Genomics 103 (2014) 229–238

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Methods

Origin, duplication and reshuffling of plasmid genes: Insights from *Burkholderia vietnamiensis* G4 genome

Isabel Maida [a], Marco Fondi [a], Valerio Orlandini [a], Giovanni Emiliani [b], Maria Cristiana Papaleo [a], Elena Perrin [a], Renato Fani [a],*

[a] *Laboratory of Microbial and Molecular Evolution, Department of Biology, Via Madonna del Piano 6, University of Florence, I-50019 Sesto F.no, Firenze, Italy*
[b] *Tree and Timber Institute, National Research Council, Via Madonna del Piano, 10 I-50019 Sesto F.no, Firenze, Italy*

ARTICLE INFO

Article history:
Received 18 August 2013
Accepted 13 February 2014
Available online 24 February 2014

Keywords:
Blast2Network
Gene sharing
Molecular remodelling
Gene duplication
Molecular habitat

ABSTRACT

Using a computational pipeline based on similarity networks reconstruction we analysed the 1133 genes of the *Burkholderia vietnamiensis* (*Bv*) G4 five plasmids, showing that gene and operon duplication played an important role in shaping the plasmid architecture. Several single/multiple duplications occurring at intra- and/or interplasmids level involving 253 paralogous genes (stand-alone, clustered or operons) were detected. An extensive gene/operon exchange between plasmids and chromosomes was also disclosed. The larger the plasmid, the higher the number and size of paralogous fragments. Many paralogs encoded mobile genetic elements and duplicated very recently, suggesting that the rearrangement of the *Bv* plastic genome is ongoing. Concerning the "molecular habitat" and the "taxonomical status" (the Preferential Organismal Sharing) of *Bv* plasmid genes, most of them have been exchanged with other plasmids of bacteria belonging (or phylogenetically very close) to *Burkholderia*, suggesting that taxonomical proximity of bacterial strains is a crucial issue in plasmid-mediated gene exchange.

1. Introduction

In the last decade, the total number of completely sequenced prokaryotic genomes has grown exponentially and, to date, more than 12,000 publicly listed bacterial and Archaeal genome projects (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) at different stages of progress are reported.

Currently, the in silico analysis of available genomic data has provided significant advances in our understanding of a number of important themes, including bacterial diversity and population characteristics. These approaches can also help in gaining a deeper understanding of the evolutionary forces that have shaped genomes architecture, from the origin(s) and evolution of new genes to their grouping into clusters and/or operons [34]. The reconstruction of the main evolutionary steps of each gene or gene clusters is usually achieved through phylogenomics [16]. It is worth noticing that both phylogenomics and comparative genomics approaches have highlighted the importance of non-vertical transmission in shaping genomes, that is the possibility that genes may not follow classical vertical inheritance but, rather, may be horizontally transferred between different cells. This process is usually referred to as horizontal gene transfer (HGT) and, despite its extent is still under debate [14,15], it has played a major role (at least) in the early stages of bacterial evolution [40–42].

HGT is usually mediated by the mobile gene pool (the so-called "mobilome") that comprises plasmids, transposons and bacteriophages (all of which usually referred to as mobile genetic elements, MGEs) [7,28]. Plasmids and other MGEs can be transferred between microorganisms and within different DNA molecules inhabiting the same cytoplasm, representing natural vectors for the gene transfer and functions they code for [7]. Usually MGEs do not accommodate any of the "core" genes required by the cell for basic growth and division, but rather they carry traits that may be useful periodically to enable the cell to exploit particular environmental conditions, such as the survival in the presence of a potentially lethal antibiotic [4]. This flexibility is mostly due to the abundance of transposable elements that may facilitate intra- and intermolecular recombination by creating homology regions. In this way, a single DNA fragment (possibly embedding one or more coding genes) can be exchanged between the MGE harbouring it and other informational molecules (including chromosomes and/or other MGEs). In this context, it is particularly interesting the finding that, in some cases, chromosomes and plasmids inhabiting the same cell can share sequences possessing a very high degree of similarity, probably as the result of recombination events [23]. As recently shown, also chromosomes and plasmids belonging to different strains/species share a number of homologous sequences, probably as the result of one (or more) HGT event(s). This has important biological drawbacks since it may allow the transfer of previously plasmid-encoded functions

* Corresponding author.
E-mail address: renato.fani@unifi.it (R. Fani).

http://dx.doi.org/10.1016/j.ygeno.2014.02.004
0888-7543/© 2014 Elsevier Inc. All rights reserved.

to the chromosome(s) and, in turn, permit to the corresponding genes to be spread in the bacterial community through vertical inheritance [23].

Despite the key-role of plasmids in the prokaryotic biology and evolution, their evolutionary dynamics has been poorly explored, mainly because of the lack of extensive similarities between them, except for genes involved in replication and transfer functions [8,21], which hampers classical phylogenetic analyses based on gene genealogy and syntheny [5]. Up to now, phylogenomics and comparative genomics approaches have been mostly applied to the analysis of large datasets of genomes belonging to (more or less) distantly related microorganisms. These studies, although providing fundamental advances in the understanding of the overall dynamics of microbial evolution, rarely tried to provide a detailed census of the major evolutionary steps occurring in single genomes. As a consequence, very little is known on the molecular mechanisms involved in plasmid construction as well as the interrelationships existing between DNA molecules inhabiting the same cytoplasm. Particularly interesting might be the understanding of the role that gene duplication and the incorporation of exogenous DNA stretches have had in the construction of plasmid molecules, an issue that, at least to our knowledge, has been poorly investigated. Gene duplication has been recognized as one of the major mechanism allowing the increase in genome size and the acquisition of new metabolic abilities [18,24,25], thus driving the evolution of genes and genomes. Genes originated via duplication of an ancestral one are called paralogs [22]. In general, paralogous genes perform different, although similar, functions within the same (micro)organism. The terms paralogous and orthologous genes were introduced to classify different types of homologous genes (genes that evolved from a single ancestral sequence). However, gene duplication may generate many copies of genes with the same function, thereby enabling the production of a large quantity of rRNAs or proteins [19]. Therefore, the evolution of paralogs does not reflect organismal evolution, which is accomplished by orthologous genes, i.e. genes that evolved from the same feature in their last common ancestor, that do not necessarily retain the ancestral function ([18] and references therein). In case paralogs undergo multiple rounds of duplication they give raise to paralogous gene families of different sizes [18,33]. In spite of the large body of information available on the role played by gene duplication in shaping the bacterial chromosome, little is known about the role that gene duplication and other mechanisms, such as the introgression of external genes might have played in the evolution of bacterial plasmids (especially the largest ones, which overall resembles bacterial chromosomes [31]). Useful hints on these issues might be inferred by a deep analysis of intra- and intermolecular relationships. To this purpose, a computational biology approach (Blast2Network) based on similarity networks reconstruction and phylogenetic profiling has been proposed and applied to very different study-cases, i.e. to depict the similarities among plasmids from *Enterobacteriaceae* [7],to analyse the *Acinetobacter* pan-plasmidome [23] and the cross-talk between plasmids and chromosomes in the cyanobacterium *Synechococcus* [35]. Moreover, it was recently implemented in a more comprehensive computational pipeline in order to study the extent and the dynamics of HGT of antibiotic resistance determinants within the whole bacterial community [26].

Thus, the aim of this work was to analyse the interrelationships existing between plasmids and chromosomes inhabiting the same cytoplasm by applying the abovementioned workflow to the analysis of the whole genome of *Burkholderia vietnamiensis* G4 genome, a β-proteobacterium possessing a complex genome consisting of three chromosomes and five plasmids (whose sequences are publicly available since 2007). This bacterium was isolated from wastewater in Pensacola, USA, [36] and it is well-known because of its role in trichloroethene co-oxidation [27]; moreover, this strain has been used in a number of polluted sites to aid clean-up of ground water. The *B. vietnamiensis* strains are also known for their rhizosphere colonizing behaviour and their ability to fix atmospheric N_2 [9]. Besides these abilities, strains belonging to this species are well-known for their role in the infection of immuno-compromised patients [9]. For its multiplicity of ability and characteristics it can be considered as an excellent model microorganism to study the gene flow between different DNA molecules inhabiting the same cytoplasm.

2. Results

2.1. Overall strategy

A total of 7617 protein sequences compose the *B. vietnamiensis* G4 genome (1133 from the five plasmids and 6484 from the three chromosomes) (Table S1). The protein sequence dataset was used for the construction of networks using the software B2N [7] accounting for the sequences identity at either intra- or intermolecular level, that is:

i) intra-molecular networks, i.e. connecting homologous (most likely paralogous) genes within the same plasmid;
ii) inter-molecular networks showing homologous genes harboured by different plasmids;
iii) “higher-level” inter-molecular identity networks, describing the putative flux of genes between plasmid(s) and chromosome(s).

In each network nodes represent proteins and links the degree of sequence identity (expressed as percentage) between shared proteins. The analysis of networks may allow the identification of paralogous genes on the same plasmid or between different plasmid(s) and between plasmids and chromosomes inhabiting the same cytoplasm. Moreover, the networks allow the identification of single/multiple duplication events involving stand-alone genes, cluster of genes and/or operons or parts thereof. The identification of the function performed by the duplicated genes might reveal the existence of genes particularly prone to duplication. Networks construction was reiterated at different sequence identity thresholds, ranging from ≥40% up to 100%. Assuming that the higher the degree of amino acid sequence identity between two proteins, the more recent the duplication event responsible for the origin of the two paralogous encoding genes, it should be possible to establish a sort of diachrony (a “temporal scan”) of the duplication events. Similarly to Dagan et al. [13] and, later, to Halary et al. [30] and Tamminen et al. [38], this allows the interpretation of the resulting networks under a molecular clock-based assumption, that is, under the hypothesis that proteins with the highest percentages of identity were likely to be more recently shared than the ones with less identity. In the present context, proteins with 95% identity were considered more recently shared than those with 70%.

Basing on these assumptions, each network (obtained at the different threshold) was analysed in order to answer the following questions:

1. Which plasmids harbour paralogs and how many genes are duplicated?
2. Is there any correlation between the number/type of paralogs and plasmid size?
3. Which is the size of the paralogous gene families?
4. How are paralogous genes arranged (tandem or scattered) and organized (stand-alone, clustered or operonically) onto their plasmid backbone?
5. Which is the function performed by paralogs?
6. Is it possible to establish the temporal scan of the duplications events?

2.2. Analysis of networks

2.2.1. The flow of genes within and/or between plasmids (Questions 1–5)

We firstly analysed the presence of paralogous genes within each of the *B. vietnamiensis* G4 plasmids. Thus, thirty-five networks were obtained by reiterating the analysis using seven different identity thresholds (≥40, ≥50, ≥60, ≥70, ≥80, ≥90, and 100%) for each plasmid. We adopted a minimum of 40% amino acid sequence identity threshold, a value that is generally accepted to be shared by proteins encoded by paralogous genes [29,39]. The networks constructed at a threshold ≥40% are reported in Fig. 1, whereas the entire sets of networks are reported in Fig. S1. The analysis of these networks revealed that:

i) In each plasmid, at least one paralogous gene pair was present.
ii) An increase in the total number of paralogs with the increase of the plasmid size was detected (Fig. 2).
iii) The dimension of paralogous gene family increased with the plasmid size (Table 1).
iv) Concerning the arrangement and organisation of paralogous genes onto the plasmids backbone, we found both tandem and scattered duplications. Moreover, we observed that duplication of larger gene arrays are more abundant in larger plasmids in respect to smaller ones; in the smallest plasmids, pBVIE05-pBVIE03, just stand-alone paralogous genes were disclosed, whereas paralogous gene clusters/operons were found in pBVIE02-pBVIE01 (Fig. 1 and Table 1).
v) Concerning the function of intra-plasmid paralogous genes (Tables S2 and S3) most of them (73.3%) coded for proteins involved in DNA transposition/mobilization (see also Table S4).

To check the evolutionary relationships existing among (all) the *B. vietnamiensis* G4 plasmids, we constructed the relative intermolecular networks using all the 1133 plasmid-encoded proteins. The networks obtained at ≥40% identity threshold is shown in Fig. 1 (the entire set of networks constructed at ≥40, ≥50, ≥60, ≥70, ≥80, ≥90, 100% sequence identity thresholds is reported in Fig. S2). The analysis of the networks revealed that:

i) Each of the five G4 plasmids is interconnected (although at different extent) at least to another plasmid and many genes (about 41%of them) are shared by at least three plasmids (Fig. 1

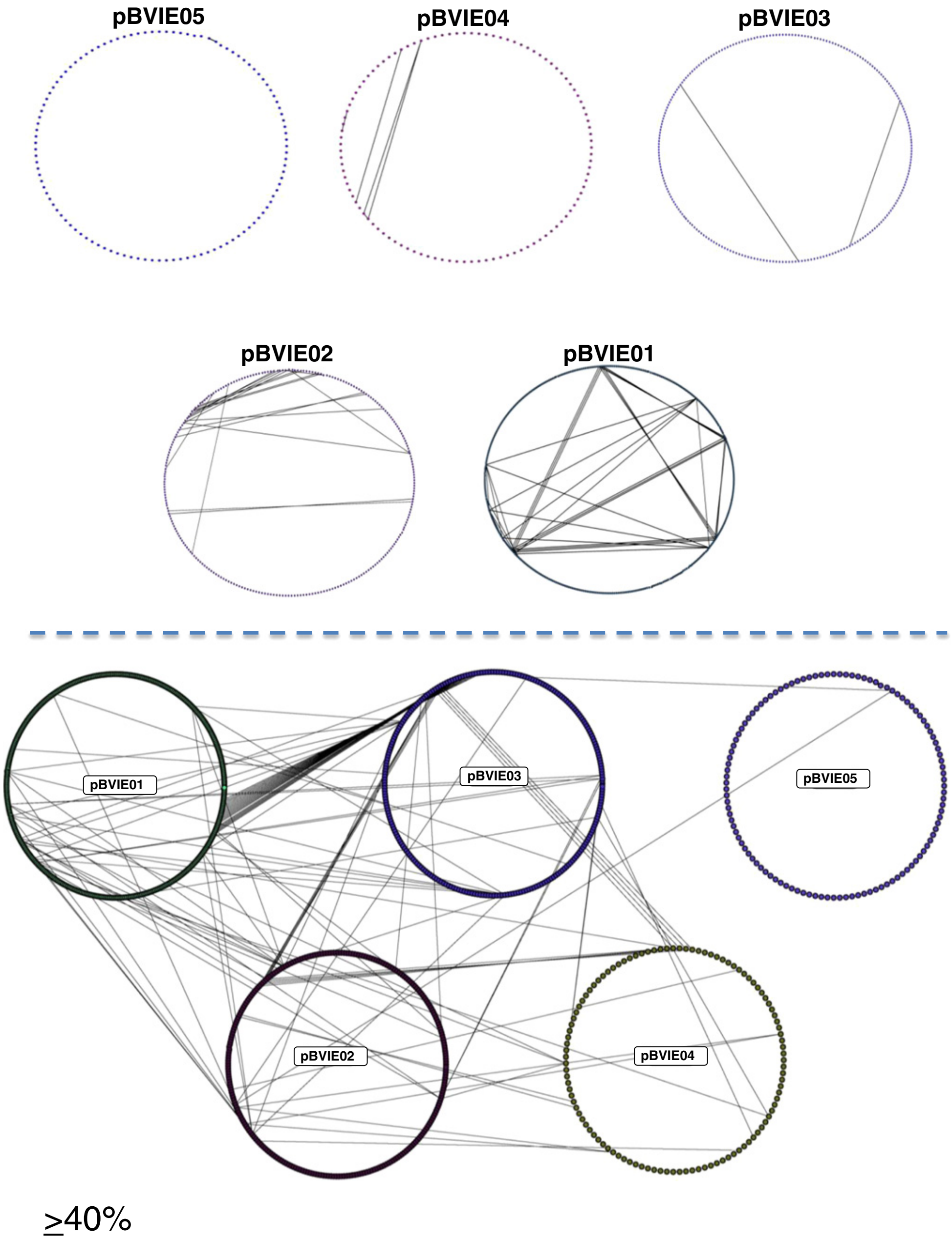

Fig. 1. Identity based networks at intra-plasmid (upper) and inter-plasmid (lower) levels. All the proteins encoded by the same plasmid (nodes) are circularly arranged and are linked to the others according to their identity value. The resulting pictures for identity threshold correspond to ≥40% are shown.

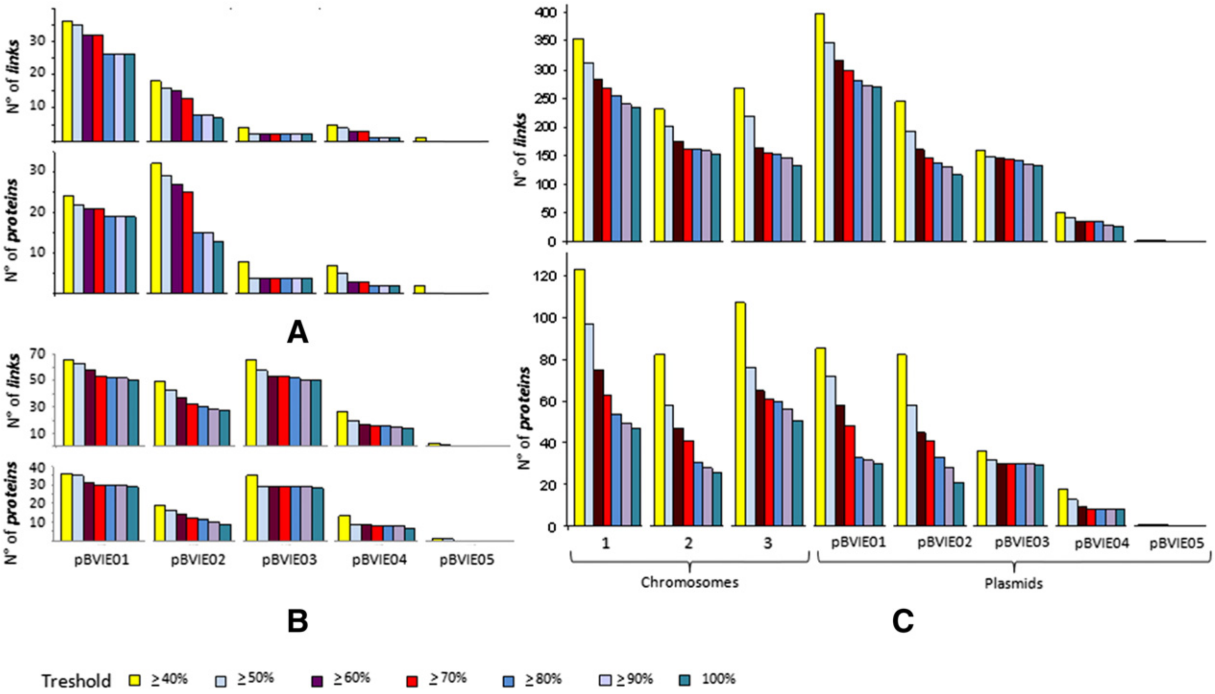

Fig. 2. Correlation between the plasmid/chromosome sizes and the no. of links interconnecting paralogous proteins (upper) or number of paralogous proteins. In each section we consider a different level of analysis, A: intra-plasmids, B: between plasmids, and C: between plasmids and chromosomes. Colours indicate the different threshold used.

and Table 1), suggesting the existence of an intense gene flow between different DNA replicons; this idea was also supported by the analysis of plasmid–chromosome networks (see below).

ii) The number of links decreases with the increase of the identity threshold (Figs. 2 and 3). At ≥40% of sequence identity, 104 (about 15%) out of the 1133 plasmid-encoded proteins are interconnected. However, at 100% identity threshold, the number of interconnected nodes remained unexpectedly high (73), suggesting a very recent and ongoing genetic exchange between different plasmids (Table S5).

iii) About 34% of the 104proteins connected at a ≥40% threshold are involved in DNA transposition and some of them exhibited a very high degree of sequence identity among themselves (100%) (Tables S2 and S5), suggesting a recent gene exchange between plasmids harbouring them.

iv) At 100% identity threshold, a cluster of genes, shared by pBVIE03 and pBVIE01, coding for proteins involved in different functions including chromate transporter, transposase and other metabolic functions was identified.

Summarizing, the analysis of all the networks reported in this paragraph revealed that:

1. Paralogous genes were found in each of the five *B. vietnamiensis* G4 plasmids (*Question 1: Which plasmids harbour paralogs and how many genes are duplicated?*).
2. As shown in Fig. 2 the increase of the number of both links and connected proteins was parallel to the increase of plasmid size (*Question 2: Is there any correlation between the number/type of paralogs and plasmid size?*).
3. Concerning Question 3 (*Which is the size of the paralogous gene families?*), we found duplication of both stand-alone and cluster of genes, some of which corresponding to or containing operons. These genes or gene clusters underwent single or multiple duplications both at intra- and inter-plasmid level.
4. The paralogous copies were scattered or tandemly arranged (in the case of stand-alone genes). No tandemly arranged cluster was detected. The reason of this is unclear, however, it might be related to the difficulty by which the in tandem-duplications of long DNA

Table 1
Summary of paralogous duplication events detected within each of the five plasmids (intra-molecular section), between different plasmids, and between plasmids and chromosomes (inter-molecular section) at the threshold identity of ≥40%.

		Stand-alone genes			Gene Clusters/ Putative Operons		Total of Genes	
		Tandem	Scattered		Scattered			
		Single	Single	Multiple	Single	Multiple		
Intra-molecular paralogous gene families	pBVIE05	1	0	0	0	0	2	75
	pBVIE04	0	2	1	0	0	7	
	pBVIE03	2	2	0	0	0	8	
	pBVIE02	1	11	2	1	0	34	
	pBVIE01	0	3	1	0	1	24	
Inter-molecular paralogous gene families	Between Plasmids	0	10	9	3	1	104	
	Between Plasmids and Chromosomes	0	29	7	14	12	534	

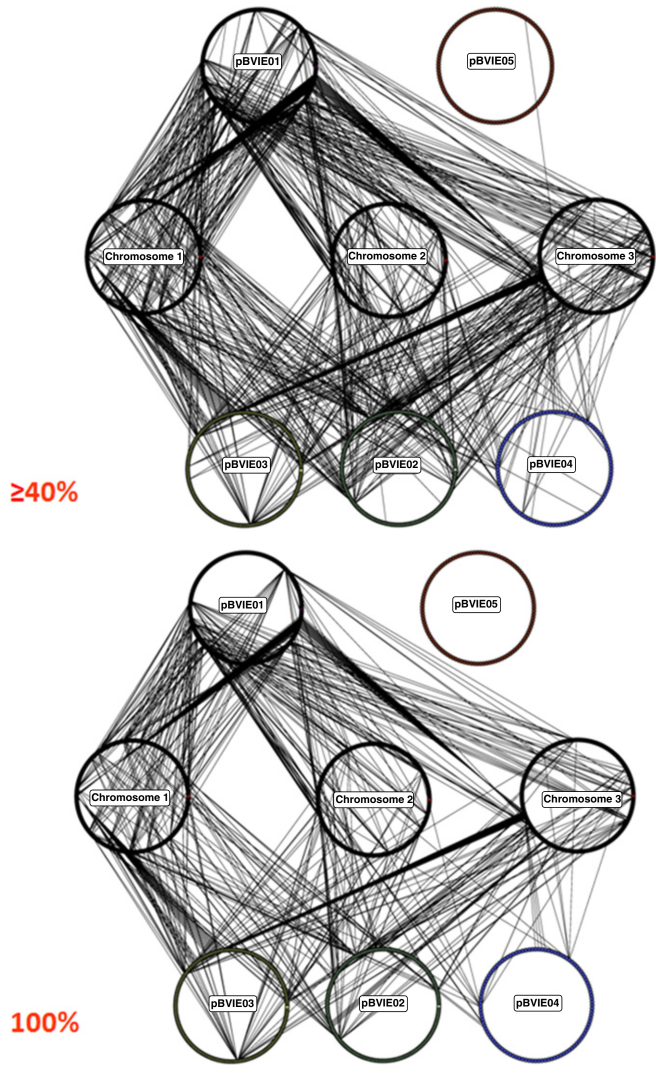

Fig. 3. Identity based networks showing the inter-molecular relationships existing between the five *Burkholderia vietnamiensis* G4 plasmids and the three chromosomes.

stretches can be fixed in the genomes [3] (*Question 4: How are paralogous genes arranged (tandem or scattered) and organized (stand-alone, clustered or operonically) onto their plasmid backbone?*).

5. The analysis of the function performed by duplicated genes (Tables S2, S4 and S5) revealed that most of them code for proteins involved in DNA transposition/mobilization. It is worth noticing that either all or most of the most recent, multiple, and operon duplications concerned only transposition/mobilization related elements at intra and inter-plasmid levels respectively (*Question 5: Which is the function performed by paralogs?*). The main presence of transposition/mobilization elements in the paralogous gene families is probably due to their structure with many homology regions that could facilitate their recombination [28] and consequently their duplication.

2.2.2. Genes flowing between plasmids and chromosomes

To check the existence of a gene flux between *B. vietnamiensis* G4 plasmids and chromosomes, we constructed the network using all the 1133 plasmid proteins and the 6484 chromosomal proteins. The networks obtained at sequence identity threshold ≥40% and 100% are shown in Fig. 3 (the entire set of networks obtained, i.e. at ≥40, ≥50, ≥60, ≥70, ≥80, ≥90, 100% thresholds, is reported in Fig. S3).

The analysis of Fig. 3 revealed that plasmids and chromosomes shared a high number of genes, and all the five plasmids shared at least one gene with one (or more) chromosomes. The total number of links and connected proteins decreased with the increase of sequence identity (Figs. 2 and 3), ranging from 853 links and 534 proteins (threshold ≥ 40%) to 543 links and 212 proteins (threshold 100%) and exhibited a positive correlation with plasmid size. Furthermore, at ≥40% sequences identity, 222 out of the 534 connected proteins belong to plasmids and 312 belong to chromosomes. The percentage of chromosome-encoded proteins connected with plasmids, were 3.7, 3.9, and 9.6% for chromosome 1, 2 and 3, respectively.

The DNA regions exchanged between plasmids and chromosomes were in many cases large and in most cases embedded more than one gene. The number of paralogous clusters and putative operons shared by plasmid(s) and chromosome(s) is higher than that found within and between plasmids (Table 1). Concerning the organisation of paralogous genes (Table 1), the analysis of networks revealed a quite complex scenario. Indeed, as for the other networks, single or multiple duplications of stand-alone genes, gene clusters and operons were found; however, a reshuffling of one or more single genes embedded in operons was also detected.

Concerning the function of the connected proteins (Tables S2 and S4), genes involved in DNA transposition/mobilization represented the most frequent class; however, the number of genes coding for proteins involved in other metabolic functions was higher than that disclosed within or between plasmids. Interestingly, long DNA regions may flow through plasmids and chromosomes. This is the case of the gene cluster shared between plasmids pBVIE03 and pBVIE01 and chromosomes 1 and 3, which contained genes coding for proteins involved in different functions including chromate transporter, transposases and other metabolic activities.

The whole body of data revealed that the "gene movement" involved 253 genes identified as paralogs at a threshold ≥40%. Besides, as shown in Fig. 4 most of paralogous genes are shared by at least two different DNA molecules. The core set of paralogous genes is represented by 16 genes. However, one of the most striking differences is the finding that just 7 out of the 104 paralogous genes are exchanged only between plasmids; this implies that when a gene is exchanged between different plasmids, this is parallel to at least another duplication (intraplasmid or between plasmids and chromosomes). The reason of such behaviour is unclear.

Most of the 253 paralogs codes for transposition/mobilization related elements (36%) or proteins with unknown function (40%).

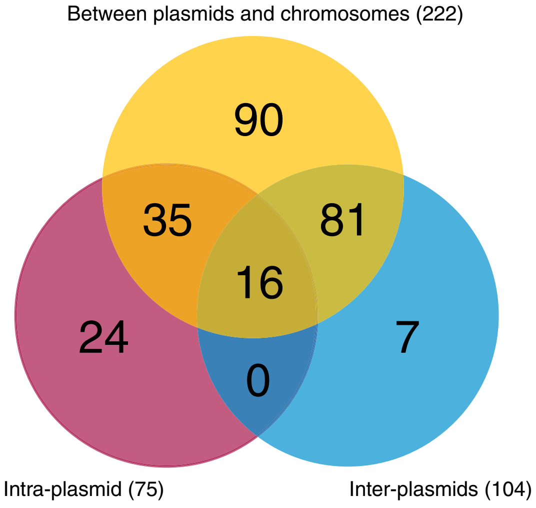

Fig. 4. Schematic representation of core, accessory and unique set of intra- and inter-molecular *Burkholderia vietnamiensis* G4 paralogous genes.

Besides, most of the genes involved in generic metabolic function are mainly found in paralogous gene families involving the chromosomes, instead the presence of many transposition/mobilization related elements were found in paralogous gene families including proteins involved in more than one exchange at the same time, in particular all the 16 members of the group 7 are transposition/mobilization related elements. This finding supports the idea that these elements can promote the "communication" between different DNA molecules.

2.3. Temporal scan of duplications events (Question 6: Is it possible to establish the temporal scan of the duplications events?)

On the basis of the degree of sequence identity shared by each protein pair, it might be possible to infer the time-line of gene duplications. To this purpose all the 105 networks constructed at the different thresholds were analysed. Even though the number of nodes and links decreased with the increase of sequence identity threshold, several proteins remained connected at very high degree of sequence identity (i. e. 100%), supporting the idea that these links connect genes that underwent very recent duplication events. In order to try to trace the diachrony of the duplication events we constructed new sets of networks using three threshold intervals (≥40–60%; >60–95%; >95–100%) that are shown in Fig. 5. The analysis of networks revealed that most of duplications occurred (very) recently. Interestingly, all the proteins jumping on the same molecules and (most of) those that are connected between plasmids and between plasmids and chromosomes at 100% threshold identity belong to the class of MGE. These data suggest i) that MGE very rarely remain located in the original site of transposition, ii) that these elements play a major role in promoting recombination event on and/or between different DNA molecules inhabiting the same cytoplasm, and iii) that these events are still ongoing in the cytoplasm of *B. vietnamiensis* G4.

2.4. Putative origin of B. vietnamiensis G4 plasmid genes

On the basis of the presence of paralogs in the *B. vietnamiensis* G4 genome, the 1133 plasmid genes were split into two clusters: i) the first one includes 253 paralogous (connected) genes, and ii) the second embedding the 880 not connected (isolated) genes, that is genes that do not have a homolog in the *B. vietnamiensis* G4 genome.

Regarding the "origin" of plasmid genes, this question can be split in two sub-questions: 1) which is their preferential "molecular habitat"? and 2) from a cellular viewpoint, which is the "preferential organismal sharing" (POS)? We define POS as the strongest evolutionary relatedness (based on sequence similarity) of each gene in the replicons with taxonomically and/or ecologically correlated organisms.

2.4.1. Identification of the "molecular habitat" of B. vietnamiensis G4 plasmid genes

In order to check the (putative) molecular habitat of the 1133 plasmid genes, we adopted an ad hoc developed computational pipeline (see Materials and methods). This approach allows the discrimination of the four different scenarios that can be depicted for genes "molecular habitat": plasmid (P), chromosomal (C), viral (V), and no preference (NP) (i. e. genes that do not have a significant percentage, or an equally shared, of match in any group). Data obtained are shown in Table 2, whose analysis revealed that in both groups (connected and isolated) the majority of genes has a likely plasmid molecular habitat, suggesting

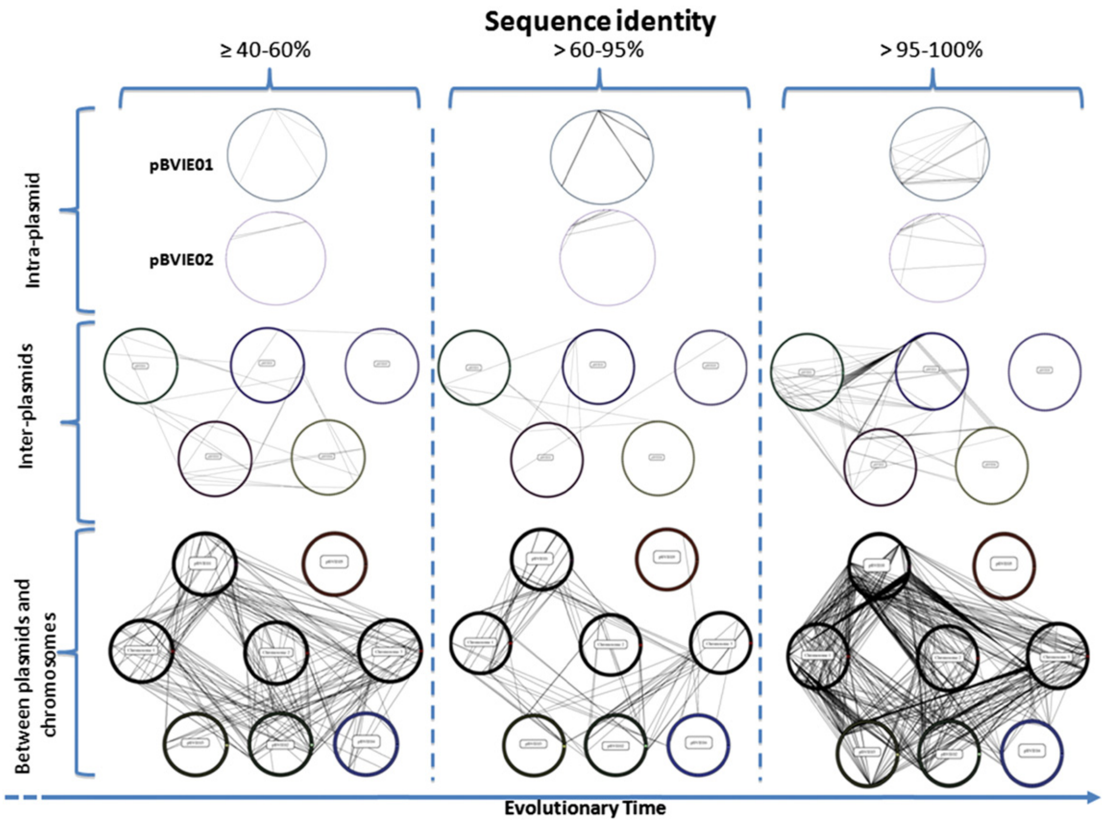

Fig. 5. Temporal scan of the duplication events occurred in plasmids pBVIE01 and pBVIE02, (upper part), between plasmids (middle part) and plasmids and chromosomes (lower part). In the left section are connected by links the proteins that display an identity threshold <60% (in order to trace the oldest events); in the middle part those that display an identity threshold between 60% and <95% (to trace the events that took place not far back in time); on the right, those exhibiting an identity threshold between 95% and 100% (to trace the very recent events).

Table 2
"Molecular habitat" of *Burkholderia vietnamiensis* G4 plasmid genes.

Genes		Molecular habitat				Plasmid
	Total number	P	C	V	NP	
Connected	93	56	25	0	12	pBVIE01
	90	70	8	3	9	pBVIE02
	42	30	6	1	5	pBVIE03
	25	18	3	1	3	pBVIE04
	3	3	0	0	0	pBVIE05
	253	179	42	5	27	
	%	70.8	16.6	1.9	10.7	
Isolated	310	192	73	43	2	pBVIE01
	173	134	24	13	2	pBVIE02
	207	132	44	29	2	pBVIE03
	82	68	6	4	4	pBVIE04
	108	54	24	27	3	pBVIE05
	880	580	171	116	13	
	%	65.9	19.4	13.2	1.5	

Abbreviations: P, plasmid; C, chromosomal; V, viral; and NP, no preference.

that plasmid genes preferentially undergo rearrangements with other plasmids rather than with other different DNA molecules. Similar percentages (16.6–19.4%) of genes having a putative chromosomal origin were detected in both sets of genes. The major difference in the two gene sets concerned genes with a hypothetical viral origin; indeed, paralogous genes with a putative viral origin are much less represented (1.9%) in respect to isolated genes (13.2%). This finding might suggest that viral genes introgressing in plasmid molecules are less prone to duplicate in respect to genes having another origin, although the reason of this is still unclear.

2.4.2. Identification of the "Preferential Organismal Sharing" (POS) of plasmid genes

By assuming that genes can be shared by different plasmids, which, in turn, can flow between (micro)organisms belonging to the same or to different species/genus, and by other DNA molecules that can exchange DNA stretches with plasmids, the "cellular" origin of plasmid genes cannot be easily identified. In other words, the presence of paralogous genes shared by different DNA molecules harboured by the same or different (micro)organisms cannot give any indication about the direction of gene exchange, that is which (micro)organism is the "donor" or the "recipient". However, it should be possible to identify the "Preferential Organismal Sharing" (hereinafter POS), that is the group of organisms that, on the basis of evolutionary relatedness (vertical transmission) and/or physical proximity (ecological niche sharing, HGT) were and/or are exchanging DNA stretches.

To identify POS, the following analysis was carried out: the amino acid sequence of each of the 1133 plasmid-encoded protein was used as seed to probe the database containing completely sequenced genomes. Once discarded the sequence retrieved from *B. vietnamiensis* G4 genome (and corresponding to the query sequence), the first BLAST hit was recovered. The (micro)organism having the first BLAST hit sequence was considered as the possible preferential organism sharing the plasmid genes analysed. Data obtained revealed that:

1. No Archaeal sequence was retrieved using the parameters described above, suggesting that no gene from Archaea has been recently exchanged with *B. vietnamiensis* G4 plasmids.
2. Interestingly, one eucaryotic sequence was retrieved at significant e-values. It belongs to the "isolated" set of plasmid protein and is a protein from *Populus trichocarpa* (GI:222874892) sharing a 99% sequence identity with protein GI 134287726 from plasmid pBVIE02. This protein belongs to the TniQ trasposon-like protein family, which has orthologs in a limited number of other bacteria, mainly belonging to β-proteobacteria. The molecular habitat of this sequence is plasmid, and it is quite possible that it might have been very recently exchanged between the plant *P. trichocarpa* and one of the β-proteobacteria harbouring this gene.
3. Concerning the isolated proteins, about one third of them (32%) shared the highest degree of sequence identity with proteins belonging to other *Burkholderia* strains. Interestingly, 6% of them shared the highest degree of sequence similarity with *Methylibium*, 4% with *Ralstonia*, 4% with *Pseudomonas* and 4% with *Cupriavidus*. The remaining 21% had the highest degree of sequence identity with different microorganisms affiliated to almost thirty different genera most of which belonging to β-proteobacteria.
4. A similar scenario was also disclosed for plasmid proteins having paralogs. Indeed, most of them (54%) shared the highest degree of sequence identity with proteins belonging to other *Burkholderia* strains, followed by proteins from *Ralstonia* (11%), *Cupriavidus* (8%), and *Alcaligenes* (4%). The remaining 23% shared the highest degree of sequence identity with proteins belonging to other microorganisms, most of which affiliated to β-proteobacteria.

This finding strongly suggested that *B. vietnamiensis* G4 plasmid genes can be preferentially exchanged between bacteria belonging to phylogenetically close taxa.

In order to check whether genes from the five plasmids exhibited the same distribution of molecular habitat and/or POS and to map them on the plasmid backbone, the five plasmids were analysed separately, considering also the GC content of each of them, which was analysed using a 350 bp window. Data obtained are shown in Fig. 6, where the outermost circle represents the POS, followed by the molecular habitat of isolated genes and the molecular habitat of paralogs; lastly, the inner circle represents the GC content.

The analysis of Fig. 6 revealed that the five plasmids can be split into two groups, the first one including pBVIE01, pBVIE03, and pBVIE05, and the second one comprising pBVIE02 and pBVIE04 exhibiting a different POS, with plasmids pBVIE02 and pBVIE04 being more homogeneous than the other three. Indeed, a very high percentage of their genes are shared preferentially with other *Burkholderia* strains belonging to the same or to different species; interestingly, these genes are not randomly distributed on the plasmid backbone, and they cover an almost continuous region of the plasmid itself. On the other side, the other three plasmids (pBVIE01, pBVIE03, and pBVI05) have a mosaic-like distribution of their genes (regarding their POS), with genes shared with other *Burkholderia* strains/species representing a small amount. Besides, the latter genes are scattered on the plasmid backbone. In spite of this difference, it is quite interesting that in all five plasmids genes that are not shared with other *Burkholderia* strains, are preferentially shared with bacteria belonging to phylogenetically close taxa, such as *Cupriviadus*, or *Ralstonia*, and overall belonging to β-proteobacteria. This suggests that genes from each of the five *B. vietnamiensis* G4 plasmids are preferentially exchanged between phylogenetically close genera and/or species. This *per se* does not imply that these plasmids cannot be transferred also between bacteria belonging to phylogenetically distant taxa; indeed, the existence of broad-host range plasmids able to flow between microorganisms belonging to phylogenetically distant taxa supports this possibility. However, data obtained in this work, strongly suggests that the exchange of genes between plasmids and other DNA molecules is more probable if they are likely exchanged between bacteria belonging to phylogenetically close taxa [2].

The major heterogeneity of plasmids pBVIE01, pBVIE03, and pBVIE05 in respect to plasmids pBVIE02 and pBVIE04 is parallel to a more heterogeneous "molecular habitat" of their genes. Indeed, the three plasmids showed a percentage of genes with a chromosomal or viral origin much higher than that found in plasmids pBVIE02 and pBVIE04 (see also Table 2), which are intermixed with genes with a plasmid origin with no apparent rule.

Concerning the duplication events, there is no apparent relationship between paralogs and their localization on the plasmid backbone.

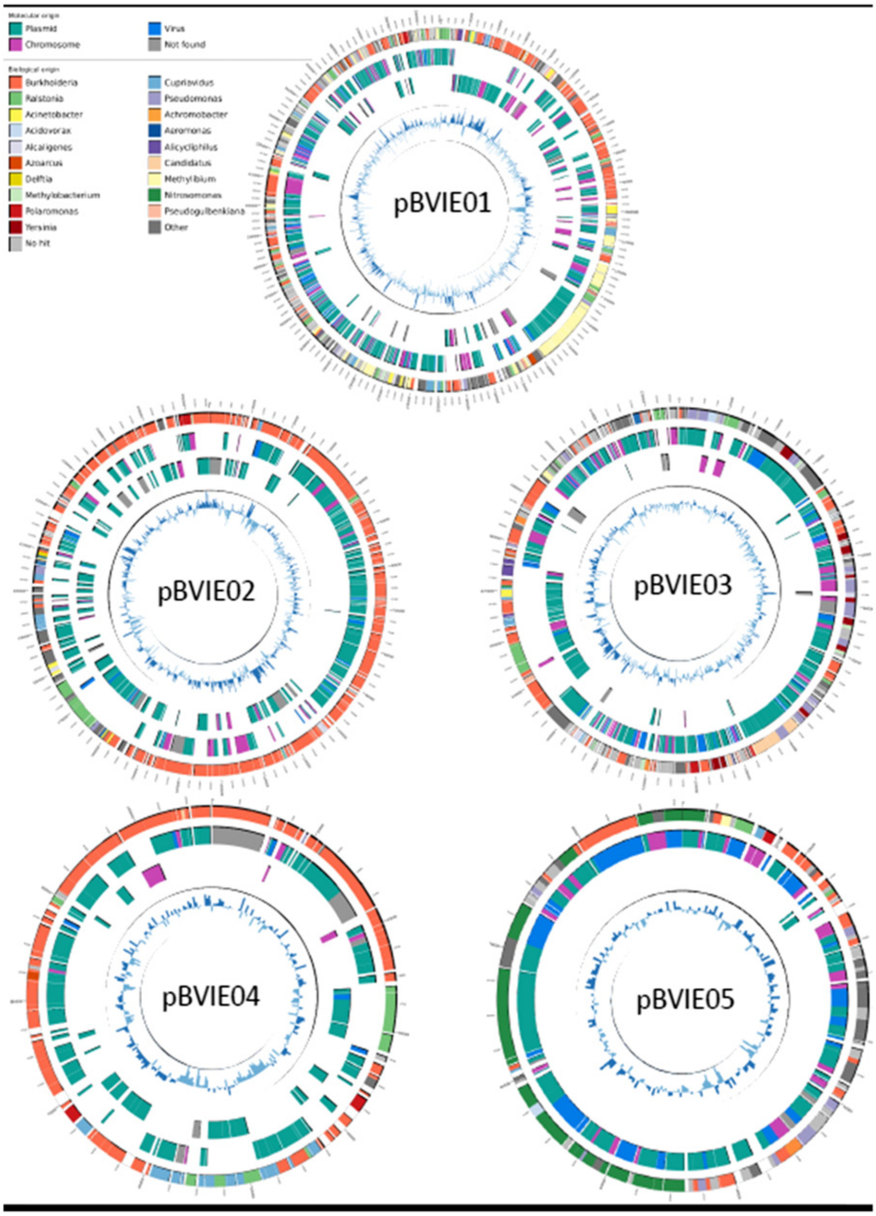

Fig. 6. Schematic representation of *Burkholderia vietnamiensis* G4 plasmids obtained using the software Circos, in each circle there are represented from the outside inwards: the cellular origin, the molecular isolated origin, the molecular paralogs origin and in the inner part it is represented the GC content. For what concern the cellular/molecular origins each colour corresponds to different cellular/molecular origin as it is reported.

Lastly, we detected the exchange of entire gene clusters between *B. vietnamiensis* G4 plasmids and other bacteria; this is particularly evident in the case of plasmid pBVIE01 embedding a gene cluster (involved in the biogenesis of the sex *pilus* assembly) shared with plasmid RPME01 from the bacterium *Methylibium petroleiphilum* PM1.

3. Discussion

The aim of this work was to analyse the gene flow existing between different DNA molecules (three chromosomes and five plasmids of different sizes) inhabiting the same cytoplasm and the molecular mechanisms responsible for the construction of plasmid molecules using the β-proteobacterium *B. vietnamiensis* G4 as a model system. Data obtained suggested that very likely the five plasmids experienced different and complex evolutionary pathways. However, the whole body of data reported in this work revealed that the plasmid(s) structure has been shaped through at least two different mechanisms: i) the acquisition of genes from different sources (exogenous plasmids, chromosomes, viruses) and ii) duplication of DNA regions of different lengths. In addition to this, an ongoing "cross-talk" between genes belonging to i) the same plasmid, ii) different plasmids, and iii) plasmid(s) and chromosome(s) of the same cell, was disclosed. In particular:

1. Intra-molecular paralogous DNA regions of different sizes and complexities in the five *B. vietnamiensis* G4 plasmids analysed in this work were found. The duplication events involved either single genes or entire operons and in some cases it may be possible to give the timing to all these duplication events.
2. Inter-plasmid paralogous genes. The five plasmids exchanged, at variable extent, 9.1% (104) of their whole gene set. The exchange may involve a single gene, operons or gene clusters. The gene flow between these plasmids might have been facilitated by the presence of genes coding for proteins involved in DNA transposition/mobilization.

3. A gene flux between plasmids and chromosomes was also detected; however, the percentage of genes exchanged was different for the three chromosomes (a total of 312 chromosomal proteins connected with the plasmids that correspond to 3.7, 3.9 and 9.6% for chromosome 1, 2 and 3, respectively). This is in agreement with the idea that secondary chromosomes are more plasmid-like than primary ones [12,31]. The gene flow mainly involved genes belonging to the largest plasmids (pBVIE01, pBVIE02, pBVIE03) and was greater than that occurring between plasmids. This situation is different from that found, for example, in the cyanobacterium *Synechococcus* sp. PCC 7002 [35], in which the plasmid genes were most prone to recombine between plasmids than with chromosomes and some plasmids harboured genes encoding proteins that do not share any link neither between them nor with other plasmid proteins in the networks. This finding suggests that different forces might drive the assembly and the gene flux between DNA molecules inhabiting the same cytoplasm in different microorganisms.
4. Several intra- or inter-molecular duplications occurred recently, at least on the basis of sequence identity values and it can be argued that the "cross-talk" in this cytoplasm is a process still ongoing, in agreement with the idea that the microorganisms belonging to the genus *Burkholderia* possess a highly flexible genome [11,32]. On the basis of these data, this flexibility might be due to the high "recombination rate" between genes harboured by different molecules and to the introgression of foreign genes from viruses, plasmids, and/or chromosomes (possibly) from different sources.
5. The finding that the majority of plasmid-borne genes (880) do not have any paralog on the other molecules sharing the same cytoplasm, might suggest that they have exchanged with external foreign DNA molecules through recombination/rearrangement event. Indeed, the most likely source molecule of most of the 1133 plasmid genes is a plasmid, suggesting that they might preferentially be exchanged between plasmids. Finally, most of the genes involved (at different extent) in duplications are related to integration or transposition. This finding suggests that mobile genetics elements are playing (and might have played) a central role in shaping the architecture of the *B. vietnamiensis* G4 genome. Accordingly, these elements promote not only HGT, but also the gene flux between different molecules inhabiting the same cytoplasm.
6. The analysis of the POS of *B. vietnamiensis* G4 plasmid-encoded proteins revealed that most of them are shared with other *Burkholderias* and/or β-proteobacteria, suggesting that next to the proximity in the environment also the phylogenetic proximity might play a central role in the HGT.

Data obtained here revealed that gene duplication has played and is still playing a role in the construction and rearrangement of plasmid molecules. However, the percentage of plasmid genes belonging to paralogous pair or family is much lower (about 10%) than that reported for bacterial chromosomes (about 50%). This finding raises the intriguing question of the biological significance of this relatively small fraction of paralogs in plasmid molecules. Two different scenarios can be depicted to explain this finding: a) it reflects a different (and unknown) evolutionary pathway in the construction of plasmids and chromosomes, or b) it is due to the size of DNA molecules, indeed the percentage of paralogous genes increases with the increase of DNA molecule size. A possible evolutionary pathway predicts that in the very first stage of plasmid construction, these molecules might acquire genes from different DNA molecules inhabiting the same or different microorganisms (even though preferentially phylogenetically close); the introgression of new genes into the initial plasmid backbone might result in an increase of plasmid size. This, in turn, might also increase the probability of interaction (and possibly of gene exchange) between plasmids and also larger DNA molecules (chromids and/or chromosomes), giving rise to mosaic-like structure of plasmids. The interaction between plasmids and other DNA molecules might have facilitated by the presence of MGE, such as transposons. The increase of plasmid size would also increase the rate of intra-molecular duplication events, in addition to inter-molecular paralogs formation. The finding that the five G4 plasmids share the highest percentage of paralogs with the secondary chromosome (chromosome 3) might support the idea that secondary chromosomes might have been originated from the acquisition of plasmid genes [17]. The increase of plasmid size might be also due to (and/or might be facilitated by) the introgression of larger DNA segments embedding entire gene clusters and/or operons. This is in agreement with data shown in Table 2 and in Supplementary Figs. 4–6, where the number and the type of most paralogous gene clusters/operons are reported. The analysis of *B. vietnamiensis* plasmid paralogous operons revealed that most of intra- and interplasmids paralogous operons included genes related to DNA transposition/mobilization (29%). However, clusters/operons shared by plasmids and/or by plasmids and chromosomes include also genes involved in general metabolic functions, transport and DNA binding (39%) and genes coding for proteins with an unknown function (32%). Particularly interesting is the finding of a cluster of 15 genes (Supplementary Fig. S6), which has been detected in four copies in plasmids pBVIE01, pBVIE03, chromosome 1 and chromosome 3 containing genes involved in resistance to chromate. The genes belonging to these paralogous family are connected also at a threshold of 100%, suggesting very recent duplication events. It is also evident that cluster/operon duplication events can occur not only between different molecules inhabiting the same cytoplasm, but also between (at least) plasmids harboured by different strains belonging to different species. The introgression of entire cluster/operons might be particularly important for the spreading of entire metabolic pathways. The importance of operon duplication in the origin and evolution of metabolic pathways has been already recognized ([19,20,25] and references therein). Indeed, since most, if not all, of operons embed the entire set of genes involved in a metabolic route, their duplication and spreading through horizontal gene transfer events (mostly mediated by MGE) may facilitate their dissemination in the microbial world and the gaining of new and diverse metabolic abilities.

Lastly, data obtained in this work are in agreement with a recently proposed model for operon formation and propagation [37]. According to this idea (the so-called Scribbling Pad hypothesis), plasmids have been used by bacteria for "*genetic experimentation and, in particular, for the construction of operons*". In our opinion, the finding that entire gene clusters and/or operons are frequently re-shuffled within the same plasmid and/or between different plasmids, between plasmids and chromosomes, and between *B. vietnamiensis* G4 plasmids and DNA molecules of other bacteria, as well as the finding that gene clusters located in one (or more) *B. vietnamiensis* G4 plasmid(s) embedded genes that are scattered on other DNA molecules (data not shown), strongly support this hypothesis.

4. Materials and methods

4.1. Sequence data source

The dataset used in this work embeds all the proteins encoded by the completely sequenced *B. vietnamiensis* G4 genome (three chromosomes and five plasmids) that were retrieved from the NCBI ftp websites ftp://ftp.ncbi.nih.gov/genomes/Bacteria/ and ftp://ftp.ncbi.nih.gov/refseq/release/plasmid (Table S1).

4.2. Networks construction

Similarity, identity based, networks were constructed using the tools implemented in the Blast2Network (B2N) software as described elsewhere [7]. Briefly, a file containing protein sequences in standard NCBI fasta format was used as an input to gather information on source sequences from the NCBI website. Input sequences identity was then analysed one against each other using BLAST [1]. B2N transforms a

BLAST output file into a (sequence similarity) network in a Visone readable format (http://visone.info/), a freely available software for network visualization and analysis. In this similarity network the nodes represent proteins whereas the links indicate the existence of a given degree of sequence similarity between them. Moreover, in the resulting network, all the nodes belonging to the same plasmid source are circularly arranged and filled with the same colour.

4.3. Analysis of plasmid genes origin

In order to identify the most likely source molecule (either chromosomal, plasmid or viral) of *B. vietnamiensis* G4 genes we adopted a similarity-oriented computational pipeline developed by Bosi et al. [6]. Briefly, each of the ORFs was used as a query for a BLAST search against three different databases, each of which embedding 100,000 sequences retrieved from NCBI plasmids, phages and chromosomes, respectively. For each BLAST search, only the Best BLAST Hit (BBH) was considered, in order to reduce any possible bias due to the presence of closely related sequences in the database that would falsely increase the number of homologs for a given ORF. This strategy was repeated 100 times for each sequence and, for each of the 100 runs, new plasmid, chromosome and viral databases were assembled, randomly sampling 100,000 sequences from the NCBI database. Finally, the putative source molecule was identified according to the database (chromosome, plasmid or phage) that produced the highest number of best hits after the 100 BLAST probing.

4.4. Functional assignment

The software Blast2GO (version 2.3.4) [10] was used, with default parameters, to obtain the functional annotation of the plasmid genes as well as the related gene ontology (GO) terms. Blast2GO was also used for GO functional enrichment analysis of genes, by performing Fisher's exact test with robust false discovery rate (FDR) correction to obtain an adjusted p-value between certain test gene groups and the whole annotation.

4.5. Circos software

This software allows the visualization of the data information in a circular layout consisting of a set of concentric circles. It was used for the construction of Fig. 6, in order to show graphically molecular habitat and putative origin of *B. vietnamiensis* G4 plasmid genes. Furthermore it was used to map them on the plasmid backbone, considering also the GC content of each of the plasmids, which was analysed using a 350 bp window. The software is available at (http://circos.ca/software/).

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2014.02.004.

Acknowledgments

Marco Fondi and Elena Perrin are financially supported by a FEMS Advanced Fellowship (FAF2012) and a "Buzzati-Traverso" Foundation fellowship, respectively. Part of this work was presented at the International Plasmid Biology Conference 2012 held in Santander (Spain), 12–16 September 2012. Lastly, we are very grateful to the two anonymous referees for their suggestions in improving the manuscript.

References

[1] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.

[2] C.P. Andam, J.P. Gogarten, Biased gene transfer and its implications for the concept of lineage, Biol. Direct 6 (2011) 47.

[3] D.I. Andersson, D. Hughes, Gene amplification and adaptive evolution in bacteria, Annu. Rev. Genet. 43 (2009) 167–195.

[4] P.M. Bennett, Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria, Br. J. Pharmacol. 153 (Suppl. 1) (2008) S347–S357.

[5] S.D. Bentley, J. Parkhill, Comparative genomic structure of prokaryotes, Annu. Rev. Genet. 38 (2004) 771–792.

[6] E. Bosi, R. Fani, M. Fondi, The mosaicism of plasmids revealed by atypical genes detection and analysis, BMC Genomics 12 (2011) 403.

[7] M. Brilli, A. Mengoni, M. Fondi, M. Bazzicalupo, P. Lio, R. Fani, Analysis of plasmid genes by phylogenetic profiling and visualization of homology relationships using Blast2Network, BMC Bioinforma. 9 (2008) 551.

[8] M.A. Cevallos, R. Cervantes-Rivera, R.M. Gutierrez-Rios, The repABC plasmid family, Plasmid 60 (2008) 19–37.

[9] T. Coenye, P. Vandamme, Diversity and significance of *Burkholderia* species occupying diverse ecological niches, Environ. Microbiol. 5 (2003) 719–729.

[10] A. Conesa, S. Gotz, J.M. Garcia-Gomez, J. Terol, M. Talon, M. Robles, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, Bioinformatics 21 (2005) 3674–3676.

[11] B.A. Conway, E.P. Greenberg, Quorum-sensing signals and quorum-sensing genes in *Burkholderia vietnamiensis*, J. Bacteriol. 184 (2002) 1187–1191.

[12] V.S. Cooper, S.H. Vohr, S.C. Wrocklage, P.J. Hatcher, Why genes evolve faster on secondary chromosomes in bacteria, PLoS Comput. Biol. 6 (2010) e1000732.

[13] T. Dagan, Y. Artzy-Randrup, W. Martin, Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution, Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 10039–10044.

[14] T. Dagan, W. Martin, The tree of one percent, Genome Biol. 7 (2006) 118.

[15] T. Dagan, W. Martin, Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 870–875.

[16] E. Desmond, S. Gribaldo, Phylogenomics of sterol synthesis: insights into the origin, evolution, and diversity of a key eukaryotic feature, Genome Biol. Evol. 1 (2009) 364–381.

[17] E.S. Egan, M.A. Fogel, M.K. Waldor, Divided genomes: negotiating the cell cycle in prokaryotes with multiple chromosomes, Mol. Microbiol. 56 (2005) 1129–1138.

[18] R. Fani, Gene duplication, ASM Press, Washington, D. C., 2004

[19] R. Fani, The origin and evolution of metabolic pathways: why and how did primordial cells construct metabolic routes? Evol. Educ. Outreach 5 (2012) 367–381.

[20] R. Fani, M. Fondi, Origin and evolution of metabolic pathways, Phys. Life Rev. 6 (2009) 23–52.

[21] R. Fernandez-Lopez, M.P. Garcillan-Barcia, C. Revilla, M. Lazaro, L. Vielva, F. de la Cruz, Dynamics of the IncW genetic backbone imply general trends in conjugative plasmid evolution, FEMS Microbiol. Rev. 30 (2006) 942–966.

[22] W.M. Fitch, Distinguishing homologous from analogous proteins, Syst. Zool. 19 (1970) 99–113.

[23] M. Fondi, G. Bacci, M. Brilli, M.C. Papaleo, A. Mengoni, M. Vaneechoutte, L. Dijkshoorn, R. Fani, Exploring the evolutionary dynamics of plasmids: the *Acinetobacter* pan-plasmidome, BMC Evol. Biol. 10 (2010) 59.

[24] M. Fondi, G. Emiliani, R. Fani, Origin and evolution of operons and metabolic pathways, Res. Bicrobiol. 160 (2009) 502–512.

[25] M. Fondi, G. Emiliani, P. Lio, S. Gribaldo, R. Fani, The evolution of histidine biosynthesis in Archaea: insights into the his genes structure and organization in LUCA, J. Mol. Evol. 69 (2009) 512–526.

[26] M. Fondi, R. Fani, The horizontal flow of the plasmid resistome: clues from inter-generic similarity networks, Environ. Microbiol. 12 (2010) 3228–3242.

[27] M.R. Fries, L.J. Forney, J.M. Tiedje, Phenol- and toluene-degrading microbial populations from an aquifer in which successful trichloroethene cometabolism occurred, Appl. Environ. Microbiol. 63 (1997) 1523–1530.

[28] L.S. Frost, R. Leplae, A.O. Summers, A. Toussaint, Mobile genetic elements: the agents of open source evolution, Nat. Rev. Microbiol. 3 (2005) 722–732.

[29] F.A. Gonzalez, E. Bonapace, I. Belzer, I. Friedberg, L.A. Heppel, Two distinct receptors for ATP can be distinguished in Swiss 3 T6 mouse fibroblasts by their desensitization, Biochem. Biophys. Res. Commun. 164 (1989) 706–713.

[30] S. Halary, J.W. Leigh, B. Cheaib, P. Lopez, E. Bapteste, Network analyses structure genetic diversity in independent genetic worlds, Proc. Natl. Acad. Sci. U. S. A. 107 (2010) 127–132.

[31] P.W. Harrison, R.P. Lower, N.K. Kim, J.P. Young, Introducing the bacterial 'chromid': not a chromosome, not a plasmid, Trends Microbiol. 18 (2010) 141–148.

[32] T.G. Lessie, W. Hendrickson, B.D. Manning, R. Devereux, Genomic complexity and plasticity of *Burkholderia cepacia*, FEMS Microbiol. Lett. 144 (1996) 117–128.

[33] W.H. Li, D. Graur (Eds.), Fundamentals of Molecular Evolution, 1991.

[34] M. Lynch, The frailty of adaptive hypotheses for the origins of organismal complexity, Proc. Natl. Acad. Sci. U. S. A. 104 (Suppl. 1) (2007) 8597–8604.

[35] I. Maida, M. Fondi, M.C. Papaleo, E. Perrin, R. Fani, The gene flow between plasmids and chromosomes: insights form bioinformatic analyses, Open Appl. Inform. J. 5 (2011) 62–76.

[36] M.J. Nelson, S.O. Montgomery, W.R. Mahaffey, P.H. Pritchard, Biodegradation of trichloroethylene and involvement of an aromatic biodegradative pathway, Appl. Environ. Microbiol. 53 (1987) 949–954.

[37] V. Norris, A. Merieau, Plasmids as scribbling pads for operon formation and propagation, Res. Microbiol. 164 (2013) 779–787.

[38] M. Tamminen, M. Virta, R. Fani, M. Fondi, Large-scale analysis of plasmid relationships through gene-sharing networks, Mol. Biol. Evol. 29 (2012) 1225–1240.

[39] W. Tian, J. Skolnick, How well is enzyme function conserved as a function of pairwise sequence identity? J. Mol. Biol. 333 (2003) 863–882.

[40] C. Woese, The universal ancestor, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 6854–6859.

[41] C.R. Woese, Interpreting the universal phylogenetic tree, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 8392–8396.

[42] C.R. Woese, On the evolution of cells, Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 8742–8747.