



VNIVERSITAT
E VALÈNCIA



Towards minimal cells and beyond, the development and application of bioinformatic tools for large-scale genomic data analysis of endosymbiotic bacteria of insects

June 2023
B. Mariana
Reyes-Prieto



VNIVERSITAT
E VALÈNCIA

Towards minimal cells and beyond, the development and application of bioinformatic tools for large-scale genomic data analysis of endosymbiotic bacteria of insects

B. Mariana Reyes-Prieto

June 2023

Supervised by Andrés Moya
and Rosario Gil

Doctoral Program in Biodiversity
and Evolutionary Biology





VNIVERSITAT DE VALÈNCIA

FACULTAD DE CIENCIAS BIOLÓGICAS

INSTITUTO DE BIOLOGÍA INTEGRATIVA DE SISTEMAS

PROGRAMA DE DOCTORADO EN BIODIVERSIDAD Y BIOLOGÍA
EVOLUTIVA

**Towards minimal cells and beyond, the development and
application of bioinformatic tools for large scale genomic
data analysis of endosymbiotic bacteria of insects**

MEMORIA PRESENTADA POR BERTHA MARIANA REYES PRIETO PARA OPTAR
AL GRADO DE DOCTORA POR LA UNIVERSIDAD DE VALENCIA

CODIRECTORES:
PROF. ANDRÉS MOYA SIMARRO Y PROF. ROSARIO GIL GARCÍA

VALENCIA, Junio 2023

El PROF. ANDRÉS MOYA SIMARRO, Catedrático de Genética de la Universitat de València, y la PROF. ROSARIO GIL GARCÍA, Profesora Titular de Genética de la Universitat de València,

CERTIFICAN que el trabajo para optar al grado de Doctor en el Programa de Biodiversidad y Biología Evolutiva, y que lleva por título “Towards minimal cells and beyond, the development and application of bioinformatic tools for large scale genomic data analysis of endosymbiotic bacteria of insects”, ha sido realizado bajo su dirección en el Instituto Cavanilles de Biodiversidad y Biología Evolutiva y el Instituto de Biología Integrativa de Sistemas por BERTHA MARIANA REYES PRIETO.

Y para que conste, en el cumplimiento de la legislación vigente, firman el presente certificado.

MOYA
SIMARRO
ANDRES -
19829106R

Firmado digitalmente por MOYA SIMARRO ANDRES - 19829106R
Fecha: 2023.06.01 17:07:30 +02'00'

Dr. Andrés Moya Simarro

GIL GARCIA
MARIA
ROSARIO -
18949844Y

Firmado digitalmente por GIL GARCIA MARIA ROSARIO - 18949844Y
Fecha: 2023.06.01 18:09:53 +02'00'

Dra. Rosario Gil García

Valencia, Junio de 2023

"Nothing in Biology makes sense except in the light of evolution."

Theodosius Dobzhansky (1973)

Agradecimientos

Después de tantos años en esta aventura, no sé por dónde comenzar. Han sido años intensos, los mejores y los peores de mi vida donde he cometido errores y los mejores aciertos que jamás me hubiera imaginado.

Quiero agradecer en primer lugar, a mis padres. Tengo muchísimo que agradecerles a ellos en lo personal, pero me centro en lo académico. A mi madre que me enseñó la perseverancia. Que con el ejemplo me enseñó lo que es ser una mujer científica, con todo lo bueno y lo malo que conlleva. Que muy a su manera y sin saberlo, me hizo entender la necesidad del feminismo en todos los ámbitos de la vida, en especial en la academia. Que con su lucha constante me enseñó a luchar por mí misma, a nunca rendirme y a poner la mejor cara ante cualquier adversidad. A mi padre, que siempre fue y sigue siendo el motor impulsando las ideas científicas y dando ánimos y apoyo cuando más lo necesito, aun cuando en momentos significó cambiar su propio entorno. Por los consejos más sabios, que a veces aún me cuestan entender, pero que siempre tienen todo el sentido, gracias.

A mis hermanos. Miguel, una de las personas más inteligentes que conozco y con una visión muy clara sobre las cosas importantes de la vida. Para las decisiones más difíciles e importantes, siempre me apoyo en él. Gracias, hermano. Y a Nidia, mi hermana y mejor amiga, el apoyo incondicional que siempre, siempre está ahí. Mi hermanita con la que he reído, llorado, y a la que he cansado con tantas quejas y bajones a lo largo de la tesis, y que ha sido tan importante para la realización de esta, tan así que igual y no la hubiera terminado sin ella. Gracias por venir a visitarme en los momentos difíciles y darme ánimos. Gracias por venir a cuidar a tus sobrinos mientras yo trabajaba. Gracias por la terapia no remunerada que siempre me brindas, gracias por tanto, tanto. Los adoro.

En el ámbito profesional, mi admiración y gratitud eterna a mi director de tesis, Andrés Moya, y a Amparo Latorre, por todo su apoyo durante todos estos años. Gracias DOC por apostar en mí, aceptarme en su laboratorio y confiar en mi trabajo. Gracias por toda la ayuda, no solo en lo profesional si no en lo personal. Jamás olvidaré lo que ha hecho por mí y por mi familia, lo llevo grabado. Hay pocos jefes en este mundo que puedan ser tan comprensivos y empáticos como usted, y después de mi historia, solo me queda agradecer la paciencia y buena voluntad que ha tenido conmigo. Amparo, gracias por ser la *mama duck* que tanto necesité en momentos críticos de la tesis y por todos tus consejos. Seguiré pidiéndolos.

A mi segunda directora de tesis, Sari. Llegaste en un momento clave para mí, estaba perdida y destrozada y me rescataste sin parpadear. Muchísimas gracias por toda la enseñanza, por el *tough love*, por la presión necesaria para terminar los artículos y esta tesis. Te has vuelto mi modelo a seguir, me gusta mucho la manera en la que trabajas y nos haces trabajar a los demás. Me gusta que eres humana y entiendes cuándo se puede empujar un poco más o cuando necesitas dar un poco de espacio. Estoy muy agradecida contigo, te admiro muchísimo y espero seguir colaborando contigo, pues creo que ya llegamos a un equilibrio excelente de trabajo donde podemos dar mucho ambas. Gracias, gracias por tanto.

También quiero agradecer a los compañeros de laboratorio con los cuales compartí espacio, seminarios, congresos, viajes, proyectos, experiencias, ideas y risas. Han sido tantos años en esta travesía que son muchos los nombres que tengo que mencionar y espero no dejarme a nadie: desde mi amiga, compi de piso y compañera cazadora de cucarachas, Lexi, el metalhead Diego Santos, los

más veteranos, Elisa, Quelo, Rafa, Sergio, Vanesa, Mariano, y la mexabanda Manzano, Carlos, Tania y Ana, hasta los que llegaron después, Jesús, María y Emilio. Aunque no compartimos espacio en el mismo laboratorio, pero si en el adyacente, quiero agradecer a mi amiga Cristina Vilanova (mi gitana favorita), por siempre estar al pendiente mío y darme tanto apoyo hasta la fecha, y a Flavia Viana, mi amiga portuguesa que conocí gracias a Symbiomics, por las mismas razones. Gracias a todos.

A los colaboradores: A mi gran amigo Omar Ortúzar por la creación (y edición varias veces) del logo de mi base de datos. ¡Me encanta!

A Mario Fares, mi mentor por un breve tiempo. En ese poco espacio que compartimos pude aprender mucho de ti, de tu filosofía de vida y de lo relajada y divertida que puede ser la carrera científica. Siempre me quedaré con las ganas de seguir trabajando juntos. Gracias por todo, te seguimos recordando.

A Luis Delaye, por las fructíferas colaboraciones que hemos tenido y también por las discusiones tan interesantes que hemos compartido.

A Mercè Llabrés y Pere Palmer. Los colaboradores que vinieron a rescatarme de una entrecruzada considerable al no saber por dónde continuar con las redes metabólicas. Llegaron con sus maravillosas e innovadoras ideas, personas admirables con los cuales he tenido las discusiones más enriquecedoras y a los que les debo la mitad de esta tesis. Gracias por siempre impulsarme a seguir, en el ámbito profesional, pero también por estar pendiente de mi en lo personal. ¡¡¡Espero que sigamos estas fructíferas colaboraciones y sigan los viajes Mallorca-Valencia que tanto disfrutamos!!! A mi equipo de trabajo actual en FISABIO, que también me han apoyado y animado a terminar esta tesis, inclusive dándome tiempo para enfocarme en esto, gracias Llúcia, Inma, Gris, Loreto, Vicente y en especial a Giuseppe que siempre me inspira e impulsa a ser mejor y dar lo mejor de mí.

Me queda agradecer a mi familia, mexicana y española. Mi familia mexicana que está lejos pero siempre al pendiente. Mis suegros y las abuelas de mi marido que han sido mi familia lejos de mi familia, me han incluido como una más, y me han hecho sentir parte de una tribu que me encanta. Siempre preocupados por nosotros, cuidándonos y mimándonos.

A mi marido Juan Antonio, mi persona favorita de este mundo, que ha sabido estar en las buenas y en las muy malas conmigo. Que ha sabido navegar entre mi euforia y mi mal genio latino. Que ha sido paciente, amable, cariñoso y un soporte invaluable durante todo este tiempo. Siempre me he sentido segura y valorada a tu lado, no sé cómo hubiera logrado sobrevivir estos años si no fuera por ti. Gracias por ser mi refugio. No podría pedir más de lo que eres. Gracias amor.

A mis hijos. A mi cariñoso, inteligente, y noble hijo Cruz Javier, el niño más increíble que haya pisado este planeta. Y a mi loquita hija Mariel, eco-chuky, traviesa, ingeniosa e inteligente como ninguna otra. Gracias por ser mi mundo. Gracias por regresarme la vida, y por hacerme la mamá más feliz de este planeta. Todo esto es por y para ustedes. Los adoro con todo mi corazón.

A mis padres, Bertha Prieto Gómez y Cruz Reyes Vázquez, por su amor incondicional. Los amo y mis logros son suyos también.

Vuela alto mami.



Table of contents

I. Summary	1
II. Publications from this thesis	5
III. Introduction	9
1. <i>Symbiosis throughout the Tree of Life and as an evolutionary driving force</i>	9
2. <i>Symbiosis in the insect world</i>	12
3. <i>Intracellular obligate endosymbionts, organelle-like organisms, symbionelles... Are we there yet?</i>	15
4. <i>A vast necessity for the organization and compilation of symbiotic genomic information</i>	20
5. <i>Minimal cells and minimal metabolisms' analysis</i>	22
6. <i>Evolutionary traces in metabolic networks of endosymbiotic bacteria of insects</i>	24
IV. Objectives	31
V. General Materials and Methods	33
Chapter 1. Conceptual framework: Scanty microbes, the "symbionelle" concept, and the evolution of small prokaryotic genomes	39
1.1. <i>Opinion article: Scanty microbes, the 'symbionelle' concept</i>	41
1.2. <i>Evolution of small prokaryotic genomes</i>	50
1.2.1 What is the minimal genome size for extant free-living prokaryotes?	51
1.2.2 Drivers of genome reduction among free-living prokaryotes	55
1.2.3 The streamlining hypothesis	57
1.2.4 Accelerated rates of protein evolution	60
1.2.5 Increased rates of mutation hypothesis	61
1.2.6 The Black Queen Hypothesis	63
1.2.7 Host-associated prokaryotes with reduced genomes	67
1.2.8 Early obligated intracellular symbiosis	69
1.2.9 Paradoxically large G+C content in two highly reduced genomes	74
1.2.10 Unexpected loss of genomic stability	76
1.2.11 Novel hypothesis to explain the reassignment of STOP to Trp (tryptophan) codon	78
1.2.12 Clues to early life?	80
1.2.13 The role of horizontal gene transfer in the evolution of intracellular symbiosis	81
1.2.14 Biochemical complementarity and convergent evolution of co-resident symbionts	84
1.2.15 Genome reduction in bacterial symbionts of fungi, a relatively unexplored world	92
1.2.16 Drivers of genome reduction in host-associated bacteria	98
1.2.17 Conclusion	104
Chapter 2. Organizing and optimizing access to big data of endosymbiotic bacteria of insects	107
2.1. <i>SymbioGenomesDB: a database for the integration and access to knowledge on host-symbiont relationships</i>	110

2.1.1 Data collection	112
2.1.2 Database architecture and web interface	115
2.1.3 Database features	115
2.1.4 Find Organisms	117
2.1.5 Find Genomes	119
2.1.6 Find Genes	122
2.1.7 Discussion and future directions	124
2.1.8 The idea behind an update on SymGenDB	125
2.2. An update on the Symbiotic Genomes Database (SymGenDB): a collection of metadata, genomic, genetic, and protein sequences, orthologs, and metabolic networks of symbiotic organisms	126
2.2.1 Database contents—an overview	128
2.2.2 Module Organisms	129
2.2.3 Module Genomes	131
2.2.4 Module Genes	133
2.2.5 Increased data content	134
2.2.6 New module—MetaDAGs	135
2.2.7 Availability of web interface and services	142
2.2.8 Future directions	143
2.2.9 Statistics and demographic overview of SymGenDB	143
Chapter 3. The minimal metabolism	147
3. The Metabolic Building Blocks of a minimal cell	149
3.1. Inference of Minimal Metabolic Networks	154
3.2 Reconstruction of the Directed Acyclic Graph of metabolic networks	155
3.3 Theoretical Minimal Metabolic Network	156
3.4 The MetaDAG methodology: analysis of the composition and connectivity of a network at a glance	162
3.5 The m-DAG of “Candidatus Nasuia deltocephalinicola”	167
3.6 The first semisynthetic viable cell and its m-DAG’s reconstruction	169
3.7 Resemblance of the MBBs of the minimal m-DAGs	173
3.8 Conclusions	175
Chapter 4. Metabolic networks of insects' endosymbiotic bacteria preserve evolutionary traces	177
4. Evolutionary traces in metabolic networks of endosymbiotic bacteria of insects	181
4.1 Genomes of endosymbiotic bacteria of insects and their free-living relatives	184
4.2 Metabolic networks and metabolic DAGs	187
4.3 Topological parameters of the MNs	189

4.4 Comparative evolution of the genome and metabolism	190
4.5 Genome vs. metabolic evolution of endosymbiotic bacteria of insects	192
4.6 Convergence of the metabolite- and the reaction-based methods	195
4.7 Core m-DAG and pan-mDAG of endosymbiotic bacteria of insects	197
4.8 Smallest and biggest MNs of symbiotic bacteria of insects	198
4.9 Conclusions	205
VI. Conclusions	210
VII. Resumen en castellano	216
Introducción	216
Objetivos de la tesis	227
Materiales y Métodos	228
Resultados y Discusión	230
Conclusiones	235
VIII. References	242
IX. Appendices	274
<i>Appendix A - Original publication first-page reprints</i>	274
<i>Appendix B - Supplementary material</i>	282
<i>Appendix C - List of figures and tables in this dissertation</i>	314

I. Summary

Over the last couple of decades, symbioses between insects and bacterial endosymbionts have been the focus of remarkable empirical studies. Models of symbiosis between specific bacterial lineages and their hosts have been described, but to our knowledge, no large-scale analyses have been done in order to begin deciphering the overall evolutionary path of the endosymbiosis phenomenon. Insects represent about 85% of animal diversity, and about 60% maintain symbiotic relationships with microbes, which mainly allow their hosts to live in niches otherwise unavailable to them by providing them with nutrients, protection, and even new forms of energy. Bacterial endosymbionts often live within specialized cells in insects called bacteriocytes; they generally have a base compositional bias towards A+T in their genomes, undergo genomic shrinkage, and have an accelerated sequence evolution, all of which are convergent attributes with organelles resembling their extended and combined evolutionary histories; so, in this work, we have proposed the term "*symbionelles*" for long-term obligate bacterial endosymbionts of insects. Most of these organisms have the smallest genomes found in nature. This makes them good models for studying minimal cells through genomic and metabolomic analyses, which is the subject of two chapters of this thesis.

Recent changes in technology have made it necessary to look for new and creative ways to handle and process large amounts of data. With completely sequenced and annotated genomes from endosymbiotic bacteria of insects,

databases are indispensable tools for organizing and easily accessing specific biological information. We constructed and published a composite database that includes the genomic data of symbiotic relationships between bacteria and insects, as well as all symbiotic relationships found in primary databases since the process for making this database was the same for both. Our database includes the confirmation of the symbiotic relationships (validated in literature), the associated publication (original journal article link where the association was first described), the organization and availability of the sequences of all genes, genomes, and orthologs of each prokaryotic symbiont, and the metabolic network of all organisms included in this repository as an m-DAG, a new reaction-based methodology applied on genomic data to create contracted metabolic diagrams by connecting directed acyclic graphs and creating metabolic building blocks as nodes of a metabolic network. By comparing these reaction-based metabolic networks to standard metabolite-based ones, we were able to look at the differences and similarities between organisms that have evolved in different ways, such as being more, or less involved in endosymbiosis.

II. Publications from this thesis

- **Reyes-Prieto, Mariana**, Amparo Latorre, and Andrés Moya. "Scanty microbes, the 'symbionelle' concept." *Environmental Microbiology* 16.2 (2014): 335-338.
- **Reyes-Prieto, Mariana**, David J. Martínez-Cano, Esperanza Martínez-Romero, Laila P. Partida-Martínez, Amparo Latorre, Andrés Moya, and Luis Delaye. "Evolution of small prokaryotic genomes." *Frontiers in Microbiology* 5 (2015): 742.
- **Reyes-Prieto, Mariana**, Carlos Vargas-Chávez, Amparo Latorre, and Andrés Moya. "SymbioGenomesDB: a database for the integration and access to knowledge on host-symbiont relationships." *Database* 2015 (2015).
- **Reyes-Prieto, Mariana**, Carlos Vargas-Chávez, Mercè Llabrés, Pere Palmer, Amparo Latorre, and Andrés Moya. "An update on the Symbiotic Genomes Database (SymGenDB): a collection of metadata, genomic, genetic and protein sequences, orthologs and metabolic networks of symbiotic organisms." *Database* 2020 (2020).
- **Reyes-Prieto, Mariana**, Rosario Gil, Mercè Llabrés, Pere Palmer-Rodríguez, and Andrés Moya. "The Metabolic Building Blocks of a Minimal Cell." *Biology* 10, no. 1 (2021): 5.
- **IN REVISION: Reyes-Prieto, Mariana**, David J. Martínez-Cano, Mercè Llabrés, Pere Palmer, Carlos Vargas-Chávez, Luis Delaye, Rosario Gil and Andrés Moya. "Evolutionary traces in metabolic networks of endosymbiotic bacteria of insects".

During my doctoral studies, the following papers about collaborations in symbiosis research were published, but they are not in this dissertation:

- **Reyes-Prieto, Mariana**, Alejandro Ocegüera-Figueroa, Sara Snell, Alicia Negredo, Emilio Barba, Luis Fernández, Andrés Moya, and Amparo Latorre. "DNA barcodes reveal the presence of the introduced freshwater leech *Helobdella europaea* in Spain." *Mitochondrial DNA* 25, no. 5 (2014): 387-393.
- Lozano-Ojalvo, Daniel, Alicia Rodríguez, Mirian Cordero, Victoria Bernáldez, **Mariana Reyes-Prieto**, and Juan J. Córdoba. "Characterisation and detection of

spoilage mold responsible for black spots in dry-cured fermented sausages." *Meat Science* 100 (2015): 283-290.

- Alía, Alberto, María J. Andrade, Alicia Rodríguez, **Mariana Reyes-Prieto**, Victoria Bernaldez, and Juan J. Córdoba. "Identification and control of molds responsible for black spot spoilage in dry-cured ham." *Meat Science* 122 (2016): 16–24.
- Gutiérrez-Preciado, Ana, Carlos Vargas-Chávez, **Mariana Reyes-Prieto**, Omar F. Ordoñez, Diego Santos-García, Tania Rosas-Pérez, Jorge Valdivia-Anistro, et al. "The genomic sequence of *Exiguobacterium chiriqhucha* str. N139 reveals a species that thrives in cold waters and extreme environmental conditions." *PeerJ* 5 (2017): e3162.
- Zouari, Sana, Monia Kamel Ben Halima, **Mariana Reyes-Prieto**, Amparo Latorre, and Rosario Gil. "Natural occurrence of secondary bacterial symbionts in aphids from Tunisia, with a focus on genus *Hyalopterus*." *Environmental Entomology* 47, no. 2 (2018): 325-333.
- Cocco, Nicoletta, Mercè Llabrés, **Mariana Reyes-Prieto**, and Marta Simeoni. "MetNet: A two-level approach to reconstructing and comparing metabolic networks." *PloS one* 16, no. 2 (2021): e0246962.
- Saldierna Guzmán, J. Paola, **Mariana Reyes-Prieto**, and Stephen C. Hart. "Characterization of *Erwinia gerundensis* A4, an almond-derived plant growth-promoting endophyte." *Frontiers in Microbiology* 12 (2021): 687971.
- Garzón, María José, **Mariana Reyes-Prieto**, and Rosario Gil. "The Minimal Translation Machinery: What We Can Learn from Naturally and Experimentally Reduced Genomes." *Frontiers in Microbiology* 13 (2022).
- Smith, Gilbert, Alejandro Manzano-Marín, **Mariana Reyes-Prieto**, Cátia Sofia Ribeiro Antunes, Victoria Ashworth, Obed Nanjul Goselle, Abdulhalem Abdulsamad A. Jan et al. "Human follicular mites: Ectoparasites becoming symbionts." *Molecular Biology and Evolution* 39, no. 6 (2022): msac125.

The following papers were published during my doctoral studies as collaborations but are not related to the symbiosis research area nor included in this dissertation:

- Mengual-Chuliá, Beatriz, Andrés Alonso-Cordero, Laura Cano, M. del Mar Mosquera, Patricia de Molina, Roser Vendrell, **Mariana Reyes-Prieto**, et al. "Whole-genome analysis surveillance of influenza A virus resistance to polymerase complex inhibitors in Eastern Spain from 2016 to 2019." *Antimicrobial Agents and Chemotherapy* 65, no. 6 (2021): e02718-20.
- Zhang, Feiyu, Macarena Ferrero, Ning Dong, Giuseppe D'Auria, **Mariana Reyes-Prieto**, Alejandro Herreros-Pomares, Silvia Calabuig-Fariñas, et al. "Analysis of the gut microbiota: an emerging source of biomarkers for immune checkpoint blockade therapy in non-small cell lung cancer." *Cancers* 13, no. 11 (2021): 2514.
- Díaz-Regañón, David, Mercedes García-Sancho, Alejandra Villaescusa, Ángel Sainz, Beatriz Agulla, **Mariana Reyes-Prieto**, Antonio Rodríguez-Bertos, and Fernando Rodríguez-Franco. "Characterization of the Fecal and Mucosa-Associated Microbiota in Dogs with Chronic Inflammatory Enteropathy." *Animals* 13, no. 3 (2023): 326.
- Parra, Manuela, Rosa de los Ángeles Bayas-Rea, Teresa Guerrero, Stuart Torres, Giuseppe D'Auria, **Mariana Reyes-Prieto**, and Sonia Zapata. "Complete Genome Sequences of Two Lytic Phages of *Salmonella enterica*." *Microbiology Resource Announcements* (2023): e01048-22.
- José Miguel Sahuquillo-Arce, **Mariana Reyes-Prieto**, Alicia Hernández-Cabezas, José Miguel Molina-Moreno, Juan Antonio Saez-Nieto, María del Pilar Marín, María Isabel Alcoriza-Balaguer, Agustín Lahoz, Jaime Sanz, Lola González-Tarancón, María Loreto Ferrús, Llúcia Martínez-Priego, Adolfo Magraner-Martínez; José Luis López-Hontangas. "Starkeya nomas sp. nov., a prosthecate and budding bacterium isolated from an immunocompromised patient." *International Journal of Systematic and Evolutionary Microbiology* (2023), IN PRESS.

III. Introduction

1. Symbiosis throughout the Tree of Life and as an evolutionary driving force

It has been almost a century and a half since the term “symbiosis” (derived from the Ancient Greek *συμβίωσις*, *symbiōsis*, "living together", from *σύν*, *sýn*, "together", and *βίωσις*, *bíōsis*, "living") was first introduced by Heinrich Anton de Bary in his work “Die Erscheinungen des symbiose” where he described it as “the living together of dissimilarly named organisms” (de Bary, 1879). From the initial interest in these relationships, microbes were equivalent to illness, but the focus slowly changed as more non-pathogenic relationships across all kingdoms of life were discovered, recognizing them as significant factors driving evolution (Douglas, 2014). After much debate, researchers now agree with the term as an intimate, outcome-independent interaction between species, including mutualistic, parasitic, and commensal associations (Saffo, 1993). Terms that address particular outcomes for the host or symbiont, *e.g.*, costs, benefits, fitness, and so on, are also considered in the language of symbiosis research, but defining the type of interaction and the consequences of these is a much more difficult task (Martin & Schwab, 2012).

Symbiosis has been observed throughout the Tree of Life in many forms, showing different degrees of the relationships’ dependence (Figure 1). It has been present for millions of years, probably since the beginning of life itself, and has become a major driver of evolution in the development of life as we know it. These relationships are fundamental for the functioning of ecosystems and have

been crucial for the development of biological complexity. Symbionts allow their hosts to live in natural niches otherwise unavailable to them by supplying nutrients, providing protection, and even helping them to use alternative energy sources (Kleiner et al., 2012; Moran et al., 2003). For instance, there are some relationships between chemosynthetic bacteria and marine animals that live in the deep sea, where very few nutrients are available, where these symbionts allow the animals to grow on inorganic energy and carbon sources like sulfide and CO₂ (Kleiner et al., 2012). As another example, plants have mycorrhizal fungi that control how much micro- and macronutrients they can get from the soil (Bati et al., 2015), and nitrogen-fixing bacteria (Rhizobia) that make nodules in the plants they live in, reproduce inside them and change nitrogen in the air into ammonium for the benefit of their host (Rogel et al., 2011). Furthermore, some resistance to herbivory has been determined to be associated with the fungal endophytes that live in leaves (Clay, 1988). As for humans, approximately 50% of the cells in our body are known to be microbial (Sender et al., 2016). The human microbiota has been described to play an important role in basic biological processes and in the development and progression of major human diseases (Turnbaugh et al., 2007; Wang et al., 2017). Gut microbiota also allows mammals to thrive on otherwise toxic diets by degrading toxic plant secondary compounds (Kohl et al., 2014). Marine invertebrates are dependent on chemosynthetic symbiosis with bacteria in the deep sea, cold swells, whale and food falls, shallow-water coastal sediments, and continental margins (Dubilier et al., 2008). And as a last example, insects represent about 85% of animal diversity

and about 60% maintain symbiotic relationships with microbes; of those, 15-20% harbor primary endosymbionts, *i.e.*, bacteria that are strictly dependent on their host for survival (and vice versa) (Feldhaar & Gross, 2009; Gil & Latorre, 2019).

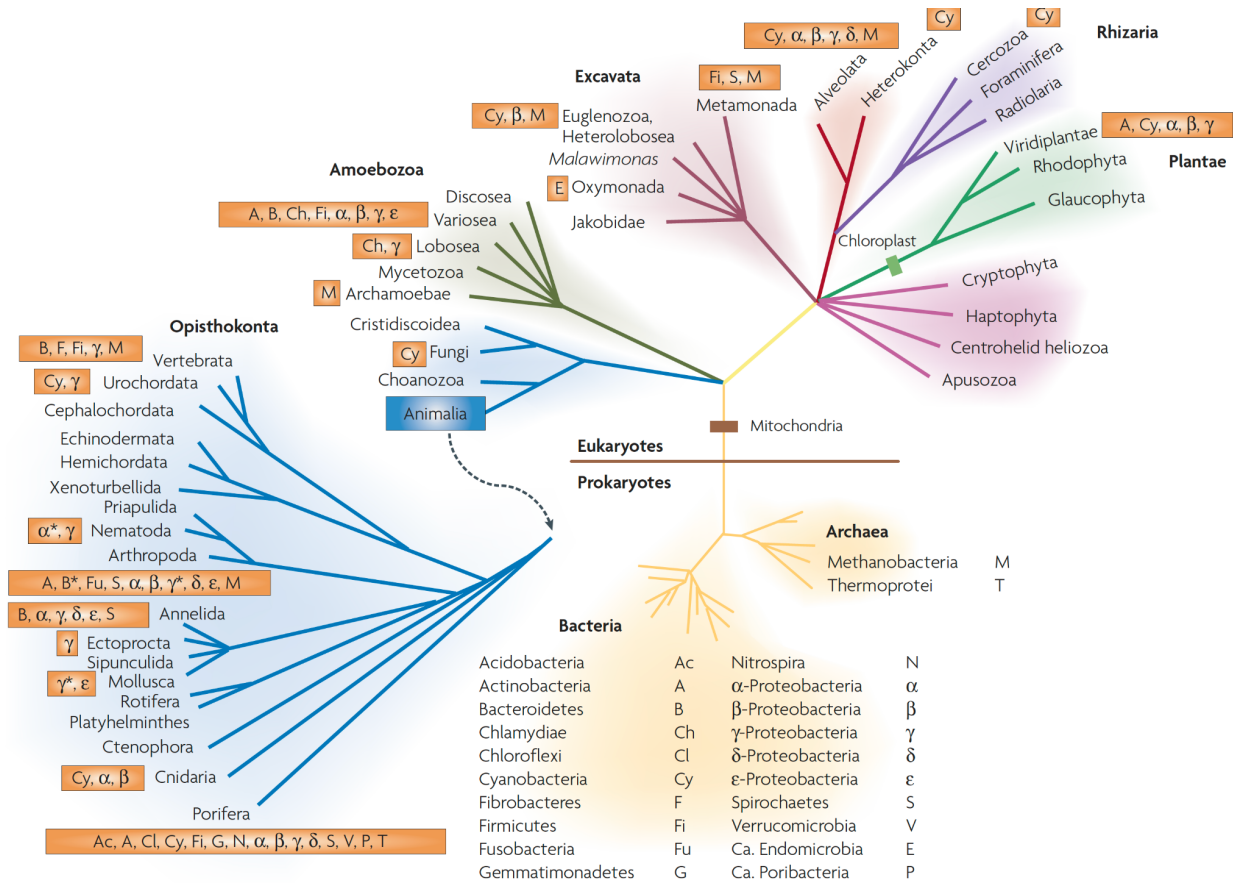


Figure 1. Overview of stable symbioses throughout the Tree of Life (Moya et al., 2008). Reprinted with permission from Springer Nature and the Copyright Clearance Center with license number 5433131267927.

These examples only contemplate symbiosis between eukaryotes and prokaryotes, but many more types, including eukaryote-eukaryote interactions, as well as with archaea participants, are also common (Allemand & Furla, 2018; Martinson, 2020; Scannerini et al., 2013; Wrede et al., 2012). Thus, it is safe to say

that the convergence of symbiosis evolution throughout the tree of life, in different forms and ecosystems, denotes the importance of coexisting, sharing, and supporting each other for the greater good, allowing the prevalence of life and the stability of communities throughout time.

2. Symbiosis in the insect world

As previously stated, insects are among the most successful animals on Earth regarding their biomass and biodiversity, calculations range between four and ten million existing species (Miller et al., 2002). They are considered the most successful taxonomic class of multicellular organisms and are found thriving in all terrestrial ecosystems, including the most extreme. Their evolutionary success may be largely explained by their capability to engage symbiotic bacteria (Dimijian, 2000). The notion of bacteria living within insect tissue (*i.e.*, 'endosymbiotic' bacteria) dates back to the 1930s (Mansour, 1934), and modern investigation techniques have unraveled that endosymbiosis is the rule rather than the exception in insects, contrary to what was first believed. About half of the insect world harbors endosymbiotic bacteria within their tissues, and about 20% directly rely on endosymbiotic bacteria to ensure their development and reproduction (Batra & Buchner, 1968; Douglas, 1989; Gil & Latorre, 2019; Medina et al., 2020; Moran & Telang, 1998).

Symbiotic associations in insects can be defined depending on the localization of the symbiont, the effects of their symbiotic relationship, and the interrelatedness of the organisms. The most common classification is based on

the level of integration of the bacterial symbiont into the host's physiology, distributing them between obligate and facultative endosymbionts, depending on whether the host can survive with or without the symbiont (Guo et al., 2017; Wilkinson et al., 2007). Furthermore, facultative endosymbionts can become obligate within a given time and have different outcomes susceptible to the dependence of the host on the symbiont (Husnik & Keeling, 2019). Facultative symbionts can be horizontally or vertically transmitted. In contrast, obligate ones are all vertically transmitted, usually by transovarial transfer from mother to offspring within female germ cells for intracellular species, and by brood contamination for extracellular species (Szklarzewicz & Michalik, 2017). Both obligate and facultative symbionts have a maintenance toll on their hosts but also lead to many ecological advantages.

The symbiosis between insects and bacterial endosymbionts has been the focus of remarkable empirical studies for the last couple of decades, and some general characteristics have been described for such associations (summarized here, but in detail in the following sections of this introduction):

- Convergence in the appearance of insect-bacteria symbiosis in the evolutionary history of these systems has allowed insects the novel acquisition of beneficial functions (Sudakaran et al., 2017), which have frequently derived in their adaptation to diverse ecosystems (Mitter et al., 1988; Raffa et al., 2008).

- ➔ Bacteria supply a broad range of advantageous functions to insects, some of which are resistance to stress (Heyworth & Ferrari, 2016), protection against other organisms (Kaltenpoth et al., 2005; Oliver et al., 2003), insecticide resistance (Kikuchi et al., 2012), and the most common and important, the supply of essential and non-essential nutrients such as amino acids and vitamins (Baumann, 2005; Douglas, 1998, 2016; Moran et al., 2003).
- ➔ Endosymbiotic bacteria of insects go through a genome-reduction process when becoming endosymbionts as a result of the strict vertical transmission within host lineages (Latorre & Manzano-Marín, 2017; Toft & Andersson, 2010), often becoming completely dependent on their host (Fisher et al., 2017).
- ➔ Resemblances between endosymbionts and pathogens include the loss of metabolic diversity and DNA repair functions (Wernegreen, 2017).
- ➔ Bacterial endosymbionts often, but not always, live within specialized cells in insects called bacteriocytes (Alarcón et al., 2022).
- ➔ There is an accelerated sequence evolution and in most described cases, a base compositional bias towards A+T in endosymbiotic bacteria of insects (Rispe et al., 2004; Sabater-Muñoz et al., 2017; Van Leuven & McCutcheon, 2012; Wernegreen, 2015).

→ The natural occurrence of more than one endosymbiotic bacteria lineage in an insect host has also been described (*i.e.*, co-symbionts), where each partner of the symbiosis plays its own important role (McCutcheon & Moran, 2010; Nakabachi et al., 2013). Even metabolic complementation between co-symbionts has been observed and is of vital importance to the survival of its host (Gosalbes et al., 2008; Ponce-de-Leon et al., 2017). Furthermore, complete endosymbiotic replacements have been observed in some cases (Gil & Latorre, 2019; Sudakaran et al., 2017).

3. Intracellular obligate endosymbionts, organelle-like organisms, symbionelles... Are we there yet?

An organelle is a specialized cellular part that has a specific function and is considered analogous to an organ. There is now compelling documentation that indicates that all eukaryotes are the product of a symbiosis between a primitive eukaryote and an intracellular bacteria, specifically belonging to the *Rickettsiales*, that evolved into the mitochondria (Gabaldón, 2018; Margulis, 1970; Williams et al., 2007), an organelle essential for the production of energy necessary for a eukaryotic cell's survival and function. This is not an isolated case, as evidence indicates that another major symbiotic event between a eukaryote and a cyanobacterium gave rise to the chloroplast, an organelle in algae and plants that captures sunlight and transforms carbon dioxide into carbohydrates, releasing oxygen, and splitting water, *i.e.*, is responsible for the process of photosynthesis (Bonen & Doolittle, 1975; McFadden, 2001; Sánchez-Baracaldo et al., 2017). The

origins of these organelles, the mitochondria and the chloroplast (among other plastids that are now considered cellular organelles of modern eukaryotes), which started as bacterial symbionts, are considered part of the most significant events in the history of modern living beings. Mitochondria are indispensable for the functioning of the eukaryotic cell, and the Earth's biosphere is sustained by photosynthesis. Without the 'powerhouse' of cells or oxygenic photosynthesis, the evolution of life on earth would be, to say the least, different.

There are many reasons why free-living bacteria evolve into symbiotic bacteria and, eventually, into intracellular obligate symbionts. Interestingly, some of the attributes of these endosymbionts are also considered indispensable characteristics of organelles. As briefly mentioned in the previous section of this work, there is a strong dependence on their host because they thrive on nutritionally unbalanced diets. Also, intracellular obligate endosymbionts have become necessary for their host's development in their environmental conditions and have total prevalence in the host insect populations (Douglas, 1989; Moran & Telang, 1998). Furthermore, endosymbionts are sequestered inside bacteriocytes, they have a strict vertical transmission, and most of those studied supplement the insects' diets by providing essential amino acids and/or vitamins, or by recycling nitrogen in exchange for energy supply and other nutrients (e.g., nonessential amino acids) (Douglas, 2016; Patiño-Navarrete et al., 2014; Ponce-de-Leon et al., 2017; Skidmore & Hansen, 2017).

Other characteristics of obligate endosymbiotic organisms are that their genomes go through an extreme genome reduction process and end up small in size, one of the most extreme known cases being '*Candidatus* Nasonia deltocephalinicola' *NAS ALF* with a 102 kb genome (Bennett & Moran, 2013). These extremely reduced genomes are massively gene-dense even with overlapping genes, increased ortholog length variation, and loss of large accessory proteins, and most of them have a bias toward A-T bases (Figure 2) (Gil et al., 2002; Kenyon & Sabree, 2014; McCutcheon & Moran, 2012; Moran & Bennett, 2014; Nowack, 2014). All these features correlate with their intracellular lifestyle. By becoming endosymbiotic the bacteria are in a highly stable environment, with selective pressure only acting upon genes that are beneficial for the symbiotic interaction. Mutations that are deleterious or neutral for the specific interaction, and also those that the bacteria does not rely on anymore due to the absence of changes in its new niche, are slowly pseudogenized or excised by large deletions and chromosomal rearrangements (Moran & Mira, 2001; Sabater-Muñoz et al., 2017). Furthermore, among these losses of genetic material, there are significant losses of informational genes such as repair genes, translation factors, tRNAs, rRNAs, RNA modification genes, and ribosomal proteins (Bennett & Moran, 2013; Garzón et al., 2022; Husnik & McCutcheon, 2016; McCutcheon et al., 2009b). Also, the spatial delimitation of the bacteria ensures no possibility of recombination or horizontal gene transfer, losing the possibility of exchanging genetic material with other unrelated bacterial lineages (McCutcheon & Moran, 2012).

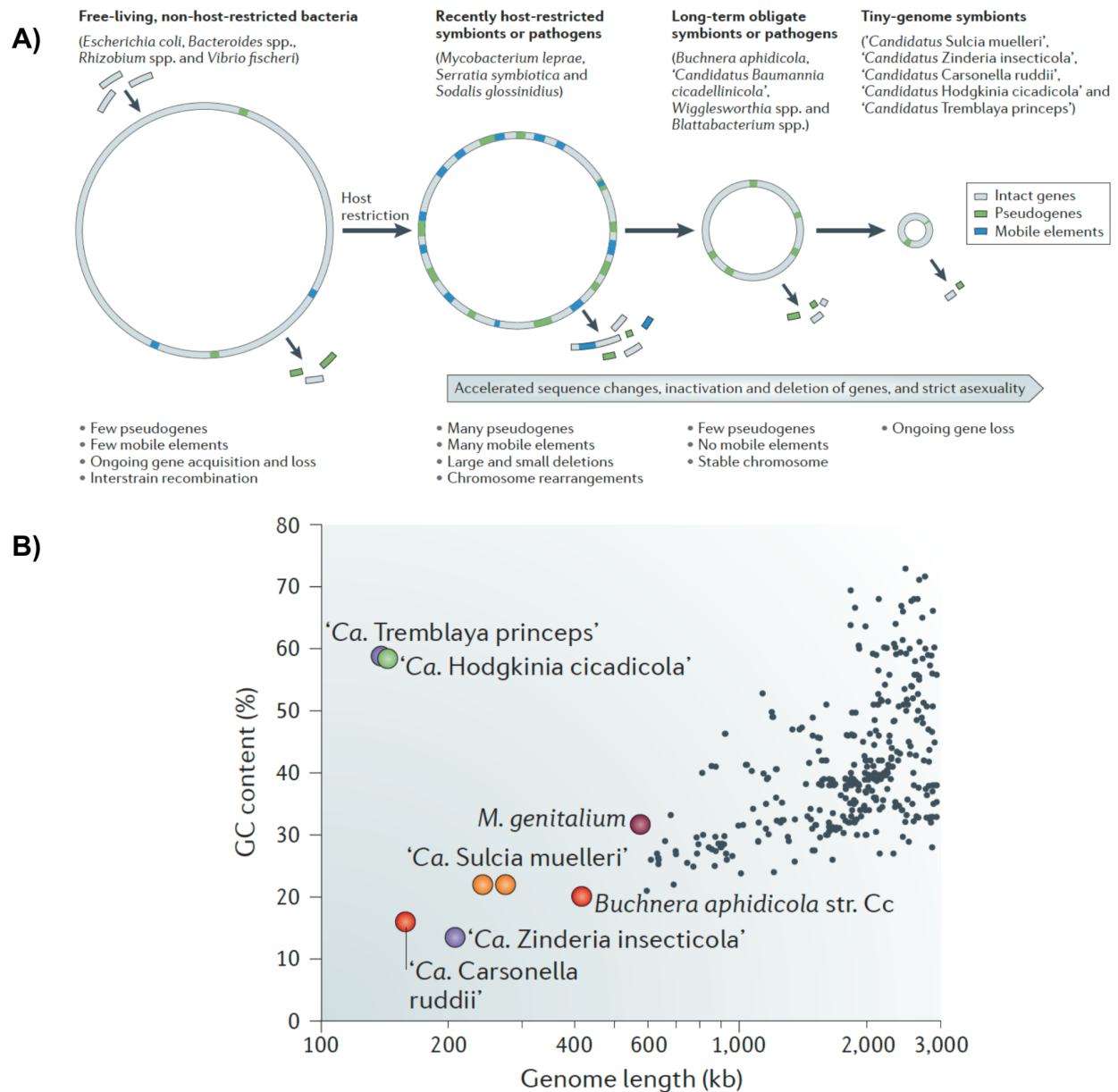


Figure 2. A) The genome reduction process of obligate intracellular endosymbionts, and B) Genome size versus GC content of endosymbiotic bacterial genomes from McCutcheon & Moran (2012). Reprinted with permission from Springer Nature and the Copyright Clearance Center with license number 5433140857649.

Moreover, long-term functional integration of many functions has been observed in symbiotic relationships. Not only as cross-feeding, where the exchange of essential components between host and endosymbiont has been observed (amino acids, vitamins, ATP, sugars, nucleotides, etc.) (Douglas, 2016; Duncan et al., 2014; Hansen & Moran, 2011; McCutcheon & Moran, 2012) but also as a complex scenario where pathways can be incomplete and shared between the symbiotic partners. For example, within sap-feeding insects with co-endosymbionts (two or more bacterial lineages living within their bacteriocytes), metabolic complementation of several biosynthetic pathways, such as those for tryptophan, biotin, tetrahydrofolate biosynthesis, and even energy metabolism has been described (Pérez-Brocal et al., 2006; Ponce-de-Leon et al., 2017; Van Leuven et al., 2014). These examples make it evident that all participants of the symbiosis contribute to a single pathway, meaning that the level of integration must involve the transport of enzymes and/or metabolites across intermediate membranes, although most of such mechanisms are still unclear.

Another commonly required attribute for organelles is that they have transferred genes to their host, and become dependent on a dedicated targeting system to re-import their protein products (Cavalier-Smith & Lee, 1985; Keeling & Archibald, 2008); *i.e.*, the organelle-like endosymbiont rely on its host to keep the essential genetic informational machinery. Studies have shown that this phenomenon is also common among endosymbiotic bacteria, specifically and experimentally proven in *Buchnera*, chromatophore, and *Kinetoplastibacterium*

(Alves et al., 2013; Morales et al., 2016; Nowack et al., 2016; Sloan et al., 2014). Not only that but it has also been demonstrated that they rely on proteins encoded by the host genomes that have a bacterial origin, something that has been observed in the endosymbiotic lineages of *Nasuia*, *Sulcia*, *Tremblaya*, *Buchnera*, *Carsonella*, *Portiera*, among others (Kelly, 2021; López-Madriral Sergio et al., 2011; Mao et al., 2018; Nakabachi et al., 2014; Nikoh et al., 2010; Sloan et al., 2014).

4. A vast necessity for the organization and compilation of symbiotic genomic information

Due to the huge technological advances in recent years, specifically in the area of genome sequencing, with first, second, and third-generation sequencers available and producing an immense amount of data, computer databases are indispensable tools for organizing and easily accessing biological information. Multiple databases offer information on organisms' genomes and genes, such as NCBI (Wheeler et al., 2007); <https://www.ncbi.nlm.nih.gov/>) or ENA (Leinonen et al., 2011); <https://www.ebi.ac.uk/ena/browser/home>), which have grown exponentially since the beginning of the sequencing era (Figure 3).

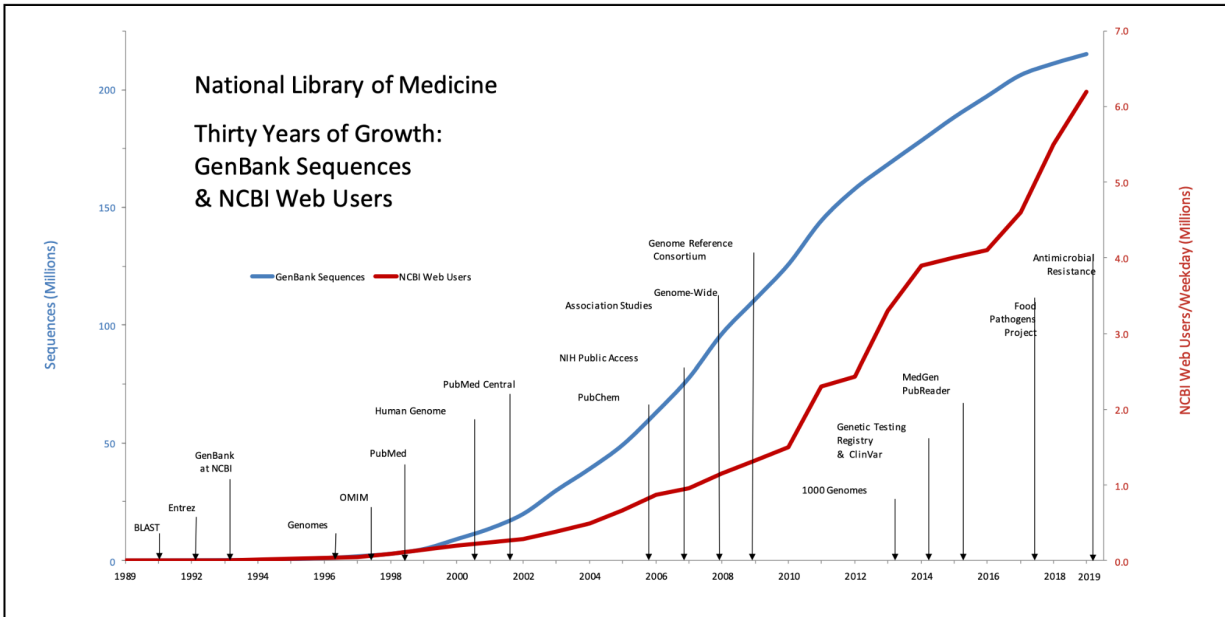


Figure 3. Growth of GenBank sequences (Benson et al., 2017) and NCBI (Wheeler et al., 2007) web users through 2019. Figure from the Department of Health and Human Services, NIH. Website (https://www.nlm.nih.gov/about/2021CJ_NLM.pdf).

These are primary databases, also known as archival databases, because they archive the experimental results submitted by scientists, and, very importantly, the archived data are not curated (modified or reviewed in detail) by other peers. This means that the data are obtained at a laboratory and made accessible to users without any change (Selzer et al., 2018).

Another type of database is a secondary database, where data stored are the analyzed result of the uncurated data obtained from primary databases. The raw data are submitted through pipelines and workflows to become meaningful and informative data for specific purposes. The information stored in secondary databases is highly curated (i.e., processed) and arguably contains more valuable information than primary databases. Some examples include InterPro ((Hunter et

al., 2009); <https://www.ebi.ac.uk/interpro/>) or UniProt ((UniProt Consortium, 2019); <https://www.uniprot.org/>) databases, which include motifs, domains, sequence families, and functional information on proteins.

The last type of biological database refers to composite or tertiary databases. The data entered in these databases are compared and then filtered on desired criteria. Basically, data are taken from primary databases and then merged based on certain conditions and are highly curated. They are very helpful for searching specific queries and sequences rapidly because they contain non-redundant data.

At the beginning of the studies presented in this thesis, to our knowledge, there were no secondary or composite databases available for endosymbiotic bacteria of insects or, nonetheless, for any large number of symbiotic organisms publicly available. Such big data repositories for organisms in these relationships would be of great interest to scientists from different research areas working on symbiotic genomics.

5. Minimal cells and minimal metabolisms' analysis

Knowing, describing, and understanding every molecular process in a living cell would give us an idea of the fundamental principles of life. It would open many possibilities for applied sciences; one of the most ambitious and important would be to design and create artificial organisms. Since this is not possible yet, and we are nowhere near knowing everything there is to know about genomic data in every organism on the planet, natural minimal cells can help to guide us through

the needs for this venture since they contain only the smallest amount of information needed for their survival. To this end, intracellular obligate symbionts with extreme genome reduction are among the main studied organisms to aid in the goal of successfully synthesizing minimal cells. Gil and coworkers (2014) defined in our group the core of a minimal gene set needed for a minimal bacterial cell. These authors performed a comprehensive analysis of computational and experimental attempts to define a minimal genome by comparing free-living, parasitic, and mutualistic endosymbiotic bacteria (Gil et al., 2004). Glass and coworkers, from the group of Craig Venter, also ventured into this topic and presented their list of essential genes for bacteria in 2006 (Glass et al., 2006), taking into account the previously mentioned work. Later, both lists of genes were explored by Gabaldón and coworkers to propose a network depicting an inferred minimal metabolism needed for life (Gabaldon et al., 2007).

The metabolism of an organism is the chemical system that produces the essential molecular components for life, and it is completely correlated with the genome of a given organism. This metabolism consists of metabolic pathways for the biosynthesis of amino acids, nucleic acids, lipids, carbohydrates, and so on, for the optimal functioning of cellular processes, and the enzymes needed for such functions are encoded in the genome. Even for minimal cells, natural or synthetic, the depiction of metabolism by conventional methods (i.e., tracking and mapping all components of a genome in metabolic pathway maps) has broad implications for its computational examination and visual plotting. We, as

humans, are not able to process or grasp the amount of information included in those models. To alleviate this problem, Alberich and coworkers developed a methodology called MetaDAG (Alberich et al., 2017; an open-source web page with the software tool available to any user is pending) based on graph representation to study the robustness, modularity, and connectivity of metabolic networks. They used directed graphs because of their simplicity and their ability to model topological information of the networks. Also, because they allow to easily depict strongly connected components of those directed graphs, which reduces the metabolic networks to study the pieces easily and, to easily visualize the topology as a whole. They introduced the term *metabolic building blocks (MBBs)* for strongly connected components of the metabolism being studied, and *m-DAGs* for the directed acyclic graphs resulting from the union of MBBs by the enzymes present in said metabolism. The MetaDAG methodology provides all the information of interest within the networks but also reduces their size to facilitate their analysis and visualization; therefore, it can be applied to specific symbiotic organisms, to theoretically defined minimal metabolisms such as the one resulting from (Gabaldon et al., 2007), and to large-scale conjoint analyses of a set of endosymbiotic bacteria of interest.

6. Evolutionary traces in metabolic networks of endosymbiotic bacteria of insects

With the availability of many completely sequenced and annotated genomes from endosymbiotic bacteria of insects, as well as by applying different methods and tools to such data, we can begin to elucidate some more specific questions on

the process of symbiotic integration within eukaryotes, and how its evolution has taken place begins to become a little clearer (Borenstein & Feldman, 2009; Martínez-Cano et al., 2014). By reconstructing metabolic networks from endosymbionts' genomes, we can comprehensively understand the functional properties of organisms by themselves and then, with comparative metabolomics, review the variations and differences between organisms in different evolutionary contexts (for example, being for more or less time involved in an endosymbiosis). This metabolomic comparison focuses on the differences between the nodes and edges of metabolic networks of each organism, or a collection of organisms; these differences are given through evolution, conditioned by the selective pressures acting upon the organism(s) genome, owing to distinct topological features (Yamada & Bork, 2009).

For endosymbiotic bacteria of insects, the evolutionary context relies on the dependence on their host, which includes the isolation of the bacteria in bacteriocytes or in specific host tissue (Moran & Telang, 1998), the biochemical environment to which they are tied (Hoffmeister & Martin, 2003), the metabolic dependence to the host and/or other cosymbionts (Gosalbes et al., 2008; Rao et al., 2015), or a combination of these phenomena, which added to the evolutionary rates of their genes determine their metabolic networks and crucially act on their composition.

It is very likely that symbiotic partnerships provide a distinctive signature because of their interconnected and highly reduced metabolic networks, which could then result in particular ecological, topological, and dynamic patterns.

Additionally, it may be possible to track the evolution of metabolic networks on a phylogeny thanks to the availability of taxonomically related bacterial genomes with comparable evolutionary constraints. This method can help to clarify the evolutionary processes and constraints that have an impact on metabolic network evolution (Mithani et al., 2010).

*“The most important discoveries of the laws methods and progress of Nature have nearly
always sprung from the examination of the smallest objects which she contains.”*

J. B. Lamarck, 1809

IV. Objectives

The main goal of this thesis was to search for evolutionary patterns in the reduced genomes of endosymbiotic bacteria of insects and to explore if convergent evolution is a common feature of their long-term evolution through the use and development of bioinformatics tools, comparative genomics, and the study of the properties of their metabolic networks.

Specific objectives included:

- 1) Review and proposal of the new concept *symbionelle*, based on the notion of intracellular bacteria of insects becoming organelles.
- 2) Construction of a public composite database of symbiotic genomes, with all genomic information available for download, including genes, genomes, orthologs, and metabolic reconstructions, in order to study these relationships at a large scale.
- 3) Development of specific public software tools for big-data analyses on endosymbiotic genomes, for analyzing symbiotic metabolic networks, and new methodologies on network reconstructions.
- 4) Comparative metabolomics: development of experiments *in silico* for coevolution testing and to compare the modularity of metabolic interactions between symbiotic bacteria of insects.

V. General Materials and Methods

The present thesis involved extended bibliographic research to propose an adequate conceptual framework, plus the use and development of bioinformatic tools for the analysis of endosymbiotic bacteria of insects' big data (Table 1).

Table 1. List of the bioinformatic software that has been employed for this dissertation and their applications. The detailed explanation of each purpose and the parameters used in each case are described within every chapter of this thesis. The asterisks (*) denote the software we have developed/or that is under development by us.

Software	Use	Website	Software references
R	Implementation of custom R scripts, for analysis, figures, and construction of the database background and link tables for chapters 1, 2, 3, and 4. Applied in Reyes-Prieto et al., 2014, 2015; Reyes-Prieto, Gil, et al., 2020; Reyes-Prieto, Vargas-Chávez, et al., 2020. Next publication in revision.	https://www.r-project.org/	Team & Others, 2008
SymGenDB*	Composite database we created for the organization and compilation of data of organisms in symbiotic relationships, used in chapters 2, 3, and 4. Applied in Reyes-Prieto et al., 2015; Reyes-Prieto, Gil, et al., 2020; Reyes-Prieto, Vargas-Chávez, et al., 2020. Next publication in revision.	http://symbiogenomesdb.uv.es/	Reyes-Prieto et al., 2015; Reyes-Prieto, Vargas-Chávez, et al., 2020
NCBI	Primary database where original research for SymGenDB was done for chapter 2.	https://www.ncbi.nlm.nih.gov/	Wheeler et al., 2007
GOLD	Primary database where original research for SymGenDB was done for chapter 2.	http://www.genomesonline.org/	Mukherjee et al., 2017
IMG	Primary database where original research for SymGenDB was done for chapter 2.	https://img.jgi.doe.gov	Markowitz et al., 2012
KEGG	Database used to download the	https://www.genome.jp/kegg/	Kanehisa & Goto, 2000

	metabolic information of organisms for chapter 2, 3 and 4.		
Microbial Genomes Database	Database used to download information for our database, used in chapter 2.	https://mbgd.nibb.ac.jp/	Uchiyama et al., 2019
Shiny	Construction of the database's web interface, chapter 2. Applied in Reyes-Prieto et al., 2015; Reyes-Prieto, Vargas-Chávez, et al., 2020.	https://shiny.rstudio.com/	RStudio, 2013
MetaDAG*	Tool to create compressed metabolic networks, used for analysis in chapters 2, 3, and 4.	Publication in process.	Based on the methodology by Alberich et al., 2017
PhyloPhlan	Software was used to create a phylogenetic tree for chapter 4.	https://huttenhower.sph.harvard.edu/phylophlan/	Segata et al., 2013
Mr.Bayes	Software used to construct phylogenetic trees for chapter 4.	https://nbisweden.github.io/MrBayes/	Hulslenbeck & Ronquist, 2001; Ronquist & Huelsenbeck, 2003
Cytoscape	Software was used to construct the metabolite-based metabolic networks of endosymbiotic bacteria of insects for chapter 4.	https://cytoscape.org/	Shannon et al., 2003
Cytoscape Network Analyzer	Software to calculate the topological parameters of the metabolic network for chapter 4.	http://manual.cytoscape.org/en/stable/Network_Analyzer.html	Doncheva et al., 2012
MUSCLE	Software was used in chapter 4 for the alignment of sequences.	https://github.com/rcedgar/muscle	Edgar, 2004
MAFFT	Software was used in chapter 4 for the alignment of sequences.	https://mafft.cbrc.jp/alignment/software/	Katoh & Standley, 2013
FastTree	Software was used in chapter 4 for the construction of phylogenetic trees.	http://www.microbesonline.org/fasttree/	Price et al., 2010
FigTree	Software to visualize and edit phylogenetic trees for chapter 4.	http://tree.bio.ed.ac.uk/software/figtree/	No publication available
BayesTraits	Software was used to measure coevolution of parameters of metabolic networks and genome size of endosymbiotic bacteria of insects in chapter 4.	http://www.evolution.reading.ac.uk/BayesTraitsV3/BayesTraitsV3.html	Pagel & Meade, n.d.
JModelTest2	Software was used to find the best evolutionary model for the construction of phylogenetic trees in chapter 4.	https://github.com/ddarriba/jmodeltest2	Darriba et al., 2012

VI. Results

Chapter 1. Conceptual framework: Scanty microbes, the "symbionelle" concept, and the evolution of small prokaryotic genomes

This chapter reproduces the following published papers in their entirety:

Reyes-Prieto, Mariana, Amparo Latorre, and Andrés Moya. "Scanty microbes, the 'symbionelle' concept." *Environmental Microbiology* 16.2 (2014): 335-338.
doi:10.1111/1462-2920.12220

Reyes-Prieto, Mariana, David J. Martínez-Cano, Esperanza Martínez-Romero, Laila P. Partida-Martínez, Amparo Latorre, Andrés Moya, and Luis Delaye. "Evolution of small prokaryotic genomes." *Frontiers in Microbiology* 5 (2015): 742.
doi:10.3389/fmicb.2014.00742

1.1. Opinion article: Scanty microbes, the 'symbionelle' concept

Mutualistic symbiosis occurs when two different species interact closely with each other and benefit from living and working together. However, not all symbiotic associations are of mutual benefit because there are also forms of parasitism (when one organism benefits but the other is adversely affected) and commensalism (when only one of the organisms involved in the association benefits, but the other is not affected); notwithstanding, the very fact that specific entities can exist together means that natural selection may guide them to live with each other. Endosymbiosis is a special case of symbiosis in which one partner, generally a prokaryote symbiont, lives sequestered inside specialized eukaryotic cells called bacteriocytes.

The notion of microbes becoming organelles of eukaryotic systems through evolution has been widely accepted because Lynn Margulis put forward her serial endosymbiotic theory of eukaryotic cell evolution (Margulis & University of Massachusetts Amherst Massachusetts Lynn Margulis, 1993). Indeed, this is the origin of mitochondria and chloroplasts. There is compelling evidence to support that these two eukaryotic organelles are the product of symbiotic events between prokaryotes and primitive eukaryotes (Latorre et al., 2011). Their original alpha-proteobacterial (mitochondria ancestor) and cyanobacterial (chloroplast ancestor) genomes have been drastically reduced, with a portion of the protein-encoded genes and even RNA genes being transferred to the eukaryotic nuclear genome. Other genes have simply been lost, and their function is replaced by the hosts. Since the proposal of these two

canonical endosymbioses, symbiotic associations between prokaryotes and unicellular and multicellular eukaryotes have been documented in practically every major branch of the tree of life, which reinforces the role played by symbiosis in the emergence of evolutionary innovations (Moya et al., 2008).

Endosymbiosis in insects is a captivating example of the aforementioned phenomenon. Insects are particularly well suited to establishing intracellular symbiosis with bacteria, which provides them with the metabolic capabilities they lack and enables them to live in almost any environment. At present, there are several well-documented cases of insect endosymbionts at different stages of symbiotic integration (Figure C1.1). Insect endosymbiosis commonly consists of an obligate mutualistic association, where bacteria produce essential nutrients that are absent in the insect's diet, and the insect, in turn, provides the bacteria with a safe environment and permanent food supply (Baumann, 2005). These endosymbiotic bacteria are vertically transmitted across host generations. Their metabolic role is renowned, and most insect endosymbiotic systems are largely convergent towards these functions regardless of the lifestyle or genomic repertoire of their free-living ancestor (López-Sánchez et al., 2008; McCutcheon et al., 2009b; McCutcheon & Moran, 2010; Sabree et al., 2013). A new symbiotic relationship, which represents a source of novel complexity, must overcome the obvious problem posed by the fact that both partners must be able to survive together despite differences in biology, particularly generation times and reproduction. Moreover, considering that these organisms generally possess different population genetics and are under different evolutionary pressures,

they need to establish a certain trade-off to acquire the evolutionary novelty represented by their stable coexistence (Delage & Moya, 2010; McCutcheon & Moran, 2012). Thus, important genetic and biochemical modifications are required in these bacteria compared with their free-living state. The eukaryotic host, on the other hand, must develop ways of controlling the bacterial population, engulfing them in specialized cells –the aforesaid bacteriocytes – and/or changing immune responses to recognize these bacteria as non-pathogenic.

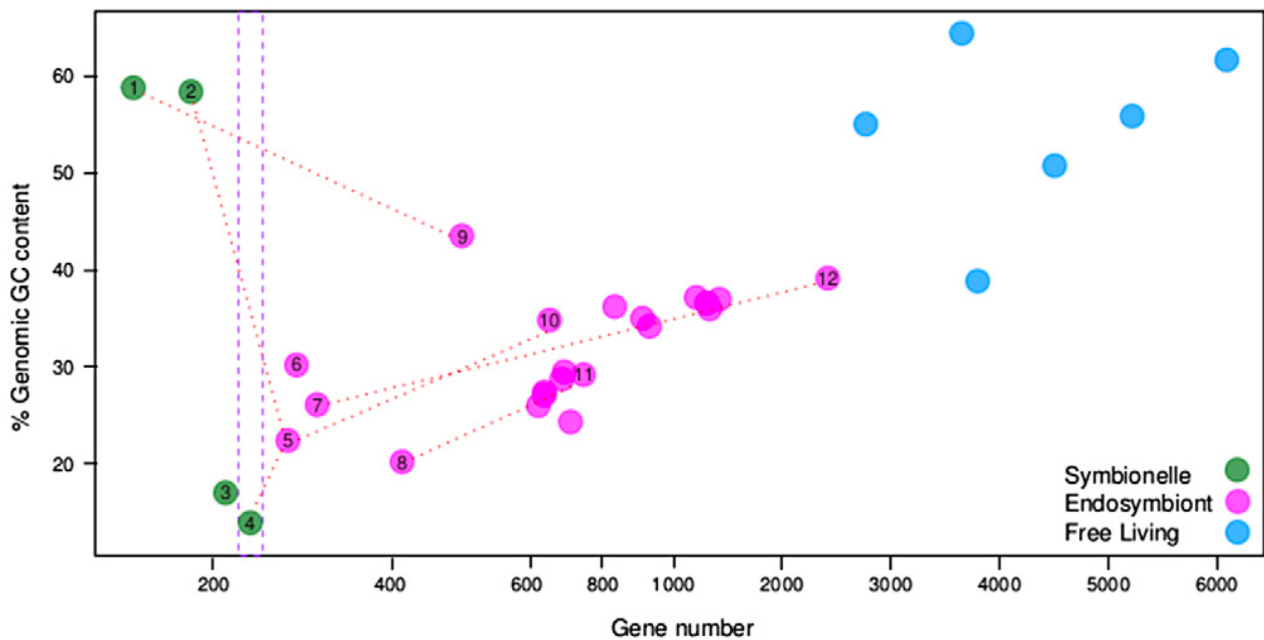


Figure C1.1 | Genomic GC content (%) versus gene number in several symbionelles, endosymbionts, and free-living bacteria. The two dashed purple vertical lines delimit the minimal gene set between 223 and 244 genes. The dotted red lines indicate pairs of symbiotic associations. Symbionelles: (1) '*Ca. Tremblaya princeps*' PCVAL, (2) '*Ca. Hodgkinia cicadicola*' DSEM, (3) '*Ca. Carsonella ruddii*' PV, (4) '*Ca. Zinderia insecticola*' CARI. Endosymbionts: (5) '*Ca. Sulcia muelleri*' GWSS, (6) '*Ca. Uzinura diaspidicola*' str. ASNER, (7) '*Ca. Portiera*

aleyrodidarum' BT-QVLC, (8) *B. aphidicola* BCc, (9) 'Ca. Moranella endobia' PCIT, (10) 'Ca. Baumannia cicadellinicola' str. Hc, (11) *S. symbiotica* str. Cc and (12) 'Ca. Hamiltonella defensa' 5AT.

One of the most important and well-known features of endosymbiotic bacteria is that they provide extreme examples of genomic shrinkage by undergoing a process called the 'genomic reduction syndrome'. Hence, prokaryotic genomes of endosymbionts are examples of a particular type of naturally evolved minimal cell, with insect endosymbionts being those with the smallest genomes reported to date (Figure C1.1). In fact, as much as 90% of their genetic matter can be lost in advanced or extreme cases of symbiotic integration.

It is remarkable to mention that the previously described genome reduction process correlates with the time period of the symbiotic association. Furthermore, many of the typical genome features of free-living bacteria, like *Escherichia coli*, such as the presence of pseudogenes and mobile elements, as well as a constant exchange of genetic material by recombination or horizontal gene transfer events, are lost in intracellular endosymbionts. As they inhabit a stable and nutrient-rich environment, they lose genes that are redundant or non-essential because they are provided by the host. This situation leads to the irreversible loss of bacterial genes and, thus, unnecessary metabolic capabilities strictly following a 'use it or lose it' tendency in evolution (Allen et al., 2009; Wernegreen, 2005). Finally, it is worth noting that a rather limited number of bacteria is vertically inherited by the host compared with the number that can be achieved during insect development. These dynamics are translated into systematic bottlenecks, which determine an ample effect of genetic drift with

respect to natural selection in the evolution of bacterial endosymbionts (Delaye et al., 2010; Moran et al., 2009).

Although organelles and endosymbionts with extremely reduced genomes present some commonalities, there are several important differences. Both share their extremely reduced genome sizes, are maternally inherited, and are completely dependent on their host for survival, mutually providing essential functions. Regarding differences, organelles have a double membrane enforced by engulfing the primitive cells inside the host cell, and genes of the organelle-becoming bacteria are transferred to the hosts' DNA, which acquires sophisticated transport mechanisms to transfer protein products from host to organelle and vice versa. Additionally, the host cell takes over the regulation of the organelle's division, synchronizing it with the cell's own division (S. B. Gould et al., 2008; Keeling & Archibald, 2008), leaving the organelle without cell status. In contrast, endosymbionts appear to lack a cell wall and possibly depend on host-derived membranes (or membranes synthesized by another symbiont present in the consortium), and to date, there is no evidence of horizontal gene transfer to the host nuclear genome. Moreover, endosymbionts have evolved together with multicellular eukaryotic organisms, plausibly making their cellular status less questionable than in the case of organelles.

In this opinion article, we propose the 'symbionelle' concept for endosymbionts that have partially lost their symbiotic role as a consequence of strong genome shrinkage and go beyond what is theoretically considered a minimal cell (Gil, 2015; Gil et al., 2004; Luisi et al., 2002). A minimal cell is formed

by a core minimal genome of protein-coding and RNA genes, which guarantee the following three major functions: (i) genetic machinery composed of virtually complete DNA replication and translation apparatus, and a simple DNA repair system; (ii) an energetic and intermediary metabolism, in which energy is obtained via substrate-level phosphorylation and the basic elements are provided by the environment to synthesize the essential cell components; and (iii) a cell envelope that encloses the genetic and metabolic pieces of machinery, controlling interaction with the environment, and growing and dividing to allow the formation of daughter cells. According to these functions, Gil (2013) proposes that the minimal gene-set machinery is composed of 188–206 protein-coding genes and 35–38 RNA genes. Figure C1.1 demarcates a zone ranging from between 223 and 244 genes (the strict theoretical minimal genome), which separates endosymbionts having minimal cell status (on the right of the dotted line) from **symbionelles**, a term coined to describe bacteria that fail to reach the minimal gene set (shown either on the left or inside the zone). According to this criterion, *Candidatus Tremblaya princeps* (155 genes), '*Ca. Hodgkinia cicadicola*' (189 genes), '*Ca. Carsonella ruddii*' (213 genes) and, probably also, '*Ca. Zinderia insecticola*' (232 genes) are examples of the term duded here as *symbionelles*. By contrast '*Ca. Sulcia muelleri*' (264 genes), '*Ca. Uzinura diaspidicola*' (272 genes), '*Ca. Portiera*' aleyrodidarum (292 genes), *Buchnera aphidicola* BCc (402 genes) and *Moranella endobia* (481 genes) meet the minimal-cell criteria (McCutcheon & Moran, 2012; Pérez-Brocal et al., 2006; Sabree et al., 2013; Santos-Garcia Diego et al., 2012).

Manifestly, both endosymbionts and organelles live in the very rich intracellular medium of their host. This type of heterotrophic environment has also been used to define the chemical environment of a minimal cell and contains glucose, phosphate, fatty acids, nitrogenous bases, amino acids, nucleotides, vitamins, inorganic ions, and several cofactors (Gil, 2015).

Another interesting feature of endosymbionts with reduced genomes is the formation of consortia among two or more symbiotic bacteria, whereby they complement each other metabolically to fulfill their symbiotic role. However, on occasions, as previously mentioned, some endosymbionts (the newly defined symbionelles), go beyond the minimal cell state, and while there are cases of the coexistence of two minimal cells or a minimal cell with a symbionelle, there is no evidence of two symbionelles living together. Examples can be found in the metabolic complementation among two minimal endosymbiotic cells in the sharpshooter *Homalodisca coagulata*, in the cicada *Diceroprocta semicincta* and in the spittlebug *Clastoptera arizonana*. All three insects have the endosymbiont '*Ca. Sulcia muelleri*' (minimal cell), which needs to be complemented by *Baumannia cicadellinicola* (minimal cell), '*Ca. Hodgkinia cicadicola*' (symbionelle) and '*Ca. Zinderia insecticola*' (symbionelle), respectively (McCutcheon et al., 2009b). It is worth mentioning that '*Ca. Sulcia muelleri*' might lie on the boundary dividing a minimal cell from a symbionelle, as it has lost some genes that would be considered 'essential', like missing genes encoding aminoacyl-tRNA synthetases (McCutcheon & Moran, 2007). Another intriguing case is the nested endosymbiosis of mealybugs of the subfamily Pseudococcinae, such as *Planococcus*

citri, where the co-endosymbiont, *Moranella endobia* (minimal cell) is located inside '*Ca. Tremblaya princeps*' (symbionelle). In this case, the complementation involves not only metabolic but also informational functions, as '*Ca. Tremblaya princeps*' appears to be a mere factory for amino acid synthesis and translating proteins, using precursors provided by *M. endobia*, including those for informational proteins (López-Madriral Sergio et al., 2011; McCutcheon & von Dohlen, 2011).

'*Ca. Carsonella ruddii*' and '*Ca. Uzinura diaspidicola*', endosymbionts of the psyllid *Pachypsylla venusta* and the armored scale insects (family Diaspididae), respectively, are endosymbionts with highly reduced genome sizes living alone in their respective hosts. However, the genome content of '*Ca. Uzinura diaspidicola*' seems to meet the requirements to be considered a minimal cell fulfilling its symbiotic role (Sabree et al., 2013), the case of '*Ca. Carsonella ruddii*' (symbionelle) is striking as it lacks not only the genes necessary for its symbiotic role but also several important genes involved in DNA replication, transcription, and translation, e.g. a ligase activity (Nakabachi et al., 2006; Tamames et al., 2007). It has been hypothesized that the host has taken over the role of the missing genes or that some of the genes have evolved novel functions (Sloan & Moran, 2012a).

Endosymbionts of white flies and aphids also reveal additional cases of metabolic complementation between bacterial genomes having cellular status. '*Ca. Portiera aleyrodidarum*' (292 genes), the endosymbiotic minimal cell of the white fly *Bemisia tabacci*, always coexists with other endosymbionts harboring

higher gene numbers (Santos-Garcia Diego et al., 2012). *Buchnera aphidicola* BCc (402 genes), the primary endosymbiotic minimal cell of the aphid *Cinara cedri*, has conserved all the necessary genes for its own replication, transcription, and translation, as well as a simplified metabolic network to produce energy. This particular *Buchnera* strain has partially lost its symbiotic role and complements with '*Ca. Serratia symbiotica*' (772 genes), which is considered to be a co-endosymbiont (Gosalbes et al., 2008; Pérez-Brocal et al., 2006).

It is noteworthy that metabolic complementation does not exclude the minimal cell status. A minimal endosymbiotic cell can survive if it is complemented by another minimal cell, a symbionelle, or by the host, which provides a function it cannot produce. This is not the case, however, of symbionelles, which are completely dependent on at least one other minimal cell.

In summary, symbionelles represent extreme cases of genome reduction in bacterial endosymbionts and cannot be considered minimal cells. No rich heterotrophic environment can be envisioned where such symbionelles can survive without the help of an additional cell. However, other endosymbionts present a number of genes and basic functional categories that enable them to be included in the theoretical definition of a minimal cell. Organelles and symbionelles represent, up to a point, a case of evolutionary convergence, although their evolutionary scenarios are completely different because organelles evolved before multicellularity appeared, and symbionelles evolved later, particularly in insect evolution.

1.2. Evolution of small prokaryotic genomes

As revealed by genome sequencing, the biology of prokaryotes with reduced genomes is strikingly diverse. These include free-living prokaryotes with ~800 genes as well as endosymbiotic bacteria with as few as ~140 genes. Comparative genomics is revealing the evolutionary mechanisms that led to these small genomes. In the case of free-living prokaryotes, natural selection directly favored genome reduction, while in the case of endosymbiotic prokaryotes neutral processes played a more prominent role. However, new experimental data suggest that selective processes may be in operation as well for endosymbiotic prokaryotes at least during the first stages of genome reduction. Endosymbiotic prokaryotes have evolved diverse strategies for living with reduced gene sets inside a host-defined medium. These include the utilization of host-encoded functions (some of them coded by genes acquired by gene transfer from the endosymbiont and/or other bacteria); metabolic complementation between co-symbionts; and forming consortiums with other bacteria within the host. Recent genome sequencing projects of intracellular mutualistic bacteria showed that previously believed universal evolutionary trends like reduced G+C content and conservation of genome synteny are not always present in highly reduced genomes. Finally, the simplified molecular machinery of some of these organisms with small genomes may be used to aid in the design of artificial minimal cells. Here we review recent genomic discoveries of the biology of prokaryotes endowed with small gene sets and discuss the evolutionary mechanisms that have been proposed to explain their peculiar nature.

Darwin proposed an externalist theory of evolution where organisms provide the raw material and the environment selects it (S. J. Gould, 2002). The outcome of this process is a fine adjustment of organisms to the environment. The evolution of prokaryotes with reduced genomes is not an exception to this Darwinian principle. Host-associated bacteria and archaea evolved the smallest genomes in nature other than those of organelles and viruses. The rationale of this pattern is simple. Prokaryotes living in a protected and chemically rich medium can afford to lose more genes than those coping with the vagaries of a free-living lifestyle (Morowitz, 1993). On the other hand, different lineages of free-living bacteria, most of them in marine environments, evolved reduced genomes likely by the direct action of natural selection (Giovannoni et al., 2014).

1.2.1 What is the minimal genome size for extant free-living prokaryotes?

Previous surveys indicated that free-living prokaryotes had no less than ~1,300 genes (Delaye et al., 2010; Islas et al., 2004; Podar et al., 2008). However, recent metagenomic sequencing suggests that there are free-living Actinobacteria with approximately 800 genes. This was discovered in the Mediterranean Sea and the bacteria were named "*Candidatus Actinomarina minuta*" (Ghai et al., 2013). Surprisingly, it is also one of the smallest cells with a cell volume of only ~0.013 μm^3 . If further sequencing of its complete genome confirms this estimate (and it is very likely that it will do), it will sensibly change our knowledge about the minimum number of genes a cell needs to survive in present free-living conditions, in a similar fashion to the discovery of "*Candidatus Carsonella*

ruddii”, that shook our belief of the minimal gene set required for cells in 2006 (McCutcheon & Moran, 2012; Nakabachi et al., 2006). Meanwhile, as reviewed below, there exists a diversity of lineages of free-living prokaryotes that converged to approximately 1,300 genes despite their varying phylogenetic origins and nutritional strategies.

Nowadays, *Methanothermus fervidus* with a genome coding for 1,311 proteins and 50 RNA genes, stands as the free-living archaeon (that does not grow associated with another cell) with the smallest sequenced genome. This organism is a methanogen and was isolated from an anaerobic Icelandic spring (Anderson et al., 2010). As mentioned above, other groups of free-living prokaryotes evolved similar genome sizes, several of them from marine environments.

For instance, α -proteobacteria from clade SAR11, which is the most abundant group of heterotrophic bacteria in the oceans, are endowed with genomes ranging from 1,321 to 1,541 protein-coding genes (Grote et al., 2012). Among them, is “*Candidatus Pelagibacter ubique*” HTCC1062, which is one of the most studied members of clade SAR11 and is an oligotroph, with 1,354 protein-coding genes, that generates energy by respiration (Giovannoni et al., 2005).

Another group of marine prokaryotes that evolved similar genome sizes is the β -proteobacteria from clade OM43. Specifically strains HTCC2181 and HIMB624 have 1,377 and 1,381 protein-coding genes respectively (Giovannoni et al., 2008; Huggett et al., 2012). These are marine and freshwater bacteria that live

heterotrophically by using methylated compounds as carbon sources (Giovannoni et al., 2008; Huggett et al., 2012).

Contrasting with the previously mentioned heterotrophs whose smallest genomes have ~1,300 genes, photoautotrophic free-living bacteria have larger genomes. For instance, some strains from *Prochlorococcus marinus*, the most abundant photosynthetic organism on Earth, have genomes coding for as few as 1,716 protein-coding genes (Dufresne et al., 2003; Rocap et al., 2003; Scanlan et al., 2009).

And finally, non-marine bacteria with small genomes include the mollicute *Acholeplasma laidlawii*, which is a saprophyte and opportunistic parasite found in a wide variety of environments and has a genome coding for 1,380 protein-coding genes (Lazarev et al., 2011; Windsor et al., 2010); the lactobacilli *Weissella koreensis* KACC 15510 which is a heterotroph that participates in the fermentation of kimchi (a representative Korean fermented food) and has a genome coding for 1,335 predicted protein-coding sequences (S. H. Lee et al., 2011); the dehalorespiration *Dehalococcoides* sp. BAV1 with a genome coding for 1,371 protein-coding genes and a member of the Chloroflexi (He et al., 2003; Löffler et al., 2013); and the Chrenarchaeon *Desulfurococcus mucosus* O7/1 with its ability for sulfur respiration with a genome with 1,371 protein-coding genes (Wirth et al., 2011).

Why do different lineages of free-living cultivable bacteria have genomes with no less than 1,300 genes? One possible explanation is that extant biotic and abiotic environments exert a selective pressure against simpler cells, therefore

imposing an ecological limit on the minimum complexity necessary for a cell to survive (S. J. Gould, n.d.). The idea is that free-living cells with fewer genes are outcompeted by cells with a more complete genetic arsenal unless associated with other organisms. However intuitive this idea is, it still requires experimental validation.

However, it is important to take into consideration that it is possible that our sample of genome sequences from cultivable organisms does not accurately represent the distribution of genome sizes that exist in nature (Giovannoni et al., 2014). Additionally, as suggested by metagenomic data, there may exist a whole biodiversity of uncultivable bacteria with genomes with less than 1,300 genes, as seems to be the case of “*Ca. Actinomarina minuta*” (Ghai et al., 2013).

In this direction, a note of caution regarding the limit of ~1,300 genes for free-living prokaryotes is given by *Lactobacillus fermentum* CECT 5716, a hetero-fermentative lactic acid bacterium inhabiting human mucosal surfaces and breast milk. This bacterium has a genome with 1,109 protein-coding genes (Jiménez et al., 2010). And, although this organism would be classified as a symbiont because it is naturally associated with humans, it grows well under laboratory conditions (Jiménez, personal communication) thus blurring the distinction between free-living and host-associated microorganisms. The discovery of “*Ca. Actinomarina minuta*,” as well as the existence of *L. fermentum* CECT 5716, indicates that in the near future, we will probably discover free-living bacteria with smaller genomes.

1.2.2 Drivers of genome reduction among free-living prokaryotes

As we will describe below, different mechanisms have been proposed to account for genome reduction among free-living prokaryotes (Figure C1.2; (Dufresne et al., 2005; Giovannoni et al., 2005; Marais et al., 2008; Mira et al., 2001).

Free-living

Streamlining hypothesis

Natural selection favors smaller genomes and low G+C content as a way to cellular economization



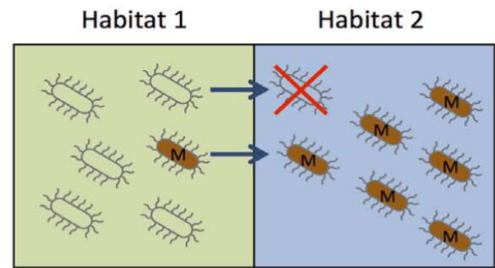
Requisites:
Large population size
Low nutrient environment

Outcome:
Low G+C ; small intergenic regions; small cell size

Selection for mutator strain hypothesis

Niche colonization favors mutator strains; as a consequence, genes with low fitness contribution are lost from the genome

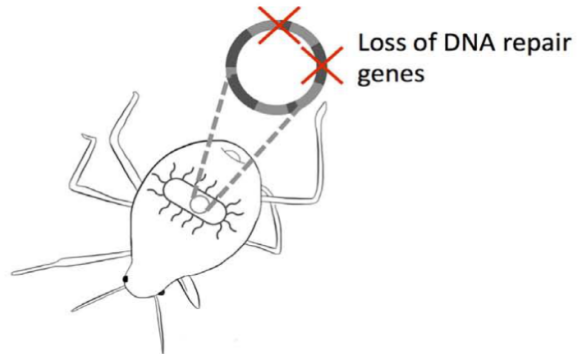
Outcome:
Genome reduction and low G+C content are side effects of selection for larger mutation rates



Host-associated

Loss of DNA repair genes hypothesis

Outcome:
All classes of genes will experience about the same increase in mutation rate and the d_N/d_S ratio will be larger for endosymbiont protein coding genes than for their free-living homologous counterparts



Muller's ratchet

Deletions accumulate by Muller's ratchet

Requisites:
Strong bottlenecks, lack of recombination, obligate symbiosis providing a rich nutrient environment

Outcome:
As a result of relaxation of natural selection there is accumulation of slightly deleterious mutations and low G+C content

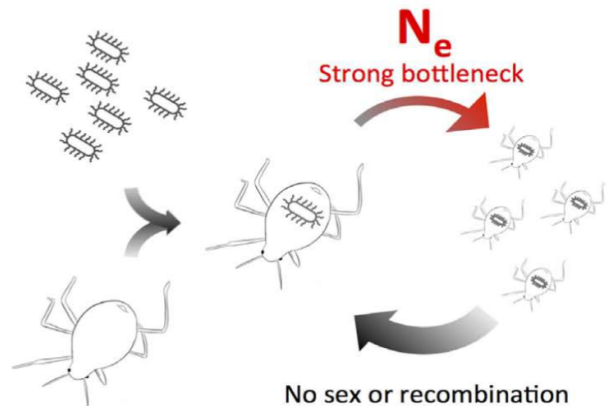


Figure C1.2 | Streamlining and Muller's ratchet hypotheses are commonly used to explain genome reduction in free-living and host-associated bacteria respectively. Alternative hypotheses include "selection for mutator strains" and "loss of DNA repair genes".

1.2.3 The streamlining hypothesis

Genome reduction by a process known as streamlining, in which smaller genomes are favored directly by selection as a way to cellular economization, is perhaps the most popular explanation (Dufresne et al., 2005; Mira et al., 2001). According to this hypothesis, natural selection directly favors genome reduction in free-living prokaryotes living in low-nutrient environments (Grote et al., 2012; Mira et al., 2001). This claim is based on the basic idea that superfluous genes are eliminated because they confer a fitness cost to the bacterium. This is especially effective in large population sizes. The reasoning is as follows: very large population sizes render negligible the effect of genetic drift and more importantly, render highly efficient the process of natural selection. Then, when a fitness-increasing deletion occurs it is quickly fixed in the population, especially under high selective pressures such as low-nutrient environments where small genomes evolve as a way to economize on matter and energy for cell maintenance (Dufresne et al., 2005; Koskiniemi et al., 2012).

For instance, streamlining was suggested to be responsible for genome reduction in the high-light-adapted marine cyanobacteria *P. marinus* MED4 and the low-light-adapted *P. marinus* SS120 (Dufresne et al., 2003, 2005). According to the streamlining hypothesis, the following characteristics are consistent with natural selection acting to economize cellular metabolism. First, the small G+C

content of these genomes (~30 to 36%) contributes to fewer requirements for phosphorus and nitrogen for DNA synthesis which are scarce in the environment where *P. marinus* MED4 lives (Dufresne et al., 2005). Second, the small cell volume of *P. marinus* SS120 (~0.1 μm^3) is suggested to improve photosynthetic efficiency by reducing self-shading and enhancing nutrient uptake by increasing the surface-to-volume ratio for the cell, and also it is in itself an adaptation that has, in turn, exerted an evolutionary pressure for a smaller genome (Dufresne et al., 2003). This follows the logic that a smaller bacterial cell has a smaller volume and can only contain a small amount of DNA, otherwise too much of the internal space is devoted to DNA storing and the remaining volume would not be sufficient for other cellular components (Trevors & Masson, 2011). Cell volume has a wide range of ~0.013 to 400 μm^3 , however, a cell is considered small if its volume is less than 0.6 μm^3 (Ghai et al., 2013; Koch, 1996). And third, the streamlining hypothesis suggests that the effects of economization are observable as an increase in fitness.

Therefore, genome reduction is believed to have had a favorable effect on fitness in *Prochlorococcus* species. In the oceanic ecosystem, the diversity of photosynthetic prokaryotes is mostly represented by two genera: *Prochlorococcus* and *Synechococcus*. While *Synechococcus* are ubiquitous owing to their flexibility and adaptability to various marine environments, the *Prochlorococcus* have had apparently better ecological success in oligotrophic areas where the conditions are more stable (Partensky et al., 1999). The success of *Prochlorococcus* is believed to be due to the differential distribution across the vertical axis of the water

column of specialized ecotypes which are genetically and physiologically distinct populations distributed in accordance with the quality of light (Scanlan et al., 2009).

Genome streamlining was suggested also to explain reductive genome evolution in the case of the α -proteobacterium “*Ca. Pelagibacter ubique*” HTCC1062 and other members from the SAR11 clade (Giovannoni et al., 2005; Grote et al., 2012). As described above, this proposal is also based on several features of its streamlined genome. For instance, “*Ca. Pelagibacter ubique*” was reported to have a median space size between coding genes of only three nucleotides, the smallest among the analyzed genomes. In addition, no pseudogenes, phage genes, or recent gene duplications were found (Giovannoni et al., 2005). And similar to the case of *P. marinus* MED4 and SS120, it was suggested that the small cell size of “*Ca. Pelagibacter ubique*” (0.019–0.039 μm^3) evolved by natural selection. In this case based on a theory proposed by Button (Button, 1991). Accordingly to this theory, selection optimized surface-to-volume ratio so that the capacity of the cytoplasm to process substrates matches transport rates (Giovannoni et al., 2005; Steindler et al., 2011).

The β -proteobacteria from Clade OM43 is another lineage where genome reduction by streamlining was suggested. As in the two cases described above, the small proportion of non-coding DNA in the reduced genome of strain HTCC2181 was interpreted as evidence of streamlining selection (Giovannoni et al., 2008). And as in the case of the strains MED4 and SS120 from *P. marinus* and

in “*Ca. Pelagibacter ubique*,” the β -proteobacteria HIMB624 also has a small cell size of about 0.1–0.3 μm wide and 0.6–1.8 μm long (Huggett et al., 2012).

Finally, streamlining could be suggested also for “*Ca. Actinomarina minuta*.” Its small genome is contained in an incredibly small cell (0.013 μm^3). The median length of its intergenic sequences is of only three bases, the same as of “*Ca. Pelagibacter ubique*.” “*Ca. Actinomarina minuta*” lives in aquatic environments where nutrients are scarce (Ghai et al., 2013).

Supporting the streamlining hypothesis, Koskiniemi et al. (Koskiniemi et al., 2012) found that under laboratory conditions, selection can drive genome reduction. In order to do this, they devised a method that would report large deletions in the genome of *Salmonella enterica*. When they measured fitness against the wild type, they observed that several mutant strains showed an increase in fitness and concluded that fitness increases were common following deletions on specific genomic loci. Additionally, they performed a serial passage experiment and observed that selection could be a significant driver of gene loss. They suggested that in naturally occurring populations, fixation of deletion could occur very fast.

1.2.4 Accelerated rates of protein evolution

Returning to *P. marinus* strains MED4 and SS120, accelerated rates of protein evolution have been observed in these cyanobacteria (Dufresne et al., 2005). This is similar to what is observed in symbiotic bacteria with reduced genomes (McCutcheon & Moran, 2012). However, the cause of this acceleration in

free-living bacteria seems to be different. According to the streamlining hypothesis, this is a consequence of an increase in the mutation rate due to the loss of repair genes and not a direct consequence of selection. Supporting this hypothesis is the fact that both strains lack the *ada* gene, which encodes 6-O-methylguanine-DNA methyltransferase among other repair genes (Dufresne et al., 2005). In addition, the lack of this gene can lead to G:C to A:T transversions which, as discussed above, can be adaptive in low phosphorus environments (Dufresne et al., 2005; Mackay et al., 1994). Nevertheless, differing from the marine picocyanobacteria described above, "*Ca. Pelagibacter ubique*" codes for the DNA repair enzyme 6-O-methylguanine-DNA methyltransferase while showing a G+C content as low as 29% (Giovannoni et al., 2005). This suggests that the loss of this enzyme is not a necessary prerequisite to evolving high levels of A+T and that the direct action of selection favoring high levels of A+T could be the cause.

1.2.5 Increased rates of mutation hypothesis

An alternative explanation for genome reduction has been proposed which includes the previously mentioned accelerated rates of protein evolution for *P. marinus* strains (Marais et al., 2008). This explanation suggests that genome reduction occurs as a byproduct when an increased mutation rate becomes advantageous, like in the cases of novel niche colonization. And indeed, *P. marinus* MED4 and SS120 colonized high-light and low-light niches of the water column respectively between 150 and 80 million years ago (Dufresne et al., 2005).

The argument is as follows. According to classical population genetic models, the fate of an allele is determined by selection if the product of the effective population size (N_e) by the coefficient of selection (s) is larger than one (*i.e.*, $N_e s > 1$); and is determined by genetic drift if it is smaller than one: $N_e s < 1$ (Figure C1.3; (Gillespie, 1998)). However, this model applies only when the mutation rate (μ) is negligible (Marais et al., 2008).

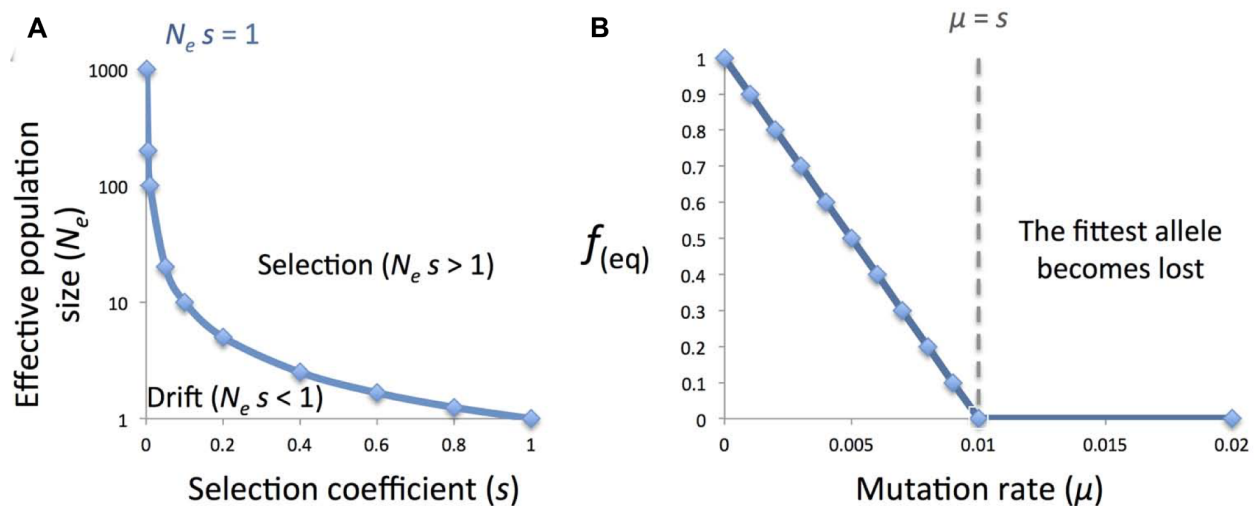


Figure C1.3 | Population genetic models. (A) The fate of an allele is determined by natural selection if the product of the effective population size (N_e) and the coefficient of selection (s) is larger than one, and by genetic drift otherwise; (B) however, when mutation is taken into account, the equilibrium frequency (f_{eq}) of the fittest allele is 0 when the mutation rate (μ) is larger than the selection coefficient (s).

When the mutation rate is not negligible and taken into account, the equilibrium frequency of the fittest allele becomes $(1 - \mu/s)$ if $\mu < s$, and 0 if $\mu > s$. This is according to a simple model developed by Eigen (Eigen, 1971). Therefore, when the mutation rate is larger than the selection coefficient, the fittest allele

will have an equilibrium frequency of 0 and therefore, will be lost in the population (Marais et al., 2008).

As mentioned above, high mutation rates can be advantageous when bacteria colonize new habitats. In natural populations, some strains often develop increased mutation rates compared to the wild type due to the loss of repair genes. These strains are called mutator strains. Accordingly, mutator strains were selected during novel niche colonization by *P. marinus* MED4 and SS120. These increased μ over s and favored the loss of genes that have only a modest contribution to fitness thus reducing the genome (Marais et al., 2008).

In agreement with this hypothesis, the proteins from MED4 and SS120 *Prochlorococcus* have similar sizes to their homologs from *Prochlorococcus* with larger genomes. This is contrary to what would be expected if natural selection directly favored genome minimization as predicted by the streamlined theory (Marais et al., 2008). However, the same pattern can be accounted for by the streamlining hypothesis if natural selection has not been strong enough to select smaller proteins.

1.2.6 The Black Queen Hypothesis

Our discussion of the mechanisms of genome reduction would not be complete if we did not include a recent illuminating proposal known as the Black Queen Hypothesis (BQH; (Morris et al., 2012)). The BQH introduces gene loss and thus genomic reduction as a community-dependent adaptive event. In order for genomic reduction to occur according to the BQH there are three main

components required in a community: a public good (PG), a helper, and beneficiary organisms. A PG is a function or product that is energetically or nutritionally expensive to do or make, and that is required and accessible by the whole community and not only by the producing organism. A helper organism is an organism capable of producing the PG and whether actively or passively is capable of leaking it to the community. A beneficiary is an organism that utilizes the PG but is incapable of producing it itself. Genome reduction occurs in the beneficiary, and accordingly, there must be a selective advantage to lose the function and thus, the genes that code for it. Importantly, the benefit of losing the function is frequency-dependent, thus once the function becomes too scarce it is no longer advantageous to lose it, such that the presence of helpers is guaranteed in the population and permits the function to remain active for all the community (Figure C1.4; (Morris et al., 2012)).

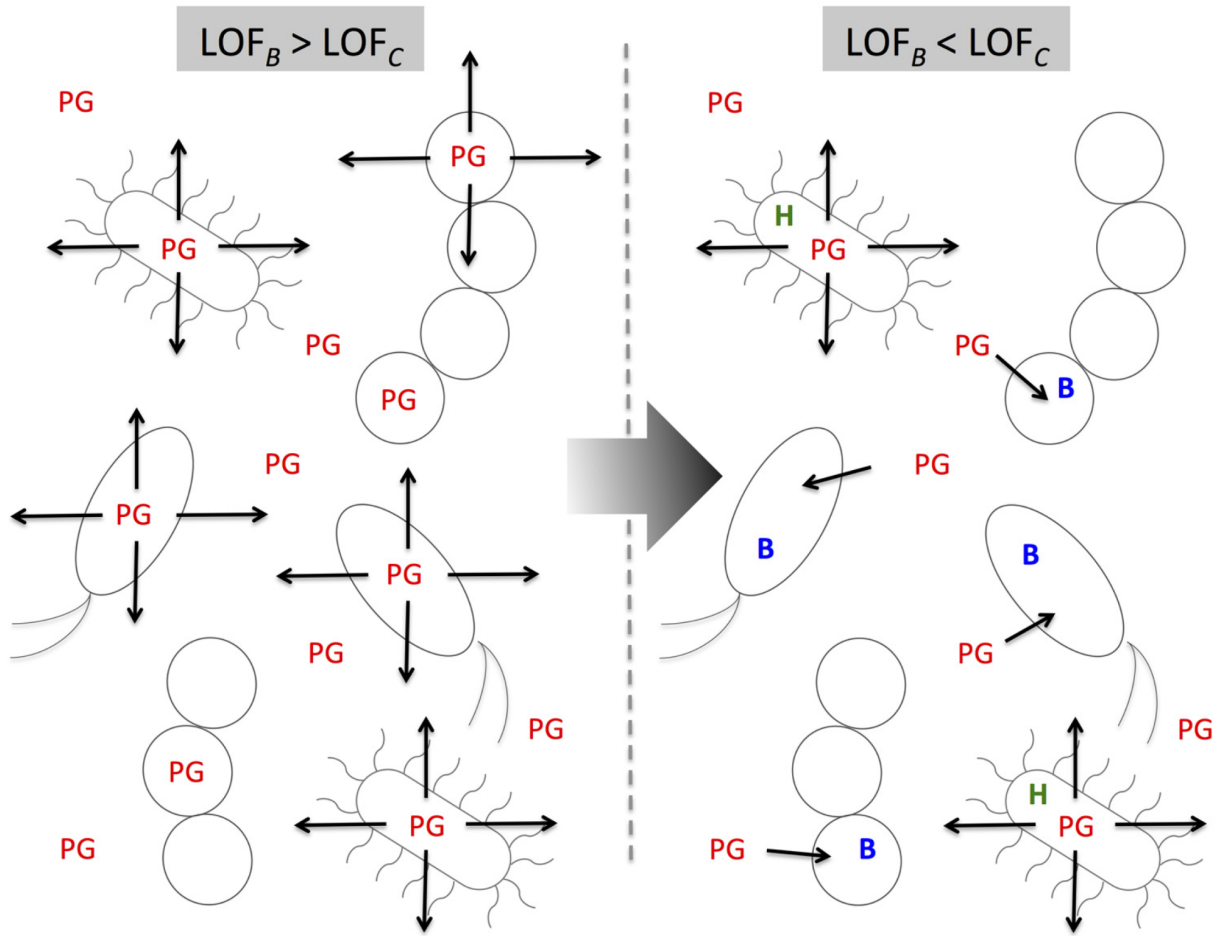


Figure C1.4 | The Black Queen Hypothesis predicts that if there is a community where different species produce an expensive and diffusible public good (PG), the system will evolve toward a scenario where only a few of their members will continue with the production of the PG, only if the benefit of losing the production of the function outweighs the cost of losing it. As such, community producers of the PG become helpers (H) and the rest become beneficiaries (B). LOF_B is the benefit of losing the function producing the PG; LOF_C is the cost of losing the function producing the PG.

As pointed out by Morris et al. (Morris et al., 2012) and Sachs and Hollowell (Sachs & Hollowell, 2012), *Prochlorococcus* strains cannot grow axenically in their own habitat. This is because these bacteria lack the gene for

















catalase-peroxidase to eliminate hydrogen peroxide that is produced by photooxidation up to toxic levels by sunlight. For *Prochlorococcus* (a beneficiary) to grow under these conditions, it is necessary the presence of other bacteria (a helper) that code for catalase-peroxidase (a PG) to detoxify the environment. This is possible because hydrogen peroxide is permeable, thus the helper organisms catalyze the removal of hydrogen peroxide and reduce natural concentrations below the toxic level in marine environments (Morris et al., 2012).













Another example is that of highly reduced bacteria "*Ca. Pelagibacter ubique*." Similarly, as the above, this bacterium could not grow on artificial media until recently, when Carini et al. (Carini et al., 2013) were capable of uncovering the mystery behind this organism's amazing nutritional strategy. They found that this bacterium uses a balanced supply of organic matter, which included methionine, glycine, and pyruvate. These could be replaced by other metabolites, which included some common osmolytes. As such, they suggested that "*Ca. Pelagibacter ubique*" had evolved to efficiently utilize various low molecular weight metabolites of phytoplankton origin produced in low but continuous concentrations (Carini et al., 2013). Other suggested examples of genetic loss by BQH are related to nitrogen fixation, inorganic nutrient acquisition, and biofilm matrix deposition, but it is likely that the list is much longer (Morris et al., 2012). Therefore, reductive genome evolution has to be reanalyzed in light of the BQH. The community-dependent nature of the BQH would also suggest an important role in understanding the evolution of host-associated prokaryotes.

1.2.7 Host-associated prokaryotes with reduced genomes

Symbiosis can be defined as “an intimate, close association between species in which the large majority or entire life cycle of one species occurs within or in very close association with another” (Holland & Bronstein, 2008). As we mentioned earlier, symbiotic organisms tend to evolve smaller genomes amongst which host-associated intracellular mutualistic prokaryotes from different phyla have evolved the smallest cellular genomes known (Table C1.1). Although the term symbiosis is sometimes confounded with mutualism, this intimate and close association can be mutualistic or not. For intracellular prokaryotes, these relationships are of parasitic, commensal as well as mutualistic nature. In particular, the biology of mutualistic prokaryotes with highly reduced genomes is strikingly diverse. Moreover, recent advances in genome sequencing have revealed novel and unexpected evolutionary trends, as briefly reviewed below.

Table C1.1 | Mutualistic prokaryotes with reduced genomes. The information on genes and gene size was obtained from Genbank. The class is shown in taxonomy unless specified. “ns”: values unknown from the not-sequenced organism. “*”: Values include the gene content and genomic size of strain-specific plasmids. “¶”: Values as reported in the original article (Rosas-Pérez et al., 2014). “†”: Value as reported in the original article (Boscaro et al., 2013). Host drawings are kindly provided by Sofia Delaye.

Symbiotic prokaryotes	Taxonomy	Protein-coding Genes	RNA-coding genes	Genome size (Kbp)	Host	Host
<i>Nasuia deltocephalinicola</i> str. NAS-ALF	β-Proteobacteria	137	32	112	<i>Macrostelus quadrilineatus</i> (aster leafhopper)	
<i>Sulcia muelleri</i> str. Sulcia-ALF	Flavobacteriia	188	35	191		
" <i>Ca. Tremblaya princeps</i> " PCVAL	β-Proteobacteria	116	20	139	<i>Planococcus citri</i> (citrus mealybug)	
" <i>Ca. Moranella endobia</i> " PCVAL	γ-Proteobacteria	411	47	538		
" <i>Ca. Hodgkinia cicadicola</i> " Dsem	α-Proteobacteria	169	19	144	<i>Diceroprocta semicincta</i> (cicada)	
" <i>Ca. Sulcia muelleri</i> " SMDSEM	Flavobacteriia	242	33	277		
" <i>Ca. Carsonella ruddii</i> " PV	γ-Proteobacteria	182	31	160	<i>Pachypsylla venusta</i> (psyllid)	
" <i>Ca. Sulcia muelleri</i> " GWSS	Flavobacteriia	227	36	246	<i>Homalodisca coagulata</i> (sharpshooter)	
<i>Baumannia cicadellincola</i> Hc	γ-Proteobacteria	595	46	686		
" <i>Ca. Zinderia insecticola</i> " CARI	β-Proteobacteria	202	29	209	<i>Clastoptera arizonana</i> (spittlebug)	
" <i>Ca. Sulcia muelleri</i> " CARI	Flavobacteriia	246	34	277		
" <i>Ca. Walzuchella monophlebidarum</i> "	Flavobacteriia	271¶	36¶	309	<i>Llaveia axin axin</i> (scale insect "Nii")	
Enterobacterial endosymbiont	γ-Proteobacteria	ns	ns	ns		
" <i>Ca. Carsonella ruddii</i> " CE	γ-Proteobacteria	190	31	163	<i>Ctenarytaina eucalypti</i> (psyllid)	
Secondary endosymbiont of <i>Ctenarytaina eucalypti</i>	γ-Proteobacteria	918	45	1441		
" <i>Ca. Carsonella ruddii</i> " HC	γ-Proteobacteria	192	31	166	<i>Heteropsylla cubana</i> (psyllid)	
Secondary endosymbiont of <i>Heteropsylla cubana</i>	γ-Proteobacteria	576	44	1122		
" <i>Ca. Uzinura diaspidicola</i> " ASNER	Flavobacteriia	227	35	263	<i>Aspidiotus nerii</i> Bouché (oleander scale)	
" <i>Ca. Portiera aleyrodidarum</i> " BT-QVLC	γ-Proteobacteria	246	38	357	<i>Bemisia tabaci</i> (Mediterranean whiteflies)	
" <i>Ca. Hamiltonella defensa</i> " MED	γ-Proteobacteria	1474	42	1843		
<i>Buchnera aphidicola</i> BCc	γ-Proteobacteria	362*	37	422*	<i>Cinara cedri</i> (cedar aphid)	
<i>Serratia symbiotica</i> str. 'Cinara cedri'	γ-Proteobacteria	672	42	1763		
<i>Blattabacterium cuenoti</i> Cpu	Flavobacteriia	548*	38	610*	<i>Cryptocercus punctulatus</i> (wood roaches)	
" <i>Ca. Riesia pediculicola</i> " USDA	γ-Proteobacteria	555*	40	582*	<i>Pediculus humanus corporis</i> (body louse)	
<i>Profftella armatura</i>	β-Proteobacteria	372*	37	465*	<i>Diaphorina citri</i> (Asian citrus psyllid)	
" <i>Ca. Carsonella ruddii</i> " DC	γ-Proteobacteria	207	31	174		
<i>Nanoarchaeum equitans</i> Kin4-M	Phylum: Nanoarchaeota	540	45	491	<i>Ignicoccus hospitalis</i> KIN4/I	
<i>Ignicoccus hospitalis</i> KIN4/I	Thermoprotei	1434	53	1298		

<i>Ishikawaella capsulata</i> Mpkobe	γ-Proteobacteria	623*	48	755*	<i>Megacopta punctatissima</i> (stinkbug)	
<i>Blochmannia vafer</i> BVAf	γ-Proteobacteria	587	42	723	<i>Camponotus vafer</i> (ant)	
<i>Wigglesworthia glossinidia</i> endosymbiont of <i>Glossina morsitans</i>	γ-Proteobacteria	618	44	720	<i>Glossina morsitans</i> (tsetse fly)	
<i>Sodalis glossinidius</i> str. 'morsitans'	γ-Proteobacteria	2516*	91	4293*		
"Ca. Kinetoplastibacterium oncopeltii" TCC290E	β-Proteobacteria	694	52	810	<i>Strigomonas oncopelti</i> (tripanosome)	
"Ca. Endolissoclinum faulkneri" L2	α-Proteobacteria	1498	48	1481	<i>Lissoclinum patella</i> (tunicate)	
"Ca. Endomicrobium sp. Rs-D17"	Phylum: Elusimicrobia	776*	48	1149*	<i>Trichonympha agilis</i> (termite protist)	
"Ca. Azobacteroides pseudotrichonymphae genomovar CFP2"	Bacteroidia	852*	42	1225*	<i>Pseudotrichonympha grassii</i> (termite protist)	
<i>Wolbachia</i> endosymbiont TRS of <i>Brugia malayi</i>	α-Proteobacteria	805	37	1080	<i>Brugia malayi</i> (filarial nematode)	
"Ca. Vesicomysocius okutanii" Ha	γ-Proteobacteria	937	38	1022	<i>Calyptogena okutanii</i> (deep-sea clam)	
"Ca. Ruthia magnifica" str. Cm	γ-Proteobacteria	976	41	1161	<i>Calyptogena magnifica</i> (giant clam)	
<i>Cyanobacterium</i> sp. UCYN-A	Subclass: Oscillatoriophyci deae	1199	42	1440	Prymnesiophyte (unicellular eukaryote)
"Ca. Midichloria mitochondrii" IricVA	α-Proteobacteria	1211	38	1184	<i>Ixodes ricinus</i> (tick)	
<i>Polynucleobacter necessarius</i> subspecies <i>necessarius</i> STIR1	β-Proteobacteria	1279†	44	1560	<i>Euplotes aediculatus</i> (ciliated protist)	

1.2.8 Early obligated intracellular symbiosis

The transition from free-living to endosymbiotic mutualistic lifestyle was recently studied by comparative genomics of free-living bacteria and their counterparts, living either in a protist or in insects (weevils or aphids; (Boscaro et

al., 2013; Clayton et al., 2012; Manzano-Marín & Latorre, 2014). These studies showed that in both cases, gene inactivation, genome rearrangement, and loss of some of the repair mechanisms played important roles. However, there were also important differences. In particular, there is an extreme proliferation of MGEs (mobile genetic elements) in symbiotic bacteria associated with the early stages of the integration process in insects, but not in the bacteria associated with the protist. This difference was attributed to the fact that in the case of the insect, the population of the bacteria undergoes recursive bottlenecks that lower the efficacy of selection allowing the proliferation of MGEs (Mira et al., 2001). Furthermore, the importance of mobile elements in the transition from free-living to an obligate endosymbiotic state involves their participation in gene inactivation and genome size reduction in recent endosymbiont genomes, as observed from comparative studies between ancient endosymbionts having lost all mobile elements, and related free-living bacteria, with a controlled number of these, presumably by natural selection (Manzano-Marín & Latorre, 2014; Oakeson et al., 2014).

Observational as well as experimental data indicate that genome reduction in host-associated bacteria occurs at a fast rate once the symbiosis is obligate. For instance, it is estimated that the obligate endosymbiont of the rice weevil *Sitophilus oryzae*, "*Candidatus Sodalis pierantonius*" str. SOPE lost 55% of their genes in just 28,000 years (Oakeson et al., 2014). In agreement with previous observations, experimental evidence shows that under laboratory conditions, similar to those of intracellular bacteria (strong bottlenecks and absence of HGT),

genome reduction can occur very rapidly on an evolutionary time scale (Nilsson et al., 2005).

"Candidatus Sodalis pierantonius" str. SOPE lives inside the rice weevil's bacteriocytes and has a relatively large genome coding for about 2,309 protein-coding genes and 1,771 pseudogenes. Perhaps the most striking characteristic of the genome of this bacterium is that it is plagued with MGEs, about 18% of the genome consists of ISs. These MGEs have contributed in this organism to genome rearrangement, partial genome duplications, deletogenic rearrangements, and ~10% of gene inactivation. This is dramatically exemplified by the constant perturbations of the G+C skew in its chromosome that reveals multiple changes in leading versus lagging strand orientation (Oakeson et al., 2014).

The early evolution of intracellular mutualistic symbiosis was also studied in aphids. Manzano-Marín and Latorre (Manzano-Marín & Latorre, 2014) compared the genomes of bacteria in different stages of the process of adaptation to intracellular lifestyle. This comparison included three strains of *"Candidatus Serratia symbiotica"* which represented an early facultative stage (*"Ca. Serratia symbiotica"* from *Acyrtosiphon pisum*, SAp) a later facultative bordering on early obligate stage (*"Ca. Serratia symbiotica"* from *Cinara tujaefilina*, SCt), and a co-obligate stage (*"Ca. Serratia symbiotica"* from *Cinara cedri*, SCc). Strain SCc has obligate endosymbiotic characteristics, such as the lack of MGEs, high A+T content, and no genetic redundancy. However, it possesses large intergenic regions, and remnants of ancient pseudogenes that are still not degraded

(Lamelas, Gosalbes, Manzano-Marín, et al., 2011). On the other hand, strain SCt, while being phylogenetically and genomically very closely related to the facultative SAp strain (Burke & Moran, 2011), shows a variety of metabolic, genetic, and architectural features which point toward this endosymbiont being one step closer to an obligate intracellular lifestyle. By studying the genome rearrangements and the impact that MGEs have had on the genome architecture of these two *Serratia* endosymbionts (SCt and SAp), it was determined that those genes belonging to IS (insertion sequence) families have been the key factor promoting massive rearrangements. These MGEs have also mediated inactivation in various genes, sometimes creating long stretches of inactivated proteins in tandem (Manzano-Marín & Latorre, 2014).

In the case of bacteria associated with the protist, the genome sequences of free-living and symbiotic strains of the β -proteobacteria *Polynucleobacter necessarius* were compared. The free-living strain is a common inhabitant of lentic freshwater ecosystems, while the symbiotic strain lives as an intracellular symbiont of the ciliated protist *Euplotes aediculatus*. These strains diverged very recently, as shown by their similarity at the level of the 16S rRNA gene which is >99% (Boscaro et al., 2013). The genome of the free-living strain contains 2,088 protein-coding genes (Meincke et al., 2012) and is itself a reduced genome (Boscaro et al., 2013). The genome of the symbiotic strain contains 1,279 protein-coding genes and is mostly a subset of the free-living strain (*i.e.*, only 105 genes are not shared with its free-living relative). The symbiotic strain also contains between 231 and 460 pseudogenes. Of course, the ultimate cause of this

genome reduction is the endosymbiotic lifestyle of *P. necessarius*. However, the proximal cause of the genome reduction is less well understood. Furthermore, the metabolic bases of the obligate symbiosis between the bacteria and the ciliate are not known. However, it was suggested that *P. necessarius* complements some metabolic deficiencies in *E. aediculatus* (Boscaro et al., 2013). Nevertheless, it was suggested that genome reduction in the symbiotic strain was caused by illegitimate recombination and loss of mismatch repair genes. In addition, it was suggested that the early loss of the gene coding for the translation DNA polymerase exerted further evolutionary pressure for a smaller genome and favored polyploidy, which is one of the main differences between both strains such that the symbiotic strain contains several nucleoids, each one containing one copy of the genome (Boscaro et al., 2013).

As exemplified by *P. necessarius*, rapid gene loss can occur in the absence of the proliferation of MGEs. The endosymbiotic strain of *P. necessarius* has already lost over 40% of its coding capacity with 13–18% still observable as pseudogenes. This loss occurred in the absence of MGEs, and as mentioned above, the lack of MGEs in *P. necessarius* has been attributed to a larger population size relative to that of “*Ca. Sodalis pierantonius*” str. SOPE (Boscaro et al., 2013).

In addition to the previously mentioned examples, the lack of proliferation of MGEs in the case of *P. necessarius* contrasts with other obligate symbioses. These include *Burkholderia rhizoxinica* the endosymbiont of the fungus *Rhizopus microsporus* with 6% of their encoded proteins similar to transposases (Lackner, Moebius, Partida-Martinez, Boland, et al., 2011); the γ 1 symbiont of the marine

Oligochaeta *Olavius algarvensis* that has a genome coding for 20% of transposases (Woyke et al., 2006); *NoAz*, the extracellular mutualistic endosymbiotic cyanobacteria of the water-fern *Azolla filiculoides* with ~600 IS (insertion sequence) elements (Ran et al., 2010); and the facultative symbiont of the whitefly "*Candidatus Cardinium hertigii*" possessing ~200 MGEs (Santos-Garcia et al., 2014).

The proliferation of MGEs can occur also in the absence of massive genome rearrangement, as shown in the case of the intracellular bacteria "*Candidatus Amoebophilus asiaticus*" that maintains a regular G+C skew along its genome despite that 24% of its genes code for MGEs (Schmitz-Esser et al., 2010, 2011). Nevertheless, this last comparison has to be taken with caution since "*Ca. Amoebophilus asiaticus*" has not adapted recently to the intracellular lifestyle, has received several genes by HGT (horizontal gene transfer), and as a parasite is under different evolutionary pressures than mutualistic endosymbionts (Schmitz-Esser et al., 2010).

1.2.9 Paradoxically large G+C content in two highly reduced genomes

As in the case of free-living bacteria with reduced genomes, host-obligated bacteria with reduced genomes also show large levels of A+T content (Delaye et al., 2010; Moran, 2003; Moya et al., 2008). It is hypothesized that the loss of repair enzymes due to genome reduction, in combination with reduced efficacy of natural selection, and a universal G:C to A:T mutational bias in bacteria, is

responsible for the observed trend (Hershberg & Petrov, 2010; Hildebrand et al., 2010; Van Leuven & McCutcheon, 2012).

However, there are notable exceptions to the above rule. The cicada *Diceroprocta semicincta* contains in its bacteriome two symbiotic bacteria, the α -Proteobacteria "*Candidatus* Hodgkinia cicadicola" and the Flavobacteria "*Candidatus* Sulcia muelleri" (McCutcheon et al., 2009a, 2009b). The genome of "*Ca. Hodgkinia cicadicola*" is in itself a paradox for molecular evolutionary theory because it has relatively large G+C content (~58%) despite having an extremely small genome (~144 kbp). This high G+C content is not due to biased gene conversion, since this bacterium lacks repair enzymes. Also, the high G+C content is not due to an A:T to G:C mutational bias since it is known that "*Ca. Hodgkinia cicadicola*" suffers from the same G:C to A:T mutational pressure universally present in bacteria (Van Leuven & McCutcheon, 2012).

One possibility that was suggested to explain this unexpectedly large G+C content is that the demography of its host *D. semicincta* inflates the population size of "*Ca. Hodgkinia cicadicola*" making natural selection more efficient to counterbalance the G:C to A:T mutational bias (Van Leuven & McCutcheon, 2012). However, it is not clear what benefit could confer single G:C over A:T polymorphisms in these bacteria. Other possibilities are a selection for better DNA replication, and/or DNA packing on C+G-rich genomes (Hershberg & Petrov, 2010). A similar situation is found in "*Candidatus* Tremblaya princeps," symbiont of the citrus mealybug *Planococcus citri*, which also shows an extremely

reduced genome of ~139 kbp and a relatively high G+C content (López-Madrigal, Latorre, et al., 2013; McCutcheon & von Dohlen, 2011).

1.2.10 Unexpected loss of genomic stability

During the early stages of reduction, genome architecture is quite unstable as discussed above. Many rearrangements occur over relatively short periods of time. As an example of this, the comparison of the genome architectures of six free-living *Serratia* and three “*Ca. Serratia symbiotica*” endosymbionts, showed surprising amounts of genomic rearrangements suffered between the three *S. symbiotica* lineages (Manzano-Marín & Latorre, 2014).

However, as the endosymbiosis evolves toward the last stages of reduction, genomic stability increases to a stalemate such that gene order conservation was considered one of the hallmarks of the genomes of obligate mutualistic bacteria. This was first noticed among genomes of *Buchnera aphidicola* from different strains with more than 50 million years of divergence. Initially, this high degree of conservation in gene order was attributed to the absence of the *recA* gene in these bacteria. Its product, RecA, is the key enzyme in homologous recombination repair (Shigenobu et al., 2000; Tamas et al., 2002; van Ham et al., 2003).

Homologous recombination is a high-fidelity DNA repair mechanism for double-strand break. The broken DNA is processed at the ends by several possible pathways producing a 3'-tailed duplex onto which RecA is loaded. RecA is an ATP-dependent multifunctional enzyme, which has recombinase activity.

RecA assembles into a filament and then sequesters the template double-stranded DNA, where it looks for the homologous loci, exchanges DNA strands, and forms joints between recombining molecules which allows the recombination and repair of the broken chromosome (Spies, 2013; Wigley, 2013).

However, the hypothesis that genome stasis is the result of loss of *recA* lost support when it was found that genome sequences from other endosymbionts like *Blattabacterium*, *Carsonella*, and *Wigglesworthia* showed similar levels of synteny conservation despite coding for *recA* (reviewed in (Sloan & Moran, 2013)). Additionally, recent genome sequencing projects showed that lack of repair and recombination genes may not be the cause of genome stability. “*Candidatus Portiera aleyrodidarum*” the primary endosymbiont of whiteflies (*Bemisia tabaci*) shows genome structural polymorphisms (i.e., lack of synteny) despite lacking *recA* and having one of the most reduced repair and recombination gene sets. These polymorphisms are demonstrated to be present even within bacteria inhabiting individual hosts and likely within individual bacterial cells. The presence of such structural polymorphisms was attributed to recombination events between large intergenic regions and repetitive elements that in turn are maintained by gene conversion (Sloan & Moran, 2012b, 2013).

A similar case of loss of genome stability was found in “*Ca. Tremblaya princeps*.” This bacterium shows a 7,032 bp region flanked by inverted repeats that are found in both orientations in the population. And similar to “*Ca. Portiera aleyrodidarum*,” this bacterium also codes for a highly reduced set of DNA

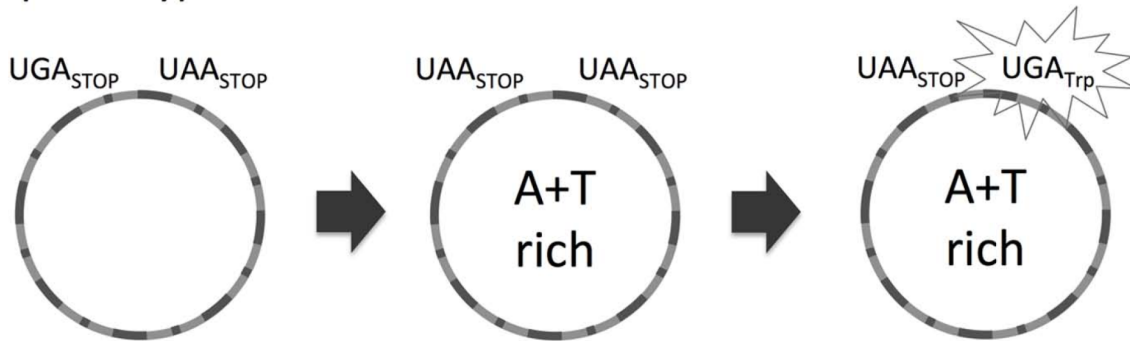
replication, recombination, and repair enzymes (López-Madrigal, Latorre, et al., 2013; McCutcheon & von Dohlen, 2011).

1.2.11 Novel hypothesis to explain the reassignment of STOP to Trp (tryptophan) codon

The obligate endosymbionts “*Candidatus* Nasuia deltocephalinicola,” “*Candidatus* Zinderia insecticola,” and “*Ca.* Hodgkinia cicadicola” evolved an alternative genetic code in which the codon UGA is reassigned from coding to stop (UGA_{stop}; UGA stop-coding codon) to code for tryptophan (UGA_{Trp}; UGA tryptophan-coding codon). This codon reassignment has been observed also in mycoplasmas and some mitochondrial genomes (Bennett & Moran, 2013; McCutcheon et al., 2009a).

The evolution from UGA_{stop} to UGA_{Trp} has been explained with the “capture” hypothesis (Figure C.1.5). According to this model, all UGA codons mutate first to its synonymous codon UAA in A+T rich genomes. This change does not affect protein length or fitness. Then, when UGA re-appears through mutation, it is free to be “captured” by an amino acid, in this case, Trp (tryptophan). The fact that almost all reassignments of the UGA_{stop} to UGA_{Trp} evolved in A+T-rich genomes supports this hypothesis (Osawa & Jukes, 1989).

Capture hypothesis



Loss of EF2 hypothesis

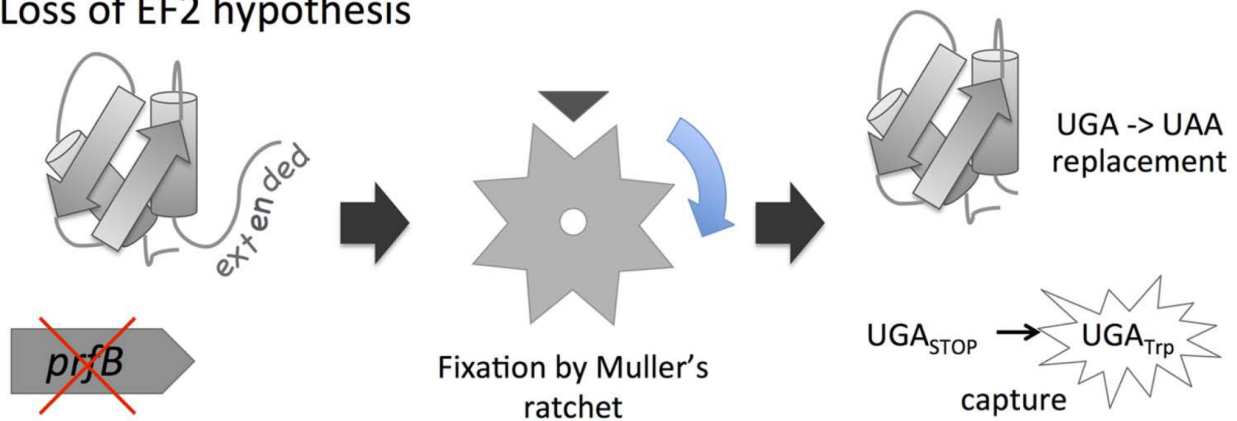


Figure C1.5 | The reassignment of the stop codon (UGA_{stop}) to code for tryptophan (UGA_{Trp}) is explained using the “capture hypothesis.” However, the large G+C content in “*Ca. Hodgkinia cicadicola*” makes the capture hypothesis unlikely in this organism. Instead, it is hypothesized that the loss of translational release factor RF2 triggered the evolution of this rearrangement.

Nonetheless, “*Ca. Hodgkinia cicadicola*” has a relatively high G+C content of about 58.4% making the “capture” hypothesis an unlikely explanation for this phenomenon. Therefore, a new hypothesis was proposed to explain the reassignment of UGA_{stop} to UGA_{Trp} in this bacterium. The new hypothesis proposes that the loss of translational release factor RF2 (encoded by *prfB*) that

recognizes UGA as a stop codon triggered the evolution of this rearrangement (McCutcheon et al., 2009a). According to this scenario, the loss of *prfB* caused that some proteins were translated with an extended sequence. These “extended” proteins were not necessarily lethal, perhaps just slightly deleterious. This is supported by experiments showing that extended proteins can increase fitness under stress conditions in yeast (Halfmann et al., 2012). Since symbiotic bacteria from insects are subject to recurrent bottlenecks, these slightly deleterious mutations become fixed by genetic drift under a Mullerian ratchet process. Then, once they are fixed, natural selection favored the replacement of UGA_{stop} codons with functional UAA or UAG, thus restoring the original size of proteins. This in turn allows UGA to be captured by tRNA-Trp (McCutcheon et al., 2009a).

1.2.12 Clues to early life?

The genome of “*Candidatus* Riesia pediculicola” the symbiont of the body louse *Pediculus humanus corporis*, codes for what seems to be a minimal tRNA decodification set (Kirkness et al., 2010). This endosymbiont lost all the enzymes that modify the tRNA body and kept only those genes that make modifications of the anticodon-stem-loop which are essential for mRNA decoding. Therefore, it was suggested that the minimal tRNA decodification set of this bacterium could resemble that of very ancient cells that existed during the early evolution of life on Earth (Kirkness et al., 2010). In support of this hypothesis, pseudouridine synthase A (encoded by *truA*), which is the enzyme responsible for the formation of pseudouridine at positions 38, 39, and 40 in the anticodon stem-loop of tRNAs

was suggested to be present in the last common ancestor of all extant life, or cenancestor (Ouzounis et al., 2006). However, at this moment, there is no evidence that the rest of the enzymes of the tRNA codification set of “*Ca. Riesia pediculicola*” are as ancient as *truA*. Nevertheless, this tRNA modification set exemplifies how simpler cellular systems can work. Which is highly relevant for synthetic biology approaches to the minimal cell. Whether other symbionts offer clues to early life on Earth still has to be carefully discussed.

1.2.13 The role of horizontal gene transfer in the evolution of intracellular symbiosis

One of the most striking peculiarities of host-associated bacteria with reduced genomes is how these organisms perform their symbiotic function and all the necessary processes to maintain themselves with such a reduced gene set. There are at least three non-mutually exclusive possibilities (McCutcheon & von Dohlen, 2011). In the first place, modifications of some genes coded in the reduced genome could allow the endosymbiont to cope with the loss of otherwise essential genes; second, the presence of complementary genes in the genomes of co-symbionts (if any) may compensate for gene losses in the endosymbiont; and third, genes coded in the genome of the host compensate for gene losses in the genome of the endosymbiont. From an evolutionary point of view, this last group of genes could be of host origin, or originally from the endosymbiont and transferred to the host, or horizontally transferred from

unrelated organisms not participating in the symbiosis to the host genome or its endosymbionts (McCutcheon, 2010; McCutcheon & von Dohlen, 2011).

Horizontal gene transfer (HGT) is one of the main forces in prokaryotic evolution (Zhaxybayeva & Doolittle, 2011). Recent discoveries show that HGT has played a role in the evolution of some obligate mutualistic symbiosis. For instance, "*Ca. Carsonella ruddii*" the obligate symbiont from the psyllid *Pachypsylla venusta* has one of the smallest genomes with 213 genes and ~160 kbp. This organism lives in the absence of other co-symbionts (Nakabachi et al., 2006). And as expected for such a small genome, a detailed analysis of its gene content indicated that several functions considered essential for a cell are missing (Tamames et al., 2007), raising the question of how these bacteria accomplish their symbiotic function. A recent transcriptomic analysis showed that the biosynthesis of essential and non-essential amino acids is performed collaboratively by the symbiont and by genes expressed in the bacteriocytes, some of them of bacterial origin, and at least one of them directly acquired from "*Ca. Carsonella ruddii*" (Sloan et al., 2014).

Similarly, the genome of the pea aphid (*A. pisum*) does contain genes of bacterial origin that are highly expressed in the bacteriocytes and likely participate in the symbiosis with *B. aphidicola* Aps (International Aphid Genomics Consortium, 2010; Nikoh & Nakabachi, 2009). These functional genes in *A. pisum* were acquired from bacteria other than its primary endosymbiont *B. aphidicola* Aps (Nikoh et al., 2010). Noteworthy, it was recently shown that the protein RplA4, coded by one of these genes, is targeted to the cytoplasm of the *B.*

aphidicola Aps (Nakabachi et al., 2014). This finding has been interpreted as blurring the distinction between endosymbionts and organelles (McCutcheon & Keeling, 2014). In addition, the only genes of *B. aphidicola* Aps origin in the genome of *A. pisum* are two highly truncated pseudogenes (Nikoh et al., 2010). Furthermore, experimental evidence has shown that the biosynthesis of some of the essential amino acids provided by *B. aphidicola* is performed partially by host enzymes (Russell et al., 2013).

Another case is found in the citrus mealybug *P. citri* where at least six distinct lineages of bacteria contributed with horizontally transferred genes to its nucleus. These genes code for protein products that complement the biosynthesis of essential amino acids, vitamins, and peptidoglycan in their endosymbionts “*Ca. Tremblaya princeps*” and “*Candidatus Moranella endobia*” (Husnik et al., 2013). Also, HGT contributed to the acquisition of toxicity in other symbiotic systems. The genome of “*Candidatus Proffotella armatura*” the symbiont of the Asian citrus psyllid (*Diaphorina citri*) acquired by HGT genes for the synthesis of cytotoxic polyketides. In this tripartite symbiosis, “*Ca. Proffotella armatura*” produces the polyketides, while another bacterium from the genus *Carsonella* provides the host with essential amino acids (Nakabachi et al., 2013).

The recent genome sequencing of the filarial nematode *Brugia malayi* showed that approximately 10.6% of the genome of its symbiont, *Wolbachia* wBM has been transferred to the eukaryotic genome. Interestingly, there is evidence that some of the genes coded in these regions are transcribed in particular stages

of the life cycle of the nematode suggesting functionality. However, their role in symbiosis still has to be determined (Ioannidis et al., 2013).

And finally, the synthesis of essential amino acids in symbiont-harboring trypanosomes is carried in part by genes of bacterial origin coded in the genome of the protist (Alves et al., 2013). These data clearly show that HGT is an important force in the evolution of some intracellular symbiosis (Sloan et al., 2014).

However, not all symbiosis shows evidence of HGT. The genome sequence of the body louse *P. h. corporis* does not appear to contain any genes of prokaryotic origin, indicating no transfer from its endosymbiont "*Ca. Riesia pediculicola*" strain USDA, nor other bacteria (Kirkness et al., 2010). Similarly, "*Candidatus Endolissoclinum faulkneri*" a defensive symbiont that also produces cytotoxic polyketes and that inhabits *Lissoclinum patella*, a colonial filter-feeding tunicate, does not seem to have acquired this capacity through HGT (Kwan et al., 2012).

1.2.14 Biochemical complementarity and convergent evolution of co-resident symbionts

In cases where there are more than one species of the obligate mutualistic symbiont, the biosynthesis of relevant metabolites for the host often requires the participation of enzymes that are coded in both co-symbionts and in some cases, as reviewed above, in the insect host (McCutcheon & von Dohlen, 2011). Furthermore, there are some occasions where different symbiotic systems

conformed by two co-symbionts have converged independently to the same division of labor regarding the biosynthesis of amino acids that are provided to their hosts.

In this sense, *B. aphidicola* BCc, from the aphid *C. cedri*, requires the co-symbiont “*Ca. Serratia symbiotica*” to provide Trp to its host. In this symbiotic system, the first two genes of Trp biosynthesis (*trpEG*) are located in a plasmid in *B. aphidicola* BCc, while the rest of the genes (*trpDCBA*) are located on the main chromosome of “*Ca. Serratia symbiotica*” (Figure C1.6; (Gosalbes et al., 2008; Pérez-Brocal et al., 2006). The same phenomena occur in the symbiotic system of the psyllid *Heteropsylla cubana*, where one of the symbionts (tentatively classified as a secondary symbiont) lost nearly all genes for the biosynthesis of essential amino acids except those of the (*trpDCBA*) operon, however, complementing the Trp biosynthetic capabilities of “*Ca. Carsonella ruddii*” HC (Sloan & Moran, 2012a). As exemplified above, the secondary symbiont of *H. cubana* and “*Ca. Serratia symbiotica*” have both evolved convergently to code for the same genes required for the biosynthesis of Trp (*trpDCBA*).

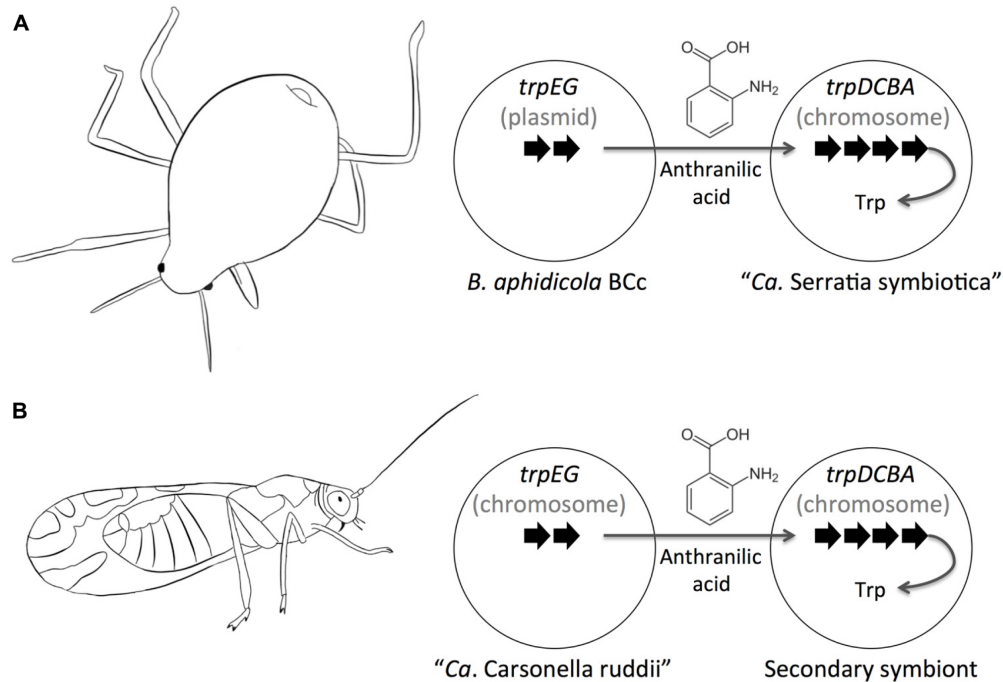


Figure C1.6 | Convergent evolution of Trp biosynthesis. The biosynthesis of Trp is performed cooperatively by two different endosymbionts in the aphid *Cinara cedri* (A) and in the psyllid *Heteropsylla cubana* (B). Strikingly, the same division of labor evolved in both systems.

A similar biosynthesis of Trp is observed in the case of *B. aphidicola* BCt, the symbiont of the aphid *C. tujaefilina*, where the first two genes of Trp biosynthesis (*trpEG*) are also located in a plasmid together with the structural genes for leucine synthesis. However, differing from *B. aphidicola* BCc, the biosynthesis of Trp does not occur cooperatively between two symbionts since the rest of the genes of Trp biosynthesis are located in the main chromosome of *B. aphidicola* BCt (Gil, Sabater-Muñoz, et al., 2006; Lamelas, Gosalbes, Moya, et al., 2011).

Another remarkable case of metabolic complementarity is provided by the co-symbionts *Baumannia cicadellinicola* and "*Ca. Sulcia muelleri*" Hc from the

xylem-feeding glassy-winged sharpshooters *Homalodisca coagulata*. For instance, *B. cicadellinicola* is able to make cysteine from homoserine and coenzyme A from 2-ketovaline but is unable to make homoserine and 2-ketovaline. And “*Ca. Sulcia muelleri*” Hc is able to make homoserine from aspartate and 2-ketovaline by the valine biosynthetic pathway. This kind of complementarity extends also to the biosynthesis of fatty acids, where *B. cicadellinicola* has all the genes necessary for the synthesis of fatty acids except for *fabF*, which in turn is encoded in “*Ca. Sulcia muelleri*” Hc. However, it is not clear how these molecules are transported across membranes since “*Ca. Sulcia muelleri*” Hc possesses few transporters (McCutcheon et al., 2009b).

Additionally, “*Ca. Sulcia muelleri*” Hc provides its host with eight of the ten essential amino acids (arginine, phenylalanine, tryptophan, lysine, threonine, isoleucine, leucine, and valine), while *B. cicadellinicola* produces the remaining two which are methionine and histidine. At the same time, the same scheme of amino acid provision by co-symbionts is found in the cicada *D. semicineta*. Both, *D. semicineta* and *H. coagulata* share the endosymbiont “*Ca. Sulcia muelleri*” which provides the same eight essential amino acids. However, *H. coagulata* and *D. semicineta* differ in their co-symbiont accompanying “*Ca. Sulcia muelleri*.” While *H. coagulata* contains *B. cicadellinicola*, *D. semicineta* hosts “*Ca. Hodgkinia cicadicola*.” This suggests that *B. cicadellinicola* and “*Ca. Hodgkinia cicadicola*” have evolved convergently to provide methionine and histidine to their host (McCutcheon et al., 2009b; Wu et al., 2006).

It is important to note that, as expected from convergent evolution, there are metabolic differences between *B. cicadellinicola* and “*Ca. Hodgkinia cicadicola*”, particularly in the biosynthesis of methionine. While *B. cicadellinicola* uses the cobalamin (vitamin B12)-independent version of methionine synthase, “*Ca. Hodgkinia cicadicola*” uses the cobalamin-dependent version of the enzyme. Both bacteria differ as well in the vitamin and cofactor biosynthetic capabilities (McCutcheon et al., 2009b).

In a similar fashion, the spittlebug *Clastoptera arizonana* contains in its bacteriome “*Ca. Sulcia muelleri*” CARI and “*Ca. Zinderia insecticola*” (McCutcheon & Moran, 2010). And as in the cases described above, both kinds of bacteria are needed to provide the ten essential amino acids to their host. However, in this case, “*Ca. Sulcia muelleri*” CARI cannot make Trp. Instead, this amino acid is synthesized by “*Ca. Zinderia insecticola*” in addition to methionine and histidine (McCutcheon & Moran, 2010).

Finally, the phloem-feeding aster leafhopper (ALF) *Macrostelus quadrilineatus* contains in separated cells within its bacteriome the symbionts “*Ca. Sulcia muelleri*” ALF and “*Ca. Nasuia deltocephalinicola*” ALF (Bennett & Moran, 2013). The genome of “*Ca. Sulcia muelleri*” ALF with 190,733 bp and 188 protein-coding genes is the smallest among sequenced genomes from this genus of bacterial symbionts. And the genome of “*Ca. Nasuia deltocephalinicola*” ALF with 112,091 bp and 137 protein-coding genes is the smallest bacterial genome sequenced so far. Similarly as above, “*Ca. Sulcia muelleri*” ALF codes for the genes necessary to synthesize the same eight essential amino acids and “*Ca.*

Nasuia deltocephalinicola” ALF codes for the genes required to produce methionine and histidine (Bennett & Moran, 2013).

It seems that “*Ca. Sulcia muelleri*” infected the ancestor of a large group of sap-feeding insects (Auchenorrhyncha) including the sharpshooter and cicada > 260 million years ago (McCutcheon & Moran, 2007). This is clearly seen by the fact that the genomes of “*Ca. Sulcia muelleri*” from *D. semicincta* and *H. coagulata* are almost collinear despite having diverged several million years ago (McCutcheon et al., 2009b).

Interestingly, *Nasuia* and *Zinderia* appear to be sister clades, which suggested the existence of an ancient lineage of β -proteobacterial endosymbionts hosted at least since the divergence of Cicadomorpha from Fulgoroidea 200 million years ago which are the only two clades in the suborder Auchenorrhyncha. This also suggests the loss and latter independent acquisition of *B. cicadellincola* and “*Ca. Hodgkinia cicadicola*” in the lineage leading to the sharpshooter and the cicada, respectively. These exchange in hosted symbionts was hypothesized to correlate with the new nutritional needs related to the diversification in the host diet (Bennett & Moran, 2013).

Similarly, metabolic complementation is suggested for the amino acid biosynthesis in the flavobacterium “*Ca. Walczuchella monophlebidarum*” where it was observed that many missing genes and pseudogenes in “*Ca. Walczuchella monophlebidarum*” were present in the γ -proteobacterial (Enterobacteriaceae) co-symbiont (Rosas-Pérez et al., 2014). Similarly as observed in the co-speciation between “*Ca. Sulcia muelleri*” and sap-feeding insects, co-speciation was

observed between Flavobacteria and several scale insects but not so with the Enterobacteriaceae symbiont suggesting a similar trend in metabolic complementarity (Rosenblueth et al., 2012).

Metabolic complementarity is also observed in the obligate symbionts of the Asian citrus psyllid *D. citri* which are “*Ca. Carsonella ruddii*” DC and “*Ca. Proffttella armatura*.” The metabolic capacities of both symbionts are largely non-redundant. For instance, the genome of “*Ca. Proffttella armatura*” encodes 16 genes for coenzyme transport and metabolism, including those for the synthesis of riboflavin and biotin, and the genome of “*Ca. Carsonella ruddii*” DC completely lacks these genes (Nakabachi et al., 2013).

Another extraordinary case of metabolic convergence is provided by *Blattabacterium* strain BGe which is the primary endosymbiont of the German cockroach *Blattella germanica* and the carpenter ant endosymbionts from the genus *Blochmannia* spp. (López-Sánchez et al., 2009). While *Blattabacterium* strain BGe is a member of the Bacteroidetes, *Blochmannia* spp. belongs to Proteobacteria. Despite the different phylogenetic origins of these bacteria, they resemble each other at the broad level of functional gene categories. A situation not found when *Blattabacterium* is compared with other insect endosymbionts like *Wolbachia* sp. and *Sulcia muelleri*. It seems that both bacteria converged due to the omnivorous diets of their hosts (Feldhaar et al., 2007; Patiño-Navarrete et al., 2014).

Finally, and perhaps one of the most striking examples of complementary is provided by one of the most extreme cases of symbiosis in nature. “*Ca.*

Tremblaya princeps,” a prokaryote with one of the smallest genomes, that as described above, lives inside the bacteriocytes of the mealybug *P. citri* and contains itself the bacteria “*Ca. Moranella endobia*” (López-Madrigal, Latorre, et al., 2013; McCutcheon & von Dohlen, 2011). In “*Ca. Tremblaya princeps*” most of its genes are devoted to RNA metabolism, the assembly of iron–sulfur [Fe–S] clusters, and to the partial biosynthesis of some essential amino acids. However, its genome does not code for any complete pathway. Therefore, “*Ca. Tremblaya princeps*” seems to depend for almost all basic functions on the coding capacities of “*Ca. Moranella endobia*” and likely from host-encoded proteins. In fact, it was proposed that “*Ca. Tremblaya princeps*” acquires the necessary cell components to function by sporadic cell lysis of “*Ca. Moranella endobia*” (McCutcheon & von Dohlen, 2011). The distribution of coded tRNAs in the genomes of “*Ca. Tremblaya princeps*” and “*Ca. Moranella endobia*” supports this hypothesis (López-Madrigal, Latorre, et al., 2013). However, immunohistochemistry assays with polyclonal antibodies to identify the location of the channel protein MscL coded only by “*Ca. Moranella endobia*” and GroEL coded by both bacteria, did not find evidence of massive and constitutive protein movement from the cytoplasm of “*Ca. Moranella endobia*” to the cytoplasm of “*Ca. Tremblaya princeps*” (López-Madrigal, Balmand, et al., 2013).

1.2.15 Genome reduction in bacterial symbionts of fungi, a relatively unexplored world

Bacteria, and fungi are commonly co-inhabitants of a large variety of niches on Earth. Both groups of microorganisms are the major colonizers of terrestrial and aquatic environments, and both have been revealed as intracellular guests of eukaryotic hosts (Bonfante & Genre, 2008; Gibson & Hunter, 2010; Moran et al., 2008). Despite the fact that bacterial-fungal interactions are ubiquitous and relevant for industry, agriculture, and medicine (Frey-Klett et al., 2011), intracellular bacterial symbionts of fungi remain largely unexplored.

Bacteria living inside fungal cells were first well documented in the AM (arbuscular mycorrhizal) fungus *Geosiphon pyriformis*. The symbiosis of this fungus with the filamentous cyanobacteria *Nostoc punctiforme* is the only known example of fungal endocyanosis to date. Interestingly, this intracellular symbiosis is cyclical, which means that the incorporation of free-living cyanobacteria within the fungal cytosol occurs periodically and only when the cyanobacteria are in the appropriate developmental state (Mollenhauer et al., 1996; E. Wolf & SCHUBERT, 2005). This cyclical transmission resembles the one that exists between plants and rhizobia, and plants and mycorrhizal fungi. *N. punctiforme* has the capability of establishing symbiosis not only with *Geosiphon* but also with gymnosperm cycad *Macrozamia* sp. (Rippka, 1992). Despite its endosymbiotic nature, the genome size of *N. punctiforme* is quite large (8.9 Mb), where 29% of the predicted genes seemed to be unique to this free-living and symbiotic cyanobacterium (Meeks et al., 2001).

Besides the endocyanosis, *G. pyriformis* harbors another type of intracellular bacteria, which were previously referred to as “bacterial-like organisms.” Recent molecular studies on these bacterial-like organisms have revealed that these endofungal bacteria belong to a monophyletic clade of the *Mollicutes*, which form a sister group to *Mycoplasmatales* and *Entoplasmatales* (Naumann et al., 2010). These obligate and heritable bacterial symbionts are widespread in several lineages of AM fungi including *Archeosporales*, *Diversisporales*, and *Glomerales*, which suggests that this symbiosis preceded the diversification of AM fungi and thus, that it may be as old (about 400 million years) as the symbiosis of AM fungi with plants (Naumann et al., 2010). Currently, no genomic information about these diverse *Mollicutes*-related symbionts is available. Furthermore, and largely due to their unculturability, their characteristics, functional roles, and capacities remain cryptic.

Among AM fungi, the best-studied system in terms of their bacterial endosymbionts is the AM fungus *Gigaspora margarita* BEG34. Pioneering studies on this AM fungus clearly showed that it harbored an intrahyphal, gram-negative, and vertically transmitted β -proteobacteria from the *Burkholderiaceae* family named “*Candidatus* Glomeribacter *gigasporarum*” (Bianciotto et al., 1996, 2003, 2004). In this relationship, the fungus is an obligate symbiont of plants, but facultative with respect to its bacterial symbiont (Desirò et al., 2014). Studies made with cured/endobacterium-free fungal spores indicated that *G. margarita* BEG34 can survive, without its bacterial partner, albeit showing signs of diminished ecological fitness (Lumini et al.,

2007). However, “*Ca. Glomeribacter gigasporarum*,” although it can be extracted from the fungus, remains unculturable in laboratory settings (Salvioli et al., 2008). It has been shown that in several strains of *G. margarita* both, the gram-positive *Mollicutes*-related bacteria as well as the gram-negative “*Ca. Glomeribacter gigasporarum*,” may be present (Desirò et al., 2014).

The genome of “*Ca. Glomeribacter gigasporarum*” revealed that these endobacteria possess a reduced genome (1.72 Mbp) with a relatively high G+C content of 54.82%, and a notorious dependency on its host for carbon, phosphorus, and nitrogen as well as energy, which is likely obtain from transporting and degrading amino acids (Ghignone et al., 2012). Phylogenetically, “*Ca. Glomeribacter gigasporarum*” closest relative is the also endofungal bacteria *B. rhizoxinica*, but analyses based on the metabolic pathways and their completion grouped this bacterium closer to some other endosymbionts of insects like “*Candidatus Hamiltonella defensa*,” spp. and *Wigglesworthia glossinidia*, suggesting that both fungal and insect intracellular symbionts had undergone convergent evolution (Ghignone et al., 2012).

A contrasting bacterial–fungal symbiosis was reported on *R. microsporus*, a zygomycete fungus, which was known as the causal agent of rice seedling blight thanks to the production of the toxin and potent antitumor agent rhizoxin (Gho et al., 1978; Takahashi et al., 1987). (Partida-Martinez & Hertweck, 2005) demonstrated that rhizoxin is not produced by the fungus, but by intracellular living bacteria from the Genus *Burkholderia*. The successful isolation and cultivation of the bacteria without its fungal host provided direct evidence of the

bacterial origin of rhizoxin and their derivatives (Partida-Martinez & Hertweck, 2005, 2007; Scherlach et al., 2006), and paved also the way for establishing a model for bacterial-fungal symbioses. All endosymbiotic bacteria associated with toxinogenic *R. microsporus* formed a defined cluster among *Burkholderia* and were later taxonomically defined as *B. rhizoxinica* and *Burkholderia endofungorum* (Partida-Martinez, Groth, et al., 2007), the latter being the type strain which besides rhizoxin, is able to produce the cyclopeptide toxin rhizonin A (Partida-Martinez, de Looß, et al., 2007). Strikingly, it was also demonstrated that cured, that is *R. microsporus* strains in which the bacteria have been eliminated by antibiotic treatment, are unable to form asexual sporangia and sporangiospores, meaning that the fungal host depends on endobacteria not only for the production of “mycotoxins,” but also for its asexual reproduction. These experiments also confirmed that bacteria are transmitted vertically, as single bacterial cells were encapsulated in fungal spores, warranting the maintenance of the symbiosis along generations (Partida-Martinez, Monajembashi, et al., 2007).

Further evolutionary studies on the *Rhizopus–Burkholderia* symbiosis have suggested that it is not as antique as the one between the AM fungi *G. margarita* and “*Ca. Glomeribacter gigasporarum*” (Castillo & Pawlowska, 2010) and that it has likely emerged as a shift from parasitism to mutualism, as mucoromycotina and some branches of the Ascomycota independently evolved rhizoxin resistant mechanisms before the establishment of the symbiosis (Schmitt et al., 2008).

Analyses of the first published endofungal genome of the type strain of *B. rhizoxinica* revealed a relatively reduced genome of 3.75 Mbp. These analyses

suggested that *B. rhizoxinica* is in an early phase of endosymbiosis, as many pseudogenes, MGEs, and transposons are present in its genome (Lackner, Moebius, Partida-Martinez, & Hertweck, 2011; Lackner, Moebius, Partida-Martinez, Boland, et al., 2011). Interestingly, and in contrast to other symbiotic systems which are also predicted to be in an early phase of evolution such as the chromatophore in *Paulinella chromatophora* (Nowack et al., 2008), the genome of *B. rhizoxinica* has shown that around 30.8% of its coding genes had no homology to any other organisms sequenced to date. This means that *B. rhizoxinica*'s genome is not simply a subset of genes derived from free-living *Burkholderia* spp. (Lackner, Moebius, Partida-Martinez, Boland, et al., 2011).

Despite this moderate genome reduction and evidence of a genome in transition, it is striking the great dependency of both partners on each other. These facts prompt questions such as: how fast can intracellular mutualisms between bacteria and fungi be established? Which molecular mechanisms contributed to this dependency? Under which circumstances would the *Burkholderia* endosymbionts evolve to a completely reduced genome close to the status of an organelle? Are host switching and/or population effective size contributing to genome stability in this symbiosis?

Very recently, the discovery of *Mollicutes*-related endobacteria not only in *Glomeromycota* but also inside several strains of *Endogone* (Desirò et al., 2015), a fungal genus belonging to Mucoromycotina which forms intimate associations with the earliest groups of land plants, has raised further questions about the ecological and evolutionary relevance of endofungal bacterial symbionts in the

establishment of the symbiosis of plants with fungi and ultimately, in the establishment in land by plants (Bidartondo et al., 2011). All these reports lend support to the idea that bacterial endosymbiosis in fungi is ancient and may have started within ancestral fungal members characterized by coenocytic mycelium (Desirò et al., 2015).

Certainly, further comprehensive molecular studies from the aforementioned, as well as other bacterial–fungal symbiosis known to date, will shed light on the evolutionary patterns of genome reduction and stability in these systems. Recent reports on intracellular gram-negative bacteria producing N-homoserine lactones in the zygomycete *Mortierella alpina* A-178 (Kai et al., 2012), together with diverse bacteria associated with endophytic fungi of the Ascomycetes (Hoffman & Arnold, 2010), and the endobacterial communities associated with the ectomycorrhizal fungus *Laccaria bicolor* (Bertaux et al., 2005) are expanding the universe of close interactions between bacteria and fungi, enabling a deeper understanding of the commonalities and differences between these symbioses and the best known bacteria-insect models. Moreover, some genome projects from fungi and bacteria engaged in intracellular symbioses had been recently undertaken as indicated on the website of the Department of Energy Joint Genome Institute (Nordberg et al., 2014). Thus, the consequences of such endosymbioses may be soon evaluated from both sides of the partnership.

1.2.16 Drivers of genome reduction in host-associated bacteria

The process of genome reduction in host-associated bacteria is largely determined by the intracellular environment in which they live. Specifically, it is reasoned that genes unnecessary for living in intracellular conditions are not maintained by selection and are lost along evolution. The process of genome reduction has been documented and seems to follow common trends from one stage of reduction to the next (Toft & Andersson, 2010). However, the mechanisms that drive these changes are far from established. In this section, we will review some of the more prominent and recent hypotheses about what drives genome reduction in host-associated bacteria.

Currently, one of the most prominent and widely accepted hypotheses to explain genome reduction is based on the process known as Muller's ratchet, which states that in populations undergoing constant bottlenecks and no recombination, genome reduction occurs through the accumulation of slightly deleterious mutations (McCutcheon & Moran, 2012; Moran, 1996). Under these conditions, selection fails to retain genes which then, by the constant accumulation of mutations, become inactive and are eventually deleted from the genome. As a result, several of the typical characteristics of these genomes, like their large A+T content or their small genomes, reflect known mutational biases (i.e., G:C to A:T mutations and deletions over insertions) rather than adaptations evolved by selection (McCutcheon & Moran, 2012; Moran, 2003; Moya et al., 2008).

In agreement with this hypothesis, theoretical studies suggest that proteins in *Buchnera* are less stable as a consequence of accumulating slightly deleterious mutations over large periods of time (van Ham et al., 2003). Additionally, proteomic studies demonstrate that the chaperonin GroEL is one of the most abundant proteins in *Buchnera* (Poliakov et al., 2011). As such, it is believed that GroEL plays a central role by stabilizing an otherwise unstable proteome (Fares et al., 2002; Moran, 1996). Furthermore, it has been shown that GroEL has suffered substitutions due to positive natural selection in two important functional regions of the protein which were suggested to be involved in the optimization of the ability to bind and prevent inappropriate folding of GroEL in *Buchnera* spp. and Flavobacteria endosymbionts (Fares et al., 2005). Also supporting this hypothesis, the pattern of nonsynonymous (dN) versus synonymous (dS) substitutions (dN/dS: nonsynonymous versus synonymous substitutions) among 42 pairs of closely related bacteria is consistent with genetic drift driving the process of genome reduction. Accordingly, dN/dS is consistently larger in organisms with smaller genomes (Kuo et al., 2009).

However, other processes have been suggested to explain the evolution of reduced genomes. For instance, (Itoh et al., 2002) suggested that the acceleration of molecular evolution experienced in these genomes is due to a general increase in the mutation rate rather than to Muller's ratchet mechanism. This hypothesis is based on: (a) the fact that the genomes of obligate mutualistic bacteria often lack DNA repair genes; (b) Muller's ratchet hypothesis is not consistent with the fact that the genomes of mutualistic endosymbionts, like those of *Buchnera* spp.

have 100s of millions of years of existence; and (c) the pattern of acceleration of molecular evolution of *Buchnera* spp. proteins is consistent with the increase of the mutation rate and not with the relaxation of purifying selection. In fact, the authors propose that loss of DNA repair genes is one of the necessary prerequisites to evolve a reduced genome. This hypothesis resembles the one proposed by (Marais et al., 2008) which also suggests that the increased mutation rate causes genomic reduction in free-living bacteria. However, the lack of repair genes coupled with recurrent bottlenecks and no recombination sets the conditions for evolution by Muller's ratchet.

Whatever the cause of the acceleration of the rate of evolution is, the hypothesis that these bacteria accumulate slightly deleterious mutations has to explain how these organisms manage to survive despite millions of years of existence. In this sense, compensatory evolution has to be part of the answer which suggests that a mutation may be compensated by a second mutation which returns the system to a working state (Fares et al., 2005; Kern & Kondrashov, 2004; McCutcheon & Moran, 2012). An example of which is the aforementioned chaperon overexpression which helps the organism to tolerate more mutations by lowering the threshold of the free energy necessary to fold properly compensating for the introduction of destabilizing mutations and making the system more robust (Gros & Tenaillon, 2009).

Importantly, Muller's ratchet hypothesis, although very compelling in the latter stages of reduction, falls short in the early stages where symbionts lack many of the prerequisites for this process to occur. Such as in facultative

pathogens, which due to their ability to return to a free-living state can evade bottlenecks and have larger population sizes. In addition, the early stages of reduction in pathogens are characterized by the acquisition of genes by HGT as well as their rapid modification by recombination (Toft & Andersson, 2010). Traditionally, genome reduction in host-associated bacteria has been linked to this view of evolution based on the relaxed or neutral selection coupled with genetic drift, since it better explains the presence of non-functional DNA such as ancient pseudogenes and intergenic regions commonly found in these bacteria (Dutta & Paul, 2012; McCutcheon & Moran, 2012; Moran & Mira, 2001). However, a recent shift in this view has started to appear in the form of empirical evidence (D'Souza et al., 2014; Koskiniemi et al., 2012; M.-C. Lee & Marx, 2012), and a new hypothesis that views genome reduction of host-associated bacteria as a selection-based process, at least on its early stages (Bliven & Maurelli, 2012; Mendonça et al., 2011; Morris et al., 2012). Here we will review a few of these hypotheses.

The first is an interesting novel hypothesis that suggests selective gene loss based on the loss of robustness in predictable environments. In this case, robustness is defined as the ability of an organism to withstand harsh and variable environments, as well as cope with internal changes and perturbations in the inner workings of the cell. The hypothesis predicts that under predictable environments such as the interior of a host's cell, this robustness is not required and thus, genomic reduction would be observed. In correlation with this, the authors found empirical evidence of the existence of a selective drive to retain

protein family diversity by sacrificing the redundancy of functions. Here, redundancy of function is a form of robustness. In other words, reduced genomes tend to have more protein families, but each family tends to have very few members. Thus, the authors suggest that the probability of losing a gene is higher if multiple copies of redundant genes exist, but very small if the function is unique. This also indicates that only those paralogs with similar functions will be lost. Finally, they suggest that other forms of robustness such as network redundancy may be similarly affected. For instance, the protein family composed of the transketolases TktA (EC2.2.1.1), TktB (EC 2.2.1.1), and the 1-deoxyxylulose-5-phosphate synthase Dsx (EC2.2.1.7) in *Escherichia coli* provides an example of this. The transketolases are 99% identical to each other but only 29% with respect to Dsx. In *B. aphidicola*, only one transketolase and one 1-deoxyxylulose-5-phosphate synthase remain (Mendonça et al., 2011).

Although originally proposed for free-living organisms, the BQH may also play an important role in the genome reduction of host-associated bacteria. A study by (D'Souza et al., 2014) showed that 76% of 949 sequenced bacteria were auxotrophic for at least one of 25 different metabolites needed for growth (20 amino acids, 3 vitamins, and 2 nucleosides), of which endosymbiotic bacteria were the most commonly observed auxotrophs (91% of endosymbiotic bacteria where auxotrophs for at least one of the 25 metabolites as opposed to 85% for free-living and 64% gut-inhabiting bacteria). Additionally, they observed that when supplemented with the metabolite, auxotroph strains of *E. coli* and *Acinetobacter baylyi* showed a significant increase in fitness as compared to the

wild type. The selective advantage depended on the concentration of the metabolite, the metabolite in question, and the absence or presence of a competitor (D'Souza et al., 2014).

And finally, another hypothesis that supports selection as the driver of reduction is the anti-virulence gene (AVG) hypothesis. This theory, proposed for pathogens, states that once a pathogen colonizes a new niche, its new role as a pathogen may be hindered by the expression of genes present and required in its previous environment. In order to better adapt and fulfill their role as a pathogen, these AVGs are selected against and end up inactivated or deleted. This theory is based on the concept of antagonistic pleiotropy, which states that the same gene may have adverse fitness in different environments. And thus, the AVG hypothesis may be considered not only for pathogens but also for other forms of symbiosis. An example of an AVG gene is that of *speG* in *Shigella* species. This gene codes for a spermidine acetyltransferase that generates N-acetylspermidine from spermidine. The loss of this gene prevents spermidine metabolism and allows for high levels of this compound in the cell. High spermidine concentration is correlated to higher survival to oxidative stress, which is of particular importance for *Shigella* spp. since part of its life cycle includes being swallowed by macrophages, and withstanding severe oxidative stress. Thus, the loss of this gene confers a higher fitness (Bliven & Maurelli, 2012).

1.2.17 Conclusion

In (Monterroso, 1959) the Latin-American writer Augusto Monterroso wrote one of the smallest stories in Spanish language called “The dinosaur”: “Cuando despertó, el dinosaurio aún estaba allí.” An approximate English translation would be “When he awoke, the dinosaur was still.” The story is composed of two parts separated by a coma. In the first one, a tacit subject awakes. In the second one, the subject realizes that a dinosaur, an explicit subject, “was still.” Despite its small size, all the elements of a story (i.e., character, setting, plot, conflict, and theme) are present in these just seven words, although some of these elements are implicit and left to the imagination of the reader.

Similarly, in the case of host-associated bacteria, extreme genome reduction is possible by metabolic and functional integration with the host and with other co-symbionts. And in the case of free-living bacteria, the BQH suggests that selection will favor the loss of those genes that code for expensive functions that are anyway provided as PGs by other species (Morris et al., 2012). In both cases, the outcome is dependence between different cellular lineages. And, as in the case of Augusto Monterroso’s story, in which some parts of the story are left implicit to the reader, parts of the functions required by the cells are performed outside their boundaries.

What cellular status do prokaryotes with extremely reduced genomes deserve? Perhaps the symbionelle concept is part of the answer (Reyes-Prieto et al., 2014). The symbionelle concept was constructed to accommodate those cases of endosymbionts that fail to reach a minimal gene set (Gil et al., 2004). They

possess evolved genomes with so few genes that they are not able to perform the three basic functions of present-day cells without the presence of a host and/or other co-symbionts, and so, represent a new category. These symbionelles present evolutionary convergence with organelles exhibiting clear and important similarities and distinctions, although each evolved in completely different evolutionary scenarios, where organelles evolved before multicellular life and symbionelles distinctly throughout insect evolution (Reyes-Prieto et al., 2014).

A concluding remark is made by (Y. I. Wolf & Koonin, 2013), in which they suggest that genome reduction is the dominant form of evolution in a two-phase genomic model where a short phase of abrupt increase in complexity, and thus genomic size, permits innovation while a long phase defined by genomic reduction allows for adaptation. Once again, demonstrates the importance of dependence as a form of adaptation born from symbiotic interactions. And finally, as pointed out by (Sloan & Moran, 2013), the biology of obligate intracellular mutualistic bacteria offers the opportunity to study the evolutionary process acting on different levels of biological organization. Thus, the development of a multilevel theory of causation stands at the frontier of evolutionary theory (S. J. Gould, 2002).

Chapter 2. Organizing and optimizing access to big data of endosymbiotic bacteria of insects

This chapter reproduces the following published papers in their entirety:

Reyes-Prieto, Mariana, Carlos Vargas-Chávez, Amparo Latorre, and Andrés Moya. "SymbioGenomesDB: a database for the integration and access to knowledge on host-symbiont relationships." *Database* 2015 (2015).

Reyes-Prieto, Mariana, Carlos Vargas-Chávez, Mercè Llabrés, Pere Palmer, Amparo Latorre, and Andrés Moya. "An update on the Symbiotic Genomes Database (SymGenDB): a collection of metadata, genomic, genetic and protein sequences, orthologs and metabolic networks of symbiotic organisms." *Database* 2020 (2020).

In the first decade of the genomic era, several publicly available databases offered information on specific organisms or models, but none offered a global understanding of relationships between organisms, their interactions, and capabilities within their niche, as well as their role as part of a system, in this case, their role in symbiosis. With an evident need for the optimization and organization of big data of endosymbiotic bacteria of insects, we began the task of creating the first publicly available composite database including symbiotic organisms. It was originally called SymbioticGenomesDB (<http://symbiogenomesdb.uv.es/>), now called SymGenDB, for short. It consisted of three modules where users could search for bacteria involved in a specific symbiotic relationship, their genomes, and their genes (including their orthologs).

Over the years, the ultimate goal of SymGenDB has been maintained, which was to host and support the growing and vast symbiotic–host relationship information, to help uncover the genetic basis of such associations. SymGenDB was originally thought of as a database to support information on symbiotic systems of insects, but since the workflow for symbionts of insects was the same for all organisms, fully sequenced and available in primary databases, it grew to maintain a comprehensive organization of information on genomes of symbionts from diverse hosts throughout the Tree of Life. The catalog of relationships was generated using computational tools, custom R scripts, and manual integration of data available in public literature. Our database first became publicly known

with its publication in the journal *Database* in 2015, and an update released in the same journal in 2020.

2.1. SymbioGenomesDB: a database for the integration and access to knowledge on host-symbiont relationships

Symbiotic relationships occur naturally throughout the Tree of Life, either in a commensal, mutualistic, or pathogenic manner. The genomes of multiple organisms involved in symbiosis are rapidly being sequenced and becoming available, especially those from the microbial world. Currently, there are numerous databases that offer information on specific organisms or models, but none offer a global understanding of relationships between organisms, their interactions, and capabilities within their niche, as well as their role as part of a system, in this case, their role in symbiosis. We have developed the SymbioGenomesDB as a community database resource for laboratories that intend to investigate and use the information on the genetics and the genomics of organisms involved in these relationships. The ultimate goal of SymbioGenomesDB is to host and support the growing and vast symbiotic–host relationship information, to uncover the genetic basis of such associations. SymbioGenomesDB maintains a comprehensive organization of information on genomes of symbionts from diverse hosts throughout the Tree of Life, including their sequences, their metadata, and their genomic features. This catalog of relationships was generated using computational tools, custom R scripts, and manual integration of data available in public literature. As a highly curated and comprehensive systems database, SymbioGenomesDB provides web access to all

information on symbiotic organisms, their features, and links to the central database NCBI. Three different tools can be found within the database to explore symbiosis-related organisms, their genes, and their genomes. Also, we offer an orthology search for one or multiple genes in one or multiple organisms within symbiotic relationships, and every table, graph, and output file is downloadable and easy to parse for further analysis. The robust SymbioGenomesDB will be constantly updated to cope with all the data being generated and included in major databases, in order to serve as an important, useful, and time-saving tool.

Database URL: <http://symbiogenomesdb.uv.es>

Symbiotic relationships are ubiquitous on our planet. They happen within every niche observable in our world, even within our bodies, and they are crucial in the maintenance of every ecosystem. Although the term symbiosis is sometimes confounded with mutualism, this intimate association can also be parasitic or commensal (Martin & Schwab, 2012).

Generally, symbiosis occurs between a eukaryotic partner forming this close relationship with one or multiple species of bacteria. Furthermore, the number of sequenced genomes is growing rapidly, especially in the microbial world (Quast et al., 2013). For this reason, the amount of information has become so immense that we need to find better ways to make it comprehensive and useful. The aim of SymbioGenomesDB is to facilitate research through the gathering of information from organisms involved in symbiotic relationships.

Initially designed to comprise all the newly sequenced and annotated genomes of endosymbionts of insects, SymbioGenomesDB has grown to include all the bacteria and even some eukaryotes, which are known symbionts worldwide. The catalog of these symbionts includes just above 1050 genomes, while the list of hosts comprises 216 organisms, most of them Eukaryotes.

The central mission of SymbioGenomesDB is to host and support the growing and vast symbiotic–host relationship information, to uncover the genetic basis of such associations. As a highly curated and comprehensive systems database, SymbioGenomesDB provides web access to this complete catalog of fully sequenced and annotated genomes of symbiotic organisms and their features, including genetic and genomic sequences, their genomic characteristics, and the symbiotic interaction in which they participate, as well as links to the central database NCBI (<http://www.ncbi.nlm.nih.gov/>) (Sayers et al., 2009). SymbioGenomesDB will be routinely updated to keep up with the growth rate of genomic sequences available, in order to serve as a recognized authority and a comprehensive data integration site and repository for symbionts’ genetic, genomic, and phenotypic data, derived from major data providers.

2.1.1 Data collection

The workflow we followed for the collection of the data is depicted in Figure C.2.1. Furthermore, the steps in the figure are thoroughly explained next:

1. First, we created a list of all of the symbiotic relationships denoted in GOLD (<http://www.genomesonline.org/>) (Pagani et al., 2012) and IMG (<https://img.jgi.doe.gov>) (Markowitz et al., 2012), through metadata searches. Both are highly curated databases that contain every registry of all genomes available up to date and those that are in progress, with highly curated and detailed metadata.

2. Next, we correlated the aforementioned list of symbionts to the genomes available in KEGG (Kanehisa et al., 2014; Kanehisa & Goto, 2000) to have the most information on each genome. We only included those genomes that are classified as finished or permanent drafts, to get the most complete set of information on each genome.

3. Following, we downloaded the genetic information of the genomes included in the Microbial Genomes Database archive (Uchiyama et al., 2010), and cross-referenced it to our list of symbiotic relationships.

4. We also completed the host field as an association to each symbiont with the metadata from GOLD and IMG, although most of the cases (80%) did not have a host associated. We manually curated the link between host–symbiont of all of the organisms without a host through literature searches.

5. This catalog was then validated when compared to a small dataset of 80 endosymbiotic bacteria of insects we had manually curated in our lab (unpublished data) which we achieved based on literature and found every relationship in accordance.

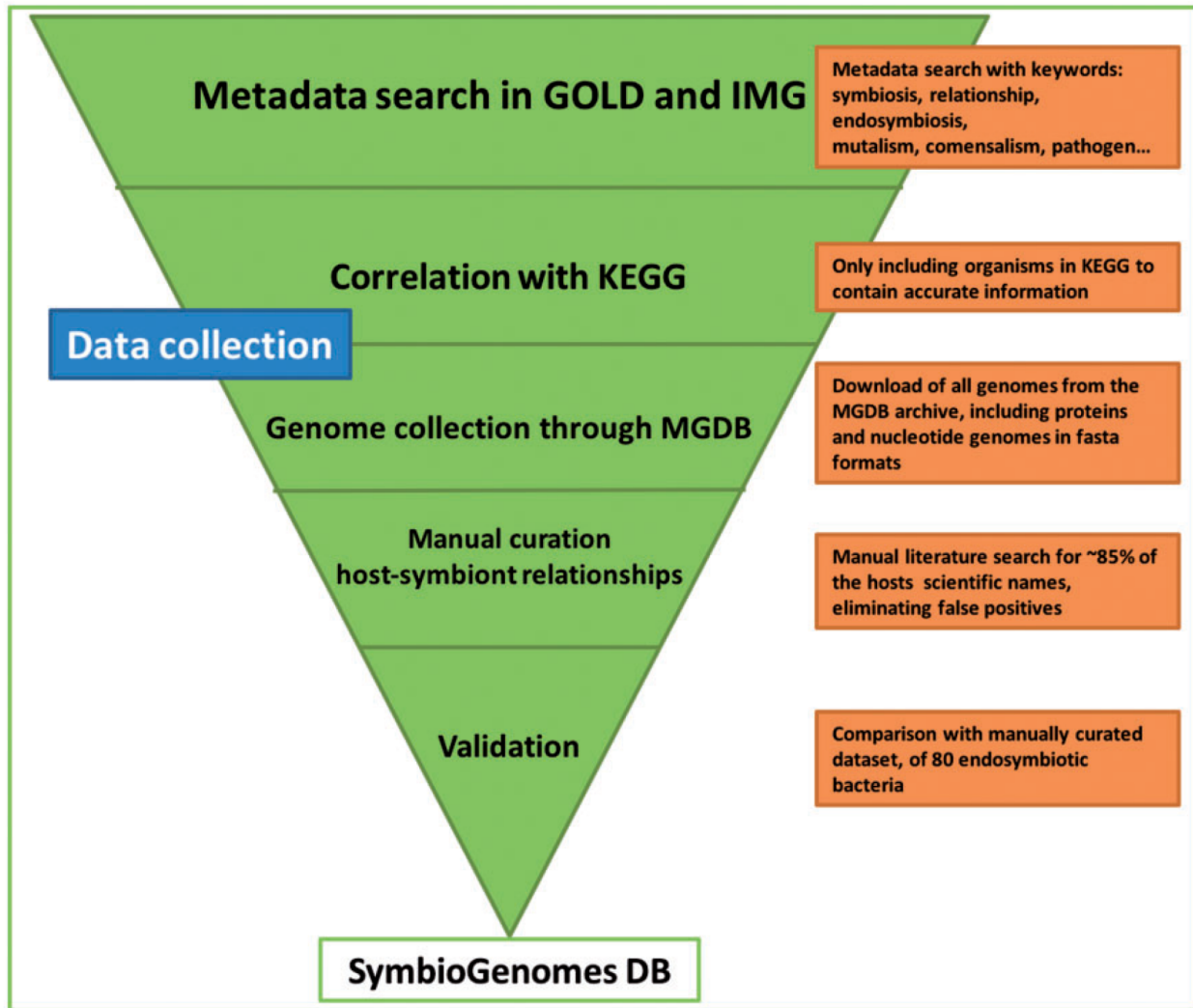


Figure C2.1 | Schematic representation of the data collection workflow. The graphic illustrates the major modules of the data collection and the action taken in each step. The pipeline is highly tunable, and every update will be easier and shorter since this first manual curation has been so thorough. The triangle shape is meant to convey that each step acts as a filter of data we are not interested in until we end up with the highly curated catalog that SymbioGenomesDB currently offers.

2.1.2 Database architecture and web interface

SymbioGenomesDB consists of an HTML front web interface, and the database itself is built using custom R (Team, 2009) scripts for the integration of data and parsing of the genomes for the database, coupled with the R Package Shiny from RStudio (RStudio, 2013). Shiny is a powerful web framework that allows us to build interactive web applications with R. It incorporates the computation power of R with the interactivity of web pages. Applications built with Shiny respond automatically as users type their queries and ask for their outputs by pressing buttons. Thus, when new queries are written into SymbioGenomesDB, the organisms, genomes, or genes searched for are automatically displayed on the screen. Also, for each genome, the size and metrics such as gene number and GC content, are automatically computed and shown as output. Users with no experience in programming languages can use the database without difficulty. The HTML web page was designed with Sandvox for Mac, and the server is published in the `symbiogenomesdb.uv.es` server hosted by the University of Valencia. SymbioGenomesDB is freely accessible at <http://symbiogenomesdb.uv.es>.

2.1.3 Database features

We have designed a user-friendly web interface for our database. Upon entrance, users will find the welcome page and the information to understand the full use and capabilities of the database, as well as everything it has to offer (Figure C2.2).

(a)

SymbioGenomesDB
Symbiotic Genomes Database for the integration and access to knowledge on host-symbiont relationships
Evolutionary Genetics Group, Cavanilles Institute of Biodiversity and Evolutionary Biology
University of Valencia, Spain

HOME | **FUNCTIONS** | QUICK TOUR | ACKNOWLEDGEMENTS AND CONTACT | DATABASE

Overview

LAST UPDATE: 06/03/2015

SymbioGenomes Database

The aim of the SymbioGenomesDB is to make research easier by compiling the available information of organisms in symbiosis, to support the growing and vast symbiotic-host relationship data, to help uncover the genetic basis of such associations.

SymbioGenomesDB maintains a catalogue of sequenced/finished genomes of symbionts from organisms throughout the tree of life, as well as their genomic features. The catalogue was generated with computational tools, custom R scripts, and manual integration of information available in public literature. As a highly curated and comprehensive systems database, SymbioGenomesDB provides web access to the catalogue of sequenced and annotated symbiotic genomes of symbiotic organisms, their features including genomic sequences and metrics, as well as links to the central database NCBI.

Because more data is being generated and included in major databases, SymbioGenomesDB will be updated and capable of providing information on these symbiotic organisms, as to become an important, useful and timesaving tool.

Metrics

1056 symbiotic genomes associated to more than 200 different hosts.

Overview of phylums:		
4 Cyanobacteria	102 Chlamydiae	1 Agricomplexa
488 Proteobacteria	7 Euryarchaeota	2 Microsporidia
173 Firmicutes	4 Fusobacteria	1 Thaumarchaeota
1 Nanoarchaeota	4 Ascomycota	2 Basidiomycota
136 Actinobacteria	2 Basidiomycota	1 Verrucomicrobia
39 Spirochaetes	50 Tenereutes	46 Bacteroidetes

60,322 ortholog clusters

Institut Cavanilles de Biodiversitat i Biologia Evolutiva
Universitat de València
Authors: Mariana Reyes-Prieto and Carlos Vargas-Chavez
Contact: mariana.reyes@uv.es, carlos.vargas@uv.es

UNIVERSITAT DE VALÈNCIA | Institut Cavanilles de Biodiversitat i Biologia Evolutiva

(b)

DATABASE

(c)

Functions

In SymbioGenomesDB, users are able to find all the symbiotic relationships that have been recorded in literature up to the last update, as well as the information available for each association. It consists of three tools separated in different tabs, each of which is designed for a specific search.

In the first tab, FIND ORGANISMS, it is possible for users to search and get an overview of specific associations between symbiotic organisms. In the second tab, FIND GENOMES, a list of the genome(s) involved in a specific symbiotic relationship of interest can be displayed. In the third tab, FIND GENES, users are able to search for orthologous genes of interest included in the symbionts of a given organism(s) included in our catalog. Each tab is fully explained next, for examples, please check out the quick tour.

(d)

Acknowledgements and contact

We thank LCG Leonardo Collado-Torres, Dr. Diego Santos-Garcia, Dr. Ana Gutierrez-Preciado and Dr. Pablo Yarzsa for their helpful suggestions. The whole team at the [Evolutionary Genomics group in the Cavanilles Institut of Biodiversity and Evolutionary Biology](#), Omar Ortuzar for the logo design. Special thanks to Dr. Ikuo Uchiyama, from the [MRGD database](#) for his collaboration and his time.

This project has been funded by CONACYT México, the EU Marie Curie Initial Training Network (ITN) Symbiotics: Molecular ecology and evolution of bacterial symbionts [FP7-PEOPLE-2010-ITN], and projects DFLU2012-002-01, cofinanced by FEDER funds and SAF2012-31187.

In case of any problems, questions or suggestions, please use the following form to get in touch with us, and we will be sure to respond as soon as possible. Thank you!

Figure C2.2 | Database overview. In HOME (a), there is a complete overview of the importance of the database and its purpose in detail. The line of buttons in the above green menu, as well as the menu on the right, denote the different parts of the web interface, with special importance to the button (b) “Enter Database”, which will open the database in Shiny from the R Studio. An explanation of the functions (c) of the database is also included, as well as a quick tour explained in detail throughout this article and the Acknowledgments (d) for the support of this work.

As aforementioned, in SymbioGenomesDB users are able to find all the symbiotic relationships that have been recorded in literature, as well as the information available for each association. It consists of three tools separated in different tabs, each of which is designed for a specific search. In the first tab,

FIND ORGANISMS, it is possible for users to search and get an overview of specific associations between symbiotic organisms. In the second tab, FIND GENOMES, a list of the genome(s) involved in a specific symbiotic relationship of interest can be displayed. In the third tab, FIND GENES, users are able to search for orthologous genes of interest included in the symbionts of a given organism(s) included in our catalog. These tools are fully described in the next section.

2.1.4 Find Organisms

In this tab, users are able to search for a specific symbiotic association of interest, by entering the common name, the species scientific name, the genus, the class, or any taxonomy level keyword based on NCBI's taxonomy. Users are also encouraged to enter any random organism they can think of, to explore and get a better understanding of the searches SymbioGenomesDB does, including shuffling between the taxonomic levels, since this is a useful way to start any type or analysis of these organisms. Users can explore through broad searches, from wider taxonomy levels, such as Kingdom or Families, to narrow searches, to the level of species and strains. Figure C2.3 is an example of a specific search of our interest.

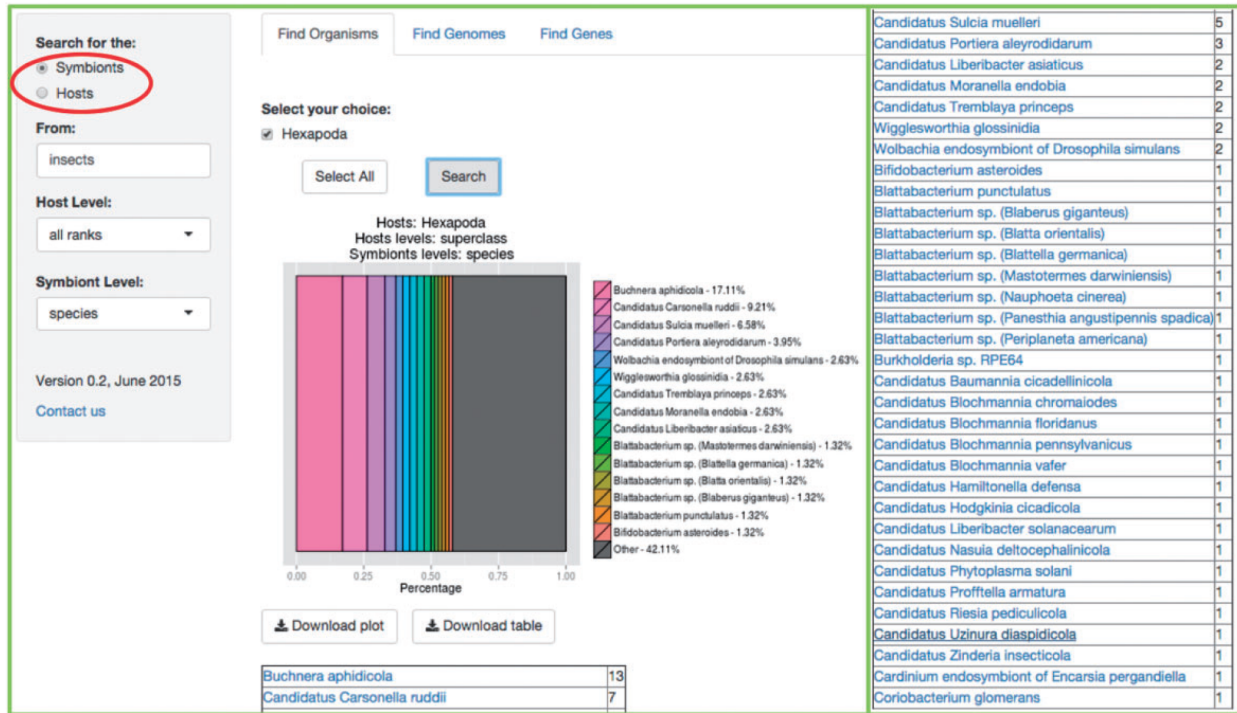


Figure C2.3 | Example of the Find Organisms tab. In our lab at the University of Valencia, we work with insect symbiosis (Martínez-Cano et al., 2014; Moya et al., 2008; Reyes-Prieto et al., 2014). This example shows the result of searching for the symbionts of ‘insecta’ with the default ‘all ranks’ host level and the ‘species’ level selected for symbionts. Be aware that the matching result indicates that the host level of the query is class since it is where a match was found. The results include an abundance graph including the first most representative 14 matches, plus a summary of the rest of the matches, a table of organisms matching the user's query, and the number of matches for each species, phylum, class, or whichever taxonomy level explored. The rest of the resulting table is at the right of the figure for space limitations.

It is important to denote which partner of the symbiosis will be searched for. It can either be the symbionts of your host of interest, or the other way around, and selecting the proper circle available indicates it (Red circle in Figure C.2.3).

Selection of the host(s) taxonomy level is available in the scrolling menu below the search bar (the default 'all ranks' will search in every taxonomic level and as a result, show the level where the match is found). In the same manner, a selection of the symbiont(s) taxonomy level of interest is also available (the default 'all ranks' is equal in every search). Typically, the results of the query entered will be shown in just milliseconds. Searching for an organism with a large number of associated symbionts (e.g. human or pig, since there are a lot of pathogenic bacteria associated with these species), can take a few seconds.

Both the resulting plot and the resulting list of organisms can be downloaded. Also, the resulting list of organisms (in Figure C.2.3, all the species involved in symbiotic relationships with insects) includes links to their respective species at NCBI Taxonomy.

2.1.5 Find Genomes

This search will retrieve the genomes associated with a specific host. Users can write the common name, the scientific species name, the class, the genus, etc., of their genome(s) of choice according to the taxonomical level of interest, in the search box (the default 'all ranks' will search in every taxonomic level). Figure C2.4 shows an example.

Search for genomes related to:

Level:

Version 0.2, June 2015
[Contact us](#)

[Find Organisms](#) [Find Genomes](#) [Find Genes](#)

Select your choice:

Hexapoda

Candidatus Proffotella armatura

Candidatus Riesia pediculicola USDA

Candidatus Sulcia muelleri CARI

Candidatus Sulcia muelleri DMIN

Candidatus Sulcia muelleri GWSS

Candidatus Sulcia muelleri SMDSEM

Candidatus Sulcia muelleri str. *Sulcia*-ALF

Candidatus Tremblaya princeps PCIT

 Displaying 4 of 76 available genomes:

Name:	<i>Candidatus Riesia pediculicola</i> USDA
Isolated from:	<i>Pediculus humanus</i>
Taxonomy ID:	515618
CDS with orthologs:	462
Strain-specific CDS:	136
Number of replicons:	2
Replicon name:	chromosome 1
Replicon size:	574,390
Replicon GC content:	28%
CDS	547
rRNA	12
tmRNA	1
tRNA	66
Replicon name:	plasmid pPAN
Replicon size:	7,737
Replicon GC content:	35%
CDS	12

Name:	<i>Candidatus Sulcia muelleri</i> CARI
Isolated from:	<i>Clastoptera arizonana</i>
Taxonomy ID:	706194
CDS with orthologs:	246
Strain-specific CDS:	39
Number of replicons:	1
Replicon name:	chromosome 1
Replicon size:	276,511
Replicon GC content:	21%
CDS	255
ncRNA	1
rRNA	6
tmRNA	1
tRNA	59

Name:	<i>Candidatus Tremblaya princeps</i> PCIT
Isolated from:	<i>Planococcus citri</i>
Taxonomy ID:	891398
CDS with orthologs:	127
Strain-specific CDS:	20
Number of replicons:	1
Replicon name:	chromosome 1
Replicon size:	138,927
Replicon GC content:	59%
CDS	129
rRNA	12
tmRNA	1
tRNA	24

Name:	<i>Candidatus Proffotella armatura</i>
Isolated from:	<i>Diaphorina citri</i>
Taxonomy ID:	669502
CDS with orthologs:	349
Strain-specific CDS:	23
Number of replicons:	2
Replicon name:	chromosome 1
Replicon size:	459,399
Replicon GC content:	24%
CDS	366
rRNA	3
tRNA	34
Replicon name:	plasmid unnamed
Replicon size:	5,458
Replicon GC content:	24%
CDS	6

Figure C2.4 | Example of the Find Genomes tab. The result of searching for the genomes associated with ‘insects’ with the default ‘all ranks’ taxonomic level. First, users get a scrolling menu from which they can select genomes of interest, or all genomes available in this search, which in turn displays a table that shows the names of the organisms selected in the scrolling menu, the precise host with which the symbiotic relationship exists, as well as metrics and characteristics of their genomes. The rest of the resulting table is at the right of the figure for space limitations.

In this case, the search can be done for a specific symbiont or a specific host. If the search is for a host, another menu listing all its symbionts' genomes will become available, where the genomes of interest can be selected.

In a matter of seconds, tables consisting of the metrics (genome size, NCBI's taxon id, CDSs, GC content, etc) of each genome selected, the symbionts' host, plasmids, and all chromosomes, in cases where more than one chromosome is present, will be shown.

An important feature embedded in this search is that users can download a table with all of the orthologous genes from the genomes they have searched for (at least two), with just one click of a button. This table can be easily parsed for further analysis. The complete set was obtained from the MBGD: Microbial Genomes Database (<http://mbgd.genome.ad.jp>) (Uchiyama et al., 2010). This orthology table includes six different fields of information for each orthology found, including the gene name of each ortholog, the description of the gene, and their id in four different databases, including MBGD, COG (Tatusov et al., 2003), KEGG, and TIGR (Haft et al., 2003). Then, each organism of the search is displayed as a column (abbreviated with their KEGG organism id as the header of the column), with a list of each ortholog shared with at least one other organism in the subset selected. The table(s), the FASTA file(s), and/or the GFF file(s) of the organisms resulting from the search are also available for download.

2.1.6 Find Genes

In this tab, users can search for genes in any symbiont present in our catalog, or a table of orthologous genes included in two or more genomes of interest, involved in symbiosis. These orthologs have been calculated with several algorithms, according to the MBGD: Microbial Genomes Database (Uchiyama et al., 2010).

Users need to write the gene of interest in the first search bar. If more than one gene is searched for, users must use commas to separate searches. In the second search bar, the common name, the scientific name, the class, the genus, etc., of the genome(s) of interest according to the taxonomical level of choice, must be written (the default 'all ranks' will search in every taxonomic level). As a result, two menus will be available, to select the gene(s) and genome(s) of interest, accordingly. The search can be done for a specific symbiont or a specific host. If searching for a host, another menu listing all its symbionts' genomes becomes available. Figure C2.5 shows an example.

Search for the genes:

In the genomes related to:

Level:

Version 0.2, June 2015
[Contact us](#)

[Find Organisms](#) [Find Genomes](#) [Find Genes](#)

Select your genes:

- trpA tryptophan synthase subunit alpha
- trpB tryptophan synthase subunit beta
- trpC bifunctional indole-3-glycerol phosphate synthase/phosphoribosylanthranilate isomerase
- trpD anthranilate phosphoribosyltransferase
- trpE anthranilate synthase component I
- trpG anthranilate synthase component II
- trpS tryptophanyl-tRNA synthetase

Select your choice:

- Buchnera
- Serratia

- Buchnera aphidicola str. 5A (Acyrtosiphon pisum)
- Buchnera aphidicola str. Ak (Acyrtosiphon kondoi)
- Buchnera aphidicola str. APS (Acyrtosiphon pisum)
- Buchnera aphidicola str. Bp (Baizongia pistaciae)
- Buchnera aphidicola str. JF98 (Acyrtosiphon pisum)
- Buchnera aphidicola str. JF99 (Acyrtosiphon pisum)
- Buchnera aphidicola str. LL01 (Acyrtosiphon pisum)
- Buchnera aphidicola str. Sg (Schizaphis graminum)

Displaying 8 of 24 available genomes:

Gene ID	Gene Description	buc	bap	bau	bab	bcc	baj	bas	ssz
<i>trpA</i>	tryptophan synthase subunit alpha	BU277	BUAP5A_272	BUAPTUC7_274	BBP257		BCTU_18	BUSG266	SCC_380
<i>trpB</i>	tryptophan synthase subunit beta	BU278	PL265	PL265	BBP258		BCTU_18	BUSG267	SCC_379
<i>trpC</i>	bifunctional indole-3-glycerol phosphate synthase/phosphoribosylanthranilate isomerase	BU279	BUAP5A_274	BUAPTUC7_276	BBP259		BCTU_18	BUSG268	SCC_378
<i>trpD</i>	anthranilate phosphoribosyltransferase	BU280	BUAP5A_275	BUAPTUC7_277	BBP260		BCTU_18	BUSG269	SCC_377

Figure C2.5 | Example of a FIND GENES search. (a) Searching for all the genes included in the tryptophan biosynthesis in the genomes related to insects. We get two scrolling menus and selected all the tryptophan genes and the genomes of several species of the *Buchnera* genus, as well as the species *Serratia symbiotica*. (b) The resulting table lists the orthologs found between

the genomes we selected, including the bacteria working as cosymbionts in the aphid *Cinara cedri*, that participate in an exceptional metabolic complementation of the tryptophan metabolic pathway (Gosalbes et al., 2008). (c) Even though the table shows the abbreviated genome names from KEGG (Kanehisa et al., 2014; Kanehisa & Goto, 2000), if you scroll over the name, you get the complete species name in a little box below the pointer. The resulting table will be a table including the genes and the genomes selected in the menus. Every output is available for download, as flat files for further and easy parsing and analysis.

2.1.7 Discussion and future directions

The boom in microbial genomes publications has necessitated the development of several tools to categorize and better organize the data we are given, to fully appreciate it for analyses, and gain as much knowledge as possible from it. As a highly curated and comprehensive systems database, SymbioGenomesDB provides access to the complete catalog of fully sequenced and annotated genomes of symbiotic organisms and their features including genomic sequences and metrics, which can be useful for research, reducing the time-consuming search for biotic relationships and the features of the organisms involved.

SymbioGenomesDB is a core component of an extensive set of genome informatics resources that comprises several microbiology databases. It is linked to NCBI (Sayers et al., 2009), and these systems conjoined will provide an intensively integrated and accessible data resource representing the highest quality and most comprehensive consensus and experimental views of host–symbiont relationships as experimental subjects.

Furthermore, SymbioGenomesDB allows users to search for specific sets of genes, which can also be useful when working with an organism(s) metabolism, comparative genomics, complementation within organisms, etc. We are open to implementing more features if requested by users.

Finally, we foresee that the continued development of high-throughput sequencing technologies and their dropping prices will result in the sequencing of numerous new organisms involved in symbiotic relationships. As more data is generated and published, SymbioGenomesDB will be updated and therefore, capable of providing information on the genomic features and the sequences of the organisms involved.

2.1.8 The idea behind an update on SymGenDB

With the discovery of the MetaDAG methodology previously mentioned (Alberich et al., 2017) and its proven usefulness, we saw a huge opportunity for its application to data we had already worked on, all the symbiotic genomes available in the SymGenDB (<http://symbiogenomesdb.uv.es/>). To this end, we collaborated with the MetaDAG developers and constructed an additional module to the three already available on the web, where we included the m-DAGs (described in the last section), for each organism available in the repository. The new module provided unique opportunities to explore the metabolism of each organism and/or evaluate the shared and joint metabolic capabilities of organisms of the same genera included in the catalog, to allow

users to construct predictive analysis of metabolic associations and complementations between symbiotic relationships.

Also, as part of the upgrade of the database, we reported a ~25% increase in the manually curated content within the database, including more than 2300 bacterial genomes associated with almost 500 hosts, improving its usefulness.

2.2. An update on the Symbiotic Genomes Database (SymGenDB): a collection of metadata, genomic, genetic, and protein sequences, orthologs, and metabolic networks of symbiotic organisms

The Symbiotic Genomes Database (now SymGenDB; <http://symbiogenomesdb.uv.es/>) is a public resource of manually curated associations between organisms involved in symbiotic relationships, maintaining a catalog of completely sequenced/finished bacterial genomes exclusively. It originally consisted of three modules where users could search for the bacteria involved in a specific symbiotic relationship, their genomes, and their genes (including their orthologs). In this update, we present an additional module that includes a representation of the metabolic network of each organism included in the database, as Directed Acyclic Graphs (MetaDAGs). This module provides unique opportunities to explore the metabolism of each individual organism and/or to evaluate the shared and joint metabolic capabilities of the organisms of the same genera included in our listing, allowing users to construct predictive analyses of metabolic associations and complementation between systems. We also report a ~25% increase in manually curated content in the database, i.e. bacterial genomes and their associations, with a final count of 2328 bacterial

genomes associated with 498 hosts. We describe new querying possibilities for all the modules, as well as new display features for the MetaDAGs module, providing a relevant range of content and utility. This update continues to improve SymGenDB and can help elucidate the mechanisms by which organisms depend on each other.

Symbiotic relationships between bacteria and eukaryotes occur naturally and ubiquitously in nature. It is a general principle in the evolution of eukaryotes and an important selective force behind evolution. The correct functioning of all ecosystems depends on these interactions (Bennett & Moran, 2015; Margulis et al., 1991; Moran, 2007; Moya et al., 2008). To our knowledge, this is the first database completely devoted to symbiotic interactions. The Symbiotic Genomes Database (SymGenDB, previously named SymbioticGenomesDB; <http://symbiogenomesdb.uv.es/>) is a public resource that provides information on symbiotic relationships throughout the Tree of Life (Reyes-Prieto et al., 2015). We use bioinformatic tools and manual curation to create a comprehensive list of associations between symbiotic organisms, and match them to their metadata, their genomic and genetic content, their association to orthologous genes, and their metabolic networks in a novel format known as Directed Acyclic Graphs (MetaDAGs) (Alberich et al., 2017). SymGenDB focuses on accessibility, so we use ids from the primary databases, NCBI and KEGG (Kanehisa et al., 2016, 2017; Kanehisa & Goto, 2000; NCBI Resource Coordinators, 2018), for taxonomy, organism names, and genes ids, so the accessibility to information is as complete as possible and enables data to be compared across species/strains, among other

advantages. In addition, we have included links to the same primary databases in the modules of SymGenDB, as well as links to each organism's scientific literature. This process of compiling a comprehensive list with community-accepted controlled ids and accession identifiers ensures that the content of SymGenDB is cohesive, controllable, and computable, and also complies with the FAIR principle (Findable, Accessible, Interoperable, and Reusable) (M. D. Wilkinson et al., 2016).

In this work, we provide the first SymGenDB update and describe our newly released module that focuses on the metabolism of each organism included in the database. We also comment on the advanced query searches we have implemented and the new tools that allow us to visualize metabolic interactions in a new manner. Researchers in the world of symbiosis and other fields (microbiome, for example), can benefit from SymGenDB to explore associations between organisms and quickly generate data and testable hypotheses about the molecular mechanisms beneath these types of relationships.

2.2.1 Database contents—an overview

SymGenDB consists of four modules that serve different purposes. The first three, organisms, genomes, and genes, have been previously described in the first publication of this database (Reyes-Prieto et al., 2015). Although a small overview is listed in the next segments for each module, we invite users to refer to the original paper for detailed functionality and to browse the database, and

watch the videos of the quick tour we offer. This update maintains the same general architecture of the previous version of the database, although the scripts required major modifications to adapt to the new data source. While the previous version relied on MGDB (Uchiyama et al., 2015), this version uses KEGG as the single source of data. The raw files from KEGG were parsed using custom scripts written in R to format the data and organize it in efficient objects that allow quick retrieving the desired information.

2.2.2 Module Organisms

First, we encounter the module 'Organisms', where users can get an overview of an organism(s) involved in a symbiotic relationship. One of the best features of this database is that users can search for either a host or a symbiont, at any taxonomic level, where the default = all taxonomic levels, as 'all ranks' (in accordance with NCBI (Benson et al., 2017; Wheeler et al., 2007)). The output of this search consists of a chart containing the relative abundance of the organisms associated with the searched query, and a list of organisms by taxonomy level resulting from the search, with links to the Taxonomy Browser of NCBI (Figure C2.6). The chart and the list are both available for download.

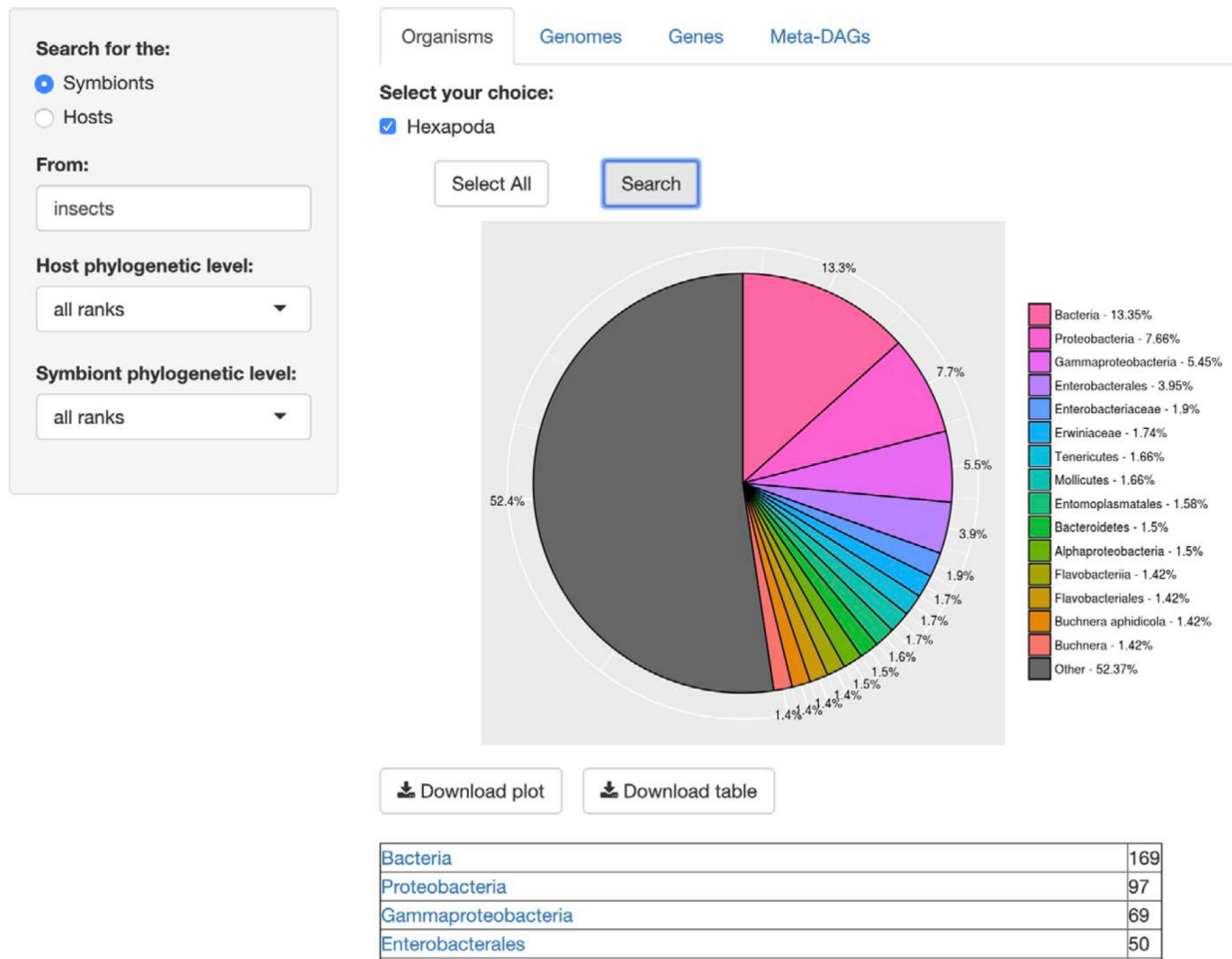


Figure C2.6 | SymGenDB's Organisms module. For this example, we searched for the symbionts of insects included in SymGenDB. It is worth noting that we used the term 'insects' and the output shows the scientific name 'Hexapoda' because the database includes a hefty list of synonyms, so users do not have to know/search only for scientific names. We searched for both the host and the symbiont's phylogenetic level as default, and the result shows an abundance chart (only the first 15 most abundant hits are displayed in color, the rest are encompassed in grey) and a list of organisms showing all the hits. Each resulting 'organism' (or family, genus, class, etc.) is a link to its NCBI taxonomy page. The resulting list shown in this figure is cut short for spacing purposes.

2.2.3 Module Genomes

In the next module, 'Genomes', searches retrieve all the metadata included in the database for each resulting symbiont: Name (the complete name of the bacteria), Host (described as the isolation source, to avoid complications regarding the host diversity), Taxonomy ID (NCBI's ID), the number of CDS with orthologs, the Strain-specific CDS, the number of replicons within this genome and a link to the organism's scientific literature (the first paper where the symbiotic relationship was described). For each replicon in the genome, we also provide information on the replicon name, its size, GC content, CDSs, genes, ncRNAs, rRNAs, tmRNAs, and tRNAs. The search can also be performed at all different taxonomic levels, with the default = all taxonomic levels as 'all ranks' (Figure C2.7). Moreover, a table with all the metadata of the genomes included in the results, a fasta file with all genomes, a gff file of all genomes, and the orthologs shared by all these genomes are available for download in this module.

Search for genomes related to:

Level:

Organisms Genomes Genes Meta-DAGs

Select your choice:

Hexapoda

Blattabacterium sp. (*Nauphoeta cinerea*)

Blattabacterium sp. (*Periplaneta americana*) str. BPLAN

Blochmannia endosymbiont of *Camponotus* (*Colobopsis*) *obliquus*

Blochmannia endosymbiont of *Polyrhachis* (*Hedomyrma*) *turneri*

Buchnera *aphidicola* (*Aphis glycines*)

Buchnera *aphidicola* BCc

Buchnera *aphidicola* (*Cinara tujaefilina*)

Buchnera *aphidicola* str. 5A (*Acyrthosiphon pisum*)

Displaying 3 of 169 available genomes:

Name:	Name:	Name:
Name: <i>Blochmannia endosymbiont of Polyrhachis turneri</i>	Name: <i>Blattabacterium</i> sp. (<i>Nauphoeta cinerea</i>)	Name: <i>Buchnera</i> <i>aphidicola</i> (<i>Cinara tujaefilina</i>)
Isolated from: Polyrhachis turneri	Isolated from: Nauphoeta cinerea	Isolated from: Cinara tujaefilina
Taxonomy ID: 1505596	Taxonomy ID: 1316444	Taxonomy ID: 261317
CDS with orthologs: 616	CDS with orthologs: 496	CDS with orthologs: 380
Strain-specific CDS: -27	Strain-specific CDS: 00	Strain-specific CDS: -21
Pubmed: 2015	Pubmed: 2013	Pubmed: 2011
Number of replicons: 1	Number of replicons: 2	Number of replicons: 1
Replicon name: chromosome	Replicon name: chromosome	Replicon name: chromosome
Replicon size: 749,321	Replicon size: 622,952	Replicon size: 444,925
Replicon GC content: 29%	Replicon GC content: 26%	Replicon GC content: 23%
CDS: 589	CDS: 581	CDS: 359
gene: 1	rRNA: 3	gene: 10
ncRNA: 2	tmRNA: 1	rRNA: 3
rRNA: 3	tRNA: 32	tmRNA: 1
tmRNA: 1	Replicon name: unnamed	tRNA: 31
tRNA: 38	Replicon size: 3,674	
	Replicon GC content: 21%	
	CDS: 5	
	rRNA: 0	
	tmRNA: 0	
	tRNA: 0	

Figure C2.7 | SymGenDB's Genomes module. Continuing with the example on symbionts of insects, we searched for those genomes and selected 3 of the 169 available in SymGenDB. The resulting table (viewed horizontally for spacing purposes, although it is presented vertically in the database), consists of all the metadata of the genomes of our choice, with a link to the host's taxonomy id from NCBI. It is important to show that one of the downloads available in this module is the orthology table of the chosen genomes (shown in red) which is very helpful for evolutionary research. Furthermore, the literature where this genome was first described is also made available to users (shown in orange).

2.2.4 Module Genes

In the third module, 'Genes', users are able to search for specific genes in the genomes of symbiotic organisms. The search involves the input of two queries, first, the name of the gene(s) to search for, and then, the name of the symbiotic organisms to search for, or the name of the host of the symbiotic organism to search for. The output is a presence/absence list of genes with their KEGG id and description, in every genome available as a result. Each resulting gene is also linked to its KEGG gene web page (Figure C2.8). The 'Gene ID' and the 'Gene Description' features are both linked to the KEGG ortholog web page of each result. Both the gene table and the amino acid sequences of the results are downloadable.

Search for the genes:

In the genomes related to:

Level:

Organisms Genomes **Genes** Meta-DAGs

Select your genes:

- tRNA-Trp tRNA Trp
- trpA tryptophan synthase alpha chain
- trpB tryptophan synthase beta chain
- trpB tryptophan synthase beta chain
- trpCF indole-3-glycerol phosphate synthase / phosphoribosylanthranilate
- trpC indole-3-glycerol phosphate synthase
- trpD anthranilate phosphoribosyltransferase

Select All

Select your choice:

- Hexapoda

Select All

Available genomes:

- Acetobacter pomorum
- Acetobacter tropicalis
- Acinetobacter larvae
- Bacillus bombysepticus str. Wang
- Bacillus kochii
- Bacillus sp. SDL1
- Bacillus sp. YP1

Select All Search

Displaying 169 of 169 available genomes:

Gene ID	Gene Description	ala	apom
<i>tRNA-Trp</i>	tRNA Trp	ala:BFG52_11675 ala:BFG52_15060	
<i>trpA</i>	tryptophan synthase alpha chain [EC:4.2.1.20]	ala:BFG52_07120 ala:BFG52_13290	apom:CPF11_05765
<i>trpB</i>	tryptophan synthase beta chain [EC:4.2.1.20]	ala:BFG52_03390 ala:BFG52_07115	apom:CPF11_05770
<i>trpB</i>	tryptophan synthase beta chain [EC:4.2.1.20]		
<i>trpCF</i>	indole-3-glycerol phosphate synthase / phosphoribosylanthranilate isomerase [EC:4.1.1.48 5.3.1.24]		
<i>trpC</i>	indole-3-glycerol phosphate synthase [EC:4.1.1.48]	ala:BFG52_10935	apom:CPF11_01335
<i>trpD</i>	anthranilate phosphoribosyltransferase [EC:2.4.2.18]	ala:BFG52_10940	apom:CPF11_01330
<i>trpE</i>	anthranilate synthase component I [EC:4.1.3.27]	ala:BFG52_15090	apom:CPF11_01320
<i>trpEG</i>	anthranilate synthase [EC:4.1.3.27]		
<i>trpF</i>	phosphoribosylanthranilate isomerase [EC:5.3.1.24]	ala:BFG52_03385	apom:CPF11_09800
<i>trpG</i>	anthranilate synthase component II [EC:4.1.3.27]	ala:BFG52_10945	apom:CPF11_01325

Figure C2.8 | SymGen's gene module. For this example, we searched for all the genes containing the letters 'trp' in the genomes of symbionts of insects. We selected all of the 169

available genomes. The resulting table is a list in a presence/absence format, where all the present genes are shown with their KEGG id and a link to their KEGG gene web page. The 'Gene ID' and the 'Gene Description' features are both linked to their KEGG orthology web page.

A couple of remarks to keep in mind while searching in all the modules at SymGenDB are as follows: the first one is that all synonyms of NCBI are linked to the organisms' names to facilitate searches. For example, if a user wants to search for the symbionts associated with humans, he/she can write the words 'man' or 'human' or common misspellings such as 'Homo sapience' and still get the output for *Homo sapiens*, its correct taxonomic name. Also, searches of more than one organism/genome/gene can be made at the same time. The only requirement is that queries must be separated by a comma (',').

2.2.5 Increased data content

One of the most important aspects of databases in the genomic era is their maintenance and updates. SymGenDB became publicly available in 2015, and two updates to the data content of the database have been made since then. As of July 2019, SymGenDB contains 2328 symbiotic genomes, associated with 498 hosts (Table C2.1), presenting a roughly 25% increase from our previous version. The interactions between symbionts and hosts are manually curated from a cross reference of the lists of organisms included in KEGG and the JGI's GOLD Genomes Online Database (Mukherjee et al., 2017). These associations are mainly revised in peer-reviewed scientific papers from PubMed. Next, all the metadata

and genomic data associated with a symbiont in the list are retrieved from the KEGG archives.

Table C2.1 | Update on the data content of SymGenDB.

Data	Previous version	Update July 2019
Symbiotic genomes	1955	2328
Hosts	384	498
Orthologous genes	3 121 262	3 808 086
Orthologous clusters	7567	8478

2.2.6 New module—MetaDAGs

We have added a new module to SymGenDB consisting of the metabolic networks of all organisms modeled as directed acyclic graphs (MetaDAGs). The MetaDAG of every organism is obtained as a suitable reduction of the reaction graph created with the metabolic data of each genome we retrieved from KEGG. In the reaction graph model, the nodes are the reactions and there is an arc (a directed edge) between two reactions if some metabolite in the product of the source-arc reaction is in the substrate of the target-arc reaction. Next, the strongly connected components in the reaction graph are collapsed into a single node, called a metabolic building block. The result is a graph with no cycles, hence a directed acyclic graph called a MetaDAG. The MetaDAG keeps the connectivity of the metabolic network but reduces considerably the number of nodes, which facilitates its visualization. Notice that the nodes in the MetaDAG are the strongly connected components in the reaction graph, which consist of one or

many reactions. In order to easily visualize the size of every metabolic building block of the MetaDAG, we scale the size of the corresponding node depending on the number of reactions included in the node, and use two colors to distinguish these: green when the metabolic building block has only one reaction, and yellow when there are two or more reactions. To contextualize the metabolic building blocks, we present an interactive metaDAG where users are capable to scroll over each node to visualize the reactions it contains, as well as highlighting their position in the global metabolic pathway's map from KEGG. Furthermore, we distinguish cut nodes (those that, if removed, disconnect the network), by drawing an octagon instead of a circle for the node. The methodology for these calculations is described in detail in (Alberich et al., 2017).

For this last module, 'MetaDAGs', users are able to search for the genome-scale metabolic network as MetaDAGs, related to any symbiont (or first, search for the host and then select the symbiotic genomes associated with that host). As a running example, we present the search for the symbionts included in SymGenDB of the genus *Buchnera*, known symbionts of aphids (Figure C2.9). The result is a list of the bacterial strains included in the symbiotic relationship the user searched for, associated with a link. When the link of interest is clicked on, a new visualizer shows its MetaDAG in a dynamic manner, where the user can zoom in and out of any part of the graph, and click on any node to get a couple of pop-up windows. The first window has information on the reaction(s) included in that node. In this pop-up window, users can also click on the image of each reaction to see it bigger and better and, by clicking on the link 'more info',

get to KEGG's reaction web page for that reaction. We have also included information on other organisms that present the same reaction (by clicking on the link 'graphs' users get a full display), which can be very useful for orthology and evolutionary studies, among others. The second window shows the global metabolic pathways map from the KEGG's web page with the reactions of the metabolic building blocks highlighted. This implementation contextualizes the metabolic building blocks to the well-known global metabolic description map.

Search for genomes related to:

Level:

[Organisms](#) [Genomes](#) [Genes](#) **[Meta-DAGs](#)**

Select your choice:

Buchnera

Name:	metaDAG
Buchnera (core)	Buchnera_core
Buchnera (pan)	Buchnera_pan
Buchnera aphidicola (Aphis glycines)	baph
Buchnera aphidicola (Cinara tujafilina)	baj
Buchnera aphidicola BCc	bcc
Buchnera aphidicola str. 5A (Acyrtosiphon pisum)	bap
Buchnera aphidicola str. Ak (Acyrtosiphon kondoi)	bak
Buchnera aphidicola str. APS (Acyrtosiphon pisum)	buc
Buchnera aphidicola str. Bp (Baizongia pistaciae)	bab
Buchnera aphidicola str. F009 (Myzus persicae)	bapf
Buchnera aphidicola str. G002 (Myzus persicae)	bapg
Buchnera aphidicola str. JF98 (Acyrtosiphon pisum)	baw
Buchnera aphidicola str. JF99 (Acyrtosiphon pisum)	bajc
Buchnera aphidicola str. LL01 (Acyrtosiphon pisum)	bua
Buchnera aphidicola str. Sg (Schizaphis graminum)	bas
Buchnera aphidicola str. TLW03 (Acyrtosiphon pisum)	bup
Buchnera aphidicola str. Tuc7 (Acyrtosiphon pisum)	bau
Buchnera aphidicola str. Ua (Uroleucon ambrosiae)	buh
Buchnera aphidicola str. USDA (Myzus persicae)	bapu
Buchnera aphidicola str. W106 (Myzus persicae)	bapw

Figure C2.9 | SymGenDB’s new MetaDAGs module. In this example, we search for the symbionts of the genus ‘*Buchnera*’. The output is a list of organisms as bacterial strains, as well as the joint (pan) or intersecting (core) metabolism of the strains resulting in the search, included in the taxonomic level ‘genus’ (in bold). It is important to denote that in the case of the ‘pan’ and ‘core’ interacting metabolism, not only the genomes of the bacterial strains resulting from the search are presented. The complete set of strains of the same genus available in SymGenDB constitutes these MetaDAGs.

Furthermore, in the first result after searching for the MetaDAGs of interest, we also offer the core MetaDAG and the pan MetaDAG of the bacterial

strains of the same genus. The core MetaDAG is obtained by considering the common metabolic building blocks present in all the MetaDAGs of the bacterial strains of the same genus. That is, the nodes (*i.e.* metabolic building blocks) in the core MetaDAG are the intersection of the nodes of the MetaDAGs of the bacterial strains of the same genus with the corresponding directed acyclic graph topology. On the other hand, the pan MetaDAG is obtained by considering the union of the nodes (metabolic building blocks) in all the MetaDAGs of the bacterial strains of the same genus with the corresponding directed acyclic graph topology. It is important to note that in the case of the 'pan' and 'core' interacting metabolism, in certain cases where all the organisms of the same genus are not part of the symbiotic relationship searched for, not only the genomes of the bacterial strains resulting of the search are presented. The complete set of strains of the same genus available in SymGenDB constitutes these pan and core MetaDAGs.

To continue with our example, Figure C2.10 shows an example of the graphical output of our MetaDAGs module. From the list of strains resulting from the search, we clicked on *Buchnera aphidicola* from *Cinara tujaefilina*, and a preview of a graphical display of the MetaDAG is shown. This display can be viewed in another window (most MetaDAGs are big and space is needed to fully appreciate the graph, as well as to get into details), download the graph as a PDF, and/or dismiss the preview.

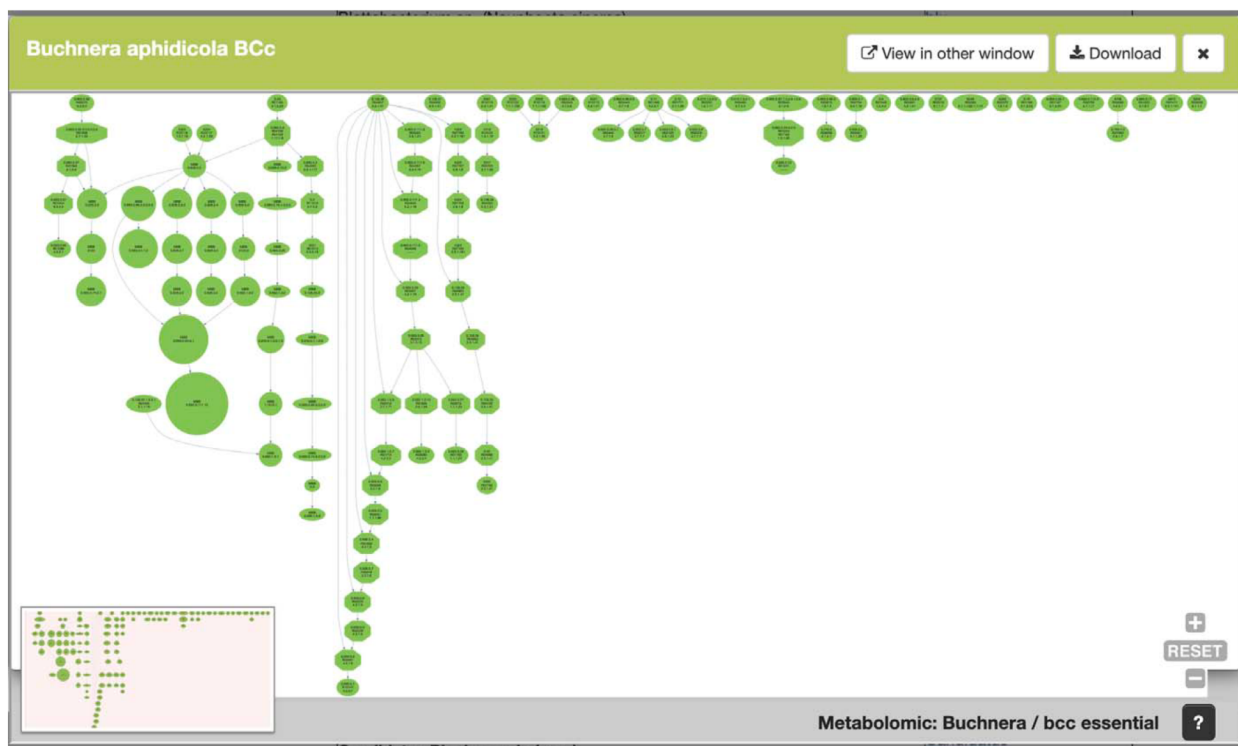


Figure C2.10 | SymGenDB's new MetaDAGs module's output. In this example, we search for the symbionts of the genus '*Buchnera*'. The output is a preview of a graphical display of the MetaDAG you get by choosing *Buchnera aphidicola* from *Cinara tujaefilina*. This dynamic display can be viewed in another window, downloaded as a PDF and all the information of the reaction(s) included in each node(s) is available by clicking on the node(s) of interest.

Lastly, Figure C2.11 is an example of the visualization of the pop-up windows that emerge from one node of the metaDAG of *B. aphidicola* from *C. tujaefilina*.

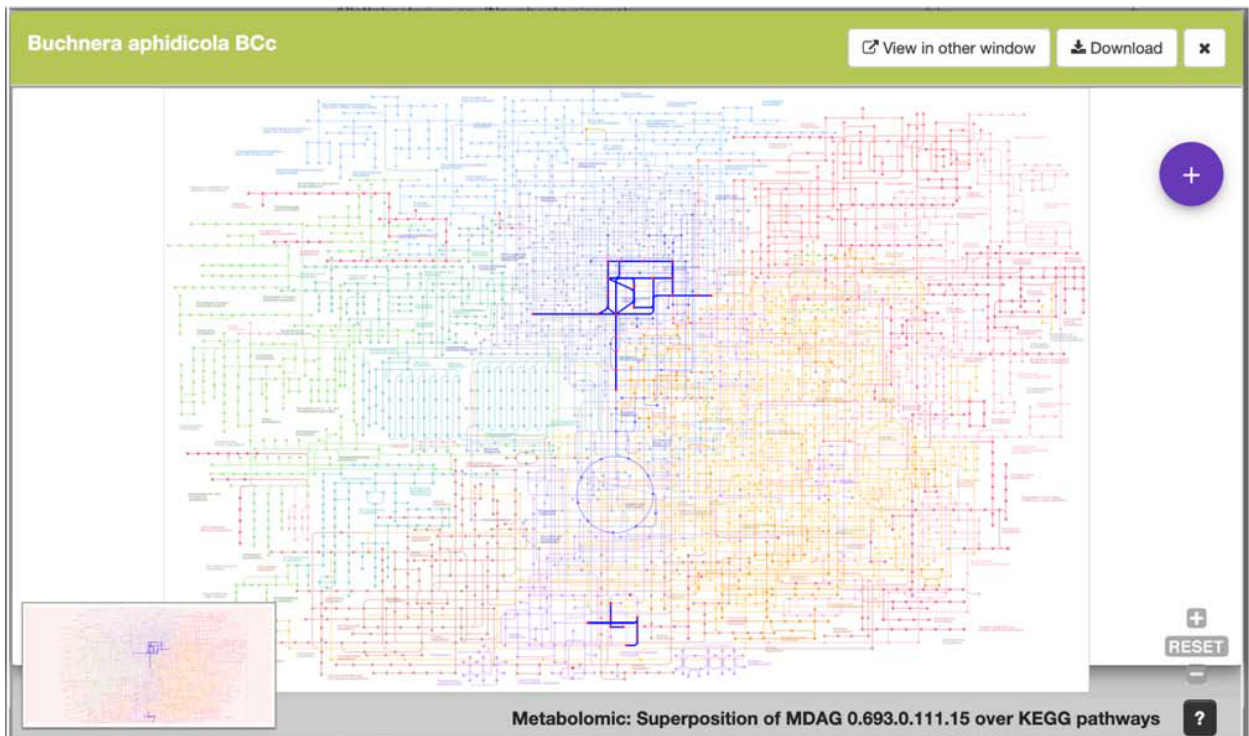
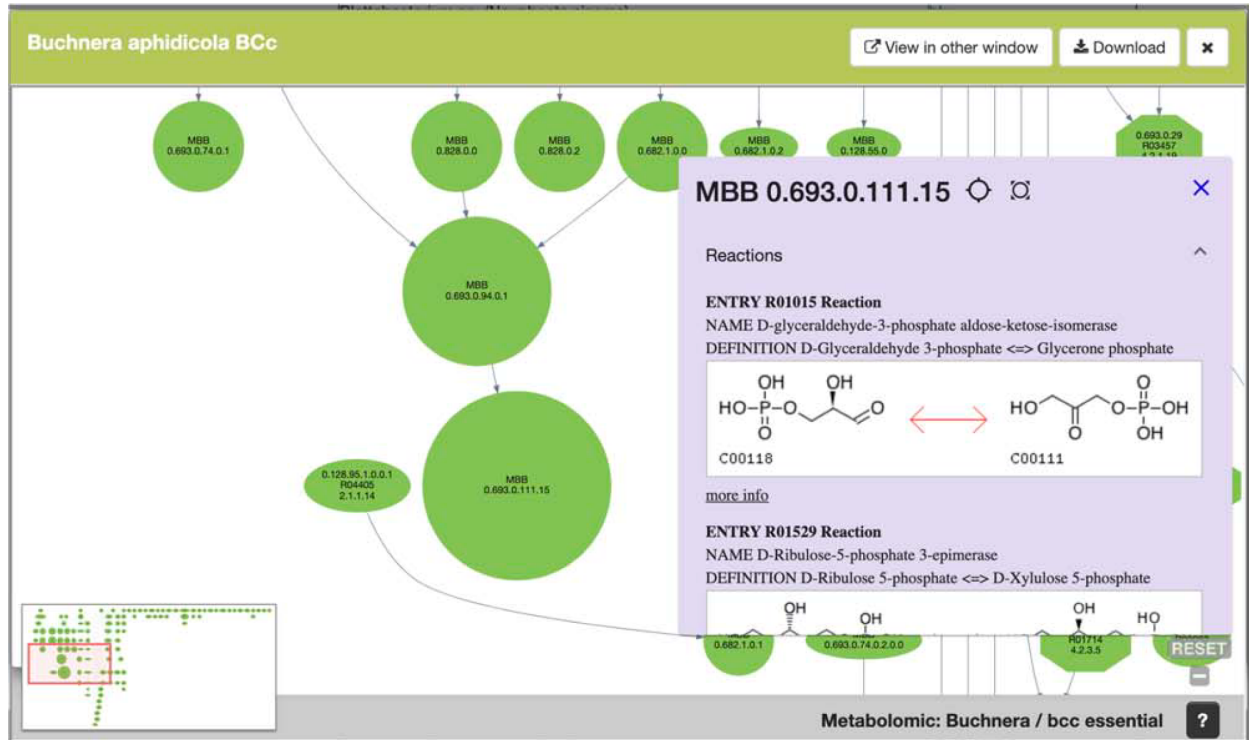


Figure C2.11 | SymGenDB's new MetaDAGs module's output where users can not only see the reactions included in their graph of interest in detail but also contextualize the reaction within the KEGG metabolic map. We have included a feature to highlight the reaction(s) of interest for this purpose.

2.2.7 Availability of web interface and services

The web interface of SymGenDB has been modified from its previous version. We have included different tabs to make the interface and services of the database more understandable:

- Home, where an overview of the purpose of the database, as well as its metrics and a phylogenetic tree including all the bacterial strains included in our catalog is shown (<http://symbiogenomesdb.uv.es/>).
- Database, where in turn, all the different modules of the database are shown in tabs. This is where all the searches and results are presented (<http://symbiogenomesdb.uv.es/database.html>).
- Functions, in the first paragraph of this tab users can find the basic description of each module of the database and below that, a detailed description of the functions and an overview of how to search in each module and the results to expect (<http://symbiogenomesdb.uv.es/functions.html>).
- Quick tour, another subset of tabs where in each tab we present a module, a video on how to search for the data in that module, and a step-by-step guide on how to search in each module, with pictures for easier interpretation (<http://symbiogenomesdb.uv.es/quick-tour.html>).

- Citation, Acknowledgements, and contact include the citation for the first article published of the database (Reyes-Prieto et al., 2015), as well as the funding received for the creation of this project and a form for users to get in touch with the authors where any problem, doubt or suggestion is received and greatly appreciated (<http://symbiogenomesdb.uv.es/acknowledgements.html>).

2.2.8 Future directions

It is the author's intent to continue to develop tools and data for the completion of the database, with new modules and organisms to include as well as updating the database.

2.2.9 Statistics and demographic overview of SymGenDB

Even though SymGenDB has been publicly available since April 2015, we have the web page metric analysis of Google Analytics on the database from June 2018 to date (<https://analytics.google.com/analytics/web/#/>). Figure C2.12 is an overview of the total number of database users since June 2018. There has been a steady flow of users, more than 2,500, which is remarkable for such a specific composite database. There was a peak in users by September 2018, and then a constant flux in the rest of the months. This data should be taken with caution since it is an estimate of metrics that could not be 100% accurate, with Google Analytics constantly changing and testing different algorithms.

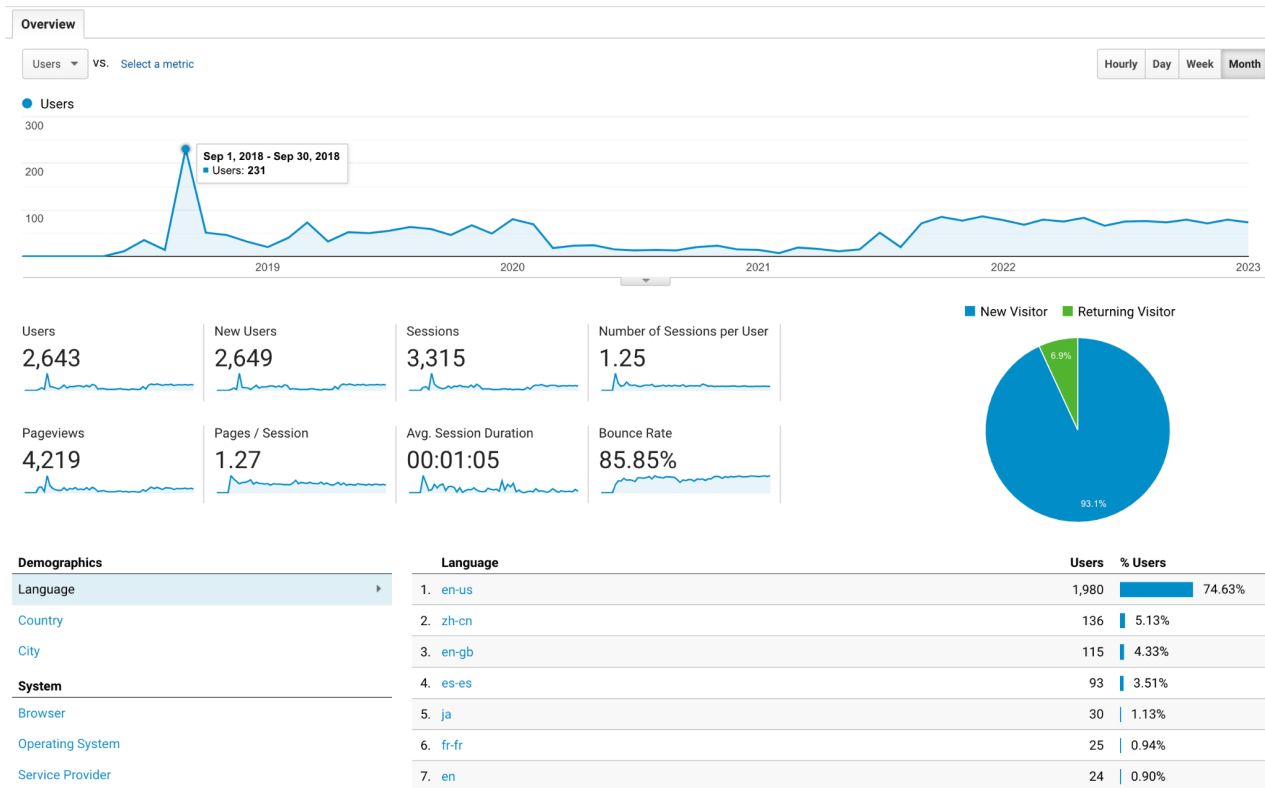


Figure C2.12 | Overview of the users of SymGenDB from June 2018 to February 2023.

The demographics of the web page, which were included in the analysis in January 2022, are shown in Figure C2.13. SymGenDB has been used in several countries around the world, with 485 users from Germany, the country with the most users. Expectedly, the language used by most users is English, using Chrome as the device model and Linux as an operating system. Finally, direct searches as opposed to organic searches are the most common, which means that users are intentionally searching for this knowledge.

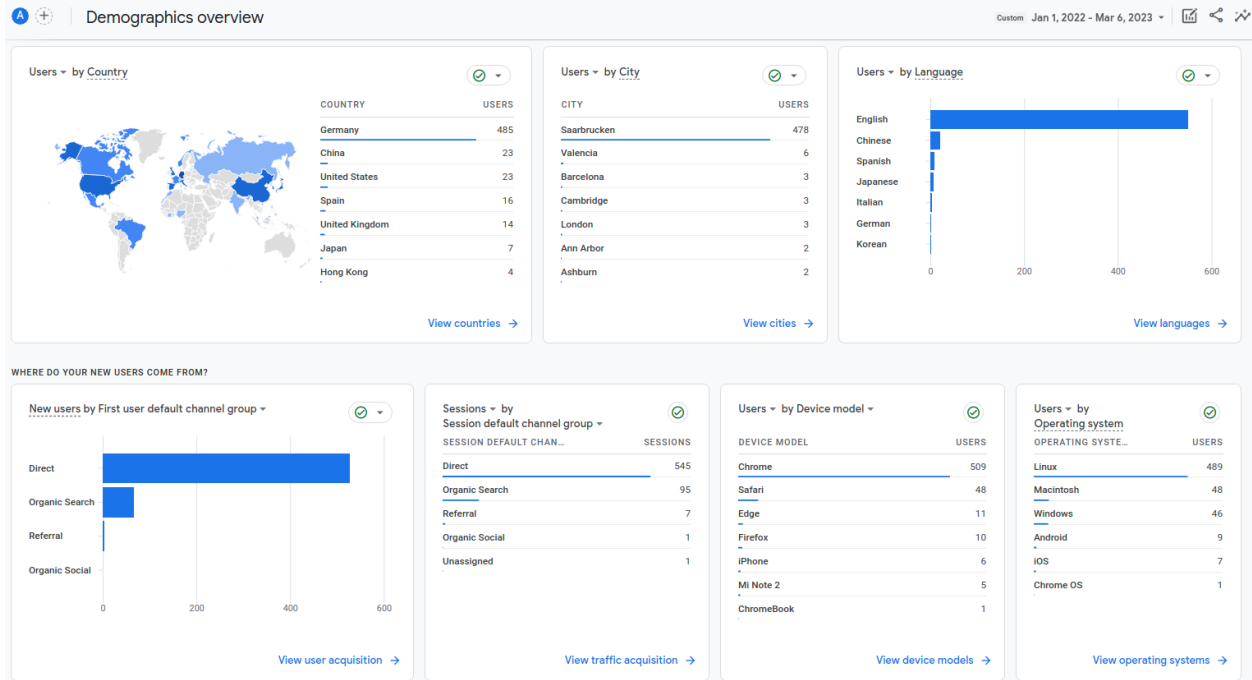


Figure C2.13. Demographic overview of SymGenDB users from Jan 2022 to March 2023.

Chapter 3. The minimal metabolism

This chapter reproduces the following published paper entirety:

Reyes-Prieto, Mariana, Rosario Gil, Mercè Llabrés, Pere Palmer-Rodríguez, and Andrés Moya. "The Metabolic Building Blocks of a Minimal Cell." *Biology* 10, no. 1 (2021): 5.

3. The Metabolic Building Blocks of a minimal cell

One of the most ambitious aspirations of modern biology is to synthesize artificial living cells. Manufacturing a cell opens endless research possibilities, both in basic and advanced sciences, and it would be critical and a turning point in fields from medicine to evolutionary biology. To reduce the levels of difficulty on this task, most efforts are focused on the synthesis of minimal cells. On the one hand, they will help by increasing our understanding of living systems; on the other hand, they can be used as capsules for the introduction of genetic material to customize cells for applied purposes (Moya et al., 2009). Several complementary paths have been followed in search of the proper techniques and methods to design this fabricated cell. The most commonly used are the bottom–up and top–down approaches (Luisi, 2002; Luisi et al., 2006; Xavier et al., 2014).

The bottom–up approach consists of the assembly, piece by piece, of each non-living biological component (i.e., a self-replicating nucleic acid, a metabolic machinery, and an encapsulating structure; (Stano & Luisi, 2011)) in order to get a system that could be considered alive. The resulting products of this approach are called “protocells” (Bedau et al., 2009; Mantri & Tanuj Sapra, 2013). No comparable system has been successfully constructed yet, but there have been developments on this front, with the design of more refined cell-like compartments (Stano, 2018).

The top–down approach consists of deconstructing living cells (Glass et al., 2017; Xavier et al., 2014). Taking modern cells with reduced genomes as a starting

point, it aims at further simplifying them by removing dispensable genetic material. Experimental (genome-wide analyses by massive transposon mutagenesis, antisense RNA, and systematic gene knockout) and computational approaches (including comparative genomics, comparative proteomics, and in silico cell modeling) have been used to characterize a set of essential and sufficient genes to compose a living cell, that is, the core of a minimal bacterial genome (Gil, 2015). Experimentally, genes are considered to be essential based on indirect evidence from systematic and genome-wide inactivation or the inhibition of each individual gene present in a genome (compiled in <http://www.essentialgene.org/> (Luo et al., 2014)). Comparative genomics has also been broadly used, assuming that genes that are common between distant organisms are prone to be essential (Acevedo-Rocha et al., 2013). In addition, naturally reduced genomes from bacteria with a host-associated lifestyle have been used for comparisons regarding gene content because they must be approaching a minimal genome (Gil et al., 2003; Mushegian & Koonin, 1996). The merging of these studies demonstrated the relevance of considering that essential functions can be performed by alternative and unrelated (non-orthologous) gene products. Comparative studies only retrieve genes involved in functions for which there is no alternative in nature (e.g., the complex translational machinery), while a minimal genome must also include all genes essential to maintain metabolic homeostasis (Gil et al., 2004).

There is a third approach for the construction of a minimal genome that searches for the biochemical and modular description of well-defined pathways

needed to perform all essential functions (Jewett & Forster, 2010). Despite some major challenges needing to be addressed, this approach allows a function-by-function debugging to reach self-replication, and it suggests a good starting point for the ultimate synthesizing of a minimal genome able to sustain an artificial minimal cell. The potentiality of chemically synthesizing genomic segments or complete genomes and confining them into pre-existing cells has revolutionized the study of minimal cells (Hutchison et al., 2016). The design of a truly minimal genome and its metabolic network can also benefit from computational whole-genome sequence rewriting and a design-build-test in silico approach, preceding the chemical synthesis of a customized genome (Venetz et al., 2019).

A cohesive metabolic network proposal can lead the path to the synthesis of minimal cells. A minimal cell would depend on a minimal set of anabolic pathways to convert and assemble its biomolecule building blocks with the use of the energy and nutrients available in the environment, to reach metabolic homeostasis, and to achieve cellular growth and reproduction. Nevertheless, there is scientific consensus regarding the existence of a variety of minimal metabolic schemes that are ecologically dependent and able to sustain a universal genetic machinery (Gabaldon et al., 2007). The simplest cell should be chemoorganoheterotrophic (i.e., an organism using organic compounds as carbon and energy sources), living in a nutrient-rich medium, in which the major metabolites (glucose, fatty acids, nitrogenous bases, amino acids, and vitamins) must be available without limitation since this cell would not be able to

synthesize them. Nevertheless, considering the adaptability of bacterial heterotrophic metabolisms, different metabolic schemes can be envisaged. The metabolic chart proposed by (Gil et al., 2004) using a top-down approach, by performing a comprehensive analysis of all previous computational and experimental attempts to define a minimal genome, was based on the metabolic functions that were preserved in highly reduced genomes completely sequenced at that time, from endosymbiotic mutualistic or parasitic bacteria. The proposed core of the minimal genome encoded the costless pathways that would allow the cell to perform the selected metabolic functions. In order to maintain a coherent metabolic functionality, some pathways that were not present in some of the reduced genomes used in the aforementioned study were also incorporated, because their lack reflected a high dependence on their hosts. Likewise, the group of Craig Venter also explored this area and presented their list of essential genes for a minimal bacterium in 2006 (Glass et al., 2006). Both sets of genes and the coherence of this metabolic network were further explored by (Gabaldon et al., 2007).

Metabolic networks determine the physiology and biochemistry of a cell. They are made of three components: the metabolic pathways, the chemical reactions involved in the metabolism, and the regulatory interactions of these reactions. Metabolic networks tend to be highly complex, even for simple organisms. For example, if we consider the metabolism of porphyrin and chlorophyll which is present in some animals, plants, fungi, bacteria, and archaea, we get a metabolic pathway map of 135 nodes and 181 edges in the

reference pathway in the KEGG database (pathway: map00860). A pathway map with so many components is very difficult to visualize, especially when we are interested in the pathway topology. To this extent, it is highly advantageous to suitably reduce the number of nodes in order to visualize the network more precisely. (Alberich et al., 2017) designed a methodology called MetaDAG, which consists of the contraction into a single node of those reactions that are strongly connected in the genome-wide reaction graphs. In this way, the resulting graph is a Directed Acyclic Graph (DAG), called a metabolic DAG (m-DAG), that preserves the network topology (i.e., the original relations between reactions) while it allows easy human exploration and visualization. One advantage of directed acyclic graphs is that they do not have cycles repeatedly producing and consuming the same metabolite. This methodology also creates reaction graphs and m-DAGs from multiple genomes, which can be used to calculate the core- and pan-metabolisms of a group of bacteria of interest as well as compare genomes by their m-DAGs in a novel manner. The MetaDAG methodology can also be of importance for large in silico analyses. By compressing metabolic networks and making them “simpler”, algorithms and computer analyses could also be less time-consuming. Just as important, fewer computational resources would be needed, making it easier for researchers to work with a large number of genome-wide m-DAGs, bacterial consortia m-DAGs, multiple symbiosis analyses, or even environmental metabolomics.

For the current work, we constructed the minimal metabolic network from the theoretical minimal gene set machinery revised in (Gabaldon et al., 2007), and

compared it to the smallest genome of a live organism known to date (Bennett & Moran, 2013), and to the genome of a semisynthetic bacteria produced by Craig Venter's group in 2016 (Hutchison et al., 2016). Despite the great efforts being done to homogenize gene and enzyme names in databases, due to how they have been discovered and described throughout history, some of their names are still associated with taxonomically related organisms. For this reason, to avoid any remaining biases toward any group of organisms and any need for synonym lists, we propose a minimal metabolic network defined by reactions and compounds instead of genes. Moreover, another of the advantages of our methodology is that it is essentially universal, since it uses homogenous identifiers and descriptors, so researchers can easily associate the involved reactions and compounds to genes of bacterial genomes with different phylogenetic backgrounds, even to synthetic genomes as proven in this study. Finally, it can also be applied to bacterial consortia in order to detect the metabolic interactions between partners and communities.

3.1. Inference of Minimal Metabolic Networks

The metabolic networks for this study were inferred from the reviewed version of the theoretical minimal genome described by (Gabaldon et al., 2007), the genome of "*Ca. Nasuia deltocephalinicola*" str. NAS-ALF (Bennett & Moran, 2013) (which is also publicly available in the new version of the SymGenDB (Reyes-Prieto, Vargas-Chávez, et al., 2020)), and the genome of JCVI-syn3.0, which is an artificial viable cell created by Hutchison and coworkers (Hutchison

et al., 2016). We first searched for all protein-coding genes in each genome for which an enzymatic activity has been assigned and then searched for the corresponding reactions in KEGG.

3.2 Reconstruction of the Directed Acyclic Graph of metabolic networks

Using the above-obtained information, which is a set of reactions for each metabolic network, we generated the corresponding reaction graph that models the relationship between reactions in terms of shared metabolites. A reaction graph, denoted by $RG = (R, ER)$, is a directed graph with a set of nodes R that are reactions and whose edges are defined as follows: there is an edge pointing from reaction R_i to reaction R_j if, and only if, a metabolite produced by reaction R_i is a substrate in reaction R_j . The fact that it is a directed graph establishes a natural production/consumption order between two reactions—that is, what is produced by R_i is then consumed by R_j . Before generating the directed graph, we manually curated it to remove redundancies (enzymes encoded by orthologous genes).

In order to analyze the reaction graph in a visually friendly manner, we used the MetaDAG methodology (Alberich et al., 2017). In a reaction graph, two reactions R_i, R_j are said to be biconnected if there is a path in each direction between them. A strongly connected component of a reaction graph is a subgraph such that every pair of reactions in it are biconnected. These strongly connected components are contracted in a single node. The reactions that are not biconnected to any other reaction become a node by themselves. Each node is called a Metabolic Building Block (MBB for short), and the MetaDAG software

automatically assigns an ID to each MBB. When each MBB is contracted to a single vertex, the resulting quotient graph is a metabolic Directed Acyclic Graph (m-DAG for short). Thus, the m-DAG is defined as follows: its nodes are the MBBs obtained from the reaction graph, and there is an edge between two MBBs, MBB1 and MBB2, if there is an edge in the reaction graph from a reaction in MBB1 to a reaction in MBB2. We denote by G_m the m-DAG, thus $G_m = (N, E)$ where N is the set of MBBs and E is the edges between them such that

$$(MBB1, MBB2) \in E, \exists Ri \in MBB1 \wedge \exists Rj \in MBB2 \mid (Ri, Rj) \in ER$$

MBBs contracting only one reaction and whose removal disconnects the reaction graph are considered essential reactions because they are crucial to maintaining the network's connectivity.

3.3 Theoretical Minimal Metabolic Network

The first step toward the creation of the minimal metabolic network was to extrapolate the list of genes and enzymes belonging to the set presented by Gabaldón and coworkers (2007) (Gabaldon et al., 2007) (Figure C3.1 and Supplementary Table C3S1) to obtain KEGG reaction identifiers (IDs). We used the complete reaction, compound, and enzyme database from KEGG and created the reaction graph by joining the reactions where metabolites were shared. The idea behind using the complete KEGG catalog is to avoid biases toward a specific phylogenetic group of bacteria.

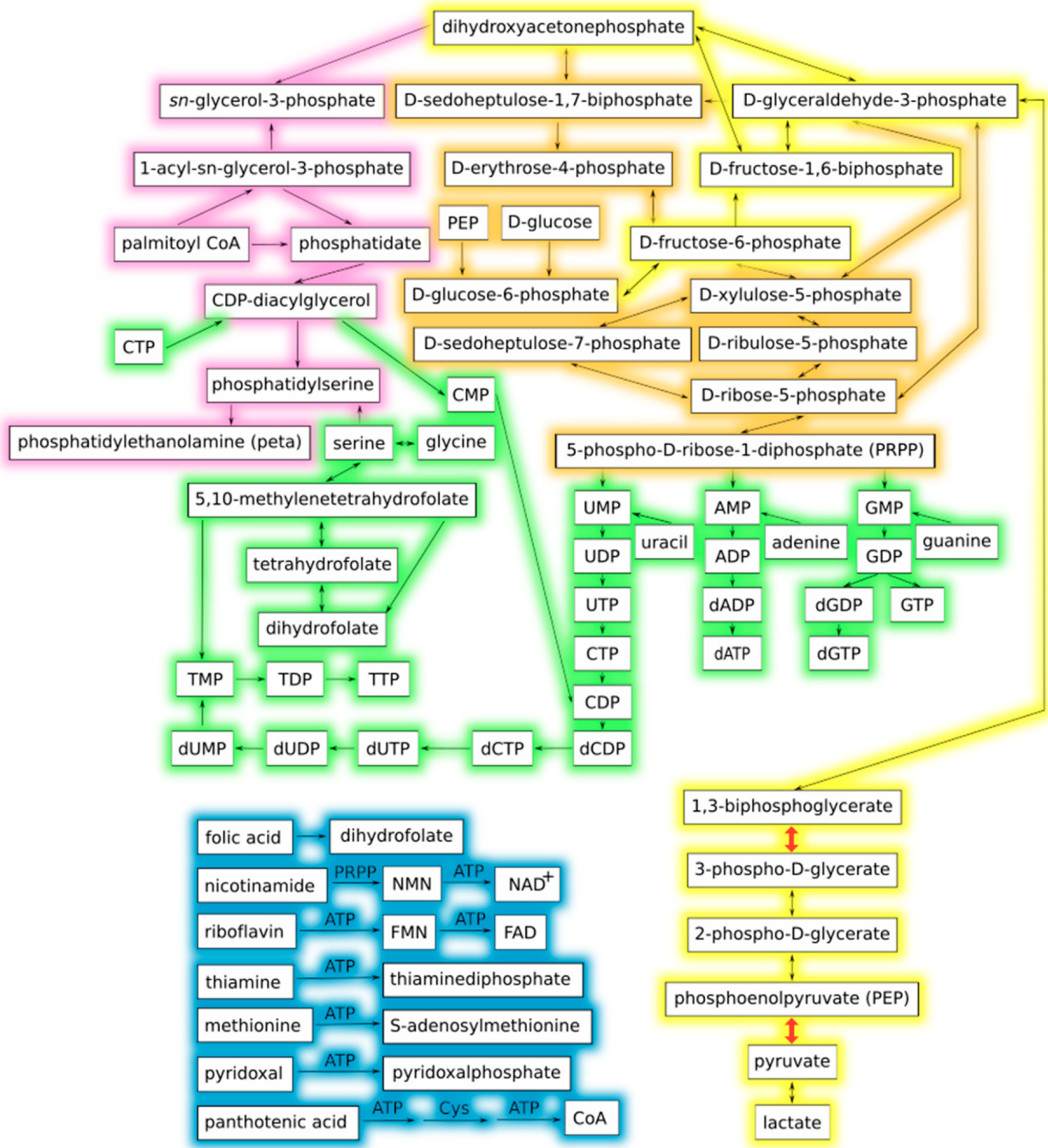


Figure C3.1 | Interaction graph of the proposed theoretical minimal metabolic network adapted from (Gabaldon et al., 2007). Line colors denote metabolic categories: yellow, glycolysis; orange, pentose phosphate pathway; pink, phospholipid metabolism; green,

nucleotide metabolism; blue, coenzyme metabolism. The two glycolytic steps in which ATP is produced by substrate-level phosphorylation are depicted with thicker red arrows and correspond to reactions R01512 and R00200 in Table C3.1. The reaction graph of this same network is presented in Figure C3.2 for comparison.

This methodology gave us a resulting reaction graph with some redundancies (i.e., different enzymes encoded by orthologous genes participating in the same metabolic pathways), so we manually curated this graph to include only one copy of each reaction and their corresponding metabolites needed for a functional cell. The reaction graph obtained is composed of 98 reactions and 80 metabolites (Figure C3.2). The fact that our model replicates almost entirely the figure of (Gabaldon et al., 2007) (Figure C3.1), validates our methodology. Table C3.1 presents the complete list of reactions, substrate, and product compounds as well as their KEGG identifiers used to reconstruct the minimal metabolic network.

Table C3.1 | Reactions, enzymes, and compounds of the minimal metabolic network presented in Figure C3.2. Reversible reactions are denoted by the superscript r. MBB IDs are the identification numbers of the metabolic building blocks to which each reaction is contracted, according to the MetaDAG analysis (Figure C3.3).

Substrate KEGG ID	Reaction ID	Enzyme Name (E.C. Number)	Definition	Product KEGG ID	MBB ID
C00020	R00127 ^r	adenylate kinase (2.7.4.3)	ATP + AMP ↔ 2 ADP	C00008	0.15
C00882	R00130	dephospho-CoA kinase (2.7.1.24)	ATP + Dephospho-CoA → ADP + CoA	C00010	0.2
C00455	R00137 ^r	nicotinamide-nucleotide adenyltransferase (2.7.7.1)	Diphosphate + NAD+ ↔ ATP + Nicotinamide D-ribonucleotide	C00003	0.80.1.0
C00015	R00156 ^r	nucleoside-diphosphate kinase (2.7.4.6)	ATP + UDP ↔ ADP + UTP	C00075	0.77.4.0
C00105	R00158 ^r	UMP/CMP kinase (2.7.4.14)	ATP + UMP ↔ ADP + UDP	C00015	0.77.4.0
C00061	R00161	FAD synthase (2.7.7.2)	ATP + FMN → Diphosphate + FAD	C00016	0.10
C00018	R00173	pyridoxal phosphatase (3.1.3.74)	Pyridoxal phosphate + H ₂ O → Pyridoxal + Orthophosphate	C00250	0.11
C00073	R00177	methionine adenosyltransferase (2.5.1.6)	ATP + L-Methionine + H ₂ O → Orthophosphate + Diphosphate + S-Adenosyl-L-methionine	C00019	0.12
C00020 + C0013	R00190 ^r	adenine phosphoribosyltransferase (2.4.2.7)	AMP + Diphosphate ↔ Adenine + 5-Phospho-alpha-D-ribose 1-diphosphate	C00147 + C00119	0.78.1.0
C00074 + C00008	R00200	pyruvate kinase (2.7.1.40)	ADP + Phosphoenolpyruvate → ATP + Pyruvate	C00022	0.9
C00144	R00332 ^r	guanylate kinase (2.7.4.8)	ATP + GMP ↔ ADP + GDP	C00035	0.77.4.0
C00044	R00430 ^r	pyruvate kinase (2.7.1.40)	GTP + Pyruvate ↔ GDP + Phosphoenolpyruvate	C00035	0.77.4.0
C00055	R00512 ^r	(d)CMP kinase (2.7.4.25)	ATP + CMP ↔ ADP + CDP	C00112	0.77.4.2
C00255	R00549	riboflavin kinase (2.7.1.26)	ATP + Riboflavin → ADP + FMN	C00061	0.13
C00112	R00570 ^r	nucleoside diphosphate kinase (2.7.4.6)	ATP + CDP ↔ ADP + CTP	C00063	0.77.4.2
C00075	R00571, R00573	CTP synthase (6.3.4.2)	ATP + UTP + Ammonia → ADP + Orthophosphate + CTP	C00063	0.77.4.6
C00378	R00619	thiamine diphosphokinase (2.7.6.2)	ATP + Thiamine → AMP + Thiamin diphosphate	C00068	0.14
C00631	R00658 ^r	enolase (4.2.1.11)	2-Phospho-D-glycerate ↔ Phosphoenolpyruvate + H ₂ O	C00074	0.77.4.0

Substrate KEGG ID	Reaction ID	Enzyme Name (E.C. Number)	Definition	Product KEGG ID	MBB ID
C00186	R00703 ^r	lactate dehydrogenase (1.1.1.27)	(S)-Lactate + NAD+ ↔ Pyruvate + NADH + H+	C00022	0.0
C00093	R00842 R00844 ^r	sn-glycerol-3-phosphate dehydrogenase (1.1.1.94)	sn-Glycerol 3-phosphate + NAD+ ↔ Glycerone phosphate + NADH + H+	C00111	0.77.4.0
C00093 + C00040	R00851	arylamine N-acetyltransferase (2.3.1.15)	sn-Glycerol 3-phosphate + Acyl-CoA → 1-Acyl-sn-glycerol 3-phosphate + CoA	C00681	0.8
C00415	R00936 R00939 ^r	dihydrofolate reductase (1.5.1.3)	Dihydrofolate + NADH + H+ ↔ Tetrahydrofolate + NAD+	C00101	0.79.0
C00037 + C00143	R00945 ^r	glycine hydroxymethyltransferase (2.1.2.1)	5,10-Methylenetetrahydrofolate + Glycine + H ₂ O ↔ Tetrahydrofolate + L-Serine	C00065 + C00101	0.79.0
C00105	R00966 ^r	uracil phosphoribosyltransferase (2.4.2.9)	UMP + Diphosphate ↔ Uracil + 5-Phospho-alpha-D-ribose 1-diphosphate	C00106 + C00119	0.77.4.0
C00117	R01049 ^r	phosphoribosylpyrophosphate synthetase (2.7.6.1)	ATP + D-Ribose 5-phosphate ↔ AMP + 5-Phospho-alpha-D-ribose 1-diphosphate	C00119	0.77.4.0
C00117	R01056 ^r	ribose-5-phosphate isomerase (5.3.1.6)	D-Ribose 5-phosphate ↔ D-Ribulose 5-phosphate	C00199	0.77.4.0
C00118	R01061 ^r	glyceraldehyde-3-phosphate dehydrogenase (1.2.1.12)	D-Glyceraldehyde 3-phosphate + Orthophosphate + NAD+ ↔ 3-Phospho-D-glyceroyl phosphate + NADH + H+	C00236	0.77.4.0
C05378	R01070 ^r	fructose-1,6-bisphosphate aldolase (4.1.2.13)	beta-D-Fructose 1,6-bisphosphate ↔ Glycerone phosphate + D-Glyceraldehyde 3-phosphate	C00111 + C00118	0.77.4.0
C00131	R01138 ^r	pyruvate kinase (2.7.1.40)	dATP + Pyruvate ↔ dADP + Phosphoenolpyruvate	C00206	0.78.1.1
C00119 + C00242	R01229 ^r	hypoxanthine phosphoribosyltransferase (2.4.2.8)	Guanine + 5-Phospho-alpha-D-ribose 1-diphosphate ↔ GMP + Diphosphate	C00144	0.77.4.0
C00361	R01858	pyruvate kinase (2.7.1.40)	dGDP + Phosphoenolpyruvate → dGTP + Pyruvate	C00286	0.6
C00008	R02017	ribonucleoside diphosphate reductase (1.17.4.1)	Thioredoxin + ADP → dADP + Thioredoxin disulfide + H ₂ O	C00206	0.78.1.2
C00035	R02019	ribonucleoside diphosphate reductase (1.17.4.1)	GDP + Thioredoxin → dGDP + Thioredoxin disulfide + H ₂ O	C00361	0.77.4.7.0

Substrate KEGG ID	Reaction ID	Enzyme Name (E.C. Number)	Definition	Product KEGG ID	MBB ID
C00112	R02024	ribonucleoside diphosphate reductase (1.17.4.1)	Thioredoxin + CDP → dCDP + Thioredoxin disulfide + H ₂ O	C00705	0.77.4.5
C00197	R01512 ^r	phosphoglycerate kinase (2.7.2.3)	ATP + 3-Phospho-D-glycerate ↔ ADP + 3-Phospho-D-glyceroyl phosphate	C00236	0.77.4.0
C00631	R01518 ^r	phosphoglycerate mutase (2,3-diphosphoglycerate-independent) (5.4.2.12)	2-Phospho-D-glycerate ↔ 3-Phospho-D-glycerate	C00197	0.77.4.0
C00199	R01529 ^r	ribose-phosphate 3-epimerase (5.1.3.1)	D-Ribulose 5-phosphate ↔ D-Xylulose 5-phosphate	C00231	0.77.4.0
C00118 + C05382	R01641 ^r	transketolase (2.2.1.1)	Sedoheptulose 7-phosphate + D-Glyceraldehyde 3-phosphate ↔ D-Ribose 5-phosphate + D-Xylulose 5-phosphate	C00117 + C00231	0.77.4.0
C00063 + C00416	R01799 ^r	phosphatidate cytidyltransferase (2.7.7.41)	CTP + Phosphatidate → Diphosphate + CDP-diacylglycerol	C00269	0.5
C00065 + C00269	R01800	phosphatidylserine synthase (2.7.8.8)	-diacylglycerol + L-Serine → CMP + Phosphatidylserine	C00055 + C02737	0.4
C00279 + C00111	R01829	fructose-1,6-bisphosphate aldolase (4.1.2.13)	Glycerone phosphate + D-Erythrose 4-phosphate → Sedoheptulose 1,7-bisphosphate	C00447	0.77.4.0
C00118 + C05345	R01830 ^r	transketolase (2.2.1.1)	beta-D-Fructose 6-phosphate + D-Glyceraldehyde 3-phosphate → D-Erythrose 4-phosphate + D-Xylulose 5-phosphate	C00231 + C00279	0.77.4.0
C00363	R02093 ^r	nucleoside diphosphate kinase (2.7.4.6)	ATP + dTDP ↔ ADP + dTTP	C00459	0.81.0
C00364	R02094 ^r	thymidine monophosphate kinase (2.7.4.9)	ATP + dTMP ↔ ADP + dTDP	C00363	0.81.0
C00365	R02098 ^r	thymidine monophosphate kinase (2.7.4.9)	ATP + dUMP ↔ ADP + dUDP	C01346	0.77.4.1
C00143 + C00365	R02101	thymidylate synthase (2.1.1.45)	dUMP + 5,10-Methylenetetrahydrofolate → Dihydrofolate + dTMP	C00364 + C00415	0.79.0
C00040 + C00681	R02241	1-acyl-sn-glycerol-3-phosphate acyltransferase (2.3.1.51)	1-Acyl-sn-glycerol 3-phosphate + Acyl-CoA → Phosphatidate + CoA	C00416	0.7
C00458	R02325	dCTP deaminase (3.5.4.13)	dCTP + H ₂ O → dUTP + Ammonia	C00460	0.77.4.4

Substrate KEGG ID	Reaction ID	Enzyme Name (E.C. Number)	Definition	Product KEGG ID	MBB ID
C00705	R02326 ^r	nucleoside diphosphate kinase (2.7.4.6)	ATP + dCDP ↔ ADP + dCTP	C00458	0.77.4.3
C01346	R02331 ^r	nucleoside diphosphate kinase (2.7.4.6)	ATP + dUDP ↔ ADP + dUTP	C00460	0.77.4.1
C02737	R02055	phosphatidylserine decarboxylase (4.1.1.65)	Phosphatidylserine → Phosphatidylethanolamine + CO ₂	C00350	0.3
C00504	R02235 R02236 ^r	dihydrofolate reductase (1.5.1.3)	Folate + NADH + H ⁺ ↔ Dihydrofolate + NAD ⁺	C00415	0.79.0
C03150	R02324	ribosylnicotinamide kinase (2.7.1.22)	ATP + Nicotinamide-beta-riboside → ADP + Nicotinamide D-ribonucleotide	C00455	0.80.0
C00031	R02738	protein-N(pi)-phosphohistidine—D-glucose phosphotransferase (2.7.1.199)	Protein N(pi)-phospho-L-histidine + D-Glucose → Protein histidine + alpha-D-Glucose 6-phosphate	C00668	0.15
C00668	R02740 ^r	glucose-6-phosphate isomerase (5.3.1.9)	alpha-D-Glucose 6-phosphate ↔ beta-D-Fructose 6-phosphate	C05345	0.77.4.0
C00831	R02971	pantetheine kinase (2.7.1.34)	ATP + Pantetheine → ADP + Pantetheine 4'-phosphate	C01134	0.16
C00864	R03018	pantothenate kinase (2.7.1.33)	ATP + Pantothenate → ADP + D-4'-Phosphopantothenate	C03492	0.19
C01134	R03035 ^r	pantetheine-phosphate adenyltransferase (2.7.7.3)	ATP + Pantetheine 4'-phosphate → Diphosphate + Dephospho-CoA	C00882	0.1
C03492	R04231 ^r	phosphopantothencysteine synthetase (6.3.2.5)	CTP + D-4'-Phosphopantothenate + L-Cysteine ↔ CMP + Diphosphate + (R)-4'-Phosphopantothencysteine	C04352	0.18
C04079	R04391 ^r	pantothenate kinase (2.7.1.33)	ATP + N-(R)-Pantothencysteine ↔ ADP + (R)-4'-Phosphopantothencysteine	C04352	3415
C05345	R04779 ^r	6-phosphofructokinase (2.7.1.11)	ATP + beta-D-Fructose 6-phosphate ↔ ADP + beta-D-Fructose 1,6-bisphosphate	C05378	0.77.4.0
C04352	R03269	phosphopantothencysteine decarboxylase (4.1.1.36)	(R)-4'-Phosphopantothencysteine → Pantetheine 4'-phosphate	C01134	0.17
C05382	R01843 ^r	6-phosphofructokinase (2.7.1.11)	ATP + Sedoheptulose 7-phosphate ↔ ADP + Sedoheptulose 1,7-bisphosphate	C00447	0.77.4.0

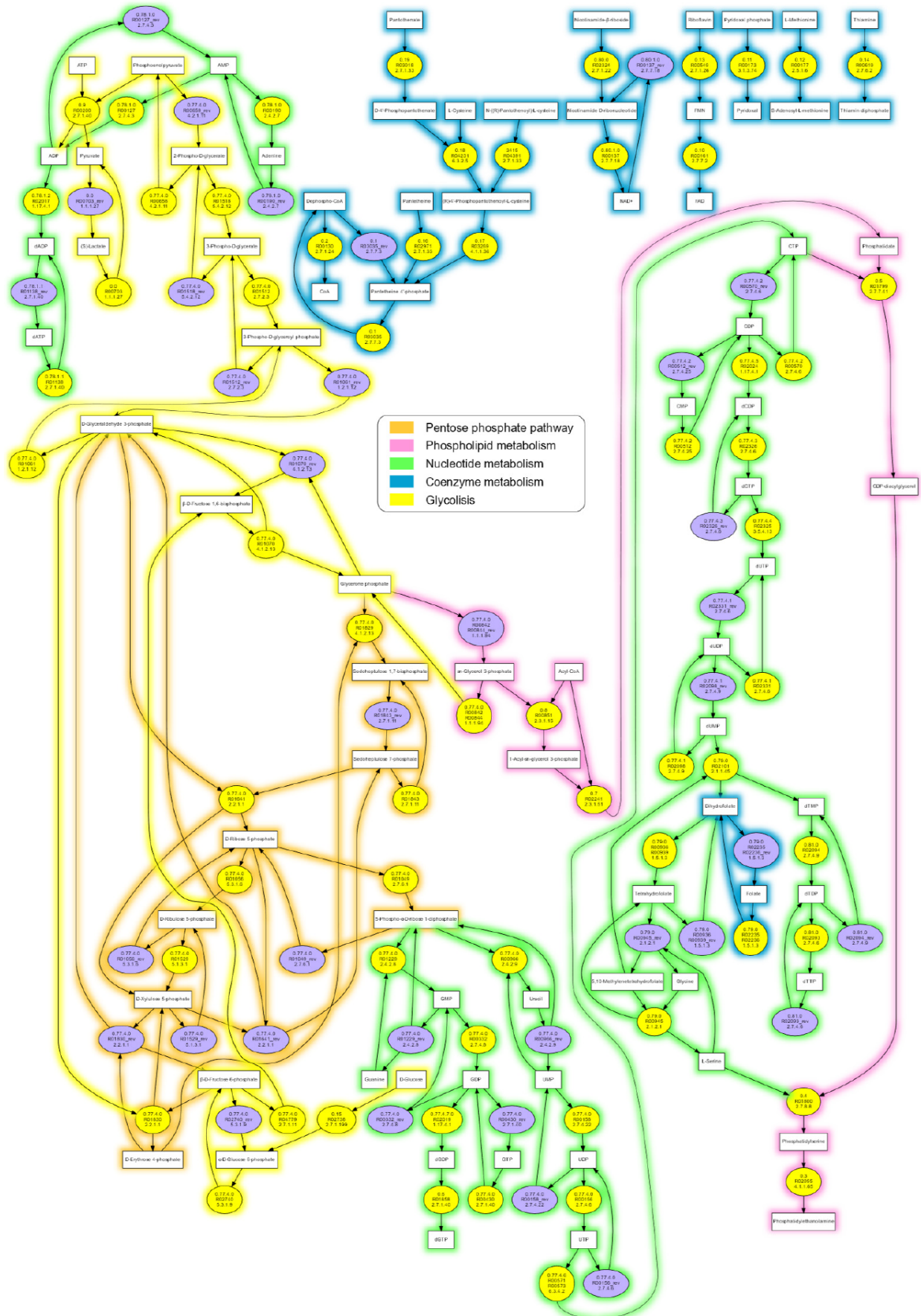


Figure C3.2 | The reaction graph of the proposed theoretical minimal metabolic network represented in Figure C3.1, obtained using data from the KEGG database. The yellow-filled circles are the reactions with their KEGG ID and E.C. numbers, and the purple-filled circles are the reverse reaction of the yellow-filled circles, when appropriate. Line colors denote metabolic categories. A full-size representation can be seen in Supplementary Figure C3S1.

3.4 The MetaDAG methodology: analysis of the composition and connectivity of a network at a glance

Despite the fact that the reaction graph of the theoretical minimal organism constructed in this work has only 98 reactions and 80 metabolites, it is difficult to visualize the detailed relationships between the reactions that make up the network's connectivity (Figure C3.2). To solve this problem, we used the MetaDAG methodology (Alberich et al., 2017) to generate an m-DAG of the manually curated reaction graph. An m-DAG is a suitable reduction of a metabolic network. Namely, the reactions that are connected by multiple paths, which are the strongly connected components of the metabolic network, are contracted into one single MBB, which can be considered a robust subgraph in the reaction graph. Moreover, those MBBs that only represent a reaction that is not biconnected to any other reaction are essential to maintain the network's connectivity. In this sense, the m-DAG provides modularity of the reaction graph that keeps the information on the robustness and connectivity of the metabolic network.

The m-DAG we obtained from the minimal metabolic reaction graph (Table C3.1, Figure C3.3) has a total of 36 nodes, 25 of them corresponding to

single reactions (yellow nodes) and 11 to contracted MBBs (gray nodes). Clearly, there are seven connected components in this network, the biggest one covering the central metabolism of the hypothetical minimal organism, while the rest are the reactions that synthesize the essential cofactors needed for the proper functionality of the complete cell.

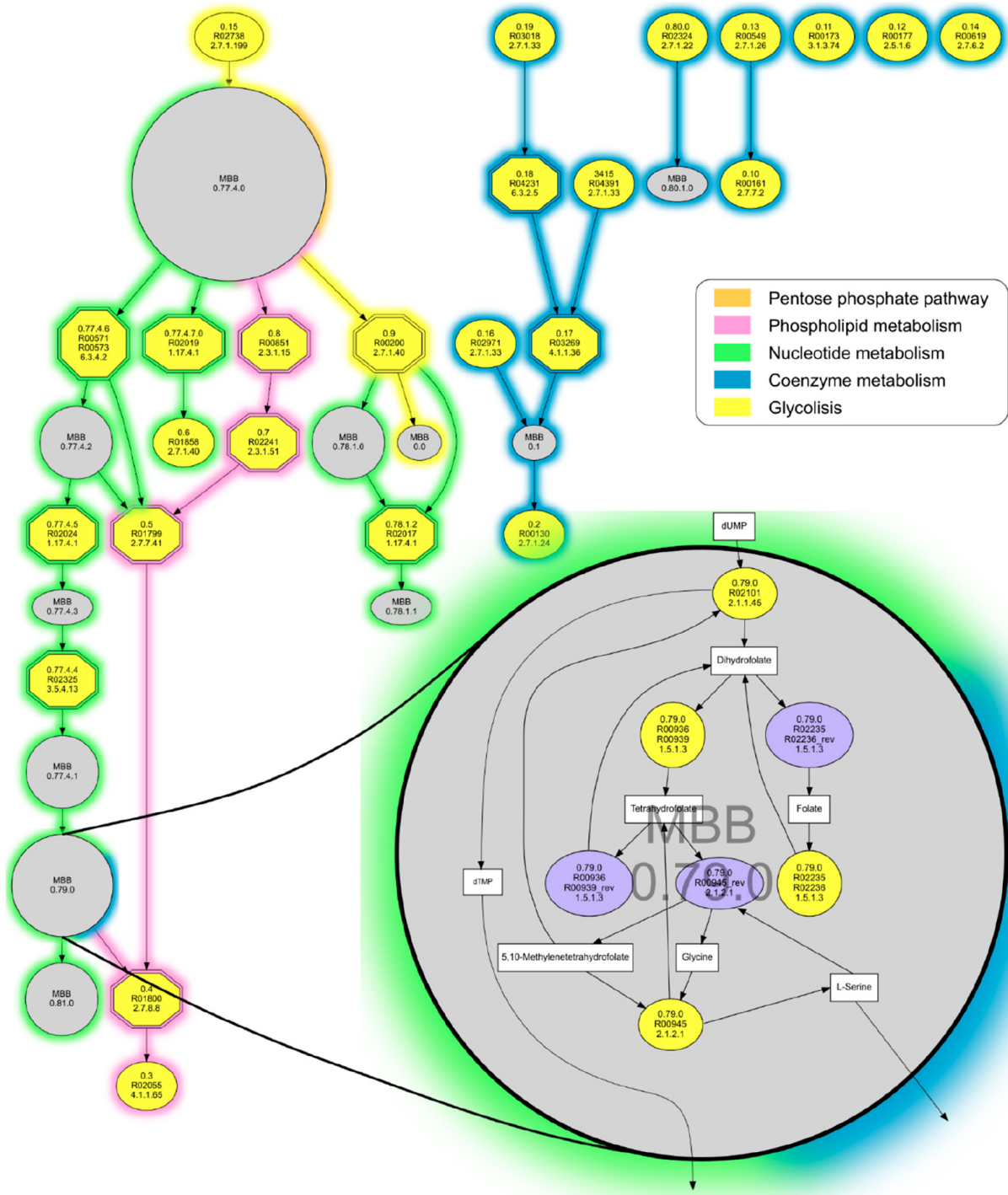


Figure C3.3 | m-DAG of the metabolism of a theoretical minimal bacterial cell. Single reactions appear in yellow, contracted MBBs in grey, and the essential reactions as hexagons with double lines. Line colors denote metabolic categories. MBB 0.79.0 is zoomed in as an

example of how a strongly connected component, which is a cyclic subgraph formed by 7 reactions and 7 compounds, is reduced to one node in our m-DAG.

In addition, essential reactions (i.e., those whose removal reduces the network's connectivity increasing the number of connected components) can be easily identified using this approach (hexagons with double lines in Figure C3.3). Table C3.2 is a list of the 12 essential reactions we found in the minimal metabolic network under study and the metabolic pathways where they participate. They are involved in purine and pyrimidine metabolism, glycerophospholipid metabolism, glycolysis and pantothenate, and CoA biosynthesis. Purines and pyrimidines are the most abundant metabolic substrates for all living organisms. They are essential components for the synthesis of DNA and RNA, and they also participate in the biosynthesis of energy nucleotides and are vital cofactors for cell survival and reproduction. Hence, purines and their by-products widely participate in biological processes. Glycerophospholipids are pivotal structural components of the cell membranes, but they are also precursors of many essential biological molecules and participate in cell signaling and other cellular processes (Alvarez & Georgellis, 2019). Glycolysis is the first step in the breakdown of glucose to extract energy for cellular metabolism by creating high-energy molecules. It is considered an ancient metabolic pathway (Romano & Conway, 1996), and its prevalence in organisms is nearly ubiquitous.

Table C3.2 | Essential reactions of the m-DAG constructed from the theoretical minimal gene set machinery needed for life.

Reaction ID	Metabolic Pathway
R02019 R02017	Purine metabolism
R00571/R00573 R02024 R02325	Pyrimidine metabolism
R00851 R02241	Glycerolipid metabolism, Glycerophospholipid metabolism
R01799 R01800	Glycerophospholipid metabolism
R00200	Glycolysis, part of the pyruvate metabolism
R04231 R03269	Pantothenate and CoA biosynthesis

We consider that what we call “essential reactions”, easily highlighted by the MetaDAG methodology, can be of crucial importance in many fields of research. Probably, the most logical and of vital importance is the idea that these reactions can help choose enzymes as potential drug targets since the removal of these reactions breaks metabolic pathways, which can lead to non-viable cells. Considering that m-DAGs take into account complete genomes, and even complementary genomes (they can be calculated for two or more genomes together, to simulate complementary metabolic pathways within consortia), the resulting essential reactions are trustworthy in the sense that researchers might overlook an enzyme doing the same job as the one highlighted and, if they find it, it would be a new discovery not previously described for a specific metabolic pathway.

3.5 The m-DAG of “*Candidatus Nasuia deltocephalinicola*”

In the case of a minimal metabolic network, each item included in the list of reactions and compounds is hypothetically essential for survival. When we extrapolate these results to living organisms possessing naturally minimized genomes, such as pathogens or mutualist endosymbiotic bacteria, we should consider that their metabolism is a patchwork dependent on the host and, in many cases, also dependent on other bacteria with which they live in consortia. Therefore, the study of their networks' connectivity has the potential of pointing out genes encoding critical steps that connect the different partners in a given pathway. Subsequently, the genes that encode those reactions can become targets for genetic engineering, and/or for mechanisms intended to regulate the cell metabolism; additionally, they might also have the potential to destroy the stability of the relationship, even killing the undesired organism in a parasitic relationship.

In order to compare the *in silico* minimal m-DAG with the m-DAG from a living organism with a naturally reduced genome, we constructed the m-DAG of “*Ca. Nasuia deltocephalinicola*” str. NAS-ALF (from now on referred to as *Nasuia* for simplicity; Supplementary Figure C3S2), an obligate endosymbiotic bacteria of the aster leafhopper *Macrosteles quadrilineatus* (Bennett & Moran, 2013). This endosymbiont possesses the smallest natural genome known so far, comprising 112,091 bp and only 138 protein-coding genes identified. The metabolic data needed to generate this m-DAG, including the complete list of its enzymes, reactions, and compounds were also obtained from the KEGG database

(Table C3S2). *Nasuia*'s m-DAG comprises 29 nodes included in 12 connected components, with 7 MBBs and 22 single reactions. Regarding the single reactions, five are essential (summarized in Table C3.3).

Table C3.3 | Essential reactions of the m-DAG of "*Candidatus Nasuia deltocephalinicola*" str. NAS-ALF.

Reaction ID	Metabolic Pathway
R09372	Selenocompound metabolism
R00443	Purine metabolism, Glycerophospholipid metabolism
R03012	Histidine metabolism
R01163	Histidine metabolism
R01288	Cysteine and methionine metabolism, Sulfur metabolism

It has been estimated that more than 60% of insects possess symbiotic bacteria inside their body tissues, and/or very often in a specialized cell type called bacteriocyte (Weinert et al., 2015). When these bacteria become endosymbionts, they lose their ability to interact with other organisms. Additionally, they become dependent on their respective hosts, and their genome is significantly reduced by the deletion of genes that become redundant or that are not needed in a rich environment such as the one they encounter within their hosts (Gil et al., 2004; Toft & Andersson, 2010). In addition, even though the niche is significantly rich for them, the insect host generally has a very incomplete diet by feeding on plant sap or seeds, or blood from mammals, so the bacteria become their helpers for the production of essential amino acids, fatty acids, or vitamins (Douglas, 2009; Vigneron et al., 2014). The essential reactions of

Nasuia's m-DAG reveal exactly that. This organism works as a factory of the vitamins and amino acids that *M. quadrilineatus* needs to survive. Moreover, this bacterium is part of a consortium with "*Candidatus Sulcia muelleri*" str. ALF (Bennett & Moran, 2013). It is widely accepted that the endosymbiotic relationship between insects and bacteria, dating from 10 to several hundred million years, allowed the proliferation of insects and their diversification in almost any ecological niche (McCutcheon et al., 2009b; Tamas et al., 2002). Obviously, if the reactions that link the metabolic routes disappear (either naturally or due to targeted modification of those genes), this association would be affected to the point of the possible death of the host.

A direct comparison between the reactions and compounds that make up the in silico m-DAGs of the theoretical minimal cell and *Nasuia* would not be significant due to their dissimilar lifestyles. What we can easily assess is the topology of the networks. At first glance, it is striking that the smallest genome found in nature has fewer nodes than the in silico m-DAG. The dependence of this endosymbiotic bacteria on its host and on its second co-obligate endosymbiont explains this phenomenon.

3.6 The first semisynthetic viable cell and its m-DAG's reconstruction

To complete our comparative analysis, we constructed the m-DAG of JCVI-syn3.0, which is an artificially designed and manufactured viable cell whose genome arose by minimizing the one from *Mycoplasma mycoides* JCVI-syn1.0 created by Hutchison et al. in 2016 (Hutchison et al., 2016). To do so,

we used the list of enzymes presented in their article and converted it into a list of reactions and compounds, compared them to our minimal metabolic network (Table C3S3), and created the reaction graph of JCVI-syn3.0 and its eventual m-DAG (Figure C3.4). JCVI-syn3.0 m-DAG is formed by 34 connected components, with a total of 70 nodes, 54 of them corresponding to single reactions, and 16 contracted MBBs. Ten reactions are essential (summarized in Table C3.4), that is, indispensable to maintain the connectivity of the network.

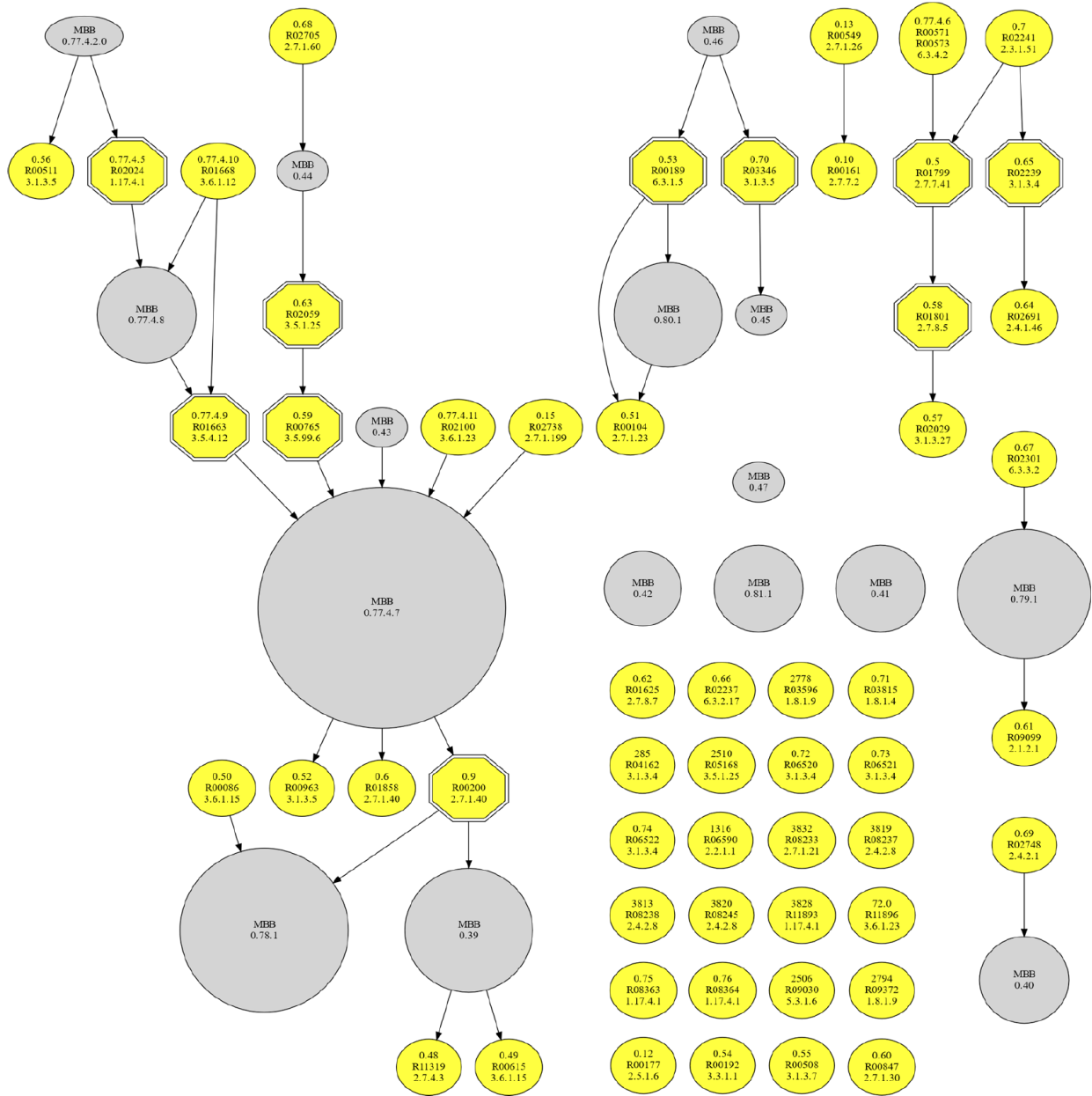


Figure C3.4 | m-DAG of the metabolism of JCVI-syn3.0. Single reactions appear in yellow, contracted MBBs in grey, and the essential reactions as hexagons with double lines.

Table C3.4 | Essential reactions of the m-DAG of JCVI-syn3.0.

Reaction ID	Metabolic Pathway
R02024 R01663	Pyrimidine metabolism
R02059 R00765	Amino sugar and nucleotide sugar metabolism
R00200	Glycolysis, part of the pyruvate metabolism
R00189 R03346	Nicotinate and nicotinamide metabolism
R01799 R01801 R02239	Glycerophospholipid metabolism

Once again, the essential reactions are involved in the metabolism of nucleotides, phospholipids, and coenzymes, even though there are significant differences between the list of reactions included in the reconstruction of JCVI-syn3.0 and the metabolic minimal network (Supplementary Table C3S3). JCVI-syn3.0 has 155 reactions included in its reaction graph, while our minimal network reaction graph has only 63 (98 when taking reverse reactions into account). The explanation for these differences is that the minimal network defined by (Gil et al., 2004) considers the minimal bacterium to live in a controlled and nutrient-rich environment, while JCVI-syn3.0 includes some metabolic pathways that are essential for the specific necessities of *M. mycoides*, its reproduction, and its survival. Interestingly enough, two reactions are essential for both networks (R02024 and R00200), while others participate closely in the same pathways (e.g., R01800 and R01801), which may be useful information for genetic engineering purposes.

3.7 Resemblance of the MBBs of the minimal m-DAGs

In order to contrast the MBBs of the three m-DAGs constructed in this study, Table C3.5 shows the correspondence among them. The list of enzymes and the definition of each reaction is presented in Supplementary Table C3S4.

Table C3.5 | Comparison of the MBBs of the three networks under study. Every row lists the reactions belonging to the corresponding MBB and the enzymes involved in those reactions. The list includes only MBBs composed of at least three reactions (reverse included) or with fewer reactions but that are shared by at least two of the networks under study.

Model Cell	MBB ID	# Reactions	Reaction ID
Minimal cell	0.77.4.0	21	R00156 (2.7.4.6) ^r , R00158 (2.7.4.22) ^r , R00332 (2.7.4.8) ^r , R00430 (2.7.1.40) ^r , R00658 (4.2.1.11) ^r , R00842 R00844 (1.1.1.94) ^r , R00966 (2.4.2.9) ^r , R01049 (2.7.6.1) ^r , R01056 (5.3.1.6) ^r , R01061 (1.2.1.12) ^r , R01070 (4.1.2.13) ^r , R01229 (2.4.2.8) ^r , R01512 (2.7.2.3) ^r , R01518 (5.4.2.12) ^r , R01529 (5.1.3.1) ^r , R01641 (2.2.1.1) ^r , R01829 (4.1.2.13), R01830 (2.2.1.1) ^r , R01843 (2.7.1.11) ^r , R02740 (5.3.1.9) ^r , R04779 (2.7.1.11)
	0.77.4.1	2	R02098(2.7.4.9) ^r , R02331 (2.7.4.6) ^r
	0.77.4.2	2	R00512(2.7.4.25) ^r , R00570 (2.7.4.6) ^r
	0.77.4.7.0	1	R02019 (1.17.4.1)
	0.78.1.0	2	R00127 (2.7.4.3) ^r , R00190 (2.4.2.7) ^r
	0.78.1.1	1	R01138(2.7.1.40) ^r
	0.78.1.2	1	R02017 (1.17.4.1)
	0.79.0	4	R00936 R00939 (1.5.1.3) ^r , R00945 (2.1.2.1) ^r , R02101 (2.1.1.45), R02235 R02236 (1.5.1.3) ^r
	0.80.1.0	1	R00137 (2.7.7.18) ^r
	0.81.0	2	R02093 (2.7.4.6) ^r , R02094 (2.7.4.9) ^r
JCVI-syn 3.0	0.39	6	R00014 (1.2.4.1), R00230 (2.3.1.8) ^r , R00315 (2.7.2.1) ^r , R02569 (2.3.1.12) ^r , R03270 (1.2.4.1), R07618 (1.8.1.4) ^r
	0.40	3	R01126 (3.1.3.5), R01132 (2.4.2.8) ^r , R01863 (2.4.2.1) ^r
	0.41	3	R02142 (2.4.2.8) ^r , R02297 (2.4.2.1) ^r , R02719 (3.1.3.5)
	0.42	2	R00921 (2.3.1.8) ^r , R01353 (2.7.2.1) ^r
	0.77.4.2.0	1	R00512 (2.7.4.25) ^r
	0.77.4.7	49	R00158 (2.7.4.22) ^r , R00289 (2.7.7.9) ^r , R00291 (5.1.3.2) ^r , R00332 (2.7.4.8) ^r , R00430 (2.7.1.40) ^r , R00505 (5.4.99.9) ^r , R00658 (4.2.1.11) ^r , R00959 (5.4.2.5) ^r , R00966 (2.4.2.9) ^r , R01015 (5.3.1.1) ^r , R01049 (2.7.6.1) ^r , R01056 (5.3.1.6) ^r , R01057 (5.4.2.7) ^r , R01058 (1.2.1.9), R01061 (1.2.1.12) ^r , R01066 (4.1.2.4) ^r , R01067 (2.2.1.1), R01068 (4.1.2.13) ^r , R01070 (4.1.2.13) ^r , R01227 (3.1.3.5), R01229 (2.4.2.7) ^r , R01229 (2.4.2.8) ^r , R01512 (2.7.2.3) ^r , R01518 (5.4.2.12) ^r , R01529 (5.1.3.1) ^r , R01641 (2.2.1.1) ^r , R01819 (5.3.1.8) ^r , R01827 (2.2.1.2) ^r , R01829 (4.1.2.13), R01830 (2.2.1.1) ^r , R01843 (2.7.1.11) ^r , R01967 (2.7.1.113) ^r , R01968 (3.1.3.5) ^r , R01969 (2.4.2.1) ^r , R02018 (1.17.4.1), R02019 (1.17.4.1), R02090 (2.7.4.8) ^r , R02098 (2.7.4.9) ^r , R02099 (2.7.1.21), R02102 (3.1.3.5), R02102 (3.1.3.89), R02147 (2.4.2.1) ^r , R02484 (2.4.2.1), R02568 (4.1.2.13) ^r , R02739 (5.3.1.9) ^r , R02740 (5.3.1.9) ^r , R02749 (5.4.2.7) ^r , R03321 (5.3.1.9) ^r , R04779 (2.7.1.11)
0.77.4.8	5	R01664 (3.1.3.5), R01664 (3.1.3.89), R01665 (2.7.4.25) ^r , R01666 (2.7.1.74), R01667 (3.6.1.12)	
0.78.1	12	R00127 (2.7.4.3) ^r , R00183 (3.1.3.5), R00185 (2.7.1.74) ^r , R00190 (2.4.2.7) ^r , R01138 (2.7.1.40) ^r , R01547 (2.7.4.11) ^r , R01547 (2.7.4.3) ^r , R01561 (2.4.2.1) ^r , R02017 (1.17.4.1), R02088 (3.1.3.5) ^r , R02089 (2.7.1.76), R02557 (2.4.2.1) ^r	
0.79.1	6	R00942 (6.3.2.17) ^r , R00945 (2.1.2.1) ^r , R01220 (1.5.1.5) ^r , R01655 (3.5.4.9) ^r , R03940 (2.1.2.9), R04241 (6.3.2.17) ^r	
0.80.1	4	R00137 (2.7.7.18) ^r , R01271 (2.4.2.12) ^r , R02294 (2.4.2.1) ^r , R02323 (3.1.3.5)	
0.81.1	4	R01567 (2.7.1.21), R01569 (3.1.3.5), R01569 (3.1.3.89), R02094 (2.7.4.9) ^r	
Nasuia	0.77.0	2	R00435 (2.7.7.6) ^r , R00441 (2.7.7.6) ^r
	0.78.0	2	R00375 (2.7.7.7) ^r , R00376 (2.7.7.7) ^r

Reversible reactions are denoted by the superscript r. Reactions depicted in blue are shared by the minimal and the synthetic (JCVI-syn3.0) metabolic networks. #: number of.

3.8 Conclusions

The construction of the minimal metabolic reaction graph and its consequent m-DAG presented in this work can be of great use in the field of synthetic biology. The composition of compounds and reactions that we present can easily be extrapolated to any phylogenetically diverse bacteria of interest considering that we did not focus specifically on genes. Chemistry and molecular biology technologies are also thriving. Thus, the in silico design of bacteria with the small number of metabolic genes described in this paper may be more feasible than previously thought.

Chapter 4. Metabolic networks of insects' endosymbiotic bacteria preserve evolutionary traces

This chapter reproduces the following manuscript entirely:

Reyes-Prieto, Mariana, David J. Martínez-Cano, Mercè Llabrés, Pere Palmer,
Carlos Vargas-Chávez, Luis Delaye, Rosario Gil, and Andrés Moya.

“Evolutionary traces in metabolic networks of endosymbiotic bacteria of insects.”

The closure of this thesis is the last manuscript I wrote, where the ideas behind have been developing for several years. For this paper, we converged all the aforementioned prefaces and experimentally analyzed the genomic information of the endosymbiotic bacteria of insects included in SymGenDB. We present a large-scale evolutionary analysis of metabolic networks where we applied two metabolic reconstruction models to all those genomes, one by conventional manners published in several studies (metabolite-based model) and one by applying the MetaDAG methodology described in the Introduction of this work. We also included the information we published for the minimal metabolism and calculated its metabolic network's topological parameters to compare with the rest of the genomes.

We found a significant correlation between the clustering coefficient, the network diameter, and the number of nodes of the networks within taxonomic groups assigned at the genus level when we conducted evolutionary tests of the network's parameters associated with the genome size of each organism and also found a correlation between the metabolic network's distances based on the MetaDAG analysis, and the taxonomic distances of a molecular phylogeny.

The major conclusion of this work is that there are evolutionary traces embedded in the metabolic networks of endosymbiotic bacteria of insects. Large-scale analyses such as this one could give a better understanding of how metabolism modulates the degree of variation among phylogenetically different organisms living in diverse niches but sharing similar evolutionary constraints.

4. Evolutionary traces in metabolic networks of endosymbiotic bacteria of insects

There are billions of bacterial species in the world, and it has been estimated that around 60% of them are associated with a host (Yarza & Munoz, 2014). Endosymbiotic bacteria of insects are a fraction within this vast pool that has been studied in detail. In these symbiotic relationships, the bacteria can provide the host with new phenotypic traits by supplementing restricted nutrients, breaking down plant polymers or detoxifying plant resistance compounds, and providing resistance to stress or defense against antagonists, among others (Hansen & Moran, 2014; Harmon et al., 2009; Sudakaran et al., 2017). As a result, hosts can expand into new niches which allows their lineage diversification. Consequently, this interaction has been considered one of the major driving forces of evolutionary innovation of this group, the most species-rich among animals (Bennett & Moran, 2015; Joy, 2013; McCutcheon et al., 2019; Sudakaran et al., 2017).

The availability of a high number of endosymbionts' genomes allows their exploration to answer questions on how the process of symbiotic integration inside eukaryotic cells took place and how its ulterior evolution has been (Borenstein & Feldman, 2009; Martínez-Cano et al., 2014). A better understanding of such questions can be obtained from the reconstruction of the metabolic networks inferred from genomic data, as part of what has been called the 'reverse ecology' approach (Borenstein & Feldman, 2009).

A metabolic network is the complete set of biochemical reactions that can be performed by an organism (Parter et al., 2007). The network captures the

functional properties of the organism, and evolutionary studies carried on multiple metabolic networks have shown that variations in their nodes and edges are subject to different selection pressures owing to distinct topological features (Yamada & Bork, 2009). Furthermore, network analyses can also provide insights about the environment in which a species evolved, because the interactions with the environment through evolution and the selection pressures that the environment imposes on the metabolic network can modify the network's organization and result in a topological 'signature' (Yamada & Bork, 2009).

Endosymbiotic bacteria of insects reduce their genomes while becoming obligatory endosymbionts (Moya et al., 2008), sometimes in an extreme manner. For instance, the 'symbionelle', refers to an organism that has suffered a drastic genome reduction, way beyond what is theoretically considered a minimal cell, and that is metabolically dependent on other organisms (Reyes-Prieto et al., 2014). In these conditions, the biochemical environment depends on the metabolic context of their corresponding hosts, the existence of other concurrent bacteria with which they establish a symbiotic consortium (as in the case of *Buchnera* and *Serratia*, *Tremblaya* and *Moranella*, or *Sulcia* and *Baumannia*, to name a few), or both (Borenstein & Feldman, 2009). (As a note, endosymbionts are often labeled *Candidatus* as the initial part of their name because they cannot be grown in a bacterial culture, followed by the genus and the species name of the bacterium; to facilitate understanding, hereupon we will only refer to endosymbiotic bacteria by their genus and species name, or just with the genus name when only one species has been described). Such associations and the

selective pressures acting upon the evolution of each species configure their metabolic networks and could significantly affect their structure. It is highly possible that, due to their interconnected and reduced metabolic networks, these symbiotic relationships generate a unique signature, which, in turn, could produce specific ecological, topological, and dynamic patterns. Additionally, the accessibility of taxonomically related bacterial genomes with similar evolutionary constraints may allow tracing metabolic evolution on a phylogeny, which can help to disentangle the evolutionary processes and constraints that affect the evolution of metabolic networks (Mithani et al., 2010).

Previous studies have focused on computational strategies to analyze the topological features of metabolic networks at a small scale, normally of one or a few organisms at a time. Other studies have concentrated on the environmental framework in which each network functions, and analyzed the effect of biochemical environments, focusing on many metabolic pathways at a time within specific models (Almaas et al., 2004; Ibarra et al., 2002). In this work, we present a large-scale evolutionary analysis of metabolic networks with a focus on endosymbiotic bacteria of insects. We studied a total of 144 bacterial genomes and compared their inferred metabolic networks to a theoretical minimal one derived from the minimal gene set proposed to sustain cellular life (Gil et al., 2004). We found that, as expected, endosymbiotic bacteria of insects present some of the smallest metabolic networks inferred so far, some even smaller than a theoretical minimal metabolic network (Reyes-Prieto, Gil, et al., 2020). We also compared the metabolic networks within taxonomic groups at the genus level,

inferred a ‘pan-metabolic’ network (the interaction of all pathways of each bacterial species in our data set, named after the concept of the ‘pangenome’ (Medini et al., 2005; Tettelin, 2020), and inferred evolutionary patterns on those taxonomic groups using two different models: a metabolite- and a reaction-based model. In the first approach, we conducted coevolutionary tests of the network’s parameters and found a significant correlation between the clustering coefficient, the network diameter, and the number of nodes of these networks within the taxonomic groups. For the second approach, we calculated the distances between the organisms of our data set after calculating their Metabolic Building Blocks (Alberich et al., 2017), and we also found a correlation between the distances observed within the MBBs and the taxonomic distances of a molecular phylogeny. The major conclusion of this work is that there are evolutionary traces embedded in the metabolic networks of endosymbiotic bacteria of insects. Large-scale analyses such as this one could give a better understanding of how metabolism modulates the degree of variation among phylogenetically different organisms living in diverse niches but sharing similar evolutionary constraints.

4.1 Genomes of endosymbiotic bacteria of insects and their free-living relatives

All genomic information of the endosymbiotic bacteria of insects included in this analysis was downloaded from the SymGenDB database (Reyes-Prieto, et al., 2020), and those of their free-living relatives from the Microbial Genome Database (Uchiyama et al., 2015). We found 169 symbiotic bacteria of insects in

SymGenDB and discarded the genomes of organisms associated with insects that were not endosymbiotic, resulting in 109 genomes. Furthermore, we selected 35 free-living bacteria to root the endosymbionts. The final list of the organisms considered for this work consists of 144 bacterial genomes (Supplementary Table C4S1). These endosymbiotic bacteria of insects belong to the following taxonomic phyla: Actinobacteria (3), Proteobacteria (80), Bacteroidetes (20), Elusimicrobia (1), Fusobacteria (1), Tenericutes (2), and Spirochaetes (3). For evolutionary analyses, endosymbiont's genomes were grouped at the genus level (Supplementary Table C4S1 and Figure C4.1).

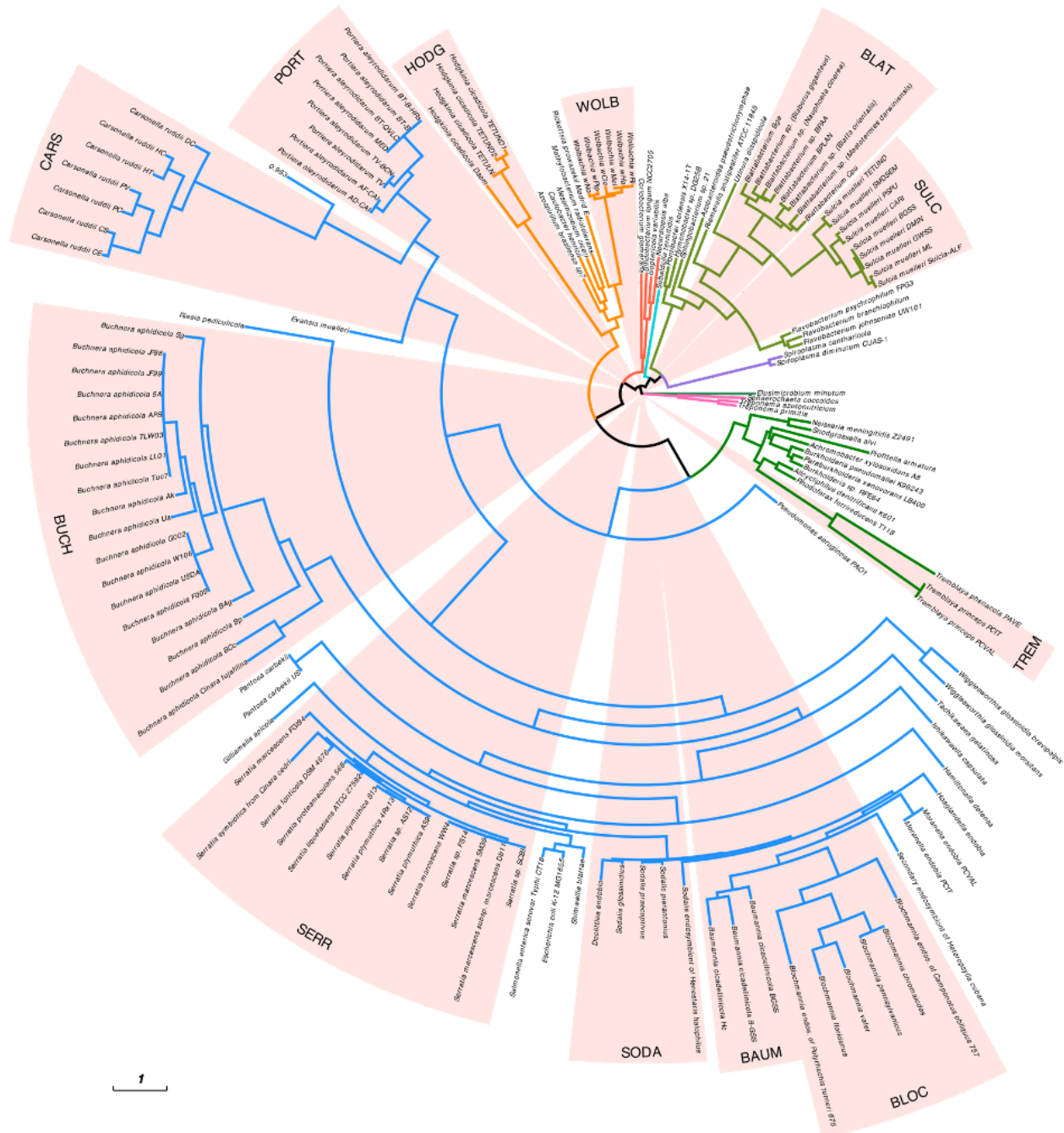


Figure C4.1 | Phylogenetic tree of the endosymbiotic bacteria of insects and free-living outgroups. Each group created for the evolutionary traits analyses is highlighted in coral with its corresponding group name (complete list in Supplementary Table C4S1). Taxonomical classes are depicted by the colors of the tree branches: blue, Gammaproteobacteria; orange,

Alphaproteobacteria; dark green, Betaproteobacteria; red, Actinobacteria; light blue, Fusobacteria; light green, Favlobacteriia; purple, Mollicutes; pink, Spirochaetia.

4.2 Metabolic networks and metabolic DAGs

We considered two different models of metabolic networks: *i*) the metabolite-based model, where nodes are metabolites; and *ii*) the reaction-based model, where nodes are reactions. For these analyses, all metabolic pathways, metabolic reactions, and compound elements of each bacterial species present in our data set were retrieved from the KEGG database (Kanehisa et al., 2016, 2017; Kanehisa & Goto, 2000). In the metabolite-based model, we inferred the metabolic network (MN) of each organism by joining their metabolic pathways considering all the interacting metabolites, meaning that if one or more metabolites of pathway A are produced or consumed by pathway B, then both pathways become linked by those metabolites. So, for each MN, the metabolites are the nodes, and the edges represent the metabolic reactions. All metabolites of a pathway are considered part of the network, regardless of their presence in any other pathway. In each MN, the edges are the reactions in all the pathways of an organism. Some reactions are reversible, so the networks are directed. When a reaction is reversible, there are two edges between the involved nodes, one in each direction. Finally, when a reaction has more than one metabolite as a substrate or product, there is one edge from each substrate metabolite to each product metabolite.

In the reaction-based model, we constructed a *reaction graph* as follows: the nodes are the metabolic reactions, and there is a directed edge from a reaction R to a reaction R' if, and only if, some product metabolite of R is also a substrate of R'. That is, the product of R and the substrate of R' have some metabolites in common. In this reaction-based model, following the methodology introduced by Alberich et al. (2017), we also constructed the *metabolic DAGs* (m-DAG for short). An m-DAG is obtained from a reaction graph by collapsing every strongly connected component in the reaction graph into a single node, called *Metabolic Building Blocks* (MBB for short). More precisely, the strongly connected components in the reaction graph are those pairs of reactions, R and R', such that there is a path from R to R' and there is also a path from R' to R in the reaction graph, i.e., generating a cycle. As a result of collapsing these *biconnected* reactions into a single node, we obtain a directed graph with no cycles, which in the graph theory setting is called a directed acyclic graph (DAG). Hence, from every reaction graph, we also constructed the corresponding m-DAG.

The two models considered here are essentially different. In the metabolite-based model, we obtain a huge and dense MN that deeply describes the metabolism, but it is difficult to handle and only the topological parameters of these networks have been determined. In the reaction-based model, we reduce the size of the network, and we deeply describe the relationships among the reactions present in the MN. In addition, by considering the m-DAGs, we were able to easily analyze and compare in detail groups of MNs of interest.

4.3 Topological parameters of the MNs

The topological parameters and centrality measures of each metabolite-based MN were calculated with the Cytoscape Network Analyzer (Shannon et al., 2003). We used 13 quantitative descriptors that have been effectively applied in other analysis (Said et al., 2004; Steele et al., 2011) which include: (1) number of nodes; (2) number of edges; (3) clustering coefficient (CI) for each MN, that is the average of the clustering coefficients for all nodes in the network (measured as the degree of interconnectivity in the neighborhood of a node); (4) connected components, where each connected component contains any pair of nodes connected by a path within the component; (5) largest connected component (or number of nodes of the biggest connected component); (6) mean of the size of the connected components; (7) shortest paths percentage, considering that the length of a path is the number of edges included in it (there can be multiple paths connecting any two given nodes, so this parameter gives the percentage of paths with the minimum length); (8) characteristic path length, that gives the expected distance between two connected nodes (the distance between two nodes is the length of the shortest path connecting them); (9) network diameter, which indicates the largest distance between two nodes; (10) the average number of neighbors, which indicates the average connectivity of a node in a network; (11) isolated nodes; (12) self-loops; and (13) multi-edge node pairs, which indicates how often neighboring nodes are linked by more than one edge (Yamada & Bork, 2009).

4.4 Comparative evolution of the genome and metabolism

Three different procedures were applied to compare genomic data with metabolisms. In the first one, we constructed a phylogenetic tree with PhyloPhlAn v3.0 (Segata et al., 2013), a software that takes advantage of whole-genome sequence data and uses 400 universal proteins to build a highly accurate and resolved phylogenomic tree. This integrated pipeline employs MUSCLE version 3.8 (Edgar, 2004) for the alignment of proteins, and FastTree version 2.1 (Price et al., 2010) to build the tree, which is generated using the ‘minimum-evolution interchanges and subtree-pruning-regrafting’ and approximate maximum likelihood joining (JTT+CAT model), as recommended by the authors for large datasets.

In the second procedure, we considered the outcome from the topological analysis and normalized each metric to derive a hierarchical clustering. In order to establish which MN metric correlates with the bacterial genome size, we used the software BayesTraits V3 (Pagel, 1994; Pagel and Meade, 2006; www.evolution.rdg.ac.uk). We did that first for the complete data set and later by taxonomical groups at the genus level, where every tree had to be rooted by an output, so we used the closely related free-living organisms. BayesTraits works by comparing two different models on the same data, one in which a correlation is assumed and one where the correlation is set to zero. We used the Continuous Random Walk model and Markov Chain Monte Carlo analysis to estimate the marginal likelihood of this comparison and then calculated the Bayes Factor (BF) between the two runs, where the model that estimated the

correlation was the complex model (the one where the correlation is not assumed). The BF interpretation is as follows: if the BF is < 2 , there is weak evidence of a correlation between traits; if it is ≥ 2 , there is positive evidence, and the higher the number, the stronger the evidence. BayesTraits software takes as input at least 50 phylogenetic trees, the result of sampling the process of the reconstruction of a Bayesian phylogeny every 100th tree, as well as a text file with the matrices of the traits data to calculate the marginal likelihood. This Bayesian phylogeny was calculated with the software Mr. Bayes (Hulsenbeck & Ronquist, 2001; Ronquist & Huelsenbeck, 2003). This software was also used for the construction of the phylogenies (and their sampling) of each taxonomical group under study since they were also needed for the evolutionary tests. The alignment file produced by PhyloPhlan was used as an input for the Bayesian phylogenetic reconstruction using the GTR model (Hulsenbeck & Ronquist, 2001), as suggested by JModelTest2 (Darriba et al., 2012). We ran two sets of 500,000 iterations, sampling and saving every 100th tree to a file, with the first 7,500 trees discarded. A majority-rule consensus of 50% of trees was constructed to examine estimates of posterior probabilities, interpreted as nodal support. Finally, we used the traits from the BayesTraits tests that resulted in positive evidence of coevolution to infer the metabolic distances between our dataset by applying a hierarchical clustering analysis, where we first standardized all the parameters individually.

For the third procedure, distances between all the reaction graphs were calculated based on the distances between the corresponding m-DAGs. The

distance between two m-DAGs described in Alberich et al. (2017) is calculated by computing the distances between their nodes, that is, their MBBs. See Alberich et al. (2017) for a complete description of these distances' calculations.

4.5 Genome vs. metabolic evolution of endosymbiotic bacteria of insects

To check for evolutionary traces embedded in the topological parameters of the MN of bacterial endosymbionts of insects, we performed co-evolution tests with the software BayesTraits (see methods section). Specifically, we wanted to know if the degradation process undergone by these bacterial genomes affects their network properties. Figure C4.1 shows a Bayesian phylogenetic tree where the groups used for the evolutionary tests are highlighted with their given name. These groups underwent evolutionary testing with Bayes Factor (see Supplementary Table C4S2). In most taxonomic classes, we found that both the network diameter and the number of nodes, as expected, are coevolving within the endosymbiotic chosen groups since they are directly proportional to the genome size. Remarkably, in six of the taxonomic genera, we were also able to detect the coevolution of the clustering coefficient and the genome size. All the MNs topological parameters are available in Supplementary Table C4S3.

In addition, as shown by (Mazurie et al., 2010), we evaluated the consistency of the pathway annotations for both of our methodologies by calculating the correlation between the number of pathways in each taxon and the logarithm of the genome size of each strain. The size of the MNs of every organism in our data set correlated with the genome size of each organism ($r^2 =$

0.95, $p = 2.2 \times 10^{-16}$). Figure C4.2 shows the MNs sizes per organism against their genome size for both methodologies, where the graph above depicts the reaction-based model, and the graph below depicts the metabolite-based model. In general, endosymbiotic bacteria have significantly smaller networks than their free-living relatives.

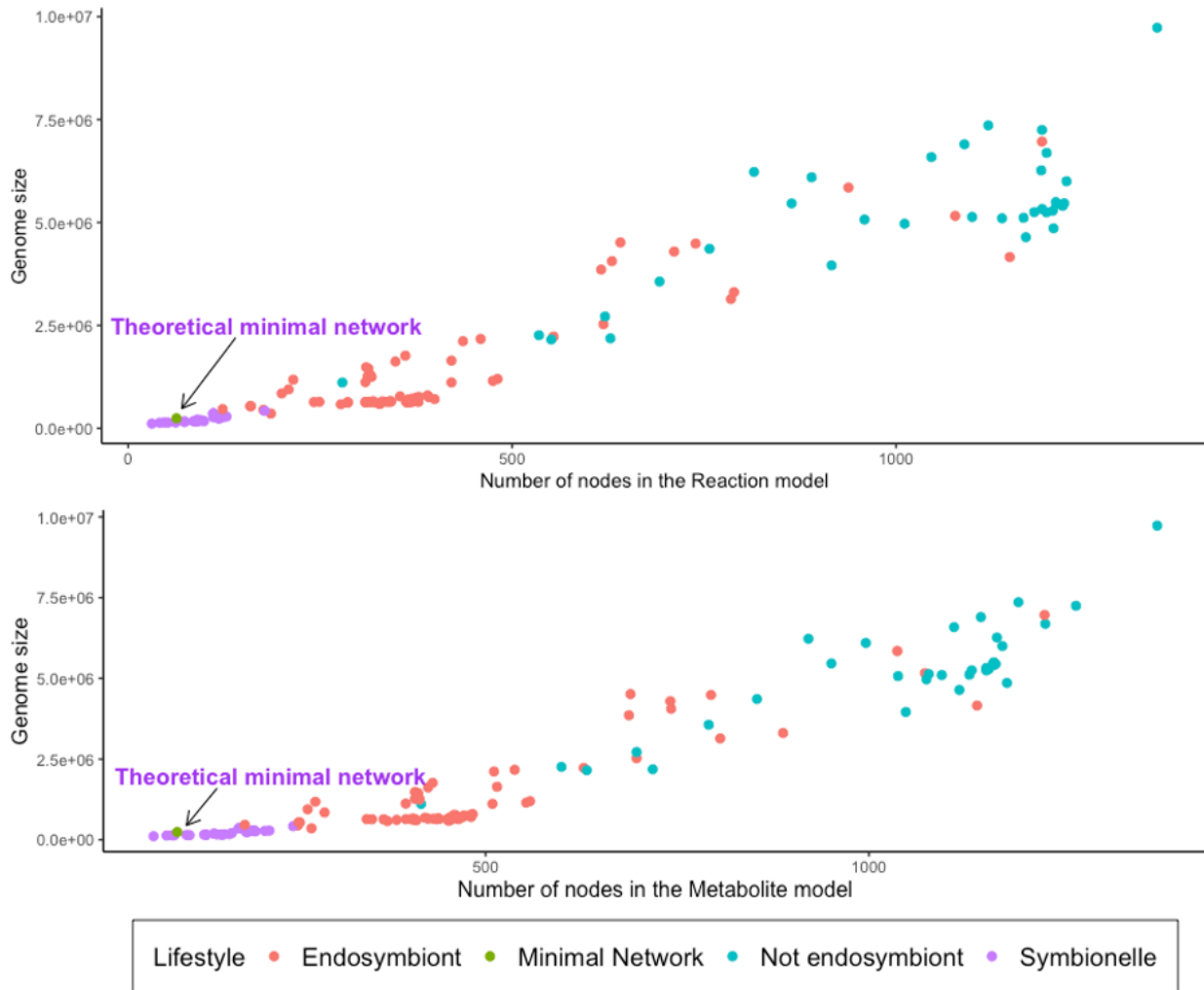


Figure C4.2 | MNs sizes per organism by lifestyle, against their genome size; above the metabolite-based model, and below the reaction-based model.

We included in this analysis a theoretical minimal MN derived from the minimal gene set proposed to sustain cellular life (Reyes-Prieto, Gil, et al., 2020). This network has a total of 98 nodes and 80 edges. Regarding its genome size calculation, we used the widely accepted concept of 1kb per gene for prokaryotic genomes (Lin Xu et al., 2006) and it was estimated to be 204,000 bp. The minimal MN was not the smallest network in our data set. Smaller networks were found in four bacterial strains (*Nasuia* NAS-ALF, *Hodgkinia* TETUND1, TETUND2, and *Tremblaya princeps* PCVAL) when looking at their number of nodes in the case of the metabolite-based model (*i.e.*, number of metabolites), and two additional *Hodgkinia* strains (Dsem and TETULN) when looking at the reaction-based model (*i.e.*, the number of reactions), with *Nasuia* NAS-ALF presenting the smallest network for both methodologies. Table C4.1 summarizes the parameters and sizes of all the organisms smaller than the minimal MN.

Table C4.1 | Sizes and parameters of the endosymbiotic bacteria of insects with metabolic networks smaller than the theoretical minimal MN. There are four genomes with fewer nodes in the reaction-based model than in the minimal network, whereas two more are added when looking at the metabolite-based model (marked with an asterisk).

Name	Class	Genome size (bp)	Number of genes	Number of nodes (metabolite-based model)	Number of nodes (reaction-based model)
<i>Nasuia deltocephalinicola</i> NAS-ALF	Betaproteobacteria	112,091	137	67	31
<i>Hodgkinia cicadicola</i> TETUND1	Alphaproteobacteria	133,698	121	84	41
<i>Tremblaya princeps</i> PCVAL	Betaproteobacteria	138,931	116	91	47
<i>Hodgkinia cicadicola</i> TETUND2	Alphaproteobacteria	140,570	140	94	52
<i>Hodgkinia cicadicola</i>	Alphaproteobacteria	150,297	170	110	62

TETULN*					
<i>Hodgkinia cicadicola</i> Dsem*	Alphaproteobacteria	143,795	169	114	62
Minimal Network	-	242,050	240	98	63

4.6 Convergence of the metabolite- and the reaction-based methods

In order to inspect the relationships between the MNs from both methods, we evaluated the distances of each by creating a hierarchical clustering for the metabolite-based method, and by assessing the distances of the reaction graphs of the reaction-based method, where a reaction graph is a step before the creation of an m-DAG as explained in methods. The resulting trees converge in the sense that all bacterial genera are localized in clusters, as they would in a phylogenetic tree (Figure C4.3) *i.e.*, bacteria from the same genera are grouped together. However, the signal is lost when looking at higher levels in taxonomy because, evidently, there is no information to resolve these branches, but the convergence in the cluster's positioning is an indication of the remnants of how organisms have evolved. The signal is picked up even in endosymbiotic organisms, whereby becoming intracellular and having lost a significant number of genes, they still possess enough information within their genome, and these results show that the evolutionary signatures can be found through the analysis of MNs.



Figure C4.3 | Distance trees resulting from the reaction- and the metabolite-based methods.

To the left, the distances resulting from the reaction graphs used to create the m-DAGs (see methods). To the right, a hierarchical clustering of the metrics derived from the topological analysis of each MN. The colored legend shows the genera of endosymbiotic bacteria of insects and free-living organisms.

4.7 Core m-DAG and pan-mDAG of endosymbiotic bacteria of insects

To evaluate the differences between the MNs of the endosymbiotic bacteria included in our data set, we analyzed their m-DAGs in depth. Namely, we considered all their metabolic building blocks, which are the nodes of an m-DAG, and we looked at whether they appeared or not in all m-DAGs, ultimately calculating the pan-mDAG of the data set, in analogy to the pangenome (Medini et al., 2005; Tettelin, 2020), and the core m-DAG of endosymbiotic bacteria of insects. We obtained MMBs that are present in all endosymbiotic bacteria, concluding that the core genome of endosymbiotic bacteria of insects only includes the reactions needed for the synthesis of DNA and RNA (Figure C4.4).

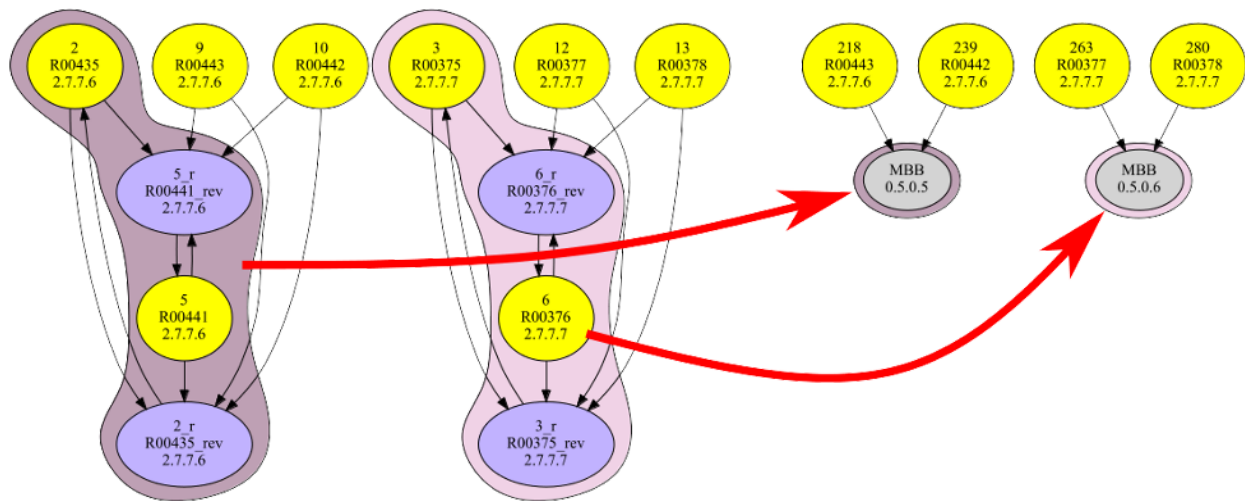


Figure C4.4 | Core MN, consisting of the reactions needed for the synthesis of DNA and RNA, of 101 genomes of endosymbiotic bacteria of insects and its corresponding m-DAG. To the left, is the reaction graph with 12 nodes. The IDs inside the yellow circles are KEGG IDs. The purple circles are their reverse reactions. To the right, the m-DAG (the simplified version of the reaction

graph) of the same core, with only 6 nodes (the contraction is shown with the violet background). The gray nodes denote the metabolic building blocks; their IDs, obtained during the calculation, are described in a text output format.

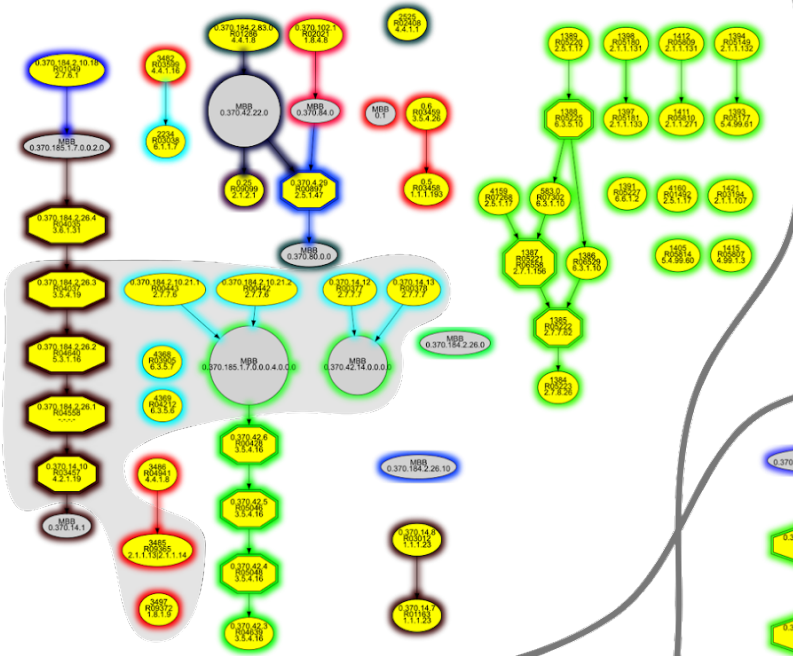
It was surprising to see that not even the purine and pyrimidine synthesis pathways, which are subunits of nucleic acids and precursors for the synthesis of nucleotide cofactors, are included in this core genome. As for the pan-mDAG, the first step of the methodology is to create a reaction graph (explained in methods). Its reaction graph consists of 2,964 reactions (nodes), where 598 are reversible reactions, and 1,883 compounds (edges) (Supplementary Figure C4S1). Next, by constructing the pan-mDAG these reactions are merged into 1,081MBBs (Supplementary figure C4S2). This pan-mDAG is not easy to read but includes every MBB of the organisms included in our data set and can be useful to search for specific reactions of interest or to explore if determined metabolic pathways are carried out by any or some endosymbiotic bacteria of insects.

4.8 Smallest and biggest MNs of symbiotic bacteria of insects

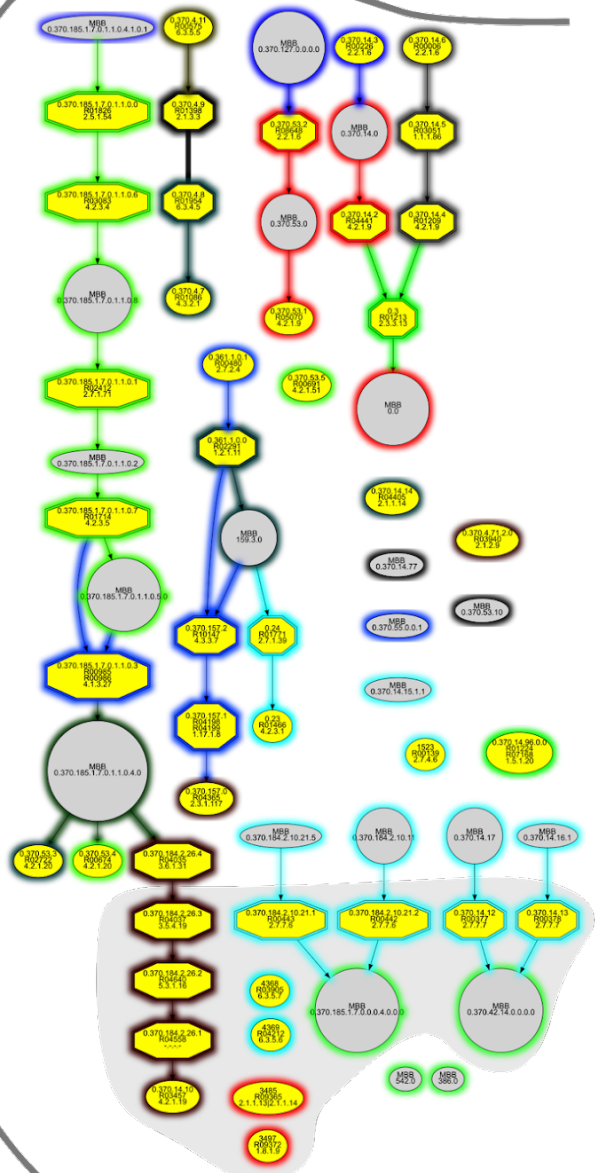
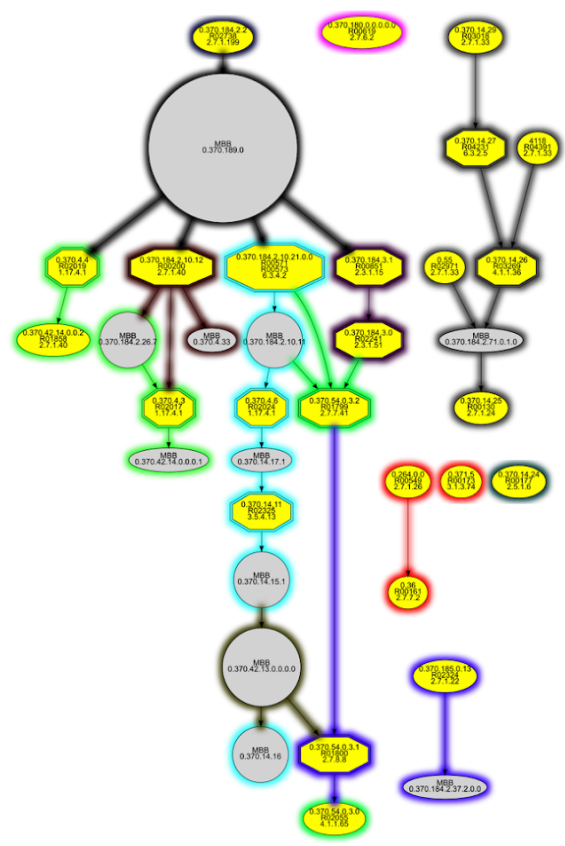
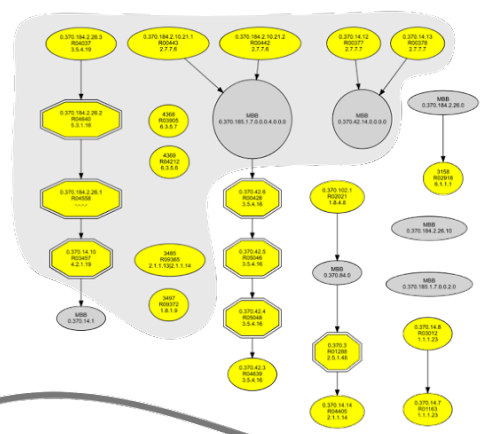
We also examined in detail the m-DAG of the six smallest genomes to account for the information absent in the proposed minimal MN and present in the endosymbiotic MNs, and vice versa. Figure C4.5 shows the m-DAG of the minimal MN and the smallest endosymbiotic bacteria. At first glance, we observe that the smallest endosymbiotic bacteria have extremely disrupted networks with respect to the minimal MN, which has only seven connected components. In fact, the m-DAG from the minimal metabolic reaction graph has 36 nodes, 11 of which are MBBs with more than one reaction, and seven connected components.

The m-DAG from *Nasuia* NAS-ALF has 29 nodes, seven of which are MBBs with more than one reaction, and 12 connected components. The m-DAG from *Tremblaya* has 41 nodes, 13 of which are MBBs with more than one reaction, and 14 connected components. The m-DAG from *Hodgkinia* has 57 nodes, of which 11 are MBBs with more than one reaction and 24 connected components.

Hodgkinia



Nasuia



Minimal

Tremblaya

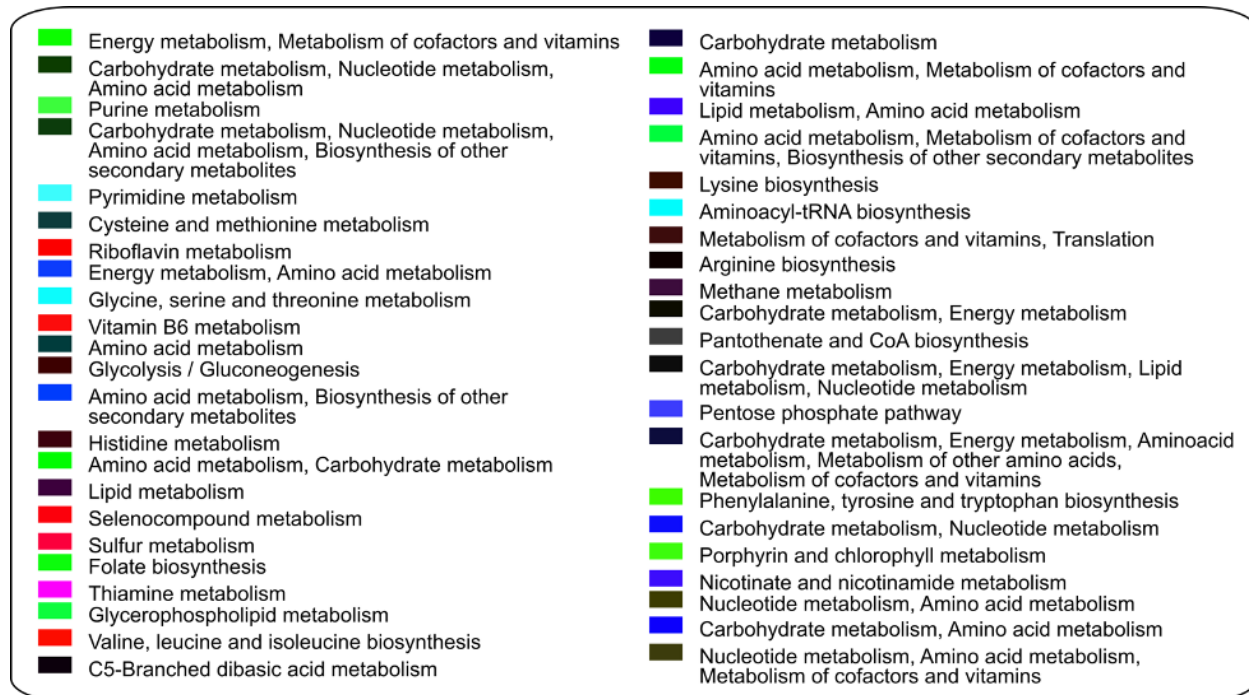


Figure C4.5 | Relation between the smallest m-DAGs under study and the minimal MN. The gray circles are MBBs that have two or more reactions, the yellow circles are single reactions, and the numbers inside the circles are ids created by the metaDAG software with information on the reactions included in each MBB. The core of all m-DAGs is denoted with a gray background.

Regarding the reactions, we first observed that 16 reactions are shared among the considered endosymbionts distributed in 14 MBBs, 12 of them with a single reaction. These reactions participate in the purine and pyrimidine metabolism, the seleno-compound metabolism involved in various redox processes, and the biosynthesis of several aminoacyl-tRNAs. Interestingly, despite sharing those reactions, the topology of their m-DAGs is quite different, since they share few MBBs, an indication that those three symbionts have different biological functions. Figure C4.6 shows a graph with the total set of MBBs shared by the three symbionts and the minimal MN. It is important to

denote that for *Hodgkinia*, we used the pan-mDAG of the four genomes that resulted smaller than the minimal MN (TETUND1, TETUND2, TETULN, and Dsem).

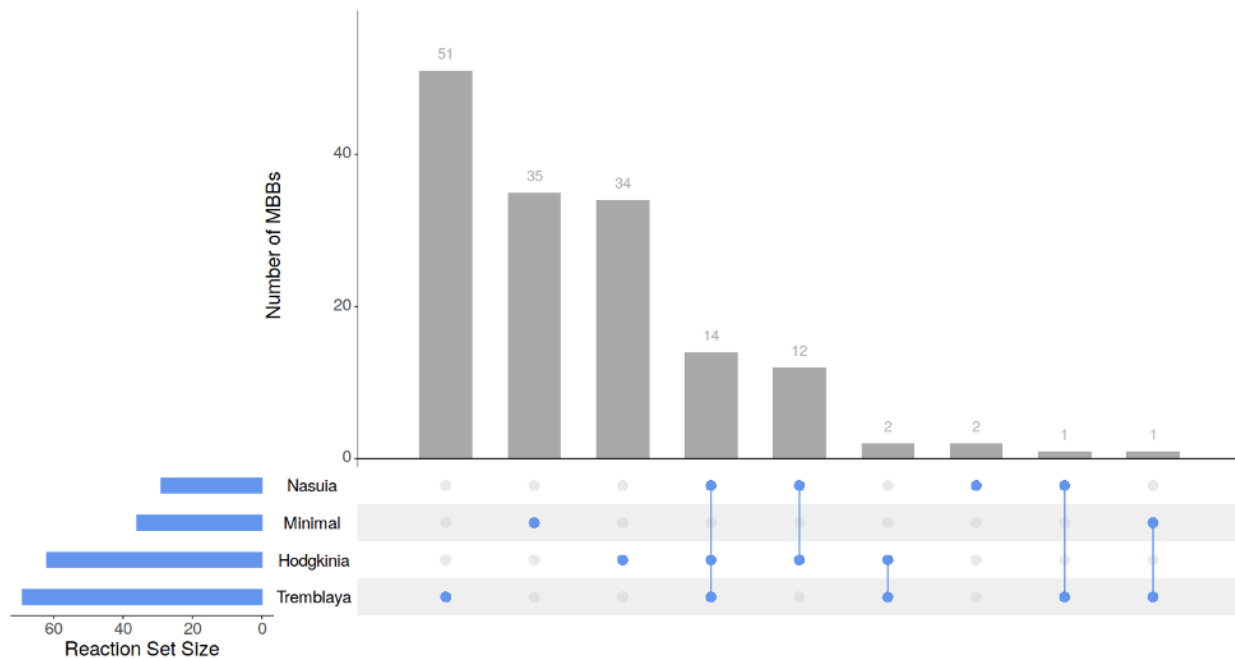


Figure C4.6 | Shared metabolism as MBBs among the m-DAGs of *Nasuia* NAS-ALF, *Tremblaya princeps* PCVAL, *Hodgkinia* (considering the pan-mDAG of strains TETUND1, TETUND2, TETULN, and Dsem), and the minimal MN. The blue dots represent the shared MBBs between the corresponding groups.

We can observe there that *Tremblaya* has 51 MBBs that are not shared with the others, the minimal MN has 35, *Hodgkinia* has 34, and *Nasuia* has only two. This result is consistent with the role they play in their respective hosts. *Nasuia* is part of an endosymbiotic consortium with *Sulcia* in the aster leafhopper *Macrostoteles quadrilineatus*. Together, they can synthesize the 10 essential amino acids for their host, but *Nasuia* alone is only capable of synthesizing two amino

acids. It also retains genes for DNA replication, transcription, and translation, but has even lost the genes needed for ATP synthesis through oxidative phosphorylation, a basic pathway in most bacteria (Bennett & Moran, 2013). It presents one of the smallest genomes found so far, with only 112 kb. Next, *Tremblaya princeps* PCVAL participates in a nested endosymbiosis shared with the gammaproteobacterium *Moranella endobia* inside the citrus mealybug *Planococcus citri*. With a 139-kb genome, which contains only 110 functional protein-coding genes, 43 of which correspond to ribosomal proteins, and 23 pseudogenes, *Tremblaya* has retained the ability to synthesize most of the essential amino acids for its insect host (López-Madrugal Sergio et al., 2011). Finally, *Hodgkinia*, with genomic sizes ranging between 133 and 144 kb, is a partner of a symbiotic consortium with *Sulcia* and plays an important role for its host, but it only has the capability to synthesize the essential amino acid histidine. It also has a few genes involved in the synthesis of other amino acids, but no other complete pathway. The most interesting genes it retains are devoted to the synthesis of cobalamin (vitamin B₁₂), described as an essential cofactor in an alternative pathway that has been preserved in this symbiotic consortium for synthesizing the essential amino acid methionine (McCutcheon et al., 2009b).

Next, we investigated the reactions present in the minimal MN but not in the endosymbiotic MNs, which accounted for 35 MBBs. They participate in the biosynthetic pathways of the non-essential amino acids: cysteine and glycine, and the two essential ones, methionine and threonine. Also, in the glycerolipid and glycerophospholipid metabolism, glycolysis, the metabolism of many

vitamins and cofactors (nicotine and nicotinamide, pantothenate and CoA, thiamine (B₁), riboflavin (B₂), and pyridoxal (B₆), as well as the synthesis of purines and pyrimidines. Even though these pathways are considered essential for a minimal cell lifeform (Gil, Pérez-Brocal, et al., 2006), assuming that an independent minimal cell has to rely on its own gene products (i.e., it can import metabolites, but not functional proteins), it is not surprising that they are not present or complete in the endosymbiotic bacteria of insects with the smallest genomes.

We can conclude that the metabolism of the smallest endosymbiotic bacteria of insects is completely disconnected, which is consistent with the notion that the metabolisms of the hosts and other bacterial partners (in the case of consortia) generate a metabolically functional connected network. Thus, the intracellular environment in which these bacteria live provides access to metabolites and even some functional proteins targeted directly to the cytoplasm of the specialized eukaryotic cells in which they live (Nakabachi et al., 2014; Sloan et al., 2014).

Regarding the biggest MNs, for instance, *Nocardiopsis alba* ATCC BAA-2165, we could observe that its m-DAG has a huge connected component which, at first glance, looks almost topologically identical to its close free-living relative *Bifidobacterium longum* NCC2705 (Supplementary figures C4S2 and C4S3). This result is congruent with the description of the modern symbiosis between *Nocardiopsis* and the honeybee *Apis mellifera* (Qiao et al., 2012). This bacterium has been frequently isolated from honeybees' guts in different studies

but does not seem ubiquitous in this host. The very large connected component corresponds to the central metabolism of the cell, while the complete network, which also has several isolated connected components, has 1,220 nodes. A large number of nodes and the small number of connected components is an indicator of the complexity of its MN, which in turn, can be a signal of the capability of this organism to have an independent life and the low probability of its essentiality for its host. Furthermore, we even found pathways associated with inhibitory bioactivities against bacterial pathogens, as described by (Patil et al., 2010; Promnuan et al., 2009). Although we did not go any further into the analysis and our main interest remains focused on endosymbionts with the most reduced genomes and the most eroded MNs, we wanted to highlight that this methodology is useful for all complexity levels of bacterial genomes.

4.9 Conclusions

We have analyzed the metabolism of endosymbiotic bacteria of insects with two different models, the metabolite-based model, and the reaction-based model. The resulting dendrograms are not identical because they contemplate different characteristics of the metabolisms: relations between compounds for the metabolite-based model, and relations between reactions for the reaction-based one. Nonetheless, with both methodologies, the organisms' clustering correlates with their taxonomic genus, which points to a considerably conserved genomic core between them.

As for the reaction graphs, we have a more detailed tool, the metaDAG, which is used to analyze the topologies of the networks. Here it is evident that the metabolic networks of endosymbiotic bacteria are completely disconnected in comparison to the topologies observed in the free-living organisms' networks, which present much greater connected components. That is, the reactions in free-living organisms are much more interconnected than the reactions in endosymbiotic bacteria. In biological terms, endosymbionts with these metabolic networks are not capable of carrying out the necessary biological functions that free-living organisms are since endosymbionts are favored by the complementary metabolism of co-symbionts and their respective hosts, while free-living organisms must be prepared for a changing environment and external stimuli such as temperature changes, stress, attacks by other organisms, and other adverse factors.

In the metabolite model, we observe coevolution among several topological parameters. Thus, the genome size of the organisms correlates with the number of nodes, the diameter, and the clustering coefficient of the networks. It should be noted that in this model, the number of edges increases considerably, but the diameter and the clustering coefficient are not affected by this growth, nor the number of nodes.

By comparing both models, we conclude that it is advisable to use the reaction graph model to handle large volumes of data because this model considers reactions as nodes and their relationships as edges and generates the

'meta-model' of the metabolism, which is usually much simpler and facilitates its analysis and visualization.

In the matter of partially answering the famous question 'Does metabolic evolution meet genomic evolution?', our results show that, in the case of endosymbiotic bacteria of insects, there is a high correlation between genomic evolution and the similarity of metabolic networks, even though these organisms live isolated within their host. Therefore, we conclude that, due to their interconnected and reduced metabolic networks, these symbiotic relationships generate a unique signature, an evolutionary footprint that is probably comparable only with pathogens and other host-dependent organisms.

VI. Conclusions

1. The idea of a symbionelle has been developed to account for endosymbionts that are unable to attain a minimal gene set. They represent a novel category due to their evolved genomes with so few genes that they cannot carry out the three essential functions of modern cells without the assistance of a host and/or additional co-symbionts. Although they each developed in very different evolutionary scenarios—organelles before multicellular life, and symbionelles specifically during insect evolution—these symbionelles show evolutionary convergence with organelles exhibiting evident and significant similarities and differences.
2. The increase in microbial genome publications has forced the design of numerous tools to categorize and better arrange data to fully appreciate them for their study and learn as much as possible from them. SymGenDB is a highly curated and comprehensive systems database, which provides access to the complete catalog of fully sequenced and annotated genomes of symbiotic organisms and their genomic sequences, metrics, and compressed metabolic networks which can be useful for research and reduce the time-consuming search for symbiotic relationships and organism features.

3. SymGenDB is part of a large genome informatics resource including multiple microbiological datasets. It is an intensely integrated and accessible data repository with the best quality and most complete consensus and experimental views of host–symbiont connections.
4. The modeling of metabolic Directed Acyclic Graph (m-DAG) based on minimal metabolisms can be a first approach for the synthesis and manipulation of minimal cells. We modeled a network depicting a minimal metabolism inferred from a minimal genome needed for life and compressed it as an m-DAG in order to better visualize its topology and to find critical reactions to maintain the network's connectivity.
5. We have also compared this minimal m-DAG to those of the smallest natural genome known until now and a synthetic minimal cell created in the laboratory, as a proof of concept that this kind of analysis can be used for comparative metabolomics.
6. The compilation of compounds and reactions that we present in our model network can easily be extrapolated to any phylogenetically diverse bacteria of interest since we did not focus specifically on genes.

7. Two models—metabolite-based and reaction-based—have been used to study insect endosymbiotic bacteria metabolism. Because the metabolite-based model considers compound relations and the reaction-based model considers reaction relations, the dendrograms are different. Both methods cluster species by taxonomic genus, implying a highly conserved genomic core, indicating the existence of phylogenetic convergence.
8. The MetaDAG analyzes network topologies from reaction graphs. The metabolic networks of endosymbiotic bacteria are fully disconnected, unlike those of free-living organisms, which have more integrated components. Endosymbionts with these metabolic networks need the complementary metabolism of co-symbionts and their hosts, while free-living organisms must be prepared for a changing environment and external stimuli like temperature changes, stress, attacks by other organisms, and other adverse factors.
9. Topological parameters coevolve in the metabolite model. Hence, organism genome size corresponds with network node count, diameter, and clustering coefficient. It should be noted that in this model, the number of edges increases considerably, but the diameter and the clustering coefficient are not affected by this growth, nor the number of nodes.

10. By comparing these models, we conclude that the reaction graph model is best for handling huge amounts of data because it treats reactions as nodes and their interactions as edges and generates the "meta-model" of the metabolism, which is simpler and easier to analyze and visualize.

11. Our data reveal that genetic evolution and metabolic network similarity are highly correlated in insect endosymbiotic bacteria, even though they reside isolated within their host. Hence, these symbiotic partnerships provide a unique evolutionary footprint, similar only to pathogens and other host-dependent organisms, due to their integrated and decreased metabolic networks.

VII. Resumen en castellano

Introducción

Heinrich Anton de Bary acuñó el término "simbiosis" para describir la convivencia de organismos con nombres diferentes (de Bary, 1879). El interés inicial en el estudio de estas relaciones se debía a que los microbios se consideraban equivalentes a enfermedad, pero el enfoque cambió lentamente a medida que se descubrieron más relaciones no patógenas en todos los reinos de la vida, reconociéndolos como unos de los principales factores que impulsan la evolución (Douglas, 2014). Los investigadores ahora están de acuerdo en que la simbiosis es una interacción íntima entre especies, independientemente de las consecuencias para los organismos implicados, de modo que incluye tanto asociaciones mutualistas como parasitarias y comensales (Saffo, 1993). La simbiosis se ha observado a lo largo del árbol de la vida en muchas formas, mostrando diferentes grados de dependencia (Figura 1 de la introducción en inglés de este trabajo). Estas relaciones son fundamentales para el funcionamiento de los ecosistemas y han sido cruciales para el desarrollo de la complejidad biológica. Los simbiositos permiten que sus hospedadores vivan en nichos naturales que de otro modo no estarían disponibles para ellos al suministrarles nutrientes, brindarles protección e incluso ayudarlos a utilizar fuentes de energía alternativas (Kleiner et al., 2012; Moran et al., 2003). Por ejemplo, las plantas tienen hongos micorrízicos que controlan la cantidad de micro y macronutrientes que pueden obtener del suelo (Bati et al., 2015), y

bacterias fijadoras de nitrógeno que transforman el nitrógeno del aire en amonio para el beneficio de su hospedador (Rogel et al., 2011); en humanos, se estima que alrededor del 50% de las células de nuestro cuerpo son microbianas (Sender et al., 2016). La simbiosis no se limita a las interacciones eucariota-procariota, sino que también incluye interacciones eucariota-eucariota (Allemand & Furla, 2018; Martinson, 2020; Scannerini et al., 2013; Wrede et al., 2012). Podemos decir que la convergencia de la aparición de la simbiosis a lo largo de la evolución, en diferentes formas y ecosistemas, pone de manifiesto la importancia de coexistir, compartir y colaborar, permitiendo la estabilidad de las comunidades a lo largo del tiempo.

Los insectos constituyen el grupo de animales más exitoso de la Tierra en términos de biomasa y biodiversidad, con una cifra estimada de cuatro a diez millones de especies existentes (Miller et al., 2002). Se encuentran en todos los ecosistemas terrestres, incluidos los más extremos, y su éxito evolutivo puede explicarse en gran medida por su capacidad para interactuar con bacterias simbióticas (Dimijian, 2000). La endosimbiosis es la regla más que la excepción en los insectos, ya que aproximadamente la mitad alberga bacterias simbióticas, y aproximadamente el 20% depende directamente de bacterias endosimbióticas para asegurar su desarrollo y reproducción (Batra & Buchner, 1968; Douglas, 1989; Gil & Latorre, 2019; Medina et al., 2020; Moran & Telang, 1998). Pueden existir millones de insectos en un solo acre de tierra, desempeñando un papel importante en la descomposición del material vegetal y animal, y constituyen una importante fuente de alimento para muchos otros animales (Hoffmann &

Frodsham, 1993). Sin embargo, tanto la cantidad como la diversidad de insectos están disminuyendo en todo el mundo debido a la pérdida de hábitat, la contaminación y el cambio climático (Raven & Wagner, 2021).

Las asociaciones simbióticas en insectos se pueden definir según la localización del simbiote, los efectos de su relación simbiótica y la interrelación de los organismos. La clasificación más común se basa en el nivel de integración del simbiote bacteriano en la fisiología del hospedador, distribuyéndolos entre endosimbiontes obligados y facultativos (Guo et al., 2017; Wilkinson et al., 2007). Los endosimbiontes facultativos pueden volverse obligados al cabo de un tiempo y tener diferentes resultados finales en función de la dependencia hospedador-simbiote (Husnik & Keeling, 2019). Pueden transmitirse horizontal o verticalmente, mientras que los simbiotes obligados se transmiten exclusivamente de manera vertical, generalmente por la transferencia transovárica de la madre a la prole (Szkwarzewicz & Michalik, 2017).

Tanto los simbiotes facultativos como los obligados suponen un peaje de mantenimiento en sus anfitriones, pero también conducen a muchas ventajas ecológicas. Los endosimbiontes bacterianos les proporcionan funciones beneficiosas que frecuentemente derivan en su adaptación a diversos ecosistemas (Sudakaran et al., 2017, Mitter et al., 1988, Raffa et al., 2008), como la resistencia al estrés (Heyworth & Ferrari, 2016), la protección contra otros organismos (Kaltenpoth et al., 2005; Oliver et al., 2003), la resistencia a los insecticidas (Kikuchi et al., 2012) y, en una importante proporción de casos estudiados, el

suministro de nutrientes esenciales, como aminoácidos y vitaminas (Baumann, 2005; Douglas, 1998, 2016; Moran et al., 2003).

Al pasar de una forma de vida libre a una forma de vida intracelular, las bacterias endosimbióticas pasan por un proceso denominado síndrome de reducción genómica, generalmente volviéndose completamente dependientes de su hospedador (Latorre & Manzano-Marín, 2017; Toft & Andersson, 2010; Fisher et al., 2017). Esta reducción genómica deriva en la pérdida de diversidad metabólica e incluso de funciones tan básicas como la reparación del DNA (Wernegreen, 2017). Otra característica del fenómeno de la endosimbiosis obligada es que estas bacterias viven dentro de células especializadas del hospedador llamadas bacteriocitos (Alarcón et al., 2022). También se ha descrito la aparición natural de más de un linaje de bacterias endosimbióticas en un insecto hospedador como co-simbiontes, ya sea en su propio bacteriocito específico o compartido, donde cada miembro de la simbiosis desempeña su propio papel esencial (McCutcheon & Moral, 2010; Nakabachi et al., 2013), comprobándose inclusive la existencia de complementación metabólica entre co-simbiontes, un fenómeno de vital importancia para la supervivencia del hospedador (Gosalbes et al., 2008; Ponce-de-Leon et al., 2017). Por último, en algunos casos se han observado reemplazos de un endosimbionte obligado por otro (Gil & Latorre 2019; Sudakaran et al., 2017).

Todos los eucariotas son el producto de una simbiosis entre un eucariota primitivo y una bacteria intracelular, específicamente perteneciente al orden

Rickettsiales, que evolucionó para dar lugar a la mitocondria, orgánulo esencial para la producción de la energía necesaria para la supervivencia y funcionamiento de una célula eucariota (Margulis, 1970; Williams et al., 2007; Gabaldón, 2018). La evidencia indica que otro evento simbiótico importante entre un eucariota y una cianobacteria dio origen al cloroplasto, el orgánulo en algas y plantas que es responsable del proceso de fotosíntesis (Bonen & Doolittle, 1975; McFadder, 2001; Sánchez-Baracaldo et al., 2017). Los orígenes de estos orgánulos como simbiontes bacterianos se consideran entre las principales transiciones evolutivas en la historia de los seres vivos modernos.

De gran interés es que algunos de los atributos de los endosimbiontes mencionados en la sección anterior, son comunes a los orgánulos. Los endosimbiontes obligados intracelulares de insectos se han vuelto imprescindibles para el mantenimiento de la eficacia biológica y desarrollo de su hospedador en sus condiciones ambientales, y se encuentran fijadas en las poblaciones (Douglas, 1989; Moran & Telang, 1998). Al encontrarse secuestrados dentro de los bacteriocitos, tienen una transmisión vertical estricta, y la mayoría de los endosimbiontes estudiados también tienen una función metabólica esencial, ya que complementan la dieta de los insectos proporcionando aminoácidos esenciales y/o vitaminas, o reciclando nitrógeno a cambio de suministro de energía y otros nutrientes (Douglas, 2016; Patiño-Navarrete et al., 2014; Ponce-de-Leon et al., 2017; Skidmore & Hansen, 2017). El proceso de reducción del genoma deriva en ocasiones en casos extremos, como el de '*Candidatus* Nasonia deltocephalinicola' NAS ALF, con un genoma de 102 kb

(Bennett & Moran, 2013), de longitud comparable al genoma de algunos orgánulos. Estos genomas extremadamente reducidos son masivamente densos en genes, incluso con genes superpuestos, con variación de la longitud de ortólogos y pérdida de proteínas accesorias grandes; la mayoría de ellos tienen además un sesgo hacia las bases A-T (Figura 2 de la introducción en Inglés) (Gil et al., 2002; Kenyon & Sabree, 2014; McCutcheon & Moran, 2012; Moran & Bennett, 2014; Nowack, 2014). Al convertirse en endosimbióticas, las bacterias se encuentran en un entorno altamente estable, con presión selectiva que sólo actúa sobre los genes que son beneficiosos para la interacción simbiótica. Las mutaciones que son perjudiciales o neutras para la interacción específica, y también aquellas que afectan a genes de los que la bacteria ya no depende debido a la ausencia de cambios en su nuevo nicho, conducen a una lenta pseudogenización o se eliminan mediante grandes deleciones y reordenamientos cromosómicos (Moran & Mira, 2001; Sabater-Muñoz et al., 2017). Además, entre estas pérdidas de material genético, se encuentran pérdidas significativas de genes informacionales como genes de reparación, factores de traducción, tRNAs, rRNAs, genes de modificación de RNA y proteínas ribosómicas (Bennett & Moran, 2013; Garzón et al., 2022; Husnik & McCutcheon, 2016; McCutcheon et al., 2009b). Finalmente, el confinamiento espacial de las bacterias asegura que no haya posibilidad de recombinación o transferencia horizontal de genes, perdiendo la posibilidad de intercambiar material genético con otros linajes bacterianos no relacionados (McCutcheon & Moran, 2012).

Las relaciones simbióticas implican la integración funcional a largo plazo de muchas funciones. Se ha observado el intercambio de componentes esenciales entre el hospedador y el endosimbionte (aminoácidos, vitaminas, ATP, azúcares, nucleótidos, etc.) (Douglas, 2016; Duncan et al., 2014; Hansen & Moran, 2011; McCutcheon & Moran, 2012), y se han descrito escenarios complejos donde algunas rutas metabólicas pueden estar incompletas en ciertos endosimbiontes, pero compartidas entre los socios simbióticos. Por ejemplo, dentro de los insectos que se alimentan de savia y que poseen co-endosimbiontes, se ha descrito la complementación metabólica de varias vías biosintéticas entre los copartícipes de la relación, como las de la biosíntesis de triptófano, biotina, tetrahidrofolato e incluso el metabolismo energético. Pérez-Brocal et al., 2006; Ponce-de-Leon et al., 2017; Van Leuven et al., 2014). Estos ejemplos hacen evidente que todos los participantes de la simbiosis contribuyen a una sola vía, lo que significa que el nivel de integración debe involucrar el transporte de enzimas y/o metabolitos a través de membranas intermedias, aunque la mayoría de estos mecanismos aún no están claros.

Otro atributo comúnmente requerido para la aparición de los orgánulos es que hayan transferido genes a su hospedador y se vuelvan dependientes de un sistema de importación específico para volver a recuperar sus productos proteicos (Cavalier-Smith & Lee, 1985; Keeling & Archibald, 2008); el endosimbionte, de forma similar a un orgánulo, depende de su hospedador para su propio funcionamiento. Este fenómeno ha sido comprobado experimentalmente en *Buchnera*, el cromatóforo y *Kinetoplastibacterium* (Alves et

al., 2013; Morales et al., 2016; Nowack et al., 2016; Sloan et al. , 2014). También se ha demostrado que estos organismos dependen de proteínas codificadas por los genomas del hospedador pero que tienen un origen bacteriano, si bien no necesariamente procedentes de su endosimbionte; tal es el caso observado en los linajes endosimbióticos de *Nasuia*, *Sulcia*, *Tremblaya*, *Buchnera*, *Carsonella*, *Portiera*, entre otros. (Kelly, 2021; López-Madrigal Sergio et al., 2011; Mao et al., 2018; Nakabachi et al., 2014; Nikoh et al., 2010; Sloan et al., 2014).

Debido a los avances tecnológicos de los recientes años, en especial en el área de la secuenciación masiva de ácidos nucleicos, las bases de datos informáticas se han convertido en herramientas indispensables para organizar y acceder fácilmente a la información biológica. Múltiples bases de datos ofrecen información sobre los genes y genomas de los organismos, como NCBI (Wheeler et al., 2007); <https://www.ncbi.nlm.nih.gov/>) o ENA (Leinonen et al., 2011); <https://www.ebi.ac.uk/ena/browser/home>), que han crecido exponencialmente desde el comienzo de la era de la secuenciación.

Las bases de datos biológicas son colecciones de datos organizados de forma sistemática a los que se puede acceder, y se pueden gestionar y actualizar rápidamente. Suelen utilizar marcos de gestión relacional y el lenguaje de consulta estándar (SQL), que permite tanto la definición como la recuperación de datos. Hay tres tipos de bases de datos biológicas: primaria, secundaria y compuesta o terciaria. Las bases de datos primarias archivan los resultados experimentales enviados por los científicos, y los datos archivados no están

seleccionados (Selzer et al., 2018). Las bases de datos secundarias comprenden datos derivados de los resultados del análisis de datos primarios, y la información almacenada en ellas está altamente seleccionada. Son ejemplos de este tipo InterPro ((Hunter et al., 2009); <https://www.ebi.ac.uk/interpro/>) y UniProt ((UniProt Consortium, 2019); <https://www.uniprot.org/>), bases de datos que incluyen secuencias de familias, motivos, dominios, e información funcional de proteínas. Las bases de datos compuestas o terciarias toman datos de las bases de datos primarias y luego los integran en base a ciertas condiciones, y los datos están altamente seleccionados.

Por lo que sabemos, al comienzo de los estudios presentados en esta tesis no existían bases de datos secundarias o terciarias disponibles públicamente para bacterias endosimbióticas de insectos u otros organismos simbióticos. Estos grandes repositorios de datos serían de gran interés para los científicos de diferentes áreas de investigación que trabajan en genómica simbiótica.

Conocer, describir y comprender cada proceso molecular en una célula viva nos daría una idea de los principios fundamentales de la vida y abriría muchas posibilidades para las ciencias aplicadas, incluido el diseño y la creación de organismos artificiales. Dado que esto aún no es posible, las células mínimas naturales pueden guiarnos en este empeño, ya que contienen solo la cantidad mínima de información necesaria para su supervivencia. Los simbioses intracelulares obligados con una reducción extrema del genoma se encuentran entre los principales organismos estudiados para ayudar en el objetivo de

sintetizar con éxito células mínimas. Con esto en mente, Gil y colaboradores (2004) definieron en nuestro grupo el núcleo de un conjunto de genes mínimo necesario para el funcionamiento de una célula bacteriana. Realizaron un análisis exhaustivo de intentos computacionales y experimentales para definir un genoma mínimo mediante la comparación de bacterias endosimbióticas mutualistas, parásitas y de vida libre. Glass y colaboradores, del grupo de Craig Venter, también incursionaron en este tema y presentaron su lista de genes esenciales para bacterias en 2006 (Glass et al., 2006), teniendo en cuenta el trabajo mencionado anteriormente. Posteriormente, ambas listas de genes fueron exploradas por Gabaldón y colaboradores para proponer una red que representa un metabolismo mínimo inferido necesario para la vida (Gabaldón et al., 2007).

Representar el metabolismo por métodos computacionales convencionales, rastreando y mapeando todos los componentes de un genoma en mapas de vías metabólicas, supone una gran complejidad que dificulta su análisis y visualización. Para aminorar este problema, Alberich y colaboradores desarrollaron una metodología denominada MetaDAG (Alberich et al., 2017, la herramienta para uso público está siendo preparada para su libre disposición en web) basada en la representación gráfica para estudiar la solidez, la modularidad y la conectividad de las redes metabólicas. La metodología MetaDAG proporciona toda la información de interés dentro de las redes pero también reduce su tamaño para facilitar su análisis, visualización y comparación con otras redes; por lo tanto, se puede aplicar a organismos simbióticos específicos, a metabolismos mínimos teóricamente definidos como el resultante de Gabaldón

(2007), y a análisis conjuntos a gran escala de un conjunto de bacterias endosimbióticas de interés.

Con la gran disponibilidad de genomas completamente secuenciados y anotados de bacterias endosimbióticas de insectos, así como aplicando diferentes métodos y herramientas a dichos datos, podemos comenzar a dilucidar algunas preguntas más específicas sobre el proceso de integración simbiótica dentro de los eucariotas, y cómo ha tenido lugar su evolución comienza a ser un poco más clara (Borenstein & Feldman, 2009; Martínez-Cano et al., 2014). Al reconstruir las redes metabólicas a partir de los genomas de los endosimbiontes, podemos comprender de manera integral las propiedades funcionales de los organismos por sí mismos y luego, con metabolómica comparativa, revisar las variaciones y diferencias entre organismos en diferentes contextos evolutivos (por ejemplo, estar más o menos tiempo involucrados en una endosimbiosis). Esta comparación metabolómica se enfoca al estudio de las diferencias entre los nodos y bordes de las redes metabólicas de cada organismo, o una colección de organismos; estas diferencias que se dan a través de la evolución, vienen condicionadas por las presiones selectivas que actúan sobre el genoma del organismo, debido a sus características topológicas (Yamada & Bork, 2009).

Para las bacterias endosimbióticas de los insectos, el contexto evolutivo está condicionado por la dependencia de su hospedador, lo que incluye el aislamiento de la bacteria en bacteriocitos o en un tejido específico del insecto (Moran & Telang, 1998), el entorno bioquímico al que están vinculados

(Hoffmeister & Martin, 2003), la dependencia metabólica al hospedador y/o otros cosimbiontes (Gosalbes et al., 2008; Rao et al., 2015), o una combinación de estos fenómenos, que sumados a las tasas evolutivas de sus genes, determinan sus redes metabólicas y actúan de manera crucial sobre su composición.

Es muy probable que las asociaciones simbióticas proporcionen una firma distintiva debido a sus redes metabólicas interconectadas y muy reducidas, lo que podría resultar en patrones ecológicos, topológicos y dinámicos particulares. Además, puede ser posible rastrear la evolución de las redes metabólicas en una filogenia gracias a la disponibilidad de genomas bacterianos taxonómicamente relacionados con limitaciones evolutivas comparables. Este método puede ayudar a aclarar los procesos evolutivos y las restricciones que tienen un impacto en la evolución de las redes metabólicas (Mithani et al., 2010).

Objetivos de la tesis

El objetivo principal de esta tesis fue buscar patrones evolutivos en los genomas reducidos de bacterias endosimbióticas de insectos y explorar si la evolución convergente es una característica común de su evolución a largo plazo, utilizando para ello herramientas bioinformáticas diseñadas específicamente en este trabajo para abordar la genómica comparativa y el estudio de las propiedades de sus redes metabólicas a gran escala.

Los objetivos específicos incluyeron:

- 1) Revisión y propuesta del nuevo concepto de *simbiónulo*, basado en la noción de bacterias intracelulares de insectos convirtiéndose en orgánulos.
- 2) Construcción de una base de datos terciaria compuesta de genomas simbióticos, con toda la información genómica disponible para descargar, incluyendo genes, genomas, ortólogos y reconstrucciones metabólicas, para estudiar estas relaciones a gran escala.
- 3) Desarrollo de herramientas de software públicas específicas para análisis de *big data* sobre genomas endosimbióticos, para analizar redes metabólicas simbióticas y nuevas metodologías para reconstrucciones de redes.
- 4) Metabolómica comparativa: desarrollo de experimentos *in silico* para investigar la coevolución y para comparar la modularidad de las interacciones metabólicas entre bacterias simbióticas de insectos.

Materiales y Métodos

Esta tesis implicó una amplia investigación bibliográfica para proponer un marco conceptual adecuado, además del uso y desarrollo de herramientas bioinformáticas para el análisis de bacterias endosimbióticas de insectos a gran escala. También se implementó el uso de programas personalizados escritos en el lenguaje de programación R (Team and Others 2008), para diversos análisis, figuras y la construcción de fondo de la base de datos SymGenDB para los

capítulos 1, 2, 3 y 4, que dieron como fruto las publicaciones Reyes-Prieto et al., 2014, 2015; Reyes-Prieto, Gil, et al., 2020; Reyes-Prieto, Vázquez-Chávez, et al., 2020, más la publicación del último capítulo de esta tesis que se encuentra actualmente en revisión.

Varios repositorios públicos también han sido utilizados para la creación, validación y compilación de datos para SymGenDB: NCBI (Wheeler et al. 2007), GOLD (Mukherjee et al. 2017), IMG (Markowitz et al. 2012), KEGG (Kanehisa and Goto 2000), Microbial Genomes Database (Uchiyama et al. 2019); así como el software Shiny (RStudio 2013) para la construcción de su interfaz web.

El uso de software bioinformático que se ha empleado para los análisis de los diferentes capítulos de esta tesis también incluyen la herramienta MetaDAG (Alberich et al. 2017) para crear redes metabólicas comprimidas, utilizada para análisis en los capítulos 2, 3 y 4; Cytoscape (Shannon et al. 2003), para construir las redes metabólicas con el modelo basado en metabolitos de bacterias endosimbióticas de insectos para el capítulo 4; Cytoscape Network Analyzer (Doncheva et al. 2012), para calcular los parámetros topológicos de la red metabólica del capítulo 4; MUSCLE (Edgar 2004), MAFFT (Kato and Standley 2013), jModelTest2 (Darriba et al. 2012), Mr. Bayes (Huelsbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003), PhyloPhlan (Segata et al. 2013), FastTree (Price et al. 2010) y FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>), para los alineamientos, construcción, validación y visualización de los árboles filogenéticos contruidos y analizados en el capítulo 4; y por último, BayesTraits (Pagel and Meade, n.d.), para medir la coevolución de los parámetros de las

redes metabólicas y el tamaño del genoma de bacterias endosimbióticas de insectos en el capítulo 4. Todos los parámetros específicos y el uso exacto de las herramientas están explicadas dentro de los capítulos donde fueron empleadas.

Resultados y Discusión

En el marco conceptual del estudio evolutivo de microbios con genomas exiguos, planteamos y analizamos el concepto de *simbiónulo* en la publicación "Scanty microbes, the 'symbionelle' concept", y realizamos una revisión sobre la evolución de genomas procariotas muy reducidos en la publicación "Evolution of small prokaryotic genomes". En ambas publicaciones se hace referencia a algunos casos extremos de bacterias endosimbiontes de insectos cuyos genomas son tan reducidos que han perdido muchos genes y rutas metabólicas y, por lo tanto, dependen de su hospedador para funciones hasta ahora consideradas esenciales para definir un organismo vivo (homeostasis, reproducción y evolución), de manera que podría ser considerado un nivel intermedio entre una célula mínima y un orgánulo, motivo por el cual se propone un nuevo concepto, el simbiónulo. Así mismo, se destaca la importancia de estudiar las comunidades endosimbiontes en insectos, ya que pueden tener un impacto significativo en la ecología y evolución de sus hospedadores, y sugerimos que comprender las interacciones entre los endosimbiontes y sus anfitriones puede proporcionar información sobre la evolución de las relaciones simbióticas y los mecanismos subyacentes a la persistencia de estas relaciones en el tiempo.

En la publicación “Evolution of small prokaryotic genomes” también se analiza la evolución de múltiples genomas procariotas caracterizados por su reducido tamaño y sus rutas metabólicas simplificadas. Se argumenta que el síndrome de reducción genómica observado en endosimbiontes es el resultado de una combinación de factores, incluida la selección natural, la deriva genética y la ausencia de transferencia horizontal de material genético. La publicación destaca la importancia de estudiar los genomas procarióticos para comprender la evolución de la vida en la Tierra, ya que los procariotas representan la mayor parte de la diversidad genética, celular y bioquímica de la vida. En el trabajo se presentan las diferentes hipótesis que se han propuesto para explicar la evolución genómica reductiva tanto en organismos de vida libre como endosimbiontes (Mira et al., 2001; Dufresne et al., 2005; Giovannoni et al., 2005; Marais et al., 2008; Morris et al., 2012). Para los organismos de vida libre, encontramos la hipótesis de optimización (*streamlining* en inglés, donde la selección natural favorece genomas pequeños y con bajo contenido en GC como forma de economización celular, y la hipótesis de cepas mutadoras, en la que la colonización del nicho favorece la persistencia de cepas mutadoras con mayor eficacia biológica, de modo que los genes con mejor eficacia se pierden. En el caso de los organismos simbioses, las dos hipótesis alternativas propuestas son la hipótesis de pérdida de genes de reparación de DNA como motor de cambio, y la hipótesis del trinquete de Muller, que causa la acumulación progresiva e irreversible de mutaciones ligeramente deletéreas. En este mismo artículo, también se discute la hipótesis de la Reina Negra (Morris et al., 2012; D’Souza et

al., 2014) en el contexto de la simbiosis, que predice que en una comunidad donde diferentes especies producen un material costoso difusible, el sistema evolucionará hacia un escenario donde solo unos pocos integrantes continuarán produciéndolo para el conjunto, dando lugar a un reparto de tareas en el que todos los integrantes se benefician.

Al inicio de esta tesis, nos vimos en la necesidad de organizar y optimizar el acceso a los datos genómicos y los metadatos de la gran cantidad de información de bacterias endosimbióticas de insectos que estaban públicamente disponibles, pero dispersos en múltiples repositorios sin facilidad de extracción. Para ello, se construyó el repositorio SymbioGenomesDB (ahora SymGenDB), una base de datos que integra y permite un fácil acceso al conocimiento sobre las relaciones hospedador-simbionte. Se trata de un recurso público que mantiene un catálogo de genomas bacterianos de organismos involucrados en relaciones simbióticas, seleccionados manualmente, que han sido completamente secuenciados, anotados y analizados funcionalmente. La primera publicación constaba de solo tres módulos en los que los usuarios podían buscar las bacterias involucradas en una relación simbiótica específica, sus genes (incluidos sus ortólogos) y genomas. Con el descubrimiento de la metodología MetaDAG (Alberich et al., 2017), para comprimir redes metabólicas con el fin de facilitar su estudio, vimos una gran oportunidad para su aplicación a todos los genomas simbióticos disponibles en SymGenDB. Por ello, colaboramos con los desarrolladores de MetaDAG y construimos un módulo adicional a los tres ya disponibles en la web , que fue

publicada en una actualización de nuestra base de datos, donde incluimos las redes comprimidas para cada organismo disponible en el repositorio. El nuevo módulo brindó oportunidades únicas para explorar el metabolismo de cada organismo y/o evaluar las capacidades metabólicas compartidas y conjuntas de organismos del mismo género incluidos en el catálogo, para permitir a los usuarios construir modelos predictivos de asociaciones metabólicas y complementaciones entre relaciones simbióticas. Además, como parte de la actualización de SymGenDB, aumentamos en alrededor del 25 % el contenido seleccionado manualmente dentro de la base de datos, que incluye ahora más de 2300 genomas bacterianos asociados con casi 500 hospedadores, lo que aumenta el catálogo y mejora su utilidad. La publicación destaca la importancia de comprender las interacciones entre el hospedador y el simbiote y cómo pueden proporcionar información sobre la evolución de las relaciones simbióticas y los mecanismos que subyacen, por lo que supone un recurso valioso para los investigadores que estudian las relaciones simbióticas y su impacto en la ecología y la evolución, tanto de los hospedadores como de los simbiotes.

Otro tema estudiado en el laboratorio de Genética Evolutiva de Simbiotes donde se desarrolló esta tesis, es el caso de los genomas mínimos capaces de sustentar una célula viva. La fabricación de células mínimas artificiales abriría infinitas posibilidades de investigación en ciencias básicas y aplicadas. Con esta motivación, muchos grupos de investigación están desarrollando metodologías para construir una célula mínima estable que sea capaz de lograr la homeostasis

metabólica, reproducirse, y evolucionar en un ambiente controlado, los tres atributos esenciales que definen la vida. Aprovechando la investigación previa sobre el genoma mínimo teórico y las redes metabólicas mínimas inferidas a partir de él por Gil (2004) y Gabaldón (2007), nos dimos a la tarea de modelar una red metabólica mínima con la metodología MetaDAG en el artículo "The Metabolic Building Blocks of a Minimal Cell". Esta red se ha comprimido aún más como un gráfico acíclico dirigido del metabolismo (m-DAG) para visualizar mejor su topología y encontrar sus reacciones esenciales (es decir, reacciones críticas para mantener la conectividad de la red). También hemos comparado este mínimo m-DAG a los del genoma natural más pequeño conocido hasta ahora y una célula mínima semisintética creada en el laboratorio (Hutchison et al., 2016). Sugerimos que una propuesta de red metabólica cohesiva puede abrir el camino hacia la síntesis de células mínimas y destacamos la importancia de comprender los componentes metabólicos básicos de una célula mínima, que incluyen la biosíntesis de aminoácidos, nucleótidos y lípidos, así como el metabolismo energético y el metabolismo central del carbono.

El capítulo final de esta tesis es el último manuscrito que escribimos, en el que reunimos toda la información descrita en los apartados previos, interrelacionando las ideas conceptuales con las herramientas bioinformáticas creadas, y analizamos experimentalmente la información genómica de las bacterias endosimbióticas de insectos incluidas en SymGenDB. Aquí se presenta un análisis evolutivo a gran escala de redes metabólicas, donde aplicamos dos

modelos de reconstrucción metabólica a todos esos genomas, uno por métodos convencionales publicados en varios estudios (modelo basado en metabolitos) y otro aplicando la metodología MetaDAG descrita en la introducción de este trabajo. También incluimos la información que publicamos para el metabolismo mínimo y calculamos los parámetros topológicos de su red metabólica para poder compararla con el resto de los genomas. Encontramos una correlación significativa entre el coeficiente de agrupamiento, el diámetro de la red y el número de nodos de las redes dentro de los grupos taxonómicos asignados a nivel de género cuando realizamos pruebas evolutivas de los parámetros de la red asociados con el tamaño del genoma de cada organismo, y también se encontró una correlación entre las distancias de las redes metabólicas basada en el análisis MetaDAG y las distancias taxonómicas de una filogenia molecular. La principal conclusión de este trabajo es que existen signos evolutivos en las redes metabólicas de bacterias endosimbióticas de insectos. Los análisis a gran escala como éste podrían brindar una mejor comprensión de cómo el metabolismo modula el grado de variación entre organismos filogenéticamente diferentes que viven en diversos nichos pero que comparten restricciones evolutivas similares.

Conclusiones

1. La noción de *simbiónulo* se ha desarrollado para referirnos a los endosimbiontes que no poseen un conjunto mínimo de genes para ser considerados una célula viva. Representan una categoría novedosa debido

a sus genomas evolucionados con tan pocos genes que no pueden llevar a cabo las tres funciones esenciales de las células modernas sin la ayuda de un hospedador y/o cosimbionte(s) adicional(es). Aunque cada uno se desarrolló en escenarios evolutivos muy diferentes (orgánulos antes de la vida multicelular, simbióntulos específicamente durante la evolución de los insectos), estos simbióntulos muestran una convergencia evolutiva con orgánulos, con los que poseen similitudes y diferencias evidentes y significativas.

2. El aumento de las publicaciones sobre genomas microbianos ha obligado al diseño de numerosas herramientas para categorizar y ordenar mejor los datos con objeto de apreciarlos plenamente para su estudio. SymGenDB es una base de datos de sistemas altamente seleccionada que brinda acceso al catálogo de genomas completamente secuenciados y anotados de organismos simbióticos, a sus secuencias genómicas, métricas y redes metabólicas comprimidas, datos que pueden ser útiles para la investigación en múltiples áreas, optimizando el tiempo de búsqueda para las relaciones simbióticas y las características de los organismos implicados.
3. SymGenDB es una gran herramienta informática para el estudio de genomas de organismos implicados en relaciones simbióticas que incluye múltiples conjuntos de datos microbiológicos. Es un repositorio de datos

terciario altamente integrado y curado, muy accesible, de alta calidad y con un catálogo completo de las conexiones hospedador-simbionte.

4. El modelado del metabolismo por medio de un gráfico acíclico dirigido (m-DAG) basado en metabolismos mínimos puede ser un primer enfoque para la síntesis y manipulación de células mínimas. Modelamos una red que representa un metabolismo mínimo inferido de un genoma mínimo necesario para la vida y la comprimimos como un m-DAG para visualizar mejor su topología y encontrar reacciones críticas para mantener la conectividad de la red.
5. También hemos comparado este m-DAG mínimo con los del genoma natural más pequeño conocido hasta ahora y con una célula mínima sintética creada en laboratorio, como prueba de concepto de que este tipo de análisis se puede utilizar para estudios de metabolómica comparativa.
6. La compilación de compuestos y reacciones que presentamos en nuestra red modelo se puede extrapolar fácilmente a cualquier bacteria de interés, independientemente de su posición filogenética, ya que no nos enfocamos específicamente en los genes.
7. Se han utilizado dos modelos, basado en metabolitos y basado en reacciones, para estudiar el metabolismo de bacterias endosimbióticas de insectos. Debido a que el modelo basado en metabolitos considera relaciones compuestas y el basado en reacciones considera relaciones de reacción, los dendrogramas resultantes son diferentes. Ambos métodos

agrupan especies por género taxonómico, lo que implica un núcleo genómico altamente conservado, indicando la existencia de convergencia filogenética.

8. La metodología MetaDAG analiza topologías de red a partir de gráficos de reacción. Las redes metabólicas de las bacterias endosimbióticas están totalmente desconectadas, a diferencia de las de los organismos de vida libre, que tienen componentes más integrados. Los endosimbiontes con estas redes metabólicas necesitan el metabolismo complementario de los cosimbiontes y sus anfitriones, mientras que los organismos de vida libre deben estar preparados para un entorno cambiante y estímulos externos como cambios de temperatura, estrés, ataques de otros organismos y otros factores adversos.
9. Los parámetros topológicos coevolucionan en el modelo de metabolitos. Por lo tanto, el tamaño del genoma del organismo se corresponde con el número de nodos de la red, el diámetro y el coeficiente de agrupamiento. Cabe señalar que en este modelo, el número de aristas aumenta considerablemente, pero el diámetro, el coeficiente de agrupamiento y el número de nodos no se ven afectados por este crecimiento.
10. Al comparar los dos modelos, concluimos que el modelo de gráfico de reacción es mejor para manejar grandes cantidades de datos, porque trata las reacciones como nodos y sus interacciones como aristas, y genera el

"metamodelo" del metabolismo, que es más simple y fácil de analizar y visualizar.

11. Nuestros datos revelan que la evolución genética y la similitud de la red metabólica están altamente correlacionadas en las bacterias endosimbióticas de insectos, a pesar de que residen aisladas dentro de su hospedador. Por lo tanto, estas asociaciones simbióticas proporcionan una huella evolutiva única, similar solo a los patógenos y otros organismos dependientes de hospedador, debido a que sus redes metabólicas también son integradas y reducidas.

VIII. References

- Acevedo-Rocha, C. G., Fang, G., Schmidt, M., Ussery, D. W., & Danchin, A. (2013). From essential to persistent genes: a functional approach to constructing synthetic life. In *Trends in Genetics* (Vol. 29, Issue 5, pp. 273–279). <https://doi.org/10.1016/j.tig.2012.11.001>
- Alarcón, M. E., Polo, P. G., Akyüz, S. N., & Rafiqi, A. M. (2022). Evolution and ontogeny of bacteriocytes in insects. *Frontiers in Physiology*, *13*, 1034066. <https://doi.org/10.3389/fphys.2022.1034066>
- Alberich, R., Castro, J. A., Llabrés, M., & Palmer-Rodríguez, P. (2017). Metabolomics analysis: Finding out metabolic building blocks. *PloS One*, *12*(5), e0177031. <https://doi.org/10.1371/journal.pone.0177031>
- Allemand, D., & Furla, P. (2018). How does an animal behave like a plant? Physiological and molecular adaptations of zooxanthellae and their hosts to symbiosis. *Comptes Rendus Biologies*, *341*(5), 276–280. <https://doi.org/10.1016/j.crvi.2018.03.007>
- Allen, J. M., Light, J. E., Perotti, M. A., Braig, H. R., & Reed, D. L. (2009). Mutational meltdown in primary endosymbionts: selection limits Muller's ratchet. *PloS One*, *4*(3), e4969. <https://doi.org/10.1371/journal.pone.0004969>
- Almaas, E., Kovács, B., Vicsek, T., Oltvai, Z. N., & Barabási, A.-L. (2004). Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature*, *427*(6977), 839–843. <https://doi.org/10.1038/nature02289>
- Alvarez, A. F., & Georgellis, D. (2019). Bacterial Lipid Domains and Their Role in Cell Processes. In *Biogenesis of Fatty Acids, Lipids and Membranes* (pp. 575–592). https://doi.org/10.1007/978-3-319-50430-8_39
- Alves, J. M. P., Klein, C. C., da Silva, F. M., Costa-Martins, A. G., Serrano, M. G., Buck, G. A., Vasconcelos, A. T. R., Sagot, M.-F., Teixeira, M. M. G., Motta, M. C. M., & Camargo, E. P. (2013). Endosymbiosis in trypanosomatids: the genomic cooperation between bacterium and host in the synthesis of essential amino acids is heavily influenced by multiple horizontal gene transfers. *BMC Evolutionary Biology*, *13*, 190. <https://doi.org/10.1186/1471-2148-13-190>
- Anderson, I., Ngatchou Djao, O. D., Misra, M., Chertkov, O., Nolan, M., Lucas, S., Lapidus, A., Glavina Del Rio, T., Tice, H., Cheng, J.-F., Tapia, R., Han, C., Goodwin, L., Pitluck, S., Liolios, K., Ivanova, N., Mavromatis, K., Mikhailova, N., Pati, A., ... Kyrpides, N. C. (2010). Complete genome sequence of *Methanothermobacter thermautotrophicus* strain (V24ST). *Standards in Genomic Sciences*, *3*(3), 315–324. <https://doi.org/10.4056/sigs.1283367>
- Bati, C. B., Santilli, E., & Lombardo, L. (2015). Effect of arbuscular mycorrhizal fungi on growth and on micronutrient and macronutrient uptake and allocation in olive plantlets growing under high total Mn levels. In *Mycorrhiza* (Vol. 25, Issue 2, pp. 97–108). <https://doi.org/10.1007/s00572-014-0589-0>
- Batra, L. R., & Buchner, P. (1968). Endosymbiosis of Animals with Plant Microorganisms. In *Mycologia* (Vol. 60, Issue 2, p. 466). <https://doi.org/10.2307/3757184>

- Baumann, P. (2005). Biology bacteriocyte-associated endosymbionts of plant sap-sucking insects. *Annual Review of Microbiology*, 59, 155–189.
<https://doi.org/10.1146/annurev.micro.59.030804.121041>
- Bedau, M. A., Parke, E. C., Tangen, U., & Hantsche-Tangen, B. (2009). Social and ethical checkpoints for bottom-up synthetic biology, or protocells. In *Systems and Synthetic Biology* (Vol. 3, Issues 1-4, pp. 65–75). <https://doi.org/10.1007/s11693-009-9039-2>
- Bennett, G. M., & Moran, N. A. (2013). Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a Phloem-feeding insect. *Genome Biology and Evolution*, 5(9), 1675–1688. <https://doi.org/10.1093/gbe/evt118>
- Bennett, G. M., & Moran, N. A. (2015). Heritable symbiosis: The advantages and perils of an evolutionary rabbit hole. *Proceedings of the National Academy of Sciences of the United States of America*, 112(33), 10169–10176. <https://doi.org/10.1073/pnas.1421388112>
- Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2017). GenBank. *Nucleic Acids Research*, 45(D1), D37–D42.
<https://doi.org/10.1093/nar/gkw1070>
- Bertaux, J., Schmid, M., Hutzler, P., Hartmann, A., Garbaye, J., & Frey-Klett, P. (2005). Occurrence and distribution of endobacteria in the plant-associated mycelium of the ectomycorrhizal fungus *Laccaria bicolor* S238N. In *Environmental Microbiology* (Vol. 7, Issue 11, pp. 1786–1795). <https://doi.org/10.1111/j.1462-2920.2005.00867.x>
- Bianciotto, V., Bandi, C., Minerdi, D., Sironi, M., Tichy, H. V., & Bonfante, P. (1996). An obligately endosymbiotic mycorrhizal fungus itself harbors obligately intracellular bacteria. *Applied and Environmental Microbiology*, 62(8), 3005–3010.
<https://doi.org/10.1128/aem.62.8.3005-3010.1996>
- Bianciotto, V., Genre, A., Jargeat, P., Lumini, E., Bécard, G., & Bonfante, P. (2004). Vertical transmission of endobacteria in the arbuscular mycorrhizal fungus *Gigaspora margarita* through generation of vegetative spores. *Applied and Environmental Microbiology*, 70(6), 3600–3608. <https://doi.org/10.1128/AEM.70.6.3600-3608.2004>
- Bianciotto, V., Lumini, E., Bonfante, P., & Vandamme, P. (2003). “*Candidatus Glomeribacter gigasporarum*” gen. nov., sp. nov., an endosymbiont of arbuscular mycorrhizal fungi. In *International Journal of Systematic and Evolutionary Microbiology* (Vol. 53, Issue 1, pp. 121–124). <https://doi.org/10.1099/ijs.0.02382-0>
- Bidartondo, M. I., Read, D. J., Trappe, J. M., Merckx, V., Ligrone, R., & Duckett, J. G. (2011). The dawn of symbiosis between plants and fungi. *Biology Letters*, 7(4), 574–577.
<https://doi.org/10.1098/rsbl.2010.1203>
- Bliven, K. A., & Maurelli, A. T. (2012). Antivirulence genes: insights into pathogen evolution through gene loss. *Infection and Immunity*, 80(12), 4061–4070.
<https://doi.org/10.1128/IAI.00740-12>
- Bonen, L., & Doolittle, W. F. (1975). On the prokaryotic nature of red algal chloroplasts. *Proceedings of the National Academy of Sciences of the United States of America*, 72(6), 2310–2314.
<https://doi.org/10.1073/pnas.72.6.2310>
- Bonfante, P., & Genre, A. (2008). Plants and arbuscular mycorrhizal fungi: an

- evolutionary-developmental perspective. *Trends in Plant Science*, 13(9), 492–498.
<https://doi.org/10.1016/j.tplants.2008.07.001>
- Borenstein, E., & Feldman, M. W. (2009). Topological signatures of species interactions in metabolic networks. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 16(2), 191–200. <https://doi.org/10.1089/cmb.2008.06TT>
- Boscaro, V., Felletti, M., Vannini, C., Ackerman, M. S., Chain, P. S. G., Malfatti, S., Vergez, L. M., Shin, M., Doak, T. G., Lynch, M., & Petroni, G. (2013). Polynucleobacter necessarius, a model for genome reduction in both free-living and symbiotic bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 110(46), 18590–18595.
<https://doi.org/10.1073/pnas.1316687110>
- Burke, G. R., & Moran, N. A. (2011). Massive genomic decay in *Serratia symbiotica*, a recently evolved symbiont of aphids. *Genome Biology and Evolution*, 3, 195–208.
<https://doi.org/10.1093/gbe/evr002>
- Button, D. K. (1991). Biochemical basis for whole-cell uptake kinetics: specific affinity, oligotrophic capacity, and the meaning of the michaelis constant. *Applied and Environmental Microbiology*, 57(7), 2033–2038. <https://doi.org/10.1128/aem.57.7.2033-2038.1991>
- Carini, P., Steindler, L., Beszteri, S., & Giovannoni, S. J. (2013). Nutrient requirements for growth of the extreme oligotroph “*Candidatus Pelagibacter ubique*” HTCC1062 on a defined medium. *The ISME Journal*, 7(3), 592–602. <https://doi.org/10.1038/ismej.2012.122>
- Castillo, D. M., & Pawlowska, T. E. (2010). Molecular evolution in bacterial endosymbionts of fungi. *Molecular Biology and Evolution*, 27(3), 622–636.
<https://doi.org/10.1093/molbev/msp280>
- Cavalier-Smith, T., & Lee, J. J. (1985). Protozoa as Hosts for Endosymbioses and the Conversion of Symbionts into Organelles^{1,2}. In *The Journal of Protozoology* (Vol. 32, Issue 3, pp. 376–379).
<https://doi.org/10.1111/j.1550-7408.1985.tb04031.x>
- Clay, K. (1988). Fungal endophytes of grasses: A defensive mutualism between plants and fungi. *Ecology*, 69(1), 10–16. <https://doi.org/10.2307/1943155>
- Clayton, A. L., Oakeson, K. F., Gutin, M., Pontes, A., Dunn, D. M., von Niederhausern, A. C., Weiss, R. B., Fisher, M., & Dale, C. (2012). A novel human-infection-derived bacterium provides insights into the evolutionary origins of mutualistic insect-bacterial symbioses. *PLoS Genetics*, 8(11), e1002990. <https://doi.org/10.1371/journal.pgen.1002990>
- Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods*, 9(8), 772.
<https://doi.org/10.1038/nmeth.2109>
- de Bary, A. (1879). *Die Erscheinung der Symbiose: Vortrag gehalten auf der Versammlung Deutscher Naturforscher und Aerzte zu Cassel*. Trübner.
https://play.google.com/store/books/details?id=7oQ_AAAAYAAJ
- Delaye, L., Gil, R., Pereto, J., Latorre, A., & Moya, A. (2010). Life with a few genes: A survey on naturally evolved reduced genomes–!2009-11-30~!2010-01-24~!2010-05-07~! *The Open Evolution Journal*, 4(1), 12–22. <https://doi.org/10.2174/1874404401004010012>
- Delaye, L., & Moya, A. (2010). Evolution of reduced prokaryotic genomes and the minimal cell

- concept: variations on a theme. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 32(4), 281–287. <https://doi.org/10.1002/bies.200900161>
- Desirò, A., Faccio, A., Kaech, A., Bidartondo, M. I., & Bonfante, P. (2015). *E ndogone* , one of the oldest plant-associated fungi, host unique Mollicutes-related endobacteria. In *New Phytologist* (Vol. 205, Issue 4, pp. 1464–1472). <https://doi.org/10.1111/nph.13136>
- Desirò, A., Salvioli, A., Ngonkeu, E. L., Mondo, S. J., Epis, S., Faccio, A., Kaech, A., Pawlowska, T. E., & Bonfante, P. (2014). Detection of a novel intracellular microbiome hosted in arbuscular mycorrhizal fungi. *The ISME Journal*, 8(2), 257–270. <https://doi.org/10.1038/ismej.2013.151>
- Dimijian, G. G. (2000, July). Evolving together: the biology of symbiosis, part 1. In *Baylor University Medical Center Proceedings* (Vol. 13, No. 3, pp. 217a-226). Taylor & Francis. <https://doi.org/10.1080/08998280.2000.11927678>
- Doncheva, N. T., Assenov, Y., Domingues, F. S., & Albrecht, M. (2012). Topological analysis and interactive visualization of biological networks and protein structures. *Nature Protocols*, 7(4), 670–685. <https://doi.org/10.1038/nprot.2012.004>
- Douglas, A. E. (1989). Mycetocyte symbiosis in insects. *Biological Reviews of the Cambridge Philosophical Society*, 64(4), 409–434. <https://doi.org/10.1111/j.1469-185x.1989.tb00682.x>
- Douglas, A. E. (1998). Nutritional interactions in insect-microbial symbioses: aphids and their symbiotic bacteria *Buchnera*. *Annual Review of Entomology*, 43, 17–37. <https://doi.org/10.1146/annurev.ento.43.1.17>
- Douglas, A. E. (2009). The microbial dimension in insect nutritional ecology. In *Functional Ecology* (Vol. 23, Issue 1, pp. 38–47). <https://doi.org/10.1111/j.1365-2435.2008.01442.x>
- Douglas, A. E. (2014). Symbiosis as a general principle in eukaryotic evolution. *Cold Spring Harbor Perspectives in Biology*, 6(2). <https://doi.org/10.1101/cshperspect.a016113>
- Douglas, A. E. (2016). How multi-partner endosymbioses function. *Nature Reviews. Microbiology*, 14(12), 731–743. <https://doi.org/10.1038/nrmicro.2016.151>
- D’Souza, G., Waschina, S., Pande, S., Bohl, K., Kaleta, C., & Kost, C. (2014). Less is more: selective advantages can explain the prevalent loss of biosynthetic genes in bacteria. *Evolution; International Journal of Organic Evolution*, 68(9), 2559–2570. <https://doi.org/10.1111/evo.12468>
- Dubilier, N., Bergin, C., & Lott, C. (2008). Symbiotic diversity in marine animals: the art of harnessing chemosynthesis. *Nature Reviews. Microbiology*, 6(10), 725–740. <https://doi.org/10.1038/nrmicro1992>
- Dufresne, A., Garczarek, L., & Partensky, F. (2005). Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biology*, 6(2), R14. <https://doi.org/10.1186/gb-2005-6-2-r14>
- Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I. M., Barbe, V., Duprat, S., Galperin, M. Y., Koonin, E. V., Le Gall, F., Makarova, K. S., Ostrowski, M., Oztas, S., Robert, C., Rogozin, I. B., Scanlan, D. J., Tandeau de Marsac, N., Weissenbach, J., Wincker, P., ... Hess, W. R. (2003). Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proceedings of the National*

- Academy of Sciences of the United States of America*, 100(17), 10020–10025.
<https://doi.org/10.1073/pnas.1733211100>
- Duncan, R. P., Husnik, F., Van Leuven, J. T., Gilbert, D. G., Dávalos, L. M., McCutcheon, J. P., & Wilson, A. C. C. (2014). Dynamic recruitment of amino acid transporters to the insect/symbiont interface. *Molecular Ecology*, 23(6), 1608–1623.
<https://doi.org/10.1111/mec.12627>
- Dutta, C., & Paul, S. (2012). Microbial lifestyle and genome signatures. *Current Genomics*, 13(2), 153–162. <https://doi.org/10.2174/138920212799860698>
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), 1792–1797. <https://doi.org/10.1093/nar/gkh340>
- Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. *Die Naturwissenschaften*, 58(10), 465–523. <https://doi.org/10.1007/BF00623322>
- Fares, M. A., Moya, A., & Barrio, E. (2005). Adaptive evolution in GroEL from distantly related endosymbiotic bacteria of insects. *Journal of Evolutionary Biology*, 18(3), 651–660.
<https://doi.org/10.1111/j.1420-9101.2004.00861.x>
- Fares, M. A., Ruiz-González, M. X., Moya, A., Elena, S. F., & Barrio, E. (2002). Endosymbiotic bacteria: groEL buffers against deleterious mutations. *Nature*, 417(6887), 398.
<https://doi.org/10.1038/417398a>
- Feldhaar, H., & Gross, R. (2009). Insects as hosts for mutualistic bacteria. *International Journal of Medical Microbiology: IJMM*, 299(1), 1–8. <https://doi.org/10.1016/j.ijmm.2008.05.010>
- Feldhaar, H., Straka, J., Krischke, M., Berthold, K., Stoll, S., Mueller, M. J., & Gross, R. (2007). Nutritional upgrading for omnivorous carpenter ants by the endosymbiont Blochmannia. *BMC Biology*, 5, 48. <https://doi.org/10.1186/1741-7007-5-48>
- Fisher, R. M., Henry, L. M., Cornwallis, C. K., Kiers, E. T., & West, S. A. (2017). The evolution of host-symbiont dependence. *Nature Communications*, 8, 15973.
<https://doi.org/10.1038/ncomms15973>
- Frey-Klett, P., Burlinson, P., Deveau, A., Barret, M., Tarkka, M., & Sarniguet, A. (2011). Bacterial-fungal interactions: hyphens between agricultural, clinical, environmental, and food microbiologists. *Microbiology and Molecular Biology Reviews: MMBR*, 75(4), 583–609.
<https://doi.org/10.1128/MMBR.00020-11>
- Gabaldón, T. (2018). Relative timing of mitochondrial endosymbiosis and the “pre-mitochondrial symbioses” hypothesis. In *IUBMB Life* (Vol. 70, Issue 12, pp. 1188–1196). <https://doi.org/10.1002/iub.1950>
- Gabaldon, T., Pereto, J., Montero, F., Gil, R., Latorre, A., & Moya, A. (2007). Structural analyses of a hypothetical minimal metabolism. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 362(1486), 1751–1762.
- Garzón, M. J., Reyes-Prieto, M., & Gil, R. (2022). The Minimal Translation Machinery: What We Can Learn From Naturally and Experimentally Reduced Genomes. *Frontiers in Microbiology*, 13, 858983. <https://doi.org/10.3389/fmicb.2022.858983>
- Ghai, R., Mizuno, C. M., Picazo, A., Camacho, A., & Rodríguez-Valera, F. (2013). Metagenomics uncovers a new group of low GC and ultra-small marine Actinobacteria. In *Scientific Reports*

- (Vol. 3, Issue 1). <https://doi.org/10.1038/srep02471>
- Ghignone, S., Salvioli, A., Anca, I., Lumini, E., Ortu, G., Petiti, L., Cruveiller, S., Bianciotto, V., Piffanelli, P., Lanfranco, L., & Bonfante, P. (2012). The genome of the obligate endobacterium of an AM fungus reveals an interphylum network of nutritional interactions. *The ISME Journal*, 6(1), 136–145. <https://doi.org/10.1038/ismej.2011.110>
- Gho N., Yaoita T., Aoyagi K., & Sato Z. (1978). Studies on the control of *Rhizopus* in the nursery cases of rice seedlings, 5: Influence of a phytotoxic substance produced by *Rhizopus* in growth of rice. *Proceedings of the Association for Plant Protection of Hokuriku*. <https://agris.fao.org/agris-search/search.do?recordID=JP19800517952>
- Gibson, C. M., & Hunter, M. S. (2010). Extraordinarily widespread and fantastically complex: comparative biology of endosymbiotic bacterial and fungal mutualists of insects. In *Ecology Letters* (Vol. 13, Issue 2, pp. 223–234). <https://doi.org/10.1111/j.1461-0248.2009.01416.x>
- Gillespie, J. H. (1998). *Population Genetics: A Concise Guide*. Johns Hopkins University Press. <https://play.google.com/store/books/details?id=eslingEACAAJ>
- Gil, R. (2015). The minimal gene-set machinery. *Encyclopedia of Molecular Cell Biology and Molecular Medicine: Synthetic Biology*, 443–478.
- Gil, R., & Latorre, A. (2019). Unity Makes Strength: A Review on Mutualistic Symbiosis in Representative Insect Clades. *Life*, 9(1). <https://doi.org/10.3390/life9010021>
- Gil, R., Pérez-Brocal, V., Latorre, A., & Moya, A. (2006). Minimal genomes required for life. In *Prokaryotic diversity* (pp. 105–122). <https://doi.org/10.1017/cbo9780511754913.007>
- Gil, R., Sabater-Muñoz, B., Latorre, A., Silva, F. J., & Moya, A. (2002). Extreme genome reduction in *Buchnera* spp.: toward the minimal genome needed for symbiotic life. *Proceedings of the National Academy of Sciences of the United States of America*, 99(7), 4454–4458. <https://doi.org/10.1073/pnas.062067299>
- Gil, R., Sabater-Muñoz, B., Perez-Brocal, V., Silva, F. J., & Latorre, A. (2006). Plasmids in the aphid endosymbiont *Buchnera aphidicola* with the smallest genomes. A puzzling evolutionary story. In *Gene* (Vol. 370, pp. 17–25). <https://doi.org/10.1016/j.gene.2005.10.043>
- Gil, R., Silva, F. J., Peretó, J., & Moya, A. (2004). Determination of the core of a minimal bacterial gene set. *Microbiology and Molecular Biology Reviews: MMBR*, 68(3), 518–537, table of contents. <https://doi.org/10.1128/MMBR.68.3.518-537.2004>
- Gil, R., Silva, F. J., Zientz, E., Delmotte, F., González-Candelas, F., Latorre, A., Rausell, C., Kamerbeek, J., Gadau, J., Hölldobler, B., van Ham, R. C. H. J., Gross, R., & Moya, A. (2003). The genome sequence of *Blochmannia floridanus*: Comparative analysis of reduced genomes. In *Proceedings of the National Academy of Sciences* (Vol. 100, Issue 16, pp. 9388–9393). <https://doi.org/10.1073/pnas.1533499100>
- Giovannoni, S. J., Cameron Thrash, J., & Temperton, B. (2014). Implications of streamlining theory for microbial ecology. *The ISME Journal*, 8(8), 1553–1565. <https://doi.org/10.1038/ismej.2014.60>
- Giovannoni, S. J., Hayakawa, D. H., Tripp, H. J., Stingl, U., Givan, S. A., Cho, J.-C., Oh, H.-M., Kitner, J. B., Vergin, K. L., & Rappé, M. S. (2008). The small genome of an abundant coastal ocean methylophile. *Environmental Microbiology*, 10(7), 1771–1782.

- <https://doi.org/10.1111/j.1462-2920.2008.01598.x>
- Giovannoni, S. J., Tripp, H. J., Givan, S., Podar, M., Vergin, K. L., Baptista, D., Bibbs, L., Eads, J., Richardson, T. H., Noordewier, M., Rappé, M. S., Short, J. M., Carrington, J. C., & Mathur, E. J. (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science*, 309(5738), 1242–1245. <https://doi.org/10.1126/science.1114057>
- Glass, J. I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M. R., Maruf, M., Hutchison, C. A., 3rd, Smith, H. O., & Venter, J. C. (2006). Essential genes of a minimal bacterium. *Proceedings of the National Academy of Sciences of the United States of America*, 103(2), 425–430. <https://doi.org/10.1073/pnas.0510013103>
- Glass, J. I., Merryman, C., Wise, K. S., Hutchison, C. A., & Smith, H. O. (2017). Minimal Cells—Real and Imagined. In *Cold Spring Harbor Perspectives in Biology* (Vol. 9, Issue 12, p. a023861). <https://doi.org/10.1101/cshperspect.a023861>
- Gosalbes, M. J., Lamelas, A., Moya, A., & Latorre, A. (2008). The striking case of tryptophan provision in the cedar aphid *Cinara cedri*. *Journal of Bacteriology*, 190(17), 6026–6029. <https://doi.org/10.1128/JB.00525-08>
- Gould, S. B., Waller, R. F., & McFadden, G. I. (2008). Plastid evolution. *Annual Review of Plant Biology*, 59, 491–517. <https://doi.org/10.1146/annurev.arplant.59.032607.092915>
- Gould, S. J. (n.d.). 1996. Full House, The Spread of Excellence from Plato to Darwin. *New York, NY: Harmony*.
- Gould, S. J. (2002). *The Structure of Evolutionary Theory*. Harvard University Press. <https://play.google.com/store/books/details?id=ILkFAwAAQBAJ>
- Gros, P.-A., & Tenaillon, O. (2009). Selection for chaperone-like mediated genetic robustness at low mutation rate: impact of drift, epistasis and complexity. *Genetics*, 182(2), 555–564. <https://doi.org/10.1534/genetics.108.099366>
- Grote, J., Thrash, J. C., Huggett, M. J., Landry, Z. C., Carini, P., Giovannoni, S. J., & Rappé, M. S. (2012). Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *mBio*, 3(5). <https://doi.org/10.1128/mBio.00252-12>
- Guo, J., Hatt, S., He, K., Chen, J., Francis, F., & Wang, Z. (2017). Nine facultative endosymbionts in aphids. A review. *Journal of Asia-Pacific Entomology*, 20(3), 794–801. <https://doi.org/10.1016/j.aspen.2017.03.025>
- Haft, D. H., Selengut, J. D., & White, O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Research*, 31(1), 371–373. <https://doi.org/10.1093/nar/gkg128>
- Halfmann, R., Jarosz, D. F., Jones, S. K., Chang, A., Lancaster, A. K., & Lindquist, S. (2012). Prions are a common mechanism for phenotypic inheritance in wild yeasts. *Nature*, 482(7385), 363–368. <https://doi.org/10.1038/nature10875>
- Hansen, A. K., & Moran, N. A. (2011). Aphid genome expression reveals host–symbiont cooperation in the production of amino acids. *Proceedings of the National Academy of Sciences*, 108(7), 2849–2854. <https://doi.org/10.1073/pnas.1013465108>
- Hansen, A. K., & Moran, N. A. (2014). The impact of microbial symbionts on host plant utilization by herbivorous insects. *Molecular Ecology*, 23(6), 1473–1496. <https://doi.org/10.1111/mec.12421>

- Harmon, J. P., Moran, N. A., & Ives, A. R. (2009). Species response to environmental change: impacts of food web interactions and evolution. *Science*, 323(5919), 1347–1350. <https://doi.org/10.1126/science.1167396>
- He, J., Ritalahti, K. M., Yang, K.-L., Koenigsberg, S. S., & Löffler, F. E. (2003). Detoxification of vinyl chloride to ethene coupled to growth of an anaerobic bacterium. *Nature*, 424(6944), 62–65. <https://doi.org/10.1038/nature01717>
- Hershberg, R., & Petrov, D. A. (2010). Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genetics*, 6(9), e1001115. <https://doi.org/10.1371/journal.pgen.1001115>
- Heyworth, E. R., & Ferrari, J. (2016). Heat Stress Affects Facultative Symbiont-Mediated Protection from a Parasitoid Wasp. *PLoS One*, 11(11), e0167180. <https://doi.org/10.1371/journal.pone.0167180>
- Hildebrand, F., Meyer, A., & Eyre-Walker, A. (2010). Evidence of selection upon genomic GC-content in bacteria. *PLoS Genetics*, 6(9), e1001107. <https://doi.org/10.1371/journal.pgen.1001107>
- Hoffmann, M. P., & Frodsham, A. (1993). Natural enemies of vegetable insect pests. Cornell University.
- Hoffman, M. T., & Arnold, A. E. (2010). Diverse bacteria inhabit living hyphae of phylogenetically diverse fungal endophytes. *Applied and Environmental Microbiology*, 76(12), 4063–4075. <https://doi.org/10.1128/AEM.02928-09>
- Hoffmeister, M., & Martin, W. (2003). Interspecific evolution: microbial symbiosis, endosymbiosis and gene transfer. *Environmental Microbiology*, 5(8), 641–649. <https://doi.org/10.1046/j.1462-2920.2003.00454.x>
- Holland, J. N., & Bronstein, J. L. (2008). Mutualism. In S. E. Jørgensen & B. D. Fath (Eds.), *Encyclopedia of Ecology* (pp. 2485–2491). Academic Press. <https://doi.org/10.1016/B978-008045405-4.00673-X>
- Huggett, M. J., Hayakawa, D. H., & Rappé, M. S. (2012). Genome sequence of strain HIMB624, a cultured representative from the OM43 clade of marine Betaproteobacteria. *Standards in Genomic Sciences*, 6(1), 11–20. <https://doi.org/10.4056/sigs.2305090>
- Hulslenbeck, J. P., & Ronquist, F. (2001). MrBayes: Bayesian inference of phylogeny. *Bioinformatics*, 17, 754–755.
- Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., ... Yeats, C. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Research*, 37(Database issue), D211–D215. <https://doi.org/10.1093/nar/gkn785>
- Husnik, F., & Keeling, P. J. (2019). The fate of obligate endosymbionts: reduction, integration, or extinction. *Current Opinion in Genetics & Development*, 58-59, 1–8. <https://doi.org/10.1016/j.gde.2019.07.014>
- Husnik, F., & McCutcheon, J. P. (2016). Repeated replacement of an intrabacterial symbiont in the tripartite nested mealybug symbiosis. *Proceedings of the National Academy of Sciences of the United States of America*, 113(37), E5416–E5424. <https://doi.org/10.1073/pnas.1603910113>

- Husnik, F., Nikoh, N., Koga, R., Ross, L., Duncan, R. P., Fujie, M., Tanaka, M., Satoh, N., Bachtrog, D., Wilson, A. C. C., von Dohlen, C. D., Fukatsu, T., & McCutcheon, J. P. (2013). Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell*, *153*(7), 1567–1578. <https://doi.org/10.1016/j.cell.2013.05.040>
- Hutchison, C. A., Chuang, R.-Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., Gill, J., Kannan, K., Karas, B. J., Ma, L., Pelletier, J. F., Qi, Z.-Q., Alexander Richter, R., Strychalski, E. A., Sun, L., Suzuki, Y., Tsvetanova, B., Wise, K. S., Smith, H. O., ... Craig Venter, J. (2016). Design and synthesis of a minimal bacterial genome. In *Science* (Vol. 351, Issue 6280). <https://doi.org/10.1126/science.aad6253>
- Ibarra, R. U., Edwards, J. S., & Palsson, B. O. (2002). *Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature*, *420*(6912), 186–189. <https://doi.org/10.1038/nature01149>
- International Aphid Genomics Consortium. (2010). Genome sequence of the pea aphid *Acyrtosiphon pisum*. *PLoS Biology*, *8*(2), e1000313. <https://doi.org/10.1371/journal.pbio.1000313>
- Ioannidis, P., Johnston, K. L., Riley, D. R., Kumar, N., White, J. R., Olarte, K. T., Ott, S., Tallon, L. J., Foster, J. M., Taylor, M. J., & Dunning Hotopp, J. C. (2013). Extensively duplicated and transcriptionally active recent lateral gene transfer from a bacterial *Wolbachia* endosymbiont to its host filarial nematode *Brugia malayi*. *BMC Genomics*, *14*(1), 639. <https://doi.org/10.1186/1471-2164-14-639>
- Islas, S., Becerra, A., Luisi, P. L., & Lazcano, A. (2004). Comparative genomics and the gene complement of a minimal cell. *Origins of Life and Evolution of the Biosphere: The Journal of the International Society for the Study of the Origin of Life*, *34*(1-2), 243–256. <https://doi.org/10.1023/b:orig.0000009844.90540.52>
- Itoh, T., Martin, W., & Nei, M. (2002). Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(20), 12944–12948. <https://doi.org/10.1073/pnas.192449699>
- Jewett, M. C., & Forster, A. C. (2010). Update on designing and building minimal cells. In *Current Opinion in Biotechnology* (Vol. 21, Issue 5, pp. 697–703). <https://doi.org/10.1016/j.copbio.2010.06.008>
- Jiménez, E., Langa, S., Martín, V., Arroyo, R., Martín, R., Fernández, L., & Rodríguez, J. M. (2010). Complete genome sequence of *Lactobacillus fermentum* CECT 5716, a probiotic strain isolated from human milk. *Journal of Bacteriology*, *192*(18), 4800. <https://doi.org/10.1128/JB.00702-10>
- Joy, J. B. (2013). Symbiosis catalyses niche expansion and diversification. *Proceedings. Biological Sciences / The Royal Society*, *280*(1756), 20122820. <https://doi.org/10.1098/rspb.2012.2820>
- Kai, K., Furuyabu, K., Tani, A., & Hayashi, H. (2012). Production of the Quorum-Sensing Molecules N-Acylhomoserine Lactones by Endobacteria Associated with *Mortierella alpina* A-178. In *ChemBioChem* (Vol. 13, Issue 12, pp. 1776–1784). <https://doi.org/10.1002/cbic.201200263>
- Kaltenpoth, M., Göttler, W., Herzner, G., & Strohm, E. (2005). Symbiotic bacteria protect wasp

- larvae from fungal infestation. *Current Biology: CB*, 15(5), 475–479.
<https://doi.org/10.1016/j.cub.2004.12.084>
- Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1), D353–D361. <https://doi.org/10.1093/nar/gkw1092>
- Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1), 27–30. <https://doi.org/10.1093/nar/28.1.27>
- Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research*, 42(Database issue), D199–D205. <https://doi.org/10.1093/nar/gkt1076>
- Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1), D457–D462. <https://doi.org/10.1093/nar/gkv1070>
- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772–780. <https://doi.org/10.1093/molbev/mst010>
- Keeling, P. J., & Archibald, J. M. (2008). Organelle evolution: what's in a name? *Current Biology: CB*, 18(8), R345–R347. <https://doi.org/10.1016/j.cub.2008.02.065>
- Kelly, S. (2021). The economics of organellar gene loss and endosymbiotic gene transfer. *Genome Biology*, 22(1), 345. <https://doi.org/10.1186/s13059-021-02567-w>
- Kenyon, L. J., & Sabree, Z. L. (2014). Obligate insect endosymbionts exhibit increased ortholog length variation and loss of large accessory proteins concurrent with genome shrinkage. *Genome Biology and Evolution*, 6(4), 763–775. <https://doi.org/10.1093/gbe/evu055>
- Kern, A. D., & Kondrashov, F. A. (2004). Mechanisms and convergence of compensatory evolution in mammalian mitochondrial tRNAs. *Nature Genetics*, 36(11), 1207–1212. <https://doi.org/10.1038/ng1451>
- Kikuchi, Y., Hayatsu, M., Hosokawa, T., Nagayama, A., Tago, K., & Fukatsu, T. (2012). Symbiont-mediated insecticide resistance. *Proceedings of the National Academy of Sciences of the United States of America*, 109(22), 8618–8622. <https://doi.org/10.1073/pnas.1200231109>
- Kirkness, E. F., Haas, B. J., Sun, W., Braig, H. R., Perotti, M. A., Clark, J. M., Lee, S. H., Robertson, H. M., Kennedy, R. C., Elhaik, E., Gerlach, D., Kriventseva, E. V., Elsik, C. G., Gaur, D., Hill, C. A., Veenstra, J. A., Walenz, B., Tubío, J. M. C., Ribeiro, J. M. C., ... Pittendrigh, B. R. (2010). Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proceedings of the National Academy of Sciences of the United States of America*, 107(27), 12168–12173. <https://doi.org/10.1073/pnas.1003379107>
- Kleiner, M., Wenstrup, C., Lott, C., Teeling, H., Wetzel, S., Young, J., Chang, Y.-J., Shah, M., VerBerkmoes, N. C., Zarzycki, J., Fuchs, G., Markert, S., Hempel, K., Voigt, B., Becher, D., Liebeke, M., Lalk, M., Albrecht, D., Hecker, M., ... Dubilier, N. (2012). Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. *Proceedings of the National Academy of Sciences of the United States of*

- America*, 109(19), E1173–E1182. <https://doi.org/10.1073/pnas.1121198109>
- Koch, A. L. (1996). WHAT SIZE SHOULD A BACTERIUM BE? A Question of Scale. *Annual Review of Microbiology*, 50(1), 317–348. <https://doi.org/10.1146/annurev.micro.50.1.317>
- Kohl, K. D., Weiss, R. B., Cox, J., Dale, C., & Dearing, M. D. (2014). Gut microbes of mammalian herbivores facilitate intake of plant toxins. *Ecology Letters*, 17(10), 1238–1246. <https://doi.org/10.1111/ele.12329>
- Koskiniemi, S., Sun, S., Berg, O. G., & Andersson, D. I. (2012). Selection-driven gene loss in bacteria. *PLoS Genetics*, 8(6), e1002787. <https://doi.org/10.1371/journal.pgen.1002787>
- Kuo, C.-H., Moran, N. A., & Ochman, H. (2009). The consequences of genetic drift for bacterial genome complexity. *Genome Research*, 19(8), 1450–1454. <https://doi.org/10.1101/gr.091785.109>
- Kwan, J. C., Donia, M. S., Han, A. W., Hirose, E., Haygood, M. G., & Schmidt, E. W. (2012). Genome streamlining and chemical defense in a coral reef symbiosis. *Proceedings of the National Academy of Sciences of the United States of America*, 109(50), 20655–20660. <https://doi.org/10.1073/pnas.1213820109>
- Lackner, G., Moebius, N., Partida-Martinez, L., & Hertweck, C. (2011). Complete genome sequence of Burkholderia rhizoxinica, an Endosymbiont of Rhizopus microsporus. *Journal of Bacteriology*, 193(3), 783–784. <https://doi.org/10.1128/JB.01318-10>
- Lackner, G., Moebius, N., Partida-Martinez, L. P., Boland, S., & Hertweck, C. (2011). Evolution of an endofungal lifestyle: Deductions from the Burkholderia rhizoxinica genome. *BMC Genomics*, 12, 210. <https://doi.org/10.1186/1471-2164-12-210>
- Lamelas, A., Gosalbes, M. J., Manzano-Marín, A., Peretó, J., Moya, A., & Latorre, A. (2011). Serratia symbiotica from the aphid Cinara cedri: a missing link from facultative to obligate insect endosymbiont. *PLoS Genetics*, 7(11), e1002357. <https://doi.org/10.1371/journal.pgen.1002357>
- Lamelas, A., Gosalbes, M. J., Moya, A., & Latorre, A. (2011). New Clues about the Evolutionary History of Metabolic Losses in Bacterial Endosymbionts, Provided by the Genome of Buchnera aphidicola from the Aphid Cinara tujafilina. In *Applied and Environmental Microbiology* (Vol. 77, Issue 13, pp. 4446–4454). <https://doi.org/10.1128/aem.00141-11>
- Latorre, A., Durbán, A., Moya, A., & Peretó, J. (2011). Role of symbiosis in eukaryotic evolution. *Origins and Evolution of Life--An Astrobiological Perspective*, Edited by: Gargaud, M., Lopez-Garcia, P., and Martin, H., Cambridge University Press, Cambridge UK, 326–339.
- Latorre, A., & Manzano-Marín, A. (2017). Dissecting genome reduction and trait loss in insect endosymbionts. *Annals of the New York Academy of Sciences*, 1389(1), 52–75. <https://doi.org/10.1111/nyas.13222>
- Lazarev, V. N., Levitskii, S. A., Basovskii, Y. I., Chukin, M. M., Akopian, T. A., Vereshchagin, V. V., Kostrjukova, E. S., Kovaleva, G. Y., Kazanov, M. D., Malko, D. B., Vitreschak, A. G., Sernova, N. V., Gelfand, M. S., Demina, I. A., Serebryakova, M. V., Galyamina, M. A., Vtyurin, N. N., Rogov, S. I., Alexeev, D. G., ... Govorun, V. M. (2011). Complete genome and proteome of Acholeplasma laidlawii. *Journal of Bacteriology*, 193(18), 4943–4953. <https://doi.org/10.1128/JB.05059-11>

- Lee, M.-C., & Marx, C. J. (2012). Repeated, selection-driven genome reduction of accessory genes in experimental populations. *PLoS Genetics*, 8(5), e1002651. <https://doi.org/10.1371/journal.pgen.1002651>
- Lee, S. H., Jung, J. Y., Lee, S. H., & Jeon, C. O. (2011). Complete genome sequence of *Weissella koreensis* KACC 15510, isolated from kimchi. *Journal of Bacteriology*, 193(19), 5534. <https://doi.org/10.1128/JB.05704-11>
- Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Ten Hoopen, P., Vaughan, R., ... Cochrane, G. (2011). The European Nucleotide Archive. *Nucleic Acids Research*, 39(Database issue), D28–D31. <https://doi.org/10.1093/nar/gkq967>
- Löffler, F. E., Yan, J., Ritalahti, K. M., Adrian, L., Edwards, E. A., Konstantinidis, K. T., Müller, J. A., Fullerton, H., Zinder, S. H., & Spormann, A. M. (2013). *Dehalococcoides mccartyi* gen. nov., sp. nov., obligately organohalide-respiring anaerobic bacteria relevant to halogen cycling and bioremediation, belong to a novel bacterial class, *Dehalococcoidia* classis nov., order *Dehalococcoidales* ord. nov. and family *Dehalococcoidaceae* fam. nov., within the phylum *Chloroflexi*. *International Journal of Systematic and Evolutionary Microbiology*, 63(Pt 2), 625–635. <https://doi.org/10.1099/ijs.0.034926-0>
- López-Madrugal, S., Balmand, S., Latorre, A., Heddi, A., Moya, A., & Gil, R. (2013). How does *Tremblaya princeps* get essential proteins from its nested partner *Moranella endobia* in the Mealybug *Planococcus citri*? *PloS One*, 8(10), e77307. <https://doi.org/10.1371/journal.pone.0077307>
- López-Madrugal Sergio, Latorre Amparo, Porcar Manuel, Moya Andrés, & Gil Rosario. (2011). Complete Genome Sequence of “*Candidatus Tremblaya princeps*” Strain PCVAL, an Intriguing Translational Machine below the Living-Cell Status. *Journal of Bacteriology*, 193(19), 5587–5588. <https://doi.org/10.1128/JB.05749-11>
- López-Madrugal, S., Latorre, A., Porcar, M., Moya, A., & Gil, R. (2013). Mealybugs nested endosymbiosis: going into the “matryoshka” system in *Planococcus citri* in depth. *BMC Microbiology Environmental Microbiology*, 10(12), 3417–3422. <https://doi.org/10.1111/j.1462-2920.2008.01776.x>
- López-Sánchez, M. J., Neef, A., Peretó, J., Patiño-Navarrete, R., Pignatelli, M., Latorre, A., & Moya, A. (2009). Evolutionary convergence and nitrogen metabolism in *Blattabacterium* strain Bge, primary endosymbiont of the cockroach *Blattella germanica*. *PLoS Genetics*, 5(11), e1000721. <https://doi.org/10.1371/journal.pgen.1000721>
- Luisi, P. L. (2002). Toward the engineering of minimal living cells. In *The Anatomical Record* (Vol. 268, Issue 3, pp. 208–214). <https://doi.org/10.1002/ar.10155>
- Luisi, P. L., Ferri, F., & Stanó, P. (2006). Approaches to semi-synthetic minimal cells: a review. In *Naturwissenschaften* (Vol. 93, Issue 1, pp. 1–13). <https://doi.org/10.1007/s00114-005-0056-z>
- Luisi, P. L., Oberholzer, T., & Lazcano, A. (2002). The notion of a DNA minimal cell: A general discourse and some guidelines for an experimental approach.
- Lumini, E., Bianciotto, V., Jargeat, P., Novero, M., Salvioli, A., Faccio, A., Bécard, G., & Bonfante,

- P. (2007). Presymbiotic growth and sporal morphology are affected in the arbuscular mycorrhizal fungus *Gigaspora margarita* cured of its endobacteria. *Cellular Microbiology*, 9(7), 1716–1729. <https://doi.org/10.1111/j.1462-5822.2007.00907.x>
- Luo, H., Lin, Y., Gao, F., Zhang, C.-T., & Zhang, R. (2014). DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements: Table 1. In *Nucleic Acids Research* (Vol. 42, Issue D1, pp. D574–D580). <https://doi.org/10.1093/nar/gkt1131>
- Mackay, W. J., Han, S., & Samson, L. D. (1994). DNA alkylation repair limits spontaneous base substitution mutations in *Escherichia coli*. *Journal of Bacteriology*, 176(11), 3224–3230. <https://doi.org/10.1128/jb.176.11.3224-3230.1994>
- Mansour, K. (1934). Memoirs: On the intracellular micro-organisms of some bostrychid beetles. *Journal of Cell Science*. <https://journals.biologists.com/jcs/article-abstract/s2-77/306/243/63346>
- Mantri, S., & Tanuj Sapra, K. (2013). Evolving protocells to prototissues: rational design of a missing link. In *Biochemical Society Transactions* (Vol. 41, Issue 5, pp. 1159–1165). <https://doi.org/10.1042/bst20130135>
- Manzano-Marín, A., & 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 Biology and Evolution*, 6(7), 1683–1698. <https://doi.org/10.1093/gbe/evu133>
- Mao, M., Yang, X., & Bennett, G. M. (2018). Evolution of host support for two ancient bacterial symbionts with differentially degraded genomes in a leafhopper host. *Proceedings of the National Academy of Sciences of the United States of America*, 115(50), E11691–E11700. <https://doi.org/10.1073/pnas.1811932115>
- Marais, G. A. B., Calteau, A., & Tenaillon, O. (2008). Mutation rate and genome reduction in endosymbiotic and free-living bacteria. *Genetica*, 134(2), 205–210. <https://doi.org/10.1007/s10709-007-9226-6>
- Margulis, L., Fester, R., & University of Massachusetts Amherst Massachusetts Lynn Margulis. (1991). *Symbiosis as a Source of Evolutionary Innovation: Speciation and Morphogenesis*. MIT Press. <https://market.android.com/details?id=book-3sKzeiHUIUQC>
- Margulis, L., & University of Massachusetts Amherst Massachusetts Lynn Margulis. (1993). *Symbiosis in Cell Evolution: Microbial Communities in the Archean and Proterozoic Eons*. Freeman. <https://play.google.com/store/books/details?id=apuHtAEACAAJ>
- Margulis, L. (1970). *Origin of Eukaryotic Cells: Evidence and Research Implications for a Theory of the Origin and Evolution of Microbial, Plant, and Animal Cells on the Precambrian Earth*. Yale University Press. <https://play.google.com/store/books/details?id=vJFnsWEACAAJ>
- Markowitz, V. M., Chen, I.-M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P., Huntemann, M., Anderson, I., Mavromatis, K., Ivanova, N. N., & Kyrpides, N. C. (2012). IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Research*, 40(Database issue), D115–D122. <https://doi.org/10.1093/nar/gkr1044>
- Martin, B. D., & Schwab, E. (2012). Current usage of symbiosis and associated terminology.

- International Journal of Biology*, 5(1). <https://doi.org/10.5539/ijb.v5n1p32>
- Martínez-Cano, D. J., Reyes-Prieto, M., Martínez-Romero, E., Partida-Martínez, L. P., Latorre, A., Moya, A., & Delaye, L. (2014). Evolution of small prokaryotic genomes. *Frontiers in Microbiology*, 5, 742. <https://doi.org/10.3389/fmicb.2014.00742>
- Martinson, V. G. (2020). Rediscovering a Forgotten System of Symbiosis: Historical Perspective and Future Potential. *Genes*, 11(9). <https://doi.org/10.3390/genes11091063>
- Mazurie, A., Bonchev, D., Schwikowski, B., & Buck, G. A. (2010). Evolution of metabolic network organization. *BMC Systems Biology*, 4, 59. <https://doi.org/10.1186/1752-0509-4-59>
- McCutcheon, J. P. (2010). The bacterial essence of tiny symbiont genomes. *Current Opinion in Microbiology*, 13(1), 73–78. <https://doi.org/10.1016/j.mib.2009.12.002>
- McCutcheon, J. P., Boyd, B. M., & Dale, C. (2019). The Life of an Insect Endosymbiont from the Cradle to the Grave. *Current Biology: CB*, 29(11), R485–R495. <https://doi.org/10.1016/j.cub.2019.03.032>
- McCutcheon, J. P., & Keeling, P. J. (2014). Endosymbiosis: protein targeting further erodes the organelle/symbiont distinction [Review of *Endosymbiosis: protein targeting further erodes the organelle/symbiont distinction*]. *Current Biology: CB*, 24(14), R654–R655. <https://doi.org/10.1016/j.cub.2014.05.073>
- McCutcheon, J. P., McDonald, B. R., & Moran, N. A. (2009a). Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. *PLoS Genetics*, 5(7), e1000565. <https://doi.org/10.1371/journal.pgen.1000565>
- McCutcheon, J. P., McDonald, B. R., & Moran, N. A. (2009b). Convergent evolution of metabolic roles in bacterial co-symbionts of insects. *Proceedings of the National Academy of Sciences of the United States of America*, 106(36), 15394–15399. <https://doi.org/10.1073/pnas.0906424106>
- McCutcheon, J. P., & Moran, N. A. (2007). Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. *Proceedings of the National Academy of Sciences of the United States of America*, 104(49), 19392–19397. <https://doi.org/10.1073/pnas.0708855104>
- McCutcheon, J. P., & Moran, N. A. (2010). Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. *Genome Biology and Evolution*, 2, 708–718. <https://doi.org/10.1093/gbe/evq055>
- McCutcheon, J. P., & Moran, N. A. (2012). Extreme genome reduction in symbiotic bacteria. *Nature Reviews. Microbiology*, 10(1), 13–26. <https://doi.org/10.1038/nrmicro2670>
- McCutcheon, J. P., & von Dohlen, C. D. (2011). An interdependent metabolic patchwork in the nested symbiosis of mealybugs. *Current Biology: CB*, 21(16), 1366–1372. <https://doi.org/10.1016/j.cub.2011.06.051>
- McFadden, G. I. (2001). Chloroplast origin and integration. *Plant Physiology*, 125(1), 50–53. <https://doi.org/10.1104/pp.125.1.50>
- Medina, P., Russell, S. L., Aswadhati, K., & Corbett-Detig, R. (2020). Deep data mining reveals variable abundance and distribution of microbial reproductive manipulators within and among diverse host species. In *bioRxiv* (p. 679837). <https://doi.org/10.1101/679837>
- Medini, D., Donati, C., Tettelin, H., Massignani, V., & Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics & Development*, 15(6), 589–594.

- <https://doi.org/10.1016/j.gde.2005.09.006>
- Meeks, J. C., Elhai, J., Thiel, T., Potts, M., Larimer, F., Lamerdin, J., Predki, P., & Atlas, R. (2001). An overview of the genome of *Nostoc punctiforme*, a multicellular, symbiotic cyanobacterium. *Photosynthesis Research*, 70(1), 85–106.
<https://doi.org/10.1023/A:1013840025518>
- Meincke, L., Copeland, A., Lapidus, A., Lucas, S., Berry, K. W., Del Rio, T. G., Hammon, N., Dalin, E., Tice, H., Pitluck, S., Richardson, P., Bruce, D., Goodwin, L., Han, C., Tapia, R., Detter, J. C., Schmutz, J., Brettin, T., Larimer, F., ... Klenk, H.-P. (2012). Complete genome sequence of *Polynucleobacter necessarius* subsp. *asymbioticus* type strain (QLW-P1DMWA-1(T)). *Standards in Genomic Sciences*, 6(1), 74–83.
<https://doi.org/10.4056/sigs.2395367>
- Mendonça, A. G., Alves, R. J., & Pereira-Leal, J. B. (2011). Loss of genetic redundancy in reductive genome evolution. *PLoS Computational Biology*, 7(2), e1001082.
<https://doi.org/10.1371/journal.pcbi.1001082>
- Miller, S. E., Novotny, V., & Basset, Y. (2002). Case studies of arthropod diversity and distribution. *Foundations of Tropical Forest Biology: Classic Papers with Commentaries*.
https://repository.si.edu/bitstream/handle/10088/3514/Miller_Novotny_and_Basset_2002_classics.pdf
- Mira, A., Ochman, H., & Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends in Genetics: TIG*, 17(10), 589–596.
[https://doi.org/10.1016/s0168-9525\(01\)02447-7](https://doi.org/10.1016/s0168-9525(01)02447-7)
- Mithani, A., Preston, G. M., & Hein, J. (2010). A Bayesian approach to the evolution of metabolic networks on a phylogeny. *PLoS Computational Biology*, 6(8).
<https://doi.org/10.1371/journal.pcbi.1000868>
- Mitter, C., Farrell, B., & Wiegmann, B. (1988). The Phylogenetic Study of Adaptive Zones: Has Phytophagy Promoted Insect Diversification? *The American Naturalist*, 132(1), 107–128.
<https://doi.org/10.1086/284840>
- Mollenhauer, D., Mollenhauer, R., & Kluge, M. (1996). Studies on initiation and development of the partner association in *Geosiphon pyriforme* (Kütz.) v. Wettstein, a unique endocytobiotic system of a fungus (Glomales) and the cyanobacterium *Nostoc punctiforme* (Kütz.) Hariot. *Protoplasma*, 193(1-4), 3–9. <https://doi.org/10.1007/bf01276630>
- Monterroso, A. (1959). *Obras completas (y otros cuentos)*. Ediciones Era.
<https://play.google.com/store/books/details?id=duEGqerNwS4C>
- Morales, J., Kokkori, S., Weidauer, D., Chapman, J., Goltsman, E., Rokhsar, D., Grossman, A. R., & Nowack, E. C. M. (2016). Development of a toolbox to dissect host-endosymbiont interactions and protein trafficking in the trypanosomatid *Angomonas deanei*. *BMC Evolutionary Biology*, 16(1), 247. <https://doi.org/10.1186/s12862-016-0820-z>
- Moran, N. A. (1996). Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 93(7), 2873–2878.
<https://doi.org/10.1073/pnas.93.7.2873>
- Moran, N. A. (2003). Tracing the evolution of gene loss in obligate bacterial symbionts. *Current*

- Opinion in Microbiology*, 6(5), 512–518. <https://doi.org/10.1016/j.mib.2003.08.001>
- Moran, N. A. (2007). Symbiosis as an adaptive process and source of phenotypic complexity. *Proceedings of the National Academy of Sciences of the United States of America*, 104 Suppl 1, 8627–8633. <https://doi.org/10.1073/pnas.0611659104>
- Moran, N. A., & Bennett, G. M. (2014). The tiniest tiny genomes. *Annual Review of Microbiology*, 68, 195–215. <https://doi.org/10.1146/annurev-micro-091213-112901>
- Moran, N. A., McCutcheon, J. P., & Nakabachi, A. (2008). Genomics and evolution of heritable bacterial symbionts. *Annual Review of Genetics*, 42, 165–190. <https://doi.org/10.1146/annurev.genet.41.110306.130119>
- Moran, N. A., McLaughlin, H. J., & Sorek, R. (2009). The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* (5912), 379–382. <https://doi.org/10.1126/science.1167140>
- Moran, N. A., & Mira, A. (2001). The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. *Genome Biology*, 2(12), RESEARCH0054. <https://doi.org/10.1186/gb-2001-2-12-research0054>
- Moran, N. A., Plague, G. R., Sandström, J. P., & Wilcox, J. L. (2003). A genomic perspective on nutrient provisioning by bacterial symbionts of insects. In *Proceedings of the National Academy of Sciences* (Vol. 100, Issue suppl_2, pp. 14543–14548). <https://doi.org/10.1073/pnas.2135345100>
- Moran, N. A., & Telang, A. (1998). Bacteriocyte-associated symbionts of insects. *Bioscience*, 48(4), 295–304.
- Morowitz, H. J. (1993). *Beginnings of Cellular Life: Metabolism Recapitulates Biogenesis*. Yale University Press. https://play.google.com/store/books/details?id=CmQDSHN_UrIC
- Morris, J. J., Lenski, R. E., & Zinser, E. R. (2012). The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *mBio*, 3(2). <https://doi.org/10.1128/mBio.00036-12>
- Moya, A., Gil, R., Latorre, A., Peretó, J., Garcillán-Barcia, M. P., & De La Cruz, F. (2009). Toward minimal bacterial cells: evolution vs. design. In *FEMS Microbiology Reviews* (Vol. 33, Issue 1, pp. 225–235). <https://doi.org/10.1111/j.1574-6976.2008.00151.x>
- Moya, A., Peretó, J., Gil, R., & Latorre, A. (2008). Learning how to live together: genomic insights into prokaryote–animal symbioses. *Nature Reviews. Genetics*, 9, 218. <https://doi.org/10.1038/nrg2319>
- Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Verezhemska, O., Isbandi, M., Thomas, A. D., Ali, R., Sharma, K., Kyrpides, N. C., & Reddy, T. B. K. (2017). Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Research*, 45(D1), D446–D456. <https://doi.org/10.1093/nar/gkw992>
- Mushegian, A. R., & Koonin, E. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 93(19), 10268–10273. <https://doi.org/10.1073/pnas.93.19.10268>
- Nakabachi, A., Ishida, K., Hongoh, Y., Ohkuma, M., & Miyagishima, S.-Y. (2014). Aphid gene of bacterial origin encodes a protein transported to an obligate endosymbiont. *Current Biology: CB*, 24(14), R640–R641. <https://doi.org/10.1016/j.cub.2014.06.038>

- Nakabachi, A., Ueoka, R., Oshima, K., Teta, R., Mangoni, A., Gurgui, M., Oldham, N. J., van Echten-Deckert, G., Okamura, K., Yamamoto, K., Inoue, H., Ohkuma, M., Hongoh, Y., Miyagishima, S.-Y., Hattori, M., Piel, J., & Fukatsu, T. (2013). Defensive bacteriome symbiont with a drastically reduced genome. *Current Biology: CB*, 23(15), 1478–1484. <https://doi.org/10.1016/j.cub.2013.06.027>
- Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H. E., Moran, N. A., & Hattori, M. (2006). The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science*, 314(5797), 267. <https://doi.org/10.1126/science.1134196>
- Naumann, M., Schüssler, A., & Bonfante, P. (2010). The obligate endobacteria of arbuscular mycorrhizal fungi are ancient heritable components related to the Mollicutes. *The ISME Journal*, 4(7), 862–871. <https://doi.org/10.1038/ismej.2010.21>
- NCBI Resource Coordinators. (2018). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 46(D1), D8–D13. <https://doi.org/10.1093/nar/gkx1095>
- Nikoh, N., McCutcheon, J. P., Kudo, T., Miyagishima, S.-Y., Moran, N. A., & Nakabachi, A. (2010). Bacterial genes in the aphid genome: absence of functional gene transfer from *Buchnera* to its host. *PLoS Genetics*, 6(2), e1000827. <https://doi.org/10.1371/journal.pgen.1000827>
- Nikoh, N., & Nakabachi, A. (2009). Aphids acquired symbiotic genes via lateral gene transfer. *BMC Biology*, 7, 12. <https://doi.org/10.1186/1741-7007-7-12>
- Nilsson, A. I., Koskiniemi, S., Eriksson, S., Kugelberg, E., Hinton, J. C. D., & Andersson, D. I. (2005). Bacterial genome size reduction by experimental evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 102(34), 12112–12116. <https://doi.org/10.1073/pnas.0503654102>
- Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., Smirnova, T., Grigoriev, I. V., & Dubchak, I. (2014). The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Research*, 42(Database issue), D26–D31. <https://doi.org/10.1093/nar/gkt1069>
- Nowack, E. C. M. (2014). Paulinella chromatophora – rethinking the transition from endosymbiont to organelle. In *Acta Societatis Botanicorum Poloniae* (Vol. 83, Issue 4, pp. 387–397). <https://doi.org/10.5586/asbp.2014.049>
- Nowack, E. C. M., Melkonian, M., & Glöckner, G. (2008). Chromatophore genome sequence of *Paulinella* sheds light on acquisition of photosynthesis by eukaryotes. *Current Biology: CB*, 18(6), 410–418. <https://doi.org/10.1016/j.cub.2008.02.051>
- Nowack, E. C. M., Price, D. C., Bhattacharya, D., Singer, A., Melkonian, M., & Grossman, A. R. (2016). Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of *Paulinella chromatophora*. *Proceedings of the National Academy of Sciences of the United States of America*, 113(43), 12214–12219. <https://doi.org/10.1073/pnas.1608016113>
- Oakeson, K. F., Gil, R., Clayton, A. L., Dunn, D. M., von Niederhausern, A. C., Hamil, C., Aoyagi, A., Duval, B., Baca, A., Silva, F. J., Vallier, A., Jackson, D. G., Latorre, A., Weiss, R.

- B., Heddi, A., Moya, A., & Dale, C. (2014). Genome degeneration and adaptation in a nascent stage of symbiosis. *Genome Biology and Evolution*, 6(1), 76–93.
<https://doi.org/10.1093/gbe/evt210>
- Oliver, K. M., Russell, J. A., Moran, N. A., & Hunter, M. S. (2003). Facultative bacterial symbionts in aphids confer resistance to parasitic wasps. *Proceedings of the National Academy of Sciences of the United States of America*, 100(4), 1803–1807.
<https://doi.org/10.1073/pnas.0335320100>
- Osawa, S., & Jukes, T. H. (1989). Codon reassignment (codon capture) in evolution. *Journal of Molecular Evolution*, 28(4), 271–278. <https://doi.org/10.1007/BF02103422>
- Ouzounis, C. A., Kunin, V., Darzentas, N., & Goldovsky, L. (2006). A minimal estimate for the gene content of the last universal common ancestor--exobiology from a terrestrial perspective. *Research in Microbiology*, 157(1), 57–68.
<https://doi.org/10.1016/j.resmic.2005.06.015>
- Pagani, I., Liolios, K., Jansson, J., Chen, I.-M. A., Smirnova, T., Nosrat, B., Markowitz, V. M., & Kyrpides, N. C. (2012). The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, 40(Database issue), D571–D579. <https://doi.org/10.1093/nar/gkr1100>
- Pagel, M., & Meade, A. (n.d.). BayesTraits v. 2.0. Reading: University of Reading.
- Partensky, F., Blanchot, J., & Vaultot, D. (1999). Differential distribution and ecology of Prochlorococcus and Synechococcus in oceanic waters: a review. *Bulletin-Institut Oceanographique Monaco-Numero Special-*, 457–476.
https://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers15-02/010019783.pdf#page=469
- Parter, M., Kashtan, N., & Alon, U. (2007). Environmental variability and modularity of bacterial metabolic networks. *BMC Evolutionary Biology*, 7, 169.
<https://doi.org/10.1186/1471-2148-7-169>
- Partida-Martinez, L. P., de Looß, C. F., Ishida, K., Ishida, M., Roth, M., Buder, K., & Hertweck, C. (2007). Rhizonin, the First Mycotoxin Isolated from the Zygomycota, Is Not a Fungal Metabolite but Is Produced by Bacterial Endosymbionts. In *Applied and Environmental Microbiology* (Vol. 73, Issue 3, pp. 793–797). <https://doi.org/10.1128/aem.01784-06>
- Partida-Martinez, L. P., Groth, I., Schmitt, I., Richter, W., Roth, M., & Hertweck, C. (2007). Burkholderia rhizoxinica sp. nov. and Burkholderia endofungorum sp. nov., bacterial endosymbionts of the plant-pathogenic fungus Rhizopus microsporus. *International Journal of Systematic and Evolutionary Microbiology*, 57(Pt 11), 2583–2590.
<https://doi.org/10.1099/ijs.0.64660-0>
- Partida-Martinez, L. P., & Hertweck, C. (2005). Pathogenic fungus harbours endosymbiotic bacteria for toxin production. *Nature*, 437(7060), 884–888.
<https://doi.org/10.1038/nature03997>
- Partida-Martinez, L. P., & Hertweck, C. (2007). A gene cluster encoding rhizoxin biosynthesis in “Burkholderia rhizoxina”, the bacterial endosymbiont of the fungus Rhizopus microsporus. *Chembiochem: A European Journal of Chemical Biology*, 8(1), 41–45.

<https://doi.org/10.1002/cbic.200600393>

- Partida-Martinez, L. P., Monajembashi, S., Greulich, K.-O., & Hertweck, C. (2007). Endosymbiont-Dependent Host Reproduction Maintains Bacterial-Fungal Mutualism. In *Current Biology* (Vol. 17, Issue 9, pp. 773–777). <https://doi.org/10.1016/j.cub.2007.03.039>
- Patil, P. B., Zeng, Y., Coursey, T., Houston, P., Miller, I., & Chen, S. (2010). Isolation and characterization of a *Nocardiopsis* sp. from honeybee guts. *FEMS Microbiology Letters*, 312(2), 110–118. <https://doi.org/10.1111/j.1574-6968.2010.02104.x>
- Patiño-Navarrete, R., Piulachs, M.-D., Belles, X., Moya, A., Latorre, A., & Peretó, J. (2014). The cockroach *Blattella germanica* obtains nitrogen from uric acid through a metabolic pathway shared with its bacterial endosymbiont. *Biology Letters*, 10(7), 20140407. <https://doi.org/10.1098/rsbl.2014.0407>
- Pérez-Brocal, V., Gil, R., Ramos, S., Lamelas, A., Postigo, M., Michelena, J. M., Silva, F. J., Moya, A., & Latorre, A. (2006). A small microbial genome: the end of a long symbiotic relationship? *Science*, 314(5797), 312–313. <https://doi.org/10.1126/science.1130441>
- Podar, M., Anderson, I., Makarova, K. S., Elkins, J. G., Ivanova, N., Wall, M. A., Lykidis, A., Mavromatis, K., Sun, H., Hudson, M. E., Chen, W., Deciu, C., Hutchison, D., Eads, J. R., Anderson, A., Fernandes, F., Szeto, E., Lapidus, A., Kyrpides, N. C., ... Stetter, K. O. (2008). A genomic analysis of the archaeal system *Ignicoccus hospitalis*-*Nanoarchaeum equitans*. *Genome Biology*, 9(11), R158. <https://doi.org/10.1186/gb-2008-9-11-r158>
- Poliakov, A., Russell, C. W., Ponnala, L., Hoops, H. J., Sun, Q., Douglas, A. E., & van Wijk, K. J. (2011). Large-scale label-free quantitative proteomics of the pea aphid-*Buchnera* symbiosis. *Molecular & Cellular Proteomics: MCP*, 10(6), M110.007039. <https://doi.org/10.1074/mcp.M110.007039>
- Ponce-de-Leon, M., Tamarit, D., Calle-Espinosa, J., Mori, M., Latorre, A., Montero, F., & Pereto, J. (2017). Determinism and Contingency Shape Metabolic Complementation in an Endosymbiotic Consortium. In *Frontiers in Microbiology* (Vol. 8). <https://doi.org/10.3389/fmicb.2017.02290>
- Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. *PloS One*, 5(3), e9490. <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0009490>
- Promnuan, Y., Kudo, T., & Chantawannakul, P. (2009). Actinomycetes isolated from beehives in Thailand. In *World Journal of Microbiology and Biotechnology* (Vol. 25, Issue 9, pp. 1685–1689). <https://doi.org/10.1007/s11274-009-0051-1>
- Qiao, J., Chen, L., Li, Y., Wang, J., Zhang, W., & Chen, S. (2012). Whole-genome sequence of *Nocardiopsis alba* strain ATCC BAA-2165, associated with honeybees. *Journal of Bacteriology*, 194(22), 6358–6359. <https://doi.org/10.1128/JB.01522-12>
- Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(Database issue), D590–D596. <https://doi.org/10.1093/nar/gks1219>
- Raffa, K. F., Aukema, B. H., Bentz, B. J., Carroll, A. L., Hicke, J. A., Turner, M. G., & Romme, W.

- H. (2008). Cross-scale Drivers of Natural Disturbances Prone to Anthropogenic Amplification: The Dynamics of Bark Beetle Eruptions. *Bioscience*, 58(6), 501–517. <https://doi.org/10.1641/B580607>
- Ran, L., Larsson, J., Vigil-Stenman, T., Nylander, J. A. A., Ininbergs, K., Zheng, W.-W., Lapidus, A., Lowry, S., Haselkorn, R., & Bergman, B. (2010). Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. *PLoS One*, 5(7), e11486. <https://doi.org/10.1371/journal.pone.0011486>
- Rao, Q., Rollat-Farnier, P.-A., Zhu, D.-T., Santos-Garcia, D., Silva, F. J., Moya, A., Latorre, A., Klein, C. C., Vavre, F., Sagot, M.-F., Liu, S.-S., Mouton, L., & Wang, X.-W. (2015). Genome reduction and potential metabolic complementation of the dual endosymbionts in the whitefly *Bemisia tabaci*. *BMC Genomics*, 16(1), 226. <https://doi.org/10.1186/s12864-015-1379-6>
- Raven, P. H., & Wagner, D. L. (2021). Agricultural intensification and climate change are rapidly decreasing insect biodiversity. *Proceedings of the National Academy of Sciences*, 118(2), e2002548117. <https://doi.org/10.1073/pnas.2002548117>
- Reyes-Prieto, M., Gil, R., Llabrés, M., Palmer-Rodríguez, P., & Moya, A. (2020). The Metabolic Building Blocks of a Minimal Cell. *Biology*, 10(1). <https://doi.org/10.3390/biology10010005>
- Reyes-Prieto, M., Latorre, A., & Moya, A. (2014). Scanty microbes, the “symbionelle” concept. *Environmental Microbiology*, 16(2), 335–338. <https://doi.org/10.1111/1462-2920.12220>
- Reyes-Prieto, M., Vargas-Chávez, C., Latorre, A., & Moya, A. (2015). SymbioGenomesDB: a database for the integration and access to knowledge on host-symbiont relationships. *Database: The Journal of Biological Databases and Curation*, 2015. <https://doi.org/10.1093/database/bav109>
- Reyes-Prieto, M., Vargas-Chávez, C., Llabrés, M., Palmer, P., Latorre, A., & Moya, A. (2020). An update on the Symbiotic Genomes Database (SymGenDB): a collection of metadata, genomic, genetic and protein sequences, orthologs and metabolic networks of symbiotic organisms. *Database: The Journal of Biological Databases and Curation*, 2020. <https://doi.org/10.1093/database/baz160>
- Rippka, R. (1992). Pasteur culture collection of cyanobacterial Strains in axenic culture. *Catalogue and Taxonomic Handbook, Catalogue of Strains 1992/1993*, 1, 1–103. <https://cir.nii.ac.jp/crid/1571980075695210752>
- Rispe, C., Delmotte, F., van Ham, R. C. H. J., & Moya, A. (2004). Mutational and selective pressures on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids. *Genome Research*, 14(1), 44–53. <https://doi.org/10.1101/gr.1358104>
- Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., Arellano, A., Coleman, M., Hauser, L., Hess, W. R., Johnson, Z. I., Land, M., Lindell, D., Post, A. F., Regala, W., Shah, M., Shaw, S. L., Steglich, C., Sullivan, M. B., ... Chisholm, S. W. (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature*, 424(6952), 1042–1047. <https://doi.org/10.1038/nature01947>
- Rogel, M. A., Ormeño-Orrillo, E., & Martínez Romero, E. (2011). Symbiovars in rhizobia reflect bacterial adaptation to legumes. *Systematic and Applied Microbiology*, 34(2), 96–104. <https://doi.org/10.1016/j.syapm.2010.11.015>

- Romano, A. H., & Conway, T. (1996). Evolution of carbohydrate metabolic pathways. In *Research in Microbiology* (Vol. 147, Issues 6-7, pp. 448–455).
[https://doi.org/10.1016/0923-2508\(96\)83998-2](https://doi.org/10.1016/0923-2508(96)83998-2)
- Ronquist, F., & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12), 1572–1574.
<https://www.ncbi.nlm.nih.gov/pubmed/12912839>
- Rosas-Pérez, T., Rosenblueth, M., Rincón-Rosales, R., Mora, J., & Martínez-Romero, E. (2014). Genome sequence of “Candidatus Walczuchella monophlebidarum” the flavobacterial endosymbiont of *Llaveia axin axin* (Hemiptera: Coccoidea: Monophlebidae). *Genome Biology and Evolution*, 6(3), 714–726. <https://doi.org/10.1093/gbe/evu049>
- Rosenblueth, M., Sayavedra, L., Sámano-Sánchez, H., Roth, A., & Martínez-Romero, E. (2012). Evolutionary relationships of flavobacterial and enterobacterial endosymbionts with their scale insect hosts (Hemiptera: Coccoidea). *Journal of Evolutionary Biology*, 25(11), 2357–2368.
<https://doi.org/10.1111/j.1420-9101.2012.02611.x>
- RStudio, I. (2013). shiny: web application framework for R. *R Package*.
- Russell, C. W., Bouvaine, S., Newell, P. D., & Douglas, A. E. (2013). Shared metabolic pathways in a coevolved insect-bacterial symbiosis. *Applied and Environmental Microbiology*, 79(19), 6117–6123. <https://doi.org/10.1128/AEM.01543-13>
- Sabater-Muñoz, B., Toft, C., Alvarez-Ponce, D., & Fares, M. A. (2017). Chance and necessity in the genome evolution of endosymbiotic bacteria of insects. *The ISME Journal*, 11(6), 1291–1304. <https://doi.org/10.1038/ismej.2017.18>
- Sabree, Z. L., Huang, C. Y., Okusu, A., Moran, N. A., & Normark, B. B. (2013). The nutrient supplying capabilities of *Uzinura*, an endosymbiont of armoured scale insects. *Environmental Microbiology*, 15(7), 1988–1999. <https://doi.org/10.1111/1462-2920.12058>
- Sachs, J. L., & Hollowell, A. C. (2012). The Origins of Cooperative Bacterial Communities. In *mBio* (Vol. 3, Issue 3). <https://doi.org/10.1128/mbio.00099-12>
- Saffo, M. B. (1993). Coming to Terms with a Field-Words and Concepts in Symbiosis (Vol 14, Pg 29, 1993). In *Symbiosis* (Vol. 15, Issues 1-2, pp. 181–181). INT SCIENCE SERVICES/BALABAN PUBLISHERS PO BOX 2039, REHOVOT, ISRAEL.
- Said, M. R., Begley, T. J., Oppenheim, A. V., Lauffenburger, D. A., & Samson, L. D. (2004). Global network analysis of phenotypic effects: protein networks and toxicity modulation in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America*, 101(52), 18006–18011. <https://doi.org/10.1073/pnas.0405996101>
- Salvioli, A., Lumini, E., Anca, I. A., Bianciotto, V., & Bonfante, P. (2008). Simultaneous detection and quantification of the unculturable microbe *Candidatus Glomeribacter gigasporarum* inside its fungal host *Gigaspora margarita*. *The New Phytologist*, 180(1), 248–257.
<https://doi.org/10.1111/j.1469-8137.2008.02541.x>
- Sánchez-Baracaldo, P., Raven, J. A., Pisani, D., & Knoll, A. H. (2017). Early photosynthetic eukaryotes inhabited low-salinity habitats. In *Proceedings of the National Academy of Sciences* (Vol. 114, Issue 37). <https://doi.org/10.1073/pnas.1620089114>
- Santos-Garcia Diego, Farnier Pierre-Antoine, Beitia Francisco, Zchori-Fein Einat, Vavre Fabrice,

- Mouton Laurence, Moya Andrés, Latorre Amparo, & Silva Francisco J. (2012). Complete Genome Sequence of “Candidatus Portiera aleyrodidarum” BT-QVLC, an Obligate Symbiont That Supplies Amino Acids and Carotenoids to Bemisia tabaci. *Journal of Bacteriology*, 194(23), 6654–6655. <https://doi.org/10.1128/JB.01793-12>
- Santos-Garcia, D., Rollat-Farnier, P.-A., Beitia, F., Zchori-Fein, E., Vavre, F., Mouton, L., Moya, A., Latorre, A., & Silva, F. J. (2014). The genome of Cardinium cBtQ1 provides insights into genome reduction, symbiont motility, and its settlement in Bemisia tabaci. *Genome Biology and Evolution*, 6(4), 1013–1030. <https://doi.org/10.1093/gbe/evu077>
- Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., ... Ye, J. (2009). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 37(Database issue), D5–D15. <https://doi.org/10.1093/nar/gkn741>
- Scanlan, D. J., Ostrowski, M., Mazard, S., Dufresne, A., Garczarek, L., Hess, W. R., Post, A. F., Hagemann, M., Paulsen, I., & Partensky, F. (2009). Ecological genomics of marine picocyanobacteria. *Microbiology and Molecular Biology Reviews: MMBR*, 73(2), 249–299. <https://doi.org/10.1128/MMBR.00035-08>
- Scannerini, S., Smith, D., Bonfante-Fasolo, P., & Gianinazzi-Pearson, V. (2013). *Cell to Cell Signals in Plant, Animal and Microbial Symbiosis*. Springer Science & Business Media. <https://play.google.com/store/books/details?id=Vg7pCAAQBAJ>
- Scherlach, K., Partida-Martinez, L. P., Dahse, H.-M., & Hertweck, C. (2006). Antimitotic rhizoxin derivatives from a cultured bacterial endosymbiont of the rice pathogenic fungus Rhizopus microsporus. *Journal of the American Chemical Society*, 128(35), 11529–11536. <https://doi.org/10.1021/ja062953o>
- Schmitt, I., Partida-Martinez, L. P., Winkler, R., Voigt, K., Einax, E., Dölz, F., Telle, S., Wöstemeyer, J., & Hertweck, C. (2008). Evolution of host resistance in a toxin-producing bacterial–fungal alliance. *The ISME Journal*, 2(6), 632–641. <https://doi.org/10.1038/ismej.2008.19>
- Schmitz-Esser, S., Penz, T., Spang, A., & 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 Evolutionary Biology*, 11, 270. <https://doi.org/10.1186/1471-2148-11-270>
- Schmitz-Esser, S., Tischler, P., Arnold, R., Montanaro, J., Wagner, M., Rattei, T., & Horn, M. (2010). The genome of the amoeba symbiont “Candidatus Amoebophilus asiaticus” reveals common mechanisms for host cell interaction among amoeba-associated bacteria. *Journal of Bacteriology*, 192(4), 1045–1057. <https://doi.org/10.1128/JB.01379-09>
- Segata, N., Börnigen, D., Morgan, X. C., & Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature Communications*, 4, 2304. <https://doi.org/10.1038/ncomms3304>
- Selzer, P. M., Marhöfer, R. J., & Koch, O. (2018). Biological Databases. In P. M. Selzer, R. J. Marhöfer, & O. Koch (Eds.), *Applied Bioinformatics: An Introduction* (pp. 13–34). Springer

- International Publishing. https://doi.org/10.1007/978-3-319-68301-0_2
- Sender, R., Fuchs, S., & Milo, R. (2016). Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. *Cell*, 164(3), 337–340. <https://doi.org/10.1016/j.cell.2016.01.013>
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), 2498–2504. <https://doi.org/10.1101/gr.1239303>
- Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., & Ishikawa, H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature*, 407(6800), 81–86. <https://doi.org/10.1038/35024074>
- Skidmore, I. H., & Hansen, A. K. (2017). The evolutionary development of plant-feeding insects and their nutritional endosymbionts. In *Insect Science* (Vol. 24, Issue 6, pp. 910–928). <https://doi.org/10.1111/1744-7917.12463>
- Sloan, D. B., & Moran, N. A. (2012a). Genome reduction and co-evolution between the primary and secondary bacterial symbionts of psyllids. *Molecular Biology and Evolution*, 29(12), 3781–3792. <https://doi.org/10.1093/molbev/mss180>
- Sloan, D. B., & Moran, N. A. (2012b). Endosymbiotic bacteria as a source of carotenoids in whiteflies. *Biology Letters*, 8(6), 986–989. <https://doi.org/10.1098/rsbl.2012.0664>
- Sloan, D. B., & Moran, N. A. (2013). The evolution of genomic instability in the obligate endosymbionts of whiteflies. *Genome Biology and Evolution*, 5(5), 783–793. <https://doi.org/10.1093/gbe/evt044>
- Sloan, D. B., Nakabachi, A., Richards, S., Qu, J., Murali, S. C., Gibbs, R. A., & Moran, N. A. (2014). Parallel histories of horizontal gene transfer facilitated extreme reduction of endosymbiont genomes in sap-feeding insects. *Molecular Biology and Evolution*, 31(4), 857–871. <https://doi.org/10.1093/molbev/msu004>
- Spies, M. (2013). There and back again: new single-molecule insights in the motion of DNA repair proteins. *Current Opinion in Structural Biology*, 23(1), 154–160. <https://doi.org/10.1016/j.sbi.2012.11.008>
- Stano, P. (2018). Is Research on “Synthetic Cells” Moving to the Next Level? In *Life* (Vol. 9, Issue 1, p. 3). <https://doi.org/10.3390/life9010003>
- Stano, P., & Luisi, P. L. (2011). On the Construction of Minimal Cell Models in Synthetic Biology and Origins of Life Studies. In *Design and Analysis of Biomolecular Circuits* (pp. 337–368). https://doi.org/10.1007/978-1-4419-6766-4_16
- Steele, J. A., Countway, P. D., Xia, L., Vigil, P. D., Beman, J. M., Kim, D. Y., Chow, C.-E. T., Sachdeva, R., Jones, A. C., Schwalbach, M. S., Rose, J. M., Hewson, I., Patel, A., Sun, F., Caron, D. A., & Fuhrman, J. A. (2011). Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *The ISME Journal*, 5(9), 1414–1425. <https://doi.org/10.1038/ismej.2011.24>
- Steindler, L., Schwalbach, M. S., Smith, D. P., Chan, F., & Giovannoni, S. J. (2011). Energy starved *Candidatus Pelagibacter ubique* substitutes light-mediated ATP production for

- endogenous carbon respiration. *PloS One*, 6(5), e19725.
<https://doi.org/10.1371/journal.pone.0019725>
- Sudakaran, S., Kost, C., & Kaltenpoth, M. (2017). Symbiont Acquisition and Replacement as a Source of Ecological Innovation. *Trends in Microbiology*, 25(5), 375–390.
<https://doi.org/10.1016/j.tim.2017.02.014>
- Szklarczyk, T., & Michalik, A. (2017). Transovarial Transmission of Symbionts in Insects. *Results and Problems in Cell Differentiation*, 63, 43–67.
https://doi.org/10.1007/978-3-319-60855-6_3
- Takahashi, M., Iwasaki, S., Kobayashi, H., Okuda, S., Murai, T., Sato, Y., Haraguchi-Hiraoka, T., & Nagano, H. (1987). Studies on macrocyclic lactone antibiotics. XI. Anti-mitotic and anti-tubulin activity of new antitumor antibiotics, rhizoxin and its homologues. In *The Journal of Antibiotics* (Vol. 40, Issue 1, pp. 66–72). <https://doi.org/10.7164/antibiotics.40.66>
- Tamames, J., Gil, R., Latorre, A., Peretó, J., Silva, F. J., & Moya, A. (2007). The frontier between cell and organelle: genome analysis of *Candidatus Carsonella ruddii*. In *BMC Evolutionary Biology* (Vol. 7, Issue 1, p. 181). <https://doi.org/10.1186/1471-2148-7-181>
- Tamas, I., Klasson, L., Canbäck, B., Näslund, A. K., Eriksson, A.-S., Wernegreen, J. J., Sandström, J. P., Moran, N. A., & Andersson, S. G. E. (2002). 50 million years of genomic stasis in endosymbiotic bacteria. *Science*, 296(5577), 2376–2379.
<https://doi.org/10.1126/science.1071278>
- Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., & Natale, D. A. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4, 41.
<https://doi.org/10.1186/1471-2105-4-41>
- Team, R. D. C. (2009). A language and environment for statistical computing.
<http://www.R-Project.org>. <https://cir.nii.ac.jp/crid/1570854175843385600>
- Team, R. D. C., & Others. (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Tettelin, H. (2020). *The Pangenome*. Springer Nature.
<https://play.google.com/store/books/details?id=hcfgDwAAQBAJ>
- Toft, C., & Andersson, S. G. E. (2010). Evolutionary microbial genomics: insights into bacterial host adaptation. *Nature Reviews. Genetics*, 11(7), 465–475. <https://doi.org/10.1038/nrg2798>
- Trevors, J. T., & Masson, L. (2011). How much cytoplasm can a bacterial genome control? *Journal of Microbiological Methods*, 84(1), 147–150. <https://doi.org/10.1016/j.mimet.2010.11.009>
- Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The human microbiome project. *Nature*, 449(7164), 804–810.
<https://doi.org/10.1038/nature06244>
- Uchiyama, I., Higuchi, T., & Kawai, M. (2010). MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. *Nucleic Acids Research*, 38(Database issue), D361–D365. <https://doi.org/10.1093/nar/gkp948>
- Uchiyama, I., Mihara, M., Nishide, H., & Chiba, H. (2015). MBGD update 2015: microbial

- genome database for flexible ortholog analysis utilizing a diverse set of genomic data. *Nucleic Acids Research*, 43(Database issue), D270–D276. <https://doi.org/10.1093/nar/gku1152>
- Uchiyama, I., Mihara, M., Nishide, H., Chiba, H., & Kato, M. (2019). MBGD update 2018: microbial genome database based on hierarchical orthology relations covering closely related and distantly related comparisons. *Nucleic Acids Research*, 47(D1), D382–D389. <https://doi.org/10.1093/nar/gky1054>
- UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1), D506–D515. <https://doi.org/10.1093/nar/gky1049>
- van Ham, R. C. H. J., Kamerbeek, J., Palacios, C., Rausell, C., Abascal, F., Bastolla, U., Fernández, J. M., Jiménez, L., Postigo, M., Silva, F. J., Tamames, J., Viguera, E., Latorre, A., Valencia, A., Morán, F., & Moya, A. (2003). Reductive genome evolution in *Buchnera aphidicola*. *Proceedings of the National Academy of Sciences of the United States of America*, 100(2), 581–586. <https://doi.org/10.1073/pnas.0235981100>
- Van Leuven, J. T., & McCutcheon, J. P. (2012). An AT mutational bias in the tiny GC-rich endosymbiont genome of *Hodgkinia*. *Genome Biology and Evolution*, 4(1), 24–27. <https://doi.org/10.1093/gbe/evr125>
- Van Leuven, J. T., Meister, R. C., Simon, C., & McCutcheon, J. P. (2014). Sympatric speciation in a bacterial endosymbiont results in two genomes with the functionality of one. *Cell*, 158(6), 1270–1280. <https://doi.org/10.1016/j.cell.2014.07.047>
- Venetz, J. E., Del Medico, L., Wölfle, A., Schächle, P., Bucher, Y., Appert, D., Tschan, F., Flores-Tinoco, C. E., van Kooten, M., Guennoun, R., Deutsch, S., Christen, M., & Christen, B. (2019). Chemical synthesis rewriting of a bacterial genome to achieve design flexibility and biological functionality. In *Proceedings of the National Academy of Sciences* (Vol. 116, Issue 16, pp. 8070–8079). <https://doi.org/10.1073/pnas.1818259116>
- Vigneron, A., Masson, F., Vallier, A., Balmand, S., Rey, M., Vincent-Monégat, C., Aksoy, E., Aubailly-Giraud, E., Zaidman-Rémy, A., & Heddi, A. (2014). Insects Recycle Endosymbionts when the Benefit Is Over. In *Current Biology* (Vol. 24, Issue 19, pp. 2267–2273). <https://doi.org/10.1016/j.cub.2014.07.065>
- Wang, B., Yao, M., Lv, L., Ling, Z., & Li, L. (2017). The Human Microbiota in Health and Disease. *Proceedings of the Estonian Academy of Sciences: Engineering*, 3(1), 71–82. <https://doi.org/10.1016/J.ENG.2017.01.008>
- Weinert, L. A., Araujo-Jnr, E. V., Ahmed, M. Z., & Welch, J. J. (2015). The incidence of bacterial endosymbionts in terrestrial arthropods. In *Proceedings of the Royal Society B: Biological Sciences* (Vol. 282, Issue 1807, p. 20150249). <https://doi.org/10.1098/rspb.2015.0249>
- Wernegreen, J. J. (2005). For better or worse: genomic consequences of intracellular mutualism and parasitism. *Current Opinion in Genetics & Development*, 15(6), 572–583. <https://doi.org/10.1016/j.gde.2005.09.013>
- Wernegreen, J. J. (2015). Endosymbiont evolution: predictions from theory and surprises from genomes. *Annals of the New York Academy of Sciences*, 1360(1), 16–35. <https://doi.org/10.1111/nyas.12740>
- Wernegreen, J. J. (2017). In it for the long haul: evolutionary consequences of persistent

- endosymbiosis. *Current Opinion in Genetics & Development*, 47, 83–90.
<https://doi.org/10.1016/j.gde.2017.08.006>
- Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L. Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Miller, V., ... Yaschenko, E. (2007). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 35(Database issue), D5–D12.
<https://doi.org/10.1093/nar/gkl1031>
- Wigley, D. B. (2013). Bacterial DNA repair: recent insights into the mechanism of RecBCD, AddAB and AdnAB. *Nature Reviews. Microbiology*, 11(1), 9–13.
<https://doi.org/10.1038/nrmicro2917>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>
- Wilkinson, T. L., Koga, R., & Fukatsu, T. (2007). Role of host nutrition in symbiont regulation: impact of dietary nitrogen on proliferation of obligate and facultative bacterial endosymbionts of the pea aphid *Acyrtosiphon pisum*. *Applied and Environmental Microbiology*, 73(4), 1362–1366. <https://doi.org/10.1128/AEM.01211-06>
- Williams, K. P., Sobral, B. W., & Dickerman, A. W. (2007). A robust species tree for the alphaproteobacteria. *Journal of Bacteriology*, 189(13), 4578–4586.
<https://doi.org/10.1128/JB.00269-07>
- Windsor, H. M., Windsor, G. D., & Noordergraaf, J. H. (2010). The growth and long term survival of *Acholeplasma laidlawii* in media products used in biopharmaceutical manufacturing. *Biologicals: Journal of the International Association of Biological Standardization*, 38(2), 204–210. <https://doi.org/10.1016/j.biologicals.2009.11.009>
- Wirth, R., Chertkov, O., Held, B., Lapidus, A., Nolan, M., Lucas, S., Hammon, N., Deshpande, S., Cheng, J.-F., Tapia, R., Han, C., Goodwin, L., Pitluck, S., Liolios, K., Ioanna, P., Ivanova, N., Mavromatis, K., Mikhailova, N., Pati, A., ... Klenk, H.-P. (2011). Complete genome sequence of *Desulfurococcus mucosus* type strain (O7/1T). *Standards in Genomic Sciences*, 4(2), 173–182. <https://doi.org/10.4056/sigs.1644004>
- Wolf, E., & SCHUBetaLER, A. (2005). Phycobiliprotein fluorescence of *Nostoc punctiforme* changes during the life cycle and chromatic adaptation: characterization by spectral confocal laser scanning microscopy and spectral unmixing. *Plant, Cell & Environment*, 28(4), 480–491. <https://doi.org/10.1111/j.1365-3040.2005.01290.x>
- Wolf, Y. I., & Koonin, E. V. (2013). Genome reduction as the dominant mode of evolution. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 35(9), 829–837.
<https://doi.org/10.1002/bies.201300037>
- Woyke, T., Teeling, H., Ivanova, N. N., Huntemann, M., Richter, M., Gloeckner, F. O., Boffelli, D., Anderson, I. J., Barry, K. W., Shapiro, H. J., Szeto, E., Kyrpides, N. C., Musmann, M.,

- Amann, R., Bergin, C., Ruehland, C., Rubin, E. M., & Dubilier, N. (2006). Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature*, 443(7114), 950–955. <https://doi.org/10.1038/nature05192>
- Wrede, C., Dreier, A., Kokoschka, S., & Hoppert, M. (2012). Archaea in Symbioses. In *Archaea* (Vol. 2012, pp. 1–11). <https://doi.org/10.1155/2012/596846>
- Wu, D., Daugherty, S. C., Van Aken, S. E., Pai, G. H., Watkins, K. L., Khouri, H., Tallon, L. J., Zaborsky, J. M., Dunbar, H. E., Tran, P. L., Moran, N. A., & Eisen, J. A. (2006). Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. *PLoS Biology*, 4(6), e188. <https://doi.org/10.1371/journal.pbio.0040188>
- Xavier, J. C., Patil, K. R., & Rocha, I. (2014). Systems Biology Perspectives on Minimal and Simpler Cells. In *Microbiology and Molecular Biology Reviews* (Vol. 78, Issue 3, pp. 487–509). <https://doi.org/10.1128/mnbr.00050-13>
- Yamada, T., & Bork, P. (2009). Evolution of biomolecular networks: lessons from metabolic and protein interactions. *Nature Reviews. Molecular Cell Biology*, 10(11), 791–803. <https://doi.org/10.1038/nrm2787>
- Yarza, P., & Munoz, R. (2014). Chapter 3 - The All-Species Living Tree Project. In M. Goodfellow, I. Sutcliffe, & J. Chun (Eds.), *Methods in Microbiology* (Vol. 41, pp. 45–59). Academic Press. <https://doi.org/10.1016/bs.mim.2014.07.006>
- Zhaxybayeva, O., & Doolittle, W. F. (2011). Lateral gene transfer. *Current Biology: CB*, 21(7), R242–R246. <https://doi.org/10.1016/j.cub.2011.01.045>

XI. Funding

Manuscripts in Chapter 1 were supported by projects BFU2012-39816-CO2-01 and Prometeo/2009/092 from the Ministerio de Economía y Competitividad, Spain, and Generalitat Valenciana, Spain, CONACYT-México Ciencia Básica (project: CB-157220), and by BFU2012-39816-CO2-01, co-financed by FEDER funds, SAF 2012-31187, the PrometeoII/2014/065, Generalitat Valenciana, Spain, and the EU Marie Curie Initial Training Network (ITN) Symbiomics: Molecular ecology and evolution of bacterial symbionts [FP7-PEOPLE-2010-ITN].

Manuscripts in Chapter 2 were supported by BFU2012-39816-CO2-01, co-financed by FEDER funds, SAF-2012-31187, the EU Marie Curie Initial Training Network (ITN) Symbiomics: Molecular ecology and evolution of bacterial symbionts [FP7-PEOPLE-2010-ITN], and PROMETEOII/2014/065; Funding for open access charge: PROMETEOII/2014/065. The National Board of Science and Technology of Mexico (CONACYT) (contract number 538243); FPI fellowship from the Ministry of Economy and Competitiveness (BES-2013-063548); Spanish Ministry of Science, Innovation and Universities and the European Regional Development Fund through projects DPI2015-67082-P and PGC2018-096956-B-C43 (FEDER/MICINN/AEI) and the Spanish Ministry of Economy and Competitiveness (projects SAF2012-31187, SAF2013-49788-EXP, SAF2015-65878-R and BFU2015-64322-C2-1-R, co-financed by FEDER funds); Carlos III Institute of Health (projects PIE14/00045 and AC15/00022); Generalitat Valenciana (project Prometeo/2018/A/133); Asociación Española contra el Cáncer

(project AECC 2017-1485) and co-financed by the European Regional Development Fund (ERDF).

The manuscript in Chapter 3 was supported by the Spanish Ministry of Science, Innovation and Universities (MICINN/AEI, projects DPI2015-67082-P and PGC2018-096956-B-C43) the Spanish Ministry of Economy and Competitiveness (projects SAF2015-65878-R and PGC2018-099344-B-I00), Generalitat Valenciana (project Prometeo/2018/A133), co-financed by the European Regional Development Fund (ERDF), and The National Board of Science and Technology of México (CONACYT) [grant number 538243].

The manuscript in Chapter 4 was funded by the Spanish Ministry of Science, Innovation and Universities and the 428 European Regional Development Fund through projects PGC2018-096956-B-C43 and PID2019-429105969GB-I00 (FEDER/MICINN/AEI) and Generalitat Valenciana (Prometeo/2018/133).

IX. Appendices

Appendix A - Original publication first-page reprints

Opinion

Scanty microbes, the ‘symbionelle’ concept

Mariana Reyes-Prieto, Amparo Latorre and
Andrés Moya*

*Institut Cavanilles de Biodiversitat i Biologia Evolutiva,
Universitat de València, Calle Catedrático Agustín
Escardino 9, 46100 Paterna, València, Spain.*

Mutualistic symbiosis occurs when two different species interact closely with each other and benefit from living and working together. However, not all symbiotic associations are of mutual benefit because there are also forms of parasitism (when one organism benefits but the other is adversely affected) and commensalism (when only one of the organisms involved in the association benefits, but the other is not affected); notwithstanding, the very fact that specific entities can exist together means that natural selection may guide them to live with each other. Endosymbiosis is a special case of symbiosis in which one partner, generally a prokaryote symbiont, lives sequestered inside specialized eukaryotic cells called bacteriocytes.

The notion of microbes becoming organelles of eukaryotic systems through evolution has been widely accepted because Lynn Margulis put forward her serial endosymbiotic theory of eukaryotic cell evolution (Margulis, 1993). Indeed, this is the origin of mitochondria and chloroplasts. There is compelling evidence to support that these two eukaryotic organelles are the product of symbiotic events between prokaryotes and primitive eukaryotes (Latorre *et al.*, 2011). Their original alpha-proteobacterial (mitochondria ancestor) and cyanobacterial (chloroplast ancestor) genomes have been drastically reduced, with a portion of the protein-encoded genes and even RNA genes being transferred to the eukaryotic nuclear genome. Other genes have simply been lost, and their function replaced by the hosts. Since the proposal of these two canonical endosymbioses, symbiotic associations between prokaryotes and unicellular and multicellular eukaryotes have been documented in practically every major branch of the tree of life, which reinforces the role

played by symbiosis in the emergence of evolutionary innovations (Moya *et al.*, 2008).

Endosymbiosis in insects is a captivating example of the aforementioned phenomenon. Insects are particularly well suited to establishing intracellular symbiosis with bacteria, which provide them with the metabolic capabilities they lack and enable them to live in almost any environment. At present, there are a number of well-documented cases of insect endosymbionts at different stages of symbiotic integration (Fig. 1). Insect endosymbiosis commonly consists of an obligate mutualistic association, where bacteria produce essential nutrients that are absent in the insect's diet, and the insect, in turn, provides the bacteria with a safe environment and a permanent food supply (Baumann, 2005). These endosymbiotic bacteria are vertically transmitted across host generations. Their metabolic role is renowned, and most insect endosymbiotic systems are largely convergent towards these functions regardless of the lifestyle or genomic repertoire of their free-living ancestor (López-Sánchez *et al.*, 2008; McCutcheon *et al.*, 2009; McCutcheon and Moran, 2010; Sabree *et al.*, 2013). A new symbiotic relationship, which represents a source of novel complexity, has to overcome the obvious problem posed by the fact that both partners must be able to survive together despite differences in biology, particularly generation times and reproduction. Moreover, considering that these organisms generally possess different population genetics and are under different evolutionary pressures, they need to establish a certain trade-off to acquire the evolutionary novelty represented by their stable coexistence (Delays and Moya, 2010; McCutcheon and Moran, 2012). Thus, important genetic and biochemical modifications are required in these bacteria compared with their free-living state. The eukaryotic host, on the other hand, must develop ways of controlling the bacterial population, engulfing them in specialized cells –the aforesaid bacteriocytes – and/or changing immune responses to recognize these bacteria as non-pathogenic.

One of the most important and well-known features of endosymbiotic bacteria is that they provide extreme examples of genomic shrinkage by undergoing a process called the ‘genomic reduction syndrome’. Hence, prokaryotic genomes of endosymbionts are examples of a particular type of naturally evolved minimal cell, with insect

Received 1 July, 2013; revised 11 July, 2013; accepted 18 July, 2013.
*For correspondence. E-mail andres.moya@uv.es; Tel. (+34) 96 354 3480; Fax (+34) 96 354 3670.



Evolution of small prokaryotic genomes

David J. Martínez-Cano^{1†}, Mariana Reyes-Prieto^{2†}, Esperanza Martínez-Romero³,
Laila P. Partida-Martínez¹, Amparo Latore², Andrés Moya² and Luis Delayo^{1*}

¹ Departamento de Ingeniería Genética, Cinvestav Unidad Irapuato, Irapuato, Mexico

² Institut Cavanilles de Biodiversitat i Biologia Evolutiva, Universitat de València, Valencia, Spain

³ Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Mexico

Edited by:

Ana E. Escalante, Universidad Nacional Autónoma de México, Mexico

Reviewed by:

Olivier Antoine Tenillon, Institut

National de la Santé et de la

Recherche Médicale, France

Luis David Alcaez, Universidad

Nacional Autónoma de México,

Mexico

Zakae Sabree, Ohio State University,

USA

*Correspondence:

Luis Delayo, Departamento de Ingeniería Genética, Cinvestav Unidad Irapuato, Kilómetro 9.6, Libramiento Norte, Carretera Irapuato-León, Irapuato, Guanajuato 36821, Mexico
e-mail: ldelayo@ira.cinvestav.mx

[†]These authors have contributed equally to this work.

As revealed by genome sequencing, the biology of prokaryotes with reduced genomes is strikingly diverse. These include free-living prokaryotes with ~800 genes as well as endosymbiotic bacteria with as few as ~140 genes. Comparative genomics is revealing the evolutionary mechanisms that led to these small genomes. In the case of free-living prokaryotes, natural selection directly favored genome reduction, while in the case of endosymbiotic prokaryotes neutral processes played a more prominent role. However, new experimental data suggest that selective processes may be at operation as well for endosymbiotic prokaryotes at least during the first stages of genome reduction. Endosymbiotic prokaryotes have evolved diverse strategies for living with reduced gene sets inside a host-defined medium. These include utilization of host-encoded functions (some of them coded by genes acquired by gene transfer from the endosymbiont and/or other bacteria); metabolic complementation between co-symbionts; and forming consortiums with other bacteria within the host. Recent genome sequencing projects of intracellular mutualistic bacteria showed that previously believed universal evolutionary trends like reduced G+C content and conservation of genome synteny are not always present in highly reduced genomes. Finally, the simplified molecular machinery of some of these organisms with small genomes may be used to aid in the design of artificial minimal cells. Here we review recent genomic discoveries of the biology of prokaryotes endowed with small gene sets and discuss the evolutionary mechanisms that have been proposed to explain their peculiar nature.

Keywords: reductive genome evolution, endosymbiosis, minimal genome size, streamlining evolution, Black Queen Hypothesis, Muller's ratchet, robustness-based selective reduction, symbionelle

INTRODUCTION

Darwin proposed an externalist theory of evolution where organisms provide the raw material and the environment selects (Gould, 2002). The outcome of this process is a fine adjustment of organisms to the environment. The evolution of prokaryotes with reduced genomes is not an exception to this Darwinian principle. Host-associated bacteria and archaea evolved the smallest genomes in nature other than those of organelles and viruses. The rationale of this pattern is simple. Prokaryotes living in a protected and chemically rich medium can afford losing more genes than those coping with the vagaries of a free-living lifestyle (Morowitz, 1993). On the other hand, different lineages of free-living bacteria, most of them in marine environments, evolved reduced genomes likely by the direct action of natural selection (Giovannoni et al., 2014).

WHAT IS THE MINIMAL GENOME SIZE FOR EXTANT FREE-LIVING PROKARYOTES?

Previous surveys indicated that free-living prokaryotes had no less than ~1,300 genes (Islas et al., 2004; Podar et al., 2008;

Delayo et al., 2010). However, recent metagenomic sequencing suggests that there are free-living Actinobacteria with approximately 800 genes. This was discovered at the Mediterranean Sea and the bacteria were named “*Candidatus Actinomarina minuta*” (Ghai et al., 2013). Surprisingly, it is also one of the smallest cells with a cell volume of only ~0.013 μm^3 . If further sequencing of its complete genome confirms this estimate (and it is very likely that it will do), it will sensibly change our knowledge about the minimum number of genes a cell needs to survive in present free-living conditions, in a similar fashion than the discovery of “*Candidatus Carsonella ruddii*” shook our belief of the minimal gene set required for cells in 2006 (Nakabachi et al., 2006; McCutcheon and Moran, 2012). Meanwhile, as reviewed below, there exists a diversity of lineages of free-living prokaryotes that converged to approximately 1,300 genes despite their varying phylogenetic origins and nutritional strategies.

Nowadays, *Methanothermobacter thermautotrophicus* stands as the free-living archaeon (that does not grow associated to another cell) with the smallest sequenced genome. This organism is a methanogen and was isolated from an anaerobic Icelandic spring (Anderson et al., 2010). As mentioned above, other groups of free-living

Abbreviations: AM, arbuscular mycorrhizal; AVG, anti-virulence gene; dn/dS, non-synonymous versus synonymous substitutions; HGT, horizontal gene transfer; IS, insertion sequence; MGEs, mobile genetic elements; Trp, tryptophan; UGA_{stop}, UGA stop-coding codon; UGA_{Trp}, UGA tryptophan-coding codon.



Database tool

SymbioGenomesDB: a database for the integration and access to knowledge on host–symbiont relationships

Mariana Reyes-Prieto^{1,†}, Carlos Vargas-Chávez^{1,†}, Amparo Latorre^{1,2,3}
and Andrés Moya^{1,2,3,*}

¹Institut Cavanilles de Biodiversitat i Biologia Evolutiva, Universitat de València, Calle Catedrático José Beltrán 2, 46980 Paterna, València, Spain, ²Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana (FISABIO)-Salud Pública, Avenida de Catalunya 21 46020, València, Spain and ³CIBER en Epidemiología y Salud Pública (CIBEResp), Madrid, Spain

*Corresponding author: Tel: +34 96 354 3480; Fax: +34 96 354 3670; Email: andres.moya@uv.es

[†]These authors contributed equally to this work.

Citation details: Reyes-Prieto, M., Vargas-Chávez, C., Latorre, A. *et al.* SymbioGenomesDB: a database for the integration and access to knowledge on host–symbiont relationships. *Database* (2015) Vol. 2015: article ID bav109; doi:10.1093/database/bav109

Received 25 November 2014; Revised 27 July 2015; Accepted 10 September 2015

Abstract

Symbiotic relationships occur naturally throughout the tree of life, either in a commensal, mutualistic or pathogenic manner. The genomes of multiple organisms involved in symbiosis are rapidly being sequenced and becoming available, especially those from the microbial world. Currently, there are numerous databases that offer information on specific organisms or models, but none offer a global understanding on relationships between organisms, their interactions and capabilities within their niche, as well as their role as part of a system, in this case, their role in symbiosis. We have developed the SymbioGenomesDB as a community database resource for laboratories which intend to investigate and use information on the genetics and the genomics of organisms involved in these relationships. The ultimate goal of SymbioGenomesDB is to host and support the growing and vast symbiotic–host relationship information, to uncover the genetic basis of such associations. SymbioGenomesDB maintains a comprehensive organization of information on genomes of symbionts from diverse hosts throughout the Tree of Life, including their sequences, their metadata and their genomic features. This catalog of relationships was generated using computational tools, custom R scripts and manual integration of data available in public literature. As a highly curated and comprehensive systems database, SymbioGenomesDB provides web access to all the information of symbiotic organisms, their features and links to the central database NCBI. Three different tools can be found within the database to explore symbiosis-related organisms, their genes and their genomes. Also, we offer an orthology search for one or multiple genes in one or multiple organisms within symbiotic relationships, and every table, graph and output file is downloadable and easy to parse for further



Database update

An update on the Symbiotic Genomes Database (SymGenDB): a collection of metadata, genomic, genetic and protein sequences, orthologs and metabolic networks of symbiotic organisms

Mariana Reyes-Prieto^{1,2,*}, Carlos Vargas-Chávez^{1,3}, Mercè Llabrés⁴, Pere Palmer⁴, Amparo Latorre^{1,5} and Andrés Moya^{1,5,6}

¹Evolutionary Systems Biology of Symbionts, Institute for Integrative Systems Biology (I²SysBio), Universitat de València, Paterna, València, Spain ²Sequencing and Bioinformatics Service, Foundation for the Promotion of Sanitary and Biomedical Research of the Valencia Region (FISABIO), València, Spain ³Functional and Evolutionary Genomics, Institute of Evolutionary Biology (IBE), CSIC-Universitat Pompeu Fabra, Barcelona, Spain ⁴Department of Mathematics and Computer Science, University of the Balearic Islands, Palma, Balearic Islands, Spain ⁵Genomic and Health Area, Foundation for the Promotion of Sanitary and Biomedical Research of the Valencia Region (FISABIO), València, Spain ⁶CIBER in Epidemiology and Public Health (CIBEResp), Madrid, Spain

*Corresponding author: Email: bertmare@uv.es, mariana3131@gmail.com, Andres.Moya@uv.es

Citation details: Reyes-Prieto, M., Vargas-Chávez, C., Llabrés, M. *et al.* An update on the Symbiotic Genomes Database (SymGenDB): a collection of metadata, genomic, genetic and protein sequences, orthologs and metabolic networks of symbiotic organisms. *Database* (2020) Vol. 2020: article ID baz160; doi:10.1093/database/baz160




Received 12 December 2018; Revised 20 July 2019; Accepted 31 December 2019

Abstract

The Symbiotic Genomes Database (SymGenDB; <http://symbiogenomesdb.uv.es/>) is a public resource of manually curated associations between organisms involved in symbiotic relationships, maintaining a catalog of completely sequenced/finished bacterial genomes exclusively. It originally consisted of three modules where users could search for the bacteria involved in a specific symbiotic relationship, their genomes and their genes (including their orthologs). In this update, we present an additional module that includes a representation of the metabolic network of each organism included in the database, as Directed Acyclic Graphs (MetaDAGs). This module provides unique opportunities to explore the metabolism of each individual organism and/or to evaluate the shared and joint metabolic capabilities of the organisms of the same genera included in our listing, allowing users to construct predictive analyses of metabolic associations and complementation between systems. We also report a ~25% increase in manually curated content in the database, i.e. bacterial genomes and their associations, with a final count of 2328 bacterial genomes associated to 498 hosts. We describe new querying possibilities for all the modules, as well as new display features for the MetaDAGs module, providing a relevant range of content and utility. This update continues to

Article

The Metabolic Building Blocks of a Minimal Cell

Mariana Reyes-Prieto ^{1,2}, Rosario Gil ¹ , Mercè Llabrés ³, Pere Palmer-Rodríguez ³  and Andrés Moya ^{1,4,5,*} 

- ¹ Evolutionary Systems Biology of Symbionts, Institute for Integrative Systems Biology, University of Valencia and Spanish Research Council, Paterna, 46980 Valencia, Spain; reyes_ber@gva.es (M.R.-P.); rosario.gil@uv.es (R.G.)
- ² Sequencing and Bioinformatics Service, Foundation for the Promotion of Sanitary and Biomedical Research of the Valencia Region, 46020 Valencia, Spain
- ³ Department of Mathematics and Computer Science, University of Balearic Islands, 07122 Palma de Mallorca, Spain; merce.llabres@uib.es (M.L.); pere.palmer@uib.es (P.P.-R.)
- ⁴ Genomic and Health Area, Foundation for the Promotion of Sanitary and Biomedical Research of the Valencia Region, 46020 Valencia, Spain
- ⁵ Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública, 28029 Madrid, Spain
- * Correspondence: andres.moya@uv.es; Tel.: +34-963-543-480

Simple Summary: Manufacturing artificial living cells would open endless research possibilities in basic and applied sciences. With this motivation, many research groups are developing methodologies to construct a stable minimal cell that is capable of achieving metabolic homeostasis, reproducing, and evolving in a controlled environment. Using as a template the gene set for a minimal cell proposed previously by Gil and coworkers, we have put together a network depicting its inferred minimal metabolism needed for life. This network has been further compressed as a metabolic Directed Acyclic Graph (m-DAG) in order to better visualize its topology and to find its essential reactions (i.e., critical reactions to maintain the network's connectivity). We have also compared this minimal m-DAG to those of the smallest natural genome known until now and a synthetic minimal cell created in the laboratory. The modeling of m-DAGs based on minimal metabolisms can be a first approach for the synthesis and manipulation of minimal cells.



Citation: Reyes-Prieto, M.; Gil, R.; Llabrés, M.; Palmer-Rodríguez, P.; Moya, A. The Metabolic Building Blocks of a Minimal Cell. *Biology* **2021**, *10*, 5. <https://dx.doi.org/10.3390/biology10010005>

Received: 23 November 2020

Accepted: 21 December 2020

Published: 24 December 2020

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright © 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Defining the essential gene components for a system to be considered alive is a crucial step toward the synthesis of artificial life. Fifteen years ago, Gil and coworkers proposed the core of a putative minimal bacterial genome, which would provide the capability to achieve metabolic homeostasis, reproduce, and evolve to a bacterium in an ideally controlled environment. They also proposed a simplified metabolic chart capable of providing energy and basic components for a minimal living cell. For this work, we have identified the components of the minimal metabolic network based on the aforementioned studies, associated them to the KEGG database and, by applying the MetaDAG methodology, determined its Metabolic Building Blocks (MBB) and reconstructed its metabolic Directed Acyclic Graph (m-DAG). The reaction graph of this metabolic network consists of 80 compounds and 98 reactions, while its m-DAG has 36 MBBs. Additionally, we identified 12 essential reactions in the m-DAG that are critical for maintaining the connectivity of this network. In a similar manner, we reconstructed the m-DAG of JCVI-syn3.0, which is an artificially designed and manufactured viable cell whose genome arose by minimizing the one from *Mycoplasma mycoides* JCVI-syn1.0, and of *Candidatus* Nasuia deltocephalinicola, the bacteria with the smallest natural genome known to date. The comparison of the m-DAGs derived from a theoretical, an artificial, and a natural genome denote slightly different lifestyles, with a consistent core metabolism. The MetaDAG methodology we employ uses homogeneous descriptors and identifiers from the KEGG database, so that comparisons between bacterial strains are not only easy but also suitable for many research fields. The modeling of m-DAGs based on minimal metabolisms can be the first step for the synthesis and manipulation of minimal cells.

Keywords: metabolic networks; minimal gene set machinery; directed acyclic graphs; minimal cells

Appendix B - Supplementary material

Chapter 3: Available at <https://www.mdpi.com/2079-7737/10/1/5/s1>

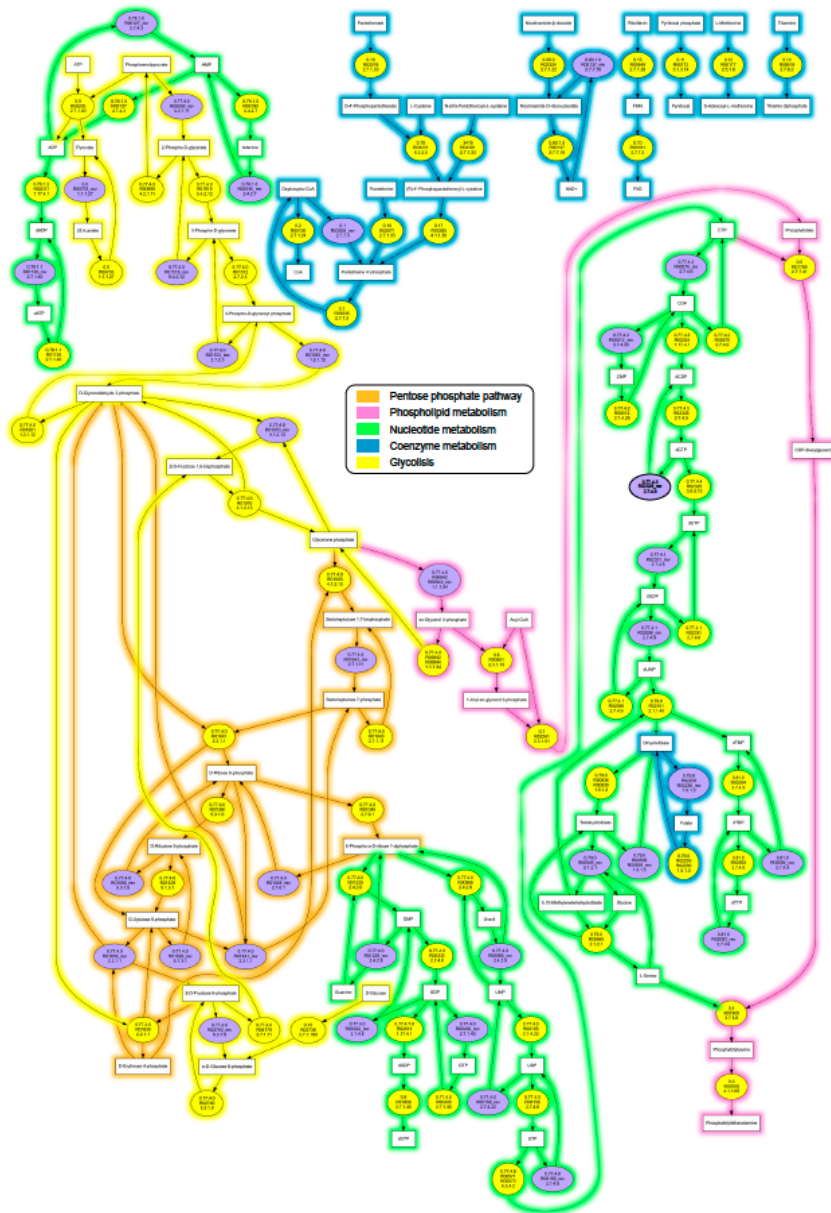


Figure C3S1 | Full-size representation of the reaction graph of the proposed theoretical minimal metabolic network represented in Figure C3.2.

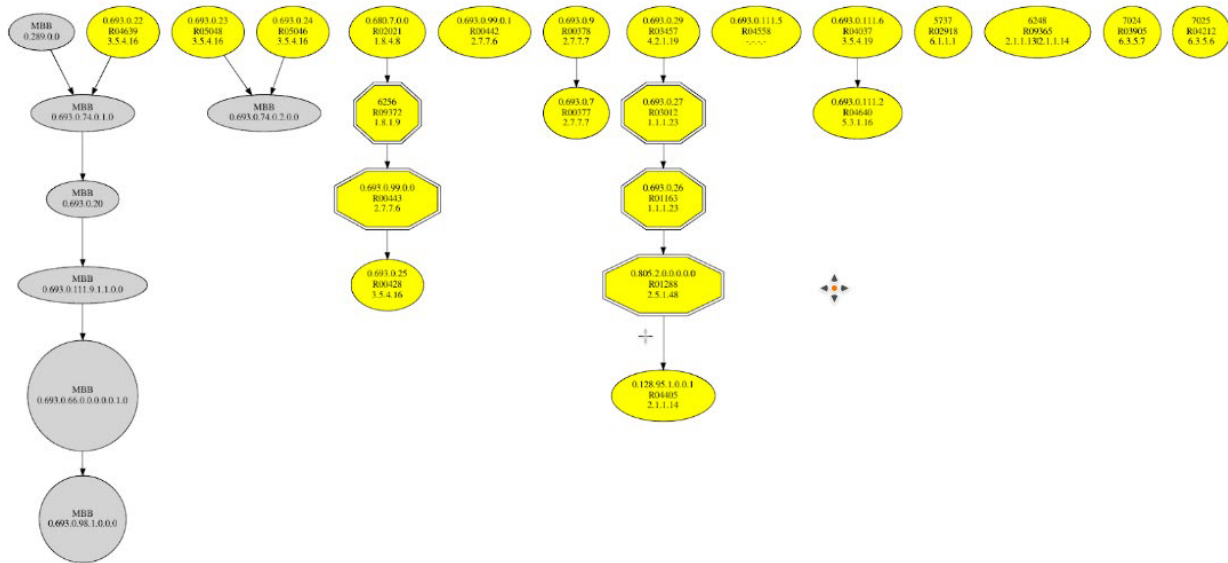


Figure C3S2 | The m-DAG of “Ca. Nasuia deltocephalinicola” str. NAS-ALF.

Table C3S1 | List of enzymes and reactions modified from Gabaldon et. al. (2007). n.i.: non-identified.

E.C. number	Name	Reaction	Gil et. al. 2004	Glass et. al. 2006
2.7.1.69	phosphotransferase system	$glc + pep \rightarrow g6p + pyr$	PTS	MG041, 069, 429
5.3.1.9	glucose-6-phosphate isomerase	$g6p \leftrightarrow f6p$	PGI	MG111
2.7.1.11	6-phosphofructokinase	$f6p + atp \rightarrow fbp + adp$	PFK	MG215
4.1.2.13	fructose-1,6-bisphosphate aldolase	$fbp \leftrightarrow gdp + dhp$	FBA	MG023
5.3.1.1	triose-phosphate isomerase	$gdp \leftrightarrow dhp$	TPI	MG431
1.2.1.12	glyceraldehyde-3-phosphate dehydrogenase	$gdp + nad + p \leftrightarrow bpg + nadh$	GAP	MG301
2.7.2.3	phosphoglycerate kinase	$bpg + adp \leftrightarrow 3pg + atp$	PGK	MG300
5.4.2.1	phosphoglycerate mutase	$3pg \leftrightarrow 2pg$	GPM	MG430
4.2.1.11	enolase	$2pg \leftrightarrow pep$	ENO	MG407
2.7.1.40	pyruvate kinase	$pep + adp \rightarrow pyr + atp$	PYK	MG216
1.1.1.27	lactate dehydrogenase	$pyr + nadh \leftrightarrow lac + nad$	LDH	MG460
1.1.1.94	sn-glycerol-3-phosphate dehydrogenase	$dhp + nadh \rightarrow g3p + nad$	GPS	n.i.
2.3.1.15	sn-glycerol-3-phosphate acyltransferase	$g3p + pal \rightarrow mag$	PLSb	n.i.
2.3.1.51	1-acyl-sn-glycerol-3-phosphate	$mag + pal \rightarrow dag$	PLSc	MG212

	acyltransferase			
2.7.7.41	phosphatidate cytidyltransferase	$\text{dag} + \text{ctp} \rightarrow \text{cdp-dag} + \text{pp}$	CDS	MG437
2.7.8.8	phosphatidylserine synthase	$\text{cdp-dag} + \text{ser} \rightarrow \text{pser} + \text{cmp}$	PSS	n.i.
4.1.1.65	phosphatidylserine decarboxylase	$\text{pser} \rightarrow \text{peta}$	PSD	n.i.
4.1.2.13	fructose-1,6-bisphosphate aldolase	$\text{gdp} + \text{e4p} \leftrightarrow \text{sbp}$	FBA2	MG023
3.1.3.37	sedoheptulose-1,7-bisphosphatase	$\text{sbp} \rightarrow \text{s7p} + \text{p}$	SPH	n.i.
2.2.1.1	transketolase	$\text{gdp} + \text{s7p} \leftrightarrow \text{rip} + \text{xip}$	TKT	MG066
2.2.1.1	transketolase	$\text{e4p} + \text{xip} \leftrightarrow \text{f6p} + \text{gdp}$	TKT2	MG066
5.1.3.1	ribulose-phosphate 3-epimerase	$\text{xip} \leftrightarrow \text{rup}$	RPE	MG112
5.3.1.6	ribose-5-phosphate isomerase	$\text{rup} \leftrightarrow \text{rip}$	RPI	MG396
2.7.6.1	phosphoribosylpyrophosphate synthetase	$\text{rip} + \text{atp} \rightarrow \text{prpp} + \text{amp}$	PRS	MG058
2.4.2.8	hypoxanthine phosphoribosyltransferase	$\text{prpp} + \text{ade} \rightarrow \text{amp} + \text{pp}$	HPT	MG276
2.4.2.8	hypoxanthine phosphoribosyltransferase	$\text{prpp} + \text{gua} \rightarrow \text{gmp} + \text{pp}$	HPT2	MG458
2.4.2.9	uracil phosphoribosyltransferase	$\text{prpp} + \text{ura} \rightarrow \text{ump} + \text{pp}$	UPP	MG030
3.6.1.1	inorganic pyrophosphatase	$\text{pp} \rightarrow 2\text{p}$	PPA	MG351
2.7.4.3	adenylate kinase	$\text{amp} + \text{atp} \rightarrow 2\text{adp}$	ADK	MG171
2.7.4.8	guanylate kinase	$\text{gmp} + \text{atp} \rightarrow \text{gdp} + \text{adp}$	GMK	MG107
2.7.4.14	cytidylate kinase	$\text{ump} + \text{atp} \rightarrow \text{udp} + \text{adp}$	CMK	MG330
2.7.4.14	cytidylate kinase	$\text{cmp} + \text{atp} \rightarrow \text{cdp} + \text{adp}$	CMK2	MG330
2.7.4.6	nucleoside diphosphate kinase	$\text{gdp} + \text{atp} \leftrightarrow \text{gtp} + \text{adp}$	NDK	MG216

2.7.4.6	nucleoside diphosphate kinase	$\text{udp} + \text{atp} \leftrightarrow \text{utp} + \text{adp}$	NDK2	
2.7.4.6	nucleoside diphosphate kinase	$\text{dadp} + \text{atp} \leftrightarrow \text{datp} + \text{adp}$	NDK3	MG216
2.7.4.6	nucleoside diphosphate kinase	$\text{dgdp} + \text{atp} \leftrightarrow \text{dntp} + \text{adp}$	NDK4	MG216
2.7.4.6	nucleoside diphosphate kinase	$\text{ctp} + \text{adp} \leftrightarrow \text{cdp} + \text{atp}$	NDK5	
2.7.4.6	nucleoside diphosphate kinase	$\text{dcdp} + \text{atp} \leftrightarrow \text{dctp} + \text{adp}$	NDK6	
2.7.4.6	nucleoside diphosphate kinase	$\text{dudp} + \text{adp} \leftrightarrow \text{dudp} + \text{atp}$	NDK7	
2.7.4.6	nucleoside diphosphate kinase	$\text{tdp} + \text{adp} \leftrightarrow \text{ttp} + \text{adp}$	NDK8	MG034
1.17.4.1	ribonucleoside diphosphate reductase	$\text{adp} + \text{nadh} \rightarrow \text{dadp} + \text{nad}$	NRD	MG229–MG231
1.17.4.1	ribonucleoside diphosphate reductase	$\text{gdp} + \text{nadh} \rightarrow \text{dgdp} + \text{nad}$	NRD2	MG229–MG231
1.17.4.1	ribonucleoside diphosphate reductase	$\text{cdp} + \text{nadh} \rightarrow \text{dcdp} + \text{nad}$	NRD3	MG229–MG231
6.3.4.2	CTP synthase	$\text{utp} \rightarrow \text{ctp}$	PYR	n.i.
3.5.4.13	dCTP deaminase	$\text{dctp} \rightarrow \text{dudp}$	DCD	n.i.
2.7.4.9	thymidylate kinase	$\text{dudp} + \text{adp} \leftrightarrow \text{dudp} + \text{atp}$	TMK	MG006
2.7.4.9	thymidylate kinase	$\text{tmp} + \text{atp} \leftrightarrow \text{tdp} + \text{adp}$	TMK2	MG006
2.1.1.45	thymidylate synthase	$\text{dudp} + \text{mthf} \rightarrow \text{dhf} + \text{tmp}$	THY	MG227
1.5.1.3	dihydrofolate reductase	$\text{dhf} + \text{nadh} \leftrightarrow \text{thf} + \text{nad}$	DFR	MG228
2.1.2.1	glycine hydroxymethyltransferase	$\text{ser} + \text{thf} \leftrightarrow \text{gly} + \text{mthf}$	GHT	MG394

Table C3S2 | Reactions, and compounds that make up *Ca. Nasuia deltocephalinicola*'s m-DAG. Reversible reactions are denoted by the superscript *r*.

Substrate KEGG id	ReactionID (E.C.number)	Definition	Product KEGG id
C00002	R00435' (2.7.7.6)	ATP + RNA ↔ Diphosphate + RNA	C00046
C00131	R00375' (2.7.7.7)	dATP + DNA ↔ Diphosphate + DNA	C00039
C00044	R00441' (2.7.7.6)	GTP + RNA ↔ Diphosphate + RNA	C00046
C00286	R00376' (2.7.7.7)	dGTP + DNA ↔ Diphosphate + DNA	C00039
C00075	R00443 (2.7.7.6)	UTP + RNA → Diphosphate + RNA	C00046
C00063	R00442 (2.7.7.6)	CTP + RNA → Diphosphate + RNA	C00046
C00458	R00377 (2.7.7.7)	dCTP + DNA → Diphosphate + DNA	C00039
C00459	R00378 (2.7.7.7)	dTTP + DNA → Diphosphate + DNA	C00039
C01118+C00283	R01288 (2.5.1.48)	O-Succinyl-L-homoserine + Hydrogen sulfide → L-Homocysteine + Succinate	C00155 + C00042
C00155	R04405 (2.1.1.14)	5-Methyltetrahydropteroyltri-L-glutamate + L-Homocysteine → Tetrahydropteroyltri-L-glutamate + L-Methionine	C00073
C02739	R01071' (2.4.2.17)	1- (5-Phospho-D-ribose)-ATP + Diphosphate ↔ ATP + 5-Phospho-alpha-D-ribose 1-diphosphate	C00119
C01929	R01163 (1.1.1.23)	L-Histidinal + H ₂ O + NAD ⁺ → L-Histidine + NADH + H ⁺	C00135
C00860	R03012 (1.1.1.23)	L-Histidamol + NAD ⁺ → L-Histidinal + NADH + H ⁺	C01929
C01100	R03243' (2.6.1.9)	L-Histidinol phosphate + 2-Oxoglutarate ↔ 3- (Imidazol-4-yl)-2-oxopropyl phosphate + L-Glutamate	C01267
C04666	R03457 (4.2.1.19)	D-erythro-1- (Imidazol-4-yl)glycerol 3-phosphate → 3- (Imidazol-4-yl)-2-oxopropyl phosphate + H ₂ O	C01267
C04896	R04640 (5.3.1.16)	5- (5-Phospho-D-ribosylaminoformimino)-1- (5-phosphoribosyl)-imidazole-4-carboxamide → N- (5'-Phospho-D-1'-ribulosylformimino)-5-amino-1- (5"-phospho-D-ribose)-4-imidazolecarboxamide	C04916
C02741	R04037 (3.5.4.19)	Phosphoribosyl-AMP + H ₂ O → 5- (5-Phospho-D-ribosylaminoformimino)-1- (5-phosphoribosyl)-imidazole-4-carboxamide	C04916
C04916	R04558 (-.-.-)	N- (5'-Phospho-D-1'-ribulosylformimino)-5-amino-1- (5"-phospho-D-ribose)-4-imidazolecarboxamide + L-Glutamine → D-erythro-1- (Imidazol-4-yl)glycerol 3-phosphate + 1- (5'-Phosphoribosyl)-5-amino-4-imidazolecarboxamide + L-Glutamate	C04666 + C04677
C00166	R00694' (2.6.1.9)	Phenylpyruvate + L-Glutamate ↔ L-Phenylalanine + 2-Oxoglutarate	C00079
C00082	R00734' (2.6.1.9)	L-Tyrosine + 2-Oxoglutarate ↔ 3- (4-Hydroxyphenyl)pyruvate + L-Glutamate	C01179

C05698	R09365 (2.1.1.13 2.1.1.14)	Selenohomocysteine + 5-Methyltetrahydropteroyltri-L-glutamate → L-Selenomethionine + Tetrahydropteroyltri-L-glutamate	C05335
C18902	R09372 (1.8.1.9)	2 NADPH + 2 H ⁺ + Methylselenic acid → 2 NADP ⁺ + 2 H ₂ O + Methaneselenol	C05703
C06148	R04639 (3.5.4.16)	2,5-Diamino-6- (5'-triphosphoryl-3',4'-trihydroxy-2'-oxopentyl)-amino-4-oxopyrimidine → 7,8-Dihydroneopterin 3'-triphosphate + H ₂ O	C04895
C05923	R05048 (3.5.4.16)	2,5-Diaminopyrimidine nucleoside triphosphate → 2,5-Diamino-6- (5'-triphosphoryl-3',4'-trihydroxy-2'-oxopentyl)-amino-4-oxopyrimidine	C06148
C05922	R05046 (3.5.4.16)	Formamidopyrimidine nucleoside triphosphate + H ₂ O → 2,5-Diaminopyrimidine nucleoside triphosphate + Formate	C05923
C00044	R00428 (3.5.4.16)	GTP + H ₂ O → Formamidopyrimidine nucleoside triphosphate	C05922
C00283	R00858' (1.8.1.2)	Hydrogen sulfide + 3 NADP ⁺ + 3 H ₂ O ↔ Sulfite + 3 NADPH + 3 H ⁺	C00094
C00053	R02021 (1.8.4.8)	Thioredoxin + 3'-Phosphoadenylyl sulfate → Thioredoxin disulfide + Sulfite + Adenosine 3',5'-bisphosphate	C00094
C00082	R02918 (6.1.1.1)	ATP + L-Tyrosine + tRNA (Tyr) → AMP + Diphosphate + L-Tyrosyl-tRNA (Tyr)	C02839
C02282	R03905 (6.3.5.7)	Glutamyl-tRNA + L-Glutamate + Orthophosphate + ADP → L-Glutamyl-tRNA (Gln) + L-Glutamine + ATP + H ₂ O	C06112
C03402	R04212 (6.3.5.6)	L-Asparaginyl-tRNA (Asn) + L-Glutamate + Orthophosphate + ADP ↔ L-Aspartyl-tRNA (Asn) + L-Glutamine + ATP + H ₂ O	C06113

Table C3S3 | Reactions included in the reconstruction of the JCVI-syn3.0 reaction graph and the minimal organism constructed for this work and the pathways in which each reaction (can) participates.

Reacción	JCVI-syn3.0	This work	Pathways in which they can participate
R02691(2.4.1.46)	Yes	No	1,2-diacylglycerol 3-beta-galactosyltransferase [EC:2.4.1.46]
R02241(2.3.1.51)	Yes	Yes	1-acyl-sn-glycerol-3-phosphate acyltransferase [EC:2.3.1.51]//lysophosphatidate acyltransferase [EC:2.3.1.51]//lysocardiolipin and lysophospholipid acyltransferase [EC:2.3.1.- 2.3.1.51]//lysophospholipid acyltransferase 1/2 [EC:2.3.1.51 2.3.1.-]//lysophospholipid acyltransferase [EC:2.3.1.51 2.3.1.23 2.3.1.-]//lysophosphatidic acid acyltransferase / lysophosphatidylinositol acyltransferase [EC:2.3.1.51 2.3.1.-]//TAG lipase / steryl ester hydrolase / phospholipase A2 / LPA acyltransferase [EC:3.1.1.3 3.1.1.13 3.1.1.4 2.3.1.51]//lysophosphatidate acyltransferase [EC:2.3.1.51]//1-acylglycerol-3-phosphate O-acyltransferase [EC:2.3.1.51]
R00508(3.1.3.7)	Yes	No	3'(2), 5'-bisphosphate nucleotidase [EC:3.1.3.7]//bifunctional oligoribonuclease and PAP phosphatase NrnA [EC:3.1.3.7 3.1.13.3]//3'(2), 5'-bisphosphate nucleotidase / inositol polyphosphate 1-phosphatase [EC:3.1.3.7 3.1.3.57]//inositol monophosphatase 3 [EC:3.1.3.25 3.1.3.7]
R01968(3.1.3.5)	Yes	No	5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//2',3'-cyclic-nucleotide 2'-phosphodiesterase / 3'-nucleotidase / 5'-nucleotidase [EC:3.1.4.16 3.1.3.6 3.1.3.5]//5'-deoxynucleotidase [EC:3.1.3.89]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase / UDP-sugar diphosphatase [EC:3.1.3.5 3.6.1.45]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]
R02088(3.1.3.5)	Yes	No	5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//2',3'-cyclic-nucleotide 2'-phosphodiesterase / 3'-nucleotidase / 5'-nucleotidase [EC:3.1.4.16 3.1.3.6 3.1.3.5]//5'-deoxynucleotidase [EC:3.1.3.89]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase / UDP-sugar diphosphatase [EC:3.1.3.5 3.6.1.45]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]
R01569(3.1.3.5)	Yes	No	5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//2',3'-cyclic-nucleotide 2'-phosphodiesterase / 3'-nucleotidase / 5'-nucleotidase [EC:3.1.4.16 3.1.3.6 3.1.3.5]//5'-deoxynucleotidase [EC:3.1.3.89]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase / UDP-sugar diphosphatase [EC:3.1.3.5 3.6.1.45]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]
R01664(3.1.3.5)	Yes	No	5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//2',3'-cyclic-nucleotide 2'-phosphodiesterase / 3'-nucleotidase / 5'-nucleotidase [EC:3.1.4.16 3.1.3.6 3.1.3.5]//5'-deoxynucleotidase [EC:3.1.3.89]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase / UDP-sugar diphosphatase [EC:3.1.3.5 3.6.1.45]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]
R02102(3.1.3.5)	Yes	No	5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//2',3'-cyclic-nucleotide 2'-phosphodiesterase / 3'-nucleotidase / 5'-nucleotidase [EC:3.1.4.16 3.1.3.6 3.1.3.5]//5'-deoxynucleotidase [EC:3.1.3.89]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase / UDP-sugar diphosphatase [EC:3.1.3.5 3.6.1.45]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]
R00183(3.1.3.5)	Yes	No	5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//2',3'-cyclic-nucleotide 2'-phosphodiesterase / 3'-nucleotidase / 5'-nucleotidase [EC:3.1.4.16 3.1.3.6 3.1.3.5]//5'-deoxynucleotidase [EC:3.1.3.89]//5'-nucleotidase / UDP-sugar diphosphatase [EC:3.1.3.5 3.6.1.45]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]
R01126(3.1.3.5)	Yes	No	5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//2',3'-cyclic-nucleotide 2'-phosphodiesterase / 3'-nucleotidase / 5'-nucleotidase [EC:3.1.4.16 3.1.3.6 3.1.3.5]//5'-deoxynucleotidase [EC:3.1.3.89]//5'-nucleotidase / UDP-sugar diphosphatase [EC:3.1.3.5 3.6.1.45]//IMP and pyridine-specific 5'-nucleotidase [EC:3.1.3.99 3.1.3.-]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]

R01227(3.1.3.5)	Yes	No	5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//2',3'-cyclic-nucleotide 2'-phosphodiesterase / 3'-nucleotidase / 5'-nucleotidase [EC:3.1.4.16 3.1.3.6 3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase / UDP-sugar diphosphatase [EC:3.1.3.5 3.6.1.45]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]
R02719(3.1.3.5)	Yes	No	5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//2',3'-cyclic-nucleotide 2'-phosphodiesterase / 3'-nucleotidase / 5'-nucleotidase [EC:3.1.4.16 3.1.3.6 3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase / UDP-sugar diphosphatase [EC:3.1.3.5 3.6.1.45]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]
R00511(3.1.3.5)	Yes	No	5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//2',3'-cyclic-nucleotide 2'-phosphodiesterase / 3'-nucleotidase / 5'-nucleotidase [EC:3.1.4.16 3.1.3.6 3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase / UDP-sugar diphosphatase [EC:3.1.3.5 3.6.1.45]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]
R00963(3.1.3.5)	Yes	No	5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//2',3'-cyclic-nucleotide 2'-phosphodiesterase / 3'-nucleotidase / 5'-nucleotidase [EC:3.1.4.16 3.1.3.6 3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase / UDP-sugar diphosphatase [EC:3.1.3.5 3.6.1.45]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]
R02323(3.1.3.5)	Yes	No	5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase / UDP-sugar diphosphatase [EC:3.1.3.5 3.6.1.45]//DMP and pyridine-specific 5'-nucleotidase [EC:3.1.3.99 3.1.3.-]//pyrimidine and pyridine-specific 5'-nucleotidase [EC:3.1.3.-]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]
R03346(3.1.3.5)	Yes	No	5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase / UDP-sugar diphosphatase [EC:3.1.3.5 3.6.1.45]//DMP and pyridine-specific 5'-nucleotidase [EC:3.1.3.99 3.1.3.-]//pyrimidine and pyridine-specific 5'-nucleotidase [EC:3.1.3.-]//5'-nucleotidase [EC:3.1.3.5]
R01569(3.1.3.89)	Yes	No	5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//2',3'-cyclic-nucleotide 2'-phosphodiesterase / 3'-nucleotidase / 5'-nucleotidase [EC:3.1.4.16 3.1.3.6 3.1.3.5]//5'-deoxynucleotidase [EC:3.1.3.89]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase / UDP-sugar diphosphatase [EC:3.1.3.5 3.6.1.45]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]
R01664(3.1.3.89)	Yes	No	5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//2',3'-cyclic-nucleotide 2'-phosphodiesterase / 3'-nucleotidase / 5'-nucleotidase [EC:3.1.4.16 3.1.3.6 3.1.3.5]//5'-deoxynucleotidase [EC:3.1.3.89]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase / UDP-sugar diphosphatase [EC:3.1.3.5 3.6.1.45]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]
R02102(3.1.3.89)	Yes	No	5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]//2',3'-cyclic-nucleotide 2'-phosphodiesterase / 3'-nucleotidase / 5'-nucleotidase [EC:3.1.4.16 3.1.3.6 3.1.3.5]//5'-deoxynucleotidase [EC:3.1.3.89]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase / UDP-sugar diphosphatase [EC:3.1.3.5 3.6.1.45]//5'-nucleotidase [EC:3.1.3.5]//5'-nucleotidase [EC:3.1.3.5]
R04779(2.7.1.11)	Yes	Yes	6-phosphofructokinase 1 [EC:2.7.1.11]//6-phosphofructokinase 2 [EC:2.7.1.11]//ATP-dependent phosphofructokinase / diphosphate-dependent phosphofructokinase [EC:2.7.1.11 2.7.1.90]//6-phosphofructokinase [EC:2.7.1.11]
R00315(2.7.2.1)	Yes	No	acetate kinase [EC:2.7.2.1]
R01353(2.7.2.1)	Yes	No	acetate kinase [EC:2.7.2.1]//propionate kinase [EC:2.7.2.15]//propionate kinase [EC:2.7.2.15]
R04378(2.4.2.7)	Yes	No	adenine phosphoribosyltransferase [EC:2.4.2.7]
R00190(2.4.2.7)	Yes	Yes	adenine phosphoribosyltransferase [EC:2.4.2.7]//hypoxanthine phosphoribosyltransferase [EC:2.4.2.8]
R01229(2.4.2.7)	Yes	Yes	adenine phosphoribosyltransferase [EC:2.4.2.7]//hypoxanthine phosphoribosyltransferase [EC:2.4.2.8]//xanthine phosphoribosyltransferase [EC:2.4.2.22]//xanthine phosphoribosyltransferase [EC:2.4.2.22]//bifunctional protein

			TiS/HprT [EC:6.3.4.19 2.4.2.8]//hypoxanthine phosphoribosyltransferase [EC:2.4.2.8]
R01229(2.4.2.8)	Yes	Yes	adenine phosphoribosyltransferase [EC:2.4.2.7]//hypoxanthine phosphoribosyltransferase [EC:2.4.2.8]//xanthine phosphoribosyltransferase [EC:2.4.2.22]//xanthine phosphoribosyltransferase [EC:2.4.2.22]//bifunctional protein TiS/HprT [EC:6.3.4.19 2.4.2.8]//hypoxanthine phosphoribosyltransferase [EC:2.4.2.8]
R00185(2.7.1.74)	Yes	No	adenosine kinase [EC:2.7.1.20]//deoxycytidine kinase [EC:2.7.1.74]//deoxyadenosine/deoxycytidine kinase [EC:2.7.1.76 2.7.1.74]
R00086(3.6.1.15)	Yes	No	adenosinetriphosphatase [EC:3.6.1.3]//apyrase [EC:3.6.1.5]//nucleoside triphosphate diphosphatase [EC:3.6.1.9]//nucleoside-triphosphatase [EC:3.6.1.15]//apyrase [EC:3.6.1.5]//golgi apyrase [EC:3.6.1.5]
R11319(2.7.4.3)	Yes	No	adenylate kinase [EC:2.7.4.3]
R01547(2.7.4.11)	Yes	No	adenylate kinase [EC:2.7.4.3]//adenylate kinase [EC:2.7.4.3]//adenylate/nucleoside-diphosphate kinase [EC:2.7.4.3 2.7.4.6]
R01547(2.7.4.3)	Yes	No	adenylate kinase [EC:2.7.4.3]//adenylate kinase [EC:2.7.4.3]//adenylate/nucleoside-diphosphate kinase [EC:2.7.4.3 2.7.4.6]
R01801(2.7.8.5)	Yes	No	CDP-diacylglycerol---glycerol-3-phosphate 3-phosphatidyltransferase [EC:2.7.8.5]
R01800(2.7.8.8)	No	Yes	CDP-diacylglycerol---serine O-phosphatidyltransferase [EC:2.7.8.8]//CDP-diacylglycerol---serine O-phosphatidyltransferase [EC:2.7.8.8]
R00512(2.7.4.25)	Yes	Yes	CMP/dCMP kinase [EC:2.7.4.25]//pantoate ligase / CMP/dCMP kinase [EC:6.3.2.1 2.7.4.25]//UMP-CMP kinase [EC:2.7.4.14]//UMP-CMP kinase 2, mitochondrial [EC:2.7.4.14]
R01665(2.7.4.25)	Yes	No	CMP/dCMP kinase [EC:2.7.4.25]//pantoate ligase / CMP/dCMP kinase [EC:6.3.2.1 2.7.4.25]//UMP-CMP kinase [EC:2.7.4.14]//UMP-CMP kinase 2, mitochondrial [EC:2.7.4.14]
R00158(2.7.4.22)	Yes	Yes	CMP/dCMP kinase [EC:2.7.4.25]//uridylylase [EC:2.7.4.22]//pantoate ligase / CMP/dCMP kinase [EC:6.3.2.1 2.7.4.25]//UMP-CMP kinase [EC:2.7.4.14]//UMP-CMP kinase 2, mitochondrial [EC:2.7.4.14]
R00571(6.3.4.2)	Yes	Yes	CTP synthase [EC:6.3.4.2]
R00573(6.3.4.2)	Yes	Yes	CTP synthase [EC:6.3.4.2]
R01663(3.5.4.12)	Yes	No	dCMP deaminase [EC:3.5.4.12]
R02325(3.5.4.13)	No	Yes	dCTP deaminase [EC:3.5.4.13]
R01667(3.6.1.12)	Yes	No	dCTP diphosphatase [EC:3.6.1.12]
R01668(3.6.1.12)	Yes	No	dCTP diphosphatase [EC:3.6.1.12]
R02089(2.7.1.76)	Yes	No	deoxyadenosine kinase [EC:2.7.1.76]//deoxyadenosine/deoxycytidine kinase [EC:2.7.1.76 2.7.1.74]
R01666(2.7.1.74)	Yes	No	deoxycytidine kinase [EC:2.7.1.74]//deoxyadenosine/deoxycytidine kinase [EC:2.7.1.76 2.7.1.74]//cytidine kinase [EC:2.7.1.213]
R01967(2.7.1.113)	Yes	No	deoxyguanosine kinase [EC:2.7.1.113]//deoxyguanosine kinase [EC:2.7.1.113]
R02235(1.5.1.3)	No	Yes	dihydrofolate reductase [EC:1.5.1.3]//dihydrofolate reductase / thymidylate synthase [EC:1.5.1.3 2.1.1.45]//dihydrofolate

			reductase (trimethoprim resistance protein) [EC:1.5.1.3]//dihydrofolate reductase (trimethoprim resistance protein) [EC:1.5.1.3]//dihydrofolate reductase (trimethoprim resistance protein) [EC:1.5.1.3]//dihydrofolate reductase (trimethoprim resistance protein) [EC:1.5.1.3]
R02094(2.7.4.9)	Yes	Yes	dTMP kinase [EC:2.7.4.9]
R02098(2.7.4.9)	Yes	Yes	dTMP kinase [EC:2.7.4.9]
R11896(3.6.1.23)	Yes	No	dUTP pyrophosphatase [EC:3.6.1.23]
R00161(2.7.7.2)	Yes	No	FAD synthetase [EC:2.7.7.2]//riboflavin kinase / FMN adenyltransferase [EC:2.7.1.26 2.7.7.2]//FAD synthetase [EC:2.7.7.2]//FAD synthetase [EC:2.7.7.2]
R00942(6.3.2.17)	Yes	No	folypolyglutamate synthase [EC:6.3.2.17]//dihydrofolate synthase / folypolyglutamate synthase [EC:6.3.2.12 6.3.2.17]//dihydrofolate synthase / folypolyglutamate synthase [EC:6.3.2.12 6.3.2.17]
R04241(6.3.2.17)	Yes	No	folypolyglutamate synthase [EC:6.3.2.17]//dihydrofolate synthase / folypolyglutamate synthase [EC:6.3.2.12 6.3.2.17]//dihydrofolate synthase / folypolyglutamate synthase [EC:6.3.2.12 6.3.2.17]
R02237(6.3.2.17)	Yes	No	folypolyglutamate synthase [EC:6.3.2.17]//dihydrofolate synthase / folypolyglutamate synthase [EC:6.3.2.12 6.3.2.17]//dihydrofolate synthase [EC:6.3.2.12]//dihydrofolate synthase / folypolyglutamate synthase [EC:6.3.2.12 6.3.2.17]//dihydrofolate synthase / dihydropterolate synthase [EC:6.3.2.12 2.5.1.15]
R00765(3.5.99.6)	Yes	No	glucosamine-6-phosphate deaminase [EC:3.5.99.6]
R02301(6.3.3.2)	Yes	No	glutamate formiminotransferase / 5-formyltetrahydrofolate cyclo-ligase [EC:2.1.2.5 6.3.3.2]//5-formyltetrahydrofolate cyclo-ligase [EC:6.3.3.2]
R01058(1.2.1.9)	Yes	No	glyceraldehyde-3-phosphate dehydrogenase (NADP+) [EC:1.2.1.9]//glyceraldehyde-3-phosphate dehydrogenase [NAD(P)+] [EC:1.2.1.90]
R00847(2.7.1.30)	Yes	No	glycerol kinase [EC:2.7.1.30]
R00851(2.3.1.15)	No	Yes	glycerol-3-phosphate O-acyltransferase 1/2 [EC:2.3.1.15]//glycerol-3-phosphate O-acyltransferase [EC:2.3.1.15]//glycerol-3-phosphate O-acyltransferase [EC:2.3.1.15]//glycerol-3-phosphate O-acyltransferase 3/4 [EC:2.3.1.15]//glycerol-3-phosphate O-acyltransferase / dihydroxyacetone phosphate acyltransferase [EC:2.3.1.15 2.3.1.42]//glycerol-3-phosphate acyltransferase [EC:2.3.1.15 2.3.1.198]
R00945(2.1.2.1)	Yes	Yes	glycine hydroxymethyltransferase [EC:2.1.2.1]
R09099(2.1.2.1)	Yes	No	glycine hydroxymethyltransferase [EC:2.1.2.1]
R00332(2.7.4.8)	Yes	Yes	guanylate kinase [EC:2.7.4.8]
R02090(2.7.4.8)	Yes	No	guanylate kinase [EC:2.7.4.8]
R01625(2.7.8.7)	Yes	No	holo-[acyl-carrier protein] synthase [EC:2.7.8.7]//4'-phosphopantetheinyl transferase [EC:2.7.8.-]
R08237(2.4.2.8)	Yes	No	hypoxanthine phosphoribosyltransferase [EC:2.4.2.8]
R08238(2.4.2.8)	Yes	No	hypoxanthine phosphoribosyltransferase [EC:2.4.2.8]

R08245(2.4.2.8)	Yes	No	hypoxanthine phosphoribosyltransferase [EC:2.4.2.8]
R01132(2.4.2.8)	Yes	No	hypoxanthine phosphoribosyltransferase [EC:2.4.2.8]//bifunctional protein Tls/HprT [EC:6.3.4.19 2.4.2.8]//hypoxanthine phosphoribosyltransferase [EC:2.4.2.8]
R02142(2.4.2.8)	Yes	No	hypoxanthine phosphoribosyltransferase [EC:2.4.2.8]//xanthine phosphoribosyltransferase [EC:2.4.2.22]//xanthine phosphoribosyltransferase [EC:2.4.2.22]//bifunctional protein Tls/HprT [EC:6.3.4.19 2.4.2.8]//hypoxanthine phosphoribosyltransferase [EC:2.4.2.8]
R02100(3.6.1.23)	Yes	No	inosine triphosphate pyrophosphatase [EC:3.6.1.-]//dUTP pyrophosphatase [EC:3.6.1.23]//XTP/dITP diphosphohydrolase [EC:3.6.1.66]
R00703(1.1.1.27)	No	Yes	L-lactate dehydrogenase [EC:1.1.1.27]
R03940(2.1.2.9)	Yes	No	methionyl-tRNA formyltransferase [EC:2.1.2.9]
R01655(3.5.4.9)	Yes	No	methylenetetrahydrofolate dehydrogenase (NADP+) / methylenetetrahydrofolate cyclohydrolase / formyltetrahydrofolate synthetase [EC:1.5.1.5 3.5.4.9 6.3.4.3]//methylenetetrahydrofolate dehydrogenase (NADP+) / methylenetetrahydrofolate cyclohydrolase [EC:1.5.1.5 3.5.4.9]//methylenetetrahydrofolate cyclohydrolase [EC:3.5.4.9]//methylenetetrahydrofolate dehydrogenase(NAD+) / 5,10-methylenetetrahydrofolate cyclohydrolase [EC:1.5.1.5 3.5.4.9]
R01220(1.5.1.5)	Yes	No	methylenetetrahydrofolate dehydrogenase (NADP+) / methylenetetrahydrofolate cyclohydrolase / formyltetrahydrofolate synthetase [EC:1.5.1.5 3.5.4.9 6.3.4.3]//methylenetetrahydrofolate/methylenetetrahydrofolate/methylenetetrahydropteridine dehydrogenase (NADP+) [EC:1.5.1.5 1.5.1.-]//methylenetetrahydrofolate dehydrogenase (NADP+) / methylenetetrahydrofolate cyclohydrolase [EC:1.5.1.5 3.5.4.9]
R05168(3.5.1.25)	Yes	No	N-acetylgalactosamine-6-phosphate deacetylase [EC:3.5.1.25]
R02059(3.5.1.25)	Yes	No	N-acetylglucosamine-6-phosphate deacetylase [EC:3.5.1.25]
R02705(2.7.1.60)	Yes	No	N-acylmannosamine kinase [EC:2.7.1.60]//bifunctional UDP-N-acetylglucosamine 2-epimerase / N-acylmannosamine kinase [EC:3.2.1.183 2.7.1.60]//N-acylmannosamine-6-phosphate 2-epimerase / N-acylmannosamine kinase [EC:5.1.3.9 2.7.1.60]
R00104(2.7.1.23)	Yes	No	NAD+ kinase [EC:2.7.1.23]
R00189(6.3.1.5)	Yes	No	NAD+ synthase [EC:6.3.1.5]
R01271(2.4.2.12)	Yes	No	nicotinamide phosphoribosyltransferase [EC:2.4.2.12]
R00137(2.7.7.18)	Yes	No	nicotinamide-nucleotide adenyltransferase [EC:2.7.7.1]//nicotinate-nucleotide adenyltransferase [EC:2.7.7.18]//nicotinamide mononucleotide adenyltransferase [EC:2.7.7.1 2.7.7.18]//HTH-type transcriptional regulator, transcriptional repressor of NAD biosynthesis genes [EC:2.7.7.1 2.7.1.22]//bifunctional NMN adenyltransferase/mudix hydrolase [EC:2.7.7.1 3.6.1.-]//nicotinamide-nucleotide adenyltransferase [EC:2.7.7.1]
R03005(2.7.7.18)	Yes	No	nicotinamide-nucleotide adenyltransferase [EC:2.7.7.1]//nicotinate-nucleotide adenyltransferase [EC:2.7.7.18]//nicotinamide mononucleotide adenyltransferase [EC:2.7.7.1 2.7.7.18]//HTH-type transcriptional regulator, transcriptional repressor of NAD biosynthesis genes [EC:2.7.7.1 2.7.1.22]//bifunctional NMN adenyltransferase/mudix hydrolase [EC:2.7.7.1 3.6.1.-]//nicotinamide-nucleotide adenyltransferase [EC:2.7.7.1]

R00615(3.6.1.15)	Yes	No	nucleoside-triphosphatase [EC:3.6.1.15]///ribosome biogenesis GTPase / thiamine phosphate phosphatase [EC:3.6.1.-3.1.3.100]
R00921(2.3.1.8)	Yes	No	phosphate acetyltransferase [EC:2.3.1.8]///phosphate acetyltransferase [EC:2.3.1.8]///phosphate propanoyltransferase [EC:2.3.1.222]///putative phosphotransacetylase [EC:2.3.1.8]
R00230(2.3.1.8)	Yes	No	phosphate acetyltransferase [EC:2.3.1.8]///phosphotransacetylase///phosphate acetyltransferase [EC:2.3.1.8]///putative phosphotransacetylase [EC:2.3.1.8]
R01799(2.7.7.41)	Yes	Yes	phosphatidate cytidyltransferase [EC:2.7.7.41]
R04162(3.1.3.4)	Yes	No	phosphatidate phosphatase [EC:3.1.3.4]
R02029(3.1.3.27)	Yes	No	phosphatidylglycerophosphatase GEP4 [EC:3.1.3.27]///phosphatidylglycerophosphatase A [EC:3.1.3.27]///phosphatidylglycerophosphatase B [EC:3.1.3.27 3.1.3.81 3.1.3.4 3.6.1.27]///phosphatidylglycerophosphatase C [EC:3.1.3.27]
R01512(2.7.2.3)	Yes	Yes	phosphoglycerate kinase [EC:2.7.2.3]
R01969(2.4.2.1)	Yes	No	purine-nucleoside phosphorylase [EC:2.4.2.1]///purine-nucleoside phosphorylase [EC:2.4.2.1]
R02557(2.4.2.1)	Yes	No	purine-nucleoside phosphorylase [EC:2.4.2.1]///purine-nucleoside phosphorylase [EC:2.4.2.1]
R02748(2.4.2.1)	Yes	No	purine-nucleoside phosphorylase [EC:2.4.2.1]///purine-nucleoside phosphorylase [EC:2.4.2.1]
R02294(2.4.2.1)	Yes	No	purine-nucleoside phosphorylase [EC:2.4.2.1]///purine-nucleoside phosphorylase [EC:2.4.2.1]
R02295(2.4.2.1)	Yes	No	purine-nucleoside phosphorylase [EC:2.4.2.1]///purine-nucleoside phosphorylase [EC:2.4.2.1]
R01561(2.4.2.1)	Yes	No	purine-nucleoside phosphorylase [EC:2.4.2.1]///purine-nucleoside phosphorylase [EC:2.4.2.1]///purine/pyrimidine-nucleoside phosphorylase [EC:2.4.2.1 2.4.2.2]
R01863(2.4.2.1)	Yes	No	purine-nucleoside phosphorylase [EC:2.4.2.1]///purine-nucleoside phosphorylase [EC:2.4.2.1]///purine/pyrimidine-nucleoside phosphorylase [EC:2.4.2.1 2.4.2.2]
R02147(2.4.2.1)	Yes	No	purine-nucleoside phosphorylase [EC:2.4.2.1]///purine-nucleoside phosphorylase [EC:2.4.2.1]///purine/pyrimidine-nucleoside phosphorylase [EC:2.4.2.1 2.4.2.2]
R02297(2.4.2.1)	Yes	No	purine-nucleoside phosphorylase [EC:2.4.2.1]///purine-nucleoside phosphorylase [EC:2.4.2.1]///xanthosine phosphorylase [EC:2.4.2.-]///purine/pyrimidine-nucleoside phosphorylase [EC:2.4.2.1 2.4.2.2]
R02484(2.4.2.1)	Yes	No	pyrimidine-nucleoside phosphorylase [EC:2.4.2.2]///uridine phosphorylase [EC:2.4.2.3]///thymidine phosphorylase [EC:2.4.2.4]///purine-nucleoside phosphorylase [EC:2.4.2.1]///purine-nucleoside phosphorylase [EC:2.4.2.1]
R00200(2.7.1.40)	Yes	Yes	pyruvate kinase [EC:2.7.1.40]///pyruvate kinase isozymes R/L [EC:2.7.1.40]
R00430(2.7.1.40)	Yes	Yes	pyruvate kinase [EC:2.7.1.40]///pyruvate kinase isozymes R/L [EC:2.7.1.40]
R01138(2.7.1.40)	Yes	Yes	pyruvate kinase [EC:2.7.1.40]///pyruvate kinase isozymes R/L [EC:2.7.1.40]
R01858(2.7.1.40)	Yes	Yes	pyruvate kinase [EC:2.7.1.40]///pyruvate kinase isozymes R/L [EC:2.7.1.40]

R00549(2.7.1.26)	Yes	No	riboflavin kinase [EC:2.7.1.26]///riboflavin kinase / FMN adenyltransferase [EC:2.7.1.26 2.7.7.2]///riboflavin kinase / FMN hydrolase [EC:2.7.1.26 3.1.3.102]
R08363(1.17.4.1)	Yes	No	ribonucleoside-diphosphate reductase subunit M1 [EC:1.17.4.1]///ribonucleoside-diphosphate reductase subunit M2 [EC:1.17.4.1]
R11893(1.17.4.1)	Yes	No	ribonucleoside-diphosphate reductase subunit M1 [EC:1.17.4.1]///ribonucleoside-diphosphate reductase subunit M2 [EC:1.17.4.1]
R02018(1.17.4.1)	Yes	No	ribonucleoside reductase, class II [EC:1.17.4.1]///ribonucleoside-diphosphate reductase alpha chain [EC:1.17.4.1]///ribonucleoside-diphosphate reductase beta chain [EC:1.17.4.1]
R02017(1.17.4.1)	Yes	Yes	ribonucleoside reductase, class II [EC:1.17.4.1]///ribonucleoside-diphosphate reductase alpha chain [EC:1.17.4.1]///ribonucleoside-diphosphate reductase beta chain [EC:1.17.4.1]///ribonucleoside-diphosphate reductase subunit M1 [EC:1.17.4.1]///ribonucleoside-diphosphate reductase subunit M2 [EC:1.17.4.1]
R02019(1.17.4.1)	Yes	Yes	ribonucleoside reductase, class II [EC:1.17.4.1]///ribonucleoside-diphosphate reductase alpha chain [EC:1.17.4.1]///ribonucleoside-diphosphate reductase beta chain [EC:1.17.4.1]///ribonucleoside-diphosphate reductase subunit M1 [EC:1.17.4.1]///ribonucleoside-diphosphate reductase subunit M2 [EC:1.17.4.1]
R02024(1.17.4.1)	Yes	Yes	ribonucleoside reductase, class II [EC:1.17.4.1]///ribonucleoside-diphosphate reductase alpha chain [EC:1.17.4.1]///ribonucleoside-diphosphate reductase beta chain [EC:1.17.4.1]///ribonucleoside-diphosphate reductase subunit M1 [EC:1.17.4.1]///ribonucleoside-diphosphate reductase subunit M2 [EC:1.17.4.1]
R01049(2.7.6.1)	Yes	Yes	ribose-phosphate pyrophosphokinase [EC:2.7.6.1]
R07618(1.8.1.4)	Yes	No	m00010: Glycolysis / Gluconeogenesis///m00020: Citrate cycle (TCA cycle)///m00280: Valine, leucine and isoleucine degradation///m00620: Pyruvate metabolism///m00640: Propionate metabolism///m01100: Metabolic pathways///m01110: Biosynthesis of secondary metabolites
R03270(1.2.4.1)	Yes	No	m00010: Glycolysis / Gluconeogenesis///m00020: Citrate cycle (TCA cycle)///m00620: Pyruvate metabolism
R00014(1.2.4.1)	Yes	No	m00010: Glycolysis / Gluconeogenesis///m00020: Citrate cycle (TCA cycle)///m00620: Pyruvate metabolism
R02569(2.3.1.12)	Yes	No	m00010: Glycolysis / Gluconeogenesis///m00020: Citrate cycle (TCA cycle)///m00620: Pyruvate metabolism
R01070(4.1.2.13)	Yes	Yes	m00010: Glycolysis / Gluconeogenesis///m00030: Pentose phosphate pathway///m00051: Fructose and mannose metabolism///m01100: Metabolic pathways///m01110: Biosynthesis of secondary metabolites///m01120: Microbial metabolism in diverse environments///m01200: Carbon metabolism///m01230: Biosynthesis of amino acids
R02740(5.3.1.9)	Yes	Yes	m00010: Glycolysis / Gluconeogenesis///m00030: Pentose phosphate pathway///m00520: Amino sugar and nucleotide sugar metabolism///m01100: Metabolic pathways///m01110: Biosynthesis of secondary metabolites///m01120: Microbial metabolism in diverse environments///m01200: Carbon metabolism
R02739(5.3.1.9)	Yes	No	m00010: Glycolysis / Gluconeogenesis///m00030: Pentose phosphate pathway///m01100: Metabolic pathways///m01110: Biosynthesis of secondary metabolites///m01120: Microbial metabolism in diverse environments
R01015(5.3.1.1)	Yes	No	m00010: Glycolysis / Gluconeogenesis///m00051: Fructose and mannose metabolism///m00562: Inositol phosphate metabolism///m00710: Carbon fixation in photosynthetic organisms///m01100: Metabolic pathways///m01110: Biosynthesis of secondary metabolites///m01120: Microbial metabolism in diverse environments///m01200: Carbon

R00615(3.6.1.15)	Yes	No	nucleoside-triphosphatase [EC:3.6.1.15]///ribosome biogenesis GTPase / thiamine phosphate phosphatase [EC:3.6.1.-3.1.3.100]
R00921(2.3.1.8)	Yes	No	phosphate acetyltransferase [EC:2.3.1.8]///phosphate acetyltransferase [EC:2.3.1.8]///phosphate propanoyltransferase [EC:2.3.1.222]///putative phosphotransacetylase [EC:2.3.1.8]
R00230(2.3.1.8)	Yes	No	phosphate acetyltransferase [EC:2.3.1.8]///phosphotransacetylase///phosphate acetyltransferase [EC:2.3.1.8]///putative phosphotransacetylase [EC:2.3.1.8]
R01799(2.7.7.41)	Yes	Yes	phosphatidate cytidyltransferase [EC:2.7.7.41]
R04162(3.1.3.4)	Yes	No	phosphatidate phosphatase [EC:3.1.3.4]
R02029(3.1.3.27)	Yes	No	phosphatidylglycerophosphatase GEP4 [EC:3.1.3.27]///phosphatidylglycerophosphatase A [EC:3.1.3.27]///phosphatidylglycerophosphatase B [EC:3.1.3.27 3.1.3.81 3.1.3.4 3.6.1.27]///phosphatidylglycerophosphatase C [EC:3.1.3.27]
R01512(2.7.2.3)	Yes	Yes	phosphoglycerate kinase [EC:2.7.2.3]
R01969(2.4.2.1)	Yes	No	purine-nucleoside phosphorylase [EC:2.4.2.1]///purine-nucleoside phosphorylase [EC:2.4.2.1]
R02557(2.4.2.1)	Yes	No	purine-nucleoside phosphorylase [EC:2.4.2.1]///purine-nucleoside phosphorylase [EC:2.4.2.1]
R02748(2.4.2.1)	Yes	No	purine-nucleoside phosphorylase [EC:2.4.2.1]///purine-nucleoside phosphorylase [EC:2.4.2.1]
R02294(2.4.2.1)	Yes	No	purine-nucleoside phosphorylase [EC:2.4.2.1]///purine-nucleoside phosphorylase [EC:2.4.2.1]
R02295(2.4.2.1)	Yes	No	purine-nucleoside phosphorylase [EC:2.4.2.1]///purine-nucleoside phosphorylase [EC:2.4.2.1]
R01561(2.4.2.1)	Yes	No	purine-nucleoside phosphorylase [EC:2.4.2.1]///purine-nucleoside phosphorylase [EC:2.4.2.1]///purine/pyrimidine-nucleoside phosphorylase [EC:2.4.2.1 2.4.2.2]
R01863(2.4.2.1)	Yes	No	purine-nucleoside phosphorylase [EC:2.4.2.1]///purine-nucleoside phosphorylase [EC:2.4.2.1]///purine/pyrimidine-nucleoside phosphorylase [EC:2.4.2.1 2.4.2.2]
R02147(2.4.2.1)	Yes	No	purine-nucleoside phosphorylase [EC:2.4.2.1]///purine-nucleoside phosphorylase [EC:2.4.2.1]///purine/pyrimidine-nucleoside phosphorylase [EC:2.4.2.1 2.4.2.2]
R02297(2.4.2.1)	Yes	No	purine-nucleoside phosphorylase [EC:2.4.2.1]///purine-nucleoside phosphorylase [EC:2.4.2.1]///xanthosine phosphorylase [EC:2.4.2.-]///purine/pyrimidine-nucleoside phosphorylase [EC:2.4.2.1 2.4.2.2]
R02484(2.4.2.1)	Yes	No	pyrimidine-nucleoside phosphorylase [EC:2.4.2.2]///uridine phosphorylase [EC:2.4.2.3]///thymidine phosphorylase [EC:2.4.2.4]///purine-nucleoside phosphorylase [EC:2.4.2.1]///purine-nucleoside phosphorylase [EC:2.4.2.1]
R00200(2.7.1.40)	Yes	Yes	pyruvate kinase [EC:2.7.1.40]///pyruvate kinase isozymes R/L [EC:2.7.1.40]
R00430(2.7.1.40)	Yes	Yes	pyruvate kinase [EC:2.7.1.40]///pyruvate kinase isozymes R/L [EC:2.7.1.40]
R01138(2.7.1.40)	Yes	Yes	pyruvate kinase [EC:2.7.1.40]///pyruvate kinase isozymes R/L [EC:2.7.1.40]
R01858(2.7.1.40)	Yes	Yes	pyruvate kinase [EC:2.7.1.40]///pyruvate kinase isozymes R/L [EC:2.7.1.40]

			metabolism///m01230: Biosynthesis of amino acids
R00959(5.4.2.5)	Yes	No	m00010: Glycolysis / Gluconeogenesis///m00052: Galactose metabolism///m00520: Amino sugar and nucleotide sugar metabolism///m01100: Metabolic pathways///m01110: Biosynthesis of secondary metabolites///m01120: Microbial metabolism in diverse environments
R01518(5.4.2.12)	Yes	Yes	m00010: Glycolysis / Gluconeogenesis///m00260: Glycine, serine and threonine metabolism///m00680: Methane metabolism///m01100: Metabolic pathways///m01110: Biosynthesis of secondary metabolites///m01120: Microbial metabolism in diverse environments///m01200: Carbon metabolism///m01230: Biosynthesis of amino acids
R00658(4.2.1.11)	Yes	Yes	m00010: Glycolysis / Gluconeogenesis///m00680: Methane metabolism///m01100: Metabolic pathways///m01110: Biosynthesis of secondary metabolites///m01120: Microbial metabolism in diverse environments///m01200: Carbon metabolism///m01230: Biosynthesis of amino acids
R01061(1.2.1.12)	Yes	Yes	m00010: Glycolysis / Gluconeogenesis///m00710: Carbon fixation in photosynthetic organisms///m01100: Metabolic pathways///m01110: Biosynthesis of secondary metabolites///m01120: Microbial metabolism in diverse environments///m01200: Carbon metabolism///m01230: Biosynthesis of amino acids
R03321(5.3.1.9)	Yes	No	m00010: Glycolysis / Gluconeogenesis///m01100: Metabolic pathways///m01110: Biosynthesis of secondary metabolites///m01120: Microbial metabolism in diverse environments
R01529(5.1.3.1)	Yes	Yes	m00030: Pentose phosphate pathway///m00040: Pentose and glucuronate interconversions///m00710: Carbon fixation in photosynthetic organisms///m01100: Metabolic pathways///m01110: Biosynthesis of secondary metabolites///m01120: Microbial metabolism in diverse environments///m01200: Carbon metabolism///m01230: Biosynthesis of amino acids
R01057(5.4.2.7)	Yes	No	m00030: Pentose phosphate pathway///m00230: Purine metabolism///m01100: Metabolic pathways
R01641(2.2.1.1)	Yes	Yes	m00030: Pentose phosphate pathway///m00710: Carbon fixation in photosynthetic organisms///m01100: Metabolic pathways///m01110: Biosynthesis of secondary metabolites///m01120: Microbial metabolism in diverse environments///m01200: Carbon metabolism///m01230: Biosynthesis of amino acids
R01056(5.3.1.6)	Yes	Yes	m00030: Pentose phosphate pathway///m00710: Carbon fixation in photosynthetic organisms///m01100: Metabolic pathways///m01110: Biosynthesis of secondary metabolites///m01120: Microbial metabolism in diverse environments///m01200: Carbon metabolism///m01230: Biosynthesis of amino acids
R01066(4.1.2.4)	Yes	No	m00030: Pentose phosphate pathway///m01100: Metabolic pathways
R02749(5.4.2.7)	Yes	No	m00030: Pentose phosphate pathway///m01100: Metabolic pathways
R01819(5.3.1.8)	Yes	No	m00051: Fructose and mannose metabolism///m00520: Amino sugar and nucleotide sugar metabolism///m01100: Metabolic pathways///m01110: Biosynthesis of secondary metabolites
R02568(4.1.2.13)	Yes	No	m00051: Fructose and mannose metabolism///m01100: Metabolic pathways///m01120: Microbial metabolism in diverse environments
R09030(5.3.1.6)	Yes	No	m00051: Fructose and mannose metabolism///m01100: Metabolic pathways///m01120: Microbial metabolism in diverse environments
R00291(5.1.3.2)	Yes	No	m00052: Galactose metabolism///m00520: Amino sugar and nucleotide sugar metabolism///m01100: Metabolic pathways

R00505(5.4.99.9)	Yes	No	rm0052: Galactose metabolism//rm00520: Amino sugar and nucleotide sugar metabolism//rm01100: Metabolic pathways
R00127(2.7.4.3)	Yes	Yes	rm00230: Purine metabolism//rm01100: Metabolic pathways//rm01110: Biosynthesis of secondary metabolites
R03815(1.8.1.4)	Yes	No	rm00260: Glycine, serine and threonine metabolism
R00192(3.3.1.1)	Yes	No	rm00270: Cysteine and methionine metabolism//rm01100: Metabolic pathways
R00177(2.5.1.6)	Yes	No	rm00270: Cysteine and methionine metabolism//rm01100: Metabolic pathways//rm01110: Biosynthesis of secondary metabolites//rm01230: Biosynthesis of amino acids
R08364(1.17.4.1)	Yes	No	rm00480: Glutathione metabolism
R00771(5.3.1.9)	Yes	No	rm00500: Starch and sucrose metabolism//rm01100: Metabolic pathways//rm01110: Biosynthesis of secondary metabolites
R02087(5.1.3.9)	Yes	No	rm00520: Amino sugar and nucleotide sugar metabolism//rm01100: Metabolic pathways
R02239(3.1.3.4)	Yes	No	rm00561: Glycerolipid metabolism//rm00564: Glycerophospholipid metabolism//rm01100: Metabolic pathways//rm01110: Biosynthesis of secondary metabolites
R02055(4.1.1.65)	No	Yes	rm00564: Glycerophospholipid metabolism//rm01100: Metabolic pathways//rm01110: Biosynthesis of secondary metabolites
R00842(1.1.1.94)	No	Yes	rm00564: Glycerophospholipid metabolism//rm01110: Biosynthesis of secondary metabolites
R06522(3.1.3.4)	Yes	No	rm00600: Sphingolipid metabolism
R06520(3.1.3.4)	Yes	No	rm00600: Sphingolipid metabolism//rm01100: Metabolic pathways
R06521(3.1.3.4)	Yes	No	rm00600: Sphingolipid metabolism//rm01100: Metabolic pathways
R00939(1.5.1.3)	No	Yes	rm00670: One carbon pool by folate//rm00790: Folate biosynthesis//rm01100: Metabolic pathways
R01068(4.1.2.13)	Yes	No	rm00680: Methane metabolism//rm00710: Carbon fixation in photosynthetic organisms//rm01100: Metabolic pathways//rm01120: Microbial metabolism in diverse environments//rm01200: Carbon metabolism
R01829(4.1.2.13)	Yes	Yes	rm00710: Carbon fixation in photosynthetic organisms//rm01100: Metabolic pathways//rm01120: Microbial metabolism in diverse environments//rm01200: Carbon metabolism
R06590(2.2.1.1)	Yes	No	rm01051: Biosynthesis of ansamycins//rm01110: Biosynthesis of secondary metabolites
R03596(1.8.1.9)	Yes	No	thioredoxin reductase (NADPH) [EC:1.8.1.9]//thioredoxin reductase (NADPH) [EC:1.8.1.9]
R09372(1.8.1.9)	Yes	No	thioredoxin reductase (NADPH) [EC:1.8.1.9]//thioredoxin reductase (NADPH) [EC:1.8.1.9]
R01567(2.7.1.21)	Yes	No	thymidine kinase [EC:2.7.1.21]
R02099(2.7.1.21)	Yes	No	thymidine kinase [EC:2.7.1.21]
R08233(2.7.1.21)	Yes	No	thymidine kinase [EC:2.7.1.21]
R02101(2.1.1.45)	No	Yes	thymidylate synthase [EC:2.1.1.45]//dihydrofolate reductase / thymidylate synthase [EC:1.5.1.3 2.1.1.45]
R01827(2.2.1.2)	Yes	No	transaldolase [EC:2.2.1.2]//transaldolase / glucose-6-phosphate isomerase [EC:2.2.1.2 5.3.1.9]

R01830(2.2.1.1)	Yes	Yes	transketolase [EC:2.2.1.1]
R01067(2.2.1.1)	Yes	No	transketolase [EC:2.2.1.1]
R00966(2.4.2.9)	Yes	Yes	uracil phosphoribosyltransferase [EC:2.4.2.9]//pyrimidine operon attenuation protein / uracil phosphoribosyltransferase [EC:2.4.2.9]
R00289(2.7.7.9)	Yes	No	UTP--glucose-1-phosphate uridylyltransferase [EC:2.7.7.9]//UDP-sugar pyrophosphorylase [EC:2.7.7.64]//UTP--glucose-1-phosphate uridylyltransferase [EC:2.7.7.9]
R00156(2.7.4.6)	No	Yes	rm00240: Pyrimidine metabolism//rm01100: Metabolic pathways
R00570(2.7.4.6)	No	Yes	rm00240: Pyrimidine metabolism//rm01100: Metabolic pathways
R02326(2.7.4.6)	No	Yes	rm00240: Pyrimidine metabolism//rm01100: Metabolic pathways
R02331(2.7.4.6)	No	Yes	rm00240: Pyrimidine metabolism//rm01100: Metabolic pathways
R02093(2.7.4.6)	No	Yes	rm00240: Pyrimidine metabolism//rm01100: Metabolic pathways
R02738(2.7.1.199)	Yes	Yes	rm00010: Glycolysis / Gluconeogenesis//rm00520: Amino sugar and nucleotide sugar metabolism//rm01100: Metabolic pathways
R01843(2.7.1.11)	Yes	Yes	
R02324	No	Yes	rm00760: Nicotinate and nicotinamide metabolism//rm01100: Metabolic pathways
R00137	No	Yes	rm00760: Nicotinate and nicotinamide metabolism//rm01100: Metabolic pathways
R00549	No	Yes	rm00740: Riboflavin metabolism//rm01100: Metabolic pathways//rm01110: Biosynthesis of secondary metabolites
R00161	No	Yes	rm00740: Riboflavin metabolism//rm01100: Metabolic pathways//rm01110: Biosynthesis of secondary metabolites
R00619	No	Yes	rm00730: Thiamine metabolism//rm01100: Metabolic pathways
R00177	No	Yes	rm00270: Cysteine and methionine metabolism//rm01100: Metabolic pathways//rm01110: Biosynthesis of secondary metabolites//rm01230: Biosynthesis of amino acids
R00173	No	Yes	rm00750: Vitamin B6 metabolism//rm01100: Metabolic pathways
R02971	No	Yes	rm00770: Pantothenate and CoA biosynthesis
R03035	No	Yes	rm00770: Pantothenate and CoA biosynthesis//rm01100: Metabolic pathways
R00130	No	Yes	rm00770: Pantothenate and CoA biosynthesis//rm01100: Metabolic pathways
R03018	No	Yes	rm00770: Pantothenate and CoA biosynthesis//rm01100: Metabolic pathways
R04391	No	Yes	rm00770: Pantothenate and CoA biosynthesis
R04231	No	Yes	rm00770: Pantothenate and CoA biosynthesis//rm01100: Metabolic pathways

Table C3S4 | Names of the enzymes and definition of each reaction involved in the comparison of the MBBs of the three networks under study.

Reaction	Enzyme name	Definition
R00156	ATP:UDP phosphotransferase	ATP + UDP ↔ ADP + UTP
R00158	ATP:UMP phosphotransferase	ATP + UMP ↔ ADP + UDP
R00332	ATP:GMP phosphotransferase	ATP + GMP ↔ ADP + GDP
R00430	GTP:pyruvate 2-O-phosphotransferase	GTP + Pyruvate ↔ GDP + Phosphoenolpyruvate
R00658	2-phospho-D-glycerate hydro-lyase (phosphoenolpyruvate-forming)	2-Phospho-D-glycerate ↔ Phosphoenolpyruvate + H ₂ O
R00842	sn-Glycerol-3-phosphate:NAD ⁺ 2-oxidoreductase	sn-Glycerol 3-phosphate + NAD ⁺ ↔ Glycerone phosphate + NADH + H ⁺
R00844	sn-Glycerol 3-phosphate:NADP ⁺ 2-oxidoreductase	sn-Glycerol 3-phosphate + NADP ⁺ ↔ Glycerone phosphate + NADPH + H ⁺
R00966	UMP:diphosphate phospho-alpha-D-ribosyltransferase	UMP + Diphosphate ↔ Uracil + 5-Phospho-alpha-D-ribose 1-diphosphate
R01049	ATP:D-ribose-5-phosphate diphosphotransferase	ATP + D-Ribose 5-phosphate ↔ AMP + 5-Phospho-alpha-D-ribose 1-diphosphate
R01056	D-ribose-5-phosphate aldose-ketose-isomerase	D-Ribose 5-phosphate ↔ D-Ribulose 5-phosphate
R01061	D-glyceraldehyde-3-phosphate:NAD ⁺ oxidoreductase (phosphorylating)	D-Glyceraldehyde 3-phosphate + Orthophosphate + NAD ⁺ ↔ 3-Phospho-D-glyceroyl phosphate + NADH + H ⁺
R01070	beta-D-fructose-1,6-bisphosphate D-glyceraldehyde-3-phosphate-lyase (glycerone-phosphate-forming)	beta-D-Fructose 1,6-bisphosphate ↔ Glycerone phosphate + D-Glyceraldehyde 3-phosphate
R01229	GMP:diphosphate 5-phospho-alpha-D-ribosyltransferase	GMP + Diphosphate ↔ Guanine + 5-Phospho-alpha-D-ribose 1-diphosphate
R01512	ATP:3-phospho-D-glycerate 1-phosphotransferase	ATP + 3-Phospho-D-glycerate ↔ ADP + 3-Phospho-D-glyceroyl phosphate
R01518	D-phosphoglycerate 2,3-phosphomutase	2-Phospho-D-glycerate ↔ 3-Phospho-D-glycerate
R01529	D-Ribulose-5-phosphate 3-epimerase	D-Ribulose 5-phosphate ↔ D-Xylulose 5-phosphate
R01641	sedoheptulose-7-phosphate-D-glyceraldehyde-3-phosphate glycolaldehyde transferase	Sedoheptulose 7-phosphate + D-Glyceraldehyde 3-phosphate ↔ D-Ribose 5-phosphate + D-Xylulose 5-phosphate
R01829	sedoheptulose 1,7-bisphosphate D-glyceraldehyde-3-phosphate-lyase	Sedoheptulose 1,7-bisphosphate ↔ Glycerone phosphate + D-Erythrose 4-phosphate
R01830	beta-D-Fructose 6-phosphate-D-glyceraldehyde-3-phosphate glycolaldehyde transferase	beta-D-Fructose 6-phosphate + D-Glyceraldehyde 3-phosphate ↔ D-Erythrose 4-phosphate + D-Xylulose 5-phosphate
R01843	ATP:Sedoheptulose 7-phosphate 1-phosphotransferase	ATP + Sedoheptulose 7-phosphate ↔ ADP + Sedoheptulose 1,7-bisphosphate
R02740	alpha-D-Glucose 6-phosphate ketol-isomerase	alpha-D-Glucose 6-phosphate ↔ beta-D-Fructose 6-phosphate
R04779	ATP:D-fructose-6-phosphate 1-phosphotransferase	ATP + beta-D-Fructose 6-phosphate ↔ ADP + beta-D-Fructose 1,6-bisphosphate
R02098	ATP:dUMP phosphotransferase	ATP + dUMP ↔ ADP + dUDP
R02331	ATP:dUDP phosphotransferase	ATP + dUDP ↔ ADP + dUTP
R00512	ATP:CMP phosphotransferase	ATP + CMP ↔ ADP + CDP
R00570	ATP:CDP phosphotransferase	ATP + CDP ↔ ADP + CTP
R02019	2'-Deoxyguanosine 5'-diphosphate:oxidized-thioredoxin 2'-oxidoreductase	dGDP + Thioredoxin disulfide + H ₂ O ↔ GDP + Thioredoxin
R00127	ATP:AMP phosphotransferase	ATP + AMP ↔ 2 ADP
R00190	AMP:diphosphate phospho-D-ribosyltransferase	AMP + Diphosphate ↔ Adenine + 5-Phospho-alpha-D-ribose 1-diphosphate
R01138	dATP:pyruvate 2-O-phosphotransferase	dATP + Pyruvate ↔ dADP + Phosphoenolpyruvate
R02017	2'-Deoxyadenosine 5'-diphosphate:oxidized-thioredoxin 2'-oxidoreductase	dADP + Thioredoxin disulfide + H ₂ O ↔ Thioredoxin + ADP
R00936	5,6,7,8-tetrahydrofolate:NAD ⁺ oxidoreductase	Tetrahydrofolate + NAD ⁺ ↔ Dihydrofolate + NADH + H ⁺
R00939	5,6,7,8-tetrahydrofolate:NADP ⁺ oxidoreductase	Tetrahydrofolate + NADP ⁺ ↔ Dihydrofolate + NADPH + H ⁺
R00945	5,10-Methylenetetrahydrofolate:glycine hydroxymethyltransferase	5,10-Methylenetetrahydrofolate + Glycine + H ₂ O ↔ Tetrahydrofolate + L-Serine
R02101	5,10-Methylenetetrahydrofolate:dUMP C-methyltransferase	dUMP + 5,10-Methylenetetrahydrofolate ↔ Dihydrofolate + dTMP
R02235	dihydrofolate:NAD ⁺ oxidoreductase	Dihydrofolate + NAD ⁺ ↔ Folate + NADH + H ⁺
R02236	dihydrofolate:NADP ⁺ oxidoreductase	Dihydrofolate + NADP ⁺ ↔ Folate + NADPH + H ⁺
R00137	ATP:nicotinamide-nucleotide adenyltransferase	ATP + Nicotinamide D-ribonucleotide ↔ Diphosphate + NAD ⁺
R02093	ATP:dTDP phosphotransferase	ATP + dTDP ↔ ADP + dTTP
R02094	ATP:dTMP phosphotransferase	ATP + dTMP ↔ ADP + dTDP
R00014	pyruvate:thiamin diphosphate acetaldehydetransferase (decarboxylating)	Pyruvate + Thiamin diphosphate ↔ 2-(alpha-Hydroxyethyl)thiamine diphosphate + CO ₂
R00230	acetyl-CoA:phosphate acetyltransferase	Acetyl-CoA + Orthophosphate ↔ CoA + Acetyl phosphate
R00315	ATP:acetate phosphotransferase	ATP + Acetate ↔ ADP + Acetyl phosphate
R02569	acetyl-CoA:enzyme N6-(dihydrolipoyl)lysine S-acetyltransferase	Acetyl-CoA + Enzyme N6-(dihydrolipoyl)lysine ↔ CoA + [Dihydrolipoyllysine-residue acetyltransferase] S-acetyldihydrolipoyllysine
R03270	pyruvate dehydrogenase	2-(alpha-Hydroxyethyl)thiamine diphosphate + Enzyme N6-(lipoyl)lysine ↔ [Dihydrolipoyllysine-residue acetyltransferase] S-acetyldihydrolipoyllysine + Thiamin diphosphate

R07618	enzyme N6-(dihydropolipoyl)lysine:NAD+ oxidoreductase	Enzyme N6-(dihydropolipoyl)lysine + NAD+ ↔ Enzyme N6-(lipoyl)lysine + NADH + H+
R01126	inosine 5'-monophosphate phosphohydrolase	IMP + H2O ↔ Inosine + Orthophosphate
R01132	IMP:diphosphate phospho-D-ribosyltransferase	IMP + Diphosphate ↔ Hypoxanthine + 5-Phospho-alpha-D-ribose 1-diphosphate
R01863	inosine:phosphate alpha-D-ribosyltransferase	Inosine + Orthophosphate ↔ Hypoxanthine + alpha-D-Ribose 1-phosphate
R02142	XMP:pyrophosphate phosphoribosyltransferase	Xanthosine 5'-phosphate + Diphosphate ↔ Xanthine + 5-Phospho-alpha-D-ribose 1-diphosphate
R02297	Xanthosine:orthophosphate ribosyltransferase	Xanthosine + Orthophosphate ↔ Xanthine + alpha-D-Ribose 1-phosphate
R02719	xanthosine 5'-phosphate phosphohydrolase	Xanthosine 5'-phosphate + H2O ↔ Xanthosine + Orthophosphate
R00921	propanoyl-CoA:phosphate propanoyltransferase	Propanoyl-CoA + Orthophosphate ↔ Propanoyl phosphate + CoA
R01353	ATP:propanoate phosphotransferase	ATP + Propanoate ↔ ADP + Propanoyl phosphate
R00512	ATP:CMF phosphotransferase	ATP + CMP ↔ ADP + CDP
R00158	ATP:UMP phosphotransferase	ATP + UMP ↔ ADP + UDP
R00289	UTP:alpha-D-glucose 1-phosphate unidyltransferase	UTP + D-Glucose 1-phosphate ↔ Diphosphate + UDP-glucose
R00291	UDP-glucose 4-epimerase//UDP-alpha-D-glucose 4-epimerase	UDP-glucose ↔ UDP-alpha-D-galactose
R00332	ATP:GMP phosphotransferase	ATP + GMP ↔ ADP + GDP
R00430	GTP:pyruvate 2-O-phosphotransferase	GTP + Pyruvate ↔ GDP + Phosphoenolpyruvate
R00505	UDP-alpha-D-galactopyranose furanomutase	UDP-alpha-D-galactose ↔ UDP-alpha-D-galactofuranose
R00658	2-phospho-D-glycerate hydro-lyase (phosphoenolpyruvate-forming)	2-Phospho-D-glycerate ↔ Phosphoenolpyruvate + H2O
R00959	alpha-D-Glucose 1-phosphate 1,6-phosphomutase	D-Glucose 1-phosphate ↔ alpha-D-Glucose 6-phosphate
R00966	UMP:diphosphate phospho-alpha-D-ribosyltransferase	UMP + Diphosphate ↔ Uracil + 5-Phospho-alpha-D-ribose 1-diphosphate
R01015	D-glyceraldehyde-3-phosphate aldose-ketose-isomerase	D-Glyceraldehyde 3-phosphate ↔ Glycerone phosphate
R01049	ATP:D-ribose-5-phosphate diphosphotransferase	ATP + D-Ribose 5-phosphate ↔ AMP + 5-Phospho-alpha-D-ribose 1-diphosphate
R01056	D-ribose-5-phosphate aldose-ketose-isomerase	D-Ribose 5-phosphate ↔ D-Ribulose 5-phosphate
R01057	D-Ribose 1,5-phosphomutase	alpha-D-Ribose 1-phosphate ↔ D-Ribose 5-phosphate
R01058	D-glyceraldehyde 3-phosphate:NADP+ oxidoreductase	D-Glyceraldehyde 3-phosphate + NADP+ + H2O ↔ 3-Phospho-D-glycerate + NADPH + H+
R01061	D-glyceraldehyde-3-phosphate:NAD+ oxidoreductase (phosphorylating)	D-Glyceraldehyde 3-phosphate + Orthophosphate + NAD+ ↔ 3-Phospho-D-glyceroyl phosphate + NADH + H+

R01066	2-deoxy-D-ribose-5-phosphate acetaldehyde-lyase (D-glyceraldehyde-3-phosphate-forming)	2-Deoxy-D-ribose 5-phosphate ↔ D-Glyceraldehyde 3-phosphate + Acetaldehyde
R01067	D-Fructose 6-phosphate:D-glyceraldehyde-3-phosphate glycolaldehyde transferase	D-Fructose 6-phosphate + D-Glyceraldehyde 3-phosphate ↔ D-Erythrose 4-phosphate + D-Xylulose 5-phosphate
R01068	D-fructose-1,6-bisphosphate D-glyceraldehyde-3-phosphate-lyase (glycerone-phosphate-forming)	D-Fructose 1,6-bisphosphate ↔ Glycerone phosphate + D-Glyceraldehyde 3-phosphate
R01070	beta-D-fructose-1,6-bisphosphate D-glyceraldehyde-3-phosphate-lyase (glycerone-phosphate-forming)	beta-D-Fructose 1,6-bisphosphate ↔ Glycerone phosphate + D-Glyceraldehyde 3-phosphate
R01227	guanosine 5'-monophosphate phosphohydrolase	GMP + H2O ↔ Guanosine + Orthophosphate
R01229	GMP:diphosphate 5-phospho-alpha-D-ribosyltransferase	GMP + Diphosphate ↔ Guanine + 5-Phospho-alpha-D-ribose 1-diphosphate
R01229	GMP:diphosphate 5-phospho-alpha-D-ribosyltransferase	GMP + Diphosphate ↔ Guanine + 5-Phospho-alpha-D-ribose 1-diphosphate
R01512	ATP:3-phospho-D-glycerate 1-phosphotransferase	ATP + 3-Phospho-D-glycerate ↔ ADP + 3-Phospho-D-glyceroyl phosphate
R01518	D-phosphoglycerate 2,3-phosphomutase	2-Phospho-D-glycerate ↔ 3-Phospho-D-glycerate
R01529	D-Ribulose-5-phosphate 3-epimerase	D-Ribulose 5-phosphate ↔ D-Xylulose 5-phosphate
R01641	sedoheptulose-7-phosphate:D-glyceraldehyde-3-phosphate glycolaldehyde transferase	Sedoheptulose 7-phosphate + D-Glyceraldehyde 3-phosphate ↔ D-Ribose 5-phosphate + D-Xylulose 5-phosphate
R01819	D-mannose-6-phosphate aldose-ketose-isomerase	D-Mannose 6-phosphate ↔ beta-D-Fructose 6-phosphate
R01827	sedoheptulose-7-phosphate:D-glyceraldehyde-3-phosphate glyceronetransferase	Sedoheptulose 7-phosphate + D-Glyceraldehyde 3-phosphate ↔ D-Erythrose 4-phosphate + beta-D-Fructose 6-phosphate
R01829	sedoheptulose 1,7-bisphosphate D-glyceraldehyde-3-phosphate-lyase	Sedoheptulose 1,7-bisphosphate ↔ Glycerone phosphate + D-Erythrose 4-phosphate
R01830	beta-D-Fructose 6-phosphate:D-glyceraldehyde-3-phosphate glycolaldehyde transferase	beta-D-Fructose 6-phosphate + D-Glyceraldehyde 3-phosphate ↔ D-Erythrose 4-phosphate + D-Xylulose 5-phosphate
R01843	ATP:Sedoheptulose 7-phosphate 1-phosphotransferase	ATP + Sedoheptulose 7-phosphate ↔ ADP + Sedoheptulose 1,7-bisphosphate
R01967	ATP:deoxyguanosine 5'-phosphotransferase	ATP + Deoxyguanosine ↔ ADP + dGMP
R01968	2'-deoxyguanosine 5'-monophosphate phosphohydrolase	dGMP + H2O ↔ Deoxyguanosine + Orthophosphate
R01969	Deoxyguanosine:orthophosphate ribosyltransferase	Deoxyguanosine + Orthophosphate ↔ Guanine + 2-Deoxy-D-ribose 1-phosphate
R02018	2'-Deoxyuridine 5'-diphosphate:oxidized-thioredoxin 2'-oxidoreductase	dUDP + Thioredoxin disulfide + H2O ↔ Thioredoxin + UDP
R02019	2'-Deoxyguanosine 5'-diphosphate:oxidized-thioredoxin 2'-oxidoreductase	dGDP + Thioredoxin disulfide + H2O ↔ GDP + Thioredoxin
R02090	ATP:dGMP phosphotransferase	ATP + dGMP ↔ ADP + dGDP
R02098	ATP:dUMP phosphotransferase	ATP + dUMP ↔ ADP + dUDP

R02099	ATP:deoxyuridine 5'-phosphotransferase	ATP + Deoxyuridine ↔ ADP + dUMP
R02102	2'-deoxyuridine 5'-monophosphate phosphohydrolase	dUMP + H ₂ O ↔ Deoxyuridine + Orthophosphate
R02102	2'-deoxyuridine 5'-monophosphate phosphohydrolase	dUMP + H ₂ O ↔ Deoxyuridine + Orthophosphate
R02147	guanosine phosphate alpha-D-ribosyltransferase	Guanosine + Orthophosphate ↔ Guanine + alpha-D-Ribose 1-phosphate
R02484	deoxyuridine:orthophosphate 2-deoxy-D-ribosyltransferase://deoxyuridine:orthophosphate ribosyltransferase	Deoxyuridine + Orthophosphate ↔ Uracil + 2-Deoxy-D-ribose 1-phosphate
R02568	D-fructose 1-phosphate D-glyceraldehyde-3-phosphate-lyase	D-Fructose 1-phosphate ↔ Glycerone phosphate + D-Glyceraldehyde
R02739	alpha-D-Glucose 6-phosphate ketol-isomerase	alpha-D-Glucose 6-phosphate ↔ beta-D-Glucose 6-phosphate
R02740	alpha-D-Glucose 6-phosphate ketol-isomerase	alpha-D-Glucose 6-phosphate ↔ beta-D-Fructose 6-phosphate
R02749	2-deoxy-D-ribose 1-phosphate 1,5-phosphomutase	2-Deoxy-D-ribose 1-phosphate ↔ 2-Deoxy-D-ribose 5-phosphate
R03321	beta-D-Glucose 6-phosphate ketol-isomerase	beta-D-Glucose 6-phosphate ↔ beta-D-Fructose 6-phosphate
R04779	ATP:D-fructose 6-phosphate 1-phosphotransferase	ATP + beta-D-Fructose 6-phosphate ↔ ADP + beta-D-Fructose 1,6-bisphosphate
R01664	2'-deoxycytidine 5'-monophosphate phosphohydrolase	dCMP + H ₂ O ↔ Deoxycytidine + Orthophosphate
R01664	2'-deoxycytidine 5'-monophosphate phosphohydrolase	dCMP + H ₂ O ↔ Deoxycytidine + Orthophosphate
R01665	ATP:dCMP phosphotransferase	ATP + dCMP ↔ ADP + dCDP
R01666	ATP:deoxycytidine 5'-phosphotransferase	ATP + Deoxycytidine ↔ ADP + dCMP
R01667	dCDP nucleotidohydrolase	dCDP + H ₂ O ↔ dCMP + Orthophosphate
R00127	ATP:AMP phosphotransferase	ATP + AMP ↔ 2 ADP
R00183	adenosine 5'-monophosphate phosphohydrolase	AMP + H ₂ O ↔ Adenosine + Orthophosphate
R00185	ATP:adenosine 5'-phosphotransferase	ATP + Adenosine ↔ ADP + AMP
R00190	AMP:diphosphate phospho-D-ribosyltransferase	AMP + Diphosphate ↔ Adenine + 5-Phospho-alpha-D-ribose 1-diphosphate
R01138	dATP:pyruvate 2-O-phosphotransferase	dATP + Pyruvate ↔ dADP + Phosphoenolpyruvate
R01547	ATP:dAMP phosphotransferase	ATP + dAMP ↔ ADP + dADP
R01547	ATP:dAMP phosphotransferase	ATP + dAMP ↔ ADP + dADP
R01561	adenosine phosphate alpha-D-ribosyltransferase	Adenosine + Orthophosphate ↔ Adenine + alpha-D-Ribose 1-phosphate
R02017	2'-Deoxyadenosine 5'-diphosphate oxidized-thioredoxin 2'-oxidoreductase	dADP + Thioredoxin disulfide + H ₂ O ↔ Thioredoxin + ADP
R02088	2'-deoxyadenosine 5'-monophosphate phosphohydrolase	dAMP + H ₂ O ↔ Deoxyadenosine + Orthophosphate

R02089	ATP:deoxyadenosine 5'-phosphotransferase	ATP + Deoxyadenosine ↔ ADP + dAMP
R02557	Deoxyadenosine:orthophosphate ribosyltransferase	Deoxyadenosine + Orthophosphate ↔ Adenine + 2-Deoxy-D-ribose 1-phosphate
R00942	Tetrahydrofolate:L-glutamate gamma-ligase (ADP-forming)	ATP + Tetrahydrofolate + L-Glutamate ↔ ADP + Orthophosphate + THF-L-glutamate
R00945	5,10-Methylenetetrahydrofolate:glycine hydroxymethyltransferase	5,10-Methylenetetrahydrofolate + Glycine + H ₂ O ↔ Tetrahydrofolate + L-Serine
R01220	5,10-methylenetetrahydrofolate:NADP+ oxidoreductase	5,10-Methylenetetrahydrofolate + NADP+ ↔ 5,10-Methylenetetrahydrofolate + NADPH
R01655	5,10-Methylenetetrahydrofolate 5-hydrolase (decyclizing)	5,10-Methylenetetrahydrofolate + H ₂ O ↔ 10-Formyltetrahydrofolate + H+
R03940	10-Formyltetrahydrofolate:L-methionyl-tRNA N-formyltransferase	L-Methionyl-tRNA + 10-Formyltetrahydrofolate ↔ Tetrahydrofolate + N-Formylmethionyl-tRNA
R04241	tetrahydropteroyl-gamma-polyglutamate:L-glutamate gamma-ligase (ADP-forming)	ATP + THF-polyglutamate(n) + L-Glutamate ↔ ADP + Orthophosphate + THF-polyglutamate(n+1)
R00137	ATP:nicotinamide nucleotide adenyltransferase	ATP + Nicotinamide D-ribonucleotide ↔ Diphosphate + NAD+
R01271	nicotinamide-D-ribonucleotide:diphosphate phospho-alpha-D-ribosyltransferase	Nicotinamide D-ribonucleotide + Diphosphate ↔ Nicotinamide + 5-Phospho-alpha-D-ribose 1-diphosphate
R02294	N-Ribosylnicotinamide:orthophosphate ribosyltransferase	Nicotinamide-beta-riboside + Orthophosphate ↔ Nicotinamide + alpha-D-Ribose 1-phosphate
R02323	nicotinamide ribonucleotide phosphohydrolase	Nicotinamide D-ribonucleotide + H ₂ O ↔ Nicotinamide-beta-riboside + Orthophosphate
R01567	ATP:thymidine 5'-phosphotransferase	ATP + Thymidine ↔ ADP + dTMP
R01569	thymidylate 5'-phosphohydrolase	dTMP + H ₂ O ↔ Thymidine + Orthophosphate
R02094	ATP:dTMP phosphotransferase	ATP + dTMP ↔ ADP + dTDP
R00435	ATP:polynucleotide adenyltransferase://ATP:RNA adenyltransferase	ATP + RNA ↔ Diphosphate + RNA
R00441	GTP:RNA guanylyltransferase (DNA-directed)//GTP:RNA guanylyltransferase (RNA-directed)	GTP + RNA ↔ Diphosphate + RNA
R00375	Deoxyadenosine 5'-triphosphate:DNA deoxynucleotidyltransferase (DNA-directed)	dATP + DNA ↔ Diphosphate + DNA
R00376	Deoxyguanosine 5'-triphosphate:DNA deoxynucleotidyltransferase (DNA-directed)	dGTP + DNA ↔ Diphosphate + DNA

Chapter 4

Note: All supplementary figures in this chapter are HUGE and can only be visualized correctly with the proper PDF. Contact the author for the complete original files.



Figure C4S1 | The reaction graph of the pan-metabolic network of endosymbiotic bacteria of insects consisting of 2,964 reactions (nodes), where 598 are reversible reactions, and 1,883 compounds (edges). This PDF file is very heavy because users can zoom into any of the reactions and compounds to explore if determined metabolic pathways are carried out by any or some endosymbiotic bacteria of insects.

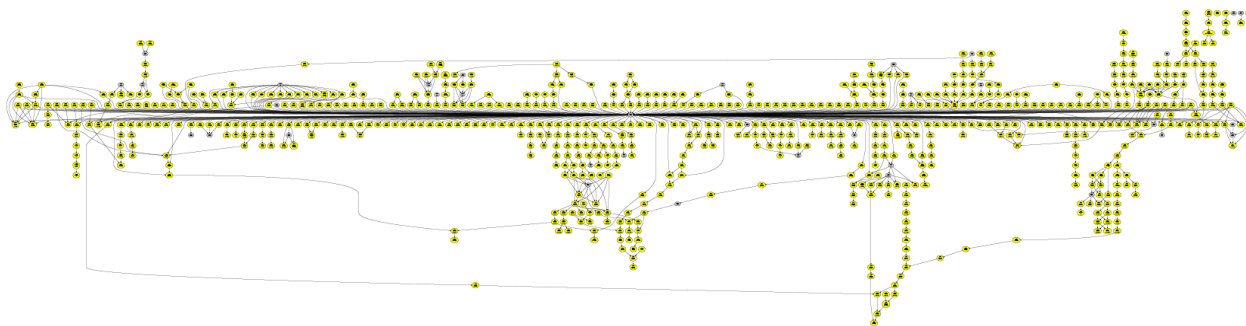


Figure C4S2 | The pan-mDAG of endosymbiotic bacteria of insects consisting of 1,081 MBBs, where 1034 MBBs have only one reaction and 47 have more than one. This PDF file is very heavy because users can zoom into any of the reactions and MBBs for further details.

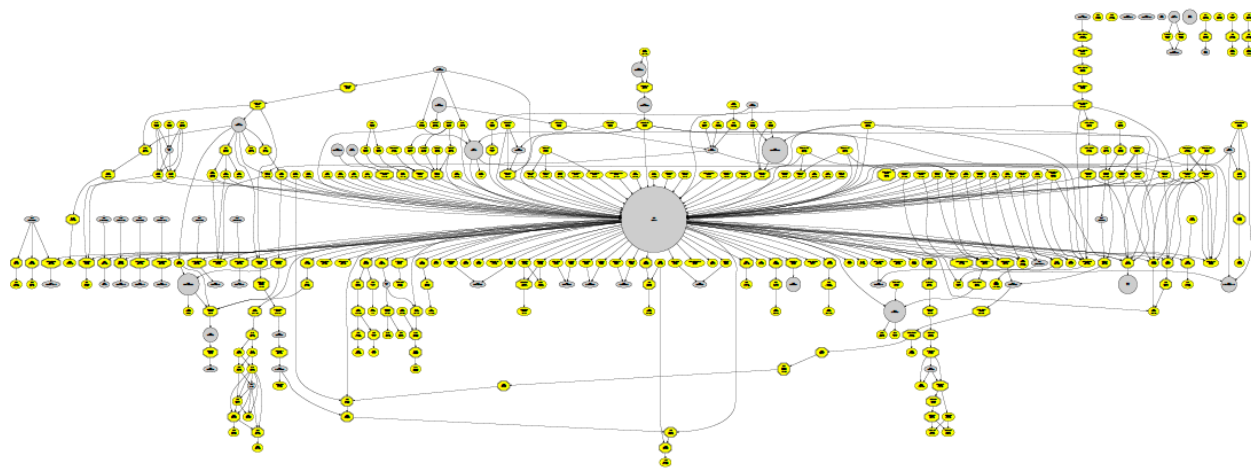


Figure C4S3 | M-DAG of *Nocardioopsis alba* ATCC BAA-2165, to compare with supplementary figure C4S3. This m- DAG has a total of 533 MBBs including 1,219 reactions and 954 compounds.

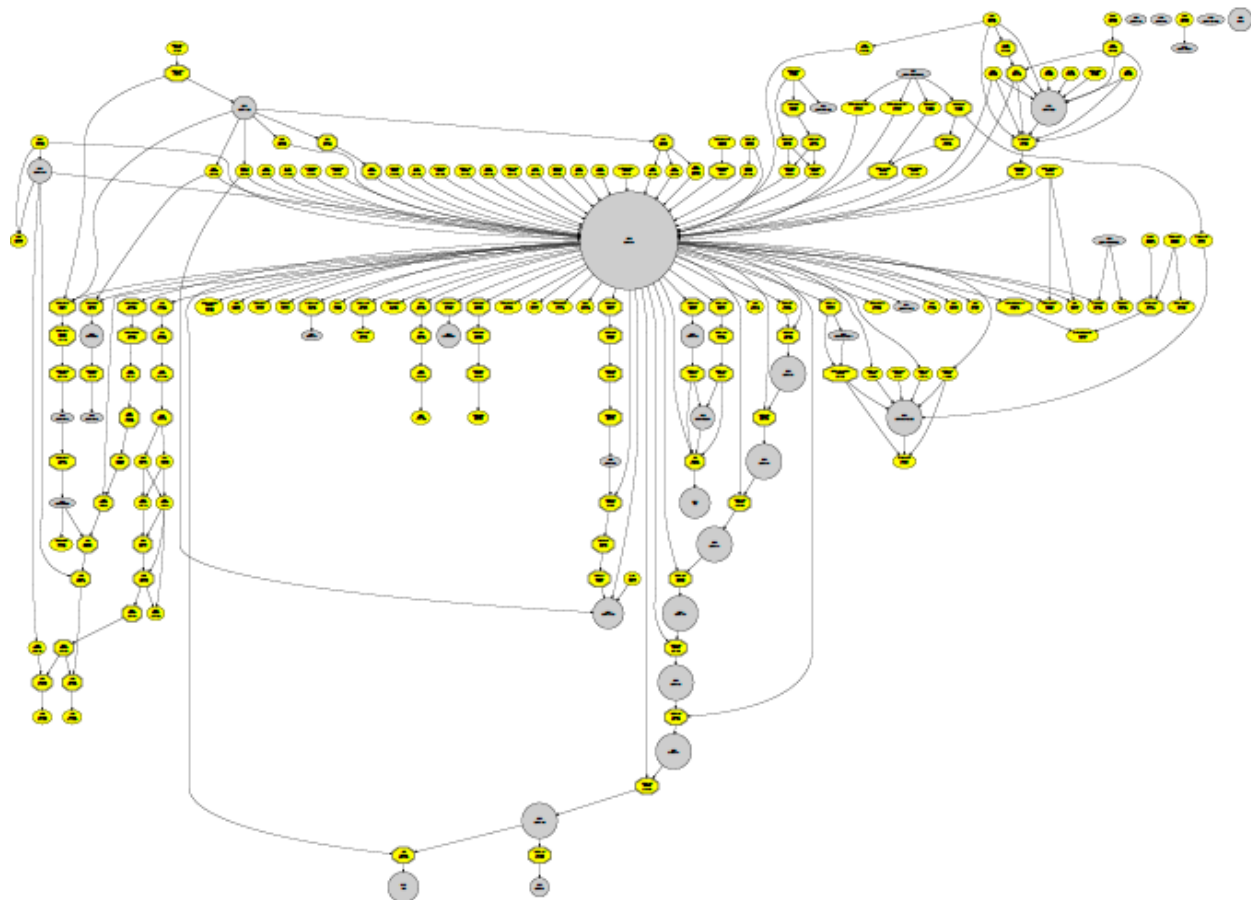


Figure C4S4 | M-DAG of *Bifidobacterium longum* NCC2705, to compare with supplementary figure C4S2. This m- DAG has a total of 290 MBBs including 743 reactions and 552 compounds.

Table C4S1 | Strains that are included in our data set. The first column depicts the group to which each strain has been assigned, and a dash indicates the strain does not belong to any group. Lastly, two dashes in the host column mean the strain is not a symbiont of insects.

Group	Strain	Taxonomy	Genome size	Number genes	Host
BAUM	Candidatus <i>Baumannia cicadellincola</i> B-GSS	Gammaproteobacteria	636137	550	<i>Graphocephala atropunctata</i>
BAUM	<i>Baumannia cicadellincola</i> Hc (<i>Homalodisca coagulata</i>)	Gammaproteobacteria	686194	595	<i>Homalodisca coagulata</i>
BAUM	Candidatus <i>Baumannia cicadellincola</i> BGSS	Gammaproteobacteria	759425	689	<i>Graphocephala atropunctata</i>
BLAT	<i>Blattabacterium</i> sp. (<i>Mastotermes darwiniensis</i>)	Bacteroidetes	590336	544	<i>Mastotermes darwiniensis</i>

BLAT	<i>Blattabacterium</i> Cpu (<i>Cryptocercus punctulatus</i>)	Bacteroidetes	609561	548	<i>Cryptocercus punctulatus</i>
BLAT	<i>Blattabacterium</i> sp. (<i>Panesthia angustipennis spadica</i>) BPAA	Bacteroidetes	632490	575	<i>Panesthia angustipennis spadica</i>
BLAT	<i>Blattabacterium</i> sp. (<i>Blaberus giganteus</i>)	Bacteroidetes	632588	577	<i>Blaberus giganteus</i>
BLAT	<i>Blattabacterium</i> sp. (<i>Blatta orientalis</i>)	Bacteroidetes	638184	579	<i>Blatta orientalis</i>
BLAT	<i>Blattabacterium</i> BPLAN (<i>Periplaneta americana</i>)	Bacteroidetes	640442	582	<i>Periplaneta americana</i>
BLAT	<i>Blattabacterium</i> sp. (<i>Nauphoeta cinerea</i>)	Bacteroidetes	626626	586	<i>Nauphoeta cinerea</i>
BLAT	<i>Blattabacterium</i> Bge (<i>Blattella germanica</i>)	Bacteroidetes	640935	590	<i>Blattella germanica</i>
BLOC	Candidatus <i>Blochmannia floridanus</i> (<i>Camponotus floridanus</i>)	Gammaproteobacteria	705557	583	<i>Camponotus floridanus</i>
BLOC	<i>Blochmannia</i> endosymbiont of <i>Camponotus</i> (<i>Colobopsis</i>) <i>obliquus</i> 757	Gammaproteobacteria	773940	584	<i>Colobopsis</i>
BLOC	Candidatus <i>Blochmannia vafer</i> (<i>Camponotus vafer</i>)	Gammaproteobacteria	722585	587	<i>Camponotus vafer</i>
BLOC	<i>Blochmannia</i> endosymbiont of <i>Polyrhachis</i> (<i>Hedomyrma</i>) <i>turneri</i> 675	Gammaproteobacteria	749321	589	<i>Polyrhachis turneri</i>
BLOC	Candidatus <i>Blochmannia chromaiodes</i> (<i>Camponotus chromaiodes</i>)	Gammaproteobacteria	791219	609	<i>Camponotus chromaiodes</i>
BLOC	Candidatus <i>Blochmannia pennsylvanicus</i> (<i>Camponotus pennsylvanicus</i>)	Gammaproteobacteria	791654	610	<i>Camponotus pennsylvanicus</i>
BUCH	<i>Buchnera aphidicola</i> (<i>Cinara tujafilina</i>)	Gammaproteobacteria	444925	359	<i>Cinara tujafilina</i>
BUCH	<i>Buchnera aphidicola</i> BCc	Gammaproteobacteria	422434	362	<i>Cinara cedri</i>
BUCH	<i>Buchnera aphidicola</i> JF98 (<i>Acyrtosiphon pisum</i>)	Gammaproteobacteria	641771	477	<i>Acyrtosiphon pisum</i>
BUCH	<i>Buchnera aphidicola</i> Bp (<i>Baizongia pistaciae</i>)	Gammaproteobacteria	618379	507	<i>Baizongia pistaciae</i>
BUCH	<i>Buchnera aphidicola</i> Ua (<i>Uroleucon ambrosiae</i>)	Gammaproteobacteria	627953	538	<i>Uroleucon ambrosiae</i>
BUCH	<i>Buchnera aphidicola</i> Sg (<i>Schizaphis graminum</i>)	Gammaproteobacteria	641454	545	<i>Schizaphis graminum</i>
BUCH	<i>Buchnera aphidicola</i> Tuc7 (<i>Acyrtosiphon pisum</i>)	Gammaproteobacteria	641895	553	<i>Acyrtosiphon pisum</i>
BUCH	<i>Buchnera aphidicola</i> 5A (<i>Acyrtosiphon</i>	Gammaproteobacteria	642122	555	<i>Acyrtosiphon pisum</i>

	<i>pisum</i>)				
BUCH	<i>Buchnera aphidicola</i> (Aphis glycines) BA9	Gammaproteobacteria	638851	562	<i>Aphis glycines</i>
BUCH	<i>Buchnera aphidicola</i> Ak (<i>Acyrtosiphon kondoi</i>)	Gammaproteobacteria	653223	569	<i>Acyrtosiphon kondoi</i>
BUCH	<i>Buchnera aphidicola</i> G002 (<i>Myzus persicae</i>)	Gammaproteobacteria	651316	572	<i>Myzus persicae</i>
BUCH	<i>Buchnera aphidicola</i> TLW03 (<i>Acyrtosiphon pisum</i>)	Gammaproteobacteria	641770	573	<i>Acyrtosiphon pisum</i>
BUCH	<i>Buchnera aphidicola</i> W106 (<i>Myzus persicae</i>)	Gammaproteobacteria	651304	573	<i>Myzus persicae</i>
BUCH	<i>Buchnera aphidicola</i> F009 (<i>Myzus persicae</i>)	Gammaproteobacteria	651310	573	<i>Myzus persicae</i>
BUCH	<i>Buchnera aphidicola</i> APS (<i>Acyrtosiphon pisum</i>)	Gammaproteobacteria	655725	574	<i>Acyrtosiphon pisum</i>
BUCH	<i>Buchnera aphidicola</i> USDA (<i>Myzus persicae</i>)	Gammaproteobacteria	651306	575	<i>Myzus persicae</i>
BUCH	<i>Buchnera aphidicola</i> LL01 (<i>Acyrtosiphon pisum</i>)	Gammaproteobacteria	641799	577	<i>Acyrtosiphon pisum</i>
BUCH	<i>Buchnera aphidicola</i> JF99 (<i>Acyrtosiphon pisum</i>)	Gammaproteobacteria	641716	590	<i>Acyrtosiphon pisum</i>
CARS	Candidatus <i>Carsonella ruddii</i> HT (<i>Heteropsylla texana</i>)	Gammaproteobacteria	157543	180	<i>Heteropsylla texana</i>
CARS	Candidatus <i>Carsonella ruddii</i> PV (<i>Pachypsylla venusta</i>)	Gammaproteobacteria	159662	182	<i>Pachypsylla venusta</i>
CARS	Candidatus <i>Carsonella ruddii</i> PC (<i>Pachypsylla celtidis</i>)	Gammaproteobacteria	159923	182	<i>Pachypsylla celtidis</i>
CARS	Candidatus <i>Carsonella ruddii</i> CS (<i>Ctenarytaina spatulata</i>)	Gammaproteobacteria	162504	190	<i>Ctenarytaina spatulata</i>
CARS	Candidatus <i>Carsonella ruddii</i> CE (<i>Ctenarytaina eucalypti</i>)	Gammaproteobacteria	162589	190	<i>Ctenarytaina eucalypti</i>
CARS	Candidatus <i>Carsonella ruddii</i> HC (<i>Heteropsylla cubana</i>)	Gammaproteobacteria	166163	192	<i>Heteropsylla cubana</i>
CARS	Candidatus <i>Carsonella ruddii</i> DC (<i>Diaphorina citri</i>)	Gammaproteobacteria	174014	207	<i>Diaphorina citri</i>
HODG	Candidatus <i>Hodgkinia cicadicola</i> TETUND1	Alphaproteobacteria	133698	121	<i>Tettigades undata</i>
HODG	Candidatus <i>Hodgkinia cicadicola</i> TETUND2	Alphaproteobacteria	140570	140	<i>Tettigades undata</i>
HODG	Candidatus <i>Hodgkinia cicadicola</i> Dsem	Alphaproteobacteria	143795	169	<i>Diceroprocta semicincta</i>

HODG	Candidatus <i>Hodgkinia cicadicola</i> TETULN	Alphaproteobacteria	150297	170	<i>Tettigades ulnaria</i>
PORT	Candidatus <i>Portiera aleyrodidarum</i> BT-QVLC	Gammaproteobacteria	357472	247	<i>Bemisia tabaci</i>
PORT	Candidatus <i>Portiera aleyrodidarum</i> MED	Gammaproteobacteria	357461	255	<i>Bemisia tabaci</i>
PORT	Candidatus <i>Portiera aleyrodidarum</i> BT-B	Gammaproteobacteria	358242	256	<i>Bemisia tabaci</i>
PORT	Candidatus <i>Portiera aleyrodidarum</i> TV-BCN	Gammaproteobacteria	280822	268	<i>Trialeurodes vaporariorum</i>
PORT	Candidatus <i>Portiera aleyrodidarum</i> TV	Gammaproteobacteria	280663	269	<i>Trialeurodes vaporariorum</i>
PORT	Candidatus <i>Portiera aleyrodidarum</i> BT-B-HRs	Gammaproteobacteria	351658	273	<i>Bemisia tabaci</i>
PORT	Candidatus <i>Portiera aleyrodidarum</i> AD-CAI	Gammaproteobacteria	290195	278	<i>Aleurodicus dispersus</i>
PORT	Candidatus <i>Portiera aleyrodidarum</i> AF-CAI	Gammaproteobacteria	290376	278	<i>Aleurodicus floccissimus</i>
SERR	<i>Serratia symbiotica</i> <i>Cinara cedri</i>	Gammaproteobacteria	1762765	672	<i>Cinara cedri</i>
SERR	<i>Serratia marcescens</i> FGI94	Gammaproteobacteria	4858216	4361	--
SERR	<i>Serratia</i> sp. SCBI	Gammaproteobacteria	5101896	4672	--
SERR	<i>Serratia plymuthica</i> 4Rx13	Gammaproteobacteria	5403731	4695	--
SERR	<i>Serratia marcescens</i> subsp. <i>marcescens</i> Db11	Gammaproteobacteria	5113802	4709	--
SERR	<i>Serratia</i> sp. FS14	Gammaproteobacteria	5249875	4761	--
SERR	<i>Serratia marcescens</i> WW4	Gammaproteobacteria	5244703	4809	--
SERR	<i>Serratia liquefaciens</i> ATCC 27592	Gammaproteobacteria	5282719	4894	--
SERR	<i>Serratia proteamaculans</i> 568	Gammaproteobacteria	5495657	4942	--
SERR	<i>Serratia plymuthica</i> AS9	Gammaproteobacteria	5442880	4952	--
SERR	<i>Serratia</i> sp. AS12	Gammaproteobacteria	5443009	4952	--
SERR	<i>Serratia marcescens</i> SM39	Gammaproteobacteria	5326023	4970	--
SERR	<i>Serratia plymuthica</i> S13	Gammaproteobacteria	5467306	4991	--
SERR	<i>Serratia fonticola</i> DSM 4576	Gammaproteobacteria	6000511	5098	--
SODA	<i>Sodalis endosymbiont of Henestaris halophilus</i>	Gammaproteobacteria	1622395	713	<i>Henestaris halophilus</i>
SODA	Candidatus <i>Sodalis pierantonius</i>	Gammaproteobacteria	4513140	2309	<i>Sitophilus oryzae</i>
SODA	<i>Sodalis glossinidius</i> (<i>Glossina</i> spp.)	Gammaproteobacteria	4292502	2516	<i>Glossina morsitans</i>

					<i>morsitans</i>
SODA	<i>Sodalis praecaptivus</i>	Gammaproteobacteria	5159425	4282	--
SULC	Candidatus <i>Sulcia muelleri</i> ML	Bacteroidetes	190405	187	<i>Dalbulus maidis</i>
SULC	Candidatus <i>Sulcia muelleri</i> Sulcia-ALF	Bacteroidetes	190733	188	<i>Macrosteles quadrilineatus</i>
SULC	Candidatus <i>Sulcia muelleri</i> DMIN	Bacteroidetes	243933	226	<i>Draeculacephala</i>
SULC	Candidatus <i>Sulcia muelleri</i> BGSS	Bacteroidetes	244618	227	<i>Graphocephala atropunctata</i>
SULC	Candidatus <i>Sulcia muelleri</i> GWSS	Bacteroidetes	245530	227	<i>Homalodisca vitripennis</i>
SULC	Candidatus <i>Sulcia muelleri</i> SMDSEM	Bacteroidetes	276984	242	<i>Diceroprocta semicincta</i>
SULC	Candidatus <i>Sulcia muelleri</i> CARI	Bacteroidetes	276511	246	<i>Clastoptera arizonana</i>
SULC	Candidatus <i>Sulcia muelleri</i> TETUND	Bacteroidetes	270029	247	<i>Tettigades undata</i>
SULC	Candidatus <i>Sulcia muelleri</i> PSPU	Bacteroidetes	285352	251	<i>Philaenus spumarius</i>
TREM	Candidatus <i>Tremblaya princeps</i> PCVAL	Betaproteobacteria	138931	116	<i>Planococcus citri</i>
TREM	Candidatus <i>Tremblaya phenacola</i> PAVE	Betaproteobacteria	171500	175	<i>Phenacoccus avenae</i>
WOLB	<i>Wolbachia</i> wHa (<i>Drosophila simulans</i>)	Alphaproteobacteria	1295804	1009	<i>Drosophila simulans</i>
WOLB	<i>Wolbachia</i> wNo (<i>Drosophila simulans</i>)	Alphaproteobacteria	1301823	1040	<i>Drosophila simulans</i>
WOLB	<i>Wolbachia</i> wRi (<i>Drosophila simulans</i>)	Alphaproteobacteria	1445873	1150	<i>Drosophila simulans</i>
WOLB	<i>Wolbachia</i> wMel (<i>Drosophila melanogaster</i>)	Alphaproteobacteria	1267782	1195	<i>Drosophila melanogaster</i>
WOLB	<i>Wolbachia</i> wCle (<i>Cimex lectularius</i>)	Alphaproteobacteria	1250060	1216	<i>Cimex lectularius</i>
WOLB	<i>Wolbachia</i> wPip (<i>Culex quinquefasciatus</i>)	Alphaproteobacteria	1482455	1275	<i>Culex quinquefasciatus</i>
-	<i>Bifidobacterium longum</i> NCC2705	Actinobacteria	2260266	1728	--
-	<i>Coriobacterium glomerans</i>	Actinobacteria	2115681	1768	<i>Pyrrhocoris apterus</i>
-	<i>Isoptericola variabilis</i>	Actinobacteria	3307740	2881	<i>Mastotermes darwiniensis</i>
-	<i>Nocardopsis alba</i> ATCC BAA-2165	Actinobacteria	5848211	5508	<i>Apis mellifera</i>
-	<i>Rickettsia prowazekii</i> Madrid E	Alphaproteobacteria	1111523	843	--
-	<i>Caulobacter henricii</i>	Alphaproteobacteria	3957288	3616	--
-	<i>Azospirillum brasilense</i> sp7	Alphaproteobacteria	6587527	5669	--
-	<i>Mesorhizobium ciceri</i>	Alphaproteobacteria	6690028	6264	--

-	<i>Methylobacterium radiotolerans</i>	Alphaproteobacteria	6899110	6431	--
-	Candidatus <i>Uzinura diaspidicola</i> str. ASNER	Bacteroidetes	263431	227	<i>Aspidiotus nerii</i>
-	Candidatus <i>Azobacteroides pseudotrichonymphae</i> (<i>Coptotermes formosanus</i>)	Bacteroidetes	1114206	852	<i>Coptotermes formosanus</i>
-	<i>Riemerella anatipestifer</i> ATCC 11845 = DSM 15868	Bacteroidetes	2155121	1972	<i>Anas platyrhynchos</i>
-	<i>Flavobacterium psychrophilum</i> FPG3	Bacteroidetes	2715909	2305	--
-	<i>Hymenobacter</i> sp. DG25B	Bacteroidetes	4360029	3153	--
-	<i>Pontibacter korlensis</i> X14-1T	Bacteroidetes	5462537	4116	--
-	<i>Flavobacterium johnsoniae</i> UW101	Bacteroidetes	6096872	5017	--
-	<i>Sphingobacterium</i> sp. 21	Bacteroidetes	6226409	5169	--
-	Candidatus <i>Nasuia deltocephalinicola</i> str. NAS-ALF	Betaproteobacteria	112091	137	<i>Macrosteles quadrilineatus</i>
-	Candidatus <i>Zinderia insecticola</i> CARI	Betaproteobacteria	208564	206	<i>Clastoptera arizonana</i>
-	Candidatus <i>Proftiella armatura</i>	Betaproteobacteria	464857	372	<i>Diaphorina citri</i>
-	<i>Neisseria meningitidis</i> Z2491 (serogroup A)	Betaproteobacteria	2184406	1909	--
-	<i>Snodgrassella alvi</i> wkB2	Betaproteobacteria	2527978	2299	<i>Apis mellifera</i>
-	<i>Rhodoferax ferrireducens</i> T118	Betaproteobacteria	4969784	4418	--
-	<i>Alicyclophilus denitrificans</i> K601	Betaproteobacteria	5070751	4696	--
-	<i>Burkholderia pseudomallei</i> K96243	Betaproteobacteria	7247547	5727	--
-	<i>Burkholderia</i> sp. RPE64	Betaproteobacteria	6964487	6732	<i>Riptortus pedestris</i>
-	<i>Achromobacter xylosoxidans</i> A8	Betaproteobacteria	7359146	6815	--
-	<i>Paraburkholderia xenovorans</i> LB400	Betaproteobacteria	9731138	8702	--
-	<i>Elusimicrobium minutum</i> Pei191	Elusimicrobia	1643562	1529	<i>Pachnoda</i>
-	<i>Sebaldella termitidis</i> ATCC 33386	Fusobacteriia	4486650	4150	<i>Reticulitermes lucifugus</i>
-	Candidatus <i>Evansia muelleri</i>	Gammaproteobacteria	357498	330	<i>Xenophyes cascus</i>
-	Candidatus <i>Moranella endobia</i> PCIT	Gammaproteobacteria	538294	406	<i>Planococcus citri</i>
-	Candidatus <i>Moranella endobia</i> PCVAL	Gammaproteobacteria	538203	411	<i>Planococcus citri</i>
-	Candidatus <i>Hoaglandella endobia</i>	Gammaproteobacteria	636713	517	<i>Trionymus perrisii</i>

-	Candidatus <i>Riesia pediculicola</i> USDA	Gammaproteobacteria	582127	556	<i>Pediculus humanus</i>
-	Candidatus <i>Doolittlea endobia</i>	Gammaproteobacteria	846562	568	<i>Maconellicoccus hirsutus</i>
-	<i>Secondary endosymbiont of Heteropsylla cubana</i>	Gammaproteobacteria	1121596	576	<i>Heteropsylla cubana</i>
-	<i>Wigglesworthia glossinidia brevipalpis</i> (<i>Glossina brevipalpis</i>)	Gammaproteobacteria	697724	611	<i>Glossina brevipalpis</i>
-	Candidatus <i>Tachikawaea gelatinosa</i>	Gammaproteobacteria	708439	614	<i>Urostylis westwoodii</i>
-	<i>Wigglesworthia glossinidia morsitans</i> (<i>Glossina morsitans</i>)	Gammaproteobacteria	719535	618	<i>Glossina morsitans</i>
-	Candidatus <i>Ishikawaella capsulata</i> Mpkobe	Gammaproteobacteria	754729	623	<i>Megacopta punctatissima</i>
-	Candidatus <i>Pantoea carbekii</i> US	Gammaproteobacteria	1196948	801	<i>Halyomorpha halys</i>
-	Candidatus <i>Pantoea carbekii</i>	Gammaproteobacteria	1151074	825	<i>Halyomorpha halys</i>
-	Candidatus <i>Hamiltonella defensa</i> (<i>Acyrtosiphon pisum</i>)	Gammaproteobacteria	2169363	2155	<i>Acyrtosiphon pisum</i>
-	<i>Gilliamella apicola</i>	Gammaproteobacteria	3139412	2803	<i>Apis mellifera</i>
-	<i>Shimwellia blattae</i> DSM 4481 = NBRC 105725	Gammaproteobacteria	4158725	3904	Blattodea
-	<i>Escherichia coli</i> K-12 MG1655	Gammaproteobacteria	4641652	4140	--
-	<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar <i>Typhi</i> CT18	Gammaproteobacteria	5133713	4473	--
-	<i>Pseudomonas aeruginosa</i> PAO1	Gammaproteobacteria	6264404	5572	--
-	<i>Spiroplasma diminutum</i> CUAS-1	Mollicutes	945296	858	<i>Culex annulus</i>
-	<i>Spiroplasma cantharicola</i>	Mollicutes	1179577	1017	<i>Cantharis</i>
-	<i>Sphaerochaeta coccoides</i> DSM 17374	Spirochaetia	2227296	1822	<i>Neotermes castaneus</i>
-	<i>Treponema azotonutricium</i> ZAS-9	Spirochaetia	3855671	3474	<i>Zootermopsis angusticollis</i>
-	<i>Treponema primitia</i> ZAS-2	Spirochaetia	4059867	3522	<i>Zootermopsis angusticollis</i>

Table C4S2 | Results of the Bayes Traits evolutionary tests. In yellow, the positive evidence of a correlation between each metric and the average genome size of the groups is taken into consideration ($BF \geq 2$). Headers with the groups formed by specific genus described in Supplementary Table C4S1.

Network metrics	Data set	BLAT	BLOC	BUCH	CARS	HODG	PORT	SERR	SULC	WOLB	TREM
Clustering1	-1423.86	-61.32	-83.32	-181.20	-101.99	-63.45	-102.16	-187.92	-112.41	-64.07	-53.29
Clustering2	-1425.43	-62.84	-84.26	-181.12	-102.97	-64.88	-103.92	-189.72	-113.37	-62.16	-54.31
logBF	3.13	3.04	1.88	-0.16	1.96	2.86	3.51	3.59	1.92	-3.82	2.047
ConnectedComponents1	-2358.35	-100.65	-135.91	-302.80	-152.40	-89.58	-173.02	-311.88	-184.16	-93.30	-81.79
ConnectedComponents2	-2358.72	-100.69	-135.83	-301.10	-152.24	-89.58	-172.68	-311.34	-184.07	-94.45	-81.80
logBF	0.75	0.08	-0.16	-3.40	-0.31	0.007	-0.69	-1.08	-0.07	2.28	0.019
NetworkDiameter1	-2283.14	-89.25	-126.26	-302.35	-141.41	-86.31	-174.48	-283.16	-181.53	-91.77	-74.68
NetworkDiameter2	-2283.97	-88.97	-127.27	-300.57	-142.89	-86.27	-175.86	-284.14	-181.57	-92.80	-75.71
logBF	1.67	-0.55	2.02	-3.55	2.96	-0.08	2.77	1.97	0.08	2.04	2.07
ShortestPaths1	-3433.86	-144.20	-199.49	-456.24	-222.51	-115.71	-257.85	-465.57	-261.82	-132.61	-115.53
ShortestPaths2	-3433.51	-144.20	-199.50	-454.60	-222.48	-115.66	-258.05	-466.15	-261.83	-132.15	-115.50
logBF	-0.71	-0.005	0.01	-3.30	-0.07	-0.10	0.40	1.17	0.02	-0.93	-0.07
ShortestPaths%1	-2341.05	-96.56	-128.13	-303.29	-149.11	-82.20	-174.54	-290.59	-182.12	-95.07	-79.84
ShortestPaths%2	-2341.32	-96.49	-128.01	-301.00	-149.12	-82.31	-174.37	-290.65	-182.19	-94.88	-79.85
logBF	0.55	-0.14	-0.23	-4.57	0.09	0.22	-0.32	0.11	0.13	-0.38	0.03
CharacteristicPathLength1	-2110.29	-81.06	-117.65	-270.91	-132.13	-81.17	-161.10	-248.40	-169.99	-85.54	-74.08
characteristicPathLength2	-2110.05	-82.02	-117.72	-270.62	-132.44	-81.11	-161.34	-248.56	-170.02	-85.62	-74.14
logBF	-0.49	1.90	0.14	-0.57	0.63	-0.12	0.47	0.33	0.06	0.15	0.13
AvgNumNeighbors1	-1786.32	-76.86	-102.26	-230.03	-119.53	-71.23	-131.99	-223.71	-133.18	-75.52	-66.90
AvgNumNeighbors2	-1786.17	-75.51	-102.34	-229.69	-118.80	-71.48	-130.53	-223.30	-133.12	-75.13	-67.11
logBF	-0.30	-2.71	0.16	-0.67	-1.46	0.50	-2.92	-0.82	-0.11	-0.78	0.42
NumNodes1	-2631.87	-117.27	-157.47	-346.32	-172.09	-97.20	-201.90	-363.12	-208.18	-109.43	-90.67
NumNodes2	-2632.93	-117.23	-158.48	-347.87	-172.08	-98.26	-203.29	-363.97	-208.19	-108.18	-90.53
logBF	2.11	-0.08	2.01	3.11	-0.02	2.14	2.77	1.71	0.02	-2.49	-0.28
Multi-edgeNodePairs1	-2606.86	-116.41	-156.58	-347.23	-171.09	-95.03	-196.14	-362.82	-204.39	-107.25	-89.90
Multi-edgeNodePairs2	-2606.56	-116.42	-156.55	-345.87	-171.13	-94.93	-196.20	-361.54	-204.43	-107.03	-89.93
logBF	-0.60	0.03	-0.05	-2.70	0.08	-0.19	0.12	-2.54	0.086	-0.43	0.07

Table C4S3 | Topological properties of the metabolic networks of the endosymbiotic bacteria of insects, and their free-living relatives of our data set calculated with the software Cytoscape.

Strain	# nodes	# edges	Clustering coefficient	Connected components (CC)	Large st CC	Mean size CC	Shortest paths %	Characteristic path length	Network diameter	Avg. # of neighbors	Isolated nodes	Self loops	Multi-edge node pairs
Candidatus <i>Nasuia deltocephalinicola</i> str. NAS-ALF	67	82	0.037	21	9	3.19	4	2.157	6	1.433	0	0	34
Candidatus <i>Hodgkinia cicadicola</i> TETUND1	84	105	0.032	26	9	3.23	3	2.212	6	1.452	0	0	44
Candidatus <i>Tremblaya phenacola</i> PAVE	91	136	0.027	20	26	4.65	9	4.41	13	1.648	0	0	61
Candidatus <i>Hodgkinia cicadicola</i> TETUND2	94	113	0.02	31	9	3.03	2	2.097	6	1.383	0	0	48
Candidatus <i>Hodgkinia cicadicola</i> TETULN	110	143	0.023	31	12	3.54	3	2.485	8	1.491	0	0	61
Candidatus <i>Hodgkinia cicadicola</i> Dsem	114	156	0.037	30	14	3.8	3	2.56	7	1.548	0	0	65
Candidatus <i>Carsonella ruddii</i> PV (<i>Pachypsylla venusta</i>)	133	199	0.026	28	43	4.75	8	5.429	18	1.729	0	0	84
Candidatus <i>Carsonella ruddii</i> HT (<i>Heteropsylla texana</i>)	136	204	0.026	29	45	4.68	8	5.506	18	1.721	0	0	87
Candidatus <i>Carsonella ruddii</i> PC (<i>Pachypsylla celtidis</i>)	136	203	0.026	29	43	4.68	8	5.409	18	1.721	0	0	86
Candidatus <i>Sulcia muelleri</i> ML	146	242	0.029	24	65	6.08	18	5.983	16	1.863	0	0	106
Candidatus <i>Sulcia muelleri</i> Sulcia-ALF	146	242	0.029	24	65	6.08	18	5.983	16	1.863	0	0	106
Candidatus <i>Carsonella ruddii</i> HC (<i>Heteropsylla cubana</i>)	150	230	0.041	30	47	5	8	5.657	18	1.747	0	0	99
Candidatus <i>Carsonella ruddii</i> CE (<i>Ctenarytaina eucalypti</i>)	156	233	0.034	34	42	4.58	7	5.237	14	1.705	0	0	100
Candidatus <i>Carsonella ruddii</i> CS (<i>Ctenarytaina spatulata</i>)	156	233	0.034	34	42	4.58	7	5.237	14	1.705	0	0	100
Minimal network	159	258	0.064	37	57	4.29	9	4.754	18	1.862	0	0	110
Candidatus <i>Carsonella ruddii</i> DC (<i>Diaphorina citri</i>)	166	261	0.045	31	47	5.35	10	6.246	18	1.771	0	0	114
Candidatus <i>Uzinura</i>	169	236	0.026	40	24	4.25	4	3.667	13	1.598	0	0	101

<i>diaspidicola</i> str. ASNER													
Candidatus <i>Portiera aleyrodidarum</i> BT-B-HRs	178	263	0.029	37	80	4.81	17	8.604	25	1.719	0	0	110
Candidatus <i>Portiera aleyrodidarum</i> MED	178	262	0.029	37	80	4.81	17	8.605	25	1.719	0	0	109
Candidatus <i>Portiera aleyrodidarum</i> BT-QVLC	179	264	0.028	37	83	4.83	18	9.213	28	1.721	0	0	110
Candidatus <i>Portiera aleyrodidarum</i> BT-B	181	266	0.028	38	83	4.76	18	9.21	28	1.713	0	0	111
Candidatus <i>Proffella armatura</i>	186	285	0.04	44	34	4.22	6	3.589	10	1.796	0	0	118
Candidatus <i>Portiera aleyrodidarum</i> TV-BCN	187	281	0.026	37	79	5.05	15	8.05	25	1.743	0	0	118
Candidatus <i>Sulcia muelleri</i> BGSS	188	314	0.023	29	69	6.48	14	5.414	16	1.851	0	0	140
Candidatus <i>Portiera aleyrodidarum</i> TV	189	285	0.026	37	86	5.1	18	9.045	28	1.746	0	0	120
Candidatus <i>Sulcia muelleri</i> DMIN	189	317	0.022	28	85	6.75	19	7.226	23	1.862	0	0	141
Candidatus <i>Sulcia muelleri</i> GWSS	189	317	0.022	28	85	6.75	19	7.226	23	1.862	0	0	141
Candidatus <i>Sulcia muelleri</i> TETUND	193	316	0.022	32	71	6.03	13	6.282	16	1.855	0	0	137
Candidatus <i>Sulcia muelleri</i> SMDSEM	196	326	0.022	31	106	6.32	25	9.053	24	1.878	0	0	142
Candidatus <i>Portiera aleyrodidarum</i> AD-CAI	198	314	0.035	37	100	5.35	23	9.284	27	1.828	0	0	133
Candidatus <i>Portiera aleyrodidarum</i> AF-CAI	198	314	0.035	37	100	5.35	23	9.284	27	1.828	0	0	133
Candidatus <i>Tremblaya princeps</i> PCVAL	200	310	0.025	40	54	5	10	5.806	16	1.79	1	1	130
Candidatus <i>Sulcia muelleri</i> CARI	212	344	0.017	36	71	5.88	11	5.334	16	1.83	0	0	150
Candidatus <i>Sulcia muelleri</i> PSPU	218	366	0.021	37	71	5.89	11	5.222	16	1.89	0	0	160
<i>Buchnera aphidicola</i> BCc	249	449	0.039	36	90	6.91	13	7.083	25	2.032	0	0	196
<i>Buchnera aphidicola</i> (<i>Cinara tujaefilina</i>)	255	461	0.038	33	110	7.72	16	8.288	25	2.047	0	0	200
Candidatus <i>Moranella</i>	256	420	0.036	50	93	5.12	12	7.661	20	1.82812	0	0	186

<i>endobia</i> PCIT										5			
Candidatus <i>Moranella endobia</i> PCVAL	258	422	0.035	51	93	5.05	12	7.659	20	1.822	0	0	187
<i>Spiroplasma diminutum</i> CUAS-1	268	460	0.049	45	73	5.95	9	5.168	15	2.03	1	1	187
Candidatus <i>Evansia muelleri</i>	273	473	0.037	37	140	10	21	9.691	28	1.963	1	1	204
<i>Spiroplasma cantharicola</i>	278	469	0.052	48	110	5.79	12	6.235	19	1.986	1	2	191
Candidatus <i>Doolittlea endobia</i>	290	516	0.054	42	154	7.02	18	10.074	36	2.007	0	0	225
<i>Buchnera aphidicola</i> JF98 (<i>Acyrtosiphon pisum</i>)	345	590	0.022	48	173	7.18	16	10.275	27	1.901	0	0	261
Candidatus <i>Hoaglandella endobia</i>	352	625	0.06	52	220	6.77	27	11.969	33	2.011	0	0	271
Candidatus <i>Baumannia cicadellinicola</i> B-GSS	367	709	0.045	42	271	8.9	41	11.594	32	2.142	1	1	315
Candidatus <i>Riesia pediculicola</i> USDA	372	703	0.041	40	276	9.3	38	12.779	36	2.102	1	1	311
<i>Buchnera aphidicola</i> Bp (<i>Baizongia pistaciae</i>)	384	693	0.041	51	217	7.53	25	10.094	27	2.046875	1	1	299
<i>Buchnera aphidicola</i> LL01 (<i>Acyrtosiphon pisum</i>)	396	733	0.034	45	242	8.8	30	11.094	34	2.061	1	1	324
Secondary endosymbiont of <i>Heteropsylla cubana</i>	396	724	0.042	47	283	8.42	30	11.433	32	2.045	1	1	318
<i>Buchnera aphidicola</i> TLW03 (<i>Acyrtosiphon pisum</i>)	404	733	0.037	52	220	7.77	18	9.014	29	2.03	1	1	322
<i>Buchnera aphidicola</i> Tuc7 (<i>Acyrtosiphon pisum</i>)	404	746	0.034	49	241	8.24	28	10.006	27	2.059	1	1	329
<i>Buchnera aphidicola</i> Sg (<i>Schizaphis graminum</i>)	405	751	0.034	51	236	7.94	30	11.091	32	2.054	1	1	334
<i>Buchnera aphidicola</i> 5A (<i>Acyrtosiphon pisum</i>)	407	762	0.035	48	244	8.48	31	10.111	28	2.093	1	1	335
<i>Buchnera aphidicola</i> JF99 (<i>Acyrtosiphon pisum</i>)	407	756	0.035	49	238	8.31	29	10.631	30	2.074	1	1	333
<i>Buchnera aphidicola</i> Ua (<i>Uroleucon ambrosiae</i>)	407	734	0.035	55	222	7.4	26	9.324	25	2.029	1	1	320

<i>Wolbachia</i> wMel (<i>Drosophila melanogaster</i>)	408	754	0.044	59	245	6.9	27	9.1	22	2.06	1	1	332
<i>Wolbachia</i> wPip (<i>Culex quinquefasciatus</i>)	408	748	0.039	60	239	6.8	26	9.047	22	2.054	1	1	328
<i>Wolbachia</i> wNo (<i>Drosophila simulans</i>)	409	758	0.042	59	248	7.03	26	9.221	24	2.088	1	1	330
<i>Blattabacterium</i> Cpu (<i>Cryptocercus punctulatus</i>)	410	768	0.042	54	250	7.59	29	9.818	26	2.073	1	1	342
<i>Wolbachia</i> wRi (<i>Drosophila simulans</i>)	412	756	0.042	60	228	6.86	27	9.493	24	2.053	1	1	332
<i>Wolbachia</i> wHa (<i>Drosophila simulans</i>)	413	760	0.042	60	248	6.88	27	9.171	22	2.058	1	1	334
<i>Wolbachia</i> wCle (<i>Cimex lectularius</i>)	414	762	0.046	58	249	7.14	26	10.208	33	2.058	1	1	335
<i>Rickettsia prowazekii</i> Madrid E	416	685	0.039	80	170	5.2	8	6.868	29	1.856	1	1	298
<i>Baumannia cicadellinicola</i> Hc (<i>Homalodisca coagulata</i>)	421	838	0.047	41	321	10.26	46	10.03	31	2.2	1	1	374
<i>Buchnera aphidicola</i> APS (<i>Acyrtosiphon pisum</i>)	425	769	0.039	55	252	7.72	30	9.902	28	2.038	1	1	335
<i>Sodalis endosymbiont of Henestaris halophilus</i>	425	807	0.041	45	319	9.44	40	12.569	36	2.108	1	1	358
<i>Serratia symbiotica</i> <i>Cinara cedri</i>	431	809	0.045	55	288	7.83	35	10.501	26	2.079	1	1	360
<i>Buchnera aphidicola</i> Ak (<i>Acyrtosiphon kondoi</i>)	432	793	0.038	56	239	7.71	27	9.451	25	2.056	1	1	348
<i>Buchnera aphidicola</i> (<i>Aphis glycines</i>) BAg	436	801	0.037	54	260	8.07	30	11.16	32	2.05	1	1	353
<i>Buchnera aphidicola</i> W106 (<i>Myzus persicae</i>)	438	809	0.037	55	262	7.96	31	10.339	27	2.064	1	1	356
<i>Buchnera aphidicola</i> F009 (<i>Myzus persicae</i>)	439	814	0.037	54	265	8.13	31	10.348	28	2.077	1	1	357
<i>Buchnera aphidicola</i> G002 (<i>Myzus persicae</i>)	439	814	0.037	54	265	8.13	31	10.348	28	2.077	1	1	357
<i>Buchnera aphidicola</i> USDA (<i>Myzus persicae</i>)	439	814	0.037	54	265	8.13	31	10.348	28	2.077	1	1	357
<i>Blattabacterium</i> sp. (<i>Blatta orientalis</i>)	450	855	0.039	49	327	9.18	40	9.925	26	2.116	1	1	378

<i>Blattabacterium</i> BPLAN (<i>Periplaneta</i> <i>americana</i>)	451	859	0.039	48	328	9.4	40	9.935	26	2.12	1	1	380
<i>Blattabacterium</i> sp. (<i>Mastotermes</i> <i>darwiniensis</i>)	452	867	0.045	50	249	8.03	40	9.706	26	2.128	1	1	385
<i>Blattabacterium</i> sp. (<i>Nauphoeta cinerea</i>)	452	867	0.045	50	328	9.04	40	9.706	26	2.128	1	1	385
<i>Blattabacterium</i> sp. (<i>Panesthia</i> <i>angustipennis</i> <i>spadica</i>) BPAA	453	862	0.046	52	327	8.71	39	9.669	26	2.115	1	1	382
<i>Wigglesworthia</i> <i>glossinidia brevipalpis</i> (<i>Glossina brevipalpis</i>)	454	975	0.046	45	335	10.0 8	41	10.892	28	2.132	1	1	390
<i>Blattabacterium</i> sp. (<i>Blaberus giganteus</i>)	456	875	0.044	50	331	9.12	40	9.729	26	2.132	1	1	388
<i>Blochmannia</i> endosymbiont of <i>Camponotus</i> (<i>Colobopsis</i>) <i>obliquus</i> 757	459	875	0.049	49	340	9.36	44	11.42	34	2.131	1	1	385
Candidatus <i>Baumannia</i> <i>cicadellinicola</i> BGSS	461	900	0.035	46	350	10.0 2	45	10.989	34	2.148	1	1	404
<i>Wigglesworthia</i> <i>glossinidia morsitans</i> (<i>Glossina morsitans</i>)	462	896	0.045	44	349	10.5	41	10.587	29	2.143	1	1	400
<i>Blattabacterium</i> Bge (<i>Blattella germanica</i>)	465	890	0.043	50	335	9.3	39	9.753	26	2.129	1	1	394
Candidatus <i>Blochmannia</i> <i>floridanus</i> (<i>Camponotus</i> <i>floridanus</i>)	470	901	0.047	53	328	8.81	38	9.829	29	2.136	1	1	398
Candidatus <i>Blochmannia vafer</i> (<i>Camponotus vafer</i>)	470	899	0.051	51	324	9.22	38	9.929	29	2.128	1	1	398
<i>Blochmannia</i> endosymbiont of <i>Polyrhachis</i> (<i>Hedomyrma</i>) <i>turneri</i> 675	472	913	0.053	52	348	9.06	38	9.544	29	2.157	1	1	403
Candidatus <i>Ishikawaella</i> <i>capsulata</i> Mpkobe	479	923	0.051	61	310	7.85	36	8.474	27	2.146	1	1	408
Candidatus <i>Tachikawaea</i> <i>gelatinosa</i>	481	933	0.056	49	339	9.82	41	9.594	28	2.183	1	1	407
Candidatus <i>Blochmannia</i> <i>chromaiodes</i> (<i>Camponotus</i> <i>chromaiodes</i>)	483	942	0.048	48	368	10.0 6	47	10.921	34	2.174	1	1	416

<i>Candidatus Blochmannia pennsylvanicus</i> (<i>Camponotus pennsylvanicus</i>)	483	942	0.048	48	368	10.06	47	10.921	34	2.174	1	1	416
<i>Candidatus Azobacteroides pseudotriconymphae</i> (<i>Coptotermes formosanus</i>)	509	949	0.5	57	373	8.93	37	10.104	26	2.051	1	1	426
<i>Caulobacter henricii</i>	511	938	0.035	75	333	6.82	33	9.219	26	2.055	1	1	412
<i>Coriobacterium glomerans</i>	515	910	0.042	88	300	5.85	26	9.314	26	1.996	1	1	395
<i>Candidatus Hamiltonella defensa</i> (<i>Acyrtosiphon pisum</i>)	538	1037	0.042	53	394	11.41	42	11.48	37	2.16	1	1	455
<i>Candidatus Pantoea carbekii</i>	553	1092	0.048	54	420	10.24	45	9.406	28	2.21	1	1	480
<i>Candidatus Pantoea carbekii</i> US	558	1108	0.048	54	425	10.33	45	9.333	28	2.219	1	1	488
<i>Bifidobacterium longum</i> NCC2705	599	1131	0.045	85	406	7.04	37	8.85	23	2.117	1	1	496
<i>Sphaerochaeta coccoides</i> DSM 17374	628	1176	0.039	91	381	6.9	31	8.55	24	2.121	1	1	509
<i>Riemerella anatipestifer</i> ATCC 11845 = DSM 15868	632	1226	0.055	80	431	7.9	35	9.113	31	2.158	1	1	543
<i>Treponema azotonutricium</i> ZAS-9	687	1315	0.038	85	477	8.08	34	8.421	25	2.143	1	1	578
<i>Candidatus Sodalis pierantonius</i>	689	1342	0.048	72	489	9.57	41	10.135	32	2.2	1	1	583
<i>Flavobacterium psychrophilum</i> FPG3	697	1343	0.046	81	478	8.6	36	8.862	30	2.158	1	1	590
<i>Snodgrassella alvi</i> wKB2	697	1354	0.041	79	509	8.82	45	9.921	29	2.155	1	1	602
<i>Neisseria meningitidis</i> Z2491 (serogroup A)	718	1379	0.045	85	524	8.54	44	9.318	29	2.164	1	1	601
<i>Sodalis glossinidius</i> (<i>Glossina</i> spp.)	741	1494	0.053	72	579	10.29	51	9.654	30	2.262	1	1	655
<i>Treponema primitia</i> ZAS-2	742	1340	0.036	104	465	7.13	29	8.664	26	2.038	1	1	583
<i>Escherichia coli</i> K-12 MG1655	791	1506	0.045	96	552	8.24	36	9.093	31	2.159	1	1	651
<i>Seibaldella termitidis</i> ATCC 33386	794	1549	0.042	94	563	8.47	39	9.214	26	2.184	1	1	681
<i>Gilliamella apicola</i>	806	1619	0.047	79	638	10.2	49	9.716	30	2.27	1	1	703

<i>Hymenobacter</i> sp. DG25B	854	1598	0.04	118	555	7.24	35	9.697	34	2.115	1	1	694
<i>Isopterocola variabilis</i>	888	1681	0.046	127	585	6.99	35	8.871	25	2.131	1	1	734
<i>Sphingobacterium</i> sp. 21	921	1745	0.05	122	624	7.55	38	9.336	30	2.122	1	2	766
<i>Pontibacter korlensis</i> X14-1T	951	1800	0.042	134	630	7.1	36	9.57	34	2.122	1	1	790
<i>Flavobacterium johnsoniae</i> UW101	996	1870	0.037	136	645	7.32	32	8.995	31	2.122	1	2	811
<i>Nocardioopsis alba</i> ATCC BAA-2165	1037	1962	0.035	142	695	7.3	36	9.752	38	2.106	1	2	868
<i>Alicyclophilus denitrificans</i> K601	1038	1989	0.035	134	691	7.75	36	8.753	25	2.16	1	1	867
Candidatus <i>Zinderia insecticola</i> CAR1	1048	1887	0.04	164	601	6.39	26	8.899	26	2.042	1	1	816
<i>Sodalis praecaptivus</i>	1073	2134	0.049	125	789	8.6	44	9.395	29	2.252	1	2	924
<i>Rhodoferax ferrireducens</i> T118	1075	2075	0.038	138	745	7.79	39	9.159	31	2.182	1	1	901
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar <i>Typhi</i> CT18	1078	2228	0.05	102	841	10.57	49	9.355	30	2.323	1	1	975
<i>Serratia</i> sp. SCBI	1095	2251	0.049	101	860	10.84	50	9.24	28	2.321	1	2	978
<i>Azospirillum brasilense</i> sp7	1111	2155	0.04	147	732	7.56	36	8.818	27	2.169	1	1	949
<i>Elusimicrobium minutum</i> Pei191	1118	2336	0.049	101	888	11.07	51	9.238	30	2.351	1	2	1020
<i>Serratia marcescens</i> subsp. <i>marcescens</i> Db11	1131	2285	0.047	119	845	9.5	46	9.07	27	2.285	1	2	991
<i>Serratia</i> sp. FS14	1134	2312	0.051	113	859	10.03	47	9.145	29	2.303	1	2	1004
<i>Shimwellia blattae</i> DSM 4481 = NBRC 105725	1141	2309	0.048	126	827	9.05	43	9.259	30	2.24	1	1	1030
<i>Methylobacterium radiotolerans</i>	1146	2192	0.043	152	734	7.53	34	9.874	39	2.131	1	1	970
<i>Serratia marcescens</i> SM39	1153	2338	0.049	121	872	9.53	47	9.174	28	2.291	1	2	1015
<i>Serratia marcescens</i> WW4	1153	2352	0.05	113	865	10.2	46	9.215	29	2.3	1	2	1024
<i>Serratia liquefaciens</i> ATCC 27592	1156	2364	0.047	112	894	10.32	49	9.334	29	2.315	1	2	1024
<i>Serratia plymuthica</i> 4Rx13	1162	2381	0.047	113	907	10.28	50	9.46	29	2.325	1	2	1028

<i>Serratia plymuthica</i> S13	1162	2398	0.048	111	908	10.47	50	9.394	29	2.339	1	2	1037
<i>Serratia proteamaculans</i> 568	1163	23850	0.045	114	910	10.2	50	9.395	29	2.325	1	2	1031
<i>Serratia plymuthica</i> AS9	1165	2376	0.046	112	902	10.44	48	9.461	29	2.314	1	2	1026
<i>Serratia</i> sp. AS12	1165	2376	0.046	112	902	10.4	48	9.461	29	2.314	1	2	1026
<i>Pseudomonas aeruginosa</i> PAO1	1167	2336	0.047	118	845	9.89	43	9.557	32	2.264	1	1	1014
<i>Serratia fonticola</i> DSM 4576	1174	2406	0.053	112	911	10.48	49	9.517	29	2.319	1	2	1043
<i>Serratia marcescens</i> FGI94	1180	2373	0.042	118	882	10	46	9.512	28	2.268	1	1	1034
<i>Achromobacter xylosoxidans</i> A8	1195	2285	0.03	145	806	8.24	36	8.861	29	2.151	1	1	999
<i>Burkholderia</i> sp. RPE64	1229	2402	0.037	146	812	8.41	35	8.713	26	2.181	1	1	1061
<i>Mesorhizobium ciceri</i>	1230	2413	0.038	153	824	8.04	34	8.985	29	2.195	1	1	1062
<i>Burkholderia pseudomallei</i> K96243	1270	2476	0.042	145	897	8.76	40	9.556	32	2.209	1	2	1071
<i>Paraburkholderia xenovorans</i> LB400	1376	2701	0.038	159	938	8.94	38	9.37	32	2.18	1	2	1199

Appendix C - List of figures and tables in this dissertation

Figures

Introduction

- **Figure 1.** Overview of stable symbioses throughout the Tree of Life (Moya et al., 2008). *Reprinted with permission from Springer Nature and the Copyright Clearance Center with license number 5433131267927.*
- **Figure 2.** A) The genome reduction process of obligate intracellular endosymbionts, and B) Genome size versus GC content of endosymbiotic bacterial genomes from McCutcheon & Moran (2012). *Reprinted with permission from Springer Nature and the Copyright Clearance Center with license number 5433140857649.*
- **Figure 3.** Growth of GenBank sequences (Benson et al., 2017) and NCBI (Wheeler et al., 2007) web users through 2019. *Figure from the Department of Health and Human Services, NIH. Website https://www.nlm.nih.gov/about/2021CJ_NLM.pdf.*

Chapter 1

- **Fig. C1.1 | Genomic GC content (%) versus gene number in several symbionelles, endosymbionts, and free-living bacteria.** The two dashed purple vertical lines delimit the minimal gene set between 223 and 244 genes. The dotted red lines indicate pairs of symbiotic associations. Symbionelles: (1) '*Ca. Tremblaya princeps*' PCVAL, (2) '*Ca. Hodgkinia cicadicola*' DSEM, (3) '*Ca. Carsonella ruddii*' PV, (4) '*Ca. Zinderia insecticola*' CARI. Endosymbionts: (5) '*Ca. Sulcia muelleri*' GWSS, (6) '*Ca. Uzinura diaspidicola*' str. ASNER, (7) '*Ca. Portiera aleyrodidarum*' BT-QVLC, (8) *B. aphidicola* BCc, (9) '*Ca. Moranella*

endobia' PCIT, (10) '*Ca. Baumannia cicadellinicola*' str. Hc, (11) *S. symbiotica* str. Cc and (12) '*Ca. Hamiltonella defensa*' 5AT.

- **Figure C1.2 | Streamlining and Muller's ratchet hypotheses** are commonly used to explain genome reduction in free-living and host-associated bacteria respectively. Alternative hypotheses include "selection for mutator strains" and "loss of DNA repair genes".
- **Figure C1.3 | Population genetic models.** (A) The fate of an allele is determined by natural selection if the product of the effective population size (N_e) and the coefficient of selection (s) is larger than one, and by genetic drift otherwise; (B) however, when mutation is taken into account, the equilibrium frequency (f_{eq}) of the fittest allele is 0 when the mutation rate (μ) is larger than the selection coefficient (s).
- **Figure C1.4 | The Black Queen Hypothesis** predicts that if there is a community where different species produce an expensive and diffusible public good (PG), the system will evolve toward a scenario where only a few of their members will continue with the production of the PG, only if the benefit of losing the production of the function outweighs the cost of losing it. As such, community producers of the PG become helpers (H) and the rest become beneficiaries (B). LOF_B , the benefit of losing the function producing the PG; LOF_C , the cost of losing the function producing the PG.
- **Figure C1.5 | The reassignment of the stop codon (UGA_{stop}) to code for tryptophan (UGA_{Trp}) is explained using the "capture hypothesis."** However, the large G+C content in "*Ca. Hodgkinia cicadicola*" makes the capture hypothesis unlikely in this organism. Instead, it is hypothesized that the loss of translational release factor RF2 triggered the evolution of this rearrangement.

- **Figure C1.6 | Convergent evolution of Trp biosynthesis.** The biosynthesis of Trp is performed cooperatively by two different endosymbionts in the aphid *Cinara cedri* (A) and in the psyllid *Heteropsylla cubana* (B). Strikingly, the same division of labor evolved in both systems.

Chapter 2

- **Figure C2.1 | Schematic representation of the data collection workflow.** The graphic illustrates the major modules of the data collection and the action taken in each step. The pipeline is highly tunable, and every update will be easier and shorter since this first manual curation has been so thorough. The triangle shape is meant to convey that each step acts as a filter of data we are not interested in until we end up with the highly curated catalog that SymbioGenomesDB currently offers.
- **Figure C2.2 | Database overview.** In HOME (a), there is a complete overview of the importance of the database and its purpose in detail. The line of buttons in the above green menu, as well as the menu on the right, denote the different parts of the web interface, with special importance to the button (b) “Enter Database”, which will open the database in Shiny from the R Studio. An explanation of the functions (c) of the database is also included, as well as a quick tour explained in detail throughout this article and the Acknowledgments (d) for the support of this work.
- **Figure C2.3 | Example of the Find Organisms tab.** In our lab at the University of Valencia, we work with insect symbiosis (Martínez-Cano et al., 2014; Moya et al., 2008; Reyes-Prieto et al., 2014). This example shows the result of searching for the symbionts of ‘insecta’ with the default ‘all ranks’ host level, and the ‘species’

level selected for symbionts. Be aware that the matching result indicates that the host level of the query is class since it is where a match was found. The results include an abundance graph including the first most representative 14 matches, plus a summary of the rest of the matches, a table of organisms matching the user's query, and the number of matches for each specie, phylum, class, or whichever taxonomy level explored. The rest of the resulting table is at the right of the figure for space limitations.

- **Figure C2.4 | Example of the Find Genomes tab.** The result of searching for the genomes associated with 'insects' with the default 'all ranks' taxonomic level. First, users get a scrolling menu from which they can select genomes of interest, or all genomes available in this search, which in turn displays a table that shows the names of the organisms selected in the scrolling menu, the precise host with which the symbiotic relationship exists, as well as metrics and characteristics of their genomes. The rest of the resulting table is at the right of the figure for space limitations.
- **Figure C2.5 | Example of a FIND GENES search.** (a) Searching for all the genes included in the tryptophan biosynthesis in the genomes related to insects. We get two scrolling menus and selected all the tryptophan genes and the genomes of several species of the *Buchnera* genus, as well as the species *Serratia symbiotica*. (b) The resulting table lists the orthologs found between the genomes we selected, including the bacteria working as cosymbionts in the aphid *Cinara cedri*, that participate in an exceptional metabolic complementation of the tryptophan metabolic pathway (Gosalbes et al., 2008). (c) Even though the table shows the abbreviated genome names from KEGG (Kanehisa et al., 2014; Kanehisa & Goto, 2000), if you scroll over the name, you get the complete species name in a little

box below the pointer. The resulting table will be a table including the genes and the genomes selected in the menus. Every output is available for download, as flat files for further and easy parsing and analysis.

- **Figure C2.6 | SymGenDB's Organisms module.** For this example, we searched for the symbionts of insects included in SymGenDB. It is worth noting that we used the term 'insects' and the output shows the scientific name 'Hexapoda' because the database includes a hefty list of synonyms so users do not have to know/search only for scientific names. We searched for both the host and the symbiont's phylogenetic level as default, and the result shows an abundance chart (only the first 15 most abundant hits are displayed in color, the rest are encompassed in grey) and a list of organisms showing all the hits. Each resulting 'organism' (or family, genus, class, etc.) is a link to its NCBI taxonomy page. The resulting list shown in this figure is cut short for spacing purposes.
- **Figure C2.7 | SymGenDB's Genomes module.** Continuing with the example on symbionts of insects, we searched for those genomes and selected 3 of the 169 available in SymGenDB. The resulting table (viewed horizontally for spacing purposes, although it is presented vertically in the database), consists of all the metadata of the genomes of our choice, with a link to the host's taxonomy id from NCBI. It is important to show that one of the downloads available in this module is the orthology table of the chosen genomes (shown in red) which is very helpful for evolutionary research. Furthermore, the literature where this genome was first described is also made available to users (shown in orange).
- **Figure C2.8 | SymGen's gene module.** For this example, we searched for all the genes containing the letters 'trp' in the genomes of symbionts of insects. We selected all of the 169 available genomes. The resulting table is a list in a

presence/absence format, where all the present genes are shown with their KEGG id and a link to their KEGG gene web page. The 'Gene ID' and the 'Gene Description' features are both linked to their KEGG orthology web page.

- **Figure C2.9 | SymGenDB's new MetaDAGs module.** In this example, we search for the symbionts of the genus '*Buchnera*'. The output is a list of organisms as bacterial strains, as well as the joint (pan) or intersecting (core) metabolism of the strains that resulted in the search, included in the taxonomic level 'genus' (in bold). It is important to denote that in the case of the 'pan' and 'core' interacting metabolism, not only the genomes of the bacterial strains resulting from the search are presented. The complete set of strains of the same genus available in SymGenDB constitutes these MetaDAGs.
- **Figure C2.10 | SymGenDB's new MetaDAGs module's output.** In this example, we search for the symbionts of the genus '*Buchnera*'. The output is a preview of a graphical display of the MetaDAG you get by choosing *Buchnera aphidicola* from *Cinara tujafilina*. This dynamic display can be viewed in another window, downloaded as a PDF and all the information of the reaction(s) included in each node(s) is available by clicking on the node(s) of interest.
- **Figure C2.11 | SymGenDB's new MetaDAGs module's output** where users can not only see the reactions included in their graph of interest in detail but also contextualize the reaction within the KEGG metabolic map. We have included a feature to highlight the reaction(s) of interest for this purpose.
- **Figure C2.12 | Overview of the users of SymGenDB** from June 2018 to February 2023.
- **Figure C2.13 | Demographic overview of SymGenDB users** from Jan 2022 to March 2023.

Chapter 3

- **Figure C3.1 | Interaction graph of the proposed theoretical minimal metabolic network adapted from** (Gabaldon et al., 2007). Line colors denote metabolic categories: yellow, glycolysis; orange, pentose phosphate pathway; pink, phospholipid metabolism; green, nucleotide metabolism; blue, coenzyme metabolism. The two glycolytic steps in which ATP is produced by substrate-level phosphorylation are depicted with thicker red arrows and correspond to reactions R01512 and R00200 in Table C3.1. The reaction graph of this same network is presented in Figure C3.2 for comparison.
- **Figure C3.2 | The reaction graph of the proposed theoretical minimal metabolic network represented in Figure C3.1**, obtained using data from the KEGG database. The yellow-filled circles are the reactions with their KEGG ID and E.C. numbers, and the purple-filled circles are the reverse reaction of the yellow-filled circles, when appropriate. Line colors denote metabolic categories. A full-size representation can be seen in Supplementary Figure C3S1.
- **Figure C3.3 | m-DAG of the metabolism of a theoretical minimal bacterial cell.** Single reactions appear in yellow, contracted MBBs in grey, and the essential reactions as hexagons with double lines. Line colors denote metabolic categories. MBB 0.79.0 is zoomed in as an example of how a strongly connected component, which is a cyclic subgraph formed by 7 reactions and 7 compounds, is reduced to one node in our m-DAG.
- **Figure C3.4 | m-DAG of the metabolism of JCVI-syn3.0.** Single reactions appear in yellow, contracted MBBs in grey, and the essential reactions as hexagons with double lines.

Chapter 4

- **Figure C4.1 | Phylogenetic tree of the endosymbiotic bacteria of insects and free-living outgroups.** Each group created for the evolutionary traits analyses is highlighted in coral with its corresponding group name (complete list in Supplementary Table C4S1). Taxonomical classes are depicted by the colors of the tree branches: blue, Gammaproteobacteria; orange, Alphaproteobacteria; dark green, Betaproteobacteria; red, Actinobacteria; light blue, Fusobacteria; light green, Favlobacteriia; purple, Mollicutes; pink, Spirochaetia.
- **Figure C4.2 | MNs sizes per organism by lifestyle, against their genome size;** above the metabolite-based model, and below the reaction-based model.
- **Figure C4.3 | Distance trees resulting from the reaction- and the metabolite-based methods.** To the left, the distances resulting from the reaction graphs used to create the m-DAGs (see methods). To the right, a hierarchical clustering of the metrics derived from the topological analysis of each MN. The colored legend shows the genera of endosymbiotic bacteria of insects and free-living organisms.
- **Figure C4.4 | Core MN,** consisting of the reactions needed for the synthesis of DNA and RNA, of 101 genomes of endosymbiotic bacteria of insects and its corresponding m-DAG. To the left, is the reaction graph with 12 nodes. The IDs inside the yellow circles are KEGG IDs. The purple circles are their reverse reactions. To the right, the m-DAG (the simplified version of the reaction graph) of the same core, with only 6 nodes (the contraction is shown with the violet background). The gray nodes denote the metabolic building blocks; their IDs, obtained during the calculation, are described in a text output format.

- **Figure C4.5 | Relation between the smallest m-DAGs under study and the minimal MN.** The gray circles are MBBs that have two or more reactions, the yellow circles are single reactions, and the numbers inside the circles are ids created by the metaDAG software with information of the reactions included in each MBB. The core of all m-DAGs is denoted with a gray background.
- **Figure C4.6 | Shared metabolism as MBBs among the m-DAGs of *Nasuia NAS-ALF*, *Tremblaya princeps PCVAL*, *Hodgkinia* (considering the pan-mDAG of strains TETUND1, TETUND2, TETULN, and Dsem), and the minimal MN.** The blue dots represent the shared MBBs between the corresponding groups.

Tables

General Materials and Methods

- **Table 1.** List of the bioinformatic software that has been employed for this dissertation and their applications.

Also available in Spanish in part IX of this thesis.

Chapter 1

- **Table C1.1 | Mutualistic prokaryotes with reduced genomes.** The information on genes and gene size was obtained from Genbank. The class is shown in taxonomy unless specified. “ns”: values unknown from the not-sequenced organism. “*”: Values include the gene content and genomic size of strain-specific plasmids. “¶”: Values as reported in the original article (Rosas-Pérez et al., 2014). “†”: Value as reported in the original article (Boscaro et al., 2013).

Chapter 2

- **Table C2.1** | Update on the data content of SymGenDB.

Chapter 3

- **Table C3.1** | **Reactions, enzymes, and compounds of the minimal metabolic network are presented in Figure C3.2.** Reversible reactions are denoted by the superscript r. MBB IDs are the identification numbers of the metabolic building blocks to which each reaction is contracted into, according to the MetaDAG analysis (Figure C3.3).
- **Table C3.2** | **Essential reactions** of the m-DAG constructed from the theoretical minimal gene set machinery needed for life.
- **Table C3.3** | **Essential reactions** of the m-DAG of "*Candidatus* Nasuia deltocephalinicola" str. NAS-ALF.
- **Table C3.4** | **Essential reactions** of the m-DAG of JCVI-syn3.0.
- **Table C3.5** | **Comparison of the MBBs of the three networks under study.** Every row lists the reactions belonging to the corresponding MBB and the enzymes involved in those reactions. The list includes only MBBs composed of at least three reactions (reverse included) or with fewer reactions but that are shared by at least two of the networks under study.

Chapter 4

- **Table C4.1** | **Sizes and parameters of the endosymbiotic bacteria of insects with metabolic networks smaller than the theoretical minimal MN.** There are four genomes with fewer nodes in the reaction-based model than in the minimal

network, whereas two more are added when looking at the metabolite-based model (marked with an asterisk).

Supplementary Material

Chapter 3, available at <https://www.mdpi.com/2079-7737/10/1/5/s1>

- **Figure C3S1 | Full-size representation of the reaction graph of the proposed theoretical minimal metabolic network** represented in Figure C3.2.
- **Figure C3S2 | The m-DAG of “*Ca. Nasuia deltocephalinicola*” str. NAS-ALF.**
- **Table C3S1 | List of enzymes and reactions** modified from Gabaldon et. al. (2007). n.i.: non identified.
- **Table C3S2 | Reactions, and compounds that make up *Ca. Nasuia deltocephalinicola*’s m-DAG.** Reversible reactions are denoted by the superscript *r*.
- **Table C3S3 | Reactions** included in the reconstruction of the JCVI-syn3.0 reaction graph and the minimal organism constructed for this work and the pathways in which each reaction (can) participates.
- **Table C3S4 | Names of the enzymes and definition of each reaction** involved in the comparison of the MBBs of the three networks under study.

Chapter 4

- **Figure C4S1 | The reaction graph of the pan-metabolic network of endosymbiotic bacteria of insects** consisting of 2,964 reactions (nodes), where 598 are reversible reactions, and 1,883 compounds (edges). This PDF file is very heavy because users can zoom into any of the reactions and compounds to explore if determined metabolic pathways are carried out by any or some endosymbiotic bacteria of insects.
- **Figure C4S2 | The pan-mDAG of endosymbiotic bacteria of insects** consisting of 1,081 MBBs, where 1034 MBBs have only one reaction and 47 have more than

one. This PDF file is very heavy because users can zoom into any of the reactions and MBBs for further details.

- **Figure C4S3 | M-DAG of *Nocardiosis alba* ATCC BAA-2165**, to compare with supplementary figure C4S3. This m- DAG has a total of 533 MBBs including 1,219 reactions and 954 compounds.
- **Figure C4S4 | M-DAG of *Bifidobacterium longum* NCC2705**, to compare with supplementary figure C4S2. This m- DAG has a total of 290 MBBs including 743 reactions and 552 compounds.
- **Table C4S1 | Strains that are included in our data set.** The first column depicts the group to which each strain has been assigned, and a dash indicates the strain does not belong to any group. Lastly, two dashes in the host column mean the strain is not a symbiont of insects.
- **Table C4S2 | Results of the Bayes Traits evolutionary tests.** In yellow, the positive evidence of a correlation between each metric and the average genome size of the groups is taken into consideration ($BF \geq 2$). Headers with the groups formed by specific genus described in Supplementary Table C4S1.
- **Table C4S3 | Topological properties of the metabolic networks** of the endosymbiotic bacteria of insects, and their free-living relatives of our data set calculated with the software Cytoscape.



VNIVERSITAT
DE VALÈNCIA