

# UNIVERSITÀ DEGLI STUDI DI MILANO

Dipartimento di Scienze veterinarie e sanità pubblica (DIVET)  
VET/06

Scuola di dottorato in Terra, Ambiente e Biodiversità  
Curriculum Biologia Animale



## **Evolution of Wolbachia symbiosis in arthropods and nematodes: insights from phylogenetics and comparative genomics**

Francesco COMANDATORE

Tutor: Prof. Claudio BANDI

Coordinatore Dottorato: Prof. Claudio BANDI

A.A. 2013 - 2014

## Abstract

*Wolbachia* is a bacterium observed in relationship with a wide array of arthropod and nematode species. This is an obligate intracellular symbiont, maternally transferred through the host oocytes. In arthropods *Wolbachia* is able to manipulate reproduction, using multiple strategies to increase the fitness of infected females. In nematodes the bacterium has a fundamental, and not completely understood, role in larvae development. *Wolbachia* infects ~50% of all the arthropod species worldwide, and in some of them it can be considered the most important sex determination factor. In contrast, *Wolbachia* presence is much more limited in nematodes, being present in a limited number of filarial species. The taxonomic status within the *Wolbachia* genus is highly debated, with the current classification dividing all strains in 14 'supergroups'.

During my Ph.D. I studied the evolution of the symbiotic relationship between *Wolbachia* and its arthropod and nematode hosts, using genomic approaches.

Indeed, during the evolution of the *Wolbachia*-host relationship, genetic signs have been left in the *Wolbachia* genomes. I worked to identify these genomic signs and to evaluate them within an evolutionary frame, in order to obtain a better understanding of how the *Wolbachia*-host symbiosis evolved.

The work here presented can be organized in three major sections: i) the sequencing and analysis of the genome of the filarial nematode *Dirofilaria immitis* and of its symbiotic *Wolbachia* strain, wDi; ii) the sequencing of the genome of *Wolbachia* endosymbiont of *Litomosoides sigmodontis*, and the phylogenomic reconstruction of the *Wolbachia* supergroups A-D; iii) a comparison of the genomes of 26 *Wolbachia* strains spanning the A to F supergroups.

Here a schematic summary of the results is reported:

1. *Dirofilaria immitis* and the *Wolbachia* symbiont wDi show metabolic complementarity for fundamental pathways
2. The metabolic pathway for the synthesis of wDi membrane proteins is one evolving the fastest in the genome of the bacterium
3. Nematode *Wolbachia* belonging to supergroups C and D are monophyletic, indicating that a single transition to mutualism likely occurred during the evolution of *Wolbachia*
4. *Wolbachia* strains of the C supergroup show genomic features that are unique in the

genus, such as a much higher level of synteny compared to the rest of *Wolbachia* supergroups, and a newly generated pattern of GC skew curves, typically observed in free-living bacteria genomes

5. *Wolbachia* supergroups show conserved genomic features, which suggest genomic isolation among them

# Index

1. Introduction	1
1.1. <i>Wolbachia</i> hosts	1
1.2. <i>Wolbachia</i> spreading	2
1.3. Evolutionary traits of the <i>Wolbachia</i> -host symbiotic relationships	2
1.3.1. <i>Wolbachia</i> in arthropods	3
1.3.2. <i>Wolbachia</i> in nematodes	7
1.4. The <i>Wolbachia</i> lineages	9
Bibliography	12
2. The genome of the heartworm, <i>Dirofilaria immitis</i> , reveals drug and vaccine targets	14
3. Phylogenomics and analysis of shared genes suggest a single transition to mutualism in <i>Wolbachia</i> of nematodes	27
4. Supergroup C <i>Wolbachia</i> , mutualist symbionts of filarial nematodes, have a distinct genome structure	35
5. Summary of results and conclusions	75
Appendix	78

# 1. Introduction

In 1924, at the Harvard University, Marshall Hertig and Burth Wolbach were studying *Culex pipiens* mosquitoes and observed, for the first time, a “*rickettsia*-like bacterium” within their ovaries. In 1936 Hertig formally described that bacterium, and decided to label it “*Wolbachia pipientis*”, in honor of his collaborator (Hertig 1936).

*Wolbachia pipientis* is an alpha-proteobacterium belonging to the Anaplasmataceae family, within the order Rickettsiales. Species classified within this order are all characterized by intracellular lifestyles, however they present an array of different symbiotic relationships with the hosts. The Anaplasmataceae family contains, in addition to *Wolbachia*, the *Ehrlichia*, *Anaplasma*, *Aegyptianella*, *Neorickettsia*, *Candidatus Neoehrlichia* (Pruneau et al. 2014) and the newly described *Arcanobacter* (Martijn et al. 2015) genera. Species responsible for major human and animal infection diseases belong to this family, such as the pathogenic agent of the anaplasmosis in cattle and sheep, *Anaplasma phagocytophilum*, and the human monocytotropic ehrlichiosis agent, *Ehrlichia chaffeensis*. *Wolbachia* however is unable to infect vertebrates, and lives in symbiosis with an array of invertebrate hosts.

Currently, *Wolbachia pipientis* is the sole species described within the genus, for this reason, in this manuscript, I will refer to the entire diversity of this genus as “*Wolbachia*”.

## 1.1. *Wolbachia* hosts

Reconstructing the microbiota composition of an adequate number of insect species, we will find that more than 60% of them result to be infected by bacteria belonging to the *Wolbachia* genus (de Oliveira et al. 2015). Widening the sampling, including also non-insect arthropods, we will discover that *Wolbachia* bacteria can live in association with a wide array of arthropod lineages (Weinert et al. 2015), including isopods, spiders, mites, springtails and termites (Werren et al. 2008).

*Wolbachia* had been considered exclusively an endosymbiont of arthropods, until Sironi and collaborators (in 1995) reported the first evidence that individuals of *Dirofilaria immitis*,

a filarial nematode, were infected by bacteria belonging to the *Wolbachia* genus. The identified *Wolbachia* strain was labeled *Wolbachia* endosymbiont of *Dirofilaria immitis*, or wDi (Sironi et al. 1995).

On the basis of the current estimation, we know that *Wolbachia* infects ~ 50% of all the arthropod species (Weinert et al. 2015), but this data cannot be considered as definitive. On the other hand *Wolbachia* has a limited diffusion among nematodes: it has been described in symbiosis with a limited number of filarial nematodes and with the plant-parasitic nematode *Radopholous similis* (Sironi et al. 1995; Bandi et al. 1998; Werren et al. 2008; Haegeman et al. 2009).

## **1.2. *Wolbachia* spreading**

*Wolbachia* is an intracellular bacterium incapable of surviving outside the host. It depends on the host not just for surviving and reproduction, but also for the diffusion through the host population (Werren et al. 2008). *Wolbachia* is transmitted from one host generation to the next through the maternal line, passing from mothers to offspring within the oocytes (vertical heritage). Vertical heritage represents a major way for *Wolbachia* spreading, and it occurs both in arthropods and nematodes (Gerth et al. 2013; Werren et al. 2008).

*Wolbachia* is also able to spread through the host population by horizontal transmission: in arthropods, experimental and phylogenetic studies showed that the bacterium is able to be transferred among adults (in rare cases, belonging to different species) (Vavre et al. 1999; Gerth et al. 2013; Cordaux et al. 2001). This mode of transmission still has to be completely clarified, and thus the mechanism (or mechanisms) followed by *Wolbachia* to “jump” from a individual to another is not completely understood. Le Clec'h and collaborators suggested that predation and cannibalism could represent suitable ways for *Wolbachia* horizontal transmission in isopods (Le Clec'h et al. 2013).

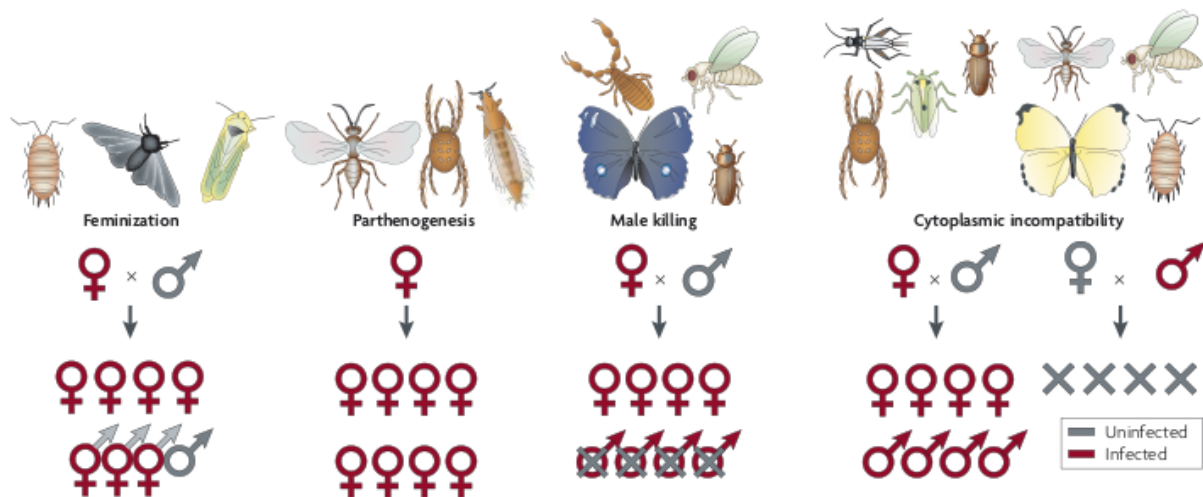
## **1.3. Evolutionary traits of the *Wolbachia*-host symbiotic relationships**

*Wolbachia* deeply affects reproduction and lifespan of most infected hosts, producing consequences at the single organism and population levels. For this reason it is difficult to

classify the *Wolbachia*-host symbiotic relationships, using the simple “pathogenic vs nonpathogenic” or “host-useful vs. host-detrimental” categories. Below I will summarize the most important effects caused by *Wolbachia* to the arthropod and nematode hosts.

### 1.3.1. *Wolbachia* in arthropods

Currently, we know that *Wolbachia* has a crucial role in sex determination of the infected hosts (Werren et al. 2008). Below, I report a description of the reproduction manipulations that *Wolbachia* can induce. Furthermore, the effects of *Wolbachia* infection are schematically represented in Figure 1.



**Figure 1.**

*Wolbachia* is able to affect the host reproduction causing an array of phenotypes: feminization, parthenogenesis, male killing and cytoplasmic incompatibility. These phenotypes are here schematically represented (figure was retrieved from Werren et al. 2008)

#### *Cytoplasmic Incompatibility:*

During the first half of the '900, several cross-mating experiments were performed on a wide array of insect species, collected worldwide. The results of these experiments highlighted the existence of an isolating mechanism that doesn't exactly correspond to

species classification. White-Smith and Woodhill summarized this problem as follows.

*“The isolating mechanisms which exist between closely allied species include a range of types similar to those known in Drosophila. They may be genetic, mechanical, ecological, physiological, or behaviouristic in nature. Such isolating mechanisms usually operate in both reciprocal directions between males and females of the species concerned. Differences in fertility between reciprocal crosses, however, are known to occur between races, subspecies or species in several genera (Toumanoff, 1939, 1950; Downs and Baker, 1949; 1953; Bonnet, 1950; Perry, 1950; Woodhill, 1949, 1950; Marshall, 1938; Laven, 1951, Dobrotworsky and Drummond, 1953) and it is apparent that the phenomenon is widespread in the Culicidae. It has introduced difficulties and complexities in the appreciation of specific and subspecific categories, and it has a significant bearing on problems of medical entomology. It is also significant to genetical and evolutionary theory.” (White-Smith and Woodhill 1954).*

Yen and Barr described the phenomenon in *C. pipiens* as follows:

*“If males of one strain of Culex pipiens are crossed with females of a strain from a different geographical area, the number of offspring may be either normal or small, or there may be none at all. These three results have been called compatible, partially compatible and incompatible, respectively. Reciprocal crosses may give the same or different results. For example, strain A males may be fully compatible with strain B females, but strain B males may be compatible, incompatible, or only partially compatible, with strain A females.” (Yen & Barr 1971).*

The name given to this intriguing phenomenon was “Cytoplasmic incompatibility” (CI). Until the '70s, the most reliable hypotheses to explain the CI involved insect nuclear factors or mitochondrial-like factors (Kitzmilller 1976). Yen and Barr proposed, in 1971, that the cause of CI could be the bacterium *W. pipientis*, an external factor not directly associated to the mosquito genetics (Yen & Barr 1971). During the following years the fundamental role of *Wolbachia* on the CI phenomenon was clarified, albeit the molecular mechanism still remains unknown.

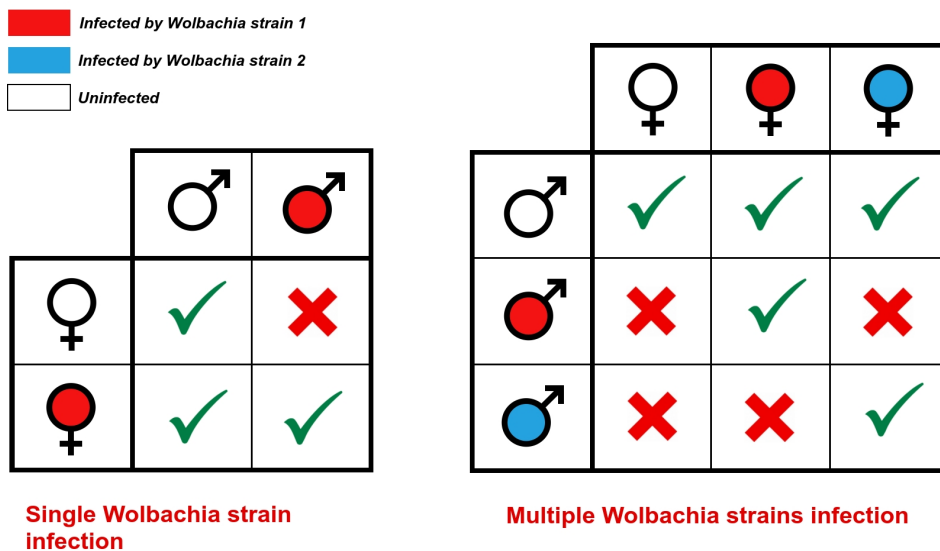
The CI has been reported from insect populations infected by one or more *Wolbachia*



strains at the same time.

If a single strain of *Wolbachia* infects the host population the CI pattern results as outlined in Figure 2: *Wolbachia* inhibits the reproduction in the cross between infected males and uninfected females, without effects on all the other infected/uninfected crosses. The final effect is that *Wolbachia*-infected males sterilize the uninfected females, drastically reducing their reproductive fitness (Werren 1997).

Currently, the most reliable hypothesis for the CI mechanism is the mod/resc model: *Wolbachia* alters the host males producing a *modification* factor (mod, possibly a poison-like molecule) within the sperm, and the host females producing the *relative rescue* factor (resc, possibly an antidote-like) within the oocytes. When a *Wolbachia*-free female ( $\text{♀}/\text{resc-}$ ) is fecundated by a *Wolbachia*-infected male ( $\text{♂}/\text{mod+}$ ), the poison-like mod factor reaches the oocytes sterilizing them. In case of [ $\text{♂}/\text{mod+} + \text{♀}/\text{resc+}$ ], the antidote-like rescue factor produced by *Wolbachia* within the host oocytes prevents the mod activity, and thus the sterilization. The following combinations [ $\text{♂}/\text{mod-} + \text{♀}/\text{resc+}$ ] and [ $\text{♂}/\text{mod-} + \text{♀}/\text{resc-}$ ] do not produce the CI phenotype (Werren et al. 2008; Engelstädter & Telschow 2009) (Figure 2).



**Figure 2.**

Schematic representation of the uni and bi directional cytoplasmic incompatibility (CI) respectively due to single and multiple strain infections of *Wolbachia*. The two tables show success (green tick marks) or failure (red crosses) of offspring production of crosses between parents with different infection states.

When multiple strains of *Wolbachia* infect an host population, each strain produces a specific pair of mod/resc factors. For this reason matings result fertile in two cases: i) if both male and female have the same infective status (uninfected or infected by the same *Wolbachia* strain); ii) if the female only is infected (Figure 2).

Currently, the molecular mechanism (or mechanisms) used by *Wolbachia* to induce CI is unknown.

#### *Parthenogenesis:*

In arthropods with arrhenotokous development, such as mites, Hymenopteras and Thysanoptera, *Wolbachia* can induce parthenogenesis. In species with arrhenotokous development, unfertilized eggs (haploid) develop into males and fertilized eggs (diploid) into females. *Wolbachia* is able to interfere with the early stages of the embryonic development, leading to the production of diploid cells (which will develop to females) from the haploid unfertilized eggs (which normally develop to males) (Werren et al. 2008).

At population level, infection by parthenogenetic *Wolbachia* induces the increasing of the frequency of females, with a significant distortion of the sex ratio. This effect represent an important benefit for *Wolbachia*, considering that it is principally transmitted through the maternal line (Werren 1997; Gerth et al. 2013).

#### *Male killing:*

In species belonging to the Coleoptera, Diptera, Lepidoptera and Pseudoscorpiones orders, *Wolbachia* strains able to kill males have been observed. The male-killing phenotype likely occurs as consequence of a lethal feminization of genetic males (Werren et al. 2008). Indeed, the genetic males produced by *Wolbachia*-infected hosts result feminized before they die, during the larval development (Werren et al. 2008). At the same time, the genetic females produced by hosts deprived of *Wolbachia* die during the larval development (Werren et al. 2008).

#### *Feminization:*

*Wolbachia* can affect the host sex determination, converting the genetic males to phenotypic females (Werren et al. 2008). This phenomenon has been observed in several

isopods from the order Oniscidea and in insects, such as *Eurema hecabe* and *Zyginidia pullula*. In isopods the mechanism has been elucidated: *Wolbachia* invades the host androgenic glands, causing hypertrophy and inhibition, and thus resulting in the feminilization of genetic males.

In conclusion, the vertical transmission is likely the most important way of diffusion for *Wolbachia* (Gerth et al. 2013). It takes place only through the females, and the phenotypes produced by *Wolbachia* host manipulation (Figure 1) increase the fitness of *Wolbachia*-infected females. This is likely one of the main causes of the *Wolbachia* success in arthropods (Werren 1997).

Manipulation of host reproduction, is however not the sole *Wolbachia*-induced effect that can explain its evolutionary success in arthropods. Indeed, evidences for *Wolbachia* anti-viral host protection have been reported, and experimental data suggest that *Wolbachia* could increase the surviving fitness of arthropod hosts (Hedges et al. 2008; Rainey et al. 2014).

### **1.3.2. *Wolbachia* in nematodes**

The discovery of the symbiosis between *Wolbachia* and a non-arthropod host (see section 1.1) could be assessed as a paradigm shift in the *Wolbachia* research field. Until 1995, *Wolbachia* had been considered a bacterium strictly associated to arthropods, likely as consequence of a long co-evolution. The work of Sironi and collaborators (1995) showed that at least one host shift occurred during the evolution of *Wolbachia*, opening an array of intriguing questions: e.g. “Can *Wolbachia* manipulate the reproduction of nematodes?”, “The first '*Wolbachia*-host' (wide sense) was a nematode-like organism, an arthropod-like organism, or neither?”.

The early studies on the nematode *Wolbachia* strains tried to face the question “How does *Wolbachia* interact with the nematode host?”. Indeed, similar studies had been performed on arthropods, evaluating the effects of *Wolbachia* on the biology of the host (typically on its reproduction) by removing the bacterium with tetracycline treatment. Thus, Bandi and collaborators (Bandi et al. 1999) subjected birds infected by the filaria *Brugia pahangi* and dogs infected by *D. immitis* to tetracycline treatments. The treated filarial nematodes

showed a strong inhibition of the development of the microfilaraemia, observed with optical microscopy and transmission electron microscopy (TEM), (Bandi et al 1999). Furthermore, in jirds the inhibition of the development of third-stage *B. pahangi* larvae to the adult stage was observed.

These results, and further studies, indicate that *Wolbachia* has a crucial role in the filarial host larvae development, and suggest that this symbiotic relationship between the bacterium and the host could be classified as mutualistic.

The analyses of *Wolbachia* prevalence in infected filarial nematodes show that the bacterium is present in ~ 100% of all the individuals (Werren et al. 2008). This result is consistent with the hypothesis of mutualistic relationship between *Wolbachia* and nematode hosts.

In a mutualistic relationship, we should expect congruence between the host and symbiont phylogenies. These kind of relationships likely originated as consequence of a long period of co-evolution, which produces congruent phylogenetic signals in host and symbiont genomes (Yamamura 1993). Indeed, the nematode *Wolbachia* phylogeny is congruent with host nematode phylogeny (Bandi et al. 1998; Casiraghi et al. 2001; Comandatore et al. 2013; Gerth et al. 2014) , supporting the mutualistic nature of this symbiosis.

The *Wolbachia*-nematode symbiotic relationship cannot be easily studied through experiments due to the complexity of the biological model, in particular for two reasons: i) filarial nematodes are parasites with life cycles that are very difficult to replicate in laboratory; ii) currently, it is not possible to have pure cultures of *Wolbachia*.

Genome sequencing technologies allow to partially bypass these problems, studying the *Wolbachia*-nematode relationship without the necessity of maintaining this biological model in laboratory. Foster and collaborators sequenced the genome of the *Wolbachia* strain endosymbiont of *Brugia malayi* (the *Wolbachia* strain was labeled wBm) (Foster et al. 2005) . Genes contained in the genome were annotated and the metabolic pathways were inferred on the basis of these detected genes.

The metabolic reconstruction showed that the bacterium is not metabolic auto-sufficient, and suggests metabolic complementary between wBm and the host. In particular, the

pathways mainly involved in this symbiotic metabolic complementarity are the following: i) the pathway for de novo biosynthesis of purines and pyrimidines; ii) the pathway for the riboflavin and flavin adenine dinucleotide biosynthesis; ii) the pathway for the heme cofactor (Foster et al. 2005). Indeed, the wBm genome contains all the genes involved in the purine and pyrimidine biosynthesis but it does not include the gene for the ADP/ATP translocase protein, which is involved in the nucleotide-triphosphates uptake from the host. This absence suggests that the bacterium could synthesize purine and pyrimidine not only for internal consumption, but also to supplement the host production, likely during expansive host metabolic stages, such as oogenesis and embryogenesis. Furthermore, the wBm genome encodes for all the enzymes involved in the riboflavin and flavin adenine dinucleotide biosyntheses, suggesting a fundamental role for the bacterium as essential provider of coenzymes to the host. The wBm genome contains all the genes involved in the heme biosynthesis, a crucial pathway absent in the host metabolism. On the other hand, wBm lacks all the enzymes for de novo biosynthesis of many vitamins and cofactors (e.g. Coenzyme A, NAD, biotin etc.), for which it depends from the host (Foster et al. 2005).

In conclusion, the results of antibiotic treatment studies (Bandi et al. 1999), prevalence screenings, phylogenetic reconstructions (Bandi et al. 1998; Casiraghi et al. 2001; Comandatore et al. 2013; Gerth et al. 2014) and metabolic analysis (Foster et al. 2005) coherently suggest that the symbiotic relationship between *Wolbachia* and nematode hosts can be classified as mutualistic.

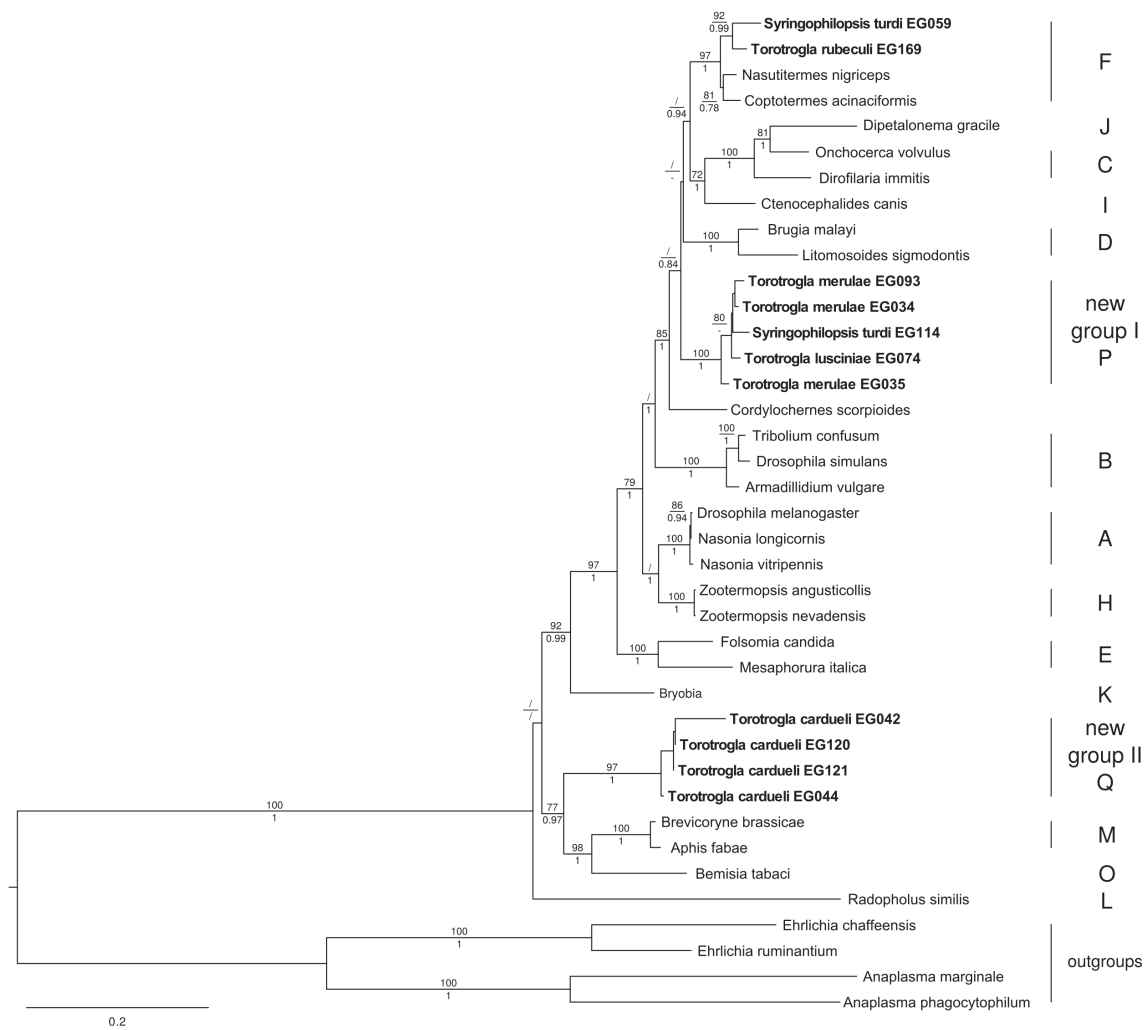
#### **1.4. The *Wolbachia* lineages**

*Wolbachia pipientis* is the sole species described within the genus, but phylogenetic reconstructions suggest the existence of high, and uneven, genetic variability among the strains of the clade (Lo et al. 2002; Bandi et al. 1998; Bordenstein et al. 2009; Fenn et al. 2006). During the evolution of *Wolbachia* a certain number of distinct lineages originated (Bandi et al. 1998; Lo et al. 2002), but the low number of *Wolbachia* sequences available in database and the difficulty in obtaining a robust phylogenetic reconstruction, led to the classification of all the strains within the same species (Lo et al. 2002). *Wolbachia* strains were then organized into monophyletic clusters labeled “supergroups” followed by a capital letter (e.g. supergroup A, supergroup B, etc.) (Lo et al. 2002).

The two main reasons that prevented the researches from obtaining a robust phylogenetic reconstruction of *Wolbachia* genus evolution are: i) the star radiation that occurred at the origin of the genus, during which *Wolbachia* moved towards a fast variability increase, and likely during which many of the lineages originated; ii) the high number of horizontal gene transfer events occurred during the evolution of many *Wolbachia* lineages (Bordenstein et al. 2009; Fenn et al. 2006). During my Ph.D. period I faced the problem of phylogenetic reconstruction of *Wolbachia* evolution using a genomic approach (see section 3).

Currently, 14 different supergroups (A-Q) have been described with the genus (Glowska et al. 2015; Augustinos et al. 2011) (Figure 3). *Wolbachiae* which belong to the supergroups A and B have been observed in association only with arthropods, in opposite to the C and D *Wolbachia* strains the live in association only with nematodes (Bandi et al. 1998; Lo et al. 2002), E *wolbachiae* have been described in springtails and the F supergroup includes *Wolbachia* strains observed in arthropods and in nematodes.

Taxonomic ranking of the supergroups remains an open topic in the *Wolbachia* research field. Currently, almost 30 genomes of *Wolbachia* strains from A-F supergroups are available and the question of supergroups ranking level can be faced. During my Ph.D. period I compared the genomes of several *Wolbachia* strains, spanning from A to F supergroups, in order to test if the supergroup grouping is consistent with the genomic features of these symbionts (see section 4).



**Figure 3.**

Maximum likelihood reconstruction of the *Wolbachia* supergroups phylogeny. Supergroups from A to Q are reported with capital letters. *Wolbachia* strains belonging to supergroups P and Q, the last described, are reported in bold. The tree was published in Glowska et al. 2015.

# Bibliography

- Augustinos AA et al. 2011. Detection and characterization of Wolbachia infections in natural populations of aphids: is the hidden diversity fully unraveled? *PLoS One*. 6:e28695.
- Bandi C et al. 1999. Effects of tetracycline on the filarial worms *Brugia pahangi* and *Dirofilaria immitis* and their bacterial endosymbionts Wolbachia. *Int. J. Parasitol.* 29:357–64.
- Bandi C et al. 1998. Phylogeny of Wolbachia in filarial nematodes. *Proc. Biol. Sci.* 265:2407–13.
- Bordenstein SR et al. 2009. Parasitism and mutualism in Wolbachia: What the phylogenomic trees can and cannot say. *Mol. Biol. Evol.* 26:231–241.
- Casiraghi M et al. 2001. A phylogenetic analysis of filarial nematodes: comparison with the phylogeny of Wolbachia endosymbionts. *Parasitology*. 122 Pt 1:93–103.
- Le Clec'h W et al. 2013. Cannibalism and predation as paths for horizontal passage of Wolbachia between terrestrial isopods. *PLoS One*. 8:e60232.
- Comandatore F et al. 2013. Phylogenomics and analysis of shared genes suggest a single transition to mutualism in Wolbachia of nematodes. *Genome Biol. Evol.* 5:1668–1674.
- Cordaux R et al. 2001. Wolbachia infection in crustaceans: novel hosts and potential routes for horizontal transmission. *J. Evol. Biol.* 14:237–243.
- Engelstädter J and Telschow A. 2009. Cytoplasmic incompatibility and host population structure. *Heredity (Edinb)*. 103:196–207.
- Fenn K et al. 2006. Phylogenetic relationships of the Wolbachia of nematodes and arthropods. *PLoS Pathog.* 2:e94.
- Foster J et al. 2005. The Wolbachia genome of *Brugia malayi*: endosymbiont evolution within a human pathogenic nematode. *PLoS Biol.* 3:e121.
- Gerth M et al. 2014. Phylogenomic analyses uncover origin and spread of the Wolbachia pandemic. *Nat. Commun.* 5:5117.
- Gerth M et al. 2013. Tracing horizontal Wolbachia movements among bees (*Anthophila*): a combined approach using multilocus sequence typing data and host phylogeny. *Mol. Ecol.* 22:6149–62. doi: 10.1111/mec.12549.
- Glowska E, Dragun-Damian A, Dabert M, Gerth M. 2015. New Wolbachia supergroups detected in quill mites (*Acari: Symbionidae*). *Infect. Genet. Evol.* 30:140–6.
- Haegeman A et al. 2009. An endosymbiotic bacterium in a plant-parasitic nematode: member of a new Wolbachia supergroup. *Int. J. Parasitol.* 39:1045–54.
- Hedges LM et al. 2008. Wolbachia and virus protection in insects. *Science*. 322:702.
- Hertig M. 2009. The Rickettsia, Wolbachia pipientis (gen. et sp.n.) and Associated Inclusions of the Mosquito, *Culex pipiens*. *Parasitology*. 28:453.
- Kitzmiller JB. 1976. Genetics, cytogenetics, and evolution of mosquitoes. *Adv. Genet.* 18:315–433.



- Lo N et al. 2002. How many wolbachia supergroups exist? *Mol. Biol. Evol.* 19:341–6. <http://www.ncbi.nlm.nih.gov/pubmed/11861893> (Accessed March 4, 2015).
- Martijn J et al. 2015. Single-cell genomics of a rare environmental alphaproteobacterium provides unique insights into Rickettsiaceae evolution. *ISME J.*
- De Oliveira CD et al. 2015. Broader prevalence of Wolbachia in insects including potential human disease vectors. *Bull. Entomol. Res.* 1–11.
- Pruneau L et al. 2014. Understanding Anaplasmataceae pathogenesis using “Omics” approaches. *Front. Cell. Infect. Microbiol.* 4:86.
- Rainey SM, Shah P, Kohl A, Dietrich I. 2014. Understanding the Wolbachia-mediated inhibition of arboviruses in mosquitoes: progress and challenges. *J. Gen. Virol.* 95:517–30.
- Sironi M et al. 1995. Molecular evidence for a close relative of the arthropod endosymbiont Wolbachia in a filarial worm. *Mol. Biochem. Parasitol.* 74:223–7.
- Smith-White S and Woodhill AR. 1954. The nature and significance of nonreciprocal fertility in *Aedes scutellaris* and other mosquitoes. *Proc. Linn. Soc. N.S.W.* 79: 163-176
- Vavre F et al. 1999. Phylogenetic evidence for horizontal transmission of Wolbachia in host-parasitoid associations. *Mol. Biol. Evol.* 16:1711–23.
- Weinert LA et al. 2015. The incidence of bacterial endosymbionts in terrestrial arthropods. *Proc. R. Soc. B Biol. Sci.* 282:20150249–20150249.
- Werren JH. 1997. Biology of Wolbachia. *Annu. Rev. Entomol.* 42:587–609.
- Werren JH et al. 2008. Wolbachia: master manipulators of invertebrate biology. *Nat. Rev. Microbiol.* 6:741–751.
- Yamamura N. 1993. Vertical Transmission and Evolution of Mutualism from Parasitism. *Theor. Popul. Biol.* 44:95–109.
- Yen JH and Barr AR. 1971. New hypothesis of the cause of cytoplasmic incompatibility in *Culex pipiens* L. *Nature.* 232:657–8.

## Section 2

“The genome of the heartworm, *Dirofilaria immitis*, reveals drug and vaccine targets”

## The genome of the heartworm, *Dirofilaria immitis*, reveals drug and vaccine targets

Christelle Godel,<sup>\*,†,‡,1</sup> Sujai Kumar,<sup>§,1</sup> Georgios Koutsovoulos,<sup>§,2</sup> Philipp Ludin,<sup>\*,†,2</sup> Daniel Nilsson,<sup>¶</sup> Francesco Comandatore,<sup>#</sup> Nicola Wrobel,<sup>||</sup> Marian Thompson,<sup>||</sup> Christoph D. Schmid,<sup>\*,†</sup> Susumu Goto,<sup>\*\*</sup> Frédéric Bringaud,<sup>††</sup> Adrian Wolstenholme,<sup>‡‡</sup> Claudio Bandi,<sup>#</sup> Christian Epe,<sup>‡</sup> Ronald Kaminsky,<sup>‡</sup> Mark Blaxter,<sup>§,||</sup> and Pascal Mäser<sup>\*,†,3</sup>

<sup>\*</sup>Swiss Tropical and Public Health Institute, Basel, Switzerland; <sup>†</sup>University of Basel, Basel, Switzerland; <sup>‡</sup>Novartis Animal Health, Centre de Recherche Santé Animale, St. Aubin, Switzerland; <sup>§</sup>Institute of Evolutionary Biology and <sup>||</sup>The GenePool Genomics Facility, School of Biological Sciences, University of Edinburgh, Edinburgh, UK; <sup>¶</sup>Department of Molecular Medicine and Surgery, Science for Life Laboratory, Karolinska Institutet, Solna, Sweden; <sup>#</sup>Dipartimento di Scienze Veterinarie e Sanità Pubblica, Università degli studi di Milano, Milan, Italy; <sup>\*\*</sup>Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, Japan; <sup>††</sup>Centre de Résonance Magnétique des Systèmes Biologiques, Unité Mixte de Recherche 5536, University Bordeaux Segalen, Centre National de la Recherche Scientifique, Bordeaux, France; and <sup>‡‡</sup>Department of Infectious Diseases and Center for Tropical and Emerging Global Disease, University of Georgia, Athens, Georgia, USA

**ABSTRACT** The heartworm *Dirofilaria immitis* is an important parasite of dogs. Transmitted by mosquitoes in warmer climatic zones, it is spreading across southern Europe and the Americas at an alarming pace. There is no vaccine, and chemotherapy is prone to complications. To learn more about this parasite, we have sequenced the genomes of *D. immitis* and its endosymbiont *Wolbachia*. We predict 10,179 protein coding genes in the 84.2 Mb of the nuclear genome, and 823 genes in the 0.9-Mb *Wolbachia* genome. The *D. immitis* genome harbors neither DNA transposons nor active retrotransposons, and there is very little genetic variation between two sequenced isolates from Europe and the United States. The differential presence of anabolic pathways such as heme and nucleotide biosynthesis hints at the intricate metabolic interrelationship between the heartworm and *Wolbachia*. Comparing the proteome of *D. immitis* with other nematodes and with mammalian hosts, we identify families of potential drug targets, immune modulators, and vaccine candidates. This genome sequence will support the development of new tools against dirofilariasis and aid efforts to combat related human pathogens, the causative agents of lymphatic filariasis and river blindness.—Godel, C., Kumar, S., Koutsovoulos, G., Ludin, P., Nilsson, D., Comandatore, F., Wrobel, N., Thompson, M., Schmid, C. D., Goto, S., Bringaud, F., Wolstenholme, A., Bandi,

C., Epe, C., Kaminsky, R., Blaxter, M., Mäser, P. The genome of the heartworm, *Dirofilaria immitis*, reveals drug and vaccine targets. *FASEB J.* 26, 4650–4661 (2012). [www.fasebj.org](http://www.fasebj.org)

**Key Words:** comparative genomics · filaria · transposon · Wolbachia

THE HEARTWORM *DIROFILARIA IMMITIS* (Leidy, 1856) is a parasitic nematode of mammals. The definitive host is the dog; however, it also infects cats, foxes, coyotes, and, very rarely, humans (1). Dirofilariasis of dogs is a severe and potentially fatal disease. Adult nematodes of 20 to 30 cm reside in the pulmonary arteries, and the initial damage is to the lung. The spectrum of subsequent pathologies related to chronic heartworm infection is broad, the most serious manifestation being heart failure. Recent rapid spread of *D. immitis* through the United States and southern Europe (2, 3) is being favored by multiple factors. Global warming is expand-

<sup>1</sup> These authors contributed equally to this work.

<sup>2</sup> These authors contributed equally to this work.

<sup>3</sup> Correspondence: Swiss Tropical and Public Health Institute, Socinstrasse 57, 4002 Basel, Switzerland. E-mail: pascal.maeser@unibas.ch

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/us/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

doi: 10.1096/fj.12-205096

This article includes supplemental data. Please visit <http://www.fasebj.org> to obtain this information.

Abbreviations: EST, expressed sequence tag; HMM, hidden Markov model; HSP, high-scoring pair; LTR, long terminal repeat; RNA-Seq, transcriptome shotgun sequencing; SOCS, suppressor of cytokine signaling; wBm *Wolbachia* endosymbiont of *Brugia malayi*; wDi, *Wolbachia* endosymbiont of *Dirofilaria immitis*

ing the activity season of vector mosquitoes, increasing their abundance and the likelihood of transmission of the parasite, and there are growing numbers of pets, reservoir animals, and “traveling” dogs (2, 3).

*D. immitis* is an onchocercid filarial nematode, related to important parasites of humans, such as *Onchocerca volvulus*, the agent of river blindness. The *D. immitis* lifecycle is typical for Onchocercidae. Microfilariae, shed into the bloodstream by adult females, are ingested by a mosquito (various species, including *Aedes*, *Anopheles*, and *Culex* spp.) where they develop into third-stage larvae (L3) and migrate to the labium. Feeding by an infected mosquito introduces L3 into the skin. The prepatent period in the newly bitten dog is 6–9 mo, during which the injected larvae undergo two further molts and migrate *via* muscle fibers to the pulmonary vasculature, where the adult nematodes develop. At present, diagnosis is effective only for patent infections, because it is based on detection of circulating microfilariae or antigens from mature females. Treatment of dirofilariasis is also problematic, because the arsenical melarsomine dihydrochloride, the only adulticide approved by the U.S. Food and Drug Administration, can cause adverse neurological reactions. Treatment carries a significant risk of lethality due to blockage of the pulmonary artery by dead nematodes. No vaccine is available. These issues, together with the alarming increasing spread of *D. immitis*, prompted the American Heartworm Society to recommend year-round chemoprophylactic treatment of dogs (4) to kill the larval stages before they develop into adults. This requires monthly administration of anthelmintics, predominantly macrocyclic lactones, such as ivermectin, milbemycin, or moxidectin.

Human-infective parasites related to *D. immitis* cause subcutaneous filariasis and river blindness and are endemic in tropical and subtropical regions around the globe, with an estimated 380 million people affected (5). Improved diagnostics, new drugs, and, ultimately, effective vaccines are sorely needed. The sequencing of the *Brugia malayi* genome provides a platform for rational drug design, but by itself this single sequence cannot distinguish between idiosyncratic and shared targets that could be exploited for control (6).

Most of the filarial nematodes that cause diseases in humans and animals, including *D. immitis*, *O. volvulus*, *Wuchereria bancrofti*, and *B. malayi*, have been shown to harbor intracellular symbiotic bacteria of the genus *Wolbachia* (e.g., refs. 7–10). These bacteria are vertically transmitted to the nematode progeny, *via* transovarial transmission. In most of the infected nematode species, all individuals are infected (reviewed in ref. 11). Even though the exact role of *Wolbachia* in filarial biology has not yet been determined, these bacteria are thought to be beneficial to the nematode host. Indeed, antibiotics that target *Wolbachia* have been shown to have deleterious effects on filarial nematodes, blocking reproduction, inducing developmental arrest, and killing adult nematodes (e.g., refs. 7–9). This has led to development of research projects with the aim of developing anti-*Wolbachia* chemotherapy as a novel strategy for the control of filarial diseases. *Wolbachia* has also been implicated in the immuno-

pathogenesis of filarial diseases, with a role in the development of pathological outcomes, such as inflammation and clouding of the cornea that is typical of river blindness (12). The genome of *Wolbachia* is thus an additional source of potential drug targets (7–10), but a single genome cannot reveal shared *vs.* unique biochemical weaknesses.

The human pathogenic Onchocercidae do not represent an attractive market for the pharmacological industry, because projected incomes from impoverished communities in developing endemic nations would be unlikely to cover the costs of drug development. The heartworm may hold a possible solution to this problem, because the market potential for novel canine anthelmintics is big, given the costs for heartworm prevention of \$75–100/dog/yr and the estimated number of 80 million dogs in the United States (13). Choosing drug targets that are likely to be conserved in related, human pathogenic species may benefit both canine and human medicine. Here we present the draft genome sequences of *D. immitis* and its *Wolbachia* endosymbiont (wDi) and use these data to investigate the relationship between nematode and endosymbiont and identify new drug and vaccine targets.

## MATERIALS AND METHODS

### *D. immitis* isolates and DNA sequencing

We sequenced two canine *D. immitis* isolates, one from Pavia, Italy, and the other from Athens, Georgia, USA. The Pavia isolate was established in a laboratory lifecycle after primary isolation from an infected dog. Adult Pavia nematodes used for DNA extraction were recovered after necropsy of dogs infected as a control group in ongoing investigations (permit FR401e/08 from the Veterinary Office Canton de Fribourg, Switzerland). The Athens nematodes used for DNA and RNA extraction were from a naturally infected dog necropsied as part of routine clinical surveillance and were not from an established strain. Genomic DNA was extracted (QIAamp DNA extraction kit; Qiagen, Valencia, CA, USA) from individual adult female nematodes from Pavia and Athens isolates, and RNA was extracted (RNeasy kit; Qiagen) from individual female and male nematodes from Athens. Whole-genome shotgun sequences were generated at The GenePool Genomics Facility (University of Edinburgh, Edinburgh, UK) and at Fasteris SA (Geneva, Switzerland) using Illumina GAIIx and HiSeq2000 instruments (Illumina, Inc., San Diego, CA, USA). Several short insert (100- to 400-bp) paired-end amplicon libraries and long insert (3- to 4-kb) mate-pair amplicon libraries were made, and data from four of these were used in the final assembly (details are given on the Web site <http://www.dirofilaria.org>). These yielded a raw data total of 28 Gb in 295 million reads [European Bioinformatics Institute (EBI) Short Read Archive, accession number ERA032353; <http://www.ebi.ac.uk/ena/>]. After trimming low-quality bases (Phred score <20) and filtering out reads with uncalled bases or length <35 b, 271 million reads were used for assembly (Supplemental Table S1).

### Nuclear, mitochondrial, and *Wolbachia* genome assemblies

The short-read data were assembled using ABySS 1.2.3 (14). A number of test assemblies were performed using other assem-

blers, and a range of parameters was tested within ABySS, and the final, optimal assembly was performed using a k-mer length of 35 and scaffolding with the paired-end data only. Assembly qualities were assessed using summary statistics including maximizing the N50 (the contig length at which 50% of the assembly span was in contigs of that length or greater), maximum contig length, and total number of bases in contigs (see Supplemental Table S1) and using biological optimality assessment, such as maximizing the coverage of published *D. immitis* expressed sequence tag (EST) sequences and maximizing the number of *B. malayi* genes matched and the completeness of representation of core eukaryotic genes (using CEGMA; ref. 15). Redundancy due to allelic polymorphism was reduced with CD-HIT-EST (16), merging contigs that were  $\geq 97\%$  identical over the full length of the shorter contig. The mitochondrial genome was assembled by mapping the reads to the published *D. immitis* mitochondrial genome (17) and predicting a consensus sequence of the mitochondrial genomes of the Athens and Pavia nematodes separately. The wDi genome was assembled by first identifying likely wDi contigs in the whole assembly with BLASTn (18) using all *Wolbachia* genomes from EMBL-Bank, and then collecting all raw reads (and their pairs;  $n=6,912,659$ ) that mapped to these putative wDi genome fragments. The reduced set of likely wDi reads was then assembled using an independently optimized ABySS parameter set, using mate-pair information where available. Mitochondrial and wDi contigs were removed from the full assembly to leave the final nuclear assembly.

### Transcriptome shotgun sequencing (RNA-Seq) assembly

The preparation of amplicon libraries and RNA-Seq analysis were performed following standard Illumina TruSeq protocols. A total of 11,019,886 (male) and 21,643,293 (female) read pairs of length 54 bp were produced on the Illumina GAIIx platform (ArrayExpress accession number E-MTAB-714; ENA study accession number ERP000758). After quality filtering, the remaining 31,396,183 pairs were assembled with TransABySS using k-mer values from 23 to 47 in steps of 4 (Supplemental Table S1).

### *D. immitis* nuclear genome protein-coding gene prediction and analysis

Repeats in the *D. immitis* genomic assembly were identified and masked using RepeatMasker 3.2.9 (19), including all "Nematoda" repeats in the RepBase libraries (20). The MAKER 2.08 annotation pipeline (21) was used to identify protein-coding genes based on evidence from the RNA-Seq assembly, alignments to the *B. malayi* proteome (WormBase release WS220; <http://wormbase.sanger.ac.uk/>), predictions made by the *ab initio* gene finder SNAP (22), and predictions from the *ab initio* gene finder Augustus (23) based on the Augustus hidden Markov model (HMM) profiles for *B. malayi*. MAKER predicted 11,895 gene models, and, with alternative splicing, a total of 12,872 transcripts and peptides. We compared the nuclear proteome of *D. immitis* with those of four other species for which complete genome data are available and which span the phylogenetic diversity of the phylum Nematoda (*B. malayi*, *Ascaris suum*, *Caenorhabditis elegans*, and *Trichinella spiralis*). The complete proteomes were compared using all-against-all BLAST, and then clustered using OrthoMCL (24). OrthoMCL clusters were postprocessed to classify clusters by their species content and analyzed with reference to the robust molecular phylogeny of the Nematoda (25). The prediction of *D. immitis* orthologs from *B. malayi*, *C. elegans*, *Homo sapiens*, and *Canis lupus* to identify drug targets was performed with InParanoid (26).

### Analyses of orthology and divergence in filarial *Wolbachia*

The wDi genome was annotated with the RAST server (27), an online resource that uses best-practice algorithms to perform both gene finding and gene functional annotation. Selected metabolic pathways were annotated based on enzyme lists from the KEGG Pathway database (28), after an HMM profile was generated for each enzyme (29) from a ClustalW (30) multiple alignment of a redundancy-reduced set of all the manually curated entries in UniProt (31). Analysis of orthology was performed using the BLAST reciprocal best-hits algorithm (32), with the following cutoff values: *E* value 0.1 and ID percentage 60%. Protein distance for each pair of orthologs was calculated using ProtDist in Phylip 3.69 (33) with the Dayhoff PAM matrix option. Proteins were allocated to functional categories using BLAST against the COG database. Protein distances were then analyzed based on COG categories: within each category we calculated the average distances of protein pairs. To evaluate whether some categories were significantly more variable than others, we performed the Kruskal-Wallis test on COG categories containing more than one ortholog pair. The pairwise Mann-Whitney test was then performed to detect pairs of COG categories that displayed significant differences in their average variation.

### Identification of *Wolbachia* insertions in nematode genomes

To identify potential lateral genetic transfers from *Wolbachia* to the host nuclear genome, the nuclear genome was queried against the 921-kbp wDi genome using the dc-megablast option in BLASTN (NCBI-blast+2.2.25) with default settings. All high-scoring pairs (HSPs) longer than 100 bp with  $>80\%$  identity were kept. Overlapping HSP coordinates on the nuclear genome were merged, and sequences from these coordinates were extracted to obtain putative nuclear *Wolbachia* DNA elements. The *B. malayi* nuclear genome was screened with the *B. malayi* *Wolbachia* (wBm) genome in the same way. The small numbers of *Wolbachia* insertions identified in the nuclear genomes of *Acanthocheilonema viteae* and *Onchocerca flexuosa* (34) were surveyed for matches to the wDi and wBm genomes and cross-compared with the insertion sets from the complete *D. immitis* and *B. malayi* genomes using reciprocal best BLAST searches and filtering alignments shorter than 100 bp. Reciprocal best BLAST matches were isolated and single-linkage clustered.

## RESULTS

### Genome assembly of *D. immitis* and its *Wolbachia* symbiont

Genome sequence was generated from single individuals of *D. immitis* isolated from naturally infected dogs, one from Athens, Georgia (USA) and the other from Pavia (Italy). A total of 16 Gb of raw data was retained after rigorous quality checks, corresponding to  $\sim 170$ -fold coverage of the *D. immitis* nuclear genome (likely to be  $\sim 95$  Mb, similar to related Onchocercidae). The ABySS (35) assembler performed best based on statistical and biological measures (Supplemental Table S1). The mitochondrial and *Wolbachia* wDi genomes were assembled independently. The final nuclear assembly

contained 84.2 Mb of sequence in 31,291 scaffolds with an N50 of 10,584 bases (**Table 1**). The draft genome of wDi consists of 2 scaffolds spanning 0.92 Mb. We identified 99% of previously deposited genome survey sequences putatively from wDi (GenBank accession numbers ET041559 to ET041665) within our wDi assembly. The wDi genome was 16% smaller than that of wBm (1.08 Mb; GenBank accession number NC\_006833), and there was significant breakage of synteny between the two genomes, as has been observed between other *Wolbachia*.

The *D. immitis* and wDi genomes, the annotations we have made on these, and additional technical details and analyses are available through a dedicated genome browser (<http://www.dirofilaria.org>).

### Lack of genetic diversity between the sequenced *D. immitis* isolates

Even though the two sequenced *D. immitis* came from independent isolates from different continents, they showed low genetic differentiation, allowing the raw sequencing data from both nematodes to be co-assembled. We mapped the reads from each nematode back to the draft assembly and identified only 32,729 high-quality single-nucleotide variations, a very low per-nucleotide diversity rate of 0.04%. We identified the sequences corresponding to 11 polymorphic microsatellite loci used previously to analyze the *D. immitis* population structure in North America (36) and genotyped our two isolates *in silico* by counting the predicted numbers of microsatellite repeats at each locus. Both our nematodes could be classified within the diversity of the eastern United States population. The mitochondrial genomes of the two isolates differed at only 6 sites (and were thus >99.9% identical). Surprisingly, compared with the published, Australian *D. immitis* mitochondrion (17), both had many shared differences (each was only 99.5% identical to the published *D. immitis* mitochondrion). Because the ~70 differences were often clumped and were unique in the published *D. immitis*

TABLE 1. Comparison of the genome assemblies of *D. immitis*, *B. malayi*, and *C. elegans*

Characteristic	<i>D. immitis</i>	<i>B. malayi</i>	<i>C. elegans</i>
Assembly size (Mb)	84.2	93.6 <sup>a</sup>	100.3
Protein-coding gene models	11,375	11,434	20,517
Genes per megabase	135	122	205
Predicted proteins	12,344	11,460	31,249
Protein-coding sequence (%)	18.0	13.8	25.4
Median exons per gene	5	5	6
Median exon size (b)	142	139	147
Median intron size (b)	226	213	73
Overall GC content (%)	28.3	30.2	35.4
Exon GC content (%)	37.4	39.4	43.4
Intron GC content (%)	26.6	27.2	32.5

<sup>a</sup>*B. malayi* data are from the GenBank RefSeq dataset; *C. elegans* data from the WS230 dataset. <sup>a</sup>70.8 Mb scaffolds + 17.5 Mb short contigs.

mitochondrial genome compared both with our two genomes and with the genomes of five other filarial nematodes, we suggest that many of these are sequencing errors in the published genome.

### A metazoan genome without active transposable elements

The *D. immitis* genome was surveyed for the three main classes of transposable elements [DNA transposons, long terminal repeat (LTR) retrotransposons, and non-LTR retrotransposons] with tBLASTn (18) using the transposon-encoded proteins as queries. No traces of active or pseudogenized DNA transposons or non-LTR retrotransposons were found, but 376 fragments of LTR retrotransposons of the BEL/Pao family (37) were identified. None of these fragments were predicted to be functional, because all contained frame shifts and stop codons in the likely coding sequence. The *D. immitis* Pao pseudogenes were most similar to Pao family retrotransposons from *B. malayi* (6). In *B. malayi*, several of the Pao retrotransposons are likely to be active, because they have complete open reading frames and LTRs. Overall, however, *B. malayi* has a lower density of Pao elements and fragments (3.4 Pao/Mb, 8.3% of which are predicted to be functionally intact) compared with *D. immitis* (4.6 Pao/Mb, none of which were intact).

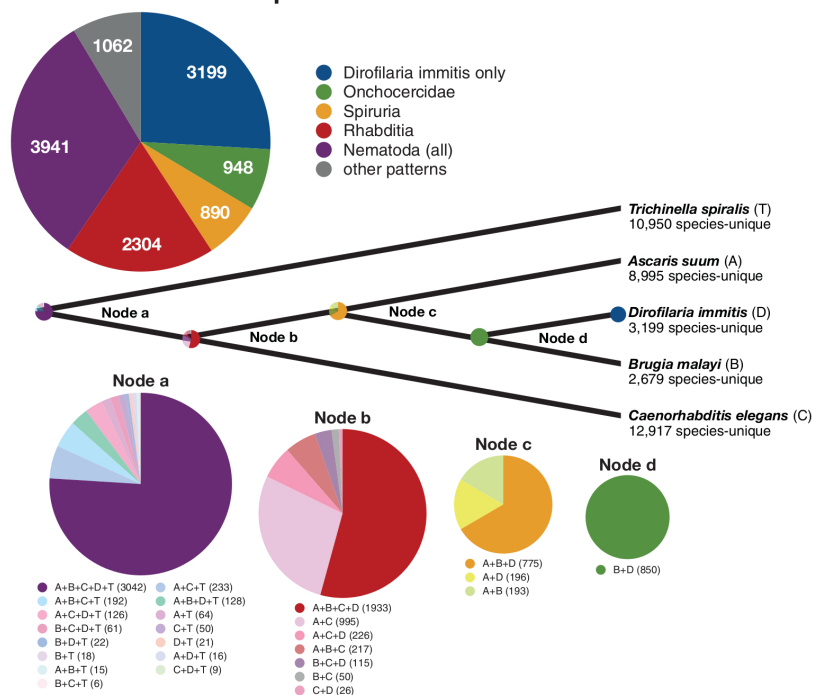
### *D. immitis* nuclear proteome

Protein coding genes were predicted in the nuclear assembly using the MAKER pipeline (21), integrating evidence-based (RNA-Seq and known protein mapping) and *ab initio* methods. Of the 11,375 gene models, 897 were predicted to generate alternate transcripts (Table 1). The total number of predicted proteins of length  $\geq 100$  aa was 10,179, similar to the 9807 predicted in *B. malayi*. Based on matches to *D. immitis* ESTs and core eukaryotic genes (15), the *D. immitis* proteome was likely to be near-complete. Protein-coding exons occupy ~18% of the genome of *D. immitis* and 14% of the genome of *B. malayi* (Table 1), but in *C. elegans* there are nearly twice as many genes, and exons cover ~30% of the genome. The median global identity between a *D. immitis* protein and its best match (as determined by BLASTp) in *B. malayi* was 75%.

*D. immitis* proteins were clustered with the complete proteomes of four other nematode species. These clusters were classified and mapped onto the phylogenetic tree of the five species based on the placement of the deepest node that linked the species that contributed members (**Fig. 1**). The *D. immitis* proteome included 3199 proteins (31% of the total proteome) that were unique to this species, a proportion similar to that found in *B. malayi* (27%), but many fewer (and a lower proportion) compared with those for the other species (for example, *C. elegans* had 63% of its proteome in species-unique

## A *Dirofilaria immitis* protein classification

**Figure 1.** Conserved and novel genes in *D. immitis*. The *D. immitis* proteome was clustered with those of *B. malayi*, *A. suum*, *C. elegans*, and *T. spiralis*. Clusters were then classified based on the membership from the five species according to the current phylogeny of the phylum Nematoda. A) Pie chart showing the distribution of classification of *D. immitis* proteins: *D. immitis* only, singletons and clusters only found in *D. immitis*; Onchocercidae, clusters with members only from *D. immitis* and *B. malayi*; Spiruria, clusters with members only from Onchocercidae and *A. suum*; Rhabditiida clusters with members only from Spiruria and *C. elegans*; Nematoda, clusters with members from all five species (i.e., Rhabditiida and *T. spiralis*); and other patterns, clusters with members not fitting simply into the phylogenetic schema (probably arising from gene loss, lack of predictions, or failure to cluster in one or more species). B) Cluster numbers and patterns of conservation mapped onto the phylogeny of the five species.



## B Cluster origins mapped onto nematode phylogeny

clusters). This difference may be partly due to the 850 proteins in clusters uniquely shared by the relatively closely related *D. immitis* and *B. malayi*, but these clusters only raise the proportion of proteins in phylogenetically local clusters to 47%.

## *D. immitis* genes homologous to known antinematode drug targets

An array of drugs are effective against nematode parasites (**Table 2**). Of these, flubendazole (38), mebendazole

**TABLE 2.** Candidate drug targets, top-down search: current anthelmintics and their known targets in *C. elegans* and orthologs in *D. immitis*

Chemical class	Drug	Target	<i>C. elegans</i>	<i>D. immitis</i>	
Benzimidazole	Albendazole	$\beta$ -Tubulin	BEN-1	DIMM36740	
	Flubendazole Mebendazole				
Imidazothiazole	Levamisole	nACh receptor	LEV-1	DIMM30000 DIMM45965 DIMM08405 DIMM16610	
			LEV-8		
			UNC-29		
			UNC-38		
			UNC-63		
Macrocyclic lactone	Ivermectin Milbemycin Moxidectin Selamectin	Glutamate receptor	AVR-14	DIMM25280, DIMM21120 DIMM22030 DIMM57890	
			AVR-15		
			GLC-1		
			GLC-2		
			GLC-3		
			GLC-4		
			GABA receptor		EXP-1
			GAB-1		
			UNC-49		
			SLO-1		
Cyclodepsipeptide	Emodepside	K <sup>+</sup> channel Latrophilin GPCR	LAT-1	DIMM33210 DIMM33710 DIMM37270, DIMM37275 DIMM17690	
			LAT-2		
			ACR-23		
Aminoacetonitrile derivative	Monepantel	nACh receptor	DES-2		

nAChR, nicotinic acetylcholine; GPCR, G protein-coupled receptor.

(39), levamisole (40), ivermectin, milbemycin, moxidectin, and selamectin (41, 42) have been demonstrated to be active against *D. immitis*. Many drug targets have been identified, particularly through forward genetics in the model nematode *C. elegans* (43) (Table 2). Prominent among these targets are neuronal membrane proteins, highlighting the importance of the neuromuscular junction as a hotspot of anthelmintic drug action. *D. immitis* appears to lack some known targets, notably members of the DEG-3 subfamily of acetylcholine receptors, which contains the presumed targets of monepantel (44). This contrasts with *B. malayi*, which possesses orthologs of DEG-3 and DES-2 (45). In *C. elegans*, the target space of levamisole and ivermectin comprises a large number of ligand-gated ion channels. Although these drugs are effective against heartworm, some of these ion channels do not have an ortholog in *D. immitis* (Table 2), indicating that those present are sufficient to confer drug susceptibility. The identified *D. immitis* orthologs of the known anthelmintic targets can now be monitored in suspected cases of drug resistance.

### New drug target candidates in *D. immitis*

New potential drug targets were identified *in silico* through an exclusion-inclusion strategy (46, 47). Start-

ing from the complete set of predicted *D. immitis* proteins, we excluded proteins that had an ortholog in the dog or human proteome or had multiple paralogs in *D. immitis*. We included proteins that had a *C. elegans* ortholog essential for survival or development (based on RNAi phenotypes) and had predicted function as an enzyme or receptor. Among the 20 candidates identified (Table 3) were several proven drug targets, such as RNA-dependent RNA polymerase (antiviral), apurinic/aprimidinic endonuclease and hedgehog proteins (anticancer; ref. 48), UDP-galactopyranose mutase (against mycobacteria, ref. 49; and kinetoplastids, ref. 50), sterol-C24-methyltransferase (antifungal; ref. 51), and the insecticide target chitin synthase (52). The *D. immitis* orthologs of these enzymes may serve as starting points for the development of new anthelmintics.

### Immune modulators and vaccine candidates

Filarial nematodes modulate the immune systems of their mammalian hosts to promote their own survival and fecundity, but the exact mechanisms used remain enigmatic. Proteases such as leucyl aminopeptidase and protease inhibitors such as serpins and cystatins have been implicated in disruption of immune signal processing (53), and we identified *D. immitis* leucyl amino-

TABLE 3. Candidate drug targets, bottom-up search

<i>D. immitis</i> protein	Predicted function	<i>B. malayi</i> ortholog	<i>H. sapiens</i> log <sub>10</sub> (E)	<i>C. lupus</i> log <sub>10</sub> (E)	<i>C. elegans</i> RNAi
Nucleic acid synthesis and repair					
DIMM09370	RNA-dependent RNA polymerase	BM06623	0.28	0.11	Lethal
DIMM23395	Apurinic/aprimidinic endonuclease	BM17151	>1	-0.12	Lethal
Glycosylation and sugar metabolism					
DIMM15580	dTDP-4-dehydrorhamnose 3,5-epimerase	BM18305	0.23	0.04	Lethal
DIMM03355	β-1,4-Mannosyltransferase	BM20353	0.95	0.94	Lethal
DIMM44525	UDP-galactopyranose mutase	BM01820	0.36	0.08	Molt defective
DIMM36945	Chitin synthase	BM18745, BM02779	-3.52	-4.00	Lethal
Lipid metabolism					
DIMM52545	Lipase	BM01258, BM03783	0.08	-1.60	Lethal
DIMM13730	Sterol-C24-methyltransferase (Erg11)	BM20515	-3.10	-4.00	Lethal
DIMM28375	Methyltransferase	BM18889	-0.03	-0.15	Lethal
Transport					
DIMM21065	Aquaporin	BM04673	-2.05	-0.52	Lethal
Signal transduction					
DIMM13570	Nuclear hormone receptor		-2.40	-4.52	Lethal
DIMM11130	G protein-coupled receptor	BM19106	-1.06	-0.59	Lethal
DIMM32415	G protein-coupled receptor		-1.26	-1.57	Lethal
DIMM39455	G protein-coupled receptor		-2.70	-1.96	Lethal
DIMM13630	Groundhog protein		>1	0.04	Lethal
DIMM47150	Warthog protein	BM01098	>1	0.78	Lethal
DIMM03220	Warthog protein	BM01043, BM17326, BM08657	>1	0.32	Lethal
DIMM11410	Haloacid dehalogenase-like hydrolase	BM19541	-3.22	-4.00	Lethal
DIMM13420	Apoptosis regulator CED-9	BM01838	0.77	-1.02	Lethal

Potential drug targets were filtered from the predicted *D. immitis* proteome using the following criteria: 1) presence of an ortholog in *C. elegans* that has as an RNAi phenotype lethal, L3\_arrest, or molt\_defective; 2) absence of a significant BLAST match ( $E > 10^{-5}$ ) in the predicted proteomes of *H. sapiens* and *C. lupus familiaris*; and 3) predicted function as an enzyme or receptor.



peptidase, as well as 3 cystatins, and many serpins (Table 4). Another route to modulation is through recruitment of nematode homologs of ancient system molecules that have been redeployed in the mammalian immune system, such as TGF- $\beta$  and macrophage migration inhibition factor (MIF). In *D. immitis*, we identified 2 MIF genes, orthologs of the MIF-1 and MIF-2 genes of *B. malayi* and *O. volvulus* and 4 TGF- $\beta$  homologs (Table 4). Another proposed route to modulation is by mimicry of immune system signals. We identified a homolog of suppressor of cytokine signaling 5 (SOCS5), a negative regulator of the JAK/STAT pathway and inhibitor of the IL-4 pathway in T-helper cells, promoting TH1 differentiation (54). Several viruses induce host SOCS protein expression for immune evasion and survival (55). Interestingly, SOCS5 homologs were also identified in the animal-parasitic nematodes *B. malayi*, *D. immitis*, *Loa loa*, *A. suum*, and *T. spiralis*, but were absent from the free-living *C. elegans*, the necromenic *Pristionchus pacificus*, and the plant parasitic *Meloidogyne* spp. *D. immitis* and other filarial nematodes (56) may use SOCS5 homologs to mimic host SOCS5. We also identified a homolog of IL-16, a PDZ domain-containing, pleiotropic cytokine (57). In mammals, IL-16 acts *via* the CD4 receptor to modulate the activity of a wide range of immune effector cells, including T cells and dendritic cells (58). Again, this molecule was only present in parasitic nematodes (in-

cluding *A. suum*; ref. 59) and was absent from genomes of free-living and plant parasitic species. We suggest that these molecules and perhaps other mimics of cytokines and modulators belong to the effector toolkit used by filarial nematodes to build an immunologically compromised niche.

We surveyed the *D. immitis* genome for molecules currently proposed as vaccine candidates in other onchocercids (60, 61) and identified homologs for all 14 classes of molecules (Table 4).

#### Analysis of the wDi genome: the *D. immitis*-*Wolbachia* symbiosis

wDi genes were predicted using the RAST online server. We performed an orthology analysis comparing wBm and wDi and found 538 shared proteins. There were 259 (with 8 duplicated) and 329 (with 4 duplicated) unique genes, respectively, for wBm and wDi. COG analysis showed that the total number of genes in each COG category was similar in the two organisms. Analysis of pairwise protein distances between wDi and wBm in different COG categories indicated that there was significant variation (Kruskal-Wallis  $P=0.00077$ ) and pairwise Mann-Whitney tests identified 2 of the 14 high-level COG categories as having elevated divergence between the two *Wolbachia*. The COG categories showing elevated divergence were M (cell wall, mem-

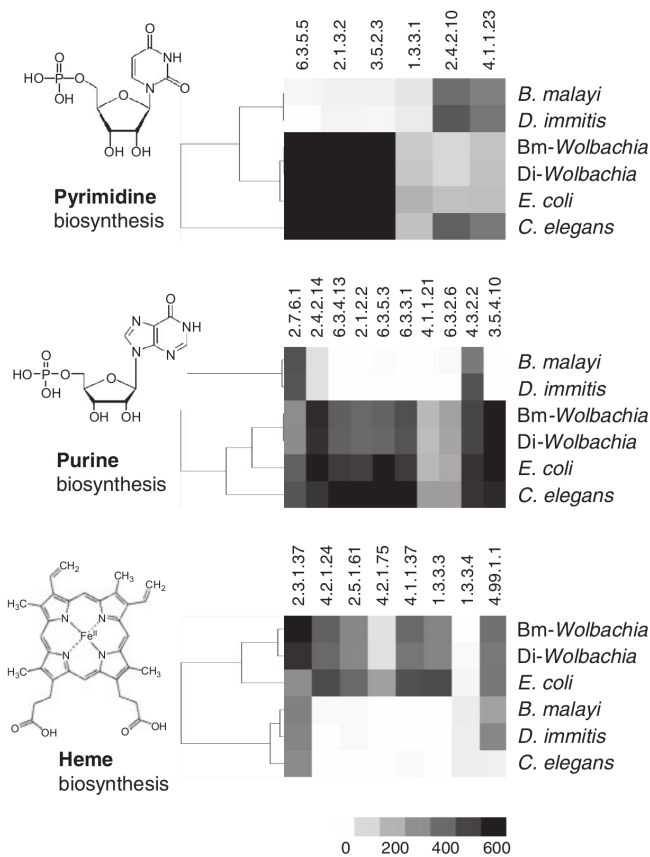
TABLE 4. *D. immitis* potential immune modulators and orthologs of onchocercid vaccine candidates

<i>D. immitis</i> protein	<i>B. malayi</i> ortholog	Description	Potential
DIMM39040, DIMM39045	BM18548	Pi-class glutathione S-transferase (GSTP)	VC
DIMM29150	BM02625	Tropomyosin (TMY)	VC
DIMM29270	BM00759, BM19824	Fatty acid and retinoic acid binding protein (FAR)	VC
DIMM47055	BM03010	Fructose biphosphate aldolase (FBA)	VC
DIMM59360		Astacin metalloprotease MPI	VC
DIMM37935, DIMM46475	BM01859, BM09541, BM14520	Chitinase (CHI)	VC
DIMM48695	BM21967, BM08119	Abundant larval transcript 1 (ALT); unknown function (also known as SLAP)	VC
DIMM48700	BM20051	"RAL-2," unknown function; DUF148 superfamily (also known as SXP-1)	VC
DIMM62215, DIMM45570, DIMM58880	BM03177, BM05783, BM16294	Activation associated proteins [ASP, also known as venom allergen homologs (VAH)]	VC
DIMM58690	BM02480	"OV103" <i>Onchocerca</i> vaccine candidate of unknown function	VC
DIMM12355	BM07484, BM22082	"B8" <i>Onchocerca</i> vaccine candidate of unknown function	VC
DIMM55190, DIMM50565, DIMM48395	BM00175, BM14240, BM04930, BM07956	"B20" <i>Onchocerca</i> vaccine candidate of unknown function	VC
DIMM56580	BM05118	Cysteine proteinase inhibitor 2 (CPI-2)	VC/IM
DIMM18905	BM04900	Cysteine proteinase inhibitor 3 (CPI-3)	VC/IM
DIMM11425	BM21284	Interleukin-16-like (IL16)	IM
DIMM57180	BM00325	Leucyl aminopeptidase (LAP)	IM
DIMM28945	BM06847	Suppressor of cytokine signaling 5 (SOCS5)	IM
DIMM42430	BM07480	Macrophage migration inhibitory factor (MIF-1)	IM
DIMM40455	BM16561	Macrophage migration inhibitory factor 2 (MIF-2)	IM
DIMM23225	BM17713	Transforming growth factor $\beta$ (TGF) homolog of <i>C. elegans</i> TIG-2	IM
DIMM37585	BM20852	TGF homologue of <i>C. elegans</i> DAF-7	IM
DIMM29335	BM21753	TGF homologue of <i>C. elegans</i> DBL-1/CET-1	IM
DIMM61250	BM18112	TGF homologue of <i>C. elegans</i> UNC-129	IM

*B. malayi* orthologs are referred to by their designation in WormBase WS230. IM, immune modulator; VC, vaccine candidate.

brane, and envelope biogenesis) and S (function unknown).

The relationship between filarial nematodes and their *Wolbachia* endosymbionts is thought to be a mutualistic symbiosis (62), because extended treatment of infected mammals with tetracycline and other antibiotics results in clearance of the nematodes. The bases of this symbiosis remain unclear. It has been proposed that wBm provides *B. malayi* with additional sources of critical metabolites such as heme and riboflavin (63). We interrogated the wDi genome to examine the symbiont's biochemical capabilities. *C. elegans* and other nematodes (including *B. malayi*, and, on the basis of the genome sequence presented here, *D. immitis*) are deficient in heme synthesis but wBm has an intact heme pathway (Fig. 2) and a CcmB heme exporter, suggesting that it may support its host by providing heme. wBm has a complete pathway from succinyl-CoA to heme (one apparently missing component, HemG, may be substituted by a functional HemY). wDi lacks both HemY and HemG (and the recently described HemJ that can perform the same transformation). This step



**Figure 2.** Anabolic pathways in *Wolbachia* and *Dirofilaria*. Selected pathways were identified by screening the predicted proteomes with HMM profiles representing each enzyme in the pathway using HMMer (29). The proteomes were hierarchically clustered (77) based on city block distance between the vectors consisting of the best scores (represented as a heat plot) obtained against each profile. A complete prediction of *D. immitis* metabolic pathways is available online at the Draft Genomes page of the Kyoto Encyclopedia of Genes and Genomes ([http://www.genome.jp/kegg/catalog/org\\_list1.html](http://www.genome.jp/kegg/catalog/org_list1.html)).

in the heme pathway is apparently absent in other bacteria, and so this may not indicate a nonfunctional heme synthesis pathway. Further anabolic pathways absent in *D. immitis* but present in wDi are purine and pyrimidine *de novo* synthesis (Fig. 2).

wBm is deficient in folate synthesis because it lacks dihydrofolate reductase and dihydroneopterin aldolase. wDi has both these genes, suggesting that it can use dihydroneopterin as an input to folate metabolism. *Wolbachia* wMel from *Drosophila melanogaster* has both these enzymes, and they are variably present in other alphaproteobacteria. Whether this pathway contributes to the nematode symbiosis is unclear, but it does highlight another component of *Wolbachia* metabolism that may be accessible to drug development. Further wDi gene products that might be exploited as drug targets include nucleic acid synthesis and cell division proteins, such as FtsZ and DnaB, the fatty acid synthesis enzymes FabZ and AcpS, components of the Sec protein secretion system, and, possibly, the peptidoglycan synthesis enzymes of the *Mur* operon. All these are unique proteins in wDi, do not have counterparts in mammals, and are being developed as antibiotic drug targets for bacterial infections (64–68).

Horizontal gene transfer from *Wolbachia* to host nuclear genomes is common in animals harboring this endosymbiont (63), and it has been proposed that these transfers may confer new functionality to the nuclear genome (34, 69, 70), although this is unlikely (71). We identified 868 elements, spanning 219 kb, of >100 b with ≥80% identity to wDi. The *Wolbachia* origin of these elements was supported by clustering based on the frequency distribution patterns (Supplemental Fig. S1) of tetramer palindromes (72). We did not identify the putative complex *Wolbachia* insertion discussed by Dunning-Hotopp *et al.* (70) involving the antigen *Dg2* gene. We found a version of the *Dg2* gene in our predicted transcriptome that contained standard nematode introns, but no evidence of the construct previously described that had the introns of *Dg2* largely replaced with sequences that match 100% to the wDi genome. It is likely that this sequence is a laboratory or computational artifact, especially because the construct includes a cloning vector sequence in addition to *Wolbachia*.

Only 9 of our identified elements matched >80% of the length of a wDi open reading frame and were not interrupted by frame-shifting insertions or deletions or stop codons. Only one of these putative lateral gene transfers had a match to a *Wolbachia* protein of known function (transcription termination factor, NusB). We found no evidence of transcription of this gene in the male and female RNA-Seq data. We applied the same procedure for finding *Wolbachia* insertions to the *B. malayi* genome and identified 654 insertions spanning 327 kb. Only 31 pairs of insertions that were probably derived from homologous *Wolbachia* genes were found (in both of the two genomes). None of these shared insertions had complete open reading frames. Comparison with the *Wolbachia* insertions in the partial ge-

nomes of *A. viteae* and *O. flexuosa*, onchocercid nematodes that have lost their symbionts (34), revealed no insertions shared by all four species. Only 48 insertions were shared by 2 species and 5 were shared by 3. The number of shared fragments was as would be expected from homoplasious, random insertion of *Wolbachia* fragments independently into their host genomes. If ~25% of the genome was randomly transferred in all species, the number of shared fragments expected by chance would be ~45 ( $0.25 \times 0.25 \times 750$  fragments). We thus tentatively conclude that, although elements from wDi have transferred to the nuclear genome, there is no evidence of their functional integration into nematode biology.

## DISCUSSION

The *D. immitis* genome sequence described here is only the second to be determined for an onchocercid nematode, despite the social and economic importance of these parasites. Three genomes were cosequenced: the mitochondrial (at ~4000-fold read coverage of the 13.6-kb genome; this had been determined previously; ref. 17); the genome of the *Wolbachia* symbiont wDi (at ~1000-fold coverage of the 0.9-Mb genome); and the nuclear genome (at ~150-fold coverage of the estimated 95-Mb genome). We used high-throughput, short-read Illumina technology, stringent quality filtering and optimized assembly methods to derive genomes of good draft quality (73). After redundancy reduction, the span of the nuclear assembly was 84.2 Mb, slightly smaller than the 88.3 Mb assembled for *B. malayi* (6). Overall, although the number of scaffolds was approximately equivalent, the contiguity of the *D. immitis* genome assembly was lower than that of *B. malayi*, because of the availability of long-range scaffolding information for the latter species. The predicted nuclear gene set was much smaller than that of *C. elegans*, but of a size similar to that of *B. malayi*. The two onchocercid nematodes also have a lower proportion of species-unique proteins. These two differences may be a feature of the Onchocercidae, because the unpublished *L. loa* genome has only 15,444 predicted proteins (Filarial Worms Sequencing Project, Broad Institute of Harvard and MIT; <http://www.broadinstitute.org/>). Another possibility is that the richer analytic environment for *C. elegans* in particular has permitted the identification of many unique genes using biological evidence (such as transcript information). We will continue to develop and improve the assembly and annotation of *D. immitis* and wDi as additional tools and biological resources become available.

Two peculiarities of the assembled *D. immitis* genome are striking: the lack of genetic diversity and the lack of active transposable elements. The lack of diversity was convenient, in that it allowed us to pool data obtained from two different *D. immitis* isolates, one from Pavia, Italy, and the other from Athens, Georgia, USA. Polymorphisms called from the independent sequencing of

the two isolates yielded a per-nucleotide diversity of 0.04%. Both sequenced isolates fall within the single eastern United States population defined by microsatellite analyses (36). The hypovariability may be a result of the recent admixture of European and American heartworm populations through movement of domestic animals or arise from the very recent introduction of heartworm into the New World by Europeans (74). The first report of dirofilariasis in the United States dates from only 1847, as opposed to a 1626 observation from Italy. The lack of genetic diversity in the nuclear genome will make identification of mutations conferring drug resistance much easier. The lack of DNA transposons and active retrotransposons in *D. immitis* is a strong negative result, because active elements are easy to identify (they are present in multiple, highly similar copies). We identified only fragmented and functionally inactivated segments of Pao-type retrotransposons, similar to those found in and probably still active in *B. malayi*. To our knowledge, this is the first metazoan genome devoid of active transposable elements. The presence of putatively active Pao elements in *B. malayi* suggests that their loss was an evolutionary recent event in *D. immitis*.

The *Wolbachia* wDi genome, with 823 predicted proteins, complements the *D. immitis* nuclear genome in that it encodes enzymes for anabolic pathways that are missing in the latter, *e.g.*, biosynthesis of heme, purine, or pyrimidines (Fig. 2). In contrast to wBm, wDi also carries the genes for folate synthesis, suggesting that folate too might be supplied by the endosymbiont. However, essential metabolites could also be taken up from the mammalian or insect host, and so it remains to be shown whether such metabolites are actually delivered from wDi to *D. immitis*. Analysis of orthology between wBm and wDi revealed that both organisms possess many unique genes (approximately one-third of the total gene complement of each genome). The representation of genes in the different COG categories was similar for wBm and wDi, suggesting that most gene losses occurred before the split of the two lineages or that there have been no biases in gene losses/acquisition after the evolutionary separation. Analysis of protein distances revealed that proteins involved in cell wall/membrane biogenesis (COG category M) displayed more variation between the two organisms compared with the other functional categories. It is reasonable to conclude that the interface between the symbiotic bacterium and the host environment is a place where evolutionary rates are elevated, either as part of an arms race underpinning conflict between the two genomes or as a feature of the dynamic exploitation of the interface in adaptation of the symbiosis. In any case, the endosymbiont, being essential for proliferation of *D. immitis*, represents a target for control of the heartworm. Screening the predicted wDi proteome returned expected antibiotic drug targets such as Fts and Sec proteins, but also the products of the *Mur* operon required for peptidoglycan synthesis.

Many of the anthelmintics used in human medicine

were originally developed for the veterinary sector. We pursued two approaches to identify potential drug targets in *D. immitis*: top-down, starting from the known anthelmintic targets of *C. elegans* (Table 2), and bottom-up, narrowing down the predicted *D. immitis* proteome to a list of essential, unique, and druggable targets (Table 3). Although the majority of the current anthelmintics activate their target (thereby interfering with synaptic signal transduction), the aim of the second approach was to identify inhibitable targets. The criteria applied—presence of an essential ortholog in *C. elegans*, absence of any significantly similar protein in human or dog, and absence of paralogs in *D. immitis*—admittedly missed many of the known anthelmintic targets, e.g., proteins that are not conserved in *C. elegans* or that possess a mammalian ortholog. The aim of the approach was to maximize the specificity of *in silico* target prediction at the cost of low sensitivity. Our goal was to end up with a manageable, rather than complete, list of unique *D. immitis* proteins that are likely to be essential and druggable. Some of the candidates identified are worth further investigation, based on their presumed role in signal transduction, e.g., the nematode-specific G protein-coupled receptors or hedgehog proteins (Table 3). Others have already been validated as drug targets in other systems: sterol-C-24-methyltransferase (EC 2.1.1.41) is a target of sinefungin, chitin synthase (EC 2.4.1.16) is the target of the insecticide lufenuron, and the mannosyltransferase bre-3 is required for interaction of *Bacillus thuringiensis* toxin with intestinal cells (52). The discovery of new *D. immitis* drug targets would be timely because resistance to macrocyclic lactones has recently been reported from the southern United States (75).

Filarial nematodes modulate the immune systems of their hosts in complex ways that result in an apparently intact immune system that ignores a large parasite residing, sometimes for decades, in tissues or the bloodstream. They may also require intact immune systems to develop properly (76). Often immune responses result in a pathologic condition for the host in addition to parasite clearance, and *Wolbachia* may exacerbate these responses (12). We identified a wide range of putative immunomodulatory molecules and, in addition, highlight two *D. immitis* products that may deflect or distract the host immune response: one similar to SOCS5 and the other similar to IL-18. The host-encoded versions of both of these molecules have been implicated in antifilarial immune responses. Strategies for development of a vaccine against filariases depend on delivering the correct antigens to the right arm of the immune system, avoiding induction of dangerous responses, and deflecting or stopping immune suppression by the parasite. We identified homologs of all the current roster of filarial vaccine candidates in our genome, and these can now be moved rapidly into testing in the dog heartworm model. In addition, we defined a large number of potentially secreted *D. immitis* proteins that may con-

tribute to the host-parasite interaction and also be accessible to the host immune system.

Onchocercid parasites share not only a fascinating biology involving immune evasion, arthropod vectors, and *Wolbachia* endosymbionts but also a pressing need for new drugs, improved diagnostic methods, and, ideally, vaccines. We hope that the genome sequence of the heartworm presented here will contribute to an increased understanding of its biology and to new leads for control. EJ

The authors thank Claudio and Marco Genchi (University of Milan, Milan, Italy) for providing *D. immitis* from Pavia, Italy, and Andrew Moorhead (University of Georgia, Athens, GA, USA) for the *D. immitis* from Athens, GA, USA. *D. immitis* genomic and RNA sequencing was performed by FASTER SA (Geneva, Switzerland), and The GenePool Genomics Facility (Edinburgh, UK). The authors are grateful for financial support from the Novartis Fellowship Program (C.G.), the School of Biological Sciences, University of Edinburgh (S.K. and G.K.), the UK Biotechnology and Biological Sciences Research Council (G.K.), the Swedish Research Council (D.N.), the UK Medical Research Council (G0900740; M.B., N.W., and M.T.), the UK Natural Environment Research Council (R8/H10/561; M.B.), the Centre National de Recherche Scientifique (F.B.), the European collaboration Enhanced Protective Immunity Against Filariasis (EPIAF), a focused research project (Specific International Cooperative Action) of the EU (award 242131; M.B.) and the Swiss National Science Foundation (P.M.).

## REFERENCES

1. Lee, A. C., Montgomery, S. P., Theis, J. H., Blagburn, B. L., and Eberhard, M. L. (2010) Public health issues concerning the widespread distribution of canine heartworm disease. *Trends Parasitol.* **26**, 168–173
2. Genchi, C., Rinaldi, L., Mortarino, M., Genchi, M., and Cringoli, G. (2009) Climate and *Dirofilaria* infection in Europe. *Vet. Parasitol.* **163**, 286–292
3. Traversa, D., Di Cesare, A., and Conboy, G. (2010) Canine and feline cardiopulmonary parasitic nematodes in Europe: emerging and underestimated. *Parasit. Vectors* **3**, 62
4. American Heartworm Society (2010) *Diagnosis, Prevention, and Management of Heartworm (Dirofilaria immitis) Infection in Dogs*, American Heartworm Society, Wilmington, DE, USA
5. World Health Organization (2008) Global programme to eliminate lymphatic filariasis. Progress report and conclusions of the meeting of the technical advisory group on the global elimination of lymphatic filariasis. *Wkly. Epidemiol. Rec.* **83**, 333–348
6. Ghedin, E., Wang, S., Spiro, D., Caler, E., Zhao, Q., Crabtree, J., Allen, J. E., Delcher, A. L., Giuliano, D. B., Miranda-Saavedra, D., Angiuoli, S. V., Creasy, T., Amedeo, P., Haas, B., El-Sayed, N. M., Wortman, J. R., Feldblyum, T., Tallon, L., Schatz, M., Shumway, M., Koo, H., Salzberg, S. L., Schobel, S., Pertea, M., Pop, M., White, O., Barton, G. J., Carlow, C. K., Crawford, M. J., Daub, J., Dimmic, M. W., Estes, C. F., Foster, J. M., Ganatra, M., Gregory, W. F., Johnson, N. M., Jin, J., Komuniecki, R., Korf, I., Kumar, S., Laney, S., Li, B. W., Li, W., Lindblom, T. H., Lustigman, S., Ma, D., Maina, C. V., Martin, D. M., McCarter, J. P., McReynolds, L., Mitreva, M., Nutman, T. B., Parkinson, J., Peregrin-Alvarez, J. M., Poole, C., Ren, Q., Saunders, L., Sluder, A. E., Smith, K., Stanke, M., Unnasch, T. R., Ware, J., Wei, A. D., Weil, G., Williams, D. J., Zhang, Y., Williams, S. A., Fraser-Liggett, C., Slatko, B., Blaxter, M. L., and Scott, A. L. (2007) Draft genome of the filarial nematode parasite *Brugia malayi*. *Science* **317**, 1756–1760
7. Slatko, B. E., Taylor, M. J., and Foster, J. M. (2010) The *Wolbachia* endosymbiont as an anti-filarial nematode target. *Symbiosis* **51**, 55–65

8. Hoerauf, A., Nissen-Pahle, K., Schmetz, C., Henkle-Duhrsen, K., Blaxter, M. L., Buttner, D. W., Gallin, M. Y., Al-Qaoud, K. M., Lucius, R., and Fleischer, B. (1999) Tetracycline therapy targets intracellular bacteria in the filarial nematode *Litomosoides sigmodontis* and results in filarial infertility. *J. Clin. Invest.* **103**, 11–18
9. Bandi, C., McCall, J. W., Genchi, C., Corona, S., Venco, L., and Sacchi, L. (1999) Effects of tetracycline on the filarial worms *Brugia pahangi* and *Dirofilaria immitis* and their bacterial endosymbionts *Wolbachia*. *Int. J. Parasitol.* **29**, 357–364
10. Bazzocchi, C., Mortarino, M., Grandi, G., Kramer, L. H., Genchi, C., Bandi, C., Genchi, M., Sacchi, L., and McCall, J. W. (2008) Combined ivermectin and doxycycline treatment has microfilaricidal and adulticidal activity against *Dirofilaria immitis* in experimentally infected dogs. *Int. J. Parasitol.* **38**, 1401–1410
11. Taylor, M. J., Bandi, C., and Hoerauf, A. (2005) *Wolbachia* bacterial endosymbionts of filarial nematodes. *Adv. Parasitol.* **60**, 245–284
12. Saint Andre, A., Blackwell, N. M., Hall, L. R., Hoerauf, A., Brattig, N. W., Volkmann, L., Taylor, M. J., Ford, L., Hise, A. G., Lass, J. H., Diaconu, E., and Pearlman, E. (2002) The role of endosymbiotic *Wolbachia* bacteria in the pathogenesis of river blindness. *Science* **295**, 1892–1895
13. American Pet Products Association (2012) *APPA National Pet Owners Survey*, American Pet Products Association, Greenwich, CT, USA
14. Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y. S., Newsome, R., Chan, S. K., She, R., Varhol, R., Kamoh, B., Prabhu, A. L., Tam, A., Zhao, Y., Moore, R. A., Hirst, M., Marra, M. A., Jones, S. J., Hoodless, P. A., and Birol, I. (2010) De novo assembly and analysis of RNA-seq data. *Nat. Methods* **7**, 909–912
15. Parra, G., Bradnam, K., and Korf, I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067
16. Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682
17. Hu, M., Gasser, R. B., Abs El-Osta, Y. G., and Chilton, N. B. (2003) Structure and organization of the mitochondrial genome of the canine heartworm, *Dirofilaria immitis*. *Parasitology* **127**, 37–51
18. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410
19. Chen, N. (2004) Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* Chapter 4, Unit 4.10
20. Kapitonov, V. V., and Jurka, J. (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genetics* **9**, 411–412; author reply 414
21. Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A., and Yandell, M. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196
22. Korf, I. (2004) Gene finding in novel genomes. *BMC Bioinform.* **5**, 59
23. Stanke, M., and Morgenstern, B. (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467
24. Li, L., Stoeckert, C. J., Jr., and Roos, D. S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178–2189
25. Blaxter, M. L., De Ley, P., Garey, J. R., Liu, L. X., Scheldeman, P., Vierstraete, A., Vanfleteren, J. R., Mackey, L. Y., Dorris, M., Frisse, L. M., Vida, J. T., and Thomas, W. K. (1998) A molecular evolutionary framework for the phylum Nematoda. *Nature* **392**, 71–75
26. Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D. N., Roopra, S., Frings, O., and Sonnhammer, E. L. (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* **38**, D196–203
27. Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formosa, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G. D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., and Zagnitko, O. (2008) The RAST server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75
28. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **27**, 29–34
29. Eddy, S. R. (1995) Multiple alignment using hidden Markov models. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3**, 114–120
30. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680
31. Mulder, N. J., Kersey, P., Pruess, M., and Apweiler, R. (2008) In silico characterization of proteins: UniProt, InterPro and In-Integr8. *Mol. Biotechnol.* **38**, 165–177
32. Remm, M., Storm, C. E., and Sonnhammer, E. L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314**, 1041–1052
33. Felsenstein, J. (1989) PHYLIP—Phylogeny inference package (version 3.2). *Cladistics* **5**, 164–166
34. McNulty, S. N., Foster, J. M., Mitreva, M., Dunning Hotopp, J. C., Martin, J., Fischer, K., Wu, B., Davis, P. J., Kumar, S., Brattig, N. W., Slatko, B. E., Weil, G. J., and Fischer, P. U. (2010) Endosymbiont DNA in endobacteria-free filarial nematodes indicates ancient horizontal genetic transfer. *PLoS One* **5**, e11029
35. Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., and Birol, I. (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123
36. Belanger, D. H., Perkins, S. L., and Rockwell, R. F. (2010) Inference of population structure and patterns of gene flow in canine heartworm (*Dirofilaria immitis*). *J. Parasitol.* **97**, 602–609
37. Eickbush, T. H., and Malik, H. S. (2002) Origins and evolution of retrotransposons. In *Mobile DNA II* (Craig, A. G., Craigie, R., Gellert, M., and Lambowitz, A. M., eds), ASM Press, Washington, DC
38. Guerrero, J., Campbell Seibert, B. P., Newcomb, K. M., Michael, B. F., and McCall, J. W. (1983) Activity of flubendazole against developing stages of *Dirofilaria immitis* in dogs. *Am. J. Vet. Res.* **44**, 2405–2406
39. McCall, J. W., and Crouthamel, H. H. (1976) Prophylactic activity of mebendazole against *Dirofilaria immitis* in dogs. *J. Parasitol.* **62**, 844–845
40. Carlisle, C. H., Atwell, R. B., and Robinson, S. (1984) The effectiveness of levamisole hydrochloride against the microfilaria of *Dirofilaria immitis*. *Aust. Vet. J.* **61**, 282–284
41. Campbell, W. C. (1982) Efficacy of the avermectins against filarial parasites: a short review. *Vet. Res. Commun.* **5**, 251–262
42. McCall, J. W. (2005) The safety-net story about macrocyclic lactone heartworm preventives: a review, an update, and recommendations. *Vet. Parasitol.* **133**, 197–206
43. Sangster, N. C., Song, J., and Demeler, J. (2005) Resistance as a tool for discovering and understanding targets in parasite neuromusculature. *Parasitology* **131** (Suppl.), S179–S190
44. Kaminsky, R., Ducray, P., Jung, M., Clover, R., Rufener, L., Bouvier, J., Weber, S. S., Wenger, A., Wieland-Berghausen, S., Goebel, T., Gauvry, N., Pautrat, F., Skripsky, T., Froelich, O., Komoin-Oka, C., Westlund, B., Sluder, A., and Mäser, P. (2008) A new class of anthelmintics effective against drug-resistant nematodes. *Nature* **452**, 176–180
45. Williamson, S. M., Walsh, T. K., and Wolstenholme, A. J. (2007) The cys-loop ligand-gated ion channel gene family of *Brugia malayi* and *Trichinella spiralis*: a comparison with *Caenorhabditis elegans*. *Invert. Neurosci* **7**, 219–226
46. Doyle, M. A., Gasser, R. B., Woodcroft, B. J., Hall, R. S., and Ralph, S. A. (2010) Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. *BMC Genomics* **11**, 222
47. Holman, A. G., Davis, P. J., Foster, J. M., Carlow, C. K., and Kumar, S. (2009) Computational prediction of essential genes in an unculturable endosymbiotic bacterium, *Wolbachia* of *Brugia malayi*. *BMC Microbiol.* **9**, 243
48. Abbotts, R., and Madhusudan, S. (2010) Human AP endonuclease 1 (APE1): from mechanistic insights to druggable target in cancer. *Cancer Treat. Rev.* **36**, 425–435
49. Borrelli, S., Zandberg, W. F., Mohan, S., Ko, M., Martinez-Gutierrez, F., Partha, S. K., Sanders, D. A., Av-Gay, Y., and Pinto,

- B. M. (2010) Antimycobacterial activity of UDP-galactopyranose mutase inhibitors. *Int. J. Antimicrob. Agents* **36**, 364–368
50. Oppenheimer, M., Valenciano, A. L., and Sobrado, P. (2011) Biosynthesis of galactofuranose in kinetoplastids: novel therapeutic targets for treating leishmaniasis and Chagas' disease. *Enzyme Res.* 2011.415976
  51. Ganapathy, K., Kanagasabai, R., Nguyen, T. T., and Nes, W. D. (2011) Purification, characterization and inhibition of sterol C24-methyltransferase from *Candida albicans*. *Arch. Biochem. Biophys.* **505**, 194–201
  52. Griffiths, J. S., Huffman, D. L., Whitacre, J. L., Barrows, B. D., Marroquin, L. D., Muller, R., Brown, J. R., Hennet, T., Esko, J. D., and Aroian, R. V. (2003) Resistance to a bacterial toxin is mediated by removal of a conserved glycosylation pathway required for toxin-host interactions. *J. Biol. Chem.* **278**, 45594–45602
  53. Maizels, R. M., Gomez-Escobar, N., Gregory, W. F., Murray, J., and Zang, X. (2001) Immune evasion genes from filarial nematodes. *Int. J. Parasitol.* **31**, 889–898
  54. Yoshimura, A., Naka, T., and Kubo, M. (2007) SOCS proteins, cytokine signalling and immune regulation. *Nat. Rev. Immunol.* **7**, 454–465
  55. Akhtar, L. N., and Benveniste, E. N. (2011) Viral exploitation of host SOCS protein functions. *J. Virol.* **85**, 1912–1921
  56. Ludin, P., Nilsson, D., and Mäser, P. (2011) Genome-wide identification of molecular mimicry candidates in parasites. *PLoS One* **6**, e17546
  57. Baier, M., Bannert, N., Werner, A., Lang, K., and Kurth, R. (1997) Molecular cloning, sequence, expression, and processing of the interleukin 16 precursor. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 5273–5277
  58. Cruikshank, W. W., Kornfeld, H., and Center, D. M. (2000) Interleukin-16. *J. Leukoc. Biol.* **67**, 757–766
  59. Wang, J., Czech, B., Crunk, A., Wallace, A., Mitreva, M., Hannon, G. J., and Davis, R. E. (2011) Deep small RNA sequencing from the nematode *Ascaris* reveals conservation, functional diversification, and novel developmental profiles. *Genome Res.* **21**, 1462–1477
  60. Lustigman, S., James, E. R., Tawe, W., and Abraham, D. (2002) Towards a recombinant antigen vaccine against *Onchocerca volvulus*. *Trends Parasitol.* **18**, 135–141
  61. Makepeace, B. L., Jensen, S. A., Laney, S. J., Nfon, C. K., Njongmeta, L. M., Tanya, V. N., Williams, S. A., Bianco, A. E., and Trees, A. J. (2009) Immunisation with a multivalent, subunit vaccine reduces patent infection in a natural bovine model of onchocerciasis during intense field exposure. *PLoS Negl. Trop. Dis.* **3**, e544
  62. Fenn, K., and Blaxter, M. (2004) Are filarial nematode *Wolbachia* obligate mutualist symbionts? *Trends Ecol. Evol.* **19**, 163–166
  63. Fenn, K., and Blaxter, M. (2006) *Wolbachia* genomes: revealing the biology of parasitism and mutualism. *Trends Parasitol.* **22**, 60–65
  64. Li, Z., Garner, A. L., Gloeckner, C., Janda, K. D., and Carlow, C. K. (2011) Targeting the *Wolbachia* cell division protein FtsZ as a new approach for antifilarial therapy. *PLoS Negl. Trop. Dis.* **5**, e1411
  65. Ma, S. (2012) The development of FtsZ inhibitors as potential antibacterial agents. *ChemMedChem* **7**, 1161–1172
  66. Chan, D. I., and Vogel, H. J. (2010) Current understanding of fatty acid biosynthesis and the acyl carrier protein. *Biochem. J.* **430**, 1–19
  67. Segers, K., and Anne, J. (2011) Traffic jam at the bacterial sec translocase: targeting the SecA nanomotor by small-molecule inhibitors. *Chem. Biol.* **18**, 685–698
  68. Katz, A. H., and Caufield, C. E. (2003) Structure-based design approaches to cell wall biosynthesis inhibitors. *Curr. Pharm. Des.* **9**, 857–866
  69. Foster, J., Ganatra, M., Kamal, I., Ware, J., Makarova, K., Ivanova, N., Bhattacharyya, A., Kapral, V., Kumar, S., Posfai, J., Vincze, T., Ingram, J., Moran, L., Lapidus, A., Omelchenko, M., Kyrpidides, N., Ghedin, E., Wang, S., Goltsman, E., Joukov, V., Ostrovskaya, O., Tsukerman, K., Mazur, M., Comb, D., Koonin, E., and Slatko, B. (2005) The *Wolbachia* genome of *Brugia malayi*: endosymbiont evolution within a human pathogenic nematode. *PLoS Biol.* **3**, e121
  70. Dunning-Hotopp, J. C., Clark, M. E., Oliveira, D. C., Foster, J. M., Fischer, P., Munoz Torres, M. C., Giebel, J. D., Kumar, N., Ishmael, N., Wang, S., Ingram, J., Nene, R. V., Shepard, J., Tomkins, J., Richards, S., Spiro, D. J., Ghedin, E., Slatko, B. E., Tettelin, H., and Werren, J. H. (2007) Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* **317**, 1753–1756
  71. Blaxter, M. (2007) Symbiont genes in host genomes: fragments with a future? *Cell Host Microbe* **2**, 211–213
  72. Lamprea-Burgunder, E., Ludin, P., and Maser, P. (2010) Species-specific typing of DNA based on palindrome frequency patterns. *DNA Res.* **18**, 117–124
  73. Chain, P. S., Grafham, D. V., Fulton, R. S., Fitzgerald, M. G., Hostetler, J., Muzny, D., Ali, J., Birren, B., Bruce, D. C., Buhay, C., Cole, J. R., Ding, Y., Dugan, S., Field, D., Garrity, G. M., Gibbs, R., Graves, T., Han, C. S., Harrison, S. H., Highlander, S., Hugenholtz, P., Khouri, H. M., Kodira, C. D., Kolker, E., Kyrpidides, N. C., Lang, D., Lapidus, A., Malfatti, S. A., Markowitz, V., Metha, T., Nelson, K. E., Parkhill, J., Pitluck, S., Qin, X., Read, T. D., Schmutz, J., Sozhamannan, S., Sterk, P., Strausberg, R. L., Sutton, G., Thomson, N. R., Tiedje, J. M., Weinstock, G., Wollam, A., and Detter, J. C. (2009) Genomics. Genome project standards in a new era of sequencing. *Science* **326**, 236–237
  74. Bowman, D. D., and Atkins, C. E. (2009) Heartworm biology, treatment, and control. *Vet. Clin. North Am. Small Anim. Pract.* **39**, 1127–1158, vii
  75. Bourguinat, C., Keller, K., Bhan, A., Peregrine, A., Geary, T., and Prichard, R. (2011) Macrocyclic lactone resistance in *Dirofilaria immitis*. *Vet. Parasitol.* **181**, 388–392
  76. Babayan, S. A., Read, A. F., Lawrence, R. A., Bain, O., and Allen, J. E. (2010) Filarial parasites develop faster and reproduce earlier in response to host immune effectors that determine filarial life expectancy. *PLoS Biol.* **8**, e1000525
  77. Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 14863–14868

Received for publication February 27, 2012.

Accepted for publication July 30, 2012.

## Section 3

“Phylogenomics and analysis of shared genes  
suggest a single transition to mutualism in  
*Wolbachia* of nematodes”

# Phylogenomics and Analysis of Shared Genes Suggest a Single Transition to Mutualism in *Wolbachia* of Nematodes

Francesco Comandatore<sup>1</sup>, Davide Sasseria<sup>1</sup>, Matteo Montagna<sup>1</sup>, Sujai Kumar<sup>2,8</sup>, Georgios Koutsovoulos<sup>2</sup>, Graham Thomas<sup>2</sup>, Charlotte Repton<sup>2</sup>, Simon A. Babayan<sup>3,9</sup>, Nick Gray<sup>3</sup>, Richard Cordaux<sup>4</sup>, Alistair Darby<sup>5</sup>, Benjamin Makepeace<sup>6</sup>, and Mark Blaxter<sup>2,7,\*</sup>

<sup>1</sup>DIVET, Università degli Studi di Milano, Milano, Italy

<sup>2</sup>Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom

<sup>3</sup>Centre for Immunity, Infection and Evolution, University of Edinburgh, Edinburgh, United Kingdom

<sup>4</sup>Université de Poitiers, UMR CNRS 7267 Ecologie et Biologie des Interactions, Equipe Ecologie Evolution Symbiose, Poitiers Cedex, France

<sup>5</sup>Centre for Genomics Research, Institute of Integrative Biology, The University of Liverpool, Liverpool, United Kingdom

<sup>6</sup>Institute of Infection and Global Health, University of Liverpool, Liverpool, United Kingdom

<sup>7</sup>Edinburgh Genomics, University of Edinburgh, Edinburgh, United Kingdom

<sup>8</sup>Present address: Department of Zoology, University of Oxford, Oxford, United Kingdom

<sup>9</sup>Present address: Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow & Moredun Research Institute, Glasgow, United Kingdom

\*Corresponding author: E-mail: mark.blaxter@ed.ac.uk.

Accepted: August 14, 2013

**Data deposition:** Sequence fastq files for *Litomosoides sigmodontis* have been submitted to the European Read Archive with accession ERP001496. Assemblies of wLs and wDi have been deposited in ENA with accessions PRJEB4155 and PRJEB4154 respectively. The ortholog clustering data and alignment files used in the analyses have been deposited with DataDryad with accession doi:10.5061/dryad.4nt8m.

## Abstract

*Wolbachia*, endosymbiotic bacteria of the order Rickettsiales, are widespread in arthropods but also present in nematodes. In arthropods, A and B supergroup *Wolbachia* are generally associated with distortion of host reproduction. In filarial nematodes, including some human parasites, multiple lines of experimental evidence indicate that C and D supergroup *Wolbachia* are essential for the survival of the host, and here the symbiotic relationship is considered mutualistic. The origin of this mutualistic endosymbiosis is of interest for both basic and applied reasons: How does a parasite become a mutualist? Could intervention in the mutualism aid in treatment of human disease? Correct rooting and high-quality resolution of *Wolbachia* relationships are required to resolve this question. However, because of the large genetic distance between *Wolbachia* and the nearest outgroups, and the limited number of genomes so far available for large-scale analyses, current phylogenies do not provide robust answers. We therefore sequenced the genome of the D supergroup *Wolbachia* endosymbiont of *Litomosoides sigmodontis*, revisited the selection of loci for phylogenomic analyses, and performed a phylogenomic analysis including available complete genomes (from isolates in supergroups A, B, C, and D). Using 90 orthologous genes with reliable phylogenetic signals, we obtained a robust phylogenetic reconstruction, including a highly supported root to the *Wolbachia* phylogeny between a (A + B) clade and a (C + D) clade. Although we currently lack data from several *Wolbachia* supergroups, notably F, our analysis supports a model wherein the putatively mutualist endosymbiotic relationship between *Wolbachia* and nematodes originated from a single transition event.

**Key words:** *Wolbachia*, phylogenomics, mutualism, *Litomosoides sigmodontis*, endosymbiosis.

## Introduction

Bacteria of the order Rickettsiales have an intracellular lifestyle and are involved in a variety of associations with eukaryotic

hosts, from protists to vertebrates. These bacteria present distinctive genomic features that are likely to be driven by their intracellular lifestyle, including genome size and gene content

© The Author(s) 2013. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)



reduction, distorted nucleotide composition, and rapid gene evolution (Darby et al. 2007; Renvoise et al. 2011). *Wolbachia pipientis* is one of the most studied members of the Rickettsiales. Symbiotic associations with *Wolbachia* are widespread in arthropods, but have also been identified in nematodes: the animal-parasitic filarial nematodes and the plant-parasitic nematode *Radopholous similis* (Bandi et al. 1998; Werren et al. 2008; Haegeman et al. 2009). The molecular diversity within the single nominal species *Wolbachia pipientis* (Lo et al. 2007) has been used to define a series of 13 supergroups (monophyletic clades; labeled alphabetically, A to N) that show different lifestyles and host ranges (Doudoumis et al. 2012). The A and B supergroups were the first to be described (Werren et al. 1995), followed by the C and D (Bandi et al. 1998). These four are also the most widely investigated *Wolbachia* supergroups. The A and B supergroup strains are associated with arthropods, whereas C and D are associated with filarial nematodes.

In arthropods, *Wolbachia* normally have a patchy distribution among species and populations, and infection is generally associated with alterations of host reproduction, such as parthenogenesis, killing of male embryos, feminization of genetic males, and cytoplasmic incompatibility (Werren et al. 2008; Cordaux et al. 2011). In a few cases, *Wolbachia* has been demonstrated to be essential for the reproduction of the arthropod host (Starr and Cline 2002; Pannebakker et al. 2007). All *Wolbachia* lineages are vertically inherited (from mother to offspring), but horizontal transmission is evident between hosts for numerous strains of the A and B supergroups. The phylogenies of A and B supergroup *Wolbachia* do not track their hosts' phylogenies, suggesting frequent host switching (Werren et al. 1995).

The characteristics of the symbiosis are different in filarial nematodes, where available evidence indicates that the symbionts are beneficial to their hosts. *Wolbachia* usually have 100% prevalence in positive species (Taylor et al. 2005; Ferri et al. 2011), and are strictly vertically inherited, with phylogenies largely congruent with that of their hosts (Bandi et al. 1998; Casiraghi et al. 2001). In addition, they appear to be essential for host survival, as *Wolbachia* elimination with tetracyclines harms the host (Bandi et al. 1999; Hoerauf et al. 1999). Supergroup C and D *Wolbachia* have smaller genomes, and fewer genes, than the parasitic supergroup A and B *Wolbachia* (Foster et al. 2005; Werren et al. 2008) as would be expected from closer integration of host and symbiont genomes. Comparative metabolic reconstruction from the genomes of sequenced *Wolbachia* from filarial nematodes has not revealed an unequivocal signal of the essential symbiotic partnership. Currently favoured models include heme and riboflavin biosynthesis (Foster et al. 2005; Godel et al. 2012), but energy provisioning and immunomodulatory models may be more realistic (Darby et al. 2012).

The origins of the mutualistic relationships of C and D supergroup *Wolbachia* with filarial nematodes are of particular

interest. *Wolbachia* have evolved from intracellular symbionts (Rickettsiales), and the closest related taxa are generally considered to be pathogens, such as the arthropod-infecting A and B supergroups. Are filarial-infecting mutualists monophyletic, implying a single origin of mutualism, or has mutualism arisen independently multiple times? Are the filarial *Wolbachia* more closely related to A or B supergroups? Several studies have highlighted the critical importance, and difficulty, of rooting *Wolbachia* supergroup phylogeny to the solution of this question (Lo et al. 2002, 2007; Fenn et al. 2006; Bordenstein et al. 2009). Two well-known artifacts likely explain the difficulty of obtaining a well-resolved phylogeny: long-branch attraction (LBA), caused by the large distances to the nearest outgroup taxa *Anaplasma* spp. and *Ehrlichia* spp., and a basal, star-like evolutionary radiation of the genus *Wolbachia* (Bordenstein et al. 2009). Fenn et al. (2006), analyzing 42 protein-coding genes from five taxa in A, C, and D supergroups, proposed rooting *Wolbachia* between A and (C + D). Bordenstein et al. (2009) used 21 protein-coding genes from 18 *Wolbachia* taxa, representing the A, B, C, D, E, F, and H supergroups, but did not find unequivocal support for the position of the root, and suggested that reliable resolution of the *Wolbachia* phylogenetic tree would require improved taxon and gene sampling. It is becoming clear that careful selection of loci before analysis is key to robust and believable resolution of many phylogenetic questions when multigene data sets are used (Salichos and Rokas 2013). In particular, coanalysis of genes with different underlying patterns of substitution, horizontal gene transfer, acquisition by hybridization, and hidden paralogy can confound strong signal within data sets.

We have determined the genome sequence of an additional supergroup D *Wolbachia* from the filarial nematode *Litomosoides sigmodontis*. Here, we revisit the selection of single-copy orthologs from completely sequenced genomes for *Wolbachia* phylogenomics, and use an extended gene data set to develop a robust hypothesis of *Wolbachia* relationships.

## Materials and Methods

*Litomosoides sigmodontis* DNA was extracted from nematodes grown in gerbils (*Meriones unguiculatus*) as previously described (Diagne et al. 1990). Short-insert paired-end libraries with 300 and 600 bp inserts were prepared by the GenePool Genomics Facility and sequenced on the Illumina HiSeq2000 with V3 reagents. Reads were corrected using SOAPec (Luo et al. 2012), digitally normalized using khmer (Brown et al. 2012), and preliminary assemblies produced using velvet (Zerbino and Birney 2008). These assemblies were screened for *Wolbachia*-derived sequence using taxon-annotated GC%-coverage plots (Kumar and Blaxter 2011) and the 18 likely *Wolbachia*-derived contigs and their reads selected for stringent reassembly using ABySS (Simpson et al. 2009) (using

a kmer of 83, default coverage cutoff and a minimum of 3 read pairs to join contigs). Joins in the assembly that had low coverage were validated using polymerase chain reaction (PCR). The assembly (wLs.2.0) is available through <http://litomosoides.nematod.es> (last accessed September 5, 2013).

The genomes of 11 *Wolbachia* strains and 4 outgroups were retrieved from the databases (see fig. 1). For wDi from the nematode *Diriofilaria immitis*, we reassembled the genome (Godel et al. 2012) using improved informatic routines, extracting additional read data from the raw genome sequence for *D. immitis*. The new assembly is improved (in that it has many fewer contigs). The contiguity of this new assembly (wDi.2.2) was verified by directed PCR and is available from <http://diriofilaria.nematod.es> (last accessed September 5, 2013).

Ortholog detection was performed using OrthoMCL 1.4 with default settings (Chen et al. 2006). All sequences of each putative orthologous cluster were automatically annotated (using BLASTP with an E-value cutoff of  $10^{-5}$ ) against the Clusters of Orthologous Groups (COG) database (Tatusov et al. 2000). An orthologous cluster was selected for subsequent analyses if all sequences of the cluster were coherently annotated by comparison to the COG database. Orthologous clusters containing members from all 16 genomes and lacking within-genome duplicates were selected, and amino acid sequences aligned using Muscle (Edgar 2004) with default settings. Nucleotide alignments were obtained by retro-translation of these amino acid alignments. The Pairwise Homoplasy Index (PHI) and MaxChi were calculated for each nucleotide alignment with PhiPack (Bruen et al. 2006) with 1,000 permutations and window dimensions of 30, 60, and 100 bases. To detect potential recombination events, recombination analyses were repeated, for each alignment, considering the sequences of the strains belonging to the A + B, C + D, and A + B + C + D supergroup sets. Alignments presenting no evidence of recombination were subjected to mutational saturation analysis with Xia's method (Xia and Xie 2001).

Poorly aligned positions and divergent regions of nucleotide and amino acid alignments were eliminated with Gblocks (Castresana 2000), allowing gap positions (-b5 = all option). Nucleotide and amino acid alignments for the 90 genes were concatenated. Phylogenetic reconstructions were estimated on nucleotide and on amino acid unpartitioned concatenates with Maximum Likelihood (ML) and Bayesian methods using models chosen using jModelTest (Darriba et al. 2012) and Prottest3 (Darriba et al. 2011). The best-fit model for unpartitioned analyses of the concatenated nucleotide alignment was GTR, whereas for unpartitioned analyses of amino acid alignment JTT was identified as optimal. The models selected for partitioned analyses of the alignments are given in [supplementary table S3, Supplementary Material](#) online. The GTR model identified as best-fitting for 30 of 90 nucleotide alignments was used for ML and Bayesian analyses for all partitions.

The amino acid alignment was split into five partitions, grouping all genes that shared the best model among those implemented in MrBayes, and ML and Bayesian analyses were performed on this partitioned concatenate using the best-fit model for each partition.

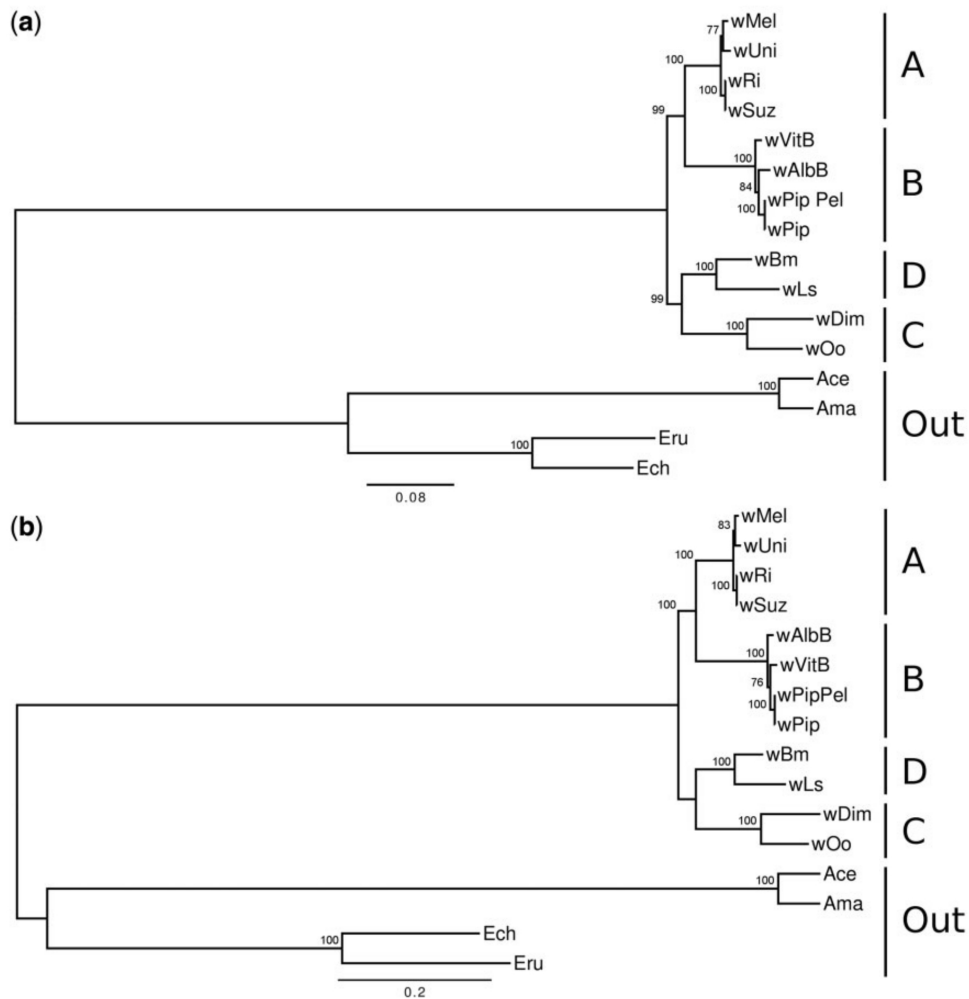
ML phylogenetic analyses were executed with 1,000 rapid bootstrap replicates within RaxML 7.2.8 (Stamatakis et al. 2008). Bayesian phylogenetic analyses were carried out on unpartitioned concatenates with MrBayes 3.2 (Ronquist et al. 2012) on the web-based Bioportal (Kumar et al. 2009). Bayesian Markov chain Monte Carlo analyses were implemented in two parallel analyses, each composed of one cold and five incrementally heated chains that were run for 10 million generations. Trees were sampled every 1,000 generations and burn-in fraction was calculated according to lnL stationary analyses.

Gene presence-absence information for wUni (*Wolbachia* endosymbiont of *Muscidifurax uniraptor*) was removed from all orthologous clusters before performing gene presence-absence analysis, because the wUni genome is not yet complete (Klasson et al. 2009). The ortholog presence-absence matrix was derived from the ortholog cluster data and used to calculate the Bray-Curtis dissimilarity matrix. The Bray-Curtis dissimilarity index evaluates the gene fraction not shared between two taxa with the formula  $1 - \frac{2 \times (A \cap B)}{A + B}$ , where A and B represent the gene sets of the two taxa. Fingerprint Analysis with Missing Data (Schlüter and Harris 2006) was used to perform UPGMA analysis and the heatmap was drawn with R.

## Results

*Litomosoides sigmodontis* is an onchocercid filarial parasite of cotton rats (Hoffmann et al. 2000). The *L. sigmodontis* *Wolbachia*, wLs, was assembled from data generated as part of the ongoing *L. sigmodontis* genome project (Koutsovoulos G, Kumar S, Babayan SA, Blaxter M, unpublished data; see <http://litomosoides.nematod.es>, last accessed September 5, 2013). The wLs genome assembly was generated by identifying contigs in initial genome assemblies that contained *Wolbachia* genes, extracting the raw data that mapped to these contigs and performing independent assembly. The genome was refined through cycles of additional read identification and assembly and validation of some joins by PCR. The raw data have been submitted to INSDC databases under project accession ERP001496.

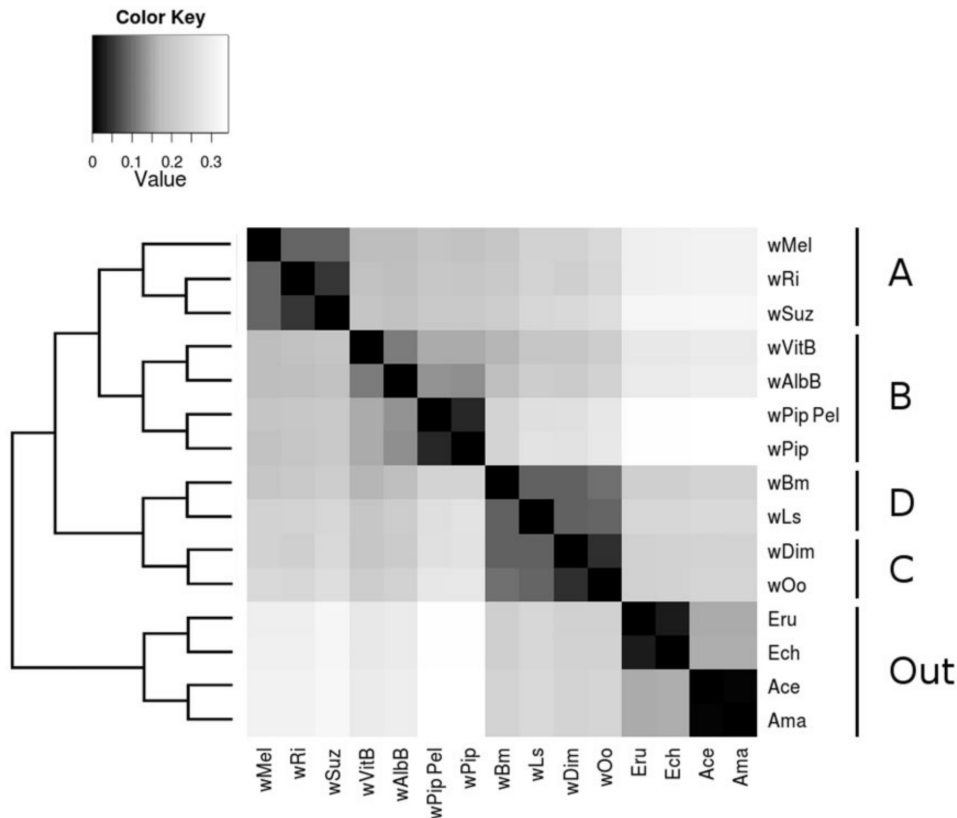
We retrieved whole genome-derived gene data for 11 additional *Wolbachia* strains and four outgroup species from ENA (see Materials and Methods). The genes used in phylogenomic analyses were selected from sets of orthologs, identified as reciprocal best BLAST hits and validated through comparison with the COG database. A subset of orthologs present in all taxa was identified. Ortholog sets showing evidence of paralog duplication, evidence of recombination



**FIG. 1.**—Phylogenomic analysis of *Wolbachia*. Phylogenetic trees generated with RaxML based on amino acid (A) and nucleotide (B) partitioned concatenates. ML bootstrap values are reported above each node of the trees. The corresponding trees generated with MrBayes, showing completely congruent topologies and posterior probability of 1 for each node, are reported in [supplementary figure S1, Supplementary Material](#) online. The strains analyzed are *Wolbachia* endosymbiont of *Drosophila melanogaster*, wMel; *Wolbachia* endosymbiont of *Drosophila simulans*, wUni; *Wolbachia* endosymbiont of *Drosophila sukuzii*, wSuz; *Wolbachia* endosymbiont of *Muscidifurax uniraptor*, wUni; *Wolbachia* endosymbiont of *Culex quinquefasciatus* JHB, wPip; *Wolbachia* endosymbiont of *Culex quinquefasciatus* Pel, wPip Pel; *Wolbachia* endosymbiont of *Nasonia vitripennis*, wVitB; *Wolbachia* endosymbiont of *Aedes albopictus*, wAlbB; *Wolbachia* endosymbiont of *Brugia malayi*, wBm; *Wolbachia* endosymbiont of *Onchocerca ochengi*, wOo; *Wolbachia* endosymbiont of *Dirofilaria immitis*, wDi; *Anaplasma centrale* str. Israel; *Anaplasma marginale* str. Florida; *Ehrlichia chaffeensis* str. Arkansas; *Ehrlichia ruminantium* str. Gardel. Letters A, B, C, and D indicate *Wolbachia* supergroup memberships.

between genomes, or evidence of nucleotide substitution saturation were removed. We identified 1,677 ortholog clusters, 1,519 of which were coherently annotated. The sixteen bacterial genomes shared 390 of the 1,519 gene clusters, 341 of which presented no evidence of duplication. Of these 341 clusters, 126 showed no evidence of recombination, and, of these, 90 showed no evidence of nucleotide substitution saturation. These 90 clusters were retained and used for phylogenomic analysis (see [supplementary table S1, Supplementary Material](#) online, for a complete list of these 90 genes). Maximum likelihood (ML; [fig. 1](#)) and Bayesian phylogenetic inference ([supplementary fig. S1, Supplementary Material](#)

online) using nucleotide and amino acid alignments of these 90 genes differed only in the relative position of the strains wAlbB and wVitB within supergroup B. Other than this disagreement, all nodes received high joint support. Importantly, the length of the branches between the three genera *Wolbachia*, *Anaplasma*, and *Ehrlichia* were reasonably homogeneous and did not suggest the presence of LBA artifacts. Analysis of individual alignments showed that (C + D) monophyly was supported by 48 of the 90 loci, and (A + B) monophyly was supported by 43 loci (see [supplementary information, Supplementary Material](#) online). None of the nodes in the concatenated analysis is supported by a low



**FIG. 2.**—Gene presence–absence analysis of *Wolbachia* genomes. An UPGMA tree (left) was inferred based on the Bray–Curtis dissimilarity matrix calculated on the presence–absence matrix of genes in the examined genomes. The heatmap to the right of the tree represents the values of the Bray–Curtis dissimilarity matrix. Strain abbreviations are as given in figure 1.

number of individual genes, and the majority rule consensus of the individual locus phylogenies is the same as the concatenated analysis. The analyses supported a root placement between the A and B supergroups and the C and D supergroups, yielding a monophyletic filarial mutualist clade.

A presence–absence matrix was constructed from the 1,519 coherently annotated orthologous clusters from the complete *Wolbachia* genomes (i.e., excluding wUni; see Materials and Methods). From this matrix, a pairwise Bray–Curtis dissimilarity matrix was calculated, and this dissimilarity matrix was subjected to UPGMA phenetic analysis (fig. 2). The phenetic analysis was congruent with the sequence-based phylogenomic analyses, linking the A and B and the C and D supergroups.

## Discussion

Previously published molecular phylogenetic reconstructions have not revealed the number of independent transitions to mutualism in filarial nematode *Wolbachia* (Casiraghi et al. 2005; Fenn et al. 2006; Bordenstein et al. 2009). This has been due to limited phylogenetic signal present in few loci (Casiraghi et al. 2005; Bordenstein et al. 2009), a lack of

genomic data from a representative diversity of strains (Fenn et al. 2006), and LBA due to extreme divergence from the nearest outgroup taxa (Bordenstein et al. 2009). We sequenced a new supergroup D genome, wLs of *L. sigmodontis* and collated a 90-gene phylogenomic data set using a custom pipeline designed to remove all loci likely to contain phylogenetic noise. Salichos and Rokas (2013) have recently explored issues of data incongruity in phylogenomic analyses, using a deep phylogeny of yeasts as their model. We concur with their proposals to eliminate rigorously from consideration loci with aberrations in phylogenetic signature, assessed independently of the derivation of the phylogeny. We note that our phylogenetic question is less problematic than their model, with fewer taxa overall and some unquestioned groupings (such as the monophyly of clades A, B, C, and D). Thus, we reduced our original set of more than 400 putative single copy orthologs to a core set of 90 genes with validated behavior. Only 3 of the 21 genes of the Bordenstein set (Bordenstein et al. 2009) passed the stringent assessment for inclusion in our database. ML and Bayesian analyses of nucleotide and amino acid alignments yielded congruent topologies with high statistical support. Importantly, the branch lengths observed between the two genera in the outgroup (*Anaplasma*

and *Ehrlichia*) was comparable to the branch length observed between the outgroup clade (*Anaplasma* + *Ehrlichia*) and the *Wolbachia* clade, suggesting the absence of LBA effects on the phylogenies. Individual locus phylogenies tended to support this hypothesis. Our phylogenies provide strong evidence, with high statistical support, for the monophyletic origin of arthropod (A and B) and nematode (C and D) *Wolbachia* strains (fig. 1 and supplementary fig. S1, Supplementary Material online). Phenetic analysis of gene presence–absence data also supported this set of relationships.

In summary, our analyses demonstrate that arthropod (A and B) and nematode (C and D) *Wolbachia* originated after the split of an ancestral lineage. We cannot determine whether this ancestral lineage was associated with nematodes, arthropods, or another host group, and we cannot derive conclusions on the nature of the symbiosis (parasitic vs. mutualistic) of this ancestor. Considering the phylogenetic position of *Wolbachia* within the order Rickettsiales, we can reasonably infer that it was an intracellular bacterium. A monophyletic origin of the C and D supergroups is congruent with the idea that the characteristics shared by these two supergroups (strict association with the host, strict vertical transmission, and evidence for a beneficial contribution to host biology) originated only once during evolution, in the lineage that led to the *Wolbachia* of filarial nematodes. As noted by authors in previous studies, analysis of *Wolbachia* diversity is compromised by partial sampling across the known supergroups (Casiraghi et al. 2005; Fenn et al. 2006; Bordenstein et al. 2009). We have been able to use complete genome data from only four supergroups and eagerly await emerging data for strains from other supergroups. Of particular interest will be genomic data from supergroup F strains, as these are reported to infect both arthropods and filarial nematodes (Lefoulon et al. 2012). The placement of supergroup F strains in phylogenies is variable, but they are often associated with supergroup C and D (Lefoulon et al. 2012). An exciting possibility is that supergroup F is sister taxon to C and D, and this may represent the lifestyle of the last common C and D ancestor.

The origin of the relationship between nematodes and *Wolbachia* is interesting from both evolutionary and medical standpoints. Nematode *Wolbachia* represent important targets for the treatment of human and animal filariases (Slatko et al. 2010). Our analyses suggest that an endosymbiotic relationship with nematodes is a plesiomorphic character of the (C + D) clade. In this scenario, an ancestral *Wolbachia* strain invaded the first filarid host, evolved a mutualistic association that included strict vertical inheritance, and the C and D supergroups originated through ancient host lineage divergence. Under this model, all these *Wolbachia* strains probably share common metabolic traits that underpin their mutualistic relationship with the filarial hosts. This in turn suggests the possible presence of common anti-*Wolbachia*

pharmacological targets for the control of their pathogenic filarial hosts.

## Supplementary Material

Supplementary figure S1 and tables S1–S3 are available at *Genome Biology and Evolution* online (<http://www.gbe.oxfordjournals.org/>).

## Acknowledgments

This work was supported by postgraduate fellowships awarded to S.K. and G.K. from the School of Biological Sciences, the University of Edinburgh, and to G.S. from the Wellcome Trust. Genome sequencing was supported by an award from the EU FP7 programme (EU Specific International Cooperation Action [SICA] reference 242131 Enhanced Protective Immunity Against Filariasis) awarded to Prof. David Taylor. R.C. was supported by an European Research Council Starting Grant (FP7/2007–2013, grant 260729 EndoSexDet).

## Literature Cited

- Bandi C, Anderson TJ, Genchi C, Blaxter ML. 1998. Phylogeny of *Wolbachia* in filarial nematodes. *Proc Biol Sci.* 265:2407–2413.
- Bandi C, et al. 1999. Effects of tetracycline on the filarial worms *Brugia pahangi* and *Dirofilaria immitis* and their bacterial endosymbionts *Wolbachia*. *Int J Parasitol.* 29:357–364.
- Bordenstein SR, et al. 2009. Parasitism and mutualism in *Wolbachia*: what the phylogenomic trees can and cannot say. *Mol Biol Evol.* 26: 231–241.
- Brown C, Howe A, Zhang Q, Pyrkosz A, Brom T. 2012. A reference-free algorithm for computational normalization of shotgun sequencing data, arXiv:1203.4802.
- Bruen TC, Philippe H, Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172:2665–2681.
- Casiraghi M, Anderson TJ, Bandi C, Bazzocchi C, Genchi C. 2001. A phylogenetic analysis of filarial nematodes: comparison with the phylogeny of *Wolbachia* endosymbionts. *Parasitology* 122:93–103.
- Casiraghi M, et al. 2005. Phylogeny of *Wolbachia pipientis* based on *gltA*, *groEL* and *ftsZ* gene sequences: clustering of arthropod and nematode symbionts in the F supergroup, and evidence for further diversity in the *Wolbachia* tree. *Microbiology* 151:4015–4022.
- Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17:540–552.
- Chen F, Mackey AJ, Stoeckert CJ Jr, Roos DS. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* 34:D363–D368.
- Cordaux R, Bouchon D, Greve P. 2011. The impact of endosymbionts on the evolution of host sex-determination mechanisms. *Trends Genet.* 27:332–341.
- Darby AC, Cho NH, Fuxelius HH, Westberg J, Andersson SG. 2007. Intracellular pathogens go extreme: genome evolution in the Rickettsiales. *Trends Genet.* 23:511–520.
- Darby AC, et al. 2012. Analysis of gene expression from the *Wolbachia* genome of a filarial nematode supports both metabolic and defensive roles within the symbiosis. *Genome Res.* 22:2467–2477.
- Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27:1164–1165.
- Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods.* 9:772.

- Diagne M, Petit G, Liot P, Cabaret J, Bain O. 1990. The filaria *Litomosoides galizai* in mites; microfilarial distribution in the host and regulation of the transmission. *Ann Parasitol Hum Comp.* 65:193–199.
- Doudoumis V, et al. 2012. Detection and characterization of *Wolbachia* infections in laboratory and natural populations of different species of tsetse flies (genus *Glossina*). *BMC Microbiol.* 12(Suppl 1):S3.
- Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.
- Fenn K, et al. 2006. Phylogenetic relationships of the *Wolbachia* of nematodes and arthropods. *PLoS Pathog.* 2:e94.
- Ferri E, et al. 2011. New insights into the evolution of *Wolbachia* infections in filarial nematodes inferred from a large range of screened species. *PLoS One* 6:e20843.
- Foster J, et al. 2005. The *Wolbachia* genome of *Brugia malayi*: endosymbiont evolution within a human pathogenic nematode. *PLoS Biol.* 3:e121.
- Godel C, et al. 2012. The genome of the heartworm, *Dirofilaria immitis*, reveals drug and vaccine targets. *FASEB J.* 26:4650–4661.
- Haegeman A, et al. 2009. An endosymbiotic bacterium in a plant-parasitic nematode: member of a new *Wolbachia* supergroup. *Int J Parasitol.* 39:1045–1054.
- Hoerauf A, et al. 1999. Tetracycline therapy targets intracellular bacteria in the filarial nematode *Litomosoides sigmodontis* and results in filarial infertility. *J Clin Invest.* 103:11–18.
- Hoffmann W, et al. 2000. *Litomosoides sigmodontis* in mice: reappraisal of an old model for filarial research. *Parasitol Today.* 16:387–389.
- Klasson L, et al. 2009. The mosaic genome structure of the *Wolbachia* wRi strain infecting *Drosophila simulans*. *Proc Natl Acad Sci U S A.* 106:5725–5730.
- Kumar S, Blaxter ML. 2011. Simultaneous genome sequencing of symbionts and their hosts. *Symbiosis* 55:119–126.
- Kumar S, et al. 2009. AIR: a batch-oriented web program package for construction of supermatrices ready for phylogenomic analyses. *BMC Bioinformatics* 10:357.
- Lefoulon E, et al. 2012. A new type F *Wolbachia* from Splendidofiliariinae (Onchocercidae) supports the recent emergence of this supergroup. *Int J Parasitol.* 42:1025–1036.
- Lo N, Casiraghi M, Salati E, Bazzocchi C, Bandi C. 2002. How many *wolbachia* supergroups exist? *Mol Biol Evol.* 19:341–346.
- Lo N, et al. 2007. Taxonomic status of the intracellular bacterium *Wolbachia pipientis*. *Int J Syst Evol Microbiol.* 57:654–657.
- Luo R, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1:18.
- Pannebakker BA, Loppin B, Elemans CP, Humblot L, Vavre F. 2007. Parasitic inhibition of cell death facilitates symbiosis. *Proc Natl Acad Sci U S A.* 104:213–215.
- Renvoise A, Merhej V, Georgiades K, Raoult D. 2011. Intracellular Rickettsiales: insights into manipulators of eukaryotic cells. *Trends Mol Med.* 17:573–583.
- Ronquist F, et al. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol.* 61:539–542.
- Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497:327–331.
- Schlüter PM, Harris SA. 2006. Analysis of multilocus fingerprinting data sets containing missing data. *Mol Ecol Notes.* 6:569–572.
- Simpson JT, et al. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Res.* 19:1117–1123.
- Slatko BE, Taylor MJ, Foster JM. 2010. The *Wolbachia* endosymbiont as an anti-filarial nematode target. *Symbiosis* 51:55–65.
- Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML Web servers. *Syst Biol.* 57:758–771.
- Starr DJ, Cline TW. 2002. A host parasite interaction rescues *Drosophila* oogenesis defects. *Nature* 418:76–79.
- Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28:33–36.
- Taylor MJ, Bandi C, Hoerauf A. 2005. *Wolbachia* bacterial endosymbionts of filarial nematodes. *Adv Parasitol.* 60:245–284.
- Werren JH, Baldo L, Clark ME. 2008. *Wolbachia*: master manipulators of invertebrate biology. *Nat Rev Microbiol.* 6:741–751.
- Werren JH, Zhang W, Guo LR. 1995. Evolution and phylogeny of *Wolbachia*: reproductive parasites of arthropods. *Proc Biol Sci.* 261:55–63.
- Xia X, Xie Z. 2001. DAMBE: software package for data analysis in molecular biology and evolution. *J Hered.* 92:371–373.
- Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18:821–829.

Associate editor: John McCutcheon

## Section 4

“Supergroup C *Wolbachia*, mutualist symbionts of filarial nematodes, have a distinct genome structure”

# **Supergroup C *Wolbachia*, mutualist symbionts of filarial nematodes have a distinct genome structure**

Francesco Comandatore<sup>1,2</sup>, Richard Cordaux<sup>3</sup>, Claudio Bandi<sup>1</sup>, Mark L Blaxter<sup>4</sup>, Alistair Darby<sup>5</sup>, Benjamin Makepeace<sup>6</sup>, Matteo Montagna<sup>7</sup>, Davide Sassera<sup>2,\*</sup>

Author information:

1. Dipartimento di Scienze Veterinarie e Sanità Pubblica (DIVET), Università degli Studi di Milano, Milan, Italy
2. Dipartimento di Biologia e Biotecnologie, Università degli Studi di Pavia, Pavia, Italy
3. Université de Poitiers, UMR CNRS 7267 Ecologie et Biologie des Interactions, Equipe Ecologie Evolution Symbiose, Poitiers, France
4. Institute of Evolutionary Biology and Centre for Immunity, Infection and Evolution, The School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3TF, UK
5. Institute of Integrative Biology and the Centre for Genomic Research, University of Liverpool, Liverpool L69 7ZB, UK
6. Institute of Infection & Global Health, University of Liverpool, Liverpool L3 5RF, UK
7. Dipartimento di Scienze Agrarie e Ambientali, Università degli Studi di Milano, Milano, Italy

\*Author for Correspondence: Davide Sassera, Dipartimento di Scienze Agrarie e Ambientali, Università degli Studi di Milano, Milano, Italy, [davide.sassera@unipv.it](mailto:davide.sassera@unipv.it)



# Abstract

*Wolbachia pipientis* is possibly the most widespread endosymbiont of arthropods and nematodes. While all *Wolbachia* strains are currently assigned to a single species, 16 monophyletic clusters of diversity (called supergroups) have been described. Different supergroups have distinct host ranges and symbiotic relationships, ranging from mutualism to manipulation of host reproduction. The evolutionary pathway that led to this wide host range is currently the focus of several studies, and the evolutionary relationships among the supergroups are being clarified by phylogenomic approaches. These phylogenetic reconstructions are characterized by long branches originating from a star-like radiation, likely the signs of major, ancestral, ecological transitions. Here we show that this evolutionary radiation left clear genomic signatures, sorts of 'genomic bauplans' that characterize current supergroups. This result emerged after the comparison of the genomes of 26 *Wolbachia* strains, spanning the diversity of the major supergroups (A-F), analysing codon usage, GC skew, gene loss, transposable element content, horizontal gene transfer and intragenic recombinations. The results of these analyses show patterns coherent with the *Wolbachia* phylogeny, providing new insights on the origin and evolution of supergroups, and of the entire *Wolbachia* genus. Our results suggest that *Wolbachia* supergroups are genetically isolated and, in particular, that strains belonging to the C supergroup present newly evolved genomic characteristics, reminiscent of free-living bacteria. This evidence for genomic isolation suggests that more than one species could be attributed to the genus *Wolbachia*.

## Key words

Wolbachia, species concept, GC skew origin

## Introduction

*Wolbachia* is one of the most widespread and studied genera of intracellular bacteria, encompassing endosymbionts of arthropods and nematodes (Werren et al. 2008; Bandi et al. 1998). All *Wolbachia* strains are currently classified into a single species, *Wolbachia pipientis* (Hertig 2009; Lo et al. 2007). This species, however, on the basis of single gene and multi-locus phylogenies (Werren et al. 1995; Bordenstein et al. 2009), has been divided into 16 monophyletic supergroups, labelled A to Q (although supergroup G is possibly an artefact) (Augustinos et al. 2011; Glowska et al. 2015; Lo et al. 2002). The monophyly of the most-studied supergroups has recently been confirmed, albeit only on a limited number of strains, using whole-genome phylogenetic approaches (Comandatore et al. 2013; Gerth et al. 2014).

The *Wolbachia* supergroups are associated with distinct sets of hosts in Arthropoda and Nematoda. The nature of the association between *Wolbachia* strains and their hosts also varies greatly. The symbiosis between C and D supergroup strains and their filarial nematode hosts presents features associated with mutualism, including 100% prevalence (Taylor et al. 2005), strict vertical inheritance (Bandi et al. 1998;

Casiraghi et al. 2001), and metabolic integration (Gill et al. 2014; Hosokawa et al. 2010; Darby et al. 2012) . In contrast, A and B supergroup strains, infecting arthropod hosts, have <100% prevalence, display evidence of rampant lateral transfer and induce a variety of reproductive manipulation phenotypes, including cytoplasmic incompatibility (CI), parthenogenesis, killing of male embryos and feminization of genetic males (Werren et al. 2008; Cordaux et al. 2011) pointing to a more parasitic role. However, within supergroups A and B, evidence has been found for more mutualistic effects. For example, these *wolbachiae* may provide an advantage to their hosts by modulating immune responses that protect against viral infections (Kambris et al. 2009; Kollenberg et al. 2014).

Is *Wolbachia pipientis* one species? While they are all intracytoplasmic symbionts, *Wolbachia* strains occupy distinct ecological niches in hosts that are phylogenetically extremely distant. Analyses of supergroup A and B strains coinfecting the same *Drosophila simulans* host showed a lack of genetic exchange between supergroups (Ellegaard et al. 2013). Contrasting biological relationships, host-range limitation and genomic distinctness has led to suggestions that *Wolbachia* supergroups should be considered separate species (Pfarr et al. 2007). If the different supergroups are evolutionarily independent lineages, distinct at the species level or evolving towards being distinct, we can expect their genomes to present a mix of shared and derived features consequent on the different selection pressures they have experienced in their different environments.

To better understand the genomic distinctiveness of the *Wolbachia* supergroups and

further explore whether they are 'irreversibly' separated, we explored the biology of 26 representative, currently available *Wolbachia* genomes. Using measures of gene presence/absence, codon usage, GC skew curve, genome rearrangements, transposable elements, horizontal gene transfer (HGT) and intragenic recombination, we found that the *Wolbachia* supergroups, and in particular those from supergroup C, have distinct patterns of genome evolution, likely the result of extensive periods of independent evolution.

## Materials and methods

### Dataset

We analysed the genomes of 26 strains of *Wolbachia* (*wAlbB*, *wAna*, *wBm*, *wBol1*, *wCitri*, *wCle*, *wDi*, *wFol*, *wHa*, *wLs*, *wMel*, *wMen*, *wNo*, *wOc*, *wOo*, *wOv*, *wOvC*, *wPipJHB*, *wPipPel*, *wRi*, *wSim*, *wSuz*, *wUni*, *wVitB*, *wWb*, *wWil*; *Wolbachia* strains are named by two to six letter names derived from the specific names of their hosts) and three outgroups (the alphaproteobacterial pathogens *Anaplasma centrale*, *Ace*, and *Ehrlichia chauffensis*, *Ech*; and the gammaproteobacterial aphid mutualist symbiont *Buchnera aphidicola*, *BAp*). Sources are given in Table 1. The *Wolbachia* genomes were divided into high-quality (10 or fewer contigs or scaffolds) and draft-quality (more than 10 contigs or scaffolds) assemblies (see Table 1).

## **Synteny conservation**

A multi-genome alignment of the high-quality *Wolbachia* strain genomes *wBm*, *wCle*, *wDi*, *wLs*, *wOo*, *wOvC*, *wMel* and *wPipPel* was generated with Mauve (Darling et al. 2004). Conservation of synteny was visualised using the genoPlotR R library.

## **Origin of replication**

The position of the origin of replication (ORI) of *wMel* and *wBm* have been previously inferred by Ioannidis and colleagues (Ioannidis et al. 2007). On the basis of these *wMel* and *wBm* ORI positions, we inferred the position of the ORI in the high-quality genomes of *wCle*, *wDi*, *wHa*, *wNo*, *wOo*, *wOvC*, *wPipPel*, and *wRd* by aligning them to *wBm* and *wMel* genomes with progressiveMauve (Darling et al. 2004), and identifying the conserved regions around the *wMel* and *wBm* ORIs. The position of the terminus of replication (TER) was set at halfway through the genome, starting from the ORI position.

Wolbachia strains (short name)	Hosts	Sources	Scaffolds numbers	Supergroups
wMel	<i>Drosophila melanogaster</i>	PRJNA57851		1 A
wRi	<i>Drosophila simulans</i>	PRJNA13364		1 A
wHa	<i>Drosophila simulans</i>	PRJNA198768		1 A
wPipPel	<i>Culex quinquefasciatus</i>	PRJNA61645		1 B
wNo	<i>Drosophila simulans</i>	PRJNA198767		1 B
wOo	<i>Onchocerca ochengi</i>	PRJEA171829		1 C
wBm	<i>Brugia malayi</i>	PRJNA58107		1 D
wOvC	<i>Onchocerca volvulus</i>	PRJNA224116		1 C
wCle	<i>Cimex lectularius</i>	Nikoh et al. 2014		1 F
wDi	<i>Diriofilaria immitis</i>	http://diriofilaria.nematod.es		2 C
wLs	<i>Litomosoides sigmodontis</i>	http://litomosoides.nematod.es		10 D
wPipJHB	<i>Culex quinquefasciatus</i>	PRJNA55557		21 B
wFol	<i>Folsomia candida</i>	Gerth et al. 2014		98 E
wCiri	<i>Diaphorina citri</i>	PRJNA199833		104 B
wSuz	<i>Drosophila suzukii</i>	PRJEB596		110 A
wOe	<i>Osmia caerulescens</i>	Gerth et al. 2014		140 F
wBoll	<i>Hypolimnas bolina</i>	PRJNA199705		144 B
wAlbB	<i>Aedes albopictus</i>	CAGB01000001-165		165 B
wUni	<i>Muscidifurax uniraptor</i>	PRJNA213628		256 A
wWill	<i>Drosophila willistoni</i>	PRJNA16739		260 A
wOv	<i>Onchocerca volvulus</i>	PRJNA43537		341 C
wAna	<i>Drosophila ananassae</i>	PRJNA13365		464 A
wVitB	<i>Nasonia vitripennis</i>	PRJNA74529		523 B
wSim	<i>Drosophila simulans</i>	PRJNA13364		629 A
wWb	<i>Wuchereria bancrofti</i>	PRJNA43539		763 D
wMen	<i>Mengenilla moldrzyki</i>	Contigs selected from PRJNA72521 (Michael Gerth personal communication)		1664 F
<b>Outgroup strains (short name)</b>				
	<b>Strain names</b>	<b>Sources</b>		
Ace	<i>Anaplasma centrale</i> str. Israel	PRJNA42155		1 ---
Ech	<i>Ehrlichia chaffeensis</i> str. Arkansas	PRJNA57933		1 ---
Bap	<i>Buchnera aphidicola</i> str. Bp ( <i>Baizongia pistaciae</i> )	NC_004545.1		1 ---

**Table 1. Genomes analysed in this report.**

List of the genomes included in this study. For each genome, information about the strain, the corresponding host and the

## GC skew

The strand bias of guanine versus cytosine nucleotides (GC skew) along the wBm, wCle, wDi, wHa, wMel, wNo, wOo, wOvC, wPipPel, wRi and Ace genomes was studied with a custom Perl script, reconstructing the CG cumulative curves with a window size of 10,000 and step size of 100. For wBm, wCle, wHa, wMel, wNo, wOo, wOvC, wPipPel, wRi and Ace, the potential effect of genomic rearrangements on the current GC skew distribution was evaluated by aligning the genome against the wDi genome with progressiveMauve, sorting and orienting the synteny blocks according to the wDi order, and calculating GC skew as above for these re-oriented genomes. For each window, the GC skew for the reoriented genomes was compared to the GC

skew of the wDi genome as

$| \text{GCSkew}_{\text{wDi}} - \text{GCSkew}_{\text{strain}} | / \sum \text{GCSkew}_{\text{wDi}}$ . The pairwise Wilcoxon test with Bonferroni post-hoc correction was applied to compare the average index values obtained for the wBm, wCle, wHa, wMel, wNo, wOo, wOvC, wPipPel, wRi and Ace genomes.

The effect of substitutional bias on the guanine and cytosine distribution in the wBm, wCle, wDi, wLs, wMel, wOo, wOvC, wPipPel and wRi genomes was evaluated on a set of 551 single-copy, orthologous genes identified with OrthoMCL (Li et al. 2003) and Perl scripts. Gene alignments were generated using Muscle (Edgar 2004), and a GC composition bias index was calculated for third position residues as follows:

nucleotide difference towards G / nucleotide difference towards C

where nucleotide difference towards G indicates all mutations

$A_{\text{w1}} \rightarrow G_{\text{w2}}, T_{\text{w1}} \rightarrow G_{\text{w2}}, C_{\text{w1}} \rightarrow G_{\text{w2}}$  in *Wolbachia*2 (W2) in respect to *Wolbachia*1 (W1).

The single-copy orthologues were subdivided into two groups: a “right” group including the genes placed after ORI and before TER, and a “left” group including the remaining genes. Student's *t*-test was used to compare the GC composition bias index of the “right” and “left” genes, considering each *Wolbachia* strain pair.

## Gene loss

Gene loss events were inferred from analysis of gene presence/absence patterns mapped onto the monophyletic *Wolbachia* supergroups and monophyletic clusters of supergroups. The gene presence/absence pattern of each of 12 high-quality

*Wolbachia* genomes (*wOo*, *wOvC*, *wDi*, *wLs*, *wBm*, *wCle*, *wMel*, *wHa*, *wRi*, *wPip Pel*, *wPipJHB* and *wNo*) was determined using OrthoMCL and an in-house Perl pipeline. We only considered genes lost by all members of a supergroup or cluster, and annotated candidates using BLAST against the COG database (with  $10^{-5}$  e-value threshold).

### **Codon usage**

Codon preference matrices were derived for each of the *wAlbB*, *wAna*, *wBm*, *wBol1*, *wCitri*, *wCle*, *wDi*, *wFol*, *wHa*, *wLs*, *wMel*, *wMen*, *wNo*, *wOc*, *wOo*, *wOv*, *wOvC*, *wPipJHB*, *wPipPel*, *wRi*, *wSim*, *wSuz*, *wUni*, *wVitB*, *wWb* and *wWill* *Wolbachia* genomes, and Relative Synonymous Codon Usage (RSCU) indices were calculated with CodonW (<http://codonw.sourceforge.net/>). RSCU values were subjected to hierarchical clustering analysis (with the “complete” method) and Principal Component Analysis (PCA) using R software. On the basis of the clustering result, codons were categorised as “high”, “medium” and “low” frequency groups.

The AT-richness of codons in the three groups was compared with the Wilcox test with post-hoc Bonferroni correction. For each complete *Wolbachia* genome available in the database (*wBm*, *wCle*, *wDi*, *wHa*, *wMel*, *wNo*, *wOo*, *wOvC*, *wPipPel* and *wRi*) the leading and the lagging strands were determined on the basis of the position of the ORI. CodonW was used to perform a Correspondence analysis (COA) on RSCU indexes calculated on leading and lagging genes.



## **Transposable elements**

Insertion sequences (IS) and group II introns were identified and annotated in *wDi* (C supergroup), *wLs* (D supergroup) and *wCle* (F supergroup). Group II introns were identified following the methods of Leclercq and colleagues (Leclercq et al. 2011). IS elements were identified using ISSaga (Varani et al. 2011), followed by manual curation of ISSaga output files. For *wLs*, most ISSaga hits were short and often formed groups of 2-4 hits located next to each other. This is typical of pseudogenized and degraded IS elements. We attributed two consecutive hits to the same or to distinct IS copies using the following rules:

(1) IS family: if the two hits belong to different IS families, they belong to distinct copies. Otherwise: go to criterion (2),

(2) Orientation: if the two hits are in opposite orientation, they belong to distinct copies. Otherwise: go to criterion (3),

(3) Physical distance: if distance between the two hits is >300 bp, they belong to distinct copies. Otherwise: they belong to the same copy.

## **Lateral gene transfer and recombination**

Orthologous genes shared among *wAlbB*, *wBm*, *wCle*, *wDi*, *wFol*, *wHa*, *wLs*, *wMel*, *wNo*, *wOc*, *wOo*, *wOvC*, *wPipJHB*, *wRi*, *Ech* and *Ace* genomes were identified with OrthoMCL. For each orthologous gene, the corresponding nucleotide sequences were retrieved and aligned based on their amino acid translation. Phylogenetic analyses were performed on each single gene alignment and on a global

concatenate of all genes, using RAxML (Stamatakis 2014) with the CAT-GTR model and 1,000 pseudo-bootstraps parameters. Gene trees with minimum bootstrap support of 70% on each node were selected and manually compared to the global concatenate tree. Each gene was examined for evidence of intragenic recombination using Phipack (Bruen et al. 2006) and GENECONV (Sawyer 1989), following the approach of Ellegaard and colleagues (Ellegaard et al. 2013).

## Results

We are interested in the evolutionary dynamics of *Wolbachia*, an important genus of intracellular bacteria. Here we explore the signatures in 26 *Wolbachia* genomes from supergroups A to F, including codon usage, synteny conservation, transposable element content, distribution pattern of GC composition and origin of replication, to derive hypotheses of their contrasting evolutionary histories.

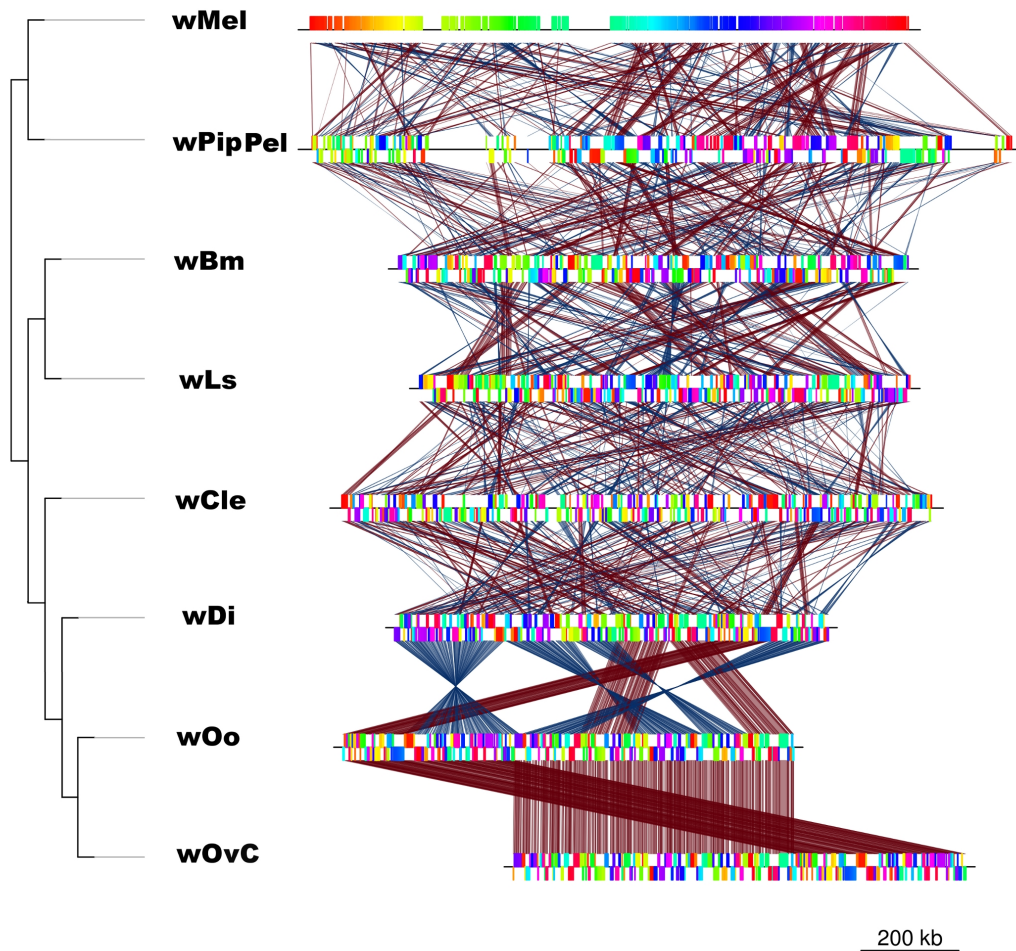
### Synteny

*Wolbachia* genomes have been reported to have undergone extensive rearrangement in comparison with other *Rickettsiales* (Klasson et al. 2008). These analyses were based on limited numbers of strains, and in particular only included *wBm* from supergroup D as a representative of *Wolbachia* infecting filarial nematodes. We analysed five high-quality genome assemblies from filarial *Wolbachia* strains from both C and D supergroups (Foster et al. 2005; Darby et al.

2012; Godel et al. 2012; Comandatore et al. 2013) alongside high quality genomes from arthropod hosts in supergroup F (*wCle*), A (*wMel*) and B (*wPipPel*). An alignment of these high-quality genomes revealed conservation of synteny among the *wDi*, *wOo* and *wOvC* genomes (all in supergroup C) (Figure 1). This conservation contrasts with very low levels of synteny within and between the other supergroups.

### **Transposable elements**

Synteny breakage and recombination is often associated with repeats and transposable elements. We therefore screened the *Wolbachia* genomes for classes of transposable element (Supplementary Table 1 and Figure 2). We found no group II introns in the *wDi* (C supergroup) and *wLs* (D supergroup) genomes. However, insertion sequences (IS) had a striking, disjunct pattern of presence. While *wDi* had only a single IS (similar to ISWpi16), *wLs* contained 210 IS copies. Supergroup A and B arthropod *Wolbachia* genomes also have many IS elements (Cerveau et al. 2011), albeit fewer than *wLs*. IS elements cover nearly 12% of the *wLs* genome, a higher percentage than in any other *Wolbachia* genome sequenced to date. Despite their high copy number, all *wLs* IS copies appear to be degraded and there is no apparent “live” transpositional activity. Remarkably, 97% of the *wLs* IS copies (204/210) belong to a single IS type (ISWpi10). The 6 remaining copies belong to ISWpi5. Interestingly, the genome of *wCle* (F supergroup) is characterized by a high density (10%) and diversity (11 different types) of IS elements and the presence of group II introns (Supplementary Table S1).



**Figure 1:** Synteny conservation in supergroup C Wolbachia A graphic representation of Mauve analysis output is shown on the right. Conserved synthetic blocks detected by Mauve are connected by coloured lines: red lines display transpositions, while blue lines display inversions. The known phylogenetic relationships among the Wolbachia strains are shown on the left.

In the D supergroup genomes, no IS copy was found to be inserted at an orthologous site, despite the high number of IS copies in these genomes. By contrast, in supergroup C, the single IS copy found in *wDi* is orthologous to the

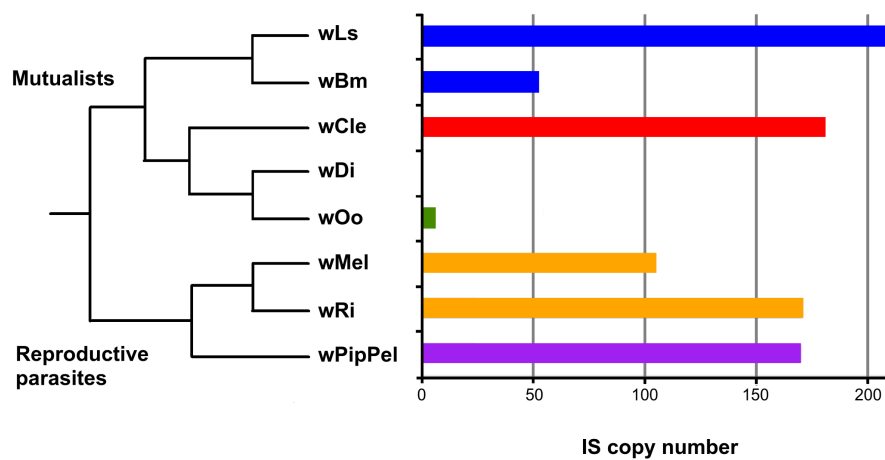
ISWpi16 copy found in *wOo*.

### **GC skew and ORI determination**

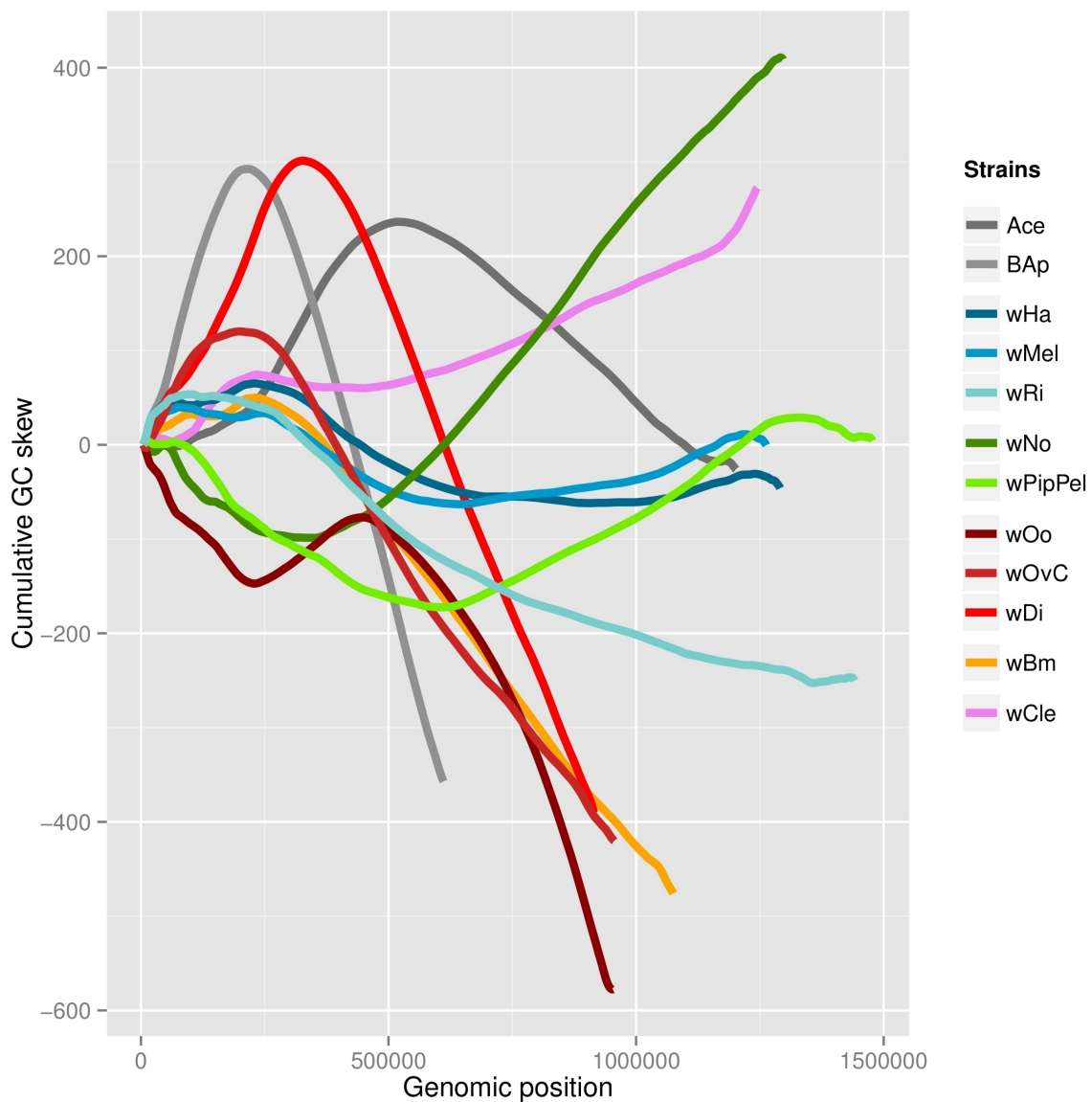
Another feature described as characteristic of arthropod *Wolbachia* genomes is the absence of strong GC skew (Klasson et al. 2008), contrasting with the pattern commonly observed in most free-living bacteria, and in other endosymbiotic bacteria such as *B. aphidicola* (Lobry 1996; Klasson & Andersson 2006). The GC skew of the nine completely sequenced *Wolbachia* genomes and two outgroups (*wBm*, *wCle*, *wDi*, *wMel*, *wOo*, *wOvC*, *wPipPel*, *wHa*, *wNo*, *wRi*, and the outgroups *Bap* and *Ace*) were analysed. In agreement with previous analyses on a smaller dataset (Klasson et al. 2008), most *Wolbachia* genomes do not present any genome-wide pattern of GC skew (Figure 3). However, the *wDi* genome has a unique GC skew pattern (Figure 3), comparable to those typically observed in free-living bacteria, in pathogenic Rickettsiales (here *Ace*) and in *B. aphidicola*.

This pattern of GC skew in *wDi* could have originated uniquely in *wDi* or could be an ancestral feature of *Wolbachia*, lost by most lineages. To test the hypothesis that the *wDi* GC skew pattern is ancestral, we evaluated whether its absence in the *wBm*, *wCle*, *wMel*, *wOo*, *wOvC*, *wPipPel*, *wHa*, *wNo*, and *wRi* genomes could have been caused by genome rearrangement. We reordered each genome to conform to the *wDi* gene order and recalculated the GC skew on the “pseudo-ancestral” genome (Figure 4 and Supplementary Figure 1). While rearrangement of supergroup A, B, C and F genomes did not reveal any hidden GC skew pattern, in the rearranged *wOo*

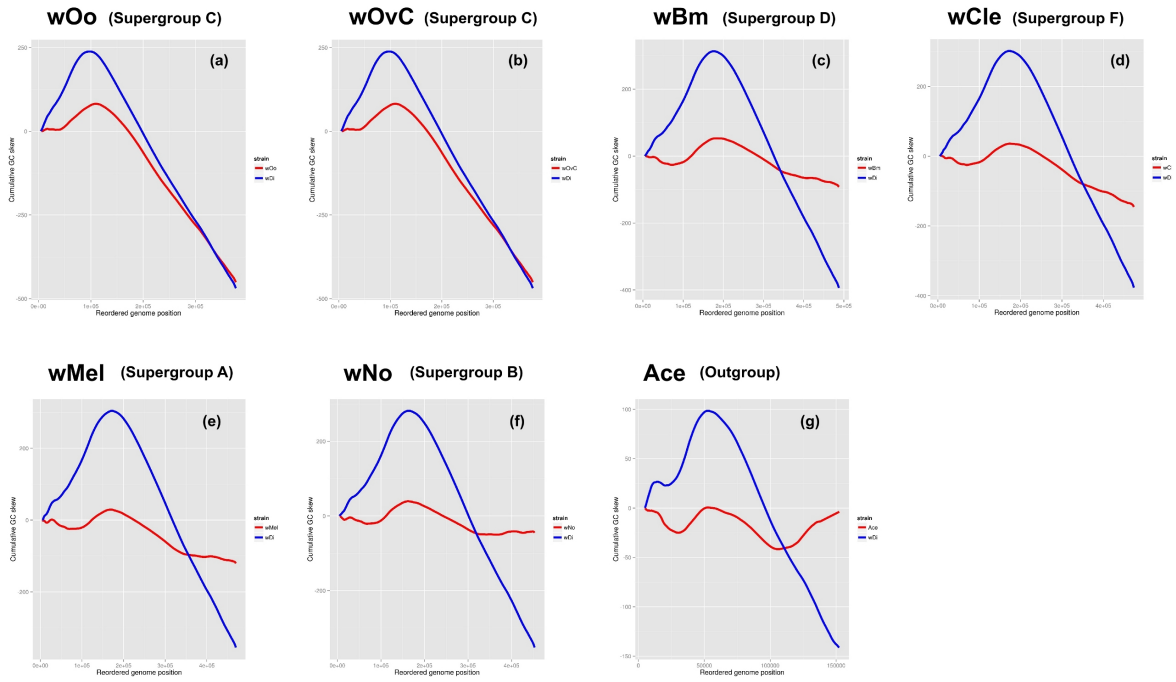
and *wOvC* genomes we observed a trend similar to that of *wDi*. The average differences between the *wDi* native GC skew curve and both rearranged *wOo* and *wOvC* curves were significantly lower than the native-native comparisons ( $p$ -value < 0.001). No better fit was observed between native *wDi* and the rearranged *wBm*, *wCle*, *wMel*, *wPipPel*, *wHa*, *wNo* or *wRi* GC skew curves (Supplementary Figure 2).



**Figure 2:** Insertion sequences in Wolbachia genomes. Results of Insertion Sequence (IS) analyses performed on the *wLs*, *wBm*, *wDi*, *wOo*, *wCle*, *wMel*, *wRi* and *wPipPel* Wolbachia strains are displayed as a histogram (right) showing IS quantification for each Wolbachia strain. The known phylogenetic relationships among the Wolbachia strains are shown on the left. For each Wolbachia strain the corresponding supergroup is colour-coded: orange, A; violet, B; green, C; blue, D; black, E; and red, F.



**Figure 3:** Cumulative GC skew curves of Wolbachia and compared genomes. GC skew was calculated with window size of 10,000 nucleotides and step size of 100 nucleotides. The curves for *Buchnera aphidicola* and *Anaplasma centrale* are coloured in grey hues, while the curves for Wolbachia strains are coloured by supergroup: A, blue hues; B, green hues; C, red hues; D, orange; and F, pink.



**Figure 4:** Cumulative GC skew curves of *Wolbachia* genomes reordered based on the *wDi* genome Cumulative GC skew curves of six reoriented *Wolbachia* genomes (in red) are compared to the *wDi* genome (in blue). Genomes were reordered on the basis on the *wDi* gene order based on a progressiveMauve genome alignment.

Based on the GC skew analysis presented above, the occurrence of genome rearrangements could explain the difference in GC distribution between *wDi* and the other two C supergroup *Wolbachia* genomes (*wOo* and *wOvC*), but cannot explain the differences between *wDi* and the genomes of strains belonging to other supergroups. We thus hypothesized that, during the evolution of the C supergroup, a mutational bias led to the asymmetric distribution of GC observed in *wDi* genome. For example, a strong mutational bias between the two genome halves delimited by the origin (ORI) and terminus of replication (TER) (here called “right” and “left” genome halves), could have led to the GC distribution pattern observed in *wDi* (Figure 3).



GC skew is thought to arise from biased substitution processes driven by the replicational structure of the circular chromosome. The effect of this process were investigated by first inferring the position of the ORI in complete *Wolbachia* genomes (*wCle*, *wDi*, *wHa*, *wNo*, *wOo*, *wOvC*, *wPipPel* and *wRi*) on the basis of the *wMel* and *wBm* ORI positions ((Ioannidis et al. 2007) - Supplementary Table S2), and then exploring GC substitution bias in genes to the right and left of the ORI.

No left-right substitution bias was observed between the C supergroups *wOo*, *wOvC* and *wDi* genomes, but a highly significant mutational bias (Student's *t*-test *p*-value < 0.001) was apparent when comparing any these genomes to any of the other *Wolbachia* genomes. No left-right substitution bias was observed in any other genome pairs, suggesting that it is likely to have originated during C supergroup evolution (Table 2).

### **Gene loss**

*Wolbachia* genomes differ in size from ~1.5 Mb to 0.9 Mb. These size differences could arise from different ratios of loss/acquisition of genomic material (including transposable elements and phages). We identified putative gene losses in each *Wolbachia* supergroup, and in clusters of supergroups, based on orthologue clustering (Table 3). The most striking result is the extremely limited number of genes inferred to have been lost on the branches leading to the arthropod-infecting supergroups A and B: a single gene was lost by the ancestor of supergroup A

(COG2142). All other putative losses were in the ancestors of supergroups C, D and F, and in the shared ancestors of (C+F), and (C+D+F). Within the losses inferred in the ancestor of (C+D+F), ten of the thirteen orthologous groups are phage-associated proteins, and likely reflect the loss of WO bacteriophage from these wolbachiae. We note that the fossil *Wolbachia* in the nuclear genome of the nematode *Dictyocaulus viviparus* contains phage fragments (Koutsovoulos et al. 2014), suggesting that this loss may have been convergent in D and (C+F).

		W2								
		wMel	wRi	wPip_PEL	wOo	wOvC	wDi	wCle	wLs	wBm
W1	wMel	-	-	-	+	+	+	-	-	-
	wRi	-	-	-	+	+	+	-	-	-
	wPipPel	-	-	-	+	+	+	-	-	-
	wOo	-	-	-	-	-	+	-	-	-
	wOvC	-	-	-	-	-	+	-	-	-
	wDi	-	-	-	-	-	-	-	-	-
	wCle	-	-	-	+	+	+	-	-	-
	wLs	-	-	-	+	+	+	-	-	-
	wBm	-	-	-	+	+	+	-	-	-

**Table 2. Mutational bias in Wolbachia genomes.**

For each pair of Wolbachia genomes, the average mutational bias measured on the genes homologous to “right” wDi genes, is compared to the mutation bias measured on the genes homologous to “left” wDi genes. Comparisons were performed with Student's t-test, with Bonferroni correction. Significantly different comparisons (p-value < 0.001) are reported with “+”; not significant with “-”.

### Codon usage

All available *Wolbachia* genomes have a strong compositional bias towards AT, a characteristic common in intracellular bacteria (Klasson & Andersson 2006).

*Wolbachia* genomes encode a limited number of tRNAs, typically 32. We thus

explored how these two features might influence *Wolbachia* codon usage. The codon usage of all 26 *Wolbachia* genomes was compared, calculating the Relative Synonymous Codon Usage (RSCU) index, hierarchical clustering and PCA analysis.

Common ancestor	COG code	COG Annotation
A	[C] COG2142	Succinate dehydrogenase, hydrophobic anchor subunit
C	[H] COG0307	Riboflavin synthase alpha chain
C	[J] COG0030	Dimethyladenosine transferase (rRNA methylation)
C	[L] COG0468	RecA/RadA recombinase
C	[MG] COG0702	Predicted nucleoside-diphosphate-sugar epimerases
C	[NU] COG1450	Type II secretory pathway, component PulD
C	[P] COG2193	Bacterioferritin (cytochrome b1)
C	[R] COG1268	Uncharacterized conserved protein
C	[S] COG5590	Uncharacterized conserved protein
CDF	[L] COG3344	Retron-type reverse transcriptase
CDF	[M] COG0818	Diacylglycerol kinase
CDF	[OU] COG0616	Periplasmic serine proteases (ClpP class)
CDF	[R] COG0220	Predicted S-adenosylmethionine-dependent methyltransferase
CDF	[R] COG3497	Phage tail sheath protein FI
CDF	[R] COG3498	Phage tail tube protein FII
CDF	[R] COG3499	Phage protein U
CDF	[R] COG3628	Phage baseplate assembly protein W
CDF	[R] COG3948	Phage-related baseplate assembly protein
CDF	[R] COG4540	Phage P2 baseplate assembly protein gpV
CDF	[R] COG5004	P2-like prophage tail protein X
CDF	[R] COG5511	Bacteriophage capsid protein
CDF	[R] COG5525	Bacteriophage tail assembly protein
CDF	[S] COG5283	Phage-related tail protein
CF	[L] COG3335	Transposase and inactivated derivatives
D	[C] COG0371	Glycerol dehydrogenase and related enzymes
D	[H] COG0262	Dihydrofolate reductase
D	[H] COG0294	Dihydropteroate synthase and related enzymes

**Table 3.** Genes lost during *Wolbachia* evolution.

Both hierarchical clustering and PCA analysis grouped *Wolbachia* strains largely coherently by supergroup, with the exception of supergroup D wLs, which clustered within supergroup C strains (Figure 5 and Supplementary Figure 3). Codons were grouped into three main clusters, characterized by low ( $n = 31$ ), medium ( $n = 29$ ) or

high ( $n = 4$ ) relative frequency. The 29 codons belonging to the medium cluster have a higher AT richness than the 31 low frequency codons (Wilcox test with Bonferroni correction  $p$ -value  $< 0.01$ ). No significant correspondence (McNemar test) was found between the tRNA anticodons found in the genomes and codon usage frequency.

The codon usages of genes transcribed on the leading and lagging strands of the more complete genomes of *wBm*, *wCle*, *wDi*, *wHa*, *wMel*, *wNo*, *wOo*, *wOvC*, *wPipPel* and *wRi* were compared to identify any effects of genome position (Figure 6 and Supplementary Figure 4). *wDi* and *wOo* were the only *Wolbachia* strains analysed that showed two different codon usages for genes on the leading and lagging genome strands.

### **Lateral gene transfer and recombination**

We identified 467 single-copy and globally shared orthologous genes, inferring phylogenetic trees for each gene and for a concatenated alignment. Only 20 of the gene trees showed a minimum ML bootstrap support of 70, and 14 of these had a topology inconsistent with the global concatenate, perhaps indicating lateral rather than vertical inheritance.

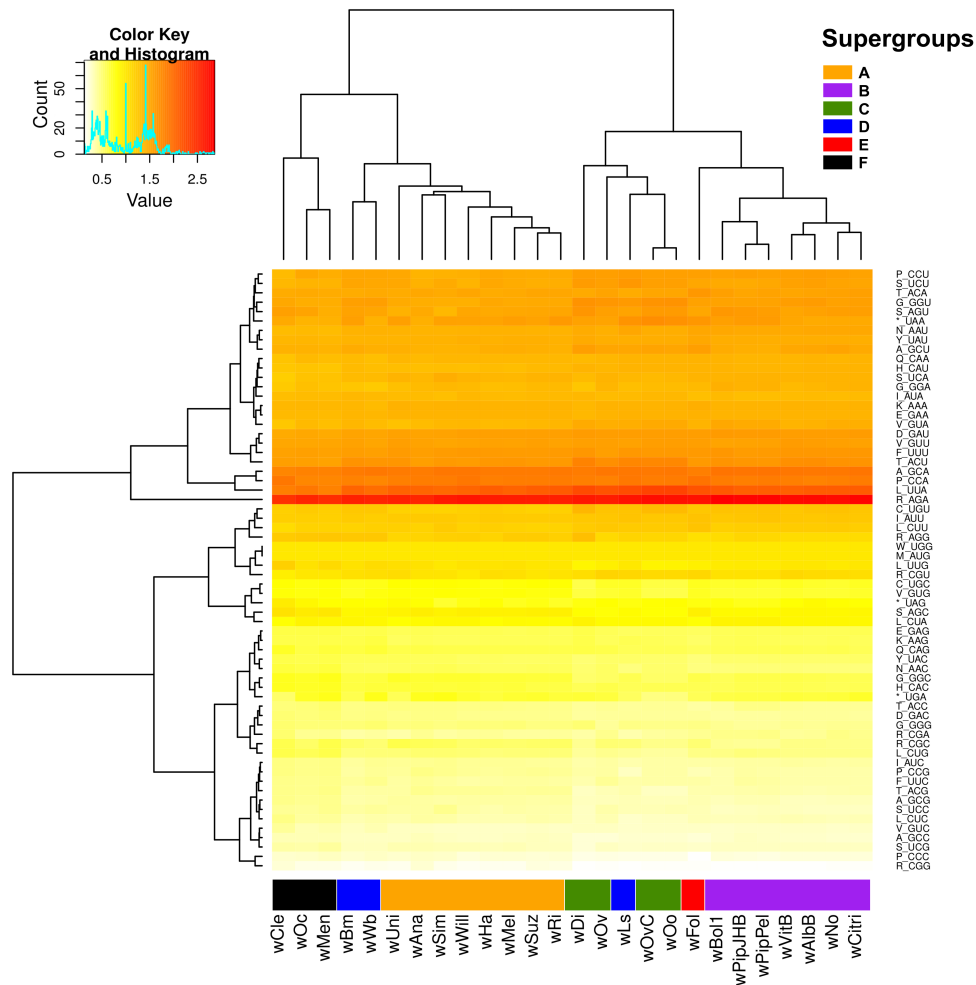
HGT and recombination signals among 14 *Wolbachia* strains (belonging to the A-F supergroups) were studied, with the exclusion of the pair A and B, previously described as highly recombined (Ellegaard et al. 2013). We found evidence for HGT between the ancestors of the current *Wolbachia* lineages and supergroups (Table 4).

Four of the highly supported gene trees present a signal of HGT between the common ancestor of the monophylum ((C,F),D) and the supergroup B ancestor. One tree suggests the presence of HGT between ((C,F),D) and supergroup A ancestors. HGT between the supergroup D ancestor and the supergroup F ancestor was suggested by three gene trees. Two trees suggested HGTs between the supergroup D ancestor and an external bacterial lineage, and one suggests that HGT occurred between the supergroup D ancestor and the common ancestor of supergroups A and B.

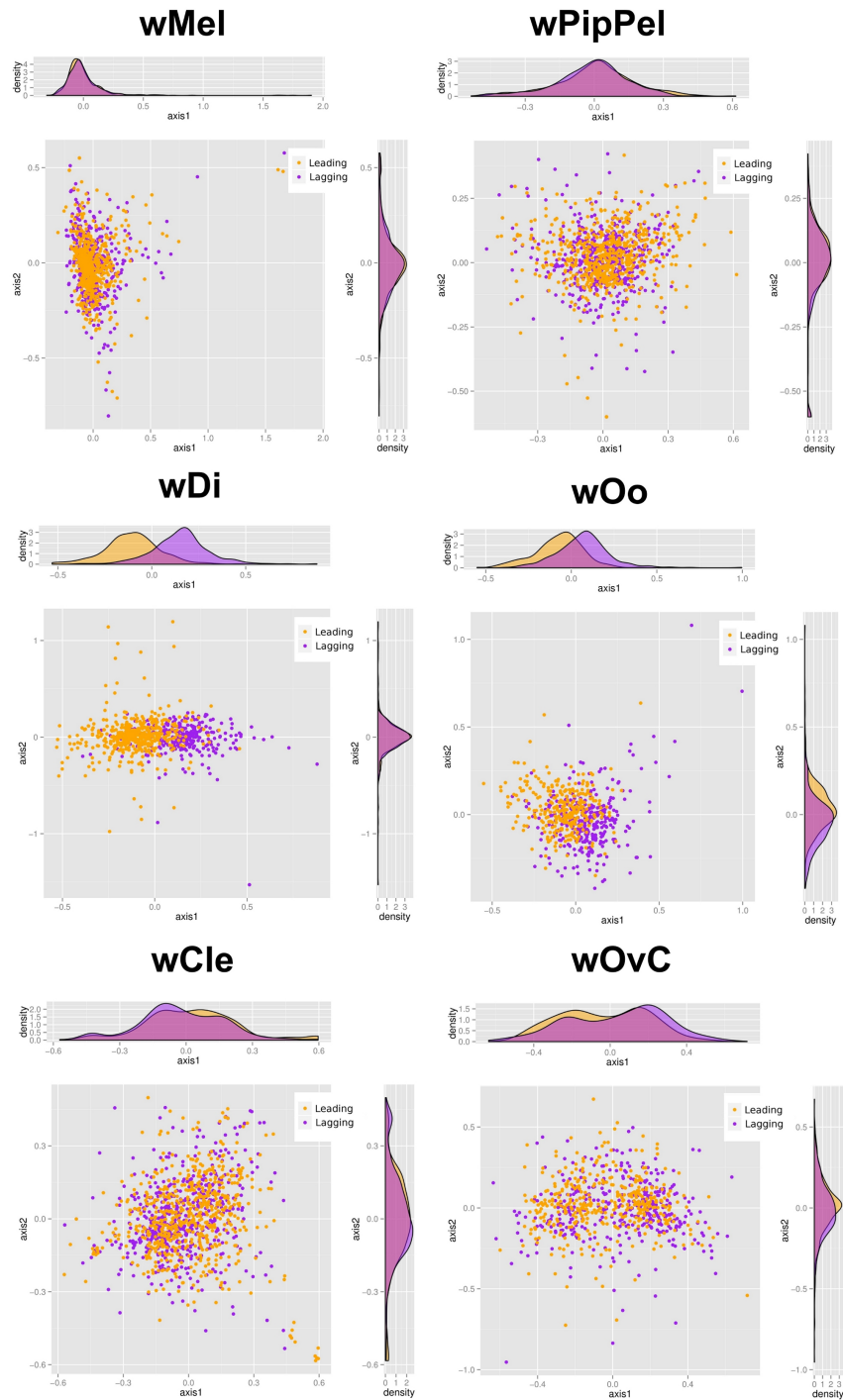
Intra-gene recombination analyses identified 33 recombination events (Table 4), of which 26 were intra-supergroup events, equally divided between supergroups A and B. The five inter-supergroup recombinations were all between supergroups A and B.

Recombination type	Supergroups involved	COG code	COG annotation
Partially transferred	A-CDF	[P] COG0475	Kef-type K <sup>+</sup> transport systems, membrane components
Partially transferred	B-CDF	[C] COG1034	NADH dehydrogenase/NADH:ubiquinone oxidoreductase 75 kD subunit (chain G)
Partially transferred	B-CDF	[J] COG1185	Polyribonucleotide nucleotidyltransferase (polynucleotide phosphorylase)
Partially transferred	B-CDF	[L] COG2812	DNA polymerase III, gamma/tau subunits
Partially transferred	B-CDF	[L] COG0749	DNA polymerase I - 3'-5' exonuclease and polymerase domains
Partially transferred	B-F	[L] COG0592	DNA polymerase sliding clamp subunit (PCNA homolog)
Partially transferred	B-F	[F] COG0519	GMP synthase, PP-ATPase domain/subunit
Partially transferred	C-D	[U] COG0541	Signal recognition particle GTPase
Partially transferred	D-AB	[R] COG0496	Predicted acid phosphatase
Partially transferred	D-F	[L] COG0188	Type IIA topoisomerase (DNA gyrase/topo II, topoisomerase IV), A subunit
Partially transferred	D-F	[J] COG0482	Predicted tRNA(5-methylaminomethyl-2-thiouridylylate) methyltransferase, contains the PP-loop ATPase domain
Partially transferred	D-F	[O] COG0760	Parvulin-like peptidyl-prolyl isomerase
Partially transferred	D-OUT	[U] COG3704	Type IV secretory pathway, VirB6 components
Partially transferred	D-OUT	[O] COG1219	ATP-dependent protease Clp, ATPase subunit
Entirely transferred	A	[F] COG0138	AICAR transformylase/IMP cyclohydrolase PurH (only IMP cyclohydrolase domain in AfuI)
Entirely transferred	A	[C] COG1249	Pyruvate/2-oxoglutarate dehydrogenase complex, dihydroliipoamide dehydrogenase (E3) component, and related enzymes
Entirely transferred	A	[C] COG1249	Pyruvate/2-oxoglutarate dehydrogenase complex, dihydroliipoamide dehydrogenase (E3) component, and related enzymes
Entirely transferred	A	[J] COG0008	Glutamyl- and glutaminyl-tRNA synthetases
Entirely transferred	A	[J] COG0060	Isoleucyl-tRNA synthetase
Entirely transferred	A	[J] COG1185	Polyribonucleotide nucleotidyltransferase (polynucleotide phosphorylase)
Entirely transferred	A	[J] COG0495	Leucyl-tRNA synthetase
Entirely transferred	A	[J] COG0495	Leucyl-tRNA synthetase
Entirely transferred	A	[J] COG0751	Glycyl-tRNA synthetase, beta subunit
Entirely transferred	A	[J] COG0751	Glycyl-tRNA synthetase, beta subunit
Entirely transferred	A	[G] COG0061	Predicted sugar kinase
Entirely transferred	A	[U] COG3504	Type IV secretory pathway, VirB9 components
Entirely transferred	A	[D] COG0849	Actin-like ATPase involved in cell division
Entirely transferred	A-B	[F] COG0034	Glutamine phosphoribosylpyrophosphate amidotransferase
Entirely transferred	A-B	[U] COG0706	Preprotein translocase subunit YidC
Entirely transferred	A-B	[C] COG0045	Succinyl-CoA synthetase, beta subunit
Entirely transferred	A-B	[J] COG0525	Valyl-tRNA synthetase
Entirely transferred	A-B	[O] COG1138	Cytochrome c biogenesis factor
Entirely transferred	B	[CP] COG0651	Formate hydrogenlyase subunit 3/Multisubunit Na <sup>+</sup> /H <sup>+</sup> antiporter, MnhD subunit
Entirely transferred	B	[U] COG0541	Signal recognition particle GTPase
Entirely transferred	B	[U] COG3736	Type IV secretory pathway, component VirB8
Entirely transferred	B	[J] COG2913	Small protein A (tmRNA-binding)
Entirely transferred	B	[U] COG0706	Preprotein translocase subunit YidC
Entirely transferred	B	[L] COG3893	Inactivated superfamily I helicase
Entirely transferred	B	[C] COG0567	2-oxoglutarate dehydrogenase complex, dehydrogenase (E1) component, and related enzymes
Entirely transferred	B	[C] COG0045	Succinyl-CoA synthetase, beta subunit
Entirely transferred	B	[J] COG0173	Aspartyl-tRNA synthetase
Entirely transferred	B	[C] COG1071	Pyruvate/2-oxoglutarate dehydrogenase complex, dehydrogenase (E1) component, eukaryotic type, alpha subunit
Entirely transferred	B	[J] COG0525	Valyl-tRNA synthetase
Entirely transferred	B	[J] COG0525	Valyl-tRNA synthetase
Entirely transferred	B	[O] COG0465	ATP-dependent Zn proteases
Entirely transferred	C-B	[L] COG3893	Inactivated superfamily I helicase
Entirely transferred	F-OUT	[U] COG3704	Type IV secretory pathway, VirB6 components

**Table 4.** Genes whose sequences originated, entirely or partially, by horizontal transfer among ancestors of *Wolbachia* lineages



**Figure 5:** RCSU index clustering across Wolbachia strains The heatmap shows hierarchical clustering of codons and genomes based on the Relative Synonymous Codon Usage index values. The RSCU values are reported with a colour range from white to red in the heatmap. The dendrogram reported on the left of the heatmap represents the result of the clustering analysis on the codons. The dendrogram above the heatmap reports the result of the clustering analysis performed on the organisms. For each Wolbachia strain the corresponding supergroup is reported as follow: orange, A; violet, B; green, C; blue, D; black, E; and red, F.



**Figure 6:** Genes on leading and lagging strands Coordinates Analysis (COA) of genes located on the leading and lagging strands in four analysed Wolbachia genomes. Leading genes are shown in orange, lagging genes in violet. The distribution graphs corresponding to the X and Y axes are reported respectively above and on the right of each panel. This same analysis performed on four other Wolbachia genomes is reported in Supplementary Figure 3.



## Discussion

The alphaproteobacterial genus *Wolbachia* has been classified into 16 supergroups, mainly on the basis of 16S rDNA phylogenetic analyses. This classification groups the *Wolbachia* strains coherently with the host taxonomy and ecology, opening the question of whether they could be classified as different species. We performed a set of genomic analyses to investigate this hypothesis.

Phylogenomic analyses have further organised *Wolbachia* diversity into two monophyletic clusters of supergroups: (A+B) and (C+D+F) (Comandatore et al. 2013; Gerth et al. 2014). While recombination has been observed between strains from the same supergroup, the supergroups may be relatively isolated genetically. Thus, no recombination was detected between *wHa* (supergroup A) and *wNo* (supergroup B), despite their infecting the same arthropod host species (Ellegaard et al. 2013). We sought to identify genomic structural and compositional differences between supergroups, with a particular focus on the (C+D+F) cluster.

Early comparisons of *Wolbachia* genomes revealed an extreme lack of synteny between supergroup A and B arthropod symbionts, and *wBm* (supergroup D) (Klasson and Andersson, 2006). Several additional supergroup C, D and F *Wolbachia* genomes are now available: *wDi*, *wOo*, *wOvC* (supergroup C), *wLs* (supergroup D), and *wCle* (supergroup F), permitting reanalysis of this pattern in the (C+D+F) cluster. We find that the genomes of supergroup C show an elevated level

of synteny, but supergroup D genomes are highly rearranged. This disjunct pattern suggests that supergroup D *Wolbachia* genomes may be evolving very differently from those of other *Wolbachia*.

IS elements are known to promote intragenomic recombination. *Wolbachia* genomes vary dramatically in terms of their IS content. Supergroup C genomes show a paucity of IS elements whereas supergroups A, B, D, and F genomes have many IS elements, a pattern consistent with a possible role for IS in synteny breakage in some *Wolbachia* genomes. IS element diversity also contrasts between supergroups. Supergroup D genomes contain many elements (and element fragments) deriving from very few distinct types, while the supergroup A and B genomes have many elements from diverse IS types (Cerveau et al. 2011). This might be explained by the lifestyle of the *Wolbachia* strains, as the mutualistic supergroup D strains are only vertically inherited in their nematode hosts, whereas supergroup A and B strains experience a combination of vertical and horizontal transmission. Horizontal transmission should enable more frequent contact and genetic exchanges with other microorganisms, and thus generate higher IS diversity. The supergroup F genome (from *wCle*) is also from a strain exhibiting mutualist interactions with its host, but *wCle* displays high IS diversity, like the parasitic supergroup A and B strains. This is consistent with the notion that *wCle* has recently shifted to mutualism and still shows transposable element patterns of its non-mutualistic ancestor. Within supergroup D, the IS elements of *wLs* and *wBm* appear to have expanded independently.

The homologous recombination pathway gene *recA*, involved in genome stability, is lacking in all supergroup C genomes (Darby et al. 2012; Badawi et al. 2014). By contrast, in supergroup D, the homologous recombination pathway is complete in the only closed genome available, *wBm* (Foster et al. 2005; Badawi et al. 2014), a result that supports genome plasticity. However, *wBm* may be exceptional, as other supergroup D genomes appear to have a deficient homologous recombination pathway (Badawi et al. 2014). Additional complete genome sequences from supergroup D strains will be needed to determine whether *wBm* is unusual in both its *recA* status and rearrangement history.

Conserved synteny means that the genomes of supergroup C *Wolbachia* also have other conserved genomic features. GC skew builds up in stable genomes through the differential mutagenic exposure of DNA on the leading and lagging strands during chromosomal replication (Rocha 2004). Rearrangement randomises the cumulative effect of this mutation pressure, and genome-wide GC skew is thus only observed in relatively evolutionarily stable genomes. Such a trend is commonly observed in free-living bacterial genomes (Grigoriev 1998), but was also previously reported for the supergroup C *wOo* genome (Darby et al. 2012). Klasson and Andersson (Klasson & Andersson 2006) described similar GC skew trend in the aphid endosymbiont *B. aphidicola*, and hypothesized that the lack of *recA* and mutational bias could be the causes of this GC distribution pattern. The GC skew also affects codon usage of genes depending on their position and orientation in the genome. We identified a strong genome-wide GC skew in *wDi*, and differences in codon usage between genes localized on leading versus lagging strands in

supergroup C genomes. As the GC skew pattern was strongest in *wDi*, implying that this genome has not rearranged for the longest time, we reordered the other *Wolbachia* genomes compared to *wDi* to identify any residual ancestral GC skew signature in the rearranged genomes that had not yet been erased during subsequent evolution.

The re-oriented *wOo* genome showed stronger GC skew than the natively ordered genome, albeit less pronounced than that of *wDi*, and was more similar to the *wDi* curve than that of other re-oriented *Wolbachia* genomes. The supergroup C genomes also had a significantly different mutational bias compared to other *Wolbachia*. The “retained” GC skew and the extended blocks of synteny among *wDi*, *wOo* and *wOvC*, lead us to hypothesize that a limited number genomic rearrangements have occurred in the *wOo/wOvC* lineage and that the ancestral gene order is largely conserved in *wDi*.

Is the *wDi* genome representative of the ancestor of all *Wolbachia*? We suggest not. It is likely that the loss of the *recA* pathway in the last common ancestor of supergroup C, and the general loss of IS elements, resulted in a halt to genome rearrangement, and this stability then permitted a build-up of GC skew and mutational bias in the stabilised genome. Limited subsequent rearrangements in some supergroup C strains (here represented by *wOo* and *wOvC*) have obscured but not erased the signatures of its evolutionary stability.

The process of gene loss is one of the most important phenomena in the evolution of

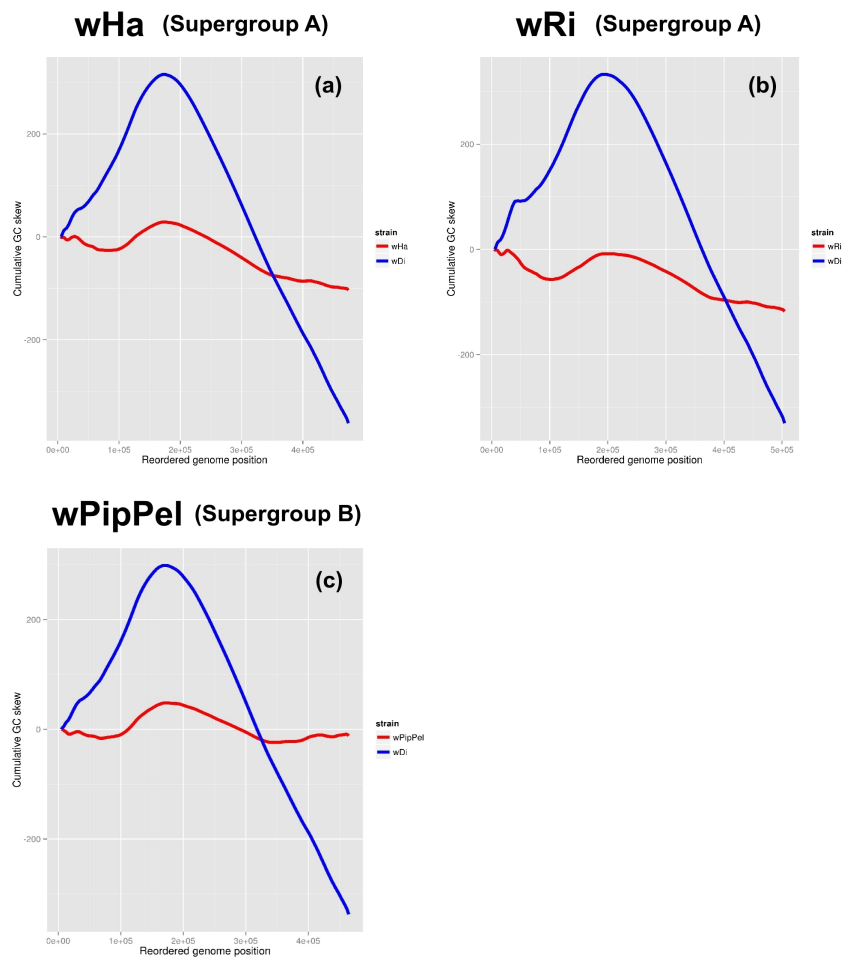
intracellular bacteria (Koonin & Wolf 2012), an importance exacerbated in filarial *Wolbachia* strains where gene acquisition from other bacterial species has not been described. The pattern of gene loss across *Wolbachia* genomes identified a set of WO-phage genes as being eliminated from extant supergroup C, D and F genomes, but the fossil supergroup F *Wolbachia* identified in the nuclear genome of *D. viviparus* contains phage components: loss of phage may have been gradual in the (C+D+F) cluster. In our global analysis, *recA* was identified as being lost from supergroup C, as expected, but we also identified a number of other losses in the supergroup C lineage associated with a variety of other processes. The physiological linkage between these gene losses, if any, is unclear.

The inferred HGTs between the ancestor of the (C+D+F) cluster and the ancestors of supergroups A and B (Table 4) suggest that the lineage that led to these *Wolbachia* supergroups was competent for acquisition of foreign DNA. HGT events between the ancestor of supergroup D and the supergroup F *wCle* suggest that the ability to recombine was maintained during early divergence of this cluster, at least within the lineage leading to the supergroup F, and was lost secondarily during genome reduction of the C and D lineages.

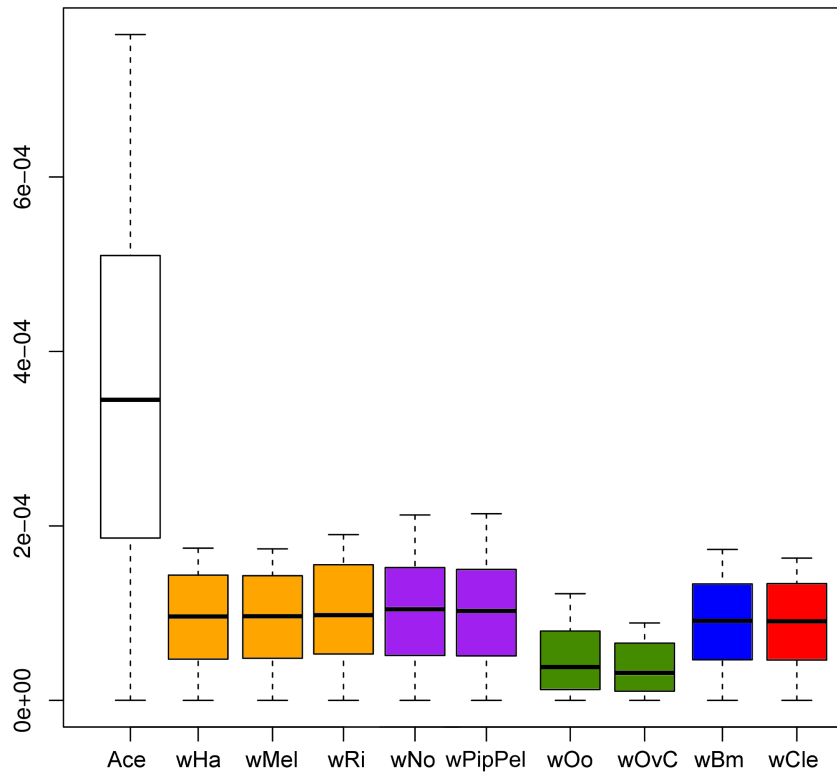
In conclusion, our analyses suggest that *Wolbachia* supergroups are not just phylogenetic lineages. The six supergroups analysed present coherence in multiple features. For example, codon usage clusters *Wolbachia* strains coherently in supergroups, and supergroups C and D are well differentiated in terms of IS presence, GC skew pattern, gene presence/absence, synteny and codon usage.

Thus, there is strong evidence for genomic isolation between living strains of *Wolbachia* that belong to different supergroups (this study and (Ellegaard et al. 2013)). In particular, supergroup C strains share a suite of genomic features (very low number of genomic rearrangements, paucity of IS elements, freeliving-like GC skew curve and codon usage that is different between genes located on leading and lagging strands) that is more commonly observed in free-living bacteria and associated with genome structure stability. These considerations support the contention that the A-F supergroups should be ranked at the species level (Pfarr et al. 2007; Ellegaard et al. 2013).

# Supplementary material

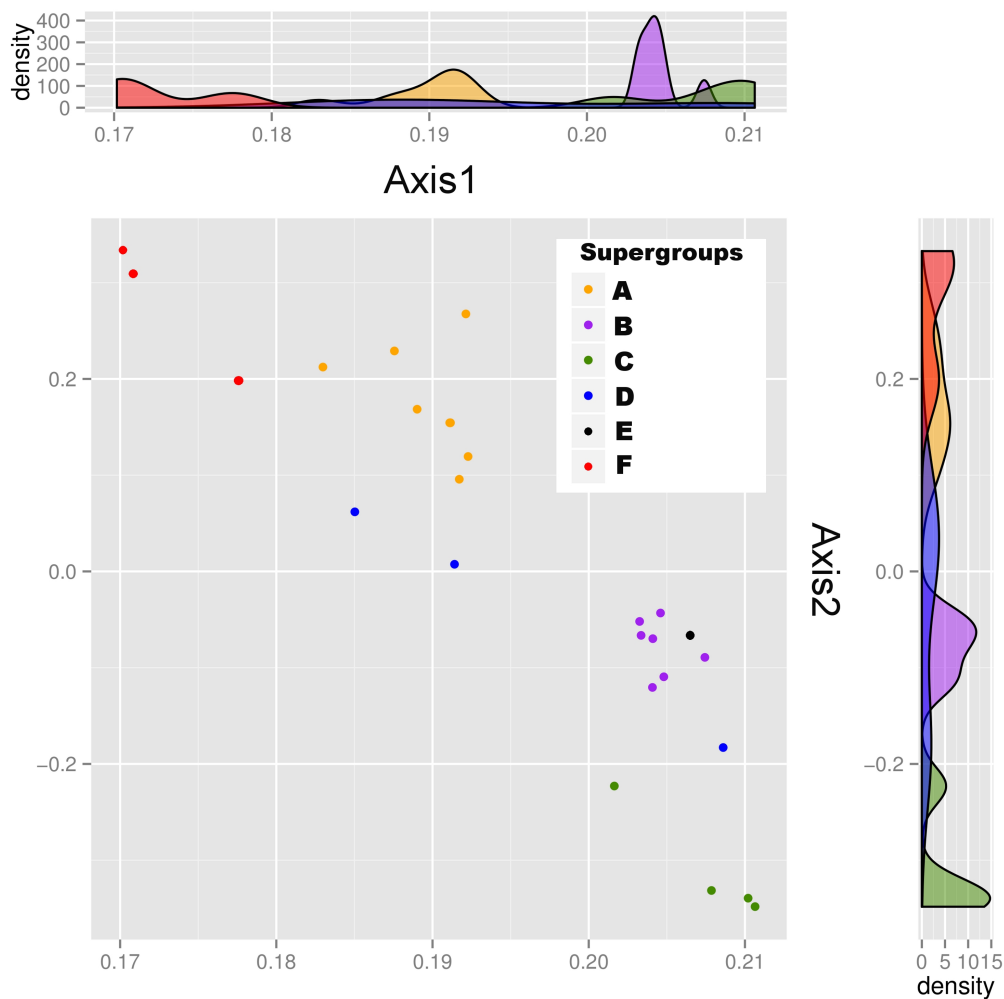


**Supplementary Figure 1.** Cumulative GC skew curves of *Wolbachia* genomes reordered based on wDi genome. Cumulative GC skew curves of the wHa (a), wRi (b) and wPipPel (c) reoriented *Wolbachia* genomes (in red) are compared to the wDi genome (in blue). Genomes were reordered

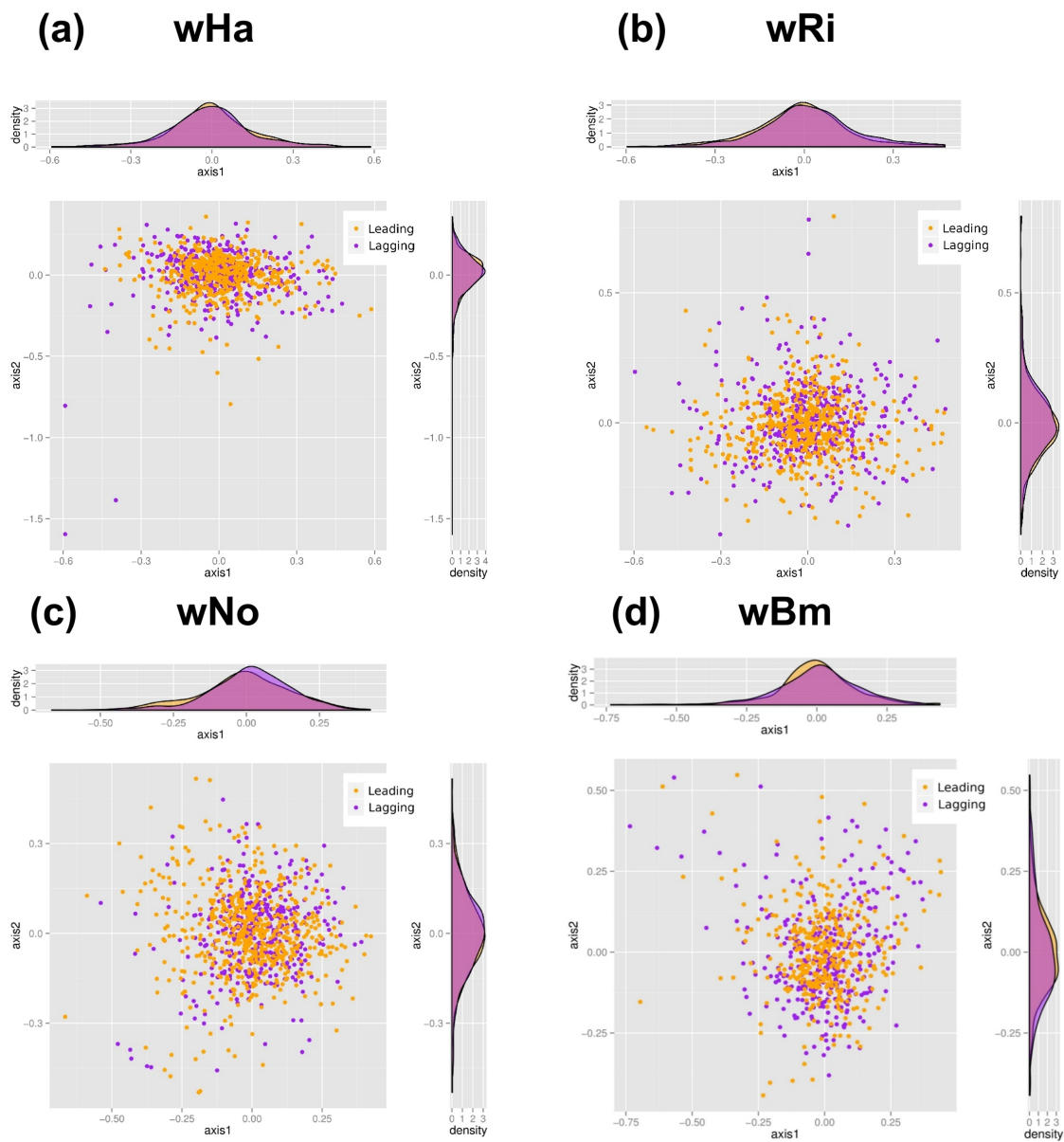


**Supplementary Figure 2.** Comparisons of the cumulative GC skew curves of *Wolbachia* genomes. Box plot of the distances between the cumulative GC skew curves of each reordered *Wolbachia* genome and the GC skew curve curve of the wDi genome. The wOo and wOvC GCskew curves are more similar to the wDi GCskew curve, in comparison to all the others (Wilcox test, p-value 0.01). For each *Wolbachia* strain the corresponding supergroup is highlighted as follow: orange, A; violet, B; green, C; blue, D; black, E; and red, F.





**Supplementary Figure 3.** Principal Component Analysis (PCA) of the Relative Synonymous Codon Usage (RSCU) indexes Scatter plot reporting the two most significant axes resulting from the Principal Component Analysis (PCA), performed of the Relative Synonymous Codon Usage (RSCU) indexes calculated on 23 *Wolbachia* strains from supergroups A-N. Each dot corresponds to a *Wolbachia* strain, colour-coded based on the supergroup: orange, A; violet, B; green, C; blue, D; black, E; and red, F. On each axis, the supergroup density distribution graph is reported. Axis1 explains 99.6% of the variation, while axis2 explains only 0.001% of the variation.



**Supplementary Figure 4.** Coordinates Analysis (COA) of genes on the leading and lagging strands in wHa, wRi, wNo and wBm genomes. Leading genes are represented in orange, lagging genes are represented in violet. In parallel with the X and Y axes, the corresponding distribution graphs are reported.

	<b>wDi</b>	<b>wOo</b>	<b>wLs</b>	<b>wBm</b>	<b>wCle</b>	<b>wMel</b>	<b>wRi</b>	<b>wPel</b>
Supergroup	C	C	D	D	F	A	A	B
Genome size (kb)	921	958	1049	1080	1,250	1268	1446	1482
IS copy number	1	6	210	52	181	105	171	170
Genome coverage	0.1%	0.5%	11.9%	2.6%	9.8%	6.1%	11.0%	8.4%
References	This study	(Cordaux 2009)	This study	(Cordaux 2009; C	This study	(Cerveau et al. 20	(Cerveau et al. 20	(Cerveau et al. 20
Group II intron copy number	0	0	0	0	7	16	14	6
Genome coverage	0	0	0	0	0,2%	1.9%	1.8%	0.8%
References	This study	(Darby et al. 2012	This study	(Leclercq et al. 20	This study	(Leclercq et al. 20	(Leclercq et al. 20	(Leclercq et al. 20

**Supplementary Table 1:** Transposable element content (IS elements and group II introns) in *Wolbachia* genomes.

<b>Wolbachia strain</b>	<b>ORI position range</b>	<b>ORI position (based on GCskew)</b>
wHa	1017366 - 1017769	
wRi	1055974 - 1056377	
wNo	580758 - 581164	
wPipPel	920589 - 920992	
wDi	2162 - 2646	-34377
wOo	374794 - 375255	
wCle	652958-653442	
wOvC	10858-11318	

**Supplementary Table 2:** Positions of the ORI and TER in *Wolbachia* genomes.

## References

- Augustinos AA et al. 2011. Detection and characterization of Wolbachia infections in natural populations of aphids: is the hidden diversity fully unraveled? PLoS One. 6:e28695. doi: 10.1371/journal.pone.0028695.
- Badawi M, Giraud I, Vavre F, Grève P, Cordaux R. 2014. Signs of neutralization in a redundant gene involved in homologous recombination in Wolbachia endosymbionts. Genome Biol. Evol. 6:2654–64. doi: 10.1093/gbe/evu207.
- Bandi C, Anderson TJ, Genchi C, Blaxter ML. 1998. Phylogeny of Wolbachia in filarial nematodes. Proc. Biol. Sci. 265:2407–13. doi: 10.1098/rspb.1998.0591.
- Bordenstein SR et al. 2009. Parasitism and mutualism in Wolbachia: What the phylogenomic trees can and cannot say. Mol. Biol. Evol. 26:231–241. doi: 10.1093/molbev/msn243.
- Bruen TC, Philippe H, Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. Genetics. 172:2665–81. doi: 10.1534/genetics.105.048975.
- Casiraghi M, Anderson TJ, Bandi C, Bazzocchi C, Genchi C. 2001. A phylogenetic analysis of filarial nematodes: comparison with the phylogeny of Wolbachia endosymbionts. Parasitology. 122 Pt 1:93–103. <http://www.ncbi.nlm.nih.gov/pubmed/11197770> (Accessed February 23, 2015).
- Cerveau N, Leclercq S, Leroy E, Bouchon D, Cordaux R. 2011. Short- and long-term evolutionary dynamics of bacterial insertion sequences: insights from Wolbachia endosymbionts. Genome Biol. Evol. 3:1175–86. doi: 10.1093/gbe/evr096.
- Comandatore F et al. 2013. Phylogenomics and analysis of shared genes suggest a single transition to mutualism in Wolbachia of nematodes. Genome Biol. Evol. 5:1668–1674. doi: 10.1093/gbe/evt125.
- Cordaux R. 2009. Gene conversion maintains nonfunctional transposable elements in an obligate mutualistic endosymbiont. Mol. Biol. Evol. 26:1679–82. doi: 10.1093/molbev/msp093.
- Cordaux R, Bouchon D, Grève P. 2011. The impact of endosymbionts on the evolution of host sex-determination mechanisms. Trends Genet. 27:332–41. doi: 10.1016/j.tig.2011.05.002.
- Darby AC et al. 2012. Analysis of gene expression from the Wolbachia genome of a filarial nematode supports both metabolic and defensive roles within the symbiosis. Genome Res. 22:2467–77. doi: 10.1101/gr.138420.112.
- Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14:1394–403. doi: 10.1101/gr.2289704.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and

- high throughput. *Nucleic Acids Res.* 32:1792–7. doi: 10.1093/nar/gkh340.
- Ellegaard KM, Klasson L, Näslund K, Bourtzis K, Andersson SGE. 2013. Comparative genomics of *Wolbachia* and the bacterial species concept. *PLoS Genet.* 9:e1003381. doi: 10.1371/journal.pgen.1003381.
- Foster J et al. 2005. The *Wolbachia* genome of *Brugia malayi*: endosymbiont evolution within a human pathogenic nematode. *PLoS Biol.* 3:e121. doi: 10.1371/journal.pbio.0030121.
- Gerth M, Gansauge M-T, Weigert A, Bleidorn C. 2014. Phylogenomic analyses uncover origin and spread of the *Wolbachia* pandemic. *Nat. Commun.* 5:5117. doi: 10.1038/ncomms6117.
- Gill AC, Darby AC, Makepeace BL. 2014. Iron necessity: the secret of *Wolbachia*'s success? *PLoS Negl. Trop. Dis.* 8:e3224. doi: 10.1371/journal.pntd.0003224.
- Glowska E, Dragun-Damian A, Dabert M, Gerth M. 2015. New *Wolbachia* supergroups detected in quill mites (Acari: Syringophilidae). *Infect. Genet. Evol.* 30:140–6. doi: 10.1016/j.meegid.2014.12.019.
- Godel C et al. 2012. The genome of the heartworm, *Dirofilaria immitis*, reveals drug and vaccine targets. *FASEB J.* 26:4650–4661.
- Grigoriev A. 1998. Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res.* 26:2286–90. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=147580&tool=pmcentrez&rendertype=abstract> (Accessed February 23, 2015).
- Hertig M. 2009. The *Rickettsia*, *Wolbachia pipientis* (gen. et sp.n.) and Associated Inclusions of the Mosquito, *Culex pipiens*. *Parasitology.* 28:453. doi: 10.1017/S0031182000022666.
- Hosokawa T, Koga R, Kikuchi Y, Meng X-Y, Fukatsu T. 2010. *Wolbachia* as a bacteriocyte-associated nutritional mutualist. *Proc. Natl. Acad. Sci. U. S. A.* 107:769–74. doi: 10.1073/pnas.0911476107.
- Ioannidis P et al. 2007. New criteria for selecting the origin of DNA replication in *Wolbachia* and closely related bacteria. *BMC Genomics.* 8:182. doi: 10.1186/1471-2164-8-182.
- Kambris Z, Cook PE, Phuc HK, Sinkins SP. 2009. Immune activation by life-shortening *Wolbachia* and reduced filarial competence in mosquitoes. *Science.* 326:134–6. doi: 10.1126/science.1177531.
- Klasson L et al. 2008. Genome evolution of *Wolbachia* strain wPip from the *Culex pipiens* group. *Mol. Biol. Evol.* 25:1877–87. doi: 10.1093/molbev/msn133.
- Klasson L, Andersson SGE. 2006. Strong asymmetric mutation bias in endosymbiont genomes coincide with loss of genes for replication restart pathways. *Mol. Biol. Evol.* 23:1031–9. doi: 10.1093/molbev/msj107.
- Kollenberg M, Winter S, Götz M. 2014. Quantification and localization of Watermelon chlorotic stunt virus and Tomato yellow leaf curl virus (Geminiviridae) in populations of *Bemisia tabaci* (Hemiptera, Aleyrodidae) with differential virus transmission characteristics. *PLoS One.* 9:e111968. doi: 10.1371/journal.pone.0111968.
- Koonin E V, Wolf YI. 2012. Evolution of microbes and viruses: a paradigm shift in

- evolutionary biology? *Front. Cell. Infect. Microbiol.* 2:119. doi: 10.3389/fcimb.2012.00119.
- Koutsovoulos G, Makepeace B, Tanya VN, Blaxter M. 2014. Palaeosymbiosis revealed by genomic fossils of *Wolbachia* in a strongyloidean nematode. *PLoS Genet.* 10:e1004397. doi: 10.1371/journal.pgen.1004397.
- Leclercq S, Giraud I, Cordaux R. 2011. Remarkable abundance and evolution of mobile group II introns in *Wolbachia* bacterial endosymbionts. *Mol. Biol. Evol.* 28:685–97. doi: 10.1093/molbev/msq238.
- Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13:2178–89. doi: 10.1101/gr.1224503.
- Lo N et al. 2007. Taxonomic status of the intracellular bacterium *Wolbachia pipientis*. *Int. J. Syst. Evol. Microbiol.* 57:654–7. doi: 10.1099/ijs.0.64515-0.
- Lo N, Casiraghi M, Salati E, Bazzocchi C, Bandi C. 2002. How many *wolbachia* supergroups exist? *Mol. Biol. Evol.* 19:341–6. <http://www.ncbi.nlm.nih.gov/pubmed/11861893> (Accessed March 4, 2015).
- Lobry JR. 1996. Origin of replication of *Mycoplasma genitalium*. *Science.* 272:745–6. <http://www.ncbi.nlm.nih.gov/pubmed/8614839> (Accessed February 23, 2015).
- Pfarr K, Foster J, Slatko B, Hoerauf A, Eisen JA. 2007. On the taxonomic status of the intracellular bacterium *Wolbachia pipientis*: should this species name include the intracellular bacteria of filarial nematodes? *Int. J. Syst. Evol. Microbiol.* 57:1677–8. doi: 10.1099/ijs.0.65248-0.
- Rocha EPC. 2004. The replication-related organization of bacterial genomes. *Microbiology.* 150:1609–27. doi: 10.1099/mic.0.26974-0.
- Sawyer S. 1989. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* 6:526–38. <http://www.ncbi.nlm.nih.gov/pubmed/2677599> (Accessed February 23, 2015).
- Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* 30:1312–3. doi: 10.1093/bioinformatics/btu033.
- Taylor MJ, Bandi C, Hoerauf A. 2005. *Wolbachia* bacterial endosymbionts of filarial nematodes. *Adv. Parasitol.* 60:245–84. doi: 10.1016/S0065-308X(05)60004-8.
- Varani AM, Siguier P, Goubeyre E, Charneau V, Chandler M. 2011. ISSaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biol.* 12:R30. doi: 10.1186/gb-2011-12-3-r30.
- Werren JH, Baldo L, Clark ME. 2008. *Wolbachia*: master manipulators of invertebrate biology. *Nat. Rev. Microbiol.* 6:741–751. doi: 10.1038/nrmicro1969.
- Werren JH, Zhang W, Guo LR. 1995. Evolution and phylogeny of *Wolbachia*: reproductive parasites of arthropods. *Proc. Biol. Sci.* 261:55–63. doi: 10.1098/rspb.1995.0117.

## 5. Summary of results and conclusions

During my Ph.D. I studied the evolution of the *Wolbachia*-host symbiotic relationship using genomic approaches. I took part to the projects for the sequencing and analysis of the genomes of the filarial nematode *Dirofilaria immitis*, of the *Wolbachia* endosymbiont of *Dirofilaria immitis* – wDi and of the *Wolbachia* endosymbiont of *Litomosoides sigmodontis* – wLs. I reconstructed the phylogeny of the supergroups A-D, using a novel phylogenomic approach, and I performed a comparison of the genomes of 26 *Wolbachia* strains spanning the supergroups from A to F.

The genomic analyses of *D. immitis* and wDi allowed to understand that the nematode and the bacterium are reciprocally dependent for the synthesis of an array of metabolites. In particular, the metabolic complementarity regards two fundamental pathways: i) the de novo biosynthesis of purine and pyrimidine pathway, absent in wDi and present in *D. immitis*; ii) the pathway for heme biosynthesis, which is present in wDi and absent in the host.

The comparison of the sequences of all the orthologous proteins shared between wBm and wDi shows that the enzymes involved in the membrane biosynthesis are likely subjected to a strong positive selective pressure. Considering that these proteins compose the contact surface between wDi and the host, this signal of fast evolution suggests a fundamental role for these proteins in the wDi-*D. Immitis* symbiosis.

The high rate of gene recombination that involved *Wolbachia* strains drastically reduces the phylogenetic signal contained in each *Wolbachia* gene sequence. For this reason, a robust phylogenetic reconstruction of the *Wolbachia* genus was not obtained before the genome sequencing of a sufficient number of strains. The sequencing of the wDi and wLs genomes allowed to have a number of *Wolbachia* genomes sufficient to perform phylogenetic analysis based on genomic sequences. I included in this analysis the genomes from 14 *Wolbachia* strains, spanning the supergroups from A to D. I selected a dataset of 90 not recombined and not saturated high-quality genes, that I then used in the phylogenetic reconstruction. This approach allowed me to obtain a robust *Wolbachia* phylogenetic tree, which showed that nematode *Wolbachia* strains originated from a single

ancestor. All the nematode *Wolbachia* strains included in the analysis have a clear mutualistic relationship with the host, in opposite to arthropod *Wolbachia* strains. Thus, this phylogenetic result suggests that mutualistic behavior originated only once during the evolution of the *Wolbachia* genus. The dataset of 90 high-quality genes that I generated resulted to be useful to the scientific community studying *Wolbachia*. Indeed Gerth et al. 2014 used it to reconstruct a more complete *Wolbachia* phylogenomic, which included, in addition to the genome that I used, the newly sequenced genomes from *Wolbachia* strains belonging to supergroups E, F and H.

I spent the last year of my Ph.D. period performing the comparison of 26 genomes from *Wolbachia* strains spanning from A to F supergroups. In this analysis I considered a set of genomic features, including GC skew curve pattern, genomic synteny, codon usage, transposable elements presence and gene loss. The results of this genomic comparison brought me to infer that the *recA* gene was lost in correspondence to the origin of the *Wolbachia* supergroup C. The RecA protein has a crucial role in the recombination pathway, often involved in chromosomal translocation. In my hypothesis, this gene loss has deeply affected the genomic evolution of *Wolbachia* strains of the C supergroup. The typical absence of synteny in *Wolbachia* genomes suggests that translocation has an important role in *Wolbachia* genome evolution. Without the *recA* gene, the ancestor of the C *Wolbachia* supergroup (and the current *Wolbachia* strains that belong to this lineage) was not able to complete a chromosomal translocation process. Thus, the role of translocation in the evolution of the C *Wolbachia* supergroup resulted to be drastically reduced. The most evident consequences of this genome stability are the conserved synteny among the strains of the lineage, and the “regeneration” of the free-living-like CG skew pattern in these genomes.

During the chromosomal replication process there is a bias during the elongation activity of the DNA Polymerase: it tends to accumulate more G than C in the leading strand, and more C than G in the lagging strand. The effects of this phenomenon are not detectable in high re-arranged genomes, but are evident in C *Wolbachia* strains. The molecular mechanism of this process is unknown, and it is not clear if it provides an advantage to the bacterium. Likely, a regular distribution of the G and C nucleotides along the genome could have an advantageous effect during the replication process.

It is very intriguing to note that the *recA* gene is present in the nematode *Wolbachia* strains



of the supergroup D. In contrast to what I observed in genomes of the C supergroup, D strains have a highly re-arranged genome structure, more similar to the arthropod *Wolbachia* strains. Furthermore, D *Wolbachia* strains resulted to be rich in transposable elements, in opposite to the C *Wolbachia* genomes that contain very low numbers.

The results of the genome comparison provided evidence for genomic isolation among supergroups: I found a very low rate of inter-supergroup gene recombination and high codon usage conservation within the supergroups. These evidences suggest that these monophyletic lineages, called supergroups, indeed have biological meaning, possibly warranting their classification at the species level. The taxonomic classification of lineages within the *Wolbachia* genus remains an open and intriguing topic in this research field.

# **Appendix:**

papers from parallel research lines

# Draft Genome of *Klebsiella pneumoniae* Sequence Type 512, a Multidrug-Resistant Strain Isolated during a Recent KPC Outbreak in Italy

Francesco Comandatore,<sup>a</sup> Paolo Gaibani,<sup>b</sup> Simone Ambretti,<sup>b</sup> Maria Paola Landini,<sup>b</sup> Daniele Daffonchio,<sup>c</sup> Piero Marone,<sup>d</sup> Vittorio Sambri,<sup>b</sup> Claudio Bandi,<sup>a</sup> Davide Sasseria<sup>a</sup>

Dipartimento di Scienze Veterinarie e Sanità Pubblica (DIVET)<sup>a</sup> and Dipartimento di Scienze per gli Alimenti, la Nutrizione e l'Ambiente (DeFENS),<sup>c</sup> Università degli Studi di Milano, Milan, Italy; Unit of Clinical Microbiology, St. Orsola Malpighi University Hospital, Bologna, Italy<sup>b</sup>; Fondazione Policlinico IRCCS San Matteo, Pavia, Italy<sup>d</sup>

F.C. and P.G. contributed equally to this article.

**Here, we present the draft genome sequence of *Klebsiella pneumoniae* subsp. *pneumoniae* sequence type 512 (ST512) isolated during a KPC-producer outbreak. This strain is resistant to  $\beta$ -lactams, cephalosporins, fluoroquinolones, aminoglycosides, macrolides, tetracyclines, and carbapenems but susceptible to colistin. The ST512-K30BO genome is composed of 289 contigs for 5,392,844 bp with 56.9% G+C content.**

Received 12 October 2012 Accepted 25 October 2012 Published 15 January 2013

**Citation** Comandatore F, Gaibani P, Ambretti S, Landini MP, Daffonchio D, Marone P, Sambri V, Bandi C, Sasseria D. 2013. Draft genome of *Klebsiella pneumoniae* sequence type 512, a multidrug-resistant strain isolated during a recent KPC outbreak in Italy. *Genome Announc.* 1(1):e00035-12. doi:10.1128/genomeA.00035-12.

**Copyright** © 2013 Comandatore et al. This is an open-access article distributed under the terms of the [Attribution 3.0 Unported Creative Commons License](http://creativecommons.org/licenses/by/3.0/).

Address correspondence to Davide Sasseria, [davide.sasseria@unimi.it](mailto:davide.sasseria@unimi.it).

*Klebsiella pneumoniae* is responsible for an increasing number of healthcare-related infections, mostly in patients with impaired immunity, including bloodstream and wound infections, pneumonia, and abscesses. The rapid diffusion of this pathogen is due mainly to the emergence of a number of multidrug-resistant strains (1). In particular, the first report of carbapenem-resistant *K. pneumoniae* in 2001 was followed by a worldwide spread of different types of carbapenemase producers, including the most widespread, *K. pneumoniae* carbapenemase (KPC) and New Delhi metallo- $\beta$ -lactamase (NDM) (2). The first Italian outbreak of *K. pneumoniae* KPC producers was reported recently (3).

The *K. pneumoniae* isolate ST512-K30BO was isolated using a central venous catheter from a hospitalized patient at the St. Orsola Malpighi University Hospital in Bologna, Italy. Antimicrobial susceptibility testing was performed according to the European Committee on Antimicrobial Susceptibility Testing guidelines (4). The isolate ST512-K30BO showed multiple resistances to clinically used antibiotics, including  $\beta$ -lactams, cephalosporins, carbapenems, fluoroquinolones, macrolides, aminoglycosides, and tigecycline. The strain was susceptible to colistin. Whole DNA was extracted using the Qiagen DNeasy kit and subjected to quality controls. Next-generation sequencing was performed on an Illumina HiSeq 2000 platform (5) with 300-base distant paired ends. Overall, 29,008,494 paired sequences were generated, for a total of more than 5.7 gigabases and a mean length of 199 bases per pair.

The genome assembly was performed using MIRA 3.4 (6) after quality selection and trimming via a specifically designed PerlScript. The assembly was manually checked using the Gap4 software of the Staden package (7). The resulting assembly consists of 289 contigs, with a G+C content of 56.9% for a total of 5,392,844 bp. Multilocus sequence type (MLST) analysis was performed using the Center for Biological Sequence Analysis (CBS)

server online tool (<http://www.cbs.dtu.dk/services/MLST/>). The sequenced genome was of strain 512. This strain, which is highly similar to the most widespread multidrug-resistant ST258 strain (8), has been reported previously as carbapenem-resistant and epidemic in Israel (9).

Genome annotation was performed automatically on the Rapid Annotation using Subsystem Technology (RAST) server (10) using Glimmer for base calling. Additionally, all open reading frames obtained from the RAST annotation were subjected to BLAST analysis against the Antibiotic Resistance Database (ARDB) (11) and the Comprehensive Antibiotic Resistance Database (CARD) (<http://arpcard.mcmaster.ca>). All of the genes indicated by at least one database as being implicated in antibiotic resistance were manually controlled. This approach highlighted the presence of 164 genes related to antibiotic resistance, including *bla*<sub>CTX-M9</sub>, *bla*<sub>TEM-33</sub>, *bla*<sub>SHV-2</sub>, *bla*<sub>KPC-3</sub>, *ant*(3'')-Ia, *ant*(2'')-Ia, *marA*, *macA*, *macB*, and *tetR*. Comparative genomic analyses will be performed to highlight similarities and differences between ST512 and other *K. pneumoniae* strains with different antimicrobial susceptibility patterns.

**Nucleotide sequence accession number.** The genome sequence was deposited in the European Bioinformatics Institute (EBI) under accession no. [CAJMO1000000](http://www.ebi.ac.uk/ena/submit/).

## ACKNOWLEDGMENTS

This study was supported by Grants RFO 2010 and RFO 2011 from the University of Bologna to V.S. and by funds from the Fondazione IRCCS Policlinico San Matteo to P.M.

## REFERENCES

1. Chong Y, Ito Y, Kamimura T. 2011. Genetic evolution and clinical impact in extended-spectrum  $\beta$ -lactamase-producing *Escherichia coli* and *Klebsiella pneumoniae*. *Infect. Genet. Evol.* 11:1499–1504.

2. Nordmann P, Gniadkowski M, Giske CG, Poirel L, Woodford N, Miriagou V, European Network on Carbapenemases. 2012. The European network on carbapenemases. Identification and screening of carbapenemase-producing *Enterobacteriaceae*. *Clin. Microbiol. Infect.* 18: 432–438.
3. Gaibani P, Ambretti S, Berlingeri A, Gelsomino F, Bielli A, Landini MP, Sambri V. 2010. Rapid increase of carbapenemase-producing *Klebsiella pneumoniae* strains in a large Italian hospital: surveillance period 1 March–30 September 2010. *Euro Surveill.* 16:ii, 19800.
4. European Committee on Antimicrobial Susceptibility Testing. 2011. Breakpoint tables for interpretation of MICs and zone diameters. Version 1.3. EUCAST, Basel, Switzerland.
5. Bennett S. 2004. Solexa Ltd. *Pharmacogenomics* 5:433–438.
6. Chevreux B, Wetter T, Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information, p 45–56. In *Proceedings of the German Conference on Bioinformatics*, Hanover, Germany.
7. Staden R, Beal KF, Bonfield JK. 2000. The Staden package. *Methods Mol. Biol.* 1998:132:115–130.
8. Woodford N, Turton JF, Livermore DM. 2011. Multiresistant Gram-negative bacteria: the role of high-risk clones in the dissemination of antibiotic resistance. *FEMS Microbiol. Rev.* 35:736–755.
9. Warburg G, Hidalgo-Grass C, Partridge SR, Tolmasky ME, Temper V, Moses AE, Block C, Strahilevitz J. 2012. A carbapenem-resistant *Klebsiella pneumoniae* epidemic clone in Jerusalem: sequence type 512 carrying a plasmid encoding *aac(6′)-Ib*. *J. Antimicrob. Chemother.* 67:898–901.
10. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 8:75.
11. Liu B, Pop M. 2009. ARDB—Antibiotic Resistance Genes Database. *Nucleic Acids Res.* 37:D443–D447.

# Draft Genome Sequences of Two Multidrug Resistant *Klebsiella pneumoniae* ST258 Isolates Resistant to Colistin

Francesco Comandatore,<sup>a</sup> Davide Sasseria,<sup>a</sup> Simone Ambretti,<sup>b</sup> Maria Paola Landini,<sup>b</sup> Daniele Daffonchio,<sup>c</sup> Piero Marone,<sup>d</sup> Vittorio Sambri,<sup>b</sup> Claudio Bandi,<sup>a</sup> Paolo Gaibani<sup>b</sup>

Dipartimento di Scienze Veterinarie e Sanità Pubblica (DIVET), Università degli Studi di Milano, Milano, Italy<sup>a</sup>; Unit of Clinical Microbiology, St. Orsola University Hospital, Bologna, Italy<sup>b</sup>; Dipartimento di Scienze per gli Alimenti, la Nutrizione e l'Ambiente (DeFENS), Università degli Studi di Milano, Italy<sup>c</sup>; Fondazione Policlinico IRCCS San Matteo, Pavia, Italy<sup>d</sup>

F.C. and D.S. contributed equally to this article.

**Sequence type 258 (ST258) is the most widespread multidrug resistant (MDR) *Klebsiella pneumoniae* strain worldwide. Here, we report the draft genome sequences of two colistin-resistant MDR *K. pneumoniae* ST258 clinical strains isolated from hospital patients in Italy. These strains are resistant to  $\beta$ -lactams, cephalosporins, fluoroquinolones, aminoglycosides, macrolides, tetracyclines, carbapenems, and colistin.**

Received 7 November 2012 Accepted 12 November 2012 Published 24 January 2013

**Citation** Comandatore F, Sasseria D, Ambretti S, Landini MP, Daffonchio D, Marone P, Sambri V, Bandi C, Gaibani P. 2013. Draft genome sequences of two multidrug resistant *Klebsiella pneumoniae* ST258 isolates resistant to colistin. *Genome Announc.* 1(1):e00113-12. doi:10.1128/genomeA.00113-12.

**Copyright** © 2013 Comandatore et al. This is an open-access article distributed under the terms of the [Attribution 3.0 Unported Creative Commons License](http://creativecommons.org/licenses/by/3.0/).

Address correspondence to Paolo Gaibani, [paolo.gaibani@unibo.it](mailto:paolo.gaibani@unibo.it).

In recent years, the rapid spread of *Klebsiella pneumoniae* showing multidrug resistant (MDR) phenotypes has been observed worldwide (1). *K. pneumoniae* carbapenemase (KPC)-producing *K. pneumoniae* isolates are resistant to carbapenems, cephalosporins, fluoroquinolones, and aminoglycosides. These MDR pathogens usually remain susceptible to colistin (2, 3). As a consequence of the increased use of colistin to treat infections provoked by these MDR strains, several outbreaks of colistin-resistant *K. pneumoniae* have been reported (4–6). Here, we present the draft genome sequences of two MDR colistin-resistant *K. pneumoniae* ST258 isolates in Italy.

The two *K. pneumoniae* ST258 isolates, ST258-K26BO and ST258-K28BO, were isolated from two patients hospitalized in the St. Orsola-Malpighi University Hospital in Bologna, Italy. An evaluation of their antimicrobial susceptibilities was performed following the European Committee on Antimicrobial Susceptibility Testing (7). The two isolates showed identical profiles, which included resistance to all  $\beta$ -lactams, cephalosporins, carbapenems, fluoroquinolones, macrolides, aminoglycosides, tigecycline, and colistin.

Next-generation sequencing was performed on the Illumina Hi-Seq 2000 platform (8) with 300-base distant paired-ends. Paired sequences (30,914,425 and 29,937,249) were generated, for a total of over 6.2 and 5.9 gigabases, respectively. Both data sets had mean lengths of 199 bases per pair.

Genome assembly was performed using MIRA 3.4 (9) after quality selection and trimming were done via a specifically designed PerlScript. The two assemblies were manually corrected using the Gap4 software of the Staden package (10). The assembly of ST258-K26BO consists of 193 contigs, with a G+C content of 57%, for a total of 5,526,679 bp. The assembly of ST258-K28BO consists of 168 contigs, with a G+C content of 57.1%, for a total of 5,663,706 bp.

Multilocus sequence typing (MLST) analysis was performed on the Component Build Service (CBS) server online tool (11). Both genomes were of the well-known sequence type 258.

Genome annotation was performed for both isolates on the Rapid Annotation using System Technology (RAST) server (12) using the Glimmer option for open reading frame (ORF) calling. Additionally, all ORFs obtained from the RAST annotation were subjected to BLAST analysis against the Antibiotic Resistance Database ARDB (13) and the Comprehensive Antibiotic Resistance Database (CARD) (14). This approach highlighted the presence of genes related to antibiotic resistance, which were 145 and 152 for ST258-K26BO and ST258-K28BO, respectively, including *bla*<sub>CTX-M9</sub>, *bla*<sub>TEM-33</sub>, *bla*<sub>SHV-2</sub>, *bla*<sub>KPC-3</sub>, *ant*(3'')-Ia, *ant*(2'')-Ia, *marA*, *macA*, *macB*, and *tetR*. Comparative genomic analyses will be performed in order to compare these and other *K. pneumoniae* strains to try to shed light on the mechanism of colistin resistance in this pathogen.

**Nucleotide sequence accession numbers.** The genome sequences were deposited at the European Bioinformatics Institute (EBI) under the accession no. [CANR01000000](http://www.ebi.ac.uk/ena/browser/view/CANR01000000) and [CANS01000000](http://www.ebi.ac.uk/ena/browser/view/CANS01000000).

## ACKNOWLEDGMENTS

This study was supported by RFO 2010 and RFO 2011 from University of Bologna to V.S. and by funds from Fondazione IRCCS Policlinico S. Matteo Pavia to P.M.

## REFERENCES

1. Nordmann P, Gniadkowski M, Giske CG, Poirel L, Woodford N, Miriagou V, European Network on Carbapenemases. 2012. Identification and screening of carbapenemase-producing Enterobacteriaceae. *Clin. Microbiol. Infect.* 18(5):432–438.
2. Gaibani P, Ambretti S, Berlingeri A, Gelsomino F, Bielli A, Landini MP, Sambri V. 2010. Rapid increase of carbapenemase-producing *Klebsiella pneumoniae* strains in a large Italian Hospital: surveillance period 1 March - 30 September 2010. *Euro Surveill.* 16(8):19800.

3. Comandatore F, Gaibani P, Ambretti S, Landini MP, Daffonchio D, Marone P, Sambri V, Bandi C, Sasseria D. In press. Draft genome of *Klebsiella pneumoniae* sequence type 512, a multidrug resistant strain isolated during a recent KPC outbreak in Italy. *Genome. Announc.*1(1):e00035-12.
4. Bogdanovich T, Adams-Haduch JM, Tian GB, Nguyen MH, Kwak EJ, Muto CA, Doi Y. 2011. Colistin-resistant, *Klebsiella pneumoniae* carbapenemase (KPC)-producing *Klebsiella pneumoniae* belonging to the international epidemic clone ST258. *Clin. Infect. Dis.* 53(4):373–376.
5. Antoniadou A, Kontopidou F, Poulakou G, Koratzanis E, Galani I, Papadomichelakis E, Kopterides P, Souli M, Armaganidis A, Giamarelou H. 2007. Colistin-resistant isolates of *Klebsiella pneumoniae* emerging in intensive care unit patients: first report of a multiclonal cluster. *J. Antimicrob. Chemother.* 59(4):786–790.
6. Mezzatesta ML, Gona F, Caio C, Petrolito V, Sciortino D, Sciacca A, Santangelo C, Stefani S. 2011. Outbreak of KPC-3-producing, and colistin-resistant, *Klebsiella pneumoniae* infections in two Sicilian hospitals. *Clin. Microbiol. Infect.* 17(9):1444–1447.
7. EUCAST. 2011. Breakpoint tables for interpretation of MICs and zone diameters. Version 1.3. **European Committee on Antimicrobial Susceptibility Testing**, Switzerland.
8. Bennett S. 2004. Solexa Ltd. *Pharmacogenomics* 5(4):433–438.
9. Chevreux B, Wetter T, Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information, p 45–56, In German Conference on Bioinformatics, Hannover, Germany.
10. Staden R, Beal KF, Bonfield JK. 2000. The Staden package. *Methods Mol. Biol.* 1998:132:115–130.
11. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Pontén T, Ussery DW, Aarestrup FM, Lund O. 2012. Multilocus sequence typing of total genome sequenced bacteria. *J. Clin. Microbiol.* 50:1355–1361.
12. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 9:75.
13. Liu B, Pop M. 2009. ARDB--Antibiotic resistance genes Database. *Nucleic Acids Res.* 37(Database issue):D443–D447.
14. The Comprehensive Antibiotic Resistance Database Pilot Project. Michael G. DeGroot Institute for Infectious Disease Research McMaster University, Hamilton, Ontario, Canada. <http://arpcard.mcmaster.ca>.

# Draft Genome Sequence of *Stenotrophomonas maltophilia* Strain EPM1, Found in Association with a Culture of the Human Parasite *Giardia duodenalis*

Davide Sasseria,<sup>a</sup> Iacopo Leardini,<sup>a</sup> Laura Villa,<sup>b</sup> Francesco Comandatore,<sup>a</sup> Claudio Carta,<sup>c</sup> André Almeida,<sup>d</sup> Maria do Céu Sousa,<sup>e</sup> Stefano Gaiarsa,<sup>a</sup> Piero Marone,<sup>f</sup> Edoardo Pozio,<sup>b</sup> Simone M. Cacciò<sup>b</sup>

DIVET, Università degli Studi di Milano, Milan, Italy<sup>a</sup>; Department of Infectious, Parasitic, and Immunomediated Diseases, Istituto Superiore di Sanità, Rome, Italy<sup>b</sup>; National Center for Rare Diseases, Istituto Superiore di Sanità, Rome, Italy<sup>c</sup>; Instituto de Ciências e Tecnologias Agrárias e Agroalimentares da Universidade do Porto, Porto, Portugal<sup>d</sup>; Faculdade de Farmácia/CEF, Universidade de Coimbra, Coimbra, Portugal<sup>e</sup>; Fondazione Policlinico IRCCS San Matteo, Pavia, Italy<sup>f</sup>

**We report the draft genome sequence of the *Stenotrophomonas maltophilia* strain EPM1, found in association with a culture of *Giardia duodenalis*. The draft genome sequence of *S. maltophilia* strain EPM1, obtained with Roche 454 GS-FLX Titanium technology, is composed of 19 contigs totaling 4,785,869 bp, with a G+C content of 66.37%.**

Received 11 March 2013 Accepted 13 March 2013 Published 18 April 2013

**Citation** Sasseria D, Leardini I, Villa L, Comandatore F, Carta C, Almeida A, do Céu Sousa M, Gaiarsa S, Marone P, Pozio E, Cacciò SM. 2013. Draft genome sequence of *Stenotrophomonas maltophilia* strain EPM1, found in association with a culture of the human parasite *Giardia duodenalis*. *Genome Announc.* 1(2):e00182-13. doi:10.1128/genomeA.00182-13.

**Copyright** © 2013 Sasseria et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 3.0 Unported license](http://creativecommons.org/licenses/by/3.0/).

Address correspondence to Simone M. Cacciò, [simone.caccio@iss.it](mailto:simone.caccio@iss.it).

*Stenotrophomonas maltophilia* is an aerobic, nonfermentative gammaproteobacterium commonly found in water and soil and in association with plants. *S. maltophilia* has several beneficial effects on plants, including protection from pathogens, promotion of growth, and biodegradation of pollutants (1). Furthermore, it is recognized as an emerging opportunistic human pathogen that is spread easily in hospital settings (2). In immunocompromised patients, *S. maltophilia* can lead to nosocomial infections with a significant fatality-to-case ratio (3), and up to 15% of patients with cystic fibrosis are colonized with *S. maltophilia* (4). Different *S. maltophilia* strains exhibit resistances to common antibiotics due to chromosomally encoded multidrug resistance proteins, such as antibiotic-inactivating enzymes and efflux pumps (5). These resistances are suggested to be acquired in the environment (6).

*S. maltophilia* strain EPM1 was found during the sequencing of two human-derived strains of *Giardia duodenalis*, a flagellated protozoan that parasitizes the small intestine of mammals. In its vegetative stage, this parasite can be grown in a medium supplemented with antibiotics and antifungals. The presence of *S. maltophilia* is likely the result of laboratory contamination of this culture medium. The environmental origin of the EPM1 strain is reinforced by its occurrence in other *G. duodenalis* strains that were propagated in the same medium. Here, we present the sequence and annotation of the genome of *S. maltophilia* EPM1. Whole DNA was extracted using a commercial kit (Qiagen) and was subjected to quality controls. Next-generation sequencing was performed on a full plate of the Roche 454 GS-FLX Titanium platform (7). A total of 1,290,645 reads were generated, with an average length of 477 bases, for a total of 591 megabases. A first assembly was performed by feeding all the reads to MIRA 3.4 (8). The 1,804 contigs obtained were subjected to BLAST analysis against *Giardia* and *Stenotrophomonas* databases. Reads belonging

to contigs exhibiting an E value of  $<10^{-4}$  against the *Giardia* database but not against the *Stenotrophomonas* database were excluded from subsequent analyses. The remaining 816,591 reads were assembled on MIRA 3.4. The resulting assembly of 258 contigs (with an average coverage of 70.44-fold in contigs of  $>5,000$  bp) was subjected to an in-house finishing procedure using the Gap4 (9) and NUCmer (10) software, allowing a total of 239 unequivocal contig joins. The resulting assembly consists of 19 contigs with a G+C content of 66.37%, for a total of 4,785,869 bp.

Multilocus sequence typing (MLST) analysis was performed on PubMLST (<http://pubmlst.org>) (11) and showed that *S. maltophilia* EPM1 differed from codified sequence types, yet it clustered with clinical isolates in genogroup 6 in phylogenetic analyses (12).

Annotation was performed automatically on the RAST server (13) using Glimmer base calling. The genome includes 4,334 predicted coding sequences and 75 RNAs. Reading frames obtained from the RAST annotation were subjected to BLAST analysis against the Comprehensive Antibiotic Resistance Database (CARD; Michael G. DeGroot, Institute for Infectious Disease Research, McMaster University, Hamilton, Ontario, Canada [<http://arpcard.mcmaster.ca>]). This approach highlighted the presence of 154 genes related to antibiotic resistance.

**Nucleotide sequence accession numbers.** This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession no. [AMXM000000000](https://www.ncbi.nlm.nih.gov/nuccore/AMXM000000000). The version described in this paper is the first version, accession no. [AMXM010000000](https://www.ncbi.nlm.nih.gov/nuccore/AMXM010000000).

## ACKNOWLEDGMENTS

This study was supported by the European Commission (contract SANCO/2006/FOODSAFETY/032). André Almeida is supported by the

grant SFRH/BPD/79539/2011 from the Fundação Para a Ciência e a Tecnologia.

## REFERENCES

- Ryan RP, Monchy S, Cardinale M, Taghavi S, Crossman L, Avison MB, Berg G, van der Lelie D, Dow JM. 2009. The versatility and adaptation of bacteria from the genus *Stenotrophomonas*. *Nat. Rev. Microbiol.* 7:514–525.
- Brooke JS. 2012. *Stenotrophomonas maltophilia*: an emerging global opportunistic pathogen. *Clin. Microbiol. Rev.* 25:2–41.
- Falagas ME, Kastoris AC, Vouloumanou EK, Rafailidis PI, Kapaskelis AM, Dimopoulos G. 2009. Attributable mortality of *Stenotrophomonas maltophilia* infections: a systematic review of the literature. *Future Microbiol.* 4:1103–1109.
- Abbott IJ, Slavin MA, Turnidge JD, Thursky KA, Worth LJ. 2011. *Stenotrophomonas maltophilia*: emerging disease patterns and challenges for treatment. *Expert Rev. Anti Infect. Ther.* 9:471–488.
- Crossman LC, Gould VC, Dow JM, Vernikos GS, Okazaki A, Sebahia M, Saunders D, Arrowsmith C, Carver T, Peters N, Adlem E, Kerhornou A, Lord A, Murphy L, Seeger K, Squares R, Rutter S, Quail MA, Rajandream MA, Harris D, Churcher C, Bentley SD, Parkhill J, Thomson NR, Avison MB. 2008. The complete genome, comparative and functional analysis of *Stenotrophomonas maltophilia* reveals an organism heavily shielded by drug resistance determinants. *Genome Biol.* 9:R74.
- Berg G, Eberl L, Hartmann A. 2005. The rhizosphere as a reservoir for opportunistic human pathogenic bacteria. *Environ. Microbiol.* 7:1673–1685.
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380.
- Chevreur B, Wetter T, Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information, p 45–56. *In* Computer science and biology: proceedings of the German Conference on Bioinformatics (GCB) '99. Universität Trier, Hannover, Germany.
- Bonfield JK, Smith KF, Staden R. 1995. A new DNA sequence assembly program. *Nucleic Acids Res.* 23:4992–4999. <http://dx.doi.org/10.1093/nar/23.24.4992>.
- Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL. 1999. Alignment of whole genomes. *Nucleic Acids Res.* 27:2369–2376.
- Jolley KA, Chan MS, Maiden MC. 2004. mlstdbNet—distributed multi-locus sequence typing (MLST) databases. *BMC Bioinformatics* 5:86.
- Kaiser S, Biehler K, Jonas D. 2009. A *Stenotrophomonas maltophilia* multilocus sequence typing scheme for inferring population structure. *J. Bacteriol.* 19:2934–2943.
- Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 9:75.



# Acetic Acid Bacteria Genomes Reveal Functional Traits for Adaptation to Life in Insect Guts

Bessem Chouaia<sup>1</sup>, Stefano Gaiarsa<sup>1,2</sup>, Elena Crotti<sup>1</sup>, Francesco Comandatore<sup>2</sup>, Mauro Degli Esposti<sup>3</sup>, Irene Ricci<sup>4</sup>, Alberto Alma<sup>5</sup>, Guido Favia<sup>4</sup>, Claudio Bandi<sup>2</sup>, and Daniele Daffonchio<sup>1,\*</sup>

<sup>1</sup>Department of Food, Environmental, and Nutritional Sciences (DeFENS), University of Milan, Italy

<sup>2</sup>Dipartimento di Scienze Veterinarie e Sanità Pubblica (DIVET), University of Milan, Italy

<sup>3</sup>Italian Institute of Technology (IIT), Genoa, Italy

<sup>4</sup>Scuola di Bioscienze e Biotecnologie, Università degli Studi di Camerino, Camerino, Italy

<sup>5</sup>Dipartimento di Scienze Agrarie (DISAFA), Forestali e Alimentari, University of Turin, Grugliasco, Italy

\*Corresponding author: E-mail: daniele.daffonchio@unimi.it.

Accepted: March 12, 2014

**Data deposition:** This project has been deposited at EMBL/GenBank under the accession numbers CBLX010000001–CBLX010000027 for *Asaia platycodi* genome and CBLY010000001–CBLY010000009 for *Saccharibacter* sp. genome.

## Abstract

Acetic acid bacteria (AAB) live in sugar rich environments, including food matrices, plant tissues, and the gut of sugar-feeding insects. By comparing the newly sequenced genomes of *Asaia platycodi* and *Saccharibacter* sp., symbionts of *Anopheles stephensi* and *Apis mellifera*, respectively, with those of 14 other AAB, we provide a genomic view of the evolutionary pattern of this bacterial group and clues on traits that explain the success of AAB as insect symbionts. A specific pre-adaptive trait, cytochrome *bo*<sub>3</sub> ubiquinol oxidase, appears ancestral in AAB and shows a phylogeny that is congruent with that of the genomes. The functional properties of this terminal oxidase might have allowed AAB to adapt to the diverse oxygen levels of arthropod guts.

**Key words:** symbiosis, acetic acid bacteria, cytochrome oxidase.

## Introduction

Besides plant tissues and food matrices, acetic acid bacteria (AAB) live in symbiosis with insects (reviewed in Crotti et al. 2010). Several research teams have investigated the relationship between AAB and their host (Crotti et al. 2010) focusing on the insect gut. In addition to the intestine, AAB could also be localized in other insect body compartments. For instance, the acetic acid bacterium *Asaia* colonizes not only the gut but also the salivary glands and the male and female reproductive systems, which are crucial sites for the bacterial transmission by horizontal and vertical routes (Damiani et al. 2008; Crotti et al. 2009; Gonella et al. 2012). Studies aiming to understand the nature of AAB symbiosis focused on the role of or potential advantages given by AAB to their respective hosts (Ryu et al. 2008; Chouaia et al. 2012; Lee et al. 2013). Key traits for intimately interacting with the insect host include, among others, the capacity to colonize host tissues and the interaction with the innate immunity and the developmental pathways of

the host (Ryu et al. 2008; Gross et al. 2009; Douglas et al. 2011; Shin et al. 2011; Lee et al. 2013; Login and Heddi 2013). For instance, *Asaia* exerts a beneficial role during the development of mosquito larvae (Chouaia et al. 2012; Mitraka et al. 2013) affecting the expression of genes related to the cuticle formation (Mitraka et al. 2013). In *Drosophila*, AAB are involved in the modulation of innate immunity, which keeps pathogenic strains under control (Ryu et al. 2008). Moreover, in the same host, *Acetobacter pomorum* modulates the insulin signaling, a pathway involved in the regulation of development, body size, energy metabolism, and intestinal stem cell activity of the host (Shin et al. 2011).

There are actually 14 genomes of AAB deposited in the databases but a genomic analysis of the evolutionary factors driving the association with insects is lacking. By including novel genome sequences of two AAB, *Asaia platycodi* and *Saccharibacter* sp., respectively, isolated from the malaria vector *Anopheles stephensi* and the honeybee *Apis mellifera*,

we present a genomic evolutionary analysis of AAB for assessing traits associated with the success of some of their members as insect symbionts. We discuss the potential role of alternative terminal oxidases as symbiotic factors favoring the adaptation of AAB to the insect hosts.

## Results and Discussion

### Several Potential Symbiotic Traits Are Present in AAB

Annotation of the *A. platycodi* and *Saccharibacter* sp. genomes revealed a series of traits compatible with a symbiotic life style in the insect gut. *Asaia platycodi* and *Saccharibacter* present several secretion system (Sec-SRP and Tat for both genomes and type IV in the case of *A. platycodi*) and ABC transporters (in the case of *A. platycodi*) that may have roles in the cross talk between the bacterium and the host. A series of bacterial components for motility and cell surface structures can be implicated in the colonization of the gut epithelium by *A. platycodi* and *Saccharibacter*. These include the genes for the flagellar machinery (e.g., *MotA*, *MotB*, *FlaA*, *FlaB*, *FlgC*, *FlgD*, *FlgE2*, *FlgH*, *FtsI*) as well as genes encoding for fimbriae (*sF-Chap* and *sF-UshP*) and glycan biosynthesis. Although these features may help in the establishment of a symbiotic relationship, they are not essential for it. The presence of these traits was not associated with the ability to establish symbiosis. In fact, genes for the flagellar machinery were also present in *Ac. aceti*, *Gluconacetobacter diazotrophicus*, *Gluconobacter frateurii*, *G. morbifer*, *G. oxydans*, and *G. thailandicus* but not *Ac. pomorum* and *Commensalibacter intestini*. This trend was also observed for the other traits.

Both genomes contain the operon for the production of acetoin and 2,3-butandiol: These molecules have been shown to play a role in insects' pheromone signaling (Tolasch et al. 2003). 2,3-Butandiol has been implicated in the modulation of the innate immunity response of vertebrate hosts, facilitating tissue colonization by pathogenic bacteria (Bari et al. 2011). We can thus speculate that the production of metabolites potentially interfering with insect physiology and innate immunity might have provided AAB with a pre-adaptive feature toward symbiosis with insect hosts. This trait was observed in other AAB including most of those described as insect symbionts (i.e., *Ac. tropicalis*, *C. intestini*, *Glucona diazotrophicus*, *Glucona europaeus*, *Glucona oboediens*, *G. frateurii*, *G. morbifer*, *G. Oxydans*, and *G. thailandicus*).

### Adaptation to Diverse Oxidic Conditions

AAB are aerobic organisms, consistent with their lifestyle in oxygen rich environments (Kersters et al. 2006). On the other hand, the oxygen levels in the guts of many arthropods may vary from aerobic to completely anoxic (Sudakaran et al. 2012). We have thus focused our attention on the presence and distribution of the oxygen-reacting systems of the electron transport chain (terminal oxidases). The genomes of both

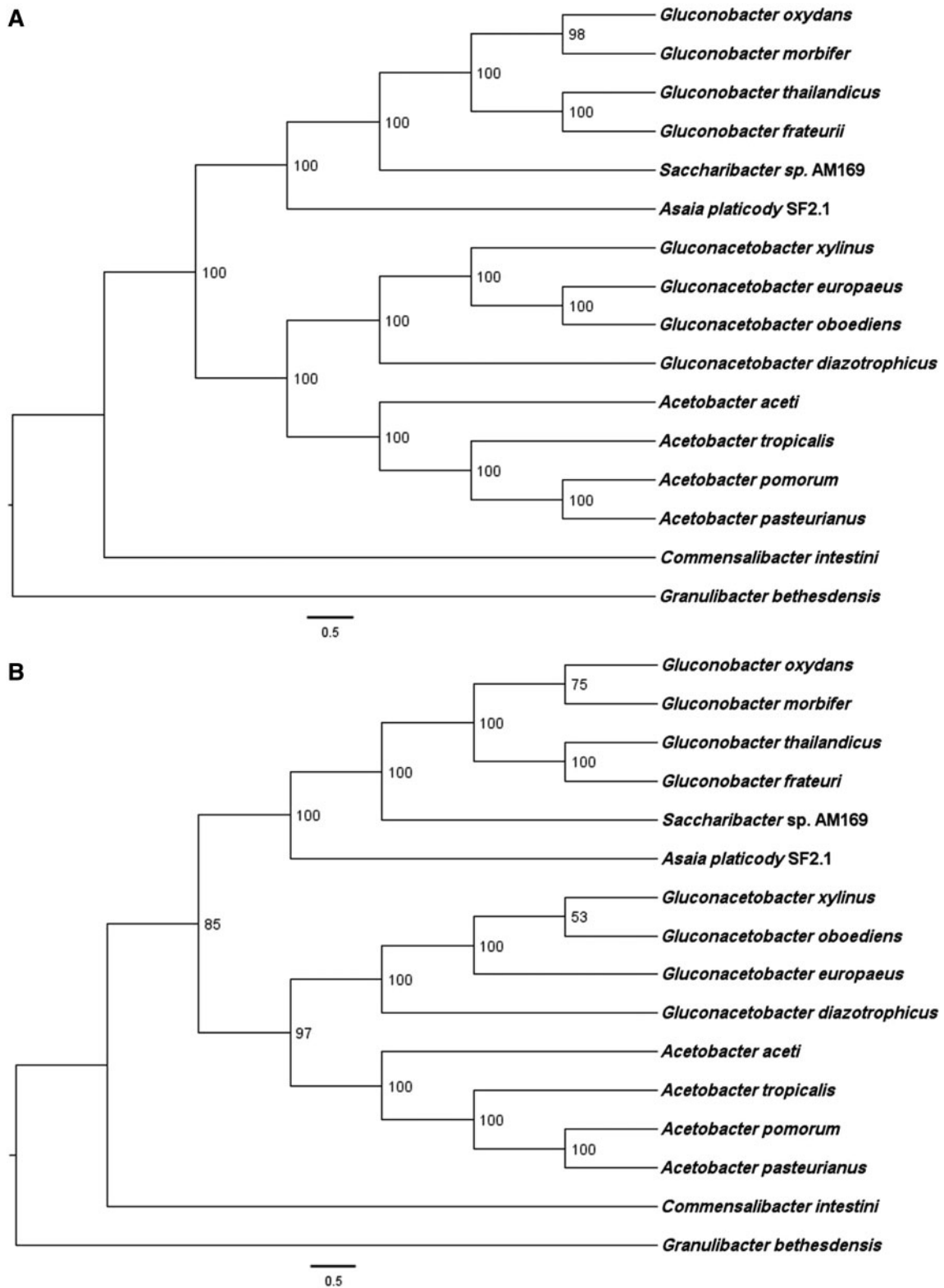
*A. platycodi* and *Saccharibacter* sp. present all the genes for the operons of cytochrome *bo*<sub>3</sub> (*CyoA-D*) and *bd* (*CydAB*) ubiquinol oxidase, which have high affinity for oxygen. The operons of both cytochrome *bo*<sub>3</sub> and *bd* oxidase are present in all AAB genomes, often with two different versions for the *bd* oxidase. This implies that both *A. platycodi* and *Saccharibacter* sp., as well as all the other AAB, have the capacity to respire through an aerobic respiratory chain independent from the terminal cytochrome *aa*<sub>3</sub> oxidase, an enzyme with low affinity for oxygen, that is absent in AAB. Therefore, AAB have the potential to thrive at low oxygen concentrations like the enterobacteria colonizing animal guts, which also do not possess the cytochrome *aa*<sub>3</sub> oxidase.

The phylogenetic tree of the protein subunits of cytochrome *bo*<sub>3</sub> oxidase of AAB matches that inferred using 70 proteins from the core genome (supplementary table S1, Supplementary Material online, and fig. 1), indicating that *bo*<sub>3</sub> oxidase evolution followed the differentiation of AAB species from a common ancestor. On the other hand, the phylogenetic tree of the cytochrome *bd* oxidases has a topology that is different from the phylogenomics inferred from core genes (fig. 2). In the *bd* oxidase tree, *Saccharibacter* branches with *Acetobacter* species rather than with *Gluconobacter* (fig. 2) suggesting that lateral gene transfer of *bd* oxidase genes might have occurred along the evolution of AAB.

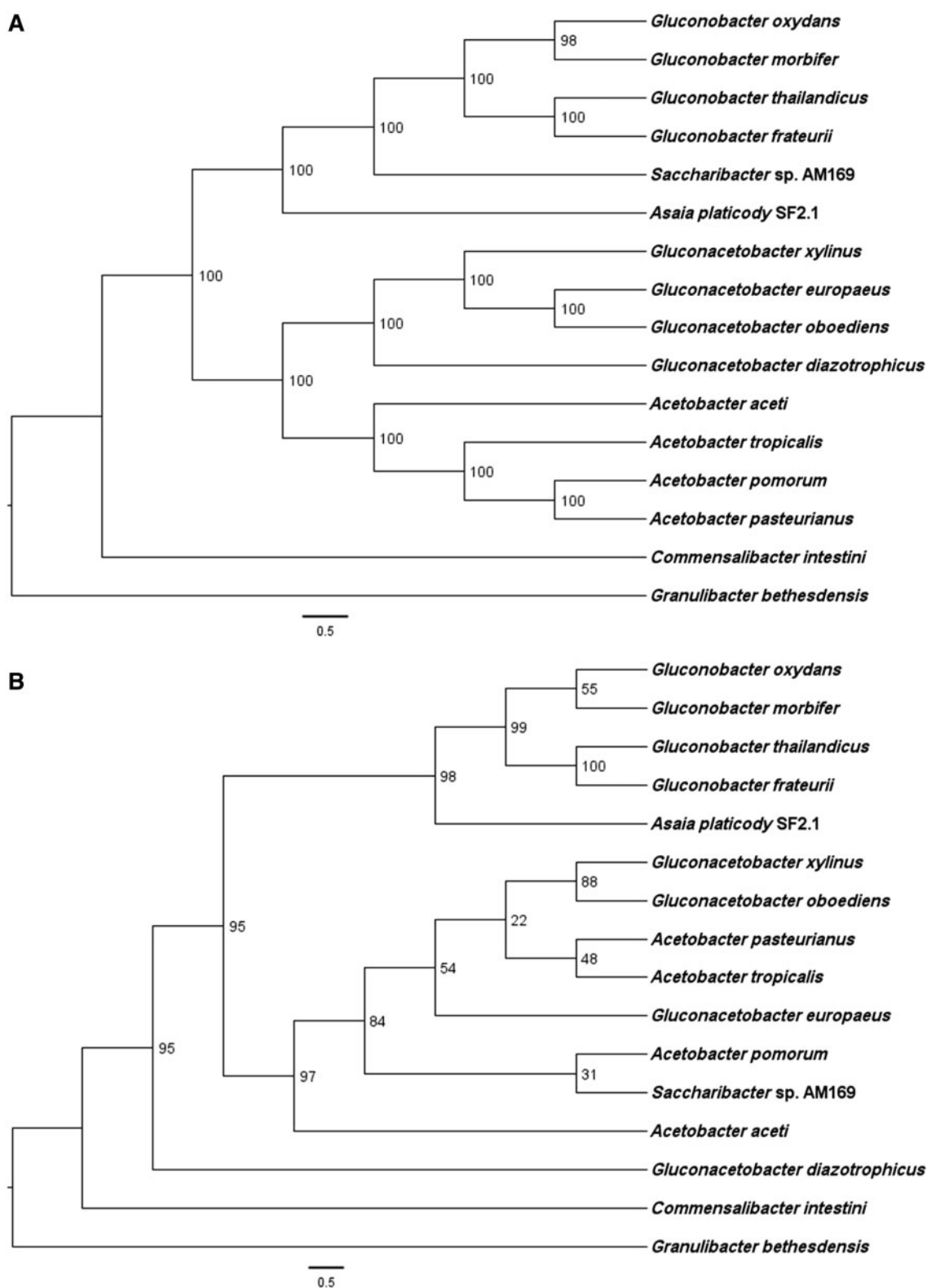
In sum, although AAB are usually described as strictly aerobic organisms thriving in normoxic environments, our results show that most of these organisms also possess ubiquinol oxidases that should allow their survival under micro-oxic conditions, such as those existing in the insect gut (Sudakaran et al. 2012). Moreover, phylogenetic comparisons show that these terminal oxidases were present in the common ancestor of AAB, thereby constituting an ancestral character. We thus propose that the capacity to thrive at low oxygen concentration conferred by ubiquinol oxidases has provided AAB organisms with a constitutive propensity for thriving in micro-oxic environments including the insect gut, an environment with ample variation in its oxygen levels (Sudakaran et al. 2012). The deep branching of the AAB family contains pathogens such as *Granulibacter bethesdensis* (figs. 1 and 2) further supports that capacity to establish intimate associations with animal hosts is an ancestral trait in these bacteria. On the other hand, the association of AAB with phylogenetically diverse insect species (Crotti et al. 2010) can be considered rather recent, in view of the phylogenetic proximity of symbiont and free-living bacteria (Chouaia et al. 2010).

The analysis of the genomes presented here thus provides new clues indicating ancient pre-adaptation traits to symbiosis in AAB organisms that might have helped, and are still helping, establishing association with insects.

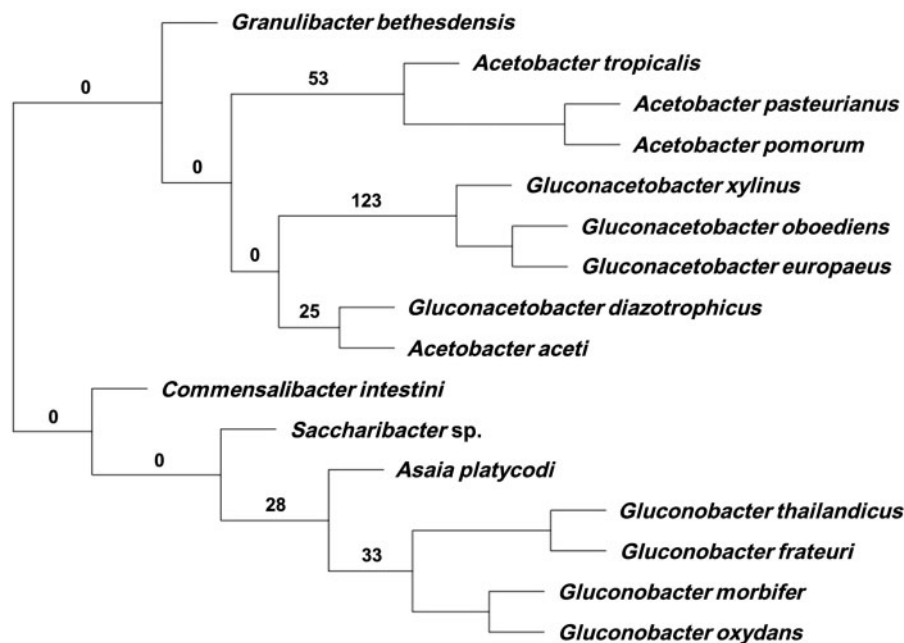
A cluster analysis carried on the ortholog gene groups that were not part of the core genes of the 16 AAB species showed that, in terms of gene acquisitions and losses, there was a coherence at the genus level: All of the members of a given



**Fig. 1.**—Comparison of species (AAB) (A) and operon (ubiquinol oxidase  $bo_3$ ) (B) in phylogenetic trees. The scientific names reported at the terminal nodes are those of the bacterial species. The tree of AAB (top) is based on the results of 70 concatenated protein (supplementary table S1, Supplementary Material online) phylogenetic analyses. Operon tree (bottom) was derived from the phylogeny inferred from the  $bo_3$  operon. The topology shown was obtained by the program RAxML using a partitioned ML model after reconstruction with 1,000 rapid bootstrap.



**Fig. 2.**—Comparison of species (AAB) (A) and operon (cytochrome oxidase *bc*) (B) in phylogenetic trees. The scientific names at the terminal nodes are those of the bacteria species. The tree of AAB (top) is based on the results of 70 concatenated protein (supplementary table S1, Supplementary Material online) phylogenetic analyses. Operon tree (bottom) was derived from the phylogeny inferred from the *bc* operon. The tree topology and other details were as given in figure 1.



**Fig. 3.**—Cluster analysis carried on the total number of ortholog groups after removal of those present in all genomes (i.e., core genome). The analysis shows that groups cluster at the genus level. Numbers on the branches indicate the number of ortholog groups specific to the cluster.

genus clustered together, except for *G. diazotrophicus* and *A. aceti* (fig. 3). However, when only orthologs present in at least 50% of the genomes are considered, the clustering resulting from the analysis of shared genes is congruent to the phylogeny based on the 70 coding sequences (CDS), including the positioning of *G. diazotrophicus* and *A. aceti* (figs. 1A and 4). In other words, phenetic analysis based on gene presence/absence produced a tree comparable to that generated by phylogenetics. This result supports the robustness of the phylogenomics here presented for AAB. It is noteworthy that the phylogenomic tree based on 70 CDS was not congruent with 16S rRNA-based phylogeny (supplementary fig. S1, Supplementary Material online)

### Electron Chain Transport and Symbiotic Traits

The analysis of the phylogeny of the cytochrome oxidase operons *bo<sub>3</sub>* and *bd* showed that both of them were ancestral. The analysis showed also that cytochrome oxidase *bo<sub>3</sub>* had an evolutionary history similar to that of the AAB genomes. These results suggest that these two operons may have played a role in the symbiotic potential of the AAB.

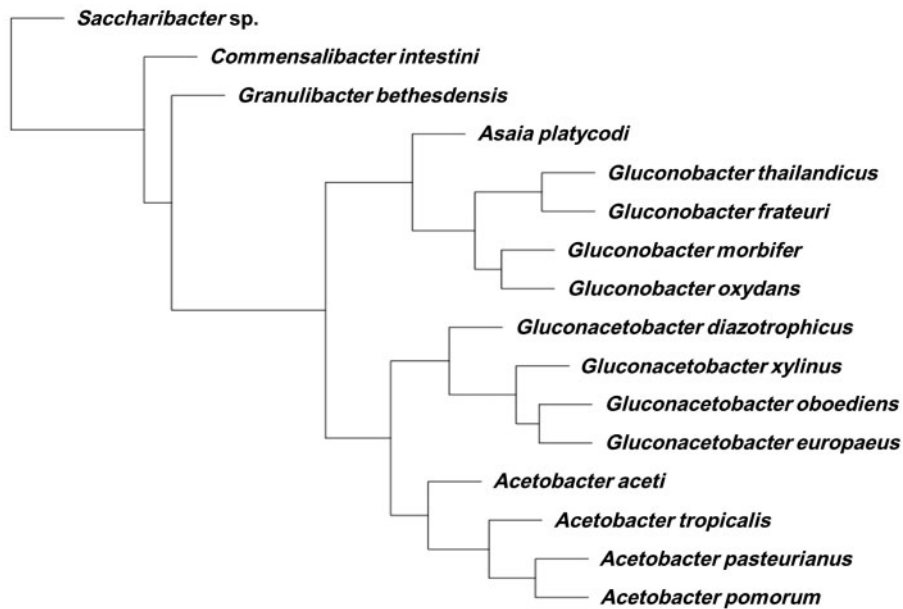
In order to investigate the possible implication of other proteins of the electron transport chain in the pre-adaptation of AAB to a symbiotic life, a further cluster analysis was carried on the different orthologs involved in the electron transport chain that were present in the different genomes. This analysis showed that the pattern of gain, loss, and duplication of these

genes was coherent and allowed to identify two groups (fig. 5), although there was no correlation between the presence of certain groups of orthologs and the ability to establish symbiosis. The two groups that were identified were the same that emerged from the phylogenomic study. The first group was formed by members of the *Acetobacter* and *Gluconacetobacter* genera, whereas the second group was formed by members of the *Gluconobacter* genus in addition to *Asaia* and *Saccharibacter*.

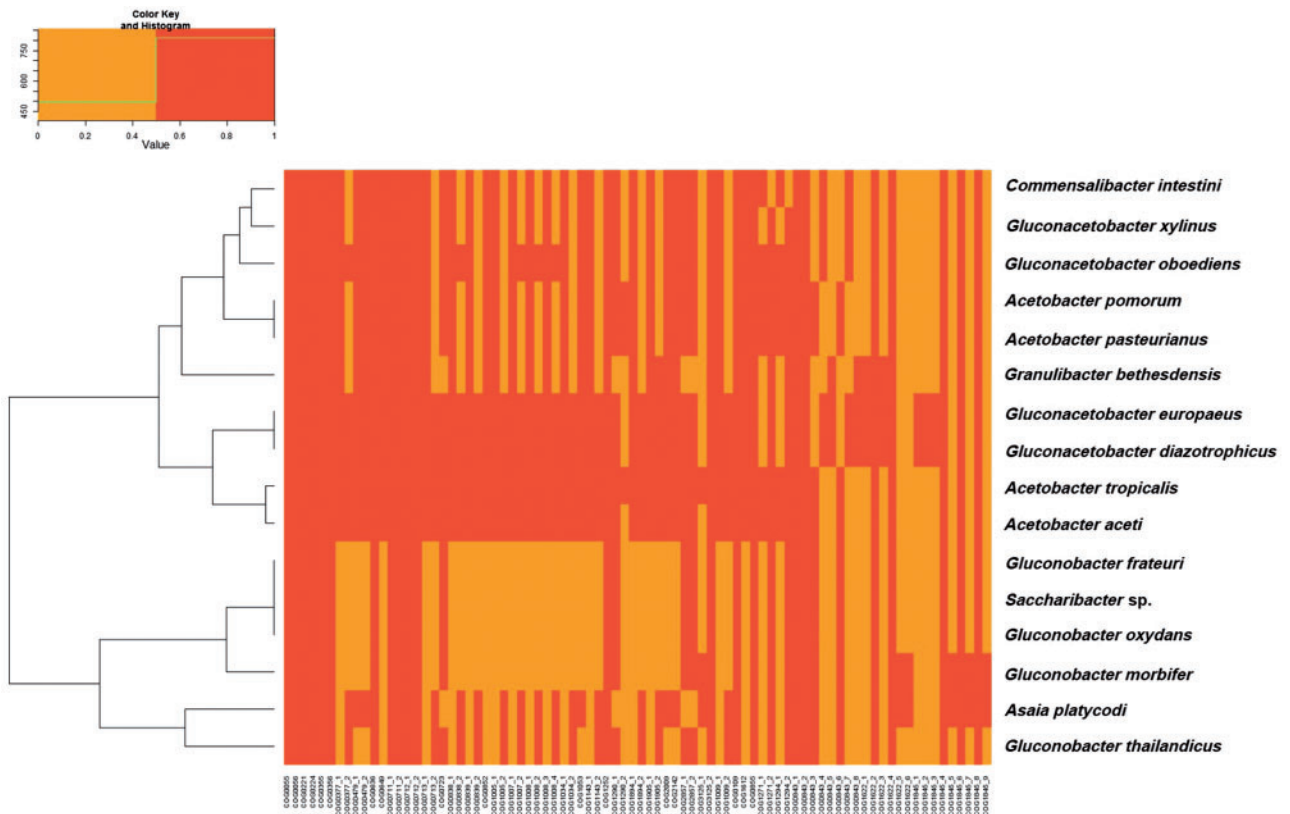
## Materials and Methods

### Strains

*Asaia platycodi* strain SF2.1 was isolated from *An. stephensi* (Favia et al. 2007). *Saccharibacter* sp. strain AM169 was isolated from an adult gut of *Ap. mellifera* using the pre-enrichment medium ABEM (2.0% D-sorbitol, 0.5% peptone, 0.3% yeast extract pH 3.5; Favia et al. 2007) supplemented with 100  $\mu\text{g ml}^{-1}$  of cycloheximide, followed by a plating on CaCO<sub>3</sub>-containing plates (1.0% D-glucose, 1.0% glycerol, 1.0% ethanol, 1.0% peptone, 0.5% yeast extract, 0.7% CaCO<sub>3</sub>, and 1.5% agar, pH 6.8). *Saccharibacter* sp. strain AM169 colony was selected based on the capability to clear CaCO<sub>3</sub>. Both strains were characterized as aerobic, Gram-negative, and rod-shaped bacteria belonging to the family Acetobacteraceae.



**Fig. 4.**—Cluster analysis carried on the subset ortholog groups that were present in at least 50% of the genome. The analysis shows that the clustering of the different groups is congruent with the phylogenomic analysis carried on 70 CDS (fig. 1A).



**Fig. 5.**—Gene presence–absence analysis of the oxidative phosphorylation chain orthologs in the genomes of AAB. A hierarchical clustering tree (left) was inferred based on the Kulczynski dissimilarity matrix calculated on the presence–absence matrix of genes in the examined genomes. The heatmap to the right of the tree represents the values of the Kulczynski dissimilarity matrix.

### Genome Sequencing, Assembly, and Annotation

The whole genome DNAs of *A. platycodi* SF2.1 and *Saccharibacter* sp. AM169 were purified using the DNeasy® Blood and Tissue kit (QIAGEN) and sequenced by Macrogen Korea institute. The genome sequence of *A. platycodi* SF2.1 was determined using a 3-kb paired-end library (~200 × 10<sup>3</sup> reads, ~80 Mb) with the Genome Sequencer FLX system (Roche, Diagnostics, Branford, CT) and a 100-bp library (~28 × 10<sup>6</sup> reads, ~3 Gb) with Genome AnalyzerIIx (Illumina, San Diego, CA). Raw data were assembled into 27 contigs—generated using Mira (version 3.4) (Chevreux et al. 1999); total coverage over the whole genome reached ~500-fold. The draft genome was 3,420,092 bp in length and contained 3,137 open reading frames (ORFs). The G+C content of the genome was 59.9%. The genome of *Saccharibacter* sp. AM169 was obtained from a tenth if a lane of Illumina HiSeq2000 platform generating 101-pb-long pair-end reads (~24 × 10<sup>6</sup> reads, ~2.4 Gb). The nine contigs were generated using Velvet (version 1.2) (Zerbino and Birney 2008); total coverage over the whole genome reached ~1000-fold. The draft genome was 1,978,091 bp in length and contained 1,877 ORFs. The G+C content of the genome was 59.3% (table 1).

The functional annotation of the predicted genes was performed using the RAST server (Aziz et al. 2008) combined with KEGG (Kanehisa et al. 2004) and COG (Tatusov et al. 2003) databases.

Genomes of *A. platycodi* and *Saccharibacter* sp. were also checked against the genomes of *G. oxydans* and *Glucon. diazotrophicus* using the KEGG database. The search for specific genes was carried out by local BLAST against the different genomes downloaded from the NCBI website using well-characterized and annotated genes. Confirmation or rectification for genes' annotation was obtained by additional DeltaBLAST

analysis (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>, last accessed April 8, 2014).

The result of this whole-genome project has been deposited at EMBL/GenBank database under the accession numbers CBLX010000001–CBLX010000027 for *A. platycodi* and CBLY010000001–CBLY010000009 for *Saccharibacter* sp. The material described in this article corresponds to the first version of the submitted genomes.

### Phylogenomics and Phylogenetic Reconstruction of Specific Operons

For the phylogenetic reconstruction of the AAB, 14 available (table 2), complete or draft, genomes were downloaded from the NCBI database along with our two genomes. A standardized ORF calling using Prodigal (Hyatt et al. 2010) was performed on all the nucleotidic sequences and nontruncated proteins longer than 50 residues were kept for the following analysis. Orthologs present in a single copy in any given genome were then selected using OrthoMCL (Li et al. 2003)

**Table 1**

General Genome Features of *Asaia platycodi* and *Saccharibacter* sp.

Organism	<i>A. platycodi</i> SF2.1	<i>Saccharibacter</i> sp. AM169
Genome size (pb)	3,420,092	1,978,091
Number of contigs	27	9
GC%	59.9	59.3
CDS	3,134	1,877
tRNAs	56	58
rRNAs	3	3
Accession number	CBLX010000001:27	CBLY010000001:9
Isolation year	2005	2010

**Table 2**

List of Bacterial Genomes Used for the Phylogenetic Studies

Organism	Accession Number	Reference	Origin
<i>Acetobacter aceti</i> NBRC 14818	PRJNA70715/PRJDA52649	Sakurai et al. (2011)	Reference strain
<i>Ac. pasteurianus</i> IFO 3283-01	PRJNA59279/PRJDA31129	Azuma et al. (2009)	Cocoa bean heap fermentation
<i>Ac. pomorum</i> DM001	PRJNA65823/PRJNA60787	Shin et al. (2011)	<i>Drosophila melanogaster</i>
<i>Ac. tropicalis</i> NBRC 101654	PRJNA68643/PRJDA46891	Matsutani et al. (2011)	Fruits
<i>Asaia platycodi</i> SF2.1	CBLX010000001:27	This study	<i>Anopheles stephensi</i>
<i>Commensalibacter intestini</i> A911	PRJNA75109/PRJNA73359	Roh et al. (2008)	<i>Drosophila melanogaster</i>
<i>Gluconacetobacter diazotrophicus</i> PAI 5	PRJNA61587/PRJNA377	Bertalan et al. (2009)	Sugarcane plants
<i>Glucon. europaeus</i> LMG 18494	PRJNA73763/PRJEA61325	Andrés-Barrao et al. (2011)	Reference strain
<i>Glucon. oboediens</i> 174Bp2	PRJNA73765/PRJEA61333	Andrés-Barrao et al. (2011)	Spirit vinegar
<i>Glucon. xylinus</i> NBRC 3288	PRJNA46523/PRJDA64985	Ogino et al. (2011)	Vinegar
<i>Gluconobacter frateurii</i> NBRC 101659	PRJNA178735/PRJDB2	Hattori et al. (2012)	Reference strain
<i>G. morbifer</i> G707	PRJNA76941/PRJNA73361	Roh et al. (2008)	<i>Drosophila melanogaster</i>
<i>G. oxydans</i> H24	PRJNA179202/PRJNA173388	Ge et al. (2013)	Reference strain
<i>G. thailandicus</i> NBRC 3255	PRJDB753/PRJNA191942	Matsutani et al. (2013)	Strawberry
<i>Granulibacter bethesdensis</i> CGDNIH1	PRJNA58661/PRJNA17111	Greenberg et al. (2007)	Chronic granulomatous disease patient
<i>Saccharibacter</i> sp. AM169	CBLY010000001:9	This study	<i>Apis mellifera</i>

and a custom script designed to keep only those matching with a single or no COG entry. The amino acid sequences of CDS belonging to each ortholog family were aligned using MUSCLE (Edgar 2004); the alignments were subsequently retro-transcribed to their respective nucleic acid sequences, which were checked for the probability of recombination and lateral gene transfer using the phi-test under the Phi-pack (Bruen et al. 2006). At the end of this screening, 70 proteins were kept for phylogenetic analysis (listed in [supplementary table S1, Supplementary Material](#) online). The alignment for the remaining CDS was Gblocked, keeping only the proteins that had <3 misaligned residues. For each of the aligned CDS, an evolutionary model was predicted using ProtTest (Darriba et al. 2011); then all the protein sequences were concatenated and their phylogenetic tree was constructed with RAxML (Stamatakis et al. 2005) using a partitioned Maximum Likelihood model that takes into account the evolutionary model predicted for each of the CDS. The phylogenetic trees were tested with 1,000 rapid bootstraps. The phylogenetic trees of the concatenated protein subunits of the cytochrome *bo<sub>3</sub>* and *bd* oxidases were constructed with the same method. A phylogenetic tree based on 16S rRNA was also inferred (see [supplementary material, Supplementary Material](#) online)

### Cluster Analysis on the Orthologs

The data obtained from OrthoMCL (5,488 ortholog groups) were transformed into a matrix reporting the presence of each ortholog group. The orthologs present in all of the genomes (i.e., core genome) were removed from the data set, leaving 4,575 ortholog groups, and a cluster analysis (Murtagh 1985) was carried out using R ([cran.r-project.org](http://cran.r-project.org), last accessed April 8, 2014). A second cluster analysis was carried out on a subset consisting only of the ortholog groups (1,167 ortholog groups) present in at least 50% of the genomes. The pattern of presence/absence was reconstructed also for the subset of genes involved in the oxidative phosphorylation chain. The oxidative phosphorylation chain genes (listed in [supplementary table S2, Supplementary Material](#) online) were identified on the basis of KEGG annotation of the *Gluconobacter diazotrophicus* PAI 5, *G. oxydans* H24, *Ac. pasteurianus* IFO 3283-01, and *Granulibacter bethesdensis* strain genomes; the relative presence/absence informations were retrieved from the OrthoMCL matrix generated above and organized in a new matrix. This new matrix was subjected to hierarchical clustering analysis using the Kulczynski distance index and the heatmap graphic representation was generated with R.

### Supplementary Material

[Supplementary methods, tables S1 and S2, and figure S1](#) are available at *Genome Biology and Evolution* online (<http://www.gbe.oxfordjournals.org/>).

### Acknowledgments

This work was supported by the project BIODESERT GA-245746 “Biotechnology from Desert Microbial Extremophiles for Supporting Agriculture Research Potential in Tunisia and Southern Europe” (European Union) and the Prin 2009 (grant 009L27YC8\_003), from the Italian Ministry of Education, University and Research (MIUR). C.B. and B.C. thank Massimo Pajoro for inspirations.

### Literature Cited

- Andrés-Barrao C, et al. 2011. Genome sequences of the high-acetic acid-resistant bacteria *Gluconacetobacter europaeus* LMG 18890T and *G. europaeus* LMG 18494 (reference strains), *G. europaeus* 5P3, and *Gluconacetobacter oboediens* 174Bp2 (isolated from vinegar). *J Bacteriol.* 193:2670–2671.
- Aziz RK, et al. 2008. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75.
- Azuma Y, et al. 2009. Whole-genome analyses reveal genetic instability of *Acetobacter pasteurianus*. *Nucleic Acids Res.* 37:5768–5783.
- Bari W, Song YJ, Yoon SS. 2011. Suppressed induction of proinflammatory cytokines by a unique metabolite produced by *Vibrio cholerae* O1 El Tor biotype in cultured host cells. *Infect Immun.* 79:3149–3158.
- Bertalan M, et al. 2009. Complete genome sequence of the sugarcane nitrogen-fixing endophyte *Gluconacetobacter diazotrophicus* Pa15. *BMC Genomics* 10:450.
- Bruen TC, Philippe H, Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172:2665–2681.
- Chevreaux B, Wetter T, Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information. Proceedings of the German Conference on Bioinformatics (GCB), Vol. 99:p. 45–56, Hannover, Germany.
- Chouaia B, et al. 2010. Molecular evidence for multiple infections as revealed by typing of *Asaia* bacterial symbionts of four mosquito species. *Appl Environ Microbiol.* 76:7444–7450.
- Chouaia B, et al. 2012. Delayed larval development in *Anopheles* mosquitoes deprived of *Asaia* bacterial symbionts. *BMC Microbiol.* 12(Suppl 1): S2.
- Crotti E, et al. 2009. *Asaia*, a versatile acetic acid bacterial symbiont, capable of cross-colonizing insects of phylogenetically distant genera and orders. *Environ Microbiol.* 11:3252–3264.
- Crotti E, et al. 2010. Acetic acid bacteria, newly emerging symbionts of insects. *Appl Environ Microbiol.* 76:6963–6970.
- Damiani C, et al. 2008. Paternal transmission of symbiotic bacteria in malaria vectors. *Curr Biol.* 18:R1087–R1088.
- Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27:1164–1165.
- Douglas AE, Bouvaine S, Russell RR. 2011. How the insect immune system interacts with an obligate symbiotic bacterium. *Proc R Soc Lond B Biol Sci.* 278:333–338.
- Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.
- Favia G, et al. 2007. Bacteria of the genus *Asaia* stably associate with *Anopheles stephensi*, an Asian malarial mosquito vector. *Proc Natl Acad Sci U S A.* 104:9047–9051.
- Ge X, et al. 2013. Complete Genome Sequence of the Industrial Strain *Gluconobacter oxydans* H24. *Genome Announc.*, 1(1). pii: e00003-13.
- Gonella E, et al. 2012. Horizontal transmission of the symbiotic bacterium *Asaia* sp. in the leafhopper *Scaphoideus titanus* Ball (Hemiptera: Cicadellidae). *BMC Microbiol.* 12(Suppl 1): S4.
- Greenberg DE, et al. 2007. Genome sequence analysis of the emerging human pathogenic acetic acid bacterium *Granulibacter bethesdensis*. *J Bacteriol.* 189:8727–8736.



- Gross R, et al. 2009. Immunity and symbiosis. *Mol Microbiol.* 73:751–759.
- Hattori H, et al. 2012. High-temperature sorbose fermentation with thermotolerant *Gluconobacter frateurii* CHM43 and its mutant strain adapted to higher temperature. *Appl Microbiol Biotechnol.* 95: 1531–1540.
- Hyatt D, et al. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119.
- Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32: D277–D280.
- Kerstens K, Lisdiyanti P, Komagata K, Swings J. 2006. The family Acetobacteraceae: the genera *Acetobacter*, *Acidomonas*, *Asaia*, *Gluconacetobacter*, *Gluconobacter*, and *Kozakia*. In: Dworkin M, Falkow S, Rosenberg E, Schleifer K-H, Stackebrandt E, editors. *The prokaryotes*. Vol. 5, 3rd ed. New York (NY): Springer. p. 163–200.
- Lee KA, et al. 2013. Bacterial-derived uracil as a modulator of mucosal immunity and gut-microbe homeostasis in *Drosophila*. *Cell* 153: 797–781.
- Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13:2178–2189.
- Login FH, Heddi A. 2013. Insect immune system maintains long-term resident bacteria through a local response. *J Insect Physiol.* 59: 232–239.
- Matsutani M, et al. 2011. Increased number of Arginine-based salt bridges contributes to the thermotolerance of thermotolerant acetic acid bacteria, *Acetobacter tropicalis* SKU1100. *Biochem Biophys Res Commun.* 409:120–124.
- Matsutani M, Kawajiri E, Yakushi T, Adachi O, Matsushita K. 2013. Draft genome sequence of dihydroxyacetone-producing *Gluconobacter thailandicus* strain NBRC 3255. *Genome Announc.* 1(2):e00118–13.
- Mitraka E, et al. 2013. *Asaia* accelerates larval development of *Anopheles gambiae*. *Pathog Glob Health.* 107(6):305–11.
- Murtagh F. 1985. Multidimensional clustering algorithms. COMPSTAT Lectures 4. Wuerzburg: Physica-Verlag.
- Ogino H, et al. 2011. Complete genome sequence of NBRC 3288, a unique cellulose-nonproducing strain of *Gluconacetobacter xylinus* isolated from vinegar. *J Bacteriol.* 193:6997–6998.
- Roh SW, et al. 2008. Phylogenetic characterization of two novel commensal bacteria involved with innate immune homeostasis in *Drosophila melanogaster*. *Appl Environ Microbiol.* 74:6171–6177.
- Ryu JH, et al. 2008. Innate immune homeostasis by the homeobox gene caudal and commensal-gut mutualism in *Drosophila*. *Science* 319: 777–782.
- Sakurai K, Arai H, Ishii M, Igarashi Y. 2011. Transcriptome response to different carbon sources in *Acetobacter acetii*. *Microbiology* 157: 899–910.
- Shin SC, et al. 2011. *Drosophila* microbiome modulates host developmental and metabolic homeostasis via insulin signaling. *Science* 334: 670–674.
- Stamatakis A, Ludwig T, Meier H. 2005. RAXML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21:456–463.
- Sudakaran S, Salem H, Kost C, Kaltenpoth M. 2012. Geographical and ecological stability of the symbiotic mid-gut microbiota in European firebugs, *Pyrrhocoris apterus* (Hemiptera, Pyrrhocoridae). *Mol Ecol.* 21: 6134–6151.
- Tatusov RL, et al. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4:41.
- Tolasch T, Sölter S, Tóth M, Ruther J, Francke W. 2003. (R)-acetoin-female sex pheromone of the summer chafer *Amphimallon solstitialis* (L.). *J Chem Ecol.* 29:1045–1050.
- Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18:821–829.

Associate editor: Bill Martin



# Evolution of Mitochondria Reconstructed from the Energy Metabolism of Living Bacteria

Mauro Degli Esposti<sup>1\*</sup>, Bessem Chouaia<sup>2</sup>, Francesco Comandatore<sup>3</sup>, Elena Crotti<sup>2</sup>, Davide Sasseria<sup>3#a</sup>, Patricia Marie-Jeanne Lievens<sup>1#b</sup>, Daniele Daffonchio<sup>2</sup>, Claudio Bandi<sup>3</sup>

**1** Italian Institute of Technology, Genoa, Italy, **2** Department of Food, Environmental and Evolutionary Sciences, University of Milan, Milan, Italy, **3** Dipartimento di Scienze Veterinarie e Sanità Pubblica, University of Milan, Milan, Italy

## Abstract

The ancestors of mitochondria, or proto-mitochondria, played a crucial role in the evolution of eukaryotic cells and derived from symbiotic  $\alpha$ -proteobacteria which merged with other microorganisms - the basis of the widely accepted endosymbiotic theory. However, the identity and relatives of proto-mitochondria remain elusive. Here we show that methylotrophic  $\alpha$ -proteobacteria could be the closest living models for mitochondrial ancestors. We reached this conclusion after reconstructing the possible evolutionary pathways of the bioenergy systems of proto-mitochondria with a genomic survey of extant  $\alpha$ -proteobacteria. Results obtained with complementary molecular and genetic analyses of diverse bioenergetic proteins converge in indicating the pathway stemming from methylotrophic bacteria as the most probable route of mitochondrial evolution. Contrary to other  $\alpha$ -proteobacteria, methylotrophs show transition forms for the bioenergetic systems analysed. Our approach of focusing on these bioenergetic systems overcomes the phylogenetic impasse that has previously complicated the search for mitochondrial ancestors. Moreover, our results provide a new perspective for experimentally re-evolving mitochondria from extant bacteria and in the future produce synthetic mitochondria.

**Citation:** Degli Esposti M, Chouaia B, Comandatore F, Crotti E, Sasseria D, et al. (2014) Evolution of Mitochondria Reconstructed from the Energy Metabolism of Living Bacteria. PLoS ONE 9(5): e96566. doi:10.1371/journal.pone.0096566

**Editor:** Hemachandra Reddy, Oregon Health & Science University, United States of America

**Received:** January 10, 2014; **Accepted:** April 7, 2014; **Published:** May 7, 2014

**Copyright:** © 2014 Degli Esposti et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This work has been partially supported by the project BIODESERT GA-245746 "Biotechnology from Desert Microbial Extremophiles for Supporting Agriculture Research Potential in Tunisia and Southern Europe" (European Union) and the Prin 2009 (grant 009L27YC8\_003), from the Italian Ministry of Education, University and Research (MIUR). Work at IIT has been sustained by intramural funds. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: mauro.degliestposti@iit.it

#a Current address: Dipartimento di Biologia e Biotecnologie, University of Pavia, Pavia, Italy

#b Current address: Department of Life and Reproduction Sciences, University of Verona, Verona, Italy

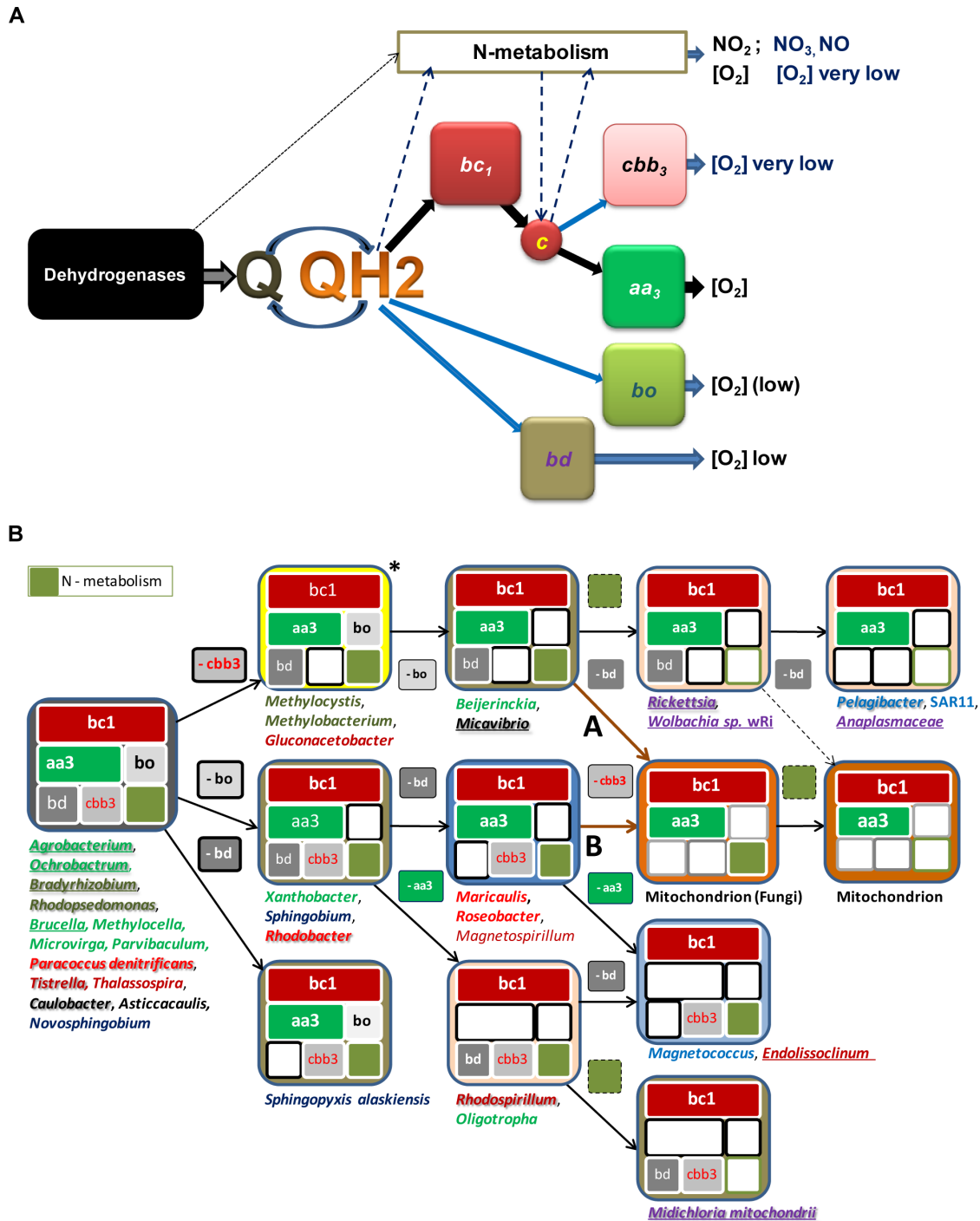
## Introduction

A major concept in biology is that the evolution of eukaryotic cell followed a symbiotic event between diverse microorganisms [1–4]. Mitochondria are the remnants of one of the original partners of this symbiotic event and in all likelihood are related to extant  $\alpha$ -proteobacteria [1–4]. However, the identity of the proto-mitochondrion remains elusive [1]. Phylogenetic studies suggested a relationship with endocellular parasites of the Rickettsiales order [4,5], which has not been confirmed in subsequent reports [6–8]. Indeed, there appears to be a "phylogenetic impasse" in the identification of the partners that merged into the ancestral symbiotic progenitor of current eukaryotic cells [9], partly due to the problem of long branch attraction blurring the true genealogy of living organisms and the fast evolution of mitochondrial DNA [1,10].

The diverse metabolic processes carried out by living bacteria provide complementary approaches to reconstruct key characteristics of the mitochondrial ancestors [11]. Although widely accepted, the reconstruction of proto-mitochondrial metabolism [12] has been partially contradicted by recent evidence suggesting that proto-mitochondria could be related to facultatively anaerobic generalists such as *Rhodobacter* [6–8,10] - which are also capable of

anoxygenic photosynthesis, an autotrophic function that must have been lost early along the evolution of mitochondria. Conversely, this evidence has recently been challenged by controversial reports that aerobic marine organisms such as *Pelagibacter ubique* may be the closest living relatives of mitochondria [13–15]. Other bacterial genera have also been considered to be phylogenetically related, or to display some analogies to the proto-mitochondrion: *Rhodospirillum* on the basis of extensive protein analysis [16]; *Paracoccus* for bioenergy considerations [1], and more recently following the evolution of complex I [17]; *Caulobacter*, on the basis of the sequence similarity of its homologues to the mitochondrial transport protein Tim44 [18]; *Micavibrio*, for its predatory ectoparasite character [19]; the Rhizobiales, *Ochrobactrum* and *Rhodopseudomonas*, for having many proteins in sister position to their mitochondrial homologues [6–8,20]; and finally *Mitichloria*, which appears to be the sole representative of the Rickettsiales retaining ancestral features typical of free-living bacteria [21]. The wide diversity of the proposed bacterial ancestors of mitochondria arises from the different approaches of molecular evolution that have been used and the inherent limits of such approaches [1–4].

This work follows a novel approach to identify proto-mitochondrial relatives among extant organisms by focusing on



**Figure 1. Bioenergetic systems of bacteria and mitochondria. A -Terminal respiratory chain of bacteria. 11.** Various bioenergetic systems - membrane redox complexes identified by their common name and different colours - carry out the oxidation of quinols (QH<sub>2</sub>) reduced by dehydrogenases. Besides oxygen (O<sub>2</sub>), nitrogen compounds can function as electron acceptors for the oxidation of dehydrogenases (dotted arrow), quinols and cytochrome c (dashed dark blue arrows), in reactions catalysed by enzyme complexes such as *Nrf* nitrite reductase [32], which are included within the N-metabolism system. Thick black arrows indicate electron transport in aerobic bacteria and mitochondria. Blue arrows indicate other electron transport pathways of facultatively anaerobic bacteria. **B - Pathways of mitochondrial bioenergetic evolution.** The bioenergetic systems illustrated in A are indicated by the coloured modules (with size proportional to their bioenergetic output) within the boxes representing the bioenergetic subset of each organism or organelle. Mitochondria of fungi and heterokont microorganisms differ from those of other eukaryotes for the presence of elements of N-metabolism. Representative taxa with fully sequenced genome are listed beneath each subset. The pathways of mitochondrial evolution are deduced by connecting these subsets with stepwise loss of a single bioenergetic system. Microorganisms underlined are symbionts or pathogens. Bacteria in embossed typeface have been proposed as ancestors or relatives of mitochondria (see Table S1 in File S1 for specific references). Dark brown arrows A and B indicate the pathways leading to fungal mitochondria. The pathway between the *Rickettsia* subset and that of mitochondria (dashed arrow) can be discounted, since the symbiotic event occurred only once [1,5,6,10,48]. \* indicates the subset from which other pathways depart (Figure S1 in File S1). doi:10.1371/journal.pone.0096566.g001

**Table 1.** Elements of N-metabolism that are shared by bacteria and eukaryotes.

Taxonomic group and organism	NAD(P)H dependent, assimilatory			PQQ-dehydrogenase
	<i>NirB</i>	<i>NirBD</i>	<i>NiaD</i> -related proteins	<i>MxaF</i>
<b>methanotrophs &amp; methylotrophs</b>				
<i>Methylocystis</i> sp. SC2	yes		1 domain	yes
<i>Methylocystis parvus</i>			precursor & 1 domain	yes
<i>Methylosinus trichosporium</i> OB3b	yes		1 domain	yes
<i>Methylosinus</i> sp. LW4			1 domain	yes
<i>Methylocella silvestris</i> BL2	yes		1 domain	yes
<i>Beijerinckia indica</i> *		yes	precursor & 2 domains	yes
<i>Microvirga</i> sp. WSM3557	yes			yes
<i>Methylobacterium extorquens</i> DM4			3 domains	yes
<i>Methylobacterium extorquens</i> PA1			3 domains	yes
<i>Methylobacterium extorquens</i> AM1			2 domains	yes
<i>Methylobacterium extorquens</i> CM4				yes
<i>Methylobacterium extorquens</i> DSM 13060				yes
<i>Methylobacterium nodulans</i> ORS 2060			2 domains	yes
<i>Methylobacterium populi</i> BJ001			2 domains	yes
<i>Methylobacterium radiotolerans</i> JCM 2831			2 domains	yes
<i>Methylobacterium mesophilicum</i> SR1.6/6			2 domains	yes
<i>Methylobacterium</i> sp. GXF4			2 domains	yes
<i>Methylobacterium</i> sp. 88A			2 domains	yes
<i>Methylobacterium</i> sp. 4–46				yes
<i>Xanthobacter autotrophicus</i> Py3	yes			yes
<i>Hyphomicrobium denitrificans</i> 1NES1	yes			yes
<b>Bradyrhizobiaceae</b>				
<i>Nitrobacter winogradskyi</i> Nb-255	yes			
<i>Nitrobacter hamburgensis</i> X14	yes			
<i>Nitrobacter hamburgensis</i> sp. Nb-255	yes			
<i>Oligotropha carboxidovorans</i> OM4 & OM5	yes			
<i>Rhodopseudomonas palustris</i> BisA53			2 domains	yes
<i>Rhodopseudomonas palustris</i> BisB18			1 domain	yes
<i>Rhodopseudomonas palustris</i> TIE-1			2 domains	
other 4 <i>Rhodopseudomonas palustris</i>			1 domain	
<b>Rhodospirillales</b>				
<i>Granulibacter bethesdensis</i> CGDNIH1	yes		2 domains	yes
<i>Commensalibacter intestini</i> A911	yes			
<i>Acidocella</i> sp. MX-AZ02	yes		1 domain	
<i>Acidiphilium multivorum</i> AIU301	yes			
<i>Acidiphilium cryptum</i> & sp. PM	yes		1 domain	
<i>Gluconobacter oxydans</i> H24		yes	precursor & 2 domains	
<i>Gluconobacter frateurii</i> NBRC 103465		yes	precursor	
<i>Gluconacetobacter oboediens</i> 174Bp2		yes	precursor & 2 domains	
<i>Acetobacter pasteurianus</i> IFO 3283-01/32		yes	precursor	
<i>Acetobacter aceti</i>		yes	precursor & 1 domains	
<i>Gluconacetobacter europaeus</i> LMG 18494		yes	precursor	
<i>Gluconacetobacter diazotrophicus</i> PA15			2 domains	
<i>Acetobacter pomorum</i> DM001		yes		
<i>Acetobacter tropicalis</i> NBRC 101654		yes		
<i>Asaia platicody</i>		yes	precursor	
<i>Saccharibacter</i> sp.	yes		2 domains	

Table 1. Cont.

Taxonomic group and organism	NAD(P)H dependent, assimilatory			PQQ-dehydrogenase
	<i>NirB</i>	<i>NirBD</i>	<i>NiaD</i> -related proteins	<i>MxaF</i>
<i>Tistrella mobilis</i> KA081020–065	yes		2 domains	
<i>Azospirillum lipoferum</i> 4B	yes		1 domain	yes
<i>Azospirillum amazonense</i> Y2	yes			
<i>Azospirillum brasilense</i> Sp245	yes			
<i>Azospirillum</i> sp. B510	yes			
<i>Caenispirillum salinarum</i> AK4	yes			
<i>Thalassospira profundimaris</i> WP0211	yes			
<i>Thalassospira xiamenensis</i> M-5	yes			
<i>Magnetospirillum magneticum</i> AMB-1	yes			
<i>Magnetospirillum</i> sp. SO-1	yes			
<i>Magnetospirillum gryphiswaldense</i> MSR-1	yes			
<b>Rhodobacterales</b>				
<i>Oceanicola granulosis</i>			1 domain	
<i>Oceanicola</i> sp. S124	yes			
<i>Octadecabacter antarcticus</i> 307	yes			
<i>Paracoccus denitrificans</i> PD1222	yes			
<i>Roseobacter denitrificans</i> OCh114	yes			
<i>Roseobacter litoralis</i> Och 149	yes			
<i>Jannaschia</i> sp. CCS1	yes			
<b>Rhizobiales (other)</b>				
<i>Marteella mediterranea</i>			precursor	
<i>Aureimonas ureilytica</i>			2 domains	
<i>Sinorhizobium meliloti</i> 1021	yes		2 domains	
<i>Rhizobium leguminosarum</i> bv. <i>trifolii</i> WSM1325	yes			
other 32 Rhizobiales	yes			
<b>Sphingomonadales &amp; Caulobacterales</b>				
<i>Novosphingobium nitrogenifigens</i>			precursor	
<i>Sphingomonas</i> sp. 17			2 domains	
<i>Sphingomonas</i> sp. PAMC26621			1 domain	
<i>Sphingopyxis alaskensis</i> RB2256	yes			
other 19 Sphingomonadales & 6 Caulobacterales	yes			
<b>total <math>\alpha</math>-proteobacteria</b>	<b>ca. 100</b>	<b>10</b>	<b>12 precursors</b>	
<b>Eukaryotes</b>				
<i>Aspergillus fumigatus</i>		yes	yes	
other 130 fungi (predominantly Ascomycetes)		yes	yes	
<i>Ectocarpus silicosus</i>		yes	yes	
plus other 8heterokonts	(1 yes)	yes	yes	
<i>Aureococcus anophagefferens</i>		yes	yes & 2 domains	
<i>Acanthamoeba castellanii</i>			yes	
<b>total Eukaryotes</b>	<b>1</b>	<b>140</b>	<b>141</b>	

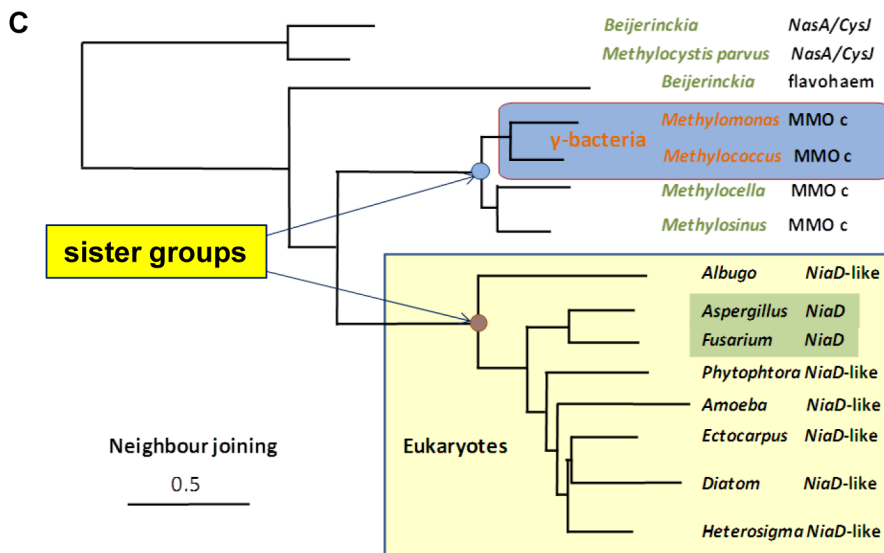
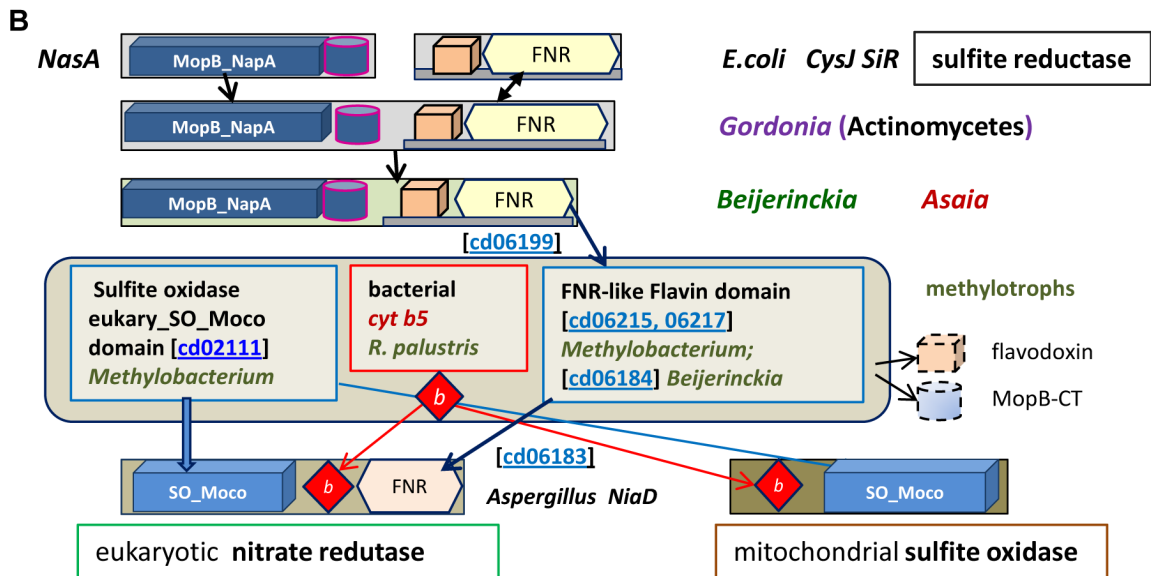
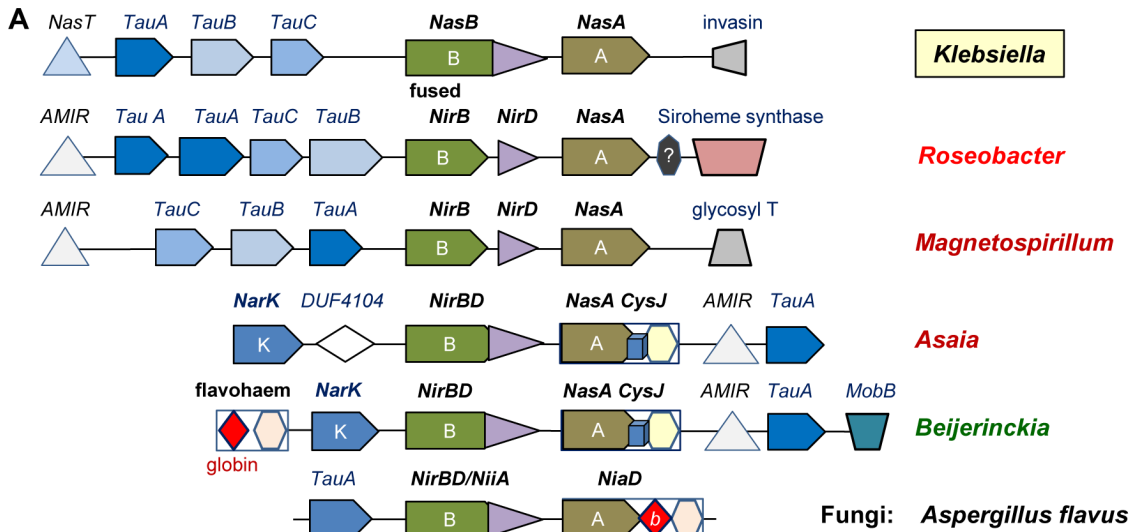
Proteins closely related to *NirB*, *NirBD*, *NiaD* and *MxaF* are annotated as **yes**, or **precursor** in the case of *Nas/CysJ* nitrate reductase (Fig. 2). The column of *NiaD*-related proteins also lists the number of *NiaD* **domains** that have homologues proteins in each organism, e.g. flavohaem (cf. Fig. 2C).

\*Its close relative *Beijerinckia mobilis* has been reported to grow on methanol and possess *MxaF*.

doi:10.1371/journal.pone.0096566.t001

the bioenergetic systems that are common between mitochondria and bacteria. An enormous increase in bioenergy production constitutes the major advantage gained in the endosymbiotic event that led to the evolution of eukaryotic cells [2]. Consequently, the mitochondrial systems that generate most cellular bioenergy must

define the minimal bioenergetic capacity of proto-mitochondria. Whereas aerobic  $\alpha$ -proteobacteria such as *Pelagibacter* present the same two bioenergetic systems of animal mitochondria [4,12], other proposed ancestors of mitochondria such as *Rhodospirillum rubrum* [6–8] possess four additional bioenergetic systems in their



**Figure 2. Graphical representation of assimilatory nitrate reduction in protists and  $\alpha$ -proteobacteria. A – The diagram shows the gene clusters of assimilatory, NAD(P)H-dependent nitrate reduction in bacteria and eukaryotes. The various elements of *Nas* operon of *Klebsiella* [36] and the *NiA-NiAD* operon in fungi [35] are colour coded as indicated in the quadrant on the top right. B – Possible molecular evolution of fungal *NiAD* nitrate reductase. Each domain is identified by a specific symbol - see the text for details. C – Representative distance tree of various proteins containing the bacterial FNR-like conserved domain. The tree was obtained with Neighbour Joining (maximal distance 0.9) using the DELTAST program [80] with methane monooxygenase subunit c of *Methylorella silvestris* (MMOC, Accession: YP\_002361598) as query. This reductase subunit of methane monooxygenase contains a FNR-like domain similar to that of assimilatory nitrate reductases [43] lying in a sister group as indicated. doi:10.1371/journal.pone.0096566.g002**

terminal respiratory chain (Fig. 1A). These systems are characteristic of bacteria living under anaerobic or micro-oxic conditions, exploiting also bioenergy-producing elements of N-metabolism which are partially retained in some eukaryotic microorganisms [10,22,23]. It is thus likely that the current bioenergetic portfolio of mitochondria has evolved from a larger genomic endowment of bioenergetic systems which has been reduced via sequential loss.

We have reconstructed the possible pathways of this sequential loss leading to the bioenergetic systems of current mitochondria by evaluating all the genomes of  $\alpha$ -proteobacteria which are currently available. Results obtained with complementary approaches then converged in indicating that methylotrophic  $\alpha$ -proteobacteria could be the closest living relatives to proto-mitochondria, while excluding the majority of bacteria previously proposed as mitochondrial relatives.

## Results and Discussion

### 1.1 Reconstructed pathways of bioenergetic evolution of bacteria into mitochondria

The bioenergetic capacity of mitochondria has been instrumental in the evolution of eukaryotic cells and complex life forms [1–3]. It is generally assumed that proto-mitochondria had an aerobic energy metabolism equivalent to that of today's mitochondria [1,4,12], with the central part of the respiratory chain consisting of ubiquinol-cytochrome *c* reductase (the cytochrome *bc*<sub>1</sub> complex) and a single terminal oxidase, cytochrome *aa*<sub>3</sub> oxidase (Fig. 1A). However, geophysical evidence indicates that proterozoic oceans were essentially anoxic during the period in which the eukaryotic cell evolved [24]. Consequently, it is likely that proto-mitochondria were adapted to different levels of environmental oxygen, exploiting also the terminal oxidases of facultatively anaerobic bacteria to obtain bioenergy [10]. For example, *Rhodospirillum rubrum* strains possess cytochrome *bd* and *bo* ubiquinol oxidases [25,26], plus an additional cytochrome *c* oxidase of the *cbb*<sub>3</sub> type [27] (Fig. 1B). Endocellular parasites have the *bd* ubiquinol oxidase either alone (in several species of *Rickettsia* [28]) or together with *cbb*<sub>3</sub> oxidase (in *Midichloria mitochondrii* [21]). Other organisms, moreover, possess proteins of the anaerobic bioenergetic process of denitrification, which are found also in mitochondria of fungi that can adapt to anaerobiosis [10,23,29].

Fungi and heterokont protists additionally possess an assimilatory nitrite reductase which is involved in ammonia fermentation, *NirB* fused with *NirD* [23,29] – hereby defined as *NirBD*. In some bacteria, this NAD(P)H-dependent enzyme forms part of the nitrogen cycle that enables their growth from the oxidation of methane or ammonia, the oxidation of C1 compounds such as methanol (methylotrophy) and ammonification of nitrite [30–32]. Because various elements of this nitrogen cycle are associated with bioenergy production [23,29–32], we have considered them within the broad bioenergetic system of N-metabolism (Fig. 1).

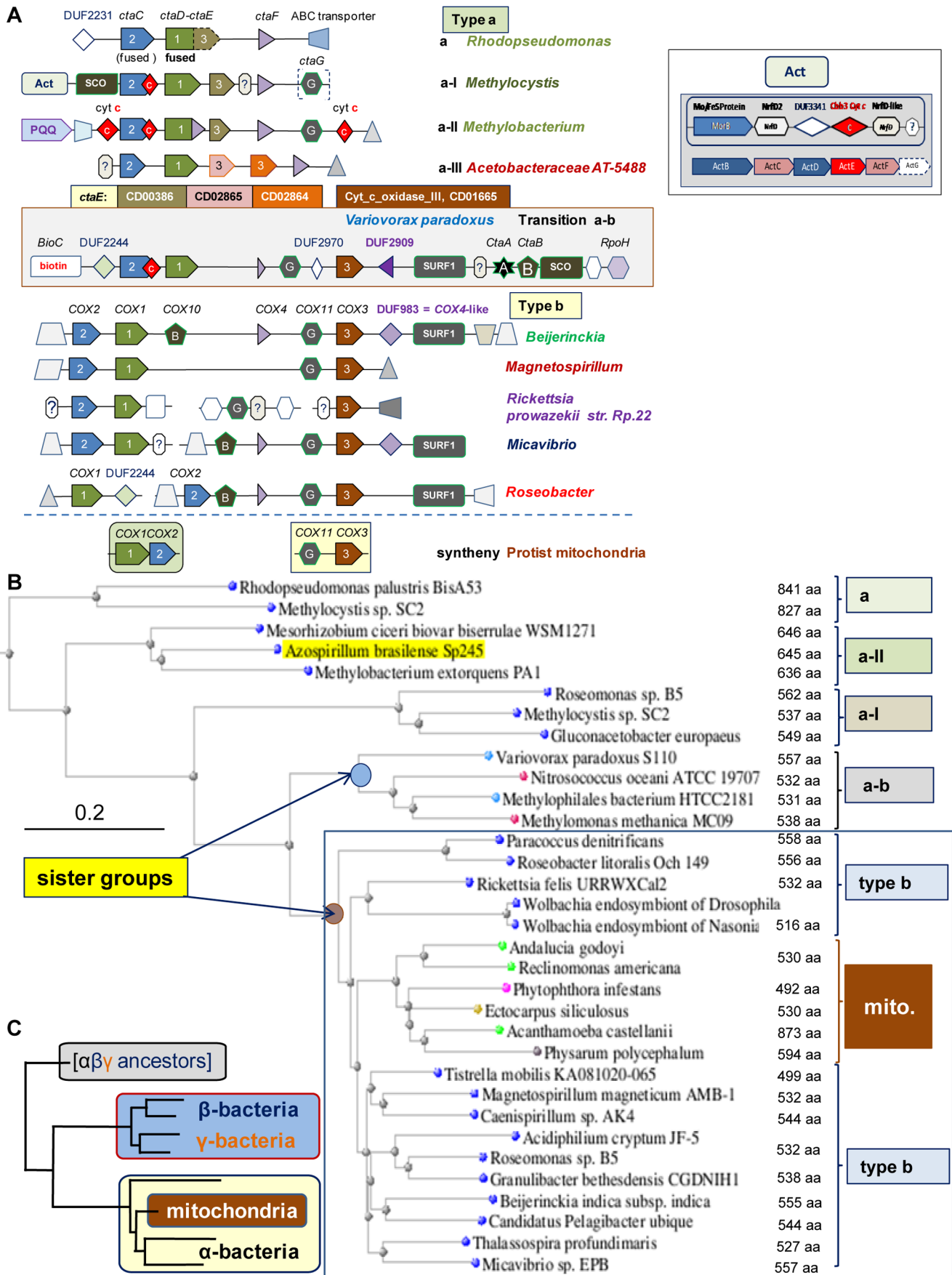
The metabolic versatility of current bacteria suggests that the ancestors of  $\alpha$ -proteobacteria had six bioenergetic systems from ubiquinol to oxygen (Fig. 1B), like diverse extant bacteria (Table S1 in File S1). To deduce the pathways of differential loss

that led to the reduced subset of current mitochondria, we have developed a model based upon the bioenergetic systems coded in all available genomes of  $\alpha$ -proteobacteria, including those we have recently sequenced (*Asaia platycody* and *Saccharibacter sp.* [22]). For parsimony, we allowed only single-step connections between the various subsets, thus obtaining two alternative pathways which directly lead to the subset of bioenergetic systems that is present in contemporary mitochondria of fungi and protists (Fig. 1B, cf. Fig. S1 in File S1). Pathway A stems from the subset present in predatory *Micavibrio* [19] and also *Beijerinckia indica*, a metabolically versatile organism closely related to methylotrophs [33] which has been shown to possess several proteins strongly related to their mitochondrial homologues [8]. Alternative pathway B originates from the subset present in some *Magnetospirillum* species and two Rhodobacterales (Fig. 1B): *Roseobacter litoralis*, which retains a functional photosynthetic apparatus, and *Maricaulis maris*, which has a dimorphic biological cycle. The loss of N-metabolism from the *Micavibrio/Beijerinckia* subset leads to the subset of *Rickettsia* [28] and *Wolbachia* organisms which retain the *bd* ubiquinol oxidase system (Fig. 1B). The loss of this bioenergetic system would also lead to the subset of metazoan (but not fungal) mitochondria, a possibility considered unlikely in view of the unique symbiotic event producing mitochondria [1,2,10]. Moreover, it occurs in related species of the same Rickettsiales order (Fig. 1B) and other taxa, for example within the *Bartonella* genus (Fig. S1 in File S1), suggesting phenomena of convergent evolution.

### 1.2 Testing the alternative pathways for mitochondrial bioenergy evolution

So, comparative genomic analysis has allowed a reconstruction of two possible reductive pathways in the bioenergetic capacity of bacteria evolving into mitochondria (Fig. 1). How can we establish which of these pathways is most likely, and thus identify extant models for proto-mitochondria? Probabilistic approaches based upon the frequency of gene loss from each subset would not produce conclusive evidence, because of the biased phylogenetic distribution of available bacterial genomes. We have then carried out the classical phylogenomic approach of computing the overall relationships of the organisms in the model of Fig. 1B by using concatenated proteins that are common to most eubacteria (cf. Ref. [21]). Although the obtained trees could be globally consistent with the sequence of either pathway A or B, they did not offer discriminatory evidence in favour of one or the other, while consistently placing *Midichloria* and other Rickettsiales close to the mitochondrial clade. This tree topology has been reported before [1,4,5,21] but is inconsistent with our new model of Fig. 1B and other evidence [1], as discussed above.

We next followed the alternative approach of exploiting the molecular diversity of key bioenergetic proteins, including their multiple duplication [34]. To enhance the discriminatory power of this approach, we have chosen proteins of energy metabolism that have a clear bacterial origin, but are encoded or located in different compartments of eukaryotic cells (cf. [34]). The hypothesis underlying our approach is that such diverse proteins,





**Figure 3.  $\alpha$ -proteobacteria have different types of COX operons and catalytic subunits of  $aa_3$  oxidase. A - Graphical representation of  $aa_3$  oxidase gene clusters.** The different COX clusters of  $\alpha$ -proteobacteria are classified by considering gene sequence variations and the features of flanking genes (see also “Classification of bacterial COX operons” in File S1). Specific graphical symbols identify COX subunits as indicated; other types of proteins are labelled as follows: white hexagon, enzyme working with RNA or DNA; red diamond with enclosed c, cytochrome c type protein; truncated triangle pointing left, ABC transporter/permease; grey sharp triangle, transcription regulator; PQQ, PQQ-dependent dehydrogenase; white diamond, protein belonging to a DUF family [41], e.g. DUF983; question mark within hexagon, completely unknown protein. Note that *SURF1* (Surfeit locus protein 1) and *SCO* (Synthesis of cytochrome c oxidase) are also involved in the biogenesis of oxidases. Distance between genes is arbitrary. COX operon **type a-I** is attached to a *Nrf*-like gene cluster, also called Alternative Complex III or Act [50], containing two homologues of the membrane subunit *NrfD* (called *NrfD2* and *NrfD*-like here, as shown at the side of the figure). The synthetic diads of protist mitochondria [48] are shown below the blue line. Each of the recognised subfamilies of COX3 [41] is represented by a different colour, as indicated in the middle of the illustration. **B - Representative distance tree of COX1 proteins.** The tree was obtained with Neighbour Joining (maximal distance 0.9) using the DELTABLAST program [80] with the COX1 protein of *Methylobacterium extorquens* PA1 (Accession: YP\_001637594) as query. The group containing bacterial and mitochondrial proteins (mito.) is enclosed in the blue square. Protein length and type of COX operon are annotated on the right of the tree. **C - Simplified pattern of typical phylogenetic trees of COX1 proteins.** The tree is modelled to match distance trees of nitrate reductase (Fig. 2C) and COX1 (part B). Branch length is arbitrary. doi:10.1371/journal.pone.0096566.g003

as well as their genetic clusters, would present transition forms between bacteria and mitochondria predominantly in those organisms that are close to the proto-mitochondrial lineage.

## 2. Molecular evolution of assimilatory N metabolism

The first bioenergetic system we considered is N metabolism, the presence or absence of which sharply determines the pathways leading to the mitochondria of fungi and metazoans (Fig. 1B). As mentioned above, fungi and heterokonts possess the assimilatory, NAD(P)H-dependent nitrite reductase *NirBD* [35], a cytosolic enzyme which is common among facultatively anaerobic  $\gamma$ -proteobacteria such as *Klebsiella*, where it was originally called *NasB* [36]. Structurally, *NirBD* is characterised by the fusion of the small protein *NirD* - belonging to the Rieske superfamily of Fe-S proteins coordinated by histidines and cysteines [37] - at the C-terminus of the *NirB* protein, which catalyses the reduction of nitrite and is structurally related to sulfite reductase (*Sir*) [38]. Interestingly, the distribution of *NirB* is restricted to a relatively narrow group of facultatively anaerobic bacteria [38,39], but that of *NirBD* is much narrower (Table 1). After finding *NirBD* in the genome of *Asaia*, we detected only ten homologous genes among  $\alpha$ -proteobacteria - compared with over one hundred in fungi (Table 1), all arranged in similar gene clusters comprising a regulator, nitrate transporters and an assimilatory nitrate reductase. The gene clusters are related to the *Nas* operon of *Klebsiella* (Fig. 2A), with its most compact version being present in fungi and Oomycetes [35].

Among the bacteria associated with pathway A and B in Fig. 1B, only *Beijerinckia* possesses *NirBD* and its cognate gene cluster. *Roseobacter litoralis* and *Magnetospirillum* have *NirB* within an operon similar to that of *Klebsiella* (Fig. 2A), whereas *Maricaulis* and *Micavibrio* do not have the same genes. This situation may well arise from secondary loss of metabolic traits in ecologically specialised organisms such as dimorphic *Maricaulis* and predatory *Micavibrio*. To gain further phylogenetic information, we then exploited the rare occurrence of *NirBD* and its associated nitrate reductase among  $\alpha$ -proteobacteria (Table 1), evaluating the molecular evolution of these modular proteins. The structure of *NirBD* is conserved in  $\alpha$ -proteobacteria and eukaryotes [35] and apparently derives from *NirB* precursors that are present in methylotrophs such as *Methylocystis* (Fig. 2, cf. [35]).

Conversely, the structure of the large protein functioning as nitrate reductase in the *NirBD* gene cluster of  $\alpha$ -proteobacteria resembles that of nitrate reductases from ancient bacteria such as *Gordonia*, which contains three redox modules formed by distinct domains. A typical Molybdenum cofactor-binding domain (Moco) occupies the N-terminus and includes a terminal part binding another molybdopterin cofactor as in *NapA* (periplasmic) and *NasA* (cytoplasmic) reductases [36–40]. This is followed by an intermediate domain homologous to the small redox protein flavodoxin

(Fig. 2B top, cf. [38]). The C-terminus then contains a flavoprotein reacting with the electron donor NAD(P)H which, in combination with flavodoxin, forms a domain closely related to sulfite reductase *CysJ* of *E.coli* (represented by a grey bar in Fig. 2B, cf. [38]). The *CysJ*-related domain belongs to the superfamily of Ferredoxin Reductase-like domains, cd 00322 FNR-like [41], which includes also the C-terminal domain of fungal nitrate reductase, *NiaD* [35,40].

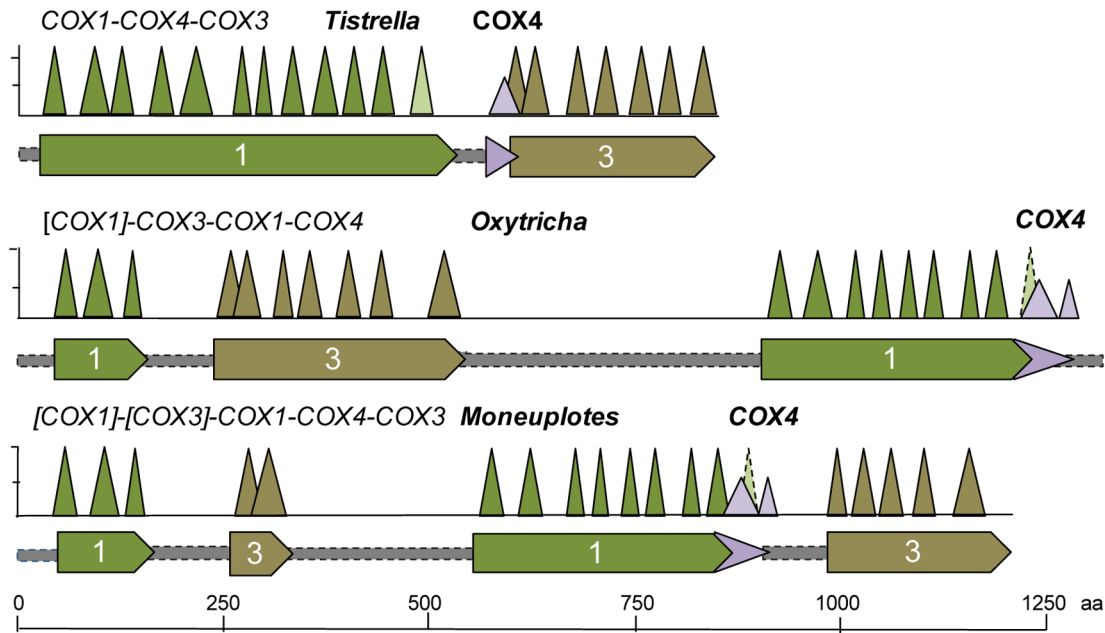
Although the fine structure of the FNR-like domain indicates two separate subfamilies, cd01699 *Sir*\_like for the *NasA/CysJ* bacterial proteins and cd06183 *cytb5\_reductase\_like* for the eukaryotic proteins, our detailed sequence comparison uncovered phylogenetic relationships with other bacterial proteins belonging to the same superfamily. In particular, flavodoxin reductases of the genus *Methylobacterium* and the reductase subunits of soluble methane monooxygenase [42,43] (MMO, present also in close relatives of *Beijerinckia* such as *Methylocella*) were consistently found in sister clades to *NiaD* and related proteins of fungi, heterokonts and *Acanthamoeba* (Fig. 2C and Table 1). Moreover, the flavohaem oxidoreductase of *Beijerinckia* (accession YP\_001833084), which contains a cytochrome *b*-related globin followed by a FNR-like domain, was found in an intermediate position between the *NiaD*-containing clade and the *NasA-CysJ* reductases of *Beijerinckia* and *Methylocystis parvus* (Fig. 2C). Notably, the gene of this protein is located at the beginning of *Beijerinckia* nitrate assimilation operon (Fig. 2A). Its Nitric Oxide dioxygenase activity is also similar to that of the hybrid nitrate reductase of microalgae from the heterokont group, e.g. *Chattonella subsalsa* (protein NR2-2/2HBnN, accession: AER70127), which possess both a cytochrome *b*<sub>5</sub> and a globin in the intermediate domain [44]. These flavoproteins, therefore, could be considered transition forms between *NapA/CisJ* reductases and eukaryotic assimilatory nitrate reductases.

In further support of the modular similarity between bacterial and eukaryotic NAD(P)H-dependent nitrate reductases, we have found that the Moco domain of *NiaD*-like eukaryotic proteins is present also in the sulfite oxidase of methylotrophs such as *Methylobacterium mesophilicum* and *extorquens* (accession: WP\_010685750 and WP\_003602739, respectively - Table 1 and Fig. 2B). Moreover, the genome of *Methylobacterium extorquens PA1* encodes a protein that is partially similar to bacterial cytochrome *b*<sub>5</sub> (accession: YP\_001638730), which is present only in *Rhodospseudomonas palustris* among  $\alpha$ -proteobacteria (Fig. 2B and data not shown). Consequently, all three functional domains of eukaryotic assimilatory reductases have homologous proteins in extant  $\alpha$ -proteobacteria, particularly among those with methylotrophic metabolism, as indicated by the presence of the signature methanol dehydrogenase *MxaF* [45] (Table 1). Hence, our data suggests that *NasA-CisJ* reductases of *Beijerinckia* and acetic acid bacteria, e.g. *Asaia*, represent the likely **precursors** of eukaryotic,

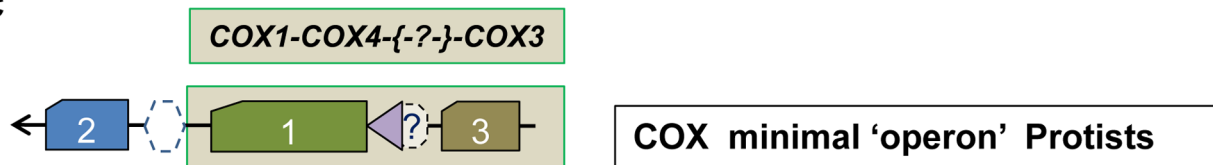
A

Beef mitos	COX3	8	TM1-----TM2-----	negative side
Paracoccus	COX3	13	PSI--WPLFGAIGAFVMLTGAVA--WMKGITFFGLPVEGFWMLIGLVGVLYVMFCWVADVVEGET-----GEHT--	
Andalucia	COX3	15	PSP--WPLFAAASVFSLVIGGVM--YM-----HSYENAN--LVFSSGLISLLYVMFVWRDIVERGT--YQGTYYQGHHTVE	
Physarum	COX3	11	PSP--WPISVSAALLGLTIGGV--SY-----FHSFNNGIYLLTSSFILLAILAGF--WRDLIREGTY-----LHNHTKE	
Thermus	COX1/3	552	PNSSFWFFYSAAITLFAFFVAVAA-----LPVNVWVWFLALFAYGLV--R--ALED--EYSHPVE-----HHTVT	
Oxytricha	COX1	247	-DLNSYIFKIVSYFNVTNVFFLLNYF--LFF--FLEKSFLLFISFSKIFNDLI--FISNFLYKIFYTI--ENF---	
Moneuplo_m	COX1	258	PSLSNSVFRFSFLTGCRIITFFYELVAVILFI--IKSLLFFFLPSTLITSASS--FFSLTWLRITSTVQGL-----COX1	
Beef mitos	COX3	73	TM3-----TM4-----	
Paracoccus	COX3	80	PAVQKGLRYGMILFIISEVLFFTFWFAFYHSSLAPT---PELGGCWPTGIHPLNPLEVPLLNTSVLLASGVISITWAHHSMLMEGTP-----	
Andalucia	COX3	80	PVVRIGLQYGFILFIMSEVMFVAFWFAFIKNALYPMGPDSPKDGWVWPEGLVTFDPWHLPLINTLILLSSGAVITWAHHAFLVLEG-----	
Physarum	COX3	76	--VQTGLRMVLLFIVSEAAFLFAFFWAFHSSSLAPNIEI---GSVWPLGIEPLSAWEVFPFLNTLILLSSGATVITWAHHAIVA--GYRKEAILAL	
Thermus	COX1/3	613	--VLLGLRLGFILFIVSEVMFFSFFWAYFHSSLSPNIEI---GSQWPPFALEVIGL-ALPVVNTVILLTSGATITVAHLAILRNKKQIAIESL-	
Oxytricha	COX1	313	-GKSNAWGMWAFIVSEVGLFALLIAGYLYRLSGAATP-----PEERPALWLALLNTFLLVSSSFTVHFAHH-DLR-----	
Moneuplo_m	COX1	975	PTLTNTIN-FEIN--LFSFVFFNMALACTLLIFARA--FPGFSIYFWP--YLCIIFNFLOQLKLFYICKLLKIFSTSFLEFQNYFF-- [7]	
Beef mitos	COX3	155	TM5-----TM6-----	negative side
Paracoccus	COX3	167	---DRKHMLQALFITITLGVYFTLLQAS--EYVEAP---FTISDGVYGSTFVATGFH--GLHVIIGSTFLIVCFPRQL--KFH-----	
Andalucia	COX3	168	---DRKTTINGLIVAVILGVCFITGLQAY--EYSHA---FGLADTVYAGAFYMATGFH--GAHVIIGTFLFVCLIRLL--KGQM-----	
Physarum	COX3	165	---ILTISLAAVFTALQIFEYATASFSLSDGIYGS-----TFMATGFH---GH--VFVGTCAVTCVLLRQIRYH-----	
Thermus	COX1/3	682	--IATIVLALVFITAIQMYEYRHAFPSISDGIYGS-----VF---YMLTGFH---GIH--VLIGTIFICVQFVRLTKDH-----	
Oxytricha	COX1	405	--RRGRFNPRFGLLVITILGVLFVLVQS--EYVQ-----FYHH--SSWQENLWTAAPFTIVGLHGLHVVIGFGGLILAYLQALRGK-----	
Moneuplo_m	COX1	1064	FFSNKNFLYNYFTFLSYFNSFFKYLGNESKFDPDLK--RFDINDFANNFGLNSYNIFKISTNNLYAVGSKTNIIRYMFKS-----	
Beef mitos	COX3	227	FISLFPREAREGSNRFLIFSF--FGLFAGIFSCFSNVETVVIIGLPFL-VTYYDWGIFENFYFTDLQLLSDIYVYVLAGLEFILMNFYLYLVILV	
Paracoccus	COX3	240	negative side TM7-----	
Andalucia	COX3	232	-----FTSNHFCFEAAA--WYWHFVDVWVLFVYSIYWWGS-----	C terminus 260
Physarum	COX3	227	-----TQKHVGFEEAAA--WYWHFVDVWVLFVVIYIWR-----	C terminus 273
Thermus	COX1/3	761	-----FTTSHHCFEAAA--WYWHFVDVWVLFVSIYWWGQ-----	C terminus 267
Oxytricha	COX1	477	-----LLSNHLCFEACA--WYWHFVDVWVLLFVIVYAYGSNAL-----	C terminus 264
Moneuplo_m	COX1	1197	-----ITLHNCLEAASMYWHLVDVWVIVITIFVW-----	C terminus 791
			.. COX1 & COX4? ---	1331
				C terminus 1203

B



C



**Figure 4. Analysis of the molecular architecture of COX3 in bacteria and protists. A – Alignment of bacterial and mitochondrial COX3 proteins.** A set of aligned COX3 sequences from bacteria and protists was initially obtained from the DELTBLAST option of multiple alignment and subsequently implemented manually following data available from the structure of beef [59,60], *Paracoccus* [61] and *Thermus* [54] *aa<sub>3</sub>* oxidase. Residues that bind phospholipids with either H or  $\pi$  bonds [60] are in yellow character and highlighted in dark grey, while those conserved are in bold character. Light grey areas indicate transmembrane helices (TM). **B – Graphical representation of COX1-3 fused proteins.** The hydrophobic peaks in the hydropathy profile of the proteins, which was obtained using the program WHAT [81] with a fixed scanning window of 19 residues, is represented by the sharp triangles, that are commensurated to the peak height (maximum in the hydrophobicity profile) and width of the predicted TM [81], which closely correspond to those observed in 3D-structures [47,54,61]. **C – Deduced sequence of the “minimal” COX operon of protists.** The arrangement of COX genes essentially corresponds to the core sequence of a COX operons of type a (cf. Fig. 3) but in the reverse order of transcription. Dashed symbol represents a protein that may intermix with other COX subunits such as a COX4-like (Fig. S2 in File S1). doi:10.1371/journal.pone.0096566.g004

*NiaD*-related nitrate reductase (Table 1 and Fig. 2B,C). The parallel evolution of mitochondrial sulfite oxidase, which shares the same cytochrome *b<sub>5</sub>* and Moco domains with eukaryotic assimilatory nitrate reductases (Fig. 2B, cf. [38,40]), underlines the intersection of this molecular reconstruction with the evolutionary trajectory of proto-mitochondria.

### 3. Evolution of COX genes and proteins from bacteria to mitochondria

To test alternative evolutionary pathways for mitochondria (Fig. 1B) we next studied the cytochrome *c* oxidase of *aa<sub>3</sub>*-type (also called COX), which appears to be the most common terminal oxidase in extant  $\alpha$  proteobacteria (Fig. 1 and Table S1). In eukaryotes, this enzyme complex is embedded in the inner mitochondrial membrane, combining catalytic subunits of bacterial origin with various nuclear-encoded subunits of unknown function. Although all *aa<sub>3</sub>*-type oxidases are of type A according to the classification of heme-copper oxygen reductases [26], the complexity of their gene clusters has not been considered before. Here, we have analysed in depth this complexity for it provides valuable phylogenetic information. Various aspects of our analysis are presented below in the following order: 1, diversity of COX operons; 2, evolution of COX operons; 3, possible COX operons of proto-mitochondria; 4, evolution of the molecular architecture of COX3; 5, phylogenetic distribution of COX operons.

**3.1 Diversity of COX operons.** We have initially undertaken a systematic analysis of the genomic diversity of *aa<sub>3</sub>*-type oxidases. The scrutiny of all the gene clusters containing proteobacterial COX subunits [46–51] suggests that they fall into three distinctive types of COX operons, which we called type a, b and a–b transition (Fig. 3A – see Table S2 and “Classification of bacterial COX operons” in File S1 for a detailed account of this classification). COX operon type a is divided in four subtypes on the basis of COX1 length and diverse adjacent genes (Fig. 3). These subtypes form coherent clades in the phylogenetic trees of their COX1 subunit (Fig. 3B). Despite the variation in gene sequence, all COX operons appear to derive from the core structure of the *ctaA-G* operon of *Bacillus subtilis* [46–51] (Fig. 3A), which consists of the catalytic subunits *ctaC* and *ctaD* (corresponding to mitochondrial COX2 and COX1, respectively) followed by the hydrophobic, non-catalytic subunit *ctaE* (corresponding to mitochondrial COX3) and *ctaF* (also called COXIV or COX4). Mitochondrial DNA (mtDNA) of eukaryotes generally encodes for COX1, COX2 and COX3 [48]. In bacteria, these principal subunits are often combined with proteins for the assembly of the metal cofactors of the oxidase: *ctaA* (heme A syntase or COX15), *ctaB* (protoporphyrin IX farnesyl transferase, or COX10) and *ctaG* (Cu-delivery protein, or COX11).

Our systematic analysis of bacterial COX subunits has revealed a novel fusion between COX1 and *ctaF*/COX4 (Fig. S2 in File S1). This fusion appears to be restricted to COX operon type a-II (Table S2 in File S1 and Fig. 3A) that often contains Pyrroloquinoline quinone (PQQ)-dependent dehydrogenases such as methanol dehydrogenase related to *MxaF* (Fig. 3A). COX4 is broadly related

to the *ctaF* subunit, which is the least conserved in the *caa<sub>3</sub>*-type oxidase of *Thermus* and *Bacillus* [47] but can be recognized as part of Cyt\_c\_ox\_IV (pfam12270 [52]). However, the diverse forms of short hypothetical proteins that intermix with COX subunits (Fig. 3A) are generally not recognized as members of this family in BLAST searches, due to the wide variation in their size and sequence [47]. Therefore, we have developed a method that quantifies the sequence similarity with the COXIV proteins from *Rhodobacter* [53] and *Thermus* [47,54], for which the 3D structure is available (see Fig. S2 in File S1 and its legend for details). Strong sequence similarity with these COX4 proteins was found in the C-terminal extension of bacterial COX1 proteins that are 630 to 670 aa long, as well as in mitochondrial COX1 of the pathogenic fungus, *Zygomoseptoria tritici* [55] (Fig. S2A in File S1). We additionally identified the sequence signatures of COX4 in small proteins previously recognized as domain with unknown function (DUF [52]) families, namely DUF2909 and DUF983 (Figs. 3 and S3 in File S1). Moreover, the C-terminal part of the mtDNA-encoded COX1 of ciliates, an ancient and diverse phylum of unicellular eukaryotes [56], shows some sequence similarity encompassing both transmembrane helices of COX4 proteins (Fig. S2A and B in File S1). Although this similarity is clearly weaker than that observed with bacterial COX1 proteins, it lies in a conserved region among ciliates (Fig. S2A in File S1 and data not shown) thereby suggesting that fusion of COX1 with COX4 might represent an additional trait shared by bacteria and mitochondria.

**3.2 Evolution of COX operons.** The identification of COX4-like proteins has been combined with phylogenetic analysis to deduce the possible evolution of COX operons. The long proteins derived from the fusion of COX1 with COX3 (hereafter called COX1-3) seem to be the most distant from their mitochondrial homologues (Fig. 3B). These proteins are characteristic of *caa<sub>3</sub>* oxidases [46,47], as well as of COX operon type a, which can therefore be considered the ancestral form of proteobacterial gene clusters for *aa<sub>3</sub>*-type oxidases (Fig. 3A). The differentiation into other types of COX operons can be evaluated also from the phylogenetic trees of the catalytic subunit COX1, the analysis of which has offered new evidence for discriminating the evolutionary pathways in Fig. 1B.

COX1 proteins fused with COX4 (see above) appear to follow the ancestral COX1-3 in phylogenetic trees and are always upstream of a major bifurcation in two large groups: one containing only proteins of COX operon a-b transition that are present in  $\beta$ - and  $\gamma$ -proteobacteria, and the other containing bacterial COX1 proteins of COX operon type b together with their mitochondrial homologues (blue square in Fig. 3B). Mitochondrial COX1 proteins cluster in a monophyletic clade that lies in sister position of closely packed bacterial sub-branches, especially that containing Rhodospirillales (Fig. 3B). This overall tree topology is consistently found with all methods, whereas the branching order within the group containing the mitochondrial clade may vary, depending upon the method and taxa used to construct the phylogenetic trees (Fig. 3B and data not shown). Nevertheless, it is noteworthy that all the

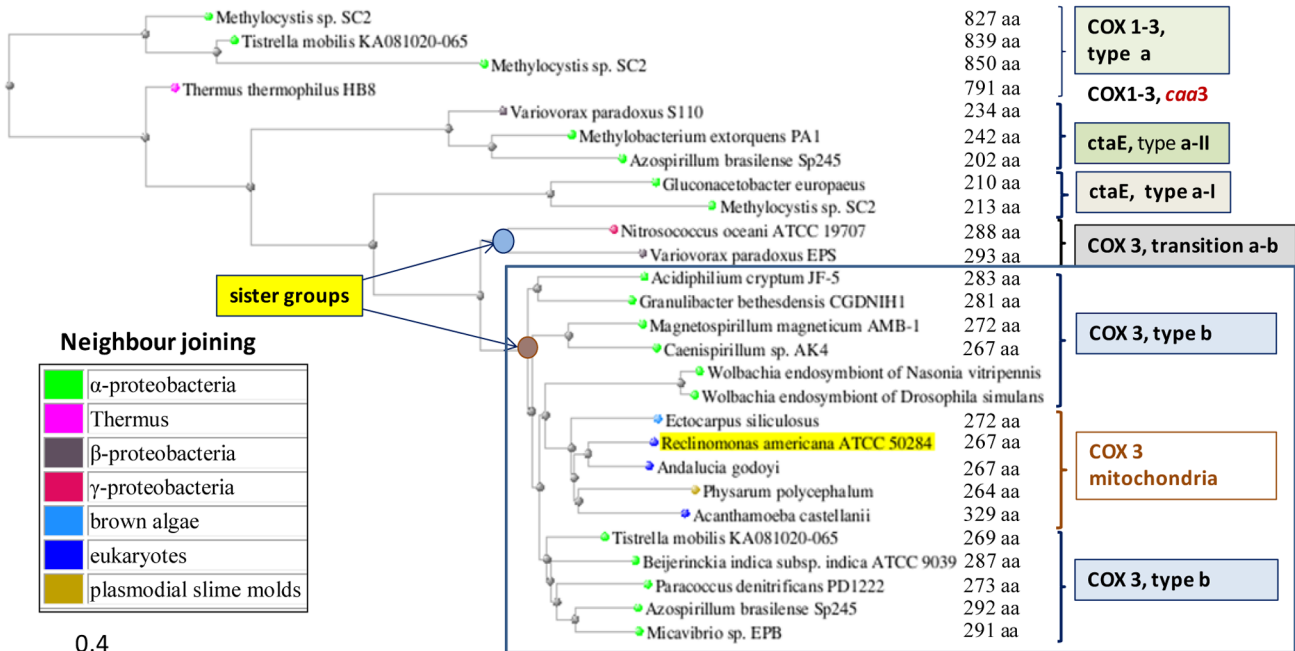
A

Organism and protein	protein	COX operon	length aa	spans TM	PL binding sites				
					E90	PG1	PG2	PE2	PE1
<i>Methylocystis</i> sp. SC2	COX1-3	a	827	6(7)		2 sites	0 site	1 site	
<i>Rhodo_palu_BisA53</i>	COX1-3	a	841	6		1 site	1 site	2 sites	
<i>Methylocystis</i> sp. SC2	COX3a	a-III	236	6		2 sites	3 sites	2 sites	
<i>Variovorax_par</i> EPS	COX3b	a-III	234	6	D	2 sites	>3 sites	2 sites	
<i>Roseomonas</i> sp. B5	COX3	a-I	217	5	yes	2 sites	>3 sites	3 sites	
<i>Gluconacetobacter_europ</i>	COX3	a-I	210	5	yes	3 sites	>3 sites	3 sites	
<i>Variovorax_par</i> S110	COX3	a-II	201	5	yes	2 sites	3 sites	2 sites	
<i>Methylocystis</i> sp. SC2	COX3	a-II	242	5	yes	2 sites	3 sites	2 sites	
<i>Methylophi_bac</i>	COX3	transition a-b	299	7	yes	>3 sites	3 sites	2 sites	
<i>Nitrosococcus_oceani</i>	COX3	transition a-b	288	7	yes	3 sites	>3 sites	>3 sites	
mitochondria, alveolata									
<i>Plasmodium_yoelii</i>	COX3	mtDNA	263	6	yes	2 sites	3 sites	1 site	
<i>Monoeuplotes_minuta</i>	COX1-3	mtDNA	1203	2&5	S	1 site	3 sites	>3 sites	
<i>Oxytricha_trifallax</i>	COX1-3	mtDNA	1331	6(7)	S	2 sites	>3 sites	2 sites	
mitochondria									
<i>Beijerinckia_indica</i>	COX3	b	287	7	yes	>3 sites	>3 sites	2 sites	
<i>Caenispirillum</i>	COX3	b	275	7	yes	>3 sites	>3 sites	>3 sites	
<i>Wolbachia_Dro_sim</i>	COX3	b broken	275	7	yes	>3 sites	>3 sites	>3 sites	
<i>Paracoccus</i>	COX3	b broken	273	7	yes	>3 sites	>3 sites	>3 sites	
mitochondria									
<i>Dictyostelium</i>	COX3	mtDNA	435	7	yes	3 sites	2 sites	3 sites	
<i>Acanthamoeba</i>	COX3	mtDNA	329	7	yes	>3 sites	>3 sites	3 sites	
<i>Andalucia</i>	COX3	mtDNA	267	7	yes	>3 sites	>3 sites	3 sites	
<i>Bos taurus</i> (beef)	COX3	mtDNA	260	7	yes	>3 sites	>3 sites	>3 sites	

Legend

binding	
conserved sites	strength
>3 sites	strong
3 sites	weak
2 sites	very weak
1 or 0 site	none

B



**Figure 5. Structure-function features of COX3 gradually evolved from bacteria to mitochondria. A – Heatmap for the strength of phospholipid binding by COX3 proteins.** The table summarises the molecular features of PL-binding sites (residues) in aligned COX 3 proteins (Table S4 in File S1); it is colour mapped according to the number of conserved sites to represent the increasing PL-binding strength along bacterial and mitochondrial protein sequences, as indicated by the legend on the right of the table. PL-binding is considered weak when less than 3 sites are

conserved for each PL, the nomenclature of which is taken from Ref. [60]. PE, phosphatidyl-ethanolamine; PG, phosphatidyl-glycerol. The list includes conserved amino acids corresponding to E90 in beef *COX3*, which lies near bound PL modulating oxygen entry into the catalytic site of the oxidase [60]. Abbreviations for organisms are: *Rhodo\_palu\_BisA53*, *R. palustris* BisA53; *Variovorax\_par*, *Variovorax paradoxus*; *Methylophi\_bac*, *Methylophilales bacterium* HTCC2181; *Wolbachia\_Dro\_sim*, *Wolbachia* endosymbiont of *Drosophila simulans*. **B - Representative distance tree of COX 3 proteins.** The tree was obtained as described in the legend of Fig. 3B, using as a query the C-terminal region of the *COX1-3* protein from *R. palustris* BisA53 (Accession: YP\_782773, residues 550 to 841) that aligns with bacterial and mitochondrial *COX3* (Fig. 4A). The group containing bacterial proteins from *COX* operon type b and their mitochondrial homologues is enclosed in a blue square as in Fig. 3B. doi:10.1371/journal.pone.0096566.g005

proteins belonging to *COX* operon type b lie in the same group containing the mitochondrial clade, as exemplified in Fig. 3C. Hence, bacteria having only *COX* operon type b cannot be the ancestors of mitochondria. This exclusion encompasses the majority of extant  $\alpha$ -proteobacteria, because the presence of other *COX* operons is restricted to a fraction of these organisms (Table S2 in File S1). We then needed additional information to identify which of the organisms containing multiple *COX* operons may be close to proto-mitochondria. To this end, we next moved to the analysis of *COX* proteins of unicellular eukaryotes.

**3.3. Possible COX operons of proto-mitochondria.** Recently, *COX11* and *COX15* have been found in the mtDNA of Jakobida, an ancient lineage of protists, despite the fact that they are normally coded by nuclear DNA in eukaryotes [48]. The syntenic *COX11COX3*, as well as that of *COX1* adjacent to *COX2* (Fig. 3A), may be considered a relic of bacterial operons that has been retained in the mDNA of eukaryotes [48]. Are these cues pointing to the original *COX* operon(s) of proto-mitochondria?

To answer this question, we searched the available mtDNA genomes of unicellular eukaryotes. Mitochondrial DNA normally contains separate genes for *COX1*, *COX2* and *COX3* [48] except for aerobic ciliates, in which *COX3* appears to be missing [56,58]. However, we have recognized the sequence signatures of the *COX3* protein within the very long *COX1* of the hypotrichous ciliate, *Oxytricha* [56] (Fig. 4). The *COX1* protein of another hypotrich, *Monoeploptes minuta* [58], appears to contain a split version of *COX3* having its initial two transmembrane helices separated from the subsequent 5-transmembrane helices domain by the major part of *COX1* (Fig. 4). The mtDNA of ciliates often contains split genes [56,58], but in this case an ancestral splitting of *COX3* must have been subsequently intermixed with the *COX1* gene. The alternative possibility would be that *COX3* splitting may reflect a fusion between precursors of mitochondrial *COX3*, since in *Monoeploptes* it occurs within the region joining the two transmembrane domains which form the V-shaped structure of the protein [53,59-61].

In any case, the novel identification of a *COX3*-like protein embedded within the long *COX1* gene of unicellular eukaryotes (Fig. 4) suggests that the primordial form of such a chimaeric gene was a *COX1-3* protein equivalent to those of bacterial *COX* operons of type a. By considering the gene order in ciliate mtDNA [56,58], we have deduced the possible sequence of the “minimal” *COX* operon that might have been present in the ancestors of ciliate mitochondria (Fig. 4C). The gene sequence closely resembles the core structure of a *COX* operon of type a - in the opposite order of transcription (cf. Fig. 3A and 4C) - and is clearly different from the sequence of *COX* operon type b (Figs. 3A and S3 in File S1). In view of the consensus that a single event of symbiosis originated all mitochondria [1-10] and considering the presence of *COX11COX3* synteny in Jakobide mitochondria [48], a feature characteristic of *COX* operon type b (Figs. 3 and S3 in File S1), we surmise that proto-mitochondria possessed two different *COX* operons: one of type a and another of type b. Differential loss of either operon might further explain some differences in the mtDNA-coded proteins of ciliates and other unicellular

eukaryotes, as well as the different types of accessory subunits of their bioenergetic complexes [1]. Of note, phenetic analysis sustains the similarity between the *COX* gene sequence of protists and bacterial *COX* operon of type a-II, in particular those lacking an isolated *COX4* as in *Methylobacterium extorquens* PA1 (Table S3 in File S1).

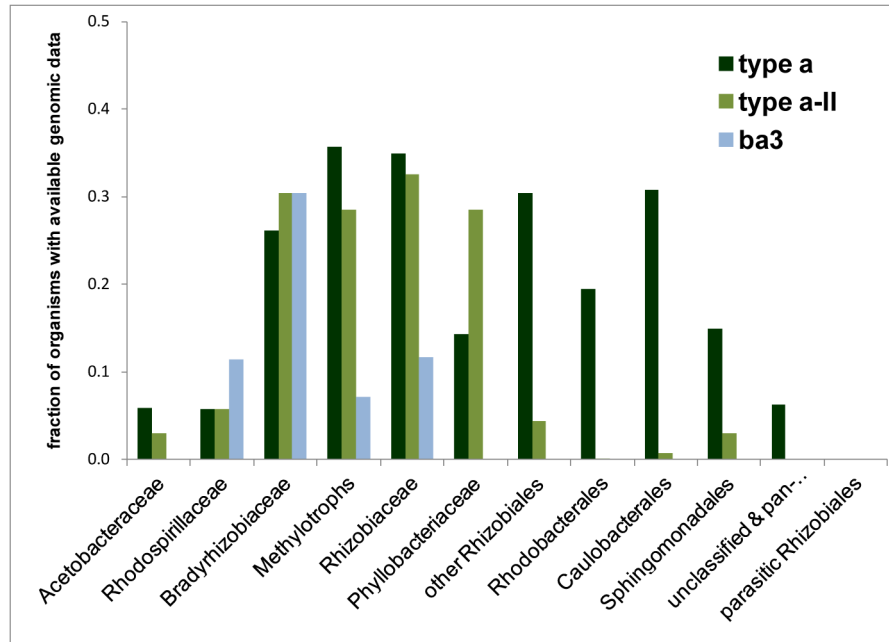
**3.4 Evolution of the molecular architecture of COX3.** In the 3D structures available for cytochrome *c* oxidases, the initial two transmembrane helices of the 7-helices *COX3* protein that is present in mitochondria and bacterial *COX* operon type b (Fig. 3A) are involved in the binding to membrane phospholipids (PL) [53,59-61]. The tight binding of two specific forms of these PL to mitochondrial *COX3* appears to modulate the entry of oxygen into the binuclear catalytic centre of the enzyme [60]. PL-binding residues are present also in other parts of the *COX3* protein that are common to all its forms and tend to be conserved [59-62]. Here, we have evaluated the amino acid substitutions of the PL-binding sites in *COX3* (Table S4 in File S1) by translating residue variation into PL-binding strength (Fig. 5A). The results of this analysis are consistent with the phylogenetic trees of *COX3*, in which a major bifurcation separates the  $\beta$ - and  $\gamma$ -proteobacterial proteins from those of  $\alpha$ -proteobacteria that are grouped together with mitochondrial *COX3* (Fig. 5B). The overall tree topology of *COX3* proteins thus matches that of *COX1* proteins, even if the internal branching of  $\alpha$ -bacteria with the mitochondrial clade appears to be different (Fig. 5B cf. Fig. 3B).

Quantitative evaluation of the PL-binding strength further refines the evolutionary relationship among *COX3* proteins. First, it shows that the 5-helices form of the protein belonging to *COX* operon type a-II occupies an intermediate position between ancestral *COX1-3* and the 7-transmembrane form of *COX3* (Fig. 5A). Secondly, it allows the comparison with the highly divergent sequence of ciliate *COX3* embedded within *COX1* (Fig. 4), which shows a PL-binding strength lying mid-way between that of *COX3* proteins of type a-II operon and those of other protists (Fig. 5A and Table S4 in File S1). Finally, bacterial *COX3* of *COX* operon type b has essentially the same PL-binding strength as that of mitochondrial *COX3* (Fig. 5A and Table S4 in File S1), thereby weakening the structural and phylogenetic significance of variable inter-group branching between  $\alpha$ -bacterial and mitochondrial *COX3* sequences (Fig. 5B and data not shown).

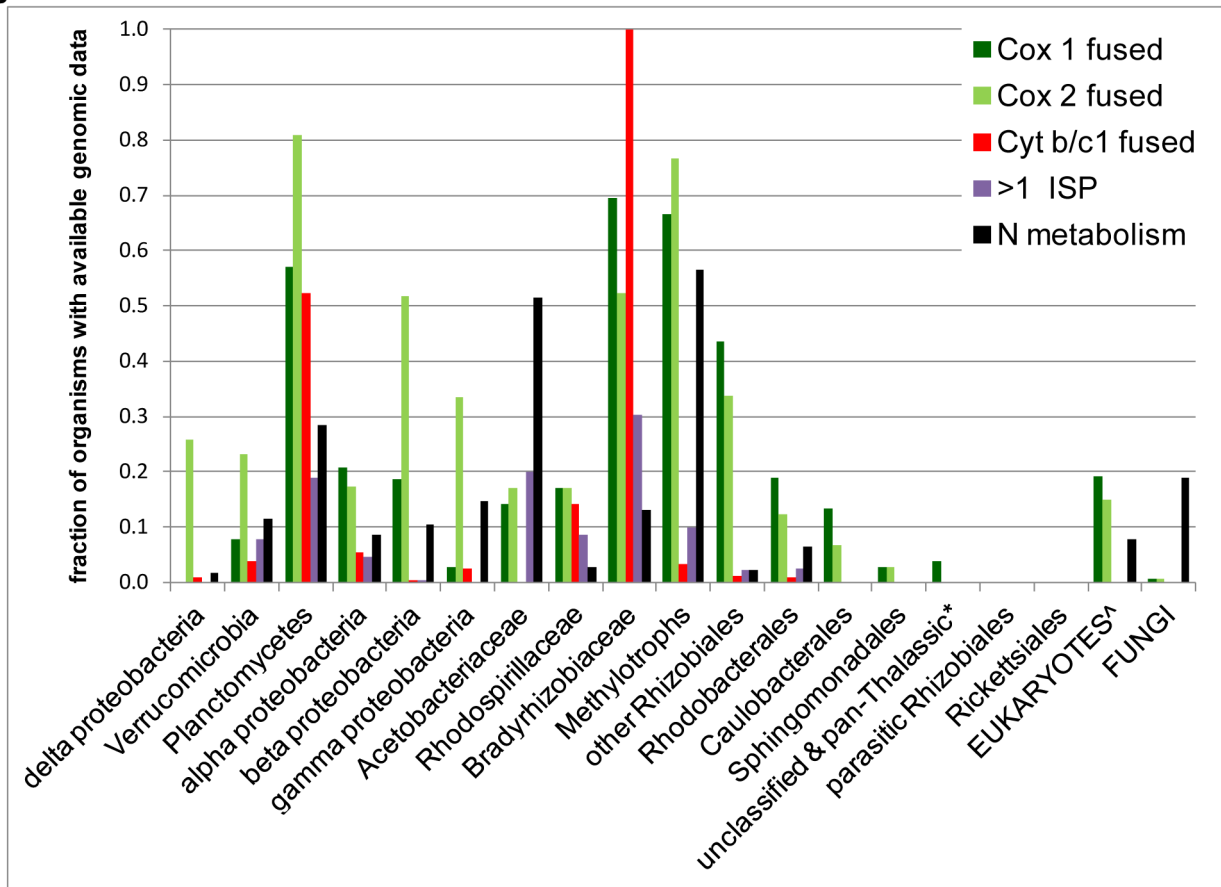
**3.5. Phylogenetic distribution of COX operons.** To acquire further information for differentiating the pathways of mitochondrial evolution in Fig. 1B, we studied the phylogenetic distribution of diverse *COX* operons. The vast majority of Rhodobacterales, Sphingomonadales and Caulobacterales, together with unclassified  $\alpha$ -proteobacteria such as *Micavibrio* and the SAR11 clade - which we include here under the generic label of ‘pan-Thalassic’ - possess only *COX* operons of type b. This implies that *Roseobacter* and *Micavibrio* cannot be related to the ancestors of mitochondria, as for *Pelagibacter* and similar marine organisms.

On the other hand, 40  $\alpha$ -proteobacterial organisms and several  $\beta$ -proteobacteria combine *COX* operon type b with a type a-II operon, the phylogenetic distribution of which is similar to that of *ba3* oxidases [26] (Fig. 6A). Conversely, *COX* operon type a-I has the broadest phylogenetic distribution among all types of *COX*

A



B



**Figure 6. Taxonomic distribution of bioenergetic systems in bacteria. A - Distribution of COX operon types in major families of  $\alpha$ -proteobacteria.** The frequency of each type of COX operon was normalised to the number of  $\alpha$ -proteobacterial organisms with genomic data that are currently available (from NCBI resources <http://www.ncbi.nlm.nih.gov/taxonomy/> - accessed 14 March 2014) [50]. See Table S2 in File S1 for a detailed list of the taxonomic distribution of diverse COX operon types. The definition 'pan-Thalassic' collects together organisms of the SAR clade

with *Magnetococcus*, *Pelagibacter* and *Micavibrio*. **B. -Distribution of fused proteins and N-metabolism elements along diverse bacterial lineages.** Fused proteins were identified with the combined resources of NCBI and the Protein Family website (PFAM 27.0 - <http://pfam.sanger.ac.uk/> [52]). Multiple forms of ISP were counted as >1 ISP. Taxa are arranged according to their approximate phylogenetic position considering also metabolic features (cf. Refs [5,31]). For each group, the frequency is normalized as in A. Eukaryotes (°) include amoebzoa, ciliates and heterokonts. N-metabolism encompasses: methane monooxygenase, ammonia monooxygenase, nitrite oxidoreductase, *Nirf* nitrite reductase and its homologues in COX operon type a-I (Fig. 3A), ammonia oxidation and anaerobic ammonia fermentation [30,32]. doi:10.1371/journal.pone.0096566.g006

operon, encompassing taxonomic groups beyond the *phylum* of proteobacteria such as Planctomycetes [50]. Indeed, the *Nif*-like gene cluster that is associated with this *COX* operon was originally discovered in ancient eubacteria including Planctomycetes [63]. Although the functional implications of the combination of a *Nif*-like operon with a *COX* gene cluster remains unknown, we are intrigued by the possibility that the overall gene sequence would produce a compact electron transport chain from quinol, or products of N metabolism, to oxygen [32,50]. Consequently, *COX* operon type a-I would represent the ultimate bioenergetic connection between cytochrome *c* oxidase and N metabolism, a fundamental concept in our approach to discern mitochondrial evolution (Fig. 1).

#### 4. Phylogenetic distribution of N metabolism and fused proteins in bacteria and mitochondria

To explore the phylogenetic dimension of the connection between *COX* operons and elements of N metabolism, we studied the taxonomic distribution of *NifD* and other key elements of the N cycle in conjunction with that of fused subunits of *aa<sub>3</sub>*-type oxidases (Fig. 6B). Indeed, *COX* operon type a-I invariably contains *COX2* fused with a *c*-type cytochrome (Figs. 3A), a fusion which is frequently present also in other *COX* operons (Fig. 3A and Fig. S3 in File S1). Fusion between catalytic subunits of bacterial heme-copper oxidases has been noted before [47,64], but considered a nuisance for phylogenetic analyses [64]. However, it constitutes a relic of ancestral bacteria adapted to harsh conditions in which the compact structure of bioenergetic systems would have been advantageous [47]. Since we have now shown that fusion between *COX* subunits is present also in the mitochondria of unicellular eukaryotes (Fig. 4) and fungi such as *Phaeosphaera* [57], we could consider them as potential relics of the evolutionary past of mitochondrial bioenergetics.

We therefore evaluated the frequency and phylogenetic distribution of fused *COX* subunits and also of the fused proteins that are present in the cytochrome *bc<sub>1</sub>* complex, the cytochrome *b* subunit of which has been previously reported to be fused with the cytochrome *c<sub>1</sub>* subunit in *Bradyrhizobium* [65]. We found the same fusion in all members of the Bradyrhizobiaceae family plus some Rhodospirillales (Fig. 6B), as well as in Planctomycetes [66].  $\alpha$ -proteobacteria show the highest frequency of fused cytochrome *b* among proteobacterial lineages, thereby suggesting that this type of protein was present before the separation of  $\beta$ - and  $\gamma$ -proteobacteria. Conversely, many more  $\beta$ -proteobacteria possess fused *COX2* proteins than  $\alpha$ -proteobacteria (Fig. 6B).

Within  $\alpha$ -proteobacteria, the distribution of fused *COX* and cytochrome *b* proteins follows a bell-shape profile along the likely evolutionary sequence of the taxonomic groups (Fig. 6B, cf. [5]). Some Sphingomonadales and Caulobacterales have fused *COX* proteins without possessing bioenergetic elements of N-metabolism (Fig. 6B). Parasitic Rhizobiales, Rickettsiales and pan-Thalassic organisms lack both fused bioenergetic proteins and elements of N-metabolism, in contrast with amoebzoa, fungi and heterokonts (Fig. 6B cf. Table 1). The absence of the above characters in parasitic and pan-Thalassic organisms could derive from their highly streamlined genomes. However, the high frequency of fused

genes in other taxa does not correlate with genome size, since acetic acid bacteria, which have a comparatively small genome, show a higher frequency of fused *COX2* proteins than, for instance, Rhodobacterales (Fig. 6). Our interpretation of the data presented in Fig. 6 is that fused bioenergetic proteins and elements of N metabolism are preserved together in phylogenetically ancient groups of  $\alpha$ -proteobacteria, from which they have been passed to proto-mitochondria but then progressively lost along the differentiation of other  $\alpha$ -proteobacteria. This implies that Methylotrophs, Bradyrhizobiaceae and several Rhodospirillales would be the oldest extant organisms of the  $\alpha$ -proteobacterial lineage, and consequently close to the distal progenitors of proto-mitochondria.

The phylogenetic distribution and similar genomic arrangement of fused bioenergetic proteins (Fig. 6) raises the question as to whether they may derive from events of Lateral Gene Transfer (LGT), for example with Planctomycetes [67]. However, detailed analysis of the molecular architecture of cytochrome *b* proteins (M. Degli Esposti, unpublished data) and the overall consistency of distance trees of fused proteins with established phylogenetic relationships (Fig. 3) indicate that LGT events have minimally contributed to the observed distribution of fused bioenergetic proteins and their diverse genomic clusters.

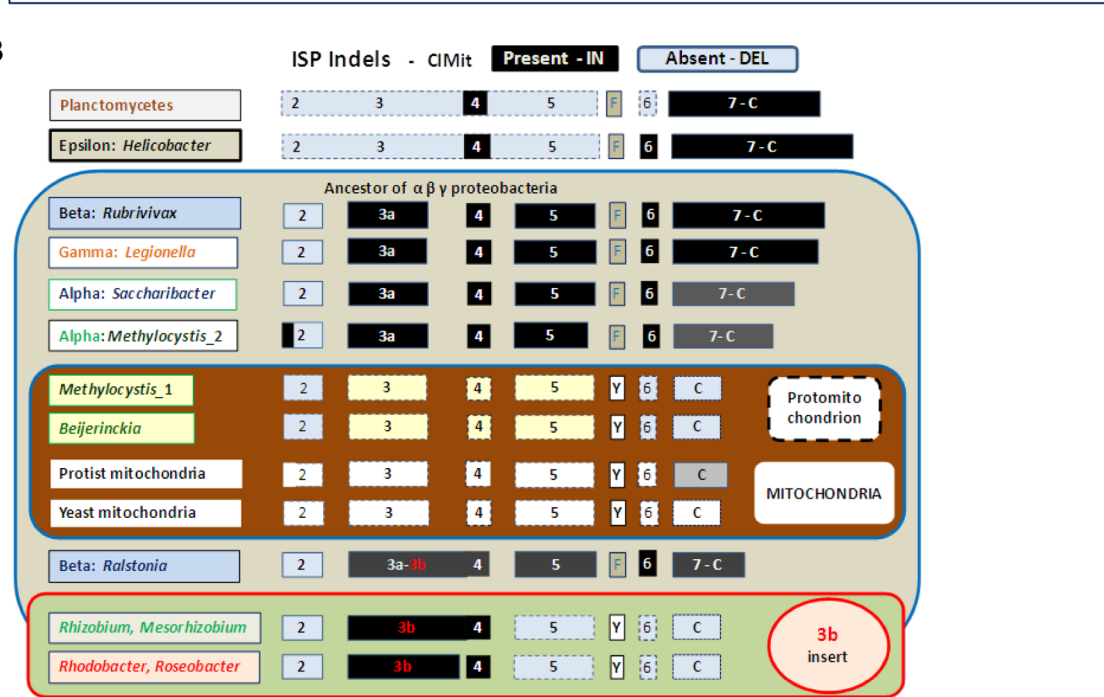
#### 5. A complementary approach: the molecular evolution of nuclear encoded ISP

To complement the above analysis of mtDNA-encoded proteins of the *aa<sub>3</sub>*-type oxidase, we next examined the molecular evolution of the “Rieske” iron sulfur subunit (ISP) of the cytochrome *bc<sub>1</sub>* complex. This ubiquitous redox protein is coded by the nuclear DNA and therefore does not suffer from the distortions due to the fast mutation rate of mtDNA-encoded proteins [16,37,48]. Its precursor form, once imported into mitochondria, matures within the intermembrane space where its catalytic core resides. After implementing structure-based alignments (Fig. S4 in File S1), we noted diverse insertions that are present in the catalytic core of ISP proteins from different lineages, which we have named CIMit - Conserved Indels vs. Mitochondria (Fig. 7 and Fig. S4 in File S1). CIMit3 is the most prominent of these insertions, lying at the surface of bacterial *bc<sub>1</sub>* complexes [68,69] with parallel inserts in the partner protein, cytochrome *b* [68–72]. This and other indels (according to the definition in Ref. [73]) seem to carry valuable phylogenetic information, enabling the resolution of relationships that are blurred in phylogenetic trees (cf. Figs. 7C and 8). For instance, only *Tistrella* ISP has no residues corresponding to the CIMit5 insertion among the proteins from Rhodospirillaceae (Fig. 7), while in distance trees these proteins appear to be equally close within a sister sub-branch of their mitochondrial homologues (Fig. 8).

*Methylocystis* sp. SC2 and a few other Rhizobiales have a second, longer ISP (ISP2) that resembles the proteins from acetic acid,  $\beta$ - and  $\gamma$ -proteobacteria, with which it clusters together in distance trees (Figs. 7, 8 and S4 in File S1). Contrary to the latter organisms, ISP2 is not present within the *petABC* operon of the *bc<sub>1</sub>* complex but in isolated gene clusters that have no common flanking genes (not shown). Hence, ISP2 may have arisen from gene duplication as reported for the  $\beta$  proteobacterium, *Rubrivivax*

A

WP_020226654	64	PGSLIKTVEW	RGKPVWIMRRTPEMIAALSGQDD	---	KLADP	NSSKDALPE	---	ALRN	PGRSER	QDLFVAIG	CT	130	Acidovorax sp. MR-S7/2	
YP_001900760	59	PGQMKVVEW	RGKPVWLLKRTPDMLLESLKKTND	---	EVADP	KSDVPTMK	TPDY	CKN	ETRSRA	EKKDLLVVVIG	CS	130	Ralstonia pickettii 123/2	
NP_275043	55	AGGLIYAEW	QKGIWLNRTDQQLKDLKGLNG	---	ELTDP	NSDAEQPE	---	YARN	NETRS	IKPNILVAIG	CT	125	Neisseria meningitidis	
WP_020564984	59	SGGLIYVW	RGKPVWLNRTPEVLATLQTLDS	---	ELRDP	LSSESIQPS	---	YTKN	NETRS	IKPEIFVAIG	CT	125	Methylosarcina fibrata	
YP_006985287	73	AGQIIVTW	RGKPVWLNRTPEMLKTLQDFAIL	---	KLKRD	SESLIFQQPK	---	DATN	WHRS	SDIGVMIG	CT	141	Gluconobacter oxydans/1	
YP_00227526	71	PGQIIVTW	RGKPVWLNRTPESLARLQDEALA	---	ARLDR	PSGALQPE	---	YARN	WHRS	IKPEGVVIG	CT	139	Gluconacetobacter diazo/2	
WP_010510616	62	PGQIIVTW	RGKPVWLNRTPESLARLQDEALA	---	GLRDR	QSNDRQPE	---	YARN	WHRS	LDPYGVVIG	CT	130	Gluconacetobacter europa/2	
CCD32085	59	PGKPVWVW	RGLPVAIFRRPDALEKLLQDPVLL	---	ELADP	DSEVLQQR	---	YARN	WHRS	IDPTVAVLVIG	CT	127	Methylocystis sp. SC2/2	
YP_577370	69	PGQIIVTW	RGLPVAIFRRPDALEKLLQDPVLL	---	QLSDP	QSSVFPQPE	---	YARN	WHRS	ANPEYGVVIG	CT	137	Nitrobacter hamburgensis/2	
WP_019172466	68	PCQIIVTW	RGLPVAIFRRPDALEKLLQDPVLL	---	ARLDR	PSQVQPE	---	YARN	WHRS	SNPQVAVLVIG	CT	136	Pseudaminobacter salicyli/2	
YP_002282274	64	PCMSLTVW	RGKPVFIIRNRTPEEVKAADVPILA	---	DLKDP	VARNANLPPEAQ	AT	GDVRS	GGKDKEN	IVMVG	CT	136	Rhizobium leguminosarum	
NP_385925	64	PCMSLTVW	RGKPVFIIRNRTPEEVKAADVPILA	---	ELKDP	VARNANLPADAES	DL	DRSAGE	GKKN	IMVVG	CT	135	Sinorhizobium meliloti/1	
WP_004610841	59	PCMSLTVW	RGKPVVVRNRTPEQEMKDEAVKLS	---	DLKDP	VARNANLPADAP	AT	DANRT	FGK	EAAMVMVQ	CT	136	Mesorhizobium opportunistum	
Rbaphaer 3D	61	EGVQIVTK	RLKLPFIIRRT	AGDLSGSLVQLG	---	QLVDT	VARNANLDAGAE	AL	DSN	TIDEAGE	NLVMG	CT	132	Rhodobacter sphaeroides
YP_004691057	59	PGIOLITW	RGKPVFIIRARTEEEIQAARATDIT	---	DLDPD	LAQANLAGDAD	AA	ENRALS	EDG	WLVVQ	MG	CT	130	Rhoseobacter litoralis/2
YP_001753157	67	DCQIIVTW	RGKLVFVRKLTAKAVADMKAAPLS	---	AMIDP	---	---	AAF	TRVKS	GHQDWLVVY	G	CT	126	Methylobacterium radio
YP_001640007	59	DCQIIVTW	RGKLVFVRKLTAKAVADMKAAPLS	---	EMIDP	---	---	AAF	TRVKS	GHQDWLVVY	G	CT	118	Methylobacterium_exto_Pa1
YP_002502576	68	EGQIIVTW	RGKLVFVRKLTAKAVADMKAAPLS	---	DLDDP	---	---	AF	QARVKE	GHQDWLVVY	G	CT	126	Methylobacterium_nodulans
WP_009763700	59	PIAEQIIV	WVWRGKLVFVRKLTAKAVADMKAAPLS	---	ASLRDP	---	---	QAD	SRVKE	GHQDWLVVY	G	CT	121	Microvirga
YP_782883	61	EGQIIVTW	RGKLVFVRKLTAKAVADMKAAPLS	---	SFRDP	---	---	QPD	SRVKE	GHQDWLVVY	G	CT	120	Rhodopseudo palu BisA53/3
YP_578436	60	EGQIIVTW	RGKLVFVRKLTAKAVADMKAAPLS	---	SLPDP	---	---	ASD	SRVKE	GHQDWLVVY	G	CT	119	Nitrobacter hamburgensis/1
WP_006590199	62	EGQIIVTW	RGKLVFVRKLTAKAVADMKAAPLS	---	ALPDP	---	---	EPD	AKRVK	---	---	CT	120	Methylocystis sp. SC2/1
YP_002364045	60	VGQIIVTW	RGKLVFVRKLTAKAVADMKAAPLS	---	ELKDP	---	---	ATD	QSRVK	---	---	CT	117	Methylocella
WP_001831301	59	EGQIIVTW	RGKLVFVRKLTAKAVADMKAAPLS	---	ELRDP	---	---	QTD	QSRVK	---	---	CT	116	Beijerinckia indica
WP_019459664	61	AGQIIVTW	RGKLVFVRKLTAKAVADMKAAPLS	---	ALKDP	---	---	ATD	QSRVK	---	---	CT	151	Roseomonas sp. B-5
YP_002299592	69	EGQIIVTW	RGKLVFVRKLTAKAVADMKAAPLS	---	SLIDP	---	---	QPD	SRVKE	GHQDWLVVY	G	CT	128	Rhodospirillum centenum/2
WP_007438995	73	EGQIIVTW	RGKLVFVRKLTAKAVADMKAAPLS	---	ELRDP	---	---	QTD	QSRVK	---	---	CT	130	Acetobacter bacter AT-5844
WP_004985728	71	EGQIIVTW	RGKLVFVRKLTAKAVADMKAAPLS	---	ELRDP	---	---	QPD	SRVKE	GHQDWLVVY	G	CT	128	Azospirillum brasilense
YP_423451	66	PGQIIVTW	RGKLVFVRKLTAKAVADMKAAPLS	---	DLRDP	---	---	QAD	ADR	R	---	CT	123	Magnetospirillum magnetium
WP_009542637	65	EGQIIVTW	RGKLVFVRKLTAKAVADMKAAPLS	---	ELRDP	---	---	QTD	EE	R	---	CT	122	Caenispirillum salinarum/2
WP_008889443	65	VGQIIVTW	RGKLVFVRKLTAKAVADMKAAPLS	---	ELVDP	---	---	QTD	D	R	---	CT	122	Thalassospira profundimaris
YP_006373746	68	AGQIIVTW	RGKLVFVRKLTAKAVADMKAAPLS	---	DLPDP	---	---	QPD	ADR	R	---	CT	125	Tistrella mobilis
XP_004356972	116	EGQIIVTW	RGKLVFVRKLTAKAVADMKAAPLS	---	DMRDP	---	---	CPD	ADR	R	---	CT	175	Acanthamoeba castellanii





**Figure 7. Molecular evolution of the Rieske subunit (ISP) of the cytochrome *bc<sub>1</sub>* complex. A – Alignment of the ISP proteins from bacteria having various *COX* operons.** ISP sequences were selected from the organisms displaying multiple *COX* operons and also ISP forms (Table S2 in File S1 and Fig. 6). The alignment was manually refined using structural information, as detailed in Fig. S4 in File S1. This alignment shows only the catalytic core of the ISP from  $\alpha$ -,  $\beta$ - and  $\gamma$ -proteobacteria, plus *Acanthamoeba* as the sole mitochondrial representative. See Fig. S4 in File S1 for a complementary alignment including the N-terminal transmembrane region and further information, including secondary structure elements (beta sheet in purple and alpha helix in green) and Conserved Indels vs. Mitochondria (CIMit). The accession codes of the proteins are shown on the left of each sequence block, while the organisms are listed on the right abbreviated as follows: *Gluconacetobacter diazo* & *\_europa*, *Gluconacetobacter diazotrophicus* PA1 5 & *europaeus*, respectively; *Pseudaminobacter salicyl*, *Pseudaminobacter salicylatoxidans*; *Methylobacterium radio* & *\_exto\_PA1*, *Methylobacterium radiotolerans* JCM 283 & *extorquens* PA1, respectively; *Rhodospseudo palu\_BisA53*, *R. palustris* BisA53; and *Acetobacter bacter* AT-5844, *Acetobacteraceae bacterium* AT-5844. ISP1 indicates the ISP form that is present in the *petABC* operon. **B – Evolutionary pattern of the conserved indels in bacterial and mitochondrial ISP.** The molecular features deduced by the structure-based alignment of ISP proteins are rendered graphically following the numerical order of conserved indels presented in A and Fig. S4 in File S1. DEletions conserved in bacterial vs. mitochondrial ISP sequences are represented in pale blue boxes with black labels, whereas INserts with respect to mitochondrial sequences are represented in black boxes with white labels.  
doi:10.1371/journal.pone.0096566.g007

*gelatinosus*, where the two forms of the proteins are interchangeable in the complex [74]. The duplicates of *Rubrivivax* ISP are closely related to each other, as in the case of the multiple ISP forms of *Roseobacter* and other Rhodobacterales (Table S1 in File S1). However, ISP2 and the in-operon ISP1 present in the same Rhizobiales organisms are separated by a deep bifurcation in phylogenetic trees, which resembles that seen in *COX1* trees (Fig. 3B,C cf. Fig. 8). Hence, ISP2 is an ancestral character of  $\alpha$ -proteobacteria equivalent to *COX* operons of type a, consistent with their similar phylogenetic distribution (Fig. 6B). Its origin can be traced to the separation of the  $\alpha\beta\gamma$  lineages, probably after the earliest proteobacterial ISP had evolved in a distinct path from its paralogues of the *bcf* complex present in Planctomycetes and Nitrospirales [75] (Fig. 8B). This ancestral form of ISP was in all likelihood devoid of the abovementioned insertions as in ISP1 of *Rhodospseudomonas palustris* BisA53 or *Nitrobacter hamburgensis*, which lie in the most distant branches of phylogenetic trees (Fig. 8A). Of note, these proteins show the single-residue deletion corresponding to CIMit6, which is shared with the ISP proteins of many  $\alpha$ -proteobacteria and their mitochondrial homologues (Figs. 7 and 8).

Importantly, the molecular features of ISP proteins provide crucial information for discriminating between the alternative pathways of mitochondrial bioenergy evolution in Fig. 1B. In particular, bacterial organisms possessing an ISP containing the CIMit3B insert (Figs. 7 and S4 in File S1) can now be excluded from mitochondrial ancestry. This applies not only to Rhodobacterales such as *Roseobacter*, but also to *Rhizobium*, *Sinorhizobium* and *Mesorhizobium* organisms that have *COX* operon type a-II (Table S2 in File S1).

## 6. Analysis of bacteria without *aa<sub>3</sub>*-type cytochrome *c* oxidase

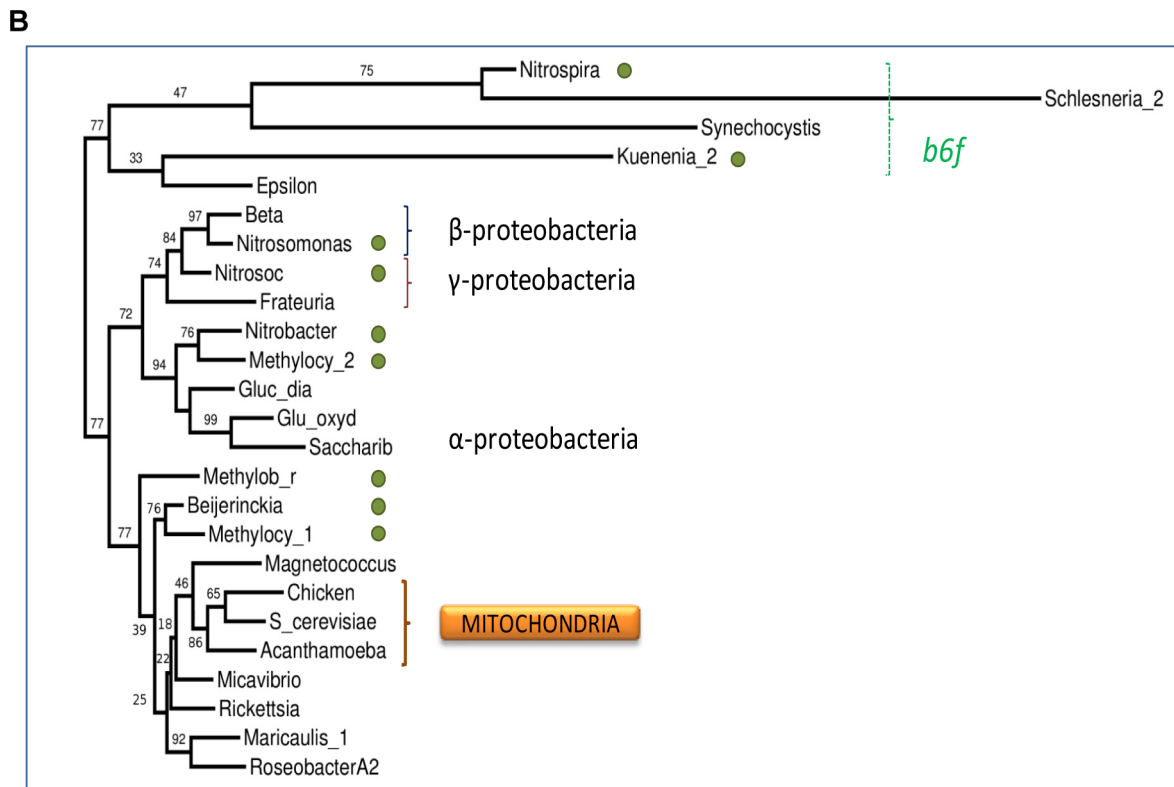
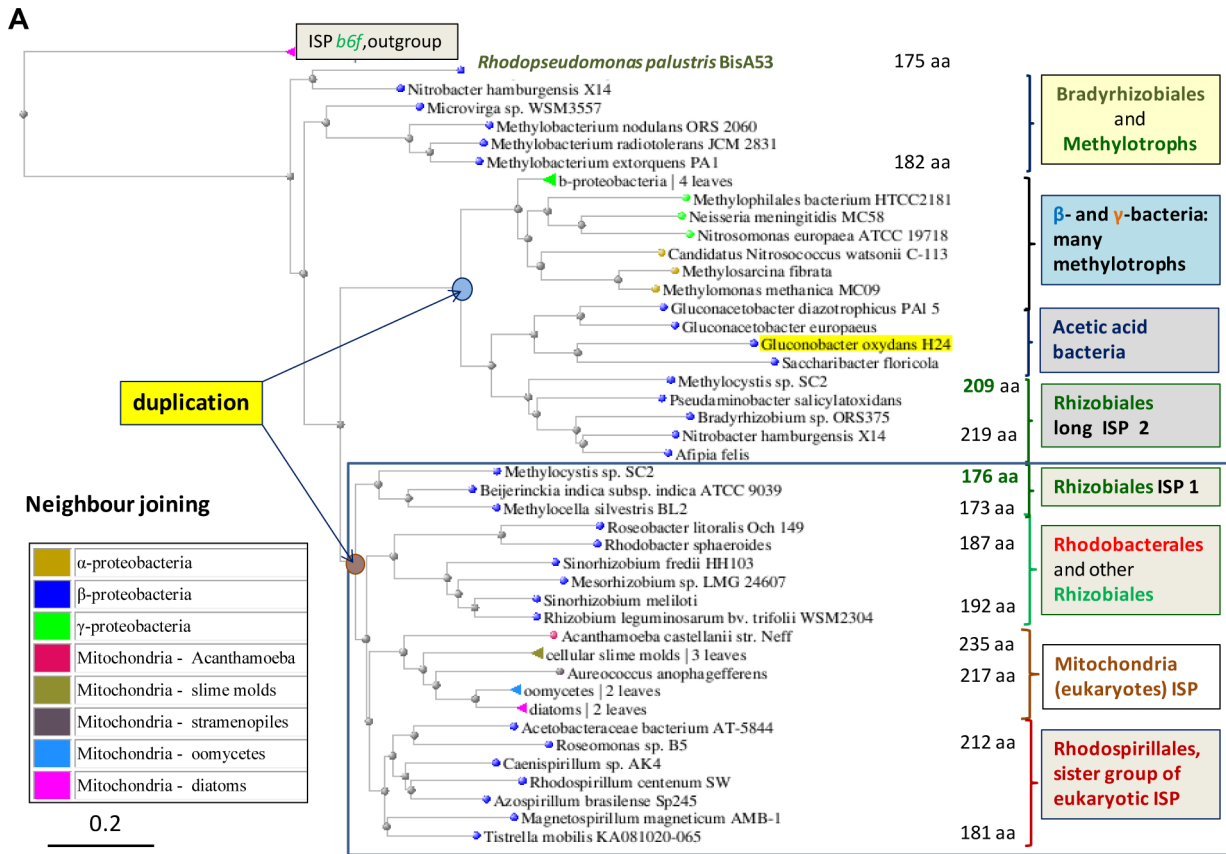
The analysis conducted so far has exploited bioenergetic systems that are not always present together in extant bacteria (Table S1 in File S1). For example, *Magnetococcus* has no functional *aa<sub>3</sub>*-type cytochrome *c* oxidase but a complete operon for the *bc<sub>1</sub>* complex and the *cbb<sub>3</sub>*-type oxidase (Table S1 in File S1, cf. Ref. [76]). Phylogenetic analysis has shown that the sequence of *Magnetococcus* ISP is rather similar to that of protists' mitochondria, even if it shows some unique amino acid changes (Figs. 8B and S4 in File S1). *Magnetococcus* lies in a deep branch of the evolutionary tree of  $\alpha$ -proteobacteria [76], similarly to *Midichloria*, which also has a *cbb<sub>3</sub>*-type oxidase instead of the *aa<sub>3</sub>*-type oxidase of other Rickettsiales [21]. *Midichloria* has an ISP protein with a unique insertion in the conserved cluster-binding region and also an unusually split version of the catalytic, *COX1*-like subunit of *cbb<sub>3</sub>*-type oxidase [21]. These molecular properties seem to indicate a side-path in the phylogenetic relationships with the mitochondrial

lineage (cf. Fig. 1B), a possibility strengthened by the analysis of the genomic and protein sequences of *cbb<sub>3</sub>*-type oxidase (data not shown). Hence, the scheme in Fig. 1B is consistent with the overall phylogenetic pattern of both *aa<sub>3</sub>*-type and *cbb<sub>3</sub>*-type terminal oxidases.

## Conclusions

Herein, we have followed novel approaches to reconstruct the possible bioenergetic characters of the bacterial ancestors of mitochondria. Rather than taking into consideration all the information that is now available from bacterial and mitochondrial genomes, we have focused on a few proteins that are crucial for bioenergy production in both bacteria and mitochondria and have multiple variants. The diverse molecular forms and genetic organization of bioenergetic systems have been hardly considered in previous studies of phylogenomics; for instance, none of the papers reviewed in Ref. [9] used proteins of energy metabolism. Conversely, recent studies on bacterial oxidases [27,64] have not considered the complexity of *COX* operons (Figs. 3 and S3 in File S1). Here we have classified this complexity and exploited its most informative aspects to reconstruct the molecular evolution of individual protein components that are encoded by either mtDNA or nuclear DNA of eukaryotes. By integrating the information thus obtained, we have excluded that several bacterial lineages previously proposed to be related to mitochondria could be in the direct line of mitochondrial ancestry, in particular the endocellular obligate parasites of the Rickettsiales group and the photosynthetic organisms *Rhodobacter* and *Rhodospirillum*. Our work indicates that mitochondrial ancestors retained bioenergetic elements of N metabolism and the *bd*-type ubiquinol oxidase, which have been subsequently lost in different paths of convergent evolution (Fig. 1B).

In concluding this work, we discuss steps of differential loss also in conjunction with the possible acquisition of systems or proteins via LGT, to provide a complete account of the remaining possibilities for the evolution of mitochondrial bioenergy production (Figure 9). Multiple lines of evidence emerging from our work lead to the conclusion that the subset of bioenergetic systems lacking the *cbb<sub>3</sub>*-type oxidase - typical of methylotrophs and *Gluconacetobacter* (Table S1 in File S1) - probably matches the bioenergetic capacity of the distal ancestors of mitochondria. This evidence includes the maximal diversity of *COX* operons and N metabolism in the abovementioned organisms (Tables S1 and S2 in File S1). The ancestral organisms from which proto-mitochondria emerged in all likelihood evolved just after the separation of  $\beta$ - and  $\gamma$ -proteobacterial lineages, a concept that is sustained, in particular, by the taxonomic distribution of fused bioenergetic proteins and key elements of N metabolism (Fig. 6). At the whole taxon level,  $\beta$ - and  $\gamma$ -proteobacteria have a much higher



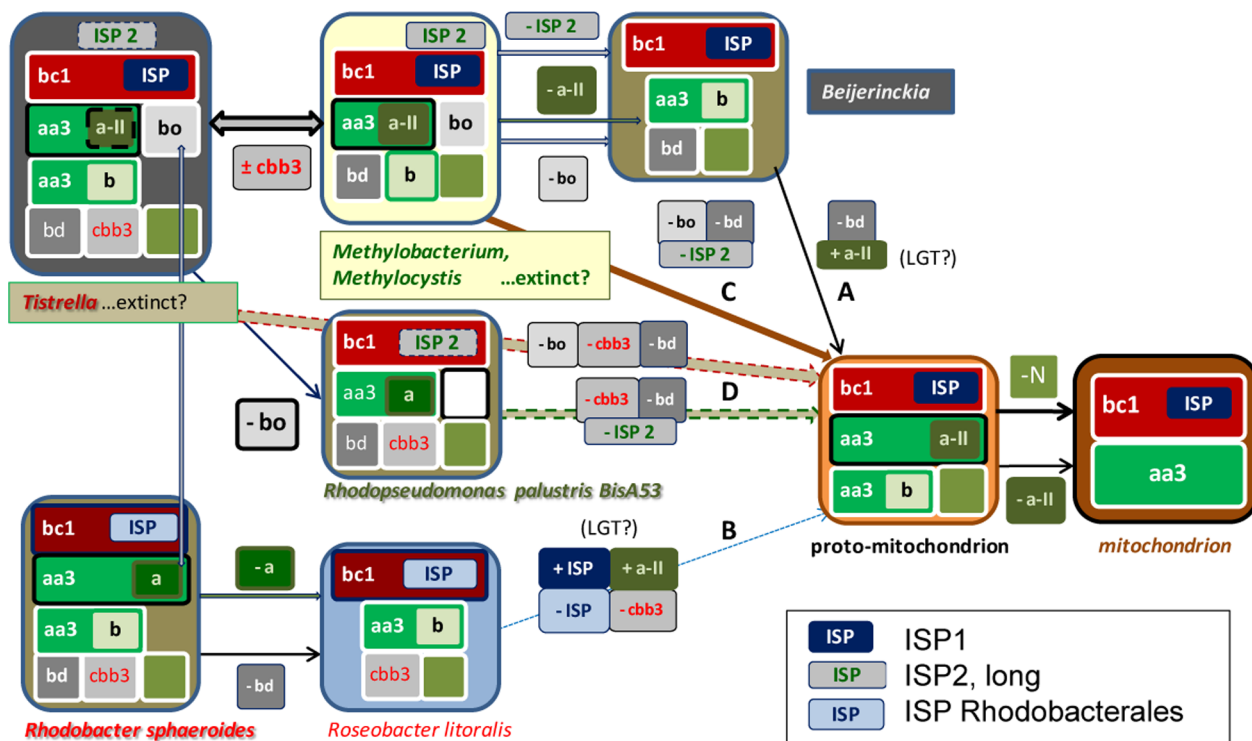
**Figure 8. Phylogenetic relationships between diverse forms of ISP. A – Distance tree encompassing proteobacteria and mitochondria.** The tree was obtained as described in the legend of Fig. 3B using the alignment of Fig. 7A and two ISP proteins from the *b<sub>6f</sub>* complex as outgroup (top). The group containing bacterial ISP1 proteins together with their mitochondrial homologs is enclosed in the blue square to

highlight a likely ancestral duplication separating it from the group with ISP2. **B – Long distance phylogenetic relationships of bacterial ISP.** The phylogenetic tree (maximal likelihood method) of ISP proteins was computed from the structure-based alignments in Fig. S4 in File S1. The small green circle indicates ancient nitrogen or methylotrophic metabolism [29–32] (Fig. 6B). The dashed green bracket indicates the paralogue proteins belonging to the *bcf* complex. Other brackets indicate proteobacterial subdivisions and mitochondria as in A. Note how the bootstrap values are much lower within the bottom branch containing mitochondrial ISP than in the upper branch containing ISP2. doi:10.1371/journal.pone.0096566.g008

frequency of these characters than  $\alpha$ -proteobacteria (Fig. 6B). However, some  $\alpha$ -proteobacteria show a high frequency of fused proteins and elements of N metabolism (Fig. 6B), namely methylotrophs - encompassing the families of Methylocystaceae, Methylobacteraceae, Beijerinckiaceae and part of Hyphomicrobiaceae, as well as Bradyrhizobiaceae such as *Afipia felis* and *Rhodospseudomonas palustris* BisA53 [77] - several Acetobacteraceae and some Rhodospirillaceae. These organisms also have a wide range of ancestral characters such as type a *COX* operons and ISP2 (Table S2 in File S1 and Fig. 8).

The information just discussed can be integrated with the timeline of bacterial evolution [31], which positions the separation of the  $\beta$ -lineage near the time at which oxygen levels dramatically increased, at least in the photic zone of marine environments and emerged land. The invention of the metabolic pathways of methane, ammonia and nitrite oxidation immediately followed,

allowing autotrophic ways of life which are now retained by a few groups of proteobacteria [30]. These bacteria also possess the largest variety of *COX* operons and molecular forms of their catalytic subunits, as the result of multiple events of operon and gene duplication. Some of these duplications are still evident in extant organisms, as indicated by the doublet of *COX3* proteins in *COX* operon type a-III (Fig. 3A) and the presence of concatenated *COX* operons in some genomes (Table S2 in File S1). Our reconstruction of the molecular evolution of *COX3* proteins and their binding strength for oxygen-modulating phospholipids (Fig. 5) seems to recapitulate a progressive adaptation to increasing levels of  $O_2$ , which had to be gauged in terms of decreasing oxygen affinity to maintain maximal efficiency of the oxidase reactions, with minimal damage by radicals and potential suicidal reactions [47,60,78]. We have also found multiple forms of other terminal oxidases in methylotrophs and Rhodospirillales, in particular for



**Figure 9. Possible progenitors for the bioenergetic evolution of mitochondria.** This diagram is modified from that in Fig. 1B to take into account the deduction that proto-mitochondria probably had two different types of *COX* operons (type a is labelled in dark olive background) and the evidence for multiple ISP forms. ISP2 is represented in a grey box while ISP1 in dark blue. Various steps of differential loss or acquisition via LGT are indicated for the possible pathways of evolution from extant or extinct  $\alpha$ -proteobacteria into proto-mitochondria. By considering the complexities arisen from our data, pathway A in Fig. 1B stemming from *Beijerinckia* would require one loss and one acquisition, while pathway B would theoretically imply two losses and two acquisitions. However, we now exclude that this pathway may have contributed to the evolution of mitochondria (see text). Pathway C, sustained by most results presented here, bypasses the *Beijerinckia* subset with the combined loss of two bioenergetic systems and ISP2. Finally, pathway D would require the combined loss of three bioenergetic systems from organisms such as *Tistrella*, but of two systems plus ISP2 for *R. palustris* BisA53, which has already lost *bo*-type oxidase (Table S1 in File S1). The obvious possibility that yet undiscovered, or extinct bacteria may be among the originators of the proto-mitochondrion is considered, as indicated. Eventual loss of photosynthesis is not shown, but it would apply only to *Methylobacterium*, *R. palustris* and *Roseobacter* among the organisms shown. The grey vertical arrow on the left indicates the possible equivalence of *COX* operon type a with dual function (cytochrome *c* and ubiquinol) oxidases in some Rhodobacterales.

doi:10.1371/journal.pone.0096566.g009

the *bd*-type ubiquinol oxidase (Table S1 in File S1). The additional forms usually correspond to the Cyanide Insensitive Oxidase (CIO) [79], which has lower affinity for oxygen than classical *bd* oxidases [25].

We believe that the large increase in ambient oxygen that occurred during the evolution of primordial proteobacteria [31] was the driving force for the genomic expansion and diversification of oxygen-reacting enzymes. High levels of O<sub>2</sub> also led to the wide availability of nitrate and nitrite that can function as alternative terminal acceptors for electron transfer and bioenergy production [22,32]. This underlines the strong link between oxygen respiration and key elements of N metabolism that we have taken in consideration here. The separation of proto-mitochondria is estimated to have occurred when oxygen levels were still very low in the oceans [10,24], where most primordial life thrived. It is therefore plausible that the distal progenitors of mitochondria were related to organisms that had experimented with a wide variety of oxygen-reacting systems and thus retained great plasticity in their adaptation to micro-oxic or even anoxic environments, a trait that is partially retained in eukaryotes adapted to anaerobic environments [10]. With this conceptual framework in mind, we can now look back to the initial approach of our work (Fig. 1) and consider the most plausible pathways for mitochondrial evolution (Fig. 9).

Following the separation of the  $\beta$ - and  $\gamma$ -proteobacterial lineages, proto-mitochondria may have branched off along one of the pathways illustrated in Fig. 9. Pathways A and B are the same as in Fig. 1B, with the additional complexities that have emerged from the detailed analysis of COX operons and ISP proteins plus possible acquisitions via LGT. Pathway A, stemming from *Beijerinckia* (we now exclude *Micavibrio* for it lacks key elements of N metabolism, cf. Fig. 2), would require one loss (*bd* oxidase) plus one acquisition (COX operon type a-II), while pathways B would theoretically require two losses and two acquisitions of bioenergetic systems. However, our results indicate that mitochondrial evolution is unlikely to have followed pathway B, since the organisms from which it departs do not have key elements of N-metabolism that are present in some eukaryotes (Figs. 2 and 6B) nor a ISP comparable to that of eukaryotes (Figs. 7 and 8). Additional pathway C bypasses the *Beijerinckia* subset with the combined loss of two bioenergetic systems and ISP2, the latter being a facile evolutionary step for only six organisms have retained ISP2 (Figs 7 and 8). This pathway stems from methylotrophic bacteria such as *Methylocystis* and *Methylobacterium*. Indeed, the analysis of three different types of bioenergy-producing systems - cytosolic nitrate assimilation, mitochondria-encoded subunits of cytochrome *c* oxidase and nuclear-encoded ISP subunit of the cytochrome *bc*<sub>1</sub> complex - converges in indicating methylotrophs as the most likely relatives to proto-mitochondria. Moreover, by combining the analysis of nitrate metabolism (Fig. 2) with that of COX (Figs. 3–6) and ISP evolution (Figs. 7, 8 and S4 in File S1), only *Tistrella* [48] and *Rhodospseudomonas palustris* [6] remain among all the bacteria that have been previously proposed as possible ancestors of mitochondria (cf. Figs. 1B and Table S1 in File S1). We have thus considered also pathway D, which would require the combined loss of three bioenergetic systems from those possessed by *Tistrella* (Fig. 9). Finally, *Rhodospseudomonas palustris* BisA53 does not have the *bo*-type oxidase as other organisms of the same genus, but possesses a methanol dehydrogenase close to that of methylotrophs (Table 1). However, it still retains a photosynthetic system, the loss of which would add to the other steps required to resemble proto-mitochondria (Fig. 9). The obvious possibility that yet undiscovered, or extinct bacteria may be among the originators of the proto-mitochondrion is also considered in Fig. 9. Yet, these

unknown organisms would probably have the subsets of bioenergy systems shown in the top part of the diagram.

Taken all our results together, methylotrophic organisms emerge as the closest living models for mitochondrial ancestors. In perspective, our work provides new means for selecting bacterial organisms that are most suitable for experimentally re-evolving proto-mitochondria with mitochondria-depleted eukaryotic cells.

## Methods

To identify genes and their products with others currently present in National Center for Biotechnology Information (NCBI) resources, we have extensively used the program DELTABLAST, Domain Enhanced Lookup time Accelerated BLAST [80], integrated with hydropathy analysis conducted with in house algorithms [72] or the program WHAT (Web-based Hydropathy, Amphipathicity and Topology <http://saier-144-21.ucsd.edu/barwhat.html> [81]). Manually refined alignments of bioenergetic proteins were subjected to phylogenetic analysis with maximum likelihood algorithm and 100 bootstrap resamplings, using the program PhyML 3.0 and evolutionary models selected with Prottest3, as described earlier [21]. The results obtained with this rigorous method essentially matched those obtained with the recent options of DELTABLAST (cf. Fig. 8). The genomes of *Asaia platicody* and *Saccharibacter sp.* (EMBL accession: CBLX010000001/27 and CBLX010000001/09, respectively) were recently reported by Chouaia *et al* [22]. See Supporting Information for additional methods and procedures of gene recognition, operon classification (cf. [82]) and sequence analysis of proteins (cf. [41,52,83]).

## Supporting Information

**File S1 We enclose File S1 with Supporting Information containing a detailed account of the classification of bacterial COX operons (2 pages), 4 additional Figures and 4 additional Tables. Figure S1, Pathways for the bioenergetic evolution of a bacterial not leading to mitochondria.** The diagram shows the additional subsets of bioenergetic systems that are not shown in Fig. 1B, including those of *Asaia* and *Saccharibacter* (Table S1B in File S1). The asterisk\* labels the same subset as in Fig. 1B (main text), but with fewer representative taxa. Underlined organisms are symbionts or pathogens. Each of the six bioenergetic systems presented in Fig. 1 was identified from its catalytic protein subunits and was considered functionally absent when one or more of these subunits were not found in their completeness, as indicated by the profile of their conserved domains (cf. [41]). The functional absence of a given system is represented by an empty square as in Fig. 1B.

**Figure S2, Sequence analysis to identify the fusion of COX4 subunit with COX1 proteins. A.** Sequences of recognised or putative COX4 were manually aligned to reference proteins having known 3D structure around the first transmembrane helix (TM1, highlighted in grey): subunit IV of *Thermus caa*<sub>3</sub> oxidase (accession: pdb|2YEV [54]) and subunit IV (COX4<sub>pro\_2</sub> super family [cl06738]) of *Rhodobacter Sphaeroides aa*<sub>3</sub> oxidase (chain D, accession: pdb|1M57 [53]). \*Residues in **bold** have positive scores ( $\geq 0$ ) in the BLOSUM62 substitution matrix [83], those **yellow-highlighted** are identical with either reference protein, while those highlighted in purple are identical to *Janibacter COXIV* (accession: ZP\_00994995) with scores  $\geq 5$  [83]. The total count of identities is also highlighted in yellow (tot) before the description of the protein on the right. It was used to identify other COX4-like proteins such as DUF983 (see Fig. 3A and the section entitled “classification of bacterial COX operons” in File S1). The

minimal count for deeming a protein as “COX4-like” was considered to be 10, but several COX1 proteins exhibited larger numbers of identities. The region of ciliate COX1 showing similarity with COX4 partially overlaps the last transmembrane region (TM12) of aligned COX1, which is well conserved among all available COX1 sequences from ciliates. However, the COX4-like region in bacterial COX1 and that of the pathogenic fungus *Zysoptoria* [55] lies outside the conserved domains of other COX1 proteins. Azospirillum\_bras, *Azospirillum brasilense*; Methylobac\_extor, *Methylobacterium extorquens*. **B** - This panel shows the alignment of COX4 subunits around the second transmembrane helix (TM2), the structure of which is known only for subunit IV of *Thermus caa3* [54] that was used as the reference for aligning bacterial COX4 and mtDNA-encoded proteins. In **bold black** are the residues that are identical in the aligned position of at least two COX4 sequences, or are positive substitutions [83] across at least three aligned COX4 sequences; they are additionally **yellow-highlighted** when identical between at least one bacterial COX4 and one mtDNA-encoded protein (cf. A). In **bold dark blue** are the residues that are positive substitutions between bacterial COX4 and mtDNA-encoded proteins, while those in **bold light blue** are identical or positive substitutions among the aligned mtDNA-encoded proteins. This colour labelling enhances the limited similarity between the sequences shown. **Figure S3, Gene sequence of additional COX operons in diverse bacteria.** The reference gene name for each cluster is indicated on the right of the figure. Symbols identify the same proteins as in Fig. 3A, with the addition of the small gray bar, protein related to nucleotide exchange factor EF-TS. These short proteins were recognised after alignment to the sequence with known 3D structure of Chain A, dimerization domain of Ef-Ts from *Thermus thermophilus* (Accession: pdb|1TFE|A) using a sequence analysis similar to that shown in Fig. S2 in File S1. Hypothetical steps in the evolution of COX operons are indicated. **Figure S4, Structure-based alignment of bacterial and mitochondrial “Rieske” ISP.** The protein sequences of various ISP of the bc<sub>1</sub> complex were aligned following structures available from various sources matching the alignment gaps or insertions with the most refined 3D data [68–71]. The limits of secondary structures (alpha helices, highlighted in green, and beta sheets, highlighted in purple) were deduced from a consensus of the latest coordinates deposited in the NCBI databanks [68–71]. Common insertions and deletions (Indels [72]) between mitochondrial and bacterial sequences are consecutively labelled CIMit1-7 (cf. Fig. 7A). The C terminus of some sequences is truncated at the residue indicated by the numeral before the slash. Key residues for the iron-sulfur cluster, including Y165 influencing its redox potential [71], are in **bold**. Note that *Nitrospira*, *Nitrosomonas*, *Nitrosococcus* and *Methylocystis* are metabolically related by ammonia/methane autothrophy. The organisms follow established phylogenetic distance from top to bottom according to the following taxonomic groups and species. **Cyanobacteria:** Synechocystis (*bef* complex), *Synechocystis* sp. PCC 6803, 192 aa; **Nitrospirales:** *Nitrospira*, *Candidatus Nitrospira defluvi* [73], 183 aa; **ε-proteobacteria:** Epsilon, *Helicobacter pylori*, 167 aa; **Planctomycetes:** *Kuenenia\_2*, *Candidatus Kuenenia stuttgartensis* (in-operon Kuste3096 [66]), 173 aa; *Schlesneria\_2*, *Schlesneria paludicola* DSM 18645 (accession: ZP\_11092182), 189 aa. **γ-proteobacteria:** Nitrosoc, *Nitrosococcus watsonii* C-113, 201 aa; Frateuria, *Frateuria aurantia*, 201 aa; **β-proteobacteria:** Nitrosomonas, *Nitrosomonas europaea* ATCC 19718, 201 aa; Beta, *Neisseria meningitidis* MC58, 193 aa. **α-proteobacteria:** Methylocy\_1 &\_2, *Methylocystis* sp. SC2 [84], \_1 in-operon, 176 aa, \_2 in isolated gene cluster, 209 aa; Methylob\_r, *Methylobacterium radiotolerans* JCM 2831, 189 aa; Nitrobacter,

*Nitrobacter hamburgensis* ISP2, 219 aa; Gluc\_dia, *Gluconacetobacter diazotrophicus* PAL 5 (in isolated gene cluster), 221 aa; Saccharib, *Saccharibacter* sp. (Chouaia *et al.* [22]), 223 aa; Glu\_oxyd, *Gluconobacter oxydans* H24, 218 aa; Beijerinckia, *Beijerinckia indica*, 172 aa; RoseobacterA2, *Roseobacter litoralis petA2* in-operon, 186 aa; Maricaulis\_1, *Maricaulis maris* in-operon, 207 aa; Micavibrio, *Micavibrio aeruginosavorus* [25], 185 aa; Magnetococcus, *Magnetococcus marinus* [76], 178 aa; Rickettsia, *Rickettsia felis*, 177 aa. **Mitochondria: Acanthamoeba**, *Acanthamoeba castellanii*, 235 aa; S\_cerevisiae, *Saccharomyces cerevisiae*, mature 185 aa (3D structure available [85]); Chicken, *Gallus gallus*, mature 192 aa (3D structure available [68]). C-terminal extensions are highlighted in pale blue with some conserved residues in gray. **Table S1, Genomic distribution of bioenergetic systems in α-proteobacteria.** **A.** The genomes of ca. 120 α-proteobacterial organisms were studied from the latest version of the genome NCBI database <http://www.ncbi.nlm.nih.gov/genome/browse/> accessed on 14 March 2014, verifying also the completeness of genomic data (\*). Reconstruction of the various bioenergetic systems (see text) was deduced by combining genomic information with biochemical and microbiological data. The organisms are listed following the left-right sequence in the model of Fig. 1B. Major types of bd oxidases are classified as bd-I or CIO [25,79]. The organisms directly shown in Fig. 1B are yellow highlighted and those proposed to be relatives of mitochondria are shown in italics with pertinent references (including [86,87]). Underlined organisms are symbionts or pathogens. **B.** This table lists the organisms that have been analysed but are not included in the model of Fig. 1B, also because they are in parallel paths of evolution with respect to the mitochondrial subset of bioenergetic systems. The organisms highlighted in pale yellow are shown in Fig. S1 in File S1, while other annotations are the same as in **A**. Complementary information is in Table S2 in File S1. **Table S2, Diverse gene clusters for aa<sub>3</sub>-type oxidase in α-proteobacteria.** The table lists the diverse types of COX operons (Fig. 3A). COX1 proteins recognised as ba3-like\_Oxidase\_I [cd01660] [41] are under the column ba3^ and correspond to class B [26]. Concatenated operons are framed in blue and connected by a thick line. Incomplete (or ‘dead’ [82]) operons, indicated by the asterisk\*, lack one or more of core subunits *ctaC-E* (Fig. 3A). Functional capacity of the oxidase has been deduced also from biochemical studies [88,89]. **Table S3, Phylogenetic distribution of the main characters of COX gene operons.** We constructed a matrix of 11 independent characters (indicated concisely on top of the columns) that could differentiate the gene sequence of COX subunits in the mitochondria of some protists from the gene sequence of bacterial COX operons. The cumulative phenetic analysis indicate that COX operon type a-II of methylotrophs and *Tistrella* (highlighted) share the largest number of characters with COX gene clusters of protist mitochondria (F. Comandatore and C. Bandi, unpublished). **Table S4, Conserved phospholipid binding sites in COX3 proteins.** The alignment in Fig. 4A was enlarged and the residues corresponding to the PL-binding sites and E90 (close to O2 entry in beef COX3 [60]) were considered conserved when producing positive substitutions [83] (bold amino acid symbols in white background). Other substitutions are highlighted in pale brown while identities are identified as **yes**. Organisms are abbreviated as in Fig. 4. (PDF)

## Acknowledgments

We thank Ed Berry (SUNY, Albany, USA), Marta d’Amora, Diego Sona, Alberto Diaspro and Roberto Cingolani (IIT, Italy) for helpful discussion and support.

## Author Contributions

Conceived and designed the experiments: MDE BC FC EC DS CB. Performed the experiments: MDE BC FC EC DS PMJL. Analyzed the data: MDE BC FC DS PMJL CB DD. Contributed reagents/materials/

analysis tools: MDE CB FC CB DD. Wrote the paper: MDE PMJL CB DD. Conceived the work and wrote the bulk of the manuscript: MDE. Inspired various aspects of the work and participated to the writing and refining of the article: CB DD.

## References

- Gray MW (2012) Mitochondrial evolution. *Cold Spring Harb Perspect Biol* 4: a011403.
- Lane N, Martin W (2010) The energetics of genome complexity. *Nature* 467: 929–934.
- Margulis L (1996) Archaeal-eubacterial mergers in the origin of Eukarya: phylogenetic classification of life. *Proc Natl Acad Sci U S A* 93: 1071–1076.
- Andersson SG, Karlberg O, Canbäck B, Kurland CG (2003) On the origin of mitochondria: a genomics perspective. *Philos Trans R Soc Lond B Biol Sci* 358: 165–177.
- Williams KP, Sobral BW, Dickerman AW (2007) A robust species tree for the alphaproteobacteria. *J Bacteriol* 189: 4578–4586.
- Abhishek A, Bavishi A, Bavishi A, Choudhary M (2011) Bacterial genome chimaerism and the origin of mitochondria. *Can J Microbiol* 57: 49–61.
- Georgiades K, Raoult D (2011) The rhizome of *Reclinomonas americana*, *Homo sapiens*, *Pediculus humanus* and *Saccharomyces cerevisiae* mitochondria. *Biol Direct* 6: 55.
- Thiergart T, Landan G, Schenk M, Dagan T, Martin WF (2012) An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. *Genome Biol Evol* 4: 466–485.
- Gribaldo S, Poole AM, Daubin V, Forterre P, Brochier-Armanet C (2010) The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse? *Nat Rev Microbiol* 8: 743–752.
- Müller M, Mentel M, van Hellemond JJ, Henze K, Woehle C, et al (2012) Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiol Mol Biol Rev* 76: 444–495.
- Searcy DG (2003) Metabolic integration during the evolutionary origin of mitochondria. *Cell Res* 13: 229–238.
- Gabaldón T, Huynen MA (2003) Reconstruction of the proto-mitochondrial metabolism. *Science* 301: 609.
- Brindefalk B, Ettema TJG, Viklund J, Thollesson M, Andersson SGE (2011) A phylometagenomic exploration of oceanic alphaproteobacteria reveals mitochondrial relatives unrelated to the SAR11 clade. *PLOS ONE* 6: e24457.
- Georgiades K, Madoui M-A, Le P Robert C, Raoult D (2011) Phylogenomic analysis of *Odyssella thessalonicensis* fortifies the common origin of Rickettsiales, *Pelagibacter ubique* and *Reclinomonas americana* mitochondrion. *PLOS ONE* 6: e24857.
- Rodríguez-Ezpeleta N, Embley TM (2012) The SAR11 group of alphaproteobacteria is not related to the origin of mitochondria. *PLOS ONE* 7: e30520.
- Esser C, Ahmadijad N, Wiegand C, Rotte C, Sebastiani F, et al (2004) A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol Biol Evol* 21: 1643–1660.
- Yip C, Harbour ME, Jayawardena K, Fearnley IM, Sazanov LA (2011) Evolution of respiratory complex I: 'supernumerary' subunits are present in the alpha-proteobacterial enzyme. *J Biol Chem* 286: 5023–5033.
- Clements A, Bursac D, Gatsos X, Perry AJ, Covicristov S, et al (2009) The reducible complexity of a mitochondrial molecular machine. *Proc Natl Acad Sci U S A* 106: 15791–15795.
- Davidov Y, Huchon D, Koval SF, Jurkevitch E (2006) A new alpha-proteobacterial clade of *Bdellovibrio*-like predators: implications for the mitochondrial endosymbiotic theory. *Environ Microbiol* 8: 2179–2188.
- Atteia A, Adrait A, Brugière S, Tardif M, van Lis R, et al (2009) A proteomic survey of *Chlamydomonas reinhardtii* mitochondria sheds new light on the metabolic plasticity of the organelle and on the nature of the alpha-proteobacterial mitochondrial ancestor. *Mol Biol Evol* 26: 1533–1548.
- Sassera D, Lo N, Epis S, D'Auria G, Montagna M, et al (2011) Phylogenomic evidence for the presence of a flagellum and *cbb(3)* oxidase in the free-living mitochondrial ancestor. *Mol Biol Evol* 28: 3285–3296.
- Chouaia B, Gaiarsa S, Crotti E, Comandatore F, Degli Esposti M, et al (2014) Acetic acid bacteria genomes reveal functional traits for adaptation to life in insect guts. *Genome Biol Evol*, in press.
- Kim SW, Fushinobu S, Zhou S, Wakagi T, Shoum H (2009) Eukaryotic *nirK* genes encoding copper-containing nitrite reductase: originating from the proto-mitochondrion? *Appl Environ Microbiol* 75: 2652–2658.
- Johnston DT, Wolfe-Simon F, Pearson A, Knoll AH (2009) Anoxygenic photosynthesis modulated Proterozoic oxygen and sustained Earth's middle age. *Proc Natl Acad Sci U S A* 106: 16925–16929.
- Borisov VB, Gennis RB, Hemp J, Verkhovskiy MI (2011) The cytochrome *bd* respiratory oxygen reductases. *Biochim Biophys Acta* 1807: 1398–1413.
- Sousa FL, Alves RJ, Ribeiro MA, Pereira-Leal JB, Teixeira M, et al (2012) The superfamily of heme-copper oxygen reductases: types and evolutionary considerations. *Biochim Biophys Acta* 1817: 629–637.
- Ducluzeau AL, Ouchane S, Nitschke W (2008) The *cbb3* oxidases are an ancient innovation of the domain bacteria. *Mol Biol Evol* 25: 1158–1166.
- McLeod MP, Qin X, Karpathy SE, Gioia J, Highlander SK, et al (2004) Complete genome sequence of *Rickettsia typhi* and comparison with sequences of other rickettsiae. *J Bacteriol* 186: 5842–5855.
- Takaya N (2009) Response to hypoxia, reduction of electron acceptors, and subsequent survival by filamentous fungi. *Biosci Biotechnol Biochem* 73: 1–8.
- Vlaeminck SE, Hay AG, Maignien L, Verstraete W (2011) In quest of the nitrogen oxidizing prokaryotes of the early Earth. *Environ Microbiol* 13: 283–295.
- Battistuzzi FU, Feijao A, Hedges SB (2004) A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol Biol* 4: 44.
- Simon J, Klotz MG (2013) Diversity and evolution of bioenergetic systems involved in microbial nitrogen compound transformations. *Biochim Biophys Acta* 1827: 114–135.
- Tamas I, Dedysh SN, Liesack W, Stott MB, Alam M, et al (2010) Complete genome sequence of *Beijerinckia indica* subsp. *indica*. *J Bacteriol* 192: 4532–4533.
- Shih PM, Matzke NJ (2013) Primary endosymbiosis events date to the later Proterozoic with cross-calibrated phylogenetic dating of duplicated ATPase proteins. *Proc Natl Acad Sci U S A* 110: 12355–12360.
- Slot JC, Hibbett DS (2007) Horizontal transfer of a nitrate assimilation gene cluster and ecological transitions in fungi: a phylogenetic study. *PLOS ONE* 2: e1097.
- Lin JT, Goldman BS, Stewart V (1994) The *nasFEDCBA* operon for nitrate and nitrite assimilation in *Klebsiella pneumoniae* M5al. *J Bacteriol* 176: 2551–2559.
- Lebrun E, Santini JM, Brugna M, Ducluzeau AL, Ouchane S, et al (2006) The Rieske protein: a case study on the pitfalls of multiple sequence alignments and phylogenetic reconstruction. *Mol Biol Evol* 23: 1180–1191.
- Moreno-Vivián C, Cabello P, Martínez-Luque M, Blasco R, Castillo F (1999) Prokaryotic nitrate reduction: molecular properties and functional distinction among bacterial nitrate reductases. *J Bacteriol* 181: 6573–6584.
- Mohan SB, Schmid M, Jetten M, Cole J (2004) Detection and widespread distribution of the *nrfA* gene encoding nitrite reduction to ammonia, a short circuit in the biological nitrogen cycle that competes with denitrification. *FEMS Microbiol Ecol* 49: 433–443.
- Zhang Y, Rump S, Gladyshev VN (2011) Comparative Genomics and Evolution of Molybdenum Utilization. *Coord Chem Rev* 255: 1206–1217.
- Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, et al (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* 39 (Database issue): D225–9.
- Lee SJ, McCormick MS, Lippard SJ, Cho US (2013) Control of substrate access to the active site in methane monooxygenase. *Nature* 494: 380–384.
- Chistoserdova L, Kalyuzhnaya MG, Lidstrom ME (2009) The expanding world of methylotrophic metabolism. *Annu Rev Microbiol* 63: 477–499.
- Stewart JJ, Coyne KJ (2011) Analysis of raphidophyte assimilatory nitrate reductase reveals unique domain architecture incorporating a 2/2 hemoglobin. *Plant Mol Biol* 77: 565–575.
- Lau E, Fisher MC, Stuedler PA, Cavanaugh CM (2013) The methanol dehydrogenase gene, *mxhF*, as a functional and phylogenetic marker for proteobacterial methanotrophs in natural environments. *PLoS ONE* 8: e56993.
- Liu X, Taber HW (1998) Catabolite regulation of the *Bacillus subtilis* *ctaBCDEF* gene cluster. *J Bacteriol* 180: 6154–6163.
- Radzi Noor M, Soulimane T (2012) Bioenergetics at extreme temperature: *Thermus thermophilus* *ba(3)-* and *caa(3)-* type cytochrome *c* oxidases. *Biochim Biophys Acta* 1817: 638–649.
- Burger G, Gray MW, Forget L, Lang BF (2013) Strikingly Bacteria-Like and Gene-Rich Mitochondrial Genomes throughout jakobid protists. *Genome Biol Evol* 5: 418–438.
- Hussain H, Grove J, Griffiths L, Busby S, Cole J (1994) A seven-gene operon essential for formate-dependent nitrite reduction to ammonia by enteric bacteria. *Mol Microbiol* 12: 153–163.
- Refojo PN, Sousa FL, Teixeira M, Pereira MM (2010) The alternative complex III: a different architecture using known building systems. *Biochim Biophys Acta* 1797: 1869–1876.
- Starkenburger SR, Larimer FW, Stein LY, Klotz MG, Chain PS, et al (2008) Complete genome sequence of *Nitrobacter hamburgensis* X14 and comparative genomic analysis of species within the genus *Nitrobacter*. *Appl Environ Microbiol* 74: 2852–2863.
- Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J (2012) The Pfam protein families database. *Nucleic Acids Res* 40 (Database issue): D290–301.
- Svensson-Ek M, Abramson J, Larsson G, Törnroth S, Brzezinski P, et al (2002) The X-ray crystal structures of wild-type and EQ(I-286) mutant cytochrome *c* oxidases from *Rhodospira sphaeroides*. *J Mol Biol* 321: 329–339.
- Lyons JA, Aragão D, Slattery O, Pislakov AV, Soulimane T, et al (2012) Structural insights into electron transfer in *caa3*-type cytochrome oxidase. *Nature* 487: 514–518.

55. Torriani SF, Goodwin SB, Kema GH, Pangilinan JL, McDonald BA (2008) Intraspecific comparison and annotation of two complete mitochondrial genome sequences from the plant pathogenic fungus *Mycosphaerella graminicola*. *Fungal Genet Biol* 45: 628–637.
56. Swart EC, Bracht JR, Magrini V, Minx P, Chen X, et al (2013) The *Oxytricha trifallax* macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. *PLoS Biol* 11: e1001473.
57. Hane JK, Lowe RG, Solomon PS, Tan KC, Schoch CL, et al (2007) Dothideomycete plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*. *Plant Cell* 19: 3347–3368.
58. de Graaf RM, van Alen TA, Dutilh BE, Kuiper JW, van Zoggel HJ, et al (2009) The mitochondrial genomes of the ciliates *Euplotes minuta* and *Euplotes crassus*. *BMC Genomics* 10: 514.
59. Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H, et al (1996) The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 Å. *Science* 272: 1136–1144.
60. Shinzawa-Itoh K, Aoyama H, Muramoto K, Terada H, Kurauchi T, et al (2007) Structures and physiological roles of 13 integral lipids of bovine heart cytochrome c oxidase. *EMBO J* 26: 1713–1725.
61. Harrenga A, Michel H (1999) The cytochrome c oxidase from *Paracoccus denitrificans* does not change the metal center ligation upon reduction. *J Biol Chem* 274: 33296–33299.
62. Qin L, Hiser C, Mulichak A, Garavito RM, Ferguson-Miller S (2006) Identification of conserved lipid/detergent-binding sites in a high-resolution structure of the membrane protein cytochrome c oxidase. *Proc Natl Acad Sci U S A* 103: 16117–16122.
63. Yanyushin MF, del Rosario MC, Brune DC, Blankenship RE (2005) New class of bacterial membrane oxidoreductases. *Biochemistry* 44: 10037–10045.
64. Brochier-Armanet C, Talla E, Gribaldo S (2009) The multiple evolutionary histories of dioxygen reductases: Implications for the origin and evolution of aerobic respiration. *Mol Biol Evol* 26: 285–297.
65. Thöny-Meyer L (1997) Biogenesis of respiratory cytochromes in bacteria. *Microbiol Mol Biol Rev* 61: 337–376.
66. Kartal B, de Almeida NM, Maalcke WJ, Op den Camp HJ, Jetten MS, et al (2013) How to make a living from anaerobic ammonium oxidation. *FEMS Microbiol Rev* 37: 428–461.
67. Budd A, Devos DP (2012) Evaluating the Evolutionary Origins of Unexpected Character Distributions within the Bacterial Planctomycetes-Verrucomicrobia-Chlamydiae Superphylum. *Front Microbiol* 3: 401.
68. Berry EA, Huang LS, Saechao LK, Pon NG, Valkova-Valchanova M, et al (2004) X-Ray Structure of *Rhodobacter Capsulatus* Cytochrome bc<sub>1</sub>(1): Comparison with its Mitochondrial and Chloroplast Counterparts. *Photosynth Res* 81: 251–275.
69. Esser L, Elberry M, Zhou F, Yu CA, Yu L, et al (2008) Inhibitor-complexed structures of the cytochrome bc<sub>1</sub> from the photosynthetic bacterium *Rhodobacter sphaeroides*. *J Biol Chem* 283: 2846–2857.
70. Zhang Z, Huang L, Shulmeister VM, Chi YI, Kim KK, et al (1998) Electron transfer by domain movement in cytochrome bc<sub>1</sub>. *Nature* 392: 677–684.
71. Kolling DJ, Brunzelle JS, Lhee S, Crofts AR, Nair SK (2007) Atomic resolution structures of rieske iron-sulfur protein: role of hydrogen bonds in tuning the redox potential of iron-sulfur clusters. *Structure* 15: 29–38.
72. Degli Esposti M, De Vries S, Crimi M, Ghelli A, Patarnello T, et al (1993). Mitochondrial cytochrome b: evolution and structure of the protein. *Biochim Biophys Acta* 1143: 243–271.
73. Valas RE, Bourne PE (2009) Structural analysis of polarizing indels: an emerging consensus on the root of the tree of life. *Biol Direct* 4: 30.
74. Ouchane S, Nitschke W, Bianco P, Vermeglio A, Astier C (2005) Multiple Rieske genes in prokaryotes: exchangeable Rieske subunits in the cytochrome bc-complex of *Rubrivivax gelatinosus*. *Mol Microbiol* 57: 261–275.
75. Lücker S, Wagner M, Maixner F, Pelletier E, Koch H, et al (2010) A *Nitrospira* metagenome illuminates the physiology and evolution of globally important nitrite-oxidizing bacteria. *Proc Natl Acad Sci U S A* 107: 13479–13484.
76. Schübbe S, Williams TJ, Xie G, Kiss HE, Brettin TS, et al (2009) Complete genome sequence of the chemolithoautotrophic marine magnetotactic coccus strain MC-1. *Appl Environ Microbiol* 75: 4835–4852.
77. Simmons SS, Isokpehi RD, Brown SD, McAllister DL, Hall CC, et al (2011) Functional Annotation Analytics of *Rhodospseudomonas palustris* Genomes. *Bioinform Biol Insights* 5: 115–129.
78. Bratton MR, Pressler MA, Hosler JP (1999) Suicide inactivation of cytochrome c oxidase: catalytic turnover in the absence of subunit III alters the active site. *Biochemistry* 38: 16236–16245.
79. Cunningham L, Pitt M, Williams HD (1997) The cioAB genes from *Pseudomonas aeruginosa* code for a novel cyanide-insensitive terminal oxidase related to the cytochrome bd quinol oxidases. *Mol Microbiol* 24: 579–591.
80. Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ, et al (2012). Domain enhanced lookup time accelerated BLAST. *Biol Direct* 7: 12.
81. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ (1998) *Bioinformatics* 14: 892–893.
82. Price MN, Arkin AP, Alm EJ (2006) The life-cycle of operons. *PLoS Genet* 2: e96.
83. Hung CL, Lee C, Lin CY, Chang CH, Chung YC, et al (2010) Feature amplified voting algorithm for functional analysis of protein superfamily. *BMC Genomics* 11 (Suppl 3): S14.
84. Dam B, Dam S, Kube M, Reinhardt R, Liesack W (2012) Complete genome sequence of *Methylocystis* sp. strain SC2, an aerobic methanotroph with high-affinity methane oxidation potential. *J. Bacteriol* 194: 6008–6009.
85. Lange C, Hunte C (2002) Crystal structure of the yeast cytochrome bc<sub>1</sub> complex with its bound substrate cytochrome c. *Proc Natl Acad Sci U S A* 99: 2800–2805.
86. Yang D, Oyaizu Y, Oyaizu H, Olsen GJ, Woese CR (1985) Mitochondrial origins. *Proc Natl Acad Sci U S A* 82: 4443–4447.
87. Fitzpatrick DA, Creevey CJ, McInerney JO (2006) Genome phylogenies indicate a meaningful alpha-proteobacterial phylogeny and support a grouping of the mitochondria with the Rickettsiales. *Mol Biol Evol* 23: 74–85.
88. Sakurai K, Arai H, Ishii M, Igarashi Y (2011) Transcriptome response to different carbon sources in *Acetobacter acetii*. *Microbiology* 157 (Pt 3): 899–910.
89. Gómez-Manzo S, Chavez-Pacheco JL, Contreras-Zentella M, Sosa-Torres ME, Arreguín-Espinosa R, et al (2010) Molecular and catalytic properties of the aldehyde dehydrogenase of *Gluconacetobacter diazotrophicus*, a quinoheme protein containing pyrroloquinoline quinone, cytochrome b, and cytochrome c. *J Bacteriol* 192: 5718–5724.

RESEARCH

Open Access

# ABC transporters are involved in defense against permethrin insecticide in the malaria vector *Anopheles stephensi*

Sara Epis<sup>1†</sup>, Daniele Porretta<sup>2†</sup>, Valentina Mastrantonio<sup>2</sup>, Francesco Comandatore<sup>1,3</sup>, Davide Sasserà<sup>3</sup>, Paolo Rossi<sup>4</sup>, Claudia Cafarchia<sup>5</sup>, Domenico Otranto<sup>5</sup>, Guido Favia<sup>4</sup>, Claudio Genchi<sup>1</sup>, Claudio Bandi<sup>1\*</sup> and Sandra Urbanelli<sup>2</sup>

## Abstract

**Background:** Proteins from the ABC family (ATP-binding cassette) represent the largest known group of efflux pumps, responsible for transporting specific molecules across lipid membranes in both prokaryotic and eukaryotic organisms. In arthropods they have been shown to play a role in insecticide defense/resistance. The presence of ABC transporters and their possible association with insecticide transport have not yet been investigated in the mosquito *Anopheles stephensi*, the major vector of human malaria in the Middle East and South Asian regions. Here we investigated the presence and role of ABCs in transport of permethrin insecticide in a susceptible strain of this mosquito species.

**Methods:** To identify ABC transporter genes we obtained a transcriptome from untreated larvae of *An. stephensi* and then compared it with the annotated transcriptome of *Anopheles gambiae*. To analyse the association between ABC transporters and permethrin we conducted bioassays with permethrin alone and in combination with an ABC inhibitor, and then we investigated expression profiles of the identified genes in larvae exposed to permethrin.

**Results:** Bioassays showed an increased mortality of mosquitoes when permethrin was used in combination with the ABC-transporter inhibitor. Genes for ABC transporters were detected in the transcriptome, and five were selected (*AnstABCB2*, *AnstABCB3*, *AnstABCB4*, *AnstABCmember6* and *AnstABCG4*). An increased expression in one of them (*AnstABCG4*) was observed in larvae exposed to the LD50 dose of permethrin. Contrary to what was found in other insect species, no up-regulation was observed in the *AnstABCB* genes.

**Conclusions:** Our results show for the first time the involvement of ABC transporters in larval defense against permethrin in *An. stephensi* and, more in general, confirm the role of ABC transporters in insecticide defense. The differences observed with previous studies highlight the need of further research as, despite the growing number of studies on ABC transporters in insects, the heterogeneity of the results available at present does not allow us to infer general trends in ABC transporter-insecticide interactions.

**Keywords:** Mosquitoes, Bioassays, Insecticide resistance, Culicidae, Vector control, ABC transporters

## Background

Malaria is a major threat to human health and socio-economic development, representing a great burden in the vast regions of the world in which this parasitosis is endemic [1-3]. WHO estimated over 200 million cases of malaria in the 99 endemic countries and around 660,000 deaths, in the year 2010 [2].

Vector control through insecticides is a core component of malaria control programmes. However, continuous use of insecticides has led to the development of resistance in many malaria vectors around the world, which poses a serious threat to the global malaria control efforts [3-5]. Research is therefore needed to understand the molecular basis of insecticide detoxification and develop even more effective methods to delay emergence of resistance [6].

In recent years, the role of ATP-binding cassette (ABC) transporters in the defense against toxic compounds as

\* Correspondence: claudio.bandini@unimi.it

†Equal contributors

<sup>1</sup>Department of Veterinary Science and Public Health, University of Milan, Milan, Italy

Full list of author information is available at the end of the article



pesticides has attracted a great deal of attention (reviewed in [7,8]). ABC transporters are ATP-dependent efflux pumps belonging to the ABC protein family located in the cellular membrane in both prokaryotic and eukaryotic organisms. In eukaryotic organisms, they mediate the efflux of compounds from the cytoplasm to the outside of the cell or into organelles. ABC proteins have been subdivided into eight subfamilies (from ABC-A to ABC-H), and can transport a wide array of different substrates across cellular membranes (e.g., amino-acids, sugars, lipids, and peptides). Most of the ABC transporters associated with the efflux of pesticides belong to the subfamilies ABC-B (also referred to as P-glycoproteins, P-gps), ABC-C and ABC-G. In some cases ABC-transporter action has also been associated with insecticide-resistant phenotypes in species of agricultural or medical importance [7,8]. In spite of an increasing awareness of the potential importance of ABC transporters in vector control, to date they have been poorly studied in detail in malaria vectors [8]. Here, we investigated the role of ABC transporters in the detoxification against the insecticide permethrin in the malaria vector *Anopheles stephensi* (Culicidae: Diptera). This mosquito species, vector of both *Plasmodium falciparum* and *Plasmodium vivax*, is one of the major vectors of human malaria in the world. *An. stephensi* occupies a geographic range that spans from the Middle East to South-East Asia [9]. These regions contribute to 15% of malaria cases worldwide, with an estimated 28 million people annually affected by the disease [2]. Permethrin belongs to the pyrethroid class of insecticides, which is by far the most commonly used in malaria vector-control interventions [2].

Pyrethroids act by modifying the gating kinetics of voltage-gated sodium channels, thereby disrupting neuron function, which leads to rapid paralysis and death of the insect [10]. They can enter into the insect body by ingestion and penetration into the hemolymph through the alimentary canal, or via contact with sensory organs of the peripheral nervous system [11]. Insect midgut is rich in ABC transporters, whose action, therefore, likely prevents permethrin to reach its target sites [8]. Furthermore, insects possess protective neural barriers (e.g. a layer of glially derived epithelial cells), where ABC transporters likely play an important role in the exchange of molecules [12,13]. In particular, inhibition of P-gp in *Schistocerca gregaria* has been shown to increase brain uptake of different drugs [13]. The involvement of ABC transporters in pyrethroid detoxification has been reported for a few insect species, such as *Helicoverpa armigera* [14-16], *Apis mellifera* [17] and *Culex pipiens* [18]. Up-regulation of ABC-transporter genes has also been reported in pyrethroid resistant strains of the bed bugs *Cimex lectularius* [19] and of the vector mosquitoes *Anopheles gambiae* [20] and *Aedes aegypti* [21]. No protein belonging to the ABC transporters

has yet been described in larvae of *An. stephensi*, nor the possible association of this class of proteins with insecticide transport has been investigated in this species. In this paper we investigated the presence and role of ABCs in transport of permethrin insecticide in larvae of *An. stephensi*: *i*) by bioassays with permethrin alone and in combination with an ABC inhibitor; *ii*) by investigating gene expression profiles in larvae exposed to permethrin treatment.

## Methods

### Mosquito samples

The mosquito larvae used in this study were obtained from adult females of a *An. stephensi* colony, derived from the Liston strain. This colony has been maintained for four years in the insectary at the University of Camerino, following standard conditions: adult insects are reared at  $28 \pm 1^\circ\text{C}$  and 85-90% relative humidity with photoperiods (12:12 L-D) with a 5% sucrose solution, and adult females are fed with mouse blood for egg laying. Eggs from this colony were put into spring water in order to obtain the larvae. Larvae were maintained in spring water and fed daily with fish food (Tetra, Melle, Germany) under the same conditions as the adults.

### Bioassays

Inhibition of ABC-transporters should lead to a higher intracellular concentration of insecticide, thus increasing larval susceptibility and insecticide efficacy [8]. In order to evaluate a potential synergy, we performed bioassays with permethrin insecticide alone and with permethrin in combination with a sub-lethal dose of the ABC-transporter inhibitor verapamil (see below for experimental determination of sub-lethal dose of verapamil). This is a calcium channel blocker, which works by competing with cytotoxic compounds for efflux by the membrane pumps [22]. All bioassays were conducted on *An. stephensi* larvae at the third instar, according to standard protocols [23].

Groups of 25 larvae were put in 250 ml plastic glasses with 100 ml of spring water and different concentrations of insecticide or insecticide + inhibitor. All tests were performed in quadruple. Additional groups of larvae, treated only with water and acetone (that was used to dilute permethrin), were used as controls. Mortality was assessed at 24 h post-treatment and the larvae were considered dead if immobile, even after a mechanical stimulus.

In the bioassays with permethrin alone (Sigma-Aldrich S.r.l., Milan, Italy), six insecticide concentrations were used (0.015, 0.047, 0.092, 0.23, 0.57, 1.44 mg/l) to have mortality in the range 1–99%. The drug was dissolved in acetone and then diluted in water to obtain the test solutions. The bioassays with permethrin in combination with verapamil were performed using permethrin at the six concentrations indicated above, plus two additional

concentrations (0.0024 and 0.0048 mg/l). The sub-lethal dose of verapamil (i.e. the dose at which no dead larvae were observed) was determined using ten different concentrations (20, 40, 80, 100, 160, 240, 320, 400, 480, 560  $\mu$ M) following the protocol above. The larval mortality data were subjected to Probit regression analysis [24] as implemented in the XLSTAT-Dose software (available at: <http://www.xlstat.com>) to estimate the LD50 values and their 95% confidence intervals (CIs). To estimate the effect on larval mortality of the ABC inhibitor at sub-lethal dose, the synergistic factor (SF) was calculated.

### Identification of ABC transporter genes

A total of 200 untreated larvae of *An. stephensi* at the third instar were pooled in 15 ml of RNAlater stabilization solution (Qiagen, Hilden, Germany) and provided to an external company (GATC Biotech AG, Costance, Germany) for one run of 2x250 paired-ends reads sequencing on the Illumina MiSeq platform. The resulting reads were assembled using Trinity with default settings [25]. The assembled contigs were compared with Blastx (evalue 0.00001) to the annotated transcriptome of *An. gambiae* available in the VectorBase database, and the sequences of ABC transporters were extracted automatically and manually controlled. Based on published results about the involvement of ABCs on multidrug resistance in several arthropods (mainly mosquitoes) [8], we selected five genes from the transcriptome of *An. stephensi*. Oligonucleotide primers were then designed from the sequence of each gene (Table 1). The sequences of ABC transporters identified in *An. stephensi* were translated to aminoacids and compared against the UniProt database [26] using Blastp. Homologous proteins were aligned using ClustalX [27] and distances among them were estimated by Dayhoff PAM matrix as implemented in the PROTDIST software of the PHYLIP package [28].

### Gene expression profile after insecticide treatment

The activity of ABC-transporters is generally modulated at gene transcriptional level: the presence of toxic compounds leads to higher transcription. In order to assess this topic, larvae of *An. stephensi* at the third instar were

exposed to permethrin and the expression of ABC-transporter genes was monitored in the surviving larvae by quantitative RT-PCR twice after insecticide treatment: 24 h (e.g. the time at which the LD50 has been estimated) and 48 h following the study of Figueira-Mansur [29] that found increased expression of ABC transporters in the mosquito *Ae. aegypti*. The larvae were treated with the LD50 (0.137 mg/l) of insecticide estimated by bioassays as described above and two pools, of ten larvae each, were collected after 24 and 48 h of insecticide-treatment. All pools of larvae were stored in RNAlater for molecular analysis and, controls (water + acetone) were collected following the same time frame.

RNA was extracted from each pool of larvae using the RNeasy Mini Kit (Qiagen, Hilden, Germany) including an on-column DNase I treatment (Qiagen, Hilden, Germany), according to the manufacturer's instructions. Total RNA was eluted into nuclease-free water and the concentration of RNA was determined at 260 nm [30] using a NanoDrop ND-1000 (Thermo Scientific, Delaware, USA). cDNAs were synthesized from 250 ng of total RNA using a QuantiTect Reverse Transcription Kit (Qiagen, Hilden, Germany) with random hexamers. The cDNA was used as template in RT-PCR reactions using the primers designed from the sequences of identified ABC genes (Table 1). The amplification fragments, obtained using standard PCR conditions and the thermal profile indicated below, were sequenced in order to confirm the specificity of the amplification.

Quantitative RT-PCRs on the target ABCs were performed using a BioRad iQ5 Real-Time PCR Detection System (Bio-Rad, California, USA), under the following conditions: 50 ng cDNA; 300 nM of forward and reverse primers; 98°C for 30 sec, 40 cycles of 98°C for 15 sec, 59°C for 30 sec; fluorescence acquisition at the end of each cycle; melting curve analysis after the last cycle. The cycle threshold (Ct) values were determined for each gene, in order to calculate gene expression levels of target genes relative to *rps7*, the internal reference gene for *An. stephensi* [31]. The expression of the ABC transporters genes in the control group was considered as the basal level (equal to 1). The estimates of the expression level of each gene in the treated larvae are reported as

**Table 1 Primer sequences of ABC transporter genes identified in *Anopheles stephensi***

Gene	Forward primer	Reverse primer	PCR product size (base pairs)	Source
<i>Anst</i> ABC2	TATCAAGTTCACGGATGTAGAGT	TATCCACCTTGCCACTGTC	185	This work
<i>Anst</i> ABC3	CAACCGTTCGGTAATACTACC	ACTGGTAGCCCAATGTGAAG	133	This work
<i>Anst</i> ABC4	GGACAAAACATTCGGGAGG	CGTAGTGAATGTTGTGGCG	109	This work
<i>Anst</i> ABCmemb6	CTGGAGACGCTGAGAGATA	TACTCCTCGGTGAAGTGG	125	This work
<i>Anst</i> ABC4	ATGAGCCCATTCGTCCTG	AGCGTGGAGAAGAAGCAG	158	This work
<i>Rps7</i>	AGCAGCAGCAGCACTTGATTG	TAAACGGCTTCTCGCTACCC	90	Capone et al. 2013 [31]

the means  $\pm$  standard deviation (SD) in Additional file 1: Table S1.

### Ethical statement

Maintenance of the mosquito colony of *An. stephensi* was carried out according the Italian Directive 116 of 10/27/92 on the “use and protection of laboratory animals” and in adherence with the European regulation (86/609) of 11/24/86 (licence no. 125/94A, issued by the Italian Ministry of Health).

## Results

### Bioassays

No mortality was observed when larvae were exposed at concentrations of verapamil up to 100  $\mu$ M; this concentration was thus used as the sub-lethal dose in the bioassays with insecticide + ABC-transporter inhibitor. The results of toxicity assays using permethrin and permethrin in combination with verapamil are reported in Table 2. The mortality data observed in bioassays well fitted the Probit dose–response model (Chi-Square probability <0.0001). The LD50 dose in permethrin assay was 0.137 mg/l while in the assay in combination with verapamil LD50 was 0.025 mg/l (Table 2). No overlapping values were observed between LD50 95% CI of insecticide alone and insecticide plus verapamil; the addition of verapamil increased the toxicity of permethrin about 5-fold (SF = 5.48).

### Isolation of ABC transporter genes and expression profile after insecticide treatment

The Illumina MiSeq platform was used to sequence the cDNA library obtained from a pool of 200 larvae of *An. stephensi*, and 16,686,276 paired-ends reads were obtained. MiSeq raw data were assembled with Trinity, obtaining 40,498 contigs. The contigs containing ABC transporter genes were extracted on the basis of the annotated transcriptome of *An. gambiae* available in database. Five sequences, respectively of 3612, 2154, 2481, 2553 and 2182 base pairs, were found to share 85-94% identity with putative ABC multidrug transporters of *An.*

*gambiae*: ABCB2 (AGAP005639) (85% identity), ABCB3 (AGAP006273) (94% identity), ABCB4 (AGAP006364) (88% identity), ABCmember6 (AGAP002278) (94% identity) and ABCG4 (AGAP001333) (85% identity). We denoted them as *Anst*ABCB2, *Anst*ABCB3, *Anst*ABCB4, *Anst*ABCmember6, *Anst*ABCG4 and we deposited them in EMBL Nucleotide Sequence Database [EMBL: LK392613 to LK392617]. The alignment of the deduced amino acidic sequences of ABC transporters identified in *An. stephensi* with sequences of homologous ABC transporters of *An. gambiae* is shown in Additional file 2: Figure S1. Dayhoff PAM distance estimates between the ABC transporters identified in *An. stephensi* and the homologous ABC transporters of mosquitoes *An. gambiae*, *Anopheles darlingi*, *Ae. aegypti* and *Culex quinquefasciatus* that showed the highest percentage of identity following Blast search are shown in Additional file 3: Table S2.

Conventional PCR amplicons obtained from each gene primer set were sequenced, confirming in all cases the sequences generated with the MiSeq experiment. The RT-PCRs were performed to investigate whether permethrin treatment at the LD50 dose (0.137 mg/l, Table 2) increased or decreased the ABC gene expression in the *An. stephensi* larvae after 24 and 48 h of insecticide-treatment. As reported in Figure 1 and Additional file 1: Table S1, after 24 h of permethrin treatment, the relative expression of all selected genes was down-regulated, with the exception of the *Anst*ABCG4 gene, that showed about three-fold increase of expression compared to the control. Similarly, after 48 h of permethrin treatment, the relative expression of all ABCB genes was down-regulated, while the *Anst*ABCG4 gene showed a ten-fold increase of expression compared to the control.

## Discussion

Bioassays and molecular data suggest the involvement of ABC transporters in the defense of *An. stephensi* larvae against the permethrin insecticide. Indeed, inhibition of ABC-transporters led to a higher susceptibility of larvae to insecticide, indicating that ABC transporters are associated with insecticide detoxification. In addition, in mosquito larvae exposed to the LD50 dose of permethrin, we observed an increased expression of *Anst*ABCG4, one of the five tested genes coding for ABC transporters.

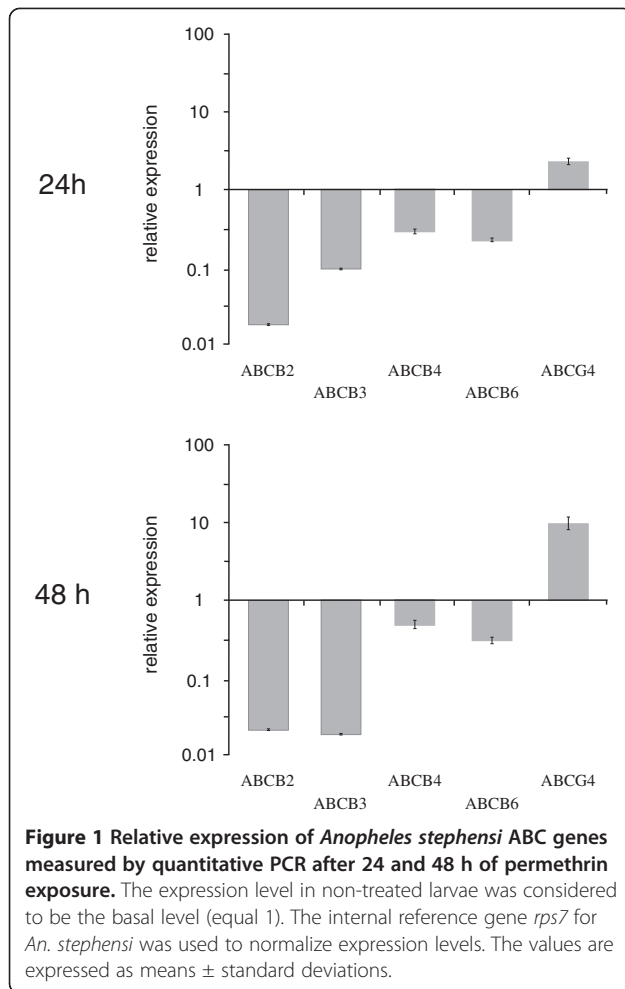
Arthropod ABCG genes are orthologous of the human gene *ABCG2*, which has been associated with resistance to cancer drugs [32,33], while data in insects show that *ABCG* transporter genes were significantly over-transcribed in response to exposure to insecticides. Microarray gene expression studies revealed that ABCG transporter genes were up-regulated in DDT resistant strains of *Drosophila melanogaster* [34] and in a *Plutella xylostella* (Lepidoptera) strain resistant to chlorpyrifos [35]. The ABCG4 transporter gene was over-transcribed in *Bemisia tabaci* whiteflies resistant to

**Table 2 Toxicity of verapamil and permethrin against *Anopheles stephensi* larvae**

Insecticide	N	Slope $\pm$ SE	LD50 (95% CI)	SF
Verapamil	600	3.846 (0.374)*	528 $\mu$ M (486-587)	
Permethrin	600	1.819 (0.125)*	0.137 mg/l (0.117-0.160)	
Permethrin + verapamil (100 $\mu$ M)	600	2.123 (0.174)*	0.025 mg/l (0.021-0.029)	5.48

LD50 and slopes of regression lines estimated from mortality data by Probit analysis are shown. N, number of larvae used in bioassays; SE, standard error; 95% CI, 95% confidence interval. SF, synergistic factor.

\*Chi-Square probability < 0.0001.



the neonicotinoid thiamethoxam [36] as well as in *Anopheles arabiensis* resistant- and sensible-strains to DDT [33]. The ABCG3 gene was found differentially expressed in *Ae. aegypti* pyrethroid resistant populations versus susceptible strains [21].

The other four genes coding for ABC transporters that we detected and tested in *An. stephensi* (*Anst*ABCB2, *Anst*ABCB3, *Anst*ABCB4 and *Anst*ABCmember6) belong to the ABCB subfamily. Several members of this subfamily have been associated with transport and/or resistance to different insecticide classes in several insect species [7,8]. In mosquitoes, the ABCB2 gene was showed by quantitative PCR to be eightfold up-regulated in larvae of a susceptible strain of *Ae. aegypti* analysed at 48 h post temephos-treatment [29]. The ABCB4 gene was showed to be over-transcribed by transcriptome and quantitative PCR analyses, in DDT and pyrethroid (permethrin and deltamethrin) resistant *Ae. aegypti* populations compared to a laboratory susceptible population [21]. Microarray analysis showed that the ABCB4 was up-regulated in different populations of DDT-resistant *An. gambiae* mosquitoes [37]. Our results showed no over-

transcription of the ABCB genes in *An. stephensi* susceptible larvae exposed to permethrin, while insecticide treatment induced an increased expression of the *Anst*ABCG4 (Figure 1, Additional file 1: Table S1).

On the whole, the results herein presented support the view of the involvement of ABC transporters in insecticide transport, although differences with previous studies have been observed. Are these differences due to the insecticides used or to the status of the analysed samples (i.e. susceptible vs. resistant)? Despite the growing number of studies on ABC transporters in insects, the heterogeneity of the data available at present does not allow to infer general trends that may underlie particular interactions between ABC transporters and insecticides. Further studies are needed to highlight these and other issues. For example, our results showed that the expression of ABCB genes in *An. stephensi* did not only increase in larvae treated with permethrin compared to those non-treated, but indeed it decreased (Figure 1, Additional file 1: Table S1). In the latter case, the study of gene expression at more time points could contribute to the understanding of whether there are temporal delays, or whether compound-specific or species-specific differences exist in their activation [38-40]. Furthermore, most studies have been conducted on larval stages [7,8]. The synergist and transcript profiles may differ between larval and adult stages, an interesting topic for future studies.

Diffusion of vectors of human diseases driven by human activities and global climate change as well as insurgence of insecticide resistance can seriously impact our ability to control vector-borne diseases [41-44]. Furthermore, environmental pollution and resistance phenomena clearly show the limits of the chemical approach for pest control and the need to delineate new strategies that optimize the use of available molecules, with the aim of reducing their impact on the environment [31,45-48].

In the last decades advances in molecular techniques have greatly improved our tools to investigate the dynamics of vector populations and of pesticide resistance insurgence [43,49-53]. More recently, next-generation sequencing technologies have offered unprecedented opportunities to investigate the molecular basis of the interaction between cellular defenses and insecticides [54]. In this context, the increasing interest about ABC transporters in transport and/or resistance against insecticides led to an increase of the information on these genes in various insect species and their association with insecticide detoxification [7,8].

## Conclusions

In this study we have demonstrated for the first time in the larvae of *An. stephensi* that verapamil increases the sensitivity to permethrin in laboratory assays; in addition,

we isolated five genes encoding for ABC transporters, and investigated their expression profile after exposure to permethrin. To analyse the potential role of ABC transporters in permethrin transport in *An. stephensi*, we performed bioassays using a sub-lethal dose of the ABC transporter inhibitor verapamil in association with permethrin. The results obtained using this approach highlight that the combination of insecticides with an ABC-transporter inhibitor can increase the efficacy of the insecticide molecule [18,29,55]. In prospect, combined treatments of insecticide plus ABC-transporter inhibitors could be proposed, with the objective of reducing the current dosages of insecticides or to prevent the development of resistance, and reduce environmental pollution [29,56]. The implementation of such a strategy would require the availability of gene- and species-specific inhibitors in order to avoid the serious consequences that would derive from a generic inhibition of ABC-transporters in non-target organisms. The study of ABC-transporters at the gene level is therefore crucial for the understanding of both their potential role as defense mechanisms and for their inhibition for vector control purposes.

## Additional files

**Additional file 1: Table S1.** Relative expression of *Anopheles stephensi* ABC genes measured by quantitative PCR after permethrin exposure. The expression level in non-treated larvae was considered to be the basal level (equal 1). The internal reference gene *rps7* for *An. stephensi* was used to normalize expression levels. The values are expressed as means  $\pm$  standard deviations.

**Additional file 2: Figure S1.** Alignment by ClustalW of deduced amino acidic sequences of ABC transporters of *Anopheles stephensi* and *Anopheles gambiae*. Asterisks: conserved amino acid residues; colons: conserved substitutions; dots: semiconserved substitutions.

**Additional file 3: Table S2.** Dayhoff PAM matrix. Estimates of distance among the ABC transporters identified in *Anopheles stephensi* and the homologous ABC transporters of other mosquitoes species are shown: *An. gambiae* (AGAP005639; AGAP006273; AGAP006364; AGAP002278; AP001333), *An. darlingi* (ETN61204; ETN66919; ETN62617; ETN64062; ETN58714), *Aedes aegypti* (AAEL010379; AAEL002468; AAEL006717; AAEL008134; AAEL003703), and *Culex quinquefasciatus* (EDS44274; EDS35382; EDS29700; EDS27088; EDS37204).

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

SE, DP, VM conceived the study, performed the experiments and contributed to data analysis, interpretation and manuscript writing. FC and DS performed the bioinformatic analyses; PR, GF and CF contributed to sample collection. DO, CG, CB and SU contributed to data interpretation and manuscript writing. All authors read and approved the final version of the manuscript.

## Acknowledgements

The authors are grateful to Dr. Valeria Mereghetti, Dr. Daniele Facchi and Dr. Massimo Pajoro for providing assistance to accomplish this research. This work was supported by MIUR Prin 2010–2011 (to C. Genchi).

## Author details

<sup>1</sup>Department of Veterinary Science and Public Health, University of Milan, Milan, Italy. <sup>2</sup>Department of Environmental Biology, University "La Sapienza"

of Rome, Rome, Italy. <sup>3</sup>Department of Biology and Biotechnology, University of Pavia, Pavia, Italy. <sup>4</sup>School of Bioscience and Veterinary Medicine, University of Camerino, Camerino, Italy. <sup>5</sup>Department of Veterinary Medicine, University of Bari, Bari, Italy.

Received: 5 June 2014 Accepted: 15 July 2014

Published: 29 July 2014

## References


1. Hay SI, Guerra CA, Tatem AJ, Noor AM, Snow RW: The global distribution and population at risk of malaria: past, present, and future. *Lancet Infect Dis* 2004, **4**:327–336. 1.
2. World Health Organization: World malaria report 2012. [www.who.int/malaria/publications/world\\_malaria\\_report\\_2012/report/en/](http://www.who.int/malaria/publications/world_malaria_report_2012/report/en/).
3. Alonso PL, Tanner M: Public health challenges and prospects for malaria control and elimination. *Nat Med* 2013, **19**(2):150–155.
4. Karunamoorthi K: Vector control: a cornerstone in the malaria elimination campaign. *Clin Microbiol Infect* 2011, **17**:1608–1616.
5. Tikar SN, Mendki MJ, Sharma AK, Sukumaran D, Veer V, Prakash S, Parashar BD: Resistance status of the malaria vector mosquitoes, *Anopheles stephensi* and *Anopheles subpictus* towards adulticides and larvicides in arid and semi-arid areas of India. *J Insect Sci* 2011, **11**:85.
6. Perry T, Batterham P, Daborn PJ: The biology of insecticidal activity and resistance. *Insect Biochem Mol Biol* 2011, **41**:411–422.
7. Buss DS, Callaghan A: Interaction of pesticides with p-glycoprotein and other ABC proteins: a survey of the possible importance to insecticide, herbicide and fungicide resistance. *Pestic Biochem Physiol* 2008, **90**:141–153.
8. Dermauw W, Van Leeuwen T: The ABC gene family in arthropods: comparative genomics and role in insecticide transport and resistance. *Insect Biochem Mol Biol* 2014, **45**:89–110.
9. Sinka ME, Bangs MJ, Manguin S, Chareonviriyaphap C, Patil AP, Temperley WH, Gething PW, Elyazar IRF, Kabaria CW, Harbach RE, Hay SI: The dominant anopheles vectors of human malaria in Asia-Pacific: occurrence data, distribution maps and bionomic précis. *Parasit Vectors* 2011, **4**:89.
10. Soderlund DM: Pyrethroids, knockdown resistance and sodium channels. *Pest Manag Sci* 2008, **64**:610–616.
11. Schleier JJ III, Peterson RKD: Pyrethrins and Pyrethroid Insecticides. In *Green Trends in Insect Control*. Edited by Lopez O, Fernandez-Bolanos JG. London, UK: Royal Society of Chemistry; 2011:94–131.
12. Mayer F, Mayer N, Chinn L, Pinsonneault RL, Kroetz D, Bainton RJ: Evolutionary conservation of vertebrate blood–brain barrier chemoprotective mechanisms in *Drosophila*. *J Neurosci* 2009, **29**:3538–3550.
13. Andersson O, Badisco L, Hakans Son Hansen A, Hansen SH, Hellman K, Nielsen PA, Olsen LR, Verdonck R, Abbott NJ, Vanden Broeck J, Andersson G: Characterization of a novel brain barrier *ex vivo* insect-based P-glycoprotein screening model. *Pharma Res Per* 2014, **2**(4):e00050.
14. Srinivas R, Shamsundar GS, Jayalakshmi SK, Sreeramulu K: Effect of insecticides and inhibitors on P-glycoprotein ATPase (M-type) activity of resistant pest *Helicoverpa armigera*. *Curr Sci* 2005, **88**:1449–1452.
15. Aurade R, Jayalakshmi SK, Sreeramulu K: Stimulatory effect of insecticides on partially purified P-glycoprotein ATPase from the resistant pest *Helicoverpa armigera*. *Biochem Cell Biol* 2006, **84**:1045–1050.
16. Aurade RM, Jayalakshmi SK, Sreeramulu K: P-glycoprotein ATPase from the resistant pest, *Helicoverpa armigera*: purification, characterization and effect of various insecticides on its transport function. *Biochim Biophys Acta* 2010, **1798**:1135–1143.
17. Hawthorne DJ, Dively GP: Killing them with kindness? In-hive medications may inhibit xenobiotic efflux transporters and endanger honey bees. *PLoS One* 2011, **6**:e26796.
18. Buss DS, McCaffery AR, Callaghan A: Evidence for p-glycoprotein modification of insecticide toxicity in mosquitoes of the *Culex pipiens* complex. *Med Vet Entomol* 2002, **16**(2):218–222.
19. Zhu F, Gujar H, Gordon JR, Haynes KF, Potter MF, Palli SR: Bed bugs evolved unique adaptive strategy to resist pyrethroid insecticides. *Sci Rep* 2013, **3**:1456.
20. Bonizzoni M, Afrane Y, Dunn WA, Atieli FK, Zhou G, Zhong D, Li J, Githeko A, Yan G: Comparative transcriptome analyses of deltamethrin-resistant and -susceptible *Anopheles gambiae* mosquitoes from Kenya by RNA-Seq. *PLoS One* 2012, **7**:e44607.
21. Bariami V, Jones CM, Poupardin R, Vontas J, Ranson H: Gene amplification, ABC transporters and cytochrome P450s: unraveling the molecular basis

- of pyrethroid resistance in the dengue vector, *Aedes aegypti*. *PLoS Negl Trop Dis* 2012, **6**:e1692.
22. Thomas H, Coley HM: **Overcoming multidrug resistance in cancer: an update on the clinical strategy of inhibiting P-glycoprotein.** *Cancer Control* 2003, **10**(02):159–165.
  23. World Health Organization: *Guidelines for Laboratory and Field Testing of Mosquito Larvicides. Communicable Disease Control, Prevention and Eradication, WHO Pesticide Evaluation Scheme.* Geneva: WHO; 2005. WHO/CDS/WHOPES/GCDPP/2005.13.
  24. Finney DJ: *Probit Analysis.* Cambridge: Cambridge University Press; 1971.
  25. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A: **Full-length transcriptome assembly from RNA-Seq data without a reference genome.** *Nat Biotechnol* 2011, **29**:644–652.
  26. Uniprot Consortium: **activities at the Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2014, **42**:D191–D198.
  27. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL\_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Res* 1997, **25**(24):4876–4882.
  28. Felsenstein J: **PHYMLIP – phylogeny inference package (version 3.2).** *Cladistics* 1989, **5**:164–166.
  29. Figueira-Mansur J, Ferreira-Pereira A, Mansur JF, Franco TA, Alvarenga ES, Sorgine MH, Neves BC, Melo AC, Leal WS, Masuda H, Moreira MF: **Silencing of P-glycoprotein increases mortality in temephos-treated *Aedes aegypti* larvae.** *Insect Mol Biol* 2013, **22**(6):648–658.
  30. Lee EJ, Schmittgen TD: **Comparison of RNA assay methods used to normalize cDNA for quantitative real-time PCR.** *Anal Biochem* 2006, **357**(2):299–301.
  31. Capone A, Ricci I, Damiani C, Mosca M, Rossi P, Scuppa P, Crotti E, Epis S, Angeletti M, Valzano M, Sacchi L, Bandi C, Daffonchio D, Mandrioli M, Favia G: **Interactions between *Asaia*, *Plasmodium* and *Anopheles*: new insights into mosquito symbiosis and implications in malaria symbiotic control.** *Parasit Vectors* 2013, **6**:182.
  32. Tarr PT, Tarling EJ, Bojanic DD, Edwards PA, Baldan A: **Emerging new paradigms for ABCG transporters.** *Biochim Biophys Acta* 2009, **1791**(7):584–593.
  33. Jones CM, Toé HK, Sanou A, Namountougou M, Hughes A, Diabaté A, Dabiré R, Simard F, Ranson H: **Additional selection for insecticide resistance in urban malaria vectors: DDT resistance in *Anopheles arabiensis* from Bobo-Dioulasso, Burkina Faso.** *PLoS One* 2012, **7**(9):e45995.
  34. Pedra JHF, McIntyre LM, Scharf ME, Pittendrigh BR: **Genome-wide transcription profile of field- and laboratory-selected dichlorodiphenyltrichloroethane (DDT) resistant *Drosophila*.** *Proc Natl Acad Sci U S A* 2004, **101**:7034–7039.
  35. You M, Yue Z, He W, Yang X, Yang G, Xie M, Zhan D, Baxter SW, Vasseur L, Gurr GM, Douglas CJ, Bai J, Wang P, Cui K, Huang S, Li X, Zhou Q, Wu Z, Chen Q, Liu C, Wang B, Li X, Xu X, Lu C, Hu M, Davey JW, Smith SM, Chen M, Xia X, Tang W, et al: **A heterozygous moth genome provides insights into herbivory and detoxification.** *Nat Genet* 2013, **45**(2):220–225.
  36. Yang N, Xie W, Yang X, Wang S, Wu Q, Li R, Pan H, Liu B, Shi X, Fang Y, Xu B, Zhou X, Zhang Y: **Transcriptomic and proteomic responses of sweetpotato whitefly, *Bemisia tabaci*, to thiamethoxam.** *PLoS One* 2013, **8**(5):e61820.
  37. Fossog Tene B, Poupardin R, Costantini C, Awono-Ambene P, Wondji CS, Ranson H, Antonio-Nkondjio C: **Resistance to DDT in an urban setting: common mechanisms implicated in both M and S forms of *Anopheles gambiae* in the city of Yaoundé Cameroon.** *PLoS One* 2013, **8**(4):e61408.
  38. Hayashi K, Schoonbeek H, Sugiura H, De Waard MA: **Multidrug resistance in *Botrytis cinerea* associated with decreased accumulation of the azole fungicide oxpoconazole and increased transcription of the ABC transporter gene *BcatrD*.** *Pestic Biochem Physiol* 2001, **70**:168–179.
  39. Urbanelli S, Della Rosa V, Punelli F, Porretta D, Reverberi M, Fabbri AA, Fanelli C: **DNA-fingerprinting (AFLP and RFLP) for genotypic identification in species of the *Pleurotus eryngii* complex.** *Appl Microbiol Biotechnol* 2007, **74**(3):592–600.
  40. Punelli F, Reverberi M, Porretta D, Nogarotto S, Fabbri A, Fanelli C, Urbanelli S: **Molecular characterization and enzymatic activity of laccases in two *Pleurotus* spp. with different pathogenic behavior.** *Mycol Res* 2009, **113**:381–387.
  41. Dantas-Torres F, Figueredo LA, Otranto D: **Seasonal variation in the effect of climate on the biology of *Rhipicephalus sanguineus* in southern Europe.** *Parasitology* 2011, **138**:527–536.
  42. Porretta D, Mastrantonio V, Bellini R, Somboon P, Urbanelli S: **Glacial history of a modern invader: phylogeography and species distribution modelling of the asian tiger mosquito *Aedes albopictus*.** *Plos One* 2012, **7**(9):e44515.
  43. Porretta D, Mastrantonio V, Amendolia S, Gaiarsa S, Epis S, Genchi C, Bandi C, Otranto D, Urbanelli S: **Effects of global changes on the climatic niche of the tick *Ixodes ricinus* inferred by species distribution modelling.** *Parasit Vectors* 2013, **6**:271.
  44. Steele J, Orsel K, Cuyler C, Hoberg EP, Schmidt NM, Kutz SJ: **Divergent parasite faunas in adjacent populations of west Greenland caribou: natural and anthropogenic influences on diversity.** *Int J Parasitol Parasites Wildl* 2013, **2**:197–202.
  45. Porretta D, Canestrelli D, Bellini R, Celli G, Urbanelli S: **Improving insect pest management through population genetic data: a case study of the mosquito *Ochlerotatus caspius* (Pallas).** *J Appl Ecol* 2007, **44**:682–691.
  46. Otranto D, Wall R: **New strategies for the control of arthropod vectors of disease in dogs and cats.** *Med Vet Entomol* 2008, **22**:291–302.
  47. Calvitti M, Moretti R, Porretta D, Bellini R, Urbanelli S: **Effects on male fitness of removing wolbachia infections from the mosquito *Aedes albopictus*.** *Med Vet Entomol* 2009, **23**(2):132–140.
  48. Bouyer F, Hamadou S, Adakal H, Lancelot R, Stachurski F, Belem AM, Bouyer J: **Restricted application of insecticides: a promising tsetse control technique, but what do the farmers think of it?** *PLoS Negl Trop Dis* 2011, **5**(8):e1276.
  49. Marcombe S, Poupardin R, Darriet F, Reynaud S, Bonnet J, Strode C, Brengues C, Yébakima A, Ranson H, Corbel V, David JP: **Exploring the molecular basis of insecticide resistance in the dengue vector *Aedes aegypti*: a case study in Martinique Island (French West Indies).** *BMC Genomics* 2009, **10**:494.
  50. Bellini R, Albieri A, Balestrino F, Carrieri M, Porretta D, Urbanelli S, Calvitti M, Moretti R, Maini S: **Dispersal and survival of *Aedes albopictus* (Diptera: Culicidae) males in Italian urban areas and significance for sterile insect technique application.** *J Med Entomol* 2010, **47**(6):1082–1091.
  51. Porretta D, Canestrelli D, Urbanelli S, Bellini R, Schaffner F, Petric D, Nascetti G: **Southern crossroads of the Western Palaearctic during the late pleistocene and their imprints on current patterns of genetic diversity: insights from the mosquito *Aedes caspius*.** *J Biogeogr* 2011, **38**:20–30.
  52. Kennedy C, Tierney K: *Xenobiotic Protection/Resistance Mechanisms in Organisms.* New York: E.A. Laws, Environ Toxicol: Springer; 2013.
  53. Porretta D, Mastrantonio V, Mona S, Epis S, Montagna M, Sasserà D, Bandi C, Urbanelli S: **The integration of multiple independent data reveals an unusual response to pleistocene climatic changes in the hard tick *Ixodes ricinus*.** *Mol Eco* 2013, **22**(6):1666–1682.
  54. Jones CM, Haji KA, Khatib BO, Bagi J, Mcha J, Devine GJ, Daley M, Kabula B, Ali AS, Majambere S, Ranson H: **The dynamics of pyrethroid resistance in *Anopheles arabiensis* from Zanzibar and an assessment of the underlying genetic basis.** *Parasit Vectors* 2013, **6**:343.
  55. Porretta D, Gargani M, Bellini R, Medici A, Punelli F, Urbanelli S: **Defence mechanism against insecticides temephos and diflubenzuron in the mosquito *Aedes caspius*: the P-glycoprotein efflux pumps.** *Med Vet Entomol* 2008, **22**:48–54.
  56. Pohl PC, Klafke GM, Carvalho DD, Martins JR, Daffre S, da Silva Vaz I Jr, Masuda A: **ABC transporter efflux pumps: a defence mechanism against ivermectin in *Rhipicephalus (Boophilus) microplus*.** *Int J Parasitol* 2011, **41**:1323–1333.

doi:10.1186/1756-3305-7-349

**Cite this article as:** Epis et al: ABC transporters are involved in defense against permethrin insecticide in the malaria vector *Anopheles stephensi*. *Parasites & Vectors* 2014 **7**:349.

# Genomic Epidemiology of *Klebsiella pneumoniae* in Italy and Novel Insights into the Origin and Global Evolution of Its Resistance to Carbapenem Antibiotics

Stefano Gaiarsa,<sup>a,b</sup> Francesco Comandatore,<sup>b,c</sup> Paolo Gaibani,<sup>d</sup> Marta Corbella,<sup>a,e</sup> Claudia Dalla Valle,<sup>a</sup> Sara Epis,<sup>b</sup> Erika Scaltriti,<sup>f</sup> Edoardo Carretto,<sup>g</sup> Claudio Farina,<sup>h</sup> Maria Labonia,<sup>i</sup> Maria Paola Landini,<sup>d</sup>  Stefano Pongolini,<sup>f</sup> Vittorio Sambri,<sup>j</sup> Claudio Bandi,<sup>b</sup> Piero Marone,<sup>a</sup> Davide Sasserà<sup>c</sup>

Microbiology and Virology Unit, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy<sup>a</sup>; Dipartimento di Scienze Veterinarie e Sanità Pubblica (DIVET), Università degli Studi di Milano, Milan, Italy<sup>b</sup>; Dipartimento di Biologia e Biotecnologie, Università degli Studi di Pavia, Pavia, Italy<sup>c</sup>; Unit of Clinical Microbiology, St. Orsola-Malpighi University Hospital, Bologna, Italy<sup>d</sup>; Biometric and Medical Statistics Unit, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy<sup>e</sup>; Sezione Diagnostica di Parma, Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna (IZSLER), Parma, Italy<sup>f</sup>; Clinical Microbiology Laboratory, IRCCS Arcispedale S. Maria Nuova, Reggio Emilia, Italy<sup>g</sup>; Microbiology Institute, AO Papa Giovanni XXIII, Bergamo, Italy<sup>h</sup>; Dipartimento di Diagnostica di Laboratorio e Trasfusionale, IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo, Italy<sup>i</sup>; Unit of Clinical Microbiology, The Greater Romagna Area-Hub Laboratory, Pievesestina, Italy<sup>j</sup>

***Klebsiella pneumoniae* is at the forefront of antimicrobial resistance for Gram-negative pathogenic bacteria, as strains resistant to third-generation cephalosporins and carbapenems are widely reported. The worldwide diffusion of these strains is of great concern due to the high morbidity and mortality often associated with *K. pneumoniae* infections in nosocomial environments. We sequenced the genomes of 89 *K. pneumoniae* strains isolated in six Italian hospitals. Strains were selected based on antibiotic types, regardless of multilocus sequence type, to obtain a picture of the epidemiology of *K. pneumoniae* in Italy. Thirty-one strains were carbapenem-resistant *K. pneumoniae* carbapenemase producers, 29 were resistant to third-generation cephalosporins, and 29 were susceptible to the aforementioned antibiotics. The genomes were compared to all of the sequences available in the databases, obtaining a data set of 319 genomes spanning the known diversity of *K. pneumoniae* worldwide. Bioinformatic analyses of this global data set allowed us to construct a whole-species phylogeny, to detect patterns of antibiotic resistance distribution, and to date the differentiation between specific clades of interest. Finally, we detected an ~1.3-Mb recombination that characterizes all of the isolates of clonal complex 258, the most widespread carbapenem-resistant group of *K. pneumoniae*. The evolution of this complex was modeled, dating the newly detected and the previously reported recombination events. The present study contributes to the understanding of *K. pneumoniae* evolution, providing novel insights into its global genomic characteristics and drawing a dated epidemiological scenario for this pathogen in Italy.**

Multidrug resistance is currently a matter of concern worldwide. At the end of the 1970s, most *Escherichia coli* and *Klebsiella pneumoniae* strains encoded ampicillin-hydrolyzing  $\beta$ -lactamases, making it necessary to use third-generation cephalosporins. In the early 1980s, the first cases of resistance to these novel antibiotics were reported in *Enterobacteriaceae* (1) and were caused by genes classified as ESBL (extended-spectrum beta-lactamases). In 1985, the United States Food and Drug Administration approved the commercialization of imipenem, a molecule that showed activity against ESBL producers. This drug, and similar compounds that quickly followed (i.e., carbapenems), then were introduced into clinical practice and widely used.

In 2001, Yigit and colleagues reported a *K. pneumoniae* strain isolated in 1996 that exhibited resistance to the carbapenems imipenem and meropenem (2). The gene responsible for the resistance was identified as a group 2f, class A, carbapenem-hydrolyzing beta-lactamase, named *Klebsiella pneumoniae* carbapenemase 1 (KPC-1). Since its discovery, carbapenem resistance caused by the *bla*<sub>KPC</sub> gene has been reported increasingly in *K. pneumoniae* isolates, initially moving through the northeastern states (3, 4) and quickly becoming the most frequently found carbapenemase in the United States (5). The spread of KPC then continued, with reports from different countries appearing ceaselessly, to the point that today this is regarded as a worldwide issue (6).

The *bla*<sub>KPC</sub> gene is carried by a plasmid; thus, horizontal trans-

fer between various *K. pneumoniae* strains, as well as other bacterial species, could be expected and was extensively reported (7–9). Nevertheless, most of the clinical reports to date have been caused by *K. pneumoniae* isolates belonging to clonal complex 258 (CC258) (10). This complex comprises sequence type 258 (ST258) and single-allele mutant STs based on multilocus sequence typing (MLST), such as ST11 and ST512. These epidemiological data suggest a dissemination starting from a single ances-

Received 4 September 2014 Returned for modification 26 September 2014  
Accepted 26 October 2014

Accepted manuscript posted online 3 November 2014

Citation Gaiarsa S, Comandatore F, Gaibani P, Corbella M, Dalla Valle C, Epis S, Scaltriti E, Carretto E, Farina C, Labonia M, Landini MP, Pongolini S, Sambri V, Bandi C, Marone P, Sasserà D. 2015. Genomic epidemiology of *Klebsiella pneumoniae* in Italy and novel insights into the origin and global evolution of its resistance to carbapenem antibiotics. *Antimicrob Agents Chemother* 59:389–396.  
doi:10.1128/AAC.04224-14.

Address correspondence to Davide Sasserà, [davide.sassera@unipv.it](mailto:davide.sassera@unipv.it).

S.G. and F.C. contributed equally.

Supplemental material for this article may be found at <http://dx.doi.org/10.1128/AAC.04224-14>.

Copyright © 2015, American Society for Microbiology. All Rights Reserved.  
doi:10.1128/AAC.04224-14

tor and that CC258 presents a genomic background that is favorable both to the acquisition of plasmids bearing the *bla*<sub>KPC</sub> gene and to the clonal spread in nosocomial environments. In 2014, Deleo and colleagues (11) presented a phylogenomic study on 85 *K. pneumoniae* isolates belonging to CC258, detecting two subclades and concluding that an ~215-kb recombination event was at the origin of the differentiation between the two. A second comparative genomic analysis, presented by Chen and colleagues (12), detected an ~1.1-Mb recombination between an ST11 recipient and an ST442 donor as the event that originated the present ST258 strain.

Since the first finding of circulation of ESBL-producing *K. pneumoniae* in Italy in 1994, a rapid and extensive dissemination of different types of ESBLs has been reported (13–15). More recently, the first Italian KPC-positive *K. pneumoniae* strain, belonging to ST258, was isolated in a hospital in Florence in 2008 from an inpatient with a complicated intra-abdominal infection (16). Since then, the diffusion of carbapenemase-producing *K. pneumoniae* in Italy has been extremely rapid and characterized mainly by isolates of CC258 (i.e., ST258 and ST512) (17–19). ST512 in particular, first reported in Israel in 2006 (20), has been spreading in southern Europe and South America (11, 19). The sporadic detection of isolates belonging to other STs (e.g., ST101 and ST147) also have characterized the epidemiology of KPC *K. pneumoniae* in Italy (19).

The aim of this study was to evaluate the geographic and phylogenetic distribution of *K. pneumoniae* isolates of different antibiotypes, both at a national and a global scale. Thus, we sequenced and analyzed the genomes from 89 *K. pneumoniae* strains, collected in six Italian hospitals from 2006 to 2013, without any *a priori* knowledge of the sequence type. We compared this national collection to all of the *K. pneumoniae* genomes available from worldwide isolations to obtain insights into both the Italian epidemiology and the global structure of the species.

## MATERIALS AND METHODS

**Strain sampling.** Eighty-nine nonduplicate *K. pneumoniae* strains, collected from six different Italian hospitals, were included in this study without prior knowledge of the sequence type. Thirty-one were KPC producers, as demonstrated using phenotypical tests (positivity with disk diffusion synergy testing using a meropenem disk alone and in combination with aminophenylboronic acid) (21) and/or genotypical analysis (in-house methods based on reference 22); 29 were ESBL producers, as demonstrated using the procedure recommended by the CLSI (23), while 29 were susceptible to third-generation cephalosporins and carbapenems. Throughout this work, we refer to this last group of isolates as susceptible. Antimicrobial susceptibility testing was performed using a Vitek2 automated system (bioMérieux), and MICs were interpreted by following the European Committee on Antimicrobial Susceptibility Testing guidelines (24). The list of isolates, year, location of isolation, sequence type, and presence of selected antibiotic resistance genes are reported in Table S1 in the supplemental material.

**Genome sequences.** DNA was extracted using a QIAamp DNA mini-kit (Qiagen) by following the manufacturer's instructions. Whole genomic DNA was sequenced using an Illumina Miseq platform with a 2 by 250 paired-end run after Nextera XT paired-end library preparation. On 24 March 2014, sequences of draft and complete genomes of *K. pneumoniae* were retrieved from the NCBI ftp site, while sequencing reads of the isolates sequenced by Deleo and coworkers (11) were retrieved from the sequence read archive (SRA) database (accession no. SRP036874).

**Genome assembly and retrieval.** Sequencing reads from the isolates obtained in this study were assembled using MIRA 4.0 software (25) with

accurate *de novo* settings. Assembled genomes are now publicly available under Bioproject (EMBL project B6543). Reads retrieved from the SRA database were checked and filtered for sequencing quality using an in-house script and then assembled using Velvet (26) with a K-mer length of 35 and automatic detection of average expected coverage and low coverage threshold.

**Resistance profile and MLST determination.** The MLST profile was obtained *in silico* by searching the characterizing gene variants on each genome, using an in-house Python script. The antibiotic resistance profile was determined using a BLAST search on a gene database comprising all of the most common resistance genes associated with resistance to beta-lactams, including ESBL- and KPC-producing phenotypes.

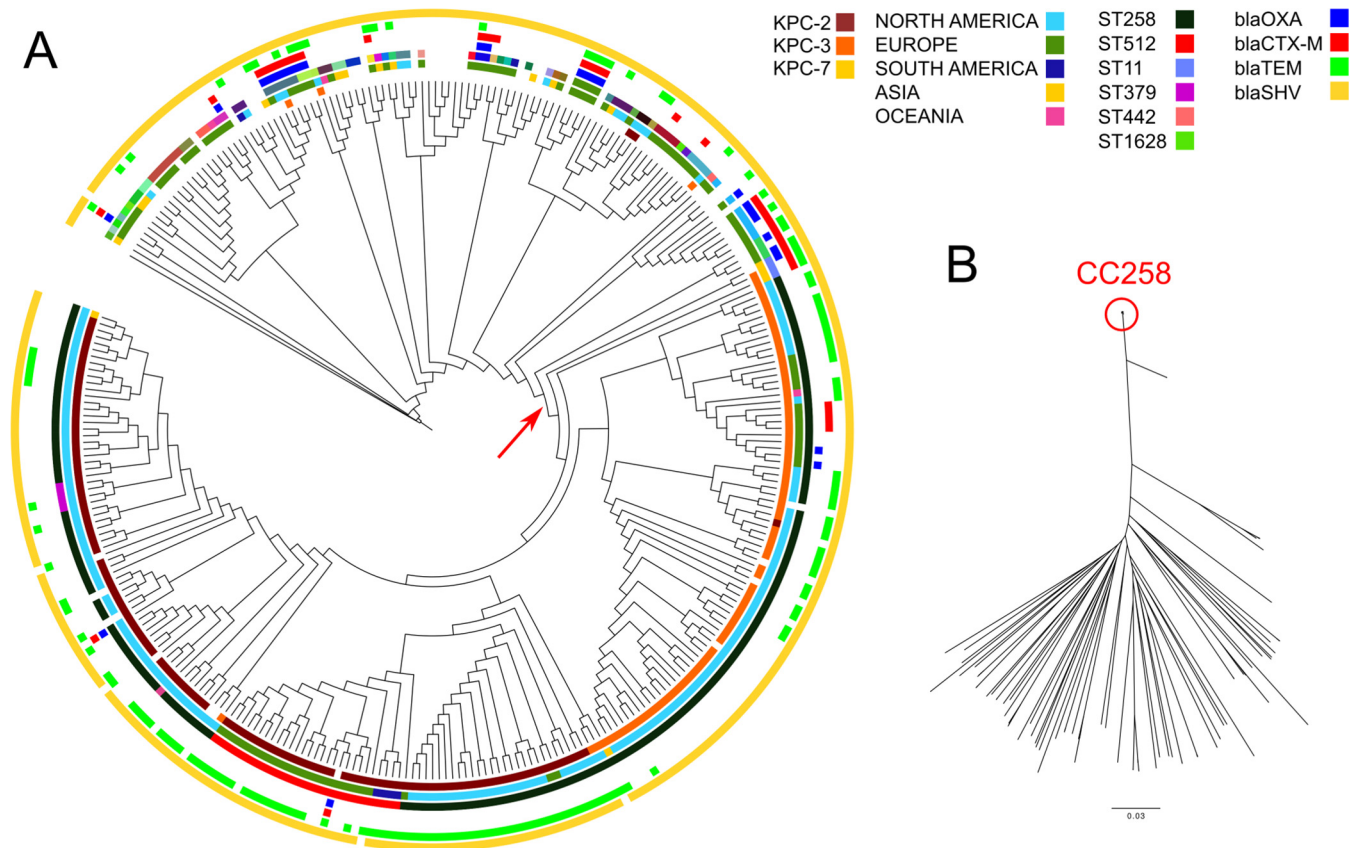
**Core SNP detection and phylogeny.** Single-nucleotide polymorphisms (SNPs) were detected using an in-house pipeline based on Mauve software (27), using the NJST258\_1 complete genome as a reference. Each genome was individually aligned to the reference, and alignments were merged with in-house scripts. Core SNPs were defined as single-nucleotide mutations flanked by identical bases present in all of the analyzed genomes. The core SNP alignment was used to perform a phylogenetic analysis using the software RAxML (28) with a generalized time-reversible (GTR) model and 100 bootstraps. The same phylogenetic approach was used to perform the analysis on three core SNP sub-data sets (i.e., nonrecombined regions and two distinct putatively recombined regions).

**Recombination.** We divided the genome alignment in 5,264 windows of 1,000 nucleotides (nt) each and calculated core SNP frequency in each window for each genome, generating a matrix. The software R then was used to generate a heatmap of SNP frequency. The newly characterized strain 46AVR was used as a reference for plotting SNPs, being a member of the sister group to CC258. In parallel, we created a sub-data set of 174 CC258 genomes and 13 closely related *K. pneumoniae* genomes, removing genomes of isolates distant from the CC258 clade ( $n = 103$ ) and the genomes within CC258 that exhibited extremely limited variability ( $n = 29$ ), such as all but one of those obtained from single outbreaks. The choice of using a relatively large number of non-CC258 genomes ( $n = 13$ ) was made in order to allow the detection of recombination events common to the whole clonal complex. We used this sub-data set of core SNPs in 187 genomes to perform a recombination detection analysis using the software BRATnextgen (29) with 100-iteration analysis, using 100 replicates for statistical significance.

**Analysis of the recombined region.** A database was created collecting protein sequences of factors previously reported to be involved in virulence and antibiotic resistance. We collected sequences from the Comprehensive Antibiotic Resistance Database (CARD) (30) and from the Antibiotic Resistance Genes Database (ARDB) (31), from proteins involved in the biosynthesis of lipopolysaccharides (LPS) and polymyxin resistance, and from the most common virulence factors and siderophores found in Gram-negative bacteria (obtained from the NCBI site). Finally, we added to our manually designed database all *K. pneumoniae* proteins described as potential virulence or resistance factors in the work by Lery and colleagues (32). Gene sequences present in the novel putative recombined region were extracted from the genome of strain NJST258\_1 using an in-house Python script. Correspondence between proteins in our database and genes in the recombined region was tested using a TBLASTN search, selecting genes covering at least 75% of the database sequence with a minimum of 75% identity. Results then were manually checked (see Table S2 in the supplemental material for a complete list).

**Molecular clock.** We created a sub-data set of 174 CC258 genomes and 3 closely related *K. pneumoniae* genomes (used as outgroups), removing genomes of isolates distant from the CC258 clade ( $n = 113$ ) and the genomes within CC258 that exhibited extremely limited variability ( $n = 29$ ), such as all but one of those obtained from single outbreaks. We used the software BEAST (33) on the core SNP alignment of the 177-genome sub-data set after removing SNPs located in the potentially recombined regions. BEAST parameters used were the following: uncorrelated log-normal relaxed clock with the GTR model, with no correction for site rate





**FIG 1** Maximum likelihood phylogeny of *Klebsiella pneumoniae*, based on 319 genomes. The phylogeny was reconstructed starting from an alignment of 94,812 core SNPs, using the software RAxML with a generalized time-reversible (GTR) model and 100 bootstraps, which are not shown for the sake of figure clarity. (A) Circular representation of the phylogeny, obtained using iTOL (itol.embl.de), ignoring branch length. Color circles indicate, from the innermost to the outermost, presence/absence of KPC variants, geographic location in terms of continents, ST based on multilocus sequence typing, and presence in the genome of genes from four beta-lactamase families. The red arrow indicates the origin of the clonal complex 258 clade. (B) Unrooted representation of the phylogeny showing the branch lengths, highlighting the genetic uniformity of clonal complex 258.

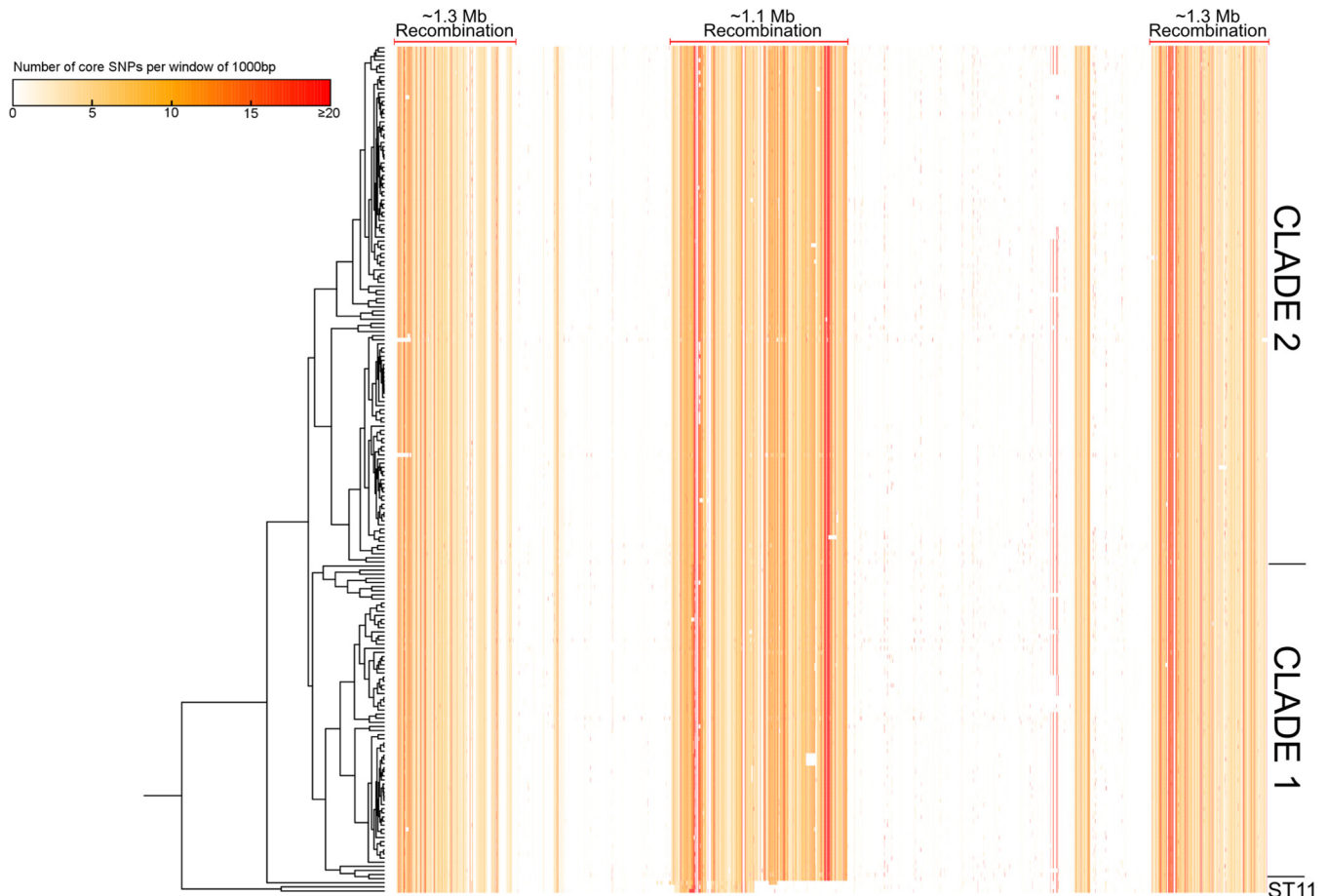
heterogeneity according to analyses performed in similar scenarios (34). The analysis was run for 1,000,000,000 steps, and at every 10,000 steps samples were taken. We discarded 250,000,000 steps as burn-in. The program TRACER (<http://beast.bio.ed.ac.uk/tracer/>) was used to evaluate the convergence of the analysis.

## RESULTS

**Sampling and genome sequencing.** Eighty-nine *K. pneumoniae* strains were collected in six Italian hospitals, chosen based on antibiograms regardless of sequence type, which was determined only afterwards. The data set was composed of 31 KPC producers, 29 ESBL producers, and 29 strains susceptible to carbapenems and third-generation cephalosporins, here referred to as susceptible. The genome of each of the 89 isolates was sequenced and assembled (average genome size, 5,551,959 nt; average N50, 154,414 nt; average coverage, 76.46×). All of the available *K. pneumoniae* genome sequences and reads then were retrieved from the databases ( $n = 230$ ) to create a global data set of 319 *K. pneumoniae* genomes. All genomes in the data set were screened for genes responsible for KPC and beta-lactam resistance phenotypes, as well as for all MLST genes. A total of 55 different MLST profiles were detected, eight of which were novel; thus, they were submitted to the curators of the *K. pneumoniae* MLST database (35). Each of the eight new profiles was represented by a single newly sequenced

Italian isolate (7 susceptible, 1 ESBL producer). Two of these isolates also presented a single novel allele, one for the gene *rpoB* and one for the gene *infB*. See Table S1 in the supplemental material for a list of all of the isolates sequenced in this study and their main characteristics.

**Global SNP phylogeny.** We used a maximum likelihood phylogenomic approach based on core SNPs to elucidate the relationships within the global genome data set comprising the newly sequenced isolates and the *K. pneumoniae* genome sequences available in the database. The presence of antibiotic resistance genes was mapped on the resulting phylogenetic tree, obtained from an alignment of 94,812 core SNPs (Fig. 1). This revealed that 97% of all KPC *K. pneumoniae* strains sequenced to date, regardless of the location of isolation, belong to a well-supported clade, corresponding to the complex CC258. On the other hand, the phylogenomic analysis showed that the isolates encoding common beta-lactam resistance genes (*bla<sub>SHV</sub>* family, *bla<sub>TEM</sub>* family, *bla<sub>OXA</sub>* family, and *bla<sub>CTX-M</sub>* family) are widespread along the tree and belong to various STs (both inside and outside CC258), with no sign of clustering. In fact, the 141 isolates encoding *bla<sub>TEM</sub>* belong to 24 different STs, the 26 isolates encoding *bla<sub>OXA</sub>* belong to 11 different STs, and the 37 isolates encoding *bla<sub>CTX-M</sub>* belong to 16 different STs.



**FIG 2** Uneven clustering of core SNPs in the clonal complex 258 clade. The phylogenetic reconstruction of the 206 representatives of the clonal complex 258 clade is shown on the left, while the core SNP frequency is shown on the right in shades of red, representing the number of core SNPs per 1,000-bp window for each genome. Detected recombinations are indicated at the top of the figure, and main clades of the clonal complex are indicated on the right side of the figure.

**Phylogeny excluding potentially recombined regions.** In a recent work by Castillo-Ramirez and coworkers (34), high-density SNPs clusters with a low ratio of nonsynonymous to synonymous evolutionary changes ( $dN/dS$ ) in closely related bacterial genomes were suggested to be indicators of recombination events. Thus, we evaluated the distribution of SNPs on the genome data set, detecting a highly uneven distribution in the genomes of CC258 isolates, as most core SNPs clustered into two main regions. The first region is located between positions 1,675,550 and 2,740,033, while the second comprises the origin of replication and spans from 4,554,906 to 629,621 in strain NJST258\_1 (Fig. 2) (for the distribution of core SNPs on the whole data set of 319 genomes, see Fig. S1 in the supplemental material). To further analyze the possible presence of recombination events in CC258, we used the software BRATnextgen (29), specifically intended for this purpose, on a reduced data set of 187 genomes of CC258 and closely related strains. This analysis (see Fig. S2) confirmed the presence of the two main recombination events, additionally indicating in what position of the phylogeny they could have occurred. The first event was placed between the entire CC258 clade and the non-KPC external isolates of different STs, while the second was between the outermost strains of ST11 and the inner CC258 clade. Details on these recombined regions are presented in the following paragraph.

We removed the two putative recombined regions from the core SNPs data set of 319 *K. pneumoniae* genomes and performed a phylogenetic analysis on the remaining 55,368 core SNPs. The resulting tree (see Fig. S3 in the supplemental material) is largely consistent with the one generated from the initial data set, confirming the widespread distribution of susceptible and ESBL isolates and the presence of the highly supported KPC CC258 clade. Indeed, both the analysis on all core SNPs and the one performed by removing recombining sites agree in clustering 97% of all KPC *K. pneumoniae* isolates sequenced in a well-supported clade (Fig. 1; also see Fig. S3). This monophyletic clade comprises 203 strains from Asia, Europe, Oceania, and North and South America, with isolation dates ranging from 2002 to 2013; 193 of these (95%) present the *bla*<sub>KPC</sub> gene. Most isolates of this clade belong to ST258 ( $n = 167$ ), but 4 other sequence types are present (i.e., ST11, SST379, ST418, and ST512), all single-nucleotide variants of ST258; thus, they belong to CC258. The second most common sequence type in the CC258 clade is 512, represented by 28 isolates that form a single monophyletic subgroup, located within the ST258 diversity. Interestingly, 24 of these 28 have been isolated in Italy, mostly in this study ( $n = 19$ ) but also in previous works (18, 36). Within the CC258 clade, two main highly supported distinct subclades are detectable, comprising the vast majority of the genomes. Three additional CC258 genomes are located in the tree as

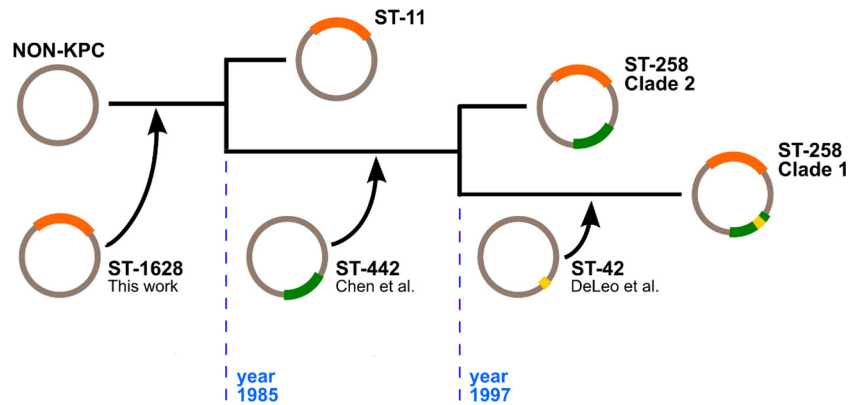


FIG 3 Hypothesis of recombinations occurring in the clonal complex 258 clade. Schematic representation based on the results of the analyses presented. Main nodes of interest are shown, highlighting the hypothesized pattern of three recombination events leading to the current state of clonal complex 258. Dates are inferred based on the molecular clock analysis depicted in Fig. 4.

sister groups of the two main clades, and all are representatives of ST11, again a single-nucleotide variation of ST258. The existence of the two main CC258 subclades was reported previously, and a single recombination event was proposed to be the cause of the differentiation between the two (11), while a subsequent work suggested multiple recombination events (37).

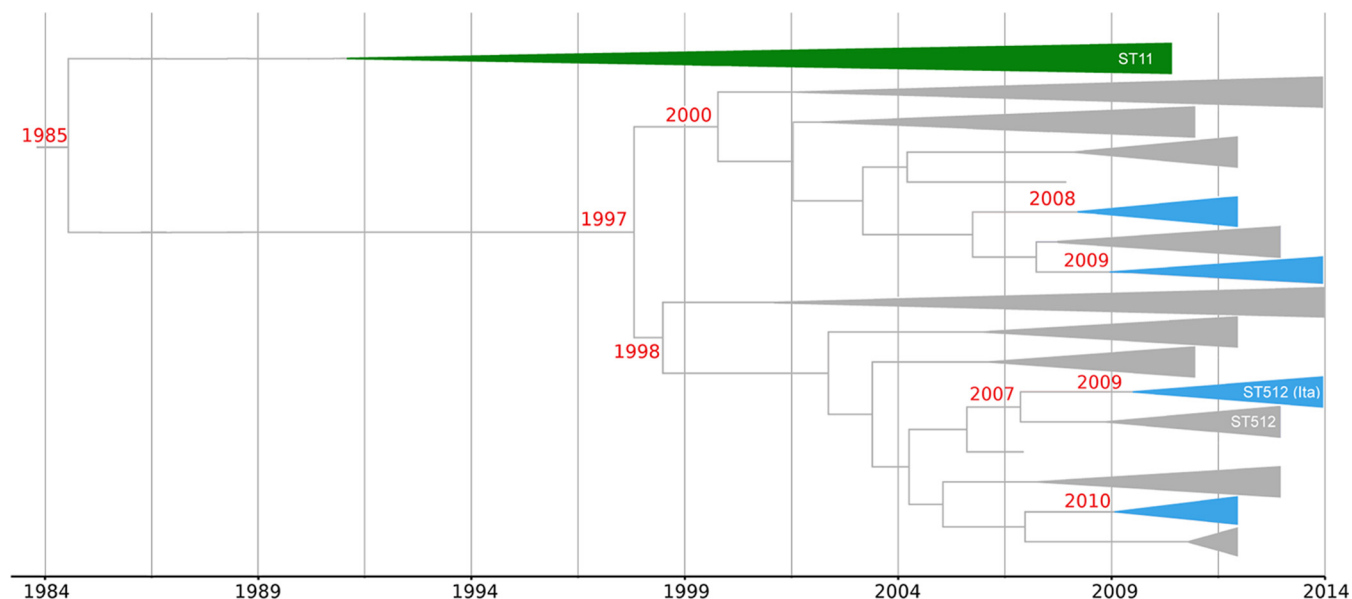
**Analysis of recombinant regions.** As described above, the SNP clustering analysis detected high SNP concentrations in two large genomic regions (Fig. 2). The smaller of the two is highly congruent with the ~1.1-Mb recombination found by Chen and colleagues (12), which represents the major evolutionary change between the members of ST11 and those in the 2 main subclades of the CC258 clade. Chen and colleagues found this region to be most similar to the corresponding region of isolate Kp13 of ST442 and suggested a recombination event, with the donor strain being a close relative of Kp13. Thus, we investigated whether a recombination event is at the origin of the second, newly detected, highly mutated genomic region, located from positions 4,554,906 to 629,621. We performed a phylogenetic analysis, including all of the 319 *K. pneumoniae* genomes examined in this work, on the core SNPs located in this region and in parallel on the core SNPs located in the ~1.1-Mb region. The phylogenetic analysis of the novel ~1.3-Mb region (see Fig. S4 in the supplemental material) confirms the recombination hypothesis, as the topology of the resulting tree clearly shows that Italian isolate 67BO, of the newly described ST1628, is the sister taxon to the entire CC258 clade, suggesting that the donor was related to this isolate. The phylogenetic tree obtained from the ~1.1-Mb recombinant region (see Fig. S5) confirms the published results, clustering the donor Kp13 as a sister taxon of the CC258 clade, with the exclusion of the outermost ST11 isolates. Thus, we propose an updated scenario in which a first recombination event gave origin to the first CC258 strains (represented by ST11), a second recombination subsequently originated ST258, and a third, smaller recombination initiated the split between the two main ST258 subclades (Fig. 3).

In order to investigate the potential effect of the newly discovered recombination on the phenotype of the acceptor CC258, the presence of genes possibly related to antibiotic resistance and virulence was investigated in the corresponding region of the genome of strain NJST258\_1, using a specifically designed database

(see Materials and Methods). Interestingly, 51 genes were detected in the region (see Table S2 in the supplemental material), grouped in three main categories: LPS modification (such as the *waa* operon), bacterial efflux transporters (i.e., efflux pumps and permeases), and regulators (e.g., *ompR-envZ* operon) (see Discussion for an analysis of the detected genes).

**Molecular clock.** In order to date the origin of the CC258 clade and its subclades, we performed a molecular clock analysis using the software BEAST (33). We produced a reduced data set of 3,615 core SNPs present in a selected subset of taxa (174 CC258 and three closely related non-KPC *K. pneumoniae* genomes used as outgroups), derived from the previously filtered data set, in which the potentially recombinant regions of the genome were excluded (Fig. 4). Compared with the dates indicated in published reports, our estimations appear to be fairly accurate. For example, the molecular clock analysis dates the appearance of ST512 to 2007, close to the first report in Israel, i.e., 2006 (20). Additionally, the molecular clock analysis dates the radiation of American and European ST258 isolates to 1997, a time point coherent with the first report of KPC-bearing *K. pneumoniae*, i.e., 1996 (2). Thus, our calibration of the evolutionary rate, superimposed on the phylogenetic tree (Fig. 4), could be used to infer unavailable dates on the global pandemic of CC258 *K. pneumoniae*. See Discussion for further discussion of the estimated dates.

**Italian strains.** The structure of the phylogenomic tree allows us to depict the scenario of the epidemiology of *K. pneumoniae* in Italy (Fig. 1 and 4). While susceptible and ESBL Italian strains are homogeneously distributed on the tree and belong to a number of different STs (24 and 15, respectively), all of the KPC strains sequenced in Italy belong to CC258, indicating a strong epidemiological prevalence of this clonal complex in the Italian hospitals. Within CC258, Italian isolates are well clustered in four monophyla, three composed mostly of isolates sequenced in this study and one encompassing two isolates from a previous study (38). Of the four Italian CC258 monophyla, the one including the most isolates is composed solely of ST512 ( $n = 24$ ), confirming the multiple reports that indicate this ST as being of great epidemiological importance, at least in this country. Our phylogenetic analysis clearly indicates that this ST512 monophylum is found within the diversity of ST258.



**FIG 4** Estimation of divergence times in clonal complex 258. A schematic version of the time-scaled phylogeny was obtained using BEAST software with an uncorrelated log-normal relaxed clock and GTR model with no correction for site rate heterogeneity. The analysis was run for 1,000,000,000 steps, with sampling every 10,000 steps and 25% burn-in. The Italian monophyla are highlighted in blue, while the sequence type 11 (ST11) Asian clade is highlighted in green. All of the phyla with no indication of ST are comprised mainly of isolates of ST258. The dates indicated in the figure, for selected branches and nodes, were inferred from the analysis described above; for a comparison with the dates of isolation of strains, see Discussion.

## DISCUSSION

***Klebsiella pneumoniae* in Italy.** We sequenced the genomes of 89 *K. pneumoniae* strains isolated in Italy, among them 31 KPC producers, 29 ESBL producers, and 29 strains susceptible to beta-lactams and carbapenems. Based on our phylogenomic analysis, the 29 genomes from susceptible *K. pneumoniae* strains isolated in Italy are scattered along the tree, showing no evident sign of clusterization. The sequencing of these isolates allowed us to expand the known diversity of the *K. pneumoniae* species, detecting seven novel MLST profiles and contributing to the overall robustness of current and future phylogenetic analyses. The genomes obtained from 29 ESBL isolates also show a considerable diversity, as they are distributed on the phylogenetic tree and belong to 15 different STs, among them a newly found ST.

Regarding KPC isolates, all Italian sequenced strains are found in CC258. Since no *a priori* selection of STs was performed, this result indicates a strong prevalence of CC258 among KPC *K. pneumoniae* isolates in Italy, even though isolates from different STs have been reported previously by nongenomic studies (e.g., reference 19), and a wider genomic sampling surely would allow us to obtain genomes of KPC isolates belonging to other STs. The genomes of KPC-producing *K. pneumoniae* strains isolated in Italy cluster in four monophyletic groups. If we consider that the first reported case of KPC in Italy occurred in 2008, we can use the dates obtained from the molecular clock to conclude that these monophyletic groups represent four different entrances of KPC *K. pneumoniae* in Italy (Fig. 4). This indicates that KPC strains can move effectively among different countries and continents, and that the current Italian scenario of widespread KPC resistance has been caused by multiple overlapping outbreaks. Additional sampling from Italian CC258 isolates could either confirm these results or detect novel monophyla, possibly discovering additional entrance events.

Among the four Italian CC258 monophyla, one is composed entirely of isolates of ST512. This KPC sequence type was first reported in Israel in 2006 (20) but has been spreading since then, mostly in Italy and South America (11, 17). In accordance with these reports, the four available ST512 genomes from South American isolates cluster in our phylogeny as a sister group of the Italian ST512 clade (Fig. 1 and 4). The molecular clock analysis dates the common ancestor of all members of ST512 to 2007, in relative agreement with the first report of this ST, i.e., 2006 (20). Considering that this ST is known to be a single-nucleotide variant of ST258, these results indicate that a mutational event occurred around 2006, giving rise to this sequence type, that then spread to Israel, South America, and Italy. Genome sequencing of isolates of this ST from Israel, currently unavailable, could allow us to perform phylogenetic analyses aimed at better understanding the geographical and temporal origin of the ST512 clade.

**Origin of the CC258 clade.** Our phylogenomic analysis, coupled with the detection of recombination events and with the molecular clock analysis, allow us to update the hypothesis regarding the origin and evolution of CC258, the most widespread bearer of KPC resistance worldwide (Fig. 3). We postulate a first recombination event that occurred before 1985 between a donor similar to ST1628 and a receiver, an ancestor of ST11. This event, which transferred a region of ~1.3 Mb to the current ST11, gave rise to the basal lineage of CC258. Since only three genomes of ST11 currently are available, all isolated from Asian patients, the current phylogeny suggests that this first recombination event occurred on the Asian continent. However, additional genome sequences of ST11 from different geographic locations are necessary to support or falsify this hypothesis. Our molecular clock analysis also can be useful to date the two subsequent, previously reported (11, 12) recombination events. The second recombination event, confirmed by our phylogenies, gave rise to ST258, having as a recipi-

ent ST11 and a donor similar to ST442 (12). Our molecular clock analysis dates this event to between 1985 and 1997. Considering that all of the known genomic CC258 diversity from the American and European continents is included within the subclade that originated in 1997 (Fig. 4), this second event could have been pivotal in the subsequent pandemic of KPC-bearing CC258. Finally, we can date the third smaller recombination event, the one that gave origin to the differentiation between the two main CC258 subclades (11), to between 1999 and 2001. Thus, we can hypothesize that these three events have produced a genomic background apt to bear and diffuse KPC plasmids, contributing to the success of the KPC pandemic.

The proposed scenario suggests that the genomic diversity of the whole *K. pneumoniae* species constitutes a reservoir of genetic variability capable of recombination events of large portions of the genome, with subsequent generation of novel variants. In this scenario, we hypothesize that large genomic recombinations are at the basis of important phenotypic/functional changes that, together with the acquisition and diffusion of plasmids bearing antibiotic resistance genes, have led to the current global epidemic. This hypothesis is supported by the multiple detected recombination events, as well as by the limited number of SNPs identified outside the recombined regions (a total of 1,086 core SNPs in the 206 analyzed CC258 genomes), and finally by the current impossibility to phenotypically differentiate the isolates of subclade ST512 from those of ST258. An alternative hypothesis is that the main reason for the diffusion of CC258 is simply the acquisition of the resistance to carbapenemic antibiotics, and that the genomic variations, whether they are recombinations or point mutations, do not provide a specific fitness benefit but are merely an example of genetic hitchhiking.

In order to investigate the importance of the recombination event described in this work, the gene content of the ~1.3-Mb region was analyzed. Fifty-one genes in this genomic context were found to be potentially related to virulence or antibiotic resistance (see Table S2 in the supplemental material). The presence of LPS synthesis genes is worth a mention because of the multiple linkages between the outer membrane and virulence (39). Genes of the operon *waa* (also known as *rfa*) are responsible for the biogenesis of the core LPS, while genes of the family *arn* control the modifications of lipid A. Modifications in membrane composition can lead to changes in surface charge and interfere with the activity of antibiotics that act on LPS, such as polymyxins and novobiocin (40). Moreover, the presence of *mia* genes in the recombined region is worth being highlighted. These genes are presumed to maintain lipid asymmetry in the Gram-negative outer membrane, as they transport phospholipids to the inner side of the membrane. *mia* genes were reported as virulence factors in *Escherichia coli* and in other Gram-negative bacteria, as mutations in these genes can lead to a change in the permeability of the outer membrane and to a subsequent variation in virulence (41). The presence of fumarate reductase genes of the family *fmr* in the recombined region suggests a link with the variation of virulence of CC258. In fact, fumarate reductase is a virulence determinant in *Helicobacter pylori*, *Mycobacterium tuberculosis*, *Actinobacillus pleuropneumoniae*, and *Salmonella enterica*, as mutants of these genes show variations in virulence (32). Finally, the *ompR-envZ* operon, present in the recombined region, is a two-component system that acts as a transcription regulator, affecting the expression of the genes *ompF* and *ompC* (42). Mutations in the *ompR* and

*envZ* genes have been shown to reduce the expression of outer membrane porins *OmpF* and *OmpC* (43). This in turn can have drastic effects on both the virulence and antibiotic resistance of mutant strains. It has been reported in particular that *OmpR* mutations can lead to reduced susceptibility to carbapenemic antibiotics in *Enterobacteriaceae* (44).

Further functional investigations aimed at unveiling the reasons for the success of the CC258 clade, possibly focusing on the detected recombinant regions, would greatly improve our understanding of the *K. pneumoniae* pandemic and would provide important tools in the fight against KPC-producing strains. Finally, our conclusions should lead to additional studies focused on the recombination potential of other STs of *K. pneumoniae*. If this capacity were found to be widespread, we should be aware that future recombination events could lead to the diffusion of novel epidemic clones.

## ACKNOWLEDGMENTS

This work was supported by Ricerca Corrente 2013 funding from Fondazione IRCCS Policlinico S. Matteo to P.M.


We thank Simone Ambretti for providing samples and Rosa Visiello for her assistance in correcting the manuscript.

## REFERENCES

1. Knothe H, Shah P, Krcmery V, Antal M, Mitsuhashi S. 1983. Transferable resistance to cefotaxime, cefoxitin, cefamandole and cefuroxime in clinical isolates of *Klebsiella pneumoniae* and *Serratia marcescens*. *Infection* 11:315–317. <http://dx.doi.org/10.1007/BF01641355>.
2. Yigit HA, Queenan M, Anderson GJ, Domenech-Sanchez A, Biddle JW, Steward CD, Alberti S, Bush K, Tenover FC. 2001. Novel carbapenem-hydrolyzing beta-lactamase, KPC-1, from a carbapenem-resistant strain of *Klebsiella pneumoniae*. *Antimicrob Agents Chemother* 45:1151–1161. <http://dx.doi.org/10.1128/AAC.45.4.1151-1161.2001>.
3. Woodford N, Tierno PM, Young K, Tysall R, Palepou MF, Ward E, Painter RE, Suber DF, Shungu D, Silver LL, Inglima K, Kornblum J, Livermore DM. 2004. Outbreak of *Klebsiella pneumoniae* producing a new carbapenem-hydrolyzing class A beta-lactamase, KPC-3, in a New York medical center. *Antimicrob Agents Chemother* 48:4793–4799. <http://dx.doi.org/10.1128/AAC.48.12.4793-4799.2004>.
4. Bratu S, Landman D, Haag R, Recco R, Eramo A, Alam M, Quale J. 2005. Rapid spread of carbapenem-resistant *Klebsiella pneumoniae* in New York City: a new threat to our antibiotic armamentarium. *Arch Intern Med* 165:1430–1435. <http://dx.doi.org/10.1001/archinte.165.12.1430>.
5. Nordmann P, Cuzon G, Naas T. 2009. The real threat of *Klebsiella pneumoniae* carbapenemase-producing bacteria. *Lancet Infect Dis* 9:228–236. [http://dx.doi.org/10.1016/S1473-3099\(09\)70054-4](http://dx.doi.org/10.1016/S1473-3099(09)70054-4).
6. Cantón R, Akóva M, Carmeli Y, Giske CG, Glupczynski Y, Gniadkowski M, Livermore DM, Miriagou V, Naas T, Rossolini GM, Samuelsen Ø, Seifert H, Woodford N, Nordmann P. 2012. Rapid evolution and spread of carbapenemases among *Enterobacteriaceae* in Europe. *Clin Microbiol Infect* 18:413–431. <http://dx.doi.org/10.1111/j.1469-0691.2012.03821.x>.
7. Richter SN, Frasson I, Bergo C, Parisi S, Cavallaro A, Palù G. 2011. Transfer of KPC-2 carbapenemase from *Klebsiella pneumoniae* to *Escherichia coli* in a patient: first case in Europe. *J Clin Microbiol* 49:2040–2042. <http://dx.doi.org/10.1128/JCM.00133-11>.
8. Luo Y, Yang J, Ye L, Guo L, Zhao Q, Chen R, Chen Y, Han X, Zhao J, Tian S, Han L. 2014. Characterization of KPC-2-producing *Escherichia coli*, *Citrobacter freundii*, *Enterobacter cloacae*, *Enterobacter aerogenes*, and *Klebsiella oxytoca* isolates from a Chinese hospital. *Microb Drug Resist* 4:264–269. <http://dx.doi.org/10.1089/mdr.2013.0150>.
9. Chen L, Chavda KD, Melano RG, Hong T, Rojzman AD, Jacobs MR, Bonomo RA, Kreiswirth BN. 2014. A molecular survey of the dissemination of two blaKPC-harboring IncFIA plasmids in New Jersey and New York hospitals. *Antimicrob Agents Chemother* 58:2289–2294. <http://dx.doi.org/10.1128/AAC.02749-13>.
10. Andrade LN, Curiao T, Ferreira JC, Longo JM, Clímaco EC, Martinez R, Bellissimo-Rodrigues F, Basile-Filho A, Evaristo MA, Del Peloso PF,

- Ribeiro VB, Barth AL, Paula MC, Baquero F, Cantón R, Darini AL, Coque TM. 2011. Dissemination of blaKPC-2 by the spread of *Klebsiella pneumoniae* clonal complex 258 clones (ST258, ST11, ST437) and plasmids (IncFII, IncN, IncL/M) among *Enterobacteriaceae* species in Brazil. *Antimicrob Agents Chemother* 55:3579–3583. <http://dx.doi.org/10.1128/AAC.01783-10>.
11. Deleo FR, Chen L, Porcella SF, Martens CA, Kobayashi SD, Porter AR, Chavda KD, Jacobs MR, Mathema B, Olsen RJ, Bonomo RA, Musser JM, Kreiswirth BN. 2014. Molecular dissection of the evolution of carbapenem-resistant multilocus sequence type 258 *Klebsiella pneumoniae*. *Proc Natl Acad Sci U S A* 111:4988–4993. <http://dx.doi.org/10.1073/pnas.1321364111>.
  12. Chen L, Mathema B, Pitout JD, DeLeo FR, Kreiswirth BN. 2014. Epidemic *Klebsiella pneumoniae* ST258 is a hybrid strain. *mBio* 5:e01355-14. <http://dx.doi.org/10.1128/mBio.01355-14>.
  13. Pagani L, Ronza P, Giacobone E, Romero E. 1994. Extended-spectrum beta-lactamases from *Klebsiella pneumoniae* strains isolated at an Italian hospital. *Eur J Epidemiol* 10:533–540. <http://dx.doi.org/10.1007/BF01719569>.
  14. Perilli M, Dell'Amico E, Segatore B, de Massis MR, Bianchi C, Luzzaro F, Rossolini GM, Toniolo A, Nicoletti G, Amicosante G. 2002. Molecular characterization of extended-spectrum beta-lactamases produced by nosocomial isolates of *Enterobacteriaceae* from an Italian nationwide survey. *J Clin Microbiol* 40:611–614. <http://dx.doi.org/10.1128/JCM.40.2.611-614.2002>.
  15. D'Andrea MM, Arena F, Pallecchi L, Rossolini GM. 2013. CTX-M-type  $\beta$ -lactamases: a successful story of antibiotic resistance. *Int J Med Microbiol* 303:305–317. <http://dx.doi.org/10.1016/j.ijmm.2013.02.008>.
  16. Giani T, D'Andrea MM, Pecile P, Borgianni L, Nicoletti P, Tonelli F, Bartoloni A, Rossolini GM. 2009. Emergence in Italy of *Klebsiella pneumoniae* sequence type 258 producing KPC-3 carbapenemase. *J Clin Microbiol* 47:3793–3794. <http://dx.doi.org/10.1128/JCM.01773-09>.
  17. Gaibani P, Ambretti S, Berlinger A, Gelsomino F, Bielli A, Landini MP, Sambri V. 2011. Rapid increase of carbapenemase-producing *Klebsiella pneumoniae* strains in a large Italian hospital: surveillance period 1 March–30 September 2010. *Euro Surveill* 16:19800.
  18. Comandatore F, Gaibani P, Ambretti S, Landini MP, Daffonchio D, Marone P, Sambri V, Bandi C, Sasser D. 2013. Draft genome of *Klebsiella pneumoniae* sequence type 512, a multidrug-resistant strain isolated during a recent KPC outbreak in Italy. *Genome Announc* 1:e00035-12. <http://dx.doi.org/10.1128/genomeA.00035-12>.
  19. Giani T, Pini B, Arena F, Conte V, Bracco S, Migliavacca R, Pantosti A, Pagani L, Luzzaro F, Rossolini GM. 2013. Epidemic diffusion of KPC carbapenemase-producing *Klebsiella pneumoniae* in Italy: results of the first countrywide survey, 15 May to 30 June 2011. *Euro Surveill* 18:20489.
  20. Warburg G, Hidalgo-Grass C, Partridge SR, Tolmasky ME, Temper V, Moses AE, Block C, Strahilevitz J. 2012. A carbapenem-resistant *Klebsiella pneumoniae* epidemic clone in Jerusalem: sequence type 512 carrying a plasmid encoding aac(6')-Ib. *J Antimicrob Chemother* 67:898–901. <http://dx.doi.org/10.1093/jac/dkr552>.
  21. Doyle D, Peirano G, Lascols C, Lloyd T, Church DL, Pitout JD. 2012. Laboratory detection of *Enterobacteriaceae* that produce carbapenemases. *J Clin Microbiol* 50:3877–3880. <http://dx.doi.org/10.1128/JCM.02117-12>.
  22. Poirel L, Walsh TR, Cuvillier V, Nordmann P. 2011. Multiplex PCR for detection of acquired carbapenemase genes. *Diagn Microbiol Infect Dis* 70:119–123. <http://dx.doi.org/10.1016/j.diagmicrobio.2010.12.002>.
  23. Clinical and Laboratory Standards Institute. 2011. Performance standards for antimicrobial susceptibility testing; 21st informational supplement. CLSI M100-S121, vol 31. Clinical and Laboratory Standards Institute, Wayne, PA.
  24. European Committee on Antimicrobial Susceptibility Testing. 2014. Breakpoint tables for interpretation of MICs and zone diameters. Version 4.0. [http://www.eucast.org/fileadmin/src/media/PDFs/EUCAST\\_files/Breakpoint\\_tables/Breakpoint\\_table\\_v\\_4.0.pdf](http://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/Breakpoint_tables/Breakpoint_table_v_4.0.pdf).
  25. Chevreaux B, Wetter T, Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information, p 45–56. Proceedings of the 1999 German Conference on Bioinformatics.
  26. Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18:821–829. <http://dx.doi.org/10.1101/gr.074492.107>.
  27. Darling AE, Mau B, Perna NT. 2010. Progressivemauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5:e11147. <http://dx.doi.org/10.1371/journal.pone.0011147>.
  28. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313. <http://dx.doi.org/10.1093/bioinformatics/btu033>.
  29. Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, Corander J. 2012. Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res* 40:e6. <http://dx.doi.org/10.1093/nar/gkr928>.
  30. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar K, Canova MJ, De Pascuale G, Ejim L, Kalan L, King AM, Koteva K, Morar M, Mulvey MR, O'Brien JS, Pawlowski AC, Piddock LJ, Spanogiannopoulos P, Sutherland AD, Tang I, Taylor PL, Thaker M, Wang W, Yan M, Yu T, Wright GD. 2013. The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother* 57:3348–3357. <http://dx.doi.org/10.1128/AAC.00419-13>.
  31. Liu B, Pop M. 2009. ARDB-Antibiotic Resistance Genes Database. *Nucleic Acids Res* 37:D443–D447. <http://dx.doi.org/10.1093/nar/gkn656>.
  32. Lery LM, Frangeul L, Tomas A, Passet V, Almeida AS, Bialek-Davenet S, Barbe V, Bengochea JA, Sansonetti P, Brisse S, Tournèze R. 2014. Comparative analysis of *Klebsiella pneumoniae* genomes identifies a phospholipase D family protein as a novel virulence factor. *BMC Biol* 12:41. <http://dx.doi.org/10.1186/1741-7007-12-41>.
  33. Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7:214. <http://dx.doi.org/10.1186/1471-2148-7-214>.
  34. Castillo-Ramirez S, Harris SR, Holden MTG, He M, Parkhill J, Bentley SD, Feil EJ. 2011. The impact of recombination on dN/dS within recently emerged bacterial clones. *PLoS Pathog* 7:e1002129. <http://dx.doi.org/10.1371/journal.ppat.1002129>.
  35. Diancourt L, Passet V, Verhoef J, Grimont PAD, Brisse S. 2005. Multilocus sequence typing of *Klebsiella pneumoniae* nosocomial isolates. *J Clin Microbiol* 43:4178–4182. <http://dx.doi.org/10.1128/JCM.43.8.4178-4182.2005>.
  36. Villa L, Feudi C, Fortini D, García-Fernández A, Carattoli A. 2014. Genomics of KPC-producing *Klebsiella pneumoniae* sequence type 512 clone highlights the role of RamR and ribosomal S10 protein mutations in conferring tigecycline resistance. *Antimicrob Agents Chemother* 58:1707–1712. <http://dx.doi.org/10.1128/AAC.01803-13>.
  37. Wright MS, Perez F, Brinkac L, Jacobs MR, Kaye K, Cober E, van Duin D, Marshall SH, Hujer AM, Rudin SD, Hujer KM, Bonomo RA, Adams MD. 2014. Population structure of KPC-producing *Klebsiella pneumoniae* isolates from midwestern U.S. hospitals. *Antimicrob Agents Chemother* 58:4961–4965. <http://dx.doi.org/10.1128/AAC.00125-14>.
  38. Comandatore F, Sasser D, Ambretti S, Landini MP, Daffonchio D, Marone P, Sambri V, Bandi C, Gaibani P. 2013. Draft genome sequences of two multidrug resistant *Klebsiella pneumoniae* ST258 isolates resistant to colistin. *Genome Announc* 1:e00113–12. <http://dx.doi.org/10.1128/genomeA.00113-12>.
  39. Heinrichs DE, Yethon JA, Whitfield C. 1998. Molecular basis for structural diversity in the core regions of the lipopolysaccharides of *Escherichia coli* and *Salmonella enterica*. *Mol Microbiol* 30:221–232. <http://dx.doi.org/10.1046/j.1365-2958.1998.01063.x>.
  40. Goldberg JB. 1999. Genetics of bacterial polysaccharides. CRC Press, London, United Kingdom.
  41. Malinverni JC, Silhavy TJ. 2009. An ABC transport system that maintains lipid asymmetry in the gram-negative outer membrane. *Proc Natl Acad Sci U S A* 12:8009–8014. <http://dx.doi.org/10.1073/pnas.0903229106>.
  42. Buckler DR, Anand GS, Stock AM. 2000. Response-regulator phosphorylation and activation: a two-way street? *Trends Microbiol* 8:153–156. [http://dx.doi.org/10.1016/S0966-842X\(00\)01707-8](http://dx.doi.org/10.1016/S0966-842X(00)01707-8).
  43. Yuan J, Wei B, Shi M, Gao H. 2011. Functional assessment of EnvZ/OmpR two-component system in *Shewanella oneidensis*. *PLoS One* 6:e23701. <http://dx.doi.org/10.1371/journal.pone.0023701>.
  44. Tängdén T, Adler M, Cars O, Sandegren L, Löwdin E. 2013. Frequent emergence of porin-deficient subpopulations with reduced carbapenem susceptibility in ESBL-producing *Escherichia coli* during exposure to ertapenem in an in vitro pharmacokinetic model. *J Antimicrob Chemother* 68:1319–1326. <http://dx.doi.org/10.1093/jac/dkt044>.

# Differential Single Nucleotide Polymorphism-Based Analysis of an Outbreak Caused by *Salmonella enterica* Serovar Manhattan Reveals Epidemiological Details Missed by Standard Pulsed-Field Gel Electrophoresis

Erika Scaltriti,<sup>a</sup> Davide Sasseria,<sup>b</sup> Francesco Comandatore,<sup>b,c</sup> Marina Morganti,<sup>a</sup> Carmen Mandalari,<sup>a</sup> Stefano Gaiarsa,<sup>c,d</sup> Claudio Bandi,<sup>c</sup> Gianguglielmo Zehender,<sup>e</sup> Luca Bolzoni,<sup>f</sup> Gabriele Casadei,<sup>a</sup>  Stefano Pongolini<sup>a,f</sup>

Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna (IZSLER), Sezione di Parma, Parma, Italy<sup>a</sup>; Dipartimento di Biologia e Biotecnologie, Università di Pavia, Pavia, Italy<sup>b</sup>; Dipartimento di Scienze Veterinarie e Sanità Pubblica (DIVET), Università degli Studi di Milano, Milan, Italy<sup>c</sup>; Fondazione IRCCS Policlinico San Matteo, Pavia, Italy<sup>d</sup>; Dipartimento di Scienze Cliniche L. Sacco, Università degli Studi di Milano, Milan, Italy<sup>e</sup>; Direzione Sanitaria – Servizio di Analisi del Rischio, Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna (IZSLER), Parma, Italy<sup>f</sup>

**We retrospectively analyzed a rare *Salmonella enterica* serovar Manhattan outbreak that occurred in Italy in 2009 to evaluate the potential of new genomic tools based on differential single nucleotide polymorphism (SNP) analysis in comparison with the gold standard genotyping method, pulsed-field gel electrophoresis. A total of 39 isolates were analyzed from patients ( $n = 15$ ) and food, feed, animal, and environmental sources ( $n = 24$ ), resulting in five different pulsed-field gel electrophoresis (PFGE) profiles. Isolates epidemiologically related to the outbreak clustered within the same pulsotype, SXB\_BS.0003, without any further differentiation. Thirty-three isolates were considered for genomic analysis based on different sets of SNPs, core, synonymous, nonsynonymous, as well as SNPs in different codon positions, by Bayesian and maximum likelihood algorithms. Trees generated from core and nonsynonymous SNPs, as well as SNPs at the second and first plus second codon positions detailed four distinct groups of isolates within the outbreak pulsotype, discriminating outbreak-related isolates of human and food origins. Conversely, the trees derived from synonymous and third-codon-position SNPs clustered food and human isolates together, indicating that all outbreak-related isolates constituted a single clone, which was in line with the epidemiological evidence. Further experiments are in place to extend this approach within our regional enteropathogen surveillance system.**

**S**almonellosis is a major food-borne disease worldwide, with an estimated 93.8 million cases occurring each year, resulting in 155,000 deaths (1). The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks (2) indicated that nontyphoid salmonellosis was the second most reported food-borne zoonosis in Europe in 2012, trailing only behind *Campylobacter jejuni* infection. The 2012 overall notification rate for human salmonellosis in the European Union (EU) was 22.2 episodes per 100,000 population, for a total of 91,034 confirmed cases, with hospitalization and mortality rates of 45.1% and 0.14%, respectively. The highest proportions of *Salmonella*-positive foodstuff samples were reported for fresh turkey, poultry, and pork at 4.4%, 4.1%, and 0.7%, respectively (2). In order to manage this food-borne infection and to limit its health and economic burdens, surveillance programs have developed and implemented DNA-based subtyping methods to identify outbreaks in a timely manner and to trace infections back to their food sources. Over the past decades, the two most intensively used protocols for *Salmonella* subtyping have been pulsed-field gel electrophoresis (PFGE) and multilocus variable-number tandem-repeat analysis (MLVA) (3). Unfortunately, these methods rely on just few features of the entire bacterial genome (rare restriction sites for PFGE or few polymorphic loci for MLVA) to assess the relatedness of different isolates. During epidemiological investigations of food-borne outbreaks, this limitation might lead to difficulties in distinguishing outbreak-related from outbreak-unrelated *Salmonella enterica* subsp. *enterica* isolates due to the high genetic homogeneity of this subspecies (4). Multilocus sequence typing (MLST) is another molecular tool for bacterial typing

based on allelic differences in the loci of specified housekeeping genes (5). While proposed as an alternative to classical serotyping (6), MLST does not seem to be discriminatory enough when all isolates being tested belong to the same serotype (7). With the aim of improving resolution in molecular epidemiology, the technological advancements of whole-genome sequencing (WGS) may provide an unprecedented opportunity to access the entire genome information at a reasonable cost, as well as to set a new series of high-resolution standards in molecular epidemiology. As PFGE and MLVA are able to resolve more genotypes within a single serovar, WGS has already proved its resolution power to detect variations within otherwise undistinguishable bacterial clones (by

Received 13 October 2014 Returned for modification 13 November 2014

Accepted 25 January 2015

Accepted manuscript posted online 4 February 2015

Citation Scaltriti E, Sasseria D, Comandatore F, Morganti M, Mandalari C, Gaiarsa S, Bandi C, Zehender G, Bolzoni L, Casadei G, Pongolini S. 2015. Differential single nucleotide polymorphism-based analysis of an outbreak caused by *Salmonella enterica* serovar Manhattan reveals epidemiological details missed by standard pulsed-field gel electrophoresis. *J Clin Microbiol* 53:1227–1238. doi:10.1128/JCM.02930-14.

Editor: D. J. Diekema

Address correspondence to Stefano Pongolini, stefano.pongolini@izsler.it.

Supplemental material for this article may be found at <http://dx.doi.org/10.1128/JCM.02930-14>.

Copyright © 2015, American Society for Microbiology. All Rights Reserved.

doi:10.1128/JCM.02930-14

PFGE or MLVA), as shown by recent examples in the literature (8, 9). Large studies based on WGS within *S. enterica* subspecies (10) and within serovars in *S. enterica* subsp. *enterica* (11, 12) contributed to the elucidation of *Salmonella* phylogenetic diversity and also accomplished important steps forward in the area of bacterial disease tracking. Moreover, serovar-specific studies on *S. enterica* subsp. *enterica* have highlighted microevolutionary differences among clinical, environmental, and food isolates in *S. enterica* serovars Montevideo (13, 14), Enteritidis (4), Newport (15), Typhimurium (16–18), and Heidelberg (12), which would have been missed by more traditional approaches.

While outbreaks of more common serovars, such as *Salmonella* Typhimurium and *Salmonella* Enteritidis, have been reported and investigated, only a few human outbreaks due to *S. enterica* serovar Manhattan have been reported (19, 20) worldwide in the past 60 years, and none have been characterized at the genomic level. Here, we present a WGS-based retrospective analysis of the only *Salmonella* Manhattan outbreak ever documented in Italy, which occurred from June to July 2009 in a relatively small geographic area in the province of Modena.

The outbreak investigation at the time of the event was carried out by international standard epidemiological techniques (21) and by PFGE on the isolates from patients and food, feed, animal, and environmental sources.

The aim of this study was 2-fold: (i) to evaluate the effectiveness of WGS to accurately identify the relationships among all the outbreak-related isolates with enough resolution to clarify the ambiguities that PFGE was not able to unravel, and (ii) to explore and test new genomic tools for bacterial molecular epidemiology based on synonymous and nonsynonymous single-nucleotide polymorphisms (SNPs) and SNPs in different codon positions.

We selected this specific *Salmonella* Manhattan outbreak to test our WGS pipeline because of three main features that made this outbreak a particularly suitable case study. First, *Salmonella* Manhattan is considered a rare serotype, as confirmed by the regional surveillance system for *Salmonella* of Emilia-Romagna, which over the past 3 years recorded a yearly average of only 5.6 sporadic cases over a total of 924 isolates per year, from a regional population of about 5,000,000 (M. Morganti, E. Scaltriti, L. Bolzoni, G. Casadei, and S. Pongolini; Enter-net Italia, unpublished data). This low prevalence of *Salmonella* Manhattan infection provides a reasonable confidence that virtually all isolates collected in the outbreak area at the time of the episode belonged to the outbreak, therefore preventing the noise effect due to unrelated isolates wrongly assigned to the epidemic. Second, the investigation conducted at the time of the outbreak was successful in tracing the infection back to a food point source using internationally coded epidemiological methods (21); bacterial isolates were also recovered not only from food (pork sausage) at the retail level but also along the food chain up to the raw meat used to prepare the implicated food (at the production establishment). Third, the regional surveillance system for *Salmonella* of Emilia-Romagna, hosted at the Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna (IZSLER), holds a full collection of *Salmonella* Manhattan strains covering the years 2001 to present. This set of isolates was pivotal in the conduct of a successful epidemiological investigation and for testing our WGS-based analyses of this rare serovar.

## CASE REPORT

The diagnostic unit of Parma of IZSLER is the Regional Reference Center for Surveillance of Enteropathogens (Enter-net) of clinical, environmental, animal, and food origins. Within this activity, a cluster of 15 human infections caused by *Salmonella* Manhattan was detected in the province of Modena from June to July 2009. All 15 isolates showed the same PFGE profile, SXB\_BS.0003, strengthening the hypothesis that the unusually high incidence of this rare serovar was due to an epidemic outbreak. Consequently, an epidemiological investigation was undertaken and, considering the rarity of the serovar involved, all 21 isolates of *Salmonella* Manhattan available from the surveillance collection of IZSLER were genotyped by PFGE to get possible clues about the source of the outbreak. Thirteen isolates from the collection had the same PFGE profile as that of the outbreak strain, but only three of them had been isolated just before the onset of the outbreak (May/June 2009). Two had been isolated from pork sausage at the establishment of an industrial producer that distributed in the outbreak area, while one had been recovered from swine intestine at an establishment near the outbreak area that processed guts for the salami industry. According to the epidemiological investigation, the gut processing establishment had no correlation with the outbreak. However, as its isolate presented the same PFGE pulsotype as that of the outbreak-related isolates, health authorities were left with a certain degree of uncertainty about its possible role. Following the results of the epidemiological and molecular analyses, food samples were collected at retail sources in the outbreak area and at the establishment producing the sausage in order to confirm the source and clonality of the outbreak strain. Two samples from retail-collected sausages, along with a sample from fresh pork supplies of the sausage producer, scored positive for the outbreak pulsotype. Based on these results, the sausage from the implicated producer was recalled, leading to the outbreak extinction.

## MATERIALS AND METHODS

**Bacterial isolates.** A total of 39 *Salmonella* Manhattan isolates were included in the study. Fifteen isolates were involved in the epidemic episode, another three isolates were collected within the epidemiological investigation, and 21 were collected between 2001 and 2009 during the surveillance activity of IZSLER (Table 1). The isolates were isolated and streak purified with standard microbiological techniques and stocked at  $-80^{\circ}\text{C}$ . They were cultured on plates with Trypticase soy agar with 5% defibrinated sheep blood (TSA-SB) and incubated overnight at  $37^{\circ}\text{C}$  before being typed by pulsed-field gel electrophoresis, according to the PulseNet protocol (22). The isolates selected for WGS were inoculated into brain heart infusion broth and cultured overnight at  $37^{\circ}\text{C}$  with agitation (200 rpm).

**Pulsed-field gel electrophoresis.** All isolates were genotyped by PFGE, according to the PulseNet protocol (22). Genomic DNA underwent XbaI restriction before electrophoresis in a Chef Mapper XA system (Bio-Rad, CA, USA). The PFGE patterns were analyzed using the BioNumerics Software version 6.6 (Applied-Maths, Sint-Martens-Latem, Belgium) and associated with isolate information in our surveillance database. Clustering of the PFGE profiles was generated using the unweighted-pair group method using averages (UPGMA) based on the Dice similarity index (optimization, 1%; band matching tolerance, 1%). Following a comparison of the electrophoretic profiles, a PFGE pattern (pulsotype) was assigned to each isolate within the Regional Surveillance Database of Emilia-Romagna.

**Whole-genome sequencing.** All outbreak-related isolates and a selection of the IZSLER *Salmonella* Manhattan collection, representative of the different pulsotypes detected, were subjected to WGS (Table 1), for a total



TABLE 1 Complete data set of *Salmonella* Manhattan isolates analyzed in this study<sup>a</sup>

Lab no.	Isolate no. (this study)	Date of isolation (DD/MM/YYYY)	Isolation place (province)	Matrix	PFGE pulsotype
160969_3	SM1 <sup>b</sup>	06/30/2009	Modena	Human	SXB_BS.0003
160969_5	SM2 <sup>b</sup>	06/30/2009	Modena	Human	SXB_BS.0003
160969_6	SM3 <sup>b</sup>	06/30/2009	Modena	Human	SXB_BS.0003
165051_2	SM4 <sup>b</sup>	07/03/2009	Modena	Human	SXB_BS.0003
165051_3	SM5 <sup>b</sup>	07/03/2009	Modena	Human	SXB_BS.0003
165051_5	SM6 <sup>b</sup>	07/03/2009	Modena	Human	SXB_BS.0003
165051_7	SM7 <sup>b</sup>	07/30/2009	Modena	Human	SXB_BS.0003
111113	SM8 <sup>b</sup>	07/03/2009	Modena	Human	SXB_BS.0003
165051_11	SM9 <sup>b</sup>	07/03/2009	Modena	Human	SXB_BS.0003
165051_12	SM10 <sup>b</sup>	07/03/2009	Modena	Human	SXB_BS.0003
180073_1	SM11 <sup>b</sup>	07/22/2009	Modena	Human	SXB_BS.0003
180073_2	SM12 <sup>b</sup>	07/22/2009	Modena	Human	SXB_BS.0003
180073_3	SM13 <sup>b</sup>	07/22/2009	Modena	Human	SXB_BS.0003
180073_4	SM14 <sup>b</sup>	07/22/2009	Modena	Human	SXB_BS.0003
180073_6	SM15 <sup>b</sup>	07/22/2009	Modena	Human	SXB_BS.0003
250920	SM42 <sup>b</sup>	08/31/2009	Milano	Pork	SXB_BS.0003
227021	SM32 <sup>b</sup>	05/06/2009	Milano	Pork sausage	SXB_BS.0003
188801	SM52 <sup>b</sup>	05/06/2009	Milano	Pork sausage	SXB_BS.0003
216630_1	SM53 <sup>b</sup>	09/03/2009	Modena	Pork sausage	SXB_BS.0003
216630_2	SM54 <sup>b</sup>	09/03/2009	Modena	Pork sausage	SXB_BS.0003
<hr/>					
226957	SM16	03/07/2006	Mantova	Swine	SXB_PR.0753
226963	SM17 <sup>b</sup>	03/20/2006	Mantova	Swine	SXB_PR.0753
226972	SM19 <sup>b</sup>	03/20/2006	Sondrio	Pork salami	SXB_PR.0753
226979_1	SM21 <sup>b</sup>	07/31/2006	Cremona	Swine gut	SXB_BS.0003
226985	SM23 <sup>b</sup>	08/03/2006	Milano	Pork sausage	SXB_BS.0003
226987	SM24 <sup>b</sup>	08/03/2006	Milano	Pork sausage	SXB_BS.0003
226993	SM26	01/22/2007	Ravenna	Hamburger	SXB_BS.0003
226998	SM27 <sup>b</sup>	06/29/2007	Milano	Pork	SXB_BS.0003
227002	SM28	09/18/2002	Pavia	Surface water	SXB_BS.0003
227009	SM29 <sup>b</sup>	09/02/2002	Bologna	Bovine sausage	SXB_PR.0754
227015	SM31	09/11/2001	Pavia	Surface water	SXB_PR.0751
227033	SM35 <sup>b</sup>	11/29/2008	Ravenna	Swine stool	SXB_BS.0003
227039	SM36 <sup>b</sup>	09/30/2008	Brescia	Swine stool	SXB_PR.0752
227052	SM38 <sup>b</sup>	09/24/2008	Milano	Swine stool	SXB_BS.0003
188806	SM48 <sup>b</sup>	06/03/2009	Reggio Emilia	Swine intestine	SXB_BS.0003
188790	SM47	10/01/2002	Pavia	Surface water	SXB_BS.0003
188795	SM49 <sup>b</sup>	03/09/2009	Brescia	Chicken farm	SXB_PR.0753
188787	SM51	09/17/2002	Pavia	Surface water	SXB_BS.0003
188781	SM50 <sup>b</sup>	07/31/2001	Modena	Minced pork	SXB_PR.0751

<sup>a</sup> The isolates above the line break are the outbreak-related isolates (15 human-origin and 5 food-origin isolates), and those below the line break are the 19 *Salmonella* Manhattan collection isolates. SM32 and SM52 were also collection isolates, but they were eventually attributed to the outbreak, following the results of this study.

<sup>b</sup> These *Salmonella* Manhattan isolates were selected for whole-genome sequencing.

of 33 isolates. Genomic DNA was extracted from overnight cultures using the Qiagen DNeasy blood and tissue kit (Qiagen) and quality controlled and quantified using a Synergy H1 hybrid multimode microplate reader (BioTek, Winooski, VT, USA). The sequencing libraries were prepared with the Nextera XT sample preparation kit (Illumina, San Diego, CA, USA), and sequencing was performed on the Illumina MiSeq platform, with a 2 × 250-bp paired-end run.

**Read quality check and assembly.** All read sets were evaluated for sequence quality and read-pair length using the softwares FastQC and Flash (23). FastQC allowed us to assess the overall quality of the generated sequences, while Flash was used to measure the distance between the sequence read pairs. All the read sets that passed the quality check (visual check for FastQC and average read pair distance > 100 nucleotides [nt] for Flash) were assembled with MIRA 4.0 (24) using accurate settings for *de novo* assembly mode.

**In silico multilocus sequence typing.** *In silico* MLST was performed using the MLST scheme optimized by the University of Warwick (<http://mlst.warwick.ac.uk/mlst/dbs/Senterica>).

**Comparative genomics by local variation calling.** In a previous work, we sequenced and published the first improved high-quality draft genome (25) of *Salmonella* Manhattan (strain 111113) (26). The 18 contigs of the *Salmonella* Manhattan 111113 genome, belonging to a human isolate of the outbreak presented here, were concatenated in a pseudochromosome and used as a reference for alignment of each of the other 32 genome assemblies included in this study, using progressiveMauve (27). A previously described bioinformatic pipeline (28) was then used to merge the results of all isolates for comparison and to extract the coordinates of all local variations spanning from SNPs to longer variations (mutations, insertions, and deletions), based on the annotation of the reference genome of strain 111113. Core SNPs were identified as single nondegenerate mutated bases flanked by identical bases and present in all 33 genomes (including that of strain 111113). Genes presenting at least one core SNP were selected and compared against the Virulence Factors Database (VFDB) (29–31), using a BLAST search with a 10<sup>-5</sup> E value cutoff.

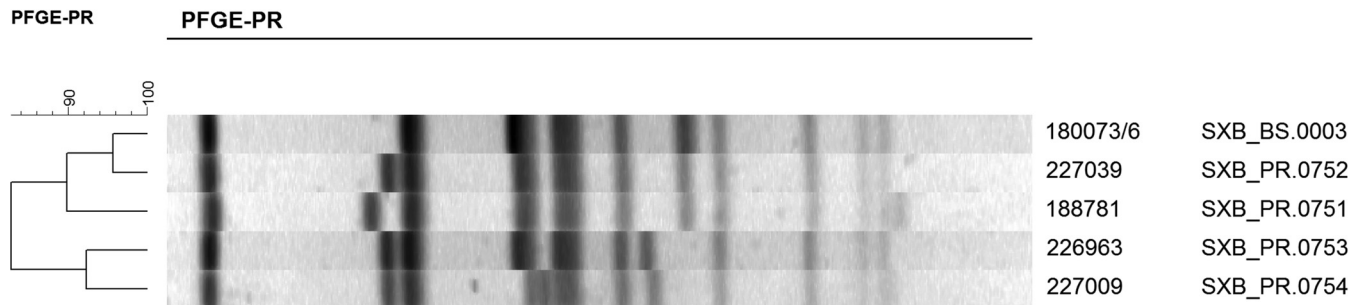


FIG 1 Similarity of *Salmonella* Manhattan isolates, examined in this study, inferred by pulsed-field gel electrophoresis profiles (PFGE-PR). The samples underwent XbaI restriction and pattern analysis according to the standard PulseNet protocol. The UPGMA dendrogram of all the profiles of the study is reported on the left; the ruler indicates the similarity values. The laboratory numbers of the isolates and their pulstotypes are reported on the right.

**Analysis of variations.** Open reading frames (ORFs) were predicted and translated on all assembled genomes (including the previously published *Salmonella* Manhattan strain 111113 genome [26]) using Prodigal (32). Next, every genomic variation (SNPs, mutations, insertions, and deletions) was parsed in order to assign it to one of the following subsets of isolates: (i) all outbreak-related isolates, irrespective of the human, food, or raw meat origin; (ii) outbreak-related human-origin-only isolates; and (iii) outbreak-related food-origin-only isolates (including those from sausage and raw meat).

**Phylogenetic analysis.** From the core SNP data set, different subsets were generated: (i) nonsynonymous SNPs, (ii) synonymous SNPs, and (iii) SNPs at the first, second, or third codon position. The core and subsets of SNPs were used as inputs for generating SNP-based phylogenies using the maximum likelihood (ML) or the Bayesian methods. Model choice was evaluated in JModelTest (33). Maximum likelihood analyses were run in PhyML (34), with a generalized time-reversible (GTR) substitution model and 100 bootstrap iterations, while Bayesian analyses were run in MrBayes (35, 36), using the same model for 2,000,000 generations, with chains sampled every 1,000 generations. The final parameter values and trees were summarized after discarding 25% of the posterior sample. The ML and Bayesian trees were displayed and edited for publication with FigTree version 1.4.0.

**Nucleotide sequence accession numbers.** The genome sequences of *Salmonella* Manhattan (strain 111113; study identification [ID], SM8) contigs were previously deposited at EBI under the accession no. CBKW010000001 to CBKW010000021 (project PRJEB1854). The newly 32 sequenced genomes (contigs) of *Salmonella* Manhattan were deposited at EBI under the project number PRJEB5339 and are summarized here in the format isolate lab no./study identification no.: WGS accession number: 160969\_3/SM1: CCBJ010000610 to CCBJ010000701, 160969\_5/SM2: CCBJ010000175 to CCBJ010000212, 160969\_6/SM3: CCBJ010000291 to CCBJ010000308, 165051\_2/SM4: CCBJ010001977 to CCBJ010002069, 165051\_3/SM5: CCBJ010002070 to CCBJ010000089, 165051\_5/SM6: CCBJ010000001 to CCBJ010000100, 165051\_7/SM7: CCBJ010004043 to CCBJ010004081, 165051\_11/SM9: CCBJ010003194 to CCBJ010003512, 165051\_12/SM10: CCBJ010000309 to CCBJ01000327, 180073\_1/SM11: CCBJ010002338 to CCBJ010002378, 180073\_2/SM12: CCBJ010003726 to CCBJ010003749, 180073\_3/SM13: CCBJ010001070 to CCBJ010001515, 180073\_4/SM14: CCBJ010001516 to CCBJ010001924, 180073\_6/SM15: CCBJ010000702 to CCBJ010000770, 250920/SM42: CCBJ010000328 to CCBJ010000609, 227021/SM32: CCBJ010004870 to CCBJ010004957, 188801/SM52: CCBJ010002097 to CCBJ010002229, 216630\_1/SM53: CCBJ010002817 to CCBJ010003193, 216630\_2/SM54: CCBJ010000213 to CCBJ010000238, 226963/SM17: CCBJ010002257 to CCBJ010002337, 226972/SM19: CCBJ010002230 to CCBJ010002256, 226979\_1/SM21: CCBJ010000101 to CCBJ010000174, 226985/SM23: CCBJ010003750 to CCBJ010004042, 226987/SM24: CCBJ010003702 to CCBJ010003725, 226998/SM27: CCBJ010000771 to CCBJ010001069, 227009/SM29: CCBJ010001925 to CCBJ010001976,

227033/SM35: CCBJ010000239 to CCBJ010000268, 227039/SM36: CCBJ010000269 to CCBJ010000290, 227052/SM38: CCBJ010002379 to CCBJ010002816, 188806/SM48: CCBJ010003540 to CCBJ010003701, 188795/SM49: CCBJ010004082 to CCBJ010004692, and 188781/SM50: CCBJ010004693 to CCBJ010004869.

## RESULTS

We present here reanalysis by WGS of an outbreak caused by *Salmonella* Manhattan in the province of Modena (Italy) in 2009. The isolates from the human cases were SM1, -2, -3, -4, -5, -6, -7, -8, -9, -10, -11, -12, -13, -14, and -15. Out of the 21 collection isolates available, all were genotyped by PFGE to search for clues on the source of infection, and SM21, -23, -24, -26, -27, -28, -32, -35, -38, -47, -48, -51, and -52 showed the outbreak pulstotype; however, SM36, -16, -17, -19, -29, -31, -49, and -50 belonged to different pulstotypes, and a selection of them were included in this study as outgroup isolates. SM42, -53, and -54 were isolated during the microbiological follow-up of the episode and presented the outbreak pulstotype.

**Pulsed-field gel electrophoresis.** The 39 *Salmonella* Manhattan isolates of the study showed five different XbaI-PFGE profiles: SXB\_BS.0003, SXB\_PR.0753, SXB\_PR.0754, SXB\_PR.0751, and SXB\_PR.0752 (Fig. 1). All the human isolates (SM1 to SM15) showed the same PFGE profile (SXB\_BS.0003), supporting the hypothesis that the unusually high incidence of this rare serovar was due to a single epidemic clone.

Another 13 isolates from the IZSLER surveillance collection belonged to genotype SXB\_BS.0003. Among these, three (SM32, SM48, and SM52) dated back to just before the outbreak period (May/June 2009) and were pivotal in guiding the epidemiological investigation. SM48 originated from an establishment near the outbreak area that processed swine guts for the salami industry. Due to this microbiological and molecular finding, the establishment was suspected of having a role in the outbreak, although no evident correlation with the human infections was made. More significantly, SM32 and SM52 were isolated just before the onset of the episode from pork sausages produced at an industrial establishment that shipped to retail stores in the outbreak area. Consequently, sausages from this producer, which were on sale in the outbreak area, were sampled along with the pork purchased by the producer. Both the sausages and the pork were positive for *Salmonella* Manhattan with the outbreak pulstotype (SXB\_BS.0003) (isolates SM53 and SM54 from the sausages and SM42 from pork). Interestingly, two *Salmonella* Manhattan isolates from our collection, isolated within the own-check hygiene procedures of

TABLE 2 Characteristic SNPs of three groups of outbreak-related isolates

Group of isolates	Amino acid change	Codon change	Position CDS <sup>a</sup>	Type of SNP	Gene	Locus →tag	Strand	Product name
All outbreak	C→R	TGT→CGT	625	Genic	<i>cobT</i>	SMA01→2283	–	Nicotinate-nucleotide–dimethylbenzimidazole phosphoribosyltransferase
	N→N	AAT→AAC	156	Genic	<i>gntR</i>	SMA01→3706	–	Gluconate utilization system Gnt-transcriptional repressor
	A→T	GCC→ACC	577	Genic	<i>ansB</i>	SMA01→3765	–	L-Asparaginase
	V→A	GTC→GCC	988	Genic	<i>dcuC</i>	SMA01→4465	–	Putative cryptic C <sub>4</sub> -dicarboxylate transporter
	K→E	AAA→GAA	70	Intergenic Genic	<i>betI</i>	SMA01→1140	+	Transcriptional regulator, TetR family
Human origin	M→T	ATG→ACG	584	Genic	<i>dsbI</i>	SMA01→0572	+	Thiol-disulfide oxidoreductase, DsbB-like
	A→T	GCC→ACC	310	Genic	<i>sthD</i>	SMA01→3447	–	β-fimbriae usher protein
	V→V	GTT→GTC	465	Genic	<i>ispH</i>	SMA01→3526	+	4-hydroxy-3-methylbut-2-enyl diphosphate reductase
	Q→STOP	CAA→TAA	252	Genic Intergenic	<i>rfbD</i>	SMA01→4557	+	UDP-galactopyranose mutase
Food origin	S→I	AGC→ATC	872	Genic	<i>fliK</i>	SMA01→2244	+	Flagellar hook-length control protein FliK
	P→L	CCT→CTT	17	Genic		SMA01→0101	+	Hypothetical protein
	A→V	GCC→GTC	1544	Genic	<i>fdrA</i>	SMA01→4374	+	Protein FdrA: acyl-CoA synthetase <sup>b</sup>

<sup>a</sup> CDS, coding sequence.<sup>b</sup> CoA, coenzyme.

the producer (SM23 and SM24) 3 years before the outbreak, presented the same genotype. Also, the surveillance collection isolates SM21, -26, -27, -28, -35, -38, -47, and -51 shared the outbreak pulsotype, but they did not seem to be correlated with the outbreak or source of infection.

Among the other nonoutbreak PFGE profiles detected, the pulsotype SXB\_PR.0752 (isolate SM36) had 95% similarity with the outbreak pulsotype, while the genotypes SXB\_PR.0751 (isolates SM31 and SM50), SXB\_PR.0753 (isolates SM16, SM17, SM19, and SM49), and SXB\_PR.0754 (isolate SM29) were less similar (90%, 84%, and 84%, respectively) (Fig. 1).

**Whole-genome sequencing.** The genomes of the 33 *Salmonella* Manhattan isolates considered for genomic analysis, including the already deposited genome of strain 111113 (26), were sequenced, quality checked, and assembled to draft status, from an average of 2,593,738 MiSeq paired-end reads per genome. The average sequenced genome characteristics were 4,678,201 nt in length, 150 large (>1,000 nt) contigs, and an  $N_{50}$  of 212,360. The genome data for each isolate are listed in Table S1 in the supplemental material. The MLST profile was determined for all draft genomes, which were found to belong to the same sequence type (ST), ST18. All assembled genomes underwent comparative and phylogenetic analyses.

**Analysis of variations.** A comparative genomic analysis was implemented to detect the differences between the *Salmonella* Manhattan genomes, in terms of nucleotide variations, exclusive to (i.e., present in all the isolates of a group and absent in all the others) the outbreak-related isolates, as divided into the following main groups: (i) all outbreak-related isolates, irrespective of the human, food, or raw meat origin; (ii) outbreak-related human-origin-only isolates; and (iii) outbreak-related food-origin-only isolates (including sausages and raw meat).

Of all the nondegenerate nucleotide variations (total 9,410) discovered by the progressiveMauve algorithm, 14 were outbreak specific, and all were core SNPs (two intergenic, two synonymous,

and 10 nonsynonymous), divided as six variations exclusive to all outbreak-related isolates, three variations characteristic of the food-origin-only outbreak-related isolates, and five characteristic of the outbreak-related human-origin-only isolates (Table 2).

**Phylogenetic analysis.** Phylogeny was reconstructed using an SNP-based approach. SNPs were extracted from the assembled genomes using a bioinformatic pipeline (28) based on progressiveMauve (27). Of the 9,410 detected variations, 953 were core SNPs, with 224 being synonymous and 467 being nonsynonymous; the remaining 262 SNPs were marked as intergenic. Among the synonymous SNPs, 6% and 94% were located in the first and third codon positions, respectively, while among the nonsynonymous SNPs, 43% were in the first, 42% in the second (total, 85% for the two positions), and 15% in the third codon position. The number of synonymous and nonsynonymous core SNPs at the first, second, and third positions were 214, 194, and 283, respectively.

The phylogenetic analysis of the study isolates was performed separately based on the different subsets of SNPs considered, namely, core, synonymous, nonsynonymous, and different codon positions using both Bayesian (Fig. 2 to 4) and maximum likelihood algorithms (see Fig. S1 and S2 in the supplemental material). Both algorithms returned the same phylogenetic results on each subset.

All data sets identified two major clades: one grouping all the isolates belonging to pulsotype SXB\_BS.0003 and the highly related SXB\_PR.0752 (95% similarity), and the other constituted by isolates with different pulsotypes (SXB\_PR.0753, SXB\_PR.0754, and SXB\_PR.0751). Interestingly, WGS analyses clustered isolate SM36 (pulsotype SXB\_PR.0752) together with the isolates of pulsotype SXB\_BS0003, meaning they are highly related compared to isolates of the other pulsotypes of the study. Therefore, we considered SXB\_PR.0752 together with SXB\_BS.0003 for the subsequent analyses of phylogeny and presence of variants.

Phylogeny based on core SNPs revealed four main groups in-

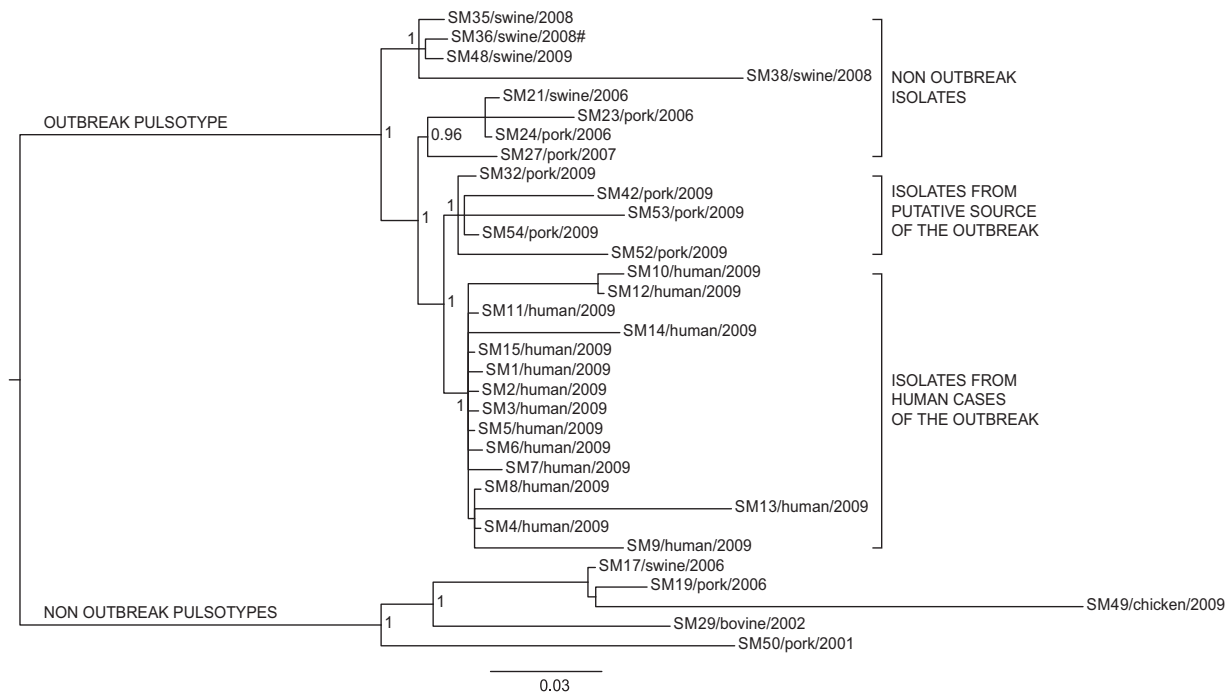


FIG 2 Bayesian phylogeny of the 33 *Salmonella* Manhattan sequenced genomes based on core SNPs. The posterior probabilities are indicated in each principal node of the tree. The scale bar units are the nucleotide substitutions per site. #, WGS analyses clustered isolate SM36 (pulsotype SXB\_PR.0752) together with the isolates of the outbreak pulsotype (SXB\_BS0003).

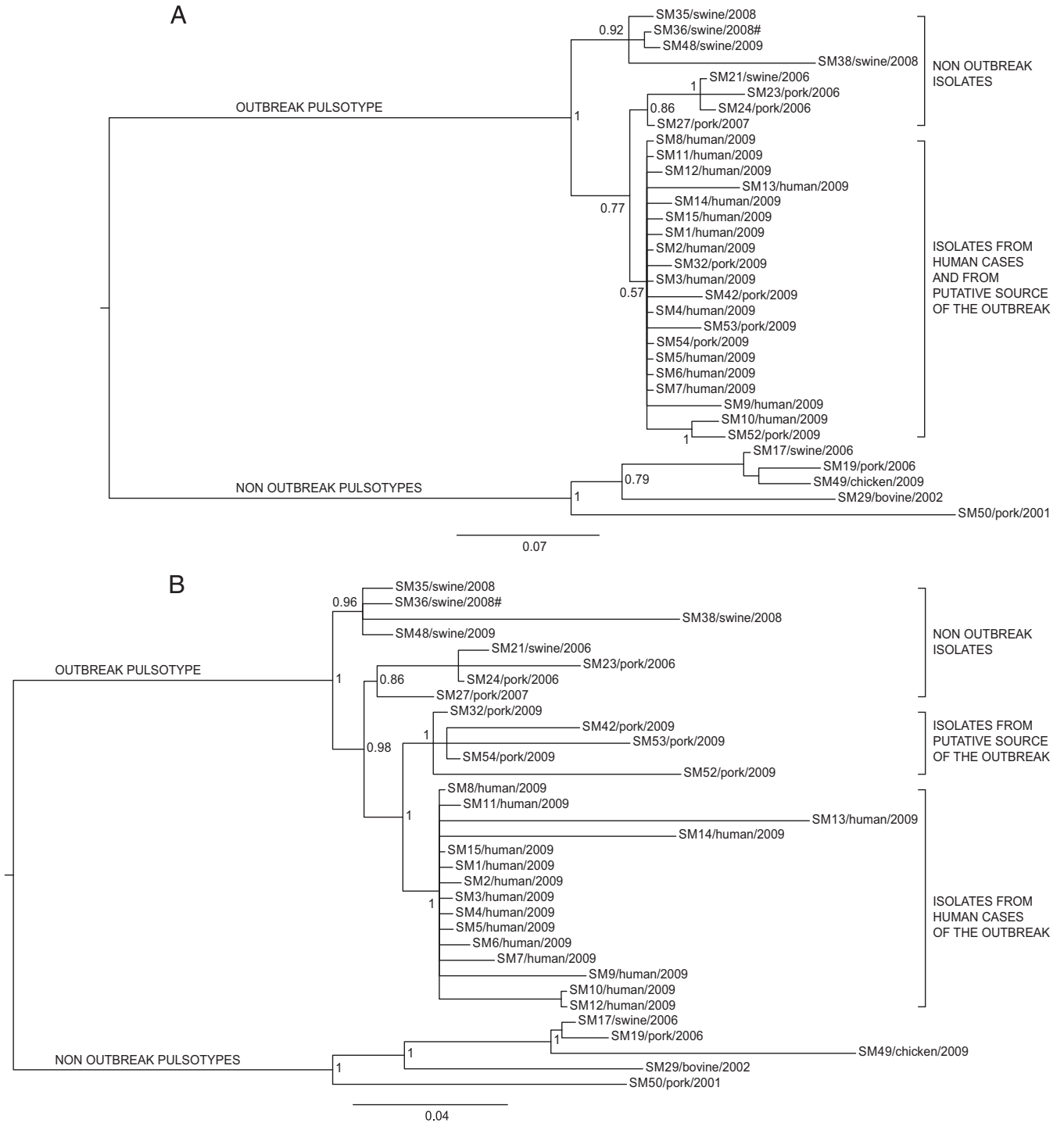
side the outbreak pulsotype. Isolates that were not epidemiologically related to the outbreak formed two monophyletic clusters, with the outermost one grouping isolates from various locations and previous years but always from swine stool within the own-check procedures of pig farms (isolates SM35, SM36, and SM38) or at food processing plants (isolate SM48). The other group included isolates collected at the sausage-producing establishment within its hygiene monitoring system 3 years before the outbreak (SM23 and SM24), along with an isolate collected on a pig farm in the same period (SM21). Isolate SM27 originated from another food processing plant in the same area of the sausage producer, but that was never linked to the outbreak.

The two innermost clusters included all the outbreak-related isolates. Five strains isolated from sausages prepared by the implicated producer (SM32, SM42, SM52, SM53, and SM54), both at a retail locations in the outbreak area and at the establishment, which were distinct from the cluster of human isolates of the outbreak (from SM1 to SM15). All outbreak-related isolates are monophyletic, confirming their derivation from a common ancestor. In order to better investigate the relationships among those isolates, we performed additional analyses on specific subsets of the core SNPs to take into account the possible effects of selective evolutionary pressure. We separately considered nonsynonymous SNPs, synonymous SNPs, and SNPs at the first, second, and third codon positions as presumptively subjected to decreasing selective pressures (37). The trees corresponding to the different subsets of SNPs are shown in Fig. 3 and 4. The trees generated by nonsynonymous SNPs and SNPs at the first plus second and second codon positions showed the same topology described by the whole data set of core SNPs, with a clear distinction between outbreak-related isolates of human and food origins. The phylogenies gen-

erated by SNPs under minor selective pressure (i.e., third position) revealed different scenarios, with the loss of a node inside the outbreak cluster showing isolates of human origin as a subgroup within the food-origin outbreak isolates. Considering synonymous SNPs only, the outbreak isolates of human and food origins are grouped in one cluster, being indicative of a single circulating clone. The phylogenetic inferences made by Bayesian and maximum likelihood algorithms gave identical results (see Fig. S1 and S2 in the supplemental material).

## DISCUSSION

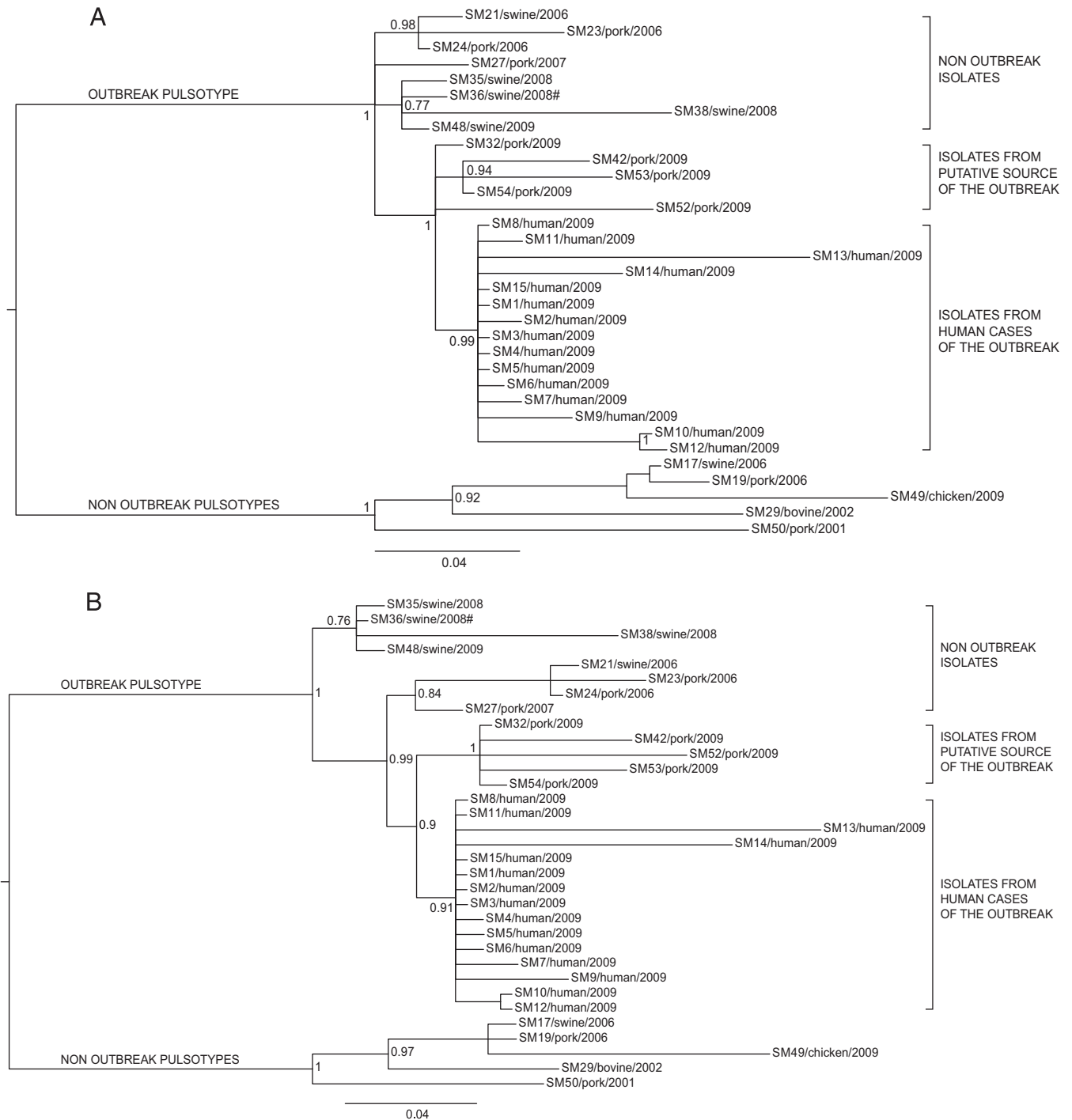
Microbiologists often need to determine the relatedness of bacterial isolates to define the network of relationships of an infectious outbreak and effectively assist epidemiological investigations. Standard protocols for typing *Salmonella* rely on internationally accepted methods, like PFGE and MLVA, which a few decades ago flanked the more limited serotyping. The possibility of accessing the vast amount of information provided by WGS of bacterial isolates promises to be the next frontier of subtyping methods, probably capable of surpassing PFGE and MLVA for molecular epidemiological purposes. In this study, we reanalyzed a well-defined *Salmonella* Manhattan outbreak detected in the summer of 2009 in the province of Modena (Italy) using WGS in order to test the power of this approach for resolving the ambiguities left by PFGE. The epidemic episode involved 15 human cases from June to July 2009, with all presenting the same PFGE profile (SXB\_BS.0003). The molecular epidemiological investigation of the outbreak involved several isolates, some from the infectious episode and others from the historic collection of the regional surveillance system of the food chain. As expected, PFGE analysis attributed the same pulsotype (SXB\_BS.0003) to all the outbreak-



**FIG 3** Phylogenetic Bayesian analysis of the 33 *Salmonella* Manhattan sequenced genomes based on synonymous (A) and nonsynonymous (B) SNP data sets. The posterior probabilities are indicated in each principal node of the tree. The scale bar units are the nucleotide substitutions per site. #, WGS analyses clustered isolate SM36 (pulsotype SXB\_PR.0752) together with the isolates of the outbreak pulsotype (SXB\_BS0003).

related isolates, but the same pulsotype was shared by many historic isolates as well. On the contrary, the WGS-based phylogeny inferred from the total core SNPs clearly showed the presence of four distinct groups of isolates (Fig. 2) within the outbreak pulsotype. The first branch of the tree, within the outbreak pulsotype, separates nonoutbreak historic isolates recovered from swine

stool at different locations and times. Among these, we find isolate SM48, which was originally suspected of being implicated in the infectious episode, based on PFGE, and eventually cleared by WGS. Interestingly, isolate SM36, which does not belong to pulsotype SXB\_BS.0003 but to the highly similar (95% similarity) pulsotype SXB\_PR.0752, is included in this clade. This is a clear



**FIG 4** Phylogenetic Bayesian analysis of the 33 *Salmonella* Manhattan sequenced genomes based on SNPs in first (A), second (B), third (C), and first plus second codon position (D) data sets. The posterior probabilities are indicated in each principal node of the tree. The scale bar units are the nucleotide substitutions per site. #, WGS analyses clustered isolate SM36 (pulsotype SXB\_PR.0752) together with the isolates of the outbreak pulsotype (SXB\_BS.0003).

discrepancy between WGS and the more limited PFGE that relies on only few genomic loci (rare restriction sites) for its typing inferences. By placing SM36 together with pulsotype SXB\_BS.0003 isolates, our WGS approach indicates that a limited genomic difference between isolates is able to jeopardize the typing outcome of PFGE. This observation confirms what Tenover et al. (38) already pointed out, the fact that as PFGE may be heavily influenced

by a single mutational event (e.g., SNP occurring in a restriction site), isolates should be considered to be possibly related even if they differ by two or three bands. However, according to this conservative interpretation of PFGE results, the vast majority of the isolates of our study should be regarded as potentially belonging to the outbreak. This would have not been sufficiently discriminatory to help the epidemiological investigations. The interpre-

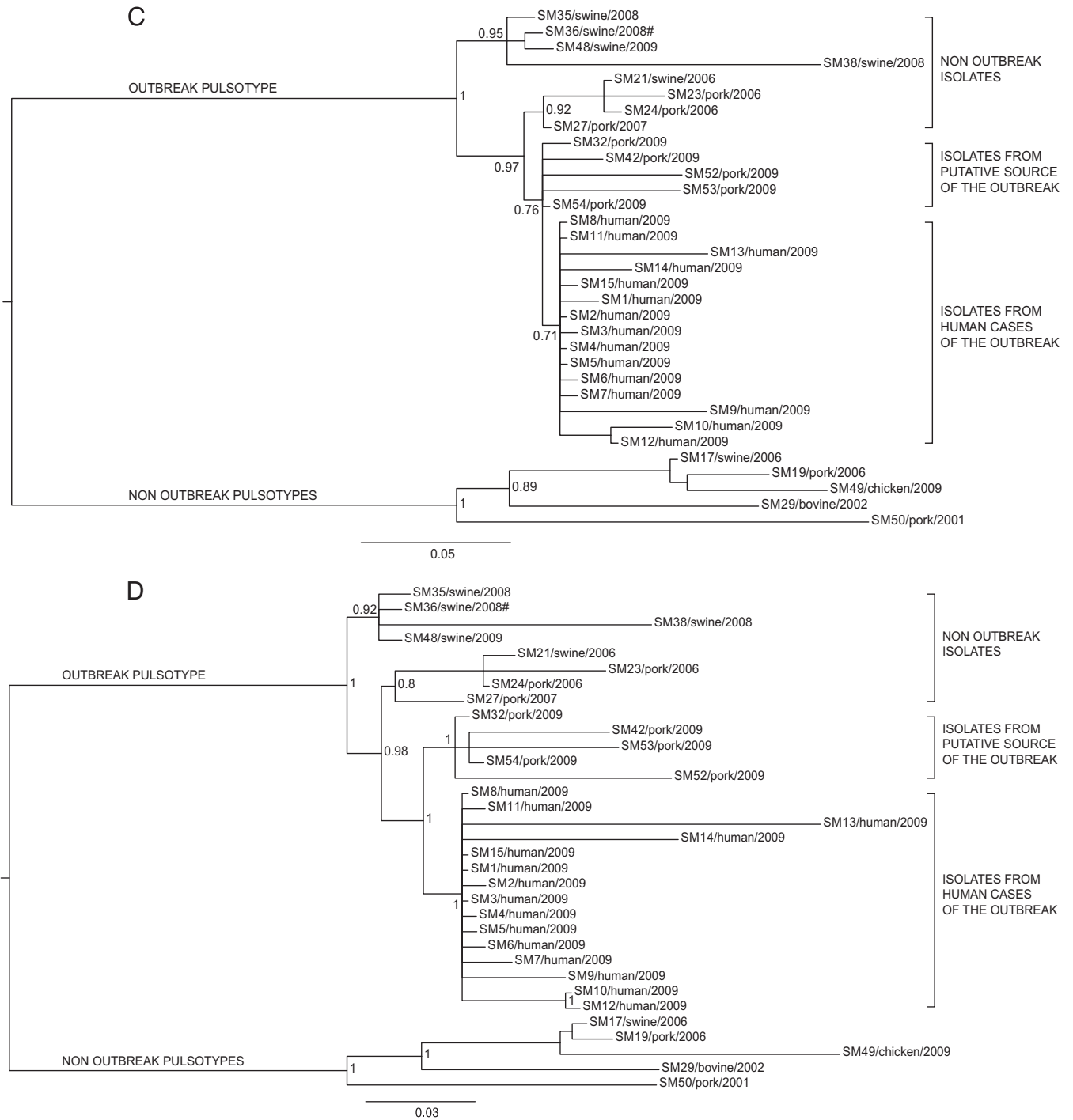


FIG 4 continued

tation criteria of Tenover et al. (38) are derived from logical considerations; as such, they are intrinsically valid, and our observations regarding isolate SM36 confirms their validity. At the same time, their use leaves molecular epidemiologists with considerable uncertainty about how to interpret PFGE results with regard to whether or not different pulsotypes are part of a single outbreak. In our case, WGS removed that uncertainty about SM36.

Moving deeper along the phylogenetic tree based on the total

core SNPs, three other groups of isolates are evident. The outermost set of this node includes isolates (SM21, -23, -24, and -27) not related to the outbreak, as they were collected 3 years before (2006). It is interesting, however, to notice that WGS-based phylogeny indicates these strains to be closer to the outbreak node (inner branch) than was the previous set of swine-stool isolates. On a better look, we were struck by the fact that SM23 and SM24 were collected in 2006, within the own-check procedures of the

sausage producer involved in the 2009 outbreak. Moreover, SM21, which is subbranched with SM23 and SM24, was routinely recovered from a local pig farm (from swine stool) at the same time as SM23 and SM24. While this specific molecular similarity was not inferred by PFGE, WGS highlighted a possible link between these two commercial entities. Moving one branch forward in the phylogenetic tree, WGS shows another bifurcation actually separating outbreak-related isolates of human origin from those of food origin. While still speculative, based on this WGS-based phylogeny, coupled with epidemiological data, we could argue that this outbreak was due to a persistent *Salmonella* Manhattan clone, which may have infected one or more pig farms and reached the food producer and the retail customers as animals arrived at the slaughterhouse in a nonclinical septic condition. This is a typical mode of transmission of *Salmonella* along the food chain, as it may asymptotically persist (thus going unnoticed) within a herd of pigs for long periods of time (even years). Sporadically, animals carrying a high level of the pathogen arrive at the slaughterhouse and contaminate a defined set of food products, thus causing an infectious outbreak as the final consumers (39, 40) become exposed to it. In this scenario, WGS seems to depict a more detailed and articulated epidemiological story. In fact, the tree inferred from core SNPs (Fig. 2) leaves a certain level of uncertainty relative to the actual causative relationship between the isolates of food origin and of human origin within the outbreak, as they cluster in two distinct groups, although very closely to each other, as evidenced by the limited number of exclusive core SNPs accumulated by the two groups (3 for food and 5 for human isolates). In the absence of epidemiological insights, we argue that the two sets of isolates are very similar to each other but still are separate entities. This substantially contradicts the epidemiological evidence that the two sets of isolates belong to the same outbreak clone. Therefore, we further investigated this apparent inconsistency of the WGS-based results by comparing new alternative phylogenies based on two different subsets of polymorphisms, synonymous and nonsynonymous, instead of the total core SNPs. The trees generated from these two subsets of SNPs were different (Fig. 3A and B). Phylogenetic analysis based on nonsynonymous SNPs (Fig. 3B) still divided the outbreak isolates of food and human origins, as in the approach based on total core SNPs. On the contrary, the tree obtained from synonymous SNPs (Fig. 3A) clustered the human isolates together with the food isolates, indicating that all outbreak-related *Salmonella* Manhattan strains constituted a single clone, in line with epidemiological evidence. While intriguing, this new outcome may have been the misleading effect of the smaller amount of data present in these new subsets than that with the total set of core SNPs, of which there were 953, whereas the number of synonymous and nonsynonymous SNPs were 224 and 467, respectively. Therefore, to confirm these results, we took a step forward in this approach by considering not just synonymous versus nonsynonymous SNPs but also taking into account the different codon position of each SNP in the core genome. *Salmonella* Manhattan synonymous SNPs were at the 3rd codon position 94% of the time, while nonsynonymous SNPs were at the 2nd 42% and at the first position 43% of the time (total, 85%). In this study, 1st, 2nd, and 3rd position SNPs accounted for 214, 194, and 283 nucleotide substitutions, respectively. The comparison of subsets of SNPs based on their codon site would then not be impaired by too-large differences in the amount of data processed by the phylogenetic algorithms. The

tree obtained from second codon position (Fig. 4B) was comparable to that of the nonsynonymous SNPs, as expected, whereas the tree obtained from third codon position showed human isolates as a subgroup of the food isolates (Fig. 4C), essentially confirming the tree based on synonymous SNPs. These results show that at least limited to our outbreak, synonymous and third-position SNPs were the only ones able to describe the causal relationship between food (source of the outbreak) and clinical isolates in a way that was consistent with the epidemiological evidence. At the same time, our results indicate that nonsynonymous and total core SNPs may have led to misleading conclusions about the relationships between the human and food isolates of the outbreak. One last aspect that caught our attention by deciphering topologies of this WGS-based retrospective analysis was that SNP-based clustering of isolates separated human from food outbreak-related isolates when considering total core SNPs (Fig. 2). As we just discussed, this topology was mainly influenced by nonsynonymous mutations, which means it is possible to find distinctive nonsynonymous SNPs for each group of isolates (human versus food). Using progressiveMauve, we identified a set of 953 core SNPs, among which we selected those that were exclusive to specific clusters of interest: six SNPs exclusive to all outbreak isolates (human and food origin), three exclusive to all food origin outbreak isolates, and only five exclusive to all human origin outbreak isolates (Table 2). The extremely limited number of exclusive SNPs in food and human isolates within the outbreak is an additional compelling element indicative of the fact that these two groups of isolates did not have enough evolutionary time to significantly differentiate, indicating they belong to the same clone. A BLAST analysis of these SNPs against the Virulence Factors Database revealed three genes of particular interest: (i) *fliK*, coding for a flagellar hook-length control protein (41), (ii) *sthD*, a gene coding for a fimbrial outer membrane usher protein (42), and (iii) *rfbD*, coding for a UDP-galactopyranose mutase precursor involved in the synthesis of the O antigen of the lipopolysaccharide (LPS). All three proteins are virulence determinants in *Salmonella* (43–46). WGS has already proved its usefulness for elucidating the evolutionary diversity of large populations of bacterial isolates (11, 47, 48). In the specific case of *Salmonella*, WGS was successfully applied to illuminate the diversity of the pathogen within a vast epidemic episode, allowing highly efficient traceback of clinical and food isolates (4, 13). The results obtained in this study underscore the power of WGS-based methods, when applied together with the most appropriate phylogenetic tools, to resolve small outbreaks characterized by few and highly clonal bacterial isolates. Our comparative genomics approach was able to correctly cluster the clinical isolates within the composite scenario of outbreak-related and collection isolates. Accurate backtracking to the source of infection at the retail and industrial levels was made possible while flagging an originally overlooked suspicious correlation with a farm supplier and clearing an originally suspect food operator. Moreover, by selectively choosing the different types of detected nucleotide variations, we were able to read the message hidden within neutral mutations as opposed to the general use of total core SNPs. Further use of the differential analysis of synonymous and nonsynonymous mutations will test the validity of this approach in deciphering the details of infection transmission in the context of other outbreaks caused by *Salmonella* and, potentially, other pathogens.



## ACKNOWLEDGMENTS

We acknowledge Elena Carra for providing some of the isolates included in the study. We thank Roberto Alfieri for technical assistance in remote analysis at the University of Parma Department of Physics and Earth Sciences.

This study was supported by Regione Lombardia grant delibera regionale 001051/22122000 and by the Italian Ministry of Health grant IZSLER-PRC2012/006.

## REFERENCES

- Majowicz SE, Musto J, Scallan E, Angulo FJ, Kirk M, O'Brien SJ, Jones TF, Fazil A, Hoekstra RM, International Collaboration on Enteric Disease 'Burden of Illness' Studies. 2010. The global burden of nontyphoidal *Salmonella* gastroenteritis. *Clin Infect Dis* 50:882–889. <http://dx.doi.org/10.1086/650733>.
- European Food Safety Authority (EFSA), European Centre for Disease Prevention and Control (ECDC). 2013. Scientific report of EFSA and ECDC: the European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2011. *EFSA J* 11: 3129–3378. <http://dx.doi.org/10.2903/j.efsa.2013.3129>.
- Wattiau P, Boland C, Bertrand S. 2011. Methodologies for *Salmonella enterica* subsp. *enterica* subtyping: gold standards and alternatives. *Appl Environ Microbiol* 77:7877–7885. <http://dx.doi.org/10.1128/AEM.05527-11>.
- Allard MW, Luo Y, Strain E, Pettengill J, Timme R, Wang C, Li C, Keys CE, Zheng J, Stones R, Wilson MR, Musser SM, Brown EW. 2013. On the evolutionary history, population genetics and diversity among isolates of *Salmonella* Enteritidis PFGE pattern JEGX01.0004. *PLoS One* 8:e55254. <http://dx.doi.org/10.1371/journal.pone.0055254>.
- Urw R, Maiden MCJ. 2003. Multi-locus sequence typing: a tool for global epidemiology. *Trends Microbiol* 11:479–487. <http://dx.doi.org/10.1016/j.tim.2003.08.006>.
- Achtman M, Wain J, Weill F-X, Nair S, Zhou Z, Sangal V, Krauland MG, Hale JL, Harbottle H, Uesbeck A, Dougan G, Harrison LH, Brisse S, *S. enterica* MLST Study Group. 2012. Multilocus sequencing typing as a replacement for serotyping in *Salmonella enterica*. *PLoS Pathog* 8:e1002776. <http://dx.doi.org/10.1371/journal.ppat.1002776>.
- Fakhr MK, Nolan LK, Logue CM. 2005. Multilocus sequence typing lacks the discriminatory ability of pulsed-field gel electrophoresis for typing *Salmonella enterica* serovar Typhimurium. *J Clin Microbiol* 43:2215–2219. <http://dx.doi.org/10.1128/JCM.43.5.2215-2219.2005>.
- Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ, Bentley SD. 2010. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 327: 469–474. <http://dx.doi.org/10.1126/science.1182395>.
- Gardy JL, Johnston JC, Sui SJH, Cook VJ, Shah L, Brodtkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJM, Brinkman FSL, Brunham RC, Tang P. 2011. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med* 364:730–739. <http://dx.doi.org/10.1056/NEJMoa1003176>.
- Desai PT, Porwollik S, Long F, Cheng P, Wollam A, Clifton SW, Weinstock GM, McClelland M. 2013. Evolutionary genomics of *Salmonella enterica* subspecies. *mBio* 4:e00579-12. <http://dx.doi.org/10.1128/mBio.00579-12>.
- Timme RE, Pettengill JB, Allard MW, Strain E, Barrangou R, Wehnes C, Van Kessel JS, Karns JS, Musser SM, Brown EW. 2013. Phylogenetic diversity of the enteric pathogen *Salmonella enterica* subsp. *enterica* inferred from genome-wide reference-free SNP characters. *Genome Biol Evol* 5:2109–2123. <http://dx.doi.org/10.1093/gbe/evt159>.
- Hoffmann M, Zhao S, Pettengill J, Luo Y, Monday SR, Abbott J, Ayers SL, Cinar HN, Muruvanda T, Li C, Allard MW, Whichard J, Meng J, Brown EW, McDermott PF. 2014. Comparative genomic analysis and virulence differences in closely related *Salmonella enterica* serotype Heidelberg isolates from humans, retail meats, and animals. *Genome Biol Evol* 6:1046–1068. <http://dx.doi.org/10.1093/gbe/evu079>.
- Allard MW, Luo Y, Strain E, Li C, Keys CE, Son I, Stones R, Musser SM, Brown EW. 2012. High resolution clustering of *Salmonella enterica* serovar Montevideo strains using a next-generation sequencing approach. *BMC Genomics* 13:32. <http://dx.doi.org/10.1186/1471-2164-13-32>.
- Lienau EK, Strain E, Wang C, Zheng J, Ottesen AR, Keys CE, Hammack TS, Musser SM, Brown EW, Allard MW, Cao G, Meng J, Stones R. 2011. Identification of a salmonellosis outbreak by means of molecular sequencing. *N Engl J Med* 364:981–982. <http://dx.doi.org/10.1056/NEJMcl100443>.
- Cao G, Meng J, Strain E, Stones R, Pettengill J, Zhao S, McDermott P, Brown E, Allard M. 2013. Phylogenetics and differentiation of *Salmonella* Newport lineages by whole genome sequencing. *PLoS One* 8:e55687. <http://dx.doi.org/10.1371/journal.pone.0055687>.
- Mather AE, Reid SWJ, Maskell DJ, Parkhill J, Fookes MC, Harris SR, Brown DJ, Coia JE, Mulvey MR, Gilmour MW, Petrovska L, De Pinna E, Kuroda M, Akiba M, Izumiya H, Connor TR, Suchard MA, Lemey P, Mellor DJ, Haydon DT, Thomson NR. 2013. Distinguishable epidemics of multidrug-resistant *Salmonella* Typhimurium DT104 in different hosts. *Science* 341:1514–1517. <http://dx.doi.org/10.1126/science.1240578>.
- Pang S, Octavia S, Feng L, Liu B, Reeves PR, Lan R, Wang L. 2013. Genomic diversity and adaptation of *Salmonella enterica* serovar Typhimurium from analysis of six genomes of different phage types. *BMC Genomics* 14:718. <http://dx.doi.org/10.1186/1471-2164-14-718>.
- Leekitcharoenphon P, Friis C, Zankari E, Svendsen CA, Price LB, Rahmani M, Herrero-Fresno A, Fashae K, Vandenberg O, Aarestrup FM, Hendriksen RS. 2013. Genomics of an emerging clone of *Salmonella* serovar Typhimurium ST313 from Nigeria and the Democratic Republic of Congo. *J Infect Dev Ctries* 7:696–706. <http://dx.doi.org/10.3855/jidc.3328>.
- Noël H, Dominguez M, Weill FX, Brisabois A, Duchazeaubeneix C, Kerouanton A, Delmas G, Pihier N, Couturier E. 2006. Outbreak of *Salmonella enterica* serotype Manhattan infection associated with meat products, France, 2005. *Euro Surveill* 11:270–273. <http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=660>.
- Fisher I, Crowcroft N. 1998. Enter-net/EPIET investigation into the multinational cluster of *Salmonella* Livingstone. *Euro Surveill* 2:pii=1271. <http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=1271>.
- European Food Safety Authority. 2012. Technical report: manual for reporting of food-borne outbreaks in accordance with Directive/99/EC from the year 2011. Supporting publication 2012:EN-265. European Safety Food Authority, Parma, Italy. <http://www.efsa.europa.eu/en/supporting/doc/265e.pdf>.
- PulseNet. 2010. One-day (24–28 h) standardized laboratory protocol for molecular subtyping of *Escherichia coli* O157:H7, non-typhoidal *Salmonella* serotypes, and *Shigella sonnei*, by pulsed field gel electrophoresis (PFGE). Centers for Disease Control and Prevention, Atlanta, GA. [http://www.cdc.gov/pulsenet/protocols/ecoli\\_salmonella\\_shigella\\_protocols.pdf](http://www.cdc.gov/pulsenet/protocols/ecoli_salmonella_shigella_protocols.pdf).
- Magoc T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27:2957–2963. <http://dx.doi.org/10.1093/bioinformatics/btr507>.
- Chevreur B, Wetter T, Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information, p 45–56. *In* Computer science and biology. Proceedings of the German Conference on Bioinformatics, GCB '99. GCB, Hannover, Germany.
- Chain PS1, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C, Cole JR, Ding Y, Dugan S, Field D, Garrity GM, Gibbs R, Graves T, Han CS, Harrison SH, Highlander S, Hugenholtz P, Khouri HM, Kodira CD, Kolker E, Kyrpides NC, Lang D, Lapidus A, Malfatti SA, Markowitz V, Metha T, Nelson KE, Parkhill J, Pitluck S, Qin X, Read TD, Schmutz J, Sozhamannan S, Sterk P, Strausberg RL, Sutton G, Thomson NR, Tiedje JM, Weinstock G, Wollam A, Genomic Standards Consortium Human Microbiome Project Jumpstart Consortium, Detter JC. 2009. Genome Project standards in a new era of sequencing. *Science* 326:236–237. <http://dx.doi.org/10.1126/science.1180614>.
- Sassera D, Gaiarsa S, Scaltriti E, Morganti M, Bandi C, Casadei G, Pongolini S. 2013. Draft genome sequence of *Salmonella enterica* subsp. *enterica* serovar Manhattan strain 111113, from an outbreak of human infections in northern Italy. *Genome Announc* 1:e00632-13. <http://dx.doi.org/10.1128/genomeA.00632-13>.
- Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5:e11147. <http://dx.doi.org/10.1371/journal.pone.0011147>.
- Gaiarsa S, Comandatore F, Gaibani P, Corbella M, Dalla Valle C, Epis S, Scaltriti E, Carretto E, Farina C, Labonia M, Landini MP, Pongolini

- S, Sambri V, Bandi C, Marone P, Sasser D. 2014. Genomic epidemiology of *Klebsiella pneumoniae* in Italy and novel insights into the origin and global evolution of its resistance to carbapenem antibiotics. *Antimicrob Agents Chemother* 59:389–396. <http://dx.doi.org/10.1128/AAC.04224-14>.
29. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q. 2005. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res* 33:D325–D328. <http://dx.doi.org/10.1093/nar/gki008>.
  30. Yang J, Chen L, Sun L, Yu J, Jin Q. 2007. VFDB 2008 release: an enhanced Web-based resource for comparative pathogenomics. *Nucleic Acids Res* 36:D539–D542. <http://dx.doi.org/10.1093/nar/gkm951>.
  31. Chen L, Xiong Z, Sun L, Yang J, Jin Q. 2011. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res* 40:D641–D645. <http://dx.doi.org/10.1093/nar/gkr989>.
  32. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119. <http://dx.doi.org/10.1186/1471-2105-11-119>.
  33. Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* 9:772. <http://dx.doi.org/10.1038/nmeth.2109>.
  34. Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML Web servers. *Syst Biol* 57:758–771. <http://dx.doi.org/10.1080/10635150802429642>.
  35. Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755. <http://dx.doi.org/10.1093/bioinformatics/17.8.754>.
  36. Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574. <http://dx.doi.org/10.1093/bioinformatics/btg180>.
  37. Bofkin L, Goldman N. 2006. Variation in evolutionary processes at different codon positions. *Mol Biol Evol* 24:513–521. <http://dx.doi.org/10.1093/molbev/msl178>.
  38. Tenover FC, Arbeit RD, Goering RV, Mickelsen PA, Murray BE, Persing DH, Swaminathan B. 1995. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. *J Clin Microbiol* 33:2233–2239.
  39. Rostagno MH. 2009. Can stress in farm animals increase food safety risk? *Foodborne Pathog Dis* 6:767–776. <http://dx.doi.org/10.1089/fpd.2009.0315>.
  40. Rostagno MH, Callaway TR. 2012. Pre-harvest risk factors for *Salmonella enterica* in pork production. *Food Res Int* 45:634–640. <http://dx.doi.org/10.1016/j.foodres.2011.04.041>.
  41. Uchida K, Aizawa SI. 2014. The flagellar soluble protein FliK determines the minimal length of the hook in *Salmonella enterica* serovar Typhimurium. *J Bacteriol* 196:1753–1758. <http://dx.doi.org/10.1128/JB.00050-14>.
  42. Waters RC, O'Toole PW, Ryan KA. 2007. The FliK protein and flagellar hook-length control. *Protein Sci* 16:769–780. <http://dx.doi.org/10.1110/ps.072785407>.
  43. Suez J, Porwollik S, Dagan A, Marzel A, Schorr YI, Desai PT, Agmon V, McClelland M, Rahav G, Gal-Mor O. 2013. Virulence gene profiling and pathogenicity characterization of non-typhoidal *Salmonella* accounted for invasive disease in humans. *PLoS One* 8:e58449. <http://dx.doi.org/10.1371/journal.pone.0058449>.
  44. Weening EH, Barker JD, Laarakker MC, Humphries AD, Tsois RM, Baumlér AJ. 2005. The *Salmonella enterica* serotype Typhimurium *lpf*, *bcf*, *stb*, *stc*, *std*, and *sth* fimbrial operons are required for intestinal persistence in mice. *Infect Immun* 73:3358–3366. <http://dx.doi.org/10.1128/IAI.73.6.3358-3366.2005>.
  45. Komoriya K, Shibano N, Higano T, Azuma N, Yamaguchi S, Aizawa S-I. 1999. Flagellar proteins and type III-exported virulence factors are the predominant proteins secreted into the culture media of *Salmonella* Typhimurium. *Mol Microbiol* 34:767–779. <http://dx.doi.org/10.1046/j.1365-2958.1999.01639.x>.
  46. Köplin R, Brisson J-R, Whitfield C. 1997. UDP-galactofuranose precursor required for formation of the lipopolysaccharide O antigen of *Klebsiella pneumoniae* serotype O1 is synthesized by the product of the *rfbDKPO1* gene. *J Biol Chem* 272:4121–4128. <http://dx.doi.org/10.1074/jbc.272.7.4121>.
  47. Leekitcharoenphon P, Lukjancenko O, Friis C, Aarestrup F, Ussery D. 2012. Genomic variation in *Salmonella enterica* core genes for epidemiological typing. *BMC Genomics* 13:88. <http://dx.doi.org/10.1186/1471-2164-13-88>.
  48. Lienau EK, Blazar JM, Wang C, Brown EW, Stones R, Musser S, Allard MW. 2013. Phylogenomic analysis identifies gene gains that define *Salmonella enterica* subspecies I. *PLoS One* 8:e76821. <http://dx.doi.org/10.1371/journal.pone.0076821>.

# Matrix-Assisted Laser Desorption Ionization–Time of Flight and Comparative Genomic Analysis of M-18 Group A *Streptococcus* Strains Associated with an Acute Rheumatic Fever Outbreak in Northeast Italy in 2012 and 2013

Paolo Gaibani,<sup>a</sup> Erika Scaltriti,<sup>b</sup> Claudio Foschi,<sup>a</sup> Enrico Baggio,<sup>a</sup> Maria Vittoria Tamburini,<sup>a</sup> Roberta Creti,<sup>c</sup> Maria Grazia Pascucci,<sup>d</sup> Marco Fagioni,<sup>e</sup> Simone Ambretti,<sup>a</sup> Francesco Comandatore,<sup>f</sup>  Stefano Pongolini,<sup>b</sup> Maria Paola Landini<sup>a</sup>

Operative Unit of Clinical Microbiology, St. Orsola–Malpighi University Hospital, Bologna, Italy<sup>a</sup>; Sezione Diagnostica di Parma, Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna, Parma, Italy<sup>b</sup>; Dipartimento di Malattie Infettive, Parassitarie ed Immunomediate, Istituto Superiore di Sanità, Rome, Italy<sup>c</sup>; Public Health Authority Emilia Romagna, Rome, Italy<sup>d</sup>; Bruker Daltonics s.r.l., Macerata, Italy<sup>e</sup>; Dipartimento di Scienze Veterinarie e Sanità Pubblica, Università degli Studi di Milano, Milan, Italy<sup>f</sup>

Acute rheumatic fever (ARF) is a postsuppurative sequela caused by *Streptococcus pyogenes* infections affecting school-age children. We describe here the occurrence of an ARF outbreak that occurred in Bologna province, northeastern Italy, between November 2012 and May 2013. Molecular analysis revealed that ARF-related group A *Streptococcus* (GAS) strains belonged to the M-18 serotype, including subtypes *emm18.29* and *emm18.32*. All M-18 GAS strains shared the same antigenic profile, including SpeA, SpeB, SpeC, SpeL, SpeM, and SmeZ. Matrix-assisted laser desorption ionization–time of flight (MALDI-TOF) analysis revealed that M-18 GAS strains grouped separately from other serotypes, suggesting a different *S. pyogenes* lineage. Single nucleotide polymorphisms and phylogenetic analysis based on whole-genome sequencing showed that *emm18.29* and *emm18.32* GAS strains clustered in two distinct groups, highlighting genetic variations between these subtypes. Comparative analysis revealed a similar genome architecture between *emm18.29* and *emm18.32* strains that differed from noninvasive *emm18.0* strains. The major sources of differences between M-18 genomes were attributable to the prophage elements. Prophage regions contained several virulence factors that could have contributed to the pathogenic potential of *emm18.29* and *emm18.32* strains. Notably, phage  $\Phi$ SPBO.1 carried erythrogenic toxin A gene (*speA1*) in six ARF-related M-18 GAS strains but not in *emm18.0* strains. In addition, a phage-encoded hyaluronidase gene (*hylP.2*) presented different variants among M-18 GAS strains by showing internal deletions located in the  $\alpha$ -helical and TS $\beta$ H regions. In conclusion, our study yielded insights into the genome structure of M-18 GAS strains responsible for the ARF outbreak in Italy, thus expanding our knowledge of this serotype.

*Streptococcus pyogenes*, group A streptococcus (GAS), is a Gram-positive bacterium responsible for a wide spectrum of diseases ranging from moderate or mild infections to severe invasive diseases such as necrotizing fasciitis and toxic shock-like syndrome (TSLs). Several GAS infections can cause severe postinfectious sequelae, including acute poststreptococcal glomerulonephritis, acute rheumatic fever (ARF), and rheumatic heart disease (1).

ARF is a systemic disorder resulting from an autoimmune disease following a GAS infection that usually occurs in children between 5 and 15 years of age (2). During the last several decades, the incidence of ARF cases has significantly declined in the United States and Western Europe, whereas it remains high in Eastern Europe, Asia, and Australia (3). However, the resurgence of ARF in several geographical areas, including United States, is a matter of concern (4).

*S. pyogenes* possesses different virulence factors such as the M protein and superantigens (SAGs) that contribute to the pathogenesis of GAS infection (1). On the basis of the high variability of the M protein among GAS strains, the 5′-terminal sequence of the *emm* gene (*emm* typing) is considered a reliable molecular marker commonly used for epidemiological studies (5).

Previous studies indicated that *emm* types 1, 3, 5, 6, 18, 19, 24, and 29 have been isolated from ARF cases, suggesting a “rheumatogenic” role of certain serotypes (2). Epidemiological study of

different ARF outbreaks in the United States revealed a strict association with serotype M-18 GAS (6).

Recently, matrix-assisted laser desorption ionization–time of flight (MALDI-TOF) mass spectrometry (MS) has been introduced in microbiological laboratories for prompt highly accurate identification and classification of bacterial species (7). Several studies have now demonstrated the ability of MALDI-TOF MS to

Received 10 December 2014 Returned for modification 12 January 2015

Accepted 20 February 2015

Accepted manuscript posted online 4 March 2015

Citation Gaibani P, Scaltriti E, Foschi C, Baggio E, Tamburini MV, Creti R, Pascucci MG, Fagioni M, Ambretti S, Comandatore F, Pongolini S, Landini MP. 2015. Matrix-assisted laser desorption ionization–time of flight and comparative genomic analysis of M-18 group A *Streptococcus* strains associated with an acute rheumatic fever outbreak in Northeast Italy in 2012 and 2013. *J Clin Microbiol* 53:1562–1572. doi:10.1128/JCM.03465-14.

Editor: S. S. Richter

Address correspondence to Paolo Gaibani, paolo.gaibani@unibo.it.

Supplemental material for this article may be found at <http://dx.doi.org/10.1128/JCM.03465-14>.

Copyright © 2015, American Society for Microbiology. All Rights Reserved.

doi:10.1128/JCM.03465-14

TABLE 1 Superantigens and *emm* type of GAS strains collected during an ARF outbreak

<i>emm</i> subtype	No. of isolates expressing an SAg gene										Antigenic profile	No. of isolates		
	<i>speA</i>	<i>speC</i>	<i>speG</i>	<i>speH</i>	<i>speI</i>	<i>speJ</i>	<i>speK</i>	<i>speL</i>	<i>speM</i>	<i>ssa</i>			<i>smeZ</i>	
6.4			32				32					32	1	32
28.0			9			9						9	2	9
18.29 <sup>a</sup>	6	6	6					6	6			6	3	6
18.32	3	3	3					3	3			3	3	3
5.3			4									4	4	4
5.6			1									1	4	1
5.18			1									1	4	1
1.0	3		4			4						4	5	4
44.0			2			2		2	2			2	6	2
89.0			3									3	4	3
3.88	1		1									1	7	1
3.1			1				1					1	1	1
9.0			1									1	4	1
102.3			1									1	4	1
12.0			1	1	1		1					1	8	1
Total	13	9	70	1	1	15	34	11	11			70		70

<sup>a</sup>Two consecutive GAS strains were isolated from the first ARF case.

type and distinguish a wide range of bacterial species at subspecies or strain level (8, 9).

On the other hand, whole-genome sequencing is a consolidated procedure for epidemiological and evolutionary purposes (10–12). This technique is highly sensitive and can identify single nucleotide polymorphisms (SNPs) throughout the genome (13). In particular, the resolution of strains genetically indistinguishable by other molecular techniques (i.e., multilocus sequence typing [MLST] and pulsed-field gel electrophoresis) made whole-genome sequencing technology a powerful tool for the epidemiological investigations of related high clonal bacterial isolates (12, 13).

We report here an epidemiological investigation based on both whole-genome sequencing and MALDI-TOF MS on serotype M-18 GAS strains collected from primary school-age children in Bologna province during an ARF outbreak in early 2013.

## MATERIALS AND METHODS

**Epidemiological investigation.** In February 2013, a notification of an ARF case following hospital admission was reported in an 11-year-old otherwise healthy Caucasian boy resident in Bologna province, Emilia-Romagna region. In this area, six ARF cases had been diagnosed in the previous 3 months. After the last notification of ARF diagnosis, the Regional Health Agency instituted active epidemiological surveillance to monitor all potential ARF contact cases following World Health Organization (WHO) criteria (14). The active surveillance protocol required both clinical evaluation and culture screening of all classmates of ARF cases. At the same time, 14 GAS strains isolated from school-age children with pharyngotonsillitis in the Bologna metropolitan area were collected. Two months later, diagnosis of ARF was notified in a 4-year-old Caucasian boy resident in the same province. Subsequently, 14 GAS-positive samples were collected from classmates of the second ARF case and from 34 symptomatic children resident in the same area. The date of isolation, mucoid trait, *emm* type, antimicrobial resistance, superantigen genes, and epidemiological linkage for each strain are listed in Table S1 in the supplemental material.

**Bacterial isolation and identification.** 70 GAS strains were isolated from throat swabs collected at the bacteriology laboratory of St. Orsola-Malpighi Hospital, Bologna, except for two 30-year-old *emm18* GAS

strains that had been deposited in the bacterial collection bank of Istituto Superiore di Sanità (ISS). Bacteria were initially identified using standard methods and confirmed by MALDI-TOF 3.1 RTC (Bruker Daltonics, GmbH, Germany) according to the manufacturer's instructions. Antimicrobial susceptibility to penicillin, ampicillin, tetracycline, chloramphenicol, erythromycin, and clindamycin was tested by MicroScan semiautomated system (Siemens, Germany), and the results were interpreted according to EUCAST criteria (15). GAS isolates were examined for the presence of a mucoid phenotype by culture visualization and categorized as either mucoid or nonmucoid.

***emm* typing and SAg genes.** Genomic DNA from 70 GAS strains were extracted from pure cell bacterial culture by using a manual DNeasy Blood & Tissue kit (Qiagen, Basel, Switzerland) according to the manufacturer's protocol. PCR amplification of the *emm* gene was performed as previously described (16). In order to assign the specific *emm* type and subtype, the first 240 nucleotides of each sequence were compared to the *S. pyogenes emm* database available at the CDC website (<http://www.cdc.gov/streplab/index.html>). SAg genes were analyzed by multiplex PCR assays, as previously described (17). The exotoxin genes *speB* and *speF* were used as PCR internal controls. The presence of the SAg genes (*speA*, *speC*, *speG*, *speH*, *speI*, *speJ*, *speK*, *speL*, *speM*, *ssa*, and *smeZ*) was confirmed by single PCRs (Table 1).

**MLST.** To determine the genetic relationship between the 11 *S. pyogenes* isolates belonging to the serotype M-18, multilocus sequence typing (MLST) based on seven housekeeping genes (*gki*, *gtr*, *murI*, *mutS*, *recP*, *xpt*, and *yiqL*) was performed (18). The allele numbers and relative sequence types were assigned by using the *S. pyogenes* MLST database (<http://spyogenes.mlst.net>).

**MALDI-TOF MS sample preparation and analysis.** Sample preparation for MALDI-TOF MS was performed as previously described with minor modifications (8). Briefly, colonies of fresh overnight culture derived from 49 GAS isolates were resuspended at 1 McFarland, and 1 ml of bacterial suspension was centrifuged at 5,000 × g for 5 min. Pellets were suspended with 300 μl of distilled water and 900 μl of absolute ethanol and pelleted again. The supernatants were then discharged, and cells were suspended in 20 μl of formic acid (70%) and 20 μl of acetonitrile. A whole-cell suspension was centrifuged at 12,000 × g for 5 min, and 1-μl portions of the supernatants were placed on a MALDI sample slide (Bruker-Daltonics, Bremen, Germany) and dried at room temperature. The sample was then overlaid with 1 μl of matrix solution (α-cyano-4-

hydroxycinnamic acid in 50% acetonitrile and 2.5% trifluoroacetic acid) and dried at room temperature. A MALDI-TOF MS measurement was performed with a Bruker MicroFlex MALDI-TOF MS (Bruker-Daltonics) using FlexControl software and a DH5- $\alpha$  *Escherichia coli* protein extract (Bruker-Daltonics) was deposited on the calibration spot of the sample slide for external calibration. Spectra collected in the positive-ion mode within a mass range of 2,000 to 20,000 Da were analyzed using a Bruker Biotyper (Bruker-Daltonics) automation control and the Bruker Biotyper 3.1 software and library (a database with 5,627 entries). The clustering analysis of the GAS strains was performed by generation of the dendrogram based on the different serotypes collected in the present study. In detail, 49 strains representing 11 different serotypes included: M-1 ( $n = 4$ ), M-3 ( $n = 2$ ), M-5 ( $n = 6$ ), M-6 ( $n = 9$ ), M-9 ( $n = 1$ ), M-12 ( $n = 1$ ), M-18 ( $n = 11$ ), M-28 ( $n = 9$ ), M-44 ( $n = 2$ ), M-89 ( $n = 3$ ), and M-102 ( $n = 1$ ). The main spectra (MSPs) of each strain were generated from 10 technical replicates prior to manual visualization inspections using Flex-Analysis 3.4 software. The relationship between MSPs obtained from each strain was visualized in a score-oriented dendrogram using the average linkage algorithm implemented in the MALDI Biotyper 3.1 software.

**Whole-genome sequencing and comparative genomics.** Whole-genome sequencing was conducted on the two *S. pyogenes* isolates from the ARF case and nine M-18 GAS strains included in the present study (see Table S1 in the supplemental materials). Libraries were prepared with a Nextera XT sample preparation kit (Illumina), and sequencing was performed on the Illumina MiSeq platform (Illumina, San Diego, CA) with a 2 $\times$ 250 paired-end run. All read sets were evaluated for sequence quality and read-pair length using FastQC (19) and then assembled with MIRA 4.0 using a *de novo* assembly mode (20).

For comparative studies, SNPs were identified with an in-house Perl pipeline based on the Mauve software (21). In this approach, the genome of *S. pyogenes*, group A strain SP665Q, was used as a reference, and the other 11 genome assemblies included in the present study were aligned against it. All of the alignments were then merged, and the coordinates of all nucleotide variations were detected on the basis of the annotated reference strain assembly (SP665Q). All of the variations were then organized in a matrix to assess the presence or absence pattern in specific subsets of strains (e.g., *emm18.29* and *emm18.32*). The core SNPs, defined as the nondegenerate SNPs present in all 12 genomes and flanked by conserved positions, were extracted and finally subjected to Bayesian phylogenetic analysis with MrBayes (22). The Bayesian analysis was run on the GTR substitution model for 2,000,000 generations with a chain sampled every 1,000th generation. The final parameter values and trees are summarized after 25% of the posterior sample was discarded. The Bayesian tree is displayed and edited using FigTree v1.4.0 (<http://tree.bio.ed.ac.uk/software/figtree>).

Informative SNPs (i.e., present in at least two strains) were extracted from core SNPs using an in-house script. In addition, among the strains with *emm* type 18.29, all genes presenting at least one core SNP were selected and compared to the virulence factors of pathogenic bacteria (VFPP) database (23), using a blast search with a 10<sup>-5</sup> value cutoff.

Prophages were detected and analyzed using the free web tool PHAST (PHAge Search Tool) (24). The comparative sequence circular maps of whole genomes and concatenated prophages of each strain of *S. pyogenes* M18GAS were generated using BRIG (25).

**Accession numbers.** The sequences of the 11 M-18 GAS genomes were deposited at EMBL/EBI under the following accession numbers: M18GASBO1065 (CDGV01000001 to CDGV01000182), M18GASBO665 (CDGO01000001 to CDGO01000182), M18GASBO9 (CDGY01000001 to CDGY01000207), M18GASBO8 (CDGW01000001 to CDGW01000183), M18GASBO7 (CDGX01000001 to CDGX01000199), M18GASBO6 (CDGM01000001 to CDGM01000081), M18GASBO5 (CDGQ01000001 to CDGQ01000096), M18GASBO4 (CDGN01000001 to CDGN01000158), M18GASBO3 (CDGS01000001 to CDGS01000294), M18GASBO2 (CDHA01000001 to CDHA01000222), and M18GASBO1 (CDHB01000001 to CDHB01000235) (study project PRJEB7108).

## RESULTS

**Characterization of GAS isolates.** Eight ARF cases were recorded in the Bologna province between November 2012 and May 2013. All patients were aged between 4 and 12 years. In recent years, the number of ARF cases in this province has ranged from one to four per year. At time of diagnosis, seven of eight ARF cases were culture negative, and only one GAS strain was isolated from a patient resident in Bologna province.

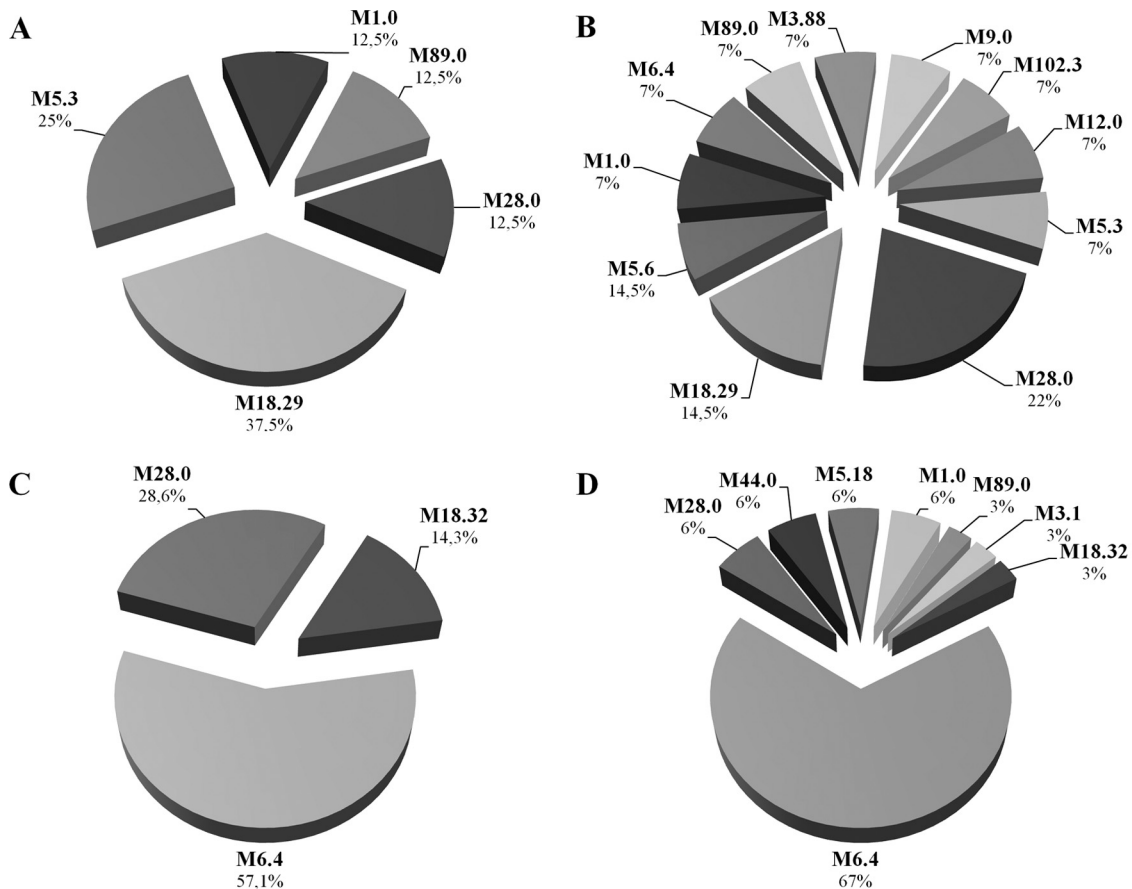
During the surveillance period, 70 GAS isolates were collected (see Table S1 in the supplemental material). In detail, two consecutive isolates were collected from the first patient with ARF (i.e., SPBO1 and SPBO2), six were isolated from classmates of the first case and 14 were isolated from symptomatic children resident in the Bologna area at the time of ARF diagnosis. Two months later, a second collection comprised 14 GAS isolates collected from classmates of the second ARF case and 34 GAS isolates obtained from symptomatic school-age children resident in the same area. However, no GAS was isolated from the second case of ARF.

Analysis of the *emm* sequence revealed that the GAS isolates obtained from patient with ARF was *emm18.29*, as two of six of the GAS isolates were obtained from contacts (Fig. 1A). At the same time, GAS isolated from symptomatic children of the community were: *emm28.0* (3 cases), *emm18.29* (2 cases), *emm89.0*, *emm1.0*, *emm56.0*, *emm9.0*, *emm6.4*, *emm102.3*, *emm12.0*, *emm5.3*, and *emm3.8* (one case each), as shown in Fig. 1B.

Molecular investigation conducted among the 14 GAS collected from the classmates of the second ARF case showed that two isolates (14.3%) were *emm18.32*, four (28.5%) were *emm28.0*, and eight (57.2%) were *emm6.4*, as shown Fig. 1C. Among the 35 GAS isolates collected from symptomatic school-age children in the community, eight different *emm* types were identified: *emm6.4* (23 cases), *emm28.0* (2 cases), *emm44.0* (3 cases), *emm1.0* (2 cases), and *emm5.3*, *emm18.32*, *emm89.0*, *emm5.1*, and *emm3.1* (one case each), as shown in Fig. 1D.

Our data indicate that *emm18.29* was the dominant serotype among the isolates collected from the class of the first case, whereas *emm18.32* spread in the class of the second case. To investigate the relationship between the specific *emm* type and the SAG profile, GAS isolates were evaluated by molecular analysis for the different GAS genes (see Table S1 in the supplemental material). Molecular analysis revealed eight different antigenic profiles among 70 GAS isolates. The chromosomally encoded *speB*, *speG*, and *smeZ* genes were present in all isolates. In addition, all *emm18* strains, including *emm18.29* and *emm18.32* and the two *emm18.0* strains derived from the ISS bank collection (SP665Q and SP1065Q), presented a characteristic SAG profile showing *speA*, *speC*, *speL*, and *speM* genes. Overall, the isolates with the same *emm* type shared a common SAG profile, with the exception of the *emm3* isolates. MLST analysis showed that all *emm18* GAS isolates belonged to sequence type 42.

To determine the association of the mucoid phenotype with *emm* type, all GAS isolates were analyzed by culture visualization. Among GAS isolates, the M-102, M-9, and M-18 GAS strains showed the highest incidence (100, 100, and 87.5%, respectively) of mucoid traits, whereas all *emm1* isolates showed a nonmucoid colony type (see Table S1 in the supplemental material). Antimicrobial susceptibility testing showed that all isolates were susceptible to penicillin, ampicillin, clindamycin, chloramphenicol, tetracycline, and clindamycin, whereas only two GAS strains



**FIG 1** Distribution of *emm* types among *S. pyogenes* strains collected during an ARF outbreak occurred in Bologna province. (A) GAS strains isolated from the ARF case and class contacts during the first episode. (B) GAS isolated from class children during the second ARF episode. (C) Distribution of *emm* types among GAS isolates collected from the community at the time of the first ARF case. (D) Distribution of *emm* types among GAS collected from school-age children of the community at time of the ARF episode.

belonging to the *emm89.0* and *emm5.3* serotypes were resistant to erythromycin.

**MALDI-TOF and clustering analysis.** MALDI-TOF MS analysis of different *S. pyogenes* strains showed that 37 of 50 (74.0%) isolates clustered in accordance with the serotype group, as shown in Fig. 2. In detail, the main spectra generated by MALDI-TOF MS analysis demonstrated a high overall discriminatory power of the strains belonging to the serotypes M-18, M-28, and M-3. However, different clustering groups were observed for M-1, M-5, M-6, M-89, and M-44 strains by showing a different protein mass spectral profiling among isolates belonging to the same serotype (Fig. 2). Notably, the score-oriented dendrogram showed that all isolates belonging to serotype M-18 formed a separate clustering group that was clearly distinguishable from other serotypes. In addition, a different cluster grouping within the M-18 serotype was observed between *emm18.0* and the *emm18.32/emm18.29* subtypes, with a critical distance of 500 (see Fig. 2), respectively corresponding to the *S. pyogenes* strains isolated 30 years ago and the isolates involved in the ARF outbreak in Bologna province during 2013. However, MALDI-TOF MS analysis was not able to distinguish among *emm18.29* and *emm18.32* subtypes.

**Whole-genome sequencing and phylogenetic analysis of M-18 GAS isolates.** The draft genomes of *emm18.29*, *emm18.32*, and *emm18.0* GAS isolates were assembled into average 185 con-

tigs with a G+C content of 38.6% for a total of 1,929,545 bp (Fig. 3). Genome annotations predicted a total of 1,881 open reading frames.

To investigate the relationship among M-18 GAS isolates, a whole-genome analysis was performed on the basis of core SNPs identified with the Mauve-based approach (593 core SNPs). Phylogenetic analysis indicated that two main clusters were highlighted with high posterior probabilities: the first cluster included the *emm18.29* strains, while the second included the *emm18.32* strains. In addition, the two *emm18.0* GAS isolates derived from the ISS bank collection (SP665Q and SP1065Q) showed a closely relationship to the GAS8232 strain. Phylogenetic analysis indicated that *emm18.0* strains and GAS8232 differed by 198 informative SNPs and were nearer to *emm18.32* strains than to *emm18.29* strains (Fig. 4). Comparison of SNPs between M-18 strains isolated during the ARF outbreak in Bologna province showed that the two consecutive GAS isolates from the ARF case (SPBO1 and SPBO2) differed by 14 SNPs (0 informative SNPs) and were closely related to strains collected from classmates (SPBO3 and SPBO4) and from the community (SPBO5 and SPBO6), as shown in Fig. 4. Comparison of *emm18.29* genomes disclosed 216 different SNPs and 5 informative SNPs in this group. In addition, all of the *emm18.32* GAS strains collected from children during the sec-

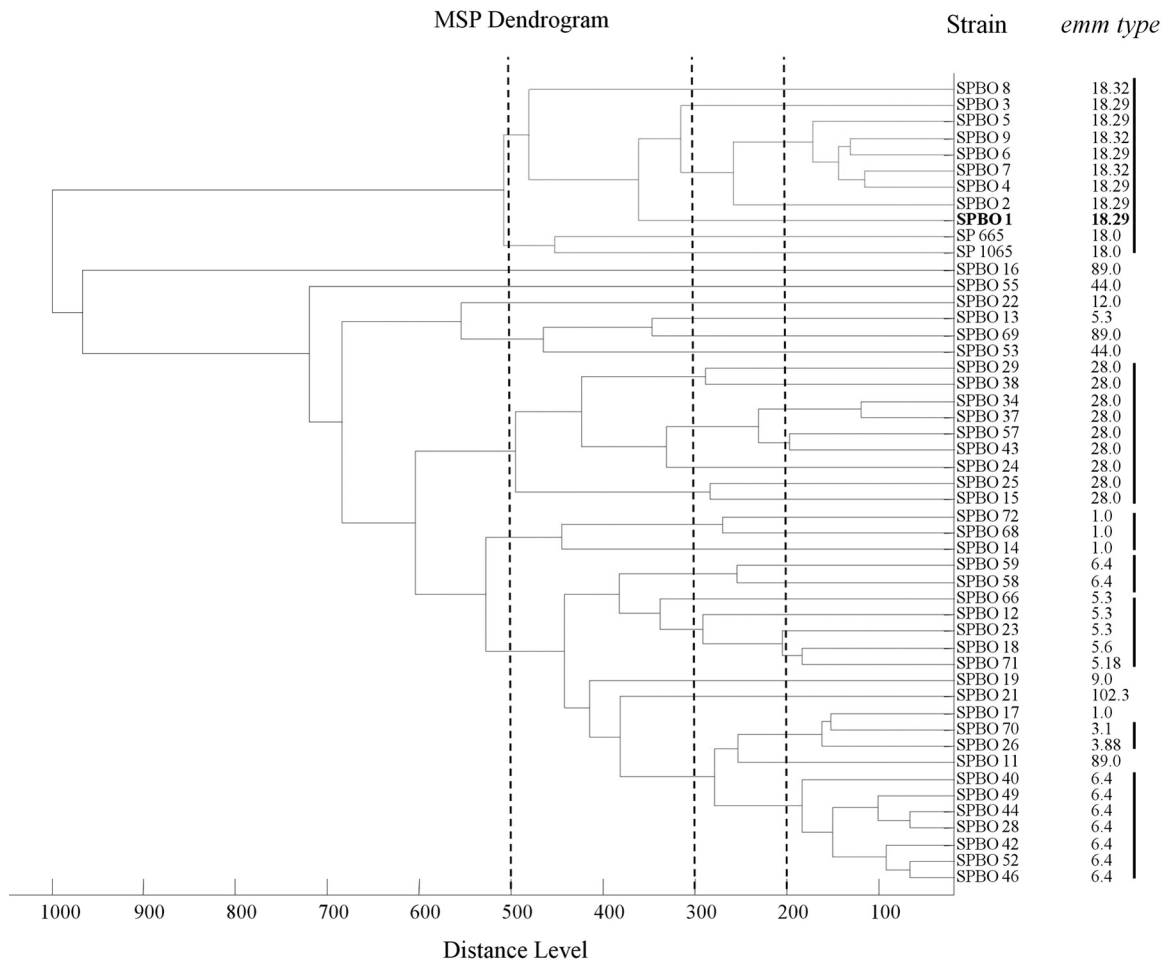


FIG 2 Score-oriented dendrogram based on the main spectra (MSP) of 50 GAS strains obtained by Bruker MALDI-TOF MS and analyzed using Biotyper 3.1 software. Correlation with the 11 different *emm* types (M-1, M-3, M-5, M-6, M-9, M-12, M-18, M-28, M-44, M-89, and M-102) shown. Dotted lines define a similarity cutoff value of 500, 300, and 200 used for clustering groups of serotypes M-18, M-1, M-3, M-5, M-28, and M-6.

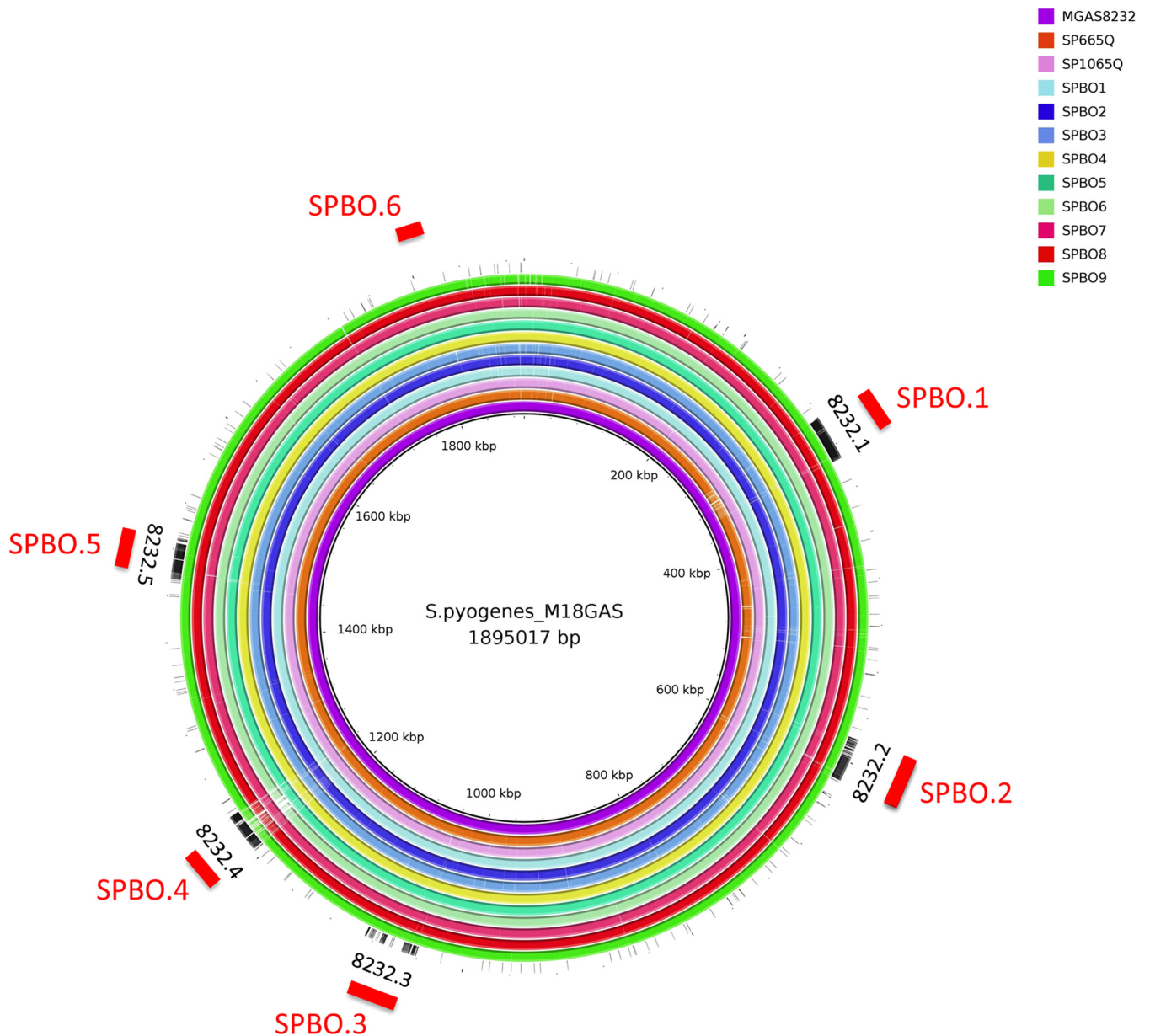
ond ARF episode were closely related (53 different SNPs and 0 informative SNPs) (Fig. 4).

Further comparison of all M-18 genomes identified 73 and 207 SNPs exclusive to *emm18.29* and *emm18.32* clusters, respectively (see Table S2 in the supplemental material). Analysis of all unique SNPs of *emm18.29* cluster showed that 36.9% ( $n = 27$ ) were synonymous, 58.9% ( $n = 43$ ) were nonsynonymous, and 4.2% ( $n = 3$ ) were located in intergenic regions. At the same time, the SNPs of *emm18.32* cluster were 41.6% ( $n = 86$ ) synonymous, 50.2% ( $n = 104$ ) nonsynonymous, and 8.2% ( $n = 17$ ) within intergenic regions. Analysis of synonymous and nonsynonymous substitutions in the coding regions of the core genome revealed that SNPs were associated with different GAS virulent factors (see Table S2 in the supplemental material) present in the VFPB database (23). Notably, 46 SNPs substitutions occurred in prophage elements, including *hyaluronidase* and *gp58*-like genes in both *emm18.29* and *emm18.32* clusters (see Table S2 in the supplemental material). In detail, five SNPs in the *hyaluronidase* (*hylP*) genes among the two cluster subtypes were synonymous and six were nonsynonymous, most of them located in the N-terminal region of *hylP* gene.

#### Comparison of phage elements in M-18 GAS strains. Analysis

of M-18 GAS chromosomes revealed that six regions contained prophage elements ( $\Phi$ SPBO.1,  $\Phi$ SPBO.2,  $\Phi$ SPBO.3,  $\Phi$ SPBO.4,  $\Phi$ SPBO.5, and  $\Phi$ SPBO.6) ranging from 7.6 to 75.8 kb (Table 2). The genome distribution of the prophages across the GAS chromosome showed that all M-18 strains shared a common localization of these elements, as shown in Fig. 3. Comparison to other GAS genomes revealed that five prophage elements ( $\Phi$ SPBO.1,  $\Phi$ SPBO.2,  $\Phi$ SPBO.3,  $\Phi$ SPBO.4, and  $\Phi$ SPBO.5) showed similar chromosome locations to that of the GAS8232 genome (11), suggesting conservative site integrations of these regions across M-18 GAS strains (Fig. 3). Moreover, examination of prophage elements showed a similar genetic architecture with GAS8232 (M-18) and MGAS315 (M-3) strains, as shown in Fig. 5 (11, 26).

Genomes of M-18 GAS strains contained prophage regions harboring several virulence factors, including exotoxin type A (SpeA), exotoxin type C (SpeC), exotoxin type L (SpeL), exotoxin type M (SpeM), mitogenic factor (DNase), and streptodornase (Sdn) (Table 2). Comparison of the prophage elements showed that the  $\Phi$ SPBO.2,  $\Phi$ SPBO.3, and  $\Phi$ SPBO.5 regions were shared among all M-18 GAS strains. These three phage regions contained different virulence factors such as genes encoding SpeC, mitogenic factor, SpeL/SpeM, *hyaluronidase*, and streptodornase. In-



**FIG 3** Comparison of the group A *Streptococcus* serotype M18 chromosomes. The circular representation shows a genome comparison from center to periphery, respectively, of strains GAS8232, SPBO1, SPBO2, SPBO3, SPBO4, SPBO5, SPBO6, SPBO7, SPBO8, SPBO9, SP665Q, and SP1065Q (see the legend for the color associations). The regions of differences within GAS genomes are indicated with white gaps. The genomic localizations of the prophage elements shared with GAS8232 are indicated as black boxes outside the circular GAS chromosome maps. The locations of the prophage regions ( $\Phi$ SPBO.1,  $\Phi$ SPBO.2,  $\Phi$ SPBO.3,  $\Phi$ SPBO.4,  $\Phi$ SPBO.5, and  $\Phi$ SPBO.6) of M-18 GAS isolates collected from Bologna province are indicated as red boxes.

terestingly, closely association between phages encoded SpeC and SpeL/speM were observed in seven M-18 GAS strains (Table 2).

Moreover, the  $\Phi$ SPBO.1 region was present in 10 of 11 of the M-18 GAS strains, lacking in the SP665Q isolate (*emm18.0* subtype). Analysis of virulence factors showed that  $\Phi$ SPBO.1 region has variants of the *speA* gene (*speA1*) in six of nine *emm18.29* and *emm18.32* strains but was absent in *emm18.0* strains (Table 2). In addition, phage  $\Phi$ SPBO.4 was absent in *emm18.32* strains, whereas it was present in all *emm18.0* and *emm18.29* strains. This region contained a gene encoding mitogenic factor. Interestingly, each phage region had a hyaluron-

idase gene (Table 2). Similar findings were observed in a previous study demonstrating that all phages in M-3 GAS strains contained a hyaluronidase gene (26).

**Comparison of phage-encoded hyaluronidase (*hylP.2*) gene among M-18 GAS strains.** To investigate the genetic variability in the *hylP.2* gene, the gene-encoded HylP.2 protein of M-18 strains was compared to previously described GAS isolates. Our analysis indicated that the *hylP.2* gene was located in phage  $\Phi$ SPBO.1 in all *emm18.29*, *emm18.32*, and *emm18.0* GAS strains (Table 2). Analysis of the *hylP.2* gene derived from the *emm18.29*, *emm18.32*, and *emm18.0* strains showed a high rate of homology with the M18



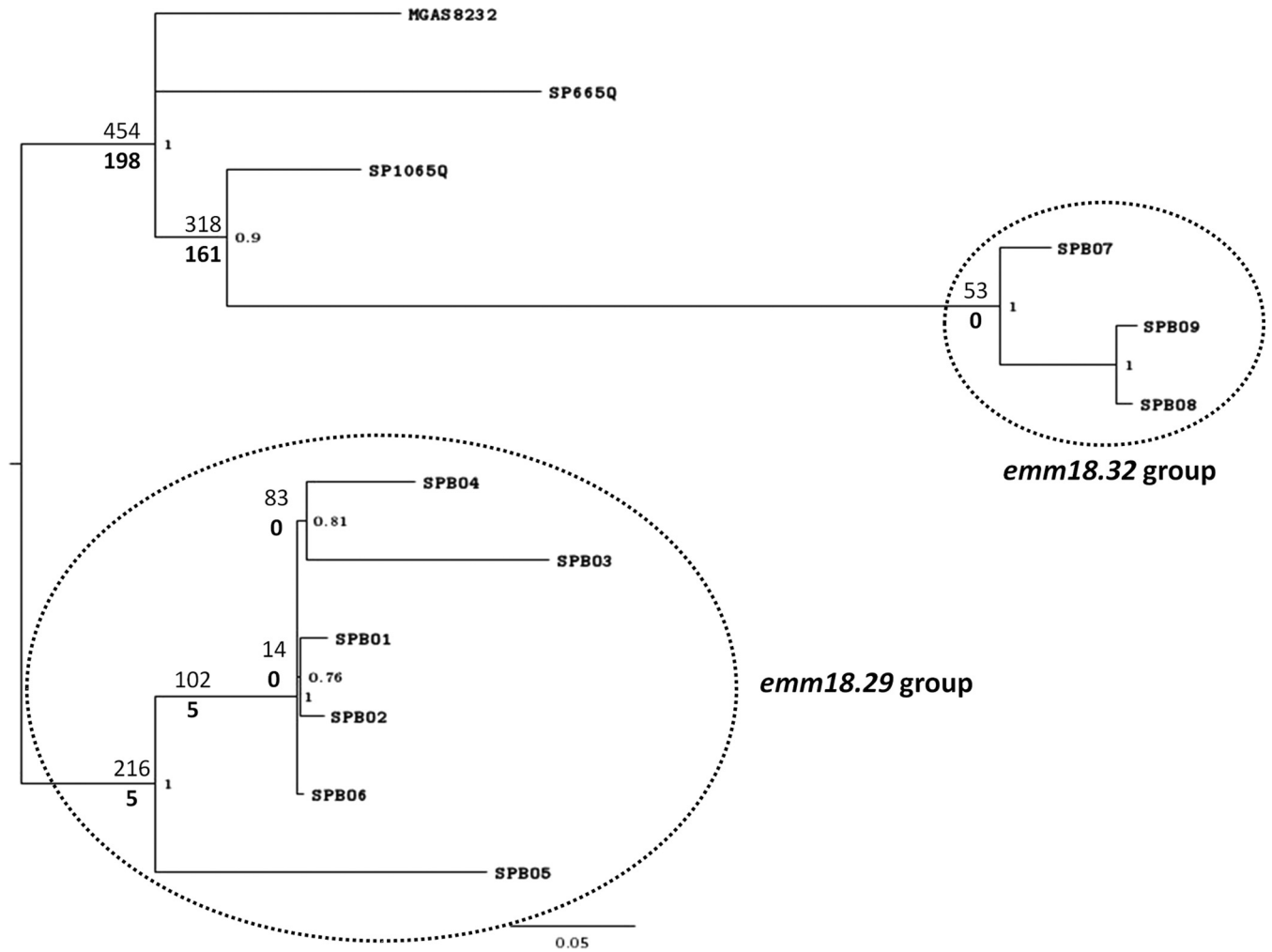


FIG 4 Bayesian tree based on core SNPs identified with the Mauve-based approach. In each node of the tree, posterior probabilities (>0.7) are indicated on the right of the node, while the numbers of different and informative SNPs located, respectively, up and down the branch are on the left of the node.

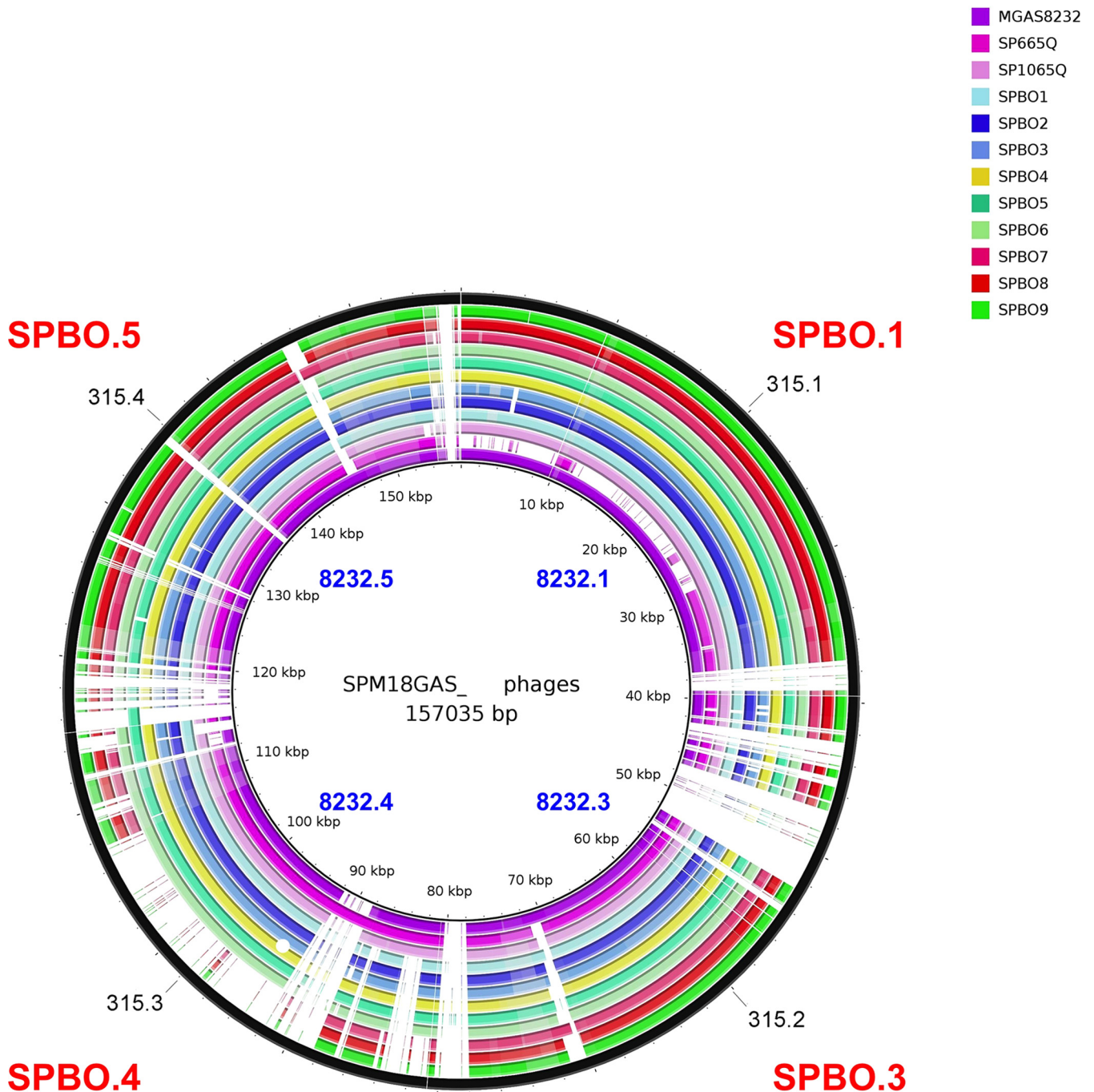
(MGAS8232) and M5 (Manfredo) GAS strains, two strains isolated from ARF cases (11, 27). Comparison of deduced amino acid sequences with M5-GAS Manfredo strain demonstrated that HylP.2 were C terminally truncated in all M-18 GAS isolates (including *emm18.29*, *emm18.32*, *emm18.0*, and GAS8232 strains) (Fig. 6). Of note, three *emm18.29* isolates (SPB01, SPB03, and

SPB06) showed an internal deletion located between the N-terminal and the TSBH domains. At the same time, the deduced amino acid sequence of *hylP.2* gene in the SPB08 isolate was truncated in the TSBH region (Fig. 6). Comparison analysis conducted on the *hylP.2* gene revealed different clustering groups according to the corresponding subtypes (data not shown).

TABLE 2 Prophage elements in M-18 GAS strains

Phage	Prophage element size (kb) in M-18 GAS strain <sup>a</sup>											Virulence factor(s)
	<i>emm18.29</i>						<i>emm18.32</i>			<i>emm18.0</i>		
	SPB01	SPB02	SPB03	SPB04	SPB05	SPB06	SPB07	SPB08	SPB09	SP1065Q	SP665Q	
ΦSPBO.1	56.4*	67.2	52.5	67.4*	64.9*	61.2*	68.7	75.8*	57.6*	-	66.1	<i>speA</i> , hyaluronidase ( <i>hylP.2</i> )
ΦSPBO.2	32.5	38.5	36.9	39.5†	38.3†	37.2†	38	38.6†	37.3†	36.4†	39.6†	<i>speC</i> mitogenic factor, hyaluronidase
ΦSPBO.3	47.8	68.5	59.4	57.‡	58.4‡	59.4‡	57.3	57.4‡	59.1‡	57.6‡	67.1‡	<i>speL</i> , <i>speM</i> , hyaluronidase
ΦSPBO.4	33	31.8	30.6	25.2	30.6	32.5				43.6	41.1	Hyaluronidase, mitogenic factor
ΦSPBO.5	60	47.3	39.3	45.7	47.2	51	48.5	42.6	48	51.3	46.6	Hyaluronidase, streptodornase
ΦSPBO.6	14.1	12.8	12.5	19.5	47.2	18.4	17		10.5		7.6	Hyaluronidase

<sup>a</sup> \*, Prophage-containing variant of *speA* (*speA1*); †, prophage-containing *speC*; ‡, prophage-containing *speL* and *speM*.



**FIG 5** Circular representation of concatenated prophage elements integrated in the genome of the M-18 GAS strains compared to M-3 GAS prophages. The outermost black circle represent the concatenated M-3 GAS prophages ( $\Phi$ 315.1,  $\Phi$ 315.2,  $\Phi$ 315.3, and  $\Phi$ 315.4). The prophages of M-18 GAS strains collected in the present study (SPBO1, SPBO2, SPBO3, SPBO4, SPBO5, SPBO6, SPBO7, SPBO8, SPBO9, SP1065Q, and SP665Q) and reference strain (M18GAS8232) are indicate in red and blue, respectively. The areas of similarity and divergence are contrasted with white gapped areas indicating regions of highest variance.

## DISCUSSION

Since late 2012, an outbreak of acute rheumatic fever (ARF) was observed in the Bologna province, Northeast Italy, where the annual frequency of ARF in resident children has ranged from 1 to 4 cases per year. From November 2012 to May 2013, eight cases of ARF were recorded, showing a significant increase of ARF in this area. Molecular analysis conducted among GAS collected from

both contacts and the community during the outbreak indicated that M-18 represented one of the most prevalent serotype within classes of two unrelated ARF episodes. Our findings showed that two distinct subtypes, i.e., *emm18.29* and *emm18.32*, were observed to spread separately across the two classes of ARF cases. Analysis of GAS isolates from children in the community showed that the majority of the isolates were subtype *emm6.4*. Phenotypic

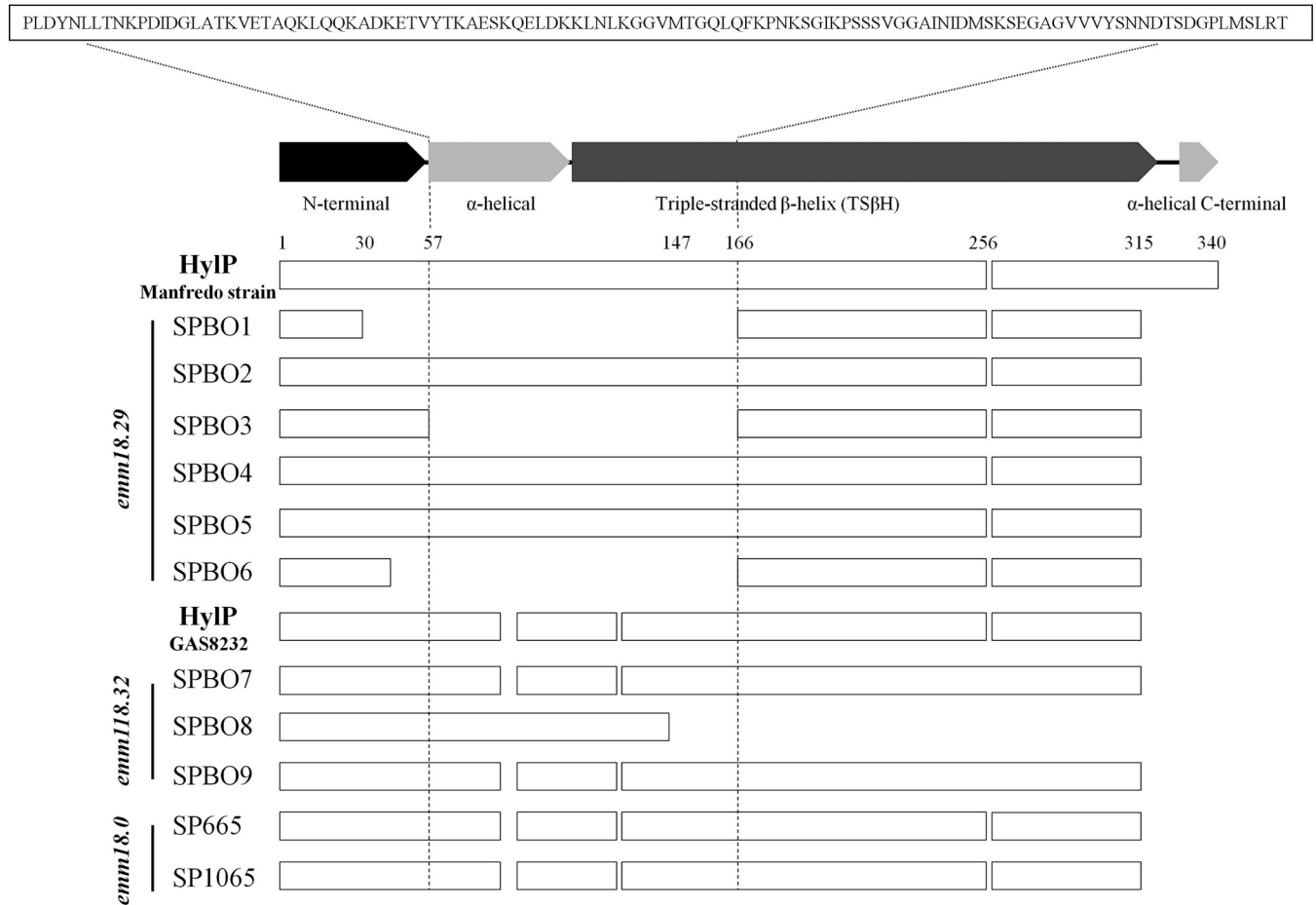


FIG 6 Alignment of the phage-encoded hyaluronidase gene (*hylP.2*) derived from M-18 GAS (*emm18.29*, *emm18.32*, and *emm18.0* subtypes). The *hylP.2* sequences from ARF-related M-18 (GAS8232) and M-5 (Manfredo) strains are shown. Dotted lines show the common deleted regions between α-helical and TSβH domains between SPBO1, SPBO3, and SPBO6 strains. The deduced amino acid sequence of a 327-nucleotide deletion in *emm18.29* strains is shown.

analysis of GAS showed that the M-18 GAS strains presented a high frequency of mucoid strains, confirming previous findings (5). Indeed, GAS mucoid isolates have been observed to correlate with invasive infections and the pathogenesis of rheumatic fever (1). Previous studies clearly demonstrated that higher capsule production has been observed in M-18 strains responsible for multiple ARF outbreaks (11). Although our study showed that GAS M-18 was responsible of an ARF case and that this serotype has spread in the two classes, we cannot exclude the possibility that others strains (i.e., serotypes) have been responsible of others ARF cases.

The present study evaluated the ability of MALDI-TOF MS to correctly identify GAS and to distinguish among the different serotypes. MALDI-TOF analysis demonstrated an excellent discrimination of M-18 GAS strains by showing a separate cluster. Overall, cluster analysis by MALDI-TOF showed a good concordance (74%) with *emm* typing methods. In detail, a 100% concordance was observed with serotypes M-28 (9 of 9) and M-3 (2 of 2), whereas 75 and 83.4% concordances were found with M-1 (3 of 4) and M-5 (5 of 6), respectively. The discrepancies observed within these serotypes were attributable to two isolates, SPBO13 (*emm5.3*) and SPBO17 (*emm1.0*), grouped separately from strains belonging to the same serotypes. These results are in accordance

with a previous study demonstrating the potential of the MALDI-TOF Biotyper system for GAS clustering analysis, thus showing a high discriminatory power among different serotypes (9). However, a poor discriminatory classification was observed for M-89 serotype. Based on these findings and for its rapidity and low cost analysis, we suggest the MALDI-TOF technique could be used successfully for the identification of the M-18 serotype among GAS strains.

In recent years, whole-genome sequencing has been applied for epidemiological purposes by showing a more accurate resolution than classical genotypic methods (13). Whole-genome sequencing has been extensively used to explore the genetic organization of bacterial genomes and to compare the rearrangements between closely related strains (10, 12, 28). The present study described the complete genomes of nine M-18 GAS strains isolated during an ARF outbreak in northeastern part of Italy and compared them to two M-18 *S. pyogenes* collected from noninvasive infections in the same area 30 years ago. SNP phylogenetic analysis revealed that *emm18.29* and *emm18.32* subtypes segregated in two separate clusters, whereas *emm18.0* GAS strains did not cluster in a distinct group. Genome polymorphism analysis showed that isolates from ARF cases and from community and class contacts were closely

related and showed a low number of informative SNPs both in *emm18.29* and *emm18.32* strains.

Our analysis revealed that M-18 strains possessed several virulence genes, including *speA*, *speC*, *speL*, *speM*, *smeZ*, *mitogenic factor*, and *hyaluronidase*, most of them located in prophage elements. Integrated prophages represent one of the most divergent tracts among GAS genomes and the majority of genetic variations among M-18 GAS strains (29).

Our results indicated that prophage elements are located in the same genomic locations among different M-18 strains collected in the present study and in the MGAS8232 reference strain. In addition, genomic organization revealed that three prophage elements ( $\Phi$ SPBO.2,  $\Phi$ SPBO.3, and  $\Phi$ SPBO.5) were common among M-18 GAS strains. Alignment of the prophage sequences showed a similar architecture between subtypes *emm18.29*, *emm18.32*, and *emm18.0*, thus revealing a similar genomic architecture of M-18 GAS strains. It has been established that hypervirulent GAS strains acquire virulence factors via prophage integrations (29). Recently, Bao et al. reported that prophage integrations represent one of the multiple genetic factors related to the pathogenic role of the M-23ND GAS strain (30). Our findings showed that three prophages ( $\Phi$ SPBO.1,  $\Phi$ SPBO.3, and  $\Phi$ SPBO.5) present in the M-18 GAS strains were similar with phages of M-3 serotypes ( $\Phi$ 315.1,  $\Phi$ 315.2, and  $\Phi$ 315.4) that have been previously associated with the emergence of virulent M-3 subclones (26) by different sequential acquisition.

We showed that the streptococcal pyrogenic exotoxin A gene (*speA1*) was located in the prophage  $\Phi$ SPBO.1 region in six of nine *emm18.29* and *emm18.32* strains, whereas this gene was absent in *emm18.0* strains. Also, we observed that the *speC* was present in all M-18 GAS strains possessing *speL* and *speM*, thus showing a strict correlation between these prophage-encoded SAg genes. Based on these findings, we hypothesize that a different combination of phage-encoded virulence factors could be related to the virulence of *emm18.29* and *emm18.32* strains isolated during focal ARF that differed from noninvasive *emm18.0* strains collected from the same area 30 years ago.

Analysis of the phage-encoded virulence factors demonstrated that phage-encoded hyaluronidase showed a higher number of synonymous and nonsynonymous substitutions than other genes within the M-18 GAS genomes. Previous studies reported that *hylP* and *hylP.2* genes were present with different alleles among different serotypes from both invasive and noninvasive GAS isolates (31, 32). However, M-18 GAS strains have been observed to possess a unique *hylP.2* gene structure among different isolates (33). Our findings demonstrated that the *hylP.2* gene possesses an internal deletion located between the N-terminal and TSBH regions in different M-18 GAS strains. Therefore, the truncated gene structures observed in several M-18 GAS isolates could be related to a different or nonfunctional activity of the HylP.2 protein. Previous study showed that inactivation in hyaluronate lyase (HylA) restored full encapsulation in partially encapsulated M-4 GAS strains, thus demonstrating the mutually exclusive interaction between the hyaluronan capsule and active hyaluronidase (32). In addition, Schommer et al. demonstrated in a mouse model that the difference in capsule size was regulated by bacterial hyaluronidase and that the high molecular mass of the hyaluronan capsule influences GAS virulence by facilitating deep tissue infections (34). Based on our findings, we hypothesize that inactivation of HylP.2 could determine a different encapsulation (i.e., capsule

sizing) of M-18 GAS strains, thus resulting in a more virulent clone. Therefore, the internal deletion in the *hylP.2* gene observed in different isolates could reflect a different virulence potential among the M-18 GAS strains.

In conclusion, we investigated the molecular and epidemiological linkage between GAS strains isolated during an ARF outbreak in Bologna province in early 2013. Our study explored the genome sequence of M-18 GAS strains, thus providing a better understanding of the genetic architecture of the M-18 serotype and expanding our knowledge of the genetic elements related to the GAS infections.

## ACKNOWLEDGMENT

This study was supported by funds from the Emilia-Romagna region.

## REFERENCES

- Cunningham MW. 2014. Rheumatic fever, autoimmunity, and molecular mimicry: the streptococcal connection. *Int Rev Immunol* 33:314–329. <http://dx.doi.org/10.3109/08830185.2014.917411>.
- Walker MJ, Barnett TC, McArthur JD, Cole JN, Gillen CM, Henningham A, Sriprakash KS, Sanderson-Smith ML, Nizet V. 2014. Disease manifestations and pathogenic mechanisms of group A *Streptococcus*. *Clin Microbiol Rev* 27:264–301. <http://dx.doi.org/10.1128/CMR.00101-13>.
- Tibazarwa KB, Volmink JA, Mayosi BM. 2008. Incidence of acute rheumatic fever in the world: a systematic review of population-based studies. *Heart* 94:1534–1540. <http://dx.doi.org/10.1136/hrt.2007.141309>.
- Wolfe RR. 2000. Incidence of acute rheumatic fever: a persistent dilemma. *Pediatrics* 105:1375. <http://dx.doi.org/10.1542/peds.105.6.1375>.
- Steer AC, Law I, Matatolu L, Beall BW, Carapetis JR. 2009. Global *emm* type distribution of group A streptococci: systematic review and implications for vaccine development. *Lancet Infect Dis* 9:611–616. [http://dx.doi.org/10.1016/S1473-3099\(09\)70178-1](http://dx.doi.org/10.1016/S1473-3099(09)70178-1).
- Smoot JC, Korgenski EK, Daly JA, Veasy LG, Musser JM. 2002. Molecular analysis of group A *Streptococcus* type *emm18* isolates temporally associated with acute rheumatic fever outbreaks in Salt Lake City, Utah. *J Clin Microbiol* 40:1805–1810. <http://dx.doi.org/10.1128/JCM.40.5.1805-1810.2002>.
- Sauer S, Kliem M. 2010. Mass spectrometry tools for the classification and identification of bacteria. *Nat Rev Microbiol* 8:74–82. <http://dx.doi.org/10.1038/nrmicro2243>.
- Mencacci A, Monari C, Leli C, Merlini L, De Carolis E, Vella A, Cacioni M, Buzi S, Nardelli E, Bistoni F, Sanguinetti M, Vecchiarelli A. 2013. Typing of nosocomial outbreaks of *Acinetobacter baumannii* by use of matrix-assisted laser desorption/ionization-time-of-flight mass spectrometry. *J Clin Microbiol* 51:603–606. <http://dx.doi.org/10.1128/JCM.01811-12>.
- Wang J, Zhou N, Xu B, Hao H, Kang L, Zheng Y, Jiang Y, Jiang H. 2012. Identification and cluster analysis of *Streptococcus pyogenes* by MALDI-TOF mass spectrometry. *PLoS One* 7:e47152. <http://dx.doi.org/10.1371/journal.pone.0047152>.
- Sassera D, Comandatore F, Gaibani P, D'Auria G, Mariconti M, Landini MP, Sambri V, Marone A. 2014. Comparative genomics of closely related strains of *Klebsiella pneumoniae* reveals genes possibly involved in colistin resistance. *Ann Microbiol* 64:887–890. <http://dx.doi.org/10.1007/s13213-013-0727-5>.
- Smoot JC, Barbian KD, Van Gompel JJ, Smoot LM, Chaussee MS, Sylva GL, Sturdevant DE, Ricklefs SM, Porcella SF, Parkins LD, Beres SB, Campbell DS, Smith TM, Zhang Q, Kapur V, Daly JA, Veasy LG, Musser JM. 2002. Genome sequence and comparative microarray analysis of serotype M18 group A *Streptococcus* strains associated with acute rheumatic fever outbreaks. *Proc Natl Acad Sci U S A* 99:4668–4673. <http://dx.doi.org/10.1073/pnas.062526099>.
- Gaiarsa S, Comandatore F, Gaibani P, Corbella M, Dalla Valle C, Epis S, Scaltriti E, Carretto E, Farina C, Labonia C, Landini MP, Pongolini S, Sambri V, Bandi C, Marone P, Sassera D. 2014. Genomic epidemiology of *Klebsiella pneumoniae*: the Italian scenario, and novel insights into the origin and global evolution of resistance to carbapenem antibiotics. *Antimicrob Agents Chemother* 59:389–396. <http://dx.doi.org/10.1128/AAC.04224-14>.

13. Aziz RK, Nizet V. 2010. Pathogen microevolution in high resolution. *Sci Transl Med* 2:16ps4. <http://dx.doi.org/10.1126/scitranslmed.3000713>.
14. Carapetis JR, Parr J, Cherian T. 2006. Standardization of epidemiologic protocols for surveillance of poststreptococcal sequelae: acute rheumatic fever, rheumatic heart disease and acute poststreptococcal glomerulonephritis. January. National Institutes of Health/National Institute of Allergy and Infectious Disease, Bethesda, MD. <http://www.niaid.nih.gov/topics/streptococcal/documents/groupasequelae.pdf>.
15. European Committee on Antimicrobial Susceptibility Testing. 2014. Breakpoint tables for interpretation of MICs and zone diameters, version 4.0. EUCAST, Basel, Switzerland.
16. Beall B, Facklam R, Thompson T. 1996. Sequencing *emm*-specific PCR products for routine and accurate typing of group A streptococci. *J Clin Microbiol* 34:953–958.
17. Friães A, Pinto FR, Silva-Costa C, Ramirez M, Melo-Cristino J. 2013. Superantigen gene complement of *Streptococcus pyogenes* relationship with other typing methods and short-term stability. *Eur J Clin Microbiol Infect Dis* 32:115–125. <http://dx.doi.org/10.1007/s10096-012-1726-3>.
18. McGregor KF, Spratt BG, Kalia A, Bennett A, Bilek N, Beall B, Bessen DE. 2004. Multilocus sequence typing of *Streptococcus pyogenes* representing most known *emm* types and distinctions among subpopulation genetic structures. *J Bacteriol* 186:4285–4294. <http://dx.doi.org/10.1128/JB.186.13.4285-4294.2004>.
19. Andrews S. 2014. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics/Babraham Institute, Cambridge, United Kingdom. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
20. Chevreux B, Wetter T, Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information. *Comput Sci Biol* 1999: 45–56.
21. Darling AE, Mau B, Perna NT. 2010. ProgressiveMauve: multiple genome alignment with gene gain, loss, and rearrangement. *PLoS One* 5:e11147. <http://dx.doi.org/10.1371/journal.pone.0011147>.
22. Huelsenbeck JP, Ronquist F. 2001. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755. <http://dx.doi.org/10.1093/bioinformatics/17.8.754>.
23. Chen LH, Xiong ZH, Sun LL, Yang J, Jin Q. 2012. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res* 40:D641–D645. <http://dx.doi.org/10.1093/nar/gkr989>.
24. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. 2011. PHAST: a fast phage search tool. *Nucleic Acids Res* 39:347–352. <http://dx.doi.org/10.1093/nar/gkq749>.
25. Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA. 2011. Blast Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12:402. <http://dx.doi.org/10.1186/1471-2164-12-402>.
26. Beres SB, Sylva GL, Barbian KD, Lei B, Hoff JS, Mammarella ND, Liu MY, Smoot JC, Porcella SF, Parkins LD, Campbell DS, Smith TM, McCormick JK, Leung DY, Schlievert PM, Musser JM. 2002. Genome sequence of a serotype M3 strain of group A *Streptococcus*: phage-encoded toxins, the high-virulence phenotype, and clone emergence. *Proc Natl Acad Sci U S A* 99:10078–10083. <http://dx.doi.org/10.1073/pnas.152298499>.
27. Holden MT, Scott A, Cherevach I, Chillingworth T, Churcher C, Cronin A, Dowd L, Feltwell T, Hamlin N, Holroyd S, Jagels K, Moule S, Mungall K, Quail MA, Price C, Rabinowitsch E, Sharp S, Skelton J, Whitehead S, Barrell BG, Kehoe M, Parkhill J. 2007. Complete genome of acute rheumatic fever-associated serotype M5 *Streptococcus pyogenes* strain Manfredo. *J Bacteriol* 189:1473–1477. <http://dx.doi.org/10.1128/JB.01227-06>.
28. Parkhill J, Wren BW. 2011. Bacterial epidemiology and biology lessons from genome sequencing. *Genome Biol* 12:230. <http://dx.doi.org/10.1186/gb-2011-12-10-230>.
29. Canchaya C, Proux C, Fournous G, Bruttin A, Brüssow H. 2003. Prophage genomics. *Microbiol Mol Biol Rev* 67:238–276. (Erratum, 67: 473, 2003.) <http://dx.doi.org/10.1128/MMBR.67.2.238-276.2003>.
30. Bao Y, Liang Z, Booyjzen C, Mayfield JA, Li Y, Lee SW, Ploplis VA, Song H, Castellino FJ. 2014. Unique genomic arrangements in an invasive serotype M23 strain of *Streptococcus pyogenes* identify genes that induce hypervirulence. *J Bacteriol* 196:4089–4102. <http://dx.doi.org/10.1128/JB.02131-14>.
31. Marciel AM, Kapur V, Musser JM. 1997. Molecular population genetic analysis of a *Streptococcus pyogenes* bacteriophage-encoded hyaluronidase gene: recombination contributes to allelic variation. *Microb Pathog* 22: 209–217. <http://dx.doi.org/10.1006/mpat.1996.9999>.
32. Mylvaganam H, Bjorvatn B, Hofstad T, Osland A. 2001. Molecular characterization and allelic distribution of the phage-mediated hyaluronidase genes *hylP* and *hylP2* among group A streptococci from western Norway. *Microb Pathog* 30:311. <http://dx.doi.org/10.1006/mpat.2001.0445>.
33. Henningham A, Yamaguchi M, Aziz RK, Kuipers K, Buffalo CZ, Dadesh S, Choudhury B, Van Vleet J, Yamaguchi Y, Seymour LM, Ben Zakour NL, He L, Smith HV, Grimwood K, Beatson SA, Ghosh P, Walker MJ, Nizet V, Cole JN. 2014. Mutual exclusivity of hyaluronan and hyaluronidase in invasive group A *Streptococcus*. *J Biol Chem* 289: 32303–32315. <http://dx.doi.org/10.1074/jbc.M114.602847>.
34. Schommer NN, Muto J, Nizet V, Gallo RL. 2014. Hyaluronan breakdown contributes to immune defense against group A *Streptococcus*. *J Biol Chem* 289:26914–26921. <http://dx.doi.org/10.1074/jbc.M114.575621>.