J. Microbiol. Biotechnol. (2019), 29(5), 794–808
https://doi.org/10.4014/jmb.1901.01040

jmb

Complete Genome Sequencing of *Bacillus velezensis* WRN014, and Comparison with Genome Sequences of other *Bacillus velezensis* Strains S

Junru Wang[1,4], Juyuan Xing[2], Jiangkun Lu[3], Yingjiao Sun[4], Juanjuan Zhao[4], Shaohua Miao[4], Qin Xiong[4], Yonggang Zhang[5*], and Guishan Zhang[4*]

[1]*BGI Education Center, University of Chinese Academy of Sciences, Shenzhen, P.R. China*
[2]*Wuhan University of Technology, Wuhan, Hubei Province, P.R. China*
[3]*School of Life Science, Beijing Institute of Technology, Beijing, P.R. China*
[4]*Key Laboratory of Microbial Resources Collection and Preservation, Ministry of Agriculture, Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing, P.R. China*
[5]*Biology Institute, Qilu University of Technology (Shandong Academy of Sciences), Jinan, Shandong Province, P.R. China*

Received: January 18, 2019
Revised: April 1, 2019
Accepted: April 24, 2019

First published online
April 27, 2019

*Corresponding authors
G.Z.
Phone: +86-10-82106223;
E-mail: gszhang86@gmail.com
Y.Z.
Phone: +86-0531-85599090;
E-mail: zhangygcq@163.com

Supplementary data for this paper are available on-line only at http://jmb.or.kr.

pISSN 1017-7825, eISSN 1738-8872

Bacillus velezensis strain WRN014 was isolated from banana fields in Hainan, China. *Bacillus velezensis* is an important member of the plant growth-promoting rhizobacteria (PGPR) which can enhance plant growth and control soil-borne disease. The complete genome of *Bacillus velezensis* WRN014 was sequenced by combining Illumina Hiseq 2500 system and Pacific Biosciences SMRT high-throughput sequencing technologies. Then, the genome of *Bacillus velezensis* WRN014, together with 45 other completed genome sequences of the *Bacillus velezensis* strains, were comparatively studied. The genome of *Bacillus velezensis* WRN014 was 4,063,541bp in length and contained 4,062 coding sequences, 9 genomic islands and 13 gene clusters. The results of comparative genomic analysis provide evidence that (i) The 46 *Bacillus velezensis* strains formed 2 obviously closely related clades in phylogenetic trees. (ii) The pan-genome in this study is open and is increasing with the addition of new sequenced genomes. (iii) Analysis of single nucleotide polymorphisms (SNPs) revealed local diversification of the 46 *Bacillus velezensis* genomes. Surprisingly, SNPs were not evenly distributed throughout the whole genome. (iv) Analysis of gene clusters revealed that rich gene clusters spread over *Bacillus velezensis* strains and some gene clusters are conserved in different strains. This study reveals that the strain WRN014 and other *Bacillus velezensis* strains have potential to be used as PGPR and biopesticide.

Keywords: *Bacillus velezensis*, comparative genomics, phylogenetic analysis, gene clusters

Introduction

Plant growth-promoting rhizobacteria (PGPR) refers to a class of bacteria that are colonized in plant roots and exert beneficial effects on plant development [1]. PGPR can promote plant growth by production of phytohormones, solubilization of inorganic phosphates, nitrogen fixation and as antagonists to soil-borne disease bacteria [2]. *Bacillus velezensis*, first isolated from the mouth of the river Vélez in Málaga (Southern Spain) in 2005 [3], is widely distributed in diversified environments such as plant rhizospheres, soil, rivers, human food, animal gut and seawater *et al.* and can easily be separated and cultured [4]. *Bacillus velezensis* is harmless to humans and animals and doesn't contaminate the environment, which means some strains have been used as biofertilizers and biopesticides commercially, such as strain SQR9 [5], RC 218 [6], LM2303 [7], FZB42 [8] *et al.* Therefore, it is an important species of

PGPR. Some characteristics of *Bacillus velezensis* about plant-growth promotion have been reported, for instance, the strain SQR9 is able to control the phytopathogenic fungus *Fusarium oxysporum f. sp. cucumerinum* J. H. Owen (FOC) [9–12], while the strain FZB42 produces secondary metabolites that suppress soil-borne plant pathogens [8].

The similarity of the 16S rRNA gene sequence between *Bacillus velezensis* and *Bacillus amyloliquefaciens* exceeds 99%[4]. At first, based on DNA hybridization analysis, *Bacillus velezensis* was deemed to be a later heterotypic synonym of *Bacillus amyloliquefaciens* [13]. From some articles, we can find that many *Bacillus velezensis* strains were earlier classified into *Bacillus amyloliquefaciens*, such as the strain FZB42 [8], SQR9 [5]. A previous study based on average nucleotide identity (ANI), DNA-DNA hybridization (DDH), and a core-genome-based phylogenetic analysis suggested that *Bacillus methylotrophicus*, *Bacillus amyloliquefaciens* subsp. *plantarum*, *Bacillus oryzicola* and *Bacillus velezensis* were all the same species [14]. Phylogenetic analysis results revealed that *Bacillus velezensis*, *Bacillus amyloliquefaciens* and *Bacillus siamensis* are closely clustered and they are referred to as the "operational group *Bacillus amyloliquefaciens*" [15]. Most *Bacillus velezensis* strains have an impressive capacity to produce secondary metabolites with antimicrobial activities, such as surfactin, fengycin, bacillomycin D, macrolactin, bacillaene, difficidin, oxidifficidin, plantazolicin, amylocyclicin, bacilysin *et al*. [6, 8, 15–17]. Because of habituating in different niches and environments, the *Bacillus velezensis* strains have their specific genomic and physiological characteristics. For instance, *Bacillus velezensis* FZB42 isolated from the *Beta vulgaris* (sugar beet) rhizosphere is used as PGPR, based on its abilities for root-colonizing and antibiotics production [18, 19]; while *Bacillus velezensis* GFP-2 isolated from the *Chiloscyllium plagiosum* (Whitespotted bamboo shark) intestine is used as a supplement to promote the growth of salmon, based on its abilities for controlling pathogenic bacteria [20].

With the development and low cost of high-throughput sequencing technologies, genome-based approaches have been used universally to obtain a comprehensive understanding of the genomic and metabolic characteristics of organisms [21]. The first available *Bacillus velezensis* genome sequence was for the strain FZB42, which was sequenced in 2007 [8]. Many studies about taxonomic analysis, phylogenetic analysis and antimicrobial activities analysis of *Bacillus velezensis* have been reported [22, 23]. So far, we can find almost 150 whole genome sequences of *Bacillus velezensis* in GenBank. In this study, we report the complete genome sequence of *Bacillus velezensis* WRN014 isolated from banana fields in Hainan, China. The assembled genome sequence has been deposited in NCBI Refseq database under accession number PJCI00000000. Forty-six complete genome sequences of the *Bacillus velezensis* were selected for comparative studies, including the newly sequenced *Bacillus velezensis* WRN014. Phylogenetic and population structure analysis suggested that 12 strains formed one clade including *Bacillus velezensis* WRN014, and the other strains formed one clade. Mutation and recombination analysis played an important role in genetic diversity. Multiple secondary metabolism clusters exist in *Bacillus velezensis* genomes, which can produce antibacterial agents. To decipher the evolutionary histories of the 46 *Bacillus velezensis* strains, gene gain and loss events have been predicted by mapping the inferred ortholog of genes to the species tree.

Materials and Methods

Bacillus velezensis WRN014 Genome Sequence and CDS Annotation

The plant growth-promoting rhizobacteria (PGPR) *Bacillus velezensis* strain WRN014 was isolated from banana fields located in Hainan Province in China. The strain was grown in Luria-Bertani (LB) broth at 30°C with moderate shaking and the cell was used for genomic extraction. The complete genome of *Bacillus velezensis* WRN014 was sequenced by combining Illumina Hiseq 2500 system and PacBio RSII high-throughput sequencing technology. The reads of the Illumina Hiseq 2500 system were assembled using an assembly pipeline called A5 (Andrew and Aaron's Awesome Assembly Pipeline) [24] and the software SPAdes genome assembler [25]. The reads of PacBio RSII were assembled into contigs using Hierarchical Genome Assembly Process 4 (HGAP4) [26] and Canu [27]. The gaps between contigs were filled by comparing the contigs that were assembled from the Illumina Hiseq 2500 system and PacBio RSII using the software MUMmer [28]. The quality of genome assembly was improved using the software Pilon [29]. The complete genome of WRN014 was annotated using the Prokaryotic Genomes Annotation Pipeline (PGAP) at NCBI [30].

Selection and Characterization of *Bacillus velezensis* Strains

Forty-six *Bacillus velezensis* and the type strain DSM7^T of *Bacillus amyloliquefaciens* were selected for a comparative genome analysis. The genome sequences and annotation information for the 46 *Bacillus* strains were downloaded from NCBI.

DDH Values and Phylogenetic Relationships

The DNA-DNA hybridization (DDH) values were obtained by means of genome-to-genome sequence comparison via GGDC 2.0 using Formula 2, and the digital variant (dDDH) was used to replace the tedious traditional approach [31, 32]. The Genome-to-Genome Distances (GGDs) were calculated by the Genome BLAST

Distance Phylogeny (GBDP) approach [33, 34]. The results were visualized using the matrix of dDDH values and heatmap. One phylogenetic tree was constructed using the web server of Composition Vector Tree Version 3 (CVTree3) based on all amino acid sequences of each strain, and 6 was the K-tuple length [35].

Population Structure

SNPs within the homologous regions of the core genes that are shared by the study strains in this article were extracted and used for population structure analyses. Population structure was identified by a Bayesian clustering approach using a Markov Chain Monte Carlo (MCMC) assignment method as implemented in the software STRUCTURE 2.3.4 [36]. The length of the burn-in period was set to 5,000 and the number of the MCMC reps after burn-in was set to 10,000. The number of populations was set from 3 to 7. The best number of populations (K) was identified using δK via the method of Evanno *et al.* [36].

Variant Calls: SNPs

Firstly, single nucleotide polymorphism (SNP) calls were performed using Snippy, which uses BWA Mem [37] to map the 250 bp single-end reads that were shredded from the contigs to the reference *Bacillus velezensis* FZB42 and then SNP calling was done with FreeBayes [38]. Whole genome alignment output from Snippy was used to identify the results about SNP distribution of the 46 *Bacillus velezensis* strains, using Gubbins (Genealogies Unbiased By Recombinations In Nucleotide Sequences) [39]. The density of SNP distribution was calculated throughout the genome sequence using a sliding-window size of 5 kb (step of the sliding window = 5 kb).

Recombination and Mutation Analysis

Gubbins was the software that was used for joint ancestral sequence reconstruction, phylogeny construction and identification of recombination. The ratio of rates at which recombination and mutation occur (ρ/θ) and relative contribution of recombination and mutation in the creation of the sample from a common ancestor (r/m) of every internal and external node of the phylogenetic tree was listed in Data Table S2.

Pan-Genome Analyses of *Bacillus velezensis*

The 46 *Bacillus velezensis* genome sequences were re-prediction and re-annotation using Prokka [40]. The GFF3 format files that were gained from Prokka were used for pan-genome analyses using the high-speed pan-genome pipeline Roary with the default value of 95% similarity among amino acid sequences [41]. The gene accumulation curve was produced via R packages ggplot2 using the results of Roary.

Identification of Functional Categories for Core and Strain-Specific Genes

The functional categories of core genes, accessory genes and strain-specific genes of the 46 *Bacillus velezensis* strains were identified, using the Clusters of Orthologous Groups (COGs) databases [42]. The best hits against the COGs databases were selected using BLAST+ as the category of the gene [43]. Each COG in the databases may be assigned to one of the 26 functional categories.

Secondary Metabolite Clusters

The secondary metabolite biosynthetic gene clusters of the 46 *Bacillus velezensis* strains were predicted using antiSMASH 4.0 [44]. The genomic homologies of the genes in the secondary metabolite clusters were identified by Roary [41]. The synteny maps of the gene clusters were generated using R package genoPlotR [45].

Gene Gain and Loss Events

A species tree for the set of genomes was inferred using a core genome alignment concatenation method. The multiple sequence alignments were constructed using MAFFT [46]. The Gblocks software with defaults settings was used to remove non-alignable regions resulting in amino acid final alignment [47]. A maximum likelihood tree was built with IQ-TREE using the GTR + I + G substitution model [48], a consensus tree was constructed from 10,000 bootstrap trees, and rooted by *Bacillus amyloliquefaciens* DSM7^T. To address the evolution history of the 46 *Bacillus velezensis* strains, ancestral family sizes were inferred using the program COUNT with Dollo parsimony [49]. Gene gain and loss events were reconstructed at both observed species and potential ancestors (leaves and nodes on the phylogenetic tree) using this method.

Nucleotide Sequence Accession Number

The sequence of *Bacillus velezensis* WRN014 was submitted to the NCBI GenBank database with the accession number of PJCI00000000.

Results

Genomic Features

A summary of the genomic features of the 46 *Bacillus velezensis* and 1 *Bacillus amyloliquefaciens* strains is shown in Table 1. The newly sequenced complete genome sequence of *Bacillus velezensis* WRN014 comprised a circular chromosome of 4,063,541 bp containing 4,062 CDSs, 27 rRNA, 86 tRNA, 9 genomic islands and 13 gene clusters with an average G + C content of 46.27% (Table 1; Fig. S1). The G + C contents of the 46 *Bacillus velezensis* genomes ranged from 45.78% to 46.80%. The genome size and number of protein-coding genes for 46 *Bacillus velezensis* strains ranged from 3.81~4.32 Mb and 3,610~4,436 genes, respectively. These genomes show little variation regarding their G + C content, genome size and amount of protein-

Table 1. Genome characteristics of 47 *Bacillus* strains.

Strain	Genome size (bp)	G+C (mol%)	CDS[a]	Country	Location	NCBI Accession No.	Plasmid
WRN014	**4,063,541**	**46.27**	**4,062**	**China**	***Musa* sp. rhizosphere**	PJCI00000000	0
FZB42	3,918,589	46.50	3,687	Germany	Sugar beet field	NC_009725.1	0
CAU B946	4,019,861	46.51	3,792	China	*Oryza sativa* rhizosphere	NC_016784.1	0
YAU B9601-Y2	4,242,774	45.85	4,042	China	*Triticum* spp. rhizosphere	NC_017061.1	0
AS43.3	3,961,368	46.60	3,669	USA	*Triticum* spp. head	NC_019842.1	0
UCMB5036	3,910,324	46.60	3,691	Tajikistan	Inner tissues of the cotton plant	NC_020410.1	0
UCMB5033	4,071,167	46.20	3,892	Tajikistan	Cotton rhizosphere	NC_022075.1	0
UCMB5113	3,889,532	46.70	3,656	Ukraine	Soil	NC_022081.1	0
NAU-B3	4,204,608	45.99	4,068	China	*Triticum* spp. rhizosphere	NC_022530.1	1
TrigoCor 1448	3,957,904	46.50	3,721	USA	Wheat rhizosphere	NZ_CP007244.1	0
SQR9	4,117,023	46.10	3,902	China	*Cucumis sativus* rhizosphere	NZ_CP006890.1	0
L-H15	3,905,973	46.60	3,781	China	Cucumber seedling substrate	NZ_CP010556.1	0
L-S60	3,903,017	46.67	3,779	China	Turfy soil	NZ_CP011278.1	0
NJN-6	4,052,546	46.60	3,842	China	*Musa* sp. rhizosphere	NZ_CP007165.1	0
JJ-D43	4,105,955	46.24	3,937	Korea	Doenjang	NZ_CP011346.1	0
YJ11-1-4	4,006,637	46.42	3,639	Korea	Doenjang	NZ_CP011347.1	0
G341	4,009,746	46.49	3,808	Korea	*Panax ginseng* rhizosphere	NZ_CP011686.1	0
B25	3,862,757	46.70	3,697	Switzerland	Inner tissues of *Platanus* x *acerifolia*	NZ_LN999829.1	1
CC09	4,167,153	46.10	3,941	China	*Cinnamomum camphora* leaf tissues	NZ_CP015443.1	0
S3-1	3,929,772	46.50	3,771	China	*Cucumis sativus rhizosphere*	NZ_CP016371.1	0
LS69	3,917,761	46.40	3,678	China	*Oryza sativa* field	NZ_CP015911.1	0
M75	4,007,450	46.60	3,790	Korea	Cotton waste	NZ_CP016395.1	0
9912D	4,241,576	45.99	4,436	China	sediment sample from Bohai Sea	NZ_CP017775.1	1
GH1-13	4,143,608	46.17	4,016	Korea	Rice paddy field	NZ_CP019040.1	1
JTYP2	3,929,789	46.50	3,656	China	*Echeveria laui* leaves	NZ_CP020375.1	0
9D-6	3,963,726	46.40	3,852	Canada	Potato rhizosphere	NZ_CP020805.1	0
CBMB205	3,929,792	46.50	3,812	Korea	Rice rhizosphere	NZ_CP011937.1	0
GQJK49	3,929,760	46.50	3,677	China	*Lycium barbarum* rhizosphere	NZ_CP021495.1	0
T20E-257	3,900,066	46.69	3,791	Korea	*Solanum lycopersicum L.* rhizosphere	NZ_CP021976.1	1
157	4,020,691	46.39	3,864	China	Bark of *Eucommia ulmoides*	NZ_CP022341.1	2
NJAU-Z9	3,872,560	46.78	3,740	China	Pepper rhizosphere	NZ_CP022556.1	2
LABIM40	3,972,310	46.50	3,882	Brazil	n.a.[b]	NZ_CP023748.1	0
L-1	4,090,582	46.52	3,934	China	*Pyrus* spp. rhizosphere	NZ_CP023859.1	0
NKG-1	4,197,217	46.30	4,138	China	Rare dormant volcanic soils	NZ_CP024203.1	0
CN026	3,995,812	46.40	3,783	Belgium	Chicken feces	NZ_CP024897.1	0
Lzh-a42	4,246,605	46.00	4,074	China	*Lycopersicon* sp. rhizophere	NZ_CP025308.1	0
GFP-2	3,975,220	46.40	3,737	China	Chiloscyllium plagiosum intestine	NZ_CP021011.1	0
QST713	4,233,757	45.90	4,159	France	n.a.	NZ_CP025079.1	0
LDO2	3,947,271	46.50	3,819	China	Peanut rhizosphere	NZ_CP029034.1	0
BS-37	4,013,888	46.50	3,889	China	Petroleum-contaminated soil	NZ_CP023414.1	0
W1	4,237,431	45.84	4,157	China	Two-spotted spider mites	NZ_CP028375.1	0
DSYZ	4,321,436	45.78	4,245	China	Garlic rhizosphere	NZ_CP030150.1	1
BIM B-439D	3,978,954	46.50	3,851	Belarus	Soil	NZ_CP032144.1	0
S141	3,974,582	46.50	3,844	Thailand	*Glycine max* rhizosphere	NZ_AP018402.1	0
JT3-1	3,929,799	46.50	3,813	China	Feces of Bos grunniens	NZ_CP032506.1	0
SB1216	3,814,720	46.80	3,610	USA	Great Salt Plains	CP015417.1	0
DSM7^T	3,980,199	46.08	3,870	Germany	n.a.	NC_014551.1	0

[a]CDS, coding sequences.

[b]n.a., not available.

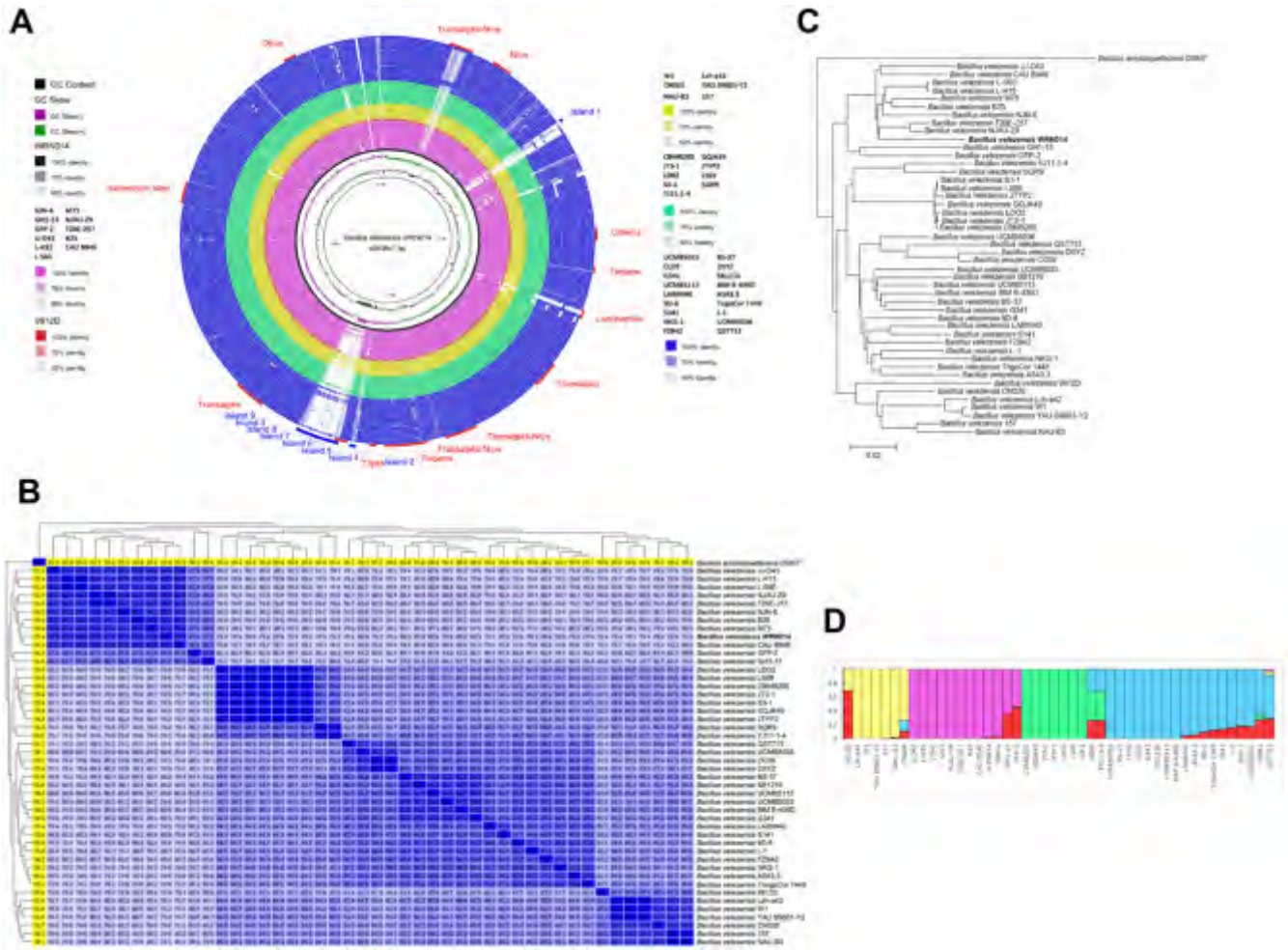

Fig. 1. Genomic, phylogenetic and population structure characteristics of *Bacillus velezensis* WRN014.
(**A**) Comparison of the genome of the WRN014 strain with the other 45 genomes of *Bacillus velezensis* strains. The three inner rings represent the DNA size, GC content and GC Skew. The 48 outer rings represent the genomes of WRN014, NJN-6, GH1-13, GFP-2, JJ-D43, L-H15, L-S60, M75, NJAU-Z9, T20E-257, B25, CAU B946, 9912D, Lzh-a42, W1, YAU B9601-Y2, 157, NAU-B3, CN026, CBMB205, GQJK49, JT3-1, JTYP2, LDO2, LS69, S3-1, SQR9, YJ11-1-4, UCMB5033, BS-37, CC09, DSYZ, G341, SB1216, UCMB5113, BIM B-439D, LABIM40, AS43.3, 9D-6, TrigoCor 1448, S141, L-1, NKG-1, UCMB5036, FZB42 and QST713 and gene clusters, genomic islands from inner to outer. (**B**) The digital DNA-DNA hybridization (dDDH) values in the 46 genomes of *Bacillus velezensis* and 1 genome of *Bacillus amyloliquefaciens* DSM7^{T}. (**C**) Phylogenetic tree of 46 *Bacillus velezensis* genomes and *Bacillus amyloliquefaciens* DSM7^{T} genome. The phylogenetic tree was constructed using all amino acid sequences of each strain on the Web Server of Composition Vector Tree Version 3 (CVTree3), and 6 was the K-tuple length. The phylogenetic tree was rooted by *Bacillus amyloliquefaciens* DSM7^{T}. (**D**) The 46 strains were divided into 5 populations (K = 5), and individuals are shown by thin vertical lines, which are divided into K-colored segments representing the estimated membership probabilities (Q) of each individual.

encoding genes, suggesting that the *Bacillus velezensis* may be a newly formed species. A map to compare the similarity of the strain WRN014 with the other 45 *Bacillus velezensis* strains was displayed in Fig. 1A using BRIG (Blast Ring Image Generator) [50].

DDH Values and Phylogenetic Analyses

The genome sequence of *B. velezensis* WRN014 was compared to the other 46 available genomic sequences by calculating the digital DNA-DNA hybridization values (dDDH) and pair-wise genome content distances. The distance matrix was displayed in Table S1. We have converted the matrix of dDDH values to a heat map and listed the dDDH values in the map (Fig. 1B). The strain WRN014 forms a cluster with 11 other *Bacillus velezensis* strains including GFP-2 (87.6%), NJN-6 (95.5%), JJ-D43

(94.4%), CAU B946 (95.1%) , L-H15 (95.1%), L-S60 (95.2%), NJAU-Z9 (95.5%), T20E-257 (94.7%), B25 (95.7%), GH1-13 (87.7%) and M75 (95.2%) with a dDDH value over 85%. The dDDH values of 46 *Bacillus velezensis* strains with *Bacillus amyloliquefaciens* DSM7^{T} are below 70% which is the standard for species definition. The 46 *Bacillus velezensis* strains formed 2 clusters in the heat map.

The whole genome-based method has been used to construct a phylogenetic tree. The type strain *Bacillus amyloliquefaciens* DSM7^{T} which has close relationship with *Bacillus velezensis* was selected as out-group. The method is alignment-free using total amino acid sequences (Fig. 1C). Supporting the previous result of dDDH analyses, the whole genome-based phylogenetic tree showed that *Bacillus velezensis* WRN014 is more related to the strains including GFP-2, NJN-6, JJ-D43, CAU B946, L-H15, L-S60, NJAU-Z9, T20E-257, B25, GH1-13 and M75 and the 46 *Bacillus velezensis* strains obviously formed 2 clades in the phylogenetic tree. Only a small difference exists between the phylogenetic tree and dDDH analysis. Even though the strains come from difference habitats, they have high similarity with each other, suggesting the short divergency time of the species *Bacillus velezensis*.

Population Analysis

The population structure of the 46 *Bacillus velezensis* genomes was analyzed by using the Bayesian clustering program STRUCTURE based on the 171,743 SNPs of core genes, with *K* changing progressively from 3-7 [36] (Fig. 1D). Analysis clearly divides the 46 *Bacillus velezensis* strains into 5 specific groups. We can find the strain WRN014 clustered together with the strains CAU B946, NJN-6, JJ-D43, M75, L-H15, L-S60, NJAU-Z9, T20E-257, B25, GH1-13, and GFP-2, and except for strains GFP-2 and GH1-13, 10 other *Bacillus velezensis* strains showed little variation. The results of population analysis show similarity with the previous phylogenetic analysis.

Single Nucleotide Polymorphisms (SNPs) Analysis

Single nucleotide polymorphisms (SNPs) are an important genetic variation in bacteria and are usually used as a genomic imprint of natural selection [51, 52]. In this study, SNP analysis is performed using all of the 46 *Bacillus velezensis* genomes. The total 245,296 polymorphic sites were identified and used to construct a phylogenetic tree using RAxML [53] (Fig. 2). The phylogenetic tree that was constructed based on the full-genome SNP sites can

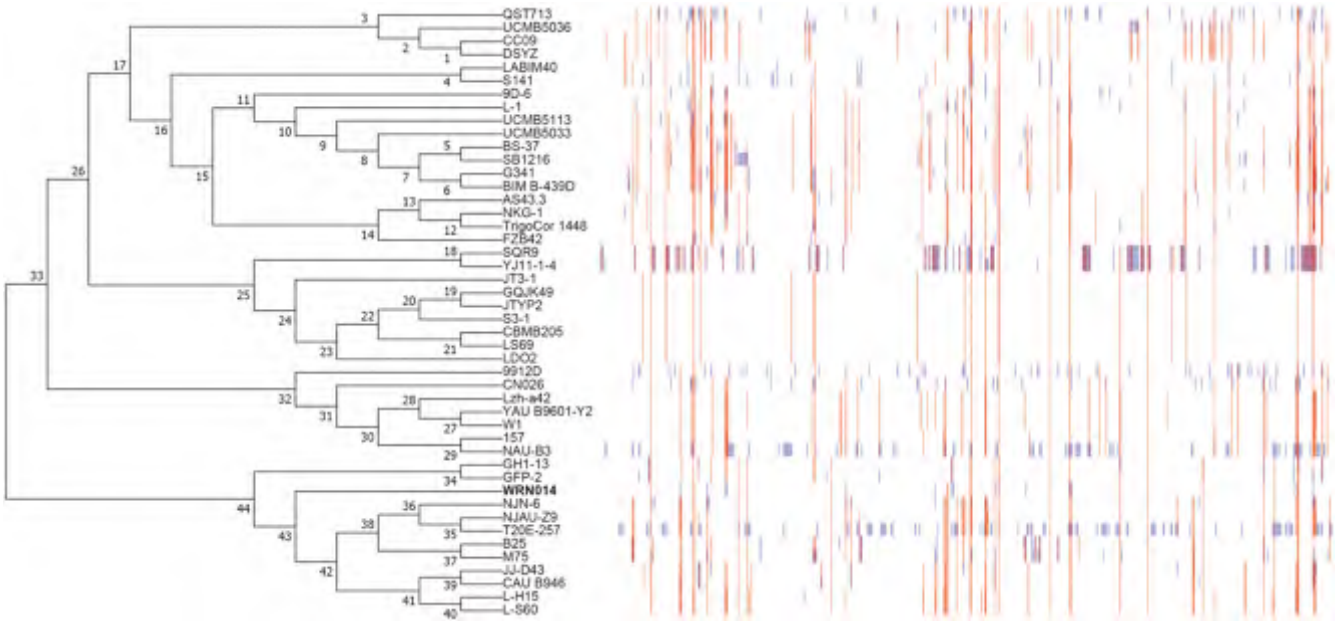

Fig. 2. Gubbins analyses of 46 *Bacillus velezensis* strains.
The maximum likelihood phylogenies generated from the whole genome alignment of 46 *Bacillus velezensis* strains using the Gubbins algorithm relying on RAxML for constructing the phylogeny in each iteration. The number on the phylogenetic tree represents the internal node. The panel represents the pattern of predicted recombinations. Each column relates to a base and each row represents a strain in the phylogeny. Red blocks indicate predicted recombinations occurring on an internal branch, which are shared by multiple isolates through common descent. Blue blocks represent recombinations that occur on terminal branches, which are unique to individual strains.

effectively distinguish the closely related strains. The result suggested that the 46 *Bacillus velezensis* strains formed 4 clades in the phylogenetic tree and the strain WRN014 was closely related to GFP-2, GH1-13, NJAU-Z9, T20E-257, JJ-D43, CAU B946, M75, B25, L-H15, L-S60, and NJN-6. The tree is basically consistent with the previous analysis.

In order to illuminate the features of SNP distribution on the genomes, we estimated SNP density throughout the 46 *Bacillus velezensis* genomes using a sliding window of 5 kb. Our results revealed that the total of 784 regions throughout the genomes were identified and SNPs were not evenly distributed among these regions (Dataset S1). The mean density of SNPs throughout the genomes is 62.6 SNPs/kb. Of all the 784 regions, 31 low density regions that have under 25 SNPs/kb and 27 high density regions that have more than 90.6 SNPs/kb were identified, and about 6 regions have almost no SNPs. The genes in the regions of low-density SNPs have functions mainly related to translation, ribosomal structure and biogenesis (J), transcription (K) and other house-keeping functions. However, of the 27 regions with high SNP density, the functions of the genes are mainly related to amino acid transport and metabolism (E), carbohydrate transport and metabolism (G), lipid transport and metabolism (I), coenzyme transport and metabolism (H), secondary metabolite biosynthesis, transport and catabolism (Q) and transcription (K). The genes concerned with the fundamental function of the bacteria may have few variations, and the SNP density is relatively low, to maintain cell stability. To adapt to various environments, many variations on the genes have occurred in relation to metabolic function.

The Recombination and Mutation Rate

Analysis of SNPs revealed that genetic diversity was identified in the 46 *Bacillus velezensis* genomes. Many different patterns of genetic variation exist in bacteria including point mutation, genetic recombination, gene duplication *et al*. Identification of recombination and its phylogenetic history is crucial for tracing the evolution of bacteria [39]. The ratio of recombination and mutation (ρ/θ) occurrence rates and relative contribution in the creation of the samples from a common ancestor (r/m) was calculated using Gubbins [39]. The results showed that the high rates of recombination occur in the strains SQR9 (ρ/θ = 0.917), YJ11-1-4 (ρ/θ = 0.686), NAU-B3 (ρ/θ = 0.716), T20E-257 (ρ/θ = 0.411), S3-1 (ρ/θ = 0.5), GQJK49 (ρ/θ = 0.2), LDO2(ρ/θ = 0.5), YAU B9601-Y2(ρ/θ = 0.231), and JT3-1(ρ/θ = 0.167), but extremely low rates of recombination occur in the other strains and clades (ρ/θ =0-0.06). The r/m value that represents the relative impact of recombination on sequence diversity is very high for SQR9 (r/m = 55.5), YJ11-1-4 (r/m = 52.1), NAU-B3 (r/m = 45.2), T20E-257 (r/m = 28.7), S3-1(r/m = 3), GQJK49 (r/m = 1.4), LDO2 (3.5), YAU B9601-Y2 (r/m = 1.5), and JT3-1 (r/m = 1.7), indicating a greater number of substitutions being introduced. However, only low r/m values for the other strains and clades were found (Fig. 2, Table S2). The results taken together suggest that the genomes of the strains SQR9, YJ11-1-4, NAU-B3, T20E-257, S3-1, GQJK49, LDO2, YAU B9601-Y2 and JT3-1 are more deeply affected by recombination while it has little effect on the other strains.

The Pan-Genome Analyses of the 46 *Bacillus velezensis* Genomes

The genome sizes and numbers of genes of the 46 *Bacillus velezensis* strains were relatively similar. Instead of the sequence comparison against a reference strain, the entire set of nucleotide sequences of the total genes were clustered to identify the ortholog clusters. Pan-genome analysis of the 46 *Bacillus velezensis* genomes revealed that 8,907 ortholog clusters that constituted the pan-genome, corresponding to more than two-fold the average for genes of the 46 genomes. The gene accumulation curve showed that the size of the *Bacillus velezensis* pan-genome may grow with the number of strains, and this pan-genome was considered in an open state (Figs. 3A and 3B). The results revealed that *Bacillus velezensis* has flexible genome contents, reflecting the diversity of metabolic functions for adapting to various environments. The numbers of core genomes, strain-specific genomes and accessory genomes are 2,952, 2,527, and 3,482 respectively. Along with the addtion of the analyzed genome, the number of core genomes converges to a constant value by the slope of exponential decay. So many strain-specific genomes and accessory genomes that are thought to contribute to the species diversity and generally provide functions that were not essential to viability may indicate the high ability of *Bacillus velezensis* to adapt to various environmental niches.

Functional Distribution of Ortholog Clusters

After obtaining the amino acid sequences of core genomes, strain-specific genomes and accessory genomes, we identified the functional categories of these clusters using the Clusters of Orthologous Groups (COGs) database [42]. However, the functional categories of only 81.9% (2418/2952) of core genes, 31.1% (787/2527) of strain-specific genes and 41.2% (1412/3428) of accessory genes could be determined using the COGs database. The fact

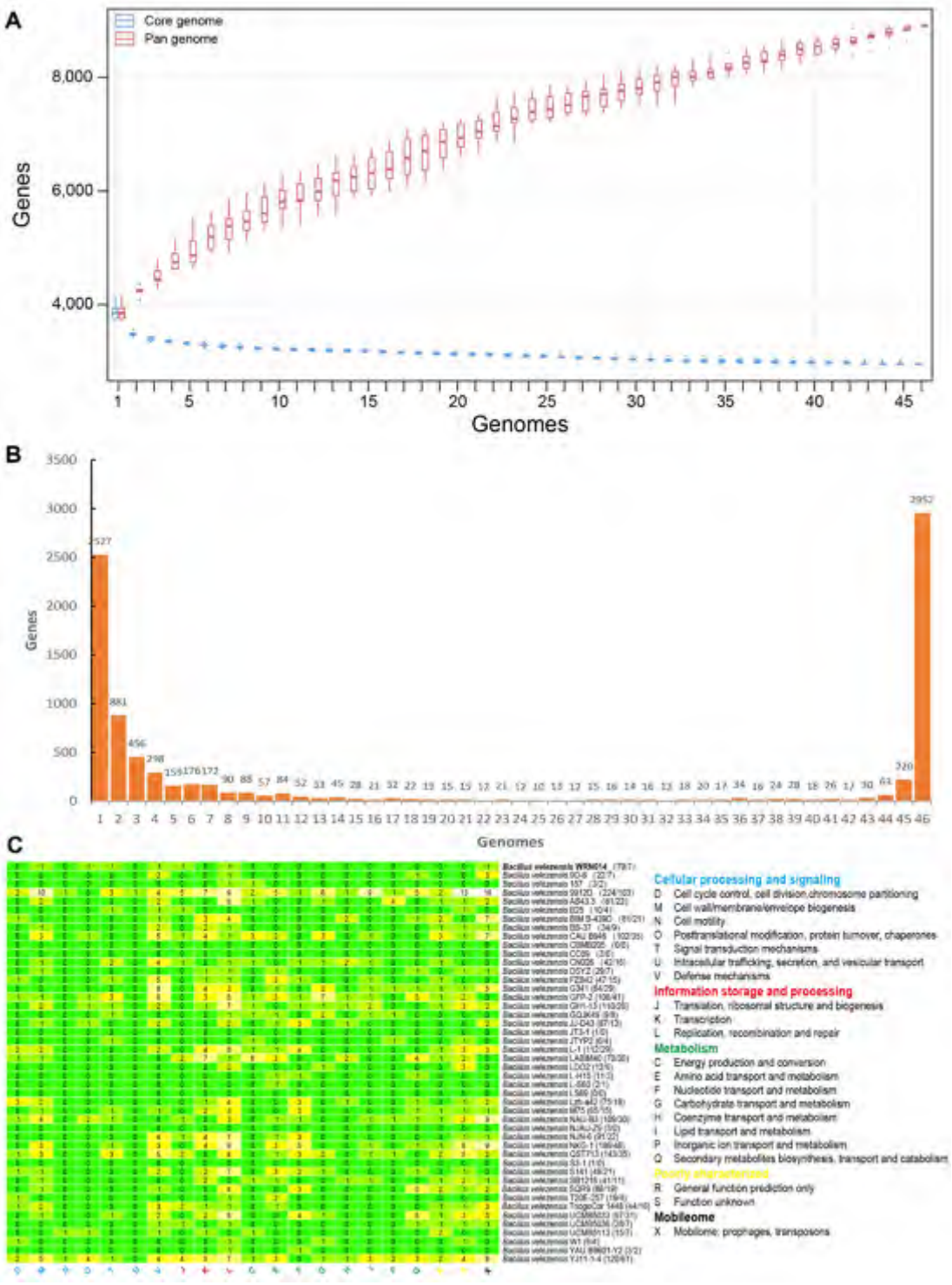

Fig. 3. Pan-genome analysis of 46 *Bacillus velezensis* strains.
(**A**) Pan-genome accumulation curve. The blue boxes denote the number of unique genes discovered with the sequential addition of new genomes. The orange boxes denote the number of core genes discovered with the sequential addition of new genomes. (**B**) Gene occurrence plot shows the core genes and additional accessory genes of *Bacillus velezensis*. (**C**) Functional classification of strain-specific genes in 46 *Bacillus velezensis* strains. The number in each square represents the COG assignment in each functional category. The annotated gene number and total specific gene number of each strain were listed behind *Bacillus velezensis* strain names.

that so many strain-specific genes and accessory genes could not be categorized using the COGs database revealed that some strains may form new functions to adapt to their specific environmental niches. The most abundant functions in the core genes of *Bacillus velezensis* are associated with metabolism. The overall proportion of genes related to metabolic functions was 45.0% (1088/2418), 27.7% (218/787) and 35.0% (494/1412) in the core genes, strain-specific genes and accessory genes, respectively (Fig. 4A). More specifically, amino acid transport and metabolism (E), carbohydrate transport and metabolism (G), and translation, ribosomal structure and biogenesis (J) are abundant in the core genes, suggesting that these genes were relatively conserved in *Bacillus velezensis*. The number of genes related to the mobilome, prophages and transposons (X) is only 7 in the core genes, however, 92 were found in the strain-specific genes and accessory genes, suggesting that the mobilome-related genes were more abundant among strain-specific genes than core genes. Prophages and transposons may present an important function in *Bacillus velezensis* in adapting to their specific environmental niches. The gene exchange occurred frequently between the *Bacillus velezensis* strains with other species in the common environment by lateral gene transferring (LGT), transfection, and other genetic information exchange processes (Fig. 4B).

The number of the specific genes among the 46 *Bacillus velzensis* genomes was 2,527 ranging from 0 to 244 for each strain. The lowest number was encoded by the strain LS-69 (0), CBMB205 (0) and the highest number was identified in the strain 9912D (149) (Fig. 3C). Although a high number of strain-specific genes (approximately 72%) was not assigned to the COG categories, the other strain-specific genes fell into different functional categories. A higher proportion of strain-specific genes in most of the strains was assigned to the mobilome: prophages, transposons (X), replication, recombination and repair (L), transcription (K), cell wall/membrane/envelope biogenesis (M) and defense mechanisms (V).

Secondary Metabolite Clusters

Through the antiSMASH genome analysis tool [44], thirteen clusters of secondary metabolites have been identified in the genome of strain WRN014, three Transatpks-Nrps, two Transatpks (trans-Acyl Transferase Polyketide Synthetase), two terpene, one Nrps (Non-Ribosomal Peptide Synthetase), one Otherks, one Lantipeptide, one T3pks, one Bacteriocin-Nrps and one otherKS. Seven clusters have been identified as being involved in the synthesis of macrolactin, bacillaene,

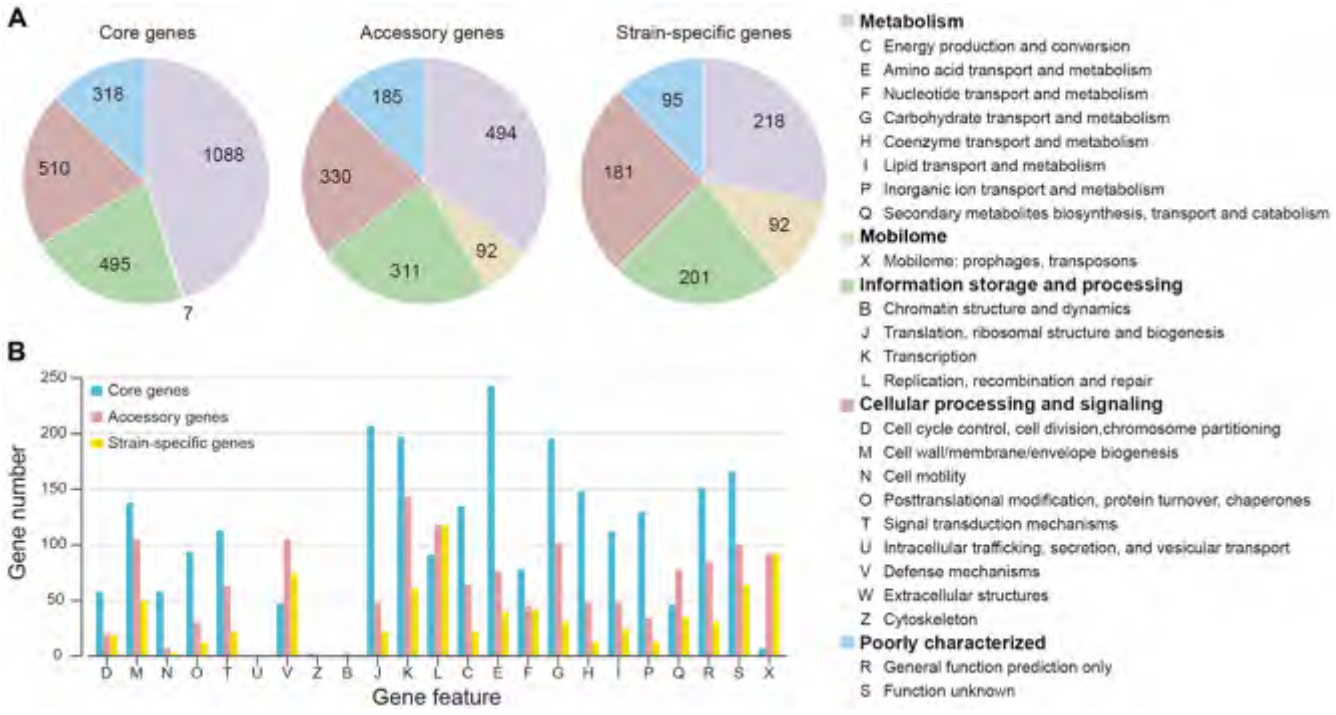

Fig. 4. Differential distribution of COG functional categories in core and strain-specific genes.
(**A**) Proportion of five classes of functional categories in core, accessory and strain-specific genes. **(B)** Functional categories in core, accessory and strain-specific genes.

fengycin, difficidin, bacillibactin (siderophore), bacilysin and surfactin. The genomes of many microorganisms contain multiple biosynthetic gene clusters (BGCs) that code for production of secondary metabolites. The secondary metabolites such as antimicrobial peptides (AMPs) from *Bacillus velezensis* play a vital role in biological control of foliar, soil-borne, and post-harvest diseases [53]. The secondary metabolite gene clusters of all the 46 *Bacillus velezensis* strains were identified using anti-SMASH 4.0 [44]. The numbers and categories of gene clusters were listed in Fig. S2 and the stations of gene clusters on their genomes were listed in Table S3. The gene distribution of gene clusters of *Bacillus velezensis* WRN014 were listed in Table S4. The result suggested that NRPs and PKs have immense structural diversity and functional diversity.

Synteny analysis and gene structure analysis of the 13 gene clusters were carried out depending on the homology and distribution of the genes in the gene clusters. Among the 13 gene clusters, 3 clusters were specific and existed only in one strain, 5 clusters existed in two strains and the other 16 clusters existed in more strains. The 46 *Bacillus velezensis* strains don't have absolutely common and identical secondary metabolite synthesis gene clusters. Even though several gene clusters are shared by some strains, the gene structure is different. On the basis of the previous SNP analysis result, the high-density SNP regions contain the genes related to secondary metabolism. The map of synteny analysis comparing the genes is shown in Fig. 5. The data suggest that several new metabolite synthesis gene clusters may be horizontally transferred from other species in the common environment and the metabolite synthesis gene clusters have changed in different strains.

Reconstructing Gene Gain and Loss Events

To decipher the evolutionary histories of the *Bacillus velezensis*, gene gain and loss events were predicted by mapping the inferred ortholog of genes to the species tree. The species tree was inferred from 2,377 single-copy core genes shared by 46 *Bacillus velezensis* strains, as well as the *Bacillus amyloliquefaciens* strain DSM7^{T} which act as out-group. Conserved blocks from multiple amino acid sequences alignment of 2,377 single-copy core genes were selected by using Gblocks[32]. *Bacillus velezensis* WRN014 is clustered with GFP-2, NJN-6, JJ-D43, CAU B946, B25, NJAU-Z9, T20E-257, L-H15, L-S60, GH1-13 and M75 and the 46 *Bacillus velezensis* strains obviously formed 2 clades in the species tree (Fig. 6). *Bacillus velezensis* is monophyletic and 46 strains share 4,916 orthologous genes with their common ancestor. We estimated that the ancestor of *Bacillus velezensis* possessed more gene families than the extant organisms. As time went on, gene loss and gain events occurred frequently in all lineages. A large number of gene gain events occurred at nodes A and B of the tree and two obvious clades formed in this step. When forming the extant organisms, many gene loss events occurred in every strain. The environmental selection leads to the loss and acquisition of the specific genes from organisms in the new microbial communities, which makes the strains better adapted to new habitats.

Discussion

In this article, we isolated a *Bacillus velezensis* strain WRN014 from banana fields in Hainan, China, and made a comparative genomic analysis with 45 other *Bacillus velezensis* strains which were sequenced previously. Through comparative genomic analysis of the 46 *Bacillus velezensis* strains, we present a global view of these genomes, and reveal that these genomes have similar genome architecture and high average nucleotide identity. Different methods were used to constructed the phylogenetic trees, including whole genome-based method, full-genome single nucleotide polymorphism (SNP) sites-based method and core genomes-based method. These phylogenetic trees showed high similarity with each other and suggested that the 46 *Bacillus velezensis* strains clustered to two obvious clades of the tree and *Bacillus velezensis* WRN014 was more related to the strains including GFP-2, NJN-6, JJ-D43, CAU B946, B25, NJAU-Z9, T20E-257, L-H15, L-S60, GH1-13, and M75. The other clade can be divided into 3 small branches, one branch includes 7 strains: 9912D, CN026, YAU B9601-Y2, W1, Lzh-a42, NAU-B3, and 157, another branch includes 9 strains: YJ11-1-4, SQR9, S3-1, LDO2, JT3-1, JTYP2, GQJK49, LS69, and CBMB205, and 17 other strains belong to the last branch. From the analyses of digital DNA-DNA hybridization, we got the all-and-all dDDH values of the strains in this article, and made clusters of these values using Euclidean Distance. Through comparison of dDDH analyses and phylogenetic analyses, we found similar results, forming 2 large clusters or 4 small clusters. Population structure analysis revealed similar results with phylogenetic analysis, forming 5 specific groups, especially 9912D forming a sole group because of diversity sequence exchanges with the environment. All results suggested that the *Bacillus velezensis* can be divided into 2 subtypes based on genome sequence.

The differential distribution of COG categories in the

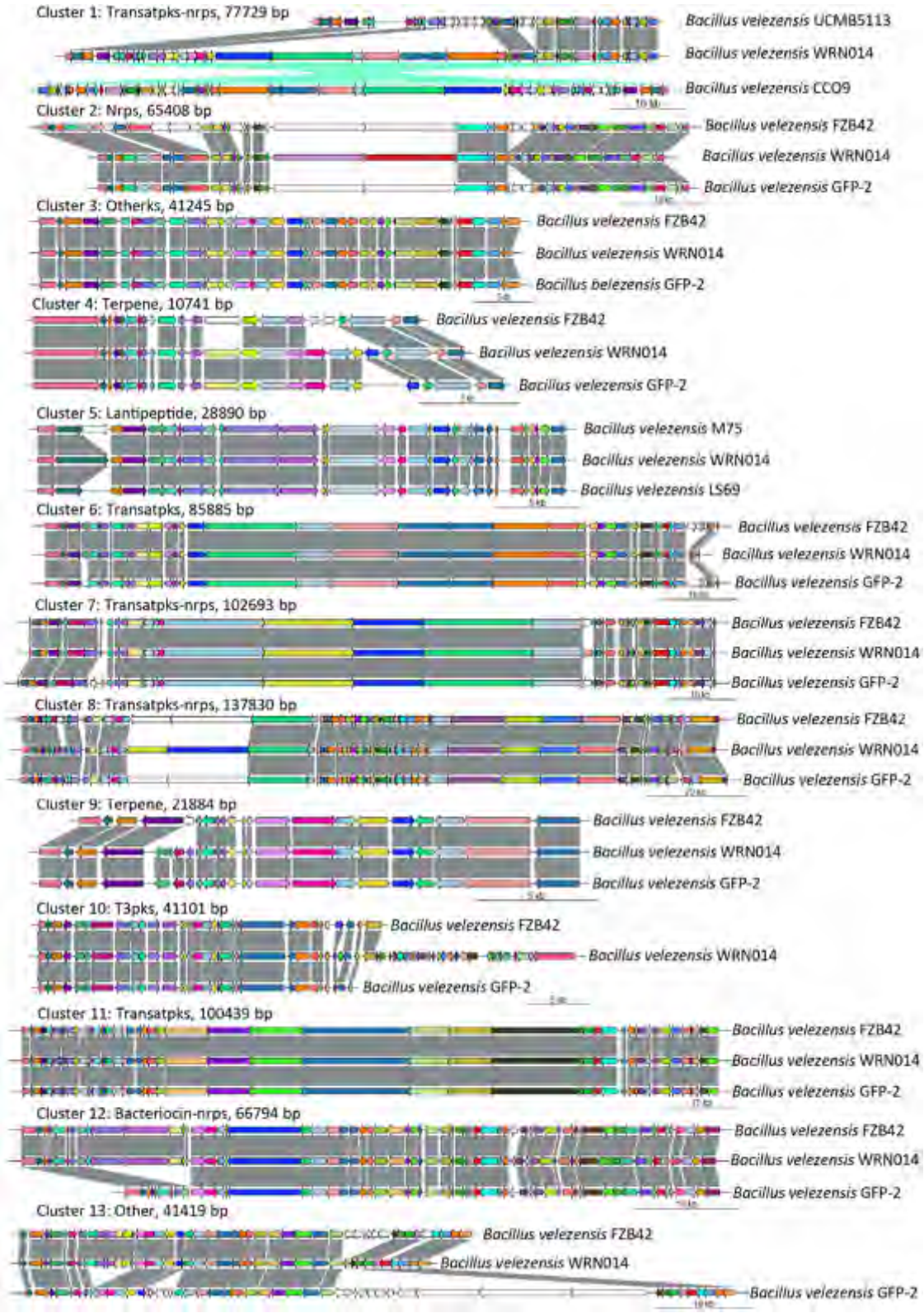

Fig. 5. Comparison of biosynthetic gene clusters from *Bacillus velezensis* WRN014 with other *Bacillus velezensis* strains.
Regions of conserved synteny were marked with grey (+) and green (-) shadow. Different genes are filled with different color, and genes with the same color are homologous to each other. The gene product was deduced by homologous blast.

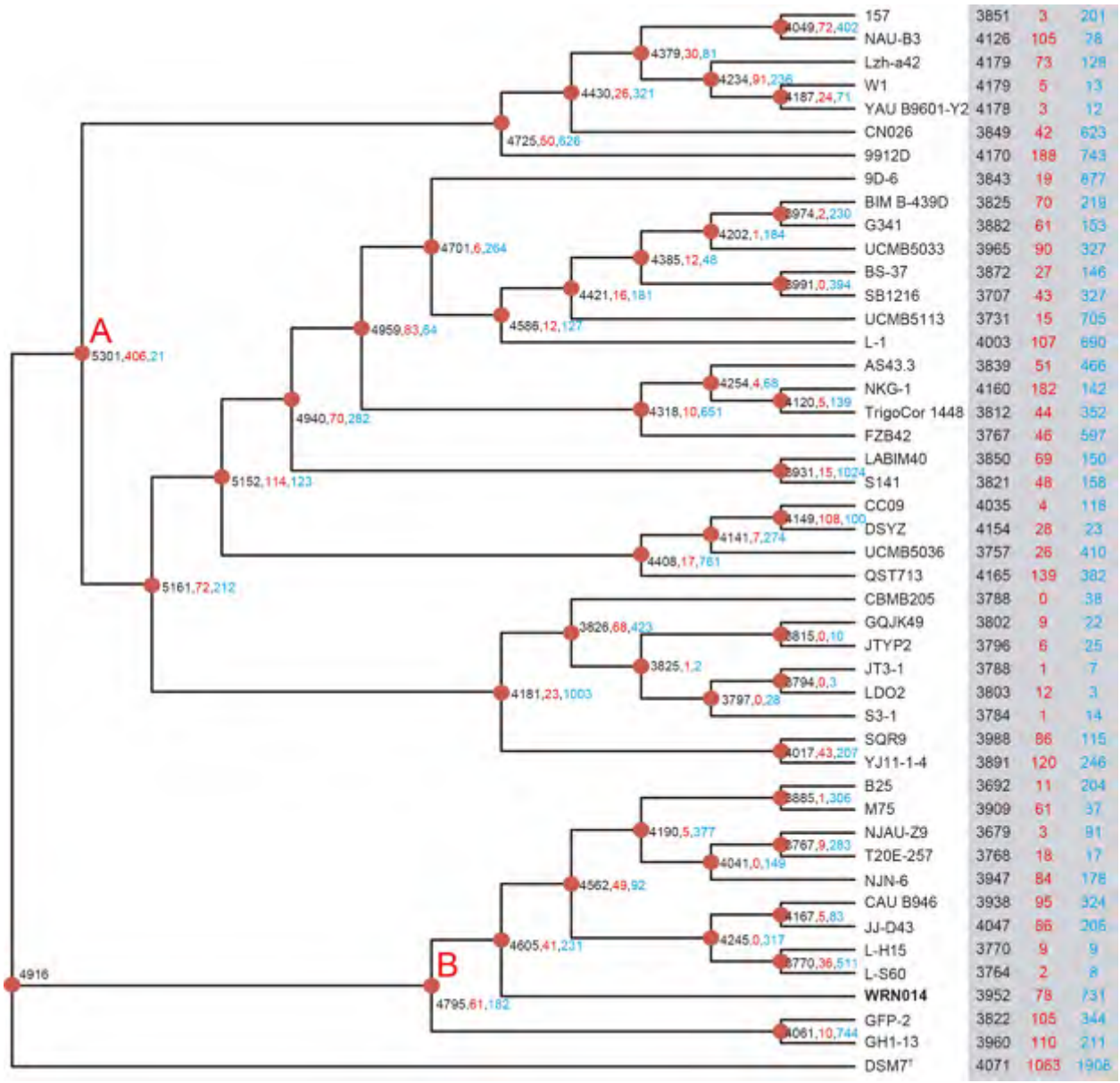

Fig. 6. Ancestral genome content reconstruction using COUNT software.
The phylogenetic tree was constructed using IQ-TREE. The numbers of gain and loss events were marked at each lineage of the tree. Reds represent gain events and blues represent loss events. The total gene numbers were marked at each lineage of the phylogenetic tree.

protein-coding genes of the 46 *Bacillus velezensis* strains was displayed in Fig. S3. By comparing the COG analysis and phylogenetic analyses, we found different results. That may be because many genes cannot be successfully categorized using the COG database, especially accessory genes.

Bacillus velezensis is always used as plant growth-promoting rhizobacteria (PGPR) to promote plant growth and control soil-borne disease. *Bacillus velezensis* WRN014 was predicted to own 13 gene clusters of secondary metabolites including three Transatpks-Nrps, two Transatpks (trans-Acyl Transferase Polyketide Synthetase), two terpene, one Nrps (Non-Ribosomal Peptide Synthetase), one Otherks, one Lantipeptide, one T3pks, one Bacteriocin-Nrps and one otherKS which produce antibacterial agents to control soil-borne disease. Comparative analysis of the gene clusters of 46 strains suggested that although the *Bacillus velezensis*

strains were isolated from different geographical locations and diverse environments, they have similar secondary metabolite gene clusters.

Based on constructing gene gain and loss events, we can know the genome of their common ancestor was larger than extant strains and some secondary metabolite gene clusters also existed in the ancestor genome. Analysis of gene function revealed that genes relevant to amino acid transport and metabolism, carbohydrate transport and metabolism, and translation, ribosomal structure and biogenesis are abundant in the core genes of the 46 *Bacillus velezensis* strains. SNP and recombination play important roles in genetic diversity of *Bacillus velezensis*. The genes in the regions of high SNP density are more related to metabolism.

To summarize, through comparative genomic analysis we found strain *Bacillus velezensis* WRN014's genetic relationship with other *Bacillus velezensis* strains, which will be instructive in guiding us to reveal more characteristics about the strain through further in vitro experiments. The fact that abundant secondary metabolite gene clusters exist in genomes of *Bacillus velezensis* illustrates that they have potential to be used as PGPR.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2017YFD0201401) and the National Natural Science Foundation of China (NSFC No. 31670113 and NSFC No. 31670011).

Conflict of Interest

The authors have no financial conflicts of interest to declare.

References

1. Lugtenberg B, Kamilova F. 2009. Plant-growth-promoting rhizobacteria. *Annu. Rev. Microbiol*. **63:** 541-556.
2. Bloemberg GV, Lugtenberg BJ. 2001. Molecular basis of plant growth promotion and biocontrol by rhizobacteria. *Curr. Opin. Plant Biol*. **4:** 343-350.
3. Ruiz-García C, Béjar V, Martínez-Checa F, Llamas I, Quesada E. 2005. *Bacillus velezensis* sp. nov., a surfactant-producing bacterium isolated from the river Vélez in Málaga, southern Spain. *Int. J. Syst. Evol. Microbiol*. **55:** 191-195.
4. Ye M, Tang X, Yang R, Zhang H, Li F, Tao F, et al. 2018. Characteristics and application of a novel species of *Bacillus*: *Bacillus velezensis*. *ACS Chem. Biol*. **13:** 500-505.
5. Zhang N, Yang D, Wang D, Miao Y, Shao J, Zhou X, *et al*. 2013. Whole transcriptomic analysis of the plant-beneficial rhizobacterium *Bacillus amyloliquefaciens* SQR9 during enhanced biofilm formation regulated by maize root exudates. *BMC Genomics* **16**: 685-694.
6. Palazzini JM, Dunlap CA, Bowman MJ, Chulze SN. 2016. *Bacillus velezensis* RC 218 as a biocontrol agent to reduce *Fusarium* head blight and deoxynivalenol accumulation: genome sequencing and secondary metabolite cluster profiles. *Microbiol.Ekd. Res*. **192:** 30-36.
7. Chen L, Heng J, Qin S, Bian KA. 2018. A comprehensive understanding of the biocontrol potential of *Bacillus velezensis* LM2303 against *Fusarium* head blight. *PLoS One* **13:** e0198560-e0198581.
8. Chen XH, Koumoutsi A, Scholz R, Eisenreich A, Schneider K, Heinemeyer I, *et al.* 2007. Comparative analysis of the complete genome sequence of the plant growth-promoting bacterium *Bacillus amyloliquefaciens* FZB42. *Nat. Biotechnol*. **25:** 1007-1014.
9. Cao Y, Zhang Z, Ling N, Yuan Y, Zheng X, Shen B, *et al*. 2011. *Bacillus subtilis* SQR 9 can control *Fusarium* wilt in cucumber by colonizing plant roots. *Biol. Fertil Soils* **47:** 495-506.
10. Qiu M, Zhang R, Xue C, Zhang S, Li S, Zhang N, *et al.* 2012. Application of bio-organic fertilizer can control *Fusarium* wilt of cucumber plants by regulating microbial community of rhizosphere soil. *Biol. Fertil Soils* **48:** 807-816.
11. Weng J, Wang Y, Li J, Shen Q, Zhang R. 2013. Enhanced root colonization and biocontrol activity of *Bacillus amyloliquefaciens* SQR9 by *abrB* gene disruption. *Appl. Microbiol. Biotechnol*. **97:** 8823-8830.
12. Xu Z, Shao J, Li B, Yan X, Shen Q, Zhang R. 2013. Contribution of bacillomycin D in *Bacillus amyloliquefaciens* SQR9 to antifungal activity and biofilm formation. *Appl. Environ. Microbiol*. **79:** 808-815.
13. Wang LT, Lee FL, Tai CJ, Kuo HP. 2008. *Bacillus velezensis* is a later heterotypic synonym of *Bacillus amyloliquefaciens*. *Int. J. Syst. Evol. Microbiol*. **58:** 671-675.
14. Dunlap CA, Kim SJ, Kwon SW, Rooney AP. 2016. *Bacillus velezensis* is not a later heterotypic synonym of *Bacillus amyloliquefaciens*; *Bacillus methylotrophicus*, *Bacillus amyloliquefaciens* subsp. plantarum and '*Bacillusoryzicola*' are later heterotypic synonyms of *Bacillus velezensis* based on phylogenomics. *Int. J. Syst. Evol. Microbiol*. **66:** 1212-1217.
15. Fan B, Blom J, Klenk HP, Borriss R. 2017. *Bacillus amyloliquefaciens*, *Bacillus velezensis*, and *Bacillus siamensis* from an "Operational Group *B. amyloliquefaciens*" within the *B. subtilis* species complex. *Front Microbiol*. **8:** 22.
16. He P, Hao K, Blom J, Rüchert C, Vater J, Mao Z, *et al*. 2013. Genome sequence of the plant growth promoting strain *Bacillus amyloliquefaciens* subsp. *plantarum* B9601-Y2 and expression of mersacidin and other secondary metabolites. *J. Biotechnol*. **164:** 281-291.

17. Rückert C, Blom J, Chen X, Reva O, Borriss R. 2011. Genome sequence of *B. amyloliquefaciens* type strain DSM7^T reveals differences to plant-associated *B. amyloliquefaciens* FZB42. *J. Biotechnol*. **155:** 78-85.
18. Borriss R, Chen XH, Rueckert C, Blom J, Becker A, Baumgarth B, *et al*. 2011. Relationship of *Bacillus amyloliquefaciens* clades associated with strains DSM 7^T and FZB42^T: a proposal for *Bacillus amyloliquefaciens* subsp. plantarum subsp. nov. based on complete genome sequence comparisons. *Int. J. Syst. Evol. Microbiol*. **61:** 1786-1801.
19. Chowdhury SP, Hartmann A, Gao X, Borriss R. 2015. Biocontrol mechanism by root-associated *Bacillus amyloliquefaciens* FZB42 - a review. *Front Microbiol*. **6:** 780.
20. Wu J, Xu G, Jin Y, Sun C, Zhou L, Lin G, *et al*. 2018. Isolation and characterization of *Bacillus* sp. GFP-2, a novel Bacillus strain with antimicrobial activities, from Whitespotted bamboo shark intestine. *AMB Express*. **8:** 84.
21. Chun BH, Kim KH, Jeong SE, Jeon CO. 2018. Genomic and metabolic features of the *Bacillus amyloliquefaciens* group - *B. amyloliquefaciens*, *B. velezensis*, and *B. siamensis* – revealed by pan-genome analysis. *Food Microbiol*. **77:** 146-157.
22. Kim Y, Koh I, Lim MY, Chung W-H, Rho M. 2017. Pan-genome analysis of Bacillus for microbiome profiling. *Sci. Rep*. **7:** 10984.
23. Yi H, Chun J, Cha C-J. 2014. Genomic insights into the taxonomic status of the three subspecies of *Bacillus subtilis*. *Syst. Appl. Microbiol*. **37:** 95-99.
24. Tritt A, Eisen JA, Facciotti MT, Darling AE. 2012. An integrated pipeline for de novo assembly of microbial genomes. *PLoS One* **7:** 9.
25. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, *et al*. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol*. **19:** 455-477.
26. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, *et al*. 2016. Phased diploid genome assembly with single molecule real-time sequencing. *Nat. Methods*. **13:** 1050-1054.
27. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. **27:** 722-736.
28. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL. 1999. Alignment of whole genomes. *Nucleic Acids Res*. **27:** 2369-2376.
29. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, *et al*. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9:** e112963.
30. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, *et al.* 2016. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res*. **44:** 6614-6624.
31. Meier-Kolthoff JP, Auch AF, Klenk HP, Göker M. 2013. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* **14:** 60.
32. Auch AF, von Jan M, Klenk HP, Göker M. 2010. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand Genomic Sci*. **2:** 117–134.
33. Henz SR, Huson DH, Auch AF, Nieselt-Struwe K, Schuster SC. 2005. Whole-genome prokaryotic phylogeny. *Bioinformatics* **21:** 2329-2335.
34. Auch, AF, Henz SR, Holland BR, Göker M. 2006. Genome BLAST distance phylogenies inferred from whole plastid and whole mitochondrion genome sequences. *BMC Bioinformatics* **7:** 350.
35. Zuo G, Hao B. 2015. CVTree3 web server for whole-genome-based and alignment-free prokaryotic phylogeny and taxonomy. *Genomics Proteomics Bioinformatics* **13:** 321-331.
36. Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol*. **14:** 2611-2620.
37. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25:** 1754-1760.
38. Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv:* 1207.3907 [q-bio.GN].
39. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, *et al*. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Bioinformatics* **43:** e15.
40. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30:** 2068-2069.
41. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, *et al*. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31:** 3691-3693.
42. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, *et al*. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4:** 41.
43. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, *et al.* 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* **10:** 421.
44. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, *et al*. 2017. AntiSMASH 4.0 - improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res*. **45:** W36-W41.
45. Guy L, Kultima JR, Andersson SGE, Quackenbush J. 2011. GenoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **27:** 2334-2335.
46. Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. **30:** 3059-3066.

47. Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol*. **17:** 540-552.
48. Minh BQ, Nguyen MA, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol*. **30:** 1188-1195.
49. Csűös M. 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*. **26:** 1910-1912.
50. Alikhan NF, Petty NK, Zakour NLB, Beatson SA. 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics*. **12:** 402.
51. Namouchi A, Didelot X, Schöck U, Gicquel B, Rocha EP. 2012. After the bottleneck: Genome-wide diversification of the Mycobacterium tuberculosis complex by mutation, recombination, and natural selection. *Genome Res*. **22:** 721-34.
52. Bart R, Cohn M, Kassen A, McCallum EJ, Shybut M, Petriello A, *et al*. 2012. High-throughput genomic sequencing of cassava bacterial blight strains identifies conserved effectors to target for durable resistance. *Proc. Natl. Acad. Sci. USA* **109:** E1972-1979.
53. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30:** 1312-1313.
54. Bechinger B, Gorr SU. 2017. Antimicrobial peptides: mechanisms of action and resistance. *J. Dent. Res*. **96:** 254-260.