Development of novel bioinformatic tools to study the diversity of *Erwinia amylovora* genomes

José Diogo Moura[1]

[1]Centre of Biological Engineering, University of Minho, 4710-057 Braga, Portugal
jddmoura@gmail.com

Abstract. Despite the remarkable genomic homogeneity of *Erwinia amylovora* (>97% similarity), the causative agent of Fire Blight disease, precise strain characterization remains crucial for phytopathological surveillance, phage therapy development, and disease management. Recent sequencing efforts have revealed subtle yet significant variations among strains, particularly in virulence-associated loci, host-specificity determinants, and phage resistance mechanisms, highlighting the need for novel typing methodologies. To address this challenge, here is presented ErwinATyper, a comprehensive Python-based *in silico* typing tool designed to detect and analyze relevant genomic differences among this species' genomes. This novel tool implements a modular framework that systematically analyzes pre-assembled *E. amylovora* genomes through multiple analytical pipelines, including CRISPR-Cas genotyping — which directly influences phage susceptibility patterns — streptomycin resistance profiling determinants, and molecular typing of critical surface polysaccharide biosynthesis clusters that can affect phage adsorption. Validation of ErwinATyper against an extensive dataset of publicly available *E. amylovora* genomes demonstrated its robust capability in delineating clade-specific characteristics, even within highly conserved genomic backgrounds, antimicrobial resistance distribution, and virulence factor architecture. The implementation of this standardized computational pipeline provides a powerful approach for distinguishing closely related strains and establishes a foundation for evidence-based surveillance strategies and targeted phage therapy development. ErwinATyper represents a decent advancement in molecular typing tools for Fire Blight research, facilitating both fundamental research and the development of targeted biocontrol approaches through phage-based interventions.

Keywords: Fire Blight, Bacterial Typing, Genomics, *Erwinia amylovora*, Python Tool

1 Context and Motivation

Fire blight, caused by the bacterial pathogen *Erwinia amylovora*, remains one of the most devastating diseases affecting commercial apple and pear production worldwide [1]. Despite its significant economic impact and research history, the pathogen continues to spread globally, with recent incursions into new regions including Korea and China [2, 3]. The disease is particularly concerning due to its rapid progression and the limited availability of effective control measures [4].

While *E. amylovora* exhibits remarkable genomic homogeneity among strains, recent genomic investigations have unveiled subtle yet significant variations that can affect virulence, host specificity, and responses to control measures [5]. Of particular interest are differences in the surface polysaccharide synthesis loci, namely the amylovoran capsule synthesis (KL) cluster, the lipopolysaccharide synthesis (OL) cluster, and the cellulose production (CL) cluster [6, 7, 8]. These loci, depicted in Figure 1, play crucial roles in bacterial survival and host-pathogen interactions.

It is also important to note that CRISPR-Cas systems, which function as bacterial adaptive immune mechanisms, have provided critical insights into the population structure of *E. amylovora*. As shown in Figure 1, the genome contains three CRISPR repeat regions (CR1–3). The genetic architecture of present spacers patterns helped define distinct phylogenetic groups within *E. amylovora*, with Amygdaloideae-infecting strains clustering into four primary groups: the Eastern North America group, the Western North America group (composed exclusively of strains from the United States), the Widely-Prevalent group (found worldwide), and another group termed *B-Group*, representing strains that exhibit substantial variation and do not align with the other groups [5, 9]. Notably, *Rubus*-infecting strains - a separate lineage - show a higher intrinsic diversity than isolates from Amygdaloideae plants, suggesting they have undergone distinct evolutionary trajectories. Notably, *Rubus*-infecting strains - a separate lineage -display higher intrinsic diversity than isolates from Amygdaloideae plants, suggesting they have undergone distinct evolutionary trajectories [5, 10].

The cell surface architecture comprises multiple complex systems, including Type III (T3SS) and Type VI (T6SS) secretion systems organized in distinct regions [11, 12], which may exhibit relevant strain-specific variations. Essential metabolic pathways, such as sorbitol utilization, are regulated by dedicated operons (SR) [13] and could demonstrate variation within different strains. Additionally, the bacterial flagellar systems, consisting of both distinct *flag-1* and *flag-3*, are encoded across multiple genomic regions and represent potential binding sites for bacteriophages [4, 14].

The emergence of streptomycin resistance [15], particularly prevalent in Western North American strains [5], coupled with the pathogen's complex population structure [16] underscores the need for comprehensive strain characterization tools. While current typing methods have revealed significant diversity in virulence factors and resistance determinants, they often lack the integration necessary to simultaneously assess multiple characteristics relevant to both basic research and applied disease management [17]. This complexity in pathogen diversity and host-pathogen interactions highlights the importance of developing new analytical tools capable of systematic strain characterization to support epidemiological surveillance and targeted intervention strategies.

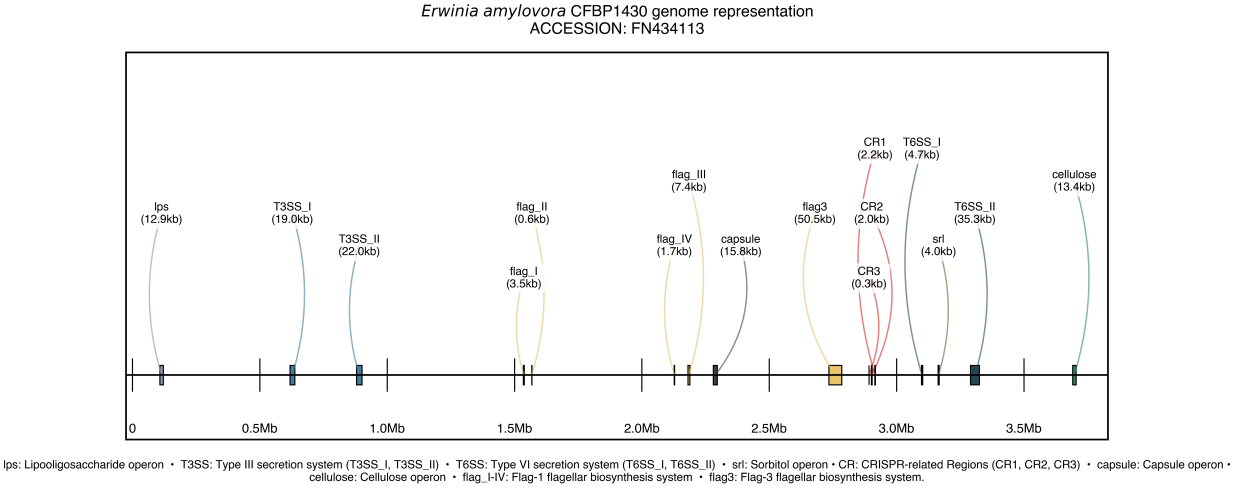

Fig. 1: Genome representation of *Erwinia amylovora* CFBP1430 (Accession: FN434113) showing the distribution of relevant loci for this study. The genome map indicates the relative positions and sizes of key regions including lipopolysaccharide (lps, purple), Type III secretion system (T3SS_I, T3SS_II, blue), Type VI secretion system (T6SS_I, T6SS_II, grey), flagellar biosynthesis systems (flag_I-IV and flag3, orange), capsule (grey), sorbitol operon (srl, grey), CRISPR-related regions (CR1-3, red), and cellulose operons (green). Different colors represent distinct functional loci categories. Distances are shown in megabases (Mb) and locus sizes are indicated in kilobases (kb).

2 Problem Analysis and Objectives

As mentioned, fire blight, caused by *Erwinia amylovora*, is one of the most destructive bacterial diseases in pomaceous fruits worldwide [18]. Despite extensive research over the past century, the pathogen remains difficult to manage due to its remarkable adaptability. On one hand, *E. amylovora* maintains high genomic similarity across strains (>97%), suggesting a stable genetic backbone; on the other hand, it has evolved clade-specific virulence factors under host selection pressures [5, 9]. Key challenges include the pathogen's rapid epiphytic growth on floral surfaces [19], its systemic colonization of xylem tissues [20], and the global emergence of streptomycin-resistant populations [15, 17]. Current management strategies face three fundamental limitations:

1. Overreliance on streptomycin, leading to dissemination of *strA-strB* resistance genes [5]
2. Lack of sufficient resolution in CRISPR-based typing schemes for major clades [9]
3. Possible variation in surface polysaccharide loci, which can reduce the effectiveness of phage therapy [6, 21]

Recent analyses highlight considerable accessory genome plasticity, including differences in plasmid content (e.g., pEA29), further complicating molecular surveillance [22, 23]. Concurrently, niche specialization has shaped the evolutionary trajectory of *E. amylovora*: *Rubus*-infecting strains have lost certain phenolic metabolism genes and acquired unique T3SS effectors [9, 24], whereas Amygdaloideae-infecting strains show altered amylovoran biosynthesis [25] and distinctive CRISPR spacer compositions [21]. Such variations impact epidemiological tracking [4] and biocontrol strategies [26], emphasizing the need for a unified typing framework.

Although comparative genomics has advanced our understanding of this pathogen [27], several critical gaps remain, such as outdated catalog of Type I-E CRISPR-Cas spacer profiles [21], no standardized multilocus sequence typing (MLST) scheme for global strain comparison, limited information on plasmid-associated virulence [28], and uncertain role of pEA29 in thiamine metabolism [25].

Objectives To address these challenges, the first MLST scheme for *E. amylovora* will be developed and integrated into an *in silico* typing pipeline. In addition to standardizing strain classification, the tool will detect streptomycin-resistance markers (*strA-strB*), identify key surface polysaccharide locus types, profile CRISPR spacers and other virulence factors (such as T3SS effectors), and include limited plasmid presence/absence screening. Specifically, the objectives are:

- *Develop and curate databases for relevant sequences.*
- *Establish a standardized MLST scheme for E. amylovora.*
- *Develop an integrated in silico tool* for automated MLST typing, CRISPR profiling, plasmid detection, streptomycin resistance screening, surface polysaccharide locus identification, and other relevant genomic features.
- *Enhance reproducibility and collaboration* via open-source release of the tool.

By pursuing these objectives within a single framework, the ambition is to facilitate improved monitoring of fire blight and support the design of targeted phage cocktails for *E. amylovora*. This direction is consistent with broader calls for prudent antimicrobial usage [18] and offers a practical path toward genomics-driven control strategies in plant pathology.

3 Methods and Materials

3.1 Genomic Dataset

The genomic dataset comprises 1,025 *Erwinia* genus genomes, including 510 *E. amylovora* genomes, retrieved from NCBI [29] as of December 22, 2024. Genome sequences were obtained using NCBI's datasets command-line tool with parameters for comprehensive assembly inclusion (complete genomes, chromosomes, scaffolds, and contigs) from both RefSeq [30] and GenBank [31] databases. The genome acquisition pipeline is publicly available here.

3.2 ErwiniaATyper Features and Typing Systems

ErwiniaATyper provides comprehensive genomic analysis capabilities for characterizing *Erwinia amylovora* strains through multiple complementary approaches (Table 1). The tool integrates essential strain information, assembly quality assessment, and extensive typing systems spanning CRISPR arrays, MultiLocus Sequence Typing (MLST), surface polysaccharides, secretion systems, and additional features such as levan synthesis and plasmid identification. This multi-faceted analysis enables detailed characterization. Each module integrated in ErwiniaATyper will be described below.

Table 1: Main Features Analyzed by ErwiniaATyper

Feature Category	**Description**
Basic Information	
Species	Species identification with ANI calculation utilizing reference genomes
Assembly Quality Metrics	
Assembly Statistics	N50, total size, and GC content
Quality Indicators	Ambiguous bases and largest contig
CRISPR Analysis	
CRISPR Arrays	CR1, CR2, and CR3 arrays with spacer counts and identifiers
Clade Assignment	Strain classification with confidence metrics based on CRISPR patterns
Core Typing Systems	
MLST Analysis	Ten-gene MLST scheme
Surface Polysaccharides	Analysis of capsule (KL), cellulose (CL), LPS (OL) loci, and Sorbitol (SR) locus
Secretion and Motility Systems	
Flagellar Systems	Analysis of *Flag-1* and *Flag-3* systems
Protein Secretion	Type III and Type VI secretion system analysis
Additional Features	
Levan Synthesis	Analysis of levan synthesis genes and associated virulence determinants
Resistance Analysis	Streptomycin resistance
Mobile Elements	Identification of plasmids

3.3 Development of Reference Loci Types Database

The development of a reference database for *E. amylovora* typing focused on key genomic regions associated with bacterial surface structures, secretion systems, and metabolic functions. Systematic identification of highly conserved genes flanking variable biosynthetic loci was performed across multiple functional systems

(Table 2). These regions encompass critical cellular processes, including capsule biosynthesis, lipopolysaccharide (LPS) synthesis, cellulose production, dual flagellar systems, Type III (T3SS) and Type VI (T6SS) secretion systems, and sorbitol metabolism. The spatial organization of these functional loci across the *E. amylovora* CFBP1430 genome is illustrated in Figure 1, highlighting their relative positions and sizes.

Table 2: Conserved genes flanking functional loci in *E. amylovora*

System	**Region**	**Flanking Genes**	**UniProt ID**
Capsule	Amylovoran synthesis (KL)	`amsL/amsG`	P33647/P33648
Cell Surface	LPS[a] synthesis (OL)	`waaD/waaQ1`	Q6XAK5/Q6XAK4
	Cellulose production (CL)	`bcsE/bcsO`	A0A6M4G7R5/A0A6M4G7S0
Flagellar (*flag-1*)	Region I (FLI)	`flgN/rne`	P0A1J2/P0A1K1
	Region II (FLII)	`fliE/rcsA`	P0A1J5/P0A1K4
	Region III (FLIII)	`fliZ/amyA`	P0A1J8/P0A1K7
	Region IV (FLIV)	`argS/flhD`	P0A1K0/P0A1K9
Flagellar *(flag-3)*	(FLT)	`cheW3/tsx`	D4HYX2/D4HZ27
T3SS[b]	Region I (TTI)	`hrpA1/rlsA`	P45466/P0DH78
	Region II (TTII)	`sipD1/orgA`	Q46654/Q46655
T6SS[c]	Region I (TSI)	`EAIL5_3103/rhs1`	A0A2S9QKN1/A0A2S9QKN3
	Region II (TSII)	`BN437_3323/EAMY_3228`	A0A2S9QKN5/A0A2S9QKN7
Metabolism	Sorbitol utilization (SR)	`srlR/srlA`	P37877/P37878

[a] Lipopolysaccharide; [b] Type III Secretion System; [c] Type VI Secretion System

An automated Python-based pipeline was implemented for systematic extraction and analysis of these loci, integrating BLAST [32] and BioPython [33]. The first step creates BLAST databases from curated sequences of the conserved flanking genes (Table 2), which were selected based on high conservation across *E. amylovora* strains and on their characterization in previous studies [34, 35, 8, 36].

The second step involves a BLASTN alignment of input genomes against these databases under stringent criteria: a minimum of 90% sequence identity and an 80% alignment coverage threshold. Once flanking genes are identified, the pipeline automatically extracts the intervening genomic regions with 200 bp extensions at both termini, thereby capturing any potential genetic variation in boundary regions.

The final step applies standardized annotation using Prokka (v1.14.5) [37] through a Docker container (staphb/prokka:latest) to ensure consistent annotation across different computing environments. This implementation utilized Docker (version 27.4.0) to manage the Prokka dependencies and execution, with custom reference databases passed as mounted volumes during the annotation process. Protein sequences from the annotated regions are analyzed by BLASTp to identify variant groups, assigning specific type designations based on unique combinations of genetic content. Type assignment is aided by Roary [38] to designate new types when gene presence/absence patterns differ or when amino acid sequence identity falls below 95% compared to existing types. Potential insertion sequences are screened using ISFinder [39], although no disruptive elements were detected in the analyzed regions.

3.4 ErwiniaATyper: Implementation

3.4.1 Overview

ErwiniaATyper is implemented as a Python-based analytical tool that processes *Erwinia amylovora* genome assemblies through a series of specialized analysis modules (Figure 2). This is a command-line tool freely available under the MIT License at Github. It is also available in Galaxy `https://galaxy.bio.di.uminho.pt/`. The software architecture provides modularity and scalability through parallel processing capabilities for the analysis of multiple genomes.

The implementation requires Python (≥3.9), Docker (≥26.0.0) [40], and BLAST (≥2.15.0+) [41] for execution, along with essential Python packages for computational (Table 3). Operating on FASTA-formatted whole-genome assemblies, the pipeline conducts sequential analyses including assembly quality assessment, species verification via Average Nucleotide Identity (ANI) calculations using pyANI (v0.2) in Docker, and typing using curated reference databases. The tool employs containerized environments for Prokka (v1.14.5) [37] annotation tasks.

The tool requires two mandatory parameters: `-input` (specifying one or more FASTA files or their directory location) and `-output_dir` (designating the output directory). Additionally, the command-line interface implements several optional parameters: the species assignment threshold (`-threshold_species`, default: 0.95) for ANI-based species determination, a binary flag for retention of intermediate loci sequences (`-keep_sequence_loci`), and an option to bypass species assignment (`-skip_species_assignment`).

The analysis pipeline comprises distinct modules for species identification, plasmid detection, genome assembly metrics calculation, MLST analysis, streptomycin resistance identification, and relevant loci typing. Results are provided in a comma-separated values (CSV) format containing 36 standardized fields per genome, ranging from basic assembly statistics to detailed typing results, with optional extracted sequences (.fasta and .gbk) available for further analysis. The tool's execution time was evaluated on a MacBook Air with Apple M1 processor, completing the analysis of a single *E. amylovora* genome in 227.289 seconds (≈3.8 minutes). This performance metric demonstrates the tool's efficiency in processing genomic datasets on consumer-grade hardware while maintaining comprehensive analysis capabilities.

Table 3: Software Dependencies and Version Requirements for ErwiniaATyper

Category	**Component**	**Description and Version**
Core Requirements	Python	≥3.9
	Docker	≥26.0.0
	BLAST+	≥2.15.0
Python Packages	NumPy	Scientific computing (1.26.4)
	Pandas	Data manipulation (2.2.0)
	BioPython	Biological computation (1.83)
	Scikit-learn	Machine learning utilities (1.4.0)
	SciPy	Scientific algorithms (1.12.0)
	Matplotlib/Seaborn	Visualization (3.8.2/0.13.2)
	Plotly	Interactive visualization (5.20.0)
Utility Packages	TQDM	Progress monitoring (4.66.1)
	PyYAML	Configuration management (6.0.1)
	Requests	HTTP functionality (2.31.0)
	Joblib	Parallel computing (1.3.2)
	Portalocker	File locking (3.1.1)

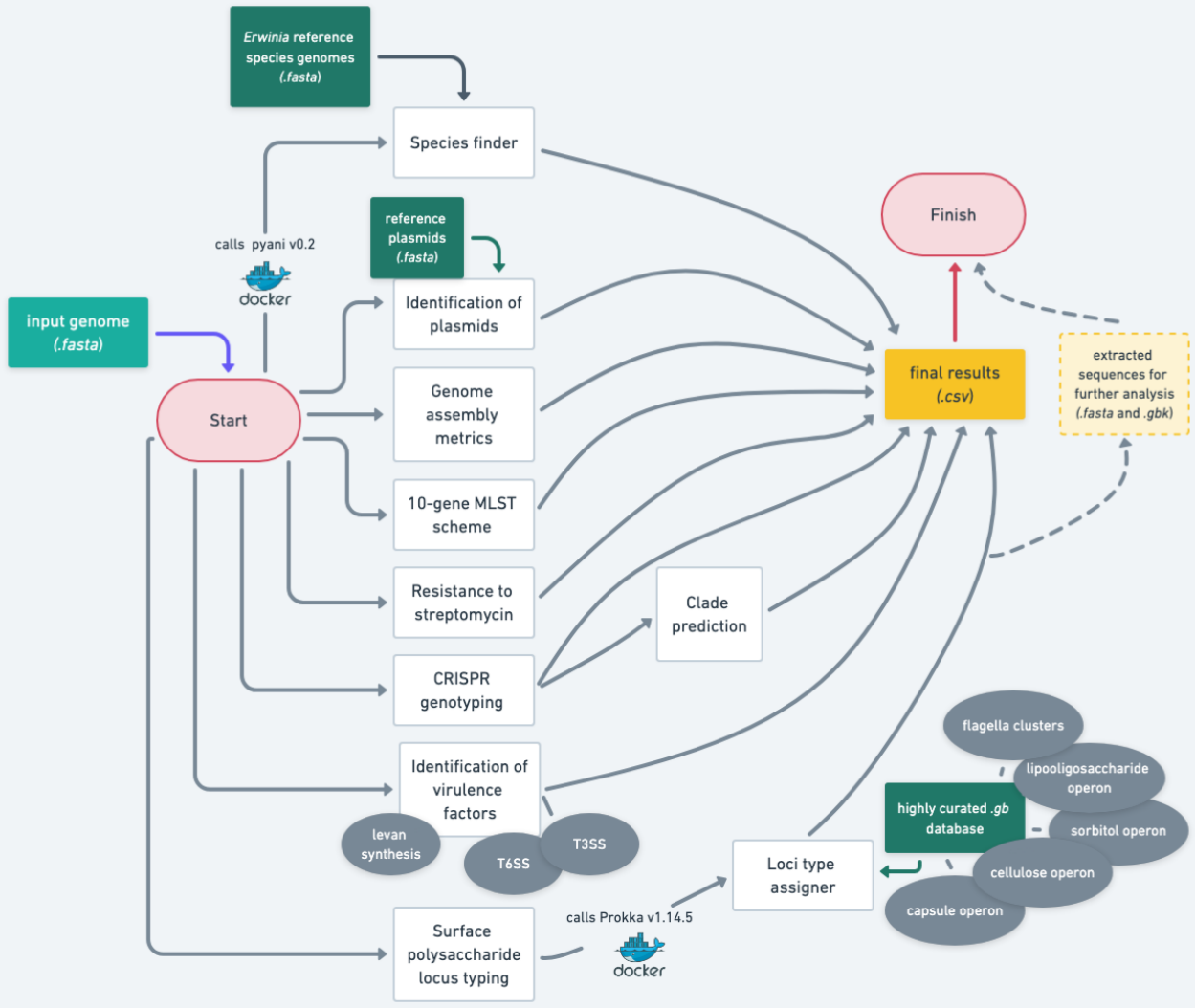

Fig. 2: Workflow diagram of ErwiniaATyper depicting the analysis modules and data flow. The pipeline starts with FASTA input and processes through multiple analysis steps. Docker containers (indicated by Docker's icon) are utilized for pyani (v0.2) [42] species identification and Prokka (v1.14.5) [37] annotation tasks. Green boxes indicate curated reference databases required for the analysis pipeline. Dashed arrows and pale yellow box represent optional outputs enabled by the `-keep_sequences_loci` parameter, which allows retention of extracted sequences (.fasta and .gbk files) for further analysis.

3.4.2 Species Assignment and Genome Metrics

Initial quality assessment evaluates assembly metrics including contig number, N50 values, and the presence of ambiguous bases. Species identification is performed using Average Nucleotide Identity (ANI) calculations against a curated set of NCBI reference genomes for the genus *Erwinia* (Table 4) through a containerized pyani [42] implementation. A threshold of $\geq 95\%$ identity is default value for species confirmation [43]. In cases where no species meets the defined threshold, the algorithm reports the best matching species and its corresponding ANI value.

Table 4: Reference genomes used for species assignment in ErwiniaATyper

Species	**Strain**	**RefSeq ID**
Pantoea beijingensis	JZB2120001	GCF_022647505.1
Candidatus E. dacicola	IL	GCF_001756855.1
Candidatus E. haradaeae	ErCicurvipes	GCF_900698925.1
E. amylovora	CFBP1430	GCF_000091565.1
E. aphidicola	JCM 21238	GCF_014773485.1
E. billingiae	Eb661	GCF_000196615.1
E. endophytica	A41C3	GCF_009295515.1
E. mallotivora	BT-MARDI	GCF_000590885.1
E. oleae	DAPP-PG531	GCF_000770305.1
E. persicina	Cp2	GCF_019844095.1
E. phyllosphera	CMYE1	GCF_019132875.1
E. piriflorinigrans	CFBP 5888	GCF_001050515.1
E. psidii	IBSBF 435	GCF_003846135.1
E. pyrifoliae	EpK1/15	GCF_002952315.1
E. rhapontici	BY21311	GCF_020683125.1
E. sorbitola	J780	GCF_009738185.1
E. tasmaniensis	Et1/99	GCF_000026185.1
E. tracheiphila	BHKY	GCF_021365465.1
E. typographi	M043b	GCF_000773975.1

3.4.3 Plasmid Detection

Plasmid identification was performed using nucleotide BLAST [41] analysis against a curated reference database with stringent thresholds: identity $\geq 90\%$ and coverage $\geq 80\%$. The reference database comprised 19 distinct plasmids previously identified in *Erwinia amylovora* strains (Table 5). These plasmids represent a diverse range of sizes (1.7 kb to 71.5 kb).

Table 5: Reference plasmids from *Erwinia amylovora* used in this study for plasmid detection.

Plasmid Name	Accession Number	Strain	Size (bp)
pEA1.7	NC_004940.1	IH3-1	1,711
pEA2.9	NZ_JAAEWU010000018.1	3446-1	2,846
pEA4.0	NZ_JAAEWD010000032.1	Ea1-00	4,087
pEAR4.3	FR719210.1	ATCC BAA-2158	4,369
pEAR5.2	NC_018985.1	ATCC BAA-2158	5,251
pEA5.8	NZ_JAAEUR010000003.1	EaIF	5,783
pEA6.0	NZ_JAAEUZ010000045.1	Ea7-96	5,955
pEa34	M95402.1	-	6,705
pEA29	NZ_CP157841.1	Ea 1/79	28,259
pEU30	NZ_JAAEWB010000012.1	Ea12	30,306
pEAR35	NZ_JAAEUQ010000042.1	EaLevo2	34,734
pEa-IncX	CP063692.1	32-10	43,634
pEAR28	NZ_JAAEVT010000068.1	Ea1-98	56,855
pEL60	NZ_JAAEUV010000003.1	EaA-11	59,865
pEM65	JQ292796.1	CFBP7517	59,940
pEA68	HG813238.1	692	59,940
pEA60	NZ_CP104024.1	Ea102	61,198
pEI70	NZ_JAAEWL010000015.1	CFBP 7130	65,238
pEA72	NZ_JAAEUS010000006.1	EaG5	71,499

Note: All plasmids were retrieved from NCBI databases

The selected plasmids were retrieved from the National Center for Biotechnology Information (NCBI) databases and represent the most comprehensive collection of known *E. amylovora* plasmids available at the time of analysis.

3.4.4 Streptomycin Resistance

Streptomycin resistance profiling employs a dual analysis approach examining both acquired resistance genes and chromosomal mutations. The analysis module integrates BLAST-based identification of resistance genes (*strA*/*strB*; UniProt Q79DN3/Q57204) with strict thresholds (identity $\geq$ 90%, coverage $\geq$ 80%) and targeted examination of the *rpsL* gene (UniProt P45809). For *rpsL* analysis, the gene is extracted with 50 bp flanking regions and annotated using Prokka to determine the amino acid at position 43, where a lysine (K) to arginine (R) mutation confers resistance. Aminoglycoside phosphotransferases *strA* and *strB* catalyze ATP-dependent phosphorylation of streptomycin, while *rpsL* gene, encoding the ribosomal protein S12, can harbor mutations that affect the binding of streptomycin to the ribosome, thus conferring resistance [5].

3.4.5 Loci Type Assignment

The core typing pipeline extracts genomic regions using flanking genes followed by protein sequence analysis using BLASTP examines translated gene content after Prokka [37] annotation. Gene differences are detected through sequence comparison metrics, including amino acid sequence identity below 100%, incomplete query or subject coverage, and deviations in sequence length ratios between query and reference genes.

The type assignment algorithm classifies loci into six hierarchical confidence levels based on cumulative analysis of sequence variations and gene content. A Perfect designation requires complete identity (100%) with the reference sequence, no gene differences, and coverage equal to 100%. Very High classification allows up to 2 genes with amino acid sequence variations at $\geq$97% coverage. High classification permits up to 4 genes showing amino acid sequence variations with 95% coverage, while Good designation accommodates up to 6 genes with sequence variations at 90% coverage. Low classification extends to sequences with up to 8 genes showing variations and 80% coverage. Very Low designation applies to sequences maintaining at least 70% coverage or having no more than 10 genes with sequence variations, truncations, or missing genes. Any sequence with absent or additional genes is assigned a Very Low confidence level. As these regions are highly conserved, genes where we have a slight minimum difference at the amino acid level are going to be reported in the final CSV as a "Flagged Gene". For further analysis, one can utilize the option to download these loci for the genome input.

3.4.6 CRISPR Analysis and Clade Assignment

CRISPR-based typing in *E. amylovora* employs a two-stage analysis approach utilizing distinct repeat-spacer arrays. The system identifies three canonical repeat sequences:

CR1: 5'-GTGTTCCCCGCGTGAGCGGGGATAAACCG-3' (29 bp)
CR2: 5'-GTGTTCCCCGCGTATGCGGGGATAAACCG-3' (29 bp)
CR3: 5'-GTTCACTGCCGTACAGGCAGCTTAGAAA-3' (28 bp)

Initial spacer identification utilizes a curated database of previously characterized spacer sequences from Rezzonico et al. (2011) [35]. Each spacer has a unique identifier and sequence as established in their work. Spacer identification employs sequence matching with stringent thresholds (identity $\geq$ 95%, coverage $\geq$ 95%). Valid spacers must meet specific length constraints: 30–35 bp for CR1/CR2 and 20–50 bp for CR3. Sequences outside these parameters are flagged as potential assembly artifacts. Clade assignment integrates a dual weighted criteria: genotypic signatures determined by spacer presence/absence patterns (70%) and spacer abundance based on CR type-specific count distributions (30%). This system delineates five major clades with characteristic spacer profiles shown in Table 6 as determined by Parcey et al. [5].

Table 6: CRISPR spacer count ranges across major clades

Clade	Subgroup	CR1	CR2	CR3
Widely Prevalent	Ia	28–36	31–35	5
(WP)	Ib	27–36	21–26	5
	II	12–19	27–35	5
Western North American (WNA)	III	44–110	25–49	5
Eastern North American (ENA)	IV	53–95	25–58	5
B-Group	I–IV	Variable	Variable	5
Rubus	I–III	Variable	Variable	3–6

The classification system employs a composite confidence scoring mechanism that combines genotype matching with spacer count analysis. This generates confidence levels categorized as Excellent for scores of ≥99%, Very High for ≥95%, High for ≥85%, Moderate for ≥70%, Low for ≥60%, and Very Low for scores below 60%.

3.4.7 Levan Synthesis

Levan synthesis pathway analysis examines three essential genes: levansucrase (*lsc*; UniProt Q4R0I7) and its regulators (*rlsA*/*rlsB*; UniProt O54509/Q8VNT8). The analysis pipeline employs BLAST-based [41] identification with specific thresholds (identity ≥ 75%, coverage ≥ 80%). Presence and sequence variations of these genes define distinct *E. amylovora* host-specificity patterns. The Widely-prevalent (Amygdaloideae-infecting) strains contain all three genes with high sequence conservation, while *E. pyrifoliae* strains lack *lsc*. *Rubus*-infecting strains show *rlsA* variants or absence, and blossom-limited strains lack *rlsB*, leading to distinct levan production phenotypes [44].

3.4.8 MLST of *Erwinia amylovora*

The MultiLocus Sequence Typing (MLST) scheme employed here was constructed by selecting genes that fulfill several key criteria for broad applicability in bacterial genotyping: they are (i) conserved among different strains of *E. amylovora*, (ii) broadly distributed throughout the bacterial genome, (iii) present as a single copy, and (iv) under neutral or near-neutral selective pressure. Specifically, four genes (`arcA` *(UniProt: D4I0J3)*, `rpoB` *(UniProt: D4HUQ8)*, `gyrB` *(UniProt: D4I4I9)*, `mdh` *(UniProt: D4HUW7)*) were drawn from a study on *E. amylovora* by Mann *et al.* [23], two (`pgi` *(UniProt: D4I2T0)*, `fusA` *(UniProt: D4I2Y1)*) were selected because they were used in an *E. coli* MLST context by Adiri *et al.* [45], and four (`infB` *(UniProt: D4HV18)*, `recA` *(for* E. amylovora*, often annotated alongside RecBCD, see D4HWI5)*, `adk` *(UniProt: D4HYK0)*, `dnaA` *(UniProt: D4I4J2)*) were taken from Diancourt *et al.*, also referenced in Guo *et al.* [46]. Each set of alleles was aligned with MUSCLE [47] (`-align` and `-output` parameters for version 5), and a distance matrix based on pairwise identity was constructed for each gene.

Hierarchical clustering was then performed using an average-linkage algorithm (via `scipy.cluster.hierarchy`). To determine the optimal number of clusters per gene, then it was incrementally tested a range of possible cluster counts (from 2 to 50) and calculated silhouette scores, selecting the clustering partition with the highest mean silhouette value. Allele assignments were subsequently merged to form a ten-locus profile `(adk-arcA-mdh-recA-dnaA-pgi-fusA-gyrB-infB-rpoB)` for each of the 494 *E. amylovora* genomes. This approach ultimately yielded 44 distinct MLST patterns, suggesting relevant allelic diversity across the analyzed strain collection.

In practice, allele assignment is performed by mapping each query genome against reference allele databases for each of the 10 genes (via BLASTN [41]). The final concatenated allele combination is then compared against a pattern library to determine the corresponding sequence type (ST).

Statistical validation using χ^2 and Fisher's exact tests with Bonferroni correction confirmed significant associations between sequence types and genomic features (Supplementary Table 9).

3.4.9 Mobile Genetic Elements and Defense Systems

Analysis of mobile genetic elements and defense systems was conducted across the *Erwinia* genus genomes using PADLOC [48] (command-line interface with default parameters). Additionally, PHASTEST [49] analysis was performed specifically on *E. amylovora* genomes using the official Docker container with default settings in *lite* annotation mode.

4 Results and Discussion

4.1 Genetic Diversity of Relevant Loci

The analyses did not reveal any novel capsule type despite previous reports that variation within the KL (capsule) cluster is correlated with differences in amylovoran production—a key determinant of both biofilm formation and host immune evasion [6]. It is possible that alternative methods (e.g., higher-resolution sequencing or functional assays) might detect subtle variations not captured here. In contrast, structural differences were observed in the OL (lipopolysaccharide) locus. The predominant OL01 configuration was found in 480 strains from Amygdaloideae-infecting isolates, whereas a distinct OL02 type was identified in 30 strains that specialize on *Rubus* hosts, suggesting that modifications in lipopolysaccharide structure may contribute to adaptation to different plant receptors [35]. Meanwhile, the CL (cellulose) cluster was highly conserved among 510 strains, underscoring its essential role in cellulose biosynthesis and the maintenance of biofilm integrity [8].

Examination of the flagellar systems (Figure 1) revealed further diversity among surface-exposed structures. The *flag-1* locus exhibited two distinct architectural variants. The predominant FL01 configuration, present in 480 strains, contains a complete set of region IV components, whereas the FL02 variant—observed in 28 strains (exclusively among *Rubus* specialists)—shows alternative features that may affect the regulation of motility [14]. In contrast, the ancestral *flag-3* system (FLT01) was retained in all strains; beyond its role in motility, this system may also function as a receptor for bacteriophages [4].

Variability was also detected within secretion systems. The Type III secretion system (T3SS) exhibits a modest level of divergence: 430 strains carry the TT01 complex, while 36 strains (mostly within the Western North American clade) harbor a TT02 variant with divergent components [11]. Similarly, the Type VI secretion system (T6SS) displays several variant configurations. A primary TS01 configuration was found in 94 genomes, whereas rare combinations (designated TS02 and TS03) were observed among Widely-prevalent strains. Additionally, a unique TS04 variant was detected exclusively in *Rubus* strains, and a very rare TS05 configuration was identified in a limited number of strains—primarily among Western North American lineages [12]. It should be noted that, because some T6SS regions are large and occasionally split across contigs, the ability to assess the complete sequence architecture in every strain was limited.

Collectively, the distribution of these virulence and metabolic loci (summarized in Table 7) indicates that, despite the overall high conservation of the *E. amylovora* core genome, focused analysis of surface-associated and secretion-related regions reveals sufficient genetic diversity to account for differences in virulence, host specificity, and phage susceptibility. These findings emphasize the importance of high-resolution molecular typing (e.g. as implemented in the ErwinATyper pipeline) for robust epidemiological surveillance and the development of targeted control strategies [4, 11].

Table 7: Distribution of the Loci Types in *E. amylovora*

System Class	Genetic Architecture	Count	Designation
Flagellar	*flag-1* complex I (FLI01/II01/III01/IV01)	480	FL01
	flag-1 complex II (FLI01/II01/III01/IV02)	28	FL02
	flag-3 system (FLT01)	496	FLT01
T6SS	Primary variant (TSI01/TSII01)	94	TS01
	Variant IIa (TSI01/TSII05)	12	TS02
	Variant IIb (TSI03/TSII01)	4	TS03
	Variant IIc (TSI01/TSII02)	2	TS04
	Variant IId (TSI04/TSII01)	2	TS05
T3SS	Primary complex (TTI01/TTII01)	430	TT01
	Secondary complex (TTI01/TTII02)	36	TT02
Core Systems	Sorbitol utilization type I	486	SR01
	Sorbitol utilization type II	22	SR02
	Capsule synthesis type I	508	KL01
	Cellulose synthesis type I	510	CL01
	O-antigen synthesis type I	480	OL01
	O-antigen synthesis type II	30	OL02

4.2 Defense Systems in *E. amylovora*

The analysis of bacterial defense systems across the *Erwinia* genus ($n = 451$) revealed that the Type I-E CRISPR-Cas system has a widespread but non-universal distribution, being present in approximately 52% of examined *Erwinia* genomes (Figure 3). When examining *E. amylovora* strains specifically ($n = 209$), the Type I-E CRISPR-Cas system showed strong conservation, being present in all analyzed genomes, suggesting its fundamental importance in this species' defense repertoire (Figure 4). However, experimental data presented by **(author?)** [50] demonstrates that despite possessing the Type I-E CRISPR-Cas system, *E. amylovora* strains remain susceptible to phage infection, with multiple phages capable of forming plaques on various *E. amylovora* strains. This indicates that while the Type I-E CRISPR-Cas system may serve as an important defense mechanism, it does not confer complete immunity against phage infection.infection.

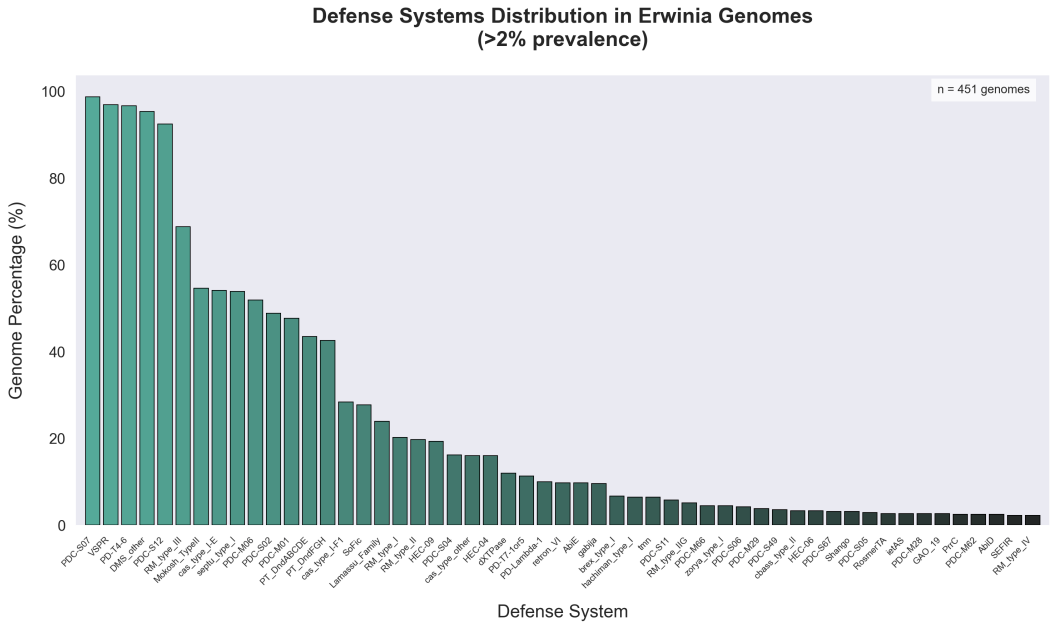

Fig. 3: Distribution of defense systems across the *Erwinia* genus ($n = 451$, systems with >2% prevalence shown). The cas_type_I-E (Type I-E CRISPR-Cas) system demonstrates significant presence at 52% prevalence across examined genomes, indicating its widespread but non-universal distribution within the genus.

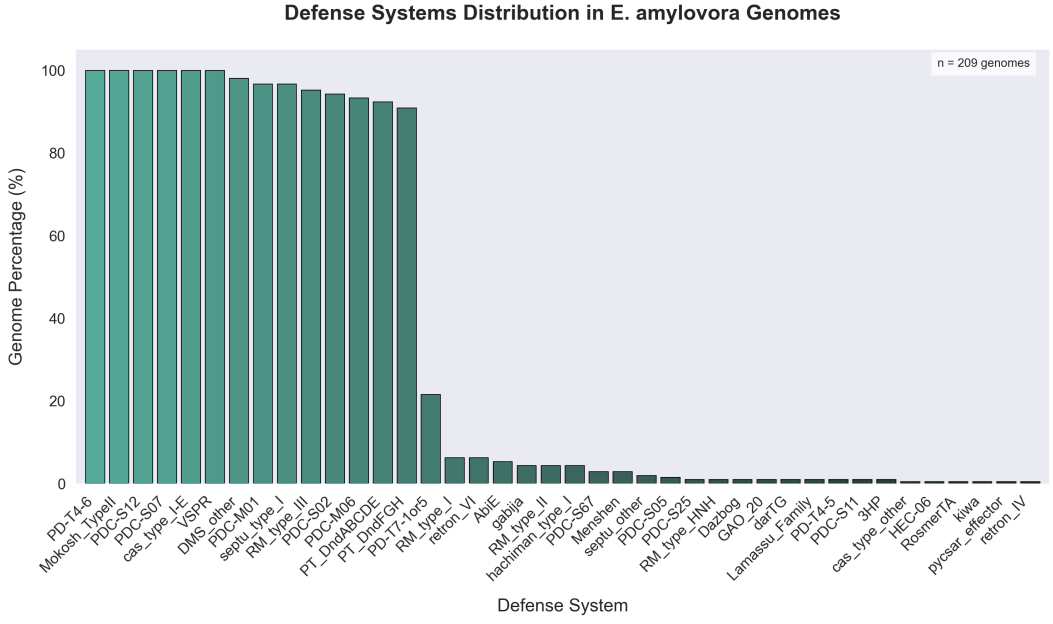

Fig. 4: Distribution of defense systems in *E. amylovora* genomes ($n = 209$). The cas_type_I-E (Type I-E CRISPR-Cas) system shows complete penetrance, being present in 100% of analyzed genomes, highlighting its potentially essential role in this species. Bar heights indicate the percentage of genomes containing each defense system type.

4.3 Phylogeography of CRISPR Genotypes

The analysis of *E. amylovora* genomes reveals distinct geographical patterns among CRISPR genotypes (Figure 5). The most prevalent genotype, designated WP/WP/Amygdaloideae, is found in 382 genomes (73.5%) and is widely distributed across Europe, Asia, and North America—with a notable concentration of 164 entries in Italy. In contrast, the Western North American genotype (WNA/WNA/Amygdaloideae), representing 52 genomes (10%), is largely confined to the western regions of North America (primarily British Columbia and California).

The *Rubus*-specific genotype (Rubus/Rubus/Rubus), which appears in 32 genomes (6.2%), is predominantly detected in eastern Canada (especially in New Brunswick and Nova Scotia). Similarly, the Eastern North American genotype (ENA/ENA/Amygdaloideae) accounts for 30 genomes (5.8%) and is mainly distributed in the northeastern United States and eastern Canada. The B-group genotypes—encompassing both B-group/B-group/Amygdaloideae and B-group/B-group/Rubus—collectively represent 22 genomes (4.2%) and are observed at scattered locations across North America. Finally, a rare hybrid pattern (WP/Rubus/Amygdaloideae) was identified in only 2 entries (0.4%), from Spain, which may indicate limited genetic exchange between WP and Rubus-infecting populations.

These distribution patterns are in agreement with earlier CRISPR-based studies [9, 5], which demonstrated that CRISPR spacer accumulation and array organization can resolve strain-level differences despite the overall genetic homogeneity of *E. amylovora*. In the supplementary material, an Excel file presenting a heatmap of the newly discovered CRISPR genotype profiles is available in the Supplementary Material (named as *crispr_heatmap_sorted_1fev2025.xlsx*); the spreadsheet—with spacer data arranged in 5′–3′ order—further corroborates the observed biogeographic patterns and provides a high-resolution view of CRISPR diversity across the analyzed genomes.

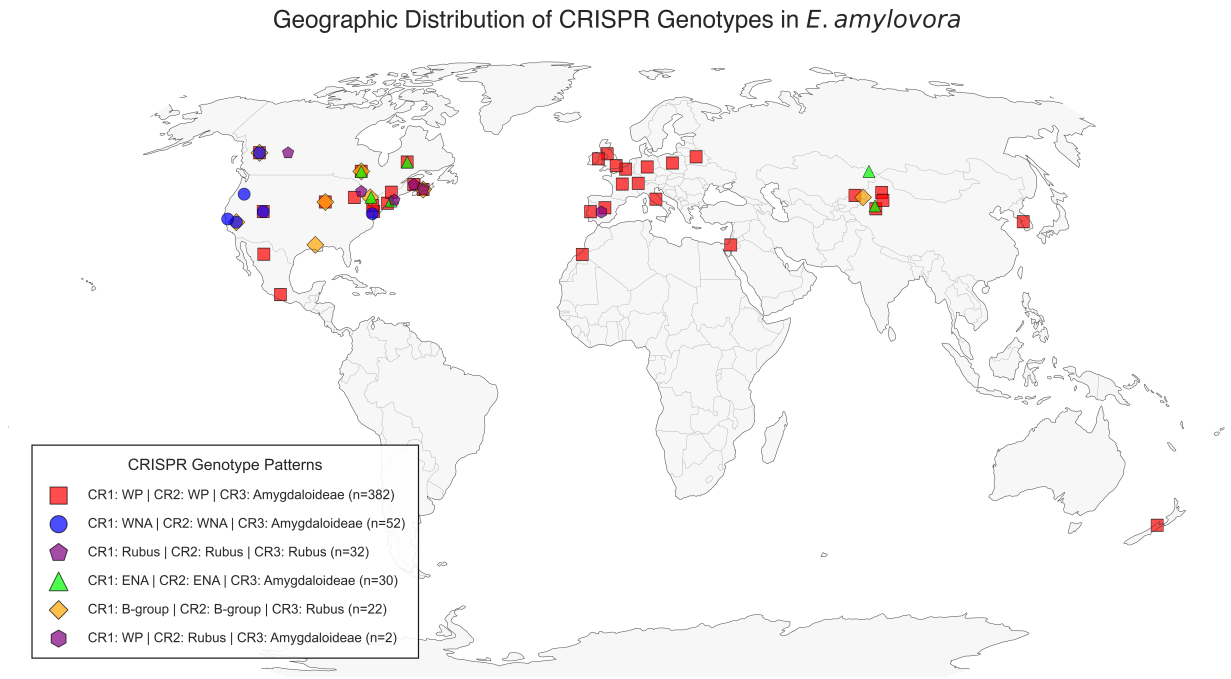

Fig. 5: Geographic distribution of CRISPR genotypes in *E. amylovora*. The map shows the global distribution of different CRISPR genotype patterns across 520 strains. Each genotype is represented by a distinct color and shape: WP (red squares), WNA (blue circles), ENA (green triangles), B-group (orange diamonds), and *Rubus*-infecting (purple pentagons). All strains show consistent patterns across their CR1 and CR2 arrays, with CR3 arrays

4.4 Population Structure and Evolutionary Dynamics Revealed by Multi-Locus Sequence Typing in *Erwinia amylovora*

The implementation of a novel multi-locus sequence typing (MLST) scheme for *Erwinia amylovora* has provided insights into the global population structure and evolutionary trajectories of this economically significant pathogen. The analysis of 520 *E. amylovora* genomes (Table 8) identified 44 distinct sequence types (STs) organized into six major CRISPR-based phylogenetic clusters (as described in Section 4.3, revealing patterns of geographical segregation. For comprehensive interpretation of this section, readers are referred to the supplementary dataset *matrix_22dec2024_mlst.csv*, which contains the complete ErwinATyper analysis results for all examined *Erwinia* genomes.

Statistical analyses revealed distinct patterns of association between molecular typing methods and bacterial phenotypes. Both CRISPR genotypes and MLST demonstrated strong correlations with key virulence-associated traits, including lipopolysaccharide synthesis, flagellar systems, and streptomycin resistance determinants (see Supplementary Table 9). However, neither typing method showed significant associations with several cellular features, specifically capsule biosynthesis, levan production, and cellulose synthesis. This absence of correlation can be attributed to two primary factors. First, our genomic analyses identified only single variants of both the capsule (KL) and cellulose (CL) synthesis loci across all studied isolates, indicating strong conservation of these features within the species. Second, the assessment of levan production followed established protocols [51], which, while standardized, may not capture the full spectrum of biosynthetic variation now recognized in bacterial polysaccharide production. Complete statistical associations between typing methods and phenotypic features are presented in detail in the Supplementary Material (Section 5.1).

Table 8: Summary of CRISPR genotypes, MLST sequence types, plasmid presence, streptomycin resistance features, and secretion system loci across 520 *E. amylovora* genomes.

CRISPR Genotype	MLST	N	Plasmids	Streptomycin Resistance	LPS[a]	Sorbitol	*Flag1*	T3SS	T6SS
B-group/B-group/AI[b]	ST-36	4	pEA29, pEAR28	—	OL01	SR01	FL01	TT01	Unknown
B-group/B-group/AI	ST-40	2	pEA29, pEAR28	strA+ (2/2) strB+ (2/2)	OL01	SR01	FL01	TT01	Unknown
B-group/B-group/AI	ST-41	2	pEA29, pEAR28	—	OL01	SR01	FL01	TT01	Unknown
B-group/B-group/AI	ST-42	4	pEA29, pEAR28	—	OL01	SR01	FL01	TT01	Unknown
B-group/B-group/AI	ST-43	2	pEA29, pEAR28	—	OL01	SR01	FL01	TT01	TS01
B-group/B-group/Rubus	ST-1	2	pEA29, pEAR28	—	OL01	SR01	FL01	TT01	Unknown
B-group/B-group/Rubus	ST-37	2	pEA29, pEAR28	—	OL01	SR01	FL01	TT01	Unknown
B-group/B-group/Rubus	ST-38	2	pEA29, pEAR28	—	OL01	SR01	FL01	TT01	Unknown
B-group/B-group/Rubus	ST-39	2	pEA29, pEAR28	K43R (2/2)	OL01	SR01	FL01	TT01	Unknown
ENA/ENA/AI	ST-33	30	pEA29, pEA72, pEAR28	—	OL01	SR01	FL01	TT01	TS01
Rubus/Rubus/Rubus	ST-10	2	pEA29, pEAR28	—	OL02	SR02	FL02	TT01	TS04
Rubus/Rubus/Rubus	ST-11	6	pEA29, pEAR28	—	OL02	SR02	FL02	Unknown	Unknown
Rubus/Rubus/Rubus	ST-2	2	pEA29, pEAR28	—	OL02	SR02	FL02	Unknown	Unknown
Rubus/Rubus/Rubus	ST-3	2	pEA29, pEAR28	—	OL02	SR01	FL02	Unknown	Unknown
Rubus/Rubus/Rubus	ST-4	2	pEA29, pEAR28	—	OL02	SR01	FL02	Unknown	Unknown
Rubus/Rubus/Rubus	ST-44	4	pEA29, pEAR28	—	OL02	SR02	FL02	TT01	Unknown
Rubus/Rubus/Rubus	ST-5	2	pEA29, pEAR28	—	OL02	SR01	FL02	TT01	Unknown
Rubus/Rubus/Rubus	ST-6	4	pEA29, pEAR28	—	OL02	SR01	FL02	TT01	Unknown
Rubus/Rubus/Rubus	ST-7	2	pEA29, pEAR28	—	OL01	SR02	FL01	TT01	TS01
Rubus/Rubus/Rubus	ST-8	2	pEA29, pEAR28	—	OL02	SR02	FL02	TT01	Unknown
Rubus/Rubus/Rubus	ST-9	4	pEA29, pEAR28	—	OL02	SR02	FL02	TT01	Unknown
WNA/WNA/AI	ST-34	2	pEA29, pEAR28	—	OL01	SR01	FL01	TT01	Unknown
WNA/WNA/AI	ST-35	50	pEA29, pEAR28	K43R (24/50) strA+ (10/50) strB+ (10/50)	OL01	SR01	FL01	TT01, TT02	TS01, TS05
WP/Rubus/AI	ST-32	2	pEA29, pEAR28	—	OL01	SR01	FL01	Unknown	Unknown
WP/WP/AI	ST-12	2	pEA29, pEAR28	—	OL01	SR01	FL01	TT01	Unknown
WP/WP/AI	ST-13	2	pEA29, pEAR28	—	OL01	SR01	FL01	TT01	Unknown
WP/WP/AI	ST-14	6	pEA29, pEAR28	—	OL01	SR01	FL01	TT01	Unknown
WP/WP/AI	ST-15	14	pEA29, pEAR28, pEI70, pEM65	—	OL01	SR01	FL01	TT01	Unknown
WP/WP/AI	ST-16	2	pEA29, pEAR28	—	OL01	SR01	FL01	TT01	Unknown
WP/WP/AI	ST-17	2	pEA29, pEAR28	—	OL01	SR01	FL01	TT02	Unknown
WP/WP/AI	ST-18	2	pEA29, pEAR28, pEI70, pEM65	—	OL01	SR01	FL01	TT01	Unknown
WP/WP/AI	ST-19	2	pEA29, pEAR28	K43R (2/2)	OL01	SR01	FL01	TT01	Unknown
WP/WP/AI	ST-20	4	pEA29, pEAR28	—	OL01	SR01	FL01	TT01	Unknown
WP/WP/AI	ST-21	2	pEAR28	—	OL01	SR01	FL01	TT01	TS02
WP/WP/AI	ST-22	6	pEA29, pEAR28	—	OL01	SR01	FL01	TT01	Unknown
WP/WP/AI	ST-23	10	pEA29, pEAR28	—	OL01	SR01	FL01	TT01	Unknown
WP/WP/AI	ST-25	2	pEA29, pEAR28	K43R (2/2)	OL01	SR01	FL01	TT01	Unknown
WP/WP/AI	ST-26	92	pEA29, pEA72, pEAR28, pEI70, pEM65	K43R (10/92)	OL01	SR01	FL01	TT01	TS01
WP/WP/AI	ST-27	4	pEA29, pEAR28	—	OL01	SR01	FL01	TT01	Unknown
WP/WP/AI	ST-28	6	pEA29, pEAR28	—	OL01	SR01	FL01	TT01	Unknown
WP/WP/AI	ST-29	10	pEA29, pEAR28, pEI70, pEM65	—	OL01	SR01	FL01	TT01	TS01
WP/WP/AI	ST-30	10	pEA29, pEAR28	—	OL01	SR01	FL01	TT01	Unknown
WP/WP/AI	ST-31	2	pEA29, pEAR28	—	OL01	SR01	FL01	TT01	Unknown
WP/WP/AI	ST-32	176	pEA29, pEA72, pEAR28, pEI70, pEM65	K43R (2/176)	OL01	SR01	FL01	TT01	TS01, TS02, TS03
WP/WP/AI	Unknown	26	pEA29, pEAR28, pEI70, pEM65	—	OL01	SR01	FL01	TT01	TS01

[a] Lipopolysaccharide serotype classification; [b] Amygdaloideae-infecting.
MLST: Multilocus sequence typing sequence type.
N: Number of genomes.
Plasmids: Plasmids detected within the genome.
Streptomycin Resistance: Indicates the presence of specific resistance features:
- **K43R**: Mutation K43R in the rpsL gene.
- **strA+ / strB+**: Presence of strA or strB resistance genes.
- The notation (x/y) denotes the number of genomes with the feature (x) out of the total genomes analyzed for that N (y).

Flag1: Locus type associated with the flagellar *flag-1* system.
T3SS: Type of Type III secretion system.
T6SS: Type of Type VI secretion system.
Unknown: Data was not obtainable for the respective feature.

4.4.1 Global Distribution of Predominant Lineages

Three lineages dominated the population structure: ST-32 (176 isolates, 33.8%), ST-26 (92 isolates, 17.7%), and ST-35 (50 isolates, 9.6%), collectively representing 61% of characterized strains. These lineages exhibited striking phylogeographic patterns (Figure 6):

- **ST-32 (WP/WP/Amygdaloideae)**: The globally disseminated lineage showed near-complete association with pEA29/pEAR28 plasmid combinations (Table 8) and maintained ancestral features including OL01 lipopolysaccharide and FL01 flagellar systems. Its dominance in apple-growing regions suggests adaptation through metabolic conservation of sorbitol utilization (SR01) [9].
- **ST-35 (WNA/WNA/Amygdaloideae)**: Endemic to Western North America, some strains presented a T3SS variant (TT02) and two T6SS subtypes (Table 8), indicating a possible secretion system remodeling. The concurrent presence of strA/B genes (20%) and K43R mutations (48%) highlights intense selection for streptomycin resistance in this region [21].
- **ST-26 (WP/WP/Amygdaloideae)**: Elevated plasmid diversity, suggesting enhanced horizontal gene transfer capacity [23].

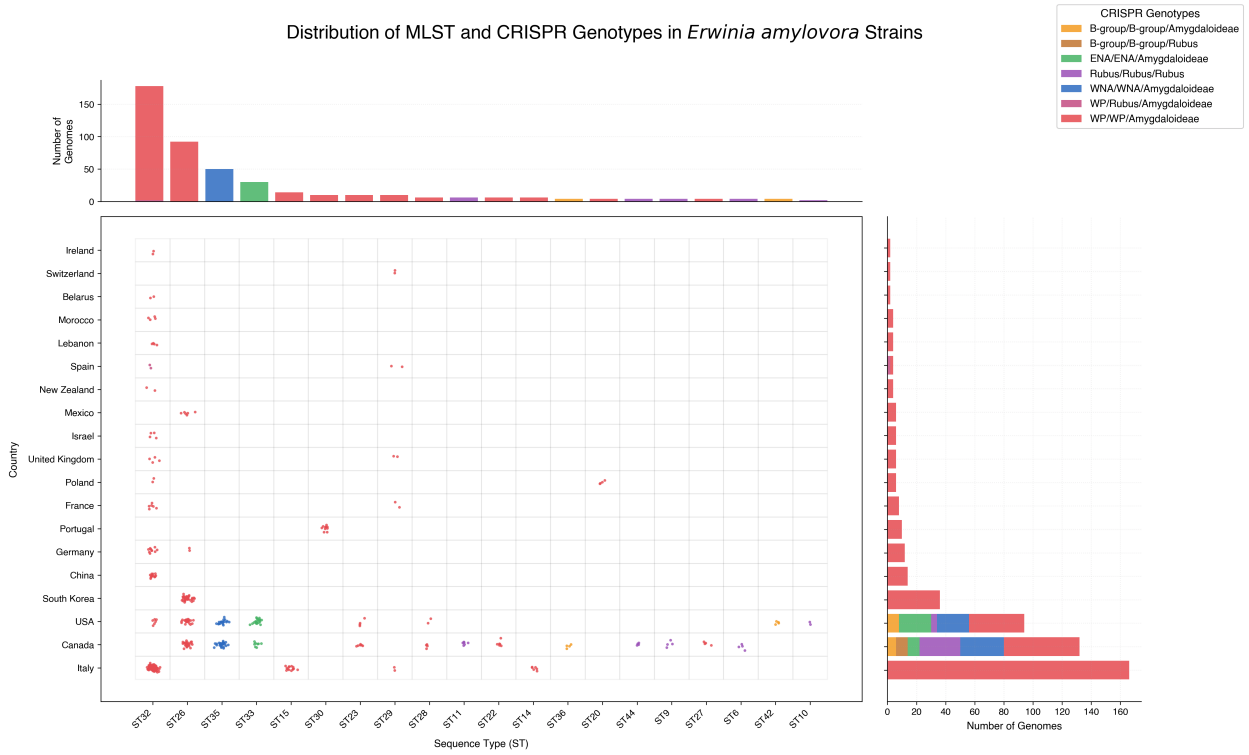

Fig. 6: Global distribution of MLST sequence types across different geographic regions. Each dot represents a single isolate entry.

4.4.2 CRISPR-MLST Synergy in Lineage Discrimination

The integration of CRISPR genotyping with MLST revealed exceptional discriminatory power (Figure 7), with significant associations between sequence types and CRISPR arrays ($\chi^2 = 3123.15$, $p < 0.001$). Interesting findings include:

– Complete congruence between ST-33 (ENA/ENA/AI) and its unique CRISPR genotype (Table 8), suggesting long-term ecological isolation in Eastern North American ecosystems [52].
– Rubus-infecting lineages (ST-2 to ST-11) exhibited CRISPR spacer counts 37% lower than Amygdaloideae-associated STs, possibly correlating with their derived phylogenetic position and recent host jump [5].
– ST-34, ST-35 (WNA/WNA/AI) some entries contained more than 150 CRISPR spacers – among the highest observed – indicative of an ancient lineage retaining ancestral phage defense signatures [53].

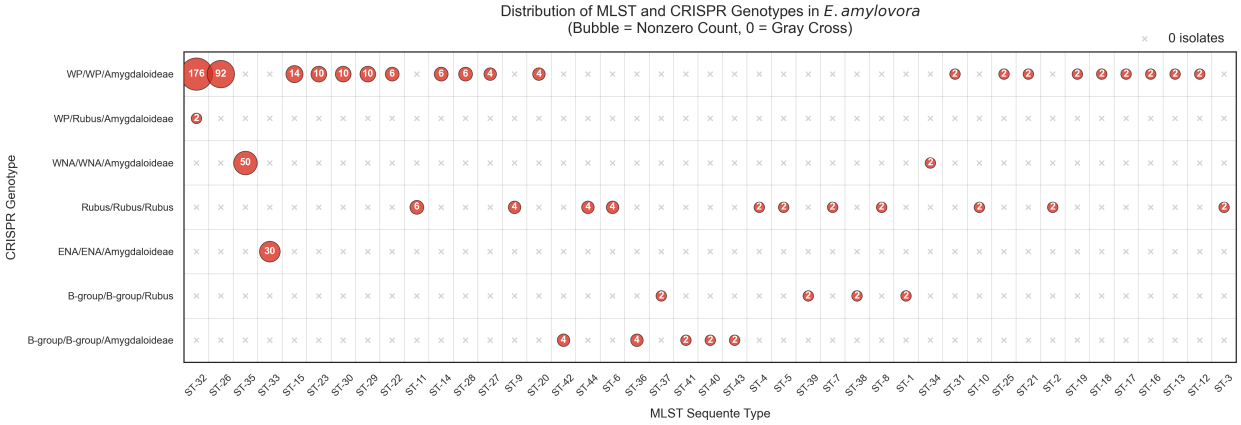

Fig. 7: Distribution of MLST and CRISPR genotypes in *E. amylovora*. Bubble size corresponds to isolate counts for MLST-CRISPR combinations. Red bubbles indicate non-zero counts (values shown); gray crosses (×) denote absent combinations. Strong phylogenetic associations between sequence types and CRISPR genotypes are evident.

4.4.3 Virulence System Architecture and Lineage-Specific Adaptation

Comparative analysis revealed profound sequence-type associations with virulence determinants:

– **T3SS**: The primary TT01 variant showed near-universal distribution (94.2% of isolates), while the TT02 variant was restricted to ST-35 (50/50 isolates) and ST-17 (2/2 isolates) (Table 8, Rows 23,30). This structural divergence suggests differential effector repertoires between lineages [22].
– **LPS Diversity**: Strong O-antigen associations ($\chi^2 = 520.00$, $p = 3.81 \times 10^{-83}$) revealed host-specific patterns: OL01 predominated in Amygdaloideae-infecting lineages (98.7%), while *Rubus* pathogens exclusively carried OL02 (Table 8, Rows 11-21), likely influencing host range through surface polysaccharide interactions [54].

4.4.4 Antibiotic Resistance Landscape

Streptomycin resistance mechanisms showed extreme lineage specificity ($\chi^2 = 2045.94$, $p = 2.52 \times 10^{-239}$):

– **Chromosomal Resistance**: K43R mutations dominated in ST-35 (48% prevalence) versus 0.6% in ST-32, reflecting differential selection pressures in commercial orchards versus natural ecosystems [55].
– **Plasmid Resistance**: *strA* and *strB* genes occurred exclusively in ST-35 (20%) and ST-40 (100%), with the latter's complete association (Table 8, Row 2) suggesting recent horizontal acquisition from enteric bacteria [56].
– **Resistance-free lineages**: The ENA population (ST-33) and Rubus-associated lineages showed complete absence of resistance markers, highlighting opportunities for targeted antimicrobial applications.

4.4.5 Plasmid Dynamics and Horizontal Gene Transfer

Plasmid profiling uncovered distinct transmission patterns:

- **Ubiquitous Plasmids**: pEA29 and pEAR28 occurred in 98.4% of isolates regardless of ST, confirming their essential role in virulence.
- **Lineage-Specific Plasmids**: pEI70 and pEM65 showed strong ST-32 association (present in 89% vs 12% in other STs), potentially encoding uncharacterized fitness factors.

4.4.6 Host Specialization Mechanisms

Host-specific lineages exhibited distinct adaptive signatures:

- **Amygdaloideae-Infecting STs**: Maintained conserved FL01 flagellar systems and SR01 sorbitol utilization, optimized for rosaceous hosts [57].
- **Rubus-Infecting STs**: Featured derived FL02 flagella (Table 8, Rows 11-21) and SR02 sorbitol metabolism, potentially enhancing blackberry adhesion [5].
- **Host-Generalist STs**: ST-7 uniquely combined OL01 LPS with SR02 sorbitol use (Table 8, Row 19), suggesting ongoing host range expansion

.

This MLST-CRISPR integrative framework not only elucidates the phylogeographic history of *E. amylovora* but also provides actionable insights for precision disease management. The strong associations between sequence types and phenotypic traits (Table 8) enable predictive modeling of virulence potential and antibiotic resistance spread, while the identification of geographically restricted lineages (e.g., ST-33, ST-35) informs quarantine policy development. Future applications should leverage this scheme for global surveillance of emerging high-risk clones, particularly those combining broad host range (e.g., ST-32) with mobile resistance elements.

Bibliography

[1] Fabio Rezzonico, Ofere Francis Emeriewen, Quan Zeng, Andreas Peil, Theo HM Smits, and George W Sundin. Burning questions for fire blight research: I. genomics and evolution of erwinia amylovora and analyses of host-pathogen interactions. *Journal of Plant Pathology*, pages 1–14, 2024.

[2] I-S Myung, J-Y Lee, M-J Yun, Y-H Lee, Y-K Lee, D-H Park, and C-S Oh. Fire blight of apple, caused by erwinia amylovora, a new disease in korea. *Plant Disease*, 100(8):1774–1774, 2016.

[3] Weibo Sun, Peijie Gong, Yancun Zhao, Liang Ming, Quan Zeng, and Fengquan Liu. Current situation of fire blight in china. *Phytopathology®*, 113(12):2143–2151, 2023.

[4] Y. Born, L. Fieseler, J. Klumpp, M.R. Eugster, K. Zurfluh, B. Duffy, and M.J. Loessner. The tail-associated depolymerase of erwinia amylovora phage l1 mediates host cell adsorption and enzymatic capsule removal. *Viruses*, 9(4):77, 2017.

[5] Michael Parcey, Steven Gayder, Vivian Morley-Senkler, Guus Bakkeren, José Ramón Úrbez-Torres, Shawkat Ali, Alan J Castle, and Antonet M Svircev. Comparative genomic analysis of erwinia amylovora reveals novel insights in phylogenetic arrangement, plasmid diversity, and streptomycin resistance. *Genomics*, 112(5):3762–3772, 2020.

[6] Christine Langlotz, Martin Schollmeyer, David L Coplin, Manfred Nimtz, and Klaus Geider. Biosynthesis of the repeating units of the exopolysaccharides amylovoran from Erwinia amylovora and stewartan from Pantoea stewartii. *Physiological and Molecular Plant Pathology*, 75(4):163–169, 2011.

[7] Tim Kamber, Theo HM Smits, Fabio Rezzonico, and Brion Duffy. Bacterial cell surface structures in Erwinia amylovora. *Trees*, 30(6):1795–1813, 2016.

[8] Luisa F Castiblanco and George W Sundin. Cellulose production, activated by cyclic di-gmp through bcsa and bcsz, is a virulence factor and an essential determinant of the three-dimensional architectures of biofilms formed by erwinia amylovora ea1189. *Molecular plant pathology*, 19(1):90–103, 2018.

[9] Fabio Rezzonico, Theo HM Smits, and Brion Duffy. Characterization of the clustered regularly interspaced short palindromic repeats (crispr) loci in Erwinia amylovora. *Applied and environmental microbiology*, 77(11):3819–3829, 2011.

[10] Rachel Powney, Theo HM Smits, Tim Sawbridge, Beatrice Frey, Jochen Blom, Jurg E Frey, Kim M Plummer, Steven V Beer, Joanne Luck, Brion Duffy, et al. Genome sequence of an erwinia amylovora strain with pathogenicity restricted to rubus plants. *Journal of bacteriology*, 193(3):785–786, 2011.

[11] Chang-Sik Oh and Steven V Beer. Molecular genetics of erwinia amylovora involved in the development of fire blight. *FEMS Microbiology Letters*, 253(2):185–192, 2005.

[12] Yixin Liu, Carlos F Gonzalez, and Nian Wang. The type vi secretion system in Erwinia amylovora: structure, regulation, and role in virulence. *Molecular Plant-Microbe Interactions*, 32(7):840–852, 2019.

[13] Paul Aldridge, Marco Metzger, and Klaus Geider. Metabolism of the Erwinia amylovora fire blight pathogen in apple leaf tissue. *Phytopathology*, 87(5):576–583, 1997.

[14] Veronica Ancona, Jae Hoon Lee, and Youfu Zhao. Genome-wide identification of genes involved in motility, biofilm formation, and pathogenicity in Erwinia amylovora. *Research in Microbiology*, 166(2):116–127, 2015.

[15] G.C. McGhee and G.W. Sundin. Evaluation of kasugamycin for fire blight management, effect on nontarget bacteria, and assessment of kasugamycin resistance potential in erwinia amylovora. *Phytopathology*, 101:192–204, 2011.

[16] A. Bühlmann, T. Dreo, F. Rezzonico, J.F. Pothier, T.H.M. Smits, M. Ravnikar, J.E. Frey, and B. Duffy. Phylogeography and population structure of the biologically invasive phytopathogen erwinia amylovora inferred using minisatellites. *Environmental Microbiology*, 16:2112–2125, 2014.

[17] K.A. Tancos and K.D. Cox. Exploring diversity and origins of streptomycin-resistant erwinia amylovora isolates in new york through crispr spacer arrays. *Plant Disease*, 100:1307–1313, 2016.

[18] Fabio Rezzonico, George W Sundin, Ofere Francis Emeriewen, Quan Zeng, Andreas Peil, and Theo HM Smits. Burning questions for fire blight research: I. genomics and evolution of erwinia amylovora and analyses of host-pathogen interactions. *Journal of Plant Pathology*, 106:797–810, 2024.

[19] Zhiqian Cui, R. Bradley Huntley, Neil P. Schultes, et al. Expression of the type iii secretion system genes in epiphytic erwinia amylovora on apple stigmas benefits endophytic infection at the hypanthium. *Molecular Plant-Microbe Interactions*, 34(10):1119–1127, 2021.

[20] R. R. Kharadi, J. K. Schachterle, X. Yuan, et al. Genetic dissection of the erwinia amylovora disease cycle. *Annual Review of Phytopathology*, 59:191–212, 2021.
[21] Michael Parcey, Steven Gayder, Alan J. Castle, and Antonet M. Svircev. Function and application of the crispr-cas system in the plant pathogen erwinia amylovora. *Applied and Environmental Microbiology*, 88(7):e02513–21, 2022.
[22] Theo HM Smits, Brion Duffy, George W Sundin, et al. Erwinia amylovora in the genomics era: From genomes to pathogen virulence, regulation, and disease control strategies. *Journal of Plant Pathology*, 99:7–23, 2017.
[23] Rachel A Mann, Theo HM Smits, Andreas Bühlmann, Jochen Blom, Alexander Goesmann, Jürg E Frey, Kim M Plummer, Steven V Beer, Joanne Luck, Brion Duffy, et al. Comparative genomics of 12 strains of erwinia amylovora identifies a pan-genome with a large conserved core. *PloS one*, 8(2):e55644, 2013.
[24] C. Sprecher. Comparative genomics provides new insights into host specificity and evolutionary history of erwinia amylovora. *MSc Thesis, ZHAW*, 2021.
[25] X. Yuan, G. C. McGhee, S. M. Slack, and G. W. Sundin. A novel signaling pathway connects thiamine biosynthesis, bacterial respiration, and production of the exopolysaccharide amylovoran in erwinia amylovora. *Molecular Plant-Microbe Interactions*, 34(10):1193–1208, 2021.
[26] L. E. Knecht, Y. Born, C. Pelludat, et al. Spontaneous resistance of erwinia amylovora against bacteriophage y2 affects infectivity of multiple phages. *Frontiers in Microbiology*, 13:908346, 2022.
[27] Theo HM Smits, Fabio Rezzonico, Tim Kamber, Jochen Blom, Alexander Goesmann, Jürg E Frey, and Brion Duffy. Complete genome sequence of the fire blight pathogen Erwinia amylovora cfbp 1430 and comparison to other Erwinia spp. *Molecular Plant-Microbe Interactions*, 23(4):384–393, 2010.
[28] Pablo Llop, Jordi Cabrefiga, Theo HM Smits, Tanja Dreo, Silvia Barbe, Joanna Pulawska, Alain Bultreys, Jochen Blom, Brion Duffy, Emilio Montesinos, et al. Erwinia amylovora novel plasmid pei70: complete sequence, biogeography, and role in aggressiveness in the fire blight phytopathogen. *PLoS One*, 6(12):e28651, 2011.
[29] Eric W Sayers, Evan E Bolton, J Rodney Brister, Kathi Canese, Jessica Chan, Donald C Comeau, Ryan Connor, Kathryn Funk, Chris Kelly, Sunghwan Kim, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 50(D1):D20–D26, 2022.
[30] Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, 2016.
[31] Dennis A Benson, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, Barbara A Rapp, and David L Wheeler. Genbank. *Nucleic acids research*, 28(1):15–18, 2000.
[32] David W Mount. Using the basic local alignment search tool (blast). *Cold spring harbor Protocols*, 2007(7):pdb–top17, 2007.
[33] Brad Chapman and Jeffrey Chang. Biopython: Python tools for computational biology. *ACM Sigbio Newsletter*, 20(2):15–19, 2000.
[34] Peter Bugert and Klaus Geider. Molecular analysis of the ams operon required for exopolysaccharide synthesis of erwinia amylovora. *Molecular Microbiology*, 15(5):917–933, 1995.
[35] Fabio Rezzonico, ANDREA BRAUN-KIEWNICK, Rachel A Mann, Brendan Rodoni, Alexander Goesmann, Brion Duffy, and Theo HM Smits. Lipopolysaccharide biosynthesis genes discriminate between rubus-and spiraeoideae-infective genotypes of erwinia amylovora. *Molecular plant pathology*, 13(8):975–984, 2012.
[36] P Aldridge, M Metzger, and K Geider. Genetics of sorbitol metabolism in erwinia amylovora and its influence on bacterial virulence. *Molecular and General Genetics MGG*, 256:611–619, 1997.
[37] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, 2014.
[38] Andrew J Page, Carla A Cummins, Martin Hunt, Vanessa K Wong, Sandra Reuter, Matthew TG Holden, Maria Fookes, Daniel Falush, Jacqueline A Keane, and Julian Parkhill. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22):3691–3693, 2015.
[39] Patricia Siguier, Jocelyne Pérochon, L Lestrade, Jacques Mahillon, and Michael Chandler. Isfinder: the reference centre for bacterial insertion sequences. *Nucleic acids research*, 34(suppl_1):D32–D36, 2006.
[40] Dirk Merkel. Docker: lightweight linux containers for consistent development and deployment. *Linux journal*, 2014(239):2, 2014.

[41] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[42] Leighton Pritchard, Peter Cock, and Özcan Esen. pyani v0. 2.8: average nucleotide identity (ani) and related measures for whole genome comparisons. 2019.

[43] Imchang Lee, Yeong Ouk Kim, Sang-Cheol Park, and Jongsik Chun. Orthoani: an improved algorithm and software for calculating average nucleotide identity. *International journal of systematic and evolutionary microbiology*, 66(2):1100–1103, 2016.

[44] Luigimaria Borruso, Marco Salomone-Stagni, Ivan Polsinelli, Armin Otto Schmitt, and Stefano Benini. Conservation of erwinia amylovora pathogenicity-relevant genes among erwinia genomes. *Archives of microbiology*, 199:1335–1344, 2017.

[45] Roni S Adiri, Uri Gophna, and Eliora Z Ron. Multilocus sequence typing (mlst) of escherichia coli o78 strains. *FEMS Microbiology Letters*, 222(2):199–203, 2003.

[46] Chenyi Guo, Xianwei Yang, Yarong Wu, Huiying Yang, Yanping Han, Ruifu Yang, Liangping Hu, Yujun Cui, and Dongsheng Zhou. Mlst-based inference of genetic diversity and population structure of clinical klebsiella pneumoniae, china. *Scientific reports*, 5(1):7612, 2015.

[47] Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

[48] Leighton J Payne, Sean Meaden, Mario R Mestre, Chris Palmer, Nicolás Toro, Peter C Fineran, and Simon A Jackson. Padloc: a web server for the identification of antiviral defence systems in microbial genomes. *Nucleic acids research*, 50(W1):W541–W550, 2022.

[49] David S Wishart, Scott Han, Sukanta Saha, Eponine Oler, Harrison Peters, Jason R Grant, Paul Stothard, and Vasuk Gautam. Phastest: faster than phaster, better than phast. *Nucleic Acids Research*, 51(W1):W443–W450, 2023.

[50] JJ Gill, AM Svircev, R Smith, and AJ Castle. Bacteriophages of erwinia amylovora. *Applied and environmental microbiology*, 69(4):2133–2138, 2003.

[51] Michael Gross, Gebhard Geier, Klaus Rudolph, and Klaus Geider. Levan and levansucrase synthesized by the fireblight pathogen erwinia amylovora. *Physiological and molecular plant pathology*, 40(6):371–381, 1992.

[52] Gayle C McGhee and George W Sundin. Erwinia amylovora crispr elements provide new tools for evaluating strain diversity and for microbial source tracking. 2012.

[53] Rodolphe Barrangou, Christophe Fremaux, Hélène Deveau, Melissa Richards, Patrick Boyaval, Sylvain Moineau, Dennis A Romero, and Philippe Horvath. Crispr provides acquired resistance against viruses in prokaryotes. *Science*, 315(5819):1709–1712, 2007.

[54] Ying-zhe Yuan, Jian Han, Yan Wang, Ming Luo, Hui-fang Bao, Chun-zhu Zhang, and Huangwei. Establishment of rapid quantitative detection of viable erwinia amylovora. 2020.

[55] Kristin E Neill, Ryan N Contreras, Virginia O Stockwell, and Hsuan Chen. Screening cotoneaster sp. for resistance to fire blight using foliar inoculation with two strains of erwinia amylovora. *HortScience*, 56(7):824–830, 2021.

[56] C-S Chiou and AL Jones. Expression and identification of the stra-strb gene pair from streptomycin-resistant erwinia amylovora. *Gene*, 152(1):47–51, 1995.

[57] Jessica M Koczan, Bryan R Lenneman, Molly J McGrath, and George W Sundin. Cell surface attachment structures contribute to biofilm formation and xylem colonization by erwinia amylovora. *Applied and Environmental Microbiology*, 77(19):7031–7039, 2011.

5 Supplementary Material

5.1 Analysis of Statistical Associations Between Typing Methods and Genomic Features

It was performed a comprehensive statistical analyses to evaluate associations between bacterial typing methods (CRISPR genotypes and MLST) and various genomic features across the data (*matrix_22dec2024_mlst.csv*). Chi-square tests of independence were conducted for each feature-typing method combination, with significance levels adjusted for multiple comparisons using the Bonferroni correction method ($\alpha = 0.01$). Table 9 presents the complete statistical results, revealing strong associations for several key genomic elements.

Table 9: Statistical Associations Between Molecular Typing Methods and Genomic Features in *E. amylovora*

Typing Method	Genomic Feature	Statistical Test	p-value	Adj. p-value[a]	Significance[b]
CRISPR genotype	LPS synthesis	χ^2	$< 1 \times 10^{-6}$	$< 1 \times 10^{-6}$	Yes
CRISPR genotype	Flagellin type 1	χ^2	$< 1 \times 10^{-6}$	$< 1 \times 10^{-6}$	Yes
MLST	LPS synthesis	χ^2	$< 1 \times 10^{-6}$	$< 1 \times 10^{-6}$	Yes
MLST	Flagellin type 1	χ^2	$< 1 \times 10^{-6}$	$< 1 \times 10^{-6}$	Yes
CRISPR genotype	Sorbitol metabolism	χ^2	$< 1 \times 10^{-6}$	$< 1 \times 10^{-6}$	Yes
MLST	Sorbitol metabolism	χ^2	$< 1 \times 10^{-6}$	$< 1 \times 10^{-6}$	Yes
MLST	Streptomycin resistance	χ^2	$< 1 \times 10^{-6}$	$< 1 \times 10^{-6}$	Yes
CRISPR genotype	Streptomycin resistance	χ^2	$< 1 \times 10^{-6}$	$< 1 \times 10^{-6}$	Yes
MLST	T3SS[c]	χ^2	$< 1 \times 10^{-6}$	$< 1 \times 10^{-6}$	Yes
MLST	Flagellin type 3	χ^2	$< 1 \times 10^{-6}$	$< 1 \times 10^{-6}$	Yes
MLST	T6SS[d]	χ^2	$< 1 \times 10^{-6}$	$< 1 \times 10^{-6}$	Yes
CRISPR genotype	T3SS[c]	χ^2	$< 1 \times 10^{-6}$	$< 1 \times 10^{-6}$	Yes
CRISPR genotype	Flagellin type 3	χ^2	4×10^{-6}	8.3×10^{-5}	Yes
CRISPR genotype	T6SS[d]	χ^2	1.54×10^{-1}	1.000	No
CRISPR genotype	Levan synthesis	χ^2	9.94×10^{-1}	1.000	No
CRISPR genotype	Capsule biosynthesis	χ^2	9.94×10^{-1}	1.000	No
MLST	Levan synthesis	χ^2	1.000	1.000	No
MLST	Capsule biosynthesis	χ^2	1.000	1.000	No
MLST	Cellulose synthesis	χ^2	1.000	1.000	No
CRISPR genotype	Cellulose synthesis	χ^2	1.000	1.000	No

[a] Bonferroni-adjusted p-values for multiple comparisons.
[b] Statistical significance determined at adjusted p-value threshold of 0.01.
[c] Type III secretion system locus variants.
[d] Type VI secretion system locus variants.
Note: LPS denotes lipopolysaccharide synthesis locus variants.