Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Genetic differentiation and pathogenicity of diverse Xanthomonas oryzae pv. oryzae strains revealed by Whole Genome Sequencing

Shanyu Chen
Yunnan University
Yun Zhang
Yunnan Academy of Agricultural Sciences
Jiaxin Xing
Yunnan Academy of Agricultural Sciences
Li Liu
Yunnan Academy of Agricultural Sciences
Tengqiong Yu
Yunnan Academy of Agricultural Sciences
Fuyou Yin
Yunnan Academy of Agricultural Sciences
Qiaofang Zhong
Yunnan Academy of Agricultural Sciences
Jinlu Li
Yunnan Academy of Agricultural Sciences
Dunyu Zhang
Yunnan Academy of Agricultural Sciences
Suqin Xiao
Yunnan Academy of Agricultural Sciences
Bo Wang
Yunnan Academy of Agricultural Sciences
Cong Jiang
Yunnan Academy of Agricultural Sciences
Zaiquan Cheng
Yunnan Academy of Agricultural Sciences
Ling Chen
cl@yaas.org.cn

Yunnan Academy of Agricultural Sciences

Article

Keywords: bacterial blight resistance genes, pathogenic system, rice diseases, TAL effectors, Xoo

Posted Date: September 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10457578/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Rice bacterial blight, caused by *Xanthomonas oryzae* pv. *oryzae* (*Xoo*), to investigate the genomic diversity and evolutionary characteristics of *Xoo*, whole-genome sequencing was performed on ten representative strains collected from different geographic regions, followed by comparative genomic analyses. Average nucleotide identity (ANI) analysis confirmed that all ten isolates belonged to *Xoo* and supported their phylogenetic classification. Phylogenomic analysis classified the strains into three distinct lineages, which were highly consistent with gene family composition, CRISPR distribution, and TAL effector (TALE) repertoires. Comparative genomic analyses revealed that the core genome, virulence genes, and antimicrobial resistance genes were highly conserved among the strains, whereas plasmid distribution, gene family composition, and TALE repertoires exhibited substantial variation. Thirty-three nonredundant core TALEs were identified, displaying extensive diversity in repeat-variable diresidue (RVD) composition, repeat number, and chromosomal organization. TALE copy numbers ranged from 12 to 18 among the genomes and showed lineage-specific distribution patterns that closely reflected phylogenetic relationships. These findings indicate that the diversification of TALE repertoires, rather than variation in conserved pathogenic systems, is a major contributor to genomic differentiation and host adaptation in *Xoo* populations. This study provides valuable genomic resources for understanding the evolution of *Xoo* pathogenicity and offers a foundation for the discovery of novel resistance genes and the development of durable strategies for bacterial blight resistance breeding

Introduction

Rice (*Oryza sativa* L.) is a vital staple food crop, serving as the primary energy source for nearly half of the global population[1]. Bacterial leaf blight (BB), caused by the Gram-negative bacterium *Xanthomonas oryzae* pv. *oryzae* (*Xoo*), is among the three most significant rice diseases. It represents the most widespread and destructive bacterial disease affecting rice, often resulting in substantial yield losses and, in severe cases, complete crop failure[2]. Among available control strategies, the identification and deployment of resistance (R) genes are widely recognized as the most cost-effective and sustainable methods for disease management[3]. *Xanthomonas oryzae* produces a range of effector proteins with virulence, avirulence, or dual functions[4]. Notably, transcription activator-like effectors (TALEs) are structurally and functionally unique proteins that are delivered into host cells via the Type III secretion system (T3SS)[5]. Once inside plant cells, TALEs bind to effector binding elements (EBEs) located in the promoters or coding regions of host genes through their specialized structural domains. The specificity of EBE recognition is determined by conserved TALE features, which include: (1) a highly conserved N-terminal region essential for Type III secretion; (2) a central repeat region composed of tandem repeats, each primarily consisting of 34 amino acids. These repeats are polymorphic at positions 12 and 13, known as repeat-variable diresidues (RVDs), which display some degeneracy and dictate EBE specificity; (3) two to three C-terminal nuclear localization signals (NLSs); (4) an acidic transcriptional activation domain (AD)[6–8]. This interaction activates host target genes at the transcriptional level, resulting in either susceptibility or resistance[9,10]. The activated genes are classified as host susceptibility (S) genes or resistance (R) genes[11]. TALE repertoires in different *Xoo* strains vary in both composition and number, indicating that TALEs contribute significantly to pathogenic variability among strains.

Currently, 50 bacterial blight resistance genes have been identified in the literature[12]. Of the 17 cloned bacterial blight resistance genes, all except *Xa3/Xa26, Xa4*, and *Xa21* are closely associated with transcription activator-like effector (TALE)-mediated interactions[1]. These resistance (R) genes are categorized according to their encoded products and mechanisms of action as receptor-like kinase (RLK), sugar transporter (SWEET), executor (E), and other gene types[3]. For instance, *Xa1* and its alleles (*Xa1, Xa2, Xa14, Xa31(t)*, and *Xa45(t)*) encode nucleotide-binding leucine-rich repeat (NB-LRR or NLR) proteins that initiate immune responses by recognizing canonical TALEs. However, this resistance can be suppressed by interfering TALEs (iTALEs), which are prevalent in Asian *Xoo* strains[13–16]. The recessive resistance gene *xa5* encodes the γ subunit of the general transcription factor TFIIA. Amino acid variation in this protein alters the interaction between TALEs and the host transcriptional complex, thereby limiting *Xoo* survival in rice[17]. Additionally, recessive resistance genes such as *xa13*, *xa25*, and *xa41* encode SWEET family sugar transporters, while their dominant alleles act as susceptibility (S) genes that are specifically induced by TALEs to promote pathogen infection[18]. In contrast, executor (E) genes such as *Xa7, Xa10, Xa23*, and *Xa27* activate defense responses upon induction by specific TALEs[19–23]. Collectively, these findings demonstrate that the intricate interaction network between TALEs and rice resistance genes is a key determinant of the durability of rice resistance and the success of disease management in agricultural settings.

Advancements in genome sequencing technologies have significantly enhanced research on the pathogenic mechanisms of bacterial blight and the evolution of virulence factors in *Xanthomonas oryzae* pv. *oryzae* (*Xoo*). Several representative *Xoo* strains have been sequenced, including the Korean strain KACC10331 (KXO85), the Japanese strain MAFF311018 (T7174), and the Philippine strain PXO99[A2,24,25]. Despite these developments, comprehensive comparisons of the composition, diversity, and evolutionary characteristics of transcription activator-like effector (TALE) repertoires among *Xoo* strains from diverse geographic regions remain scarce. Previous studies have collected and isolated highly virulent strains from the Philippines, Japan, and various regions of China. In this context, the present study selected ten representative *Xoo* strains for whole-genome sequencing and comparative genomic analysis, with a focus on characterizing TALE effector composition. Phylogenetic analyses were conducted to elucidate evolutionary relationships, thereby providing a theoretical basis for understanding TALE–resistance gene interactions and informing the deployment of bacterial blight–resistant rice varieties.

Results

Statistics of Third-Generation Sequencing (PacBio) Data

Analysis of third-generation sequencing results (Table 1) revealed substantial differences in total read numbers and sequence lengths among the ten pathogenic strains. Ym1 showed significantly higher total read counts and sequence length than the other strains, while PB showed slightly lower values. No

ambiguous bases were detected in any of the 10 pathogen sequences, indicating high base-calling accuracy and superior data quality. The read length distribution for these 10 pathogens was balanced, with uniform coverage, no unknown bases, and consistently high data quality.

Table 1
Third-generation sequencing data of the ten pathogenic strains

Total sequence number	**C5**	**Y8**	**Hzhj19**	**T7147**	**C9**	**YJdp-2**	**PB**	**PXO99^{A}**
	181,714	**189,351**	**168,323**	**178,301**	**177,857**	**238,068**	**158,376**	**233,681**
Total sequence length(bp)	2,508,671,831	2,500,364,249	2,371,206,555	2,443,514,097	2,378,306,381	3,282,370,351	2,206,858,577	2,906,423,003
Lngest(bp)	43,508	38,761	50,675	37,035	42,134	47,463	41,720	34,241
Shortest(bp)	300	1,078	68	260	1,101	103	257	316
N20(bp)	18,983	17,983	19,354	18,519	18,425	18,706	19,476	16,678
N50(bp)	14,957	14,170	15,194	14,661	14,520	14,747	15,290	13,264
N90(bp)	10,018	9,602	10,088	10,049	9,650	10,149	9,971	9,151
N number	0	0	0	0	0	0	0	0
N rate	0	0	0	0	0	0	0	0
GC content %	63.66	63.58	63.66	63.64	63.58	63.61	63.56	63.51

N20 (bp): All sequences are arranged from longest to shortest, and their lengths are cumulatively summed in this order. When the cumulative length reaches 20% of the total sequence length, the length of the last sequence added is defined as the N20 value. N50 (bp): All sequences are arranged from longest to shortest, and their lengths are cumulatively summed in this order. When the cumulative length reaches 50% of the total sequence length, the length of the last sequence added is defined as the N50 value. N90 (bp): All sequences are arranged from longest to shortest, and their lengths are cumulatively summed in this order. When the cumulative length reaches 90% of the total sequence length, the length of the last sequence added is defined as the N90 value. N number: The number of ambiguous bases. N rate: The proportion of ambiguous bases relative to the total length of the assembled sequences.

ANI analysis

Average nucleotide identity (ANI) is a widely used genome-wide metric for assessing genetic relatedness among bacterial genomes, with strains belonging to the same species typically exhibiting ANI values greater than 95%. ANI is calculated by comparing homologous genomic sequences across whole genomes, providing a quantitative measure of genomic similarity and an intuitive assessment of phylogenetic relatedness. In this study, ANI values were calculated using FastANI, and the resulting similarity matrix was visualized as a heatmap using TBtools[26]. Average nucleotide identity (ANI) analysis was performed to evaluate the genomic relatedness of the ten *Xoo* strains with representative *Xanthomonas* species. All ten strains exhibited ANI values greater than 99% with the reference *Xanthomonas oryzae* pv. *oryzae* genome, whereas ANI values with other *Xanthomonas* species ranged from approximately 86% to 91% (Fig. 2). Hierarchical clustering based on ANI values clearly separated *Xoo* from other *Xanthomonas* species, confirming the taxonomic identity of all isolates. Moreover, the highly similar ANI profiles among the ten *Xoo* strains indicate that they share a conserved genomic background despite differences observed in TALE repertoires and other accessory genomic features.

Global Genomic Features and Preliminary Classification

A preliminary assessment of whole-genome distribution among the ten pathogenic strains was conducted using circular genome maps generated with CGView (Fig. 2), which illustrated the complete genomic framework of each strain. Analysis of GC skew, GC content, coding sequence (CDS) annotation, and the positional distribution of CDS, tRNA, and rRNA revealed that the ten strains could be grouped into three categories. Strains C5, Y8, Hzhj19, and T7147 displayed highly similar genomic profiles and were classified as Group I. Strains C9 and YJdp-2 were assigned to Group II, while Ym1, PB, PXO99^{A}, and Ym187 were categorized as Group III.

Phylogenetic Tree Construction and Classification Validation Based on Core Genes

The prokaryotic whole-genome database UBCG established a bacterial core genome comprising 92 conserved genes derived from 1,429 complete genome sequences across 28 phyla[27]. In the present study, these 92 conserved genes were individually identified and extracted from the genome sequences of ten *Xoo* strains. Multiple sequence alignment and phylogenetic tree construction were then conducted (Fig. 3). The analysis demonstrated that the 10 *Xoo* strains could be divided into three distinct clades, a classification consistent with groupings based on overall genomic structural features. Specifically, strains C5, Y8, Hzhj19, and T7147 clustered into one branch; C9 and YJdp-2 formed a second branch; and Ym1, PB, PXO99^{A}, and Ym187 grouped into a third branch. Strains within the same branch exhibited the closest phylogenetic relationships and the highest sequence similarity.

Comparative Analysis of Gene Families Among the Three Strain Types

Gene families are groups of homologous genes that originate from a common ancestral gene through gene or genome duplication events and encode proteins with similar structures and functions[28]. In this study, gene family clustering analysis was performed using the predicted protein sequences of 10 *Xoo* strains (Fig. 4). The total number of gene families ranged from 4,326 to 4,443 across these strains. Distinct differences in gene family numbers were evident among the three phylogenetic clusters. Strains of Group I (C5, Y8, Hzhj19, and T7147) possessed fewer gene families, ranging from 4,326 to 4,386. Group II strains (C9 and YJdp-2) exhibited higher numbers, ranging from 4,436 to 4,473. Group III strains (Ym1, PB, PXO99^{A}, and Ym187) showed intermediate values, ranging from 4,427 to 4,443. Analysis indicated that the 10 pathogenic strains shared 3,906 common gene families, accounting for more than 80% of each strain's total gene families. However, substantial differences were observed in the number of strain-specific gene families. Group III strains contained the fewest unique gene families, ranging from 5 to 12, while no significant differences were detected between the other two groups.

Additional Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) annotations were conducted for the strain-specific genes of each strain (Fig. S1 and S2). The results demonstrated that the unique genes in Group I strains were predominantly enriched in functions related to genetic information processing, regulation, and execution. For Group II strains, the strain-specific genes of C9 were primarily enriched in quorum sensing, Type IV secretion, and regulatory processes, which are closely linked to the regulation of TALE expression and delivery. In contrast, the unique genes of YJdp-2 were mainly enriched in nitrogen metabolism, amino acid biosynthesis, and redox homeostasis, functions that may facilitate adaptation to metabolic constraints and oxidative stress resulting from non-TALE-mediated *Xa* resistance. The strain-specific genes of Group III strains were chiefly enriched in processes associated with genomic variation, adaptive evolution, and horizontal transfer of virulence genes.

Prediction of Virulence Factors

Virulence factors are properties, such as gene products, that enable microorganisms to establish themselves within a specific host species and increase their potential to cause disease[29]. In this study, BLAST was employed to compare gene-encoded protein sequences with amino acid sequences (Set B) from a virulence factor–related gene database to predict virulence-associated genes present in the genomes. The analysis indicated that all 10 strains contained numerous virulence genes, primarily distributed across 8 categories, with effector-delivery systems being the most prevalent. Further examination demonstrated that the total number of virulence genes closely corresponded with the phylogenetic clustering results. Specifically, strains of Group I (C5, Y8, Hzhj19, and T7147) contained 106–108 virulence genes; Group II strains (C9 and YJdp-2) possessed 108–111 virulence genes; and Group III strains (Ym1, PB, PXO99^{A}, and Ym187) harbored 103–105 virulence genes (Table 2).

Plasmid Prediction

Plasmids function as key vectors enabling bacteria to acquire and disseminate resistance genes, virulence factors, and metabolic traits, which may result in functional differences in pathogenicity among strains[30]. Whole-genome sequencing analysis revealed that only C5, C9, and PXO99^{A} among the ten strains carried plasmids, with C9 containing two plasmids. In contrast, Hzhj19, PB, T7147, Y8, YJdp-2, Ym1, and Ym187 lacked plasmids.

CRISPR Prediction

In certain pathogenic bacteria, virulence factors are encoded on plasmids or bacteriophages. The CRISPR-Cas system may indirectly stabilize virulence gene expression and sustain or enhance pathogenicity by preventing the loss or replacement of these exogenous genetic elements by other phages[31]. Analysis of CRISPR loci among ten pathogenic strains indicated that CRISPR sites were predominantly located on the chromosome (Table 3). The distribution of CRISPR locus start sites closely corresponded with phylogenetic clustering results. In particular, strains of Group I exhibited CRISPR locus start positions within the 853,909–892,728 bp region, Group II strains within the 867,793–867,816 bp region, and Group III strains within the 4,315,333–4,327,530 bp region. Strain YJdp-2 was distinct in possessing two independent CRISPR loci.

Table 3
CRISPR prediction results

C5	start	end	Number of Spacer	Length(bp)
	855,493	**861,705**	**94**	**6,213**
C9	867,816	874,024	94	6,209
PXO99^{A}	4,327,530	4,332,631	77	5,102
Hzhj19	883,195	887,693	68	4,499
PB	4,315,333	4,320,289	75	4,957
T7147	892,728	897,685	75	4,958
Y8	853,909	857,614	56	3,706
YJdp-2	867,793	871,240	52	3,448
	872,356	874,947	39	2,592
Ym1	4,315,441	4,320,397	75	4,957
Ym187	4,315,517	4,320,473	75	4,957

Start: the starting position of the CRISPR locus; End: the ending position of the CRISPR locus; Number of Spacers: the number of spacer sequences within the CRISPR locus; Length (bp): the total length of the CRISPR locus in base pairs.

Prediction of Antibiotic Resistance Genes

Although antibiotic resistance is not an intrinsic component of the pathogenic system, it demonstrates a synergistic relationship with pathogenicity, collectively facilitating pathogen proliferation and dissemination[32]. Analysis using the Comprehensive Antibiotic Resistance Database (CARD) (Table 4) indicated that the number of resistance genes in the ten pathogenic strains ranged from 53 to 55. This consistency in gene counts suggests that antibiotic resistance mechanisms are highly conserved across these strains. Furthermore, resistance in the ten strains is primarily achieved through modification of antibiotic target sites.

Table 4
Statistical summary of antimicrobial resistance gene prediction results

Resistance Mechanisme	C5	C9	PXO99^{A}	Hzhj19	PB	T7147	Y8	YJdp-2	Ym1	Ym187
Antibiotic efflux	21	22	22	22	22	22	20	22	22	22
Antibiotic inactivation	1	1	1	1	1	1	1	1	1	1
Antibiotic target alteration	29	29	29	28	29	29	29	29	29	29
Antibiotic target protection	1	1	1	1	1	1	1	1	1	1
Antibiotic target replacement	1	1	1	1	1	1	1	1	1	1
Resistance by absence	1	1	1	1	1	1	1	1	1	1
summer	54	55	55	54	55	55	53	55	55	55

Prediction of Protein Secretion Systems

Both pathogenic bacteria, which require protein secretion for survival, reproduction, and dissemination, and non-pathogenic bacteria, which secrete proteins for environmental adaptation, release proteins into their surroundings. The mechanisms responsible for transporting proteins across the cytoplasmic membrane are collectively termed secretion systems. These systems are classified into seven types (Type I–VII), with the Type III secretion system (T3SS) being particularly associated with the pathogenic mechanism of bacterial leaf blight[33]. Analysis revealed that all ten pathogenic strains contained a complete Type III secretion system, suggesting that each strain possesses the capacity to translocate TAL effectors into rice cells. Genomic data enable the systematic identification and annotation of TAL effectors in these strains. Additionally, a distribution matrix of TAL proteins can be constructed to facilitate comparison of TAL repertoires among the strains.

Identification and Redundancy Removal of TAL Effectors

TAL proteins from ten *Xoo* strains were identified using the TALE Prediction tool, yielding 158 predicted TAL effectors. These proteins were named according to the AnnoTALE nomenclature system, and duplicate names were subsequently detected. The corresponding sequences were aligned and analyzed phylogenetically using MEGA (Fig. S2). Redundant proteins were removed, resulting in a final set of 33 non-redundant TAL effectors.

Diversity of TAL Effector RVD Sequences

Analysis of conserved features in TAL effector repeat-variable diresidues (RVDs) enabled identification of the RVD sequences for 33 TAL proteins (Table S1). As indicated in Table S1, TAL proteins from different strains demonstrated considerable diversity in both RVD composition and repeat number. The number

of RVD repeats ranged from five units (TalLX2 from PXO99^{A}) to 36 units (TalLR2), highlighting dynamic evolutionary variation within this protein family. Regarding RVD types, in addition to common core units such as NN, HD, NI, and NG, several rare variant RVDs were detected, including N*, H*, S*, NK, NV, SS, HI, and ND. Comparative analysis showed that TAL proteins classified under the same nominal subtype (for example, TalLX2) possessed entirely different RVD sequences in different strains (Hzhj19 and PXO99^{A}), indicating that the nomenclature may primarily denote strain origin rather than functional homology. Additionally, TAL proteins annotated as "Pseudo" (such as TalHT26 and TalFS26) often began with NS or NG as frequent initial RVDs, and their overall RVD composition patterns differed substantially from those of canonical functional TAL effectors. This observation suggests that these proteins may represent pseudogenes or functionally degenerated effectors.

Phylogenetic Analysis of 33 TAL Effectors

A phylogenetic tree was constructed using the amino acid sequences of the RVD regions from 33 TAL proteins (Fig. 5). The analysis demonstrated that all examined TAL effectors were divided into eight distinct clades based on their clustering relationships, with these patterns showing strong correlation with strain origin and functional annotation. Pseudogene TAL proteins formed a separate cluster; all effectors annotated as "Pseudo" (such as TalHT26, TalFS26, and TalDR34) grouped into a distinct branch and exhibited substantial genetic distance from functional TAL proteins, indicating significant evolutionary divergence. This observation supports the hypothesis that these sequences represent pseudogenes or functionally degenerated effectors. Furthermore, TAL proteins with identical nominal designations but derived from different strains (for example, TalLX2-1 and TalLX2-2) were found in different clades, suggesting that the evolutionary diversification of TAL proteins is closely linked to strain lineage. Thus, identical nomenclature among TAL proteins may indicate strain origin rather than true functional homology.

Comparison of TAL Effector Numbers

Analysis of TALE effector numbers mapped onto the strain phylogenetic tree, constructed from core genes, demonstrated that TAL effector counts varied among *Xoo* strains but generally remained within a narrow range of 12 to 18. Strains within the same evolutionary lineage showed highly similar TALE counts, whereas notable differences were observed among distinct clades. Specifically, Group I strains contained 16 to 18 TALEs, Group II strains consistently contained 18 TALEs, and Group III strains harbored 12 to 17 TALEs.

Genomic Localization of TAL Effectors

Genomic localization analysis of TAL effector genes using Geneious[34] demonstrated highly similar distribution patterns across specific strain groups. The genomic positions of TAL proteins in C5, Hzhj19, T7147, and Y8 were highly similar; those in C9 and YJdp-2 were closely aligned; and the TAL protein locations in PB, PXO99^{A}, Ym1, and Ym187 were nearly identical. These results indicate that TALE genes are not randomly distributed across the chromosomes of different strains but are primarily concentrated in specific genomic regions, forming a distinct clustered distribution pattern. Comparative analysis among strains revealed that TALE gene distribution and structural arrangement were highly similar within a given phylogenetic lineage, whereas notable differences were observed between distinct lineages. These observations suggest that the physical localization of TALE genes is influenced by the evolutionary background of the strains.

Discussion

The reliability of genetic differentiation among the ten strains was validated using multiple analytical approaches. Global genomic features, visualized through circular genome maps, enabled the preliminary classification of the strains into three groups. This grouping was fully corroborated by core gene–based phylogenetic analysis. Although all ten strains shared nearly identical ANI values and exhibited highly conserved core genomes, remarkable diversity was observed in their TALE repertoires. This contrast suggests that strain-specific pathogenic variation is unlikely to result from large-scale genome divergence but rather from diversification of accessory virulence determinants, particularly TAL effectors. The results demonstrate that the observed differentiation reflects evolutionary divergence in conserved core genes, rather than superficial genomic variation, thereby accurately representing the genetic relationships among the strains[35]. Branch-specific patterns in gene family numbers, virulence factor abundance, and CRISPR locus distribution further underscore the biological significance of this classification. Group II strains (C9 and YJdp-2) harbor more gene families, virulence genes, and TAL effectors, which may confer greater environmental adaptability and pathogenic potential. Previous studies[2,24,36,37] and extensive laboratory inoculation experiments indicate that C9 and YJdp-2 exhibit stronger pathogenicity than the other eight strains, with YJdp-2 demonstrating virulence toward most tested rice varieties. The identification of two distinct CRISPR loci specific to YJdp-2 suggests that this strain may undergo adaptive evolution by increasing immunity against exogenous DNA[38]. This feature offers a novel perspective for understanding strain-specific evolutionary strategies. Additionally, all strains were found to contain a complete Type III secretion system, confirming that TAL effector delivery is a core pathogenic mechanism of *Xoo*, consistent with previous reports on bacterial blight pathogens[39].

The diversity and evolutionary characteristics of TAL effectors elucidate key mechanisms underlying pathogenic differentiation. High variability in RVD composition, repeat length, and phylogenetic relationships highlights the dynamic evolution of TAL proteins. The presence of rare variant RVDs and substantial differences in repeat numbers may serve as critical strategies for pathogen adaptation to immune pressures from different rice varieties[36]. The independent clustering of pseudo-TAL proteins supports the hypothesis of functional degeneration, suggesting that pathogens may adapt to ecological niches by losing specific effector functions during evolution[40]. Furthermore, the clustering patterns of TAL proteins are closely linked to strain lineage. Functional heterogeneity among identically named TAL proteins, as evidenced by significant differences in RVD sequences among strains, indicates that the current nomenclature system reflects strain origin rather than functional homology. Future research should prioritize RVD sequence characteristics for functional annotation. Additionally, the clustered chromosomal distribution and branch-specific localization of TAL genes suggest that their evolution is constrained by the genetic background of each strain, offering important insights into the origin and dissemination of TAL genes[41].

No significant correlation was found between phylogenetic grouping and the geographic origin of strains, highlighting the complex transmission and evolutionary dynamics of pathogens in modern agricultural ecosystems. Frequent cross-regional movement of rice seeds and germplasm exchange has disrupted the constraints of geographic isolation, resulting in strains from different regions sharing similar genetic backgrounds[42]. Selection pressure from rice resistance genes has driven the evolution of key pathogenic determinants, such as TAL effectors, leading to convergent evolution of virulence-associated genes among pathogens from distinct geographic areas[33]. Additionally, high-frequency horizontal transfer of TAL gene clusters mediated by mobile genetic elements, including transposons and plasmids, has accelerated genetic recombination and further diminished the impact of geographic isolation on genetic differentiation[43]. These factors explain why strains from the same geographic region may belong to different evolutionary lineages, while strains from different regions may cluster within the same phylogenetic branch.

The results of this study have significant practical implications for managing rice bacterial blight. Traditional strategies for disease forecasting and the deployment of resistant cultivars based primarily on geographic origin have clear limitations[16]. Future management should adopt a precision approach centered on core genome lineage and TAL effector functional profiling. For the deployment of resistant rice varieties, resistance genes should be selected according to the pathogenic TAL effector profiles of local strains, rather than inferring virulence types solely from geographic origin[44]. Disease surveillance should prioritize monitoring variations in TAL effectors and in the dynamics of the CRISPR system to enable early detection of emerging strain lineages capable of overcoming existing resistance genes. These strategies will provide precise targets for resistance breeding and improve the sustainability of disease control efforts.

Materials and Methods

Materials

This study utilized seven bacterial blight pathogen strains from China: C5, C9, Y8, Ym1, Ym187, Hzhj19, and YJdp-2. Of these, Y8, Ym1, Ym187, Hzhj19, and YJdp-2 are recently isolated strains. Strain T7147 was obtained from Japan, PXO99^{A} from the Philippines, and PB is a mutant derived from PXO99^{A}. PXO99^{A} served as the control and reference strain for genome sequencing. Additional details are presented in Table 5.

Table 5
Detailed information on the bacterial blight pathogens

Strain Name	Source
C5	China standard strain No.5 rice, from Sichuan
C9	China standard strain No.9 rice, from Yunnan
PXO99^{A}	International standard strain, race 6 of Philippine
PB	Mutant strain of PXO99^{A}
T7147	International standard stain, race 2 of Japan
Y8	Field-collected *Xoo* strains in Yunnan, China
Hzhj19	Field-collected *Xoo* strains in Yunnan, China
Ym1	Field-collected *Xoo* strains in Yunnan, China
Ym187	Field-collected *Xoo* strains in Yunnan, China
YJdp-2	Field-collected *Xoo* strains in Yunnan, China

Bacterial Culture

Xoo strains were cultured on NA solid medium (1 g/L yeast extract, 10 g/L sucrose, 17 g/L agar, 5 g/L peptone, and 3 g/L beef extract; pH 6.8–7.0) under inverted, dark conditions for 2–3 days. Cultured bacteria were transferred into sterilized 1.5 mL microcentrifuge tubes, and an appropriate volume of 1× PBS buffer was added. Tubes were gently inverted to disperse the cells. Bacteria were collected by centrifugation at 10,000 rpm for 10 min. The washing step was repeated 1–2 times until the supernatant became clear. After discarding the supernatant with no residual liquid, the bacterial pellet in each tube was adjusted to approximately 0.5 mL (about 0.5 g). Samples were rapidly frozen in liquid nitrogen for at least 15 min, placed in sealed bags, labeled with the sample name and date, and stored at − 80°C or immediately used for DNA extraction.

DNA Extraction

Bacterial pellets were washed with sterile water to eliminate residual culture medium components. Genomic DNA was extracted using a modified CTAB method. DNA quantification was performed with a fluorescent dye assay (Quant-iT PicoGreen dsDNA Assay Kit). DNA integrity was evaluated by 1% agarose gel electrophoresis. Only high-quality DNA (OD260/280 = 1.8–2.0; total DNA ≥ 3 μg) was selected for library construction and sequencing. The DNA working solution was stored at 4°C, and the stock solution was preserved at − 20°C.

PacBio Third-Generation Library Construction and Assembly

A minimum of 10 μg of genomic DNA was purified using 0.8× AMPure XP beads and dissolved in 100 μL of EB buffer. An additional purification step was performed using 0.6× PB beads, and the DNA was dissolved in 150 μL EB buffer. The purified DNA was transferred into a g-TUBE and fragmented by

centrifugation with an Eppendorf mini centrifuge. Fragmented DNA underwent treatment with Exonuclease VII for the initial round of DNA damage repair, followed by end-repair. After denaturation of PacBio adapters, ligation reagents were added, and the mixture was incubated overnight at 25°C to complete the ligation reaction. A double digestion with Exonuclease III and Exonuclease VII was then performed to remove linear DNA and damaged circular DNA. Genomic DNA fragments were size-selected using the BluePippin system (Sage Science, Beverly, MA, USA), typically retaining fragments larger than 10 kb. Following size selection, DNA was purified sequentially with 1× PB beads and 0.8× PB beads, then quantified using Qubit. Sequencing was subsequently performed on a PacBio sequencer.

The raw output data generated by PacBio sequencing were obtained in BAM format. Sequences were extracted from the BAM files and converted into FASTQ format using the bam2fastq software package. The read length distribution histogram was generated with ggplot2. Genome assembly was conducted using third-generation single-molecule sequencing data: PacBio raw reads were assembled with HGAP[45] and Canu[46] to generate contig sequences. High-quality second-generation sequencing data were subsequently used to correct the third-generation assembly results. The contigs were polished using Pilon[47], resulting in complete genome sequences. The assembled genomes were deposited in the NCBI GenBank database, and the corresponding accession numbers are provided in Supplementary Table S1.

Gene Prediction and Annotation

Genome sequences of *Xoo* strains were subjected to coding sequence (CDS) prediction using GeneMarkS[47]. Circular genome maps were constructed with CGView. Functional annotation of predicted CDSs was conducted using sequence alignment tools, including BLASTP, DIAMOND, Blast2GO, and KAAS, and comparisons were made against the NR, Swiss-Prot, Pfam, GO, COG, and KEGG databases[48–50].

Gene Family Identification and Analysis

Gene family analysis of the protein sequences was conducted using OrthoFinder, with DIAMOND set to an E-value threshold of 0.001 (-e 0.001).

Phylogenetic Tree Construction

Genome sequences of *Xoo* strains were obtained, and a phylogenetic tree was generated using the maximum likelihood method in MEGA 11 based on multiple sequence alignment (MSA) results. The resulting tree was visualized and refined with Evolview, and distinct evolutionary lineages were highlighted using different colors.

Identification and Analysis of TAL Effector RVD Regions

TAL effector–encoding sequences in *Xoo* strains were identified using AnnoTALE 1.5[51]. TAL proteins in each strain were detected with the TALE Prediction tool, and RVD regions were determined based on the conservation of their RVD sequences. Systematic names for TALEs were assigned using the TALE Class Assignment function in AnnoTALE, which applies a nomenclature based on TALE RVD sequences. The identified TAL proteins were named according to AnnoTALE conventions.

Declarations

Funding

This study was supported by the National Natural Science Foundation of China (Grant No. 32502441). This study was also supported by the National Key Research and Development Program of China, Key Project “Exploration and Innovative Utilization of Agricultural Biological Germplasm Resources” (No. 2021YFD1200102-02), the Yunnan Agricultural Joint Special Program – Key Project (No. 202301AT070001-149), the Yunnan Provincial Basic Research Program (General Project) (No. 202301AT070165), the Yunnan Academy of Agricultural Sciences Research Pre-Project (No. 2025KYZX-09), the Yunnan Science and Technology Talents and Platform Program—Field Scientific Observation and Research Station Construction Project (No. 202205AM340037), and the 2023 Provincial Agricultural Development Special Fund for Modern Seed Industry Development—Construction of Wild Rice Germplasm Resource Nursery (Yun Nong Tong Gao [2021] No. 8).

Acknowledements

We are grateful to Zaiquan Cheng and Ling Chen for generously providing the *Xanthomonas oryzae* pv. *oryzae* strains used in this study. We also gratefully thank the support of the Yunnan Academy of Agricultural Sciences and the Graduate School of Yunnan University.

Data Accessibility

The genome assemblies generated in this study have been deposited in the NCBI GenBank database. The accession numbers are JCBPEZ000000000 (PXO99A), JCBPFA000000000 (T7147), JCBPFB000000000 (Y8), JCBPFC000000000 (YJdp-2), JCBPFD000000000 (Ym1), JCBPFE000000000 (Ym187), JCBPFF000000000 (C5), JCBPFG000000000 (C9), JCBPFH000000000 (Hzhj19), and JCBPFI000000000 (PB). The corresponding BioProject and BioSample accession numbers are listed in Supplementary Table S1. Fig.S1,S2,S3 and Table S2 are listed in the supplementary infomation.

Author contributions

The investigation and methodology were completed by L.C. and Z.C. S.C., Y.Z., J.X., and L.L. performed the experiment, wrote an original draft, and revised the manuscript. T.Y., F.Y., and Q.F. assessed the data for potential analysis. J.L., D.Z., S.X.,B.W., and C.J. contributed reagents/materials/analytical tools and reviewed the draft.

References

1. Tauqeer, Z. et al. Unraveling rice defense against bacterial blight: Molecular pathways, resistance genes and breeding perspectives. *Microb. Pathog.* **209**, 108128 (2025).
2. Ochiai, H., Inoue, Y., Takeya, M., Sasaki, A. & Kaku, H. Genome sequence of xanthomonas oryzae pv. oryzae suggests contribution of large numbers of effector genes and insertion sequences to its race diversity. *JARQ* **39**, 275–287 (2005).
3. Jiang, N. et al. Resistance genes and their interactions with bacterial blight/leaf streak pathogens (*xanthomonas oryzae*) in rice (oryza sativa L.)—an updated review. *Rice* **13**, 3 (2020).
4. Bogdanove, A. J. & Voytas, D. F. TAL effectors: Customizable proteins for DNA targeting. *Science* **333**, 1843–1846 (2011).
5. Shah, S. M. A. et al. TALome and phenotypic analysis of pakistani *xanthomonas oryzae* pv. *oryzae* population revealed novel virulent TALEs contributing to bacterial blight of rice. *Phytopathol. Res.* **7**, 5 (2025).
6. Deng, D. et al. Structural Basis for Sequence-Specific Recognition of DNA by TAL Effectors. *Science* **335**, 720–723 (2012).
7. Hutin, M., Sabot, F., Ghesquière, A., Koebnik, R. & Szurek, B. A knowledge-based molecular screen uncovers a broad-spectrum OsSWEET14 resistance allele to bacterial blight from wild rice. *Plant J.* **84**, 694–703 (2015).
8. Mak, A. N. S., Bradley, P., Cernadas, R. A., Bogdanove, A. J. & Stoddard, B. L. The Crystal Structure of TAL Effector PthXo1 Bound to Its DNA Target. *Science* **335**, 716–719 (2012).
9. Bogdanove, A. J., Schornack, S. & Lahaye, T. TAL effectors: finding plant genes for disease and defense. *Curr. Opin. Plant. Biol.* **13**, 394–401 (2010).
10. Booher, N. J. & Bogdanove, A. J. Tools for TAL effector design and target prediction. *Methods* **69**, 121–127 (2014).
11. Yang, B. & White, F. F. Diverse Members of the AvrBs3/PthA Family of Type III Effectors Are Major Virulence Determinants in Bacterial Blight Disease of Rice. *MPMI* **17**, 1192–1200 (2004).
12. Li, M. et al. A new bacterial blight resistance gene *Xa50(t)* in the *Xa4* locus confers resistance against Xanthomonas oryzae pv. oryzae in rice. *Front. Plant. Sci.* **16**, 1657476 (2025).
13. Ji, C. et al. *Xa1* allelic R genes activate rice blight resistance suppressed by interfering TAL effectors. *Plant. Commun.* **1**, 100087 (2020).
14. Ji, Z. et al. Interfering TAL effectors of xanthomonas oryzae neutralize R-gene-mediated plant disease resistance. *Nat. Commun.* **7**, 13435 (2016).
15. Xu, X. et al. TALE-triggered and iTALE-suppressed Xa1-mediated resistance to bacterial blight is independent of rice transcription factor subunits OsTFIIAγ1 or OsTFIIAγ5. *J. Exp. Bot.* **72**, 3249–3262 (2021).
16. Zhang, B. et al. Multiple alleles encoding atypical NLRs with unique central tandem repeats in rice confer resistance to xanthomonas oryzae pv. oryzae. *Plant. Commun.* **1**, 100088 (2020).
17. Yuan, M. et al. A host basal transcription factor is a key component for infection of rice by TALE-carrying bacteria. *eLife* 5, e19605 (2016).
18. Streubel, J. et al. Five phylogenetically close rice SWEET genes confer TAL effector-mediated susceptibility to xanthomonas oryzae pv. oryzae. *New Phytol.* **200**, 808–819 (2013).
19. Chen, X. et al. Xa7, a new executor R gene that confers durable and broad-spectrum resistance to bacterial blight disease in rice. *Plant. Commun.* **2**, 100143 (2021).
20. Luo, D. et al. The *Xa7* resistance gene guards the rice susceptibility gene *SWEET14* against exploitation by the bacterial blight pathogen. *Plant. Commun.* **2**, 100164 (2021).
21. Gu, K. et al. R gene expression induced by a type-III effector triggers disease resistance in rice. *Nature* **435**, 1122–1125 (2005).
22. Tian, D. et al. The rice TAL effector-dependent resistance protein XA10 triggers cell death and calcium depletion in the endoplasmic reticulum. *Plant. Cell.* **26**, 497–515 (2014).
23. Wang, C. et al. XA23 is an executor R protein and confers broad-spectrum disease resistance in rice. *Mol. Plant.* **ssu132** 10.1093/mp/ssu132 (2014).
24. Bm, L. et al. The genome sequence of xanthomonas oryzae pathovar oryzae KACC10331, the bacterial blight pathogen of rice. *Nucleic acids research* **33**, (2005).
25. Salzberg, S. L. et al. Genome sequence and rapid evolution of the rice pathogen xanthomonas oryzae pv. oryzae PXO99A. *BMC Genom.* **9**, 204 (2008).
26. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
27. Na, S. I. et al. UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. *Journal Microbiology* **56**, (2018).
28. Xu, P. & Chen, Z. Genome-wide identification and characterization of the ALBA gene family in rapeseed (brassica napus L.) and its role in development and abiotic stress responses. *Front Plant. Sci* **16**, (2026).
29. Hautefort, I. & Hinton, J. C. D. Measurement of bacterial gene expression in vivo. *Philos. Trans. R Soc. Lond. B Biol. Sci.* **355**, 601–611 (2000).
30. Su, Z., Sakuma, S., Nitta, M., Gorafi, Y. S. A. & Nasuda, S. Extensive phenotyping and genome-wide analysis of grain number-related traits in a diverse wheat core collection. *Plant. Genome*. **19**, e70183 (2026).
31. Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
32. Zhu, B. Degradation of plasmid and plant DNA in water microcosms monitored by natural transformation and real-time polymerase chain reaction (PCR). *Water Res.* **40**, 3231–3238 (2006).
33. Tsuge, S., Furutani, A. & Ikawa, Y. Regulatory network of hrp gene expression in xanthomonas oryzae pv. oryzae. *J. Gen. Plant. Pathol.* **80**, 303–313 (2014).

34. Kearse, M. et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
35. Zhao, X. et al. Transcriptional control of carbohydrate catabolism by the CcpA protein in the ruminal bacterium streptococcus bovis. *Appl. Environ. Microbiol.* **89**, e0047423 (2023).
36. Chen, P. et al. Transcriptome and metabolome analysis of rice cultivar CBB23 after inoculation by xanthomonas oryzae pv. oryzae strains AH28 and PXO99A. *Plants (Basel)*. **13**, 1411 (2024).
37. Salzberg, S. L. et al. Erratum to: Genome sequence and rapid evolution of the rice pathogen xanthomonas oryzae pv. oryzae PXO99A. *BMC Genom.* **9**, 534 (2008).
38. Briner, A. E. et al. Occurrence and diversity of CRISPR-cas systems in the genus bifidobacterium. *PLoS One*. **10**, e0133661 (2015).
39. Kumar, A. et al. Deployment of genetic and genomic tools toward gaining a better understanding of rice-xanthomonas oryzae pv. oryzae interactions for development of durable bacterial blight resistant rice. *Front. Plant. Sci.* **11**, 1152 (2020).
40. Wang, X. Y., Zhou, L., Yang, J., Ji, G. H. & He, Y. W. The RpfB-dependent quorum sensing signal turnover system is required for adaptation and virulence in rice bacterial blight pathogen xanthomonas oryzae pv. oryzae. *Mol. Plant. Microbe Interact.* **29**, 220–230 (2016).
41. Ruh, M., Briand, M., Bonneau, S., Jacques, M. A. & Chen, N. W. G. Xanthomonas adaptation to common bean is associated with horizontal transfers of genes encoding TAL effectors. *BMC Genom.* **18**, 670 (2017).
42. Korinsak, S. et al. Identification of bacterial blight resistance loci in rice (oryza sativa L.) against diverse xoo thai strains by genome-wide association study. *Plants* **10**, 518 (2021).
43. Muñoz, M. et al. Community-acquired infection with hypervirulent clostridium difficile isolates that carry different toxin and antibiotic resistance loci: A case report. *Gut Pathog*. **9**, 63 (2017).
44. Peeters, N. et al. Repertoire, unified nomenclature and evolution of the type III effector gene set in the ralstonia solanacearum species complex. *BMC Genom.* **14**, 859 (2013).
45. Chin, C. S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods*. **10**, 563–569 (2013).
46. Koren, S. et al. Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017). Canu.
47. Besemer, J., Lomsadze, A., Borodovsky, M. & GeneMarkS A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **29**, 2607–2618 (2001).
48. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*. **12**, 59–60 (2015).
49. Conesa, A. & Götz, S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics* 619832 (2008). (2008).
50. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, W182–185 (2007).
51. Grau, J. et al. AnnoTALE: bioinformatics tools for identification, annotation and nomenclature of TALEs from Xanthomonas genomic sequences. *Sci. Rep.* **6**, 21077 (2016).

Table 2

Table 2 is available in the Supplementary Files section.

Figures

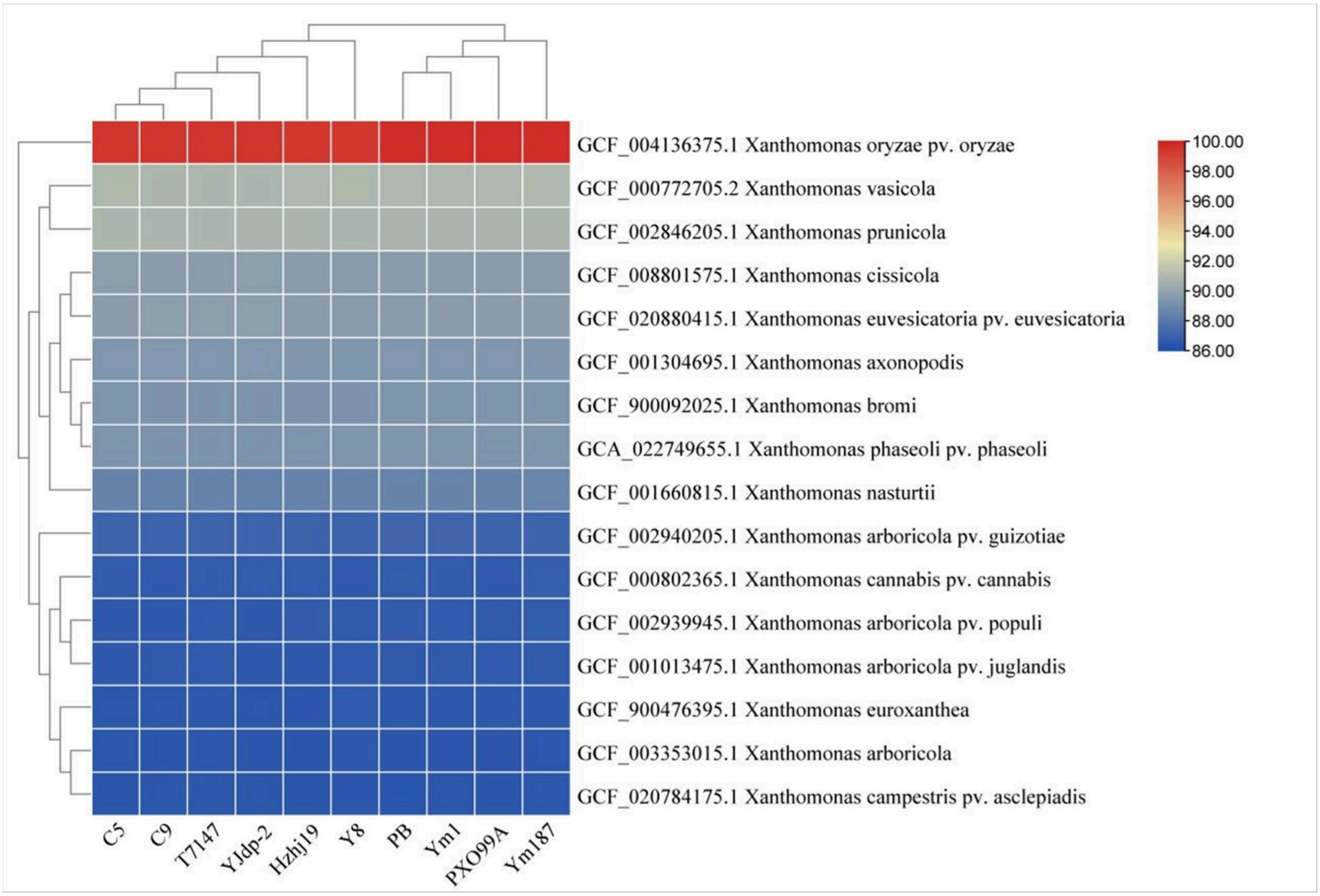


Figure 1

Average nucleotide identity (ANI) heatmap of ten *Xanthomonas oryzae* pv. *oryzae*strains against representative *Xanthomonas* species. Hierarchical clustering was performed based on ANI values calculated using FastANI. Each column represents one *Xoo* strain, and each row represents a representative genome of the *Xanthomonas* genus downloaded from the NCBI database. Color intensity indicates ANI values ranging from 86% (blue) to 100% (red). The dendrograms were generated using hierarchical clustering based on Euclidean distance.

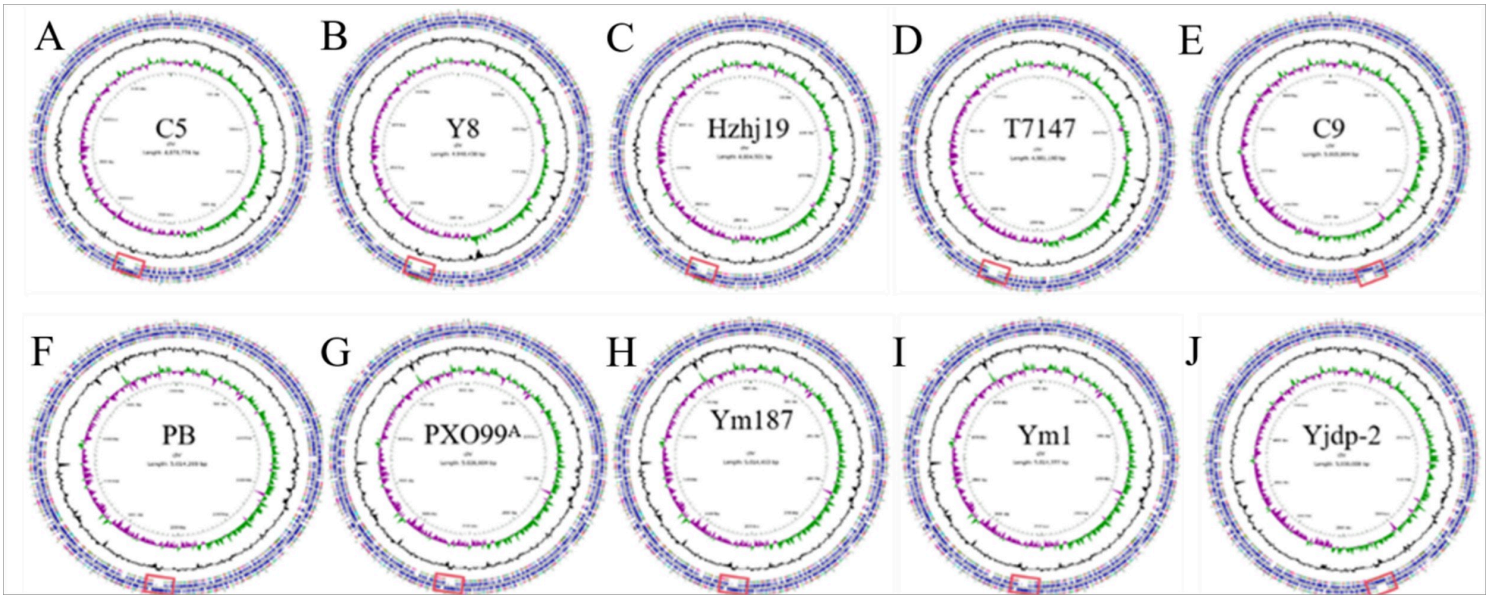


Figure 2

Circular genome maps of the ten pathogenic strains (A-J). From the inside to the outside, the first circle represents the scale; the second circle indicates GC skew, where green denotes skew+ (G content greater than C) and purple denotes skew− (G content less than C); the third circle represents GC content; and the fourth to seventh circles represent CDS annotation as well as the genomic positions of CDS, tRNA, and rRNA. Red boxes highlight regions with relatively pronounced differences.

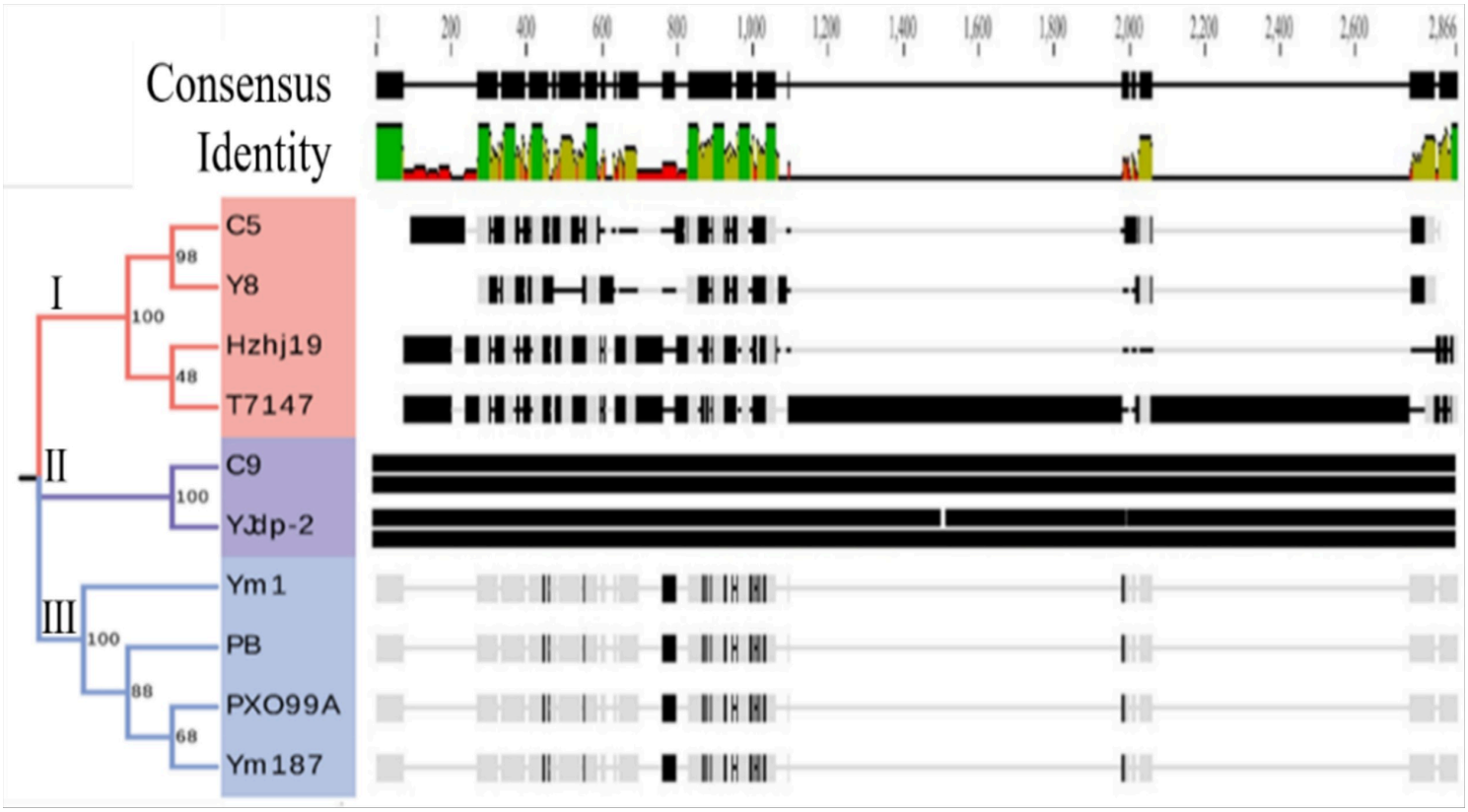


Figure 3

Core gene phylogenetic tree and sequence alignment of the ten pathogenic strains. The phylogenetic tree on the left was constructed using MEGA, while the panel on the right shows the multiple sequence alignment (MSA) results generated using Geneious. The Consensus row represents the “consensus sequence,” reflecting the most consistent sequence pattern derived from the multiple sequence alignment (the denser the black blocks, the more conserved the region across sequences). The Identity row indicates the proportion of sequence similarity among strains (the taller the bar, the higher the similarity, indicating that the corresponding position is more conserved).

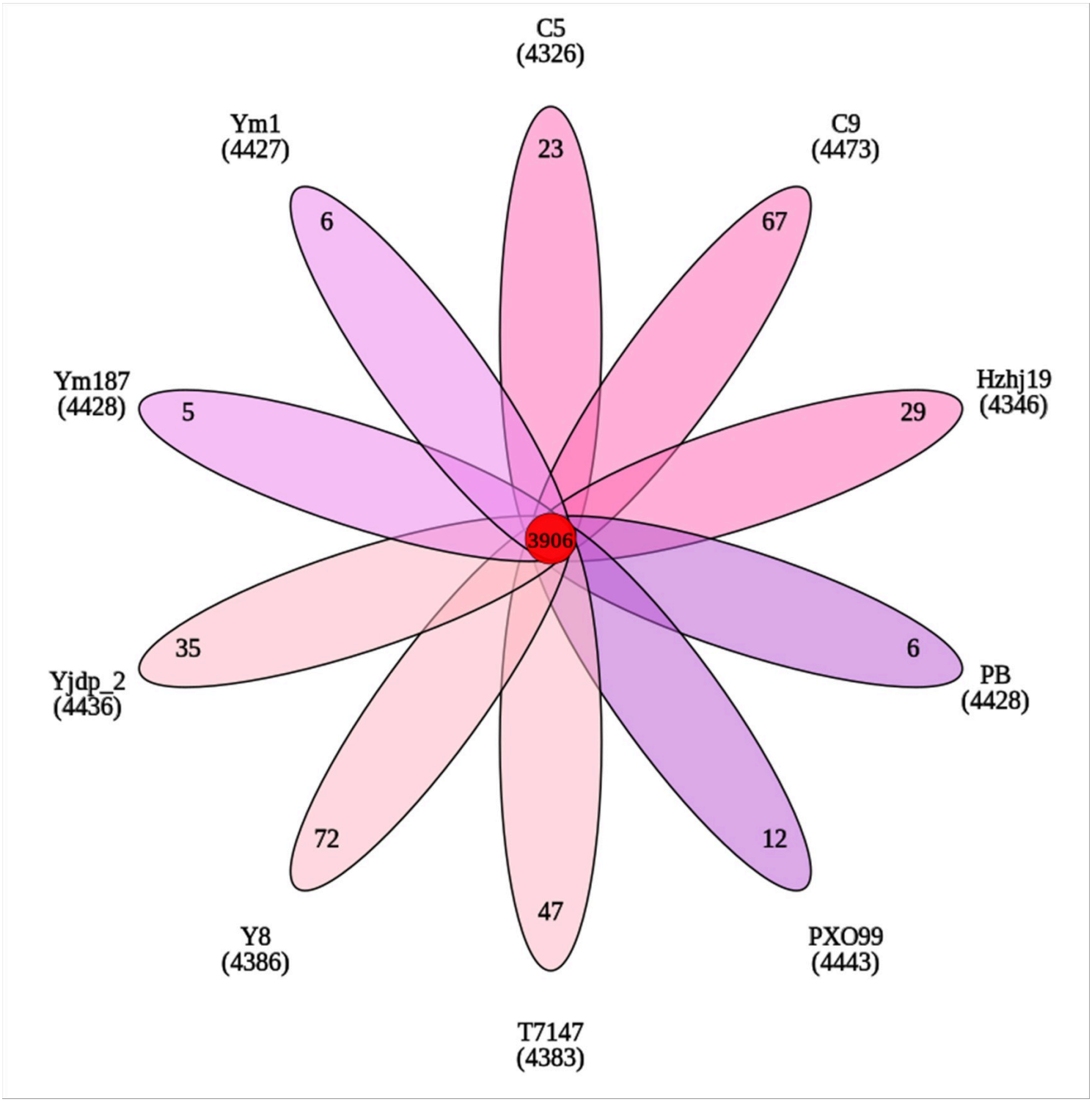


Figure 4

Venn diagram of gene family analysis. The outer circles represent the total number of strain-specific genes for each individual species, while the central circle indicates the number of gene families shared among *Xanthomonas* strains. The numbers in parentheses denote the total number of gene families present in each strain.

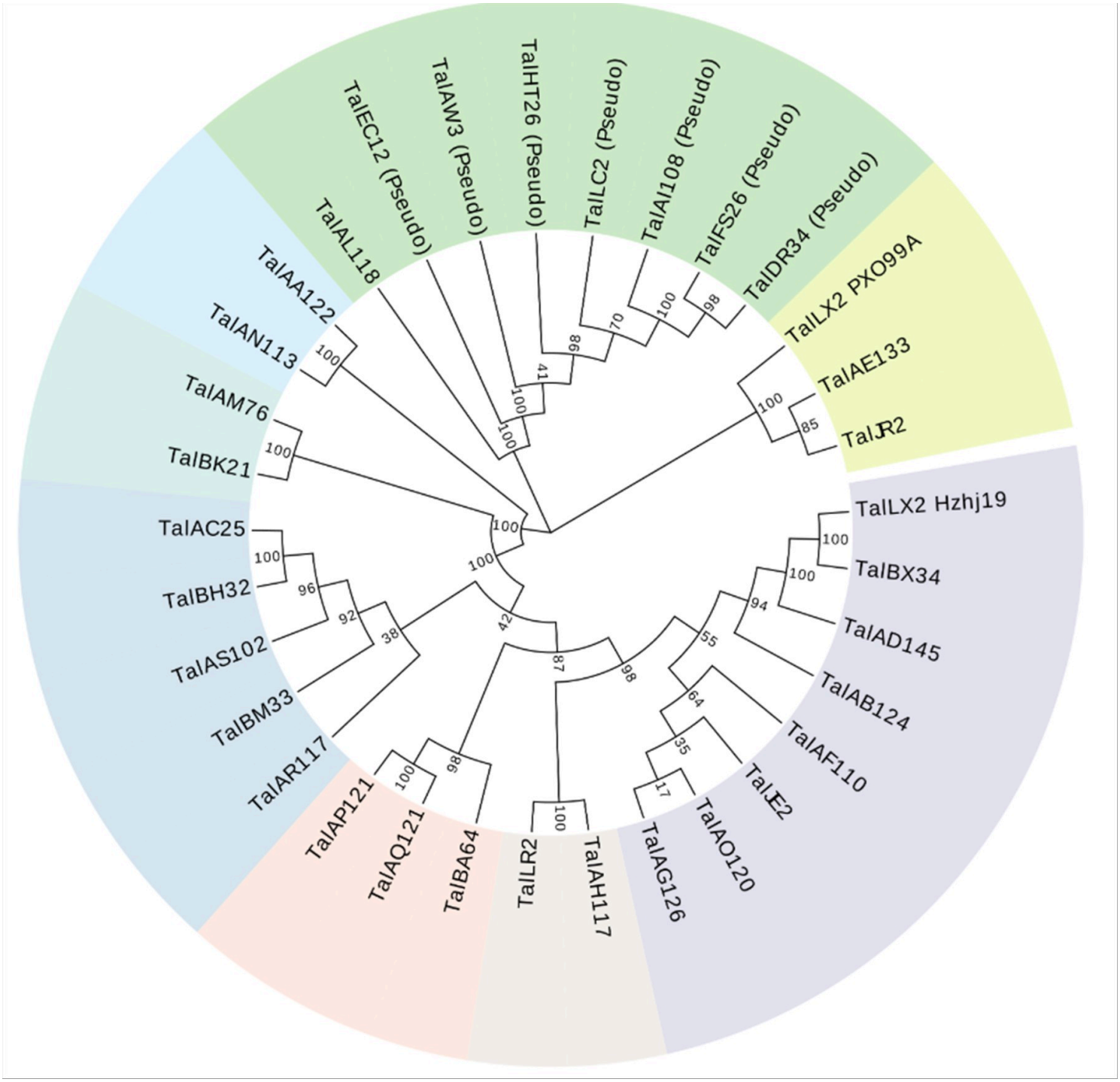


Figure 5

Species of the TAL evolutionary tree.

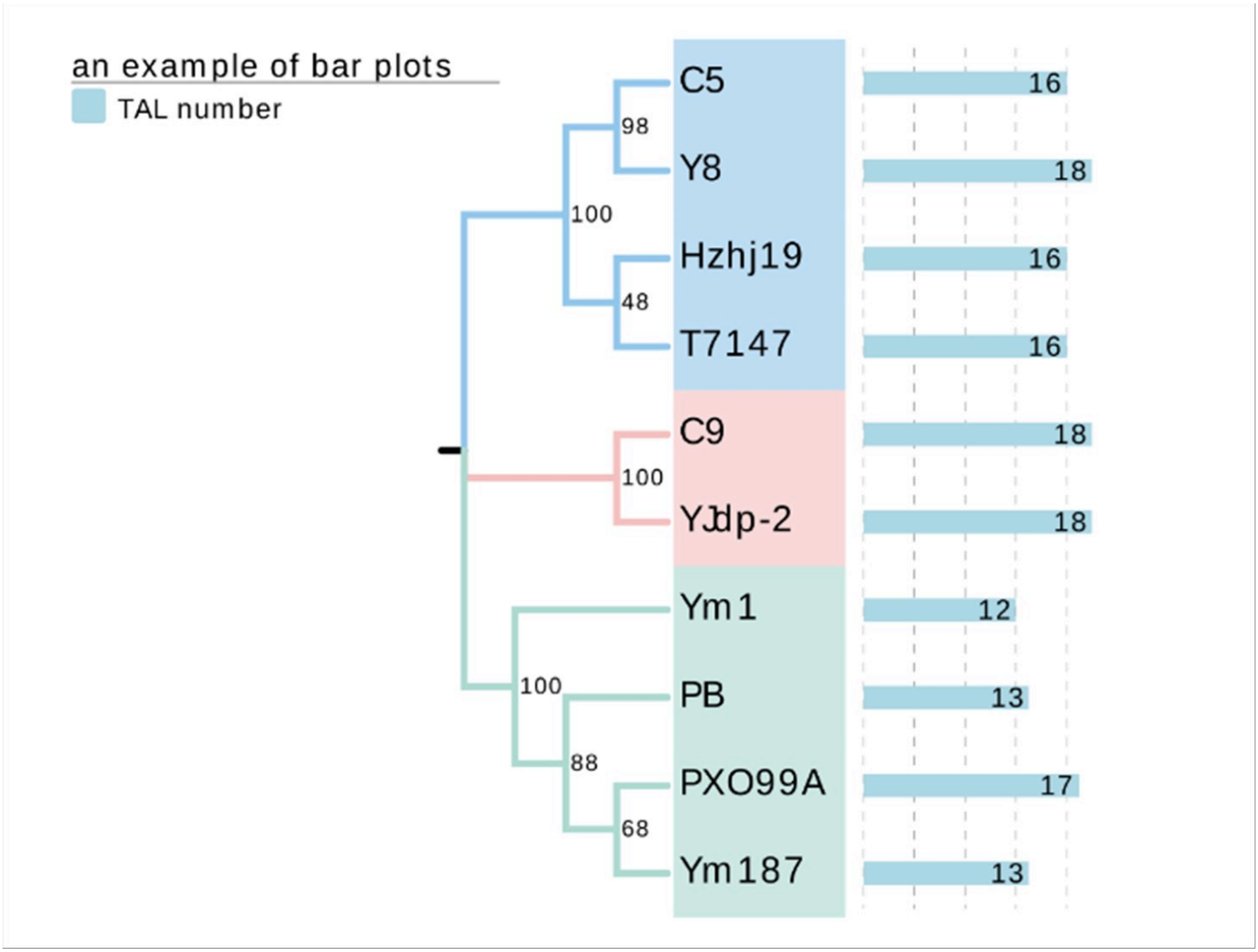


Figure 6

Comparative analysis of the number of TAL effectors in ten *Xoo* strains.

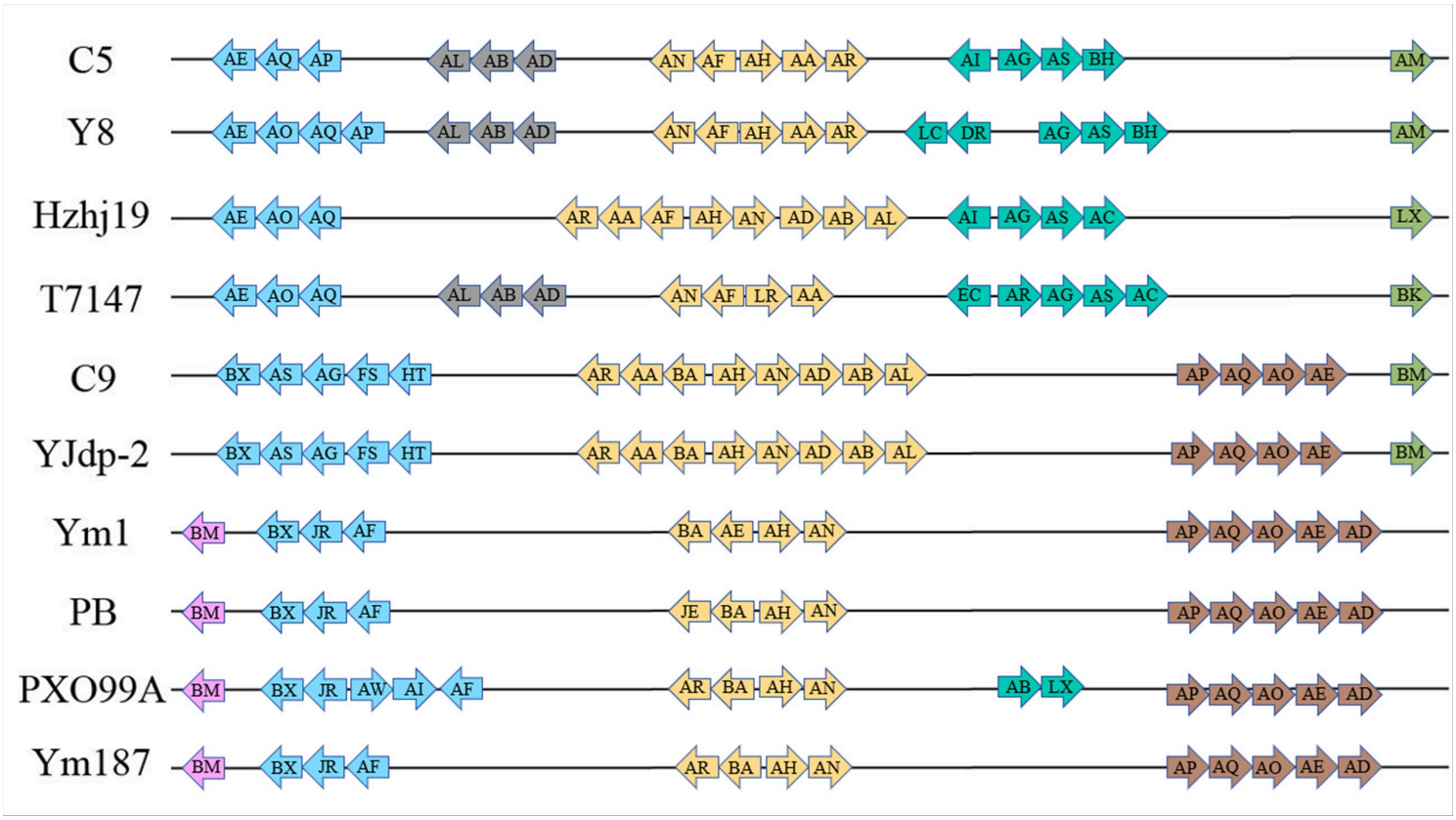


Figure 7

Chromosomal distribution of TAL effectors. The direction of the arrows indicates the transcriptional orientation of the genes (an arrow pointing to the right indicates that the tal gene is located on the positive strand of DNA and is transcribed in the 5'→3' direction; an arrow pointing to the left indicates that the tal gene is located on the complementary strand (negative strand), with transcription likewise proceeding in the 5'→3' direction).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryinformation.docx
- Table2.docx