# The geographic impact on genomic divergence as revealed by comparison of nine Citromicrobial genomes

Qiang Zheng, Yanting Liu, Christian Jeanthon, Rui Zhang, Wenxin Lin, Jicheng Yao, Nianzhi Jiao

# Geographic impact on genomic divergence as revealed by comparison of nine Citromicrobial genomes

Qiang Zheng[1], Yanting Liu[1], Christian Jeanthon[2, 3], Rui Zhang[1], Wenxin Lin[1], Jicheng Yao[4] and Nianzhi Jiao[1]

[1]State Key Laboratory for Marine Environmental Science, Institute of Marine Microbes and Ecospheres, Xiamen University, Xiamen 361102, People's Republic of China.

[2]CNRS, UMR 7144, Marine Phototrophic Prokaryotes Team, Station Biologique de Roscoff, F-29680 Roscoff, France.

[3]Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Oceanic Plankton Group, Station Biologique de Roscoff, F-29680 Roscoff, France.

[4]Shanghai Personal Biotechnology Limited Company, 218 Yindu Road, Shanghai, 200231, People's Republic of China.

Author for correspondence:

Qiang Zheng: zhengqiang@xmu.edu.cn

Nianzhi Jiao: jiao@xmu.edu.cn

# Abstract

Aerobic anoxygenic phototrophic bacteria (AAPB) are thought to be important players in oceanic carbon and energy cycling in the euphotic zone of the ocean. The genus *Citromicrobium*, widely found in oligotrophic oceans, is a member of marine alphaproteobacterial AAPB. Nine *Citromicrobium* strains isolated from the South China Sea, the Mediterranean Sea or the tropical South Atlantic were found to harbor identical 16S rRNA sequences. The sequencing of their genomes revealed high synteny in major regions. Nine genetic islands (GIs), involved mainly in type IV secretion systems, flagellar biosynthesis, prophage and integrative conjugative elements, were identified by a fine scale comparative genomics analysis. These GIs played significant roles in genomic evolution and divergence. Interestingly, the co-existence of two different photosynthetic gene clusters (PGCs) was not only found in the analyzed genomes but also confirmed, for the first time, in environmental samples. The prevalence of the coexistence of two different PGCs may suggest an adaptation mechanism for *Citromicrobium* members to survive in the oceans. Comparison of genomic characteristics (e.g., GIs, ANI, SNPs and phylogeny) revealed that strains within a marine region shared a similar evolutionary history that was distinct from that of strains isolated from other regions (South China Sea vs Mediterranean Sea). Geographic differences are partly responsible for driving the observed genomic divergences, and allow microbes to evolve through local adaptation. Three *Citromicrobium* strains isolated from the Mediterranean Sea diverged millions of years ago from other strains, and evolved into a novel group.


# Importance

Aerobic anoxygenic phototrophic bacteria are a widespread functional group in the upper ocean, and their abundance could be up to 15% of total heterotrophic bacteria. To date, a great number of studies display their biogeographic distribution patterns in the ocean, however little is understood about the geographic isolation impact on genome divergence of marine AAPB. In this study we compare nine *Citromicrobium* genomes of strains with identical 16S rRNA sequences but from

different ocean origins. Our results reveal that strains isolated from the same marine region share a similar evolutionary history that is distinct from that of strains isolated from other regions. These *Citromicrobium* strains diverged millions of year ago. In addition, the co-existence of two different PGCs is prevalent in the analyzed genomes and in environmental samples.

**Keywords**

Aerobic Anoxygenic Phototrophic Bacteria (AAPB), *Citromicrobium*, Comparative Genomes, Genetic Islands, Evolutionary Divergence, Photosynthetic Gene Cluster

## Introduction

Aerobic anoxygenic phototrophic bacteria (AAPB) are a widespread functional microbial group in the euphotic zone of the ocean and are thought to play important roles in the cycling of marine carbon and energy (1-10). AAPB can harvest light using bacteriochlorophyll *a* (BChl *a*) and various carotenoids to form a trans-membrane proton gradient for the generation of ATP (1, 11). The photosynthetic process is performed by a series of photosynthetic operons including *bch*, *crt*, *puf*, *puh* and some regulatory genes, forming a highly conserved 40 to 50 kb 'photosynthetic gene cluster' (PGC) (12, 13). The heart part of anoxygenic photosynthesis is the reaction center, encoded by the *puf* and *puh* operons (14).

The genus *Citromicrobium* belonging to the order *Sphingomonadales* in the class *Alphaproteobacteria*, is a member of the marine AAPB (15-17). Since the initial isolation of the type species, *Citromicrobium bathyomarinum* strain JF-1, from deep-sea hydrothermal vent plume waters in the Juan de Fuca Ridge (Pacific Ocean), dozens of *Citromicrobium* strains were isolated from depth of 500 to 2,379 m near this region (17, 18). Genomic analyses showed that members of the genera *Citromicrobium* and *Erythrobacter* generally contained the shortest and simplest PGC structure among all known AAPB (16). Analysis of the *C. bathyomarinum* JL354 genome led to the discovery that two different PGCs could coexist in one bacterium, with one complete cluster and the other cluster incomplete (15, 16). Horizontal gene transfer (HGT) was detected to mediate the incomplete PGC acquisition, and multiple mechanisms mediating HGT were also found through studying its genome, including gene transfer agent (GTA), prophage, integrative conjugative element (ICE) and the type IV secretion system (T4SS). *Citromicrobium* species may benefit from obtaining genes by HGT to compete and survive in natural environments. Additionally, genes encoding xanthorhodopsin which is a light-driven proton pump like bacteriorhodopsin (19-21), but more effective at collecting light, are also found in Citromicrobial genomes (22, 23).

A recent pyrosequencing analysis of *pufM* genes showed that *Citromicrobium*-like AAPB were mainly distributed in the oligotrophic ocean, and were relatively more abundant in the upper

twilight zone (150-200 m depth) than in the subsurface waters (5 m and 25 m) of the western Pacific Ocean (24). However, in this study, no sequences belonging to the incomplete PGC were detected, although the *Citromicrobium-like pufM* gene belonging to the complete PGC showed a high relative abundance.

Fortunately, a great number of Citromicrobial strains were isolated from the Mediterranean Sea, the South China Sea and the South Atlantic Ocean (16, 25). The Mediterranean Sea is an almost completely enclosed sea connected with the eastern North Atlantic Ocean by the narrow Strait of Gibraltar. The South China Sea is the largest marginal sea in the western Pacific Ocean, extending from subtropical to tropical zones. The sampled area in the South Atlantic Ocean (13.857°S, 26.018°W) is a region of tropical open oceans. Although a great number of studies showed that microbes displayed biogeographic patterns in the ocean (26, 27), little is understood about the geographic isolation impact on genome divergence of marine microbes.

The aim of this study is (i) to illustrate the evolutionary divergence of *Citromicrobium* genomes with identical 16S rRNA sequences but different geographical origins and (ii) to demonstrate the prevalence of the coexistence of two PGCs in these strains and within the marine environment.

## Materials and Methods

### Isolation of Citromicrobia

Strains JL31 (24.458°N, 118.247°E), JL354 (21.684°N, 112.918°E), JL477 (22.167°N, 115.153°E), JL1351 (17.994°N, 120.287°E) and WPS32 (17.000°N, 115.000°E) were isolated from euphotic waters from the South China Sea on plates containing Rich Organic (RO) medium (**Table 1**) (17). Strain JL2201 was isolated from South Atlantic surface water (13.857°S, 26.018°W) on RO medium plate (**Table 1**). Strains RCC1878 and RCC1885 were isolated from Ionian Sea (at middle of Mediterranean Sea) surface water (34.133°N, 18.450°E) on MiA and MAD plates, respectively, and strain RCC1897 was isolated from western Mediterranean Sea (38.633°N, 7.917°E) on MiA plate (**Table 1**) (25).

## Genome sequencing and assembly

Whole-genome sequencing of *Citromicrobium* sp. JL354 was performed by 454 as previously reported (15, 16). The genomes of JL31, JL477, JL1351, JL2201, WPS32, RCC1878, RCC1885 and RCC1897 were obtained by Illumina MiSeq system. Paired-end reads of average 250-bp length were assembled using Velvet software (v2.8) (28). The sequencing coverage ranged from 155x (JL1351) to 440x (RCC1897). The genome of strain JL477 has been completed, and the other seven genomes each possessed 14-22 contigs (**Table 1**).

## Gene prediction and annotation

The open reading frames (ORFs) were analyzed using a combination of Glimmer 3.02 (29) and GeneMark (30, 31). All predicted ORFs were then annotated using the NCBI Prokaryotic Genome Annotation Pipeline (32) and Rapid Annotation using Subsystem Technology (RAST) (33). rRNA identification was performed with RNAmmer 1.2 software (34), and tRNAscan-SE (v1.21) was used to identify the tRNA genes (35).

The genomic average nucleotide identity (ANI) was calculated by online JSpecies Web service (http://jspecies.ribohost.com/jspeciesws) (36).

## Core genome and pan-genome analyses

Orthologous clusters (OCs) were assigned by grouping all protein sequences from the nine genomes using OrthoMCL based on their sequence similarity (E-value < $10^{-5}$, >50% coverage) (37). The core and pan-genomes were analyzed according to the method described by Tettelin et al. (38). The functional proteins were classified by comparison with the COG (Cluster of Orthologous Genes) databases (39).

## SNP discovery

SNPs were detected by sequence comparisons of the 9 *Citromicrobium* genomes using MUMmer (40). Because 8 out of 9 genomes were draft genomes, the positions of SNPs from all genomes

relative to the sequence of strain JL477 were recorded. Paralogous genes and repeated regions were removed from our analysis. The synonymous SNPs of coding regions were used to roughly estimate the pairwise strain divergence time (41).

**Sequence comparison**

The genome of JL477 was compared to the other eight Citromicrobial genomes in silico using the Blast Ring Image Generator (BRIG) software (42). Regions with nucleotide sequence similarity above 70% are shown on the map. Nine genetic islands (GIs) were identified from the comparative map. The upstream and downstream regions of each GI were then retrieved and manually searched for the presence of conserved regions or signature genes (such as tRNA). Some GIs and their flanking genes from different genomes were chosen for pairwise comparison (**Table S1**). Although there were eight draft genomes with a number of contigs involved in analyses, the completeness for each genome was more than 99% as result of the high genome sequencing coverage. Gaps between contigs usually were intergenic regions or didn't contain more than three genes (**Table S1**). All gene losses (especially more than 10 kb fragment) occurred inside contigs but not between contigs (**Table S1**).

**Phylogenetic analysis**

All *pufM* gene sequences collected from NCBI database, *Citromicrobium* genomes and environmental samples were aligned using Clustal X and phylogenetic trees were constructed using the maximum likelihood and neighbour-joining algorithms of MEGA 6 software (43).The phylogenetic trees were supported by bootstrap for resampling test with 100 and 1000 replicates for the maximum likelihood and neighbour-joining algorithms, respectively.

**Environmental sample collection**

Seawater samples were collected on board during a western Pacific Ocean cruise in July 2011. Seawater was collected at two stations (P3 [129.00°E, 14.00°N] and P10 [130.00°E, 2.00°N]) and

five depths (5 m, 25 m, 75 m, 150 m and 200 m). For each sample, 2-3 L of seawater was prefiltered through a 20-μm filter, and the microorganisms were then collected onto 0.22-μm-pore-size polycarbonate filters (Millipore). Nucleic acids were extracted using hot sodium dodecyl sulphate, phenol, chloroform and isoamyl alcohol (24, 44). The high-quality DNA was stored at -20°C for future use.

### Sequence generation and processing

Considering that previous primers (2, 45) had five mismatches with *pufM* sequences belonging to the incomplete PGC in *Citromicrobium*, the following primer set was used to amplify the environmental DNA: *pufM*_Citro forward (5′-TACGGSAAYTTSTWCTAC-3′) and *pufM*_Citro reverse (5′-GCRAACCAGYANGCCCA-3′). High throughput sequencing of *pufM* gene (~240 bp) was performed using Illumina MiSeq technology. The generated high throughput sequence data were processed as described in Zheng et al (24). Briefly, after quality control, all sequences were grouped into operational taxonomic units (OTUs) using a 6% cutoff. One representative sequence for each OTU was chosen to perform local BLAST against our *pufM* sequence database (for details see Zheng et al (24)).

### Accession numbers

The complete JL477 genome sequence is available under GenBank accession number CP011344. Whole genome sequences of strains JL31, JL354, JL1351, WPS32, JL2201, RCC1878, RCC1885 and RCC1897 are available under GenBank accession numbers LAIH00000000, ADAE00000000, LAPR00000000, LAPS00000000, LARQ00000000, LARQ00000000, LBLY00000000 and LUGI01000000, respectively.

All environmental *pufM* sequences obtained in this study have been submitted to the MG-RAST public database (http:// metagenomics.anl.gov/) under ID number: 4653301.3.

## Results and Discussion

## Overview of nine *Citromicrobium* spp.

Nine *Citromicrobium* sp. strains were used to perform comparative genome analyses, which were isolated from the South China Sea (strains JL31, JL354, JL477, JL1351 and WPS32), the Mediterranean Sea (strains RCC1878, RCC1885 and RCC1897) and the tropical South Atlantic (strain JL2201) (**Table 1**). Although 1-2 base mismatches were found in the 16S rRNA sequences collected from GenBank database (strains JL354, RCC1878, RCC1885 and RCC1897) or after Sanger sequencing (the other five strains), the 16S rRNA sequences (1442 bp, one copy per genome) extracted from the nine genomes with high sequencing coverage were identical. The Sanger sequencing might induce some biases or mismatches during the 16S rRNA amplification and sequencing PCR steps.

The nine genomes displayed highly similar genomic characteristics, in terms of genome size (from 3.16 to 3.28 Mb), GC content (from 64.8 to 65.1%), gene number (from 3,056 to 3,250), COGs (from 1,934 to 2,010) and tRNA number (44 or 45) (**Table 1**).

## The pan- and core genomes of the *Citromicrobium* strains

Based on the total set of genes from the 9 sequenced strains, the *Citromicrobium* pan-genome consisted of 3,546 predicted orthologous clusters (OCs), with a conserved core genome of 2,691 OCs. The cumulative length of the core genome was approximately 2.50 Mbp, which covered >75% of each genome. The flexible genome comprises 853 OCs including 362 unique OCs and 490 shared by more than one strains but not all strains.

The core genome is mainly involved in housekeeping functions and central metabolism, from the Calvin cycle to the TCA cycle. Approximately 80% of predicted core genes are assigned to COG functional categories. The predicted core genes contain a relatively high percentage of genes assigned to the following COG categories: general function prediction only (R), amino acid transport and metabolism (E), function unknown (S), translation, ribosomal structure and

biogenesis (J), energy production and conversion (C), lipid transport and metabolism (I), cell wall/membrane/envelope biogenesis (M) and inorganic ion transport and metabolism (P). Due to a larger fraction of putative or hypothetical genes, only 36.8% of flexible genes are assigned to COG functional categories. Compared to the core genes, they include an overrepresentation of genes assigned to the following COG categories: general function prediction only (R), lipid transport and metabolism (I), replication, recombination and repair (L), intracellular trafficking, secretion, and vesicular transport (U), secondary metabolites biosynthesis, transport and catabolism (Q), cell motility (N), transcription (K), and defense mechanisms (V). Most of flexible genes were sourced from the genetic island regions.

**The genomic Average Nucleotide Identity**

The Average Nucleotide Identity (ANI) shared between genome pairs ranged from 95.96% to 100% (**Table S2**). Five genomes, JL31, JL1351, JL354, JL477 and JL2201 share more than 99.5% ANI between them but share lower values with three RCC strains (from 95.96 to 96.47%) and WPS32 (from 96.39 to 96.44%). The three RCC strains had the lowest percentages of all the genomes involved in pairwise comparisons (**Table S2**).

Genome pairs JL31 and JL1351, JL31 and JL2201, JL1351 and JL2201, JL477 and JL354, RCC1878 and RCC1885, RCC1878 and RCC1897, and RCC1885 and RCC1897 showed strikingly high ANI (almost 100%) (**Table S2**). Among them, genome pairs JL31 and JL1351, JL477 and JL354, RCC1878 and RCC1885, RCC1878 and RCC1897, and RCC1885 and RCC1897 showed high genomic percentages (>98.0%) involved in pairwise comparisons (**Table S2**), indicating closer evolutionary relationships with each other.

The proposed cut-off of the ANI between two genome sequences for a species boundary is 95-96% (36). Concerning five JL strains share 95.96~96.47% ANI with three RCC strains, therefore, three RCC strains isolated from the almost enclosed Mediterranean Sea have diverged for a long history from other strains, and tended to evolve into a novel group. However, all nine *Citromicrobium* strains have identical 16S rRNA sequences. This emphasizes that traditional diversity studies,

which classify sequences into operational taxonomic units based on the nucleotide sequence similarity, underestimate real environmental microbial information. The classification and diversity results based on environmental 16S rRNA couldn't link to *in situ* microbial functions (46, 47).

**Comparison of nine genomes**

Comparison of all nine genomes (JL477 versus the others) showed high synteny of major regions, and a significantly high level of sequence conservation (**Figure 1, Table S1**). DNA fragment insertions and deletions were detected in genome comparison (one versus other eight) (**Table S1**).

Horizontal gene transfer (HGT) is common in bacteria, contributing to the genomic plasticity and possibly to environmental adaptation (48). To better understand genome plasticity and unique genome characteristics, nine specific genomic regions larger than 10 kb (except GI07 with 9.7 kb) in size were identified based on the comparative genome map (**Figure 1**). They were absent or different in the corresponding regions of the eight other genomes (JL477 versus the others) and designated here as genomic islands (GI) GI01–GI09 (**Figure 1**). These nine GIs contribute approximately 12% each of genome size. Almost all the GIs were regarded as originating from HGT (gene gain and loss) mediated by the transposases, integrases and conjugal-transfer systems, and five of them (GI03, GI04, GI07, GI08 and GI09) were flanked by a tRNA gene. Previous studies showed that GIs frequently originate from integration events that associated with tRNA-encoding genes (49-51). These nine GIs are scattered throughout the genomes, and their general features and sequence information are summarized in **Table 2**.

**GI01, Type IV Secretion System.** GI01 mainly consists of a *trb* gene cluster, *trbBCDEJLFGI*, which is probably involved in the conjugal transfer of mobile genetic elements mediated by the Type IV Secretion System (T4SS) (51-53). In the genomes of strains JL354, RCC1878, RCC1885 and RCC1897, the highly homologous gene cluster (here denoted T4SS-I) is detected at the same chromosome position as in strain JL477 (**Figure 2A**), which is flanked by putative genes for T4SS protease (*traF*), relaxase (*virD2*) and ATPase for T-DNA transfer (*virD4*) in the upstream regions,

and for genes associated with amino acids metabolism, transmembrane transport and transcriptional regulation in the downstream region.

Interestingly, the same flanking gene organization was found in the genomes of strains JL31, JL1351 and JL2201, but with gene fragment loss in the middle of two genes (genes 7 and 8) (**Figure 2B**). There is a 770 bp deletion in the latter part of gene 7 and a 670 bp deletion in the front part of gene 8, which indicates a large DNA fragment deletion in these three genomes.

In the genomes of strains JL31, JL1351 and JL2201, a *trb* gene cluster (here denoted T4SS-II) that located in an integrase-mediated foreign DNA fragment, was also found (**Figure 2C**). The average nucleotide identity between T4SS-I and T4SS-II was low (< 50%), indicating that the T4SS-II gene cluster was acquired via HGT mediated by the integrase. In addition, a three-gene cluster coding for the Type I Restriction-modification System (TIRS) which protects microbes from the foreign DNA (e.g., bacteriophage) was detected in the integrated DNA fragment. The inserted sequence is adjacent to the tRNA-CCG gene in these three genomes. In the genomes of strains RCC1878, RCC1885, RCC1897 and WPS32, HGT derived from integration events also occurred adjacent to tRNA-CCG gene (**Figure 2D, 2E**). However, different inserted gene clusters were found in these three genomes. A TIRS gene cluster was also observed in the genomes of strains RCC1878, RCC1885 and RCC1897, but TIRS gene clusters from three RCC strains share less than 50% nucleotide identity with strains JL31, JL1351, and JL2201 (**Figure 2D**). Two genes homologous to *trwC* and *traD*, both involved in conjugative transfer, were found next to the TIRS in the inserted sequence of strains RCC1878, RCC1885 and RCC1897 (**Figure 2D**). The inserted sequence flanking tRNA-CCG in the genome of strain WPS32 contains a few genes involved in restriction-modification, anti-restriction and several hypothetical genes (**Figure 2E**). No inserted gene were found around tRNA-CCG in the genomes of JL477 and JL354 (**Figure 2F**). This suggests that different foreign DNA fragments are independently integrated into the same tRNA gene, which contributes to bacterial genome evolution and species divergence (50). No *trb* gene cluster was found in the genome of strain WPS32.

**GI02, Gene Transfer Agent.** A gene transfer agent (GTA) is an unusual bacteriophage-like

element of genetic exchange that transfers a random host genomic DNA fragment (4-14 kb in size) between closely related bacteria (54, 55). Analysis of Citromicrobial genomes found GTA gene cluster present in all nine genomes at the same chromosome position (**Figure S1**). The structure and composition of the GTA gene cluster and flanking genes are identical in all the genomes except for that of strain WPS32 (**Figure S1B**). For example, an approximately 3 kb DNA fragment mediated by transposase is inserted in the front of GTA in all genomes but was absent in strain WPS32. We speculate that in this strain, the transposase, after acquisition, mediated the gene loss of the downstream region including the ORFs of the GTA.

**GI03, Flagella and Motility.** Flagella support marine bacterial motility, and allow cells to move toward favorable living conditions in the environment, e.g., rich nutrient and light (56-58). Sometimes, flagella also contribute to adhesion (56, 59). In some special cases, flagella might also provide an advantage for bacterial competition (56, 57). Two gene clusters for flagellum biosynthesis were found in the nine genomes (**Figure 3**). The first cluster (flagella I) was common in all genomes (**Figure 3A**) while the second (35.5 kb in size, flagella II) is detected in only five genomes, JL31, JL354, JL477, JL2201, and JL1351 (**Figure 3B**). Flagella I mainly consists of two large gene clusters (*flgBCDEFGHIJKL* and *fliEFGHIJKLMNOPQR*), *motAB*, *fliDS* and some regulatory genes (**Figure 3A**). The organization of flagella II is irregular (**Figure 3B**).

An integrase mediates the acquisition of the flagella II gene cluster, and the integration event occurs adjacent to the tRNA-GGA gene. In the other four genomes, the inserted DNA sequences were also found at the same position. In the three RCC genomes, a 17.1 kb inserted fragment was detected, and only a few genes could be annotated as encoding a known function (*traG* and DNA-invertase) (**Figure 3C**). In the WPS32 genome, genes involved in the serine-glyoxylate cycle and respiration-related were found at the same position (**Figure 3D**).

**GI04.** GI04, the longest GI at 101.1 kb in size, and is mainly involved in choline and betaine uptake as well as metabolism of glycerolipids, glycerophospholipids, fatty acids and pyruvate. This large DNA fragment was integrated into JL354 and JL477 chromosomes via an integrase flanked by the tRNA-GCT gene. The other genomes except for WPS32 have the same flanking

366 genes with no GI04 sequences. GI04 contains 88 genes, 18 of which have unknown functions. The
367 GC content of GI04 (62.14%), is lower than the genomic GC content.
368
369 In the WPS32 genome, an approximately 47.9 kb inserted DNA fragment was also found at the
370 same chromosome position adjacent to the tRNA-GCT gene, and its acquisition was mediated by
371 the integrase. It also contains several fatty acid metabolism related genes, but with significantly
372 lower sequence identity (or different genes) than in strains JL477 and JL354.
373
374 **GI05.** GI05 (approximately 11.3 kb) displayed the lowest GC content (52.08%). Only found in the
375 genomes of strains JL31, JL354, JL477, JL1351 and JL2201, it contains four genes, and its coding
376 region represents less than 50% of its sequence. The only known function was a DNA polymerase
377 of family B.
378
379 **GI06, Mu-like prophage.** In a previous study, we isolated one inducible bacteriophage from
380 strain JL354, consisting of three parts: an early expression region and regions encoding heads and
381 tails (60). Nearly identical prophage sequences were observed in the genomes of strains JL354,
382 JL477 and JL2201. They share high levels of structural conservation and sequence identity and are
383 defined here as prophage type I (**Figure S2**). In addition, another type of prophage (here defined
384 as prophage type II) was found in strains JL2201, RCC1878, RCC1885 and RCC1897. The type II
385 prophage has structure modules similar to those of type I (**Figure S2**), but they share significantly
386 lower sequence identity.
387
388 All three type I prophages share the same upstream and downstream genes, and the first
389 downstream gene is a transposase. Type II prophages integrated into the host chromosome in a
390 different position with type I prophage. This indicates that these two types of prophages originate
391 from different integration events.
392
393 Two types of prophages co-exist in the genome of strain JL2201. The type I prophage in strain
394 JL2201 is identical to the prophage found in strains JL354 and JL477. However, the type II
395 prophage in strain JL2201 lost its early expression region but kept the structural genes encoding

heads and tails (**Figure S2**). Interestingly, the structural genes form duplication (approximately 26 kb × 2) centers around the last gene with less than 93% nucleotide identity (**Figure S2**). The incomplete duplicated prophage sequence might contribute to increase more viral particles production under the control of the early expression genes of prophage I in strain JL2201.

**GI07.** GI07, located between the tRNA-TGG and tRNA-CAT genes, is the shortest length (approximately 9.7 kb) among all GIs. Its GC content (54.39%) is much lower than the genomic GC content (64.8-65.1%). It contains three genes, an integrase, a hypothetical gene, and a reverse transcriptase, and the coding sequences represent approximately 50% of its length. The gene organization and composition of GI07 in the five JL genomes is identical.

In the three RCC strains, the inserted foreign DNA (approximately 11.3 kb with 60.67% GC content) is after the tRNA-TGG gene. It consists of nine genes mainly involved in the type I restriction-modification system, flavodoxin reductase, and fatty acid metabolism. The corresponding region was not detected in the WPS32 genome.

**GI08, integrative and conjugative elements.** Integrative and conjugative elements (ICEs) are defined as self-transmissible mobile genetic elements with the capacity to integrate into and excise from a host chromosome (61, 62). The core ICEs are made up of three typical genetic modules: ICE integration and excision, ICE conjugation, and ICE regulation modules (62-64). ICEs integrate characteristics of both temperate bacteriophages (the front part) and conjugative plasmids (the latter part) (**Figure 4**) (62, 65). ICEs have been reported to contain several intergenic hotspots where a diverse range of exogenous genes can be carried, including antibiotic or heavy metal resistance genes (65). ICEs mediate HGT among prokaryotes, and greatly facilitate microbial genome evolution and ecological fitness (61, 62).

All analyzed genomes except that of strain WPS32 possess an ICE. Based on the structure and gene composition, the eight ICEs could be classified into three groups namely group 1 (JL477 and JL354), group 2 (JL1351, JL2201 and JL31) and group 3 (RCC1878, RCC1885 and RCC1897).

Two intergenic hotspots carrying exogenous genes were found in the eight genomes. The first one is located between genes encoding a nuclease and a single-stranded DNA binding protein. It also contains three different exogenous gene clusters (I, II, and III) corresponding to three types of ICEs (group 1, group 2, and group3). Exogenous gene cluster I is present in the genomes of JL477 and JL354 (group 1), and mainly consists of heavy metal resistance genes (*czcCBAD*) and their transcriptional regulator (*merR*), transposase AB, copper homeostasis, and anti-restriction genes. Exogenous gene cluster II, found in group 2, contains all the genes of exogenous gene cluster I and also possesses several extra multidrug resistance genes, whose acquisition is mediated by two integrases. Interestingly, although no ICE was found in the genome of WPS32, this genome contains multidrug resistance genes and two integrases. Exogenous gene cluster III lost heavy metal resistance genes but gained restriction/modification-related genes. This hotspot might carry a limited length of foreign genes, suggesting that these microbes might carefully select foreign genes that are optimally adapted to their environment.

The second hotspot, located between gene clusters *traDI* and *traGHFN*, mostly comprises peptidase and nuclease metabolism-related genes. In ICE group 1, it contains 14 genes with known functional genes for nucleases, helicases, ATPases, peptidases, and transcriptional regulator. The group 2 ICE lost a large part of the conjugation module (*traDI* and *traGHFN-trbC-traUW- trhF*). The second hotspot in ICE group 3 is composed of only five genes related to peptidases, pyrophosphatases, and transcriptional regulators. Downstream of ICE group 3, an approximately 27 kb gene fragment mainly involved in fatty acid metabolism, butyrate metabolism, and branched-chain amino acid biosynthesis is absent in the genomes of three RCC and WPS32.

**GI09.** GI09 is an approximately 11.5 kb fragment with a GC content similar to that of the genome (65.76%). It contains three genes, a giant hypothetical gene (8.70 kb), a transcriptional regulator and a histidine kinase, which together represent 97.14% of its sequence. GI09 is adjacent to a tRNA-CGA gene. The protein for the giant hypothetical gene product comprises three domains: an immunoglobulin beta-sandwich folding domain, a cadherin-like beta sandwich domain, and an autotransporter beta-domain. Cadherins are suggestive of adhesion molecules that mediate

$Ca^{2+}$-dependent cell-cell junctions (66). Usually, bacteria or cells containing the same cadherins tend to preferentially aggregate together.

GI09 was not detected at the same position in the genomes of strains RCC1878, RCC1885 and RCC1897. However, we found a remnant short sequence predicted as a hypothetical gene (324 bp) that shares 91% (296/324) nucleotide identity with the giant hypothetical gene of strain JL477. This supports the hypothesis that three RCC strains lost the GI09 sequence.

## Co-existence of two PGCs in genomes of *Citromicrobium* isolates

Interestingly, two different (one complete and one incomplete) PGCs were found in all nine genomes. The complete PGC consists of two conserved subclusters, *crtCDF-bchCXYZ-pufBALM* and *bchFNBHLM-lhaA-puhABC* (**Figure 5A**). The complete PGC organization is identical in all nine genomes in terms of gene arrangement and composition. The incomplete PGC contains only the *pufLMC* and *puhABC* genes (**Figure 5B**). The incomplete PGC, which was proved to be obtained by HGT (15, 16), is located at the same position in all the genomes and is flanked by respiratory complex I and CoA metabolism-related genes. This indicates that the ancestral *Citromicrobium* obtained the incomplete PGC before divergence. Both the complete and incomplete PGCs are close to the GI regions, creating conditions for gain and loss of phototrophic genes (**Figure 1**).

The *pufM* sequences from the complete PGC formed a clade close to that of *Erythrobacter* species also belonging to the order *Sphingomonadales*, alpha-IV subcluster (**Figure 6A**).. The *pufM* sequences from the incomplete PGC formed a distant clade branching with *Fulvimarina pelagi* HTCC2506 (alpha-VI subcluster) (**Figure 6A**). This phylogenetic placement is in agreement with our previous finding showing that the incomplete PGC genes might have been acquired from a *Fulvimarina*-related species (16).

In both *pufM* clades, the sequences could be grouped into three clusters: three RCC strains formed

485 one cluster, WPS32 by itself was a second cluster, and the other five strains formed a third cluster
486 (**Figure 6A**).
487
488 **Co-existence of two copies of *pufM* in *Citromicrobium* environmental sequences**
489
490 A total of 540,022 good quality sequence reads were obtained from two stations at five depths (5,
491 25, 75, 150 and 200 m) using the revised primers (**Table 3**). A large proportion (29.8%) of *pufM*
492 sequences having *Citromicrobium* as the closest relative were obtained. Among them, 66,182 and
493 95,052 sequences were classified into the complete and incomplete PGC clades, respectively.
494
495 Eleven and ten OTUs (>10 sequences) were classified into the Citromicrobial complete and
496 incomplete PGC clades, respectively (**Figure 6B, 6C**). All the environmental sequences differed
497 from *pufM* sequences from the isolates. Five main OTUs (with more than 1000 sequences) were
498 retrieved, three (denovo741, denovo766 and denovo718) in the complete PGC clade and two
499 (denovo180 and denovo574) in the incomplete PGC clade (**Table 3**). Interestingly, denovo741 and
500 denovo180 showed similar positions in their phylogenetic trees (**Figure 6B, 6C**). Their
501 representative sequences shared 99.1% (230/232) and 99.6% (227/228) nucleotide identity with
502 the *pufM* sequences belonging to the complete and incomplete PGCs of strain JL477, respectively.
503 In addition, denovo741 and denovo180 demonstrated the same depth distribution pattern (**Table 3**).
504 A similar situation was observed for denovo766 and denovo574, whose representative sequences
505 shared 91.4% (212/232) and 94.3% (217/230) nucleotide identity, respectively (**Table 3**).
506
507 However, our analysis did not find an OTU corresponding to a copy of denovo718 in the
508 incomplete PGC clade (**Figure 5B**). This may suggest that some Citromicrobial strains have lost
509 the incomplete PGC or that denovo718 is a novel *Citromicrobium* relative.
510
511 **Single-nucleotide polymorphisms**
512
513 The number of SNPs of the eight genomes relative to the complete genome of strain JL477 had a
514 wide range. More than 84,000 SNPs were found in the genomes of strains RCC1878, RCC1885,

RCC1897 and WPS32, while fewer than 200 SNPs were present in strains JL354, JL31, and JL1351. In the genome of JL2201, 1,603 SNPs were found, and most of them (1,379) originated from the prophage I sequences, suggesting that viruses had much faster evolutionary rates. Approximately 90% of all SNPs are located in coding regions and are scattered throughout the genomes except in the genetic islands.

Based on the growth rate (0.72-2.13 $day^{-1}$) of AAPB in the ocean (3), their generation time should be approximately 250-750 generations per year. The estimated divergence times based on the accumulation of synonymous mutations that excluded SNPs from GIs span a long history. The divergence times among JL477, JL31, JL1351, and JL354 are in century timescales, and these four strains diverge at a millennial timescale with JL2201. The three RCC strains and the WPS32 strain diverged from the five JL strains millions of years ago.

**Geographic relationship**

The isolates used in the study originate from diverse geographic locations, including the Mediterranean Sea, the South China Sea and the South Atlantic Ocean. Water from the Atlantic Ocean refilled the Mediterranean Sea through the Strait of Gibraltar 5.33 million years ago (67, 68). Before the water poured, the Mediterranean almost entirely dried out as result of the 'Messinian salinity crisis' (67, 68). In another word, the modern Mediterranean Sea has ~5.33-million year history. Microbes in the almost enclosed Mediterranean Sea might have evolved to their unique characteristics compared to the other open ocean regions. That is consistent with the divergence time between three RCC strains and five JL strains.

Both phylogenies based on marker genes and comparisons of genome sequences revealed that strains from a same region (South China Sea or Mediterranean Sea) shared a similar evolutionary history and are distinct from those originating from other regions (South China Sea vs Mediterranean Sea). Geographic differences are partly responsible for driving the observed evolutionary divergences, and they allow microbes to diverge through local adaptation to specific environmental conditions (69-71). The divergence processes within species are traditionally

considered as micro-evolutionary. However, some specific events, such as viral infection, grazing or extreme physical events, might contribute to unusual evolutionary diversification (e.g., strain WPS32).

HGT plays an important role in *Citromicrobium* genomic plasticity. Three integration events occurred, mediated by two types of prophages (JL477 and JL354; three RCC strains; JL2201), corresponding to the three marine regions from which the strains originated. Three of the nine strains were free of viral infection. Several genes preventing viral infection are detected in their GIs, suggesting that bacteria-phage interactions are actively ongoing in their environment.

Comparison of nine *Citromicrobium* genomes that share identical 16S rRNA sequences provides new insights into bacterial microevolution and divergence under different environments. The distribution of various genetic islands plays important roles in genomic plasticity and adaptability. The information gathered by comparing *Citromicrobium* genomes shed new light on the evolution and environmental adaptations resulting from geographic isolation in *Citromicrobium* species.

## Acknowledgements

This work was supported by the NSFC project (41306126), the national key research programs (2013CB955700 and 2016YFA0601400), the SOA project (GASI-03-01-02-05), CNOOC grants CNOOC-KJ125FZDXM00TJ001-2014 and CNOOC-KJ125FZDXM00ZJ001-2014, and the China National Marine Science Talent Training Base project (2015Z01). The work performed at the Station Biologique (Roscoff) is a contribution of the BOUM (Biogeochemistry from the Oligotrophic to the Ultraoligotrophic Mediterranean) experiment (http://www.com.univ-mrs.fr/BOUM/) of the French national LEFE-CYBER program, the European IP SESAME and the international IMBER project. The BOUM experiment was coordinated by the Institut National des Sciences de l'Univers (INSU) and managed by the Centre National de la Recherche Scientifique (CNRS). We thank the four anonymous reviewers and the editor for their useful comments and suggestions.

## Conflict of Interest

The authors declare that they have no competing interests.

# References


1. **Kolber ZS, Gerald F, Lang AS, Beatty JT, Blankenship RE, VanDover CL, Vetriani C, Koblizek M, Rathgeber C, Falkowski PG.** 2001. Contribution of aerobic photoheterotrophic bacteria to the carbon cycle in the ocean. Science **292:**2492-2495.
2. **Béjà O, Suzuki MT, Heidelberg JF, Nelson WC, Preston CM, Hamada T, Eisen JA, Fraser CM, DeLong EF.** 2002. Unsuspected diversity among marine aerobic anoxygenic phototrophs. Nature **415:**630-633.
3. **Koblížek M, Mašín M, Ras J, Poulton AJ, Prášil O.** 2007. Rapid growth rates of aerobic anoxygenic phototrophs in the ocean. Environmental microbiology **9:**2401-2406.
4. **Koblížek M.** 2015. Ecology of aerobic anoxygenic phototrophs in aquatic environments. FEMS microbiology reviews **39:** 854-870.
5. **Jiao N, Zhang Y, Zeng Y, Hong N, Liu R, Chen F, Wang P.** 2007. Distinct distribution pattern of abundance and diversity of aerobic anoxygenic phototrophic bacteria in the global ocean. Environmental Microbiology **9:**3091-3099.
6. **Yutin N, Suzuki MT, Teeling H, Weber M, Venter JC, Rusch DB, Béjà O.** 2007. Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling expedition metagenomes. Environmental microbiology **9:**1464-1475.
7. **Yurkov V, Csotonyi JT.** 2009. New light on aerobic anoxygenic phototrophs. The purple phototrophic bacteria Springer, pp. 31-55
8. **Ritchie AE, Johnson ZI.** 2012. Abundance and genetic diversity of aerobic anoxygenic phototrophic bacteria of coastal regions of the Pacific Ocean. Applied and environmental microbiology **78:**2858-2866.
9. **Ferrera I, Borrego CM, Salazar G, Gasol JM.** 2014. Marked seasonality of aerobic anoxygenic phototrophic bacteria in the coastal NW Mediterranean Sea as revealed by cell abundance, pigment concentration and pyrosequencing of pufM gene. Environmental microbiology **16:**2953-2965.
10. **Stegman MR, Cottrell MT, Kirchman DL.** 2014. Leucine incorporation by aerobic anoxygenic phototrophic bacteria in the Delaware estuary. The ISME journal **8:**2339-2348.
11. **Koblížek M, Béjà O, Bidigare RR, Christensen S, Benitez-Nelson B, Vetriani C, Kolber MK, Falkowski PG, Kolber ZS.** 2003. Isolation and characterization of Erythrobacter sp. strains from the upper ocean. Archives of Microbiology **180:**327-338.
12. **Swingley WD, Sadekar S, Mastrian SD, Matthies HJ, Hao J, Ramos H, Acharya CR, Conrad AL, Taylor HL, Dejesa LC.** 2007. The complete genome sequence of Roseobacter denitrificans reveals a mixotrophic rather than photosynthetic metabolism. Journal of bacteriology **189:**683-690.
13. **Zheng Q, Zhang R, Koblížek M, Boldareva EN, Yurkov V, Yan S, Jiao N.** 2011. Diverse arrangement of photosynthetic gene clusters in aerobic anoxygenic phototrophic bacteria. PloS one **6**: e25050.
14. **Beatty JT.** 1995. Organization of photosynthesis gene transcripts. Anoxygenic photosynthetic bacteria Springer, pp. 1209-1219
15. **Jiao N, Zhang R, Zheng Q.** 2010. Coexistence of two different photosynthetic operons

621 in Citromicrobium bathyomarinum JL354 as revealed by whole-genome sequencing.
622 Journal of bacteriology **192:**1169-1170.
623 16. **Zheng Q, Zhang R, Fogg PC, Beatty JT, Wang Y, Jiao N.** 2012. Gain and loss of
624 phototrophic genes revealed by comparison of two Citromicrobium bacterial genomes.
625 PloS one **7:**e35790.
626 17. **Yurkov VV, Krieger S, Stackebrandt E, Beatty JT.** 1999. Citromicrobium
627 bathyomarinum, a novel aerobic bacterium isolated from deep-sea hydrothermal vent
628 plume waters that contains photosynthetic pigment-protein complexes. Journal of
629 bacteriology **181:**4517-4525.
630 18. **Rathgeber C, Lince MT, Alric J, Lang AS, Humphrey E, Blankenship RE, Verméglio**
631 **A, Plumley FG, Van Dover CL, Beatty JT.** 2008. Vertical distribution and
632 characterization of aerobic phototrophic bacteria at the Juan de Fuca Ridge in the Pacific
633 Ocean. Photosynthesis research **97:**235-244.
634 19. **Balashov SP, Imasheva ES, Boichenko VA, Antón J, Wang JM, Lanyi JK.** 2005.
635 Xanthorhodopsin: a proton pump with a light-harvesting carotenoid antenna. Science
636 **309:**2061-2064.
637 20. **Balashov S, Lanyi J.** 2007. Xanthorhodopsin: Proton pump with a carotenoid antenna.
638 Cellular and Molecular Life Sciences **64:**2323-2328.
639 21. **Boeuf D, Audic S, Brillet-Guéguen L, Caron C, Jeanthon C.** 2015. MicRhoDE: a
640 curated database for the analysis of microbial rhodopsin diversity and evolution. Database
641 **2015:**bav080.
642 22. **Kwon S-K, Kim BK, Song JY, Kwak M-J, Lee CH, Yoon J-H, Oh TK, Kim JF.** 2013.
643 Genomic makeup of the marine flavobacterium Nonlabens (Donghaeana) dokdonensis
644 and identification of a novel class of rhodopsins. Genome biology and evolution
645 **5:**187-199.
646 23. **Riedel T, Gómez-Consarnau L, Tomasch J, Martin M, Jarek M, González JM,**
647 **Spring S, Rohlfs M, Brinkhoff T, Cypionka H.** 2013. Genomics and physiology of a
648 marine flavobacterium encoding a proteorhodopsin and a xanthorhodopsin-like protein.
649 PloS one **8:**e57487.
650 24. **Zheng Q, Liu Y, Steindler L, Jiao N.** 2015. Pyrosequencing analysis of aerobic
651 anoxygenic phototrophic bacterial community structure in the oligotrophic western
652 Pacific Ocean. FEMS microbiology letters **362:**fnv034.
653 25. **Jeanthon C, Boeuf D, Dahan O, Gall FL, Garczarek L, Bendif EM, Lehours A-C.**
654 2011. Diversity of cultivated and metabolically active aerobic anoxygenic phototrophic
655 bacteria along an oligotrophic gradient in the Mediterranean Sea. Biogeosciences
656 **8:**1955-1970.
657 26. **Agogue H, Lamy D, Neal PR, Sogin ML, Herndl GJ.** 2011. Water mass-specificity of
658 bacterial communities in the North Atlantic revealed by massively parallel sequencing.
659 Molecular Ecology **20:**258-274.
660 27. **Caporaso JG, Lauber CL, Walters WA, Berglyons D, Lozupone C, Turnbaugh PJ,**
661 **Fierer N, Knight R.** 2011. Global patterns of 16S rRNA diversity at a depth of millions
662 of sequences per sample. Proceedings of the National Academy of Sciences of the United
663 States of America **108:**4516-4522.

28. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research **18:**821-829.
29. **Salzberg SL, Delcher AL, Kasif S, White O.** 1998. Microbial gene identification using interpolated Markov models. Nucleic acids research **26:**544-548.
30. **Borodovsky M, McIninch J.** 1993. GENMARK: parallel gene recognition for both DNA strands. Computers & chemistry **17:**123-133.
31. **Lukashin AV, Borodovsky M.** 1998. GeneMark. hmm: new solutions for gene finding. Nucleic acids research **26:**1107-1115.
32. **Angiuoli SV, Gussman A, Klimke W, Cochrane G, Field D, Garrity GM, Kodira CD, Kyrpides N, Madupu R, Markowitz V.** 2008. Toward an online repository of Standard Operating Procedures (SOPs) for (meta) genomic annotation. OMICS A Journal of Integrative Biology **12:**137-141.
33. **Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M.** 2008. The RAST Server: rapid annotations using subsystems technology. BMC genomics **9:**75.
34. **Lagesen K, Hallin P, Rødland E, Stærfeldt H, Rognes T, Ussery D.** 2007. RNammer: consistent annotation of rRNA genes in genomic sequences. Nucleic Acids Res **35:**3100-3108.
35. **Lowe TM, Eddy SR.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic acids research **25:**0955-0964.
36. **Richter M, Rosselló-Móra R.** 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proceedings of the National Academy of Sciences **106:**19126-19131.
37. **Li L, Stoeckert CJ, Roos DS.** 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome research **13:**2178-2189.
38. **Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS.** 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proceedings of the National Academy of Sciences of the United States of America **102:**13950-13955.
39. **Tatusov RL, Koonin EV, Lipman DJ.** 1997. A genomic perspective on protein families. Science **278:**631-637.
40. **Delcher AL, Salzberg SL, Phillippy AM.** 2003. Using MUMmer to identify similar regions in large sequence sets. Current Protocols in Bioinformatics**:**10.13. 11-10.13. 18.
41. **Foster JT, Beckstrom-Sternberg SM, Pearson T, Beckstrom-Sternberg JS, Chain PS, Roberto FF, Hnath J, Brettin T, Keim P.** 2009. Whole-genome-based phylogeny and divergence of the genus Brucella. Journal of bacteriology **191:**2864-2870.
42. **Alikhan N-F, Petty NK, Zakour NLB, Beatson SA.** 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC genomics **12:**402.
43. **Tamura K, Stecher G, Peterson D, Filipski A, Kumar S.** 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Molecular biology and evolution **30:**2725-2729.
44. **Fuhrman JA, Comeau DE, Hagström Å, Chan AM.** 1988. Extraction from natural planktonic microorganisms of DNA suitable for molecular biological studies. Applied and

environmental microbiology **54:**1426-1429.

45. **Yutin N, Suzuki MT, Béjà O.** 2005. Novel primers reveal wider diversity among marine aerobic anoxygenic phototrophs. Applied and environmental microbiology **71:**8958-8962.

46. **Rajendhran J, Gunasekaran P.** 2011. Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond. Microbiological research **166**: 99-110.

47. **Delgado-Baquerizo M, Giaramida L, Reich PB, Khachane AN, Hamonts K, Edwards C, Lawton LA, Singh BK.** 2016. Lack of functional redundancy in the relationship between microbial diversity and ecosystem functioning. Journal of Ecology **104**: 936–946.

48. **Dobrindt U, Hochhut B, Hentschel U, Hacker J.** 2004. Genomic islands in pathogenic and environmental microorganisms. Nature Reviews Microbiology **2:**414-424.

49. **Hacker J, Hentschel U, Dobrindt U.** 2003. Prokaryotic chromosomes and disease. Science **301:**790-793.

50. **Grozdanov L, Raasch C, Schulze J, Sonnenborn U, Gottschalk G, Hacker J, Dobrindt U.** 2004. Analysis of the genome structure of the nonpathogenic probiotic Escherichia coli strain Nissle 1917. Journal of bacteriology **186:**5432-5441.

51. **Kaneko T, Maita H, Hirakawa H, Uchiike N, Minamisawa K, Watanabe A, Sato S.** 2011. Complete genome sequence of the soybean symbiont Bradyrhizobium japonicum strain USDA6T. Genes **2:**763-787.

52. **O'Callaghan D, Cazevieille C, Allardet‐Servent A, Boschiroli ML, Bourg G, Foulongne V, Frutos P, Kulakov Y, Ramuz M.** 1999. A homologue of the Agrobacterium tumefaciens VirB and Bordetella pertussis Ptl type IV secretion systems is essential for intracellular survival of Brucella suis. Molecular microbiology **33:**1210-1220.

53. **Lawley T, Klimke W, Gubbins M, Frost L.** 2003. F factor conjugation is a true type IV secretion system. FEMS microbiology letters **224:**1-15.

54. **Marrs B.** 1974. Genetic recombination in Rhodopseudomonas capsulata. Proceedings of the National Academy of Sciences **71:**971-973.

55. **Lang AS, Beatty JT.** 2007. Importance of widespread gene transfer agent genes in α-proteobacteria. Trends in microbiology **15:**54-62.

56. **Grossart H-P, Riemann L, Azam F.** 2001. Bacterial motility in the sea and its ecological implications. Aquatic Microbial Ecology **25:**247-258.

57. **Harshey RM.** 2003. Bacterial motility on a surface: many ways to a common goal. Annual Reviews in Microbiology **57:**249-273.

58. **Stocker R.** 2012. Marine microbes see a sea of gradients. Science **338:**628-633.

59. **Svensson SL, Pryjma M, Gaynor EC.** 2014. Flagella-mediated adhesion and extracellular DNA release contribute to biofilm formation and stress tolerance of Campylobacter jejuni. PloS one **9**: e106063.

60. **Zheng Q, Zhang R, Xu Y, White III RA, Wang Y, Luo T, Jiao N.** 2014. A marine inducible prophage vB_CibM-P1 isolated from the aerobic anoxygenic phototrophic bacterium Citromicrobium bathyomarinum JL354. Scientific reports **4**.

61. **Böltner D, MacMahon C, Pembroke JT, Strike P, Osborn AM.** 2002. R391: a conjugative integrating mosaic comprised of phage, plasmid, and transposon elements.

Journal of bacteriology **184:**5158-5169.

62. **Burrus V, Marrero J, Waldor MK.** 2006. The current ICE age: biology and evolution of SXT-related integrating conjugative elements. Plasmid **55:**173-183.

63. **Ravatn R, Studer S, Springael D, Zehnder AJ, van der Meer JR.** 1998. Chromosomal integration, tandem amplification, and deamplification in Pseudomonas putida F1 of a 105-kilobase genetic element containing the chlorocatechol degradative genes from Pseudomonas sp. strain B13. Journal of bacteriology **180:**4360-4369.

64. **Hochhut B, Marrero J, Waldor MK.** 2000. Mobilization of plasmids and chromosomal DNA mediated by the SXT element, a constin found in Vibrio cholerae O139. Journal of bacteriology **182:**2043-2047.

65. **Wozniak RA, Fouts DE, Spagnoletti M, Colombo MM, Ceccarelli D, Garriss G, Déry C, Burrus V, Waldor MK.** 2009. Comparative ICE genomics: insights into the evolution of the SXT/R391 family of ICEs.

66. **Pokutta S, Weis WI.** 2007. Structure and mechanism of cadherins and catenins in cell-cell contacts. Annu Rev Cell Dev Biol **23:**237-261.

67. **Krijgsman W, Langereis C, Zachariasse W, Boccaletti M, Moratti G, Gelati R, Iaccarino S, Papani G, Villa G.** 1999. Late Neogene evolution of the Taza–Guercif Basin (Rifian Corridor, Morocco) and implications for the Messinian salinity crisis. Marine Geology **153:**147-160.

68. **Garcia-Castellanos D, Estrada F, Jiménez-Munt I, Gorini C, Fernández M, Vergés J, De Vicente R.** 2009. Catastrophic flood of the Mediterranean after the Messinian salinity crisis. Nature **462:**778-781.

69. **Whitaker RJ, Grogan DW, Taylor JW.** 2003. Geographic barriers isolate endemic populations of hyperthermophilic archaea. Science **301:**976-978.

70. **Papke RT, Ramsing NB, Bateson MM, Ward DM.** 2003. Geographical isolation in hot spring cyanobacteria. Environmental Microbiology **5:**650-659.

71. **Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JB.** 2012. Beyond biogeographic patterns: processes shaping the microbial landscape. Nature Reviews Microbiology **10:**497-506.

# Table and Figure legends

**Figure 1.** Whole genome map of *Citromicrobium* sp. JL477 compared with other eight *Citromicrobium* genomes. From the inner to outer circles: GC content plot with a grey circle representing 50%, GC skew plot, *Citromicrobium* sp. JL354, *Citromicrobium* sp. JL31, *Citromicrobium* sp. JL1351, *Citromicrobium* sp. JL2201, *Citromicrobium* sp. RCC1878, *Citromicrobium* sp. RCC1885, RCC1897 and *Citromicrobium* sp. WPS32. Genomic island regions are indicated with red line on the outermost circle from GI01 to GI09. The complete and incomplete PGCs are labeled by the black line on the outermost circle.

**Figure 2**. Organization of GI01 structural genes in *Citromicrobium* genomes. A was found in strains JL477, JL354, RCC1878, RCC1885 and RCC1897; B was found in strains JL31, JL1351 and JL2201; C was found in strains JL31, JL1351 and JL2201; D was found in the three RCC strains; E was found in strain WPS32; F was found in strains JL477 and JL354. Yellow, conserved upstream and downstream genes (from 1 to14, and from I to XII) of the GI01 gene cluster in *Citromicrobium* genomes; pink, *trb* gene cluster; red, tRNA or integrase; green, type I restriction-modification system; cyan, the other function known genes; light gray, hypothetic genes.

1, Type IV secretory pathway, protease TraF; 2, hypothetical protein; 3, Membrane-bound lytic murein transglycosylase C precursor; 4, Type IV secretory pathway, VirD2 components (relaxase); 5, hypothetical protein; 6, hypothetical protein; 7, Coupling protein VirD4, ATPase required for T-DNA transfer; 8, Asparagine synthetase [glutamine-hydrolyzing]; 9, Acylamino-acid-releasing enzyme; 10, TonB-dependent receptor; 11, RNA polymerase sigma-70 factor, ECF subfamily; 12, hypothetical protein; 13, hypothetical protein; 14, transcriptional regulator.

I, Cell division protein FtsH; II, ATPase, ParA family protein; III, Butyryl-CoA dehydrogenase; IV, Alpha-methylacyl-CoA racemase; V, Enoyl-CoA hydratase; VI, Ferrichrome-iron receptor; VII, hypothetical protein; VIII, hypothetical protein; IX, Sterol desaturase family protein; X, hypothetical protein; XI, hypothetical protein; XII, hypothetical protein.

**Figure 3**. Organization of flagellum and GI03 structural genes in *Citromicrobium* genomes. Yellow, conversed upstream and downstream genes (from 1 to 12) of the GI03 gene cluster in *Citromicrobium* genomes; pink, flagellar gene cluster; red, tRNA or integrase; cyan, the other function known genes; light gray, hypothetical genes.

**Figure 4**. Structure and composition of ICE. Two hotspots were detected in all ICEs. One contained three types of exogenous gene cluster (I, II and III), and the other two. Red, phage-related genes; pink, conjugative-related genes; cyan, the other function known genes; white, hypothetic genes.

**Figure 5.** Structure and arrangement of two PGCs in *Citromicrobium*. A, complete PGC; B, incomplete PGC. Green, *bch* genes; red, *puf* and regulators genes; pink, *puh* genes; orange, *crt* genes; blue, *hem* and *cyc* gene; yellow, *lhaA* gene; blank, uncertain or unrelated genes; grey, hypothetical protein. The horizontal arrows represent putative transcripts.

**Figure 6**. Neighbour-joining phylogenetic trees based on *pufM* gene sequences. A, phylogenetic tree containing *pufM* sequences of the nine isolates; B, the part tree containing environmental *pufM* sequences from the complete PGC; C, partial tree containing environmental *pufM* sequences from the incomplete PGC. Only bootstrap percentages (> 50%) are shown (neighbour-joining/maximum likelihood).

**Table 1.** Genome information for the nine *Citromicrobium* strains

**Table 2.** Detailed information for the nine GIs

**Table 3.** Distribution and identity of environmental *pufM* sequences retrieved from sites P3 and P10 at different depths

# Supplementary Information

**Figure S1**. Organization of GTA structural genes in *Citromicrobium* genomes. A, GTA in strains JL31, JL354, JL477, JL1351, JL2201, RCC1878, RCC1885 and RCC1897; B, GTA in strain WPS32. Yellow, conversed upstream and downstream genes (from 1 to 7) of the GTA structural

gene cluster in *Citromicrobium* genomes; red, a putative transposase; pink, functions known in GTA genes; white, hypothetical genes; gray, conserved hypothetical genes belonging to GTA.

**Figure S2.** Structure and organization of prophage in *Citromicrobium*. Pink, early expression genes; orange, heads; yellow, tails; red, transposase; green, lysozyme genes; light gray, putative proteins.

**Table S1**. Comparison of gene organization and identities for nine Citromicrobium genomes

**Table S2**. Average Nucleotide Identity by pairwise genome comparison

855
856 Table 1. Genome information for the nine strains

857

| Strains | Genome size (Mb) | Contigs | GC content | Genes | CDs | COGs | tRNA | Sequencing coverage | Isolation source | Acc. No. |
|---|---|---|---|---|---|---|---|---|---|---|
| JL31 | 3.16 | 22 | 65.1 | 3092 | 2960 | 1969 | 45 | 180x | South China sea | LAIH00000000 |
| JL1351 | 3.16 | 17 | 65.1 | 3090 | 2961 | 1981 | 45 | 155x | South China sea | LAPR00000000 |
| WPS32 | 3.16 | 16 | 64.9 | 3056 | 2925 | 1934 | 44 | 250x | South China sea | LAPS00000000 |
| **JL477** | 3.26 | 1 | 65.0 | 3168 | 3027 | 2004 | 45 | 220x | South China sea | CP011344 |
| JL354 | 3.27 | 68 | 65.0 | 3208 | 3137 | 2010 | 45 | 26x | South China sea | ADAE00000000 |
| JL2201 | 3.27 | 22 | 65.1 | 3250 | 3105 | 1975 | 45 | 245x | South Atlantic | LARQ00000000 |
| RCC1878 | 3.28 | 14 | 64.8 | 3194 | 3061 | 1966 | 45 | 205x | Mediterranean Sea | LBLZ00000000 |
| RCC1885 | 3.28 | 14 | 64.8 | 3197 | 3063 | 1975 | 45 | 190x | Mediterranean Sea | LBLY00000000 |
| RCC1897 | 3.28 | 17 | 64.8 | 3192 | 3113 | 1978 | 45 | 440x | Mediterranean Sea | LUGI01000000 |

858
859
860

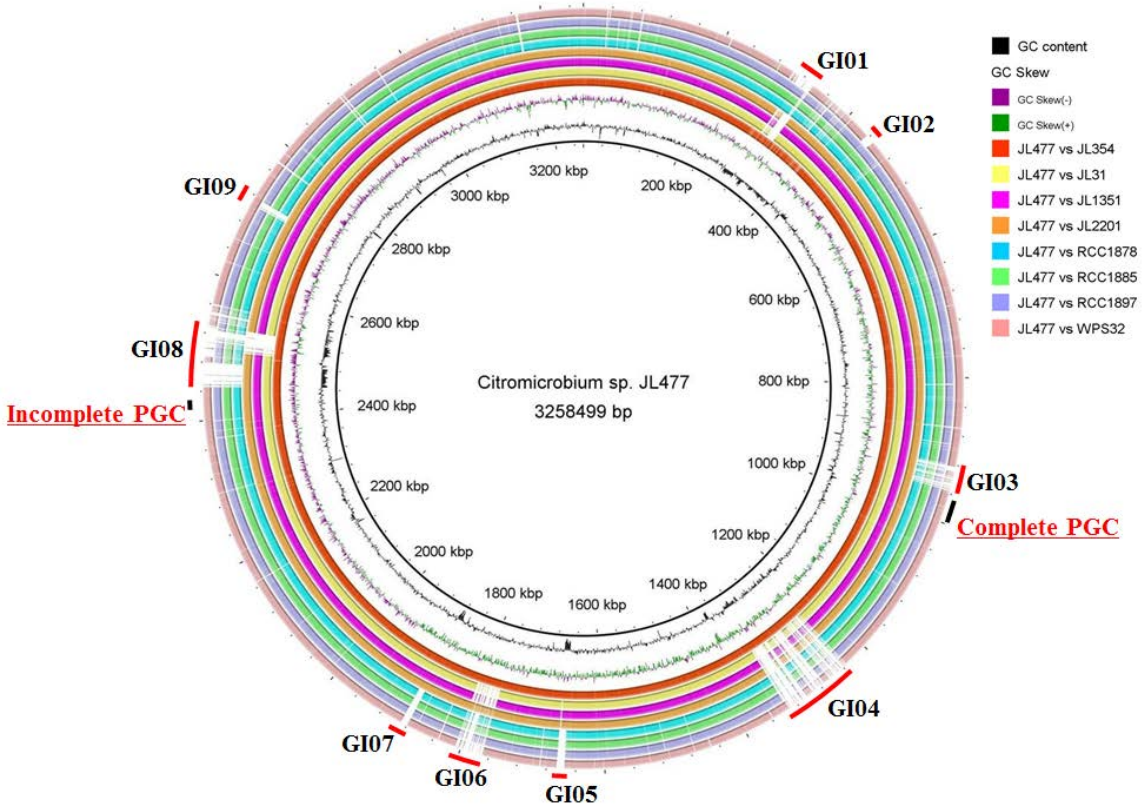
Table 2. Detailed information for the nine GIs

| GI | Size (kb) | GC content | tRNA | No. of genes | No. of transposase and integrase | Hypothetical proteins | Predicted function |
|---|---|---|---|---|---|---|---|
| 01 | 36.7 | 60.66% | | 34 | 0 | 10 | T4SS |
| 02 | 3.1/ 11.4 | 57.59%/ 66.80% | | 16 | 1 | 4 | GTA |
| 03 | 35.5 | 65.13% | tRNA-Ser-GGA | 36 | 1 | 10 | Flagellar biosynthesis |
| 04 | 101.1 | 62.14% | tRNA-Ser-GCT | 88 | 1 | 18 | Choline and Betaine Uptak;Glycerolipid and Glycerophospholipid Metabolism; Fatty acid metabolism; Pyruvate metabolism |
| 05 | 11.3 | 52.08% | | 4 | | 4 | Unkown |
| 06 | 38.1 | 65.98% | | 58 | 2 | (30) | Prophage |
| 07 | 9.7 | 54.39% | tRNA-Pro-TGG tRNA-Met-CAT | 4 | 1 | 2 | Unkown |
| 08 | 113.4 | 60.62% | tRNA-Met-CAT | 98 | 1 | 21 | ICE |
| 09 | 11.5 | 65.76% | tRNA-Ser-CGA | 3 | 0 | 0 | Flagellar hook-length control |

**Table 3.** Distribution and identity of environmental *pufM* sequences retrieved from sites P3 and P10 at different depths

| OTU ID | P10-5m | P10-25m | P10-75m | P10-150m | P10-200m | P3-5m | P3-25m | P3-75m | P3-150m | P3-200m | Total | Identities |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| denovo180 | 820 | 1517 | 7885 | 1830 | 15898 | 5382 | 14689 | 5287 | 10405 | 25756 | 89469 | 99.6% |
| denovo574 | 17 | 301 | 1956 | 15 | 81 | 2255 | 126 | 47 | 308 | 52 | 5158 | 94.3% |
| denovo741 | 721 | 460 | 3645 | 982 | 10915 | 783 | 9503 | 3423 | 7093 | 14815 | 52340 | 99.1% |
| denovo766 | 76 | 586 | 2950 | 36 | 147 | 5586 | 73 | 23 | 419 | 98 | 9994 | 91.4% |
| denovo718 | 11 | 109 | 662 | 10 | 9 | 645 | 44 | 1629 | 3 | 10 | 3132 | 88.4% |
| Total seqs | 62205 | 61988 | 42912 | 72509 | 78655 | 35183 | 73232 | 37557 | 31088 | 44693 | 540022 | |

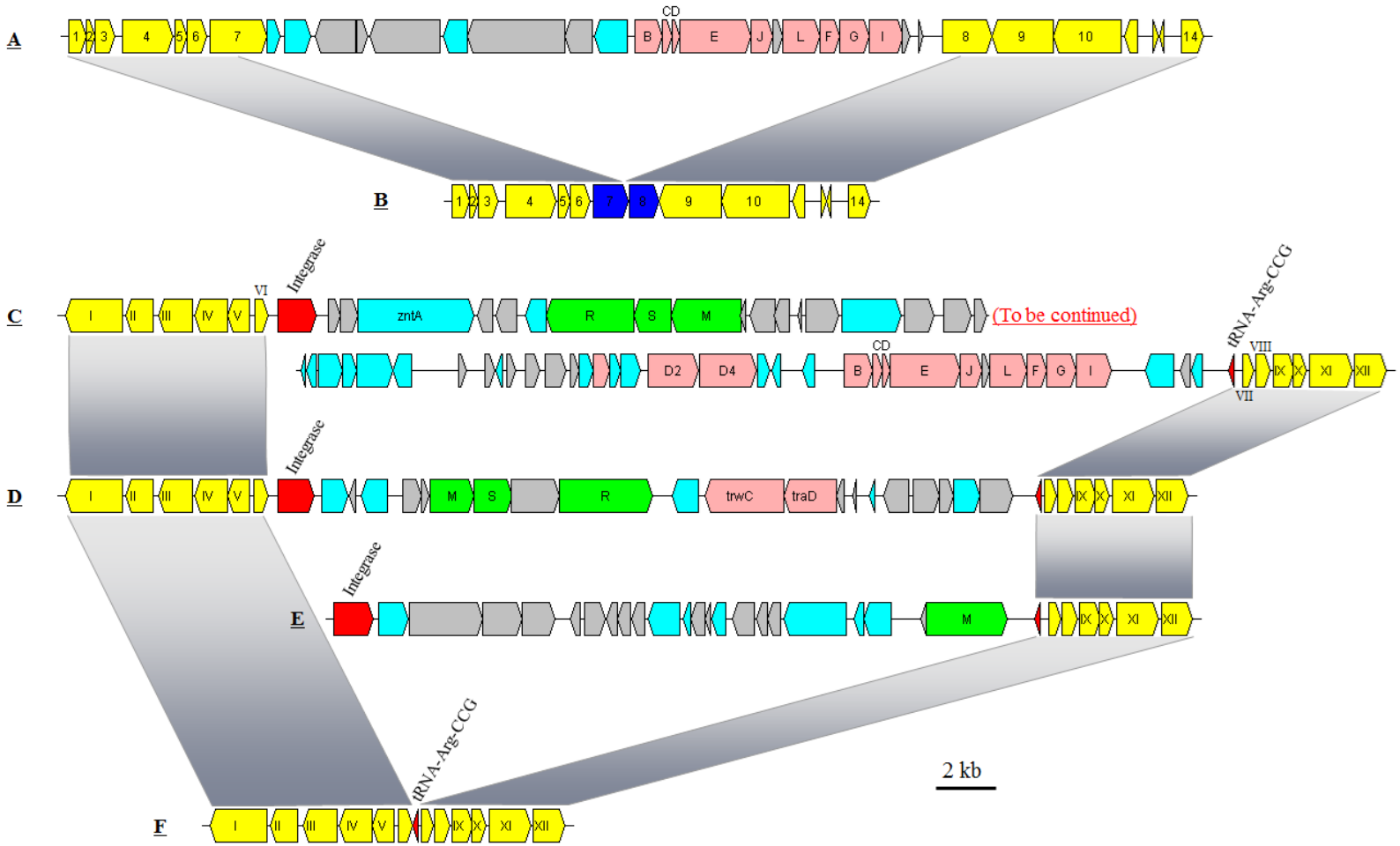
873 **Figure 1**




874

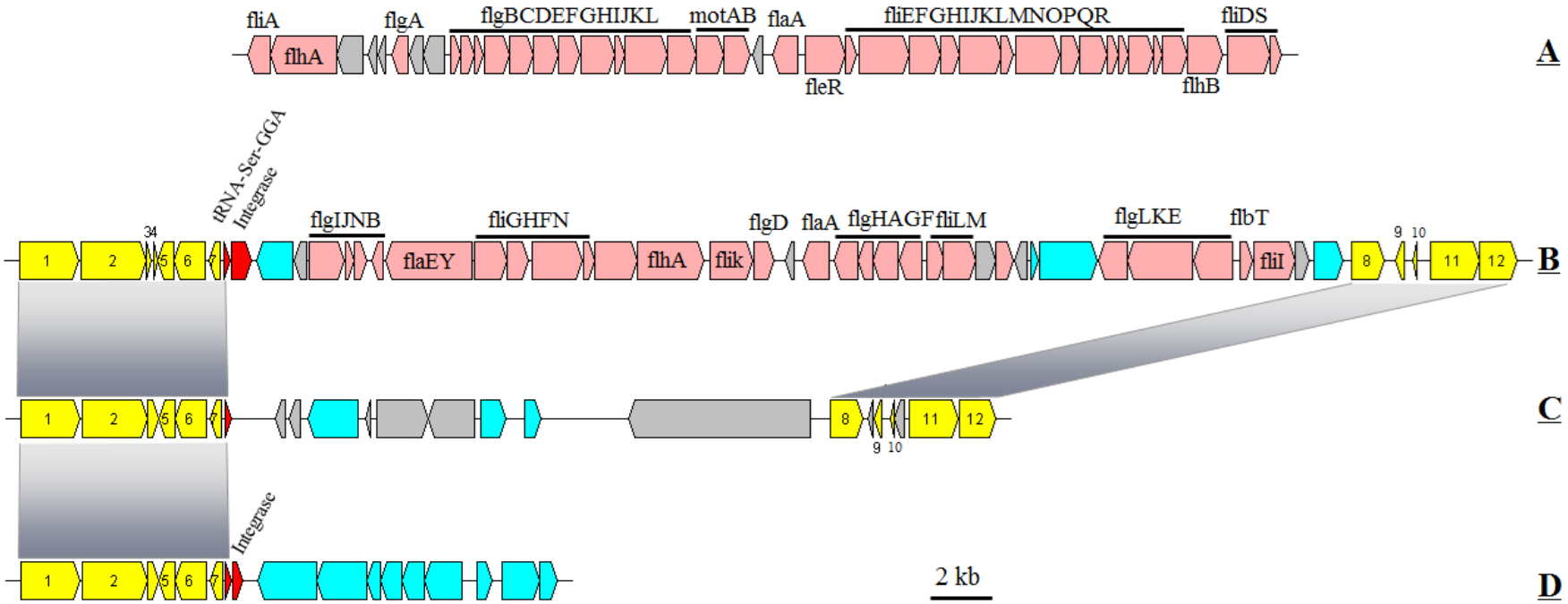
875

# Figure 2


A
B
C
D
E
F
CD
Integrase
VI
zntA
R
S
M
(To be continued)
tRNA-Arg-CCG
VIII
VII
D2
D4
trwC
traD
2 kb

878

879 **Figure 3**




880

881

**Figure 4**

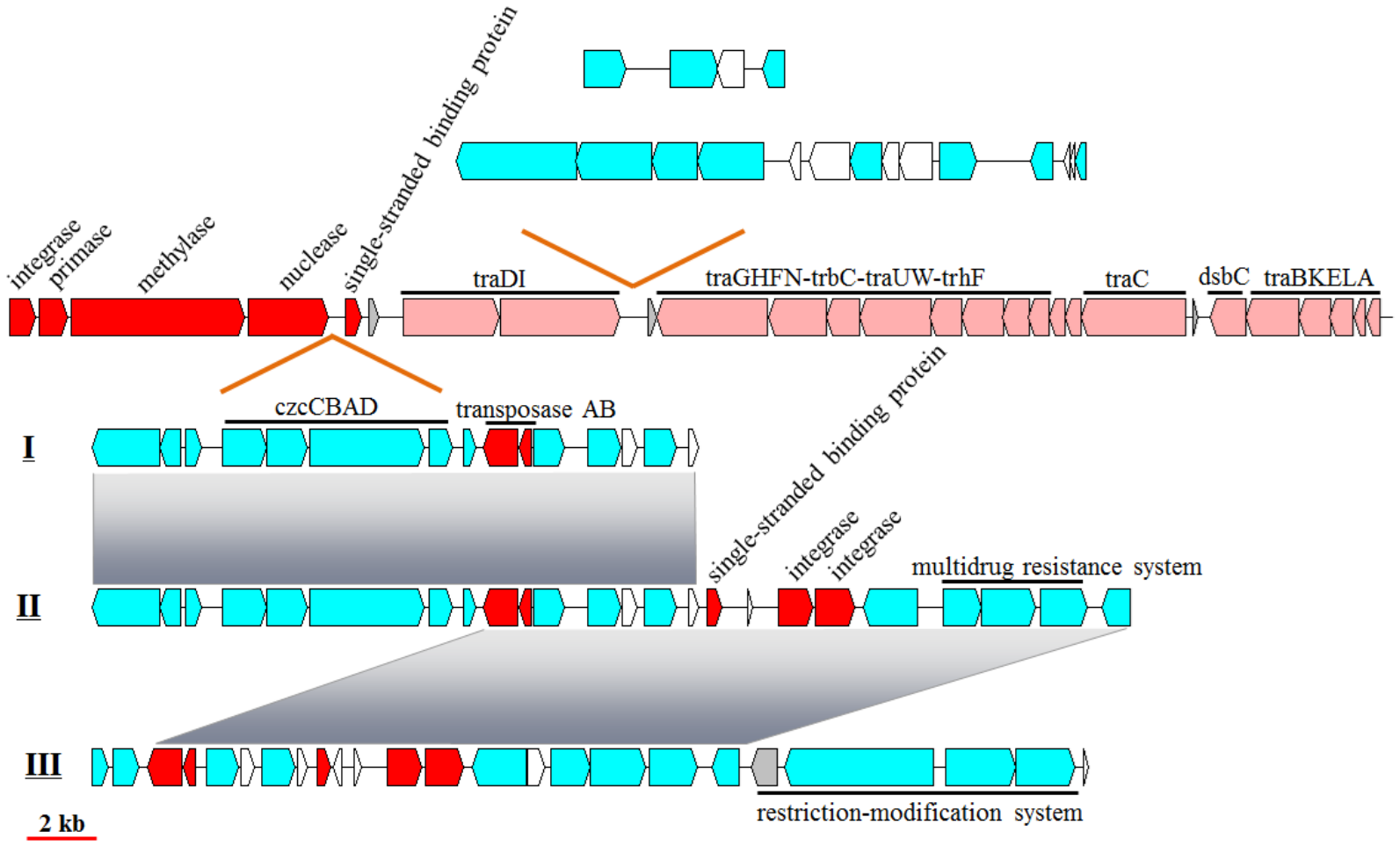
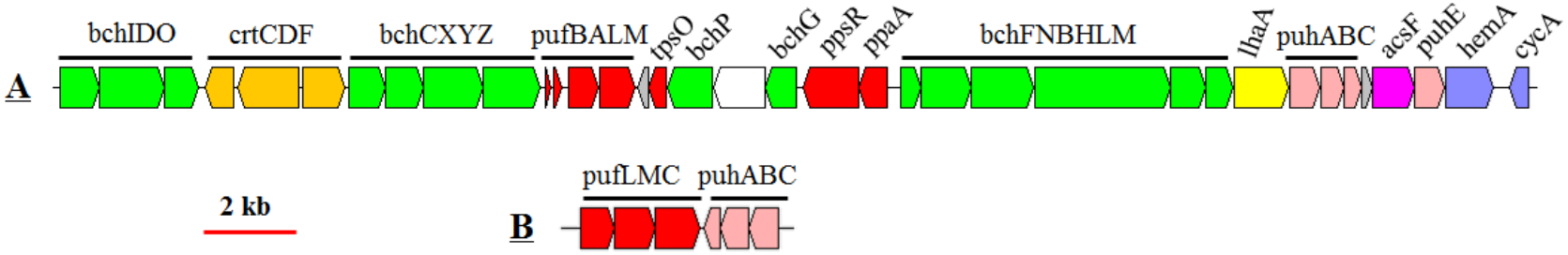
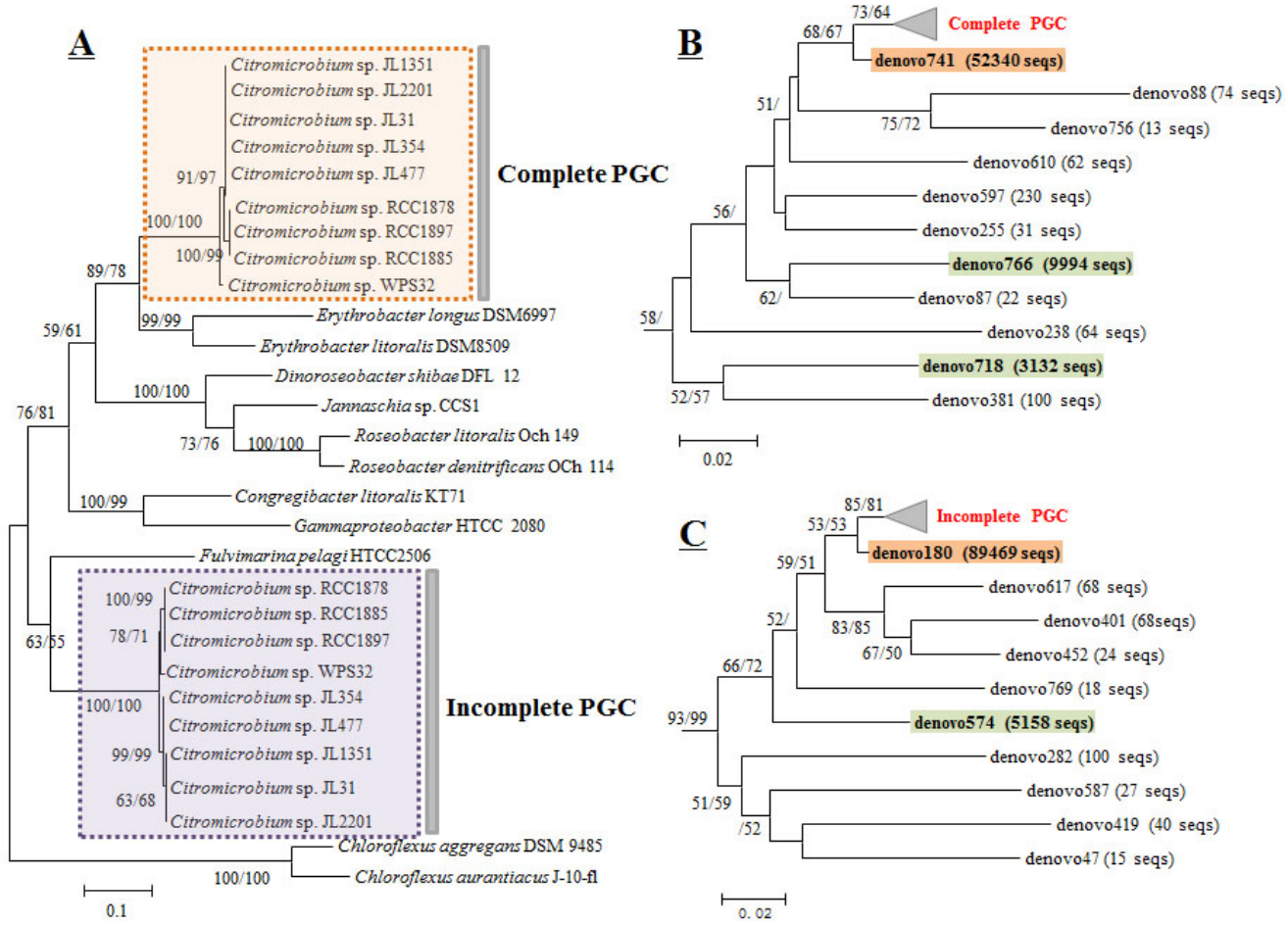

integrase
primase
methylase
nuclease
single-stranded binding protein
traDI
traGHFN-trbC-traUW-trhF
traC
dsbC
traBKELA
czcCBAD
transposase AB
I
single-stranded binding protein
integrase
integrase
multidrug resistance system
II
III
restriction-modification system
2 kb

**Figure 5**


bchIDO
crtCDF
bchCXYZ
pufBALM
tspO
bchP
bchG
ppsR
ppaA
bchFNBHLM
lhaA
puhABC
acsF
puhE
hemA
cycA
A
2 kb
pufLMC
puhABC
B

## Figure 6


A
Citromicrobium sp. JL1351
Citromicrobium sp. JL2201
Citromicrobium sp. JL31
Citromicrobium sp. JL354
Citromicrobium sp. JL477
Citromicrobium sp. RCC1878
Citromicrobium sp. RCC1897
Citromicrobium sp. RCC1885
Citromicrobium sp. WPS32
Complete PGC
Erythrobacter longus DSM6997
Erythrobacter litoralis DSM8509
Dinoroseobacter shibae DFL 12
Jannaschia sp. CCS1
Roseobacter litoralis Och 149
Roseobacter denitrificans OCh 114
Congregibacter litoralis KT71
Gammaproteobacter HTCC 2080
Fulvimarina pelagi HTCC2506
Citromicrobium sp. RCC1878
Citromicrobium sp. RCC1885
Citromicrobium sp. RCC1897
Citromicrobium sp. WPS32
Citromicrobium sp. JL354
Citromicrobium sp. JL477
Citromicrobium sp. JL1351
Citromicrobium sp. JL31
Citromicrobium sp. JL2201
Incomplete PGC
Chloroflexus aggregans DSM 9485
Chloroflexus aurantiacus J-10-fl
0.1
B
Complete PGC
denovo741 (52340 seqs)
denovo88 (74 seqs)
denovo756 (13 seqs)
denovo610 (62 seqs)
denovo597 (230 seqs)
denovo255 (31 seqs)
denovo766 (9994 seqs)
denovo87 (22 seqs)
denovo238 (64 seqs)
denovo718 (3132 seqs)
denovo381 (100 seqs)
0.02
C
Incomplete PGC
denovo180 (89469 seqs)
denovo617 (68 seqs)
denovo401 (68seqs)
denovo452 (24 seqs)
denovo769 (18 seqs)
denovo574 (5158 seqs)
denovo282 (100 seqs)
denovo587 (27 seqs)
denovo419 (40 seqs)
denovo47 (15 seqs)
0. 02