scientific data

OPEN

DATA DESCRIPTOR

Whole genome sequences of 135 "*Candidatus* Liberibacter asiaticus" strains from China

Yongqin Zheng[1,2], Jiaming Li[1,2], Mingxin Zheng[1,2], You Li[3], Xiaoling Deng[1,2] ✉ & Zheng Zheng[1,2] ✉

"*Candidatus* Liberibacter asiaticus" (CLas) is a phloem-limited alpha-proteobacteria causing Citrus Huanglongbing, the destructive disease currently threatening global citrus industry. Genomic analyses of CLas provide insights into its evolution and biology. Here, we sequenced and assembled whole genomes of 135 CLas strains originally from 20 citrus cultivars collected at ten citrus-growing provinces in China. The resulting dataset comprised 135 CLas genomes ranging from 1,221,309 bp to 1,308,521 bp, with an average coverage of 675X. Prophage typing showed that 44 strains contained Type 1 prophage, 89 strains contained Type 2 prophage, 44 strains contained Type 3 prophage, and 34 of them contained more than one type of prophage/phage. The SNP calling identified a total of 5,090 SNPs. Genome-based phylogenetic analysis revealed two major clades among CLas strains, with Clade I dominated by CLas strains containing Type 1 prophage (79/95) and Clade II dominated by CLas strains containing Type 1 or Type 3 prophage (80/95). This CLas genome dataset provides valuable resources for studying genetic diversity and evolutionary pattern of CLas strains.

Background & Summary

Citrus Huanglongbing (HLB, also called yellow shoot disease) is a destructive disease threatening citrus production worldwide[1]. HLB is caused by the unculturable phloem-limited α-proteobacterium "*Candidatus* Liberibacter spp.", mainly including "*Ca*. L. asiaticus" (CLas), "*Ca*. L. africanus" and "*Ca*. L. americanus"[1,2]. Among the three species, CLas was the most widespread species and responsible for the increasing economic losses of citrus industries in Asia and America[3]. The characteristic symptoms of citrus trees infected by CLas mainly included yellow shoots, yellowing/mottling leaves, small and malformed fruit with aborted seeds, fruit abscission, rotted roots and ultimately tree death[4]. HLB severely affected the longevity and fruit yield of citrus plants and posed a significant risk challenge for disease management due to the presence of the vector Asian citrus psyllid (ACP, *Diaphorina citri*)[5]. Literature records indicated that HLB was first observed in Chaoshan region of Guangdong province, China around 1860s and became a local epidemic around 1930s[2,6]. The spread of HLB to the major citrus producing areas in southern China was observed after 1930s[2,6]. As of now, HLB have been found in 11 out of 19 citrus growing provinces in China and widely distributed in more than 50 citrus producing countries in Asia, Africa, America, which severely limits the development of global citrus industry[3,7].

Despite CLas cannot be cultured *in vitro*, the advancement in whole genome sequencing has greatly facilitated the CLas research, mainly including genetic diversity, evolution, gene function analysis, pathogenicity and biology of CLas[8–11]. One major break-through from CLas genome sequence analyses was the discovery of CLas phages/prophages, which was further used for CLas strain characterization and biological investigations[12–14]. Currently, three large prophages, designated SC1 (Type 1), SC2 (Type 2), and P-JXGC-3 (Type 3), were identified in CLas strains[12,13], providing valuable insights into CLas biology and genetic diversity[11,15,16]. Additionally, whole genome sequence resource of CLas have also been employed for evolutionary analysis among Liberibacter species and genetic diversity among CLas strains. Comparative genomes of Liberibacter species showed the evolutionarily separation of CLas from the non-pathogenic *Liberibacter crescens*[10,17]. Genome-based analysis revealed the genetic variations among CLas strains from different geographical locations and offered insights

[1]State Key Laboratory of Green Pesticide, South China Agricultural University, Guangzhou, Guangdong, China. [2]Guangdong Province Key Laboratory of Microbial Signals and Disease Control, South China Agricultural University, Guangzhou, China. [3]Vector-borne Virus Research Center, Fujian Province Key Laboratory of Plant Virology, Fujian Agriculture and Forestry University, Fujian, China. ✉e-mail: xldeng@scau.edu.cn; zzheng@scau.edu.cn

into the possible source of CLas introduction and HLB epidemiology in the United States[10]. A recent study of genome comparison based on 35 published CLas genomes identified over 6,000 minor variations and the highly heterogeneous variations distribution across CLas genome, including four highly diverse non-prophages regions and three prophage regions[18].

As the most widely distributed species within Liberibacter genus, the current available CLas genomes resources are very limited, mainly due to the inability of *in vitro* culture. Thus, the total DNAs extracted from CLas-infected citrus plants or insect vectors became the only DNA resources for CLas genome sequencing[8,19]. However, the high ratio of host DNA as compared to CLas DNA in total DNA resulted in a very low efficiency of CLas whole genome sequencing, thereby increasing the challenge of obtaining high-quality CLas genome sequence. Efforts have been focused on getting sufficient number of CLas reads to enhance the quality of CLas genome assembly and led to two main effective strategies, including the use of host tissue with high CLas titer and increase of sequencing depth[8,19–21]. Our previous study found that the citrus fruit pith can be used as an ideal host tissue source for CLas genome sequencing, due to its ability to support the multiplication of CLas to a high level[21,22]. In addition, two non-natural host plants of CLas, the periwinkle (*Catharanthus roseus*) and dodder (*Cuscuta campestris*), were proved to be the more amenable hosts for CLas proliferation, which can serve as the surrogate host source to gain the DNA sample with high ratio of CLas DNA[20,21]. These efforts provided the suitable DNA sources for CLas whole genome sequencing, which makes it feasible to obtain the high-quality genome sequence of CLas strains from different geographical locations.

The establishment of a comprehensive CLas genome database is anticipated to greatly advance the CLas research, particularly in CLas genetic diversity, evolution, epidemiology and biology. Currently, a total of 46 CLas genomes were released, with only 13 were in complete level (https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=34021, accessed June 2024). However, there were only seven CLas genomes originally from China, among which were limited to few HLB-epidemic areas in China[11,13,19,21,23]. As the most prevalent and destructive Liberibacter species throughout HLB-affected citrus-growing areas worldwide, the limited genome resource of CLas hindered the understanding of evolution relationships of CLas population across varied geographical regions in the HLB-epidemic countries around world, especially in the historical HLB-epidemic country, such as China.

This study presents the whole genome sequencing data from 135 CLas strains originally collected from 20 commercial citrus cultivars distributed in ten HLB-endemic provinces in China. The average nucleotide coverage of 135 CLas genomes was 675X, indicating the high-quality genome sequencing and assembly. A total of 5,090 SNPs were identified among 148 CLas genomes, including 135 sequenced in this study and 13 complete CLas genomes available in NCBI database. These SNPs included 4,247 SNPs in chromosome, 383 SNPs in Type 1 prophage region, 323 SNPs in Type 2 prophage region, and 137 SNPs in Type 3 prophage region. Our CLas genome sequence dataset will not only serve as a valuable resource for further research in evolutionary pattern and genetic diversity of CLas strains from China and others worldwide HLB-endemic countries, but also facilitate the research in CLas pathogenicity and biology.

Methods

Sample collection. Over 1000 CLas samples were collected from ten HLB-epidemic provinces in China, including Fujian, Guangdong, Guangxi, Guizhou, Hainan, Hunan, Jiangxi, Sichuan, Yunnan and Zhejiang (Fig. 1). To improve the quality of CLas genome sequencing and assembly, a total of 135 representative CLas samples with high CLas titer (Ct value $<$ 25 by primer set CLas4G/HLBr) were further selected for genome sequencing (Fig. 1; Supplementary Table S1). These samples were originally collected from 20 commercial citrus cultivars (Supplementary Table S1). All samples were collected during 2017 to 2022. The leaf midribs from leaves showing HLB typical symptoms (mottled or yellowing) or fruit pith tissue from HLB-symptomatic fruit ("red-nose" fruit) were sampled and stored at −20 °C for before DNA extraction.

DNA extraction and validation. A total of 100 mg leaf midribs or 50 mg fruit pith tissue were chopped in to 2-mm section with sterile blades. Total plant DNA was extracted using the E.Z.N.A. high-performance plant DNA kit (Omega Bio-Tek, Doraville, GA, USA) according to the manufacturer's instructions. The quality of the extracted DNA was verified by 1% agarose gel electrophoresis. The concentration and purity were measured using the NanoDrop One microvolume UV-Vis spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) at a wavelength of A260/A280.

Quantification of CLas. Quantification of CLas was performed by SYBR Green Real-time PCR with the primer sets CLas4G/HLBr as described in a previous study[24]. The 20 μL of PCR reaction mixture contained 1 μL of DNA template (~25 ng), 0.5 μL of each forward and reverse primer (10 μM), 10 μL of iQ™ SYBR Green Supermix (Bio-Rad, Hercules, CA, USA) and 8 μL of ddH_2O. All PCR was conducted in CFX Connect Real-Time System (Bio-Rad, Hercules, CA, USA) under the following procedure: 95 °C for 3 min, followed by 40 cycles at 95 °C for 10 s and 60 °C for 30 s, with fluorescence signal capture at the end of each 60 °C step. The data (cycle threshold value, Ct value) were generated and analyzed using Bio-Rad CFX Manager 2.1 software with automated baseline settings and a manually set threshold at 0.1. Only CLas samples with Ct value $<$ 25 were further selected as candidate for whole genome sequencing (Supplementary Table S1).

High-throughput sequencing. Library preparation for each CLas sample was performed with the NEB Next Ultra DNA Library Prep Kit (Illumina, San Diego, CA, USA). Genome sequencing was performed on the Illumina HiSeq 3000 platform (Illumina, San Diego, CA, USA) with 150-bp paired-end reads by a commercial sequencing company. The raw data files that obtained from high-throughput sequencing were converted to raw sequences (fastq format) by Illumina CASAVA Base Calling v.1.8.2.

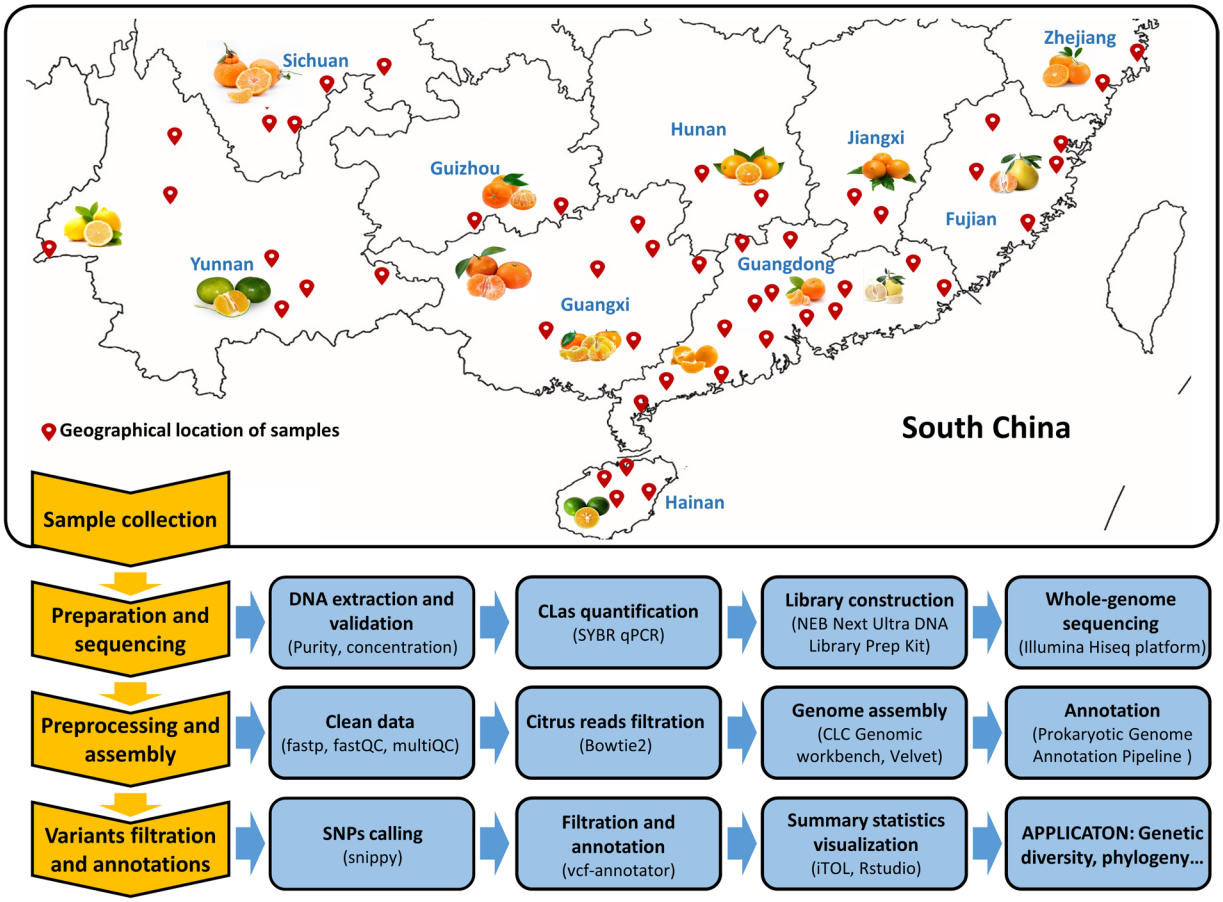

Fig. 1 Overview of sample collection, data processing and bioinformatics analysis pipeline. The name of province was in blue fonts in the map and the red points represent the sampling sources. The fruit photo indicated the main citrus cultivars resources for CLas samples collected from each province.

Data pre-processing. Raw data obtained by sequencing were filtered by removing adapter reads, reads with N (N indicates that base information could not be determined) greater than 10%, and low-quality reads (Qphred $\leq$ 20 bases accounting for more than 50% of the entire length of the reads) using fastp v.0.19.4 with default parameters to generate the clean reads[25]. The quality control of clean data was performed by fastQC v.0.11.5 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and multiQC v.1.14[26] using default parameters. All clean reads were mapped to citrus genomes (*Citrus maxima* genome: MKYQ00000001.1, *Citrus reticulata* genome: NIHA00000000.1, *Citrus sinensis* genome: AJPS00000000.1, *Citrus sinensis* mitochondrion: NC_037463.1 and *Citrus reticulata* chloroplast: KU170678.1) to filter out the citrus reads using Bowtie v.2.4.2[27]. The unmapped reads were retained for CLas genome assembly.

Genome assembly. CLas genome sequence was generated by the combination with reference-based assembly and *de novo* assembly. For reference-based assembly, three CLas genomes that contained different types of prophages, including YNBC (CP118771, contained Type 1 prophage), A4 (CP010804, contained Type 2 prophage) and JXGC (CP019958, contained Type 3 prophage), were used as reference for CLas genome assembly. The identification of prophage type for each CLas strain was based on the reads mapping to three prophage sequences, including Type 1 prophage (P-YNBC-1, ranged from 1,187,948 bp to 1,230,892 bp of strain YNBC), Type 2 prophage (P-A4-2, ranged from 1,189,877 bp to 1,603 bp of strain A4) and Type 3 prophage (P-JXGC-3, ranged from 1,192,430 bp to 1,582 bp of strain JXGC). The retained reads were mapped to the reference genomes using CLC Genomics Workbench v.20.0 with default parameters. The *de novo* assembly of each retained data was performed using Velvet v.1.2.10 by setting the minimum contig length as 1,000 bp[28]. The CLas *de novo* contigs were identified and extracted through BLAST search using CLas strain A4, YNBC and JXGC genome as queries. Gap closures of consensus sequences generated from reference-based assembly were performed using: (1) *de novo* assembly contigs that connected CLas reference-mapping contigs via BLAST + v.2.10.0 (E-value < 1e-5)[29]; (2) reads walking following the previously published method[30]. These efforts generated the high-quality CLas whole genome sequence, including the chromosomal and prophage region. All genomes were submitted to the NCBI genome database and annotated using the Prokaryotic Genome Annotation Pipeline (PGAP) v.6.5[31].

SNP calling and annotation. A total of 13 available CLas complete genomes, including A4 (CP010804), CoFLP (CP054558), GDCZ (CP118922), gxpsy (CP004005), Ishi-1 (AP014595), JRPAMB1 (CP040636), JXGC

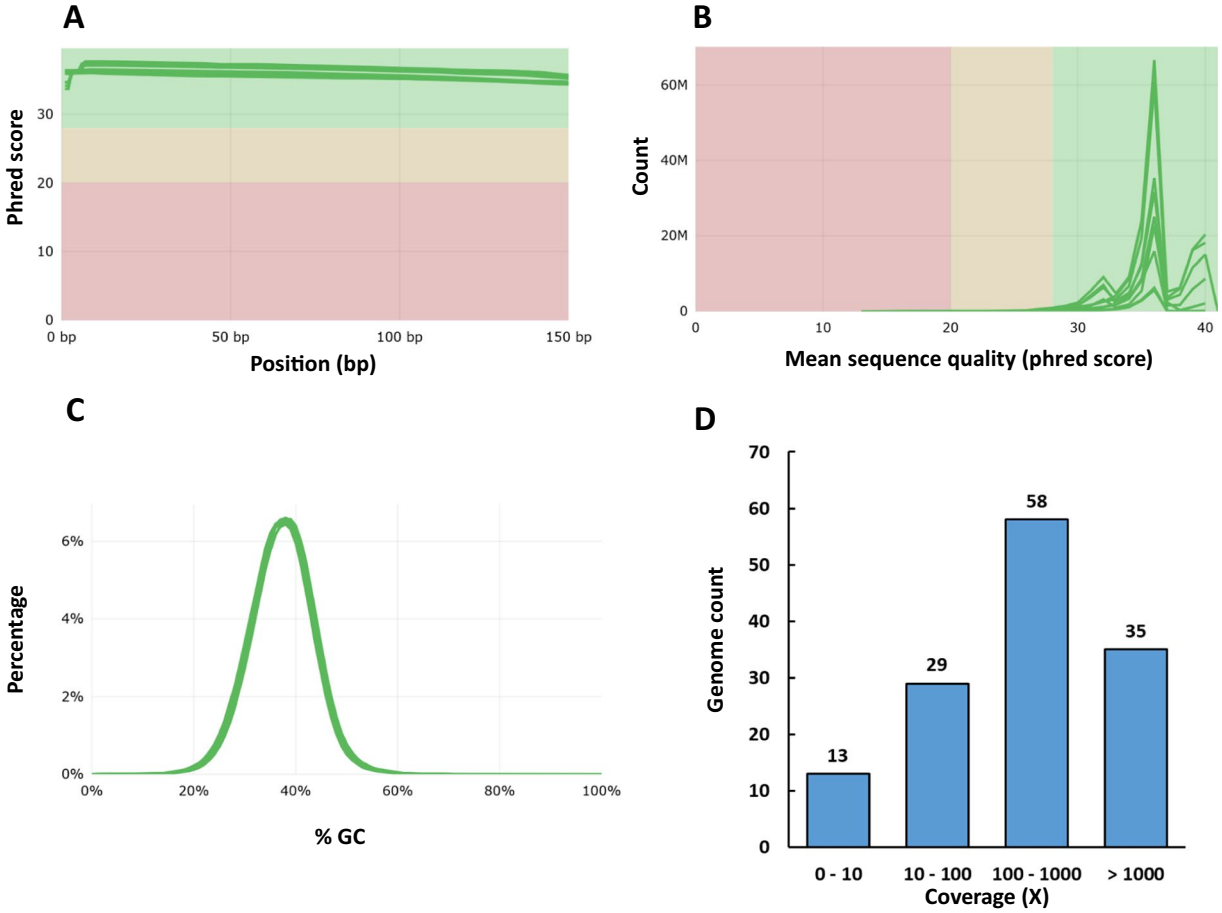

Fig. 2 Quality assessment of sequencing data and genome assembly. (**A**) Mean quality scores across each base position. (**B**) Mean quality scores per read. (**C**) GC content of reads. (**D**) Statistics of coverage (X) of genome assembly.

(CP019958), PGD (CP100754), psy62 (CP001677), PYN (CP100417), ReuSP1 (CP061535), TaiYZ2 (CP041385) and YNBC (CP118771), were downloaded from the NCBI genome database and combined with 135 CLas genome sequenced in this study as the CLas genome dataset for SNPs calling analyses. The SNPs of all above CLas genomes were identified with "snippy-multi" program in snippy v.4.6 (https://github.com/tseemann/snippy) using the A4 chromosomal sequence (ranging from 1,604 bp to 1,189,876 bp of A4 genome) and three prophage sequences (P-YNBC-1, P-A4-2 and P-JXGC-3) as the reference. The information of each SNP was extracted, including SNP base positions and alleles (.vcf files). The SNPs were annotated using vcf-annotator (https://github.com/rpetit3/vcf-annotator), including intragenic (synonymous or non-synonymous) and intergenic mutations.

Phylogeny analysis. The phylogeny (Neighbour Joining tree) of all CLas strains (including 135 sequenced in this study and 13 previously published) was constructed based on the aligned SNPs with CLC Genomics Workbench v.20.0 under the Jukes-Cantor (JC) model. The phylogenetic tree was visualized in iTOL v.6.8.1[32]. The p-distance of SNP mutation profiles was calculated using VCF2Dis v.1.50 (https://github.com/BGI-shenzhen/VCF2Dis).

Data Records

The clean sequencing data (fastq format) were deposited in NCBI Sequence Read Archive under accession number PRJNA1123441 (https://identifiers.org/ncbi/insdc.sra:SRP513632)[33]. The genome assembly (GenBank format) were deposited in NCBI's Genome Database with the accession number for Project PRJNA996237 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA996237)[34] (Supplementary Table S1). The VCF files are available on figshare[35].

Technical Validation

Quality control of sequencing data. The quality of all sequencing data was assessed by fastQC and multiQC program to investigate the mean quality score per position and the GC content. Illumina HiSeq platform yielded a total of 1,533 gigabases clean data for 135 samples, with an average of 11.4 gigabases per sample. The multiQC reports for 150 paired-end reads showed the quality value across each base position of all reads were higher than the Phred quality score of 30 (Fig. 2A) and the average quality scores of >94% of total reads was greater than the phred score of 20 (Fig. 2B, Supplementary Table S1). These statistics confirmed that the mean

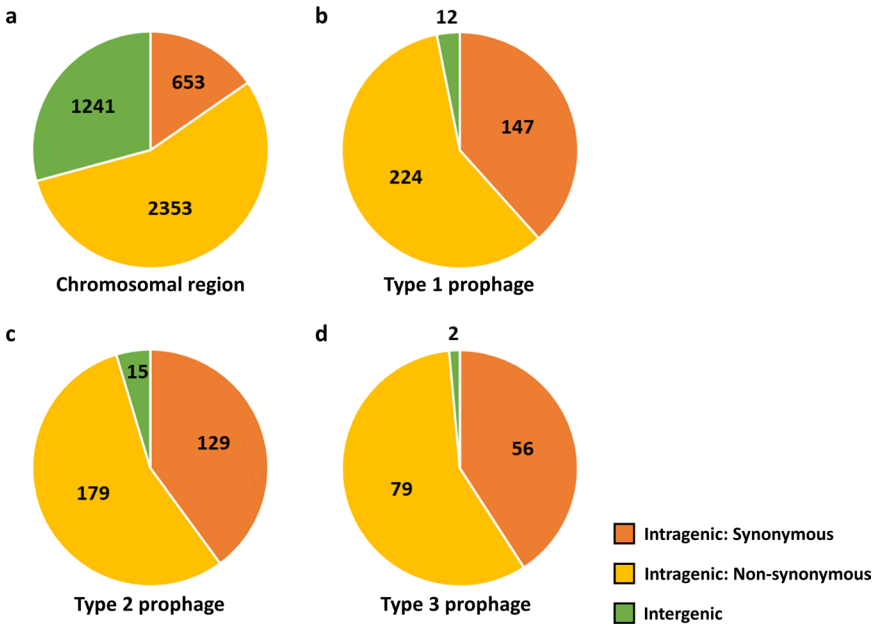

Fig. 3 Statistics of SNP annotation in chromosomal region (**a**), Type 1 prophage (**b**), Type 2 prophage (**c**) and Type 3 prophage (**d**).

quality scores and per-sequence metrics fell within the high sequence standard range for subsequent analyses. The GC content of all reads from all samples showed a stable distribution (average of 36.5%) (Fig. 2C), indicating no possible contamination during the sequencing process.

Quality evaluation of CLas genome assembly. Reads filtering with citrus host genome combined with reference-based mapping generated a total of 750,493,499 CLas reads from 135 CLas samples, with an average of 5,559,211 reads per sample. The full length of the genome assembly for 135 CLas strains showed high integrity, with the average length of 1,241,536 bp, ranging from 1,221,309 bp to 1,308,521 bp (Supplementary Table S1). The average coverage of 135 CLas genomes was about 675X. A total of 122 CLas genomes (90.4%) showed over 10X coverage. In particular, 93 CLas genomes (68.9%) showed over 100X coverage (Fig. 2D). The number of genes could also be employed as the standard for evaluating the quality of the whole genome assembly. A close number of genes were annotated in the whole genome among 135 CLas strains, ranging from 1,104 genes to 1,211 genes, indicating a high-quality of 135 CLas genomes (Supplementary Table S1). Prophage typing of 135 CLas genomes found that 44 CLas strains contained Type 1 prophage, 89 CLas strains contained Type 2 prophage, 44 CLas strains contained Type 3 prophage. It was noted that 28 CLas strains contained two types of prophages/phages and six of them contained three types of prophages/phages (Supplementary Table S1). These quality metrics indicate the completeness and contiguity of CLas genome assemblies, which enhanced the resolution and reliability for downstream analyses.

Quality control of SNP data. The high-resolution SNPs dataset for CLas genomes was generated by "snippy-multi" program in snippy v.4.6. A total of 5,090 high-quality SNPs were identified among 148 CLas genomes, included 135 sequenced in this study and 13 complete CLas genomes downloaded from NCBI database. The density of SNPs across the CLas genome was shown in Fig. 3. Specifically, a total of 4,247 SNPs were retained in chromosome, 383 SNPs were in Type 1 prophage, 323 SNPs were in Type 2 prophage, and 137 SNPs were in Type 3 prophage. Based on variants annotation, SNPs were distributed in intragenic (synonymous or non-synonymous) and intergenic regions (Fig. 3). Overall, most SNPs were located in the intragenic region of both CLas chromosomal region and three prophage regions. Among SNPs identified in intragenic region, the non-synonymous SNPs, which caused the change in the amino acids, were accounted for more than half number of total SNPs (Fig. 3). In chromosomal region, a higher number of intergenic SNPs were observed than those identified in three prophages (Fig. 3).

Phylogeny analysis of CLas genomes. The SNP mutant profiles of all strains were compared and clustered into a phylogenetic tree (Fig. 4). The 148 CLas strains showed high similarity in chromosome with the pairwise evolutionary distance less than 0.1 (Fig. 4a). Two main phylogenetic clades (Clade I and Clade II) were identified among CLas strains with a close intrinsic number within clade, i.e. 55% (82/148) in Clade I and 45% (66/148) in Clade II (Fig. 4a). The majority of CLas strains in Clade I carried Type 2 prophages (79 out of 95), while those in Clade II predominantly carried Type 1 (41 out of 50) or/and Type 3 (39 out of 45) prophages (Fig. 4b). According to the geographical origin of CLas strains, Clade I was dominant in CLas strains from Guangdong, Guangxi, and Guizhou, while Clade II was dominant in CLas strains from Hainan, Jiangxi, Sichuan, and Yunnan (Fig. 4c). Early studies had suggested the difference in CLas population structure could be associated with the bacterial environment adaptation

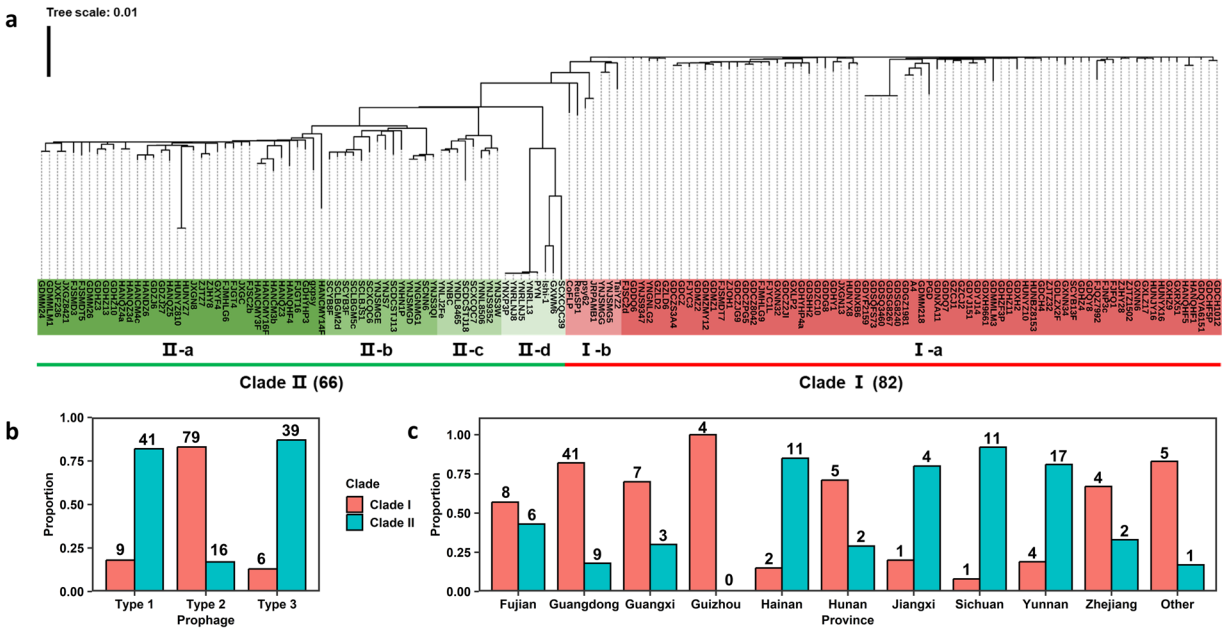

Fig. 4 Phylogeny of "*Candidatus* Liberibacter asiaticus" (CLas) strains. (**a**) The Neighbor-Joining (NJ) tree based on genomic SNPs. (**b**) The proportion of prophage types. (**c**) The proportion of CLas strains from different provincial sources.

and activity of phage[11,12,36]. Based on the comparison result of CLas genomes dataset, it was therefore interesting to hypothesize that the population differentiation of CLas could be mainly related to the CLas-phage interaction and various environment conditions from different geographical locations. However, the regional transport of CLas-infected seedlings could also lead to the genetic mixing and differentiation of CLas populations[37], reflecting the possible HLB spread and epidemiology. The phylogeny analysis suggested that the CLas whole genome data would be reliable for further genomic research and provide information for the CLas/HLB epidemiology and control.

Code availability

All the software programs used in this study (data processing and analysis) are listed with the version in the Methods section. In case of no details on parameters, the programs were used with the default settings. No custom code was generated or used for analysis of the data presented.

Received: 2 July 2024; Accepted: 3 September 2024;
Published online: 19 September 2024

References

1. Bové, J. M. Huanglongbing: a destructive, newly-emerging, century-old disease of citrus. *J. Plant Pathol.* **88**(1), 7–37 (2006).
2. Zheng, Z., Chen, J. & Deng, X. Historical perspectives, management, and current research of citrus HLB in Guangdong province of China, where the disease has been endemic for over a hundred years. *Phytopathology.* **108**(11), 1224–1236 (2018).
3. Bové, J. M. Heat-tolerant Asian HLB meets heat-sensitive African HLB in the Arabian Peninsula! Why? *J. Citrus Pathol.* **1**, 1–78 (2014).
4. Gottwald, T. R. Current epidemiological understanding of Citrus Huanglongbing. *Annu. Rev. Phytopathol.* **48**, 119–139 (2010).
5. Hall, D. G. *et al.* Asian citrus psyllid, *Diaphorina citri*, vector of Citrus Huanglongbing disease. *Entomologia Experimentalis et Applicata.* **146**, 207–223 (2013).
6. Lin, K. H. Observations on yellow shoot of citrus. *Acta Phytopathol. Sin.* **2**, 1–11 (1956).
7. Wang, N. The Citrus Huanglongbing crisis and potential solutions. *Mol. Plant.* **12**(5), 607–609 (2019).
8. Duan, Y. *et al.* Complete genome sequence of citrus Huanglongbing bacterium, '*Candidatus* Liberibacter asiaticus' obtained through metagenomics. *Mol. Plant Microbe In.* **22**(8), 1011–1020 (2009).
9. Clark, K. *et al.* An effector from the Huanglongbing-associated pathogen targets citrus proteases. *Nat. Commun.* **9**(1), 1718 (2018).
10. Thapa, S. P. *et al.* Genome-wide analyses of Liberibacter species provides insights into evolution, phylogenetic relationships, and virulence factors. *Mol. Plant Pathol.* **21**(5), 716–731 (2020).
11. Zheng, Y. *et al.* Pathogenicity and transcriptomic analyses of two "*Candidatus* Liberibacter asiaticus" strains harboring different types of phages. *Microbiol. Spectr.* **11**(3), e00754–23 (2023).
12. Zhang, S. *et al.* '*Ca*. Liberibacter asiaticus' carries an excision plasmid prophage and a chromosomally integrated prophage that becomes lytic in plant infections. *Mol. Plant Microbe In.* **24**(4), 458–468 (2011).
13. Zheng, Z. *et al.* A Type 3 prophage of '*Candidatus* Liberibacter asiaticus' carrying a restriction-modification system. *Phytopathology.* **108**(4), 454–461 (2018).
14. Zhang, L. *et al.* A novel Microviridae phage (CLasMV1) from "*Candidatus* Liberibacter asiaticus". *Front. Microbiol.* **12**, 754245 (2021).
15. Dominguez-Mirazo, M., Jin, R. & Weitz, J. S. Functional and comparative genomic analysis of integrated prophage-like sequences in "*Candidatus* Liberibacter asiaticus". *mSphere.* **4**(6), e00409–00419 (2019).
16. Zheng, Y. *et al.* Prophage region and short tandem repeats of "*Candidatus* Liberibacter asiaticus" reveal significant population structure in China. *Plant Pathol.* **70**(4), 959–969 (2021).
17. Batarseh, T. N. *et al.* Comparative genomics of the Liberibacter genus reveals widespread diversity in genomic content and positive selection history. *Front. Microbiol.* **14**, 1206094 (2023).

18. Gao, F. *et al*. Genetic diversity of "*Candidatus* Liberibacter asiaticus" based on four hypervariable genomic regions in China. *Microbiol. Spectr.* **10**(6), e02622–22 (2022).
19. Zheng, Z., Deng, X. & Chen, J. Whole-genome sequence of "*Candidatus* Liberibacter asiaticus" from Guangdong, China. *Genome Announc.* **2**(2), e00273–14 (2014).
20. Li, T. *et al*. Establishment of a *Cuscuta campestris* - mediated enrichment system for genomic and transcriptomic analyses of '*Candidatus* Liberibacter asiaticus'. *Microb. Biotechnol.* **14**(2), 737–751 (2021).
21. Zheng, Y. *et al*. Genome sequence resource for "*Candidatus* Liberibacter asiaticus" strain GDCZ from a historical HLB endemic region in China. *BMC Genomic Data.* **24**(1), 63 (2023).
22. Fang, F. *et al*. A significantly high abundance of "*Candidatus* Liberibacter asiaticus" in citrus fruit pith: in planta transcriptome and anatomical analyses. *Front. Microbiol.* **12**, 681251 (2021).
23. Lin, H. *et al*. Complete genome sequence of a Chinese strain of "*Candidatus* Liberibacter asiaticus". *Genome Announc.* **1**(2), e00184–13 (2013).
24. Bao, M. *et al*. Enhancing PCR capacity to detect '*Candidatus* Liberibacter asiaticus' utilizing whole genome sequence information. *Plant Dis.* **104**(2), 527–532 (2020).
25. Chen, S. *et al*. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics.* **34**(17), i884–i890 (2018).
26. Ewels, P. *et al*. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics.* **32**(19), 3047–3048 (2016).
27. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods.* **9**(4), 357–359 (2012).
28. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome res.* **18**(5), 821–829 (2008).
29. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics.* **10**(1), 421 (2009).
30. Shih, H. *et al*. Draft genome sequence of "*Candidatus* Sulcia muelleri" strain KPTW1 from Kolla paulula, a vector of *Xylella fastidiosa* causing Pierce's disease of grapevine in Taiwan. *Microbiol. Resour. Ann.* **8**(2), e01347–18 (2019).
31. Tatusova, T. *et al*. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **44**(14), 6614–6624 (2016).
32. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**(W1), W293–W296 (2021).
33. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP513632 (2024).
34. Zheng, Y. *et al*. Whole genome sequences of 135 "*Candidatus* Liberibacter asiaticus" strains from China. *Genbank* https://identifiers.org/ncbi/bioproject:PRJNA996237 (2024).
35. Zheng, Y. *et al*. Whole genome sequences of 135 "*Candidatus* Liberibacter asiaticus" strains from China. *figshare* https://doi.org/10.6084/m9.figshare.26112721.v1 (2024).
36. Ma, W. *et al*. Population structures of '*Candidatus* Liberibacter asiaticus' in southern China. *Phytopathology.* **104**(2), 158–162 (2014).
37. Liu, R. *et al*. Analysis of a prophage gene frequency revealed population variation of '*Candidatus* Liberibacter asiaticus' from two citrus-growing provinces in China. *Plant Dis.* **95**(4), 431–435 (2011).

Acknowledgements

This work was supported by funding from National Key Research and Development Program of China (2021YFD1400800), National Natural Science Foundation of China (31901844), Guangzhou Basic and Applied Basic Research Foundation (SL2022A04J00758) and China Agriculture Research System of MOF and MARA.

Author contributions

Y.Z., J.L., M.Z., Y.L. and Z.Z. collected samples, generated genome data, performed the bioinformatic analysis. Y.Z. and Z.Z. wrote and revised the manuscript. X.D. and Z.Z. conceived the experiments, supervised and financed the study. All authors reviewed the manuscript and made critical contributions to the manuscript drafts.

Competing interests

The author declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-03855-3.

Correspondence and requests for materials should be addressed to X.D. or Z.Z.