol.*, **3**(7):440–449. Link.

Hunter S, *et al.* 2012. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**(Database issue):D306–312. Link.

Husnik F, *et al.* 2013. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell*, **153**(7):1567–1578. Link.

Huson DH, Mitra S, Ruscheweyh HJ, Weber N, and Schuster SC. 2011. Integrative analysis of environmental sequences using MEGAN4. *Genome Res.*, **21**(9):1552–1560. Link.

Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, and Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**:119. Link.

Jarrell KF and McBride MJ. 2008. The surprisingly diverse ways that prokaryotes move. *Nat. Rev. Microbiol.*, **6**(6):466–476. Link.

Jiang ZF, *et al.* 2013. Comparison of the genome sequences of "*Candidatus* Portiera aleyrodidarum"primary endosymbionts of the whitefly *Bemisia tabaci* B and Q biotypes. *Appl. Environ. Microbiol.*, **79**(5):1757–1759. Link.

Jones P, *et al.* 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**(9):1236–1240. Link.

Käll L, Krogh A, and Sonnhammer ELL. 2007. Advantages of combined transmembrane topology and signal peptide prediction–the Phobius web server. *Nucleic Acids Res.*, **35**(Web Server issue):W429–432. Link.

Kanehisa M, Goto S, Sato Y, Furumichi M, and Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**(Database issue):D109–114. Link.

Karlsson FH, Ussery DW, Nielsen J, and Nookaew I. 2011. A closer look at bacteroides: phylogenetic relationship and genomic implications of a life in the human gut. *Microb. Ecol.*, **61**(3):473–485. Link.

Karp PD, Paley S, and Romero P. 2002. The Pathway Tools software. *Bioinformatics*, **18 Suppl 1**:S225–232. Link.

Katoh K, Misawa K, Kuma K, and Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**(14):3059–3066. Link.

Kaur SJ, Rahman MS, Ammerman NC, Beier-Sexton M, Ceraul SM, Gillespie JJ, and Azad AF. 2012. TolC-Dependent Secretion of an Ankyrin Repeat-Containing Protein of *Rickettsia typhi*. *J. Bacteriol.*, **194**(18):4920–4932. Link.

Kearns DB. 2007. Microbiology. Bright insight into bacterial gliding. *Science*, **315**(5813):773–774. Link.

Kelkar YD and Ochman H. 2013. Genome reduction promotes increase in protein functional complexity in bacteria. *Genetics*, **193**(1):303–307. Link.

Keseler IM, *et al.* 2013. EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res.*, **41**(Database issue):D605–612. Link.

Kitajima EW, Groot TVM, Novelli VM, Freitas-Astúa J, Alberti G, and de Moraes GJ. 2007. In situ observation of the *Cardinium* symbionts of *Brevipalpus* (Acari: Tenuipalpidae) by electron microscopy. *Exp. Appl. Acarol.*, **42**(4):263–271. Link.

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

Koga R and Moran NA. 2014. Swapping symbionts in spittlebugs: evolutionary replacement of a reduced genome symbiont. *The ISME journal*, **8**(6):1237–1246. Link.

Koga R, Meng XY, Tsuchida T, and Fukatsu T. 2012. Cellular mechanism for selective vertical transmission of an obligate insect symbiont at the bacteriocyte-embryo interface. *Proc. Natl. Acad. Sci. U. S. A.*, **109**(20):1230–1237. Link.

Koga R, Bennett GM, Cryan JR, and Moran NA. 2013. Evolutionary replacement of obligate symbionts in an ancient and diverse insect lineage. *Environ. Microbiol.*, **15**(7):2073–2081. Link.

Koskiniemi S, Lamoureux JG, Nikolakakis KC, t'Kint de Roodenbeke C, Kaplan MD, Low DA, and Hayes CS. 2013. Rhs proteins from diverse bacteria mediate intercellular competition. *Proc. Natl. Acad. Sci. U. S. A.*, **110**(17):7032–7037. Link.

Kreimer A, Doron-Faigenboim A, Borenstein E, and Freilich S. 2012. NetCmpt: a network-based tool for calculating the metabolic competition between bacterial species. *Bioinformatics*, **28**(16):2195–2197. Link.

Krzywinski M, *et al.* 2009. Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**(9):1639–1645. Link.

Kuechler SM, Gibbs G, Burckhardt D, Dettner K, and Hartung V. 2013. Diversity of bacterial endosymbionts and bacteria-host co-evolution in Gondwanan relict moss bugs (Hemiptera: Coleorrhyncha: Peloridiidae). *Environ. Microbiol.*, **15**(7):2031–2042. Link.

Kurtti TJ, Munderloh UG, Andreadis TG, Magnarelli LA, and Mather TN. 1996. Tick cell culture isolation of an intracellular prokaryote from the tick *Ixodes scapularis*. *J. Invertebr. Pathol.*, **67**(3):318–321. Link.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, and Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol.*, **5**(2):R12. Link.

Lamelas A, Gosalbes MJ, Moya A, and Latorre A. 2011a. 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**(13):4446–4454. Link.

Lamelas A, Gosalbes MJ, Manzano-Marín A, Peretó J, Moya A, and Latorre A. 2011b. *Serratia symbiotica* from the aphid *Cinara cedri*: a missing link from facultative to obligate insect endosymbiont. *PLoS Genet.*, **7**(11):e1002357. Link.

Lartillot N, Lepage T, and Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, **25**(17):2286–2288. Link.

Latorre A, Gil R, Silva FJ, and Moya A. 2005. Chromosomal stasis versus plasmid plasticity in aphid endosymbiont *Buchnera aphidicola*. *Heredity (Edinb).*, **95**(5):339–347. Link.

Latorre A, Durbán A, Moya A, and Peretó J. 2011. The role of symbiosis in eukaryotic evolution. In *Origins and Evoution of Life - an Astrobiological Perspective.*, pages 326–433. Cambridge University Press., Cambridge. Link.

Lee W, Park J, Lee GS, Lee S, and Akimoto S. 2013. Taxonomic status of the *Bemisia tabaci* complex (Hemiptera: Aleyrodidae) and reassessment of the number of its constituent species. *PLoS One*, **8**(5):e63817. Link.

Lefèvre C, Charles H, Vallier A, Delobel B, Farrell B, and Heddi A. 2004. Endosymbiont phylogenesis in the dryophthoridae weevils: evidence for bacterial replacement. *Molecular biology and evolution*, **21**(6):965–973. Link.

Li L, Stoeckert CJ, and Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**(9):2178–2189. Link.

Lind PA and Andersson DI. 2008. Whole-genome mutational biases in bacteria. *Proc. Natl. Acad. Sci. U. S. A.*, **105**(46):17878–17883. Link.

Linhartová I, *et al.* 2010. RTX proteins: a highly diverse family secreted by a common mechanism. *FEMS Microbiol. Rev.*, **34**(6):1076–1112. Link.

Lo N, Casiraghi M, Salati E, Bazzocchi C, and Bandi C. 2002. How Many *Wolbachia* Supergroups Exist? *Mol. Biol. Evol.*, **19**(3):341–346. Link.

López-Madrigal S, Latorre A, Porcar M, Moya A, and Gil R. 2013. Mealybugs nested endosymbiosis: going into the 'matryoshka' system in *Planococcus citri* in depth. *BMC Microbiol.*, **13**(1):74. Link.

López-Sánchez MJ, Neef A, Peretó J, Patiño Navarrete R, Pignatelli M, Latorre A, and Moya A. 2009. Evolutionary convergence and nitrogen metabolism in Blattabacterium strain Bge, primary endosymbiont of the cockroach Blattella germanica. *PLoS Genet.*, **5**(11):e1000721. Link.

Maddison WP and Maddison D. 2011. Mesquite: a modular system for evolutionary analysis. Link.

Manzano-Marín A and Latorre A. 2014. Settling Down: The Genome of *Serratia symbiotica* from the Aphid *Cinara tujafilina* Zooms in on the Process of Accommodation to a Cooperative Intracellular Life. *Genome Biol. Evol.*, **6**(7):1683–1698. Link.

Manzano-Marín A, Lamelas A, Moya A, and Latorre A. 2012. Comparative genomics of *Serratia* spp.: two paths towards endosymbiotic life. *PLoS One*, **7**(10):e47274. Link.

Marchler-Bauer A, *et al.* 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.*, **39**(Database issue):D225–229. Link.

Martin BD and Schwab E. 2012. Current Usage of Symbiosis and Associated Terminology. *Int. J. Biol.*, **5**(1):32–45. Link.

Martin J. 2007. Giant whiteflies (Sternorrhyncha, Aleyrodidae): a discussion of their taxonomic and evolutionary significance, with the description of a new species of *Udamoselis* Enderlein from Ecuador. *Tijdschr. voor Entomol.*, **150**(June):13–29. Link.

Martin J and Mound L. 2007. An annotated check list of the world's whiteflies (Insecta: Hemiptera: Aleyrodidae). *Zootaxa*, **1492**:1–84. Link.

Martin J, Hernandez-Suarez E, and Carnero A. 1997. An introduced new species of *Lecanoideus* (Homoptera: Aleyrodidae) established and causing economic impact on the Canary Islands. *J. Nat. Hist.*, **31**(8):1261–1272. Link.

Martin W and Müller M. 1998. The hydrogen hypothesis for the first eukaryote. *Nature*, **392**(6671):37–41. Link.

Matalon Y, Katzir N, Gottlieb Y, Portnoy V, and Zchori-Fein E. 2007. *Cardinium* in *Plagiomerus diaspidis* (Hymenoptera: Encyrtidae). *J. Invertebr. Pathol.*, **96**(2):106–108. Link.

Mauriello EMF, Mouhamar F, Nan B, Ducret A, Dai D, Zusman DR, and Mignot T. 2010a. Bacterial motility complexes require the actin-like protein, MreB and the Ras homologue, MglA. *EMBO J.*, **29**(2):315–326. Link.

Mauriello EMF, Mignot T, Yang Z, and Zusman DR. 2010b. Gliding motility revisited: how do the myxobacteria move without flagella? *Microbiol. Mol. Biol. Rev.*, **74**(2):229–49. Link.

McBride MJ. 2004. Cytophaga-flavobacterium gliding motility. *J. Mol. Microbiol. Biotechnol.*, **7**(1-2):63–71. Link.

McBride MJ and Zhu Y. 2013. Gliding motility and Por secretion system genes are widespread among members of the phylum bacteroidetes. *J. Bacteriol.*, **195**(2):270–278. Link.

McBride MJ, *et al.* 2009. Novel features of the polysaccharide-digesting gliding bacterium *Flavobacterium johnsoniae* as revealed by genome sequence analysis. *Appl. Environ. Microbiol.*, **75**(21):6864–6875. Link.

McCutcheon JP and Keeling PJ. 2014. Endosymbiosis: protein targeting further erodes the organelle/symbiont distinction. *Curr. Biol.*, **24**(14):R654–655. Link.

McCutcheon JP and Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.*, **10**(1):13–26. Link.

McCutcheon JP, McDonald BR, and Moran NA. 2009. Convergent evolution of metabolic roles in bacterial co-symbionts of insects. *Proc. Natl. Acad. Sci. U. S. A.*, **106**(36):15394–15399. Link.

McFall-Ngai M. 2008. Are biologists in ’future shock’? Symbiosis integrates biology across domains. *Nat Rev Micro*, **6**(10):789–792. Link.

Metzker ML. 2010. Sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**(1):31–46. Link.

Michie KA and Löwe J. 2006. Dynamic filaments of the bacterial cytoskeleton. *Annu. Rev. Biochem.*, **75**:467–492. Link.

Mignot T, Shaevitz JW, Hartzell PL, and Zusman DR. 2007. Evidence that focal adhesion complexes power bacterial gliding motility. *Science*, **315**(5813):853–856. Link.

Miller JR, Koren S, and Sutton G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics*, **95**(6):315–327. Link.

Minkin I, Patel A, Kolmogorov M, Vyahhi N, and Pham S. 2013. Sibelia: A Scalable and Comprehensive Synteny Block Generation Tool for Closely Related Microbial Genomes. In Darling A and Stoye J, editors, *Algorithms in Bioinformatics.*, volume 8126 of *Lecture Notes in Computer Science*, pages 215–229. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 978-3-642-40452-8. Link.

Montllor CB, Maxmen A, and Purcell AH. 2002. Facultative bacterial endosymbionts benefit pea aphids *Acyrthosiphon pisum* under heat stress. *Ecol. Entomol.*, **27**(2):189–195. Link.

Morag N, Klement E, Saroya Y, Lensky I, and Gottlieb Y. 2012. Prevalence of the symbiont *Cardinium* in *Culicoides* (Diptera: Ceratopogonidae) vector species is associated with land surface temperature. *FASEB J.*, **26**(10):4025–4034. Link.

Moran NA. 1996. Accelerated evolution and Muller’s rachet in endosymbiotic bacteria. *Proc. Natl. Acad. Sci. U. S. A.*, **93**(7):2873–2878. Link.

Moran NA and Bennett GM. 2014. The Tiniest Tiny Genomes. *Annu. Rev. Microbiol.*, **68**(May):195–215. Link.

Moran NA, Russell JA, Koga R, and Fukatsu T. 2005. Evolutionary relationships of three new species of Enterobacteriaceae living as symbionts of aphids and other insects. *Appl. Environ. Microbiol.*, **71**(6):3302–3310. Link.

Moran NA, McCutcheon JP, and Nakabachi A. 2008. Genomics and evolution of heritable bacterial symbionts. *Annu. Rev. Genet.*, **42**:165–190. Link.

Moran NA, McLaughlin HJ, and Sorek R. 2009. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science*, **323**(5912):379–382. Link.

Moreira D and Lopez-Garcia P. 1998. Symbiosis between methanogenic archaea and delta-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. *J. Mol. Evol.*, **47**(5):517–530. Link.

Morimoto S, Kurtti TJ, and Noda H. 2006. In vitro cultivation and antibiotic susceptibility of a Cytophaga-like intracellular symbiote isolated from the tick *Ixodes scapularis*. *Curr. Microbiol.*, **52**(4):324–329. Link.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, and Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.*, **35**(Web Server issue):W182–185. Link.

Moya A, Peretó J, Gil R, and Latorre A. 2008. Learning how to live together: genomic insights into prokaryote-animal symbioses. *Nat. Rev. Genet.*, **9**(3):218–229. Link.

Nakabachi A, *et al.* 2013. Defensive bacteriome symbiont with a drastically reduced genome. *Curr. Biol.*, **23**(15):1478–1484. Link.

Nakabachi A, Ishida K, Hongoh Y, Ohkuma M, and Miyagishima S. 2014. Aphid gene of bacterial origin encodes a protein transported to an obligate endosymbiont. *Curr. Biol.*, **24**(14):R640–R641. Link.

Nakagawa S, *et al.* 2014. Allying with armored snails: the complete genome of gammaproteobacterial endosymbiont. *ISME J.*, **8**(1):40–51. Link.

Nakamura Y, *et al.* 2009. Prevalence of *Cardinium* bacteria in planthoppers and spider mites and taxonomic revision of "*Candidatus* Cardinium hertigii"based on detection of a new *Cardinium* group from biting midges. *Appl. Environ. Microbiol.*, **75**(21):6757–6763. Link.

Nakamura Y, Gotoh T, Imanishi S, Mita K, Kurtti TJ, and Noda H. 2011. Differentially expressed genes in silkworm cell cultures in response to infection by *Wolbachia* and *Cardinium* endosymbionts. *Insect Mol. Biol.*, **20**(3):279–289. Link.

Nakane D, Sato K, Wada H, McBride MJ, and Nakayama K. 2013. Helical flow of surface protein required for bacterial gliding motility. *Proc. Natl. Acad. Sci. U. S. A.*, **110**(27):11145–11150. Link.

Nan B and Zusman DRD. 2011. Uncovering the mystery of gliding motility in the myxobacteria. *Annu. Rev. Genet.*, **45**(58):21–39. Link.

Nan B, Bandaria JN, Moghtaderi A, Sun I, Yildiz A, and Zusman DR. 2013. Flagella stator homologs function as motors for myxobacterial gliding motility by moving in helical trajectories. *Proc. Natl. Acad. Sci. U. S. A.*, **110**(16):E1508–513. Link.

Navas-Castillo J, Fiallo-Olivé E, and Sánchez-Campos S. 2011. Emerging virus diseases transmitted by whiteflies. *Annu. Rev. Phytopathol.*, **49**:219–248. Link.

Nawrocki EP. 2009. *Structural RNA Homology Search and Alignment Using Covariance Models*. Ph.D. thesis, Washington University School of Medicine. Link.

Oakeson KF, *et al.* 2014. Genome Degeneration and Adaptation in a Nascent Stage of Symbiosis. *Genome Biol. Evol.*, **6**(1):76–93. Link.

Oliver KM, Russell JA, Moran NA, and Hunter MS. 2003. Facultative bacterial symbionts in aphids confer resistance to parasitic wasps. *Proc. Natl. Acad. Sci. U. S. A.*, **100**(4):1803–1807. Link.

Oliver KM, Campos J, Moran NA, and Hunter MS. 2008. Population dynamics of defensive symbionts in aphids. *Proc. Biol. Sci.*, **275**(1632):293–299. Link.

Oliver KM, Degnan PH, Burke GR, and Moran NA. 2010. Facultative symbionts in aphids and the horizontal transfer of ecologically important traits. *Annu. Rev. Entomol.*, **55**:247–266. Link.

Paradis E, Claude J, and Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, **20**(2):289–290. Link.

Parkhill J, *et al.* 2003. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat. Genet.*, **35**(1):32–40. Link.

Parsons JB and Rock CO. 2013. Bacterial lipids: metabolism and membrane homeostasis. *Prog. Lipid Res.*, **52**:249–276. Link.

Patiño-Navarrete R, Moya A, Latorre A, and Peretó J. 2013. Comparative Genomics of *Blattabacterium cuenoti*: The Frozen Legacy of an Ancient Endosymbiont Genome. *Genome Biol. Evol.*, **5**(2):351–361. Link.

Penz T, *et al.* 2012. Comparative genomics suggests an independent origin of cytoplasmic incompatibility in *Cardinium hertigii*. *PLoS Genet.*, **8**(10):e1003012. Link.

Pérez-Brocal V, *et al.* 2006. A small microbial genome: the end of a long symbiotic relationship? *Science*, **314**(5797):312–313. Link.

Petersen L, Bollback JP, Dimmic M, Hubisz M, and Nielsen R. 2007. Genes under positive selection in *Escherichia coli*. *Genome Res.*, **17**(9):1336–1343. Link.

Petersen TN, Brunak Sr, von Heijne G, and Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, **8**(10):785–786. Link.

Petroni G, Spring S, Schleifer KH, Verni F, and Rosati G. 2000. Defensive extrusive ectosymbionts of *Euplotidium* (Ciliophora) that contain microtubule-like structures are bacteria related to Verrucomicrobia. *Proc. Natl. Acad. Sci. U. S. A.*, **97**(4):1813–1817. Link.

Poliakov A, Russell CW, Ponnala L, Hoops HJ, Sun Q, Douglas AE, and van Wijk KJ. 2011. Large-scale label-free quantitative proteomics of the pea aphid-*Buchnera* symbiosis. *Mol. Cell. Proteomics*, **10**(6):M110.007039. Link.

Pop M. 2009. Genome assembly reborn: recent computational challenges. *Brief. Bioinform.*, **10**(4):354–366. Link.

Preer JR, Preer LB, and Jurand A. 1974. Kappa and other endosymbionts in *Paramecium aurelia*. *Microbiol. Mol. Biol. Rev.*, **38**(2):113–163. Link.

Price DRG, Duncan RP, Shigenobu S, and Wilson ACC. 2011. Genome expansion and differential expression of amino acid transporters at the aphid/*Buchnera* symbiotic interface. *Mol. Biol. Evol.*, **28**(11):3113–3126. Link.

Punta M, *et al.* 2012. The Pfam protein families database. *Nucleic Acids Res.*, **40**(Database issue):D290–301. Link.

R Core Team 2014. 2014. *R: A Language and Environment for Statistical Computing*. Link.

Rader BA, Kremer N, Apicella MA, Goldman WE, and McFall-Ngai MJ. 2012. Modulation of symbiont lipid A signaling by host alkaline phosphatases in the squid-vibrio symbiosis. *MBio*, **3**(3). Link.

Rao Q, Wang S, Su YL, Bing XL, Liu SS, and Wang XW. 2012. Draft Genome Sequence of "*Candidatus* Hamiltonella defensa", an Endosymbiont of the Whitefly *Bemisia tabaci*. *J. Bacteriol.*, **194**(13):3558. Link.

Reyes-Prieto M, Latorre A, and Moya A. 2014. Scanty microbes, the 'symbionelle' concept. *Environ. Microbiol.*, **16**(2):335–358. Link.

Rispe C, Delmotte F, van Ham RCHJ, and Moya A. 2004. Mutational and selective pressures on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids. *Genome Res.*, **14**(1):44–53. Link.

Rocha EPC and Danchin A. 2002. Base composition bias might result from competition for metabolic resources. *Trends Genet.*, **18**(6):291–294. Link.

Russell CW, Bouvaine S, Newell PD, and Douglas AE. 2013. Shared metabolic pathways in a coevolved insect-bacterial symbiosis. *Appl. Environ. Microbiol.*, **79**(19):6117–6123. Link.

Russell JA, Latorre A, Sabater-Munoz B, Moya A, and Moran NA. 2003. Side-stepping secondary symbionts: widespread horizontal transfer across and beyond the Aphidoidea. *Molecular Ecology*, **12**(4):1061–1075. Link.

Russell LM. 1965. A New Species of *Aleurodicus* Douglas and Two Close Relatives (Homoptera: Aleyrodidae). *Florida Entomol.*, **48**(1):47. Link.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, and Barrell B. 2000. Artemis: sequence visualization and annotation. *Bioinformatics*, **16**(10):944–945. Link.

Sabree ZL, Huang CY, Okusu A, Moran NA, and Normark BB. 2013. The nutrient supplying capabilities of *Uzinura*, an endosymbiont of armoured scale insects. *Environ. Microbiol.*, **15**(7):1988–1999. Link.

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

Sagan L. 1967. On the origin of mitosing cells. *J. Theor. Biol.*, **14**(3):255–274.

Salje J, van den Ent F, de Boer P, and Löwe J. 2011. Direct membrane binding by bacterial actin MreB. *Mol. Cell*, **43**(3):478–487. Link.

Sandström J and 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**(11):947–955. Link.

Santos-Garcia D, *et al.* 2012. Complete genome sequence of "*Candidatus* Portiera aleyrodidarum"BT-QVLC, an obligate symbiont that supplies amino acids and carotenoids to *Bemisia tabaci*. *J. Bacteriol.*, **194**(23):6654–6655. Link.

Santos-Garcia D, Silva FJ, Moya A, and Latorre A. 2014a. No exception to the rule: *Candidatus* Portiera aleyrodidarum cell wall revisited. *FEMS Microbiol. Lett.*, page Article first published online: 19 SEP. Link.

Santos-Garcia D, *et al.* 2014b. Small but powerful, the primary endosymbiont of moss bugs, *Candidatus* Evansia muelleri, holds a reduced genome with large biosynthetic capabilities. *Genome Biol Evol*, **6**(7):1875–1893. Link.

Santos-Garcia D, *et al.* 2014c. The genome of *Cardinium* cBtQ1 provides insights into genome reduction, symbiont motility and its settlement in *Bemisia tabaci*. *Genome Biol Evol*, **6**(4):1013–1030. Link.

Sapp J. 1994. *Evolution by Association: A history of Symbiosis*. Oxford University Press: New York. ISBN 9780195088212.

Sarris PF, Ladoukakis ED, Panopoulos NJ, and Scoulica EV. 2014. A Phage Tail-Derived Element with Wide Distribution among Both Prokaryotic Domains: A Comparative Genomic and Phylogenetic Study. *Genome Biol. Evol.*, **6**(7):1739–1747. Link.

Sato K, *et al.* 2010. A protein secretion system linked to bacteroidete gliding motility and pathogenesis. *Proc. Natl. Acad. Sci. U. S. A.*, **107**(1):276–281. Link.

Saw JHW, Yuryev A, Kanbe M, Hou S, Young AG, Aizawa SI, and Alam M. 2012. Complete genome sequencing and analysis of *Saprospira grandis* str. Lewin, a predatory marine bacterium. *Stand. Genomic Sci.*, **6**(1):84–93. Link.

Scarborough CL, Ferrari J, and Godfray HCJ. 2005. Aphid protected from pathogen by endosymbiont. *Science*, **310**(5755):1781. Link.

Schattner P, Brooks AN, and Lowe TM. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.*, **33**(Web Server issue):W686–689. Link.

Schindelin J, *et al.* 2012. Fiji: an open-source platform for biological-image analysis. *Nat. Methods*, **9**(7):676–682. Link.

Schlee D. 1970. Verwandtschaftsforschung an fossilen und rezenten Aleyrodina (Insecta; Hemiptera). *Stuttgarter Beiträge zur Naturkd.*, **213**:1–72. Link.

Schmitz-Esser S, Penz T, Spang A, and Horn M. 2011. A bacterial genome in transition - an exceptional enrichment of IS elements but lack of evidence for recent transposition in the symbiont *Amoebophilus asiaticus*. *BMC Evol. Biol.*, **11**:270. Link.

Shcherbakov D. 2000. The most primitive whiteflies (Hemiptera; Aleyrodidae; Bernaeinae subfam. nov.) from the Mesozoic of Asia and Burmese amber, with an overview of Burmese amber hemipterans. *Bull. nat. Hist. Mus. Lond. (Geol.)*, **56**(June):29–37. Link.

Shendure J and Ji H. 2008. Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**(10):1135–1145. Link.

Shi G, *et al.* 2012. Age constraint on Burmese amber based on U-Pb dating of zircons. *Cretac. Res.*, **37**:155–163. Link.

Shigenobu S and Wilson ACC. 2011. Genomic revelations of a mutualism: the pea aphid and its obligate bacterial symbiont. *Cell. Mol. Life Sci.*, **68**(8):1297–1309. Link.

Shigenobu S, Watanabe H, Hattori M, Sakaki Y, and Ishikawa H. 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature*, **407**(6800):81–86. Link.

Shikuma NJ, Pilhofer M, Weiss GL, Hadfield MG, Jensen GJ, and Newman DK. 2014. Marine tubeworm metamorphosis induced by arrays of bacterial phage tail-like structures. *Science*, **343**(6170):529–533. Link.

Shrivastava A, Johnston JJ, van Baaren JM, and McBride MJ. 2013. *Flavobacterium johnsoniae* GldK, GldL, GldM, and SprA are required for secretion of the cell surface gliding motility adhesins SprB and RemA. *J. Bacteriol.*, **195**(14):3201–3212. Link.

Sibley LD. 2004. Intracellular Parasite Invasion Strategies. *Science*, **304**(5668):248–253. Link.

Sibley LD, Hâkansson S, and Carruthers VB. 1998. Gliding motility: an efficient mechanism for cell penetration. *Curr. Biol.*, **8**(1):R12–14. Link.

Silva FJ, Latorre A, and Moya A. 2001. Genome size reduction through multiple events of gene disintegration in *Buchnera* APS. *Trends Genet.*, **17**(11):615–618. Link.

Silva FJ, Latorre A, and Moya A. 2003. Why are the genomes of endosymbiotic bacteria so stable? *Trends Genet.*, **19**(4):176–180. Link.

Singh ST, *et al.* 2012. Diversity and phylogenetic analysis of endosymbiotic bacteria from field caught *Bemisia tabaci* from different locations of North India based on 16S rDNA library screening. *Infect. Genet. Evol.*, **12**(2):411–419. Link.

Skaljac M, Zanic K, Ban SG, Kontsedalov S, and Ghanim M. 2010. Co-infection and localization of secondary symbionts in two whitefly species. *BMC Microbiol.*, **10**:142. Link.

Skaljac M, Zanić K, Hrnčić S, Radonjić S, Perović T, and Ghanim M. 2012. Diversity and localization of bacterial symbionts in three whitefly species (Hemiptera: Aleyrodidae) from the east coast of the Adriatic Sea. *Bull. Entomol. Res.*, **103**(May):1–12. Link.

Sloan DB and Moran NA. 2012a. Endosymbiotic bacteria as a source of carotenoids in whiteflies. *Biology letters*, **8**(6):986–989. Link.

Sloan DB and Moran NA. 2012b. Genome reduction and co-evolution between the primary and secondary bacterial symbionts of psyllids. *Mol. Biol. Evol.*, **29**(12):3781–3792. Link.

Sloan DB and Moran NA. 2013. The evolution of genomic instability in the obligate endosymbionts of whiteflies. *Genome Biol. Evol.*, **5**(5):783–793. Link.

Sloan DB, Nakabachi A, Richards S, Qu J, Murali SC, Gibbs RA, and Moran NA. 2014. Parallel Histories of Horizontal Gene Transfer Facilitated Extreme Reduction of Endosymbiont Genomes in Sap-Feeding Insects. *Mol. Biol. Evol.*, **31**(4):857–871. Link.

Sonnhammer ELL, von Heijne G, and Krogh A. 1998. A Hidden Markov Model for Predicting Transmembrane Helices in Protein Sequences. In *Proc. 6th Int. Conf. Intell. Syst. Mol. Biol.*, ISMB '98, pages 175–182. AAAI Press. ISBN 1-57735-053-7. Link.

Spalding MD and Prigge ST. 2010. Lipoic acid metabolism in microbial pathogens. *Microbiol. Mol. Biol. Rev.*, **74**(2):200–228. Link.

Spormann AM. 1999. Gliding motility in bacteria: insights from studies of *Myxococcus xanthus*. *Microbiol. Mol. Biol. Rev.*, **63**(3):621–641. Link.

Stackebrandt E and Ebers J. 2006. Taxonomic parameters revisited: tarnished gold standards. *Microbiol. Today*, **4**(33):152–155.

Staden R, Beal KF, and Bonfield JK. 2000. The Staden package, 1998. *Methods Mol. Biol.*, **132**:115–130. Link.

Stansly PA and Mckenzie CL. 2007. Fourth International *Bemisia* Workshop International Whitefly Genomics Workshop. *J. Insect Sci.*, **8**(4):53pp. Link.

Stansly PA and Naranjo SE, editors. 2010. *Bemisia: Bionomics and Management of a Global Pest*. Springer, 2nd edition. Link.

Stewart FJ, Newton ILG, and Cavanaugh CM. 2005. Chemosynthetic endosymbioses: adaptations to oxic-anoxic interfaces. *Trends Microbiol.*, **13**(9):439–448. Link.

Su Q, *et al.* 2014. The endosymbiont *Hamiltonella* increases the growth rate of its host *Bemisia tabaci* during periods of nutritional stress. *PLoS One*, **9**(2):e89002. Link.

Sun M, Wartel M, Cascales E, Shaevitz JW, and Mignot T. 2011. Motor-driven intracellular transport powers bacterial gliding motility. *Proc. Natl. Acad. Sci. U. S. A.*, **108**(18):7559–7564. Link.

Szklarzewicz T and Moskal A. 2001. Ultrastructure, distribution, and transmission of endosymbionts in the whiteflyAleurochiton aceris Modeer (Insecta, Hemiptera, Aleyrodinea). *Protoplasma*, pages 45–53. Link.

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

Tamas I, *et al.* 2002. 50 Million Years of Genomic Stasis in Endosymbiotic Bacteria. *Science*, **296**(5577):2376–2379. Link.

Tatsuta T. 2009. Protein quality control in mitochondria. *J. Biochem.*, **146**(4):455–461. Link.

Tatusov RL, *et al.* 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**(1):41. Link.

Terraz G, Gueguen G, Arnó J, Fleury F, and Mouton L. 2014. Nuclear and cytoplasmic differentiation among Mediterranean populations of *Bemisia tabaci*: testing the biological relevance of cytotypes. *Pest Manag. Sci.*, **70**:1503–1513. Link.

Thanedar S and Margolin W. 2004. FtsZ exhibits rapid movement and oscillation waves in helix-like patterns in *Escherichia coli*. *Curr. Biol.*, **14**:1167–1173. Link.

Thao M and Baumann P. 2004a. Evolutionary relationships of primary prokaryotic endosymbionts of whiteflies and their hosts. *Appl. Environ. Microbiol.*, **70**(6):3401. Link.

Thao ML and Baumann P. 2004b. Evidence for multiple acquisition of *Arsenophonus* by whitefly species (Sternorrhyncha: Aleyrodidae). *Curr. Microbiol.*, **48**(2):140–144. Link.

Thao ML, Moran NA, Abbot P, Brennan EB, Burckhardt DH, and Baumann P. 2000. Cospeciation of psyllids and their primary prokaryotic endosymbionts. *Appl. Environ. Microbiol.*, **66**:2898–2905. Link.

Thao ML, Baumann L, Hess JM, Falk BW, Ng JCK, Gullan PJ, and Baumann P. 2003. Phylogenetic evidence for two new insect-associated Chlamydia of the family Simkaniaceae. *Curr. Microbiol.*, **47**(1):46–50. Link.

The UniProt Consortium. 2012. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**(Database issue):D71–75. Link.

Thierry M, Becker N, Hajri A, Reynaud B, Lett JM, and Delatte H. 2011. Symbiont diversity and non-random hybridization among indigenous (Ms) and invasive (B) biotypes of *Bemisia tabaci*. *Mol. Ecol.*, **20**(10):2172–2187. Link.

Timmis JN, Ayliffe MA, Huang CY, and Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.*, **5**(2):123–135. Link.

Toft C and Andersson SGE. 2010. Evolutionary microbial genomics: insights into bacterial host adaptation. *Nat. Rev. Genet.*, **11**(7):465–475. Link.

Tsuchida T, Koga R, Matsumoto S, and Fukatsu T. 2011. Interspecific symbiont transfection confers a novel ecological trait to the recipient insect. *Biol. Lett.*, **7**(2):245–248. Link.

Valmalette JC, Dombrovsky A, Brat P, Mertz C, Capovilla M, and Robichon A. 2012. Light-induced electron transfer and ATP synthesis in a carotene synthesizing insect. *Sci. Rep.*, **2**:1–8. Link.

Van Domselaar GH, *et al.* 2005. BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Res.*, **33**(Web Server issue):W455–459. Link.

Varani AM, Siguier P, Gourbeyre E, Charneau V, and Chandler M. 2011. ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biol.*, **12**(3):R30. Link.

Varma A and Young KD. 2009. In *Escherichia coli*, MreB and FtsZ direct the synthesis of lateral cell wall via independent pathways that require PBP 2. *J. Bacteriol.*, **191**(11):3526–3533. Link.

Vellozo AF, *et al.* 2011. CycADS: an annotation database system to ease the development and update of BioCyc databases. *Database (Oxford).*, **2011**:bar008. Link.

Waku Y and Endo Y. 1987. Ultrastructure and life cycle of the symbionts in a homopteran insect, *Anomoneura mori* Schwartz (Psyllidae). *Appl. Entomol. Zool. (Jpn.)*, **22**:630–637. Link.

Walsh PS, Metzger DA, and Higuchi R. 1991. Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. *Biotechniques*, **10**(4):506–513. Link.

Warnes GR, *et al.* 2013. *gplots: Various R programming tools for plotting data*. Link.

White JA, Kelly SE, Perlman SJ, and Hunter MS. 2009. Cytoplasmic incompatibility in the parasitic wasp *Encarsia inaron*: disentangling the roles of *Cardinium* and *Wolbachia* symbionts. *Heredity (Edinb).*, **102**(5):483–9. Link.

White JA, Kelly SE, Cockburn SN, Perlman SJ, and Hunter MS. 2011. Endosymbiont costs and benefits in a parasitoid infected with both *Wolbachia* and *Cardinium*. *Heredity (Edinb).*, **106**(4):585–591. Link.

Williams RAM, Westrop GD, and Coombs GH. 2009. Two pathways for cysteine biosynthesis in *Leishmania major*. *Biochem. J.*, **420**(3):451–462. Link.

Wilson SCC, *et al.* 2010. Genomic insight into the amino acid relations of the pea aphid, *Acyrthosiphon pisum*, with its symbiotic bacterium *Buchnera aphidicola*. *Insect Mol. Biol.*, **19 Suppl 2**:249–258. Link.

Wootton RJ. 1981. Palaeozoic Insects. *Annu. Rev. Entomol.*, **26**(1):319–344. Link.

Wu D, *et al.* 2006. Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. *PLoS Biol.*, **4**(6):e188. Link.

Wu M, *et al.* 2004. Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: a streamlined genome overrun by mobile genetic elements. *PLoS Biol.*, **2**(3):E69. Link.

Xie G, *et al.* 2007. Genome sequence of the cellulolytic gliding bacterium *Cytophaga hutchinsonii*. *Appl. Environ. Microbiol.*, **73**(11):3536–3546. Link.

Xie W, *et al.* 2012. Pyrosequencing the *Bemisia tabaci* Transcriptome Reveals a Highly Diverse Bacterial Community and a Robust System for Insecticide Resistance. *PLoS One*, **7**(4):e35181. Link.

Yamamoto T. 1967. Presence of rhapidosomes in various species of bacteria and their morphological characteristics. *J. Bacteriol.*, **94**(5):1746–1756. Link.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**(8):1586–1591. Link.

Yang Z, Wong WSW, and Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.*, **22**(4):1107–1118. Link.

Yoshikawa T, Nakahara M, Tabata A, Kokumai S, Furusawa G, and Sakata T. 2008. Characterization and expression of *Saprospira* cytoplasmic fibril protein (SCFP) gene from algicidal *Saprospira* spp. strains. *Fish. Sci.*, **74**(5):1109–1117. Link.

Yu XC. 1998. Inhibition of Assembly of Bacterial Cell Division Protein FtsZ by the Hydrophobic Dye 5,5'-Bis-(8-anilino-1-naphthalenesulfonate). *J. Biol. Chem.*, **273**(17):10216–10222. Link.

Zchori-Fein E and Brown JK. 2002. Diversity of Prokaryotes Associated with *Bemisia tabaci* (Gennadius) (Hemiptera: Aleyrodidae). *Ann. Entomol. Soc. Am.*, **95**(6):711–718. Link.

Zchori-Fein E and Perlman SJ. 2004. Distribution of the bacterial symbiont *Cardinium* in arthropods. *Mol. Ecol.*, **13**(7):2009–2016. Link.

Zchori-Fein E, Gottlieb Y, Kelly SE, Brown JK, Wilson JM, Karr TL, and Hunter MS. 2001. A newly discovered bacterium associated with parthenogenesis and a change in host selection behavior in parasitoid wasps. *Proc. Natl. Acad. Sci. U. S. A.*, **98**(22):12555–12560. Link.

Zchori-Fein E, Perlman SJ, Kelly SE, Katzir N, and Hunter MS. 2004. Characterization of a 'Bacteroidetes' symbiont in *Encarsia* wasps (Hymenoptera: Aphelinidae): proposal of '*Candidatus* Cardinium hertigii'. *Int. J. Syst. Evol. Microbiol.*, **54**(Pt 3):961–968. Link.

Zchori-Fein E, Lahav T, and Freilich S. 2014. Variations in the identity and complexity of endosymbiont combinations in whitefly hosts. *Front. Microbiol.*, **5**(July):1–8. Link.

Zhang J, Nielsen R, and Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.*, **22**(12):2472–2479. Link.

Zhou Y, Liang Y, Lynch KH, Dennis JJ, and Wishart DS. 2011. PHAST: a fast phage search tool. *Nucleic Acids Res.*, **39**(Web Server issue):W347–352. Link.

Part 7

Appendix

“Politicians hide themselves away
They only started the war
Why should they go out to fight?
They leave that role to the poor, yeah”

Black Sabbath

7.1 List of Figures, Tables and Acronyms

List of Figures

1.1.1 Types of symbiosis. 4
1.1.2 Origins of the eukaryotic cells 5
1.1.3 Symbiosis distribution 7
1.1.4 Genome reduction process 8

1.2.1 Human louse bacteriome 11
1.2.2 P-endosymbionts metabolic capabilities 14

1.3.1 Hemiptera simplified phylogeny 19
1.3.2 Whiteflies life cycle 20

1.4.1 *B. tabaci* 25
1.4.2 *B. tabaci* species tree 26
1.4.3 Facultative Endosymbiont Combinations 27
1.4.4 *T. vaporariorum* 27
1.4.5 *A. dispersus* 27
1.4.6 *A. floccisimus* 28

3.6.1 Assembly pipeline 1 45
3.6.2 Assembly pipeline 2 46
3.6.3 Assembly pipeline 3 47

4.1.1 Whole nymph FISH 59
4.1.2 Nymph bacteriome FISH focus 61
4.1.3 Bacteriome renderization 62
4.1.4 Endosymbionts ultrastructure 62
4.1.5 *Portiera* no cell wall 64
4.1.6 *Portiera* cell wall 65
4.1.7 *Portiera* BT-QVLC genome 69
4.1.8 *Portiera* **COG** comparison 71
4.1.9 Basic cell machinery 72
4.1.10 *Portiera* strains from *B. tabaci* 77
4.1.11 *Hamiltonella* synteny 80
4.1.12 *Hamiltonella* **COG** assignment 82
4.1.13 Metabolism integration 84

4.2.1 *Cardinium* plasmids 92
4.2.2 *Cardinium* strains phylogeny 94
4.2.3 Bacteroidetes phylogenomics 95
4.2.4 Redundancy plot 96

4.2.5 *Cardinium* cBtQ1 IS 97
4.2.6 Segmental duplications 98
4.2.7 *Cardinium* strains synteny 99
4.2.8 *Cardinium* Euler diagram 100
4.2.9 Bacteroidetes **COG** clustering 102
4.2.10 *Cardinium* LCAs comparisons 105
4.2.11 *Cardinium* metabolism 107
4.2.12 *Cardinium*'s pattern 110
4.2.13 MLCs in other Bacteria 114
4.2.14 *Cardinium*'s gliding machinery 116

4.3.1 Whiteflies diversification 124
4.3.2 *Portiera* strains phylogeny 125
4.3.3 *Portiera* strains genomes 127
4.3.4 *Portiera* strains **COG** comparisons 129
4.3.5 *Portiera* strains synteny 131
4.3.6 *Portiera* strains metabolism blueprints 135
4.3.7 BEAST2 *Portiera* tree 140
4.3.8 BEAST2 hosts tree 142
4.3.9 *Portiera* raw data 146
4.3.10 *Portiera* clean data 147
4.3.11 Genomic dN/dS 150

List of Tables

1.3.1 Whiteflies' endosymbionts 22

3.2.1 FISH probes 37

4.1.1 Simplified membrane biosynthesis capabilities. 63
4.1.2 *Portiera* membrane measurements in nm. 66
4.1.3 Genomic features of *Portiera* BT-QVLC compared to other P-endosymbionts. 69
4.1.4 *Hamiltonella* strains assemblies statistics. 78
4.1.5 *Hamiltonella* strains general genomic features. 78

4.2.1 *Cardinium* genome 91
4.2.2 *B. tabaci* QHC-VLC endosymbionts' pairwise EMO scores. 108

4.3.1 *Portiera* General features 126
4.3.2 *Portiera* lost genes 131
4.3.3 *Portiera* divergence 141
4.3.4 mtCOI dating 144

4.3.5 *Portiera* positive selection 151

List of Acronyms

A	Adenine	9
Afp	Anti-feeding Prophage	99
Arg	Arginine	85
Asp	Aspartate	85
ATP	Adenosine triphosphate	10
B1	Thiamin diphosphate	86
B2	Riboflavin	86
B5	(R)-pantothenate	86
B6	Pyridoxal 5’-phosphate	86
B7	Biotin	86
BCA	Branched-chain amino acids	85
BEB	Bayes Empirical Bayes	151
bp	base pairs	44
C	Cytosine	9
ca.	*circa*	10
Ca.	*Candidatus*	9
CDS	Coding DNA Sequence	48
Chsm	Chorismate	87
CI	Cytoplasmic Incompatibility	16
CoA	Co-enzyme A	86
COG	Cluster of Orthologous Categories	48

CTP	Cytosine triphosphate	10
dN	number of non-synonymous substitutions per non-synonymous site	54
dS	number of synonymous substitutions per synonymous site	54
DTT	Dithiothreitol	38
E4P	D-erythrose 4-phosphate	83
EC number	European Community number	51
EDTA	Edetic acid	36
EMO	Effective Metabolic Overlap	51
ES	Electron-dense Structure	115
FAD	Flavin adenine dinucleotide	86
FEP	Fibrous Electron-dense Plaque	114
FISH	Fluorescent In Situ Hybridization	36
FMN	Flavin mononucleotide	86
FPP	(2e,6E)-farnesyl diphosphate	84
G	Guanine	9
GAP	D-glyceraldehyde 3-phosphate	84
gDNA	genomic DNA	38
GGPP	Geranylgeranyl diphosphate	86
Glu	Glutamate	86
GO	Gene Ontology	48
GPP	Geranyl diphosphate	84
GTP	Guanosine triphosphate	10
h	hour	35
Hcy	Homocysteine	85

HGT	Horizontal Gene Transfer	9
His	Histidine	85
HPD	Highest Posterior Density	139
IGR	Intergenic Region	70
Ile	Isoleucine	85
indel	insertion and/or deletion	9
IS	Insertion Sequences	8
kb	kilobase pairs	70
kDa	KiloDalton	119
LCA	Last Common Ancestor	50
Leu	Leucine	85
LTE	Low TE	39
LTR	Likelihood Ratio Test	54
Lys	Lysine	85
M	molar	36
Mb	megabase pair	79
MEAM1	Middle East-Asia Minor 1	25
MED	Mediterranean	25
Met	Methionine	85
min	minute	36
ML	Maximum Likelihood	52
MLC	Microtubule-Like Complex	62
MLS	Microtubule-Like Structures	114
mRNA	messenger RNA	68
mtCOI	Mitochondrial Cytochrome C Oxidase subunit I	25

mya	million years ago	123
Myr	million years	10
NAD	Nicotinamide adenine dinucleotide	86
NADH	nicotinamide adenine dinucleotide	134
NGS	Next Generation Sequencing	39
nm	nanometre	64
O/N	overnight	36
ORF	Open Reading Frame	48
Oxa	Oxaloacetate	84
P-endosymbiont	Primary endosymbiont	11
PBS	Phosphate Buffer Saline	37
PCR	Polymerase Chain Reaction	38
PEP	Phosphoenolpyruvate	83
Phe	Phenylalanine	85
PLTS	Phage-Like-Protein-Translocation Structures	118
PMF	Proton Motive Force	112
PPyr	2-oxo-3-phenylpropanoate	86
PRPP	5-phospho-a-D-ribose-1-diphosphate	84
Pyr	Pyruvate	85
Rb5P	D-ribulose-5-phosphate	85
rpm	revolutions per minute	38
rRNA	ribosomal RNA	68
RT	Room Temperature	36
S-endosymbiont	Secondary endosymbiont	11
s	second	40

SCFP *Saprospira* cytoplasmic fibril proteins . . . 118

SDS Sodium Dodecyl Sulfate . . . 36

SET Serial Endosymbiotic Theory . . . 5

SFF Standard flowgram format . . . 43

T Thymine . . . 9

T1SS Type 1 Secretion System . . . 101

T6SS Type 6 Secretion System . . . 118

T9SS Type 9 Secretion System . . . 112

TCA tricarboxylic acid cycle . . . 74

TE Tris-EDTA . . . 39

TEM Transmission Electron Microscope . . . 36

THF Tetrahydrofolate . . . 86

Thr Threonine . . . 83

Tm Melting Temperature . . . 40

tmRNA transfer-messenger RNA . . . 68

Tris Tris(hydroxymethyl)aminomethane . . . 36

tRNA transfer RNA . . . 48

Trp Tryptophan . . . 83

TTP Thymidine triphosphate . . . 10

Tyr Tyrosine . . . 87

Val Valine . . . 85

WGA Whole Genome Amplification . . . 38

7.2 Scientific production

The results from this thesis have been (or are going to be) published in:

- Santos-Garcia D, *et al.* 2012. Complete genome sequence of *Candidatus* Portiera aleyrodidarum"BT-QVLC, an obligate symbiont that supplies amino acids and carotenoids to *Bemisia tabaci*. J. Bacteriol., 194(23):6654–5. (Sections *Portiera* BT-QVLC and *Portiera* biosynthetic capabilities).
- Santos-Garcia D, *et al.* 2014. The genome of *Cardinium* cBtQ1 provides insights into genome reduction, symbiont motility and its settlement in *Bemisia tabaci*. Genome Biol Evol, 6(4):1013–30. (Section The third passenger: *Cardinium* cBtQ1).
- Santos-Garcia D, *et al.* 2014. Small but powerful, the primary endosymbiont of moss bugs, *Candidatus* Evansia muelleri, holds a reduced genome with large biosynthetic capabilities. Genome Biol Evol, 6(7):1875–1893. (Related to Section *Portiera* BT-QVLC).
- Santos-Garcia D, *et al.* 2014. No exception to the rule: *Candidatus* Portiera aleyrodidarum cell wall revisited. FEMS Microbiol. Lett. (published online: 19 SEP 2014) (Section *B. tabaci* QHC-VLC endosymbionts).
- Santos-Garcia D, et al. 2014-2015. Genome evolution in the primary endosymbionts of whiteflies sheds light about its host divergence. In preparation. (Section Genome evolution of the genus *Portiera*).

The results from this thesis have been presented in different national and international meetings.:

- Talks:

  GDRE Comparative genomics meeting. Santos-Garcia D, Latorre A, Moya A, Beitia F, Mouton L, Silva FJ. The endosymbiotic metagenome of the whitefly *Bemisia tabaci* à la carte. International meeting: Barcelona - Spain (2010).

7th International Symbiosis Society Congress. Santos-Garcia D, Beitia F, Mouton L, Moya A, Latorre A, Silva FJ. Endosymbiont genomes of *Bemisia tabaci* QHC. International congress: Krakow - Poland (2012).

GDRE Comparative genomics meeting. Santos-Garcia D, Beitia F, Mouton L, Moya A, Latorre A, Silva FJ. Whitefly endosymbiont metagenomes. International meeting: Lyon - France (2012).

IV Biodiversity Congress. Santos-Garcia D, Beitia F,Moya A, Latorre A, Silva FJ. Symbiotic communities in *Bemisia tabaci*. National congress: Bilbao - Spain (2013).

First International Whitefly Symposium. Santos-Garcia D, Farnier P-A, Beitia F, Zchori-Fein E, Vavre F, Mouton L, Moya A, Latorre A, Silva FJ. The genome of "*Candidatus* Portiera aleyrodidarum" BT-QVLC, and obligate symbiont that supplies amino acids and carotenoids to *Bemisia tabaci*. International symposium: Kolymbari - Greece (2013).

First International Whitefly Symposium. Santos-Garcia D, Farnier P-A, Beitia F, Zchori-Fein E, Vavre F, Mouton L, Moya A, Latorre A, . *Cardinium* and *Bemisia tabaci*: is it a mutualistic relationship?. International symposium: Kolymbari - Greece (2013).

First International Whitefly Symposium. PA. Rollat-Farnier, W.W. Wang, D. Santos-Garcia, E. Zchori-Fein, S.S. Liu, M.F. Sagot, F.J. Silva, F. Vavre, L. Mouton. Evolution of *Hamiltonella defensa* genomes in phloemophagous insects. International symposium: Kolymbari - Greece (2013).

- Posters

III SESBE Congress. Santos-Garcia D, Beitia F, Moya A, Latorre A, Silva FJ. Symbiosis in *Bemisia tabaci* strain QHC: a metagenomic approach. National congress: Madrid -Spain (2011).

XXXVIII Spanish Society of Genetics Congress. Santos-Garcia D, Beitia F, Moya A, Latorre A, Silva FJ. The endosymbiotic metagenome of the whitefly *Bemisia tabaci*: QHC strain. National congress: Murcia - Spain (2011).

Arthopod Symbiosis: From fundamental research to pest and disease management (COST FAOFAO701 Final Meeting). Santos-Garcia D, Beitia F, Mouton L, Moya A, Latorre A, Silva FJ. The genome of *Candidatus* Cardinium hertigii, a secondary endosymbiont of the whitefly *Bemisia tabaci*. International meeting: St. Pierre d'Oleron - France (2012)

Arthopod Symbiosis: From fundamental research to pest and disease management (COSTFAOFAO701 Final Meeting). Moreira M, Santos-Garcia D, Latorre A, Khadem M. Detection of spiders microbial communities in Madeira Island. International meeting: St. Pierre d'Oleron - France (2012).

Arthopod Symbiosis: From fundamental researchto pest and disease management (COST FAO701 Final Meeting). Augustinos AA, Santos-Garcia D, Dionyssopoulou E, Moreira M, Papapanagiotou A, Scarvelakis M, Doudoumis V, Ramos S, Aguiar AF, Borges PAV, Khadem M, Latorre A, Tsiamis G, Bourtzis K. New Supergroups and hidden *Wolbachia* diversity in aphids. International meeting: St. Pierre d'Oleron - France (2012).

7th International Symbiosis Society Congress. Santos-Garcia D, Peris-Bondia F, D'Auria G, Moya A, Silva FJ, Latorre A. Enrichment of insect samples with bacterial symbionts using flow cytometry. International congress: Krakow - Poland (2013).

4th Meeting of the Spanish Society for Evolutionary Biology. Santos-Garcia D, Vargas-Chavez C, Moya A, Latorre A, Silva FJ. Genome Evolution in the Primary Endosymbionts of Whiteflies. National congress: Barcelona - Spain (2013).

Part 8

Resumen

“I am the Lord and Master of the Sword
See Magic in my eyes
That Force became my endless curse
Witcher is my name
Adrenaline burns me inside
All Spirits from the Past protect the souls which never rest...”

Vader

Introducción

8.1.1. Simbiosis

Dependiendo de la localización del hospedador y el simbionte, la simbiosis es considerada **ectosimbiosis** (el simbionte está localizado en la superficie externa del hospedador), **endosimbiosis** (el simbionte está dentro del hospedador). Esta última puede ser **extracelular** (el simbionte está en cavidades internas o en el espacio intercelular) o **intracelular** (dentro de las células del hospedador). La simbiosis además puede clasificarse dependiendo del tipo de relación entre simbionte y hospedador en **parasitismo** (solo se beneficia el simbionte en detrimento del hospedador), **comensalismo** (solo se beneficia el simbionte) o **mutualismo** (ambos se benefician). Por último las simbiosis pueden ser **obligadas**, el simbionte no puede sobrevivir fuera del hospedador, o **facultativas**, donde el simbionte no requiere la simbiosis para su supervivencia.

Generalmente la endosimbiosis intracelulat conlleva un proceso denominado "reducción genómica" como consecuencia del paso de una forma de vida extracelular a una intracelular. Esto proceso se debe principalmente a la relajación de la selección natural (cambio a un ambiente muy estable), la redundancia génica entre el hospedador y el simbionte y a la acumulación de mutaciones en poblaciones asexuales pequeñas (trinquete de Muller).

8.1.2. Simbiosis en insectos

Los insectos presentan unos requerimientos nutricionales similares, necesitando un aporte de los nueve aminoácidos esenciales y arginina. Se ha propuesto que las relaciones simbióticas entre insectos y bacterias puede ser una de las razones del éxito evolutivo de estos animales, ya que les permite suplir las carencias que conllevan ciertas dietas desequilibradas. Los endosimbiontes en insectos se han clasificado como obligados o primarios (**P-endosimbiontes**) a aquellos que son necesarios para la supervivencia del insecto, frente a los facultativos o secundarios (**S-endosimbiontes**) que no lo son. Los P-endosimbiontes siempre se encuentran dentro de vacuolas derivadas del hospedador en unas células especializadas del insecto llamadas bacteriocitos, que pueden agruparse formando el bacterioma. En P-endosimbiontes gram negativos esto produce una típica estructura de tres membranas (la vacuola y la pared celular del simbionte, compuesta de la membrana externa e interna). Los S-endosimbiontes pueden hallarse tanto dentro como fuera del bacteriocito, incluso ocupar bacteriocitos secundarios.

La mayor parte del trabajo sobre simbiontes proviene del orden Hemiptera de insectos, y sobre todo del suborden Sterrnorrhyncha. Estos insectos son mayoritariamente fitófagos, alimentándose principalmente del floema de las plantas, que es rico en azúcares y aminoácidos no esenciales pero deficiente en aminoácidos esenciales, vitaminas y cofactores.

Los P-endosimbiontes de este grupo presenta unas características generales como son: reducción genómica manteniendo una maquinaria celular básica y aquellos genes biosintéticos requeridos por el insecto, un porcentaje de AT elevado, ausencia de elementos móviles, estasis genómica, transferencia vertical materna estricta y co-evolución con su huésped. Debido a la irreversibilidad de la reducción genómica, en ocasiones la pérdida de genes biosintéticos en el P-endosimbionte conlleva la aparición de interdependencias metabólicas entre el P-endosimbionte y un S-endosimbionte, que puede llegar a derivar en un endosimbionte co-primario, o que el P-endosimbionte sea reemplazado por un nuevo endosimbionte menos degradado. También es posible que estas funciones sean transferidas al insecto mediante TGH o que otros enzimas adquieran esa función. La problemática actual es donde situar el límite entre un orgánulo (mitocondria o cloroplasto) y un simbionte extremadamente reducido.

Los S-endosimbiontes no presentan una transferencia vertical materna estricta, sino que pueden transferirse además horizontalmente (por ejemplo de una especie a otra). Se ha postulado que debido a los efectos negativos producidos por la presencia de los S-endosimbiontes, estos deben mantenerse por diversos mecanismos no excluyentes: incremento de la transferencia horizontal, manipulación de la reproducción del hospedador o un incremento en la eficacia biológica del insecto (generalmente dependiente del ambiente).

8.1.3. Moscas blancas

El orden Hemiptera está compuesto por cuatro subórdenes: Sternorrhyncha, Auchenorrhyncha (cícadas), Heteroptera y Coleorrhyncha (bichos del musgo). Los Sternorrhyncha se dividen en dos linajes y cuatro superfamilias: los Aphidinea que contiene a los Aphidoidea (áfidos) y Coccoidea (cochinillas), y los Psyllinea que agrupa a Psylloidea (psílidos) y Aleyrodoidea.

Los Aleyrodoidea, o moscas blancas, son de origen plaeotropical y pudieron alimentarse de gimmnospermas para luego radiar junto a las angiospermas. Su reproducción es por partenogénesis (arrenotoquia) con una determinación sexual X0. Presenta cuatro estadios ninfales, el último conocido como “pupa de ojos rojos” debido a que es un estadio quiescente.

Los Aleyrodoidea se compone de una familia (Aleyrodidae) y dos subfamilias, los Aleyrodinae (96 géneros) y los Aleurodicinae (14 géneros). Los primeros fósiles datan del Cretácico Temprano pero es posible que la radiación de las moscas blancas se iniciara en el Jurásico Tardío asociada a los bosques de gimmnospermas.

Las moscas blancas presentan un P-endosimbionte llamado *Ca.* Portiera aleyrodidarum (*Portiera*), que parece carecer de pared celular y por la tanto presenta solo dos membranas, la vacuola y la membrana interna de la bacteria. *Portiera* (Oceanospirillales:Halomonadaceae) forma, junto a *Ca.* Carsonella ruddi (*Carsonella*) y *Ca.* Evansia muelleri (*Evansia*) (P-endosimbiontes de psílidos y bichos del musgo respectivamente), un linaje de endosimbiontes emparentados con las bacterias de vida libre *Halomonas elongata* y *Chromohalobacter salexigens*. Las moscas blancas pueden presentar una gran variedad de S-endosimbiontes. Los S-endosimbiontes *Ca.* Hamiltonella defensa, *Ca.* Arsenophonus sp., *Ca.* Hemipteriphilus asiaticus y *Fritschea bemisiae* se encuentran estrictamente en el bacteriocito. Sin embargo, *Wolbachia* sp, *Rickettsia* sp y *Ca.* Cardinium hertigii pueden estar tanto dentro como fuera del bacteriocito. Es importante conocer que estos tres últimos simbiontes junto a *Ca.* Arsenophonus sp. son conocidos manipuladores de la reproducción en insectos.

Por último, las moscas blancas presentan un sistema especializado de transmisión de los simbiontes en el que un bacteriocito de la madre migra hasta el oocito en desarrollo y penetra a través del pedicelo. Al final de la oogénesis, el bacteriocito se integra al ooplasma.

8.1.4. Moscas blancas usadas en este trabajo

Aleyrodinae:

- *Bemisia tabaci*: mide cerca de 1 mm y tiene un distribución mundial desde las regiones tropicales a las subtropicales. Está considerada una de las peores especies invasores. Actualmente se le considera un complejo de especies (morfológicamente indiferenciables) divididas en la menos 24 especies. Las dos especies más distribuidas, y dañinas para la agricultura, son el denominado biotipo B o la especie Middle East-Asia Minor 1 (MEAM1), y el biotipo Q o la especie Mediterranean (MED). El biotipo Q se divide en cuatro haplotipos, siendo el Q1 y el Q2 los más distribuidos a nivel global. El Q1 se caracteriza por presentar *Hamiltonella* como S-endosimbionte, muchas veces portando además *Cardinium* o *Wolbachia*.
- *Trialeurodes vaporariorum*: mide entre 1-1,5mm y se distribuye

en climas templados e invernaderos. Los géneros de simbiontes detectados son similares a *B. tabaci*.

Aleurodicinae:

- *Aleurodicus dispersus*: mide entre 2-3 mm. Las ninfas produce secreciones características como protección frente a enemigos. Proviene del Caribe y América Central, siendo a día de hoy un problema en regiones neotropicales como las Islas Canarias.
- *Aleurodicus floccissimus* (*Lecanoideus*): posible endemismo de las Islas Canarias, donde es un problema agrícola. Se alimenta de las mismas plantas que *A. dispersus*.

Objetivos

El primer objetivo de este trabajo es describir la comunidad endosimbiótica y su relación con su hospedador en un cepa de laboratorio de *B. tabaci* a través de una aproximación metagenómica. El segundo objetivo es describir la evolución del P-simbionte de las moscas blancas y validar su uso para la datación molecular de sus hospedadores.

Material y Métodos

8.3.1. Moscas blancas usadas

Se usaron cuatro especies de moscas blancas. La cepa QHC-VLC es una cepa de *B. tabaci* criada en laboratorio. Se denominó QHC-VLC de acuerdo a los endosimbiontes secundarios que porta, *Hamiltonella* y *Cardinium*, y a la localización geográfica, Valencia. Las otras tres especies fueron capturadas en el campo y se denominaron *T. vaporariorum* TVAW-BCN, *A. dispersus* ADAW-CAI y *A. floccissimus* AFAW-CAI debido a que portaban *Arsenophonus* y *Wolbachia* y fueron recolectadas en Barcelona y las Islas Canarias respectivamente.

8.3.2. Técnicas microscópicas

Las ninfas y huevos de *B. tabaci* usados para microscopia electrónica fueron lavados rápidamente con etanol 70 % para eliminar la cera y fijados con Karnowsky en una bomba de vacío (5 ciclos de 1 min) y dejados toda la noche en el fijador. El resto de pasos fueron los comunes para este tipo de técnicas. Para medir las membranas se usaron dos muestras independientes de las que se seleccionaron tres imágenes por muestra. De cada imagen se tomaron 5 medidas para cada componente de la membrana con **Fiji**.

Las ninfas de *B. tabaci* para los análisis de hibridación fluorescente in situ (FISH) se fijaron en Carnoy toda la noche y se decoloraron con 6 % H_22O_2. La hibridación con las sondas fluorescentes se dejo toda la noche a temperatura ambiente en una solución de hibridación estándar y se lavó previamente al montaje. Las sondas usadas fueron obtenidas de la literatura y el fluoróforo fue FAM para *Portiera*, Cy3 para *Hamiltonella* y Cy5 para *Cardinium*. **Icy** se usó para obtener las imágenes.

8.3.3. Enriquecimiento de muestras en endosymbiontes

Se usaron dos técnicas:

- Protocolo de Harrison: se basa en homogeneizar la muestra para liberar el simbionte de dentro de la célula eucariota. Por sucesivos filtrados (desde 1mm hasta 5 um) se van eliminando los tejidos y restos celulares del insecto hasta obtener una muestra enriquecida en bacterias. Como paso final se incuba con DNasaI para eliminar parte del ADN del insecto en suspensión.
- Extracción de bacteriomas: los bacteriomas se extrajeron con un microcapilar a partir de pupas de cuarto estadio. Posteriormente se realizó la reacción de amplificación genómica (WGA) usando el kit GenomiPhi V2. Se realizaron varias amplificaciones de diversos bacteriomas para mezclarse previamente a la secuenciación.

8.3.4. Extracciones de ADN, PCR y cuantificación

Para las extracciones generales de ADN genómico se usaron dos kits comerciales:

- JetFlex Genomic DNA: se basa en una lisis alcalina conjunta con proteinasa K y SDS. Se elimina el SDS y los restos celulares con acetato y se recupera el ADN por precipitación con isopropanol.
- Chelex: la muestra se homogeneiza y digiere a 99°C en presencia del Chelex. La muestra es centrifugada y el ADN queda en el sobrenadante, que puede usarse para diversas aplicaciones.

Las reacciones de PCR siguieron los protocolos estándares en biología molecular. Cuando se necesitó un mayor poder de detección (muestras con poca cantidad de ADN) se utilizó una PCR cuantitativa (LightCycler 2.0). Por último, cuando las amplificaciones por PCR dieron más de un producto, se usó la técnica de PCR de colonias para obtener un solo producto por amplificación. Para purificar los amplicones de la PCR se usó el kit NucleoFast R PCR (Macherey-Nagel) o la banda de interés era cortada del gel y purificada con el kit SpinPrep Gel DNA Kit (Millipore).

Para la cuantificación de ADN se usaron técnicas de espectrofotometría (Nanodrop ND-1000) y fluorimetría (Picogreen y Qubit 2.0.).

8.3.5. Secuenciación de genomas

Los amplicones de PCR se secuenciaron por el método de Sanger. El paquete de software **Staden** se usó para procesar los ficheros de salida del secuenciador.

Para la secuenciación de los genomas de los endosimbiontes de las cuatro moscas blancas se usaron dos tecnologías: Genome Sequencer FLX+ (454 Life Sciences, Roche) y HiSeq2000 (Illumina). En cuatros casos, se generaron dos tipos de librerías: *shotgun* (no contiene información posicional) y *pair-end/mate-pair* (contienen información posicional.)

8.3.6. Ensamblaje y anotación de genomas

Para ensamblar los distintos genomas se usó una serie de pasos (o pipeline) para identificar y separar las secuencias de cada simbionte del total secuenciado (o metagenoma), que incluye al insecto y otras bacterias:

1. Pre-procesado: las secuencias en bruto deben pasar unos filtros de calidad, de forma que se eliminan todas las secuencias de baja calidad. Las secuencias de calidad son ensambladas y los *contigs* son agrupados acorde a distintos factores (contenido de GC, similitud por **BLAST** a genomas conocidos, cobertura del *contig*, **PhymMBL**,...). La agrupación de *contigs* de interés es usada para seleccionar las secuencias de calidad (mediante un mapeo) y son re-ensambladas por separado.
2. Refinamiento y edición manual del ensamblaje: las secuencias *pair-end/mate-pair* se usaron junto a programas que automáticamente ordenan e intentan cerrar los huecos en el ensamblaje. Finalmente, se usó **Gap4** para unir *contigs* de forma manual usando información sobre la redundancia del genoma (proporcionada por el ensamblador **MIRA**) y las secuencias *pair-end/mate-pair*. En este punto si el genoma no está cerrado se procede a intentar recuperar nuevas secuencias.
3. Mapeo iterativo: es un proceso iterativo para recuperar nuevas lecturas y cerrar posibles huecos. **MIRA** se usa para extender los extremos de los *contigs* mientras que **Gap4** es utilizado para cerrar las posibles uniones debido a las nuevas secuencias de los extremos. Se repite hasta que el genoma se cierra o no se recuperan nuevas secuencias.

Para la anotación de los genomas se obtuvo una primera anotación automática (**Prodigal**, **BASys** y **RAST**) que luego fue refinada en Artemis usando distintas bases de datos y programas:

- Secuencias de DNA codificantes (CDS): **Pfam**, **Uniprot**, **Interpro**, **BLAST**, **CCD** y **PHAST**.

- Dominios funcionales y clasificación de proteínas: **HMMER**, **InterProScan** (Gene Ontology (**GO**) y **TIGRFAM**), Cluster of Orthologous Genes (**COG**).
- Genes de RNA: **tRNAScan-SE**, **TFAM**, **RFAM**.
- Inferencia metabólica: **KEGG**, **KAAS**, **pathway-tools** (**EcoCyc**, **BioCyc** y **MetaCyc**).
- Secuencias de Inserción: **ISsaga**, **ISfinder** y **MEGAN4** (**BLASTX**)
- El origen de replicación se buscó con **Ori-Finder** y los genomas se dibujaron con **circos**.

8.3.7. Genómica comparativa

Las tablas de CDS ortólogas entre los organismos de interés se generaron con **OrthoMCL** y revisados manualmente. Los diagramas de Euler (**gplots** de **R**) permiten visualizar las tablas de ortología organizadas en: las CDS compartidas por todos los organismos (*core*), las compartidas solo por distintos organismos (intersecciones) y aquellas específicos de cada organismo (*strain*). La localización de cada CDS ortóloga en cada genoma (sintenia) se representó con **genoPlotR**. **MGR** se usó para calcular el número de reordenaciones genómicas necesarias para explicar las distintas arquitecturas genómicas.

Mauve y **Sibelia** se usaron para comparar bloques sinténicos de nucleótidos entre los genomas de interés. La representación gráfica se realizó con **genoPlotR** y **circos** respectivamente. **NUCmer** (**MUMmer3**) se usó para identificar y representar las regiones repetidas de los genomas (95 % identidad y al menos 500 pb).

La reconstrucción de los Últimos Ancestros Comunes (LCA) se basó en las tablas de ortología generadas por **OrthoMCL**. El paquete **ape** de **R** y la función MPR se usaron para inferir por máxima parsimonia la presencia de cada ortólogo en cada LCA. Posteriormente las reconstrucciones se refinaron manualmente con **Mesquite**. EL número de cada categoría **COG** (obtenidas previamente para cada ortólogo) se calcularon para los LCAs de acuerdo a ausencia (0), presencia (1) o imposible determinar su presencia/ausencia (0,5). El número de categorías **COG**, o su abundancia relativa, se representaron como *heatmaps* (usando en ocasiones la opción de *hierarchical clustering*) con **gplots** de **R**.

La competencia metabólica entre los distintos simbiontes se calculó usando el programa **NetCmpt** (facilita un índice de competencia que va de 0 a 1, siendo 1 que un organismo excluye al otro).

8.3.8. Métodos filogenéticos

Los alineamientos se obtuvieron con el programa **MAFFT** (nucleótidos o proteínas) y **ssu-aligner** (específico para genes ribosomales). **Gblocks** se uso para eliminar las posiciones demasiado variables del alineamiento. **JModeltest2** (nucleótidos) y **ProtTest3** (proteínas) se usaron para inferir el mejor modelo evolutivo. **PAL2NAL** se usó para generar los alineamientos basados en codones.

RaxML (con optimización de ramas y 1000 bootstrap rápidos) se usó para generar árboles filogenéticos por Máxima Verosimilitud (ML). **PhyloBayes3** se usó para el análisis Bayesiano de los árboles ML, siguiendo las recomendaciones del autor. En cada caso, el árbol seleccionado (por la regla de la mayoría) se visualizó con **Archeopterix**.

Para la estimación de la divergencia se usó **BEAST2**. El archivo xml se generó con **Beauti**, seleccionando en cada caso la partición de los datos más adecuado y el modelo dado por **JModeltest2**. El modelo de especiación seleccionado fue Yule con un reloj logarítmico relajado. Los puntos de calibración se ajustaron a un modelo uniforme. Cada grupo de datos se corrió previamente sin añadir las secuencias para comprobar que las edades de divergencia no se debían a los *priors* seleccionados. Finalmente, para cada grupo de datos se corrieron ocho cadenas independientes que se combinaron a *posteriori*. Todas las cadenas cumplieron las recomendaciones de los autores. **Phylobayes3** se usó a su vez para datar los mismos conjuntos de datos, ajustándolos a los requisitos del programa, y corroborar la reproducibilidad de las dataciones.

8.3.9. Análisis Evolutivo

CodeML (del paquete **PAML**) se usó para calcular el número de sustituciones sinónimas (dS) y no sinónimas (dN) para las distintas CDS ortólogas. Las dS y dN se calcularon bajo los modelos m0, m1 y m2, usando el Test de Razón de Verosimilitud (LTR) para seleccionar el más adecuado para cada CDS. **R** se usó para la limpieza de datos y los análisis estadísticos, así como para generar las figuras relacionadas. Se usaron dos tipos de pruebas estadísticas: T-test de Student para varianzas iguales y diferentes (procedimiento de Welch's) pero con datos normales y Kruskal-Wallis (con sus pruebas *post-hoc* y corregidos por el método Bonferroni) para varianzas iguales pero con datos no-normales. Las dS y dN genómicas se calcularon como un media aritmética ponderada.

Para detectar selección positiva se usaron aquellas CDS ortólogas con una similitud entre ellas igual o superior al 80 %. **CodeML** bajo el modelo A de *branch-site* seleccionando, en cada caso, las ramas basales y las de interés fueron ajustadas para cada comparación. Cuando la hipótesis nula (los codones de todas las ramas presentan evolución purificadora o neutra) era rechazada (LTR ajustado por Bonferroni), la hipótesis alternativa (la rama de interés presentan sitios que evolucionan bajo selección positiva), el Bayes Empirical Bayes (BEB) fue inspeccionado para corroborar la significación estadística.

Resultados y discusión

8.4.1. *Portiera* y su socia *Hamiltonella*

8.4.1.1. Endosimbiontes en *B. tabaci* QHC-VLC

Los experimentos de FISH mostraron que *Portiera* y *Hamiltonella* están siempre presentes en el bacterioma, mientras que *Cardinium* puede estar tanto dentro como fuera (fenotipo disperso). Mientras que *Hamiltonella* se localiza más cercana al núcleo del bacteriocito (formando un cinturón), *Cardinium* parece localizarse más a la periferia formando a veces densos agregados celulares. Por último, *Portiera* parece englobar a los dos S-endosimbiontes.

Durante este trabajo se encontró que *Portiera* presentaba tres membranas. Cuando se comparó la capacidad de generar membranas de *Portiera* frente a otros P-endosimbiontes reducidos con un sistema de tres membranas, incluyendo a sus parientes *Carsonella* y *Evansia*, no hubo grandes diferencias. Finalmente se procedió a la medición de la membrana vacuolar y la pared celular de *Portiera* siendo 9,52 nm la membrana externa, 7,72 la membrana interna, 6.88 el espacio perisplasmático (21,12-30,18 la pared celular en conjunto) y 12 nm la membrana vacuolar. Si bien no está claro como fabrican su membrana algunos endosimbionte como *Portiera*, *Carsonella*, *Evansia* o las cepas de *B. aphidicola* más reducidas. Se ha propuesto que los metabolitos necesarios para ello podrían obtenerse del hospedador o de algún otro endosimbionte. Además podría ser que la fragilidad de la membrana de *Portiera*, comparada con *Evansia* o *Carsonella*, tenga que ver con el sistema de transmisión de los endosimbiontes en las moscas blancas.

8.4.1.2. Portiera BT-QVLC

El nombre de *Portiera* BT-QVLC hace referencia a la especie *B. tabaci* así como al biotipo Q (MED) y la localización geográfica (Valencia). *Portiera* BT-QVLC es un P-endosimbionte extremadamente reducido con un genoma de 357 Kb, que codifica para 246 CDS, 38 genes de RNA (3 ribosomales, 33 tRNA, un tmRNA y la subunidad de RNA de la RNasa P (*rnpB*). Pese a compartir muchas características con otros P-endosimbiontes, *Portiera* presenta otras poco comunes: una densidad codificante muy baja (68 %), largas regiones intergénicas y regiones repetitivas (repeticiones en tándem principalmente).

8.4.1.2.1. Genómica comparada

La capacidades metabólicas de *Portiera* BT-QVLC se compararon con otros P-endosimbiontes en base a las categorías **COG**. Pese a tener un genoma de tamaño similar, las capacidades biosintéticas de *Portiera* frente a *Evansia* son menores. En general parece ser que *Portiera* BT-QVLC es más similar a P-endosimbiontes con un genoma de menor tamaño y que generalmente presentan una asociación con un endosimbionte co-primario.

Además se comparó la maquinaria celular básica y el metabolismo central de *Portiera* frente a *Carsonella* y *Buchnera* 5A. *Portiera* presenta el menor número de genes de replicación y reparación de todos los P-endosimbiontes conocidos, a excepción de *Uzinura*. Ambos endosimbiontes solo presentan dos genes de la DNA polimerasa, *dnaE* y *dnaB*, faltando incluso la subunidad encargada de la corrección de errores *dnaQ*. *Portiera* codifica casi toda la cadena de transporte de electrones y parte del ciclo Krebs (TCA), al contrario que *Carsonella* que ha perdido la NADH deshidrogenasa. Por otra parte, *Portiera* BT-QVLC, *Carsonella* y *Evansia* han perdido algunas aminoacil-tRNA sintetasas. *Portiera* presenta el sistema de reciclaje de proteínas ClpXP pseudogenizado. Pese a que este sistema suele estar conservado en otros P-endosimbiontes extremadamente reducidos, puede ser que en *Portiera*, haya sido sustituido por el sistema HlsUV (de función parcialmente redundante).

La pérdida de genes esenciales para la transferencia de la información en *Portiera* parece indicar que es más una entidad subcelular, o simbionelo, que un P-endosimbionte. Parece haber un umbral biológico que separa simbiontes de orgánulos y pese a que *Portiera*, *Carsonella* y *Evansia* aún cumplen su función endosimbiótica, han perdido su autonomía informacional. Pese a que parece que ciertas funciones metabólicas pueden ser transferidas del simbionte al núcleo del hospedador, aún no está claro la situación de los genes relacionados con la transferencia de información.

A día de hoy hay cuatro cepas de *Portiera* de *B. tabaci*, dos del biotipo Q (MED) y dos del B (MEAM1). Pese a algunas diferencias de anotación y ensamblaje (relacionadas con variantes estructurales), las tres cepas son idénticas.

8.4.1.2.2. *Hamiltonella* BT-QVLC

El genoma de *Hamiltonella* BT-QVLC se ensambló en 85 *scaffolds* (101 *contigs*), presentando un tamaño de 1.61 Mb. Otros dos genomas de *Hamiltonella* han sido hechos públicos, *Hamiltonella* 5A de *Acyrthosiphon pisum* y *Hamiltonella* MED de *B. tabaci* que presentan un tamaño genómico de 2.17 Mb y 1.84 Mb respectivamente. El número de CDS presentes en *Hamiltonella* 5A es de 2148 con una densidad codificante del 81 %. Pese a que *Hamiltonella* MED y *Hamiltonella* BT-QVLC presentan la misma densidad codificante que *Hamiltonella* 5A, su número de CDS, 1916 y 1839 respectivamente, es muy elevado. Esto puede ser debido a problemas de ensamblaje o a fragmentos pseudogenes reconocidos como CDS por los programas de anotación. *Hamiltonella* BT-QVLC y MED presentan una identidad nucleotídica del 99,6 %, indicando claramente que estas dos cepas han divergido recientemente y que las diferencias en cuanto a contenido y orden génico pueden ser debidas principalmente a diferencias de ensamblaje o anotación. Al comparar los bloques sinténicos entre estas dos cepas, encontramos que comparten la mayoría de su contenido genómico y estas cepas han sufrido una reducción genómica respecto a *Hamiltonella* 5A, aunque es posible detectar distintas adquisiciones de nuevo material genético, probablemente debidas a eventos de TGH. Al comparar el plásmido de *Hamiltonella* 5A frente a las *Hamiltonellas* de *B. tabaci*, este parece estar ausente aunque algunas regiones aún son detectables. Además, al igual que en *Hamiltonella* 5A, se detectó la presencia del fago APSE en *Hamiltonella* BT-QVLC y MED. Dicho fago parece conferir cierta resistencia frente a parasitoides, por lo que está función no debe ser descartada para *Hamiltonella* BT-QVLC y MED. Por último, se detectaron 230 Kb en *Hamiltonella* MED no pertenecientes a *Hamiltonella* y que son artefactos del proceso de ensamblaje (50 Kb pertenecientes a *Portiera* y 180 Kb de origen desconocido).

Al comparar las categorías **COG** de las tres *Hamiltonellas* todas parecen poseer una capacidades metabólicas similares, pero en BT-

QVLC y MED se ha producido una gran reducción en la categoría de transcripción y reparación (L), indicando la pérdida de autonomía informacional.

8.4.1.2.3. Integración metabólica

El metabolismo de *Portiera* BT-QVLC fue reconstruido y depositado en **BioCyc**. Debido al estado incompleto de los genomas, se reconstruyeron los metabolismos de *Hamiltonella* BT-QVLC y MED. Si algún pseudogen era detectado en las rutas de estudiadas, se comprobó manualmente el estado real de este gen para descartar posibles artefactos del ensamblaje. Las posibles reacciones acometidas por *B. tabaci* se infirieron en base a la base de datos **AcypiCyc**, **KEGG**, trabajos previamente publicados y buscando los enzimas seleccionados en los transcriptomas disponibles de *B. tabaci*.

Portiera BT-QVLC necesita varios metabolitos intermediarios del insecto para iniciar, o completar, sus rutas biosintéticas. *Portiera* presenta las rutas completas para sintetizar triptófano y treonina. Además es capaz de producir isoleucina, leucina y valina junto a *B. tabaci*, que codifica el último paso de esta ruta *ilvE*. También podría sintetizar arginina, siempre que *B. tabaci* codificara el último paso de la ruta (*argH*), lo que si ocurre en sus parientes cercanos los psílidos y su simbionte *Carsonella*. *Portiera* produce metionina a partir de homocisteína (*metE*). La ruta de la histidina se encuentra incompleta en *Portiera*, pero es posible que este aminoácido esencial se encuentre en le floema de las plantas, lo sintetice la microbiota o exista algún tipo de complementación entre *Portiera* y el insecto. Por último, *Portiera* es capaz de sintetizar carotenos y algunos derivados del tetrahidrofolato (B9).

Hamiltonella BT-QVLC es capaz de producir treonina, fenilalanina y diversas vitaminas a veces previa importación desde el insecto de los metabolitos iniciales de esas rutas.

La fenilalanina puede ser producida por tres vías: por *hisC* (sustituye a *aspC*) en *Portiera*, el insecto produce el último paso o *Hamiltonella* termina la síntesis a partir de los productos de *Portiera*. A su vez, la lisina puede producirse bien porque la ruta esté compartida entre *Portiera* y *Hamiltonella* (tres intercambios de metabolitos) o *Hamiltonella* junto con el insecto (un intercambio) terminan la ruta.

Pese a que *Portiera* presenta un número muy reducido de transportadores, todos los tipos de metabolitos que requiere pueden ser transportados por al menos uno de ellos muy reducidos. Por el contrario, *Hamiltonella* codifica un gran número de transportadores. Además, el insecto podría codificar ciertos transportadores para facilitar el intercambio metabólico entre el consorcio.

8.4.2. El tercer pasajero: *Cardinium* cBtQ1

El holotipo de *Ca.* Cardinium hertigii fue caracterizado en *Encarsia hispida* y su prevalencia actual en artrópodos es del 7 %. La especie está divida en cuatro supergrupos (A, B, C y D). Aunque se le considera un manipulador de la reproducción en ocasiones podría ser un mutualista, como es el caso de *Cardinium* cEper1 de *Encarsia pergandiella*.

8.4.2.1. Características generales del genoma de *Cardinium* cBtQ1

Cardinium cBtQ1 (por hallarse en *B. tabaci* biotipo Q1) presenta un cromosoma de aproximadamente 1.013 Mb (11 contigs con un N50 de 661,9 Kb) y un plásmido multicopia, pcBtQ1, de 52 Kb. El cromosoma presenta 709 CDS, 156 pseudogenes (132 son transposasas), tres genes ribosomales (*16S* + *23S*-*5S*), 35 tRNA, un tmRNA y el gen *rnpB*. El plásmido contiene 30 CDS, cuatro pseudogenes (3 transposasas y una resolvasa).

La identidad nucleotídica entre *Cardinium* cBtQ1 y cEper1 es cerca del 93 %. Aunque cEper1 presenta un mayor número de genes, esto es debido a criterios de anotación. Además, los plásmidos de ambos *Cardinium* derivan del mismo plásmido ancestral aunque actualmente conservan muy pocos genes en común (aunque mantienen el orden génico y una alta identidad nucleotídica).

8.4.2.2. Estatus taxonómico de *Cardinium* cBtQ1

Las filogenias recuperadas a partir del gen ribosomal *16S* y el gen codificante *gyrB* sitúan a *Cardinium* cBtQ1 en el mismo clado que los *Cardinium* de distintas *Encarsia* spp. (supergrupo A). Debido a que la divergencia del *16S* es del 1 % *Cardinium* cBtQ1, se puede considerar una cepa del holotipo *Ca.* Cardinium hertigii de *E. hispida*.

La robustez de la reconstrucción filogenómica permitió establecer que tanto *Cardinium* como su pariente *Amoebophilus asiaticus* (endosimbionte de amebas) forman la familia Amoebophilaceae, incluida en el orden Cytophagales y próxima a las familias Cyclobacteriaceae y Flammeovirgaceae.

8.4.2.3. Genomica comparada

8.4.2.3.1. Elementos móviles y redundancia genómica

El nivel de redundancia genómica de *Cardinium* cBtQ1 es de un 14 %, el doble que *Cardinium* cEper1 o *A. asiaticus*. La mayoría de esta redundancia se debe a la presencia de elementos móviles, pero también encontramos duplicaciones segmentales. De 20 familias de elemento móviles, solo ocho pudieron ser anotadas como IS (ISCca1-8). Tres familias fueron específicas de *Cardinium* cBtQ1 mientras que las otras estaban presentes en el resto de Amoebophilaceae. Además, algunos IS parecen provenir de eventos de TGH desde Alfa-proteobacterias. Se detectó que algunos de los IS aún son activos (ISCca4 y 5), ya que durante su transposición han inactivado genes recientemente duplicados (identidad próxima al 100 %). La actividad de los IS junto a la presencia de una maquinaria de replicación y reparación completa parecen ser los responsables las reordenaciones producidas en *Cardinium* cBtQ1 frente a cEper1. Se ha postulado que los IS pueden ofrecer cierta plasticidad genómica útil en la colonización de nuevos nichos, lo que parece indicar que *Cardinium* cBtQ1 es más un endosimbionte facultativo mientras que cEper1 sería más un P-endosimbionte.

8.4.2.3.2. Genómica comparada entre las cepas de *Cardinium* y *A. asiaticus*

Las dos cepas de *Cardinium* y *A. asiaticus* comparten 468 grupos de CDS (*core*). Las cepas de *Cardinium* comparten 140 grupos de CDS, *Cardinium* cEper1 contiene 202 grupos de CDS, *A. asiaticus* comparte 13 grupos de CDS con *Cardinium* cEper1 y 6 con cBtQ1. Por último *Cardinium* cBtQ1 contiene 65 grupos de CDS específicas entre los que cabe destacar el operon *gldKLMN* y el gen *CHV_p021* (mide 14 Kb) del plásmido y el sistema de secreción tipo 1 (T1SS) *rtxBDE-tolC*, que ha sufrido una duplicación segmental en el cromosoma. Tanto el sistema RTX como el gen *CHV_p021* provienen de eventos de TGH de *Vibrio* y *Wolbachia* respectivamente. Proteínas con dominios de repeticiones RHS similares a CHV_p021 y con repeticiones de ankirinas en el extremo C-terminal, se han asociado con toxinas insecticidas o con procesos de señalización intracelular. Los T1SS son capaces de secretar proteínas con ankirinas en el extremo C-terminal.

8.4.2.3.3. Evolución del repertorio génico en los linajes de *A. asiaticus* y *Cardinium*

La agrupación en base a categorías funcionales, como las **COG**, parece estar relacionada con el nicho ecológico. Al realizar una agrupación jerárquica basada en las abundancias relativas de las categorías **COG**, se identificaron tres grupos: el primer grupo contenía a *Cardinium*, *A. asiaticus* y los ancestros comunes (LCA) 1 y 2, un segundo grupo incluía a *Marivirga tractuosa*, *Cyclobacterium marinum* y los LCS 3-4 y 8-9 y en el tercero se agruparon el resto de especies de Cythophagales y LCAs. Mientras que el tercer grupo son mayoritariamente especies marinas de vida libra, el segundo grupo incluye algunos casos de simbiosis facultativas. El grupo de los simbiontes presenta una retención de categorías asociadas la transferencia de información, algo muy común en endosimbiotes con genoma reducido. El LCA4, común a Amoebophilaceae y Cyclobacteriaceae, era una bacteria de vida libre que pudo iniciar alguna relación simbiótica y capaz de moverse mediante deslizamiento (*gliding*) como otros Bacteroidetes. La transición del LCA4 al LCA2 (ancestro de *Cardinium* y *A. asiaticus*) es claramente reductiva produciendo un endosimbionte ancestral con unas capacidades biosintéticas ya muy reducidas. Durante el paso de LCA2 al LCA1, el *Cardinium* ancestral, prosiguió la reducción aunque a su vez hubo varios casos de TGH, lo que podría posibilitar la especialización de cada cepa en unos tipos de hospedadores.

8.4.2.4. El metabolismo de *Cardinium* cBtQ1

La mayor diferencia en el metabolismo de las dos cepas de *Cardinium* está relacionada con la biosíntesis de vitaminas. Mientras que ambas cepas son capaces de sintetizar lipoato, *Cardinium* cBtQ1 parece haber perdido recientemente la capacidad de producir piridoxal, biotina y la pérdida del enzima cistationina gamma-liasa. Al analizar las posibles interacciones entre *Portiera*, *Hamiltonella* y *Cardinium* cBtQ1 con **NetCmpt**, se vio que mientras que los dos primeros no se ven afectados por la presencia del resto de endosimbiontes, *Cardinium* se ve inhibido por la presencia de *Hamiltonella*. Esto indica que *Cardinium* cBtQ1 compite contra *Hamiltonella* por los metabolitos del ambiente (hospedador) y es posible que el fenotipo disperso de *Cardinium* cBtQ1 le confiera alguna ventaja al abandonar el bacterioma, evitando así la competencia.

8.4.2.5. Genes de "deslizamiento" (*gliding*) en *Cardinium* cBtQ1

Atendiendo a los resultados del apartado 8.4.2.3.2 y 8.4.2.3.3, parece que el operon *gldKLMN* es muy importante para *Cardinium* cBtQ1 y se ha perdido en *Cardinium* cEper, ya que el ancestro de ambos contaba con estos genes.

Con el fin de confirmar que distintas cepas de *Cardinium* de *Encarsia* ssp. sin fenotipo disperso presentan esto genes, se realizó un cribado por PCR en tres especies de *Encarsia*, sin obtener ningún resultado positivo. Esto sugiere que los genes *gldKLMN* no están presentes en estas cepas de *Cardinium* y parecen ser responsables del fenotipo disperso de *Cardinium* cBtQ1.

Esto plantearía dos hipótesis: estos genes están relacionadas con el *gliding* o bien *Cardinium* cBtQ1 presenta un el sistema de secreción 9 o PorSS (compuesto por siete genes: *gldKLMN* y *sprAET*). Si estos genes están relacionados con el *gliding* y forman (según el modelo más aceptado) un motor molecular mínimo, necesitarían la interacción del citoesqueleto y la secreción de alguna proteína que pusiera en contacto el ambiente extracelular con el motor. En *Cardinium* cBtQ1 solo se ha encontrado el operon *gldKLMN*. En este caso el sistema RTX sería el encargado de secretar proteínas con dominios eucarióticos que podrían ejercer la función de las adesinas en un ambiente pluricelular eucariotico. La segunda hipótesis considera que el operon *gldKLMN* forma parte del sistema PorSS. El no haber encontrado los genes *sprAET*, ni ninguna proteína con el dominio especifico para la secreción mediante el sistema PorSS, junto con que el sistema RTX contiene su propia ATPasa para genera la energía requerida durante la secreción, hace dudar de la existencia de un sistema PorSS en *Cardinium* cBtQ1.

8.4.2.5.1. Organización de la maquinaria de *gliding*

Se ha propuesto que los MLCs en *Cardinium* están divididos en las Estructuras Similares a Microtúbulos (MLS) del citoplasma, la Placa Fibrosa Electrodensa (FEP) cerca de la membrana interna y las Estructuras Electrodensas (ES) del periplasma. En este trabajo se propone una posible organización de esta maquinaria. El homólogo de la tubulina FtsZ y de la actina MreB conformarían el ML y el FEP respectivamente mientras que el ES estaría conformado por las proteínas GldKLMN. GldM y L conformarían el motor molecular (las únicas con dominios transmembrana) mientras que GldK y N que están situadas en la membrana externa podrían ser el nexo de unión con las proteínas secretadas (contactos focales). El sistema SecYEG sería el encargado de transportar las proteínas GldKLMN al espacio periplásmico mientras que el sistema RTX se encargaría de secretar las proteínas necesarias para formar los contactos focales externos. El MLS y el FEP se encargaría de ser el nexo entre el motor y el citoplasma, siendo el FEP donde los MLS se insertan y organizan. Además, es muy probable que el MLS y FEP tengan otras funciones celulares (se han descrito en diversas bacterias), siendo el complejo GldKLMN capaz de reclutar y estabilizar dichas estructuras.

8.4.2.5.2. Rhapidosomas en *Cardinium*

Los rhapidosomas, también llamados proteínas citoplasmáticas fibrilares de *Saprospira* (SCFP) o textitafp-like, son proteínas similares a las colas de fagos que parecen microtúbulos y que se encuentras distribuidos en gran variedad de linajes bacterianos. Además, parecen estar relacionados con el sistema de secreción tipo 6 (T6SS), por lo que podrían tener una función similar.

La relación entre los SCFP/afp y los MLC proviene de trabajos experimentales ambiguos en *Saprospira.*Además, un estudio posterior caracterizó, mediante proteómica, los rhapidosomas de *Saprospira* en células estáticas indicando que los SCFP/rhapidosomas se generan tanto en células móviles como inmóviles. Todo ello indica que la asociación de los SCFP/rhapidosomas/afp con los MLC no es concluyente, lo que no excluye que *Cardinium* presente un T6SS o que los MLC sean en realidad rhapidosomas.

8.4.2.5.3. Posibles implicaciones del *gliding*

Al hacer un cribado por PCR para detectar la presencia de los distintos simbiontes en poblaciones de *B. tabaci* salvajes en la provincia de Valencia, se vio que *Cardinium* cBtQ1 estaba fijado en la poblaciones, casi todas ellas de biotipo Q (MED) aunque también en algunas poblaciones de biotipo S (Sub-Saharan Africa). El biotipo S es muy poco común en España y además el hecho de que porte *Arsenophonus*, un simbionte que muy raramente está en combinación con *Cardinium*, sugiere que *Cardinium* cBtQ1 ha saltado al biotipo S recientemente. Además, todas las muestras fueron positivas para el cribado de los genes *gldKLMN* y *CHV_p021*. Debido a que portar un endosimbionte que no proporciona ninguna mejora en la eficiencia biológica del hospedador suele seleccionarse en contra, es posible que *Cardinium* cBtQ1 produzca algún beneficio. Una posibilidad sería que los genes *CHV_p018* y *CHV_p021* puedan tener un efecto insecticida frente a parasitoides. Además, la capacidad de *Cardinium* para moverse e invadir distintos tejidos también podría ser beneficiosa. Sin embargo, es posible que *Cardinium* cBtQ1 sea simplemente un simbionte parásito manipulador de la reproducción.

8.4.3. Evolución genómica en el género *Portiera*

El primer registro fósil de las subfamilias Aleyrodinae y Aleurodicinae se encontró en el ámbar Libanés, datado en el Cretácico Inferior (125-135 millones de años (ma.)).

A día de hoy, cinco genomas de *Portiera* de la subfamilia Aleyrodinae han sido secuenciados, cuatro de *B. tabaci* y uno de *T. vaporariorum*. En este trabajo además se presentan tres genomas más de *Portiera*, uno de *T. vaporariorum* y dos pertenecientes a la subfamilia Aleurodicinae, *A. dispersus* y *A. floccissimus*.

8.4.3.1. Características genómicas de las cepas de *Portiera*

Las tres nuevas cepas de *Portiera* presentan un genoma extremadamente reducido, de aproximadamente 281 Kb *T. vaporariorum* y 290 Kb *A. disperus* y *A. floccissimus*. Las tres contienen 34 tRNAs, tres rRNAs (16S-23S-5S), un tmRNA y *rnpB*. Todas ellas presentan un sesgo del GC estable que se ha perdido en el linaje de *Portiera* de *B. tabaci*. Aunque todas las cepas de *Portiera* presentan repeticiones en tándem, estas se han acumulado en el linaje Aleyrodinae y principalmente en *B. tabaci*. Por último, las dos cepas de *Portiera* de *T. vaporariorum* son idénticas, indicando que pese a provenir de Norte América y España, son la misma cepa y su hospedador ha sido distribuido gracias a la importación/exportación de plantas.

8.4.3.2. Genómica comparada y estasis en el género *Portiera*

Mientras que las CDS compartidas por las cepas de *Portiera* serían 240, el pan-genoma estaría compuesto por 40 CDS más. La *Portiera* ancestral contendría un genoma de 280 CDS (incluyendo el gen *alaS* completo y el pseudogen *PAQ_201* solo presente en el linaje de *B. tabaci*), por lo que *Portiera* AF-CAI es la más cercana a dicho ancestro.

Al asignar las categorías **COG** a las CDS de cada cepa quedó patente que todos las cepas tienen un metabolismo similar. Sin embargo, la mayoría de pérdidas génicas se han dado en las *Portiera* del linaje Aleyrodinae (43 genes) mientras que solo tres en el Aleurodicinae. Además, mientras las cepas de *Portiera* de *T. vaporirarium* han perdido otros cuatro genes, las de *B. tabaci* han perdido 30. Además, cuando se analizó la sintenia entre las distintas cepas quedo patente que mientras que las cepas del linaje Aleurodicinae y de *T. vaporariorum* presentan una estasis genómica desde su divergencia, el linaje de *B. tabaci* ha sufrido un gran número de reordenaciones. Lo más probable es que esta inestabilidad genómica sea una combinación de diversos factores que incluyen la pérdida de la subunidad *dnaQ* de la polimerasa (corrección de errores), la expansión de las repeticiones en tándem y el alto índice de recombinación ilegítimo que presentan estás últimas.

8.4.3.3. Los “planos” metabólicos de las cepas Portiera

Todas las cepas de *Portiera* son capaces de sintetizar carotenos, proteínas Fe-S, producir poder reductor y metabolitos intermediarios usando piruvato y producir energía mediante la cadena de transporte de electrones y la ATP sintasa. Aún así, todas las cepas de *Portiera* necesitan importar diferentes compuestos por lo que han mantenido diez transportadores distintos (excepto el linaje de *B. tabaci* que ha perdido uno de ellos, *galP*, dedicado al importe de galactosa) que pueden importar un amplio rango de compuestos.

Las cepas de *Portiera* AD-CAI y AF-CAI presentan las rutas completas para la síntesis de lisina (tres genes perdidos en BT-QVLC debido a que *Hamiltonella* puede sintetizarla), arginina (*argH* se ha perdido en BT-QVLC), treonina, triptófano, fenilalanina, isoleucina, leucina y valina (el último paso de los cuatro últimos aminoácidos es producido por el insecto). La ruta de la histidina (que podría adquirirse de la dieta o podría complementarla el insecto) está incompleta en todas las cepas, habiéndose producido además la pérdida del gen *hisE* en la cepa TV-BCN. Todas las cepas producen metionina a partir de homocisteína (*metE*). La *Portiera* ancestral, y las actuales, son incapaces de sintetizar la mayoría de las vitaminas y cofactores, por lo que es posible que siempre requieran un S-endosimbionte para completar esta función. Esto podría explicar el sistema de transmisión de simbiontes en las moscas blancas, pero no hay que olvidar que los S-simbiontes podrían ser solo necesarios en ciertos momentos y no de manera continua.

Pese a que todas las cepas de *Portiera* presentan todos los tRNA necesario para cargar todos los aminoácidos, han sufrido la pérdida de distintas aminoacil-tRNA sintetasas (*argS* y *thrS*). La función de *asnS* ha sido sustituida por la combinación de *aspS* en conjunción a *gatABC*. BT-QVLC ha perdido además *alaS*, *metG* y *trpS*, mientras que TV-BCN solo ha perdido las dos primeras. El gen *alaS* se encuentra, compuesto por los dominios funcionales aminoacil-tRNA sintetasa y de edición (*alaXp*, que corrige tRNA$^{\mathrm{Ala}}$ mal cargados, ya que son letales), se encuentra completo en AF-CAI, partido en dos CDS funcionales en AD-CAI y solo se conserva el *alaXp* en BT-QVLC y TV-BCN. Esto implica que las cepas de *Portiera* necesitan algún mecanismo alternativo para aminoacilar esos tRNAs.

8.4.3.4. Tiempos de divergencia en los linajes de *Portiera*

La divergencia de las cepas de *Portiera* se calculó usando como punto de calibración 125-135 ma. (ámbar Libanés), y usando *H. elongata* y *C. salexigens* como *outgroups*. A no ser que se especifique lo contrario, los tiempos son los obtenidos con **BEAST2**. Además se usaron dos conjuntos de datos: el A (genes *rpoB*, *rpoC*, *carB* y *dnaE*) y el B (*sucA*, *aceE*, *valS* y *leuS*). Debido a que los intervalos de probabilidad (HDP) estimados por **BEAST2** para ambos conjuntos coincidían, se unieron para estimar las medias de los tiempos de divergencia. Además, los HPD de **PhyloBayes3** y **BEAST2** parecían coincidir, lo que indica la robustez de los tiempos obtenidos. La divergencia entre las *Portieras* de *A. dispersus* y *A. floccissimus* se estimó en 18,35 ma. mientras que la de *Portiera* de *B. tabaci* y *T. vaporariorum* en 90,1 ma. La divergencia entre *Portiera* BT-QVLC (Q o MED) y BT-B (B o MEAM1) se produjo hace 0,38 ma. (según **PhyloBayes3** 0.1 o 0.7 ma., conjunto A y B respectivamente). Al parecer la separación de los linajes de *Bemisia* y *Trialeurodes* ocurrieron durante el Cretácico Inferior junto a la radiación de las angiospermas mientras que la separación de los linajes de *Aleurodicus* estudiados ocurrió durante la aparición y evolución de las plantas con flor actuales.

Para corroborar estos resultados, se dató la divergencia de varias moscas blancas usando un fragmento de 1341 pb del gen mitocondrial *COI*, con dos puntos de calibración: el usado en el punto anterior y la aparición de los Sternorrhyncha (250-270 m.). En este caso se uso *A. pisum* como *outgroup*, obteniendo otra vez una congruencia entre los resultados de **BEAST2** y **PhyloBayes3** y coincidiendo con los resultados obtenidos con los genes de *Portiera*. La divergencia entre *A. floccissimus* y el clado *A. dispersus-dugesii* fue de 20,25 ma., la de *Trialeurodes-Bemisia* de 86,07 ma. y la de *B. tabaci* Q (MED) y B (MEAM1) en torno a 0,21 ma. La inclusión de una *B. tabaci* New World (se une a las otras en la rama más basal de la filogenia) permitió estimar de forma aproximada la divergencia del complejo *B. tabaci*, en unos 18,43 ma.

Pese a que la radiación del complejo *B. tabaci* ocurrió antes de la aparición de la agricultura, no es posible descartar del todo que esta no haya intervenido en la especiación de los biotipos Q (MED) y B (MEAM1).

8.4.3.5. Tasas de substitución nucleotídica en los linajes de *Portiera*

Tras calcular el número de substituciones sinónimas por sitio sinónimo por año (dS/año) y el de no-sinónimas (dN/año) para cada grupo de CDS ortólogas, solo se mantuvieron los valores que pasaron un control de calidad (146 CDS de 240).

Para ver si había diferencias estadísticas en la distribución de dN/año entre las cepas de *Portiera* TV-BCN, AD-CAI y AF-CAI, se empleó un test de Kruskal-Wallis junto a sus pruebas a *posteriori*. AD-CAI presenta una distribución de dN/año estadísticamente diferente a TV-BCN y AF-CAI, sin embargo estos dos últimos no presentaron diferencias en su distribución de dN/año.

A nivel de dS/año, se usó un T-test de Welch para comparar AD-CAI contra TV-BCN y AF-CAI obteniendo que la distribución de AD-CAI es estadísticamente diferente de las otras dos cepas de *Porteira*. Para comparar TV-BCN y AF-CAI se usó un T-test, concluyendo que no hay diferencias a nivel de dS/año entre estas dos cepas.

No fue necesario comprobar las distribuciones de dN/año y dS/año de *Portiera* BT-QVLC, ya que al fallar el test de Levene, ya es indicativo de que su distribución es distinta al resto de cepas. *Portiera* BT-QVLC presenta los mayores valores de dN o dS por año, seguida por AD-CAI y siendo TV-BCN y AF-CAI las que presentan unos valores menores.

Por otra parte, se calcularon las ω (dN/dS) para 187 grupos de CDS ortólogas y sus distribuciones en cada cepa de *Portiera* se compararon con el test de Kruskal-Wallis. No hubo diferencias significativas concluyendo que la mayoría de los genes en estas cepas está evolucionando bajo selección purificadora (ω <1). Además, la selección natural no parece ser responsable del incremento en las tasas de dN o dS de *Portiera* BT-QVLC sino que se debe a otros parámetros. Uno de ellos sería la pérdida de *dnaQ*, que produciría un aumento de la tasa de mutación y por ende de las dN y dS.

Por último se cálculo un valor medio, o genómico, para las dN/año y dS/año. En general, *Portiera* BT-QVLC presentó unas medias tres veces superiores a AD-CAI y cuatro a AF-CAI y TV-BCN. Pese a todo, todas las cepas de *Portiera* analizadas presentan unos valores de dN/año y dS/año en el rango de otros P-endosimbiontes.

8.4.3.6. Selección positiva en *Portiera* de *B. tabaci*

Se realizaron tres pruebas de selección positiva: *Portiera* de *T. vaporariorum* frente a *Portiera B. tabaci*, *Portiera* de *B. tabaci* B (MEAM1) frente a Q (MED) y *vicecersa*. Estos dos últimos test presentan muy poco poder estadístico debido a la falta de un tercer biotipo para comparar.

En los genes que dieron positivo para las tres pruebas (han sufrido selección positiva), y pese a los falsos positivos producidos por genes muy cortos, como los ribosomales, se observa una abundancia de genes relacionados con la biosíntesis de aminoácidos. Es posible que una forma de incrementar la eficacia biológica del hospedador sea actuando sobre las rutas de síntesis de aminoácidos, pero ello no es suficiente para explicar las diferencias de eficacia entre *T. vaporariorum* y *B. tabaci*.

Conclusiones

- Las cepas de *Portiera* de *B. tabaci* Q (MED) y B (MEAM1) son idénticas, presenta tres membranas, largas regiones intergénicas y un alto número de repeticiones.

- *Portiera* BT-QVLC participa en la síntesis de todos los aminoácidos esenciales, excepto lisina que es producido por *Hamiltonella*. Mientras *Portiera* sintetiza carotenos, *Hamiltonella* produce un gran número de vitaminas y cofactores.

- *Cardinium* cBtQ1 presenta una mayor redundancia que el resto de miembros de la familia Amoebophilaceae, que además presenta un clara evolución reductiva en sus genomas debido al paso a una forma de vida intracelular.

- *Cardinium* cBtQ1 no contribuye a complementar la dieta del insecto, presentando además una fuerte competencia con *Hamiltonella*. Es posible que el fenotipo disperso en esta cepa se haya mantenido para evitar está competencia.

- El mecanismo de *gliding* parece ser el responsable del fenotipo disperso. En este caso *Cardinium* cBtQ1 presentaría una maquinaria de *gliding* mínima adaptada a un ambiente multicelular. Además, este mecanismo podría ofrecer ciertas ventajas al hospedador.

- *Portiera* ha mantenido una estasis genómica durante los últimos 125-135 ma., excepto *Portiera* de *B. tabaci* que presenta gran número de reordenaciones. La pérdida de los genes de replicación y reparación junto a la recombinación ilegal entre las repeticiones en tándem parecen ser responsables de está inestabilidad genómica.

- *Portiera* AF-CAI es la más cercana a la *Portiera* ancestral. Todas las cepas tienen un metabolismo similar y sintetizan, o participan en la síntesis de, todos los aminoácidos esencial excepto *Portiera* BT-QVLC. Todas las moscas blancas requieren un S-endosimbionte para sintetizar vitaminas y cofactores, aunque podrían no ser necesarios en cierto momentos.

- La divergencia entre las *Portiera*, y por ende entre sus hospedadores, es más reciente de lo que se había establecido. *B. tabaci* B (MEAM1) y Q (MED) divergieron hace 0,38 ma.

- *Portiera* BT-QVLC mostró la tasa más alta a nivel de substituciones nucleotídicas por año (dN y dS), seguida por *Portiera* AD-CAI y TV-BCN/AF-CAI. El valor medio, o genómico, para dichos ratios está en el rango de otros P-endosimbiontes.

- No se observaron diferencias en al distribución de los valores de ω entre las cepasde *Portiera*, por lo que la selección natural no es la causante de la evolución acelerada en *Portiera* de *B. tabaci*.

ALEXANDER PP VI VALENTINVS FERDINANDVS DEI GRA REX ARAGONVM