INTEGRATIVE HIGH-THROUGHPUT SEQUENCING TECHNIQUES FOR MICROBIAL GENOMICS: FROM TARGETED SEQUENCING OF CLINICAL SAMPLES TO WHOLE GENOME SEQUENCING AND REAL-TIME PCR TEST DEVELOPMENT

Aleksander Benčič

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Dr. Tanja Dreo, National Institute of Biology, Ljubljana, Slovenia
Co-Supervisor: Dr. Alexandra Bogožalec Košir, National Institute of Biology, Ljubljana, Slovenia

Evaluation Board:
Assist. Prof. Dr. Mojca Milavec, Chair, National Institute of Biology and Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
Dr. Maja Zagorščak, Member, National Institute of Biology, Ljubljana, Slovenia
Dr. Joël F. Pothier, Member, Zürich University of Applied Sciences, Zürich, Switzerland

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Aleksander Benčič

INTEGRATIVE HIGH-THROUGHPUT SEQUENCING TECHNIQUES FOR MICROBIAL GENOMICS: FROM TARGETED SEQUENCING OF CLINICAL SAMPLES TO WHOLE GENOME SEQUENCING AND REAL-TIME PCR TEST DEVELOPMENT

Doctoral Dissertation

UPORABA RAZLIČNIH PRISTOPOV VISOKOZMOGLJIVEGA SEKVENCIRANJA V MIKROBNI GENOMIKI: OD TARČNEGA SEKVENCIRANJA KLINIČNIH VZORCEV DO SEKVENCIRANJA CELOTNIH GENOMOV IN RAZVOJA TESTOV PCR V REALNEM ČASU

Doktorska disertacija

Supervisor: Dr. Tanja Dreo

Co-Supervisor: Dr. Alexandra Bogožalec Košir

Ljubljana, Slovenia, November 2025

Acknowledgments

This thesis would not have been possible without the cooperation of the many excellent researchers from the Department of Biotechnology and Systems Biology at the National Institute of Biology (NIB) and from various other institutions. Research was funded by the Slovenian Research and Innovation Agency (contract no. P4-0165, Young Researchers Programme, CRP V4-2415 MapQuest), the Ministry of Agriculture, Forestry and Food of the Republic of Slovenia (Professional Tasks Programme in the Field of Plant Health, Euphresco projects 2018-A275, 2021-A383 and 2023-A454, and CRP V4-2415 MapQuest), the European Union Reference Laboratories – Bacteriology (co-funded by the Ministry of Agriculture, Forestry and Food of the Republic of Slovenia) and the European Cooperation in Science and Technology (EuroXanth COST Action CA16107). Some of the equipment used in this study was financed by the Metrology Institute of the Republic of Slovenia (MIRS), which received financial support from the European Regional Development Fund.

I would like to thank my supervisor, Tanja Dreo, for her guidance and support throughout my PhD studies, and my co-supervisor, Alexandra Bogožalec Košir, for her assistance, help, and advice. Thanks also to the members of the dissertation committee: Mojca Milavec, Joël F. Pothier, and Maja Zagorščak.

This thesis would never have been completed without the help of many other people. Special thanks go to members of the Bacteriology and Metrology working group, as well as those working in diagnostics for plant pathogenic bacteria, including Aleš, Janja, Lidija, Manca P., Neža, Pia, Špela P., Vesna, and Vladimir. I would also like to thank Mojca Juteršek for her advice on my PhD thesis, Ana Vučurović, Anja Pecman, and Veronika Bukvič for their help with nanopore sequencing, Živa R. for her help with bioinformatics, and my colleagues at the Marine Biology Station in Piran for introducing me to NIB. I would also like to thank the amazing researchers from other organisations who played an important role: Nataša Toplak and Simon Koren from Omega; Viktorija Tomić and Dane Lužnik from the University Clinic Golnik; and Steve Baeyen, Jolien Venneman, and Johan Van Vaerenbergh from ILVO. Special thanks go to all my current and former colleagues, especially my roommates Špela A. and Tjaša, Vesna for her common-sense wisdom, Katja, Zala, Manca K., and Nina K., who were my first lunch-break companions, Meta, Nina T., and Sara, for all the funny moments, and to members of the PDF for keeping me in shape.

I would also like to thank all my friends for their support during my PhD studies. Finally, my greatest thanks go to my family and parents, who have always supported me on my journey and enabled me to chase my dreams.

Abstract

The development of high-throughput sequencing (HTS) technologies has enabled us to gain a much deeper insight into the mysterious and wonderful world of bacteria. These exciting technologies can be used in a variety of ways, ranging from the study of entire bacterial communities to more specific studies of individual bacterial isolates. In this thesis, we examine different HTS-based approaches to studying bacteria, applying insights from our own case studies. These include targeted HTS for metagenomic studies, whole genome sequencing (WGS) to study plant-pathogenic bacteria, and a genome-informed approach for the development of new molecular methods to detect target bacteria with high specificity.

Targeted HTS is a common approach in metagenomics, where it can be used to determine microbiomes. However, the results are often affected by the way samples are processed before sequencing, with the DNA extraction method being especially important. In our study, we examined the impact of three different DNA extraction methods on the results of microbiome and resistome determination using sputum samples as an example. We found that the DNA extraction method significantly affects the results for the diversity and composition of the sputum microbiome and resistome. Despite these differences, the results obtained with targeted HTS were repeatable. This highlights the importance of careful experimental design in microbiome research, where selecting the appropriate DNA extraction method is of crucial importance. WGS is often used to perform a deeper analysis of selected bacteria. The data generated in such an experiment is then used to perform comparative genomics studies. The main focus of our study was the plant pathogen *Pantoea stewartii* subsp. *stewartii*, the causative agent of Stewart's wilt in maize. In recent years, this quarantine bacterium has been found in Slovenia and elsewhere in Europe. To shed light on these findings, Slovenian isolates were sequenced alongside historical isolates from collections and compared to publicly available genomes. The results revealed all Slovenian isolates except one group together and are not related to other European isolates, suggesting multiple introduction events. The data generated by both metagenomic sequencing and WGS can be used to develop new molecular tests for the specific detection of target bacteria. Molecular detection methods are particularly well-suited for slow-growing bacteria. In our study, we used available genomic data to design multiple new quantitative real-time PCR (qPCR) tests to detect one such bacterium: *Xylophilus ampelinus*, the causative agent of bacterial blight in grapevines. In laboratory experiments, three of the newly designed qPCR tests were proven to be suitable for detecting *X. ampelinus* in different grapevine samples with high efficiency, specificity, and sensitivity.

The examples of HTS applications described in this thesis clearly demonstrate the versatility of this method. These include basic research, in which HTS can be used to develop and test new hypotheses regarding bacterial genomics, biology, and ecology. Other, more practical applications of HTS often include diagnostics, where it can be used to either directly analyse samples or to develop new diagnostic tests. HTS technologies will undoubtedly continue to evolve and remain an indispensable tool in the field of bacteriology.

Povzetek

Razvoj tehnologij visokozmogljivega sekvenciranja (*ang. high-throughput sequencing*, HTS) nam je ponudil nov in poglobljen vpogled v skrivnosten in očarljiv svet bakterij. Uporabnost teh tehnologij zajema številna področja, ki vključujejo vse od študij celotnih bakterijskih združb do poglobljenih raziskav izbranih bakterijskih izolatov. V tej doktorski nalogi smo na lastnih primerih preverili različne pristope uporabe HTS za preučevanje bakterij. Ti primeri vključujejo uporabo tarčnega HTS za izvedbo metagenomskih študij, sekvenciranje celotnih genomov (*ang. whole genome sequencing*, WGS) bakterijskih povzročiteljev bolezni rastlin in uporabo genomskih pristopov za načrtovanje novih molekularnih testov za specifično detekcijo tarčnih bakterij.

Tarčni HTS je pogosto uporabljen pristop, ki se uporablja v metagenomiki predvsem za določanje mikrobioma. Vendar na rezultate pogosto vpliva način obdelave vzorcev pred samim sekvenciranjem, pri čemer ima ključen pomen predvsem metoda ekstrakcije DNA. V naši študiji smo preučili vpliv treh različnih metod ekstrakcije DNA na rezultate mikrobioma in rezistoma na primeru vzorcev izpljunka. Ugotovili smo, da metode ekstrakcije DNA pomembno vplivajo na rezultate raznolikosti in sestave mikrobioma in rezistoma iz izpljunka. Kljub tem razlikam so bili rezultati, pridobljeni s tarčnim HTS, ponovljivi. To je poudarilo pomen premišljenega načrtovanja študij v raziskavah mikrobioma, kjer igra izbira ustrezne metode ekstrakcije DNA ključno vlogo. WGS je pogosto uporabljena metoda, ki nam omogoča poglobljeno preučevanje izbranih bakterij. Pri tem pridobljeni podatki se pogosto uporabijo za izvedbo študij primerjalne genomike. V naših raziskavah smo se osredotočili na rastlinski patogen, *Pantoea stewartii* subsp. *stewartii*, ki povzroča bakterijsko venenje koruze. Ta karantenska bakterija je bila v zadnjih letih večkrat zaznana tako v Sloveniji kot tudi drugod po Evropi. Da bi razložili te najdbe, smo sekvencirali genome slovenskih izolatov in zgodovinskih izolatov iz zbirk, s katerimi smo nato izvedli študije primerjalne genomike. Te so pokazale, da so slovenski izolati verjetno rezultat vsaj dveh vnosov bakterije, ki niso povezani z najdbami drugod po Evropi, kar nakazuje na več vnosov bakterije. Podatki, pridobljeni z WGS, se lahko uporabijo tudi za razvoj novih molekularnih testov za specifično odkrivanje tarčnih bakterij. Molekularne metode detekcije so še posebej primerne za bakterije, ki jih je težavno gojiti v laboratorijskih pogojih. Primer tovrstne bakterije je povzročitelj bakterijske ovelosti vinske trte *Xylophilus ampelinus*. V naši študiji smo tako uporabili razpoložljive genomske podatke za razvoj novih testov PCR v realnem času za detekcijo le-tega. V laboratorijskem testiranju smo potrdili primernost treh novih testov za odkrivanje *X. ampelinus* v različnih vzorcih vinske trte z visoko učinkovitostjo, specifičnostjo in občutljivostjo.

Metode uporabe HTS, opisane v doktorski nalogi, jasno prikazujejo raznolikost te metode. Med njimi so tako bazične raziskave, v katerih lahko HTS uporabimo za razvoj in testiranje novih hipotez, ki se navezujejo na bakterijsko genomiko, biologijo in ekologijo. Drugi bolj praktični načini uporabe se navezujejo predvsem na diagnostiko, kjer se HTS lahko uporablja tako za neposredno analizo vzorcev kot tudi za razvoj novih testov. Vse to zagotavlja nadaljnji razvoj HTS, ki bo ostalo nepogrešljivo orodje na področju bakteriologije.

Contents

Abbreviations

AMR	...	antimicrobial resistances
AMG	...	antimicrobial resistance gene
ANI	...	average nucleotide identity
AST	...	antimicrobial susceptibility testing
COPD	...	chronic obstructive pulmonary disease
CTAB	...	cetrimonium bromide
dPCR	...	digital polymerase chain reaction
DNA	...	deoxyribonucleic acid
ELISA	...	enzyme-linked immunosorbent assay
HGT	...	horizontal gene transfer
HTS	...	high-throughput sequencing
JSI	...	Jožef Stefan Institute
LPP-1	...	large *Pantoea* plasmid 1
IPS	...	International Postgraduate School
MALDI-TOF	...	matrix-assisted laser desorption/ionization time-of-flight
MGE	...	mobile genetic elements
NCBI	...	National Center for Biotechnology Information
NIB	...	National Institute of Biology
NJ	...	neighbour-joining
OTU	...	operational taxonomic unit
PCR	...	polymerase chain reaction
qPCR	...	quantitative real-time PCR
rDNA	...	ribosomal DNA
rRNA	...	ribosomal ribonucleic acid
RUCS	...	rapid identification of PCR primers for unique core sequences
SNP	...	single nucleotide polymorphism
T3SS	...	type III secretion system
TALE	...	transcription-activator-like effector
WGS	...	whole genome sequencing

Chapter 1

Introduction

1.1 Overview of Research Included in the Thesis

Within the scope of this thesis, we examined three important applications of high-throughput sequencing (HTS) and the data it generates. The first application involved the use of targeted HTS to perform metagenomic studies. We examined the use of targeted HTS for determining microbiomes and for the specific detection of clinically relevant bacteria in sputum, as this is often used in the clinical diagnosis of lung infections. Targeted HTS can similarly be used to detect genes associated with antimicrobial resistance (AMR), which is an ever-increasing threat to public health, making the ability to quickly and reliably detect such genes crucial in our fight against resistant bacteria. Targeted HTS could therefore be an important tool, as it can simultaneously detect all known bacteria present in samples, together with associated AMR genes. The second application of HTS examined in this thesis was whole genome sequencing (WGS) and subsequent comparative genomics analysis, focusing on plant-pathogenic bacteria. These bacteria are of special interest as they can cause severe yield losses of important crops, which, besides causing economic damage, could also endanger food security. In this thesis, we used WGS and comparative genomics to study two such bacteria. The first was *Pantoea stewartii* subsp. *stewartii*, the causative agent of Stewart's wilt in maize; the second was *Xanthomonas translucens* pv. *undulosa*, which causes disease in wheat. The third application that we examined was the use of HTS data for genome-informed design of a new quantitative real-time polymerase chain reaction (qPCR) test for the specific detection of bacteria. Once again, we focused on plant-pathogenic bacteria. For this application, we can use the large amount of genomic data stored in publicly available databases instead of performing HTS ourselves. This approach was used to develop multiple new qPCR tests for the specific and sensitive detection of *Xylophilus ampelinus*, which causes bacterial blight in grapevines. In an additional study, we attempted to use this approach to develop a qPCR test capable of detecting *Pseudomonas syringae* pv. *aptata*. After *in silico* design, the new qPCR tests still needed to be evaluated in laboratory testing before being put into use.

1.2 High-throughput Sequencing and Its Applications for the Study of Bacteria

Bacteria are one of the most widespread forms of life on our planet, inhabiting nearly every corner of it. The diverse and sometimes extreme conditions required for the growth of different bacteria pose significant challenges to their study and detection. Traditional microbiological methods, which rely on cultivating bacteria in growth media, are often

limited by the time and conditions required for bacteria to grow. Over the past two decades, high-throughput sequencing (HTS) methods have been developed that have helped us to overcome these challenges and provide us with unprecedented insights into the fascinating world of bacteria [1]. As a versatile technique, it can be used to study either a whole bacterial community or a single bacterium. The generic nature of HTS makes it suitable for studying virtually any bacterial community or taxon, which is one of its major advantages [2], [3]. Whole bacterial communities from different environments are usually studied by sequencing the total DNA in samples, a process known as metagenomic sequencing [4]. Conversely, to study a single bacterial isolate, the DNA is extracted and sequenced from the bacteria in a pure culture in order to obtain its complete genome. This is known as whole genome sequencing (WGS) [5]. Both of these approaches have proven to be extremely valuable, but they serve different purposes, so it is important to select the most suitable method for any given application [1].

HTS is the next step in the development of DNA sequencing methods, which are used to determine nucleic acid sequences. These form the genetic material of every organism, enabling us to decipher the universal life code hidden inside every living cell, including bacteria. The DNA molecule itself is a chain of nucleotides connected by a phosphodiester bond, and includes four different bases: adenine, cytosine, guanine, and thymine. These bases form a double helix, with two DNA molecules connected by hydrogen bonds formed by adenine and thymine, or cytosine and guanine. HTS is an umbrella term for multiple sequencing technologies, which can be broadly separated into two categories according to the length of the reads they produce. The first category comprises technologies that produce short reads, including sequencing by synthesis (Illumina, USA), ion semiconductor sequencing (Ion Torrent; Thermo Fisher Scientific, USA), and pyrosequencing (454; Roche, Switzerland). Of these, the Illumina sequencing-by-synthesis platform has proved to be the most successful and has become the industry standard. Conversely, methods that produce long reads include single-molecule real-time sequencing (PacBio; Pacific Biosciences, Menlo Park, USA) and nanopore sequencing (Oxford Nanopore Technologies, Oxford, UK). Although PacBio and nanopore sequencing enable easier genome assembly, they have some disadvantages. PacBio has relatively lower throughput and is more expensive than other platforms. Nanopore sequencing has traditionally been limited by lower accuracy, but recent advances in base-calling algorithms and sequencing chemistry have significantly narrowed the gap with Illumina [6]. The types of errors produced by DNA sequencing technologies also differ, with these errors being either systematic or random. PacBio produces mostly random errors, whereas nanopore sequencing produces systematic errors in homopolymer sequences and k-mer-specific errors [7], [8].

HTS generates vast quantities of genomic data on a regular basis. These datasets are deposited in public databases such as GenBank, European Nucleotide Archive (ENA), and DNA Data Bank of Japan (DDBJ), providing a foundation for downstream analyses [9]–[11]. Both metagenomic sequencing and WGS rely on these pre-existing databases for bioinformatic analyses and results interpretation. Sequencing data is analysed using databases containing previously annotated sequences to identify the bacterial taxa and genes present in the samples or genome under study [12], [13]. HTS-derived data further enable a broad range of analyses that deepen our understanding of bacterial biology. These include studies of mobile genetic elements, pathogenicity factors, and antimicrobial resistance [14]–[16]. Metadata like host organism, geographic location, and year of isolation are equally important as genomic sequences themselves and are crucial for interpreting and contextualising the results obtained. One important field in which genomic data can also be used is diagnostics. Here, HTS can be used directly to identify microorganisms or indirectly for genome-informed identification of novel targets, which are then used to design new molecular diagnostic tests [17].

1.2.1 Targeted high-throughput sequencing

HTS can be performed in a targeted or non-targeted manner. In the non-targeted approach, the entire DNA is sequenced, whereas in the targeted approach, only the segments of interest are sequenced. To achieve this, targeted HTS requires an additional step to select the target DNA sequences. This can be achieved by enriching for target DNA, depleting non-target DNA, or amplifying target DNA. The latter is the most widely used strategy and will be described in more detail below [18]–[20]. The advantages of targeted HTS are that unessential data is not generated, increasing sensitivity and simplifying further bioinformatic analysis while reducing the cost of the experiment. This makes it especially promising for use in clinical and diagnostic settings [21]. The obvious disadvantage of targeted HTS is that some relevant information could be missed if the wrong sequences are targeted. Another challenge is the additional bias that can be introduced by the process of selecting target sequences, which must be taken into account when interpreting results. For example, when the goal is the detection of genes associated with antimicrobial resistance that we actually target such genes present in an environment [22]. However, these challenges can be overcome through an intelligent experiment design based on prior expert knowledge.

1.2.2 Targeted HTS for bacterial metagenomics

Metagenomics is the study of all the genetic material obtained by sequencing the DNA present in a given sample. In this context, the term 'microbiome' refers to the genes and genomes of all the microorganisms present in a given environment, such as the human body or a specific ecological niche [23]. A similar term is 'microbiota', which describes all the microorganisms present in an environment. Areas covered by microbiome research include the ecology of diverse environments, the impact of microbial communities on agriculture, their potential applications in biotechnology, and microbial evolution. Metagenomic studies are often performed on environmental DNA, which helps us to monitor the presence of different organisms in the studied ecosystem. However, the area receiving the most attention is the study of the different microbiomes present in the human body and their impact on health. Multiple studies have shown that the microbiome composition can significantly impact health, with microbiome changes being either the cause or consequence of disease development. Parts of the human body that are characterised by distinct microbiome communities include the gut, the oral cavity, the urogenital system, and the respiratory tract [24]–[27]. Microbiome is often described by alpha diversity, which is a general term that describes richness (number of taxa present), evenness (commonness or rarity of taxa), and diversity (Shannon and Simpson indices) [28].

The development of HTS has greatly impacted microbiome studies by facilitating culture-independent analysis of microbial communities, thereby deepening our understanding of their complexity. Targeted HTS is a common approach in microbiome studies, where only the sequences necessary for identifying bacterial taxa are sequenced. The 16S rRNA region is most commonly sequenced for microbiome analysis due to its evolutionary conservation, the availability of reference sequences in databases, and well-developed protocols [29], [30]. Organisms detected in microbiome are often instead of species classified into so-called operational taxonomic units (OTUs), each of which includes closely related organisms (usually the 97 % sequence similarity threshold is applied) [31]. Furthermore, 16S rRNA sequences alone often lack the resolution required for species-level identification and are used primarily for bacterial classification on the genus level. To address this issue, a targeted HTS approach using species-specific amplicons can be

implemented to enable the simultaneous detection of a wide array of bacterial species [32]. This method shows promise in the field of clinical microbiology, particularly in pathogen diagnostics. Advantages of this approach include higher specificity and sensitivity; however, it is limited to the detection of fewer bacteria than untargeted or 16S rRNA approaches.

A major challenge in interpreting results is that many steps in the targeted HTS process can influence microbiome results. These steps include sampling, sample heterogeneity, sample pre-processing (filtration, enzymatic lysis, enrichment, cell lysis, etc.), the DNA extraction method, library preparation, sequencing chemistry, and the bioinformatic analysis of raw sequencing data [33]–[35]. DNA extraction has been recognised as a crucial step that influences the results of targeted HTS in metagenomic studies. It can introduce bias by overrepresenting certain groups of microorganisms, such as Gram-negative bacteria, and cause variability between experiments due to a lack of repeatability [36], [37]. Therefore, the choice of DNA extraction method is of the utmost importance when designing targeted HTS experiments.

1.2.3 Whole genome sequencing

One of the main applications of HTS in the field of microbial genomics is determining the complete DNA sequence of a bacterial genome, a process known as whole genome sequencing (WGS). The first bacterial genome to be sequenced was that of *Haemophilus influenzae* in 1995, using the chain termination method developed by Sanger [38]. WGS was initially laborious and costly but became much more accessible with the development of HTS technologies (Section 1.1). This has led to an exponential increase in the number of newly sequenced genomes, with 2.41 million bacterial genomes publicly available in the GenBank database by 2024 [39]. Today, WGS has a variety of applications, including academic and biomedical research, biotechnological applications, environmental studies, food safety, and plant pathology, among others. Data from WGS can be used for taxonomical classification of bacteria or even for their identification on a strain level (strain typing), for which this method has become the gold standard due to the high level of detail it provides [40], [41]. This can be used to identify pathogenic bacteria in clinical diagnostics, foodborne pathogens in the food industry, and plant-pathogenic bacteria in agriculture [42], [43]. By sequencing multiple bacterial strains, we can trace the outbreak and transmission pathways of a new pathogen to discover its source of spread [44]. Data from WGS can also be used in studies that examine functional characterisation of bacterial genomes, interactions between bacteria and host organisms, and in genome-wide association studies to identify specific functional links [45], [46].

The bacteria used in conventional WGS are usually first isolated in pure culture. This enables assemblies with deeper coverage to be generated and reduces the possibility of contamination in the genome. Additionally, isolating the bacteria enables us to acquire phenotype data suitable for mapping genotype–phenotype relationships. For WGS data analysis, both assembled genomes and raw reads generated in sequencing can be used. Bacterial genomes can also be assembled from reads generated in metagenomic sequencing and are usually referred to as metagenome-assembled genomes. Such genomes can be a rich source of data; however, they often lack phenotypic characterisation and adequate depth for reliable use in further analysis [47]. To perform WGS, DNA must first be extracted from the bacteria using methods that remove proteins, RNA, and other contaminants, producing DNA fragments of sufficient length. The length of the DNA fragments is especially important when long-read sequencing technologies such as nanopore sequencing are used. Library preparation in WGS sequencing is relatively straightforward and mostly

serves to prepare the DNA for sequencing. When short-read methods are used, the DNA is cut into smaller pieces, with a length of around 350 bp. Conversely, when long-read methods are used, smaller DNA fragments are often depleted during library preparation.

After sequencing, the raw sequencing reads are demultiplexed, trimmed to remove adapters that were attached to the DNA during library preparation, and subjected to quality control [48], [49]. This step is crucial to avoid errors, since the quality of the raw reads determines the quality of the final assembly. High-quality reads that pass this step are then used for genome assembly. Genome assemblies can be generated genome-guided (reference-based) or *de novo* (reference-free) [50], [51]. In WGS, the *de novo* assembly strategy is usually employed as it is free of reference bias, can identify strain-specific genes, and can be applied to any previously uncharacterised bacteria. Most current genome-assembly tools use *k*-mer–based de Bruijn graphs to construct contigs from sequencing reads [52]. Genome assemblers are adaptable for either short-read (e.g., SPAdes, SKESA), long-read (e.g., Flye, Canu, Miniasm), or hybrid (e.g., SPAdes, Unicycler, MaSuRCA) genome assembly [53]–[59]. To obtain genome assemblies of the highest quality and completeness, a hybrid approach combining short and long reads is often used. Genome assemblies are subsequently polished (tools like Pilon, Racon, Medaka) to correct errors, improve consensus accuracy, and refine contig sequences [60], [61]. Tools such as Unicycler are particularly useful as they incorporate the entire genome assembly workflow, including quality control, assembly, and polishing [58].

The quality of assembled genomes is evaluated using different parameters (genome coverage, completeness, and contamination percentages, number of contigs, and guanine-cytosine content). The average genome coverage, in combination with the sequencing method, indicates how accurate the genome sequences are. The number of contigs can provide an initial indication of genome completeness; however, it is highly dependent on the sequencing technology used. Short-read sequencing methods produce genome assemblies with a higher number of contigs than long-read technologies, but can still have a high percentage of completeness. Complete chromosomes and plasmids usually have contig sequences in a circular form, which is a useful initial indication of their completeness. A more suitable method of calculating genome completeness is to use programs that search for the presence of evolutionarily conserved genes that are expected to exist as single copies in the vast majority of species within a lineage. Examples of such programmes are BUSCO (Benchmarking Universal Single-Copy Orthologs) and CheckM [62], [63].

The next crucial step is genome annotation, during which functional elements such as genes, regulatory sequences, and other features are identified, labelled, and exported in standard formats such as GFF3 (general feature format version 3). This process enables us to identify coding DNA sequences, including protein-coding sequences, multiple types of RNAs, and pseudogenes. Genome annotation can be divided into structural annotation, which predicts coding DNA sequences and open reading frames, and functional annotation, which assigns functions to identified genes [64]. The most widely used tools for automated genome annotation are the NCBI Prokaryotic Genome Annotation Pipeline, Prokka, and Bakta [65]–[67]. Tools such as EggNOG-mapper and DIAMOND are used for functional annotation by performing sequence-similarity searches against curated reference databases [68], [69]. The data generated by genome annotation is crucial for the performance of comparative genomics studies, which will be described in the next section.

1.2.4 Comparative genomics

The availability of huge amounts of data generated by WGS has led to the development of comparative genomics, which attempts to contextualise this data. It analyses and

compares the genomes of different organisms to establish evolutionary relationships, identifying similarities and differences between taxa. In bacteriology, comparative genomics is used to explore evolutionary relationships, genome structure, horizontal gene transfer (HGT), the development of pathology, and microbial ecology, among others. Such studies are necessary to establish the genotype-phenotype relationship and help us understand the role of bacteria in an environment.

An important area of comparative genomics is the investigation of the evolutionary relationships between different bacterial taxa using phylogenetic methods, which often involves constructing a phylogenetic tree. Such analyses can also be used to track the spread of new pathogenic bacteria in the event of a disease outbreak and to discover transmission pathways [70]. Various methods can be employed to conduct phylogenetic analyses; the selection of the most suitable approach often depends on the degree of relatedness between the studied bacteria and the available data. Comparative and phylogenetic analyses commonly use whole-genome multi-locus sequence typing, *in silico* DNA–DNA hybridisation, or distance matrices derived from average nucleotide identity (ANI) or single-nucleotide polymorphisms (SNPs) to measure relatedness between strains. For closely related bacterial strains belonging to the same species or subspecies, SNP-based approaches are commonly used: reads or assemblies are aligned to identify variant sites (e.g., MAFFT, BWA-MEM), SNPs are called (using tools like Snippy or CSI Phylogeny), and the resulting SNP alignment is used to infer a tree with maximum-likelihood or neighbour-joining (NJ) methods (using tools like RAxML or IQ-TREE) [71]–[76]. Alternatively, ANI-based phylogeny calculates pairwise average nucleotide identity (using tools like PyANI or FastANI) to produce a distance matrix, which is then converted into a tree using NJ, unweighted pair group method with arithmetic mean, or minimal-evolution algorithms (using tools like FastME or PHYLIP) and finally visualised (using tools like FigTree, iTOL, or Dendroscope) [77]–[80].

Comparative genomics extends beyond mere phylogenetic analysis. One example is the study of the pan-genome, which comprises all the genes present in all strains of a given bacterial taxon [81]. The pan-genome consists of a core genome, a shell genome, and a cloud genome. The core genome consists of genes present in every genome of the examined taxa, and the term 'soft core' is often used for genes present in almost all (over 95%) genomes. Core genome genes include housekeeping genes that are essential for survival or responsible for basic cellular functions [82]. Shell genes are present in some, but not all, genomes (usually ranging from 15 to 95% of analysed genomes). These genes are not essential for bacterial survival, but they can offer an evolutionary advantage and often include genes connected to antimicrobial resistance, pathogenicity, niche adaptation, and metabolic flexibility. Cloud genes are strain-specific or present in just a few genomes (usually under 15% of genomes are used as the criterion). These genes serve functions similar to those in the shell genome [83]. Bacteria can have an open (continuous addition of new genes as more genomes are sequenced, which lowers core-gene proportion and highlights increased rate of HGT) or closed (stabilizes after sampling a sufficient number of genomes, as the number of genomes included in the pangenome increases, the total number of genes will plateau) pan-genome [84], [85]. To determine the pan-genome, bacterial genomes must first be annotated. To ensure reliable results, it is crucial that only high-quality genomes with a high completeness percentage are used in the analysis and that they are all annotated using the same method. These annotations are then used in pan-genome analysis with tools such as Roary, Anvi'o, and Get_homologues [86]–[88].

Genes present in shell and cloud genomes are often acquired through horizontal gene transfer (HGT), which is another important area of study in comparative genomics. HGT is the process by which bacteria acquire new genetic material from other organisms without reproducing. This can be achieved through the processes of transformation, transduction,

and conjugation. Genes acquired through HGT can help bacteria adapt to their environment, become more pathogenic, and develop antimicrobial resistance. The process of HGT often involves mobile genetic elements (MGEs) present in bacterial genomes, including plasmids, transposons, integrative and conjugative elements, prophages, and others [14], [89]. Comparative genomics is used to identify MGEs present in different bacterial strains and to study how they spread and the role they play in the evolution of bacterial species. By comparing MGEs present in different strains, scientists can identify HGT events that lead to their spread [90]. In addition to the tools used in pan-genome analysis, many more can be used to identify HGT and MGEs in bacteria. IslandViewer4 can identify genomic islands [91]; Phastest can identify prophage sequences and regions [92]; MetaCHIP can be used to identify HGT events within a set of genomes [93]; and MAUVE can be used to identify larger genomic rearrangements [94].

The genes that bacteria acquire through HGT could contribute to the development of pathogenicity against a specific host. The study of pathogenicity is another important area on which comparative genomics studies often focus. Pathogenicity factors can include adhesins that promote bacterial attachment to host cells, toxins that cause host tissue damage, secretion systems that deliver effectors into host cells, and immune evasion factors [95]. Tools used in comparative genomics can be used to identify known pathogenicity factors or to discover new ones by comparing the genomes of bacteria that express different phenotypes. Databases (e.g., VFDB or BacSPaD) containing information on known pathogenicity factors can be used; however, a literature survey must be conducted in cases involving less-studied bacteria [96], [97]. These databases are usually used alongside tools such as PathogenFinder and VirulenceFinder to identify pathogenicity factors present in bacterial genomes [98], [99]. This information helps us to understand the underlying mechanisms of pathogenicity and, hopefully, to predict the pathogenic potential of newly discovered bacteria.

1.2.5 Development of new molecular methods for the detection of bacteria

Traditionally, bacterial classification relied on culture-based methods, which involved growing bacteria on artificial media and isolating them in pure culture. These methods still have some advantages, such as being inexpensive and having high theoretical sensitivity, as well as being able to provide some phenotypic information on the target bacteria. However, in recent decades, they have increasingly been replaced in many areas by molecular methods due to their lower throughput, lack of specificity, and greater time-consuming nature. On the other hand, molecular methods based on the detection of specific genetic markers have become indispensable in routine diagnostics and are particularly useful for detecting target bacteria due to their specificity, sensitivity, speed, and relative ease of use. Many of these methods are based on the amplification of a specific genomic sequence, which is usually achieved using some form of the polymerase chain reaction (PCR)[100]. Variants of PCR that are often used include conventional PCR, nested PCR, colony PCR, multiplex PCR, and quantitative real-time PCR (qPCR), which enables relative quantification of the target sequence, as well as digital PCR (dPCR), which enables absolute quantification [101]–[106]. PCR amplification must be combined with another method to enable the detection of amplified DNA fragments. For conventional PCR, agarose gel electrophoresis is used to separate DNA fragments by size, which is combined with fluorescent staining to detect and isolate PCR products. qPCR and dPCR use fluorescent dyes to detect amplified DNA sequences of known size. The product of PCR amplification can also be sequenced further for more precise taxonomic classification in a

process known as DNA barcoding [107]. When such an analysis is performed on multiple loci within the bacterial genome, it is referred to as multi-locus strain typing (MLST), which is a common method used to identify new bacterial isolates [108]. Another advantage of PCR-based methods is that they can be implemented for the detection of new target bacteria. However, PCR must be able to specifically detect target bacteria, which requires prior knowledge of the specific DNA sequence. Large amounts of data generated by whole-genome sequencing are crucial for identifying species-specific sequences. However, there is often still a lack of high-quality genomic data needed for less studied bacteria. Due to the immense biodiversity on Earth, it can be challenging to design a test with zero cross-reactivity; however, intelligent design can help us to avoid false-positive results in samples in which we want to detect target bacteria.

In order to identify DNA sequences specific to target bacteria, we must first determine which bacteria we wish to detect (the positive dataset) with the new test and which we should not be detecting (the negative dataset). PCR-based methods can have a range of different specificities; they can be designed to detect all bacteria belonging to a select genus, or to be specific to a single subspecies or phylogroup within a species [109], [110]. Once the target bacterial taxa have been selected, databases such as GenBank and RefSeq are screened for genomes, which are then used to perform a phylogenetic analysis to determine the positive and the negative dataset [111]. The genomes from the positive dataset are then compared to the genomes from the negative dataset to identify genomic sequences that are specific to the target group. Non-target bacteria usually involve other closely related bacteria belonging to the same genus, or even the same species and subspecies. Tools such as RUCS (Rapid Identification of PCR Primers for Unique Core Sequences) can be used to identify such sequences [112]. Those sequences are then used to design oligonucleotides for PCR amplification using programs such as Primer3 and Primer Express [113]. Some versions of PCR also use hydrolysis probes for greater specificity and easier signal detection [114]. The newly designed oligonucleotides are tested *in silico* for specificity through BLASTn searches against nucleotide sequence databases (e.g., BLAST Core Nucleotide Database or RefSeq Genome Database). Those oligonucleotides that show no cross-reactivity *in silico* are then evaluated in laboratory tests involving mock and real samples to determine their suitability for use in diagnostics. Laboratory testing is often performed on so-called spiked samples in which the target bacteria are added or spiked in known concentrations. These tests are validated for detection within a limited concentration range in a specific matrix, since real samples often contain inhibitors of PCR amplification or may cause cross-reactivity. Using PCR-based tests in matrices for which they have not been validated can lead to false-negative or even false-positive results. After their introduction into diagnostics, the performance of tests is monitored to identify any previously unnoticed cross-reactivity or inhibition.

1.3 Sputum Microbiome and Resistome

1.3.1 Lung microbiome

The lungs were previously thought to be a sterile environment; however, improved microbiological methods have shown they host distinct, low-biomass microbial communities that differ from other body sites, such as the gut or oral cavity, and reflect the lungs' requirement for minimal biomass to maintain efficient gas exchange. A healthy lung microbiome plays an important role in the immune system's response to pathogens, while a disturbed microbiome can be either a cause or a result of disease [115]. The lungs mostly

rely on mechanical clearance mechanisms, such as mucociliary clearance and coughing, as well as the host immune response, to maintain a low microbial burden [116], [117]. Bacteria are constantly introduced into the lungs from the upper respiratory tract via aspiration, as reflected by the bacterial taxa present in the lung microbiome [118]. The most common bacteria are from the genera *Prevotella*, *Streptococcus*, *Veillonella*, *Haemophilus,* and *Neisseria*. The regular introduction of microorganisms into the lungs, followed by their clearance, leads to a more dynamic lung microbiome compared to other human microbiomes [119], [120].

An increase in microbial burden and microbiome dysbiosis is associated with pathological states, which can be either temporary or chronic. These include diseases such as cystic fibrosis, asthma, chronic obstructive pulmonary disease (COPD), and lung cancer [121]–[124]. The composition of the microbiome could impact inflammation levels in the lungs, particularly in cases of chronic lung diseases such as COPD. These diseases often lead to lung colonisation by bacteria such as *Haemophilus influenzae*, which are associated with increased airway inflammation [125]. Microbiome dysbiosis also plays an important role in exacerbating chronic lung diseases through interaction with the host immune response [126]. In diseases such as cystic fibrosis, a reduction in microbiome diversity is correlated with disease severity and can lead to exacerbation. A reduction in microbiome diversity also makes the lungs more susceptible to colonisation by pathogenic bacteria, such as *Pseudomonas* and *Burkholderia* species, which could begin to dominate the lung microbiome [127]. Due to the anatomy of the lower respiratory tract, topographical differences in the lung microbiome must also be taken into account [128].

1.3.2 Antimicrobial resistance and lung resistome

Emergence and spread of antimicrobial resistance (AMR) in pathogenic bacteria present one of the main public health concerns [129]. It causes increased mortality and morbidity, posing an additional burden on the public health system and society as a whole [130], [131]. The process of introducing new antibiotics is highly time-consuming and expensive. Even when a new antibiotic is introduced for clinical use, bacteria can quickly develop resistance. Like any other bacteria, those present in the lungs can also possess AMR against a variety of clinically used antibiotics, encoded by their corresponding genes. All the genes associated with AMR in a given environment form a 'resistome', which can be divided into a 'core' resistome and an 'accessory' resistome [132]. The core resistome includes AMR genes (AMGs) that are usually found in both healthy subjects and those with respiratory diseases. In contrast, the accessory resistome consists of AMGs that are specific to a given respiratory disease. AMR is especially problematic in the case of bacteria inhabiting the lungs, due to respiratory diseases being among the most common reasons for hospitalisation [133], [134]. The development of AMR is particularly critical in chronic lung diseases and infections requiring prolonged antibiotic treatment, which creates selective pressure for the accumulation of AMGs [135]. Another vulnerable group is immunocompromised patients, who are more likely to develop prolonged infections with opportunistic pathogens, making the emergence of AMR even more likely [136]. Bacteria use HGT to acquire new AMGs, leading to the accumulation of multiple resistances in a single bacterium and the formation of so-called multidrug-resistant bacteria, which are resistant to most of the antibiotics currently in clinical use [137]. Examples of bacteria that frequently acquire AMGs and cause lung infections include *Acinetobacter baumannii*, *Klebsiella pneumoniae*, *Pseudomonas aeruginosa*, and *Staphylococcus aureus*, which can carry resistance determinants against major antibiotic classes such as beta-lactams, macrolides, fluoroquinolones, and tetracyclines [138], [139]. AMGs encode, among other

things, enzymes that degrade antibiotics, pumps that secrete antibiotics from bacteria, proteins that bind to antibiotics instead of their targets, and modifications to antibiotic binding sites [140].

To mitigate the spread of AMR, multiple strategies have been proposed, including reducing the prescription of antibiotics when they are unjustified and prescribing antibiotics more selectively. For these approaches to be effective, antimicrobial susceptibility testing (AST) is of the utmost importance to quickly and reliably determine which antibiotics are most suitable for treating the infection. Methods of AST can be broadly divided into phenotypic and genotypic methods. Classical phenotypic methods mostly rely on in vitro cultivation of bacterial isolates in the presence of antibiotics [141]. These methods are well established; however, they can be time-consuming, especially in cases involving slow-growing bacteria, such as *Mycobacterium tuberculosis*, or bacteria that are difficult to isolate and grow in pure culture, such as *Mycoplasma pneumoniae*. Additionally, they are often limited in scope and throughput, as they can only test a limited number of antibiotics and bacteria in a single procedure. Conversely, culture-independent methods could be used for more comprehensive detection of AMGs in the lungs. Targeted molecular methods that offer high sensitivity and specificity, such as qPCR, can only detect a limited set of predetermined AMGs [142]. To overcome these obstacles, HTS is used to simultaneously determine all AMGs present in the resistome. For this purpose, both targeted and non-targeted HTS approaches can be employed and can be performed directly on DNA extracted from a clinical sample [16], [143]. By omitting the time needed to grow and isolate bacteria in culture media, the time taken for diagnosis can be shortened, which is extremely important for critically ill patients.

Despite genotypic methods and HTS presenting a huge advancement in AST, there are still some challenges in the application of these methods. The most significant factor is that the presence of some AMGs does not necessarily result in phenotypic antibiotic resistance [144]. To detect AMGs, HTS-based methods also require prior knowledge and the relevant databases (such as The Comprehensive Antibiotic Resistance Database) containing sequences associated with a resistance phenotype are needed [145], [146]. Another challenge is that AMR in bacteria can emerge through different mechanisms, including HGT or new genetic mutations [147], [148]. Resistance caused by mutations is harder to predict from genomic data due to the vast number of mutations that can result in the same phenotype [149]. Another major challenge is the difficulty in determining exactly which bacteria any detected AMG belongs to. An additional obstacle in determining the resistome by metagenomic sequencing is the large proportion of host DNA. This often constitutes the majority of the DNA extracted from a diagnostic sample, reducing the number of useful sequences obtained during sequencing. This makes it increasingly difficult to reach the sufficient sequencing depth needed for a reliable determination of AMGs [150]. For this reason, targeted HTS is an attractive option for determining the resistome. Using this approach, we enrich the sample for sequences belonging to AMGs, typically through PCR-based amplification [151], [152].

1.3.3 Determination of microbiome and resistome using sputum

Sputum is a mucus collected from the lower respiratory tract that contains bacteria present in the lungs. In clinical diagnostics, sputum culture is often used to diagnose lower respiratory tract infections (such as bacterial pneumonia) and monitor treatment effectiveness. The bacteria isolated from the sputum can then be used for antimicrobial susceptibility testing. [153]. In addition to routine diagnostic testing, sputum is, due to its simple and non-invasive collection procedure, frequently used as a source material to

determine the lung microbiome and resistome. One of the main advantages of sputum is the non-invasive collection procedure. Furthermore, sputum samples have a higher bacterial load than corresponding bronchoalveolar lavage fluid samples [154]. This is especially useful for studying microbiome changes associated with lung diseases, such as COPD and cystic fibrosis, that change microbial composition and are characterised by excessive sputum production [155], [156]. As with any other biological sample, sputum presents a few challenges in research and diagnostics. Healthy human lungs usually produce low amounts of sputum, which are not sufficient for analysis, making bronchoalveolar lavage a more suitable method of sample collection. The collection procedure is crucial to obtain sputum samples representative of the lung microbiome with minimal upper-respiratory contamination; patients typically rinse their mouths with water before expectoration, but potential oral contamination must still be considered when interpreting results [154].

In order to perform a culture-independent analysis, the method used to extract DNA from sputum is of the utmost importance and should introduce as little bias as possible into the results. It is crucial that the extraction provides DNA of a high enough yield and quality to enable further analysis. The high heterogeneity of bacteria present in sputum poses an additional challenge to DNA extraction, which is typically mitigated by adding solubilising agents, but the predominance of human host DNA remains a major obstacle to recovering microbial genetic material [157]. This is particularly problematic in metagenomic sequencing, as the majority of the generated data is not useful for the analysis and can skew the results. To overcome the problem of human host DNA, different depletion methods could be implemented; however, these could introduce additional bias in the proportions of detected AMGs present in the results [158], [159]. Another approach is to use targeted HTS to detect bacteria and AMGs in the sputum microbiome. This approach also enables straightforward analysis and interpretation of results, making it especially suitable for clinical applications.

1.4 Plant Pathogenic Bacteria

The omnipresence of bacteria on Earth is reflected in plants as well, where they form the natural microbiome and perform many ecological functions. Although, the vast majority of plant-related bacteria being harmless or mutualistic, pathogenic species that affect major crops remain a primary concern. Such bacteria include many different classes, including both Gram-negative bacteria (e.g., *Pseudomonas*, *Ralstonia,* and *Xanthomonas*) and Gram-positive bacteria (e.g., *Clavibacter* and *Streptomyces*). Their spread causes significant yield loss, which can lead to economic damage and, in the worst case, endanger our food security. Plant pathogenic bacteria differ in their host range. Bacterial species are often subdivided into so-called pathovars according to the plant host in which they cause symptoms (e.g., *Pseudomonas syringae* pv. *aptata,* which causes disease in beets) [160]. However, the classification into pathovars could be problematic as bacteria often cause disease in multiple plants. Additionally, some strains could be included in the same pathovar despite not being closely related. Plant-pathogenic bacteria infect not only crops, but also ornamental plants such as orchids, as is the case with *Dickeya fangzhongdai* [161]. Some bacteria can even cause symptoms in trees, as with *P. syringae* pv. *aesculi*, which causes bleeding canker in horse chestnuts [162].

Since plants are sessile organisms, plant pathogenic bacteria have to use different transmission pathways. Some spread through water; thus, irrigation and hydroponic systems present an especially dangerous environment for transmission in agricultural settings. Many plant-pathogenic bacteria belonging to the species *P. syringae*, the

Ralstonia solanacearum complex, the genus *Pectobacterium,* and the genus *Dickeya* have so far been reported in irrigation systems [163]. Wind does not transmit bacteria well by itself, but it can help to transmit bacteria in aerosols in rain droplets over moderate and long distances [164], [165]. Some bacteria can persist in soil or plant debris for months or even years, causing new infections. An example of this is the *R. solanecearum* complex, which can survive in soil and infect new plants through their roots [166]. Some bacteria also rely on insect vectors for transmission. Examples include *Xylella fastidiosa*, which spreads via leafhoppers (Cicadellidae), and *Pantoea stewartii*, which spreads via the corn flea beetle (*Chaetocnema pulicaria*) [167]. Bacteria usually cannot penetrate healthy plant tissue, but enter plants through natural openings, wounds, or insect bites. Contaminated tools and equipment used for pruning can cause the spread of bacteria, as with *Erwinia amylovora* [168]. Bacteria can reside within seeds and can spread locally or globally via seed trade. Human movement of infected plant material, including grafts and vegetative cuttings, further facilitates the spread, which could devastate susceptible crops. A recent example is the spread of *X. fastidiosa*, which was introduced from America to southern Italy, where it caused widespread olive tree dieback [169].

Plant-pathogenic bacteria use different mechanisms to survive and spread inside the host plant. These include different secretion systems that deliver effectors into host cells. One example is the type III secretion system, which acts like a molecular syringe, injecting effector proteins such as AvrE into host cells [170]. Plant pathogenic bacteria produce a wide array of cell wall-degrading enzymes, such as pectinases and cellulases, to break down plant tissue. These are particularly prevalent in bacteria belonging to the genera *Pectobacterium* and *Dickeya* [171], [172]. Bacteria such as *P. syringae* are also known to produce phytotoxins that can kill plant cells or disrupt plant signalling pathways [173]. Many bacteria also utilise the plant's vascular system to travel through the tissue within the water-conducting xylem vessels. Well-known examples of such bacteria are *X. fastidiosa* and *Xylophilus ampelinus* [174], [175]. These bacteria often form biofilms inside the xylem vessel, which can restrict the flow of water and nutrients, resulting in symptoms such as wilting. Biofilm formation can be regulated by quorum sensing, which is often involved in bacterial pathogenicity [176]. The disease symptoms caused by plant-pathogenic bacteria can vary and are often dependent on the mechanism of pathogenicity used by the bacteria.

Plant pathogenic bacteria can cause symptoms in different parts of plants. On leaves, for example, they can cause spots, blights, and water-soaked lesions, as well as yellowing, which can be caused by bacterial toxins and enzymes. An example is the genus *Xanthomonas*, which can cause such symptoms in a wide range of plants [177]. On stems, branches, and trunks, bacteria can cause sunken or cracked areas. In some cases, entire plants or branches can wilt due to bacteria blocking xylem vessels. One such bacterium is *Clavibacter michiganensis*, which causes bacterial canker in tomatoes [178]. Plant-pathogenic bacteria can cause root rot, which is characterised by soft, mushy, brown-to-black decay of the roots due to bacterial enzymes degrading the root tissue [179]. Bacteria from the genus *Agrobacterium* are also well known for causing tumours in plants. An example is crown gall, which is caused by *Agrobacterium tumefaciens* [180]. On fruit, bacteria can cause soft rot, which is characterised by watery, foul-smelling decay. This is known to be caused by *Pectobacterium* and *Dickeya* species [181]. Bacteria from the genus *Xanthomonas* are also known to cause sunken or raised brown-black spots on the surface of fruit [177]. Development of disease symptoms caused by plant pathogenic bacteria is often difficult to observe and study in the wild, due to the prolonged time required for the symptoms to develop, as well as the effects of other environmental factors like drought or lack of certain nutrients (e.g., magnesium, nitrogen, or phosphorus). To overcome these

challenges, the pathogenicity of bacteria on plants can be studied under controlled conditions in pathogenicity tests performed in greenhouses or growth chambers.

Various methods can be employed to control the spread of plant-pathogenic bacteria, including biological, chemical, and cultural approaches [182]. Countries often implement regulatory and quarantine measures to prevent the introduction of plant diseases to new areas. This includes testing plant material that could carry the pathogen, including seeds and plant material such as potatoes used for food production and propagation [183]. In biological control, other beneficial bacteria, such as *Bacillus* species, or bacteriophages, could be used to inhibit the growth of phytopathogenic bacteria [184], [185]. Chemical control can be effective in some cases, but the use of antibiotics is very limited due to the risk of AMR, which has led to their ban in many countries, including the European Union [186]. Cultural control methods are often employed to limit the spread of existing bacteria, including the removal of infected plants, the disinfection of tools, and crop rotation to interrupt the disease cycle. To effectively mitigate the spread of plant-pathogenic bacteria, it is crucial to combine multiple approaches, eradicating the bacteria when possible and preventing their further spread.

1.4.1 *Pantoea stewartii* subsp. *stewartii*

Pantoea stewartii subsp. s*tewartii* (Smith 1898) is a Gram-negative bacterium belonging to the Erwiniaceae family and causes Stewart's wilt in maize [187]. The bacterium is indigenous to North America, yet cases have been reported all around the world. Symptoms of the disease include wilting, which is typical of infection in young seedlings, and the appearance of water-soaked lesions on leaves. The seedlings become severely wilted, often resulting in the death of the plant. Other symptoms include blight, which manifests when mature plants are infected. This causes characteristic linear yellow-grey lesions that run parallel to the veins of the leaf. Symptoms are particularly severe in sensitive maize cultivars, including sweet maize (*Zea mays* subsp. *saccharata*) and certain elite inbred maize lines [188]. The disease is transmitted by insects, particularly the corn flea beetle (*Chaetocnema pulicaria*), which can harbour the bacterium in its gut during the winter months [167]. It is hypothesised that the dissemination of bacteria over long distances is facilitated by infected seeds, resulting in restrictions on international trade involving seeding material [189]. Mitigating the dissemination of the bacterium can be achieved through cultivating resistant maize cultivars and rigorously testing imported seeds for the bacterium's presence [190]. In addition to maize, *P. stewartii* subsp. *stewartii* has been observed to infect various other plants, including sudan grass, oats, triticale, sorghum, millet, and sugarcane [191]. Apart from subspecies *stewartii*, another subspecies, *P. stewartii* subsp. *indologenes* has been identified as a causative agent of disease in various crops, including *Allium* spp. and rice [192]. Although *P. stewartii* subsp. *indologenes* is not considered pathogenic to maize; it has recently been isolated from symptomatic maize plants [193].

The two main pathogenicity factors of *P. stewartii* subsp. *stewartii* are the Hrp type III secretion system (T3SS) and the exopolysaccharide stewartan [176], [194]. The Hrp gene cluster encodes the T3SS and its effectors, which are involved in host plant colonisation [195]. *P. stewartii* does not possess a large number of different effector proteins, indicating that it acquired the T3SS relatively recently [194]. One such effector is WtsE, a member of the AvrE family of effectors that causes cell death, resulting in water-soaking lesions and necrosis in maize plants, and is essential for pathogenesis [196]. Subspecies *stewartii* possesses two additional T3SSs: one is involved in the bacterium's ability to colonise the gut of the corn flea beetle, while the function of the other is not yet fully

elucidated [195]. Stewartan production is responsible for vascular streaking, bacterial oozing, and wilting. *P. stewartii* possesses one quorum-sensing system that regulates stewartan production and thus causes pathogenicity [176]. Another mechanism by which quorum sensing is implicated in pathogenicity is surface motility, which contributes to virulence in plants. The bacterium exhibits flagella-dependent surface motility, which is involved in biofilm development and plays a significant role in colonising the plant host [197]. The genome of *P. stewartii* contains genes that encode several putative endoglucanases, xylanases, and a β-1,4 β-1,3 mixed-linkage glucan glucanohydrolase that can digest the plant cell wall. These enzymes could play a critical role in enabling the bacteria to access carbohydrates associated with xylem cell wall structure and development [198].

The first complete genome sequence of *P. stewartii* subsp. *stewartii* is that of the widely used laboratory strain DC283. This revealed that the bacterium possesses numerous mobile genetic elements [199]. In addition to a 4.53 Mb circular chromosome, the genome contains ten circular plasmids ranging in size from 4,277 to 304,641 bp and one linear phage plasmid (ppDSJ01), which is related to the *E. coli* N15 prophage. The two smallest plasmids, pDSJ01 and pDSJ02, are present in medium copy numbers (around 30 each), while the others are present in low copy numbers (fewer than 10 copies). Furthermore, *P. stewartii* subsp. *stewartii* has a large number of repetitive transposase sequences, which present a challenge in genome assembly when only short reads are used. The chromosome also contains numerous integrative and conjugative elements and prophage sequences. These elements contribute to genetic diversity and may carry virulence factors or genes that confer resistance to environmental stresses. MGEs present in subspecies *stewartii* play critical roles in its evolution, adaptability, and pathogenicity. Notably, the majority of the MGEs present in *P. stewartii* subsp. *stewartii* are absent in other *P. stewartii* strains, suggesting that the recent evolution of *P. stewartii* subsp. *stewartii* is largely due to the acquisition of these elements [15]. One major exception is the large *Pantoea* plasmid 1 (LPP-1), which is present in various *Pantoea* species and in subspecies *stewartii* as plasmid pDSJ10 [200]. This plasmid plays a key role in expressing genes associated with metabolism, stress responses, and virulence, thereby contributing to the bacterium's ability to thrive in diverse ecological niches. Interestingly, other strains of *P. stewartii* that do not belong to subspecies *stewartii* also contain a functional type VI secretion system, which is present in other *Pantoea* species [201].

1.4.2 *Xylophilus ampelinus*

Xylophilus ampelinus is the causative agent of bacterial blight in grapevines (*Vitis vinifera*) and was first isolated in Crete, Greece, in 1969. It is an aerobic, non-spore-forming, rod-shaped, monotrichously flagellated, Gram-negative, beta-proteobacterium that produces a yellow, water-insoluble pigment and metabolises sugars oxidatively [202]. At the time of its isolation, the bacterium was designated *Xanthomonas ampelina*; however, it was later transferred to the newly established genus *Xylophilus*, based on DNA-DNA hybridisation studies [203]. The only other described member of this genus is *X. rhododendri*, which was recently isolated from *Rhododendron schlippenbachii* in South Korea [204]. However, other studies have also reported bacterial isolates that are most likely part of the *Xylophilus* genus, based on genetic identification using the 16S rRNA sequence [205], [206]. Moreover, bacteria from the genus *Xylophilus* have also been reported in multiple metagenomic studies, and most of the currently available genomic data on this genus originates from these studies [207].

X. ampelinus is widely distributed throughout the wine-growing regions of Europe, where it has been isolated in Greece, Italy, Spain, France, Moldova, and Slovenia [208]. Beyond Europe, it has also been identified in Jordan [209], South Africa [210], and Japan [174]. The bacterium can survive and overwinter in the vascular system of infected plants, causing a systemic infection of the xylem tissue. It can spread via moisture to other plants through wounds or leaf scars; however, the disease can also spread without wounding. Disease transmission can occur via infected tools and machinery used for grafting, as well as pruning knives. The bacterium can be transmitted over long distances via infected propagation material, such as cuttings used for rooting or grafting [211]. Symptoms include necrotic leaf lesions, cracks, and cankers on infected shoots, and brown discolouration of xylem tissue, as well as dead branches. The severity of the symptoms exhibited by the plants can vary significantly from year to year, depending on the prevailing climatic conditions. Latent infections are very common, with up to 50% of plants in a vineyard being asymptomatic, which makes the disease difficult to control [208]. Outbreaks are sporadic and can cause significant economic damage.

The symptoms caused by *X. ampelinus* are non-specific, making laboratory testing necessary to confirm the infection. The bacterium can be isolated from plants showing symptoms of necrosis and canker throughout the year, but it is difficult to isolate from samples collected during hot, dry periods [211]. Isolation can also be very challenging due to the bacteria's poor and slow growth and the lack of selective growing media. For these reasons, molecular methods that rely on DNA amplification are usually employed for routine testing of symptomatic grapevine plants for infection with *X. ampelinus*. Examples of these methods include a qPCR developed by Dreo et al. [208], a PCR by Manceau et al. [212], which can be combined with an enzyme-linked immunosorbent assay (ELISA)-based signal amplification system, nested PCR by Botha et al. [103], and multiplex PCR by Carminati et al. [104]. The qPCR assay by Dreo et al. and the PCR assay by Manceau et al. are based on the Xamp 1-27A fragment, whereas the nested PCR assay by Botha et al. is based on the intergenic spacer region of 16S-23S rDNA [103]. A significant aspect of the investigations pertains to international trade in plant material, where positive findings can have substantial economic consequences, such as the imposition of an import ban.

1.4.3 *Xanthomonas translucens* pv. *undulosa*

The *Xanthomonas translucens* group comprises Gram-negative, plant-pathogenic bacteria that cause disease in cereal crops and forage grasses. *X. translucens* pv. *undulosa* belongs to the translucens group of pathovars that cause bacterial leaf streak disease. This pathogen has been reported in many regions worldwide where wheat is cultivated, including the Americas, Africa, Asia, and Europe, and is believed to have spread relatively recently due to its low genomic diversity. Yield losses depend on environmental factors and the susceptibility of the wheat cultivars grown, typically amounting to less than 10%. However, losses can be as high as 40% in the event of a severe outbreak, making *X. translucens* pv. *undulosa* potential threat to global wheat production [213]. There are currently no effective chemical agents against this bacterium, and disease management primarily focuses on preventing outbreaks [214]. *X. translucens* pv. *undulosa* is best known for causing disease in wheat, but it has a fairly broad host range that includes several small grains and perennial grasses [215]. Those alternative hosts are also suspected of playing a role in the establishment and spread of the disease by serving as overwintering hosts and green bridges. However, the infected seeds are thought to be the primary means by which *X. translucens* pv. *undulosa* is introduced to new areas [213]. Symptoms of infection primarily appear on the leaves and spikes. On leaves, these manifest as water-soaked

streaks that develop into translucent lesions and bacterial ooze. When the plant is under high disease pressure, symptoms can affect the entire leaf. Type III secreted effector proteins play a crucial role in the pathogenicity of *Xanthomonas* bacteria in plant hosts, with transcription-activator-like effectors (TALEs) being among the most studied of these proteins. Following their injection into plant cells, these effector proteins localise in the nucleus, where they contribute to the development of symptoms and bacterial growth by inducing host susceptibility genes. The presence of different TALEs also affects the range of host species for the bacterium [216].

1.5 Aims of the Research

1.5.1 Evaluation of targeted HTS for the determination of microbiome and resistome in sputum samples

Target HTS is a widely used method for determining the microbiome, with the 16S rRNA region being the most frequently used target sequence for bacterial classification. However, other sequences that enable the more specific classification and characterization of bacteria can also be used in targeted HTS, with genes associated with AMR being one such example. Targeted HTS has great potential for many applications, including clinical diagnostics, but it has not yet been widely adopted. The process of determining the microbiome includes multiple steps that can introduce additional bias, affecting the results and their repeatability. One of the most critical steps in this process is DNA extraction from the sample, making the choice of extraction method especially important.

To facilitate the implementation of targeted HTS, we conducted a study in which we evaluated the effects that different DNA extraction methods have on the results of microbiome and resistome determination, as well as the detection of specific bacterial species. We selected three methods utilising different principles, including CTAB-based, silica column-based, and magnetic bead-based DNA extraction, and examined their impact on the limit of detection, diversity, and repeatability. We conducted our research on sputum, which, despite being used very often in clinical diagnostics, poses a special challenge for DNA extraction due to its high heterogeneity and rich microflora.

Our primary objective was to assess the impact of DNA extraction methods on the results of sputum microbiome and resistome analysis using targeted HTS, as well as their repeatability, based on principles from diagnostics and metrology. Using triplicates on two separate days enabled us to evaluate the repeatability of targeted HTS and its ability to reliably determine the microbiome and resistome.

1.5.2 Applying whole genome sequencing technologies to plant pathogenic bacteria to perform comparative genomic studies

The data generated by WGS can be used to perform a variety of analyses in the field of bacterial genomics. These include taxonomic classification of new isolates, phylogenetic analysis to track the spread of new pathogens, and comparative genomic studies to identify differences between strains, subspecies, species, or other taxonomic levels. In order to uncover the true value of genomic data, it is necessary to contextualise it by including metadata on the analysed strains, which enables us to answer important biological questions.

Research presented in this thesis includes two examples where WGS data were used to study plant pathogenic bacteria. In the first and major study, we analysed multiple strains of *P. stewartii* subsp. *stewartii*, a plant-pathogenic bacterium that causes Stewart's wilt in maize. This was motivated by the repeated discovery of *P. stewartii* subsp. *stewartii* in Slovenia and elsewhere in Europe in recent years. This prevalence was somewhat surprising, given that *P. stewartii* subsp. *stewartii* is classified as a quarantine microorganism in the EU and should not be present in this area. As part of our study, we sequenced the genomes of Slovenian isolates of *P. stewartii* subsp. *stewartii*, as well as multiple historical isolates of this bacterium. The newly generated genomic data was used alongside genomes already available from the public databases to perform comparative genomic studies. This study aimed to elucidate the origin of Slovenian *P. stewartii* subsp. *stewartii* isolates and identify how they compare to strains from other geographic areas. The genomes were also used to conduct a comprehensive genomic analysis of the *P. stewartii* bacterial species and its subspecies *stewartii*.

The second minor study focused on *X. translucens* pv. *undulosa*, which causes bacterial leaf streak disease in cereal crops and forage grasses. The genome of the South American isolate was sequenced and compared to other complete genomes in databases. The study aimed to elucidate the phylogenetic relationship between strains and to explore the differences in pathogenicity factors present in the studied genomes.

1.5.3 Development and validation of novel and genome-informed molecular tests for the detection of plant pathogenic bacteria

Target bacteria can be detected using molecular PCR-based tests, which detect target bacteria by amplifying species-specific sequences and offer many advantages over other methods. These methods are rapid and easy to perform in routine diagnostic settings, and with an available genome, can be applied for the detection of basically any bacterium. HTS-derived genomic data present a pool of potential unique nucleotide markers needed for the development of new specific molecular tests.

In the major study, we designed multiple new qPCR tests to detect *X. ampelinus*, a bacterial pathogen of grapevines. Due to this bacterium's fastidious growth under laboratory conditions, detection mostly relies on molecular methods. However, previously developed tests were reported to produce conflicting results, which limits their use for diagnostics. We used genomic data from publicly available databases to identify sequences unique to *X. ampelinus*, using them to design new qPCR tests. The novel tests were then evaluated in the laboratory to determine their efficiency, specificity, and sensitivity. This study aimed to design and evaluate multiple qPCR tests that would enable the specific and sensitive detection of *X. ampelinus* in various plant matrices, including leaves, roots, and xylem.

In the second minor study, we used a genome-informed approach to design a qPCR test that would be able to specifically detect *P. syringae* pv. *aptata* in samples of sugar beet and in environmental samples. Newly designed tests were evaluated for their specificity and sensitivity in plant samples and water from irrigation channels.

1.6 Research Hypotheses

1. The DNA extraction method has an impact on the results of microbiome and resistome determination of sputum samples performed using targeted HTS.

2. Results of microbiome and resistome determined by targeted HTS are repeatable when the data generated is of high enough quality.
3. Slovenian finding of *P. stewartii* subsp. *stewartii* are a consequence of multiple introduction events as determined by comparative genomics and phylogenetic analysis based on WGS data.
4. Publicly available data is sufficient to be used to design qPCR tests capable of the specific detection of *X. ampelinus* in different grapevine samples.

1.7 Publications Included and Candidate`s Contribution

The first publication (*Metrological evaluation of DNA extraction method effects on the bacterial microbiome and resistome in sputum*) examined the effects of three different DNA extraction methods on the results of microbiome and resistome determination in sputum samples performed using targeted HTS. Despite being challenging material due to its high heterogeneity and diverse microflora, sputum is very often used in clinical diagnostics. The sputum samples used in this study were collected from patients with pulmonary diseases at the University Clinic Golnik and were spiked with bacteria at known concentrations, enabling the sensitivity of targeted HTS to be determined. DNA was extracted in triplicate on two separate days and used to evaluate the repeatability of targeted HTS for determining the microbiome and resistome. DNA extraction method showed to have significant effect on the results of the diversity and composition of the sputum microbiome and resistome but had no effect on the limit of detection of spiked bacteria. Targeted HTS proved to be a reliable method for determining the microbiome and resistome; however, the DNA extraction method should be carefully chosen to introduce as little bias as possible into the results. The PhD candidate contributed to this work by preparing sequencing libraries, analysing microbiome and resistome data, writing the manuscript, and preparing figures and tables.

The second publication (*Comparative genomics and pathogenicity of* Pantoea stewartii *subsp.* stewartii *reveal multiple introductions and limited distribution in Europe*) describes a comparative genomics study of *P. stewartii* subsp. *stewartii*. This is a plant-pathogenic bacterium that causes Stewart's wilt in maize and was first confirmed in Slovenia in 2018. In our study, we performed WGS of Slovenian isolates and strains from bacterial collections belonging to *P. stewartii* subsp. *stewartii*. These newly obtained sequences were used alongside publicly available *P. stewartii* genomes to perform comparative genomics analyses. These analyses included a phylogenetic study of *P. stewartii*, which first elucidated the relationship between the Slovenian isolates and other strains and determined the position of the subspecies stewartii within the species *P. stewartii*. Furthermore, we examined the differences between Slovenian isolates and other strains, as well as the differences between *P. stewartii* subsp. *stewartii* and other *P. stewartii* strains. Particular attention was paid to the presence of various pathogenicity factors in different *P. stewartii* strains. The findings from the genomic study were also examined in a greenhouse experiment. PhD candidate contributed to this work by preparing material used in WGS, performing nanopore sequencing, analysing raw sequencing data, performing comparative genomics analysis, and conducting greenhouse experiments. He analysed the results, wrote the manuscript, and prepared figures and tables.

The third publication (*Complete genome sequence resource for* Xanthomonas translucens *pv.* undulosa *MAI5034, a wheat pathogen from Uruguay*) describes the sequencing of the *X. translucens* pv. *undulosa* genome using nanopore technology. This

was the first time an isolate of this bacterium from South America had been sequenced. *X. translucens* is a plant-pathogenic bacterium responsible for leaf streak, black chaff of small grains, and wilt of forage grasses. The newly obtained genome, consisting of one circular chromosome (4.6 Mb), was compared to existing genomes in databases to identify TALEs and to conduct phylogenetic analysis. The results showed lower TALE diversity in pathovar *undulosa*, indicating its recent worldwide spread, which is consistent with its lower genetic diversity. The PhD candidate contributed to the bioinformatic analysis of the sequencing data, including genome assembly and phylogenetic analysis.

The fourth publication (*Development of a multi-targeted real-time PCR assay for the detection of grapevine pathogen* Xylophilus ampelinus) describes the design and evaluation of new qPCR tests for the specific and sensitive detection of *X. ampelinus* in different grapevine samples. *X. ampelinus* is a bacterial grapevine pathogen known for its fastidious growth under laboratory conditions and is usually detected using molecular methods. As the current tests occasionally produce inconsistent results, we developed multiple new qPCR tests to enable reliable detection. To achieve this, we used a genome-informed approach, utilising data available in public databases, to design specific sets of primers and probes. The newly designed tests were evaluated for efficiency, specificity, and sensitivity in grapevine leaf, root, and xylem samples. The performance of the tests was further evaluated in multiple laboratories. The PhD candidate contributed to this work by performing *in silico* design, sample preparation, and laboratory testing. He analysed the results, wrote the manuscript, and prepared the figures and tables.

The fifth publication (*Irrigation systems as reservoirs of diverse and pathogenic* Pseudomonas syringae *strains endangering crop health*) describes the development and testing of a new qPCR for the detection of *P. syringae* phylogroup 2. Bacteria from the species *P. syringae* are well-known plant pathogens with a wide range of hosts and are present in a variety of environments. Those include water habitats, which present a special danger to agriculture if the bacteria are present in water used for irrigation. Multiple strains of the *P. syringae* complex used were isolated from the Danube-Tisa-Danube Hydrosystem in Serbia. Using available genomic data, new qPCR tests were designed for the specific detection of *P. syringae* phylogroup 2. The tests were evaluated using Serbian isolates and strains from the bacterial collection to determine their specificity and sensitivity. The test, which showed the most promising results, was also adapted for absolute quantification as dPCR and used for the detection of *P. syringae* phylogroup 2 in water from irrigation canals. The results obtained by the new test were compared to results from other detection methods. The PhD candidate was involved in the *in silico* design, sample preparation, and laboratory testing.

Chapter 2

Scientific Publications

2.1 Metrological Evaluation of DNA Extraction Method Effects on the Bacterial Microbiome and Resistome in Sputum

Aleksander Benčič, Nataša Toplak, Simon Koren, Alexandra Bogožalec Košir, Mojca Milavec, Viktorija Tomič, Dane Lužnik, Tanja Dreo

mSystems, 2024; 9(9): e0073524. doi: 10.1128/msystems.00735-24

This article evaluates the impact of DNA extraction methods on the results of microbiome and resistome obtained using targeted HTS. Sputum samples spiked with bacteria *Acinetobacter baumannii*, *Klebsiella pneumoniae,* and *Pseudomonas aeruginosa* at known concentrations (10^3–10^6 cells/mL) were used. DNA was extracted from each spiked sputum sample on two separate days, in triplicate, using three different methods: CTAB-based, magnetic bead-based, and silica membrane-based. The concentrations of the spiked bacteria were determined using digital PCR (dPCR). 16S rDNA sequencing was used to determine the microbiome, while species-specific amplicons were used to detect a list of clinically relevant bacteria. The resistome was determined by sequencing amplicons associated with antimicrobial resistance. Approaches from metrology and diagnostics were employed to analyse and interpret the results. All three spiked bacteria were detected in the sputum samples, and the method of DNA extraction had no significant effect on the detection results. However, the DNA extraction method was found to affect the results of the microbiome and resistome determination, with repeatable results in most cases. The results obtained using the silica membrane-based DNA extraction kit were the most consistent and demonstrated the greatest microbiome and resistome diversity. Targeted HTS was shown to be a reliable tool for determining the microbiome and resistome. However, the DNA extraction method must be carefully selected to minimise its impact on the results.

The PhD candidate is the first author of this publication. He contributed to experiment planning, laboratory work, performed statistical analysis of the microbiome and resistome results, and compared the results obtained by targeted HTS with those obtained using other methods. He wrote the manuscript, including the supplementary material, and prepared the figures and tables.

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems

| Human Microbiome | Research Article

Metrological evaluation of DNA extraction method effects on the bacterial microbiome and resistome in sputum

Aleksander Benčič,[1,2] Nataša Toplak,[3] Simon Koren,[3] Alexandra Bogožalec Košir,[1] Mojca Milavec,[1] Viktorija Tomič,[4] Dane Lužnik,[4] Tanja Dreo[1]

AUTHOR AFFILIATIONS See affiliation list on p. 15.

ABSTRACT Targeted high-throughput sequencing (HTS) has revolutionized the way we look at bacterial communities. It can be used for the species-specific detection of bacteria as well as for the determination of the microbiome and resistome and can be applied to samples from almost any environment. However, the results of targeted HTS can be influenced by many factors, which poses a major challenge for its use in clinical diagnostics. In this study, we investigated the impact of the DNA extraction method on the determination of the bacterial microbiome and resistome by targeted HTS using principles from metrology and diagnostics such as repeatability and analytical sensitivity. Sputum samples spiked with *Acinetobacter baumannii*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa* at three different concentrations (10^3–10^6 cells/mL) were used. DNA was extracted from each sample on 2 separate days in three replicates each using three different extraction methods based on cetrimonium bromide, magnetic beads, and silica membranes. All three spiked bacteria were detected in sputum, and the DNA extraction method had no significant effect on detection. However, the DNA extraction method had significant effects on the composition of the microbiome and the resistome. The sequencing results were repeatable in the majority of cases. The silica membrane-based DNA extraction kit provided the most repeatable results and the highest diversity of the microbiome and resistome. Targeted HTS has been shown to be a reliable tool for determining the microbiome and resistome; however, the method of DNA extraction should be carefully selected to minimize its impact on the results.

IMPORTANCE High-throughput sequencing (HTS) is one of the crucial new technologies that gives us insights into previously hidden parts of microbial communities. The DNA extraction method is an important step that can have a major impact on the results, and understanding this impact is of paramount importance for their reliable interpretation. Our results are of great value for the interpretation of sputum microbiome and resistome results obtained by targeted HTS. Our findings allow for a more rational design of future microbiome studies, which would lead to higher repeatability of results and easier comparison between different laboratories. This could also facilitate the introduction of targeted HTS in clinical microbiology for reliable identification of pathogenic bacteria and testing for antimicrobial resistance (AMR). As AMR is a major threat to public health, the improved methods for determining AMR would bring great benefits to both the healthcare system and society as a whole.

KEYWORDS targeted high-throughput sequencing, bacterial microbiome, resistome, bacteria detection, DNA extraction

The microbiome of the lower respiratory tract plays an important role in the development and progression of lung diseases and the development of bacterial infections (1–3). Genes associated with antimicrobial resistance (AMR) are also part of

Editor Naseer Sangwan, Cleveland Clinic, Cleveland, Ohio, USA

Address correspondence to Aleksander Benčič, aleksander.bencic@nib.si.

The authors declare no conflict of interest.

See the funding table on p. 15.

Received 17 June 2024
Accepted 9 July 2024
Published 16 August 2024

the lung microbiome and define the so-called "resistome." The term "resistome" is similar to the term "microbiome" and includes all genes associated with AMR. The role of the lung resistome is not yet well understood. The AMR genes in bacteria that make up the microbiota of healthy lungs form the core resistome, which is also a reservoir from which pathogenic bacteria can acquire AMR (4, 5).

Over the last decade, enormous progress has been made in microbiome and resistome research through the development of high-throughput sequencing (HTS) methods. These have enabled culture-independent microbiome analysis, in particular by targeted sequencing of the hypervariable regions of 16S ribosomal RNA (rRNA) (6, 7). Similarly, targeted HTS can also be used to detect genetic elements associated with AMR or horizontal gene transfer of such elements (8–10). However, when studying the lung microbiome and resistome, there can be variability in each step of the required procedures. These include sample collection, storage, DNA extraction, PCR amplification, library preparation, sequencing, and bioinformatic analysis (11, 12). DNA extraction has been shown to be one of the main sources of variability and bias in microbiome analysis and is therefore a critical step in the process. Therefore, the choice of DNA extraction method is an important part of the design of any study to determine the microbiome and resistome (7, 11, 13–17).

Sputum is an attractive source material for lung microbiome studies as it can be collected relatively easily and non-invasively (18). However, it is also a complex and heterogeneous matrix. The vast majority of DNA isolated from sputum is host DNA. To overcome these challenges, various methods for sample preparation and DNA extraction have been developed. Generally, solubilizers are used prior to DNA extraction from sputum to reduce differences between samples and increase DNA yield (19). Mechanical, enzymatic, and chemical methods can be used to ensure adequate cell lysis (16). Different approaches to sample preparation and DNA extraction can lead to differences in the determined composition of the microbiome. This also complicates the comparison of results from different studies and hinders a possible meta-analysis of these data. The comparison of different DNA extraction methods is further complicated by the fact that the actual composition of the microbial community is not known and it is also not known which DNA extraction method provides results that are closest to the actual state (20). To date, there are only a limited number of studies that have addressed the determination of the lung microbiome and resistome using sputum and there is a general lack of data on the effects of different DNA extraction methods on the composition (16). To facilitate the adoption of HTS in clinical diagnostics of respiratory diseases, it is of great importance that possible impacts of DNA extraction methods are identified and most suitable methods are recognized and implemented in practice.

The main objective of the present study was to investigate the impact of three DNA extraction methods on the results of targeted HTS and their repeatability and reproducibility from sputum samples. To this end, the principles of metrology and diagnostics were used to evaluate the results. Spiked sputum samples were used from which DNA was extracted on 2 different days in three replicates to perform targeted HTS. For DNA extraction, the following methods were used: an in-house cetrimonium bromide (CTAB) solution-based extraction and two commercially available kits for solid phase extraction, one based on magnetic beads and an extraction robot (GXT NA/Arrow), and the other on silica membranes (QIAamp DNA mini kit). Targeted HTS was used to determine the microbiome based on 16S rRNA, to determine the resistome, and for species-specific detection of bacteria. The target bacteria were *Acinetobacter baumannii*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa* (Fig. 1). The repeatability of the DNA extraction methods was determined as the coefficient of variation (CV; %) of different quantities for the samples from which the DNA was extracted on the same day using the same extraction method.

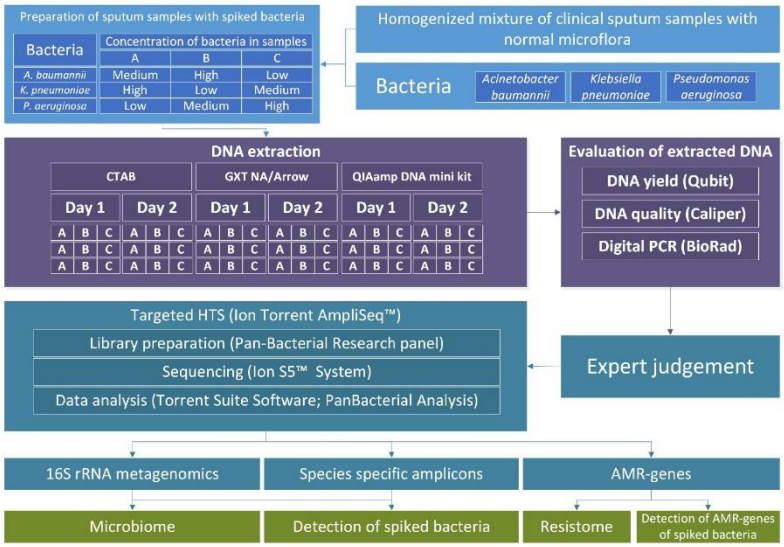


FIG 1 Scheme of the study design describing samples A, B, and C preparation from homogenized sputum samples and bacterial suspensions. DNA was extracted using three different DNA extraction methods on 2 separate days in three technical repeats. DNA yield and quality were analyzed prior to library preparation. From data obtained by targeted HTS, we detected spiked bacteria and AMR genes belonging to these bacteria, and determined microbiome and resistome.

MATERIALS AND METHODS

Sample preparation

The samples used in this study were prepared as part of the study conducted by Bogožalec Košir et al. (21). The sputum samples were collected from patients with lung diseases at the University Clinic of Respiratory and Allergic Diseases (Golnik, Slovenia). All sputum samples tested negative for the bacteria *A. baumannii*, *K. pneumoniae*, and *P. aeruginosa* with culture-based methods used in standard clinical diagnostics. In total, 16 sputum samples containing normal mixed lung microflora were pooled and digested with mucolytic reagent (1:1, vol/vol; Liquillizer; MetaSystems Hard & Software GmbH, Altlussheim, Germany). Bacterial suspensions were prepared from three bacteria: *A. baumannii* (DSMZ 30007), *K. pneumoniae* (DSMZ 30104), and *P. aeruginosa* (DSMZ 50071), which were obtained from the Deutsche Sammlung von Mikroorganismen und Zellkulturen (German Collection of Microorganisms and Cell Cultures). Spiked sputum samples were prepared by mixing aliquots of pooled and digested sputum with bacterial suspensions. Based on turbidity measurement, samples A, B, and C were prepared as randomized low, medium, and high concentrations (5.5×10^3, 5.5×10^4, and 5.5×10^5 cells/mL, respectively) of *A. baumannii*, *K. pneumoniae*, and *P. aeruginosa* (Fig. 1; Table S1). The concentrations of the bacteria in the suspensions and the spiked sputum samples were additionally determined by direct digital PCR (without DNA extraction) (see "Digital PCR" below). These were the assigned concentrations of the three bacteria in the range of 10^3 to 10^6 cells/mL, which were used in the subsequent analysis (Table S2). A detailed description of samples preparation and DNA extraction is in Fig. S1 and S2, respectively.

DNA extraction methods

DNA was extracted and purified from each spiked sample (200 µL) in three technical replicates on 2 separate days using the three different DNA extraction protocols. Negative controls (200 µL molecular grade water) were also extracted for each DNA extraction method on both extraction days. DNA extraction from the sample was performed as described by Bogožalec Košir et al. (21). The first method was a CTAB-based DNA extraction protocol adapted from Devonshire et al. (22). Prior to extraction, lysozyme (10 µL at 50 mg/mL) was added to each sample and incubated overnight at 37°C. At the end of the extraction, the pellets were rehydrated in 200 µL Tris-EDTA buffer overnight at 4°C and stored at −20°C. The second method was the GXT NA (Hain Lifescience GmbH, Tübingen, Germany) kit, which was used in combination with an automated nucleic acid extraction system (Arrow; NorDiag ASA, Bergen, Norway). DNA was eluted in 100 µL, according to the manufacturer's instructions. The third method was the QIAamp DNA mini kit (Qiagen, Hilden, Germany). The DNA was eluted from the spin columns with 100 µL Buffer AE (included in the kits).

DNA extraction from bacterial suspensions

DNA was also extracted from pure bacterial suspensions of *A. baumannii*, *K. pneumoniae*, and *P. aeruginosa* using QIAamp DNA mini kit extraction protocol as described by Bogožalec Košir et al. (21). One sample (100 µL, 10^8 cells/mL) of the suspension was extracted from each bacterium.

DNA yield, concentration, and quality

Qubit dsDNA HS assay kits and a fluorimeter (Qubit 3.0; both Invitrogen, Carlsbad, CA, USA) were used to estimate DNA extraction yields (micrograms) and to measure the concentration of extracted DNA (nanograms per microliter). The quality of the extracted DNA was determined by its fragmentation profile using the Genomic DNA analysis on the LabChip GX capillary gel electrophoresis instrument (PerkinElmer, Inc., Hopkinton, MA, USA) and capillary gel electrophoresis. DNA fragmentation is expressed as a genomic quality score (GQS) ranging from 0 to 5, with lower values corresponding to higher DNA fragmentation.

Digital PCR

Digital PCR was used to determine the bacterial concentration in the suspension prior to extraction and to determine the bacterial DNA concentration in extracts of sputum samples. The dPCR and subsequent data analysis were performed as described by Bogožalec-Košir et al. (21). The sequences of the primers and probes used for dPCR are listed in the Table S3. The dPCR experiments were performed on QX100/QX200 platform (BioRad). The dPCR mixtures had a total volume of 20 µL, which included 10 µL ddPCR supermix for probes (no dUTP), 6 µL primers and probe mix, and 4 µL DNA sample or bacterial suspension. Amplification conditions were 10 min DNA polymerase activation at 95°C, followed by 40 cycles of a two-step thermal profile of 30 s at 94°C for denaturation, and 60 s at 60°C for annealing and extension, followed by 10 min at 98°C, and then cooling to 4°C. Data were analyzed using the software package provided with the dPCR system (QuantaSoft 1.7.4.0917; BioRad) and Microsoft Excel. Reactions with droplet counts <8,000 per 20 µL PCR were excluded. Each suspension was tested in quadruplicate. Each dPCR run included a non-template control (water) and a positive control for each of the assays (synthetic DNA fragment with amplicon sequence for each assay).

Targeted high-throughput sequencing

Targeted HTS was performed with an Ion Torrent platform (Thermo Fisher Scientific, Waltham, MA, USA) using ion semiconductor chemistry. The Ion AmpliSeq Pan-Bacterial

Research Panel (Thermo Fisher Scientific), a community panel containing two primer pools for library construction, was used. The first pool consists of primer sets that enable species-specific detection of bacterial species and primers that target genes associated with antimicrobial resistance. The second pool consists of primers for the amplification of 16S rRNA regions.

Species-specific and antibiotic resistance amplicons

The first pool of the Pan-Bacterial Research Panel contains primers for 269 amplicons for the specific detection of 21 microbial species (Table S4). These include *A. baumannii*, *K. pneumoniae*, and *P. aeruginosa*, as used here. The other 18 microorganism species included are also known as possible pathogens. In addition, the first pool also contains primers for 716 amplicons belonging to 364 AMR genes associated with resistance to 31 different classes of antibiotics.

16S rRNA sequencing

The second pool of the Pan-Bacterial Research Panel contains 24 amplicons for 16S rRNA profiling, which are arranged along the entire sequence of the 16S rRNA and designed in such a way that as many different bacteria as possible can be detected. The determination of the operational taxonomic units (OTUs) is based on approximately 400,000 16S rRNA sequences from the public Greengenes database, to which the reads are mapped after sequencing.

Library preparation

The Ion AmpliSeq Library Kit 2.0 (Thermo Fisher Scientific) was used to prepare the libraries using 10 ng of DNA extracted from the sputum samples. In cases where the DNA was too diluted to provide 10 ng, the maximum possible volume of DNA was added (3 µL). This was only the case for the negative controls of the DNA extraction and the positive controls (DNA extracted from bacterial suspensions). In addition, a negative control without template was also used for amplification in order to exclude possible contamination during PCR amplification. The composition of the PCR mixture for the 10 µL reaction was 2 µL 5 × Ion AmpliSeq HiFi mix, 5 µL primer pool, 10 ng (1–3 µL) DNA, and nuclease-free water (to 10 µL). The PCR profile was as follows: activation of the enzyme at 99°C for 2 min; 15 cycles of denaturation at 99°C for 15 s; and annealing and extension at 60°C for 8 min. Amplifications of the first and second pools of the Pan-Bacterial Research Panel were performed in separate reactions, which were then combined for further library preparation. Library preparation was performed according to the manufacturer's protocol, and during these steps, the adapters containing the barcodes from the Ion Xpress Adapters barcode kits were also ligated to the amplicons. The libraries were purified with magnetic beads (Agencourt AMPure XP; Beckman Coulter, Brea, CA, USA) according to the manufacturer's protocol. All libraries were quality checked and quantified using the DNA High Sensitivity assay on the LabChip GX capillary gel electrophoresis instrument (PerkinElmer, Inc., Hopkinton, MA, USA). Emulsion PCR for clonal amplification of the libraries on ion spheres, enrichment for spheres containing library, and loading of the chip were performed using the Ion Chef System (Thermo Fisher Scientific). For ion semiconductor sequencing, the Ion 530 chip and the Ion S5 system were used (Thermo Fisher Scientific).

Data analysis

The raw reads obtained during sequencing were analyzed using the Torrent Suite software PanBacterialAnalysis plug-in, which was specially developed for the analysis of reads from the Pan-Bacterial Research Panel. Reads with a length under 70 bp were filtered out and labeled as "invalid" in the output. Reads were classified by mapping against reference sequences within the plug-in (Greengenes public database for 16S rRNA sequences). Reads with local alignment scores over 35 were used and counted in

the further analysis. Reads that were not aligned to any sequence in the database were reported as "unmapped." Read counts for AMR genes and species-specific amplicons and 16S families were normalized separately. The normalized read counts for AMR genes and species-specific amplicons are reported as the ratio between the sum of the read counts of all amplicons for the species or gene of interest and the total read counts in that sample. The normalized 16S rRNA read count is reported as the ratio between the read counts of the family and the total reads mapped to the 16S rRNA reference in that sample. Only the AMR genes and the species-specific amplicons with normalized read counts over 0.1 and the 16S families with normalized read counts over 1.0 are reported as present in the samples. For 16S rRNA metagenomics, the results are reported as bacterial OTUs detected in the samples. For species-specific amplicons and for AMR genes, results are reported as the number of reads belonging to each amplicon. After automated data analysis, a threshold of 10 reads was applied for AMR genes and species-specific amplicons. Only AMR genes and bacterial species that reached this threshold were used for further analysis.

Statistical data analysis

Kruskal-Wallis tests were performed to determine whether there was a statistically significant difference in DNA yield, genomic quality score, alpha-diversity and number of AMR genes due to the different DNA extraction methods, extraction days, and spiked sputum samples. The Benjamini-Hochberg method was used to control the false discovery rate. The beta-diversity of the microbiomes was analyzed using principal coordinate analysis (PCoA) based on Bray-Curtis dissimilarities. A permutational multivariate analysis of variance (PERMANOVA) was performed on Bray-Curtis dissimilarities to assess the effects of DNA extraction methods, extraction days, sputum samples and replicates on beta-diversity, and a pairwise PERMANOVA for the pairwise comparisons of the different groups. The same methods used to analyze and compare the microbiomes were also applied to the resistomes. All analyses were performed using the R programming language for statistical computing (23) and the Rstudio integrated development environment (24) in combination with the following packages: ggplot2 (25) and phyloseq (26) for visualization of the results, and vegan (27) and RVAideMemoire (28) for the diversity, dissimilarity, and multivariate analyses (29).

RESULTS

DNA yields and fragmentation with the different extraction methods

DNA was extracted from sputum samples in three technical replicates, each on 2 separate days, using one of three different DNA extraction methods and sequenced to determine the repeatability of the methods (Table S1). DNA yield and GQS differed significantly between the three extraction methods (P-value < 0.001; Kruskal-Wallis t-tests). The highest amount of DNA was extracted with the CTAB protocol (overall mean, 10.6 µg), followed by the GXT NA/Arrow kit (overall mean, 4.5 µg) and the QIAamp DNA mini kit (overall mean, 1.8 µg). Although the yield of the QIAamp DNA mini kit is the lowest, the quantity is sufficient for HTS analysis and is of high quality (highest GQS, overall mean of 2.4) and the highest repeatability (Fig. S3; Table S5). All three kits extracted DNA of sufficient quality and quantity to be used for library preparation for targeted HTS. Results were considered repeatable when the CV was below 25%. In terms of yield, all three extraction methods gave repeatable results. For DNA quality and fragmentation, defined as GQS using gel capillary electrophoresis, CVs were >25% in some cases for the CTAB protocol and the GXT NA/Arrow kits, indicating less consistent quality of extracted DNA (Table S5).

Analysis of targeted high-throughput sequencing and alignment of raw reads

The total number of reads obtained with the targeted was 2.60×10^7, and the average number of reads per library was 4.81×10^5 (CV, 52.1%). Both the differences in the total number of reads and the differences in the proportions of invalid reads between the DNA extraction methods were not significant (Fig. 2A and B; Table S6). There were significant differences in the proportions of mapped and unmapped reads between the three DNA extraction methods (*P*-values < 0.001, Kruskal-Wallis *t*-tests). In the samples extracted with the QIAamp DNA mini kit, the proportion of mapped reads was the highest (overall mean 52%) and the proportion of unmapped reads the lowest (30%). On the other hand, the CTAB protocol had the lowest proportion of mapped reads (overall mean 24%) and the highest proportion of unmapped reads (overall mean 54%) (Fig. 2C through E). Although samples extracted with the QIAamp DNA mini kit had the highest average percentage of mapped reads and the highest GQS, there was no significant correlation between the percentage of mapped reads and GQS in samples extracted with the same method.

Controls

Negative controls used for DNA extraction and library preparation were negative throughout the process. Amplifiable DNA was successfully extracted from all sputum samples as determined by DNA measurements (yield and quality), dPCR, and targeted HTS (specific detection of spiked bacteria). While dPCR pointed to no contaminations, HTS detected *A. baumannii* DNA in a CTAB negative control. This indicates that the contamination occurred during library preparation. The complex sputum background was followed throughout the process as a positive control. In addition, HTS sequencing of bacteria used for spiking identified species-specific and AMR amplicons (Table S7), matching genomic information (GenBank assembly accessions: GCA_009759685.1, JOOW00000000.1, GCA_001045685.1 Table S8).

Analytical sensitivity of targeted HTS for detection of bacteria in sputum samples

The sputum samples were spiked with relatively low bacterial concentrations, which are typical for the early phase of infection and persistent infections. All spiked sputum samples tested positive for the presence of spiked bacteria using dPCR (Tables S9 to S11). Targeted HTS and species-specific amplicons were used to detect spiked bacteria in sputum samples. As expected, reads belonging to spiked bacteria represent only a small fraction of the microbiome (<4%). We were able to detect *K. pneumoniae* and *A. baumannii* at high concentrations (5.4×10^5 and 1.2×10^6 cell/mL, respectively) with each of the DNA extraction methods, regardless of the days and technical replicates (Table 1). At the lowest concentrations, which are below the limit of detection for targeted HTS, we did not detect neither *A. baumannii* nor *K. pneumoniae* in any sample. The detection of *A. baumannii* and *K. pneumoniae* was inconsistent at intermediate concentrations (Table 1). In contrast to the other two spiked bacteria, we could not detect *P. aeruginosa* repeatedly, but only in a few samples (Table 1).

Analytical sensitivity of targeted HTS for detection of AMR genes from spiked bacteria in sputum samples

Since spiked bacteria only make up a small part of the microbiome, the AMR genes belonging to these bacteria also only make up a small part of the resistome (<2%). All five AMR genes that were present in spiked bacteria were also detected in spiked sputum samples with high concentrations of spiked bacteria to which AMR genes belong. All genes were detected in samples of all extraction methods except for *macB* gene in samples extracted with CTAB. The percentages of detection for AMR genes ranged from 17 to 100% for the samples with high percentage of spiked bacteria. As expected, no

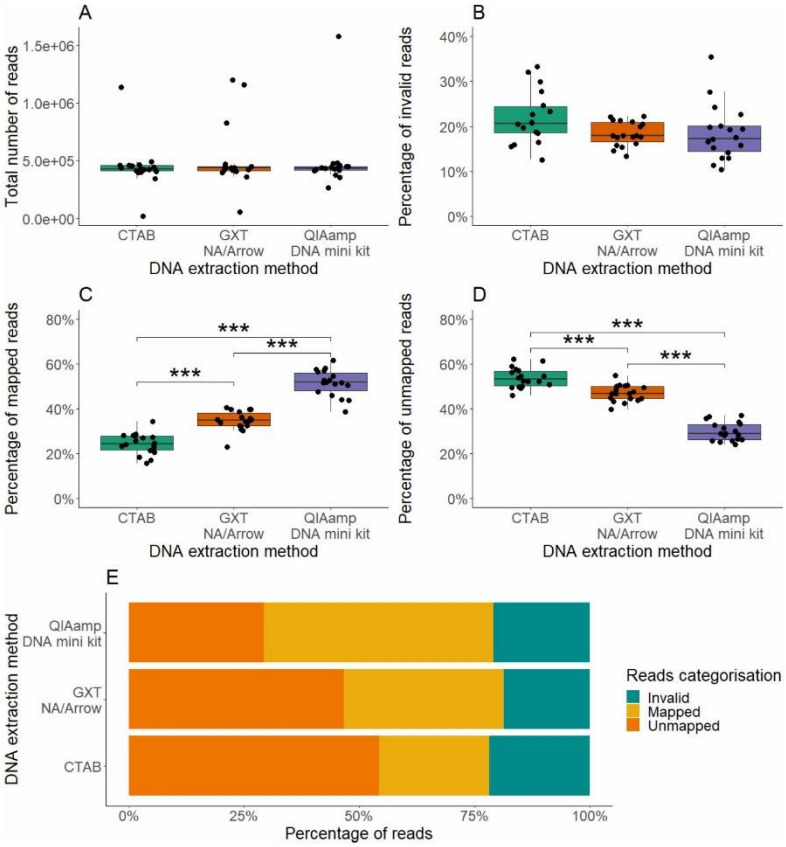


FIG 2 Results of data analysis after targeted HTS. Reads obtained for the high-throughput sequencing are colored according to the DNA extraction methods. (A–D) Boxplots showing medians and quartiles for the total number of reads (A), invalid reads (i.e., too short; not included in the analysis) (B), reads successfully mapped to reference sequences (C), and valid reads that did not map to any reference sequence (D). (E) Bar plot representations of the invalid mapped and unmapped reads for each extraction method (as indicated). ***, *P*-value < 0.001 (Kruskal-Wallis *t*-test).

AMR gene was detected in samples with low concentrations of spiked bacteria (Table 2). This indicates that these AMR genes are not part of the sputum microbiome and actually belong to the spiked bacteria. Compared to the detection of spiked bacteria with species-specific amplicons, some discrepancies were found. AMR genes were detected in a lower proportion of samples than *A. baumannii* and *K. pneumoniae*. In contrast, AMR genes from *P. aeruginosa* were detected in a much higher proportion of samples than the bacterium itself. This indicates that the overall sensitivity is related to the amplification efficiency of the different amplicons in a complex matrix. Similar to the detection of spiked bacteria, day of extraction and DNA extraction method did not show any significant effects on the detection of AMR genes from spiked bacteria.

TABLE 1 Analytical sensitivity of targeted HTS detection of bacteria added to sputums[a]

Bacteria	Concentration of bacteria (cp/mL)	DNA extraction method					
		CTAB		GXT NA/Arrow		QIAamp DNA mini kit	
		Number	%	Number	%	Number	%
Acinetobacter baumannii	1.20×10^6	6/6	100	6/6	100	6/6	100
	1.30×10^5	0/6	0	4/6	67	6/6	100
	1.30×10^4	0/6	0	0/6	0	0/6	0
Klebsiella pneumoniae	5.40×10^5	5/6	83	6/6	100	6/6	100
	5.30×10^4	0/6	0	1/6	17	2/6	33
	5.70×10^3	0/6	0	0/6	0	0/6	0
Pseudomonas aeruginosa	1.20×10^6	2/6	33	0/6	0	1/6	17
	1.30×10^5	1/6	17	0/6	0	0/6	0
	1.20×10^4	0/6	0	0/6	0	0/6	0

[a]Proportion of samples ($n = 6$) in which the spiked bacteria at the different concentrations (low, medium, high) in the sputum samples were detected, according to the DNA extraction methods.

Evaluation of microbiome with 16S rRNA metagenomics

The microbiome of spiked sputum samples was determined using targeted HTS of the 16S rRNA region. When looking at all sequenced samples together, we detected 80 different OTUs, 29 genera, and five phyla (Actinobacteria, Bacteroidetes, Firmicutes, Fusobacteria, and Proteobacteria) (Table S12; Fig. 3A and B). When looking at samples extracted by different method, a statistically significant effect on the microbiome was determined using Kruskal-Wallis *t*-test (Table S13). Beta-diversity of bacterial genera based on Bray-Curtis dissimilarities and two-dimensional PCoA were used for visualization. Samples were clustered by DNA extraction method, and there was some overlap between samples extracted with the QIAamp DNA mini kit and GXT NA/Arrow (Fig. 3). The DNA extraction method had a significant effect on the microbiome, explaining 63.2% of the total variance therein as calculated with PERMANOVA (Table S14). Richness (number of OTUs and genera) and diversity (Shannon's and Simpson's indices) were calculated for each sample. Samples extracted with the QIAamp DNA mini kit had significantly the highest richness and diversity (P-value < 0.001; Kruskal-Wallis *t*-tests, Table S15) and repeatability. In contrast, samples extracted using the CTAB protocol had the lowest richness and diversity (Fig. 4). No significant difference in richness and

TABLE 2 Analytical sensitivity of targeted HTS for detection of AMR genes[a]

Bacteria	AMR gene	Concentration of bacteria (cp/mL)	DNA extraction method					
			CTAB		GXT NA/Arrow		QIAamp DNA mini kit	
			Number	%	Number	%	Number	%
Acinetobacter baumannii	*eptA*	1.20×10^6	6/6	100	5/6	83	6/6	100
		1.30×10^5	0/6	0	0/6	0	0/6	0
		1.30×10^4	0/6	0	0/6	0	0/6	0
	uppP	1.20×10^6	4/6	67	5/6	83	4/6	67
		1.30×10^5	0/6	0	0/6	0	0/6	0
		1.30×10^4	0/6	0	0/6	0	0/6	0
Klebsiella pneumoniae	*macB*	5.40×10^5	0/6	0	1/6	17	1/6	17
		5.30×10^4	0/6	0	0/6	0	0/6	0
		5.70×10^3	0/6	0	0/6	0	0/6	0
Pseudomonas aeruginosa	*aph(3') iib*	1.20×10^6	5/6	83	2/6	33	3/6	50
		1.30×10^5	0/6	0	0/6	0	0/6	0
		1.20×10^4	0/6	0	0/6	0	0/6	0
	catB7	1.20×10^6	5/6	83	2/6	33	2/6	33
		1.30×10^5	1/6	17	0/6	0	0/6	0
		1.20×10^4	0/6	0	0/6	0	0/6	0

[a]Detection rates for the given AMR genes according to the concentrations of the bacteria and the DNA extraction methods.

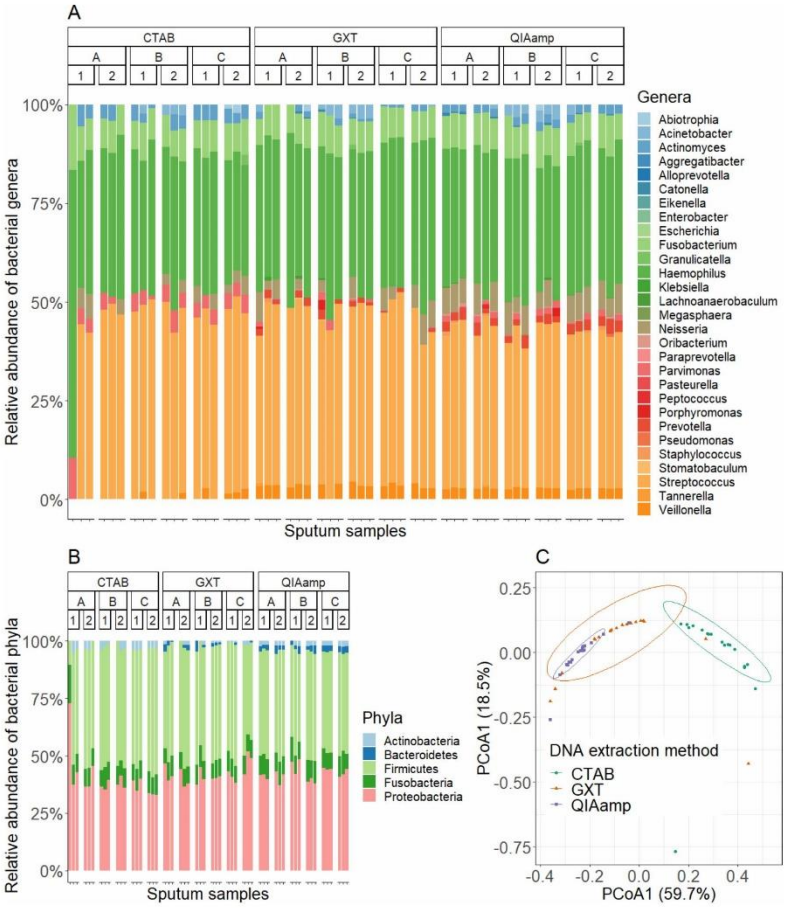


FIG 3 Comparison of microbiome richness and diversity from samples extracted with different DNA extraction methods. (A) Relative abundancies of the bacterial genera for the sputum samples according to the DNA extraction methods (CTAB, GXT, and QIAamp), sputum samples (A, B, and C), and days of extraction (1 and 2). Principal coordinate analysis in two dimensions (B) and constrained analysis of principal coordinates (C), using Bray Curtis distances and clustering of the samples according to the DNA extraction methods.

diversity was observed between extraction days, also indicating the repeatability of all three DNA extraction methods (Table S15). The method of DNA extraction also had a significant effect on the composition of the microbiome. The relative abundance of 10 bacterial genera differed significantly between the DNA extraction methods (Table S16). The proportion of Gram-positive bacteria was significantly higher in the samples extracted using the CTAB protocol (Tables S17 to S18).

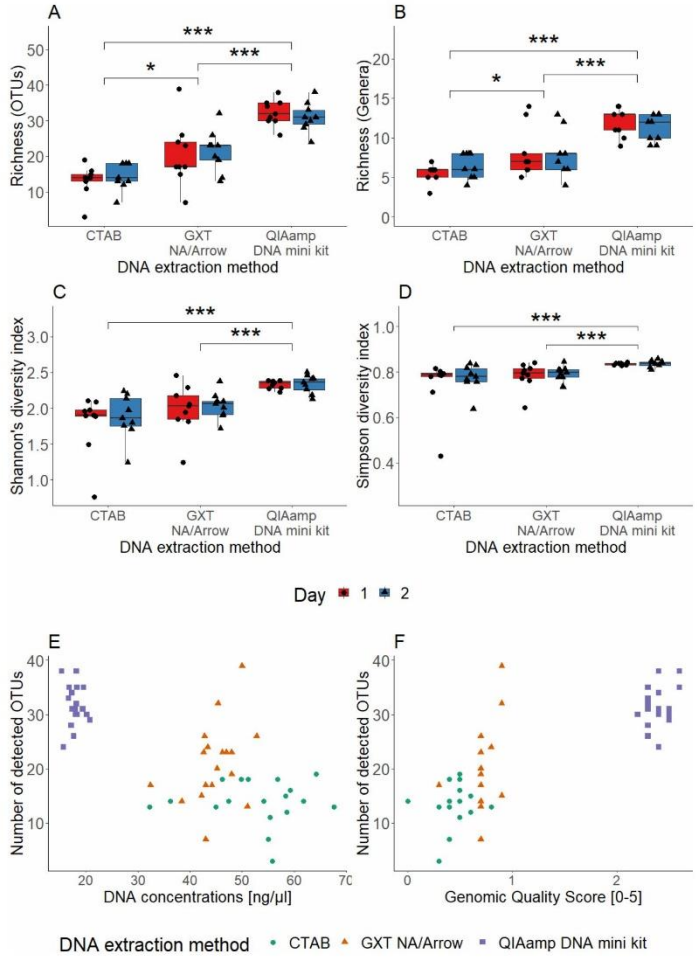


FIG 4 Alpha-diversity according to the DNA extraction methods, shown for the richness, as the number of OTUs (A), the number of genera (B), the Shannon diversity index (C), and the Simpson index (D). Effects on the OTUs detected in terms of DNA concentration (E) and genomic quality score (F) for the different samples. *, *P*-value < 0.05; ***, *P*-value < 0.001 (Kruskal-Wallis *t*-test).

Evaluation of the resistome using targeted high-throughput screening of AMR genes

Together, we detected 50 amplicons and 27 AMR genes responsible for resistance to eight classes of antimicrobials (Fig. 5A and B; Table S19). AMR genes responsible for resistance to macrolides and beta-lactams were the most abundant, accounting for 64% and 31% of the resistome, respectively. Similar to the microbiome, the DNA extraction method had a statistically significant influence on the number of AMR genes and antimicrobial classes detected in the samples (*P*-value < 0.001; Kruskal-Wallis *t*-tests;

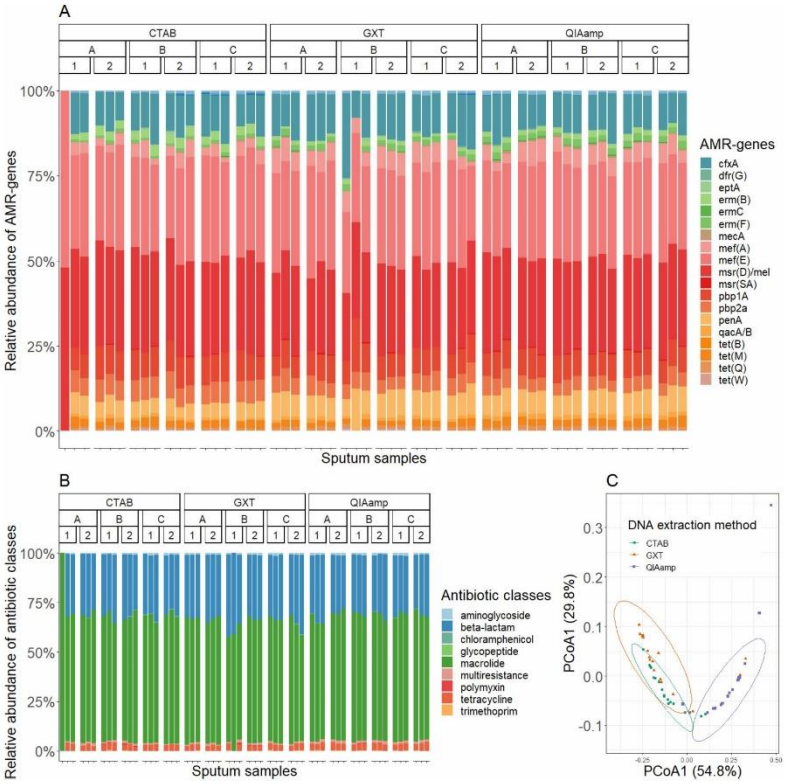


FIG 5 (A, B) Relative abundancies of the different AMR genes (A) and antimicrobial classes (B) for the sputum samples according to the DNA extraction methods (CTAB, GXT, and QIAamp), sputum samples (A, B, and C), and days of extraction (1 and 2). (C) Principal coordinate analysis using Bray-Curtis distances and clustering of the samples according to the DNA extraction methods. (D–F) The numbers of amplicons (D), AMR genes (E), and antimicrobial classes to which these genes possess resistance (F). The boxplots show median values and quartiles for each DNA extraction method and day of extraction. *, *P*-value < 0.05; **, *P*-value < 0.01; ***, *P*-value < 0.001 (Kruskal-Wallis *t*-test).

Table S20). Again, beta-diversity of AMR genes based on Bray-Curtis dissimilarities and two-dimensional PCoA was used to visualize differences between samples. Samples extracted with the QIAamp DNA mini kit were grouped separately, while samples extracted with the CTAB protocol and GXT NA/Arrow overlapped (Fig. 5). The DNA extraction method had a significant effect on resistome, accounting for 54.2% of the total variance within samples as calculated with PERMANOVA (Table S21). Resistome diversity was analyzed by examining the number of AMR genes present in the samples and the corresponding antimicrobial classes. Samples extracted with the QIAamp DNA mini kit had significantly the highest number of AMR genes (overall mean 20.3) and corresponding antimicrobial classes (overall mean 6.7) as well as the most repeatable results. In contrast, the samples extracted with the CTAB protocol had the lowest number of AMR genes (overall mean 15.0) and the samples extracted with GXT NA/Arrow had the lowest number of antimicrobial classes and the lowest repeatability of the resistome (Fig. S4; Table S22). The DNA extraction method also influenced the composition of the resistome, with 26 AMR genes showing significantly different relative abundances (Table

S23). The results described above clearly show that the DNA extraction method plays an important role in the determination of the resistome.

DISCUSSION

In this study, the impact of different DNA extraction method on the results of targeted HTS for the microbiome and the resistome was critically assessed in a complex sputum matrix. This work addresses a gap in current research, as many previous studies have focused on mock communities rather than clinically relevant samples. By incorporating sputum, a highly heterogeneous matrix, findings of this study have direct applicability to diagnostic settings, enhancing the relevance and utility of results (14, 17, 30). Such clinical mock samples are also better suited to evaluate the repeatability and consistency of DNA extraction methods. Although there are many studies investigating the effects of DNA extraction methods on the microbiome, only a few of them focus on sputum. Due to the unique characteristics of each matrix, it is difficult to transfer the results of studies conducted with different matrices which is especially true for sputum. DNA extraction protocols and sample processing prior to DNA extraction have been shown in previous studies to have a significant impact on 16S rRNA metagenomics results (15, 20). For instance, mechanical lysis by bead beating and enzymatic lysis with lysozyme can significantly increase the percentage of Gram-positive bacteria that may otherwise be underrepresented (19, 31, 32). For sputum samples, the use of liquefying agents to homogenize the sputum samples is an important step in the treatment of samples prior to DNA extraction (16, 19).

In this study, three distinct DNA extraction methods were evaluated, involving different principles for DNA extraction. Each method demonstrated unique strengths and weaknesses. The CTAB protocol had the highest DNA yield but also the highest variability and fragmentation. The GXT NA/Arrow kit is commercially available and offers automated extraction based on binding of DNA to magnetic beads. Although it is automated, it did not show the highest repeatability. The QIAamp DNA mini kit is commercially available DNA extraction kit that is based on the binding of DNA to silica membranes using spin columns for extraction. It provided the most repeatable results with the least fragmented DNA, despite lower overall yield.

Targeted HTS with species-specific amplicons achieved detection limits as low as 10^4 cells/mL for spiked bacteria, with variations in detection sensitivity among different species. *A. baumannii* was detected in the highest proportion of samples, *P. aeruginosa* in the lowest. The lower analytical sensitivity for *P. aeruginosa* than expected could be related to the lower efficiency of target amplification or the lower number of species-specific amplicons in the Pan-Bacterial Panel. Since similar differences in the detection rates of the different bacteria were found for both dPCR and targeted HTS, the lower efficiency of DNA extractions for the different bacterial species could also be the reason. The overall sensitivity is lower compared to dPCR, but with targeted HTS, a larger number of different bacterial species can be detected simultaneously.

The microbiome determination using targeted HTS of the 16S rRNA region provided repeatable results. The DNA extraction method showed a significant effect on the alpha-diversity parameters. Importantly, the QIAamp DNA mini kit consistently yielded the highest alpha-diversity and richness, suggesting it is the most suitable for comprehensive microbiome analysis. Both the CTAB protocol and the GXT NA/Arrow kit yielded a significantly lower number of detected OTUs and diversity in the samples. The presence or absence of low abundance species is one of the main reasons for the differences between the extraction methods. This could be due to the different dilution of the DNA prior to library preparation. Due to the higher concentration, the DNA extracted with the CTAB protocol had to be diluted significantly more, which could lead to a lower detection of low abundance species and consequently a lower diversity. The samples extracted with the CTAB protocol showed a higher proportion of Gram-positive bacteria which could be due to the enzymatic digestion with lysozyme (14, 16, 31). These insights are

crucial for studies aiming to capture the full breadth of microbial diversity and resistome composition in clinical samples.

Similar trends were observed in the determination of the resistome, where the samples extracted with the QIAamp DNA mini kit showed a higher average number of detected AMR genes than with the CTAB protocol and the GXT NA/Arrow kit. However, the differences between the DNA extraction methods were less pronounced for the resistome than for the microbiome. This suggests that while the QIAamp DNA mini kit is particularly effective for microbiome studies, it is also a robust choice for resistome analysis. Even if the other two methods allow a more automated extraction, as with the GXT NA/Arrow kit, or are more cost-efficient, as with the CTAB protocol, these advantages seldom outweigh the higher repeatability and microbiome diversity of the results obtained with the QIAamp DNA mini kit. The number of samples in a typical study of the microbiome is generally not so high as to require automation, and the price of DNA extraction is only a small part of the overall price of targeted HTS. However, the automation could be integrated into DNA extraction methods based on the spin columns. Furthermore, the addition of enzymatic lysis with lysozyme prior to extraction to the QIAamp DNA mini kit protocol could enhance the detection of Gram-positive bacteria, improving the overall yield and providing results that even more accurately reflect the actual composition of the microbiome and resistome. Previous studies using mock communities with known abundance of different bacterial taxa have shown that different DNA extraction methods resulted in different microbiome compositions. However, no DNA extraction method gave results that perfectly reflected the actual composition, so we must be cautious when interpreting the differences between DNA extraction methods (31, 33).

Sequencing of the pure bacterial suspensions revealed that all three species used for spiking contained AMR genes which were also detected in the spiked sputum samples. The detection rates of these AMR genes increased with the concentrations of the spiked bacteria to which the AMR genes belong; this indicates that these AMR genes indeed originate from the spiked bacteria. However, at this stage, it cannot be completely ruled out that they belong to other bacteria in the sputum samples. It must also be emphasized that the AMR genes detected for the spiked bacteria and the sputum samples are not necessarily all genes associated with AMR in the bacteria and sputum, but only those that can be detected using the Pan-Bacterial Research Panel. An advantage of this panel is that it is based on genes associated with AMR rather than single nucleotide polymorphisms, which are less reliable for predicting AMR. The use of DNA extraction blank controls and no-template controls is necessary to identify potential contaminants that may occur during sample handling, DNA extraction, and library preparation. These controls are essential for accurate data interpretation and for minimizing the risk of false positives in microbiome studies (34).

While individual steps of DNA extraction (e.g., enzymatic lysis, use of solubilizing agents, bead beating, etc.) were not examined separately, comprehensive evaluation of the overall effects of three widely used DNA extraction methods enhances the transferability of results. In contrast to studies using mock communities, the actual composition of the microbiome is not known. Therefore, it cannot be determined how similar results of this study are to the actual composition of the microbiome. However, the sputum samples used here are much more representative of actual clinical samples.

In conclusion, the results of this study should serve as a guide for future research into the bacterial microbiome and resistome using targeted HTS, with particular emphasis on the importance of appropriate DNA extraction method. It is also shown that targeted HTS can provide repeatable results when the appropriate DNA extraction method is used, which allows for a more rationalized study design, avoiding potential errors and unsatisfactory results that may lead to prolonged studies and higher costs. By adopting the principles of metrology and diagnostics to targeted HTS, researchers can achieve better repeatability and comparability of results between laboratories, enhancing

confidence in their findings and contributing to the advancement of microbiome and resistome research.

ACKNOWLEDGMENTS

This study was funded by the Slovenian Research and Innovation Agency (contract nos. P4-0165, 52760, and Z2-1860), SEPTIMET project (which has received funding from the EMPIR program co-financed by the participating states and from the European Union's Horizon 2020 research and innovation program, project no. 18HLT03). Part of dPCR equipment used in this study was financed by the Metrology Institute of the Republic of Slovenia (MIRS), with financial support from the European Regional Development Fund.

Special thanks go to Maja Zagorščak for help with the statistical analysis of the data and Denis Kutnjak for critical reading of the article.

AUTHOR AFFILIATIONS

[1]Department of Biotechnology and Systems Biology, National Institute of Biology, Ljubljana, Slovenia
[2]Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
[3]OMEGA d.o.o., Ljubljana, Slovenia
[4]University Clinic of Pulmonary and Allergic Diseases Golnik, Laboratory for Respiratory Microbiology, Golnik, Slovenia

AUTHOR ORCIDs

Aleksander Benčič http://orcid.org/0000-0001-6155-4208
Alexandra Bogožalec Košir https://orcid.org/0000-0002-4039-9963

FUNDING

Funder	Grant(s)	Author(s)
Slovenian Research and Innovation Agency (ARIS)	P4-0165, 52760, Z2-1860	Aleksander Benčič Alexandra Bogožalec Košir Mojca Milavec Tanja Dreo
EURAMET | European Metrology Programme for Innovation and Research (EMPIR)	SEPTIMET 18HLT03	Alexandra Bogožalec Košir Mojca Milavec
Metrology Institute of the Republic of Slovenia (MIRS)		Alexandra Bogožalec Košir Mojca Milavec

AUTHOR CONTRIBUTIONS

Aleksander Benčič, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft.

DATA AVAILABILITY

Data used in this study are publicly available at https://zenodo.org/records/11619401.

ETHICS APPROVAL

Written patients' consent was obtained at University Clinic Golnik, and the study was approved by the institution's review board.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental material (mSystems00735-24-S0001.docx). Supplemental figures and tables.

REFERENCES

1. Man WH, de Steenhuijsen Piters WAA, Bogaert D. 2017. The microbiota of the respiratory tract: gatekeeper to respiratory health. Nat Rev Microbiol 15:259–270. https://doi.org/10.1038/nrmicro.2017.14
2. O'Dwyer DN, Dickson RP, Moore BB. 2016. The lung microbiome, immunity, and the pathogenesis of chronic lung disease. J Immunol 196:4839–4847. https://doi.org/10.4049/jimmunol.1600279
3. Dickson RP, Erb-Downward JR, Martinez FJ, Huffnagle GB. 2016. The microbiome and the respiratory tract. Annu Rev Physiol 78:481–504. https://doi.org/10.1146/annurev-physiol-021115-105238
4. Wright GD. 2007. The antibiotic resistome: the nexus of chemical and genetic diversity. Nat Rev Microbiol 5:175–186. https://doi.org/10.1038/nrmicro1614
5. AlMS, Aogáin M, Aogáin A, Lau KJX, Cai Z, Kumar Narayana J, Purbojati RW. 2020. Metagenomics reveals a core macrolide resistome related to microbiota in chronic respiratory disease. Am J Respir Crit Care Med 202:433–447. https://doi.org/10.1007/s10198-019-01117-3
6. Yatera K, Noguchi S, Mukae H. 2018. The microbiome in the lower respiratory tract. Respir Investig 56:432–439. https://doi.org/10.1016/j.resinv.2018.08.003
7. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower C. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 31:814–821. https://doi.org/10.1038/nbt.2676
8. Urbaniak C, Sielaff AC, Frey KG, Allen JE, Singh N, Jaing C, Wheeler K, Venkateswaran K. 2018. Detection of antimicrobial resistance genes associated with the international space station environmental surfaces. Sci Rep 8:814. https://doi.org/10.1038/s41598-017-18506-4
9. Park J, Shin SY, Kim K, Park K, Shin S, Ihm C. 2018. Determining genotypic drug resistance by ion semiconductor sequencing with the Ion ampliSeq TB panel in multidrug-resistant mycobacterium tuberculosis isolates. Ann Lab Med 38:316–323. https://doi.org/10.3343/alm.2018.38.4.316
10. Su M, Satola SW, Read TD. 2019. Genome-based prediction of bacterial antibiotic resistance. J Clin Microbiol 57:1–15. https://doi.org/10.1128/JCM.01405-18
11. Brooks JP, Edwards DJ, Harwich MD, Rivera MC, Fettweis JM, Serrano MG, Reris RA, Sheth NU, Huang B, Girerd P, Strauss JF, Jefferson KK, Buck GA. 2015. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies ecological and evolutionary microbiology. BMC Microbiol 15:1–14. https://doi.org/10.1186/s12866-015-0351-6
12. Li A-D, Metch JW, Wang Y, Garner E, Zhang AN, Riquelme MV, Vikesland PJ, Pruden A, Zhang T. 2018. Effects of sample preservation and DNA extraction on enumeration of antibiotic resistance genes in wastewater. FEMS Microbiol Ecol 94:1–11. https://doi.org/10.1093/femsec/fix189
13. Whiteley AS, Jenkins S, Waite I, Kresoje N, Payne H, Mullan B, Allcock R, O'Donnell A. 2012. Microbial 16S rRNA Ion Tag and community metagenome sequencing using the Ion Torrent (PGM) platform. J Microbiol Methods 91:80–88. https://doi.org/10.1016/j.mimet.2012.07.008
14. Abusleme L, Hong BY, Dupuy AK, Strausbaugh LD, Diaz PI. 2014. Influence of DNA extraction on oral microbial profiles obtained via 16S rRNA gene sequencing. J Oral Microbiol 6. https://doi.org/10.3402/jom.v6.23990
15. Kennedy NA, Walker AW, Berry SH, Duncan SH, Farquarson FM, Louis P, Thomson JM, Satsangi J, Flint HJ, Parkhill J, Lees CW, Hold GL, UK IBD Genetics Consortium. 2014. The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. PLoS One 9:e88982. https://doi.org/10.1371/journal.pone.0088982
16. Oriano M, Terranova L, Teri A, Sottotetti S, Ruggiero L, Tafuro C, Marchisio P, Gramegna A, Amati F, Nava F, Franceschi E, Cariani L, Blasi F, Aliberti S. 2019. Comparison of different conditions for DNA extraction in sputum - a pilot study. Multidiscip Respir Med 14:6. https://doi.org/10.1186/s40248-018-0166-z
17. Willner D, Daly J, Whiley D, Grimwood K, Wainwright CE, Hugenholtz P. 2012. Comparison of DNA extraction methods for microbial community profiling with an application to pediatric bronchoalveolar lavage samples. PLoS One 7:e34605. https://doi.org/10.1371/journal.pone.0034605
18. Feigelman R, Kahlert CR, Baty F, Rassouli F, Kleiner RL, Kohler P, Brutsche MH, von Mering C. 2017. Sputum DNA sequencing in cystic fibrosis: non-invasive access to the lung microbiome and to pathogen details. Microbiome 5:20. https://doi.org/10.1186/s40168-017-0234-1
19. Terranova L, Oriano M, Teri A, Ruggiero L, Tafuro C, Marchisio P, Gramegna A, Contarini M, Franceschi E, Sottotetti S, Cariani L, Bevivino A, Chalmers JD, Aliberti S, Blasi F. 2018. How to process sputum samples and extract bacterial DNA for microbiota analysis. Int J Mol Sci 19:1–12. https://doi.org/10.3390/ijms19103256
20. Greathouse KL, Sinha R, Vogtmann E. 2019. DNA extraction for human microbiome studies: the issue of standardization. Genome Biol 20:212. https://doi.org/10.1186/s13059-019-1843-8
21. Bogožalec Košir A, Lužnik D, Tomič V, Milavec M. 2023. Evaluation of DNA extraction methods for reliable quantification of Acinetobacter baumannii, Klebsiella pneumoniae, and Pseudomonas aeruginosa. Biosensors 13:463. https://doi.org/10.3390/bios13040463
22. Devonshire AS, O'Sullivan DM, Honeyborne I, Jones G, Karczmarczyk M, Pavšič J, Gutteridge A, Milavec M, Mendoza P, Schimmel H, et al. 2016. The use of digital PCR to improve the application of quantitative molecular diagnostic methods for tuberculosis. BMC Infect Dis 16:366. https://doi.org/10.1186/s12879-016-1696-7
23. Team RC. 2021. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
24. Team Rs. 2021. RStudio: integrated development environment for R. RStudio, PBC, Boston, MA.
25. Wickham H. 2009. Ggplot2: elegant graphics for data analysis. New York, NY.
26. McMurdie PJ, Holmes S. 2013. Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8:e61217. https://doi.org/10.1371/journal.pone.0061217
27. Jari Oksanen F, GuillaumeB, Michae F, Roeland K, et al. 2020. Vegan: community ecology package. R package version 2.5-7
28. Hervé M. 2021. RVAideMemoire: testing and potting procedures for biostatistics. R package version 0.9-79
29. Xia Y, Sun J, Chen D-G. 2018. Statistical analysis of microbiome data with R. Springer Singapore, Singapore.
30. Panek M, Čipčić Paljetak H, Barešić A, Perić M, Matijašić M, Lojkić I, Vranešić Bender D, Krznarić Ž, Verbanac D. 2018. Methodology challenges in studying human gut microbiota - effects of collection, storage, DNA extraction and next generation sequencing technologies. Sci Rep 8:5143. https://doi.org/10.1038/s41598-018-23296-4
31. Teng F, Darveekaran Nair SS, Zhu P, Li S, Huang S, Li X, Xu J, Yang F. 2018. Impact of DNA extraction method and targeted 16S-rRNA hypervariable region on oral microbiota profiling. Sci Rep 8:16321. https://doi.org/10.1038/s41598-018-34294-x
32. Sinclair J, West NP, Cox AJ. 2023. Comparison of four DNA extraction methods for 16s rRNA microbiota profiling of human faecal samples. BMC Res Notes 16:169. https://doi.org/10.1186/s13104-023-06451-7
33. Douglas CA, Ivey KL, Papanicolas LE, Best KP, Muhlhausler BS, Rogers GB. 2020. DNA extraction approaches substantially influence the assessment of the human breast milk microbiome. Sci Rep 10:123. https://doi.org/10.1038/s41598-019-55568-y
34. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. 2019. Contamination in low microbial biomass microbiome studies: issues and recommendations. Trends Microbiol 27:105–117. https://doi.org/10.1016/j.tim.2018.11.003

2.2 Comparative Genomics and Pathogenicity of *Pantoea stewartii* subsp. *stewartii* Reveal Multiple Introductions and Limited Distribution in Europe

Aleksander Benčič, Joël F. Pothier, Tanja Dreo

bioRxiv, 2025.11.07.687269. https://doi.org/10.1101/2025.11.07.687269

This publication describes whole-genome sequencing and comparative genomics analysis of different strains of *Pantoea stewartii* subsp. *stewartii*. Despite its quarantine status, this bacterium was detected in Europe in recent years detected in Europe on multiple occasions and in different countries. To elucidate these findings, we sequenced Slovenian isolates and historical isolates from bacterial collections belonging to this bacterium. To obtain complete genomes hybrid approach utilising short and long read technologies was used. The obtained genomes were included in comparative genomic studies alongside those available in public databases. Phylogenetic analysis revealed that all except one Slovenian isolate form a distinct branch, indicating at least two introduction events. Italian isolates, genomes of which were also included in the analysis, showed a similar pattern and were not shown to be related to Slovenian isolates. These results showed that recent European findings of *P. stewartii* subsp. *stewartii* are the result of multiple introduction events. Further phylogenetic analysis of the whole species *P. stewartii* revealed that the subspecies *stewartii* forms a distinct branch within it. Pan-genome analysis revealed that *P. stewartii* has an open pangenome, with subspecies *stewartii* exhibiting lower diversity compared to other *P. stewartii* strains. Conversely, subspecies *stewartii* exhibited a diversity of mobile genetic elements, such as plasmids and prophages, which are absent in other *P. stewartii* strains. These plasmids often contain important pathogenicity factors, such as type III secretion systems, which are responsible for the symptoms of maize infection. The analysis also discovered that a few strains lack some of these pathogenicity factors. Further greenhouse experiments revealed that these strains do not cause the typical symptoms of infection with *P. stewartii* subsp. *stewartii* in maize.

The PhD candidate is the first author of this publication. He contributed to the preparation of bacterial DNA, whole genome sequencing, bioinformatic analysis, comparative genomic studies, greenhouse pathogenicity tests on maize, and the analysis of experimental results. He wrote the manuscript, including the supplementary material, and prepared the figures and tables.

Comparative genomics and pathogenicity of *Pantoea stewartii* subsp. *stewartii* reveal multiple introductions and limited distribution in Europe

Submission to: bioRxiv

Aleksander Benčič[1, 2], Joël F. Pothier[3], Tanja Dreo[1]

[1]Department of Biotechnology and Systems Biology, National Institute of Biology, Večna pot 121, SI-1000 Ljubljana, Slovenia

[2]Jožef Stefan International Postgraduate School, SI-1000 Ljubljana, Slovenia

[3]Environmental Genomics and Systems Biology Research Group, Institute for Natural Resource Sciences, Zurich University of Applied Sciences (ZHAW), Wädenswil, Switzerland

***Corresponding author: Tanja Dreo**

Department of Biotechnology and Systems Biology,

National Institute of Biology

Vecna pot 121

SI-1000 Ljubljana, Slovenia

E-mail: tanja.dreo@nib.si

Tel: +386-5-9232797

ABSTRACT

Pantoea stewartii subsp. *stewartii* (Pss), the causal agent of Stewart's wilt in maize, is native to North America but has been detected on several occasions also in Europe, including in Italy in association with seed maize production. These findings have raised concern due to the bacterium's quarantine status in the EU. To investigate its diversity and possible introduction routes, we performed whole-genome sequencing of Slovenian and other Pss isolates and compared them with publicly available genomes. Comparative genomics revealed that Pss forms a distinct clade within *P. stewartii*, exhibiting high average nucleotide identity (>99.9%). Most of the Slovenian isolates clustered closely together, forming a separate branch consistent with at least two independent introduction events. They were genetically distinct from recent Italian isolates. Additionally, several unique plasmids and prophages were identified in Pss isolates, with notable diversity in mobile genetic elements despite overall genomic homogeneity. Pangenome analysis confirmed an open pangenome for *P. stewartii*, with greater genomic diversity among non-Pss strains. Functional analysis identified multiple secretion systems in Pss, which are likely to contribute to its pathogenicity and insect-mediated transmission. Our findings highlight previously unrecognised diversity within the subspecies and confirm the presence of at least two distinct clades within *P. stewartii*. So far, the detections in Slovenia remain confined to the Vipava Valley with evidence of multiple introduction events. Together, these findings cannot definitely conclude whether Pss is already established in Slovenia. This suggests that eradication remains achievable through continued surveillance and preventive phytosanitary measures.

Keywords

Stewart's wilt; maize; whole genome sequencing; comparative genomics; multiple introductions in Europe; quarantine pathogen

Abbreviations

ANI - Average Nucleotide Identity

CDS - Coding DNA Sequence

DPI - Days Post-Inoculation

EPS – Exopolysaccharide

LPP-1 - Large *Pantoea* Plasmid 1

MALDI-TOF - Matrix-Assisted Laser Desorption/Ionization Time-of-Flight

NA - Nutrient Agar

PCR - Polymerase Chain Reaction

PBS - Phosphate Buffered Saline

PF - Pathogeny Factor

PT - Pathogenicity Test

QS - Quorum Sensing

rRNA - Ribosomal Ribonucleic Acid

SNP - Single Nucleotide Polymorphism

T2SS - Type II Secretion System

T3SS - Type III Secretion System

T4SS - Type IV Secretion System

T6SS - Type VI Secretion System

tRNA - Transfer Ribonucleic Acid

INTRODUCTION

Pantoea stewartii subsp. *stewartii* (Smith 1898) is a Gram-negative, xylem-inhabiting bacterium and the causal agent of Stewart's wilt of maize (*Zea mays* L.). The disease, first described in the United States in the late 19th century, is considered one of the most significant bacterial diseases of maize in its native range. Initially classified as part of the genus *Erwinia*, it was later reclassified as part of the genus *Pantoea* (1). Bacteria from this genus inhabit a variety of ecological niches and have been isolated all around the world. *P. stewartii* subsp. *stewartii* (Pss) is native to North America, but findings of the disease have been reported worldwide. Sweet corn and some elite inbred maize lines are especially susceptible to the disease, which has two main phases: wilt and leaf blight. In the first phase, known as the wilt phase, young seedlings become infected, and water-soaked lesions appear on the leaves. The seedlings then wilt severely and often die. The second phase, known as the blight phase, occurs when mature plants become infected and characteristic linear yellow-brown lesions appear, running parallel to the leaf veins. The disease is transmitted from plant to plant via the corn flea beetle (*Chaetocnema pulicaria*), which can also carry it in its gut during the winter months (2). Pss spreads over longer distances via infected seeds, which has led to restrictions on international trade in seed material. This can be mitigated by planting resistant maize cultivars and testing imported seeds for the bacterium's presence. In addition to maize, Pss can infect other plants, including sudangrass, oats, triticale, sorghum, millet, and sugarcane. Another subspecies, *P. stewartii* subsp. *indologenes* (Psi), infects a variety of crops, including *Allium* spp. and rice (3). Although it is widely accepted that Psi does not cause symptoms of infection in maize, there have been recent reports of Psi isolated from symptomatic maize plants (4).

Pss is an attractive laboratory model organism due to its fast growth rate and the relatively small number of pathogenicity mechanisms it possesses. The two main pathogenicity factors (PF) of Pss are the Hrp type III secretion system (T3SS) and the exopolysaccharide

stewartan. The Hrp gene cluster is responsible for the T3SS and the effectors involved in colonising the host plant. Pss does not possess a large number of different effector proteins, indicating that it acquired the T3SS relatively recently (5). One such protein is WtsE, an AvrE-family effector that causes cell death and leads to water-soaking lesions and necrosis in maize plants, playing an essential role in pathogenesis (6). Stewartan, on the other hand, is responsible for vascular streaking, bacterial oozing, and wilting (7). Pss possesses a quorum-sensing (QS) system, which is regulated by *esaI* (a *luxI* homologue) that encodes the signal *N*-3-oxohexanoyl homoserine lactone synthase, and by *esaR* (a *luxR* homologue) that encodes the response regulator (8). The QS system and the Rcs phosphorelay have been observed to regulate the stewartan production, thereby contributing to virulence (9). Another mechanism through which QS is involved in pathogenicity in plants is the regulation of surface motility. Pss exhibits flagella-dependent surface motility, which is involved in biofilm development and plays a significant role in colonising the plant host. The flagellum apparatus is encoded by the *fliTSDC1* locus; mutations in this locus result in loss of motility (10).

The first complete genome sequence of the Pss DC283 strain reveals a 4.53-Mb circular chromosome, ten circular plasmids ranging in size from 4,277 to 304,641 bp, and one linear phage plasmid (ppDSJ01), which is related to the *E. coli* N15 prophage (11). The two smallest plasmids, pDSJ01 and pDSJ02, have medium copy numbers, while the remaining plasmids have low copy numbers (fewer than 10 copies). Genes associated with the T3SS are located on two separate plasmids: pDSJ08 and pDSJ10. The Pss DC283 genome contains genes encoding several putative endoglucanases, xylanases, and a β-1,4–β-1,3 mixed-linkage glucan glucanohydrolase. These enzymes can digest plant cell walls and may play a critical role in enabling the bacteria to access carbohydrates associated with xylem cell wall structure and development (12). The Pss genome also contains many repetitive transposable sequences,

which present challenges in genome assembly when only short reads are used. Additionally, multiple prophage sequences were identified within the DC283 genome assembly (11).

In recent years, Pss has been detected on several occasions outside its native range, including in Italy in association with seed maize production and in Slovenia, where findings have so far been confined to the Vipava Valley, a small region with sub-Mediterranean climatic conditions bordering Italy. So far, detections have been limited to symptomatic maize plants appearing late in the growing season, with no evidence of early-season infections, suggesting that *Pss* does not overwinter or cycle naturally under conditions in the areas studied. While previous comparative studies provided valuable insights into the taxonomy of *P. stewartii*, the intraspecific relationships within the species and the genomic diversity among European isolates have remained poorly understood. In particular, it is unclear whether the recent detections represent local establishment, transient survival, or repeated introductions.

The present study aimed to clarify the genetic position and diversity of European Pss isolates and to relate these findings to their phenotypic and pathogenic characteristics. To this end, high-quality whole-genome sequences of Slovenian isolates and additional historical Pss strains were generated and compared with publicly available genomes of *P. stewartii*. Comparative genomic, phylogenetic, and pangenome analyses were combined with pathogenicity testing to assess the diversity and functional potential of Pss strains. The results provide the first detailed genomic and phenotypic characterisation of Pss in Europe, offering new evidence of multiple introduction events, limited distribution, and informing ongoing surveillance and eradication programmes.

MATERIAL AND METHODS

Preparation of material

In the scope of this study, 24 new strains of *Pantoea stewartii* subsp. *stewartii* (Pss) were selected for sequencing, including 11 Slovenian isolates and 13 isolates from bacterial strain collections (Table 1). The study includes Slovenian isolates collected between 2018 and 2022. Those were obtained from symptomatic maize plants, which were collected as part of an annual official survey conducted by the Administration of the Republic of Slovenia for Food Safety, Veterinary and Plant Protection, Ministry of Agriculture and Environment, and Phytosanitary Inspectorate. The presence of *P. stewartii* was screened in DNA extracted from symptomatic maize samples using real-time PCR (13). If a sample tested positive, an additional real-time PCR test (14) was performed to differentiate between the subspecies *stewartii* and *indologenes*. The bacterial isolates from maize extracts were identified as Pss by MALDI TOF mass spectrometry (smartfleX, Bruker, Billerica, USA) and real-time PCR tests (14-15). The DNA used for sequencing library preparation was extracted from bacteria that were grown on NA (Bacto Nutrient Agar; Difco) media at 28 °C. Bacterial suspensions were prepared in 10 mM phosphate-buffered saline (PBS; 1.08 g Na_2HPO_4, 0.4 g $NaH_2PO_4 \cdot 2H_2O$, 8 g NaCl, 1 L distilled water, pH 7.2). DNA was extracted from the bacterial suspensions using QuickPick™ SML Plant DNA kits (Bio-Nobile, Turku, Finland) and an automated KingFisher™ mL system (Thermo LabSystems), as previously described (15).

High-throughput DNA sequencing

Whole genome sequencing was performed on DNA extracted from selected bacterial strains (**Table 1**), using the Illumina platform (Illumina, San Diego, USA) alongside one of two long-read sequencing technologies: either PacBio (Pacific Biosciences, Menlo Park, USA) or Nanopore (Oxford Nanopore Technologies, Oxford, UK). Libraries for Illumina sequencing

were prepared using a Nextera XT DNA Library Preparation Kit (Illumina) and sequenced on a MiSeq instrument (Illumina) in 2 × 300 bp (V3) mode. Nanopore sequencing was performed using the MinION platform with R9.4.1 or Flongle R9.4.1 flow cells (Oxford Nanopore Technologies). Libraries for Nanopore sequencing were prepared using a Ligation Sequencing Kit (SQK-LSK109) and Native Barcode Expansions 1–12 (EXP-NBD104) and 13–24 (EXP-NBD114), following the manufacturer's protocol. Magnetic beads (Mag-Bind TotalPure NGS, Omega Bio-Tek, Norcross, USA) were used to purify the DNA during library preparation. Demultiplexing and basecalling were carried out using the Guppy basecaller v6.4.2. PacBio sequencing was performed by an external provider (Novogene GmbH, Munich, Germany). HiFi SMRTbell libraries were prepared and sequenced for PacBio sequencing using the PacBio Sequel II system (Pacific Biosciences).

Bioinformatic analysis

The basecalled nanopore reads were trimmed using Porechop v0.2.4 (16) to remove adapters, and then filtered using NanoFilt v2.8.0 (17) to remove low-quality reads. The PacBio raw reads were used to generate circular consensus sequences with the pbccs v6.4.0 package (Pacific Biosciences), which were then used for genome assembly in subsequent analyses. Hybrid genome assembly using short and long reads was performed with Unicycler v0.4.8 (18), which uses SPAdes v3.15.5 (19) to assemble contigs from short reads. These contigs were then used alongside long reads to assemble hybrid genomes using miniasm v0.3-r179 (20) and Racon v1.4.20 (21). To assemble genomes using only long reads, Flye assembler v2.8.1-b1676 (22) was used. Assemblies were manually curated to remove any duplicated contigs. If possible, gaps in hybrid assemblies were filled with sequences generated using only long read assemblies. BUSCO v5.7.1 (23) with the pantoea_odb12 dataset was used to determine the completeness of genomes. Completeness was additionally determined by CheckM2 v1.1.0 with database v3,

189 which was also used to determine contaminations within genomes (24). The assembled
190 genomes were visualised and aligned using the commercially available software Geneious
191 Prime v2024.0.7 and the Mauve plugin v1.1.3 (25). Automatic annotation of the assembled
192 genomes was performed using Bakta v1.9.3 and the v5.1 database (26). Genome assemblies of
193 newly sequenced strains were deposited in the European Nucleotide Archive under the study
194 PRJEB100601.

195

196 **Phylogenetic analysis**

197 The phylogenetic analysis used 24 newly sequenced Pss strains (Table 1), 38 published
198 genome assemblies from different strains in the NCBI GenBank database, and three genomes
199 assembled from raw data from the Sequence Read Archive (Table 2). All genomes were
200 assembled using the same protocol described in the chapter 'Bioinformatic analysis'. The
201 genomes included in Table 1 and Table 2 are referred to simply as 'genomes' through the text.
202 Genomes with incomplete metadata (host, location, and year of isolation), poor quality (high
203 fragmentation, BUSCO (C) score under 95% and low sequencing depth), duplicate genomes of
204 the same strain, and metagenome-assembled genomes were excluded from the analysis.

205 Within the NCBI database, the genomes were categorised as either subspecies *stewartii*
206 or *indologenes,* or neither. For the phylogenetic analysis of all *P. stewartii* genomes, average
207 nucleotide identity (ANI) was used. This analysis was performed using the bioinformatic tool
208 fastANI v1.33 (27) to generate an ANI matrix. A phylogenetic tree was constructed from the
209 ANI matrix with the bioinformatic tool SplitsTree App v6.5.1 (28) using the neighbour-joining
210 method. For the phylogenetic analysis of genomes belonging to the Pss, single-nucleotide
211 polymorphisms (SNPs) were utilised. The programme snippy v4.6.0 (29) was employed to
212 identify SNPs between the haploid reference genome DC283 and other Pss genomes, and to
213 perform a core genome alignment. The maximum likelihood criterion and the programme

RaxML-NG v0.9.0 were used to construct the phylogenetic tree(30). Phylogenetic analysis of subspecies Pss was additionally performed using the accessory genome (genes present in less than 95 % of analysed genomes) as well. Presence-absence genes output generated in analysis performed by software Roary v3.13.0, described in chapter 'Pangenome analysis', was used to generate a binary matrix, which was then used to calculate the Jaccard distance matrix using the R package *vegan* v2.7-2 (31). A phylogenetic tree was constructed from a distance matrix using neighbour-joining method with the R package *ape* v 5.8-1 (32). Phylogenetic networks were calculated from both core genome SNPs alignment and accessory genome binary matrix in SplitsTree App using *p*-distances and neighbour net method. To perform whole genome multi-locus strain typing (MLST) of Pss genomes, the software chewBBACA v3.4.1 (33) was used together with a training file created from genome DC283 by software Pyrodigal v3.6.3 (34, 35). A whole genome MLST scheme was generated and used to perform allele calling. Core genome MLST was determined after removing paralogs and the allele call on the number of loci present in all genomes was used to generate a minimum spanning tree which was generated in R v4.0.5 and the integrated development environment RStudio v1.4.1106 (36), using packages ape (32), tidyverse v2.0.0 (37), ggnetwork v0.5.14 (38) and igraph v 2.1.4 (39) as previously described (40).

Pangenome analysis

The pangenomes of the subspecies Pss and Psi, as well as the species *P. stewartii*, were determined using annotated .gff3 files generated by Bakta, and the Roary pipeline was used to calculate the pangenome (41). The analysis included the newly sequenced genomes of Pss and *P. stewartii* genomes from NCBI databases (Table 1, Table 2). The results were visualised using the programming language R v4.0.5 (42), in combination with the ggplot2 v3.3.3 (43), ggpubr v0.4.0.999 (44) and gridExtra v2.3 (45) packages.

239

240 **Pathogenicity factors and mobile genetic elements**

241 A literature survey was conducted to identify known and potential pathogenicity factors
242 for Pss in maize (Supplement Table: S1). Sequences of proteins involved in the type III
243 secretion system were extracted from the annotation of the published Pss DC283 genome
244 (GCF_002082215.1). A BLASTp analysis was conducted on protein sequences generated in
245 the annotation of genomes to identify the presence of different types of secretion system
246 components and other pathogenicity factors. Pathogenicity factors were considered present if
247 they were identified with a score of over 80% identity and query coverage. A secretion system
248 was considered present if over 60% of the proteins that form it were identified. Prophage-like
249 regions and genes in *P. stewartii* genomes were identified using the PHASTEST v3.0 web
250 server (PHAge Search Tool with Enhanced Sequence Translation: https://phastest.ca/) (46). To
251 determine the identity of plasmids in complete genome assemblies, a BLASTn analysis was
252 performed on contigs and the Core Nucleotide BLAST database. To identify plasmids within
253 fragmented genome assemblies at the contig level, BLASTn was used to identify sequences of
254 known plasmids within the analysed assemblies. The plasmids were first extracted from the
255 complete genome of DC283, which served as a reference. Rearrangements in the plasmids were
256 analysed using progressiveMauve v1.1.3 (25) and visualised using the R package genoPlotR
257 v0.8.11 (47).

258

259 **Pathogenicity test on plants**

260 A pathogenicity test was performed on the sweetcorn variety 'Gucio' (L'Ortolano S.r.l.,
261 Cesena, Italy) using the following *P. stewartii* strains: CFBP 1719, CFBP 3165, CFBP 3167^{T},
262 CFBP 3169, CFBP 3445, CFBP 3614^{T}, NIB Z 2806, and NIB Z 3391. Each group contained
263 ten plants, with strains CFBP 1719 and NIB Z 2806 duplicated in two groups (giving a total of
264 20 plants per group). The strains used in the pathogenicity test were blinded to ensure unbiased

reporting of symptoms. Maize seedlings inoculated with 10 mM PBS served as the negative control, while the Pss strain NIB Z 2806 served as the positive control. The maize seeds were sterilised in 1.5% sodium hypochlorite solution (v/v) before being sown into 0.46 L pots containing a special substrate (Hawita Gruppe GmbH, Vechta, Germany). The maize seedlings were then grown in controlled greenhouse chambers following a 16 h light (22 °C)/8 h dark (19 °C) day/night cycle and 60% humidity. Fourteen days after sowing, the maize seedlings in growth stages V2-3 were infected via stem inoculation under the first leaf using a sterile 25 G hypodermic needle. This was performed using a bacterial suspension of 10^8 cells/mL (as determined by optical density/McFarland measurements at a wavelength of 565 nm), prepared from an approximately 48-hour culture on King's B medium. Following inoculation, the plants were incubated at a constant temperature of 26 °C, with 85% humidity and under the same lighting conditions as before inoculation. The presence of Pss infection symptoms was examined in the plants at 3-, 7-, 14-, and 30-days post-inoculation (DPI). The severity of the symptoms was evaluated using a scale from 0 to 7: 0 represented no symptoms; 1, yellow stripes on the leaves; 2, necrotic lesions on the leaves; 3, small, localised, water-soaked lesions with ooze; 4, larger, water-soaked lesions with ooze; 5, leaf deformities with widespread lesions; 6, stunted growth with wilting; and 7, plant death. The scale was modified from that described by Mohammadi et al. (12). Photos of representative symptoms used for symptom evaluation are presented in Supplement: Table S2.

RESULTS

Genome sequencing of Pss strains

A total of 24 Pss strains were newly sequenced, comprising 11 Slovenian isolates and 13 strains from bacterial collections. All strains were sequenced using Illumina technology, with either ONT or PacBio technology used to obtain hybrid assemblies. The genome assembly

290 sizes of the newly sequenced Pss strains ranged from 5.1 Mb (CFBP 3165) to 5.5 Mb (CFBP
291 3157), with an average size of 5.4 Mb. For strains for which a complete chromosome was
292 obtained, the size ranged from 4.49 Mb (CFBP 1719) to 4.68 Mb (NIB Z 3388), with no major
293 rearrangements observed (Table 1, Supplement: Figure S1, Table S3). These results are
294 consistent with data from the publicly available assembly of strain DC283, which has a genome
295 size of 5.2 Mb and a chromosome size of 4.53 Mb. Differences in genome size are mainly due
296 to differences in plasmids present in different strains. The number of coding DNA sequences
297 (CDSs) ranged from 5,237 (CFBP 3165) to 5,944 (CFBP 1719), averaging 5,604. The GC% of
298 Pss ranged from 53.6% to 53.9%, averaging 53.7% (Table 1, Supplement: Table S3). When
299 compared to other genome assemblies assigned to *P. stewartii* in GenBank, Pss genomes had,
300 on average larger size and a higher number of CDSs. Genome sizes of other *P. stewartii* strains
301 ranged from 4.2 Mb and 4,016 CDSs (RON18713) to 5.2Mb and 4,736 (ICMP 10132), with an
302 average genome size being 4.8 Mb and an average number of CDSs 4,359 (Table 2). All newly
303 sequenced and assembled genomes had high completeness (over 98%) and low contamination
304 (under 3 %) and could be included in further analysis (Table 1, Supplement: Table S3).

305

306 **Phylogenetic analysis**

307 ***Species* P. stewartii**

308 The genomes of *P. stewartii* included in the phylogenetic analysis had previously been
309 assigned to either the Pss or Psi subspecies, or to no subspecies at all. To obtain reliable results,
310 only the complete genome assemblies and high-quality draft assemblies with BUSCO
311 completeness score (C) > 95 % were included in phylogenetic and other analyses. These
312 genomes were used to calculate an ANI matrix between the genomes, which was then used to
313 construct a phylogenetic tree. The analysis revealed that the *P. stewartii* phylogenetic tree

formed two distinct branches. Strains assigned to Pss formed a single monophyletic group (Figure 1). The second branch included all strains assigned to the Psi subspecies, as well as all unassigned strains, all of which were more closely related to Psi. The exception was strain MS1, which was isolated from jackfruit and was previously assigned to Pss but is genetically more closely related to Psi strains. In the following text, all 35 strains from the branch containing Psi strains will be referred to as Psi. Strains in the Psi branch showed much greater diversity, with ANI values among them ranging from 98.48 % to 99.99% and a mean of 98.92%. In contrast, Pss strains were found to be very closely related, with ANI values ranging from 99.7% to 99.99% and a mean of 99.93%. ANI percentages between strains in Pss in Psi clades ranged from 98.27% to 98.76% with an average ANI being 98.56% (Supplement Table S3). Results of phylogenetic analysis based on ANI clearly demonstrate Pss and Psi to form two distinct lineages within species *P. stewartii*.

Subspecies stewartii

The Pss branch comprised 30 strains, including 11 Slovenian isolates, six Italian isolates, and the remainder from the USA. To investigate the relationships between these closely related strains, phylogenetic analysis of Pss subspecies was performed on both core and accessory genomes. In core genome-based phylogeny, SNPs were identified within the core genome alignment, which was then used to construct a phylogenetic tree using the maximum likelihood method. The results showed that all Slovenian isolates, except for NIB Z 3391, form a single branch within the Pss branch (Figure 2: A, Supplement: Figure S3). Slovenian isolates were most closely related to the Pss branch containing strains CFBP 3396, CFBP 3166, CFBP 3157, and CREA DC 1899. In accessory genome-based phylogeny binary matrix of genes present in the accessory genome was generated and used to construct a phylogenetic tree using Jaccard distances and the neighbour-joining method. Results of accessory genome-based phylogeny showed similar patterns. One major difference is that the Slovenian strain NIB Z

3388 does not group with the majority of other Slovenian isolates but with another Slovenian isolate, NIB Z 3391, which is positioned separately from other Slovenian isolates (Figure 3: A). Phylogenetic network analysis showed the core genome produces results with fewer conflicts in comparison to the accessory genome. It additionally showed strain NIB Z 3388 to form a separate branch (Figure 2: B, Figure 3: B). The Italian isolates do not group with the Slovenian isolates in any of the performed phylogenetic analyses, but they exhibit a similar pattern of behaviour. Five of them form a distinct branch, while isolate CREA DC 1899 is positioned elsewhere within the Pss tree (Figure 2, Supplement: Figure S3). To further examine the relationships between Pss strains, cgMLST was performed, which used a minimum spanning tree (MST) to visualise those relationships. The number of allelic differences between nodes in MST ranges from 4 to 90, which indicates that Pss strains do not form a unified clonal complex. Slovenian strains, with the exceptions of NIB Z 3388 and NIB Z 3391, showed a smaller number of allelic differences among them (4 to 13) in comparison to the other strains (number of allelic differences over 18). A similar pattern is again seen with Italian strains, which further form a separate group except for strain CREA DC 1899 (Figure 2: C).

Pangenome

The pangenome was determined for both subspecies, as well as for the entire *P. stewartii* species. The complete pangenome of Pss contained 7,425 genes. Of these, 3,968 formed the core genome, while the extended soft core — genes present in at least 95% of genomes — consisted of an additional 443 genes. The shell contained an additional 1,584 genes present in 15% to 95% of the genomes, and the cloud contained a further 1,430 genes present in less than 15% of the genomes (Figure 4 Figure 4A, B; Table 3). The Psi clade pangenome consisted of 11,952 genes, 3,496 of which were core genes (Figure 4Figure 4 C, D), whereas the pangenome of all *P. stewartii* species consisted of 15,559 genes, of which 3,013 were core genes (Figure

4Figure 4 E, F; Table 3). Core genes represented the largest percentage of the pangenome for subspecies Pss (53.4%), whereas in Psi and the entire *P. stewartii* species, core genes represented a smaller percentage of the pangenome (29.3 and 19.4%, respectively; Figure 4Figure 4**, G**). The number of genes in the pangenome continues to increase with additional genomes added, indicating *P. stewartii* possesses open pangenomes (Figure 4 A, C, E).

Diversity of mobile genetic elements

Plasmid diversity

The genomes forming the Pss branch of the phylogenetic tree were analysed for the presence of plasmids. The analysis identified 15 different plasmids, one of which was linear (ppPSS01) and the rest circular (Supplement Table S5). Despite being closely related, the analysed Pss strains showed diversity in their plasmids, with the number of plasmids ranging from eight (CFBP 3167^{T}) to 14 (CFBP 3445 and CFBP 3157), and the majority of genomes contained more than 10 plasmids (Figure 5), with no strain containing all plasmids. The larger plasmids pPSS10 (64.9 kb), pPSS11 (72.0 kb), pPSS13 (107.8 kb), and pPSS14 (305.8 kb) were present in some form in all analysed genomes (Figure 5, Supplement Table S6). Two previously unidentified plasmids, pPSS06 and pPSS09, were discovered in newly sequenced Pss strains. Plasmid pPSS06 is 36.3 kb in size on average, has a GC% of 43.3, and is present in nine genomes, while pPSS09 is 54.5 kb in size on average, has a GC% of 48.8, and is present in 20 genomes (Figure 5, Supplement Table S5-6). Both new plasmids contain a type IV secretion system (T4SS), including VirB10/TraB/TrbI, TrbG/VirB9, and TrbC/VirB2. They also contain a P-type DNA transfer ATPase, VirB11, which is involved in horizontal gene transfer via conjugation. Some rearrangements of plasmids have been observed in different Pss strains compared to the DC283 reference genome.

388 Additionally, multiple arrangements were also observed between plasmids in different
389 strains. Most of these rearrangements were observed in the 133 kb plasmid pDSJ09 from
390 reference genome DC283, which is split into the plasmids pPSS11 and pPSS12 in all Pss
391 genomes except said reference genome and CFBP 1719. This is in agreement with the
392 phylogenetic analysis, in which those two strains are positioned next to each other. The pPSS12
393 plasmid contains an insertion absent from pDSJ09. The exception is strain CFBP 3169, which
394 lacks pPSS12 altogether. The size of both pPSS11 and pPSS12 is just over 70 kb in most Pss
395 genomes, with some notable exceptions. Strain CFBP 3165 has a smaller pPSS11 plasmid (45
396 kb). pPSS12 showed greater variability, ranging in size from 52 to 91 kb (Supplement: Table
397 S6, Figure S2). Other plasmids also showed some rearrangements and variability. The plasmid
398 pPSS14 is conserved among all Pss strains but is smaller in strains CFBP 3165 (151 kb) and
399 CFBP 3169 (144 kb) than in all other strains, in which pPSS14 is over 300 kb in size
400 (Supplement: Figure S3). In 19 strains, parts of plasmids pPSS05 and pPSS07 merged to form
401 plasmid pPSS04; however, large parts of both original plasmids were lost. This variant is
402 present in all Slovenian and Italian isolates but absent from the majority of American isolates
403 (Supplement: Table S6 and Figure S4). In 14 strains, plasmid pPSS08 has a large insertion that
404 forms a monophyletic branch in the phylogenetic tree (Supplement: Table S6 and Figure S5).
405 This variant of pPSS08 is present in all Slovenian isolates except NIB Z 3391, which is
406 consistent with the ANI-based phylogenetic analysis. Some rearrangements were observed in
407 individual strains only. In strain CFBP 3157, part of plasmid pPSS10 was inserted into plasmid
408 pDSJ03, resulting in a smaller pPSS10 (47 kb) and a larger pPSS03 (34 kb) (Supplement: Figure
409 S6). In strain CFBP 3394, part of plasmid pPSS10 was inserted into pPSS05, resulting in
410 smaller pPSS10 (48 kb) and larger pPSS10 (45 kb) (Supplement: Figure S7). Overall, Slovenian
411 strains exhibit a consistent plasmid composition among them, with eight plasmids present in all
412 analysed genomes. Similar can also be said for Italian insoles, while American isolates, on the

413 other hand, show greater diversity in plasmid presence. Of all the plasmids identified in Pss
414 genomes, only pPSS14 is also present in other *P. stewartii* strains. The pPSS14 plasmid was
415 identified in all analysed strains, but it is smaller in Psi strains, which lack the part encoding the
416 T3SS.

417 ***Prophages***

418 All the *P. stewartii* genomes used in the phylogenetic analysis were also examined for the
419 presence of prophages. This analysis revealed that Pss strains have a greater number of
420 prophage regions and genes than Psi strains. Pss genomes contain between 10 and 14 intact
421 prophage regions, ranging in size from 10.5 to 112.6 kb (Supplement Table S7-9). In addition
422 to complete prophages, all genomes contained additional incomplete prophage regions. The
423 genomes of the Pss strains exhibited a distinct prophage profile, with the prophages
424 Burkho_BcepMu_NC_005882, Entero_phiT5282H_NC_049429,
425 Haemop_SuMu_NC_019455, and Salmon_SEN34_NC_028699 being conserved across all of
426 them (see Supplementary Table S8). One prophage was found to be in the form of the linear
427 plasmid ppDSJ01, present in 29 out of 30 analysed Pss genomes. Genomes in the Psi branch
428 contained between 0 and 6 complete prophage regions, ranging in size from 10.5 to 70.1 kb.
429 Only one prophage, Erwini_ENT90_NC_019932, was conserved in the majority of strains;
430 however, it is in Pss present only in genomes of CFBP 1719 and Pss 1990 (Supplement Table
431 S8-9).

432

433 **Pathogenicity factors and secretion systems**

434 The two main pathogenicity mechanisms in Pss are the type III secretion system (T3SS)
435 and the production of the exopolysaccharide stewartan. However, Pss contains many other
436 factors that contribute to its virulence in maize. To determine the presence of different

437 pathogenicity factors (PF) in Pss and other *P. stewartii* strains, the annotations of the 24
438 genomes sequenced in the present study and the 41 genomes from the NCBI databases were
439 analysed (Table 1, Table 2).

440 ***Secretion systems***

441 The Pss genome contains three different T3SS operons: one is encoded by the pPSS14
442 plasmid (T3SS-1), and the other two are encoded by the pPSS13 plasmid (T3SS-2 and T3SS-
443 3). All Pss genomes contain at least two T3SSs; 27 genomes contain all three. T3SS-1, which
444 is required for Pss pathogenicity in maize, is absent from the genomes of CFBP 3165 and CFBP
445 3169. T3SS-2, which is linked to bacterial colonisation of the corn flea beetle gut, is present in
446 all analysed genomes. The function of T3SS-3 is not fully understood, but it is present in all
447 Pss genomes except the type strain CFBP 3167^T (Figure 6). The majority (n = 28) of strains in
448 the Psi branch also contain some type of T3SS. T3SS-1 is present in 24 Psi genomes, while the
449 other two T3SSs found in Pss are absent from the genomes of other *P. stewartii* strains. Strains
450 NCPPB 1562, NCPPB 2282, SJM_1_1, and ST25 possess a different T3SS each, indicating
451 different acquisition pathways. NCPPB 1562 and NCPPB 2282 have the same type of T3SS;
452 however, the T3SSs of SJM_1_1 and ST25 are unrelated. Strains MS1, NRRL B-13, NS381,
453 RSA13, RSA30, RSA36, and S301 lack the complete T3SS apparatus and the associated
454 effectors (Figure 6).

455 In addition to the T3SS, *P. stewartii* strains contain other types of secretion systems. A
456 type II secretion system (T2SS) is encoded by plasmid pPSS13, which is present in all analysed
457 Pss genomes, but absent in all Psi branch strains. All Pss genomes contain a type IV secretion
458 system (T4SS), which is encoded by plasmids absent in other *P. stewartii* strains. The genomes
459 of strains NCPPB 1562 and NCPPB 2282 contain a T4SS that differs from that of Pss strains.
460 Pss genomes containing the newly discovered plasmids pPSS06 and pPSS09 also contain an

461 additional T4SS involved in horizontal gene transfer by conjugation. All *P. stewartii* genomes
462 contain nonfunctional parts of a type VI secretion system (T6SS) encoded on the bacterial
463 chromosome. Unlike Pss t Psi strains also encode for at least one additional T6SS locus, which
464 is thought to play an important role in the plant pathogenicity of these strains. Nineteen genomes
465 encode an additional third T6SS locus (Figure 6).

466 ***Other pathogenicity factors***

467 In addition to secretion systems, *P. stewartii* strains also have many other genes
468 involved in pathogenicity that have been reported in the literature. A survey was conducted to
469 identify 54 known and possible PFs that were, at the time of analysis, described in the existing
470 literature. These PFs were used in further analysis (Supplement Table S1). These include genes
471 involved in bacterial adhesion and aggregation, capsule formation, carbohydrate degradation,
472 carotenoid biosynthesis, flagellum-mediated surface motility, the inducible oxidative stress
473 response, transport across the bacterial membrane, quorum sensing, the secretion of effector
474 proteins, the acquisition of iron via siderophores, and the biosynthesis of stewartan and water
475 soaking during infection. All PFs were identified by BLAST analysis in at least one strain. The
476 majority of the pathogenicity factors were conserved in the analysed Pss strains, with the genes
477 involved in quorum sensing and stewartan biosynthesis being conserved in all analysed *P.*
478 *stewartii* genomes. All Pss strains also contain cellulase (RefSeq: WP_006122111.1), which,
479 in addition to T2SS, could play a role in pathogenicity. Additionally, gene *ucp1,* located on
480 plasmid pPSS10, which enables gut colonisation in the insect vector, is present in all Pss
481 genomes except Pss 1869; however, it is absent in all Psi genomes. Conversely, *P. stewartii*
482 strains lacking T3SS-1 also lack the associated effectors WtsE and the HrpN family
483 hypersensitivity reaction elicitor. The *wtsE* gene was also not found in the genome of CREA
484 DC 1775, which otherwise contains T3SS-1 and *hrpN*; however, this may also be due to gaps
485 in the genome assembly (Supplement Table S10).

Pathogenicity test

A pathogenicity test (PT) was performed to determine the pathogenicity of different Pss strains on maize and to study the effects of the different PFs present in these strains. The strains used in the PT were selected to include those lacking different T3SSs, as well as American and Slovenian isolates. The PT also included a Psi type strain, which was expected to cause no symptoms. Plants in the positive control group exhibited symptoms indicative of Pss infection, including yellow stripes on leaves, necrotic water-soaked lesions on leaves, and leaf deformities. In contrast, plants in the negative control group exhibited only yellow stripes near the inoculation site in some plants, resulting from mechanical damage caused by the needle and buffer. Symptoms were most severe in maize seedlings inoculated with the Pss strains CFBP 1719, CFBP 3445, NIB Z 2806, and NIB Z 3391, with an average symptom intensity greater than three at 3, 7, and 14 days post-inoculation (DPI) (Figure 7, Supplement Table S11). These strains were also the only ones still showing symptoms at 30 DPI. The CFBP 3167^{T} strain, which lacks T3SS-3, exhibited milder symptoms that did not result in stunted growth or leaf deformities. These symptoms then subsided, with no symptoms observed at 30 DPI. Pss strains CFBP 3165 and CFBP 3169, which lack T3SS-1, did not exhibit symptoms indicative of Pss infection (Figure 7, Supplement Table S11). The Psi strain CFBP 3614^{T}, which contains T3SS-1 but lacks other T3SSs, exhibited minimal symptoms following inoculation (averaging 1.4), which disappeared by 14 DPI and were not indicative of Pss infection (Figure 7, Supplement Table S11). Overall, showed differences in symptom severity between strains, with some strains showing severe symptoms while others lacked symptoms of infection with Pss.

511 **DISCUSSION**

512 Although *Pantoea stewartii* subsp. *stewartii* (Pss) is native to North America; detections
513 have been reported on other continents as well. Measures have been implemented to prevent its
514 introduction into the European Union, including testing of imported seed lots. Despite these
515 measures, several new detections have been reported in Europe in recent years. Notably, *Pss*
516 has been identified in Italy in association with seed maize production (in years 2022-2024),
517 representing one of the first confirmed occurrences of the pathogen in European commercial
518 seed systems (48, 49). These detections underscore the potential for inadvertent movement of
519 the bacterium through seed trade and highlight the importance of systematic surveillance.
520 In Slovenia, multiple findings have also been reported in recent years, concentrated within the
521 Vipava Valley in the west of the country, bordering Italy. This region has a distinct sub-
522 Mediterranean climate, which differs from the continental conditions prevailing across most of
523 Slovenia. The genomes of Slovenian isolates and additional Pss strains from bacterial
524 collections were sequenced in the scope of this study, covering detections reported up to 2022.
525 Subsequent findings and isolates are currently under analysis and will be presented separately.
526 These sequences, together with genome data available in public databases, were used to perform
527 phylogenetic and comparative genomic analyses to clarify the position of Slovenian isolates
528 within Pss subspecies, and the position of Pss within the *P. stewartii* species.
529 Previous comparative genomic studies have shown *P. stewartii* to be a distinct bacterial
530 species, most closely related to *P. ananatis* and *P. allii* (50–52). However, the position of Pss
531 within *P. stewartii* has remained unclear due to the scarcity of high-quality genomic data. At
532 the time of this work, the genome of strain DC283 was the only complete, high-quality
533 assembly on which much of our knowledge of Pss genomics was based, whereas other available
534 assemblies were fragmented or of low completeness. This study generated new, high-quality
535 genomic data that enabled robust comparative analyses of the Pss subspecies and of the *P.*

536 *stewartii* species as a whole. Phylogenetic analysis based on ANI revealed that *P. stewartii*
537 strains form two distinct clades: one comprising all Pss strains, which display a close
538 relatedness (>99.9% ANI), and another containing all *P. stewartii* subsp. *indologenes* (Psi) and
539 unassigned strains. The latter clade is more diverse, consistent with its broader range of host
540 plants, whereas Pss isolates are largely restricted to maize. Strain MS1, previously assigned to
541 the Pss, was nested within the Psi clade, which supports the findings of a previous study (53).

542 Given the close relatedness among Pss strains, core-genome SNPs were used for higher-
543 resolution phylogenetic analysis. Slovenian isolates formed a distinct branch in the
544 phylogenetic tree, except for the strain NIB Z 3391, which clustered separately. This pattern
545 suggests at least two independent introduction events, one of which may have resulted in limited
546 local dissemination within Slovenia. Strain NIB Z 3388 may represent an additional
547 introduction, as it shows lower relatedness to the remaining Slovenian isolates and forms a
548 separate branch in both accessory-genome and network analyses. The Italian isolates display a
549 comparable pattern, with most forming a separate branch except for strain CREA DC 1899.
550 The currently available genomes of European Pss isolates show no clear genetic connection
551 between findings from different countries, indicating that these are likely to represent multiple,
552 independent introduction events. Broader surveys and additional genome sequencing from
553 different geographical areas and time periods will be required to clarify the pathways of entry
554 and potential persistence of Pss in Europe.

555 Despite being closely related, Pss strains displayed diversity in plasmid content. The
556 analysed genomes contained between eight (CFBP 3167^{T}) and 14 (CFBP 3445 and CFBP 3157)
557 plasmids, which is higher than in other *P. stewartii* strains. Two previously undescribed
558 plasmids, pPSS06 and pPSS09, were identified. The total plasmid count may still be
559 underestimated in some assemblies, particularly those with low average genome coverage and
560 those assembled using only long reads (49). For example, the publicly available genome of

561 strain ATCC 8199 (GCF_045159535.1), which was not included in the analysis, contains only
562 five plasmids, whereas our assembly of strain CFBP 3167^{T}, which represents the same isolate,
563 contains eight. It is improbable that several plasmids were lost during cultivation, especially as
564 one of them, pPSS14, is highly conserved among *Pantoea* species. Genome assemblies,
565 however, remain models of genome organisation, and some apparent rearrangements may not
566 reflect biological variation. Among the plasmids identified, only the pPSS14 occurs in other
567 *Pantoea* species, where it is known as the large *Pantoea* plasmid (LPP-1) and plays an
568 important role in pathogenesis (54). Slovenian isolates share a similar plasmid composition
569 among them, except for isolate NIB Z 3391, consistent with their phylogenetic grouping. Italian
570 isolates show a comparable pattern, again with the exception of CREA DC 1899. The functional
571 importance of plasmids is underscored by the pathogenicity factors they encode: the T2SS on
572 pPSS13; T3SS on pPSS11 and pPSS14; and T4SS on pPSS06 and pPSS09. The presence of
573 these systems likely contributes to the observed diversity of plasmids within Pss. In addition to
574 plasmids, other mobile genetic elements, notably prophages, are abundant in Pss genomes.
575 They contain more prophage regions than those of Psi, suggesting an important evolutionary
576 role of prophages in the diversification of the Pss lineage.

577 The phylogenetic patterns are mirrored in the pangenome analysis. Pss has a smaller
578 pangenome than Psi, with a higher proportion of core genes, reflecting its lower overall
579 diversity. The species *P. stewartii* has an open pangenome, consistent with its extensive
580 complement of mobile genetic elements and its capacity for horizontal gene transfer (51). The
581 slower increase in gene number within the Pss pangenome likely reflects its narrower host
582 range. Genes unique to Pss may be associated with its pathogenicity in maize and its capacity
583 for insect-mediated transmission.

584 The presence and composition of pathogenicity factors vary between *P. stewartii*
585 subspecies. The two principal virulence determinants of Pss are the Hrp type III secretion

system (T3SS-1) and the exopolysaccharide stewartan (8, 55). The *wce-I*, *-II,* and *-III* gene clusters for stewartan biosynthesis were present in all analysed *P. stewartii*, confirming the potential for exopolysaccharide production. Pss also contains two additional T3SSs: one implicated in colonisation of the corn flea beetle gut, and another of as yet unclear function (56). The limited repertoire of effector proteins supports the view that Pss acquired the T3SS relatively recently (55). One such effector is WtsE, a member of the AvrE family of effectors that causes cell death, resulting in water-soaked lesions and necrosis in maize plants, and is essential for pathogenesis (6). Pss also possesses a T2SS absent from Psi, whereas Psi retains functional T6SS loci that are inactive in Pss (57). Only T3SS-1 is shared by both subspecies, being located on pPSS14, homologous to LPP-1. This indicates an older evolutionary origin of T3SS-1, although its loss in some strains of both subspecies points to relatively recent separate evolutionary events. Previous work showed that *P. stewartii* produces endoglucanase, which cleaves cellulose into smaller subunits and contributes to mobility in xylem vessels and to symptom development (12). Our analysis indicates that the T3SS-3 locus also encodes a cellulase and the expansin protein YoaJ, which binds cellulose and may enhance plant colonisation (58). This could explain why strains lacking T3SS-3 cause milder symptoms that do not progress to systemic infection. However, the further studies are needed to confirm the functional roles of T3SS-3 and YoaJ. Notably, T3SS-3 is absent only from the type strain CFBP 3167^{T}, one of the oldest isolates, making it difficult to determine whether this system was lost from that strain or gained in others.

Pathogenicity tests on maize seedlings confirmed the functional significance of the secretion systems. Strains lacking T3SS-1 did not cause typical Pss symptoms, confirming its essential role in virulence on maize. A Psi strain carrying T3SS-1 caused only mild symptoms, supporting the idea that other factors are also required for pathogenicity. The Pss type strain (CFBP 3167^{T}), lacking T3SS-3, produced only transient symptoms, further confirming that this

611 system contributes to disease development. This further confirmed its role in the pathogenicity
612 of Pss in maize. Together, these results corroborate the hypothesis that the absence of specific
613 secretion systems markedly reduces virulence.

614 Pss remains an important maize pathogen, and its recent detection in new regions has
615 renewed attention to its biology and movement. This study helps fill a long-standing knowledge
616 gap by providing complete and high-quality genome assemblies, thereby enabling the
617 placement of new isolates within the broader context of *P. stewartii* diversity. The results offer
618 new insights into the genetics and biology of Pss, including the diversity of its mobile genetic
619 elements and virulence factors. These findings are critical for improving diagnostic specificity
620 and informing measures to limit further introductions. However, more research is required to
621 understand the mechanisms governing its persistence and apparent geographical dispersal. The
622 increasing number of detections in Europe remains difficult to interpret without coordinated
623 monitoring across countries.

624 The detection of genetically similar isolates in geographically proximate locations may
625 indicate limited local dissemination, but further data are required to confirm this. These findings
626 nonetheless emphasise the need for continued surveillance and phytosanitary vigilance. The
627 situation in Slovenia most likely reflects proactive monitoring rather than a unique occurrence,
628 suggesting that similar introductions may be going undetected elsewhere. Preliminary
629 phenotypic analyses using BIOLOG plates indicate that Slovenian isolates possess broad
630 metabolic versatility consistent with survival under varying environmental conditions,
631 supporting the view that these findings reflect persistence rather than active establishment.

632 Taking together the available evidence, we cannot yet conclude that Pss is already
633 established in Slovenia. In addition, the bacterium has been detected in a weed species within
634 the affected area, while limited surveys have so far not confirmed the presence of potential
635 insect vectors. The principal North American vector, the corn flea beetle (*Chaetocnema*

636 *pulicaria*), is not known to be present in Europe and has not been reported from Slovenia. The
637 absence of early-season (spring) infections and the late appearance of symptoms, typically from
638 late summer to autumn, further support the view that active overwintering and the presence of
639 effective vector-mediated transmission have not yet been showcased. Ongoing studies are
640 addressing the potential role of alternative insects and the ecology of Pss in non-crop hosts.
641 These findings collectively indicate that eradication within the affected area remains a realistic
642 objective, provided that targeted phytosanitary measures and continued monitoring are
643 maintained. Strengthening genomic, phenotypic, and ecological surveillance in other regions
644 will therefore be crucial to assess the true frequency and extent of Pss introductions in Europe.
645 Slovenia provides a valuable case study for exploring the genomic diversity of recently detected
646 European isolates. Future efforts will concentrate on refining diagnostic and isolation methods,
647 expanding phenotypic characterisation, and supporting coordinated surveillance to ensure early
648 detection and containment of any further introductions.

649

650 **ACKNOWLEDGMENTS**

651 The authors acknowledge the financial support of the Slovenian Research and Innovation
652 Agency (research core funding No. P4-0165, Research Programme on Biotechnology and
653 Systems Biology, and young researchers programme – MR Aleksander Benčič, CRP V4-2415
654 MapQuest); the Ministry of Agriculture, Forestry and Food of the Republic of Slovenia
655 (Professional Tasks Programme in the Field of Plant Health; Euphresco projects 2018-A275
656 and 2023-A454; and CRP V4-2415 MapQuest); and the European Union Reference
657 Laboratories – Bacteriology co-funded by the Ministry of Agriculture, Forestry and Food of the
658 Republic of Slovenia; and the Department of Life Sciences and Facility Management of the
659 Zurich University of Applied Sciences (ZHAW), Wädenswil, Switzerland.

660 We are grateful to Primož Pajk (Administration of the Republic of Slovenia for Food Safety,
661 Veterinary and Plant Protection) for his continuous support and constructive discussions, and
662 to colleagues from the Chamber of Agriculture and Forestry – Institute of Nova Gorica (KGZ-
663 NG) and other regional services for conducting field inspections, visual assessments, and
664 sample collections within the official survey programmes.
665 We would like to thank S. Baeyen, J. Venneman and J. Van Vaerenbergh from Flanders
666 Research Institute for Agriculture, Fisheries and Food for their help in the implementation of
667 new bioinformatic analyses.
668 We thank the laboratory staff of the National Institute of Biology, in particular A. Blatnik, V.
669 Dukić, J. Matičič, L. Matičič, M. Pirc, Š. Prijatelj Novak and N. Turnšek, for their valuable
670 technical assistance. The successful isolation of the strains described in this study would not
671 have been possible without their routine yet essential work. We also thank A. Bogožalec Košir
672 for comments on the manuscript.

673 **REFERENCES**

674 1. Mergaert J, Verdonck L, Kersters K. 1993. Transfer of Erwinia ananas (synonym,
675 Erwinia uredovora) and *Erwinia stewartii* to the genus *Pantoea* emend. as *Pantoea*
676 *ananas* (Serrano 1928) comb. nov. and *Pantoea stewartii* (Smith 1898) comb. nov.,
677 respectively, and description of *Pantoea stewartii* subsp. *indologenes* subsp. nov. Int J
678 Syst Bacteriol 43:162–173.
679 2. Roper MC. 2011. *Pantoea stewartii* subsp. *stewartii*: lessons learned from a xylem-
680 dwelling pathogen of sweet corn. Mol Plant Pathol 12:628–637.
681 3. Li P, Zhang Y, Sun Y, Wu X, Wang Z, Zhou J, Zhou X. 2019. Comparative genomic
682 analysis of subspecies of *Pantoea stewartii* reveals distinct variations. J Plant Pathol
683 101:997–1004.
684 4. Mangeli F, Sistani F, Lori Z, Azadvar M, Hosseinipour A, Shahriyari N. 2024. First
685 report of maize leaf blight disease caused by *Pantoea stewartii* subsp *indologenes* in
686 Kerman Province, Iran. Crop Prot 184:106836.
687 5. Roper MC. 2011. *Pantoea stewartii* subsp. *stewartii*: Lessons learned from a xylem-
688 dwelling pathogen of sweet corn. Mol Plant Pathol 12:628–637.
689 6. Ham JH, Majerczak DR, Arroyo-Rodriguez AS, Mackey DM, Coplin DL. 2007. WtsE,
690 an AvrE-Family Effector Protein from *Pantoea stewartii* subsp. *stewartii*, Causes
691 Disease-Associated Cell Death in Corn and Requires a Chaperone Protein for Stability.
692 https://doi.org/101094/MPMI-19-1092 19:1092–1102.
693 7. Dolph PJ, Majerczak DR, Coplin DL. 1988. Characterization of a gene cluster for
694 exopolysaccharide biosynthesis and virulence in *Erwinia stewartii*. J Bacteriol 170:865–
695 871.
696 8. Koutsoudis MD, Tsaltas D, Minogue TD, Von Bodman SB. 2006. Quorum-sensing
697 regulation governs bacterial adhesion, biofilm development, and host colonization in

Pantoea stewartii subspecies *stewartii*. Proc Natl Acad Sci U S A 103:5983–5988.

9. Von Bodman SB, Majerczak DR, Coplin DL. 1998. A negative regulator mediates quorum-sensing control of exopolysaccharide production in *Pantoea stewartii* subsp. *stewartii*. Proc Natl Acad Sci 95:7687–7692.

10. Herrera CM, Koutsoudis MD, Wang X, Von Bodman SB. 2008. *Pantoea stewartii* subsp. *stewartii* Exhibits Surface Motility, Which is a Critical Aspect of Stewart's Wilt Disease Development on Maize. / 1359 MPMI 21:1359–1370.

11. Duong DA, Stevens AM, Jensen R V. 2018. Expanded Analysis of the *Pantoea stewartii* subsp. *stewartii* DC283 Complete Genome Reveals Plasmid-borne Virulence Factors. bioRxiv 407825.

12. Mohammadi M, Burbank L, Roper MC. 2012. *Pantoea stewartii* subsp. *stewartii* Produces an Endoglucanase That Is Required for Full Virulence in Sweet Corn. https://doi.org/101094/MPMI-09-11-0226 25:463–470.

13. Tambong JT, Mwange KN, Bergeron M, Ding T, Mandy F, Reid LM, Zhu X. 2008. Rapid detection and identification of the bacterium *Pantoea stewartii* in maize by TaqMan® real-time PCR assay targeting the cpsD gene. J Appl Microbiol 104:1525–1537.

14. Pal N, Block CC, Gardner CAC. 2019. A Real-Time PCR Differentiating *Pantoea stewartii* subsp. *stewartii* From *P. stewartii* subsp. *indologenes* in Corn Seed. Plant Dis 103:1474–1486.

15. Pirc M, Ravnikar M, Tomlinson J, Dreo T. 2009. Improved fireblight diagnostics using quantitative real-time PCR detection of *Erwinia amylovora* chromosomal DNA. Plant Pathol 58:872–881.

16. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Completing bacterial genome assemblies with multiplex MinION sequencing. Microb Genomics 3:e000132.

723 17. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack:
724 visualizing and processing long-read sequencing data. Bioinformatics 34:2666–2669.
725 18. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome
726 assemblies from short and long sequencing reads. PLOS Comput Biol 13:e1005595.
727 19. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM,
728 Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A V., Sirotkin A V., Vyahhi N, Tesler G,
729 Alekseyev MA, Pevzner PA. 2012. SPAdes: A New Genome Assembly Algorithm and
730 Its Applications to Single-Cell Sequencing. J Comput Biol 19:455.
731 20. Li H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long
732 sequences. Bioinformatics 32:2103–2110.
733 21. Vaser R, Sović, I, Sović, S, Nagarajan N, Šikić1 M, Šikić1 Š. 2017. Fast and accurate de
734 novo genome assembly from long uncorrected reads.
735 22. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads
736 using repeat graphs. Nat Biotechnol 2019 375 37:540–546.
737 23. Manni M, Berkeley MR, Seppey M, Zdobnov EM. 2021. BUSCO: Assessing Genomic
738 Data Quality and Beyond. Curr Protoc 1:e323.
739 24. Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. 2023. CheckM2: a rapid, scalable
740 and accurate tool for assessing microbial genome quality using machine learning. Nat
741 Methods 20:1203–1212.
742 25. Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: Multiple Alignment of
743 Conserved Genomic Sequence With Rearrangements. Genome Res 14:1394–1403.
744 26. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. 2021.
745 Bakta: Rapid and standardized annotation of bacterial genomes via alignment-free
746 sequence identification. Microb Genomics 7:000685.
747 27. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High

throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 2018 91 9:1–8.

28. Huson DH, Bryant D. 2024. The SplitsTree App: interactive analysis and visualization using phylogenetic trees and networks. Nat Methods 21:1773–1774.

29. Seemann T. 2015. snippy: fast bacterial variant calling from NGS reads.

30. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35:4453–4455.

31. Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin RB, O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens ES and HW. 2020. vegan: Community Ecology Package. R package version 2.5-7.

32. Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35:526–528.

33. Silva M, Machado MP, Silva DN, Rossi M, Moran-Gilad J, Santos S, Ramirez M, Carriço JA. 2018. chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. Microb genomics 4:e000166.

34. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:1–11.

35. Larralde M. 2022. Pyrodigal: Python bindings and interface to Prodigal, an efficient method for gene prediction in prokaryotes. J Open Source Softw 7:4296.

36. Team Rs. 2021. RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA.

37. Wickham H, Averick M, Bryan J, Chang W, D' L, Mcgowan A, François R, Grolemund

G, Hayes A, Henry L, Hester J, Kuhn M, Lin Pedersen T, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. 2019. Welcome to the Tidyverse. J Open Source Softw 4:1686.

38. François Briatte, Michał Bojanowski, Mickaël Canouil, Zachary Charlop-Powers, Jacob C. Fisher, Kipp Johnson, Tyler Rinker. 2025. ggnetwork: Geometries to Plot Networks with "ggplot2." CRAN Contrib Packag. Comprehensive R Archive Network (CRAN).

39. Gabor Csardi, Tamas Nepusz. 2006. The igraph software package for complex network research. InterJournal Complex Systems:1695.

40. Bergsma-Vlami M, Fornefeld E, Glenz R, Kołodziejska A, Giacobbi F, Gozzi R, Loreti S, Ambrus Á, Hederics D, Debonneville C, Dreo T, Pirc M, Bencic A, Venneman J, Vaerenbergh J Van, Baeyen S, Holeva M, Glynos PE, Vreeburg R, Bosch T van den, Pel C, Raaymakers T. 2025. Exploring the genetic diversity of recent *Ralstonia pseudosolanacearum* phylotype I findings in Europe.

41. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693.

42. Team RC. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

43. Wickham H. 2009. Elegant Graphics for Data Analysis. ggplot2.

44. Kassambara A. 2023. ggplot2 Based Publication Ready Plots • ggpubr.

45. Auguie B. 2017. Miscellaneous Functions for "Grid" Graphics [R package gridExtra version 2.3]. CRAN Contrib Packag.

46. Wishart DS, Han S, Saha S, Oler E, Peters H, Grant JR, Stothard P, Gautam V. 2023. PHASTEST: faster than PHASTER, better than PHAST. Nucleic Acids Res 51:W443–W450.

798 47. Guy L, Kultima JR, Andersson SGE, Quackenbush J. 2010. genoPlotR: comparative
799 gene and genome visualization in R. Bioinformatics 26:2334–2335.
800 48. NPPO of Italy. 2022. Update on the situation of Pantoea stewartii subsp. stewartii in
801 Italy. EPPO Report Serv 06–2022.
802 49. Scala V, Pucci N, Fiorani R, L'Aurora A, Polito A, Marsico M Di, Aiese Cigliano R,
803 Barra E, Ciarroni S, De Amicis F, Fascella S, Gaffuri F, Gallmetzer A, Giacobbi F,
804 Grieco PD, Gualandri V, Mason G, Pasqua di Bisceglie D, Rizzo D, Silletti MR, Talevi
805 S, Testa M, Tocci C, Loreti S. 2025. *Pantoea stewartii* subsp. *stewartii* an Inter-
806 Laboratory Comparative Study of Molecular Tests and Comparative Genome Analysis
807 of Italian Strains. Plants 14.
808 50. de Maayer P, Aliyu H, Vikram S, Blom J, Duffy B, Cowan DA, Smits THM, Venter SN,
809 Coutinho TA. 2017. Phylogenomic, pan-genomic, pathogenomic and evolutionary
810 genomic insights into the agronomically relevant enterobacteria *Pantoea ananatis* and
811 *Pantoea stewartii*. Front Microbiol 8:287061.
812 51. Agarwal G, Gitaitis RD, Dutta B. 2021. Pan-genome of novel *Pantoea stewartii* subsp.
813 *indologenes* reveals genes involved in onion pathogenicity and evidence of lateral gene
814 transfer. Microorganisms 9:1761.
815 52. Crosby KC, Rojas M, Sharma P, Johnson MA, Mazloom R, Kvitko BH, Smits THM,
816 Venter SN, Coutinho TA, Heath LS, Palmer M, Vinatzer BA. 2023. Genomic delineation
817 and description of species and within-species lineages in the genus *Pantoea*. Front
818 Microbiol 14:1254999.
819 53. Scala V, Faino L, Costantini F, Crosara V, Albanese A, Pucci N, Reverberi M, Loreti S.
820 2023. Analysis of Italian isolates of *Pantoea stewartii* subsp. *stewartii* and development
821 of a real-time PCR-based diagnostic method. Front Microbiol 14:1129229.
822 54. De Maayer P, Chan WY, Blom J, Venter SN, Duffy B, Smits THM, Coutinho TA. 2012.

823 The large universal Pantoea plasmid LPP-1 plays a major role in biological and
824 ecological diversification. BMC Genomics 13:1–12.
825 55. Ahmad M, Majerczak DR, Pike S, Hoyos ME, Novacky A, Coplin DL. 2007. Biological
826 Activity of Harpin Produced by *Pantoea stewartii* subsp. *stewartii*.
827 https://doi.org/101094/MPMI200114101223 14:1223–1234.
828 56. Correa VR, Majerczak DR, Ammar ED, Merighi M, Pratt RC, Hogenhout SA, Coplin
829 DL, Redinbaugh MG. 2012. the bacterium *Pantoea stewartii* uses two different type III
830 secretion systems to colonize its plant host and insect vector. Appl Environ Microbiol
831 78:6327–6336.
832 57. Ibrahim R, Izera Ismail S, Yasin Ina-Salwany M, Zulperi D. 2022. Biology, Diagnostics,
833 Pathogenomics and Mitigation Strategies of Jackfruit-Bronzing Bacterium *Pantoea*
834 *stewartii* subspecies *stewartii*: What Do We Know So Far about This Culprit?
835 Horticulturae 8:702.
836 58. Kerff F, Amoroso A, Herman R, Sauvage E, Petrella S, Filée P, Charlier P, Joris B,
837 Tabuchi A, Nikolaidis N, Cosgrove DJ. 2008. Crystal structure and activity of *Bacillus*
838 *subtilis* YoaJ (EXLX1), a bacterial expansin that promotes root colonization. Proc Natl
839 Acad Sci U S A 105:16876–16881.
840

Table 1: A report on the whole genome sequencing and genome assembly for all *Pantoea stewartii* subsp. *stewartii* strains sequenced within the scope of this study. The table includes metrics on the quality of genome assemblies includes size in megabases (Mb), number of contigs, guanine-cytosine content (GC%), number of coding sequences (CDSs) and number of transfer RNAs (tRNAs), shortest contig length that needs to be included for covering 50% of the genome (N50), completeness as percentage of complete BUSCOs (BUSCO (C) %), contamination percentage as determined by CheckM2 (Cont. %), number of complete plasmids present in genome assembly (Plasmids), and the accession number of genome assemblies in European Nucleotide Archive.

Strain[a]	Genome size (Mb)	Contigs	GC%	CDSs	tRNAs	N50	BUSCO (C) %	Cont. %	Plasmids	Acc. No.
CFBP 1719	5.3	13	53.8	5,411	74	4,480,175	98.6	0.35	8	GCA_977010205
CFBP 2502	5.5	15	53.7	5,694	74	4,610,069	98.5	0.45	7	GCA_977010145
CFBP 3157^{C}	5.5	15	53.6	5,762	74	4,609,570	98.6	0.59	11	GCA_977011095
CFBP 3165	5.1	14	53.9	5,237	74	4,547,336	98.3	0.16	10	GCA_977011085
CFBP 3166^{C}	5.4	12	53.7	5,625	74	4,600,323	98.7	0.58	12	GCA_977011055
CFBP 3167^{T}	5.2	14	53.9	5,325	74	4,500,371	98.4	0.33	8	GCA_977013035
CFBP 3169^{C}	5.1	11	53.8	5,285	74	4,572,251	98.4	0.08	11	GCA_977011105
CFBP 3393^{C}	5.3	12	53.8	5,525	74	4,567,718	98.2	0.26	14	GCA_977011075
CFBP 3394	5.3	15	53.8	5,524	74	1362,061	98.6	0.31	4	GCA_977011115
CFBP 3395^{C}	5.3	12	53.8	5,477	74	4,556,511	98.5	0.31	10	GCA_977011135
CFBP 3396	5.4	62	53.6	5,578	70	243,803	98.6	0.71	10	GCA_977011145
CFBP 3445^{C}	5.5	15	53.7	5,730	75	4,598,621	98.7	0.58	8	GCA_977011165
NCPPB 3253^{C}	5.5	11	53.7	5,741	75	4,653,398	98.7	0.5	11	GCA_977011125
NIB Z 2806^{C}	5.4	13	53.7	5,614	75	4,579,505	98.7	0.5	14	GCA_977009295
NIB Z 2809^{C}	5.5	13	53.7	5,687	75	4,626,859	98.7	0.5	10	GCA_977010175
NIB Z 3057^{C}	5.4	13	53.7	5,671	75	4,619,739	98.7	0.36	11	GCA_977011175
NIB Z 3386^{C}	5.5	11	53.7	5,679	75	4,626,474	98.6	0.49	10	GCA_977011155
NIB Z 3388	5.5	16	53.6	5,730	75	4,684,951	98.5	0.51	10	GCA_977009985
NIB Z 3391	5.2	203	53.7	5,441	69	134,045	98.7	0.42	9	GCA_977011785
NIB Z 3394	5.4	55	53.7	5,644	75	498,560	98.7	2.04	10	GCA_977011765
NIB Z 3506	5.5	17	53.7	5,678	75	4,625,939	98.6	0.58	10	GCA_977011795
NIB Z 3516^{C}	5.4	11	53.7	5,669	75	4,624,962	98.7	0.58	4	GCA_977011805
NIB Z 3526	5.4	14	53.7	5,655	75	4,626,258	98.7	0.57	12	GCA_977011775
NIB Z 3532	5.4	16	53.7	5,569	75	4,558,672	98.6	0.47	12	GCA_977011755

[a] Superscript codes following the strain name correspond to: ^{C}, complete genomes; ^{T}, type strain of *Pantoea stewartii* subsp. *stewartii*

855 **Table 2:** The publicly available *Pantoea stewartii* genomes used in this study are listed below.
856 Data on genome assemblies includes: GenBank or Sequence Read Archive accession number,
857 size in megabases (Mb), number of contigs, guanine-cytosine content (GC%), number of coding
858 sequences (CDSs), shortest contig length that needs to be included for covering 50% of the
859 genome (N50), genome coverage (×) as provided in NCBI database and completeness as
860 percentage of complete BUSCOs (BUSCO (C) %). Additional information on genomes
861 assembled from raw sequences is available in Supplement Table S3.

862

863

Subspecies	Strain[a]	Acc. No.	Genome size (Mb)	Contigs	GC%	CDSs	N50	Genome coverage (×)	BUSCO (C) %
stewartii	DC283[C]	GCA_002082215.1	5.3	12	53.8	5,485	4,528,215	200	98.5
	CREA DC 1775	GCA_030370585.1	5.6	60	53.7	5,873	4,601,780	30	98.7
	CREA DC 1899	GCA_030370545.1	5.5	17	53.6	5,752	4,566,383	30	98.5
	Pss 1869[SRA]	SAMN48331321	5.3	45	53.7	5,555	1,682,138	NA	98.6
	Pss 1870[SRA]	SAMN48331320	5.4	16	53.7	5,602	4,567,131	NA	98.7
	Pss 1990[SRA]	SAMN48331319	5.4	16	53.7	5,674	4,622,825	NA	98.7
	MS1	GCA_010273335.1	4.8	59	53.7	4,407	480,566	not available	99.6
indologenes	LMG 2632[T]	GCA_000757405.2	4.7	35	53.7	4,454	304,929	99	96.7
	LMG 2671	GCA_030370575.1	4.9	3	53.6	4,524	4,443,848	30	99.7
	NCPPB 1562	GCA_017051845.1	4.9	30	53.4	4,525	332,508	200	99.5
	NCPPB 1877	GCA_017051875.1	4.8	36	53.6	4,410	286,090	200	99.7
	NCPPB 2275	GCA_017051895.1	4.8	32	53.6	4,406	411,329	200	99.5
	NCPPB 2281	GCA_017051805.1	4.7	26	53.7	4,315	354,374	200	99.7
	NCPPB 2282	GCA_017051815.1	4.9	28	53.4	4,525	332,508	200	99.5
	PANS 07-4	GCA_017052095.1	5.0	37	53.5	4,685	348,948	200	99.7
	PANS 07-6	GCA_017052115.1	5.1	37	53.5	4,740	348,948	200	99.7
	PANS 07-10	GCA_017051975.1	4.9	36	53.7	4,603	401,017	200	99.7
	PANS 07-12	GCA_017052015.1	4.9	36	53.7	4,606	401,017	200	99.7
	PANS 07-14	GCA_017051935.1	4.8	29	53.7	4,426	349,987	200	99.7
	PANS 99-15	GCA_017051945.1	4.8	36	53.8	4,425	364,980	200	99.6
	PNA 14-9	GCA_017052375.1	4.7	24	53.7	4,310	386,228	200	99.7
	PNA 14-11	GCA_017052195.1	4.7	24	53.7	4,309	573,378	200	99.7
	PNA 14-12	GCA_017052135.1	4.7	23	53.7	4,309	573,378	200	99.7
	PNA 15-2	GCA_017052175.1	4.6	26	53.7	4,254	382,660	200	99.7
	SJM_1_1[C]	GCA_040746145.1	5.2	5	53.7	4,831	4,417,030	96	99.7
no subspecies	626	GCA_013277595.1	4.8	43	53.6	4,408	638,908	79	99.7
	A206	GCA_001310285.1	4.7	19	53.8	4,299	646,058	168	99.6
	C10109	GCA_029991035.1	4.6	20	53.7	4,270	366,162	558	99.5
	HR3-48[C]	GCA_025765915.1	4.8	2	53.9	4,351	4,439,845	200	99.6
	ICPM 10132	GCA_029433915.1	4.2	13	53.6	4,877	4,209,699	40	99.6
	M009	GCA_000786255.1	4.8	56	53.9	4,446	223,175	44	99.4
	M073a	GCA_000803205.1	4.8	39	53.9	4,440	375,223	56	99.5
	NC66	GCA_035787335.1	5.0	70	53.5	4,643	607,939	70	99.7
	NS381	GCA_001476355.1	4.7	51	53.8	4,315	268,280	82	99.7
	RON18713[C]	GCA_030064655.1	4.6	2	53.9	4,227	4,334,600	180	97.8
	RSA13	GCA_001477215.1	4.8	48	53.7	4,405	167,513	96	99.5
	RSA30	GCA_001476795.1	4.8	84	53.7	4,392	211,952	65	99.6
	RSA36	GCA_001476375.1	4.8	76	53.7	4,419	285,765	139	99.6
	S301	GCA_001310295.1	4.5	30	53.9	4,127	576,093	230	99.7
	ST25	GCA_025599245.1	4.8	35	53.6	4,418	570,997	190	99.5
	ZJ-FGZX1[C]	GCA_011044475.1	5.0	3	53.5	4647	4,550,072	100	99.7%

864

865 [a] Superscript codes following the strain name correspond to: [C], complete genomes; [SRA],
866 genomes assembled from raw reads; [*], strain assigned to subspecies *stewartii* which proved to
867 be more closely related to subspecies *indologenes*; [T], type strain of *Pantoea stewartii* subsp.
868 *indologenes*

869

870 **Table 3:** Pangenome composition for *Pantoea stewartii* subspecies *stewartii* (Pss), subspecies
871 *indologenes* (Psi), and the whole species *P. stewartii*. The table includes the criteria for
872 inclusion and number of genes in core, soft core, shell, cloud and the total number of genes
873 for each group.

Pangenome	Criteria	Pss (*N* = 30)	Psi (*N* = 35)	*P. stewartii* (*N* = 65)
Core genes	From 99 to 100% strains	3,968	3,496	3,013
Soft core genes	From 95 to 100% strains	443	92	250
Shell genes	From 15 to 95% strains	1,584	1,260	3,456
Cloud genes	From 0 to 15% strains	1,430	7,104	8,840
Total genes	All strains	7,425	11,952	15,559

874

875 **FIGURES**

876 ## Figure legends.

877 **Figure 1**: **Phylogeny of the *Pantoea stewartii* species.** A phylogenetic tree based on the
878 average nucleotide identity (ANI) of 65 *P. stewartii* genomes, obtained from NCBI GenBank
879 or assembled in this study. Strains belonging to the *stewartii* (Pss) subspecies are shown in
880 cyan, those belonging to the subspecies *indologenes* (Psi) in orange, and strains not assigned to
881 any subspecies in green. The figure also shows metadata including the geographic origin
882 (country), year of isolation, and host organism. All Pss strains form a separate branch within
883 the phylogenetic while all strains without an assigned subspecies are dispersed among the Psi
884 strains within the phylogenetic tree.

885 ^{T}: strains CFBP 3167^{T} and LMG 2632^{T} are type strains of subspecies *stewartii* and *indologenes*,
886 respectively.

887 *: This indicates strain MS1, which was previously assigned to subspecies *stewartii,* but which
888 is more closely related to other *P. stewartii* strains.

889

890 **Figure 2**: **Core genome-based phylogenetic analysis of *Pantoea stewartii* subsp. *stewartii*.**
891 **A**: Phylogenetic tree constructed from high-quality *P. stewartii* subsp. *stewartii* (Pss) genomes
892 single nucleotide polymorphisms alignment and maximum likelihood method with 1000
893 bootstrap replicates. The bootstrap values are shown as percentages in the nodes of the tree.
894 The figure further shows metadata, including the geographic origin (country), year of isolation,
895 and host organism strain numbers. **B**: Phylogenetic network of Pss constructed using P
896 distances. **C:** Minimum spanning tree generated from core genome multiple locus strain typing
897 showing allelic differences between strains. Node contains strain numbers of strains as listed in
898 the table in panel A. Names of the strains are coloured according to country of origin in the
899 whole figure, with Slovenian strains shown in green, the USA strains in red, and the Italian
900 strains in blue as shown in the legend.

901

902 **Figure 3**: **Accessory genome-based phylogenetic analysis of *Pantoea stewartii* subsp.**
903 ***stewartii***. **A**: Unrooted phylogenetic tree of *P. stewartii* subsp. *stewartii* (Pss) constructed from
904 the accessory genome (genes present in under 95% genomes) binary matrix using Jaccard

905 distances and neighbour-joining clustering method. **B**: Phylogenetic network of Pss constructed
906 using *p* distances. Names of the strains are coloured according to country of origin in the whole
907 figure, with Slovenian strains shown in green, the USA strains in red, and the Italian strains in
908 blue, as shown in the legend.

909

910 **Figure 4**: **Pangenome analysis of *Pantoea stewartii* and its subspecies**. The pangenomes of
911 Pantoea stewartii subsp. stewartii (Pss), strains of *P. stewartii* subsp. *indologenes* (Psi), and the
912 whole *P. stewartii* species are shown. Panels A, C, and E illustrate how the pangenome's gene
913 count increases with each additional genome (in blue), while the core genome's count decreases
914 with each additional strain, stabilising at a set number of core genes characteristic of *P.*
915 *stewartii*. Panels B, D, and F show the number of genes present in a given number of genomes.
916 Panels A and B show the data for Pss, panels C and D show the data for non-Pss strains, and
917 panels E and F show the data for all *P. stewartii* strains. Panel G shows the pangenome
918 composition for Pss, Psi, and *P. stewartii*, expressed as a percentage of genes included in the
919 core, soft core, shell, or cloud.

920

921 **Figure 5**: **Plasmid diversity of *Pantoea stewartii* subsp. *stewartii*.** Strains are listed according
922 to their position within the core genome phylogenetic tree, which is presented as a dendrogram
923 on the left side of the figure. The presence of plasmids in strains is indicated by a green colour,
924 while an empty field indicates that the plasmid was not identified in these genome assemblies.
925 A total of 15 different plasmids were identified among the 30 *Pantoea stewartii* subsp. *stewartii*
926 (Pss) strains. The number of plasmids per strain ranges from eight (CFBP 3167^{T}) to 14 (CFBP
927 3445), and no strain contains all known plasmids. Details about the plasmids and the
928 nomenclature used in this figure can be found in Supplemental Table S4.

929

930 **Figure 6**: **Diversity of secretion systems among *Pantoea stewartii* strains.** This figure shows
931 the presence or absence of the following secretion systems, as identified using the genome
932 annotations of 65 *Pantoea stewartii* strains: type II (T2SS; turquoise), type III (T3SS-1, -2, -3,
933 -O; green), type IV (T4SS-1, -2, -3, -O; olive green), and type VI (T6SS-1, -2, -3; ochre). A
934 white field indicates the absence of a secretion system in this strain. The strains are listed in
935 order of their position in the dendrogram shown on the left.

936

937 **Figure 7: Results of pathogenicity tests on maize for different *Pantoea stewartii* strains.**
938 The severity of the symptoms caused by the different *Pantoea stewartii* strains was assessed at
939 3-, 7-, 14-, and 30-day post-inoculation. Maize seedlings in the negative control (NC) group
940 were inoculated with 10 mM phosphate-buffered saline (PBS), while those in the positive
941 control (PC) group were inoculated with strain NIB Z 2806.

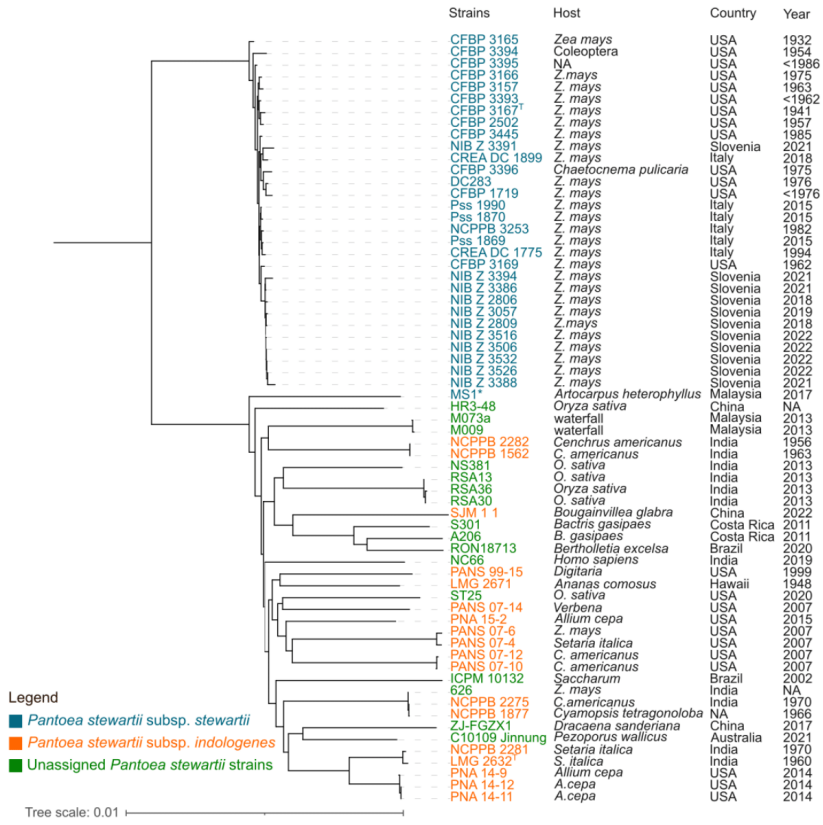
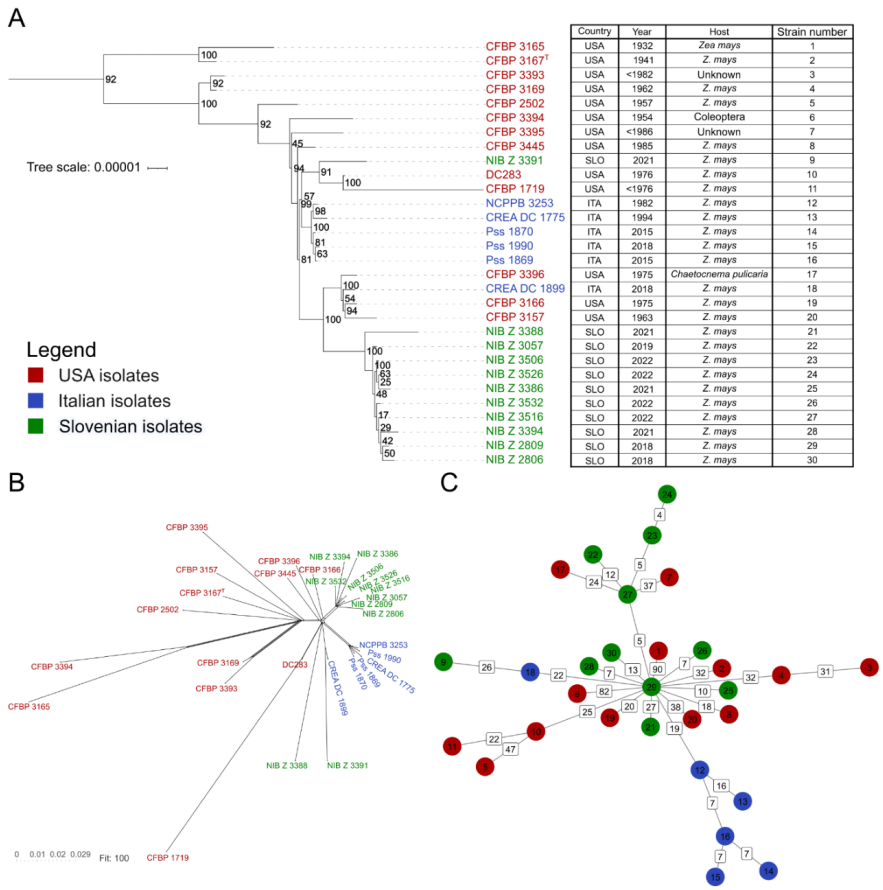
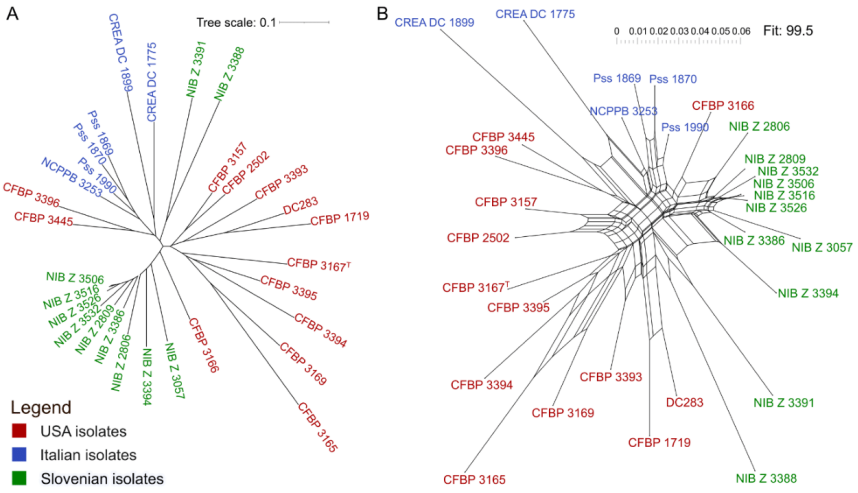


Figure 1



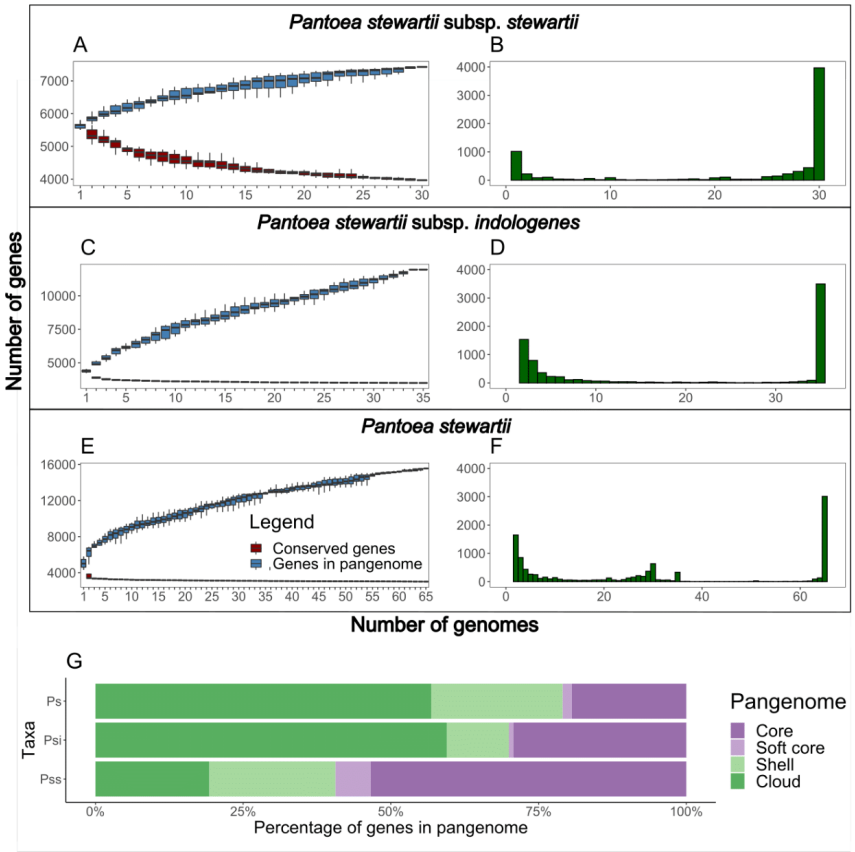
Country	Year	Host	Strain number
USA	1932	*Zea mays*	1
USA	1941	*Z. mays*	2
USA	<1982	Unknown	3
USA	1962	*Z. mays*	4
USA	1957	*Z. mays*	5
USA	1954	Coleoptera	6
USA	<1986	Unknown	7
USA	1985	*Z. mays*	8
SLO	2021	*Z. mays*	9
USA	1976	*Z. mays*	10
USA	<1976	*Z. mays*	11
ITA	1982	*Z. mays*	12
ITA	1994	*Z. mays*	13
ITA	2015	*Z. mays*	14
ITA	2018	*Z. mays*	15
ITA	2015	*Z. mays*	16
USA	1975	*Chaetocnema pulicaria*	17
ITA	2018	*Z. mays*	18
USA	1975	*Z. mays*	19
USA	1963	*Z. mays*	20
SLO	2021	*Z. mays*	21
SLO	2019	*Z. mays*	22
SLO	2022	*Z. mays*	23
SLO	2022	*Z. mays*	24
SLO	2021	*Z. mays*	25
SLO	2022	*Z. mays*	26
SLO	2022	*Z. mays*	27
SLO	2021	*Z. mays*	28
SLO	2018	*Z. mays*	29
SLO	2018	*Z. mays*	30

Figure 2



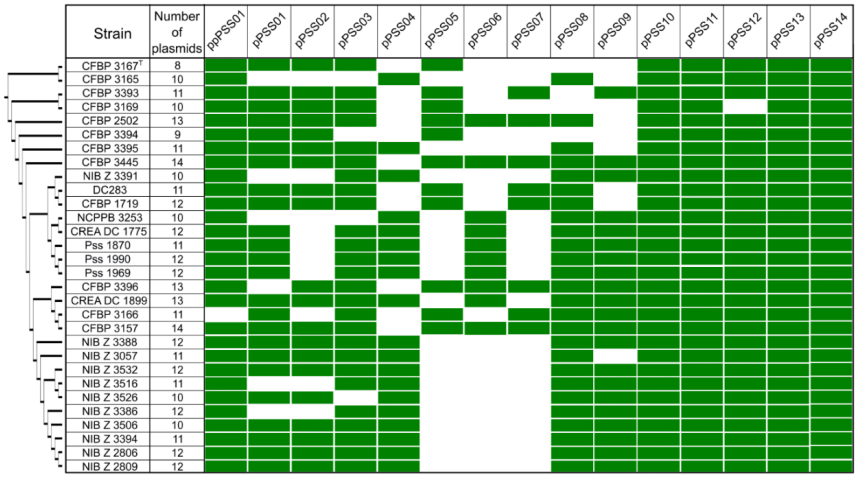
946

947 **Figure 3**



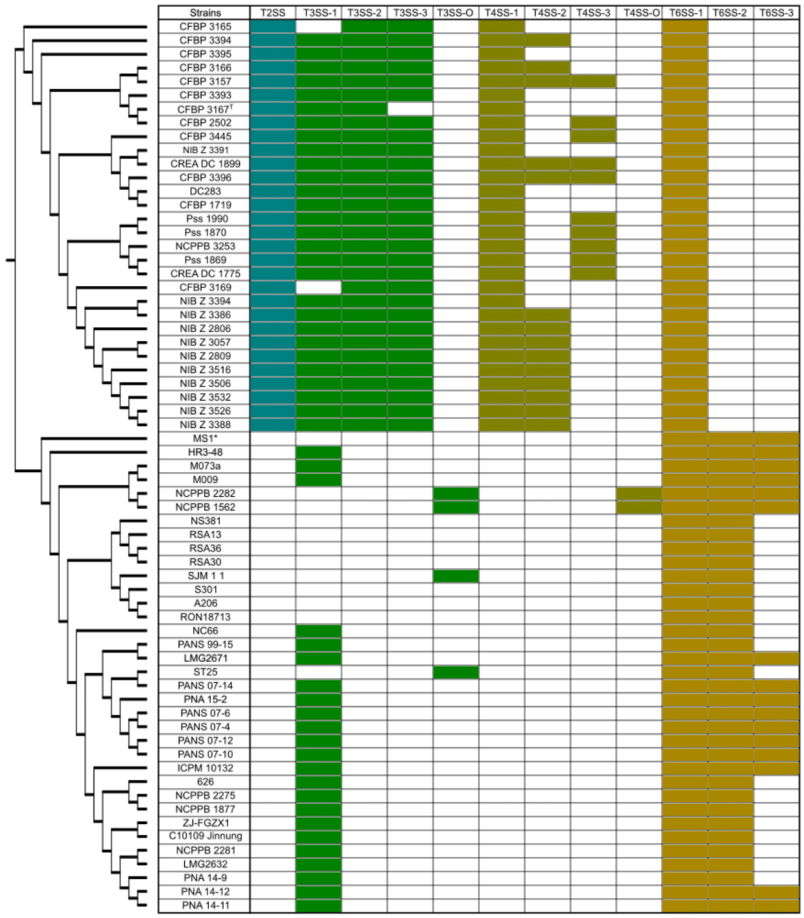
948

949 **Figure 4**



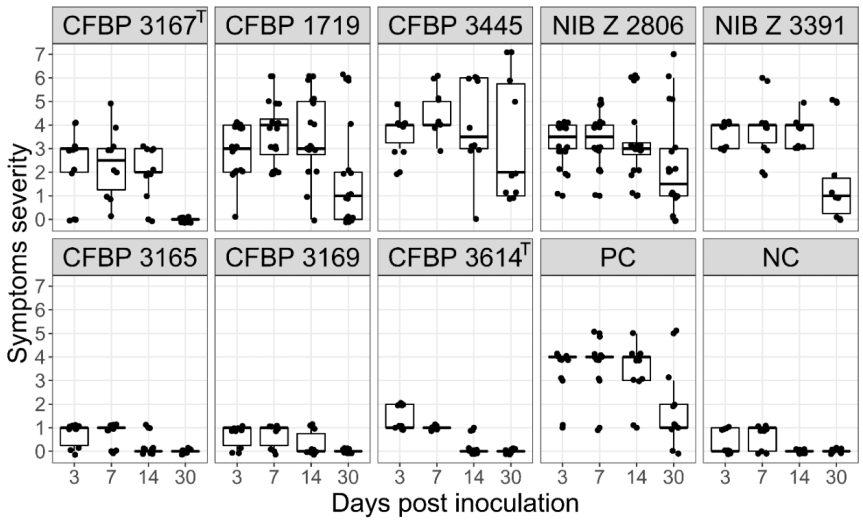
950

951 **Figure 5**



952

953 **Figure 6**



954

955 **Figure 7**

2.3 Complete Genome Sequence Resource for *Xanthomonas translucens* pv. *undulosa* MAI5034, a Wheat Pathogen from Uruguay

Felipe Clavijo, Claudia Barrera, Aleksander Benčič, Valentina Croce, Jonathan M. Jacobs, Adriana J. Bernal, Ralf Koebnik, Veronica Roman-Reyna

Phytopathology, 2022 Sep;112(9):2036-2039. doi: 10.1094/PHYTO-01-22-0025-A

This article describes the first complete genome sequence of *Xanthomonas translucens* to be isolated in South America. *X. translucens* is a plant-pathogenic bacterium responsible for bacterial leaf streak, black chaff of small grains, and bacterial wilt of forage grasses, with pathovar *undulosa* being most associated with wheat. Strain MAI5034 was isolated in 2018 from symptomatic leaf tissue and identified as *X. translucens* pv. *undulosa* based on multilocus sequence analysis. Bacterial DNA was sequenced using nanopore technology to obtain complete genomes. The reads were assembled into genomes using multiple assemblers, with the Racon-polished Flye assembly demonstrating superior continuity, completeness, and quality. Homopolish was then applied to the assembly to resolve issues with homopolymeric nucleotide runs. This yielded a single circular chromosome measuring 4.6 Mbp, with a G+C content of 64.7%, and 139× sequencing coverage. The genome of strain MAI5034 contains seven TALEs that are involved in phytopathogenicity and are widely conserved in *X. translucens* pv. *undulosa*. This indicates the recent worldwide spread of this pathovar. The low diversity of TALEs is consistent with the lower genetic diversity of the pathovar *undulosa*.

The PhD candidate contributed to this work by performing a bioinformatic analysis yielding a genome assembly. He also performed a phylogenetic analysis of *X. translucens* pv. *undulosa*, presented in the article as a phylogenetic tree.

Phytopathology® • 2022 • 112:2036-2039 • https://doi.org/10.1094/PHYTO-01-22-0025-A

RESOURCE ANNOUNCEMENT

Complete Genome Sequence Resource for *Xanthomonas translucens* pv. *undulosa* MAI5034, a Wheat Pathogen from Uruguay

e-Xtra*

Felipe Clavijo,[1] Claudia Barrera,[2] Aleksander Benčič,[3,4] Valentina Croce,[1] Jonathan M. Jacobs,[5,6] Adriana J. Bernal,[2] Ralf Koebnik,[7,†] and Veronica Roman-Reyna[5,6]

[1] Laboratorio de Microbiología Molecular, Departamento de Biociencias, Facultad de Química, Universidad de la República, Montevideo, Uruguay
[2] Department of Biological Sciences, Universidad de Los Andes, Bogotá, Colombia
[3] National Institute of Biology, Ljubljana, Slovenia
[4] Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
[5] Department of Plant Pathology, The Ohio State University, Columbus, OH, U.S.A.
[6] Infectious Diseases Institute, The Ohio State University, Columbus, OH, U.S.A.
[7] Plant Health Institute of Montpellier, University of Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France

Funding
Support was provided by the Institut de Recherche pour le Développement, France, for the North-South-South Network on Xanthomonads within the International Scientific Coordination Network-South (GDRI-Sud NSSN-X); the European Cooperation in Science and Technology (COST) Programme for support of A. Benčič and R. Koebnik through the EuroXanth COST Action CA16107, which was key in initiating this collaborative genome project; the American Malting Barley Association funding to J. M. Jacobs; the Faculty of Sciences at Universidad de los Andes to A. J. Bernal; and the Slovenian Research Agency to A. Benčič.

Keywords
evolution, genetic diversity, TAL effector, wheat, *Xanthomonas*

Genome Announcement

The bacterial species *Xanthomonas translucens* is responsible for bacterial leaf streak and black chaff of small grains and bacterial wilt of forage grasses (Egli et al. 1975; Jones et al. 1917; Sapkota et al. 2020; Vauterin et al. 1995). In particular, bacterial leaf streak of wheat represents the most limiting bacterial disease in wheat production worldwide (Duveiller et al. 1997). The *X. translucens* pathovar most commonly associated with wheat is pathovar *undulosa* (Smith et al. 1919).

Until recently, genome sequences of 61 strains were available for this species at NCBI GenBank, 16 of which are complete genomes. Only one *X. translucens* genome sequence from a South American strain, UPB787, is currently available, which was isolated from barley in Paraguay in 1990. To enlarge the geographic coverage of the pathogen's genomic resources, we present the first complete genome sequence of a South American strain of *X. translucens*, which reveals a surprising conservation of its repertoire of transcription activator-like effectors (TALEs) across continents.

Strain MAI5034 was isolated in October 2018 from symptomatic wheat leaf tissue obtained in Soriano, Uruguay (Clavijo et al. 2022). Pathogenicity on wheat was confirmed in greenhouse assays. The strain was identified as *X. translucens* pv. *undulosa* by multilocus sequence analysis of the concatenated partial sequences of four housekeeping genes (*dnaK*, *fyuA*, *gyrB*, and *rpoD*) and, through multilocus sequence typing, it was assigned to novel sequence type ST3 (Clavijo et al. 2022).

Strain MAI5034 was grown at 28°C on peptone-sucrose-agar medium (0.5% peptone, 2% sucrose, and 1.5% agar) for 24 h. Bacteria were then resuspended in 10 mM $MgCl_2$ and diluted to an optical density at 600 nm of 1.0. Cells from 2 ml were harvested by centrifugation and washed once with 10 mM $MgCl_2$, and genomic DNA was isolated using Qiagen Genomic tip 100/G (Qiagen, Hilden, Germany) according to the manufacturer's instructions.

†Corresponding author: R. Koebnik; ralf.koebnik@ird.fr

R. Koebnik and V. Roman-Reyna contributed equally to this work.

*The *e*-Xtra logo stands for "electronic extra" and indicates that one supplementary table is published online.

The author(s) declare no conflict of interest.

Accepted for publication 28 March 2022.

Table 1. Repeat-variable diresidue (RVD) sequences of completely sequenced *Xanthomonas translucens* pv. *undulosa* strains[a]

TALE class	Strains	RVD pattern
TalCT	ICMP11055	HN HD HD HD NI NI NI HN HD HD NH NN NI NN HD
	LW16, XtFa1	NN HD HD HD NI NI NI HN HD HD NN NN NI NN HD
	MAI5034, P3, XtKm12, XtKm15, XtLr8, Xtu4699	HN HD HD HD NI NI NI HN HD HD NN NN NI NN HD
TalCZ	ICMP11055, LW16, MAI5034, P3, XtFa1, XtKm12, Xtu4699	NH NN HD NN HD NH HD YK NG NH Y* HD NN NI NG QD
TalDA	ICMP11055, MAI5034, P3, XtFa1, XtKm12, XtKm15, XtLr8, Xtu4699	HD YD NI NG NG NN YK NG HD NG NG ND NG QD NH HD
	LW16	HD YD NI NG NG NN YK NG HD NG NG ND NG QD NH QD
TalDB	LW16	NN HD KG HD HD HN NF NI NN HD HD HD HN HN HD
	P3	NN HD NG HD HD HN NF NI NN NN HD HD HN HN HD
	XtKm15,[b] XtLr8	NN HD NG HD HD HN NF NI NF HD HD HD HN HN HD
	Xtu4699	NN HD NG HD HD HN NF NI NH HD HD HD HN HN HD
TalDC	MAI5034, P3, XtFa1, XtLr8, Xtu4699	NN NG HD HD HD KG NN Y* NG HD HD QD HN
TalDD	ICMP11055	NN HD NG NN HN KG NI HD NI HN HD HN HD Y* NG HD HD HN
	LW16	NN HD NG NN HN KG NI HD NI HN HD HN HD HD NI HN HN HD
	MAI5034	NN HD NG NN HN KG NI HD NI HN HD HN HD HD NI HN HD QD
	P3, XtFa1, XtKm12, XtKm15, XtLr8, Xtu4699	NN HD NG NN HN KG NI HD NI NN HD HN HD HD NI HN HD QD
	LW16, XtKm12	NN HD NG NN HN NG NI HD NI NN HD HD NN NN NI HN HD
TalDE	LW16, MAI5034	NN HD NG NN HN HN NN NI NI NH NN HD NN NH HD HD
	P3, XtFa1, XtKm15,[b] Xtu4699	NN HD NG NN HN HN NI NI NI NH NN HD NN NH HD HD
	XtKm12	NN HD NG NN HN HN NI NI NI NH NN HD HN NH HD HD
TalDF	ICMP11055, LW16, MAI5034, P3, XtKm12, XtKm15, XtLr8, Xtu4699	HD HN HN HD NH NH HG HD KG NN Y* NG HD HD HN
	XtFa1	HD HN HN HD NH NH HG HD KG NN Y* NG HD NI NH NG HD HN
TalHM	ICMP11055	NN HD NG HD HD HG HD KG NN Y* NG NG HD HD QD HN
	XtKm15[b]	NN HD NG HD HD HG HD KG
TalHN	ICMP11055	NN HD NG HD NG HD HD HG HD KG NN KG HD HN QD HN
TalJD	XtLr8	NN HD NG NN Y* NG HD HD NN NH HD HD

[a] Alignment of RVDs of transcription activator-like effectors (TALEs) of *X. translucens* pv. *undulosa* strains. RVDs that differ between otherwise highly conserved TALEs are underlined. An asterisk (*) indicates that the second amino acid of the RVD is missing.
[b] Three TALE genes of strain XtKm15(*talDB*, *talDE*, and *talHM*) are recorded as pseudogenes due to in-frame stop codons or frameshift mutations (http://www.jstacs.de/downloads/List_of_classes.txt) (Grau et al. 2016).

For library construction and DNA sequencing, performed by ohmX.bio (Gent, Belgium), 1 μg of DNA was mechanically fragmented with g-Tubes (Covaris) at approximately 13 kb. A sequencing library was prepared with the Ligation Sequencing kit (SQK-LSK110; Oxford Nanopore Technologies [ONT]) and the Native Barcode Expansion (PCR-free) (EXP-NBD114; ONT) based on the manufacturer's protocol. For multiplexing with other samples, a unique barcode (barcode ID NB19, GTTCCTCGTGCAGTGTCAAGAGAT) was ligated to the sample type using the ONT direct-DNA (SQK-LSK109) library preparation kit in combination with the Native barcoding expansion (EXP-NBD104). Upon pooling of eight libraries, samples were sequenced on a GridION, R9.4 flow cell, and sequence reads were demultiplexed by ohmX.bio and provided as FASTQ files.

Sequence reads were trimmed with Porechop (v0.2.4) and assembled using three different algorithms: Flye (version 2.8.1-b1676), Shasta, and Miniasm (Kolmogorov et al. 2020; Li 2016; Shafin et al. 2020; Wick 2017). In addition, the Flye and Miniasm assemblies were polished using Racon (Vaser et al. 2017). A comparison of these assemblies revealed the superior performance of the Racon-polished Flye assembly, as indicated by its better contiguity, completeness, and quality (Supplementary Table S1). However, manual inspection revealed significant issues with homopolymeric nucleotide runs, resulting in four frame shifts per TALE gene. Therefore, we applied Homopolish (version 0.0.1) on the Racon-polished Flye assembly (Huang et al. 2021), which resolved all of the problems at the TALE genes and reduced the number of predicted pseudogenes from 315 to 202 (see below).

This procedure yielded one circular chromosome of 4,625,916 bp with a typical G+C content of 64.7%, corresponding to 139× sequencing coverage. The chromosome was annotated with GeneMarkS-2+ (Lomsadze et al. 2018), as implemented in the NCBI Prokaryotic Genome Annotation Pipeline (https://www.ncbi.nlm.nih.gov/genome/annotation_prok/), which predicted a total of 4,001 genes, including 3,736 coding genes, 202 pseudogenes, 53 transfer RNA genes, 4 noncoding RNAs, and 2 rRNA operons (5S, 16S, and 23S).

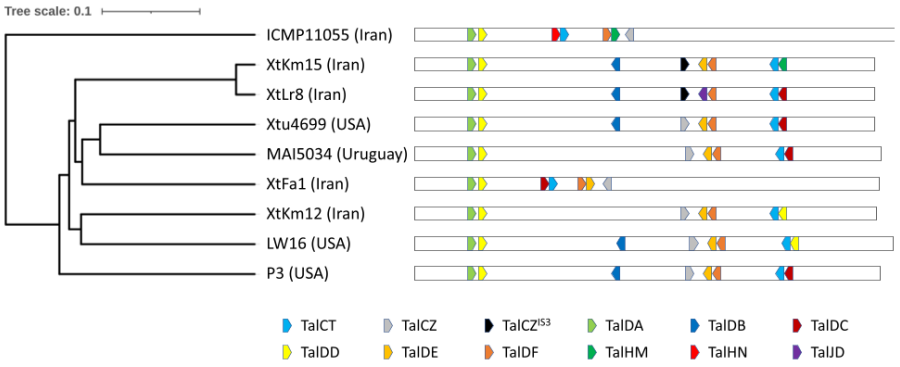


Fig. 1. Phylogenetic tree of *Xanthomonas translucens* pv. *undulosa* strains and repertoire of transcription activator-like (TAL) effector genes. Genome sequences were retrieved from GenBank (Sayers et al. 2022): ICMP 11055 (CP009750), XtKm15 (CP063997, CP063998, CP063999), XtLr8 (CP063993, CP063994, CP063995), Xtu4699 (CP008714), MAI5034 (CP089584), XtFa1 (CP063996), XtKm12 (CP064000), LW16 (CP043540), and P3 (CP043500). Calculation of genome-wide pairwise average nucleotide identities and phylogenetic analysis were performed on the enve-omics platform (http://enve-omics.ce.gatech.edu) (Rodriguez-R and Konstantinidis 2016). The unweighted pair-group method with arithmetic mean was used to build the phylogenetic tree (Sokal and Michener 1958). The interactive Tree Of Life suite was used for better visualization of the tree (https://itol.embl.de; Letunic and Bork 2019). Open rectangles symbolize the chromosomes of the *X. translucens* strains. For better comparability, genome sequences were rotated so that they start with the translation initiation codon of the *dnaA* gene. Position and orientation of TAL effector genes, as classified by the AnnoTale suite (Grau et al. 2016), are indicated by colored arrows. TAL effector genes are not drawn to scale. Genes encoding TalCZ in strains XtKm15 and XtLr8 are interrupted by the insertion of an IS element.

Type 3 effectors and, in particular, TALEs, are of specific interest when trying to understand the pathogenicity of xanthomonads and their host adaptation (Boch and Bonas 2010; Jacques et al. 2016; White et al. 2009). For this reason, TALEs were predicted and classified using the AnnoTale suite (Grau et al. 2016). In total, seven TALE genes were found, belonging to the classes TalCT, TalCZ, TalDA, TalDC, TalDD, TalDE, and TalDF, all of which are widely conserved in strains of *X. translucens* pv. *undulosa* (Table 1; Fig. 1) (Falahi Charkhabi et al. 2017; Peng et al. 2016; Shah et al. 2021), arguing for a recent worldwide expansion of this pathovar (Khojasteh et al. 2019). Notably, a homolog of the *tal8* gene in strain Xtu4699, which elevates expression of a 9-cis-epoxycarotenoid dioxygenase gene (*TaNCED-5BS*) in wheat, that encodes the rate-limiting step in the biosynthesis of the phytohormone abscisic acid for disease susceptibility (Peng et al. 2019), is present in strain MAI5034 (*talDC*). Two additional genes that contribute to virulence of strain ICMP 11055 are less conserved (Falahi Charkhabi et al. 2017). Whereas *tal4b* (*talHM*) is absent in strain MAI5034 and other strains of *X. translucens* pv. *undulosa* (Peng et al. 2016; Shah et al. 2021), strain MAI5034 encodes a novel allele of *tal2*, *talDD*.

The lower diversity of TALE repertoires in the pathovar *undulosa* as compared with the pathovar *translucens* is also mirrored by their lower genetic diversity, which was observed in a cohort of 178 strains of small-grain-infecting xanthomonads (Khojasteh et al. 2019) and may support the hypothesis that the pathovar *undulosa* emerged or originated from the pathovar *translucens*. Such a scenario is also supported by the observation that the *cbsA* gene, a genetic switch between vascular and nonvascular plant pathogenesis (Gluck-Thaler et al. 2020), is disrupted by an IS*1595*-family transposase in strain MAI5034 and other strains of *X. translucens* pv. *undulosa*, ultimately resulting in nonsystemic infection.

Data Availability

Raw reads and the complete genome were uploaded to the NCBI Sequence Read Archive and GenBank under BioProject accession PRJNA786744.

Acknowledgments

We thank J. Grau, Martin-Luther-Universität Halle-Wittenberg, Germany, for assistance with the AnnoTale suite.

Literature Cited

Boch, J., and Bonas, U. 2010. Xanthomonas AvrBs3 family-type III effectors: Discovery and function. Annu. Rev. Phytopathol. 48:419-436.

Clavijo, F., Curland, R. D., Croce, V., Lapaz, M. I., Dill-Macky, R., Pereyra, S., and Siri, M. I. I. 2022. Genetic and phenotypic characterization of Xanthomonas species pathogenic of wheat in Uruguay. Phytopathology 112:511-520.

Duveiller, E., Bragard, C., and Maraite, H. 1997. Bacterial leaf streak and black chaff caused by *Xanthomonas translucens*. Pages 25-32 in: The Bacterial Disease of Wheat: Concept and Methods of Disease Management. E. Duveiller, L. Fucikovskil, and K. Rudolph, eds. CIMMYT, D.F, Mexico.

Egli, T., Goto, M., and Schmidt, D. 1975. Bacterial wilt, a new forage grass disease. J. Phytopathol. 82:111-121.

Falahi Charkhabi, N., Booher, N. J., Peng, Z., Wang, L., Rahimian, H., Shams-Bakhsh, M., Liu, Z., Liu, S., White, F. F., and Bogdanove, A. J. 2017. Complete genome sequencing and targeted mutagenesis reveal virulence contributions of Tal2 and Tal4b of Xanthomonas translucens pv. undulosa ICMP11055 in bacterial leaf streak of wheat. Front. Microbiol. 8:1488.

Gluck-Thaler, E., Cerutti, A., Perez-Quintero, A. L., Butchacas, J., Roman-Reyna, V., Madhavan, V. N., Shantharaj, D., Merfa, M. V., Pesce, C., Jauneau, A., Vancheva, T., Lang, J. M., Allen, C., Verdier, V., Gagnevin, L., Szurek, B., Beckham, G. T., De La Fuente, L., Patel, H. K., Sonti, R. V., Bragard, C., Leach, J. E., Noël, L. D., Slot, J. C., Koebnik, R., and Jacobs, J. M. 2020. Repeated gain and loss of a single gene modulates the evolution of vascular plant pathogen lifestyles. Sci. Adv. 6:eabc4516.

Grau, J., Reschke, M., Erkes, A., Streubel, J., Morgan, R. D., Wilson, G. G., Koebnik, R., and Boch, J. 2016. AnnoTALE: Bioinformatics tools for identification, annotation, and nomenclature of TALEs from Xanthomonas genomic sequences. Sci. Rep. 6:21077.

Huang, Y. T., Liu, P. Y., and Shih, P. W. 2021. Homopolish: A method for the removal of systematic errors in nanopore sequencing by homologous polishing. Genome Biol. 22:95.

Jacques, M. A., Arlat, M., Boulanger, A., Boureau, T., Carrère, S., Cesbron, S., Chen, N. W., Cociancich, S., Darrasse, A., Denancé, N., Fischer-Le Saux, M., Gagnevin, L., Koebnik, R., Lauber, E., Noël, L. D., Pieretti, I., Portier, P., Pruvost, O., Rieux, A., Robène, I., Royer, M., Szurek, B., Verdier, V., and Vernière, C. 2016. Using ecology, physiology, and genomics to understand host specificity in Xanthomonas. Annu. Rev. Phytopathol. 54:163-187.

Jones, L. R., Johnson, A. G., and Reddy, C. S. 1917. Bacterial blight of barley. J. Agric. Res. 11:625-643.

Khojasteh, M., Taghavi, S. M., Khodaygan, P., Hamzehzarghani, H., Chen, G., Bragard, C., Koebnik, R., and Osdaghi, E. 2019. Molecular typing reveals high genetic diversity of Xanthomonas translucens strains infecting small-grain cereals in Iran. Appl. Environ. Microbiol. 85:e01518-19.

Kolmogorov, M., Bickhart, D. M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S. B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T. P. L., and Pevzner, P. A. 2020. metaFlye: Scalable long-read metagenome assembly using repeat graphs. Nat. Methods 17:1103-1110.

Letunic, I., and Bork, P. 2019. Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. Nucleic Acids Res. 47:W256-W259.

Li, H. 2016. Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32:2103-2110.

Lomsadze, A., Gemayel, K., Tang, S., and Borodovsky, M. 2018. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. Genome Res. 28:1079-1089.

Peng, Z., Hu, Y., Xie, J., Potnis, N., Akhunova, A., Jones, J., Liu, Z., White, F. F., and Liu, S. 2016. Long read and single molecule DNA sequencing simplifies genome assembly and TAL effector gene analysis of Xanthomonas translucens. BMC Genomics 17:21.

Peng, Z., Hu, Y., Zhang, J., Huguet-Tapia, J. C., Block, A. K., Park, S., Sapkota, S., Liu, Z., Liu, S., and White, F. F. 2019. Xanthomonas translucens commandeers the host rate-limiting step in ABA biosynthesis for disease susceptibility. Proc. Natl. Acad. Sci. U.S.A. 116:20938-20946.

Rodriguez-R, L. M., and Konstantinidis, K. T. 2016. The enveomics collection: A toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Preprints 4:e1900v1.

Sapkota, S., Mergoum, M., and Liu, Z. 2020. The translucens group of Xanthomonas translucens: Complicated and important pathogens causing bacterial leaf streak on cereals. Mol. Plant Pathol. 21:291-302.

Sayers, E. W., Cavanaugh, M., Clark, K., Pruitt, K. D., Schoch, C. L., Sherry, S. T., and Karsch-Mizrachi, I. 2022. GenBank. Nucleic Acids Res. 50:D161-D164.

Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S., Sedlazeck, F. J., Marschall, T., Mayes, S., Costa, V., Zook, J. M., Liu, K. J., Kilburn, D., Sorensen, M., Munson, K. M., Vollger, M. R., Monlong, J., Garrison, E., Eichler, E. E., Salama, S., Haussler, D., Green, R. E., Akeson, M., Phillippy, A., Miga, K. H., Carnevali, P., Jain, M., and Paten, B. 2020. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Nat. Biotechnol. 38:1044-1053.

Shah, S. M. A., Khojasteh, M., Wang, Q., Taghavi, S. M., Xu, Z., Khodaygan, P., Zou, L., Mohammadikhah, S., Chen, G., and Osdaghi, E. 2021. Genomics-enabled novel insight into the pathovar-specific population structure of the bacterial leaf streak pathogen Xanthomonas translucens in small grain cereals. Front. Microbiol. 12:674952.

Smith, E. F., Jones, L. R., and Reddy, C. S. 1919. The black chaff of wheat. Science 50:48.

Sokal, R. R., and Michener, C. D. 1958. A statistical method for evaluating systematic relationships. Univ. Kans. Sci. Bull. 38:1409-1438.

Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27:737-746.

Vauterin, L., Hoste, B., Kersters, K., and Swings, J. 1995. Reclassification of Xanthomonas. Int. J. Syst. Bacteriol. 45:472-489.

White, F. F., Potnis, N., Jones, J. B., and Koebnik, R. 2009. The type III effectors of Xanthomonas. Mol. Plant Pathol. 10:749-766.

Wick, R. 2017. Porechop. Github. https://github.com/rrwick/Porechop

2.4 Development of a Multi-targeted Real-time PCR Assay for the Detection of Grapevine Pathogen *Xylophilus ampelinus*

Aleksander Benčič, Alexandra Bogožalec Košir, Janja Matičič, Manca Pirc, Neža Turnšek, Tanja Dreo

Plant Methods, 2025, 21(99). doi.org/10.1186/s13007-025-01422-4

This publication describes the genome-informed design of new quantitative real-time PCR (qPCR) tests for the specific and sensitive detection of the bacterial grapevine pathogen *Xylophilus ampelinus*. Genomes from publicly available databases were used to identify sequences that are specific to *X. ampelinus*. Sets of primers and probes were designed to align with and amplify these sequences. The newly designed tests were evaluated in the laboratory for their efficiency, specificity, and selectivity. The performance of the selected tests was then evaluated for the detection of *X. ampelinus* in various grapevine samples, including leaves, roots, and xylem. Three of the newly designed tests were found to reliably and specifically detect *X. ampelinus* in different grapevine samples with high sensitivity, which was also confirmed in an international test performance study.

The PhD candidate is the first author of this publication. He contributed to the planning of experiments, the *in silico* design of qPCR tests, the preparation of samples used in the evaluation, and laboratory testing, which included qPCR, dPCR, and high-throughput qPCR. He also analysed the experimental results. He wrote the manuscript, including the supplementary material, and prepared the figures and tables.

Benčič *et al. Plant Methods* (2025) 21:99
https://doi.org/10.1186/s13007-025-01422-4

Plant Methods

RESEARCH Open Access

Development of a multi-targeted real-time PCR assay for the detection of the grapevine pathogen *Xylophilus ampelinus*

Aleksander Benčič[1,2*], Alexandra Bogožalec Košir[1], Janja Matičič[1], Manca Pirc[1], Neža Turnšek[1] and Tanja Dreo[1]

Abstract

Background *Xylophilus ampelinus* is a plant pathogenic bacterium that causes bacterial blight in grapevines, which can lead to severe yield losses and economic damage. Owing to its fastidious growth on culture media, detection is primarily based on molecular methods. However, existing tests have produced inconsistent results, particularly when used to detect latent infections and non-validated matrices. There is a risk of false-positive results, with economic consequences such as restrictions on international trade. To enhance the diagnostics of *X. ampelinus*, a genome-informed approach was utilised to identify new potential targets for specific detection. On the basis of these sequences, multiple real-time PCR assays were designed, and their specificity and sensitivity were assessed, as well as their performance validated across three different grapevine tissues, including leaves, roots, and xylem.

Results The newly designed real-time PCR assays were evaluated via high throughput testing for specificity and sensitivity and compared with a reference assay. The most promising assays were selected and validated in different grapevine tissues and included in a test performance study to validate their reproducibility and robustness. Three new assays (Xamp_BA_2, TXmp22.4, and Xamp_BA_7) demonstrated high specificity and sensitivity for *X. ampelinus* detection. The Xamp_BA_2 assay exhibited the best overall performance, offering high diagnostic sensitivity and robustness across diverse plant matrices. Importantly, the assays exhibited no cross-reactivity with non-target bacterial species and maintained high detection accuracy across diverse grapevine tissue types.

Conclusions The newly developed real-time PCR assays provide an enhanced diagnostic framework for the detection of *X. ampelinus* in various plant matrices, significantly improving the applicability of molecular testing. The Xamp_BA_2 assay demonstrates superior performance and is recommended for routine diagnostics, with other validated assays being employed for confirmation of identification. The development of these new assays represents a significant expansion of our toolkit for the precise detection of *X. ampelinus* in grapevines, with the potential to contribute to the mitigation of grapevine bacterial blight, the prevention of yield losses, and the protection of international trade in grapevine material. Further implementation of these assays will support regulatory and phytosanitary efforts to mitigate the spread of *X. ampelinus*.

Keywords *Xylophilus ampelinus*, Grapevine bacterial blight, Molecular diagnostics, *Vitis vinifera*, Real-time PCR, Genome-informed assay development

*Correspondence:
Aleksander Benčič
aleksander.bencic@nib.si

[1]Department of Biotechnology and Systems Biology, National Institute of Biology, Večna pot 121, Ljubljana SI-1000, Slovenia
[2]Jožef Stefan International Postgraduate School, Ljubljana SI-1000, Slovenia

Background

Xylophilus ampelinus is a gram-negative betaproteobacterium (order Burkholderiales) that is responsible for the bacterial blight of grapevine (*Vitis vinifera*). Its initial isolation in 1969 in Greece designated it *Xanthomonas ampelina* [1], prior to its subsequent transfer to the newly established genus *Xylophilus* [2]. The only other described member of this genus is *X. rhododendri*, which was isolated from *Rhododendron schlippenbachii*. *X. ampelinus* is widely distributed in the European Mediterranean region, where it has been isolated in Greece, Italy, France, Moldavia, Russia, Slovenia, Spain and Ukraine [3–5]. In addition to Europe, it has also been identified in Jordan [6], South Africa [7] and Japan [8].

X. ampelinus is known to cause systemic infection of the xylem tissue, where the bacterium also overwinters. The bacterium is transmitted both locally, through the use of infected tools and machinery, and over long distances, via infected cuttings used for rooting or grafting. The symptoms include necrotic lesions of the leaves, cracks and cankers along the infected shoots, brown discolouration of the xylem tissue and dead branches [9]. The severity of symptoms exhibited by plants can vary significantly from year to year, contingent on prevailing climatic conditions. Notably, latent infections are prevalent in infected vineyards [3]. The economic impact of *X. ampelinus* is significant, and the disease is listed as a regulated non-quarantine pest (Annex IV, part C of Commission Implementing Regulation (EU) 2019/2072 of 28 November 2019) and is on the A2 list of pests recommended for regulation as quarantine pests by the European and Mediterranean Plant Protection Organisation (EPPO; https://www.eppo.int/ACTIVITIES/plant_quarantine/A2_list; version 2024-09). A significant aspect of the investigations pertains to the international trade in plant material, wherein positive findings can bear substantial economic ramifications, such as the imposition of an import ban.

Owing to the non-specific symptoms of the disease, laboratory tests are required to confirm the infection. Typically, a screening test is performed to detect the pathogen, followed by the isolation of bacteria on culture media if the screening test results are positive or suspicious. However, it should be noted that bacteria may not always be isolated reliably on artificial culture media because of their slow growth rate. Consequently, molecular methods that rely on DNA amplification are predominantly employed for the detection of bacteria in plants. One notable example is the quantitative real-time PCR (real-time PCR) assay developed by Dreo et al. in 2007 [6], which was among the first real-time PCR tests to be incorporated into the EPPO protocol [3]. Other molecular detection methods include PCR [10], which can be combined with an ELISA-based signal amplification system, nested PCR [11] and multiplex PCR [12].

The grapevine samples utilised for detection can be obtained from diverse parts of the plant at varying annual growth stages and can be either symptomatic or asymptomatic. The assays are validated for detection within a limited concentration range in a specific matrix, as plant extracts often contain inhibitors for PCR amplification or may cause cross-reactivity. The use of assays in matrices for which they have not been validated can lead to problematic outcomes. In the case of *X. ampelinus*, contradictory results have been reported from different assays when vine cuttings or root offshoots were analysed for latent infections.

The validation of existing assays on new matrices prior to conducting laboratory tests, as well as the development of new assays capable of detecting *X. ampelinus* in a broader range of matrices, is imperative for reliable detection. The molecular assays currently in use are based on only a few different genetic targets, with real-time PCR by Dreo et al. (2007) and PCR by Manceau et al. (2000) based on the Xamp 1–27 A fragment [6, 13] and nested PCR by Botha et al. (2001) based on the intergenic spacer region of 16–23 S rDNA [11]. This limits the possibility for reliable detection in the event of a false-positive result with one of the currently used tests. The objective of this study was to identify new genetic targets for the specific detection of *X. ampelinus* via a genome-based approach and to utilise these targets for the development of several new real-time PCR assays. These assays were evaluated on the basis of their specificity and sensitivity, with the most promising assays subsequently validated for the detection of *X. ampelinus* in various grapevine samples (leaves, roots, and xylem). To investigate the robustness and applicability of the selected assays, a test performance study (TPS) was conducted in collaboration with other laboratories. The objectives of this study were to validate a series of new real-time PCR assays that should allow reliable detection of *X. ampelinus* in a wide range of grapevine samples and, second, to identify assays suitable for further use in diagnostics (Fig. 1).

Methods

In silico analysis and assay design

The genomes of bacteria of the genus *Xylophilus* used for the in silico analysis were obtained from the publicly accessible GenBank database (Fig. 1: A, Table S1). The average nucleotide identity (ANI) analysis was performed using the online tool enve-omics [14] to calculate the ANI matrix and construct phylogenetic trees using the neighbour-joining method. The program MEGA v11.0.13 was used to visualise the phylogenetic tree. An automated analysis for the determination of unique core sequences (UCS) was performed using a program for the rapid

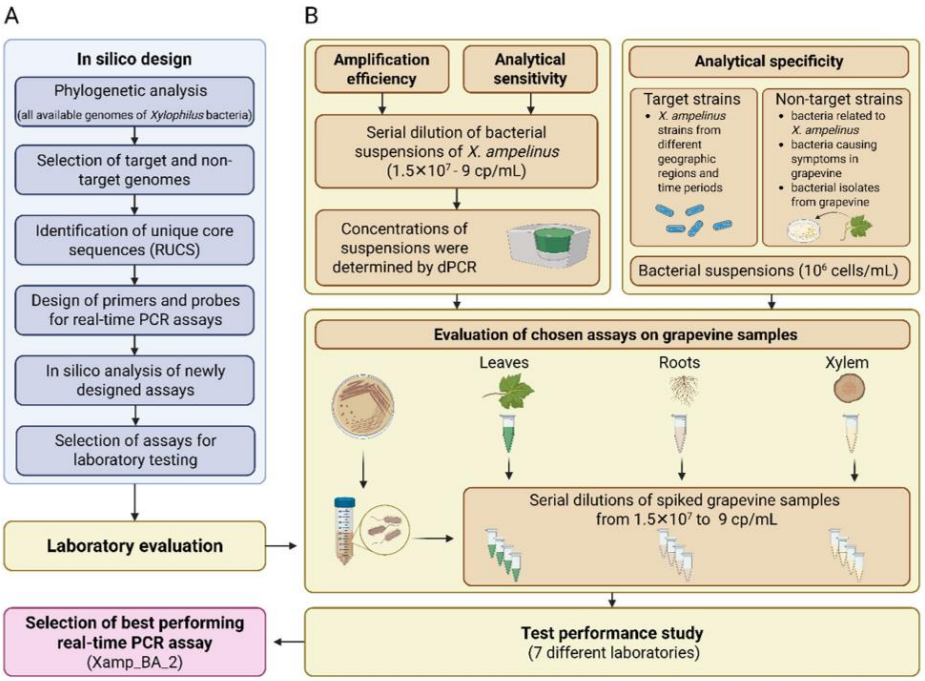


Fig. 1 Schematic overview of the complete real-time PCR design (**A**) and validation (**B**). This involved the in silico design of new assays, which included phylogenetic analysis of all the genomes of the genus *Xylophilus* available at the time. The results of phylogenetic analysis were used to determine target and non-target genomes, which were used in the identification of UCS by RUCS. The primers and probe were designed using Primer Express 2 on UCS. The evaluation of new real-time PCRs included the determination of analytical specificity and sensitivity and amplification efficiency. Five selected assays were validated on grapevine samples from leaves, roots and xylem. Four assays were then used in the TPS, which involved the participation of seven laboratories. The best performing assay was selected on the basis of the TPS and previous in-house validation

identification of PCR primers for unique core sequences: the RUCS [15] web server (https://cge.food.dtu.dk/services/RUCS/). The RUCS programme uses two datasets for the analysis. The first dataset is labelled positive and contains the genomes of all bacteria to be detected. The second dataset is labelled negative and contains the genomes of bacteria that should not be detected. These are usually closely related bacteria that do not belong to the target taxon. Initially, RUCS determines the core genome of the positive genomes, subsequently identifying the sequences present in the core genome but absent in the negative genomes. The application of RUCS was undertaken for the identification of genome sequences specific to *Xylophilus ampelinus*, using a reference genome assembly of the type strain CECT 7646^{T} (GenBank accession number GCA_003217575.1), which was the sole available genome of *X. ampelinus* at the time of analysis. Two separate analyses were performed with two different negative datasets (Table S2). The top ten sequences identified in each RUCS analysis were then used to design the real-time PCR assays. The primers and hydrolysis probes were designed using Primer Express version 2.0 (Applied Biosystems, Thermo Fisher Scientific, Waltham, USA). BLASTN [16] was performed using standard parameters to determine the position of the newly designed assays in the *X. ampelinus* CECT 7646^{T} genome and to assign them to the corresponding genes. To identify similar sequences that might cause cross-reactivity, BLASTN was performed using the Core Nucleotide BLAST database. The calculation of amplicon length, melting temperature, guanine and cytosine percentages, and secondary structures was conducted using the web tool OligoAnalyzer (idtdna.com/calc/analyzer; Integrated DNA Technologies).

Preparation of testing material

Bacterial strains

The bacterial strains used in the study are listed in Table 1. All bacteria were cultivated on nutrient agar media (Bacto Nutrient Agar; Difco), with the exception of *X. fastidiosa*, which was cultivated on BCYE media [17]. The isolates of *X. ampelinus* were cultivated at 25 °C, while the other bacterial strains were cultivated at 28 °C. The new bacterial isolates employed in the study (labelled I in Table 1) were isolated from extracts of grapevine leaves and roots. The isolates were identified by MALDI-TOF mass spectrometry (smartfleX, Bruker, Billerica, US) and 16 S rRNA barcoding using the primers described in EPPO Standard PM 7/129 [2, 18].

Plant material

Asymptomatic leaves, roots and xylem of the grapevine (*Vitis vinifera*) were used for the preparation of the plant extracts (Fig. 1: B). The vine leaves were collected by the Administration for Food Safety, Veterinary Sector and Plant Protection as part of the regular annual monitoring programme coordinated by the Phytosanitary Administration of the Republic of Slovenia in 2023. The plant material used in this study was procured from Trsničarska zadruga Vrhpolje (Vrhopolje, Slovenia). The leaf extracts were prepared by cutting selected leaves (1 g fresh weight per 5 mL buffer) and placing them in sterile phosphate buffer (0.01 M PBS; 1.08 g Na_2HPO_4, 0.4 g $NaH_2PO_4 \times 2H_2O$, 8 g NaCl, 1 L distilled water, pH 7.2) with 10% glycerol (v/v), shaken and incubated for 15 min at room temperature. After incubation, the samples were

Table 1 List of bacterial strains used in validation. The table contains target strains that belong to *X. ampelinus* and non-target strains, which include bacteria related to *X. ampelinus*, bacteria that cause symptoms similar to those of *X. ampelinus*, and bacteria isolated from grapevine. Comments: P-grapevine pathogen, T-type strain, I-strain isolated from grapevine, R-related bacterium

Type	Bacterial species	Strain	Isolated	Comments
Target	*Xylophilus ampelinus*	LMG 5856^T	Greece (1966)	P, T
		LMG 518	Greece (1966)	P
		LMG 527	France (1969)	P
		LMG 520	South Africa (1972)	P
		LMG 510	Spain (1982)	P
		CFBP 2875	France (1988)	P
		CFBP 4864	South Africa (1994)	P
		CFBP 4717	France (1998)	P
		CFBP 5787	France (2001)	P
		NIB Z 3131	Slovenia (2002)	P
		NIB Z 3147	Slovenia (2004)	P
		NIB Z 2746	Slovenia (2018)	P
Non-target	*Xylophilus rhododendri*	KACC 21,265^T	Korea (2019)	R, T
	Agrobacterium tumefaciens	NIB Z 3648	Slovenia (2023)	I
	Agrobacterium vitis	Av 13 – 2	Slovenia (2002)	P
	Ancylobacter novellus	NIB Z 3643	Iran (2017)	I
	Bacillus sp.	NIB Z 3649	Slovenia (2023)	I
	Brevundimonas nasdae	NIB Z 3642	Iran (2017)	I
	Cellulomonas sp.	NIB Z 3645	Iran (2017)	I
	Enterobacter sp.	NIB Z 3656	Slovenia (2023)	I
	Enterobacteriaceae	NIB Z 3654	Slovenia (2023)	I
	Erwinia sp.	NIB Z 3653	Slovenia (2023)	I
	Pantoea agglomerans	NIB Z 2034	Slovenia (2014)	I
	Pantoea agglomerans	NIB Z 3644	Iran (2017)	I
	Pantoea agglomerans	NIB Z 3646	Slovenia (2023)	I
	Pseudomonas syringae	CFBP 2896	Japan (1951)	P
	Pseudomonas chlororaphis	NIB Z 3647	Slovenia (2023)	I
	Pseudomonas graminis	NIB Z 2045	Slovenia (2014)	I
	Rahnella aquatilis	NIB Z 3651	Slovenia (2023)	I
	Rouxiella badensis	NIB Z 3652	Slovenia (2023)	I
	Serratia liquefaciens	NIB Z 3655	Slovenia (2023)	I
	Sodalis sp.	NIB Z 3650	Slovenia (2023)	I
	Xylella fastidiosa	CFBP 8075	USA (1997)	P
	Xylella fastidiosa	CFBP 8351	USA (1993)	P

filtered through a sterile cloth, and the filtrate was stored at -20 °C until further analysis. To produce the root extracts, the grafted vines were grown in sterile water for 3 weeks before the root offshoots were sampled. To prepare the xylem extracts, the bark was first removed from the grafted vines, and then the xylem tissue was cut into small pieces. After the sampling process, the xylem and root extracts were prepared in accordance with a uniform protocol. 1 g of plant tissue was placed in an extraction bag and macerated with a Homex 7 homogeniser (Bioreba AG, Reinach, Switzerland). Next, 5 mL of 0.01 M PBS with 10% glycerol (v/v) was added, and the mixture was incubated for 15 min. The supernatant was then separated from the plant tissue by pipetting and stored at -20 °C until further testing.

Preparation of the samples

The bacterial suspensions were prepared in sterile 0.01 M PBS with 10% glycerol (v/v). The concentrations of the bacterial suspensions were determined turbidimetrically using a densitometer (DEN-1B; BioSan, Riga, Latvia). The bacterial suspensions of the strains listed in Table 1, which were used to determine the analytical specificity, were diluted to a concentration of 10^6 cells/mL. To determine the analytical sensitivity, 10-fold serial dilutions ranging from 10^8 to 10 cells/mL of a bacterial suspension of *X. ampelinus* (strain LMG 5856^T) in 0.01 M PBS with 10% glycerol (v/v) were prepared. The effects of the plant extracts on sensitivity were evaluated, and the diagnostic sensitivity of the selected real-time PCR assays was determined. Plant extracts from leaves, roots and xylem were spiked with 10-fold serial dilutions of a bacterial suspension of *X. ampelinus* (strain LMG 5856^T). The target concentrations of *X. ampelinus* in the spiked samples were 10^8, 10^7, 10^6, 10^5, 5×10^4, 10^4, 5×10^3, 10^3, 10^2 and 10 cells/mL. The concentrations of *X. ampelinus* in the serial dilution samples and the spiked samples were additionally determined by dPCR (QX200, Bio-Rad, Pleasanton, USA) using primers and probes from Dreo et al. (2007) [9] (Fig. 1: B, Table S3).

DNA extraction

DNA was extracted from 100 µL aliquots of bacterial suspensions, plant extracts and spiked plant extracts via the magnetic bead-based QuickPick SML Plant DNA Kit (Bio-Nobile, Turku, Finland) on an automated KingFisher mL system (Thermo Fisher Scientific). Molecular grade water (100 µL) was used as a negative control for DNA extraction. Extraction was performed as previously described by Pirc et al. (2009) [19] with slight modifications (415 µL of lysate for purification) [20]. The DNA was stored at -20 °C until analysis.

Digital PCR

The DNA of *X. ampelinus* extracted from 10-fold serial dilutions of bacterial suspensions in PBS and samples used in the TPS was quantified using the QX200™ Droplet Digital™ PCR system (Bio-Rad) and primers and probe from Dreo et al. (2007). PCR reactions consisted of 10 µL of ddPCR Supermix for probes (Bio-Rad) and 4 µL of sample DNA, with the same primer and probe concentrations used for real-time PCR. After droplet generation, 40 µL of the generated droplet emulsion was transferred to a 96-well PCR plate and amplified in a thermal cycler. The amplification conditions were 10 min of DNA polymerase activation at 95 °C, followed by 40 cycles of a two-step thermal profile of 30 s at 94 °C for denaturation and 60 s at 60 °C for annealing and extension, followed by a final hold of 10 min at 98 °C for droplet stabilisation and cooling to 4 °C. The temperature ramp was set to 2.5 °C/s, and the lid was heated to 105 °C. After the thermal cycle, the plates were transferred to a droplet reader. The software package supplied with the dPCR system (QuantaSoft 1.7.4.0917, Bio-Rad) was used for data acquisition.

High-throughput real-time PCR

High-throughput real-time PCR was performed to determine the analytical specificity and sensitivity of the newly designed real-time PCR assays. This was done using a microfluidics-based platform (Standard BioTools - formerly known as Fluidigm, South San Francisco, USA) and a 48.48 Dynamic Array IFC (Integrated Fluidic Circuit) chip (Standard BioTools). Both the assays and the samples were run in two replicates, resulting in four technical replicates. The analysis included DNA extracted from bacterial suspensions (10^6 cells/mL) of the strains in Table 1 and DNA extracted from the samples in Table S3 at concentrations of 2.8×10^4, 2.8×10^5, 2.4×10^6 and 1.5×10^7 cp/mL. The assay mixtures were prepared with the following components per sample: 3 µL 2× Assay Loading Reagent (Standard BioTools, PN 85000736) and 3 µL 20× mixture of primers and probe. The sample premix was prepared with the following components: 3 µL of TaqMan Universal PCR Master Mix (Applied Biosystems), 0.3 µL of 20× GE Sample Loading Reagent (Standard BioTools) and 2.7 µL of bacterial DNA. The 48.48DA was primed with the IFC Controller MX (Standard BioTools) prior to loading the samples and assays. 5 µL of the sample mixture and 5 µL of the assay mixture were pipetted into the inlets on the chip. The chip was then inserted into the IFC controller, where it undertook the loading and mixing of the samples and assays. After loading and mixing, thermal cycling was performed using the BioMark HD (Standard BioTools) under the following conditions: 2 min at 50 °C and 10 min at 95 °C, followed by 45 cycles of 15 s at 95 °C and 60 s at 60 °C. Two non-template controls were included to control for

non-specific amplification and sample contamination. The data collected with the BioMark HD system were analysed with Fluidigm Real-Time PCR Analysis Software version 3.1.3 (Standard BioTools).

Quantitative real-time PCR

Real-time PCR was performed using newly designed real-time PCR assays and the reference assay from Dreo et al. (2007) (Table S4). All reactions were performed in three technical replicates on a ViiA 7 real-time PCR system (Applied Biosystems) under the same universal cycling conditions: 2 min at 50 °C and 10 min at 95 °C, followed by 45 cycles of 15 s at 95 °C and 60 s at 60 °C in standard temperature ramp mode (1.6 °C/s). The reaction volumes of 10 µL contained the following final concentrations: 900 nM primers (Integrated DNA Technologies, Inc.), 250 nM 6-carboxyfluorescein (FAM) and Black Hole Quencher (BHQ)-1 labelled probe (Integrated DNA Technologies, Inc.), TaqMan Universal PCR Master Mix (Applied Biosystems) and 2 µL of DNA sample. QuantStudio software v1.6 (Applied Biosystems) was utilised for the acquisition of fluorescence and the calculation of the quantification cycle (C_q) value. The baseline was set automatically, and the fluorescence threshold was manually adjusted to overlap with the linear portion of the amplification curves of all real-time PCR assays, which yielded the final C_q value for each well.

Performance characteristics of real-time PCR assays

Analytical specificity was assessed using DNA extracted from bacterial suspensions (10^6 cells/mL) of target and non-target bacterial strains (Table 1). Analytical specificity was also determined using 10 µL real-time PCR reactions, with the results interpreted as positive if at least one replicate was positive. The analytical sensitivity of the designed real-time PCR assays was determined on the basis of DNA extracted from 10-fold serial dilutions of a bacterial suspension of *X. ampelinus* (strain LMG 5856^T) in 0.01 M PBS with 10% glycerol (v/v). All samples and controls were tested in three technical replicates using the ViiA 7 real-time PCR system (Applied Biosystems). The effects of the plant matrices, leaves, roots and xylem on amplification efficiency and sensitivity were analysed via five new real-time PCR assays and the reference assay Dreo et al. (2007). The diagnostic sensitivity was evaluated on DNA extracted from samples of plant extracts of leaves, roots and xylem spiked with 10-fold serial dilutions of a bacterial suspension of *X. ampelinus* (strain LMG 5856^T). The assigned concentrations of *X. ampelinus* determined by dPCR in the spiked samples were 1.5×10^7, 2.4×10^6, 2.8×10^5, 2.8×10^4, 1.8×10^4, 6.2×10^3, 2.3×10^3, 3.7×10^2, 71 and 9 cp/mL (Table S3). To determine the cross-reactivity of the DNA extracted from the plant extracts, real-time PCR assays were used. Molecular grade water was used as a no-template control, whereas DNA extracted from *X. ampelinus* (strain LMG 5856^T) was used as a positive control.

Test performance study (TPS)

The performance of the assays selected in a previous validation (Dreo et al. (2007), TXmp22.4, Xamp_BA_2, Xamp_BA_6 and Xamp_BA_7) was evaluated in a TPS. The TPS sample panel consisted of 30 test items, including positive samples containing DNA extracted from a pure culture of *X. ampelinus* (strain LMG 5856^T); positive samples containing plant extracts from leaves, roots and xylem spiked with *X. ampelinus*; negative samples containing DNA from other non-target bacteria; and negative samples containing plant extracts without *X. ampelinus*, along with the corresponding control samples (Table S5). The quantitative reference values (target concentrations) were determined turbidimetrically and assigned to the concentrations by dPCR Dreo et al. (2007) for the test samples and controls. The stability of the test samples and controls was tested by real-time PCR by Dreo et al. (2007). The stability tests were performed under two conditions: first, at 25 °C for a period of one week, and second, at the standard storage temperature of -20 °C for the duration of the TPS. Each participant analysed all the samples in three separate reactions. A sample was designated as positive if it elicited at least one positive reaction. The collected TPS results were then analysed on the basis of their agreement with the assigned reference values (Table S5). Test records were considered valid if the results of the controls were concordant and if the overall percentage of non-concordant results did not exceed 15%. Participating laboratories reported which instruments, reagents and protocols were used to process the samples. The data for each participating laboratory were analysed on the basis of the number of positive matches (positive versus expected positive), negative matches (negative versus expected negative), positive deviations (positive versus expected negative) and negative deviations (negative versus expected positive).

Statistical data analysis

Using the programming language R for statistical computing [21] and the integrated development environment RStudio [22] in combination with the *drc* package [23], non-linear modelling of the target detection probability was calculated and the target concentration detected with 95% probability (LOD95) was determined from the real-time PCR results of the samples used to determine analytical and diagnostic sensitivity. The dichotomous positive and negative results of the real-time PCR assays were analysed via a 2-by-2 contingency table. Kruskal–Wallis tests were performed using the *vegan* package [24]

to determine whether there was a statistically significant difference between the datasets analysed.

Results

Phylogenetic analysis

Prior to the in silico design of the primers, a phylogenetic analysis was performed to decide which genomes should be included in the positive and negative datasets (Fig. 1: A). All published genomes assigned to the genus *Xylophilus* were included in the calculation of the ANI matrix and the construction of the phylogenetic tree (Fig. 2, Table S1). The analysis revealed that only the genome of the type strain CECT 7646^T belonged to the species *X. ampelinus*. Strains BgEED09 and CCH5-B3, which were assigned to *X. ampelinus* at the time of analysis, did not belong to this species, as their ANI with the genome of the type strain was 80% for both strains. These two genomes were therefore excluded from further assay design. The strain most closely related to *X. ampelinus* was identified as Leaf220, with an ANI of 93%. The remaining genomes analysed presented ANIs ranging from 78 to 83%, indicating their affiliation with the genus *Xylophilus* (Fig. 2, Figure S1).

Identification of unique core sequences and design of real-time PCR assays

New real-time PCR assays were designed to amplify segments of unique core sequences (UCS) that are present in *X. ampelinus* and absent in other related bacteria. At the time of the study, there were thirteen published genomes of the genus *Xylophilus*, only one of which belonged to the species *X. ampelinus*. A further challenge was the fact that ten of these genomes were derived from metagenomic data, which generally had lower coverage and quality (Table S1). For this reason, two separate RUCS analyses were performed. The first included only two genomes of high quality, the genome of *X. ampelinus* (CECT 7646^T) as a positive dataset and the genome of *X. rhododendri* as a negative dataset. The UCSs identified in this analysis were then used for the design of eight real-time PCR assays (TXmp22.1-7, TXmp22.10). In contrast to the first analysis, the second analysis incorporated all available genomes with the exception of BgEED09 and

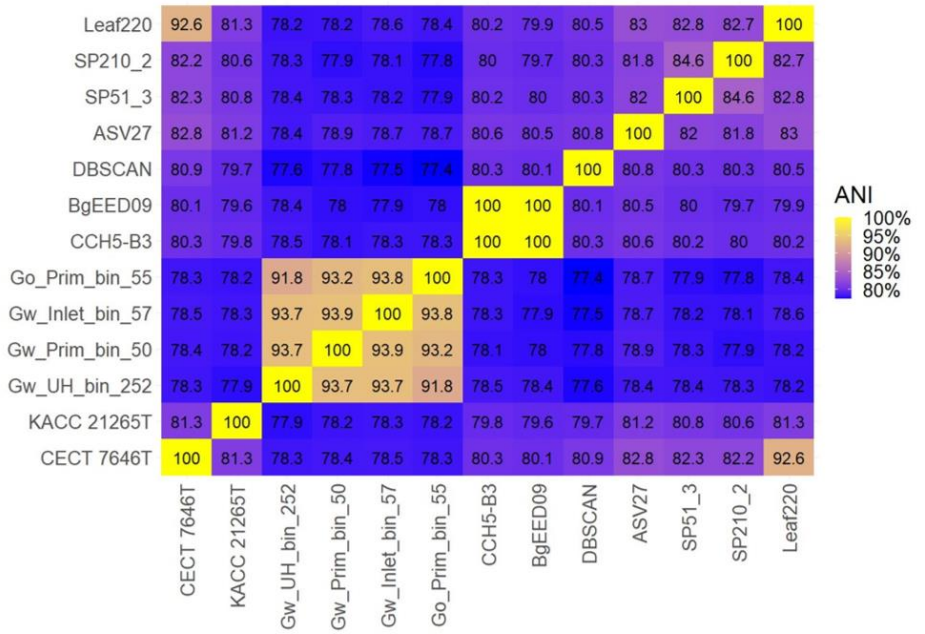


Fig. 2 Heatmap of the average nucleotide identity values matrix in percentages for he genomes of the genus *Xylophilus* available in the GenBank database at the time of this study and listed in Table S1. CECT7646^T is a genome from a type strain of *X. ampelinus*, and KACC21265^T is a genome from a type strain of *X. rhododendri*. Other genomes belong to different *Xylophilus* strains that have not yet been assigned to a specific species

CCH5-B3, with in the *X. ampelinus* genome (CECT 7646^T) again being the sole genome in the positive dataset, whereas all other genomes were allocated to the negative dataset. The UCSs identified in this analysis were then used to develop 10 real-time PCR assays (Xamp_BA_1–10) (Tables S2, S4). Each set of primers and probe was designed for different genomic sequences using Primer Express 2 and TaqMan hydrolysis probes. The melting temperatures of the primers ranged from 63 to 68 °C, whereas those of the probes ranged from 68 to 72 °C. The GC content of the primers varied from 31 to 73%, and that of the probes ranged from 50 to 67%. The amplicons ranged in size from 58 to 108 base pairs (bp), with melting temperatures ranging from 75 to 85 °C (Table S4). BLASTN analyses revealed that the new real-time PCR assays amplified non-coding sequences as well as sequences coding for known or hypothetical proteins (Table S4). It was also shown that four assays (TXmp22.1, TXmp22.2, TXmp22.6 and Xamp_BA_3) amplified prophage sequences and were therefore excluded from further analysis.

Performance and selection of real-time PCR assays

Analytical specificity

To determine the analytical specificity, the newly designed real-time PCR assays were tested using the target and non-target bacterial strains listed in Table 1. The target bacteria included various *X. ampelinus* strains isolated from different geographical regions and time periods, whereas the non-target bacteria included the related bacterium *X. rhododendri*, bacteria that cause similar symptoms in grapevines, and bacterial isolates from grapevine leaves and roots. The analytical specificity was determined using DNA extracted from a bacterial suspension (concentration of 10^6 cells/mL) (Fig. 1: B). A microfluidics-based high-throughput real-time PCR system was used to test multiple assays simultaneously. Eight newly designed real-time PCR assays were able to detect all *X. ampelinus* strains used in the study, while the others showed significantly lower specificity. No instances of false positive results were identified (Tables S6, S7). The analytical specificity determined by high-throughput real-time PCR was confirmed using 10 µL real-time PCR reactions. This confirmed the ability of eight new assays to detect *X. ampelinus* with high specificity, inclusivity and accuracy (Fig. 3, Tables S6, S7).

Analytical sensitivity

To determine the analytical sensitivity, linear range and amplification efficiency of the newly designed assays, 10-fold dilutions (10^8 – 10 cells/mL) of bacterial suspensions of the *X. ampelinus* type strain were used. The copy number of target DNA extracted from these suspensions was determined by dPCR and adjusted for losses during DNA extraction, resulting in slightly lower concentrations ranging from 1.5×10^7 – 9 cp/mL (Fig. 1: B, Table S3). For assays utilising high-throughput real-time PCR, the lowest concentrations that still yielded positive results ranged from 2.8×10^4 to 2.8×10^5 cp/mL for distinct assays (Table S8, Figure S2). For assays in 10 µL real-time PCR reactions, the lowest concentrations that yielded positive results ranged from 71 to 6.2×10^3 cp/mL (Table S9). The linear range for the majority of the assays was observed to be between 3.7×10^2 and a minimum of 1.5×10^7 cp/mL (Table 2). The LOD95, defined as the lowest target concentration that led to positive results in 95% of the reactions, was between 1.2×10^2 cp/mL (BA_Xamp_2) and 1.8×10^3 (TXmp22.4 and Xamp_BA_7) (Table 2). The amplification efficiencies ranged from 0.94 (TXmp22.3) to 1.07 (Xamp_BA_5) and were within the acceptable range for all the assays tested (Table 2). The real-time PCR assays demonstrated high sensitivity, a wide linear range, and good amplification efficiency, confirming their ability to detect *X. ampelinus* over a wide concentration range.

Testing of selected assays for the effects of different matrices

To determine which assays are suitable for testing different grapevine samples, five newly designed real-time PCR assays (TXmp22.4, TXmp22.5, BA_Xamp_2, BA_Xamp_6 and BA_Xamp_7) were selected for further evaluation (Fig. 1: B, Table 3). The selection of these assays was based on the results of analytical specificity and the approach used to identify the target sequences. Initially, assays with an inclusivity of less than 100% were excluded on the basis that they were unable to detect all strains of *X. ampelinus*. Eight assays remained, of which five were selected, taking care to include assays designed on the basis of the two different negative datasets for the RUCS analyses. The selected assays were additionally evaluated using DNA from grapevine extracts of asymptomatic leaves, roots and xylem spiked with a bacterial suspension of the *X. ampelinus* type strain (LMG 5856^T) at different concentrations (Fig. 1: B, Table S3). Grapevine extracts not spiked with bacteria were also tested, and no cross-reactivity with the extracts or the corresponding microflora was observed. A reference assay Dreo et al. (2007), which had not previously been validated for use on matrices other than leaves, was also included in additional tests. The results of the new assays and the reference assay were then compared to determine any differences between the assays and the possible effects of the plant matrices on the performance of the assays. Despite the presence of some differences, the matrices had no significant detrimental effects on amplification efficiency, diagnostic sensitivity, analytical sensitivity and overall performance of the real-time PCR assays. The exception was TXmp22.5, which demonstrated inhibition

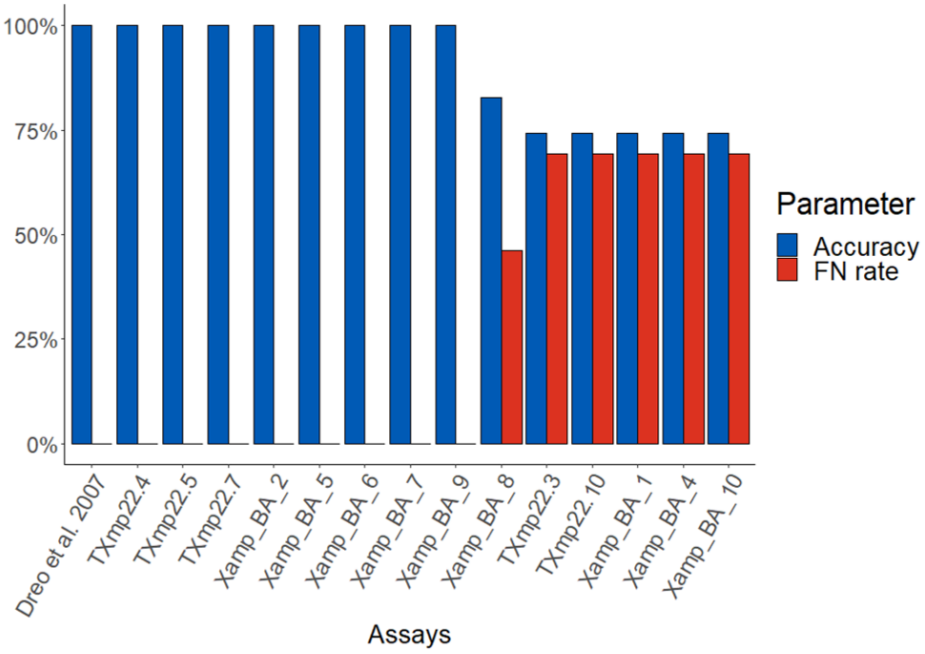


Fig. 3 Bar plot showing the accuracy (dark green bars) and false negative (FN) rate (dark red bars) expressed as percentages of different real-time PCR assays for which analytical specificity parameters were determined during validation

Table 2 Performance characteristics of real-time PCR assays with DNA extracted from *X. ampelinus* suspensions in PBS

Assays	C_q(min–max)[a]	Linear range (cp/mL)[b]	Linear regression [c]			LOD95 [cp/mL]	C_q (1 cp/rxn)
			k	R^2	E		
Dreo	18.8–36.7	3.7×10^2-1.5×10^7	-3.66	0.996	0.88	1.28×10^2	36.72
TXmp22.2	18.9–37.7	3.7×10^2-1.5×10^7	-3.75	0.997	0.85	1.28×10^2	36.54
TXmp22.4	18.9–36.1	3.7×10^2-1.5×10^7	-3.56	0.992	0.91	1.77×10^3	36.27
TXmp22.5	20.2–38.1	3.7×10^2-1.5×10^7	-3.66	0.995	0.88	7.94×10^2	38.13
TXmp22.7	19.5–37.0	3.7×10^2-1.5×10^7	-3.65	0.997	0.88	1.28×10^2	37.37
TXmp22.10	20.9–34.2	6.2×10^3-1.5×10^7	-3.74	0.997	0.85	9.91×10^2	38.45
Xamp_BA_1	20.3–37.7	3.7×10^2-1.5×10^7	-3.61	0.996	0.89	1.28×10^2	38.09
Xamp_BA_2	20.9–39.1	3.7×10^2-1.5×10^7	-3.67	0.996	0.87	1.15×10^2	38.76
Xamp_BA_4	20.8–37.7	6.2×10^3-1.5×10^7	-4.76	0.998	0.62	9.91×10^2	38.22
Xamp_BA_5	19.2–32.2	6.2×10^3-1.5×10^7	-3.68	0.997	0.87	1.53×10^3	36.48
Xamp_BA_6	22.0–41.0	6.2×10^3-1.5×10^7	-3.64	0.992	0.88	9.91×10^2	40.24
Xamp_BA_7	19.0-36.6	3.7×10^2-1.5×10^7	-3.90	0.997	0.81	1.77×10^3	36.23
Xamp_BA_8	18.0-35.4	3.7×10^2-1.5×10^7	-3.53	0.996	0.92	1.28×10^2	35.66
Xamp_BA_9	18.9–31.6	6.2×10^3-1.5×10^7	-3.59	0.997	0.90	1.53×10^3	36.28
Xamp_BA_10	18.7–35.6	3.7×10^2-1.5×10^7	-3.70	0.997	0.86	1.28×10^2	35.88

[a] Minimal and maximal C_q values that still enable detection

[b] The range of concentrations for which C_q values were in linear relationship with logarithms of concentrations (determined by exploring slope values across sections of C_q values × log number of the cells)

[c] Linear regression of all positive samples in a plot of C_q values against the logarithmic number of *X. ampelinus* cells: k = slope of the linear regression line; R^2 = average square regression coefficient; E = efficiency of amplification

Table 3 List of real-time PCR assays that were selected for further evaluation. For each assay, the target inside the genome, forward primer, reverse primer, probe with modifications, amplicon length in base pairs (bp), melting temperature (Tm), and percentage of guanine and cytosine bases (%GC) are listed

Assay (target gene)	Length [bp]	T_m [°C]	%GC
TXmp22.4 (Non-coding sequence)	65	80.8	60
CCCGCGAAGGCCATTC			
GGAGAGACGGCGGGAAAT			
56-FAM/CATCCCTTCACACGGAGTTCTTCCATGAC/3BHQ_1			
TXmp22.5 (ABC transporter permease)	61	81.6	61
AACCCCCTGCGGAAGTG			
GCGTCGCCGCGAAA			
56-FAM/AGGGCTGAGATGGAATCCTACGCACTG/3BHQ_1			
Xamp_BA_2 (alanine racemase)	64	81.1	61
GGTGGCCTCCCTGTTCAGTA			
CGGCATCCTGTTGCAGAAAT			
56-FAM/TCGATCCCGAGCAATGGCGC/3BHQ_1			
Xamp_BA_6 (Non-coding sequence)	88	79.6	49
CCCTCCAAAACTGTGATTGCTAT			
CACCCGGATTTGAATGGAAA			
56-FAM/AGCTATAAGCCGGCACTCGAAGCCG/3BHQ_1			
Xamp_BA_7 (hypothetical protein)	75	79.2	48
GATTCGGCTTCTATCCTTCACAA			
TTCGGAGTTTCCAGGAGATGA			
56-FAM/ACGGCTTCTGGCAGGTTTCACCC/3BHQ_1			

in root and xylem extracts (Table 4; Fig. 4, Table S10). The amplification efficiencies of the assays ranged from 0.45 (TXmp22.5; xylem) to 0.92 (Xamp_BA_7; PBS). The leaf and root extracts showed on average, the lowest and highest amplification efficiencies, respectively. However, the differences were not significant and did not affect the overall performance of the assays (Table 4; Fig. 4: A). No significant differences in analytical sensitivity were detected between the bacterial suspension samples and the samples of the plant extracts spiked with *X. ampelinus*. The limits of detection at 95% confidence (LOD95) values were determined via nonlinear modelling and ranged from 1.15×10^2 cp/mL (BA_Xamp_2; PBS) to 2.67×10^4 cp/mL (TXmp22.5; xylem). The lowest concentration at which we could reliably detect *X. ampelinus* in all the matrices was 3.7×10^2 cp/mL (Fig. 5). The diagnostic sensitivity of the assays ranged from 73 to 82%, with no statistically significant differences observed between the matrices (Table 4; Fig. 4: B). Overall, the new assays demonstrated high sensitivity and amplification efficiency, with the exception of TXmp22.5, which was excluded from further analysis. In comparison, Xamp_BA_2 and Xamp_BA_7 exhibited slightly superior results in terms of sensitivity across different matrices compared with the other evaluated assays.

Test performance study

The reproducibility and robustness of the newly designed assays and the reference assay were investigated through a test performance study (TPS). In this study, seven participants were tasked with testing samples provided by the study organiser. The TPS sample panel consisted of DNA from *X. ampelinus* or non-target bacteria and grapevine extracts of leaves, roots and xylem with and without the addition of *X. ampelinus* (Table S5).

The analysis incorporated only valid experimental recordings, accounting for 86% of the total results submitted by the participants. The TPS demonstrated high reproducibility for all the assays, with 96.1% of the total results being concordant. All the assays exhibited a diagnostic specificity of 100% for all the matrices, with no false positive results observed. The overall diagnostic sensitivity of the assays included in the TPS was greater than 90%, with Xamp_BA_2 having the highest (97%) and Xamp_BA_6 the lowest (90%) diagnostic sensitivity (Fig. 6, Table S11, S12). The performance of the assays was not affected by the different matrices, except for the leaves, which showed inhibition in all assays. The Xamp_BA_6 assay exhibited the highest level of inhibition in the leaf samples (19.0% of non-concordant results), and it also had the highest percentage of non-concordant results in the xylem samples (7.1%). Considering these findings, the use of the Xamp_BA_6 assay is not recommended, particularly for samples from leaves.

To assess the robustness of the assays, the participating laboratories used either the reagents and equipment recommended by the organiser or their own. They used different combinations of DNA extraction methods, PCR mastermixes and thermal cyclers. When the recommended DNA extraction method (QuickPick SML Plant

Table 4 Performance characteristics of real-time PCRs with DNA extracted from samples of plant extracts spiked with *X. ampelinus* and from suspensions of *X. ampelinus* in PBS

Assays	Matrix	Linear range (cp/mL)[a]	Linear regression[b]			LoD95 (cp/mL)[c]	Diagnostic sensitivity
			k	R^2	E		
Dreo	PBS	3.7×10^2-1.5×10^7	-3.66	0.994	0.88	1.8×10^3	75%
	leaf	6.2×10^3-1.5×10^7	-4.16	0.988	0.74	6.2×10^2	73%
	root	3.7×10^2-1.5×10^7	-3.66	0.993	0.88	6.2×10^2	73%
	xylem	3.7×10^2-1.5×10^7	-3.87	0.992	0.81	1.3×10^3	73%
TXmp22.4	PBS	3.7×10^2-1.5×10^7	-3.56	0.992	0.91	7.9×10^2	88%
	leaf	2.3×10^3-1.5×10^7	-3.91	0.991	0.8	7.1×10^2	73%
	root	3.7×10^2-1.5×10^7	-3.55	0.989	0.91	6.2×10^2	73%
	xylem	2.3×10^3-1.5×10^7	-3.92	0.988	0.8	2.7×10^4	73%
TXmp22.5	PBS	3.7×10^2-1.5×10^7	-3.66	0.995	0.88	1.2×10^2	75%
	leaf	2.3×10^3-1.5×10^7	-3.74	0.992	0.85	6.2×10^2	73%
	root	non-linear	NA	NA	NA	1.4×10^2	73%
	xylem	2.8×10^4-1.5×10^7	-6.23	0.995	0.45	7.1×10^2	45%
Xamp_BA_2	PBS	3.7×10^2-1.5×10^7	-3.67	0.996	0.87	9.9×10^2	88%
	leaf	3.7×10^2-1.5×10^7	-3.75	0.989	0.85	6.2×10^2	73%
	root	3.7×10^2-1.5×10^7	-3.63	0.988	0.89	6.2×10^2	82%
	xylem	2.3×10^3-1.5×10^7	-3.65	0.989	0.88	7.1×10^2	73%
Xamp_BA_6	PBS	6.2×10^3-1.5×10^7	-3.90	0.997	0.81	1.8×10^3	75%
	leaf	3.7×10^2-1.5×10^7	-3.87	0.988	0.81	1.3×10^3	73%
	root	3.7×10^2-1.5×10^7	-3.59	0.981	0.9	6.2×10^2	73%
	xylem	6.2×10^3-1.5×10^7	-4.02	0.993	0.77	2.2×10^3	73%
Xamp_BA_7	PBS	3.7×10^2-1.5×10^7	-3.53	0.996	0.92	1.3×10^2	88%
	leaf	3.7×10^2-1.5×10^7	-3.73	0.987	0.85	6.2×10^2	82%
	root	3.7×10^2-1.5×10^7	-3.54	0.991	0.92	1.3×10^2	73%
	xylem	3.7×10^2-1.5×10^7	-3.77	0.994	0.84	6.2×10^2	82%

[a]The range of concentrations for which C_q values were in linear relationship with logarithms of concentrations (determined by exploring slope values across sections of C_q values × log number of the cells)

[b]Linear regression of all positive samples in a plot of C_q values against the logarithmic number of *X. ampelinus* cells: k = slope of the linear regression line; R^2 = average square regression coefficient; E = efficiency of amplification

[c]LOD = limit of detection; for the purpose of this study, LOD95 was defined as the concentration at which 95% of the samples were positive

DNA Kit) was used, the results were 100% concordant regardless of other variables (Figure S3). The volume of plant extract used for DNA extraction and PCR mastermix had no effect on the results. The results of the TPS confirmed the high sensitivity of the assays and demonstrated the reproducibility and robustness of the newly designed assays in several laboratories.

Discussion

In this study, an innovative genome-informed approach was utilised to identify novel targets, which were then employed to develop a panel of real-time PCR assays capable of detecting *X. ampelinus* in various plant matrices with high sensitivity and selectivity. The sequences unique to *X. ampelinus* were identified using a bioinformatics pipeline based on RUCS, a freely available program that has already been shown to be useful for the design of specific real-time PCR assays [25]. A total of 18 primer/hydrolysis probe sets were designed, eight of which could detect all *X. ampelinus* strains without generating false positive results. Despite the fact that *X. ampelinus* is regarded as a highly homogeneous species [26], a significant proportion of assays yielded false negative results, suggesting the potential for intraspecies diversity. This could be because only one genome of *X. ampelinus* was used for the design of the primers. Consequently, the core genome could not be determined, and some shell and cloud genes were probably also included in the UCSs, which were used to design new assays. Despite the limited number of genomes utilised in the analysis, the successful design of numerous specific assays underscores the robustness of the RUCS-based pipeline.

To limit the scope of validation, five out of eight promising assays were selected for further evaluation. These were selected to include assays designed on UCS identified in both RUCS analyses. The five selected assays and the reference assay of Dreo et al. (2007) were then subjected to further evaluation on matrices, including leaves, roots and xylem. It was established that all assays demonstrated acceptable amplification efficiency and diagnostic sensitivity across all the matrices, with the exception of

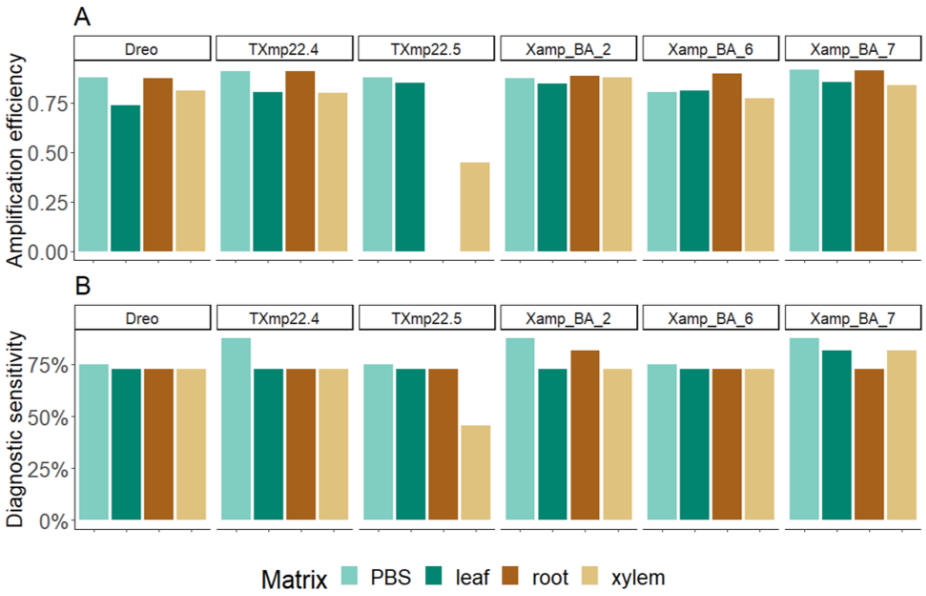


Fig. 4 Comparison of the performance expressed as amplification efficiency (**A**) and diagnostic sensitivity (**B**) of the selected real-time PCR assays TXmp22.4, TXmp22.5, BA_Xamp_2, BA_Xamp_6 and BA_Xamp_7 and the reference Dreo et al. (2007). The DNA used was isolated from samples of plant extracts of leaves, roots and xylem fortified with *X. ampelinus* and from a bacterial suspension of *X. ampelinus* in phosphate-buffered saline (PBS). For TXmp22.5, the amplification efficiency in the root extract could not be calculated because of the lack of a linear range

TXmp22.5. Further evaluation of this assay revealed inhibition in spiked xylem and root samples, resulting in its exclusion from the study. Nevertheless, some differences in assay performance were observed in different matrices, with samples from leaves showing slightly lower amplification efficiency. The remaining assays (TXmp22.4, BA_Xamp_2, BA_Xamp_6 and BA_Xamp_7) were found to be suitable for the detection of *X. ampelinus* in different plant samples and at concentrations as low as 10^2 cp/mL. The validation of the assays for the detection of *X. ampelinus* in different matrices is of great importance, as it enables the analysis of a broad spectrum of samples ranging from green plant parts to woody tissue from grafted plant material. The assays demonstrate a high level of sensitivity, enabling the detection of pathogenic bacteria at low concentrations. This capacity for detection extends to latent infections, a significant consideration given the prolonged incubation periods associated with slow-growing bacteria such as *X. ampelinus*.

The TPS in which new assays were evaluated identified the leaf extract as the most problematic matrix, with the highest percentage of non-concordant results. The Xamp_BA_6 assay demonstrated the lowest diagnostic sensitivity, particularly in the leaf and xylem matrices, and was consequently also excluded. Among the assays evaluated, Xamp_BA_2 exhibited the highest diagnostic sensitivity and is recommended for further diagnosis of *X. ampelinus* infection. The assays TXmp22.4, BA_Xamp_7 and Dreo et al. (2007) also demonstrated high sensitivity and specificity, making them suitable for detection. In instances where a positive result is obtained with Xamp_BA_2, these assays should be employed to ascertain the identity of *X. ampelinus*. To assess the robustness of the novel assays, the participating laboratories were permitted to use their own equipment and reagents to perform TPS. It was found that there were some discrepancies between the participating laboratories. For example, the leaf samples that presented the most challenges in both the initial evaluation and the TPS were found to be 100% concordant when specific DNA extraction methods were employed. This highlights the importance of the DNA extraction method, which should be carefully selected for each matrix in which target bacteria are to be detected. Nevertheless, owing to the limited number of participants, it is challenging to ascertain that these discrepancies are attributable to differing protocols. Importantly,

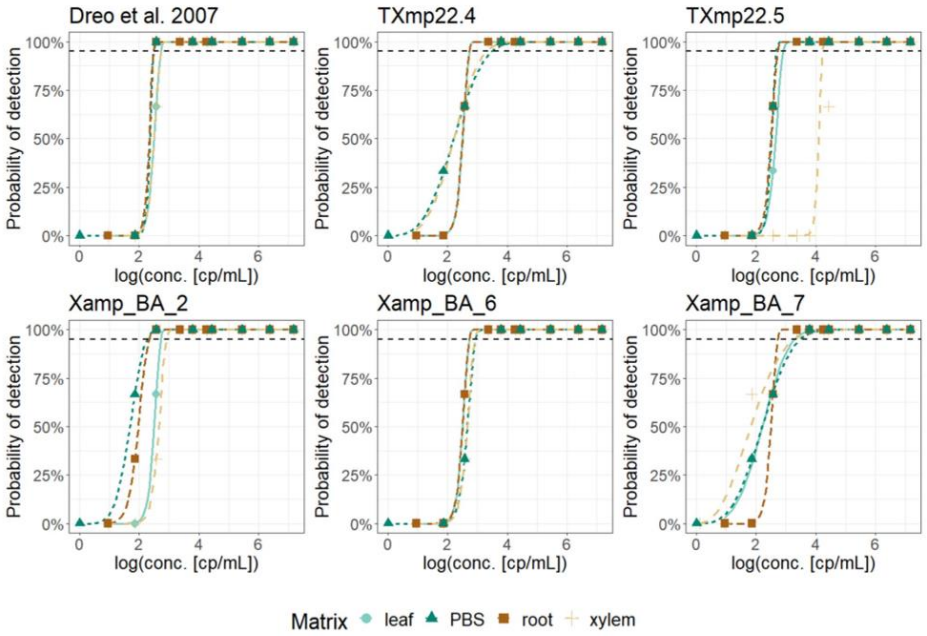


Fig. 5 Plots showing the detection probabilities for selected real-time PCR assays, TXmp22.4, BA_Xamp_2, BA_Xamp_6 and BA_Xamp_7, and reference Dreo et al. (2007), which were calculated using non-linear modelling and the Weibull type 2 model. The dashed lines indicate the 95% probability (LOD95) of detection. Probabilities were calculated for four different matrices (leaf, root, xylem and PBS)

the discrepancies between protocols were small showcasing the high robustness of the three new and reference assays.

In previous studies, molecular tests were found to be more reliable than serological tests and had a lower sensitivity and a higher proportion of false positives [27]. Although other molecular methods have shown results similar to those of real-time PCR, it offers certain advantages, as it is a rapid and straightforward method with high sensitivity that could also be adapted for use as dPCR. The primary benefit of dPCR is its capacity for absolute quantification, a particularly advantageous feature when dealing with complex samples or in scenarios where reference materials are not available, a common occurrence in many research contexts. Compared with other molecular methods, such as nested or conventional PCR, real-time PCR has the advantages of the absence of an electrophoresis step, which reduces the possibility of contamination, and lower sensitivity to inhibitors present in plant tissues.

Conclusions

In this study, five new real-time PCR assays were validated, three of which (TXmp22.4, Xamp_BA_2 and BA_Xamp_7) were found to be suitable for testing different grapevine samples for the presence of *X. ampelinus*. Additionally, the application of the reference assay of Dreo et al. (2007) was expanded to include root and xylem matrices. These assays have been demonstrated to be valuable tools for the detection of *X. ampelinus*, which remains an important pathogen in viticulture, as it can cause yield losses and restrictions in the international trade of grafted plants. Furthermore, these tests are instrumental in preventing the spread of the disease in areas where it has not yet been detected. It is recommended that the new Xamp_BA_2 assay be used in conjunction with other assays for the screening of infected plants to reduce the possibility of false negative results.

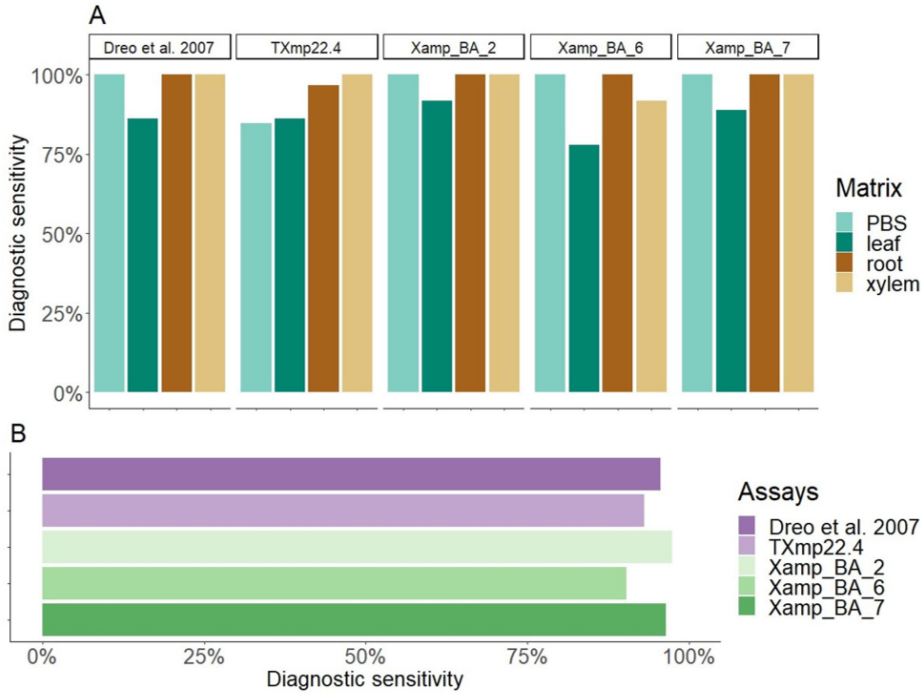


Fig. 6 Results of the test performance study (TPS) showing the diagnostic sensitivity of real-time PCR assays (TXmp22.4, TXmp22.5, BA_Xamp_2, BA_Xamp_6, BA_Xamp_7 and reference from Dreo et al. (2007)) in different matrices (**A**) and over all different matrices

Abbreviations

ANI	Average nucleotide identity
BHQ	Black hole quencher
dPCR	Digital PCR
ELISA	Enzyme-linked immunosorbent assay
EPPO	European and Mediterranean Plant Protection Organization
FAM	6-carboxyfluorescein
LOD95	Limit of detection at 95% confidence
MALDI-TOF	Matrix-assisted laser desorption/ionization time-of-flight
NA	Not applicable
PCR	Polymerase chain reaction
PBS	Phosphate buffered saline
rDNA	Ribosomal DNA
RUCS	Rapid identification of PCR primers for unique core sequences
TPS	Test performance study
UCS	Unique core sequences

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13007-025-01422-4.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

The plant material used in this study was collected by the Administration of the Republic of Slovenia for Food Safety, Veterinary and Plant Protection of the Ministry of Agriculture and Environment and Phytosanitary Inspectorate. A test performance study was performed as part of the Euphresco project 2021-A-383. The authors would like to acknowledge Aleš Blatnik, Vesna Dukić, Sara Fišer, Vladimir Grujić and Špela Prijatelj Novak for their technical support.

Author contributions

A.B. contributed in aquisition and interpreatation of data, wrote the main manuscript text and prepared figures. A.B.K. contributed in aqusition and interpretation of data. J.M. contributed in data aquisition. M.P. contributed in aqusition and interpretation of data. N.T. contributed in data aquisition. T.D. contributed in design of work and interpretation of data. All authors reviewed the manuscript.

Funding

This study was funded by the Slovenian Research and Innovation Agency (contract no. P4-0165 and MR Aleksander Benčič). Part of the dPCR equipment used in this study was financed by the Metrology Institute of the Republic of Slovenia (MIRS), with financial support from the European Regional Development Fund.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 28 February 2025 / Accepted: 12 July 2025
Published online: 18 July 2025

References

1. Panagopoulos C. 1969. The disease Tsilik Marasi of grapevine: its description and identification of the causal agent (Xanthomonas Ampelina sp. nov). Ann Inst Phytopathol Beanki 59–81.
2. Willems A, Gillis M, Kersters K, Broecke L, Van Den, Ley JD. Transfer of Xanthomonas ampelina Panagopoulos 1969 to a new genus, Xylophilus gen. Nov., as Xylophilus ampelinus (Panagopoulos 1969) comb. Nov. Int J Syst Evol Microbiol. 1987;37:422–30.
3. Xylophilus ampelinus. EPPO Bull 39:403–12.
4. Kovaleva I, Konup L, Nikolaeva N, Konup A, Chistyakova V. 2022. Eutiposis and bacterial necrosis of the vine on the vineyards of the Odessa region. Quar Plant Prot 21–5.
5. Agriculture Rministry. of. 2020. National report on the quarantine status on the territory of the Russian Federation in 2019 pp28.
6. Khlaif H, Abu-Obeid I, Werikat B. Afr J Agricultural Res Occurrence Plant Bacterial Dis Jordan. 2018;13:2104–17.
7. Erasmus HD, Matthee FN, Louw HA. A comparison between plant pathogenic species of pseudomonas, Xanthomonas and Erwinia with special reference to the bacterium responsible for bacterial blight of Vines *. Phytophylactica. 1974;6:11–8.
8. Komatsu T, Kondo N. Winter habitat of Xylophilus ampelinus, the cause of bacterial blight of grapevine, in Japan. J Gen Plant Pathol. 2015;81:237–42.
9. Dreo T, Gruden K, Manceau C, Janse JD, Ravnikar M. Development of a real-time PCR-based method for detection of Xylophilus ampelinus. Plant Pathol. 2007;56:9–16.
10. Manceau C, Grall S, Brin C, Guillaumes J. Bacterial extraction from grapevine and detection of Xylophilus ampelinus by a PCR and microwell plate detection system*. EPPO Bull. 2005;35:55–60.
11. Botha WJ, Serfontein S, Greyling MM, Berger DK. Detection of Xylophilus ampelinus in grapevine cuttings using a nested polymerase chain reaction. Plant Pathol. 2001;50:515–26.
12. Carminati G, Bianchi GL, De Amicis F, Cannistraci I, Sadallah A, Ermacora P, Torelli E, Martini M, Firrao G. Improving reliability of PCR diagnostics for Xylophilus ampelinus by metagenome-informed circumscription of the target taxon. Plant Pathol. 2025;74:249–57.
13. Manceau C, Coutaud MG, Guyon R. Assessment of subtractive hybridization to select species and subspecies specific DNA fragments for the identification of Xylophilus ampelinus by polymerase chain reaction (PCR). Eur J Plant Pathol. 2000;106:243–53.
14. Rodriguez -RLM, Konstantinidis KT. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Prepr. 2016;4:e1900v1.
15. Thomsen MCF, Hasman H, Westh H, Kaya H, Lund O. RUCS: rapid identification of PCR primers for unique core sequences. Bioinformatics. 2017;33:3917–21.
16. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol. 2000;7:203–14.
17. Wells JM, Raju BC, Nyland G, Lowe, 2 SK. Medium for isolation and growth of Bacteria associated with Plum leaf scald and phony Peach diseases. Appl Environ Microbiol. 1981;42:357–63.
18. PM 7/129. (2) DNA barcoding as an identification tool for a number of regulated pests. EPPO Bull 51:100–43.
19. Pirc M, Ravnikar M, Tomlinson J, Dreo T. Improved fireblight diagnostics using quantitative real-time PCR detection of Erwinia amylovora chromosomal DNA. Plant Pathol. 2009;58:872–81.
20. Dreo T, Pirc M, Ramšak Ž, Pavšič J, Milavec M, Žel J, Gruden K. Optimising droplet digital PCR analysis approaches for detection and quantification of bacteria: A case study of fire blight and potato brown rot. Anal Bioanal Chem. 2014;406:6513–28.
21. Team RC. R: A Language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.
22. Team Rs, RStudio. RStudio: integrated development environment for R. Boston, MA: PBC; 2021.
23. Ritz C, Baty F, Streibig JC, Gerhard D. Dose-Response Anal Using R PLoS One. 2015;10:e0146021.
24. Jari Oksanen F, Guillaume Blanchet M, Friendly R, Kindt P, Legendre D, McGlinn, Peter R, Minchin RB, O'Hara GL, Simpson P, Solymos M, Henry H. Stevens ES and HW. 2020. vegan: Community Ecology Package. R package version 2.5-7.
25. Ališ Š, Dermastia M, Burger J, Dickinson M, Pietersen G, Pietersen G, Dreo T. Genome-Informed design of a LAMP assay for the specific detection of the strain of 'candidatus Phytoplasma asteris' Phytoplasma occurring in grapevines in South Africa. Plant Dis. 2022;106:2927–39.
26. Portier P, Taghouti G, Bertrand PE, Briand M, Dutrieux C, Lathus A, Fischer-Le Saux M. Analysis of the diversity of Xylophilus ampelinus strains held in CIRM-CFBP reveals a strongly homogenous species. Microorganisms. 2022;10:1531.
27. Harrison C, Tomlinson J, Brittain I, Laurenson L, van den Berg F. Test performance study to evaluate diagnostic tests for the detection and identification of Xylophilus ampelinus in vine material spiked with cells from X. ampelinus cultures. Eur J Plant Pathol. 2024;168:1–13.

Publisher's note

2.5 Irrigation Systems as Reservoirs of Diverse and Pathogenic *Pseudomonas syringae* Strains Endangering Crop Health

Marina Anteljević, Iva Rosić, Olja Medić, Tamara Ranković, Karolina Sunjog, Margareta Kračun-Kolarević, Stoimir Kolarević, Tanja Dreo, Aleksander Benčič, Tanja Berić, Slaviša Stanković, Ivan Nikolić

Water Research X, 2025, 28, 100380. doi.org/10.1016/j.wroa.2025.100380

This publication examines the *Pseudomonas syringae* complex in the Danube-Tisa-Danube Hydrosystem (DTD) in Serbia, including its abundance, phylogenetic diversity, and pathogenic potential. *P. syringae* is a well-known foliar plant pathogen that was also found to inhabit a variety of different environments, including water habitats. Results of the phylogenetic analysis included in the study showed high diversity of *P. syringae* strains isolated in DTD, which included phylogroups 1, 2, 7, 12, and 13. Some of the isolated strains were linked to previous epidemics of sugar beet in Serbia and were proven to be capable of inducing disease. In the scope of research described in this publication, new molecular methods for the detection of *P. syringae* phylogroup 2 were also developed. That could reduce the reliance on time-consuming microbiological methods of isolation on media and phenotypic identification. To that end, new sets of primers and probes for qPCR tests were designed using a genome-informed approach. Newly designed tests were evaluated for their sensitivity and specificity using strains of *P. syringae* and other related bacteria isolated from DTD water. The most promising of the new tests was also adapted for use as dPCR. The results showed that dPCR outperformed both traditional methods of isolation on media as well as qPCR.

The PhD candidate contributed to this work by being involved in the *in silico* design of a new qPCR test, design of the evaluation process, preparation of samples used in the evaluation, laboratory testing and analysis, and interpretation of results.

Water Research X 28 (2025) 100380

Contents lists available at ScienceDirect

Water Research X

journal homepage: www.sciencedirect.com/journal/water-research-x

Irrigation systems as reservoirs of diverse and pathogenic *Pseudomonas syringae* strains endangering crop health

Marina Anteljević [a], Iva Rosić [a], Olja Medić [b,c], Tamara Ranković [b,c], Karolina Sunjog [d], Margareta Kračun-Kolarević [e], Stoimir Kolarević [e], Tanja Dreo [f], Aleksander Benčič [f,g], Tanja Berić [a,b,c], Slaviša Stanković [b,c], Ivan Nikolić [b,c,*]

[a] University of Belgrade, Institute of Physics Belgrade, National Institute of the Republic of Serbia, Laboratory for Experimental Astrobiology, Pregrevica 118, Belgrade 11080, Serbia
[b] University of Belgrade, Faculty of Biology, Department of Microbiology, Studentski trg 16, Belgrade 11000, Serbia
[c] University of Belgrade, Faculty of Biology, Center for Biological Control and Plant Growth Promotion, Studentski trg 16, Belgrade 11000, Serbia
[d] University of Belgrade, Institute for Multidisciplinary Research, Kneza Višeslava 1, Belgrade 11000, Serbia
[e] University of Belgrade, Institute for Biological Research "Siniša Stanković", National Institute of the Republic of Serbia, Department of Hydroecology and Water Protection, Bulevar Despota Stefana 142, Belgrade 11060, Serbia
[f] National Institute of Biology, Department of Biotechnology and Systems Biology, Večna pot 121, Ljubljana 1000, Slovenia
[g] Jožef Stefan International Postgraduate School, Jamova 39, Ljubljana 1000, Slovenia

ARTICLE INFO

Keywords:
Phytopathogen detection
Irrigation
qPCR
dPCR

ABSTRACT

Pseudomonas syringae (*Psy*) is a widely distributed bacterial species complex primarily recognized as a foliar pathogen but also inhabits diverse environments, including water habitats, where strains closely related to agricultural pathogens have been identified. The connection between *Psy*-caused epidemics and its potential presence in nearby irrigation systems remains underexplored. This study comprehensively examined the *Psy* complex in the Danube-Tisa-Danube Hydrosystem (DTD) in Serbia, assessing its abundance, phylogenetic diversity, and pathogenic potential. To reduce the reliance on the time-consuming steps of isolation and identification, we developed novel high-specific primers and probes for precise detection of strains belonging to phylogroup 2 within *Psy* complex. Our results demonstrated that dPCR, coupled with highly specific and sensitive primers, outperformed both traditional plating and qPCR in detecting the *Psy* complex and phylogroup 2 in irrigation waters, making *Psy* diagnostics more effective. Phylogenetic analysis indicated high strain diversity within the DTD, identifying phylogroups 1, 2, 7, 12, and 13 and haplotypes linked to strains previously encountered in epidemics on sugar beet in Serbia. Notably, 66.67% of the isolates from the DTD were capable of inducing disease. Phylogroup 2 isolates displayed a broad host range, suggesting that the dissemination of *Psy* from DTD through irrigation, poses a substantial threat to crop health and agricultural productivity.

Introduction

The condition of irrigation systems plays a vital role in supporting agricultural productivity and shaping the interactions between waterborne microorganisms and crops. Irrigation systems can act as pathways for the dispersal of plant pathogens, with repeated use of contaminated water sources posing significant risks to crop health and yield (Lamichhane and Bartoli, 2015). Within the broader water-food nexus, effective water resource management, including irrigation water quality and pathogen control, is essential for sustaining food production (Corona-Lopez et al., 2021; Salem et al., 2022). In this context, the Danube-Tisa-Danube Hydrosystem (DTD) in the northern part of Serbia, in Vojvodina Province, is important to its region. DTD is an artificial network of canals constructed for irrigation, drainage, and flood management (Blagojević et al., 2020). The total length of the main channel network is 930 km, which covers an area of around 12,000 km^2 (Blagojević et al., 2020). This hydrosystem is characterized by pronounced seasonal dynamics, influencing hydrological conditions such as water level and flow, temperature changes, water quality, biodiversity, and ecological processes. Recent initiatives aim to optimize its use for

* Corresponding author.
E-mail address: ivan.nikolic@bio.bg.ac.rs (I. Nikolić).

https://doi.org/10.1016/j.wroa.2025.100380
Received 22 April 2025; Received in revised form 9 July 2025; Accepted 14 July 2025
Available online 16 July 2025

irrigation across Vojvodina (FAO, 2021), a region renowned for its agricultural productivity and intensive crop cultivation. However, this area has also experienced numerous plant disease outbreaks (Ignjatov et al., 2017; Marković et al., 2022; Mitrović et al., 2015; Popović et al., 2013). Among the bacterial pathogens reported, members of the *Pseudomonas syringae* species complex (*Psy*) were most frequently associated with disease epidemics in crops, fruits, and vegetables in Vojvodina (Balaž et al., 2014; Ilićić et al., 2021; Popović et al., 2015a, 2015b; Stojšin et al., 2015). *Psy* pathogens are recognized globally for their substantial impact on plant health and productivity (Morris et al., 2019).

Although primarily studied in agricultural context, recent research highlights *Psy*'s broader ecological significance, particularly in water systems such as rivers, lakes, groundwaters, and irrigation networks (Berge et al., 2023; Morris et al., 2010, 2023). *Psy* has also been detected in atmospheric environments like rain, snow, and clouds, where its ice nucleation proteins are thought to play an important role in cloud formation and precipitation, thereby influencing the global water cycle (de Araujo et al., 2019). This ability to persist and disperse in both agricultural and non-agricultural environments makes the diversity of strains found in non-agricultural habitats (e.g. fresh waters and irrigation channels) even more significant than those encountered in epidemic outbreaks (Berge et al., 2014; Morris et al., 2010).

Based on Multilocus Sequence Typing analysis, the complex is divided into 13 phylogroups, each representing a distinct evolutionary lineage (Berge et al., 2014). Among the reported epidemics detected in fields near the DTD canals, most were caused by phylogroup 2 (PG02), known for its wide host range, severe symptoms, high dispersal, and broad distribution (Morris et al., 2019, 2022). Given the proximity of numerous epidemics to the DTD, it is critical to investigate the *Psy* population in this canal network. In our previous study, *Psy* strains associated with earlier epidemics in cherry orchards were detected in the Danube River Basin of Serbia, which is linked to the DTD (Anteljević et al., 2023).

Various traditional and modern diagnostic techniques have been employed to detect *Psy*, including culture isolation, microscopy, PCR, and ELISA assay (Yang et al., 2023). Molecular techniques such as qPCR have proven highly effective in detecting and quantifying *Psy* in seeds (Chai et al., 2020), while both qPCR and dPCR (digital PCR) have been successfully employed for plant samples (Barrett-Manako et al., 2021). However, these methods require further improvement to achieve the sensitivity and specificity necessary for reliable detection in water samples (Kokkoris et al., 2021).

Our study aimed to provide a comprehensive assessment of *Psy* occurrence in DTD, an important water source for irrigation and agricultural production in a region where previous epidemics have affected sugar beet and vegetable crops. The main objectives of our study were to: (i) design specific primers and probes for the highly efficient detection of PG02 of the *Psy* complex, (ii) employ and compare traditional culture methods, qPCR, and dPCR assays to determine the abundance of the *Psy* complex and PG02 in the DTD, and (iii) investigate the pathogenic potential of isolated *Psy* strains and assess the risk for future *Psy* epidemics and crop health.

2. Materials and methods

2.1. Sampling sites and collection

Water samples were collected on April 20th and 21st, 2023, from eight locations at the DTD (Fig. 1 and Supplementary Material 1, Table S1). The DTD spans the northern region of the Republic of Serbia, located between the geographical coordinates 44° 51′ and 45° 52′ N latitude and 18° 51′ and 21° 17′ E longitude. The sampling sites were chosen based on prior reports of epidemics in Serbia linked to the *Psy* species complex (Supplementary Material 1, Table S2). Samples were taken approximately 3 m from the embankment at a depth of 30 cm below the water surface. In situ measurements of pH, temperature, conductivity, and dissolved oxygen (DO) were carried out using the portable multi-parameter instrument Multi 340i (WTW, Germany). Following the initial field measurements, water samples were stored at 4 °C, transported to the laboratory in cooling boxes, and processed immediately. Mass concentrations of ammonium (NH_4^+), nitrate (NO_3^-), nitrite (NO_2^-), and phosphate (PO_4^{3-}) were measured using colorimetric kits (FLUKA Analytical, Germany). These analyses were quantified using the WINLAB® DATA LINE PHOTOMETER (Dr. Lange, Germany). The

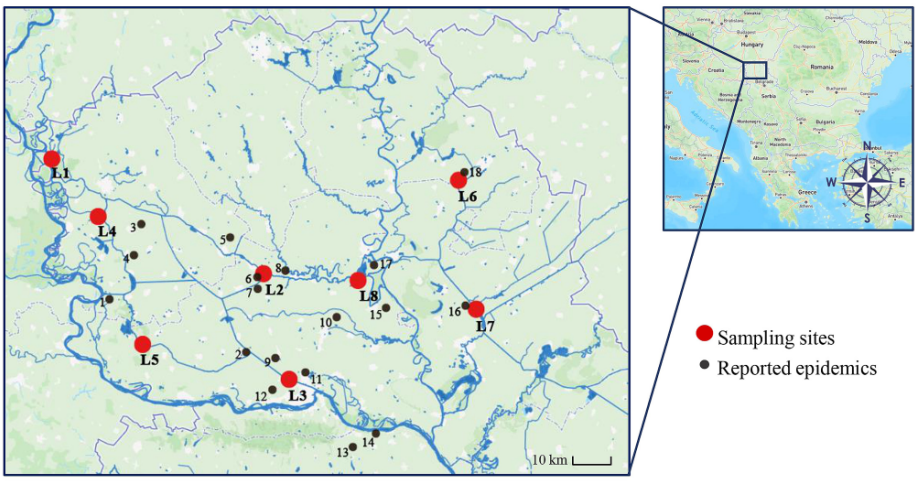


Fig. 1. Sampling locations on the Danube-Tisa-Danube Hydrosystem (red dots: L1 – Bezdan, L2 – Vrbas, L3 – Rumenka, L4 – Prigrevica, L5 – Bač, L6 – Kikinda, L7 – Jankov Most, L8 – Bačko Gradište) and nearby reported epidemics caused by P. syringae strains belonging to phylogroup 2 (black dots: 1 - epidemics on sweet cherry, 2 - oil pumpkin, 3–18 - sugar beet). Map created using Kepler.gl (https://kepler.gl).

photometric determination of NH_4^+, NO_3^-, NO_2^-, and PO_4^{3-} was based on specific chemical reactions that produce measurable color changes, with concentrations determined by comparison to calibration standards. All laboratory analyses were performed in accordance with standard protocols recommended by the manufacturers of the kits.

2.2. Sample processing – membrane filtration and microbiological analyses

One liter of water from each sampling site was filtered in two 500 mL steps using membrane filtration, with a new membrane filter used for each step. Mixed cellulose ester filters with a pore size of 0.45 μm and a diameter of 47 mm (Millipore, France) were employed. After filtration, one filter from each sample was used for total environmental DNA (eDNA) extraction using the ZymoBiomics DNA Miniprep Kit, with the eDNA stored at −20 °C for future analyses.

The second filter was immersed in 30 mL of sterile distilled water in an Erlenmeyer flask and subjected to agitation at 270 rpm for 15 min, followed by 1 min of sonication at 35 kHz in an ultrasonic bath. The 30 mL suspension was centrifuged at 2000 *g* for 20 min to concentrate the bacteria separated from the filter. The supernatant was discarded, and the bacterial pellet was resuspended in 1 mL of sterile distilled water, effectively concentrating the original sample 500-fold. This concentrated suspension was serially diluted and spread-plated onto King's medium B supplemented with cephalexin, boric acid, and cycloheximide (KBC), semiselective for *Pseudomonas* species (Mohan, 1987). The incubation step lasted for 5 days at room temperature.

Additionally, unfiltered water samples were used to quantify aerobic heterotrophic bacteria by dilution spread-plating on Nutrient Agar. The ratio of total heterotrophic bacteria (determined by pour-plating on Nutrient Agar) to oligotrophic bacteria (determined by pour-plating on 1:10 diluted Nutrient Agar) was also calculated after 4 days of incubation at room temperature.

The quantification of standard fecal indicator bacteria (SFIB), including *Escherichia coli*, total coliforms, and intestinal enterococci, was performed using the IDEXX Quanti-Tray 2000 system (IDEXX Laboratories, USA). The methodology is based on defined substrate technology (DST), which detects specific enzymatic activity related to each target organism. For *E. coli* and total coliforms, the Colilert-18 reagent was used, while Enterolert-E was used to identify and quantify intestinal enterococci. The substrate was added to 100 mL of the native or adequately diluted water sample. For total coliforms and *E. coli*, the Colilert-18 substrate is activated within 18 hours of incubation at 37 °C, where *E. coli* produces fluorescence and total coliforms induce a yellow color change. For intestinal enterococci, samples were incubated for at least 24 hours at a temperature of 44 °C. After incubation, the positive (color-changed or fluorescent) wells were counted to estimate the Most Probable Number (MPN) of SFIB in water.

2.3. PCR detection of Pseudomonas syringae complex

All *Pseudomonas* colonies from countable KBC plates (<300 colonies) were picked and subcultured to patch plates, where they were grown on fresh KBC medium for subsequent colony-PCR analysis. In the case of KBC plates with an uncountable number of colonies, only those resembling *Psy* were subcultured. We based our assessment on the description provided by Morris et al. (2022), who reported that *Psy* colonies are smooth with translucent, irregular edges, whereas some representatives exhibit a mucoid appearance but are not fully opaque. Colonies from the patch plates were subjected to a *P. syringae* species-specific PCR using primers Psy_F and Psy_R, designed by Guilbaud et al. (2016).

PCR reactions were carried out in a MiniAmp™ Thermal Cycler (Thermo Fisher Scientific, Waltham, Massachusetts, USA). The PCR mixture and conditions followed the protocol established by Guilbaud et al. (2016). Briefly, reactions were conducted in an 18 μL final volume, consisting of 10 μL of DreamTaq™ Green PCR Master Mix (2 ×), 6 μL of PCR water (Thermo Fisher Scientific, Waltham, Massachusetts, USA), and 0.55 μM of each *Psy*-specific primer. Bacterial cells from single colonies were collected using sterile pipette tips and directly transferred into PCR tubes containing the PCR mixture.

The thermal cycling conditions were as follows: an initial denaturation at 96 °C for 5 min, followed by 30 cycles of denaturation at 94 °C for 30 s, annealing at 61 °C for 30 s, extension at 72 °C for 30 s, and a final elongation step at 72 °C for 10 min.

PCR products were visualized on 1 % agarose gels containing ethidium bromide (0.4 μg/mL, SERVA, Germany) with a transilluminator (LKB, Transilluminator 2011 Microvue UV Light, Sweden). Electrophoresis was run in Tris-borate-EDTA (TBE) buffer (5.4 g Tris, 2.75 g boric acid, 4 mL of 0.5 M pH 8.0 EDTA, with dH_2O added to 1 L) at 90 V and 300 mA for 90 min. Colonies that tested positive for *Psy* (amplification of 144-bp specific fragment) were preserved in LB medium containing 20 % glycerol and stored at −80 °C. Total genomic DNA was extracted from the preserved isolates for further analysis using a modified CTAB protocol (Anteljević et al., 2023).

2.4. Phylogenetic and phenotypic analyses

To confirm the identity of *Psy* isolates from the DTD, a partial sequence of the *cts* housekeeping gene was amplified and sequenced using primers described by Morris et al. (2010). PCR amplification, purification, sequencing, and phylogenetic analysis followed the same methodology as in our previous study (Anteljević et al., 2023). The partial *cts* gene sequences (368 bp) were aligned with the Muscle program integrated into MEGA 11 software, which was also used to generate a phylogenetic tree with 1000 bootstrap replicates based on the neighbor-joining method to infer evolutionary relationships. Besides *Psy* reference strains, *P. graminis, P. rhizosphaerae, P. protegens*, and *P. aeruginosa* were included as out-groups. The tree was visualized using iTOL (Letunic and Bork, 2024) and rooted with *P. aeruginosa* PAO1. The partial *cts* gene sequences were deposited in NCBI under accession numbers PQ586931-PQ586972 (Supplementary Material 1, Table S1) and compared using the NCBI BLASTn tool to identify similarities with sequences available in the database, allowing us to investigate potential links to previous epidemics and distinguish haplotypes.

Phenotypic tests were performed to evaluate virulence-related traits of the *Psy* isolates, which were confirmed through *cts* gene analysis. These included the hypersensitive response assay on *N. benthamiana*, ice nucleation activity, pectinolytic activity, and motility assays (swimming and swarming). The preparation of bacterial suspensions and experimental procedures are in detail described in our previous work (Anteljević et al., 2023).

2.5. Host range

Pathogenicity tests were conducted on nine host plants (Table 1) to determine the host range. Plant species were chosen based on existing knowledge of previous infections in Serbia and the most commonly cultivated vegetables in the study area (FAO, 2022). We selected domestic cultivars of plants grown in Potgrond H substrate (Klasmann-Deilmann GmbH, Germany) within a greenhouse maintained at ambient temperatures. The virulence of *Psy* isolates from PG02, which tested positive in the hypersensitivity assay, was assessed. Bacterial suspensions for leaf inoculations were prepared from overnight cultures (grown in King's B medium at 30 °C, with agitation set at 180 rpm) by adjusting their optical density in fresh KB broth to $OD_{600} = 0.2$ (approximately 10^8 CFU/mL). Subsequently, suspensions were centrifugated at 13,000 g for 10 min and the precipitate was resuspended in sterile distilled water. One leaf per plant was inoculated with ~10 μL of a bacterial suspension at the abaxial side between two lateral nerves using a needleless syringae. Each isolate was tested in five replicates. Sterile distilled water was used as a negative control. Infected plants were incubated in a growth chamber with relative humidity maintained between 70–90 %, an

Table 1
Plant species and cultivars used for host range testing of Pseudomonas syringae isolates from the Danube-Tisa-Danube Hydrosystem.

Plant species	Cultivar	Age of plants at inoculation (months)	Source type
Beta vulgaris var. *saccharifera* *(Sugar beet)*	Heston	1	Seed
Solanum lycopersicum *(Tomato)*	Novosadski jabučar	1	Seed
Cucumis sativus *(Cucumber)*	Delikates	1	Seed
Pisum sativum *(Pea)*	Mali provansalac	1	Seed
Brassica oleracea var. *capitata* *(Cabbage)*	Futoški	1	Seed
Citrullus lanatus *(Watermelon)*	Slatka princeza	1	Seed
Capsicum annuum *(Chili pepper)*	Crveni ljuti	2	Seed
Petroselinum crispum *(Parsley)*	Lišćar	2	Seed
Allium cepa *(Onion)*	Kupusinski jabučar	1	Bulb

ambient temperature of 25 °C, and 12 h light/dark period. Plant-scoring was done 1, 3, 7, and 10 days after inoculation and the evaluation process was designed as an adjustment of the method used by Balthazar et al. (2022). Reactions were observed as three groups: no visible reaction, localized lesions at the inoculation zone and spreading lesions. In the case of spreading lesions, the percentage of leaf area exhibiting disease symptoms (necrosis and chlorosis around the inoculation zone) was measured by using ImageJ software (Schneider et al., 2012). Isolates were considered to have a pathogenic potential if at least three out of five replicate plants exhibited spreading lesions.

2.6. Design and validation of new qPCR tests

2.6.1. Design of qPCR tests for specific detection of P. syringae strains from phylogenetic group 2

Genomes of diverse *Pseudomonas* spp. strains were sourced from the NCBI GenBank database (Supplementary Material 2, Table S1). Employing DNA-DNA hybridization (DDH) values, the genomes were clustered using the Type Strain Genome Server (TYGS) web server (Meier-Kolthoff and Göker, 2019) for precise taxonomic classification, with the computational parameters set to default.

Genomes were categorized into "positive" and "negative" groups based on position in the phylogenetic tree. "Positive" included genomes of *Psy* strains from phylogenetic group 2 for which qPCR tests for specific detection were designed, while "negative" comprised all other strains. Five primer design analyses were conducted using different combinations of strains from both groups (Supplementary Material 2, Table S1). RUCS (Rapid Identification of PCR primers for Unique Core Sequences) webtool (Thomsen et al., 2017) was employed to identify unique core sequences (UCS) for the "positive" group. Computational parameters in RUCS webtool were set to default.

The primary RUCS output yielded a list of UCS for the "positive" genome group, organized based on their size. The top ten longest sequences were selected, and their associated genes were identified using the NCBI BLASTn tool. To ensure the exclusion of mobile genetic elements, unique core sequences not associated with such elements were considered for subsequent primer design.

These chosen UCS were then subjected to Primer Express v2.0 software (Applied Biosystems, USA) for the automated generation of qPCR primers and Taqman probes, with the default configuration of computational parameters. In instances where UCS identified by RUCS were too short for the design of the qPCR test, sequences were extended downstream. Using the CLC Main Workbench v 20.0.4 (QIAGEN) alignment tool, the augmented sequence underwent search for point mutations by comparing it to sequences of closely related strains determined through BLASTn.

The final selection of primers and probes adhered to the following criteria: the forward primer was based on the original unique core sequence, while the reverse primer and the probe were derived from the additional sequence (Supplementary Material 2, Table S2).

2.6.2. Genomic DNA isolation

All bacterial strains (Supplementary Material 2, Table S3) were cultivated on King's B agar medium before undergoing DNA isolation. The method employed for DNA isolation utilized thermal lysis in microtiter plates. A single colony of each bacterial strain was suspended in 100 µl of sterile Milli-Q water (Merck, USA). Subsequently, plates were heated at 100 °C for 5 min, then transferred to ice for 5 to 10 min. Afterward, the plates were briefly centrifugated and stored at − 20 °C. The DNA extracted in this manner was used to assess the specificity of the primers in the qPCR method (Supplementary Fig. 1).

Genomic DNA from spiked samples (sugar beet phyllosphere homogenate combined with bacterial suspension) was extracted by the magnetic-bead-based QuickPick™ SML Plant DNA Kit (Bio-Nobile) as described by Pirc et al. (2009) with a slight modification (415 µL lysate for purification). The DNA was stored at ≤− 15 °C until analysis. The extracted DNA was utilized to evaluate the sensitivity of qPCR tests (Supplementary Fig. 1).

2.6.3. Setup of the qPCR tests

All four sets of primers and probe pairs were tested in 10 µL reactions, with the following composition: TaqMan 2 × Universal PCR Master Mix (Applied BiosystemsTM, USA) 5 µL, forward and reverse primer 0.9 µM each, probe 0.25 µM, Milli-Q water 2.80 µL, and 2 µL of DNA sample (from strains described Supplementary Material 2, Table S3). All reactions were performed in three technical repeats on 384-well plates, with each plate also including no-template control (NTC), a negative control of DNA isolation, and NIB Z 1237 (CFBP 1617) DNA as the positive control.

The qPCR amplifications were performed using the ViiA™ 7 Real-Time PCR System (Applied Biosystems, USA) under the following conditions: initial incubation for 2 min at 50 °C, followed by the activation of TaqMan polymerase for 10 min at 95 °C. Subsequently, 45 cycles were performed, consisting of denaturation for 15 s at 95 °C and hybridization/elongation for 1 min at 60 °C.

2.6.4. Validation of qPCR tests: analytical specificity

Given the well-documented diversity of the *Psy* complex, we decided to include strains from multiple sources to examine qPCR test specificity. In addition to 50 *Psy* strains isolated from the Danube River and DTD, we included 49 *P syringae* pv. *aptata* strains isolated from the sugar beet epidemics in Serbia (Nikolić et al., 2018), 16 *Psy* strains from various plant hosts, and 39 isolates from the sugar beet phyllosphere microbiota (Krstić Tomić et al., 2023) (Supplementary Material 2, Table S3). The specificity of the qPCR tests was evaluated by calculating true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) among the analyzed *Psy* and non-*Psy* strains. Obtained values were used to calculate the following parameters: analytical specificity - representing the ability of the method to identify true negatives among the total samples accurately [TN/(FP+TN)], percentage of false positives - indicating the proportion of samples falsely identified as positive [FP/total number of strains], percentage of false negatives - denoting the percentage of samples wrongly classified as negative [FN/total number of strains], inclusivity - measuring the method's capability to detect all strains expected to be positive [TP/(targets number)] and exclusivity - signifying the method's effectiveness in excluding samples that are supposed to be negative [TN/(non-targets number)].

2.6.5. Validation of qPCR tests: analytical sensitivity

A mixture of sugar beet DNA and bacterial DNA at different concentrations extracted from spiked samples was utilized to evaluate the sensitivity of the qPCR method (DNA extraction described in paragraph 2.6.2). Spiked samples were generated by combining 180 µl of sugar beet phyllosphere homogenate (10 g of leaf ground using a mortar and pestle in 5 mL 1 × PBS buffer (1.08 g Na_2HPO_4, 0.4 g $NaH_2PO_4 \times 2H_2O$, 8 g NaCl, 1 L distilled water, pH 7.2)) with 20 µl of bacterial suspension to obtain concentrations ranging from 10^8 cells/mL to 10 cells/mL. Bacterial suspensions for three *P. syringae* pv. *aptata* (*Ptt*) strains (P21, P16, and NIB Z 1237 (CFBP 1617)) were prepared by resuspending colonies in 5 mL of 10 mM PBS buffer with 10 % glycerol (v/v). The concentrations of the bacterial suspensions were determined turbidimetrically using a densitometer (DEN-1B; BioSan, Riga, Latvia). The sensitivity of the method was determined by comparing Ct values across concentrations of bacteria ranging from 10^8 cells/mL to 10 cells/mL for three *Ptt* strains. Additionally, logarithmic concentrations with their corresponding Ct values were graphically connected. A linear trendline equation was calculated based on this relationship. The qPCR tests' amplification efficiency (E) was then calculated using the standard curve slope. The efficiency was determined using the formula: $E = (10^{(\ 1/\text{slope})}\ 1) \times 100$ (Supplementary Fig. 1).

2.7. Detection and quantification of Pseudomonas syringae complex and phylogroup 2 using qPCR

eDNA samples from eight sites within the DTD (Fig. 1) were analyzed using qPCR testing to detect *Psy* complex and PG02. For the reactions targeting the *Psy* complex, the primers Psy_F and Psy_R, originally designed by Guilbaud et al. (2016), were used, along with a probe (5′-FAM-TAAAGTGATCGACAAGGGCGCTGA-BHQ_1–3′) designed using the PrimerQuest™ Tool (Integrated DNA Technologies, USA). To optimize the reaction mix, we evaluated four different formulations with varying concentrations of primers and probes on strain *Ptt* P21 as a positive control (Supplementary material 3, Table S1). The final reactions were conducted on eDNA samples using a mix containing 5 µL of TaqMan 2 × Universal PCR Master Mix (Applied Biosystems™, USA), 5 µM of both forward and reverse primers, and 1.6 µM of the probe. Two concentrations of eDNA samples were tested, at 1 ng/µL and 2 ng/µL, with Milli-Q water added to achieve a final reaction volume of 10 µL. All reactions were conducted in triplicate. To assess the method's sensitivity, qPCR was conducted on serial dilutions of positive control *Ptt* P21 DNA at concentrations of 0.5, 1, 1.5, and 2 ng/µL. The resulting Ct values were plotted against the logarithmic DNA concentrations to generate a standard curve, and amplification efficiency was calculated (paragraph 2.6.5).

For the detection of PG02, the Psa_AvrE1 primer pair and probe were used (Supplementary Material 3, Table S2) with a qPCR reaction mix prepared as previously described and 2 ng/µL of eDNA sample (paragraph 2.5.3).

The qPCR reactions were executed using the ViiA™ 7 Real-Time PCR System (Applied Biosystems, USA). As described in paragraph 2.6.3, the thermal cycling conditions were identical for both target groups. qPCR was used to quantify the target DNA in the sample and to calculate the abundance of the *Psy* complex and PG02 in the DTD, as target sequences are present in one copy per genome. Copies per microliter of reaction were determined by interpolating Ct values from a standard curve. These values were then adjusted based on the total DNA extracted from the water sample and the water sample volume, allowing the final results to be expressed as cp/L (copies per liter of water).

2.8. Digital PCR (dPCR) for detection and quantification of Pseudomonas syringae complex and phylogroup 2

dPCR was performed on eight eDNA samples from sites within the DTD (Fig. 1) using a QX200 Droplet Digital System (Bio-Rad) according to the manufacturer's instructions. Each dPCR reaction mixture (total of 20 µL) consisted of 10 µL of 2 × digital PCR Supermix for Probes, 0.9 µM forward and reverse primers, 0.25 µM probe (sequences for both assays are provided in Supplementary Material 3, Table S2), and 2 µL of template DNA (resulting in a final concentration of 1 ng/µL in the reaction). Droplets were generated using the QX200 Droplet Generator, and PCR amplification was conducted on a Bio-Rad T100 thermal cycler under the following conditions: initial denaturation at 95 °C for 10 min, followed by 40 cycles of 94 °C for 30 s and 59.3 °C for 60 s (with a ramp rate of 2 °C/s), and a final extension at 98 °C for 10 min. Droplets were analyzed using the QX200 Droplet Reader, and only samples with a minimum of 10,000 droplets generated were considered for analysis. Fluorescence data were analyzed using QX Manager Software (version 2.1), with the threshold for positive droplets set manually based on the fluorescence amplitude of no-template controls and positive controls (DNA extract of strain BFD143). The Limit of Quantification (LOQ) was determined by analyzing serial dilutions of the BFD143 DNA extract, where LOQ was set as the lowest concentration, yielding a coefficient of variation (CV) $\leq$25 % across five replicates (Supplementary Material 4, Figs. S1 and S2). The Limit of Blank (LOB) was calculated using the mean value of blank samples (bm) and their standard deviation (stdev), following the formula: $\mathbf{LOB = bm + 1.65 \times stdev}$ (Clinical and Laboratory Standards Institute 2004). Samples that did not meet the LOQ thresholds were assessed qualitatively as either presence or absence of the target DNA. A sample was considered positive if its value exceeded the LOB threshold.

The number of target copies per microliter in the dPCR reactions was used to quantify the *Psy* complex and PG02. These values were then adjusted based on the total DNA extracted from the water sample and the sample volume, allowing the final results to be expressed as cp/L.

2.9. Statistical analyses

Statistical analysis of the dPCR data was conducted using GraphPad Prism (version 8.4.3). Group differences were assessed by one-way ANOVA with Tukey's HSD post hoc test, with significant differences marked at $p < 0.05$.

To assess the agreement between the *cts*-based identification and dPCR methods for quantifying *Psy* complex abundance in the DTD, a Bland-Altman analysis was performed on log-transformed data using GraphPad Prism (version 8.4.3). The percentage difference between the two methods was calculated using the formula:

$$\textit{Percentage difference} = 100 \times \frac{(A \quad B)}{\textit{Average of A and B}}$$

Where A represents the quantification of *Psy* abundance by the *cts*-based identification method, B represents the quantification by dPCR, and the average is the mean of the two methods ($\frac{A+B}{2}$). The resulting percentage differences were then plotted against the average values to assess the agreement between the two methods.

For the host range test, statistical analyses were performed using RStudio software (version 2024.09.0, R version 4.3.2) with the stats (R Core Team, 2024) and dunn.test (Dinno, 2017) packages, while the ggplot2 (Wickham, 2016) package was used for data visualization.

The normality of the data was assessed using the Shapiro-Wilk test, conducted on both non-transformed and log-transformed datasets. Since the data did not follow a normal distribution, non-parametric tests were applied. The Kruskal-Wallis test was used to compare symptom severity among groups based on bacterial strains, host plants, and sampling days. To evaluate the differences in symptom severity caused by various bacterial isolates on the same plant host across different days, Dunn's test was performed as a post-hoc analysis following the Kruskal-Wallis test. This test compares all possible pairs of bacterial isolates within each host and day combination to identify statistically significant differences in symptom severity. The resulting p-values were adjusted for

multiple comparisons using the Bonferroni correction to control for Type I error. The most significant differences were extracted for each day-host combination and visualized to highlight the key findings of the analysis. Additionally, a K-means clustering analysis was performed to group bacterial strains based on their overall aggressiveness in causing symptoms across different host plants. Clusters were visualized to observe similarities between strains.

3. Results

3.1. Water quality of the DTD hydrosystem shows spatial variability

Measurements of physicochemical and microbiological parameters were conducted on eight sampling sites of the DTD: L1 – Bezdan, L2 – Vrbas, L3 – Rumenka, L4 – Prigrevica, L5 – Bač, L6 – Kikinda, L7 – Jankov Most, L8 – Bačko Gradište (Fig. 1). Water temperature ranged from 13.3 °C at site L7 to 17.1 °C at site L5, typical for spring conditions (Supplementary Material 1, Table S3). Conductivity, which reflects ion concentration, was highest at L3 (608 μS/cm) and lowest at L7 (422 μS/cm). Oxygen levels showed considerable variation, with the highest concentration measured at L2 (19.4 mg/L) and the lowest at L6 (5.5 mg/L). The pH levels indicated slightly alkaline conditions across the sites. Ammonium, nitrate, and phosphate concentrations were higher at L6 and L7, suggesting increased nutrient inputs at these sites.

In terms of microbiological water quality, the sites L6 and L7 displayed the most concerning results, with excessive counts of total coliforms, intestinal enterococci, and aerobic heterotrophs, indicating significant fecal and organic pollution (Supplementary Material 1, Table S3). At site L1, these groups had relatively low total counts, indicating minimal contamination. The oligotroph-to-heterotroph ratio was highest at L2 (25.54), suggesting lower nutrient availability compared to other sites. In contrast, L1 (0.58), L6 (0.86), and L5 (0.76) had lower ratios, indicating higher nutrient availability and greater heterotrophic activity, often linked to pollution.

Among the eight sites where physicochemical and microbiological parameters were assessed, only L4 and L8 met the criteria for indication of good to moderate ecological potential (Class II-III), according to the standards applied to artificial water bodies in Serbia (Official Gazette of the Republic of Serbia No.74/2011 2011). The other six sites would thus classify within moderate to poor ecological potential (Class III-IV) based on parameters that exceeded optimal limits.

3.2. Isolation and cts gene-based identification revealed high abundances of Pseudomonas syringae in the Danube-Tisa-Danube Hydrosystem

To detect and isolate representatives of *Psy* from samples collected at eight different sites within the DTD canal network, a total of 651 bacterial colonies isolated on semi-selective KBC media were screened using species-specific PCR primers targeting the *Psy* complex (Supplementary Material 1, Table S1). Of the 651 colonies, 68 (10.45 %) showed

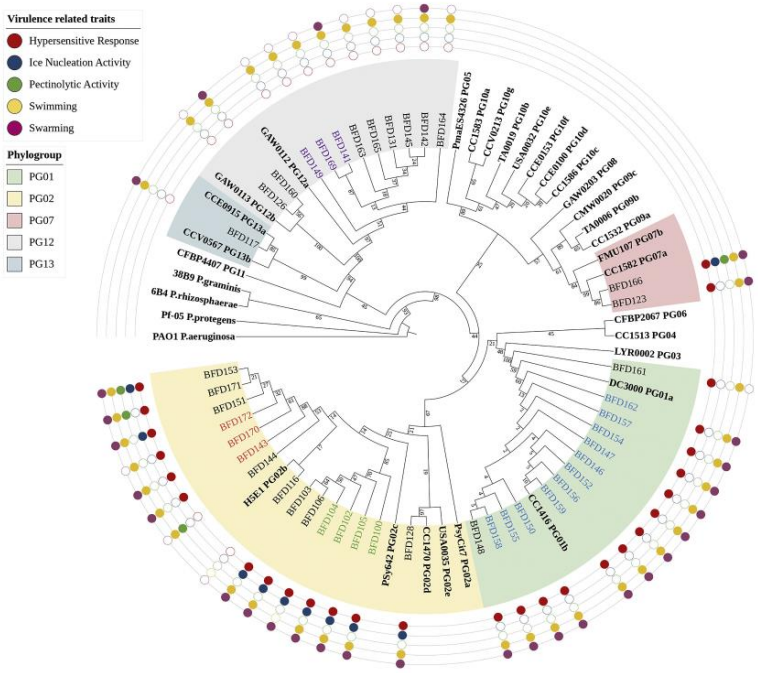


Fig. 2. The phylogenetic tree of *Pseudomonas syringae* isolates constructed based on the partial sequence of the *cts* gene. A phylogenetic relationship between 42 *Psy* isolates from the Danube-Tisa-Danube Hydrosystem (BFD labeled isolates), reference strains representing 13 phylogroups, and strains used as out-groups (in bold) is shown. The tree was generated by the neighbor-joining method. Bootstrap values (expressed as a percentage of 1000 replications) are shown at branch nodes. The presence (colored circles) or absence (empty circles) of tested phenotypic traits is displayed next to each isolate. Isolate codes labeled with the same color (red, green, blue, and purple) share the identical *cts* sequence.

successful amplification of a specific 144-bp fragment, and those were considered putative *Psy*. The identity of these isolates was further validated by DNA sequencing of the partial *cts* gene, coding for citrate-synthase, a key enzyme in regulating energy generation of mitochondrial respiration (Cheng et al., 2009). Among the initial 68 isolates, 42 (68 %) were finally confirmed as *Psy* and preserved in a strain collection (Supplementary Material 1, Table S1).

The abundance of the *Psy* complex at each site varied significantly in a range of 10^4– 10^7 CFU/L (Supplementary Material 1, Table S1). Site L5 displayed the highest *Psy* abundance, reaching 1.9×10^7 CFU/L, while the lowest confirmed abundance was at L1 with 6×10^4 CFU/L. At L6, none of the colonies were confirmed as *Psy* by *cts* sequencing, so this site was excluded from abundance calculations.

Phylogenetic analysis of the 42 confirmed *Psy* isolates, based on the *cts* gene (368 bp), revealed the presence of five distinct phylogroups within the DTD (Fig. 2): PG01 (30.9 %), PG02 (35.7 %), PG07 (4.8 %), PG12 (26.2 %), and PG13 (2.4 %). Phylogroup 2 was the most widely distributed, being detected at four different sites. The abundance of PG02 was the highest at site L5 (6×10^6 CFU/L) (Supplementary material 1, Table S4). Notably, L5 exhibited the highest diversity, with four distinct phylogroups identified at this site. We could observe the dependence of the phylogroup distribution on the sampling point. At the rest of the sites, the presence of only one or two phylogroups was confirmed (Supplementary Material 1, Table S1).

Further comparisons of the isolates' *cts* sequences with publicly available sequences in the NCBI database revealed 25 distinct haplotypes among the 42 isolates, some of which showed 100 % identity in nucleotide sequence to previously described strains (Supplementary Material 1, Table S5). In PG02, a group of isolates (BFD143, BFD170, and BFD172) shared identical *cts* sequences with the DD.1 haplotype described by Morris et al. (2023), which has been reported as a widespread haplotype in river water, possessing a highly virulent phenotype. It matches many plant pathogenic strains, including *P. syringae* pv. *aptata* from earlier epidemics on sugar beet in Serbia. Similarly, another haplotype within PG02 isolates (BFD100, BFD102, BFD104, and BFD105) corresponded to strains previously isolated from diverse environments such as freshwater in France (Berge et al., 2014), snow in Switzerland, and stone fruits in Poland.

In phylogroup 1, a larger group of isolates (BFD146, BFD147, BFD150, BFD152, BFD154, BFD155, BFD156, BFD157, BFD158, BFD159, and BFD162) displayed identical *cts* sequences. These sequences matched the reference strain *P. syringae* CMO0017, isolated from rainwater in France, and a strain isolated from kiwi fruit plants in Spain. Additionally, another set of isolates from PG12 (BFD141, BFD149, and BFD169) shared 100 % sequence identity with the reference strain *P. syringae* GAW0112, isolated from an irrigation canal network in France. Considering these concurrences, we observed that half of the isolates from our collection (21) belong to some of the haplotypes previously detected in river or rainwater.

3.3. *Pseudomonas syringae isolates from the Danube-Tisa-Danube Hydrosystem exhibited virulent traits important for pathogenic lifestyle*

The examination of hypersensitive response (HR), ice nucleation activity (INA), pectinolytic activity (PA), and motility capabilities in the isolates from our collection provided insight into their virulence potential (Fig. 2). A hypersensitive response was induced by 28 isolates (66.67 % of the collection), indicating the pathogenic potential of these strains. The isolates that caused the necrotic zones in the inoculated leaf areas belonged to phylogroups 1 (46.4 %), 2 (46.4 %), and 7 (7.2 %). In contrast, isolates from phylogroups 12 and 13 did not induce a hypersensitive response, consistent with current findings regarding these groups (Berge et al., 2014; Morris et al., 2022). Regarding INA, only ten isolates (23.8 %) promoted ice formation at temperatures ranging from −2 °C to −10 °C. One ice-nucleating isolate belonged to PG07, and the others belonged to PG02. The highest temperature at which ice formation was observed was− 4 °C, noted in isolates BFD128 and BFD153. INA was only observed in isolates that induced HR, suggesting an association between these two traits. The pectinolytic activity was detected in four isolates (9.5 %), while motility was a much more widespread trait, with swimming observed in 40 isolates (95.2 % of the collection) and swarming in 30 isolates (71.4 %). Isolates BFD104 and BFD116 from PG02 were the only ones lacking swimming ability, while swarming was predominantly absent in isolates from PG12.

Among the *Psy* isolates from the DTD, BFD153 (PG02) and BFD166 (PG07) were the only strains to test positive for all virulence traits, while BFD116 (PG02) was the only isolate negative for all traits. The most common trait combination, observed in 15 isolates from phylogroups 1 and 2, included HR, swimming, and swarming. Many non-swarming isolates could not induce HR, indicating that swarming ability may be a critical factor for disease development. Considering the observed virulence traits of the *Psy* population representatives from the DTD, we could also conclude that certain strains could negatively affect nearby crops through irrigation practices, posing a risk to agricultural productivity.

3.4. *Phylogroup 2 representatives of Psy complex induced severe symptoms in a wide range of host plant tested*

PG02 was the most commonly identified as a causative agent of diseases in crops in the region within the DTD canal network, among other phylogroups. To assess the epidemiological risk of the isolated *Psy* PG02 strains, we conducted pathogenicity tests on economically important vegetables widely cultivated for human consumption, as well as on sugar beet, an industrial crop that was severely affected by *Ptt* outbreaks in Vojvodina Province over the past decade (Nikolić et al., 2018). Out of 15 representatives of PG02 isolated from the DTD, 13 isolates that triggered a hypersensitive reaction in *Nicotiana benthamiana* (BFD128 from clade 2d, and the other twelve from clade 2b) were tested for pathogenicity on nine different plant species. Lesion spreading in some cases was already evident by day 3. Generally, the isolates exhibited a broad host range, with at least five plant species developing visible symptoms. Isolates BFD128, BFD151, BFD171, and BFD172 were the most virulent, inducing symptoms across all plant species tested. In contrast, isolate BFD102 displayed the narrowest host range (Fig. 3a). Chili pepper and watermelon exhibited the highest resistance, as more than half of the tested isolates were unable to induce disease symptoms in these hosts.

Statistically significant differences in symptom severity across plant hosts depending on the bacterial strain were confirmed (Chi-square = 1358.8, $p = 1.11 \times 10^{-118}$, df = 350), with the most significant findings attributed to BFD100 and BFD128 isolate combination. Notably, isolate BFD128 consistently induced more severe symptoms compared to BFD100, particularly on tomato ($Z = -5.88181$), sugar beet ($Z = -4.06315$), and chili pepper ($Z = -4.04011$). The pronounced aggressiveness of BFD128 highlights its high virulence relative to the other tested strains (Fig. 3b).

Based on average symptom severity, isolates were grouped into three clusters through K-means clustering: Cluster 1 (high severity, 59.08–100 % symptom severity), Cluster 2 (low severity, 0–20.06 %), and Cluster 3 (moderate severity, 21.49–36.66 %) (Fig. 3c). On the third day post-inoculation, the majority of isolates were classified within the low-severity cluster. By day 7, a more significant number of plant hosts exhibited moderate to high-severity symptoms, with isolate BFD128 notably causing severe symptoms in cabbage, tomato, and chili pepper. On day 10, the incidence of high-severity symptoms increased further. Notably, chili pepper and watermelon remained the only hosts without high-severity symptoms, underscoring their resilience. Observations of symptom severity 10 days post-inoculation indicate the presence of two groups of isolates differentiated by their aggressiveness. A less aggressive group, consisting of strains BFD100, BFD102, BFD103, BFD104, BFD105, and BFD106, was characterized by only a few instances of

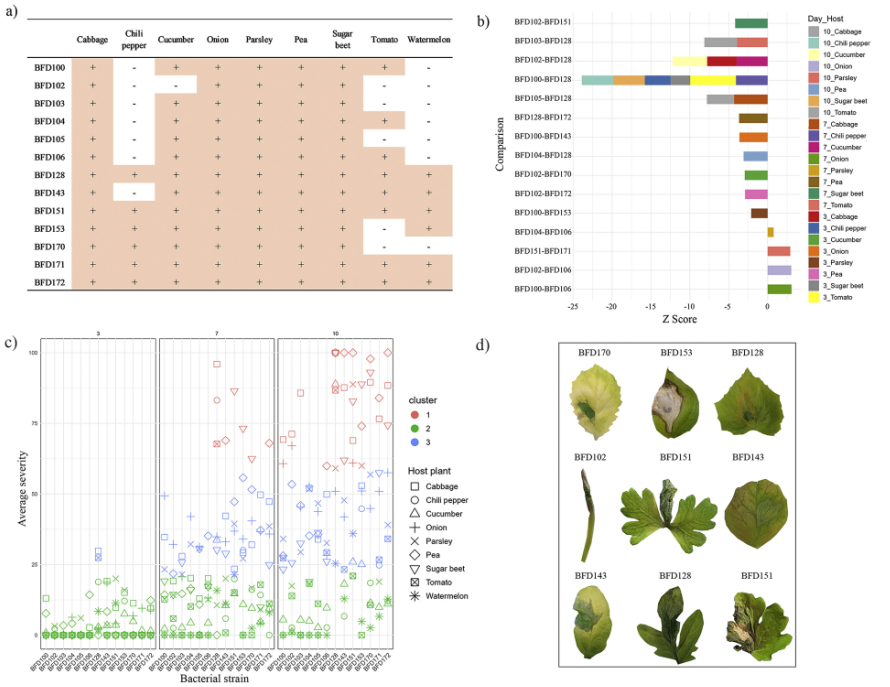


Fig. 3. Analysis of *Pseudomonas syringae* phylogroup 2 isolates from the Danube-Tisa-Danube Hydrosystem in host range tests. a) A table displaying the presence (+) or absence ( ) of disease symptoms on plant hosts infected with phylogroup 2 isolates, observed 10 days post-inoculation. b) Dunn's post hoc test results, showing the most statistically significant differences in symptom severity caused by isolate pairs for each host and day combination. Higher absolute Z-scores indicate more substantial evidence of significant differences. c) K-means clustering results categorizing isolates across days based on symptom severity percentages: Cluster 1 (high severity, 59.08–100 %), Cluster 2 (low severity, 0–20.06 %), and Cluster 3 (moderate severity, 21.49–36.66 %). d) Representative leaf specimens of each host showing symptoms caused by the most aggressive isolate on day 10 (layout of plant species tested: Upper row – cabbage, chili pepper, cucumber; middle row – onion, parsley, pea; bottom row – sugar beet, tomato, watermelon).

highly severe symptoms. Conversely, the more aggressive group, including BFD143, BFD151, BFD153, BFD170, BFD171, BFD172, and BFD128, induced highly severe symptoms in a larger number of tested hosts, underscoring the threat posed by their presence in the DTD (Fig. 3d).

3.5. Designed qPCR tests exhibit high specificity for the detection of phylogroup 2 within Pseudomonas syringae complex

Given the ubiquity and pathogenicity of PG02 representatives of the *Psy* complex, we aimed to develop specific and sensitive qPCR tests for their detection. Four qPCR primer pairs and corresponding probes were designed based on the genes *avrE1* (encoding a type III secretion system effector), *cstA* (carbon starvation protein), *pleD* (diguanylate cyclase), and *avrE2* (another type III secretion system effector) (Supplementary Material 2, Table S2). The specificity of all four sets of primers and probes was primarily evaluated for detecting PG02. All tests were conducted on 154 bacterial strains, representing various phylogroups of the *Psy* complex and other genera (Supplementary Material 2, Table S4).

The Psy_AvrE1 primer and probe set proved to be the most effective for detecting representatives of PG02, achieving 100 % analytical specificity by correctly identifying all 74 tested strains from PG02, with no false positives or negatives among the additional 80 non-target strains tested (Supplementary Material 2, Table S5). This suggests that Psy_AvrE1 could be a reliable tool for detecting PG02 strains.

Sensitivity tests using spiked samples with varying concentrations of *Psy* (PG02) strains revealed that the Psy_AvrE1 primer and probe set consistently detected bacteria across all replicates down to 5×10^3 cells/mL, confirming its high sensitivity (Supplementary Material 2, Tables S6, S7, S8, and Fig. S1). The qPCR assay using the Psy_AvrE1 primer set consistently detected bacterial DNA, demonstrating its robustness in complex sample matrices containing plant-derived DNA.

3.6. qPCR detection of Pseudomonas syringae complex and phylogroup 2 in the Danube-Tisa-Danube Hydrosystem was successful across samples tested from all sites

To detect and quantify the *Psy* complex and PG02, we performed qPCR on eDNA samples collected from all eight sites within the DTD. For the *Psy* complex, we employed the set of previously designed *Psy* primers (Guilbaud et al., 2016) and the PsyP TaqMan probe we designed to improve specificity. MIX3 was selected due to its consistent performance and reliable detection in optimization assays with positive controls at an eDNA concentration of 2 ng/μL, with a reaction efficiency of 114 %, which falls within the acceptable range for qPCR assay reliability (Supplementary Material 3, Table S1). The target DNA was detected in

each sampling site, though not consistently across all replicates. Due to variability in results, precise quantitative interpretation remains limited. The exception is the site L3 with the consistent Ct values across replicates, yielding an estimated population of 1.42×10^5 cp/L (Supplementary Material 5, Table S1).

The Psy_AvrE1 primers and probe set were employed to detect PG02 in the DTD system at the tested sites. The qPCR assay using the Psy_AvrE1 primer pair proved effective in detecting the presence of PG02 at each site, making it suitable for initial screening of their presence. However, its quantification potential was limited in most cases, as the assays with eDNA samples introduced obstacles and variations. The calculated population densities for sites L4 and L8 were on the order of 10^4 cp/L, with both sites consistently testing positive across replicates and exhibiting minimal standard deviations (Supplementary Material 5, Table S2).

3.7. dPCR outperforms the cultivation method in both the detection and quantification of the Pseudomonas syringae complex in the Danube-Tisa-Danube Hydrosystem

We applied the dPCR method to quantify the *Psy* population at eight sites within the DTD and to assess the abundance of PG02, following qPCR tests performed with the *Psy* primers described by Guilbaud et al. (2016) and the designed probe, as well as the Psa_AvrE1 primer and probe set for PG02. This transition to dPCR was made to achieve a more precise quantification (Fig. 4).

The target DNA concentration was above the verified limit of quantification (LOQ) for the *Psy* assay at each site (Supplementary Material 4, Fig. S1). Based on the concentration of the target DNA in *Psy* dPCR assay (Supplementary Material 4, Table S1), the highest abundance was recorded at the site L6 ($(1.54 \pm 0.27) \times 10^6$ cp/L), while the lowest concentration was observed at the site L1 ($(6.23 \pm 2.77) \times 10^4$ cp/L) (Supplementary Material 1, Table S1). The one-way ANOVA test showed a statistically significant difference in abundances between sites ($p < 0.01$) (Supplementary material 1, Tables S6 and S7), suggesting variable distribution of *Psy* population across the DTD Hydrosystem. A comparison of the methodologies revealed an average proportional difference of 33.42 % in calculated *Psy* abundance between the *cts*-based identification and dPCR, as determined by the Bland-Altman analysis. *Psy* population sizes determined using the traditional plating method combined with *cts* identification (Fig. 4c) were higher than those obtained by dPCR in seven out of eight sites. However, at the site L6, where

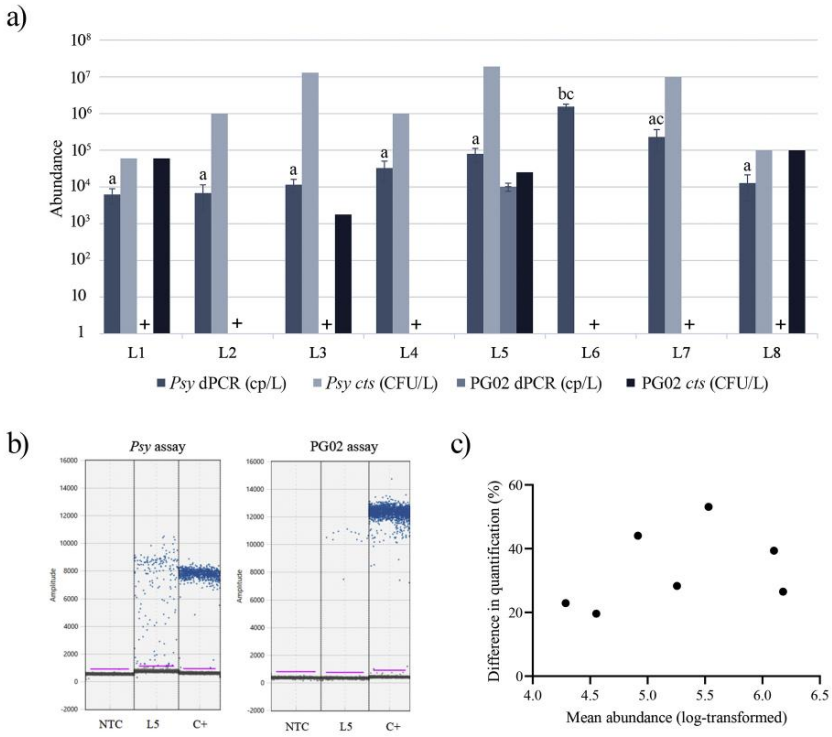


Fig. 4. The determined abundance of P. syringae complex and phylogroup 2 in the Danube-Tisa-Danube Hydrosystem assessed by dPCR and cts-based identification. a) Psy and PG02 abundances determined by both methods, presented on a logarithmic scale. Sites sharing the same letters (a, b, c) indicate no statistically significant differences in Psy abundance among these locations. PG02 was quantified by dPCR only at L5; at other sites, detection was below the quantification limit and is therefore interpreted qualitatively ('+' indicates the presence of PG02 at a given locality). b) Representative dPCR results of Psy and PG02 assays for site Bać (L5). Blue droplets indicate positive results, while grey droplets indicate negative results. NTC – no template control, *C+* – positive control strain BFD143. c) Results of the Bland-Altman analysis showing the percentage difference in the quantification of Psy complex abundance between the cts-based identification and dPCR method.

Psy could not be successfully isolated using the plating method, dPCR successfully quantified its abundance, confirming the sensitivity and effectiveness of the method.

For PG02, at the site L5 concentration of the target DNA was above LOQ (Fig. 4 and Supplementary Material 4, Fig. S2). The determined abundance for site L5 based on dPCR results was $(1.01 \pm 0.25) \times 10^4$ cp/L. At other sites, the concentration of target DNA was below LOQ, but above the determined LOB (Supplementary Material 4, Table S2), thus confirming the presence of PG02. Site L5 is therefore the only location where PG02 was both detected and quantified by dPCR. At this site, six strains belonging to PG02 were isolated. Although PG02 strains were also isolated from sites L1, L3, and L8, the combination of the highly specific Psy_AvrE1 primer set and the sensitivity of dPCR enabled PG02 detection even at sites where no isolates were obtained. This underscores the limitations of cultivation-based methods that rely on a small number of isolates, as some groups may be missed. In contrast, dPCR demonstrates superior sensitivity for detecting low-abundance groups such as PG02.

4. Discussion

Despite extensive studies of the *Psy* complex in non-agricultural contexts (Morris et al., 2023), its routine monitoring in water habitats and irrigation sources remains largely overlooked. Yet, these environments often host highly diverse populations and include key pathogenic strains (Morris et al., 2010, 2013). Our study specifically examined the *Psy* complex in the DTD, a multipurpose irrigation network of substantial economic importance to Serbia (Milanović et al., 2011), therefore a valuable target for investigating the *Psy* population in aquatic environments.

One objective of our study was to develop, validate, and optimize qPCR and dPCR assays for detecting the *Psy* complex in water samples, as molecular methods are more reliable and widely used for pathogen detection (Venbrux et al., 2023). Since the target sequences used for detection are present as single copies, we also utilized this approach to estimate bacterial abundances. This method significantly accelerates the process by bypassing cultivation and molecular identification, thus enabling more feasible and efficient regular monitoring. Compared to qPCR, dPCR demonstrated superior performance in quantification, probably due to its lower susceptibility to inhibition, enhanced analytical sensitivity, reduced variability at the detection limit, and greater precision (Tiwari et al., 2022). Unlike qPCR, which only offers relative quantification, dPCR also enables absolute quantification (Kokkoris et al., 2021). Interestingly, abundance estimated by dPCR was the highest at the site where cultivation-based isolation failed. The lack of highly selective media likely reduces isolation recovery rates and may cause underestimation of the *Psy* abundance and diversity by allowing competition from non-target strains and bias against certain *Psy* strains, as observed with KBC media (Morris et al., 2008). Increasing the number of replicate plates could provide more accurate abundance estimates, though this approach is even more time-intensive and may not be convenient for rapid diagnostics.

For PG02, although the detection with the highly specific Psy_AvrE1 primer-probe set was successful, low target DNA concentrations likely influenced and limited quantification. The common practice of diluting the sample to mitigate the adverse effects of co-extracted inhibitors, such as trace amounts of organic acids, salts, metals, household detergents, pharmaceuticals, and personal care products (Venbrux et al., 2023) may also reduce trace target DNA concentrations in environmental samples (Shi et al., 2021). Working with eDNA at low target concentrations can lead to uneven distribution across replicate reactions, a technical challenge that affects the accuracy and reproducibility of assays.

Regarding the diversity of *Psy* in the DTD, while PG02, PG07, and PG13 were previously detected along the Danube River in Serbia (Anteljević et al., 2023), this study marks the first detection of PG01 and PG12 in Serbian water systems. Phylogroup 1 includes strains isolated from both diseased plants and environmental habitats, grouped into clades 1a and 1b (Berge et al., 2014). The PG07 strains we obtained belong to clade 7a and are identified as *P. viridiflava*, a species commonly found in various environmental sources (Lipps and Samac, 2022). In Serbia, *P. viridiflava* was identified as a causative agent of pit necrosis in tomato plants, in the southern part of the country (Popović et al., 2015c). Strains among the PG12 and PG13 are not considered pathogenic, and both groups are commonly found in irrigation water, but PG13 strains have also been isolated from asymptomatic plants (Berge et al., 2014).

Phylogroup 2 stands out as the most widely distributed within the *Psy* complex, comprising representatives from diseased plants, asymptomatic plants, and non-plant substrates (Berge et al., 2014). In the area surrounding the DTD canal network, most epidemics in the last years have been caused by representatives of this phylogroup (Fig. 1 and Supplementary Material 1, Table S2). The most vulnerable plant species included sugar beet, oil pumpkin, pea, carrot, parsley, parsnip, and sweet cherry (Balaž et al., 2014; Ilićić et al., 2021; Popović et al., 2015a, 2015b; Stojšin et al., 2015). A broad host range of PG02 representatives has been demonstrated in nature and previous studies, with symptoms observed in over 75 % of tested plant species (Morris et al., 2019).

Epidemics on sugar beet, reported 10 years ago in Vojvodina (Stojšin et al., 2015; Nikolić et al., 2018) were attributed to the same haplotype now persisting in the DTD that matches *P. syringae* pv. *aptata* (Fig. 1). This haplotype, designated as DD.1, was reported as dominant in the Durance River catchment (Morris et al., 2023) and detected in river headwaters across the U.S., Europe, and New Zealand (Morris et al., 2010), and it has also been found in various plant-related contexts across four continents, which underscores its long-distance dissemination and potential for global spread (Morris et al., 2023). Its pathogenic potential is also notable, as we observed two groups of strains with differing levels of aggressiveness in our host range test, with the more aggressive group comprising isolates from the DD.1 haplotype (and others that differ by a single nucleotide from DD.1 in the *cts* gene sequence). The encounter location of highly aggressive strains aligns with nearby reported epidemic sites. In contrast, less aggressive strains were isolated from non-epidemic regions, suggesting a link between prior epidemics and the *Psy* population in the DTD. Specifically, the area around site L3 could be an important site for monitoring, as this part of the DTD harbors strains with a wide host range, capable of inducing severe symptoms in susceptible hosts, and closely related to the DD.1 haplotype, which was previously isolated from surrounding epidemics. However, no segment of the DTD should be neglected, as these strains could spread throughout the network and negatively affect a wide range of plants through irrigation.

As a leaf pathogen, *Psy* can effectively spread through sprinkling irrigation, which is the most commonly used method in Serbia (Statistical Office of the Republic of Serbia, 2024). A study demonstrated that an inoculum concentration as low as 1×10^2 CFU/mL is sufficient to induce characteristic midrib rot symptoms on lettuce when applied through overhead sprinkler irrigation (Cottyn et al., 2011). This finding highlights the pathogen's low threshold for initiating infection under favorable conditions, suggesting that some DTD sites may pose a significant threat to susceptible crops.

Understanding how *Psy* populations respond to nutrient-rich and microbially contaminated waters, such as those in the DTD system, requires further investigation and ongoing monitoring. A recent study on the Durance River catchment identified temperature as the primary factor influencing *Psy* abundance (Morris et al., 2023), but to our knowledge, the impact of organic pollutants on *Psy* populations has not been documented before.

5. Conclusions

Monitoring *Psy* populations in water systems connected to

agricultural fields is essential for assessing potential disease risks to crops. The *Psy* complex in the DTD demonstrated high strain diversity and significant pathogenic potential, which displayed a broad host range and the capacity for widespread dissemination, as evidenced by globally distributed haplotypes. In this study, we designed highly specific qPCR tests for the detection of PG02 of the *Psy* species complex. Also, we successfully utilized dPCR to detect and quantify *Psy* complex populations in irrigation water. At specific sites, *Psy* populations are abundant enough to trigger disease in susceptible hosts under favorable conditions, potentially highlighting the risk of extensive epidemics. Despite our efforts to quantify PG02 populations, low target concentrations restricted our results to qualitative analyses, emphasizing the need for improved sensitivity and specificity in detection methods. Enhanced monitoring strategies are crucial for the timely identification and management of *Psy* populations in irrigation systems to mitigate risks to crop health and agricultural productivity. Future studies should focus on refining molecular techniques and exploring the ecological dynamics of *Psy* in water systems to develop more effective disease prevention strategies.

CRediT authorship contribution statement

Marina Anteljević: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Iva Rosić:** Validation, Methodology, Investigation, Data curation, Conceptualization. **Olja Medić:** Writing – review & editing, Conceptualization. **Tamara Ranković:** Writing – review & editing, Conceptualization. **Karolina Sunjog:** Writing – review & editing, Investigation, Data curation. **Margareta Kračun-Kolarević:** Writing – review & editing, Investigation, Formal analysis. **Stoimir Kolarević:** Writing – review & editing, Visualization, Validation, Investigation, Formal analysis. **Tanja Dreo:** Supervision, Resources, Project administration, Conceptualization. **Aleksander Benčič:** Writing – review & editing, Validation, Investigation. **Tanja Berić:** Writing – review & editing, Conceptualization. **Slaviša Stanković:** Writing – review & editing, Supervision, Resources. **Ivan Nikolić:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank Mojca Milavec for her valuable contribution to establishing and supporting the collaboration that enabled this research. The content of this study is in line with following United Nations (UN) Sustainable Development Goals, under the frame of the UN Agenda for Sustainable Development (Transforming our World: The 2030 Agenda for Sustainable Development – UN, 2015): Goal 2: Zero hunger; Goal 6: Clean water and sanitation.

Funding

This work was supported by the Ministry of Science, Technological Development and Innovations of the Republic of Serbia, grants no 451-03-136/2025-03/200178, 451-03-137/2025-03/200178 (Faculty of Biology), 451-03-136/2025-03/200053 (Institute for Multidisciplinary Research), 451-03-136/2025-03/200007 (Institute for Biological Research 'Siniša Stanković' National Institute of the Republic of Serbia, University of Belgrade) and bilateral collaboration project between Serbia (Faculty of Biology, University of Belgrade) and Slovenia (National Institute of Biology), grant no 337-00-110/2023-05/47. The authors also acknowledge funding provided by the Institute of Physics, National Institute of the Republic of Serbia, through a grant by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.wroa.2025.100380.

Data availability

Data will be made available on request

References

Anteljević, M., Rosić, I., Medić, O., Kolarević, S., Berić, T., Stanković, S., Nikolić, I., 2023. Occurrence of plant pathogenic *Pseudomonas syringae* in the Danube River Basin: abundance and diversity assessment. Phytopathol. Res. 5 (1), 18. https://doi.org/10.1186/s42483-023-00174-0.

Balaž, J., Ilićić, R., Maširević, S., Jošić, D., Kojić, S., 2014. First report of *Pseudomonas syringae* pv. *syringae* causing bacterial leaf spots of oil pumpkin (*Cucurbita pepo*) in Serbia. Plant Dis. 98 (5), 684. https://doi.org/10.1094/PDIS-07-13-0714-PDN.

Balthazar, C., Novinscak, A., Cantin, G., Joly, D.L., Filion, M., 2022. Biocontrol activity of *Bacillus* spp. And *Pseudomonas* spp. Against *botrytis cinerea* and other *cannabis* fungal pathogens. Phytopathology 112 (3), 549–560. https://doi.org/10.1094/PHYTO-03-21-0128-R.

Barrett-Manako, K., Andersen, M., Martínez-Sánchez, M., Jenkins, H., Hunter, S., Reese-George, J., Montefiori, M., Wohlers, M., Rikkerink, E., Templeton, M., Nardozza, S., 2021. Real-time PCR and droplet digital PCR are accurate and reliable methods to quantify *Pseudomonas syringae* pv. *actinidiae* biovar 3 in kiwifruit infected plantlets. Plant Dis. 105 (6), 1748–1757. https://doi.org/10.1094/PDIS-08-20-1703-RE.

Berge, O., Monteil, C.L., Bartoli, C., Chandeysson, C., Guilbaud, C., Sands, D.C., Morris, C.E., 2014. A user's guide to a data base of the diversity of *Pseudomonas syringae* and its application to classifying strains in this phylogenetic complex. PLoS ONE 9 (9), e105547. https://doi.org/10.1371/journal.pone.0105547.

Berge, O., Nofal, S., Razan, F., Chandeysson, C., Guilbaud, C., 2023. Groundwater as a reservoir for plant pathogenic bacteria : the case of *Pseudomonas syringae* Complex in the alluvial aquifer of Avignon. Preprints. https://doi.org/10.20944/preprints202306.0217.v1.

Blagojević, B., Langović, M., Novković, I., Dragićević, S., Živković, N., 2020. Introduction to water resources management in Balkan countries. In: Zelenakova, M., Negm, A. M., Minea, I. (Eds.), Water Resources Management in Balkan Countries, pp. 123–247. https://doi.org/10.1007/978-3-030-22468-4_1.

Chai, A.L., Ben, H.Y., Guo, W.T., Shi, Y.X., Xie, X.W., Li, L., Li, B.J., 2020. Quantification of viable cells of *Pseudomonas syringae* pv. *tomato* in tomato seed using propidium monoazide and a real-time PCR assay. Plant Dis. 104 (8), 2225–2232. https://doi.org/10.1094/PDIS-11-19-2397-RE.

Cheng, T.L., Liao, C.C., Tsai, W.H., Lin, C.C., Yeh, C.W., Teng, C.F., Chang, W.T., 2009. Identification and characterization of the mitochondrial targeting sequence and mechanism in human citrate synthase. J. Cell. Biochem. 107 (5), 1002–1015. https://doi.org/10.1002/jcb.22200.

Clinical and Laboratory Standards Institute, 2004. Protocols For Determination of Limits of Detection and Limits of quantitation: Approved guideline (NCCLS Document EP17-A (ISBN 1-56238-551-8).

Corona-López, E., Román-Gutiérrez, A.D., Otazo-Sánchez, E.M., Guzmán-Ortiz, F.A., Acevedo-Sandoval, O.A., 2021. Water–Food Nexus assessment in agriculture: a systematic review. Int. J. Environ. Res. Public Health. 18 (9), 4983. https://doi.org/10.3390/ijerph18094983.

Cottyn, B., Baeyen, S., Pauwelyn, E., Verbaendert, I., De Vos, P., Bleyaert, P., Hofte, M., Maes, M., 2011. Development of a real-time PCR assay for *Pseudomonas cichorii*, the causal agent of midrib rot in greenhouse-grown lettuce, and its detection in irrigating water. Plant Pathol 60 (3), 453–461. https://doi.org/10.1111/j.1365-3059.2010.02388.x.

de Araujo, G.G., Rodrigues, F., Gonçalves, F.L.T., Galante, D., 2019. Survival and ice nucleation activity of *Pseudomonas syring*ae strains exposed to simulated high-altitude atmospheric conditions. Sci. Rep. 9 (1), 1–11. https://doi.org/10.1038/s41598-019-44283-3.

Dinno, A., 2017. dunn. Test: dunn's test of multiple comparisons using rank sums. R package version 1 (5), 1. https://cran.r-project.org/package=dunn.test.

FAO, 2022. FAOSTAT: Crops and livestock products. Accessed on November 2, 2024. https://www.fao.org/faostat/en/#data/QCL.

FAO, 2021. Law on Agricultural Land, Official Gazette of the Republic of Serbia, No. 11/2021. FAOLEX database. Accessed on November 2, 2024. https://faolex.fao.org/docs/pdf/srb195486.pdf.

Guilbaud, C., Morris, C.E., Barakat, M., Ortet, P., Berge, O., 2016. Isolation and identification of *Pseudomonas syringae* facilitated by a PCR targeting the whole *P. syringae* group. FEMS Microbiol. Ecol. 92 (1), 1–9. https://doi.org/10.1093/femsec/fiv146.

Ignjatov, M., Milošević, D., Nikolić, Z., Gvozdanović-Varga, J., Tatić, M., Popović, T., Ivanović, Ž., 2017. The first report was that *fusarium tricinctum* causes rot in garlic

bulbs in Serbia. Plant Dis. 101 (2), 382. https://doi.org/10.1094/PDIS-09-16-1333-PDN.
Ilićić, R., Balaž, J., Ognjanov, V., Popović Milovanović, T., 2021. Epidemiology studies of *Pseudomonas syringae* pathovars associated with bacterial canker on the sweet cherry in Serbia. Plant Protect. Sci. 57 (3), 196–205. https://doi.org/10.17221/140/2020-PPS.
Kokkoris, V., Vukicevich, E., Richards, A., Thomsen, C., Hart, M.M., 2021. Challenges using droplet digital PCR for environmental samples. Appl. Microbiol. 1 (1), 74–88. https://doi.org/10.3390/applmicrobiol1010007.
Krstić Tomić, T., Atanasković, I., Nikolić, I., Joković, N., Stević, T., Stanković, S., Berić, T., Lozo, J., 2023. Culture-dependent and metabarcoding characterization of the sugar beet (*Beta vulgaris* L.) microbiome for high-yield isolation of bacteria with plant growth-promoting traits. Microorganisms 11 (6), 1538. https://doi.org/10.3390/microorganisms11061538.
Lamichhane, J.R., Bartoli, C., 2015. Plant pathogenic bacteria in open irrigation systems: what risk for crop health? Plant Pathol 64 (4), 757–766. https://doi.org/10.1111/ppa.12371.
Letunic, I., Bork, P., 2024. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucl. Acid. Res. 52 (W1), W78–W82. https://doi.org/10.1093/nar/gkae268.
Lipps, S.M., Samac, D.A., 2022. *Pseudomonas viridiflava*: an internal outsider of the *Pseudomonas syringae* species complex. Mol. Plant Pathol. 23 (1), 3–15. https://doi.org/10.1111/mpp.13133.
Marković, S., Milić Komić, S., Jelušić, A., Iličić, R., Bagi, F., Stanković, S., Popović, T., 2022. First report of *pectobacterium versatile* causing blackleg of potato in Serbia. Plant Dis. 106 (1), 312. https://doi.org/10.1094/PDIS-06-21-1128-PDN.
Meier-Kolthoff, J.P., Göker, M., 2019. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. Nat. Commun. 10 (1), 2182. https://doi.org/10.1038/s41467-019-10210-3.
Milanović, A., Milijašević, D., Brankov, J., 2011. Assessment of polluting effects and surface water quality using water pollution index: a case study of Hydro–system Danube–Tisa–Danube, Serbia. Carpath. J. Earth Environ. Sci. 6 (2), 269–277.
Mitrović, M., Cvrković, T., Jović, J., Krstić, O., Jakovljević, M., Kosovac, A., Toševski, I., 2015. First report of 'candidatus *phytoplasma solani* 'infecting garden bean *phaseolus vulgaris* in Serbia. Plant Dis. 99 (4), 551. https://doi.org/10.1094/PDIS-10-14-1052-PDN.
Mohan, S.K., 1987. An improved agar plating assay for detecting *Pseudomonas syringae* pv. *syringae* and *P. s.* pv. *phaseolicola* in contaminated bean seed. Phytopathology 77 (10), 1390. https://doi.org/10.1094/phyto-77-1390.
Morris, C.E., Lacroix, C., Chandeysson, C., Guilbaud, C., Monteil, C., Piry, S., Rochelle Newall, E., Fiorini, S., Van Gijsegem, F., Barny, M.A., Berge, O., 2023. Comparative abundance and diversity of populations of the *Pseudomonas syringae* and Soft Rot *Pectobacteriaceae* species complexes throughout the Durance River catchment from its French Alps sources to its delta. Peer Commun. J. 3, e88, 10.24072/pcjournal.317.
Morris, C.E., Lamichhane, J.R., Nikolić, I., Stanković, S., Moury, B., 2019. The overlapping continuum of host range among strains in the *Pseudomonas syringae* complex. Phytopathol. Res. 1 (1), 1–16. https://doi.org/10.1186/s42483-018-0010-6.
Morris, C.E., Monteil, C.L., Berge, O., 2013. The life history of *Pseudomonas syringae*: linking agriculture to earth system processes. Annu. Rev. Phytopathol. 51, 85–104. https://doi.org/10.1146/annurev-phyto-082712-102402.
Morris, C.E., Ramirez, N., Berge, O., Lacroix, C., Monteil, C., Chandeysson, C., Guilbaud, C., Blischke, A., Sigurbjörnsdóttir, M.A., Vilhelmsson, O., 2022. *Pseudomonas syringae* on plants in Iceland has likely evolved for several million years outside the reach of processes that mix this bacterial complex across earth's temperate zones. Pathogens 11 (3), 357. https://doi.org/10.3390/pathogens11030357.
Morris, C.E., Sands, D.C., Vanneste, J.L., Montarry, J., Oakley, B., Guilbaud, C., Glaux, C., 2010. Inferring the evolutionary history of the plant pathogen *Pseudomonas syringae* from its biogeography in headwaters of rivers in North America, Europe, and New Zealand. MBio 1 (3), e00107–e00110. https://doi.org/10.1128/mBio.00107-10.
Morris, C.E., Sands, D.C., Vinatzer, B.A., Glaux, C., Guilbaud, C., Buffiere, A., Thompson, B.M., 2008. The life history of the plant pathogen *Pseudomonas syringae* is linked to the water cycle. ISME J. 2 (3), 321–334. https://doi.org/10.1038/ismej.2007.113.
Nikolić, I., Stanković, S., Dimkić, I., Berić, T., Stojšin, V., Janse, J., Popović, T., 2018. Genetic diversity and pathogenicity of *Pseudomonas syringae* pv. *aptata* isolated from sugar beet. Plant. Pathol. 67 (5), 1194–1207. https://doi.org/10.1111/ppa.12831.
Official Gazette of the Republic of Serbia No.74/2011 (2011). Regulation on the parameters of ecological and chemical status of surface water and parameters of the chemical and quantitative status of groundwater.
Pirc, M., Ravnikar, M., Tomlinson, J., Dreo, T., 2009. Improved fireblight diagnostics using quantitative real-time PCR detection of *Erwinia amylovora* chromosomal DNA. Plant Pathol. 58 (5), 872–881. https://doi.org/10.1111/j.1365-3059.2009.02083.x.
Popović, T., Balaž, J., Starović, M., Trkulja, N., Ivanović, Ž., Ignjatov, M., Došić, D., 2013. First report of *Xanthomonas campestris* pv. *campestris* as the causal agent of black rot on oilseed rape (*Brassica napus*) in Serbia. Plant Dis. 97 (3), 418. https://doi.org/10.1094/PDIS-05-12-0506-PDN.
Popović, T., Ivanović, Ignjatov, M., 2015c. First report of *Pseudomonas viridiflava* causing pith necrosis of tomato (*Solanum lycopersicum*) in Serbia. Plant Dis. 99 (7), 1033. https://doi.org/10.1094/PDIS-01-15-0052-PDN.
Popović, T., Ivanović, Ž., Ignjatov, M., Milošević, D., 2015a. First report of *Pseudomonas syringae* pv. *coriandricola* causing bacterial leaf spot on carrot, parsley and parsnip in Serbia. Plant Dis. 99 (3), 416. https://doi.org/10.1094/PDIS-10-14-1041-PDN.
Popović, T., Ivanović, Ž., Trkulja, N., Milosavljević, A., Ignjatov, M., 2015b. First report of *Pseudomonas syringae* pv. *syringae* on pea (*Pisum sativum*) in Serbia. Plant Dis. 99 (5), 724. https://doi.org/10.1094/PDIS-11-14-1212-PDN.
R Core Team, 2024. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.Rproject.org/.
Salem, H.S., Pudza, M.Y., Yihdego, Y., 2022. Water strategies and water–food nexus: challenges and opportunities towards sustainable development in various regions of the World. Sustain. Water Resour. Manag. 8 (4), 114. https://doi.org/10.1007/s40899-022-00676-3.
Schneider, C.A., Rasband, W.S., Eliceiri, K.W., 2012. NIH image to ImageJ: 25 years of image analysis. Nat. Method. 9 (7), 671–675. https://doi.org/10.1038/nmeth.2089.
Shi, X., Liu, G., Shi, L., Chen, M., Wu, X., Lu, Y., Zhao, J., 2021. The detection efficiency of digital PCR for the virulence genes of waterborne pathogenic bacteria. Water Supply 21 (5), 2285–2297. https://doi.org/10.2166/ws.2021.056.
Statistical Office of the Republic of Serbia, 2024. Monthly statistical bulletin: February 2024. Accessed on November 2, 2024. https://publikacije.stat.gov.rs/G2024/Pdf/G20242057.pdf.
Stojšin, V., Balaž, J., Budakov, D., Stanković, S., Nikolić, I., Ivanović, Ž., Popović, T., 2015. First report of *Pseudomonas syringae* pv. *aptata* causing bacterial leaf spot on sugar beet in Serbia. Plant Dis. 99 (2), 281. https://doi.org/10.1094/PDIS-06-14-0628-PDN.
Tiwari, A., Ahmed, W., Oikarinen, S., Sherchan, S.P., Heikinheimo, A., Jiang, G., Simpson, S.L., Greaves, J., Bivins, A., 2022. Application of digital PCR for public health-related water quality monitoring. Sci. Total Environ. 837, 155663. https://doi.org/10.1016/j.scitotenv.2022.155663.
Thomsen, M.C.F., Hasman, H., Westh, H., Kaya, H., Lund, O., 2017. RUCS: rapid identification of PCR primers for unique core sequences. Bioinformatics 33 (24), 3917–3921. https://doi.org/10.1093/bioinformatics/btx526.
Venbrux, M., Crauwels, S., Rediers, H., 2023. Current and emerging trends in techniques for plant pathogen detection. Front. Plant Sci. 14, 1–25. https://doi.org/10.3389/fpls.2023.1120968.
Wickham, H., 2016. Data analysis. Title of the Book. Springer International Publishing, pp. 189–201. https://doi.org/10.1007/978-3-319-24277-4_9.
Yang, P., Zhao, L., Gao, Y.G., Xia, Y., 2023. Detection, diagnosis, and preventive management of the bacterial plant pathogen *Pseudomonas syringae*. Plants 12 (9), 1765. https://doi.org/10.3390/plants12091765.

Chapter 3

Discussion

3.1 Targeted High-throughput Sequencing Could Be Used for the Determination of Microbiome and Resistome in Sputum

Targeted high-throughput sequencing (HTS) is a common method for studying bacterial communities. It can be used to determine the microbiome and resistome, or to detect specific bacterial taxa present in a sample. The relatively straightforward interpretation of the results makes this method attractive for use in clinical diagnostics. However, to utilise the full potential of targeted HTS, we must be confident in the reliability of the results obtained by this method. To achieve this, we must first understand how various factors and variables in sample processing can affect the results. One of the most significant of these factors is the method of DNA extraction; therefore, our study focused on the effect it has on the results of microbiome and resistome determination in sputum samples. Although targeted HTS has become a valuable tool for microbiome characterisation, only a limited number of studies have examined how the DNA extraction method influences microbiome composition results, particularly in sputum. Findings from other matrices cannot be readily transferred to sputum due to its complex composition and high degree of heterogeneity. To evaluate these effects, we applied principles from diagnostics and metrology to the study design and interpretation of the results. Those principles included: evaluation of repeatability (expressed as coefficient of variation), evaluation of sensitivity (expressed as limit of detection), and the use of controls. However, they are not yet routinely applied to microbiome analyses performed by targeted HTS. To evaluate repeatability, DNA was extracted from samples in triplicate on two separate days. To evaluate the sensitivity of targeted HTS in detecting bacteria in sputum, the samples were spiked with three bacteria (*Acinetobacter baumannii*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa*) at known concentrations. This differs from previous studies, which used mock bacterial communities, and makes our findings more applicable to the study of real samples [217]–[219]. DNA was extracted from the sputum samples using three methods based on different extraction mechanisms: the first method used CTAB and DNA precipitation, the second used DNA binding to magnetic beads, and the third used silica membranes and spin columns. These three methods are commonly used in research and diagnostic settings, which increases the transferability of the study's results. Another important practice that ensures the reliability of targeted HTS results is the inclusion of negative controls at different stages of the experiment [220]. These include blank controls for DNA extraction and no-template controls to identify potential contaminants that may occur during sample processing, DNA extraction, or library preparation.

Targeted HTS can be used to amplify essentially any known sequence within the genome. For identification purposes, however, some regions within bacterial genomes have proven to be more suitable. The chosen region depends greatly on the taxonomic level at which the bacteria present in the sample are to be identified. To determine the microbiome, multiple amplicons within the 16S rRNA region were sequenced, as this is one of the most

commonly used genetic targets for bacterial taxonomic identification. While this region allows for the identification of basically any bacteria, it often lacks specificity and can only reliably classify bacteria at the genus level. To detect exact species or even subspecies, more specific genomic regions must be used. To this end, species-specific amplicons were sequenced to enable the reliable detection of clinically relevant bacteria. To demonstrate that targeted HTS can be used for purposes other than identifying bacteria, numerous genetic sequences associated with antimicrobial resistance (AMR) were sequenced to determine the resistome in sputum samples. The ability to determine AMR using genotyping methods is important given the widespread occurrence of AMR genes (AMGs). Targeted HTS has great potential in AMR testing due to its ability to simultaneously detect a large number of AMGs that can confer resistance to various antibiotics.

The results showed that targeted HTS of the 16S rRNA region yielded consistent microbiome results. However, the DNA extraction method was found to significantly affect alpha diversity parameters and microbiome composition in these results. These findings are consistent with previous studies, which demonstrated that the DNA extraction method can influence the results of microbiome as determined by 16S rRNA sequencing [221]. Of the three DNA extraction methods tested, the silica membrane-based method produced the highest diversity and richness consistently. The main cause of the differences between the extraction methods was the presence or absence of low-abundance species. One possible reason for this may be the different DNA concentrations yielded by the various methods, since they involve different dilutions of the extracted DNA. The CTAB-based protocol also included enzymatic lysis with lysozyme, resulting in a higher proportion of Gram-positive bacteria, as observed in previous studies [208, 211]. The resistome results reflected those of the microbiome. Targeted HTS produced repeatable results, with the silica membrane-based method again yielding the highest average number of detected AMGs. However, the differences between the methods were less pronounced than in the case of the microbiome. Additionally, species-specific amplicons were sequenced to enable more precise detection of a smaller subsample of clinically relevant bacteria. In contrast to microbiome and resistome determination, the DNA extraction method did not significantly impact the detection limits for bacteria identified using species-specific amplicons. Targeted HTS of species-specific amplicons could detect spiked bacteria in sputum at concentrations as low as 10^4 cps/mL. However, the sensitivity varied between different bacterial species, potentially due to differences in the efficiency of target amplification or the number of amplicons used to detect each species. Although targeted HTS is less sensitive than other PCR-based methods, an important benefit is that it enables the simultaneous detection of a much larger number of different bacteria.

Overall, the results demonstrated the importance of the DNA extraction method for the results of microbiome and resistome determination using targeted HTS. Our study focused on three DNA extraction methods, one of which — the silica membrane-based method — proved to be the most appropriate. However, many other DNA extraction methods can be used in combination with targeted HTS, and choosing the right method is an essential part of study design. To minimise the bias introduced by DNA extraction methods, it is important to evaluate the method before use, particularly in large-scale studies. Applying principles from metrology and diagnostics enabled us to evaluate the use of targeted HTS in determining the microbiome and resistome, as well as in species-specific bacterial detection. The results proved to be repeatable, particularly when the appropriate DNA extraction method was employed. These findings facilitate the transition of this method into its mature phase, in which its results can be reliably applied to a variety of contexts. Reliable results are particularly important in fields such as clinical diagnostics, where targeted HTS still has considerable potential to be explored.

3.2 Application of Comparative Genomics for the Study of Plant Pathogenic Bacteria

Another major application of HTS is whole genome sequencing (WGS), which allows for a more in-depth study of selected bacteria and has become the preferred method for genomic research. This is particularly true in the case of bacteria, whose relatively small genomes make WGS even more accessible. The data generated by WGS can be used for various purposes, such as taxonomic classification of bacteria, phylogenetic analysis, and comparative genomics. These applications also extend to plant pathogenic bacteria, where WGS is increasingly being used to identify and characterise newly isolated strains, monitor the spread of bacterial diseases in crops, and identify possible genetic markers linked to bacterial pathogenicity in plants. In this thesis, we used the principles of WGS and comparative genomics in two studies focusing on different plant-pathogenic bacteria.

The main study focused on *Pantoea stewartii* subsp. *stewartii*, the causative agent of Stewart's wilt in maize. This bacterium is native to North America; however, multiple recent findings of *P. stewartii* subsp. *stewartii* have been reported in Slovenia and elsewhere in Europe. These findings are surprising and concerning, as it is a quarantine bacterium that is not supposed to be permanently present in Europe. Our study aimed to shed light on these recent findings, which are currently limited to the western part of Slovenia, especially the Vipava River Valley. The objectives were to try to examine how this bacterium spread to Slovenia and whether the recent detections were the result of multiple introduction events or the spread of the bacterium within Slovenia. We also aimed to determine how Slovenian isolates differ from *P. stewartii* strains isolated in other geographical areas.

To achieve the aims of our study, we performed WGS of multiple recent Slovenian isolates together with historical isolates from strain collections to obtain high-quality hybrid genome assemblies generated from short and long sequencing reads. Newly obtained genome assemblies were used together with those available in public databases to conduct comparative genomics studies. The *P. stewartii* subsp. *stewartii* genomes used in the study originated from strains isolated in either America, Slovenia, or Italy. These genomes were used for phylogenetic analysis to better understand the relationships between the strains. WGS data were crucial for this analysis due to the high percentage of nucleotide identity between *P. stewartii* subsp. *stewartii* strains (ANI >99.9%), indicating relatively recent evolution and spread. Due to strains being highly related, it was crucial to include only the genomes of high quality, as mistakes in genome assemblies could easily skew the results. The phylogenetic analysis using the core genome revealed that Slovenian isolates form a distinct branch within the phylogenetic tree. The only exception was strain NIB Z 3391, which did not cluster with the other Slovenian isolates. These results suggest at least two separate introduction events. Additionally, strain NIB Z 3388 may be the result of an additional introduction, as it is less closely related to the other Slovenian isolates and forms a separate branch in the phylogenetic tree based on the accessory genome. The Italian isolates showed similar patterns, with all of them except one forming a separate branch within the tree, and did not show a relationship to Slovenian isolates. Together, these

results showed that recent European findings of *P. stewartii* subsp. *stewartii* were the result of multiple introduction events.

We further determined the position of subspecies *stewartii* within the species *Pantoea stewartii*, which also includes the additional subspecies *idologenes*. Previous phylogenetic analyses revealed *P. stewartii* to be a distinct species within the genus *Pantoea*, with *P. ananatis* and *P. allii* being the most closely related species [15], [224], [225]. The exact position of *P. stewartii* subsp. *stewartii* was previously unclear due to a lack of high-quality genomic data; the genome of strain DC283 was the only high-quality complete genome assembly available. We performed a phylogenetic analysis using newly generated genome assemblies, alongside the genomes of *P. stewartii* available in public databases. This revealed that *P. stewartii* strains form two distinct clades in the phylogenetic tree: one containing all strains belonging to subspecies *stewartii*, and the other containing all other strains. The other clade was much more diverse and included all strains belonging to the subspecies *indologenes*, as well as all strains without an assigned subspecies. The strains in this clade also have a wider host range, affecting different plants, whereas subspecies *stewartii* is mostly restricted to causing symptoms in maize. The results of the phylogenetic analysis were also reflected in the pan-genomes of both subspecies and the whole *P. stewartii* species. The pan-genome of the *stewartii* subspecies is smaller, with core genes representing a larger proportion in comparison to the subspecies *indologenes*. *P. stewartii* possesses an open pan-genome which encompasses the diversity of mobile genetic elements, as well as the bacterium's ability to participate in horizontal gene transfer [224].

In fact, despite strains of subspecies *stewartii* being closely related, they exhibited a diversity in mobile genetic elements, including plasmids and prophages. The number of plasmids in the analysed genomes ranged from eight (CFBP 3167^{T}) to 14 (CFBP 3445 and CFBP 3157). Newly sequenced strains also contain two previously unknown plasmids, which were named pPSS06 and pPSS09. The only plasmid also present in other *Pantoea* species is pPPS14, which is also known as the large *Pantoea* plasmid (LPP-1) and plays an important role in pathogenesis [200]. Except for NIB Z 3391, Slovenian isolates have a similar plasmid composition, which is consistent with previous phylogenetic analyses. The presence or absence of different plasmids could affect strain phenotype as they encode pathogenicity factors involved in host colonisation.

We further examined the differences in the presence of pathogenicity factors among two *P. stewartii* subspecies, which in total included 65 different strains. The two major pathogenicity factors typical of the *stewartii* subspecies are the Hrp type III secretion system (T3SS) and the exopolysaccharide stewartan [176], [194]. In addition, we examined the diversity of different secretion systems present in the genomes of *P. stewartii*. Some are present only in strains of the *stewartii* subspecies, including one type II secretion system and two additional T3SSs. One T3SS is involved in the bacteria's ability to colonise the corn flea beetle gut, while the function of the other is not yet fully understood [195]. The vast majority of strains from the *stewartii* subspecies contain all three T3SSs. The limited number of effector proteins suggests that *P. stewartii* acquired the T3SSs relatively recently [194]. In contrast, the subspecies *stewartii* lacks the functional type VI secretion system, which is present in all *indologenes* strains and is believed to be one of the primary pathogenicity mechanisms of this subspecies [226]. The Hrp T3SS is present in strains from

both *P. stewartii* subspecies, and it is encoded on the plasmid pPSS14, which is homologous to the widespread LPP-1 plasmid. This indicates that the Hrp T3SS has an older evolutionary origin. However, some genomes from both subspecies lack this T3SS, suggesting that it was lost during evolution. A pathogenicity test was performed to confirm the effects of the different T3SSs. This test confirmed that strains lacking the Hrp T3SS do not cause the typical symptoms of infection with subspecies *stewartii*. The experiment also included a strain lacking a T3SS with an unclear role, which exhibited milder symptoms than strains with all three T3SS, further confirming the role of this T3SS in pathogenicity. In addition to the secretion systems, other pathogenicity factors were examined. The *wce-I*, *-II,* and *-III* gene clusters, which are involved in stewartan biosynthesis, were present in all analysed *P. stewartii* genomes, indicating these bacteria's ability to produce exopolysaccharides. Furthermore, we confirmed previous findings that *P. stewartii* genomes contain enzymes that metabolise cellulose and are involved in the bacteria's ability to colonise plants [11, 54]. However, additional laboratory testing is required to elucidate the role of these pathogenicity factors in the ability of *P. stewartii* to cause symptoms in maize.

In an additional minor study, WGS data were used to perform a comparative genomics analysis of *Xanthomonas translucens* pv. *undulosa*. In order to expand the geographic coverage of the available genomic data, the first complete genome of *X. translucens* isolated in South America was obtained. The sequenced strain, MAI5034, was identified as *X. translucens* pv. *undulosa* and was isolated in Uruguay in 2016. It has been confirmed to cause symptoms in wheat. The newly obtained genome was used alongside other complete *X. translucens* pv. *undulosa* genomes to perform a comparative genomics study. This focused primarily on the presence of different factors responsible for the pathogenicity and host adaptations of bacteria in the genus *Xanthomonas*. Type III secreted effector proteins play a prominent role in this, with transcription-activator-like effectors (TALEs) being among the most studied [228]. Analysis of the MAI5034 genome revealed seven different TALEs that are widely conserved among all *X. translucens* pv. *undulosa* strains. This supports the recent worldwide spread of *X. translucens* pv. *undulosa*. Pathovar *undulosa* also exhibits a lower diversity of TALEs than other pathovars of *X. translucens*, such as pathovar *translucens*. This is consistent with the lower genetic diversity of pathovar *undulosa*, suggesting that it emerged from pathovar *translucens*. We also participated in an international study, where the recent European findings of *Ralstonia pseudosolanecearum* phylotype I were examined. Results of this study, which is currently under review, also showed multiple introductions of this pathogen in Europe [229].

Overall, the results presented in this thesis demonstrate that data obtained through WGS can be utilised for comparative genomics analyses within the context of plant-pathogenic bacteria. This data can be used to perform phylogenetic analyses, enabling us to establish relationships between different bacterial strains and taxa. Furthermore, this helps us to track the spread of disease and identify introduction events, making it a powerful tool for managing the spread of plant diseases. Additionally, it also enables us to identify the variety of pathogenicity factors present in bacterial genomes, which could be used to study the bacteria's ability to colonise, spread, and cause disease symptoms in plants.

3.3 Genome-informed Approach for the Development of Specific and Sensitive Quantitative Real-time PCR Tests for the Detection of Plant Pathogenic Bacteria

Another important application of HTS data in bacteriology is designing new molecular tests for specific bacterial detection. Molecular methods, which mostly involve amplifying the target DNA sequence using PCR, are popular in diagnostics due to their ease of use and relatively short turnaround time. One of the most widely used PCR-based methods is quantitative real-time PCR (qPCR) as it encompasses all the aforementioned advantages. Further benefits include eliminating the need for electrophoresis, which reduces the risk of contamination and makes the method less sensitive to inhibitors present in plant tissues. Assays developed for qPCR can be adapted for use as digital PCR (dPCR), enabling absolute quantification. Both methods, qPCR and dPCR, play a crucial role in the detection of plant-pathogenic bacteria, which is of immense importance for disease management. Regular testing helps us to detect infected plants and mitigate the spread of disease via international trade of infected planting material.

A particular advantage of qPCR is in detecting slow-growing bacteria such as *Xylophilus ampelinus*, the causative agent of bacterial blight in grapevines. The fastidious growth of this bacterium under laboratory conditions makes traditional culture-based microbiological methods challenging and time-consuming. Although a few molecular tests for detecting *X. ampelinus* have been developed, conflicting results have been reported when different methods are used. As false positive or negative results could have serious repercussions, there was a growing need for a more reliable detection method. To this end, we employed a genome-informed approach to design a panel of new qPCR tests for the specific and sensitive detection of *X. ampelinus*. To achieve this, we first identified genomic sequences unique to *X. ampelinus* using a pipeline based on the open-source bioinformatic tool RUCS [230]. This identifies DNA sequences present in the target group (positive) of genomes but absent in the non-target group (negative). Ideally, only high-quality genomes should be included in the target group; however, we are often limited by the availability of genomes in public databases. Additionally, correct taxonomic classification is crucial in determining target and non-target genomes. To ensure that only correctly classified genomes are used in test design, phylogenetic analysis should be performed prior to the inclusion of genomes in either group. All 13 *Xylophilus* genomes available in public databases at the time were included in the initial phylogenetic analysis, but some were later removed after being found misclassified; only one genome proved to actually belong to *X. ampelinus*, while all others belonged to related species within the same genus. Despite this limitation, we were able to *in silico* design 18 new qPCR tests utilising a hydrolysis probe. These were then evaluated in laboratory testing to determine their amplification efficiency, analytical specificity, and sensitivity.

Eight of the new tests could detect *X. ampelinus* without generating any false negative results, and five of these were selected for further evaluation. It is somewhat surprising that a significant proportion of the new tests yielded false negative results, given that *X. ampelinus* is considered a highly homogeneous species [207]. This could be explained by the fact that only one *X. ampelinus* genome was used to determine the unique core sequences. Consequently, the core genome could not be determined, and some shell and cloud genes were possibly also included in the newly designed tests. Whenever a new test is evaluated, it is crucial to compare its performance with that of existing tests that have been in use for some time and are well characterised. To this end, all results obtained by the new tests were compared to the existing test by Dreo et al., which served as a

reference [231]. All tests selected for further testing showed similar or better performance than the reference test. Existing tests for the detection of *X. ampelinus* produced conflicting results, particularly when used on matrices for which they had not been validated. To reduce the risk of such discrepant results, the new test should be evaluated using a variety of different testing materials that are as similar as possible to those that would be used in an actual diagnostic setting. As different parts of the grapevine plant can be used for testing, we decided to evaluate the new tests using leaf, xylem, and root samples. Based on these results, one of the new tests was excluded from further use, as inhibition was observed in the xylem and root samples. Four newly designed tests and a reference test by Dreo et al. showed acceptable amplification efficiency across all matrices, with samples from leaves showing slightly lower efficiency. These tests also proved to be highly sensitive across different matrices, detecting *X. ampelinus* at concentrations as low as 10^2 cps/mL. For a test to be fully implemented in diagnostics, it must be able to reliably detect the target bacterium in a variety of laboratory settings, which is often described by the robustness and reproducibility of the tests. To this end, we further validated the new tests by assessing these two parameters and examining the performance of the tests in multiple laboratories. The results confirmed the findings of the initial evaluation performed in our laboratory, with qPCR tests demonstrating high specificity and sensitivity. The results also confirmed that leaves are the most problematic sample type for detecting *X. ampelinus*. One additional new test was excluded at this stage as it demonstrated the lowest sensitivity, particularly in xylem and leaf samples, thus leaving us with three functional new tests.

Laboratory validation proved that a qPCR test designed using a genome-informed approach could detect *X. ampelinus* in different samples with high specificity and sensitivity. The ability of the test to detect bacteria at low concentrations opens the door to possible testing for latent infections. This is of significant importance due to the prolonged incubation periods associated with slow-growing bacteria such as *X. ampelinus*. One of five new tests that were evaluated in different matrices exhibited the highest diagnostic sensitivity and is recommended for screening *X. ampelinus* infections. The additional two new tests and reference test by Dreo et al. also demonstrated high sensitivity and specificity and should be used to confirm the presence of *X. ampelinus* when positive results are obtained in screening.

An additional study used a genome-informed approach to design a qPCR test for detecting *Pseudomonas syringae* phylogroup 2. Unlike the *Xylophilus* genus, *Pseudomonas* is one of the most extensively studied bacterial genera, with the species *P. syringae* being a well-studied plant pathogen. This resulted in a much greater abundance of publicly available genomic data; however, only a few of these belonged to pathovar *aptata*. The initial aim was to design a test that could specifically detect *P. syringae* pv. *aptata*, which infects sugar beet. However, the definition of pathovars is based on the ability to cause symptoms in certain plants rather than on genetic characteristics. Since strains included in the same pathovar often do not form a monophyletic branch in a phylogenetic tree, which makes designing tests that are specific to certain pathovars is extremely difficult, and sometimes impossible. All available genomes belonging to *P. syringae* were used in the *in silico* design, with genomes belonging to the pathovar *aptata* serving as the set of positives that we wanted to detect. Four new qPCR tests were designed in total, and their specificity was evaluated on different strains of *P. syringae*. Despite the tests being designed *in silico* to specifically detect pathovar *aptata*, their specificity expanded to include pathovar 2 as a whole after the initial evaluation of analytical specificity on different strains of *P. syringae*. The sensitivity of the tests was evaluated in plant extracts from sugar beets, and the results showed the superior performance of one of the newly designed tests. This

test was also adapted for use in dPCR and evaluated using water samples from irrigation channels. Such irrigation systems are a common route of transmission for plant pathogenic bacterial infections, including those caused by *P. syringae*, which has previously been isolated from water samples. Testing of irrigation water showed that the selected newly designed test outperformed other culture-based detection methods, especially when used as dPCR. These results demonstrated that the newly designed test could be used to specifically and sensitively detect *P. syringae* phylogroup 2 strains in both plant and water samples.

Overall, the genome-informed approach to designing new PCR-based detection methods has proven capable of producing specific and sensitive tests for detecting plant-pathogenic bacteria in a variety of samples. The generic design pipeline can be applied to basically any bacterium for which genomic data is available. However, no matter how advanced the *in silico* design is, practical laboratory testing remains a crucial part of the validation process for any new test before it can be used in routine diagnostic testing.

3.4 Different Applications of HTS in Bacteriology, Current State of the Art

The scope of this thesis was to examine the various applications of HTS in bacteriology, ranging from the study of bacterial communities using metagenomic sequencing to comparative genomics studies, the results of which could be applicable in diagnostic settings. Since its inception, the field of metagenomics has been inextricably linked to advances in DNA sequencing technologies. Over the years, many different sequencing approaches have been developed for metagenomic studies. Among these, targeted HTS has gained special prominence in determining microbiomes and resistomes. However, some challenges in the application of this technology remain, especially in fields such as diagnostics, which require an exceptionally high degree of reliability. Although targeted HTS is a generic method that enables the study of almost any biological environment or sample, the methods are not always transferable due to the unique characteristics of each sample type. In our study, we focused on applying targeted HTS to determine the microbiome in sputum samples. We demonstrated that targeted HTS can generate reliable results, particularly when the appropriate DNA extraction method is employed in the processing of samples. This serves as an important contribution to the development and implementation of targeted HTS in diagnostics.

Having studied whole bacterial communities, we then focused on a more in-depth study of a target bacterial taxon of interest. The primary HTS approach in this field is WGS, the results of which are usually employed for comparative genomics studies. Recent advances in sequencing technologies have increased the accuracy of the generated sequence while also lowering the cost. Consequently, more genomic sequences are now available in public databases, facilitating additional studies and accelerating downstream research. These studies aim to establish relationships between different bacterial strains and taxa, which could help us to monitor the spread of bacterial diseases. Furthermore, we can identify the genetic determinants of pathogenicity and link specific genetic markers to particular phenotypes. In our research, we focused on plant-pathogenic bacteria, which can cause serious yield losses and are of special interest. Our primary focus was on *P. stewartii* subsp. *stewartii*, the causative agent of Stewart's wilt in maize. Despite its quarantine status, this bacterium has recently been isolated from multiple maize samples in Slovenia and Italy. Our study identified multiple introduction events of this bacterium into Europe

and showed that the subspecies *stewartii* forms a distinct branch within the species *P. stewartii*. Despite the close relationship between strains of the subspecies *stewartii*, we also identified diversity in various pathogenicity factors and mobile genetic elements across the genomes.

Genomic data obtained by HTS stored in databases like GenBank could also be used for the genome-informed development of new molecular tests capable of specifically detecting target bacteria. Such tests, which detect genetic sequences that are unique to the target bacterial taxa, have replaced classical culture-based methods in many applications in recent years. They offer many advantages, most notably the relatively short time required to obtain results and the simplicity of the procedure. In our research, we used a genome-informed approach to design qPCR tests that can specifically detect the plant-pathogenic bacteria *X. ampelinus* and *P. syringae* phylogroup 2. Crucially, the performance of the newly designed tests was evaluated in the laboratory to prove their suitability for further use.

In conclusion, the examples described in this thesis merely scratch the surface of the possible applications of HTS in bacteriology. However, they illustrate the range of applications of this technology, which will undoubtedly continue to play an important role in expanding our knowledge of the elusive world of bacteria.

Chapter 4

Conclusions

The bacterial world is immensely complex, and we are still far from completely understanding it. However, HTS represents a significant breakthrough that sheds light on its previously unknown corners. There are many applications of HTS in the study of bacteria, ranging from community-level analyses to targeted applications in clinical diagnostics and agronomy. One of the most frequently used applications of HTS is studying bacterial communities using targeted HTS; however, the results often lack reliability, which limits their potential use. In our study, we proved that the method of DNA extraction significantly affects the results of sputum microbiome and resistome as determined by targeted HTS, thus confirming the first hypothesis. This finding also showcases the importance of intelligent experimental design, in which the choice of DNA extraction method plays a significant role. Furthermore, we demonstrated that the microbiome and resistome results obtained were repeatable, particularly when the appropriate DNA extraction method was employed. This confirmed the second hypothesis and further proved that targeted HTS can be used reliably to study bacterial communities, even in complex samples such as sputum. We then moved on to the next stage of our research, focusing on the application of HTS to perform deeper studies of target bacteria of interest. To this end, we performed WGS and comparative genomics studies of *Pantoea stewartii* subsp. *stewartii*. In our research, we discovered that this maize pathogen, which was recently identified in Slovenia, forms a distinct branch within the broader *P. stewartii* species. While the strains of *P. stewartii* subsp. *stewartii* are closely related; they can differ substantially in the mobile genetic elements they contain, which can affect their pathogenicity in maize. Phylogenetic analysis revealed that all Slovenian isolates except one group together, indicating at least two separate introduction events. Slovenian isolates did not show to be related to Italian isolates, indicating multiple introductions in Europe. However, further studies are needed to better understand the status of this bacterium in Slovenia. Altogether, these findings confirmed our third hypothesis. In the final part of our research, we examined how data generated by WGS could be used to design new, specific genetic tests for detecting target bacteria. Using publicly available genomic data, we designed new qPCR tests that can specifically and sensitively detect the grapevine pathogen *Xylophilus ampelinus*. Laboratory testing proved that three of the newly designed tests could specifically detect *X. ampelinus* in different grapevine samples, including leaves, roots, and xylem. Thus, we confirmed our fourth and final hypothesis. The three major application areas that we examined in this thesis explored how HTS could be used in the field of bacteriology and helped us to answer the research questions we identified at the outset. This will facilitate the further development of HTS, which is reaching the stage of a mature technology with a wide range of applications.

Appendix A

A.1 Supplementary Material for Publication 2.1

Supplementary material is available at:
https://journals.asm.org/doi/suppl/10.1128/msystems.00735-24/suppl_file/msystems.00735-24-s0001.docx. Accessed: 21-May-2025

Benčič, A., Toplak, N., Koren, S., Bogožalec Košir, A., Milavec, M., Tomič, V., Lužnik, D., Dreo, T. (2024). Metrological evaluation of DNA extraction method effects on the bacterial microbiome and resistome in sputum. mSystems, 2024, Vol. 9, Iss. 9, Art. No. e00735-24, https://doi.org/10.1128/msystems.00735-24

A.2 Supplementary Material for Publication 2.2

Supplementary material is available at:
https://www.biorxiv.org/content/10.1101/2025.11.07.687269v1.supplementary-material. Accessed: 8-November-2025

Benčič, A., Pothier, J., Dreo, T. (2025). Comparative genomics and pathogenicity of *Pantoea stewartii* subsp. *stewartii* reveal multiple introductions and limited distribution in Europe. bioRxiv, 2025.11.07.687269. https://doi.org/10.1101/2025.11.07.687269

A.3 Supplementary Material for Publication 2.2

Supplementary material is available at:
https://plantmethods.biomedcentral.com/articles/10.1186/s13007-025-01422-4. Accessed: 18-July-2025

Benčič, A., Bogožalec Košir, A., Matičič, J., Pirc, M., Turnšek, N., Dreo, T. (2025). Development of a multi-targeted real-time PCR assay for the detection of the grapevine pathogen *Xylophilus ampelinus*. Plant Methods, 2025, Vol. 21, Art. No. 99, https://doi.org/10.1186/s13007-025-01422-4

A.4 Supplementary Material for Publication 2.5

Supplementary material is available at:
https://www.sciencedirect.com/science/article/pii/S2589914725000799?via%3Dihub. Accessed: 5-November-2025

Anteljević, M., Rosić, I., Medić, O., Ranković, T., Sunjog, K., Kračun-Kolarević, M., Kolarević, S., Dreo, T., Benčič, A., Berić, T., Stanković, S., Nikolić, I. (2025). 2.4 Irrigation systems as reservoirs of diverse and pathogenic *Pseudomonas syringae strains* endangering crop health. Water Research X, 2025, Vol. 28, Art. No. 100380. https://doi.org/10.1016/j.wroa.2025.100380

References

[1] J. G. Caporaso *et al.*, "Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms," *ISME J. 2012 68*, vol. 6, no. 8, pp. 1621–1624, Mar. 2012.

[2] N. J. Loman *et al.*, "High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity," *Nat. Rev. Microbiol. 2012 109*, vol. 10, no. 9, pp. 599–606, Aug. 2012.

[3] J. M. Di Bella, Y. Bao, G. B. Gloor, J. P. Burton, and G. Reid, "High throughput sequencing methods and analysis for microbiome research," *J. Microbiol. Methods*, vol. 95, no. 3, pp. 401–414, Dec. 2013.

[4] L. Bragg and G. W. Tyson, "Metagenomics Using Next-Generation Sequencing," *Methods Mol. Biol.*, vol. 1096, pp. 183–201, 2014.

[5] A. S. Mustafa, "Whole Genome Sequencing: Applications in Clinical Bacteriology," *Med. Princ. Pract.*, vol. 33, no. 3, pp. 185–197, Jun. 2024.

[6] T. Hu, N. Chitnis, D. Monos, and A. Dinh, "Next-generation sequencing technologies: An overview," *Hum. Immunol.*, vol. 82, no. 11, pp. 801–811, Nov. 2021.

[7] A. M. Wenger *et al.*, "Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome," *Nat. Biotechnol. 2019 3710*, vol. 37, no. 10, pp. 1155–1162, Aug. 2019.

[8] C. Delahaye and J. Nicolas, "Sequencing DNA with nanopores: Troubles and biases," *PLoS One*, vol. 16, no. 10, p. e0257521, Oct. 2021.

[9] D. A. Benson *et al.*, "GenBank," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D36–D42, Jan. 2013.

[10] R. Leinonen *et al.*, "The European Nucleotide Archive," *Nucleic Acids Res.*, vol. 39, no. suppl_1, pp. D28–D31, Jan. 2011.

[11] O. Ogasawara *et al.*, "DDBJ new system and service refactoring," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D25–D29, Jan. 2013.

[12] J. Chen and G. Coppola, "Bioinformatics and genomic databases," *Handb. Clin. Neurol.*, vol. 147, pp. 75–92, Jan. 2018.

[13] R. H. Smith, L. Glendinning, A. W. Walker, and M. Watson, "Investigating the impact of database choice on the accuracy of metagenomic read classification for the rumen microbiome," *Anim. Microbiome*, vol. 4, no. 1, pp. 1–17, Dec. 2022.

[14] M. Tokuda and M. Shintani, "Microbial evolution through horizontal gene transfer by mobile genetic elements," *Microb. Biotechnol.*, vol. 17, no. 1, Jan. 2024.

[15] P. de Maayer *et al.*, "Phylogenomic, pan-genomic, pathogenomic and evolutionary genomic insights into the agronomically relevant enterobacteria *Pantoea ananatis*

and *Pantoea stewartii*," *Front. Microbiol.*, vol. 8, no. SEP, p. 287061, Sep. 2017.

[16] T. S. Crofts, A. J. Gasparrini, and G. Dantas, "Next-generation approaches to understand and combat the antibiotic resistome," *Nat. Rev. Microbiol. 2017 157*, vol. 15, no. 7, pp. 422–434, Apr. 2017.

[17] O. Israeli *et al.*, "A Rapid high-throughput sequencing-based Approach for the Identification of Unknown Bacterial Pathogens in Whole Blood," *Futur. Sci. OA*, vol. 6, no. 6, Jul. 2020.

[18] M. Gołębiewski and A. Tretyn, "Generating amplicon reads for microbial community assessment with next-generation sequencing," *J. Appl. Microbiol.*, vol. 128, no. 2, pp. 330–354, Feb. 2020.

[19] L. Y. Ballester, R. Luthra, R. Kanagal-Shamanna, and R. R. Singh, "Advances in clinical next-generation sequencing: target enrichment and sequencing technologies," *Expert Rev. Mol. Diagn.*, vol. 16, no. 3, pp. 357–372, Mar. 2016.

[20] W. Gu *et al.*, "Depletion of Abundant Sequences by Hybridization (DASH): Using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications," *Genome Biol.*, vol. 17, no. 1, pp. 1–13, Mar. 2016.

[21] V. Grossmann *et al.*, "Robustness of Amplicon Deep Sequencing Underlines Its Utility in Clinical Applications," *J. Mol. Diagnostics*, vol. 15, no. 4, pp. 473–484, Jul. 2013.

[22] V. Guernier-Cambert, A. Chamings, F. Collier, and S. Alexandersen, "Diverse Bacterial Resistance Genes Detected in Fecal Samples From Clinically Healthy Women and Infants in Australia—A Descriptive Pilot Study," *Front. Microbiol.*, vol. 12, p. 596984, Sep. 2021.

[23] J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman, "Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products," *Chem. Biol.*, vol. 5, no. 10, pp. R245–R249, Oct. 1998.

[24] Z. Wang *et al.*, "Altered diversity and composition of the gut microbiome in patients with cervical cancer," *AMB Express*, vol. 9, no. 1, 2019.

[25] E. Caselli *et al.*, "Defining the oral microbiome by whole-genome sequencing and resistome analysis: The complexity of the healthy picture," *BMC Microbiol.*, vol. 20, no. 1, pp. 1–19, May 2020.

[26] S. A. Whiteside, H. Razvi, S. Dave, G. Reid, and J. P. Burton, "The microbiome of the urinary tract—a role beyond infection," *Nat. Rev. Urol. 2015 122*, vol. 12, no. 2, pp. 81–90, Jan. 2015.

[27] B. Dickson, Robert; Erb-Downward, Jonhn; Martines, Fernando; Gaby, "The Microbiome and the Respiratory Tract," *Annu Rev Physiol*, vol. 78, pp. 481–504, 2017.

[28] J. G. Kers and E. Saccenti, "The Power of Microbiome Studies: Some Considerations on Which Alpha and Beta Metrics to Use and How to Report Results," *Front. Microbiol.*, vol. 12, p. 796025, Mar. 2022.

[29] J. M. Janda and S. L. Abbott, "16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls," *J. Clin. Microbiol.*, vol. 45, no. 9, pp. 2761–2764, Sep. 2007.

[30] P. Yarza *et al.*, "Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences," *Nat. Rev. Microbiol. 2014 129*, vol. 12, no. 9, pp. 635–645, Aug. 2014.

[31] N. P. Nguyen, T. Warnow, M. Pop, and B. White, "A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity," *npj Biofilms Microbiomes 2016 21*, vol. 2, no. 1, pp. 16004-, Apr. 2016.

[32] B. Paul, K. Kavia Raj, T. S. Murali, and K. Satyamoorthy, "Species-specific genomic sequences for classification of bacteria," *Comput. Biol. Med.*, vol. 123, p. 103874, Aug. 2020.

[33] R. D. Bjerre, L. W. Hugerth, F. Boulund, M. Seifert, J. D. Johansen, and L. Engstrand, "Effects of sampling strategy and DNA extraction on human skin microbiome investigations," *Sci. Rep.*, vol. 9, no. 1, pp. 1–11, Dec. 2019.

[34] F. Teng *et al.*, "Impact of DNA extraction method and targeted 16S-rRNA hypervariable region on oral microbiota profiling," *Sci. Rep.*, vol. 8, no. 1, pp. 1–12, Dec. 2018.

[35] M. Oriano *et al.*, "Comparison of different conditions for DNA extraction in sputum - A pilot study," *Multidiscip. Respir. Med.*, vol. 14, no. 1, pp. 1–8, 2019.

[36] P. I. Costea *et al.*, "Towards standards for human fecal sample processing in metagenomic studies," *Nat. Biotechnol.*, vol. 35, no. 11, pp. 1069–1076, Oct. 2017.

[37] M. Y. Lim, E. J. Song, S. H. Kim, J. Lee, and Y. Do Nam, "Comparison of DNA extraction methods for human gut microbial community profiling," *Syst. Appl. Microbiol.*, vol. 41, no. 2, pp. 151–157, Mar. 2018.

[38] R. D. Fleischmann *et al.*, "Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd," *Science (80-.).*, vol. 269, no. 5223, pp. 496–512, 1995.

[39] "Genome - NCBI - NLM." [Online]. Available: https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=2&release_year=1980:2024. [Accessed: 12-Mar-2025].

[40] S. J. Salipante, D. J. SenGupta, L. A. Cummings, T. A. Land, D. R. Hoogestraat, and B. T. Cookson, "Application of whole-genome sequencing for bacterial strain typing in molecular epidemiology," *J. Clin. Microbiol.*, vol. 53, no. 4, pp. 1072–1079, Apr. 2015.

[41] L. Uelze *et al.*, "Typing methods based on whole genome sequencing data," *One Heal. Outlook 2020 21*, vol. 2, no. 1, pp. 1–19, Feb. 2020.

[42] J. C. Kwong, N. Mccallum, V. Sintchenko, and B. P. Howden, "Whole genome sequencing in clinical and public health microbiology," *Pathology*, vol. 47, no. 3, pp. 199–210, Apr. 2015.

[43] M. Salanoubat *et al.*, "Genome sequence of the plant pathogen *Ralstonia solanacearum*," *Nat. 2002 4156871*, vol. 415, no. 6871, pp. 497–502, Jan. 2002.

[44] A. F. W. Mikhail *et al.*, "Utility of whole-genome sequencing during an investigation of multiple foodborne outbreaks of *Shigella sonnei*," *Epidemiol. Infect.*, vol. 149, p. e71, 2021.

[45] G. N. Tenea and P. Hurtado, "Next-Generation Sequencing for Whole-Genome Characterization of *Weissella cibaria* UTNGt21O Strain Originated From Wild *Solanum quitoense* Lam. Fruits: An Atlas of Metabolites With Biotechnological

Significance," *Front. Microbiol.*, vol. 12, p. 675002, Jun. 2021.

[46] K. Beilsmith, M. P. M. Thoen, B. Brachi, A. D. Gloss, M. H. Khan, and J. Bergelson, "Genome-wide association studies on the phyllosphere microbiome: Embracing complexity in host–microbe interactions," *Plant J.*, vol. 97, no. 1, pp. 164–181, Jan. 2019.

[47] D. Chivian *et al.*, "Metagenome-assembled genome extraction and analysis from microbiomes using KBase," *Nat. Protoc. 2022 181*, vol. 18, no. 1, pp. 208–238, Nov. 2022.

[48] S. Andrews, "Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data." [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. [Accessed: 17-Jul-2024].

[49] W. De Coster, S. D'Hert, D. T. Schultz, M. Cruts, and C. Van Broeckhoven, "NanoPack: visualizing and processing long-read sequencing data," *Bioinformatics*, vol. 34, no. 15, pp. 2666–2669, Aug. 2018.

[50] A. Tritt, J. A. Eisen, M. T. Facciotti, and A. E. Darling, "An Integrated Pipeline for de Novo Assembly of Microbial Genomes," *PLoS One*, vol. 7, no. 9, p. e42304, Sep. 2012.

[51] R. Guo, A. Papanicolaou, and M. L. Fritz, "Validation of reference-assisted assembly using existing and novel Heliothine genomes," *Genomics*, vol. 114, no. 5, p. 110441, Sep. 2022.

[52] P. E. C. Compeau, P. A. Pevzner, and G. Tesler, "Why are de Bruijn graphs useful for genome assembly?," *Nat. Biotechnol.*, vol. 29, no. 11, p. 987, Nov. 2011.

[53] A. Bankevich *et al.*, "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing," *J. Comput. Biol.*, vol. 19, no. 5, p. 455, May 2012.

[54] A. Souvorov, R. Agarwala, and D. J. Lipman, "SKESA: strategic k-mer extension for scrupulous assemblies," *Genome Biol. 2018 191*, vol. 19, no. 1, pp. 153-, Oct. 2018.

[55] M. Kolmogorov, J. Yuan, Y. Lin, and P. A. Pevzner, "Assembly of long, error-prone reads using repeat graphs," *Nat. Biotechnol. 2019 375*, vol. 37, no. 5, pp. 540–546, Apr. 2019.

[56] S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy, "Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation," *Genome Res.*, vol. 27, no. 5, pp. 722–736, May 2017.

[57] H. Li, "Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences," *Bioinformatics*, vol. 32, no. 14, pp. 2103–2110, Jul. 2016.

[58] R. R. Wick, L. M. Judd, C. L. Gorrie, and K. E. Holt, "Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads," *PLOS Comput. Biol.*, vol. 13, no. 6, p. e1005595, Jun. 2017.

[59] A. V. Zimin, G. Marçais, D. Puiu, M. Roberts, S. L. Salzberg, and J. A. Yorke, "The MaSuRCA genome assembler," *Bioinformatics*, vol. 29, no. 21, pp. 2669–2677, Nov. 2013.

[60] B. J. Walker *et al.*, "Pilon: An Integrated Tool for Comprehensive Microbial Variant

Detection and Genome Assembly Improvement," *PLoS One*, vol. 9, no. 11, p. e112963, Nov. 2014.

[61] R. Vaser, I. Sović, S. Sović, N. Nagarajan, M. Šikić1, and Š. Šikić1, "Fast and accurate de novo genome assembly from long uncorrected reads," 2017.

[62] M. Manni, M. R. Berkeley, M. Seppey, and E. M. Zdobnov, "BUSCO: Assessing Genomic Data Quality and Beyond," *Curr. Protoc.*, vol. 1, no. 12, p. e323, Dec. 2021.

[63] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, "CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes," *Genome Res.*, vol. 25, no. 7, pp. 1043–1055, Jul. 2015.

[64] N. Beckloff, S. Starkenburg, T. Freitas, and P. Chain, "Bacterial genome annotation," *Methods Mol. Biol.*, vol. 881, pp. 471–503, 2012.

[65] T. Tatusova *et al.*, "NCBI prokaryotic genome annotation pipeline," *Nucleic Acids Res.*, vol. 44, no. 14, pp. 6614–6624, Aug. 2016.

[66] T. Seemann, "Prokka: rapid prokaryotic genome annotation," *Bioinformatics*, vol. 30, no. 14, pp. 2068–2069, Jul. 2014.

[67] O. Schwengers, L. Jelonek, M. A. Dieckmann, S. Beyvers, J. Blom, and A. Goesmann, "Bakta: Rapid and standardized annotation of bacterial genomes via alignment-free sequence identification," *Microb. Genomics*, vol. 7, no. 11, p. 000685, Nov. 2021.

[68] C. P. Cantalapiedra, A. Hernąndez-Plaza, I. Letunic, P. Bork, and J. Huerta-Cepas, "eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale," *Mol. Biol. Evol.*, vol. 38, no. 12, pp. 5825–5829, Dec. 2021.

[69] B. Buchfink, K. Reuter, and H. G. Drost, "Sensitive protein alignments at tree-of-life scale using DIAMOND," *Nat. Methods 2021 184*, vol. 18, no. 4, pp. 366–368, Apr. 2021.

[70] J. Ahrenfeldt, C. Skaarup, H. Hasman, A. G. Pedersen, F. M. Aarestrup, and O. Lund, "Bacterial whole genome-based phylogeny: Construction of a new benchmarking dataset and assessment of some existing methods," *BMC Genomics*, vol. 18, no. 1, pp. 1–13, Jan. 2017.

[71] T. Seemann, "snippy: fast bacterial variant calling from NGS reads," 2015. [Online]. Available: https://github.com/tseemann/snippy. [Accessed: 18-Jul-2024].

[72] R. S. Kaas, P. Leekitcharoenphon, F. M. Aarestrup, and O. Lund, "Solving the Problem of Comparing Whole Bacterial Genomes across Different Sequencing Platforms," *PLoS One*, vol. 9, no. 8, p. e104984, Aug. 2014.

[73] K. Katoh, K. Misawa, K. I. Kuma, and T. Miyata, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform," *Nucleic Acids Res.*, vol. 30, no. 14, pp. 3059–3066, Jul. 2002.

[74] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows–Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, Jul. 2009.

[75] A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, and A. Stamatakis, "RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference," *Bioinformatics*, vol. 35, no. 21, pp. 4453–4455, Nov. 2019.

[76] L. T. Nguyen, H. A. Schmidt, A. Von Haeseler, and B. Q. Minh, "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies," *Mol. Biol. Evol.*, vol. 32, no. 1, pp. 268–274, Jan. 2015.

[77] L. Pritchard, R. H. Glover, S. Humphris, J. G. Elphinstone, and I. K. Toth, "Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens," *Anal. Methods*, vol. 8, no. 1, pp. 12–24, Dec. 2015.

[78] C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, and S. Aluru, "High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries," *Nat. Commun. 2018 91*, vol. 9, no. 1, pp. 1–8, Nov. 2018.

[79] V. Lefort, R. Desper, and O. Gascuel, "FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program," *Mol. Biol. Evol.*, vol. 32, no. 10, pp. 2798–2800, Oct. 2015.

[80] "PHYLIP Home Page." [Online]. Available: https://phylipweb.github.io/phylip/. [Accessed: 17-Mar-2025].

[81] H. Tettelin *et al.*, "Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial 'pan-genome,'" *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 39, pp. 13950–13955, Sep. 2005.

[82] F. Rosconi *et al.*, "A bacterial pan-genome makes gene essentiality strain-dependent and evolvable," *Nat. Microbiol. 2022 710*, vol. 7, no. 10, pp. 1580–1592, Sep. 2022.

[83] F. J. Whelan, R. J. Hall, and J. O. McInerney, "Evidence for Selection in the Abundant Accessory Gene Content of a Prokaryote Pangenome," *Mol. Biol. Evol.*, vol. 38, no. 9, pp. 3697–3708, Aug. 2021.

[84] Y. Zhai and C. Wei, "Open pangenome of *Lactococcus lactis* generated by a combination of metagenome-assembled genomes and isolate genomes," *Front. Microbiol.*, vol. 13, p. 948138, Aug. 2022.

[85] X. Argemi *et al.*, "Comparative genomic analysis of *Staphylococcus lugdunensis* shows a closed pan-genome and multiple barriers to horizontal gene transfer," *BMC Genomics*, vol. 19, no. 1, pp. 1–16, Aug. 2018.

[86] A. J. Page *et al.*, "Roary: rapid large-scale prokaryote pan genome analysis," *Bioinformatics*, vol. 31, no. 22, pp. 3691–3693, Nov. 2015.

[87] A. M. Eren *et al.*, "Community-led, integrated, reproducible multi-omics with anvi'o," *Nat. Microbiol. 2020 61*, vol. 6, no. 1, pp. 3–6, Dec. 2020.

[88] P. Vinuesa and B. Contreras-Moreira, "Robust Identification of Orthologues and Paralogues for Microbial Pan-Genomics Using GET_HOMOLOGUES: A Case Study of pIncA/C Plasmids," *Methods Mol. Biol.*, vol. 1231, pp. 203–232, 2015.

[89] A. J. Weisberg and J. H. Chang, "Mobile Genetic Element Flexibility as an Underlying Principle to Bacterial Evolution," *Annu. Rev. Microbiol.*, vol. 77, no. Volume 77, 2023, pp. 603–624, Sep. 2023.

[90] A. Guidot, B. Coupat, S. Fall, P. Prior, and F. Bertolla, "Horizontal gene transfer between *Ralstonia solanacearum* strains detected by comparative genomic hybridization on microarrays," *ISME J.*, vol. 3, no. 5, pp. 549–562, May 2009.

[91] C. Bertelli *et al.*, "IslandViewer 4: Expanded prediction of genomic islands for larger-scale datasets," *Nucleic Acids Res.*, vol. 45, no. W1, pp. W30–W35, Jul. 2017.

[92] D. S. Wishart *et al.*, "PHASTEST: faster than PHASTER, better than PHAST,"

Nucleic Acids Res., vol. 51, no. W1, pp. W443–W450, Jul. 2023.

[93] W. Song, B. Wemheuer, S. Zhang, K. Steensen, and T. Thomas, "MetaCHIP: Community-level horizontal gene transfer identification through the combination of best-match and phylogenetic approaches," *Microbiome*, vol. 7, no. 1, pp. 1–14, Mar. 2019.

[94] A. C. E. Darling, B. Mau, F. R. Blattner, and N. T. Perna, "Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements," *Genome Res.*, vol. 14, no. 7, pp. 1394–1403, Jul. 2004.

[95] H. J. Wu, A. H. J. Wang, and M. P. Jennings, "Discovery of virulence factors of pathogenic bacteria," *Curr. Opin. Chem. Biol.*, vol. 12, no. 1, pp. 93–101, Feb. 2008.

[96] L. Chen *et al.*, "VFDB: a reference database for bacterial virulence factors," *Nucleic Acids Res.*, vol. 33, no. suppl_1, pp. D325–D328, Jan. 2005.

[97] S. Ribeiro *et al.*, "BacSPaD: A Robust Bacterial Strains' Pathogenicity Resource Based on Integrated and Curated Genomic Metadata," *Pathogens*, vol. 13, no. 8, Aug. 2024.

[98] S. Cosentino, M. Voldby Larsen, F. Møller Aarestrup, and O. Lund, "PathogenFinder - Distinguishing Friend from Foe Using Bacterial Whole Genome Sequence Data," *PLoS One*, vol. 8, no. 10, p. e77302, Oct. 2013.

[99] K. G. Joensen *et al.*, "Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*," *J. Clin. Microbiol.*, vol. 52, no. 5, pp. 1501–1510, 2014.

[100] A. K. Järvinen *et al.*, "Rapid identification of bacterial pathogens using a PCR- and microarray-based assay," *BMC Microbiol.*, vol. 9, no. 1, pp. 1–16, Aug. 2009.

[101] S. A. Bustin *et al.*, "The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments," *Clin. Chem.*, vol. 55, no. 4, pp. 611–622, Apr. 2009.

[102] C. Wang *et al.*, "*Pseudomonas aeruginosa* Detection Using Conventional PCR and Quantitative Real-Time PCR Based on Species-Specific Novel Gene Targets Identified by Pangenome Analysis," *Front. Microbiol.*, vol. 13, p. 820431, May 2022.

[103] W. J. Botha, S. Serfontein, M. M. Greyling, and D. K. Berger, "Detection of *Xylophilus ampelinus* in grapevine cuttings using a nested polymerase chain reaction," *Plant Pathol.*, vol. 50, pp. 515–526, 2001.

[104] G. Carminati *et al.*, "Improving reliability of PCR diagnostics for *Xylophilus ampelinus* by metagenome-informed circumscription of the target taxon," *Plant Pathol.*, vol. 74, pp. 249–257, Jan. 2025.

[105] M. Pirc, M. Ravnikar, J. Tomlinson, and T. Dreo, "Improved fireblight diagnostics using quantitative real-time PCR detection of *Erwinia amylovora* chromosomal DNA," *Plant Pathol.*, vol. 58, no. 5, pp. 872–881, Oct. 2009.

[106] A. Bogožalec Košir, T. Cvelbar, M. Kammel, H. P. Grunert, H. Zeichhardt, and M. Milavec, "Digital PCR method for detection and quantification of specific antimicrobial drug-resistance mutations in human cytomegalovirus," *J. Virol. Methods*, vol. 281, no. April, p. 113864, 2020.

[107] S. Antil *et al.*, "DNA barcoding, an effective tool for species identification: a review," *Mol. Biol. Reports 2022 501*, vol. 50, no. 1, pp. 761–775, Oct. 2022.

[108] M. C. J. Maiden *et al.*, "MLST revisited: the gene-by-gene approach to bacterial genomics," *Nat. Rev. Microbiol. 2013 1110*, vol. 11, no. 10, pp. 728–736, Sep. 2013.

[109] C. R. Jackson, P. J. Fedorka-Cray, and J. B. Barrett, "Use of a genus- and species-specific multiplex PCR for identification of enterococci," *J. Clin. Microbiol.*, vol. 42, no. 8, pp. 3558–3565, Aug. 2004.

[110] N. Pal, C. C. Block, and C. A. C. Gardner, "A real-time PCR differentiating *Pantoea stewartii* subsp. *stewartii* from *P. stewartii* subsp. *indologenes* in corn seed," *Plant Dis.*, vol. 103, no. 7, pp. 1474–1486, Jul. 2019.

[111] T. Goldfarb *et al.*, "NCBI RefSeq: reference sequence standards through 25 years of curation and annotation," *Nucleic Acids Res.*, vol. 53, no. D1, pp. D243–D257, Jan. 2025.

[112] M. C. F. Thomsen, H. Hasman, H. Westh, H. Kaya, and O. Lund, "RUCS: rapid identification of PCR primers for unique core sequences," *Bioinformatics*, vol. 33, no. 24, pp. 3917–3921, Dec. 2017.

[113] A. Untergasser *et al.*, "Primer3—new capabilities and interfaces," *Nucleic Acids Res.*, vol. 40, no. 15, pp. e115–e115, Aug. 2012.

[114] P. M. Holland, R. D. Abramson, R. Watson, and D. H. Gelfand, "Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of Thermus aquaticus DNA polymerase.," *Proc. Natl. Acad. Sci.*, vol. 88, no. 16, pp. 7276–7280, Aug. 1991.

[115] J. G. Natalini, S. Singh, and L. N. Segal, "The dynamic lung microbiome in health and disease," *Nat. Rev. Microbiol. 2022 214*, vol. 21, no. 4, pp. 222–235, Nov. 2022.

[116] M. R. Knowles and R. C. Boucher, "Mucus clearance as a primary innate defense mechanism for mammalian airways," *J. Clin. Invest.*, vol. 109, no. 5, pp. 571–577, Mar. 2002.

[117] S. A. Whiteside, J. E. McGinniss, and R. G. Collman, "The lung microbiome: Progress and promise," *J. Clin. Invest.*, vol. 131, no. 15, Aug. 2021.

[118] S. H. Choi *et al.*, "Clinical characteristics of community-acquired viridans streptococcal pneumonia," *Tuberc. Respir. Dis. (Seoul).*, vol. 78, no. 3, pp. 196–202, Jul. 2015.

[119] M. F. Moffatt and W. O. C. M. Cookson, "The lung microbiome in health and disease," *Clin. Med. (Northfield. Il).*, vol. 17, no. 6, pp. 525–529, Dec. 2017.

[120] R. P. Dickson *et al.*, "Bacterial topography of the healthy human lower respiratory tract," *MBio*, vol. 8, no. 1, Jan. 2017.

[121] A. C. Blanchard and V. J. Waters, "Microbiology of Cystic Fibrosis Airway Disease," *Semin. Respir. Crit. Care Med.*, vol. 40, no. 6, pp. 727–736, 2019.

[122] T. F. Carr, R. Alkatib, and M. Kraft, "Microbiome in Mechanisms of Asthma," *Clin. Chest Med.*, vol. 40, no. 1, pp. 87–96, Mar. 2019.

[123] E. Monsó, "Microbiome in chronic obstructive pulmonary disease," *Ann. Transl. Med.*, vol. 5, no. 12, p. 251, 2017.

[124] A. Karvela, O. Z. Veloudiou, A. Karachaliou, T. Kloukina, G. Gomatou, and E. Kotteas, "Lung microbiome: an emerging player in lung cancer pathogenesis and progression," *Clin. Transl. Oncol.*, vol. 25, no. 8, pp. 2365–2372, Aug. 2023.

[125] E. Tufvesson, H. Markstad, G. Bozovic, M. Ekberg, and L. Bjermer, "Inflammation

and chronic colonization of *Haemophilus influenzae* in sputum in COPD patients related to the degree of emphysema and bronchiectasis in high-resolution computed tomography," *Int. J. COPD*, vol. 12, pp. 3211–3219, Nov. 2017.

[126] R. P. Dickson, F. J. Martinez, and G. B. Huffnagle, "The role of the microbiome in exacerbations of chronic lung diseases," *Lancet*, vol. 384, no. 9944, pp. 691–702, Aug. 2014.

[127] L. A. Carmody *et al.*, "Changes in cystic fibrosis airway microbiota at pulmonary exacerbation," *Ann. Am. Thorac. Soc.*, vol. 10, no. 3, pp. 179–187, Jun. 2013.

[128] R. P. Dickson *et al.*, "Bacterial topography of the healthy human lower respiratory tract," *MBio*, vol. 8, no. 1, Jan. 2017.

[129] E. Coli and K. Pneumoniae, "Surveillance of antimicrobial resistance in Europe, 2023 data: executive summary."

[130] A. Cassini *et al.*, "Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis," *Lancet Infect. Dis.*, vol. 19, no. 1, pp. 56–66, Jan. 2019.

[131] C. J. Murray *et al.*, "Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis," *Lancet*, vol. 399, no. 10325, pp. 629–655, Feb. 2022.

[132] G. D. Wright, "The antibiotic resistome: The nexus of chemical and genetic diversity," *Nat. Rev. Microbiol.*, vol. 5, no. 3, pp. 175–186, 2007.

[133] R. Hubbard, "The burden of lung disease," *Thorax*, vol. 61, no. 7, p. 557, Jul. 2006.

[134] P. Burney, D. Jarvis, and R. Perez-Padilla, "The global burden of chronic respiratory disease in adults," *Int. J. Tuberc. Lung Dis.*, vol. 19, no. 1, pp. 10–20, Jan. 2015.

[135] H. Pailhoriès *et al.*, "Antibiotic resistance in chronic respiratory diseases: from susceptibility testing to the resistome," *Eur. Respir. Rev.*, vol. 31, no. 164, May 2022.

[136] A. Duhaniuc *et al.*, "Multidrug-Resistant Bacteria in Immunocompromised Patients," *Pharm. 2024, Vol. 17, Page 1151*, vol. 17, no. 9, p. 1151, Aug. 2024.

[137] M. Maeusli *et al.*, "Horizontal Gene Transfer of Antibiotic Resistance from *Acinetobacter baylyi* to *Escherichia coli* on Lettuce and Subsequent Antibiotic Resistance Transmission to the Gut Microbiome," *mSphere*, vol. 5, no. 3, Jun. 2020.

[138] A. Rodrigo-Troyano and O. Sibila, "The respiratory threat posed by multidrug resistant Gram-negative bacteria," *Respirology*, vol. 22, no. 7, pp. 1288–1299, Oct. 2017.

[139] X. Tan *et al.*, "Chronic *Staphylococcus aureus* Lung Infection Correlates With Proteogenomic and Metabolic Adaptations Leading to an Increased Intracellular Persistence," *Clin. Infect. Dis.*, vol. 69, no. 11, pp. 1937–1945, Nov. 2019.

[140] Z. Zhang *et al.*, "Assessment of global health risk of antibiotic resistance genes," *Nat. Commun. 2022 131*, vol. 13, no. 1, pp. 1–11, Mar. 2022.

[141] Z. A. Khan, M. F. Siddiqui, and S. Park, "Current and emerging methods of antibiotic susceptibility testing," *Diagnostics*, vol. 9, no. 2, 2019.

[142] R. Banerjee and R. Patel, "Molecular diagnostics for genotypic detection of antibiotic resistance: current landscape and future directions," *JAC-Antimicrobial*

Resist., vol. 5, no. 1, Dec. 2022.

[143] M. Su, S. W. Satola, and T. D. Read, "Genome-based prediction of bacterial antibiotic resistance," *J. Clin. Microbiol.*, vol. 57, no. 3, pp. 1–15, 2019.

[144] C. L. Schwan *et al.*, "Genotypic and Phenotypic Characterization of Antimicrobial Resistance Profiles in Non-typhoidal *Salmonella enterica* Strains Isolated From Cambodian Informal Markets," *Front. Microbiol.*, vol. 12, p. 711472, Sep. 2021.

[145] A. G. McArthur *et al.*, "The comprehensive antibiotic resistance database," *Antimicrob. Agents Chemother.*, vol. 57, no. 7, pp. 3348–3357, 2013.

[146] V. K. Deekshit and S. Srikumar, "'To be, or not to be'—The dilemma of 'silent' antimicrobial resistance genes in bacteria," *J. Appl. Microbiol.*, vol. 133, no. 5, pp. 2902–2914, Nov. 2022.

[147] J. M. A. Blair, M. A. Webber, A. J. Baylay, D. O. Ogbolu, and L. J. V. Piddock, "Molecular mechanisms of antibiotic resistance," *Nat. Rev. Microbiol.*, vol. 13, no. 1, pp. 42–51, 2015.

[148] R. J. Fair and Y. Tor, "Antibiotics and Bacterial Resistance in the 21st Century," *Perspect. Medicin. Chem.*, pp. 25–64, 2014.

[149] W. Liu *et al.*, "Machine learning for identifying resistance features of *Klebsiella pneumoniae* using whole-genome sequence single nucleotide polymorphisms," *J. Med. Microbiol.*, vol. 70, no. 11, p. 001474, Nov. 2021.

[150] J. Pereira-Marques *et al.*, "Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis," *Front. Microbiol.*, vol. 10, no. JUN, p. 454372, Jun. 2019.

[151] A. K. Guitor *et al.*, "Capturing the Resistome: A targeted capture method to reveal antibiotic resistance determinants in metagenomes," *Antimicrob. Agents Chemother.*, no. October, 2019.

[152] C. Urbaniak *et al.*, "Detection of antimicrobial resistance genes associated with the International Space Station environmental surfaces," *Sci. Rep.*, vol. 8, no. 1, pp. 1–13, 2018.

[153] F. Shen and C. Sergi, "Sputum Analysis," *Pocketb. Chest Physiother.*, pp. 43–43, Feb. 2023.

[154] S. M. Carney *et al.*, "Methods in Lung Microbiome Research," 2020.

[155] A. J. Dicker *et al.*, "The sputum microbiome, airway inflammation, and mortality in chronic obstructive pulmonary disease," *J. Allergy Clin. Immunol.*, vol. 147, no. 1, pp. 158–167, Jan. 2021.

[156] R. Feigelman *et al.*, "Sputum DNA sequencing in cystic fibrosis: Non-invasive access to the lung microbiome and to pathogen details," *Microbiome*, vol. 5, no. 1, pp. 1–14, Feb. 2017.

[157] A. Bogožalec Košir, D. Lužnik, V. Tomič, and M. Milavec, "Evaluation of DNA Extraction Methods for Reliable Quantification of *Acinetobacter baumannii*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa*," *Biosensors*, vol. 13, no. 4, p. 463, Apr. 2023.

[158] T. Charalampous *et al.*, "Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection," *Nat. Biotechnol.*, vol. 37, no. 7, pp. 783–792, 2019.

[159] N. A. Kok *et al.*, "Host DNA depletion can increase the sensitivity of *Mycobacterium* spp. detection through shotgun metagenomics in sputum," *Front. Microbiol.*, vol. 13, p. 949328, Oct. 2022.

[160] I. Nikolić *et al.*, "Genetic diversity and pathogenicity of *Pseudomonas syringae* pv. *aptata* isolated from sugar beet," *Plant Pathol.*, vol. 67, no. 5, pp. 1194–1207, Jun. 2018.

[161] Ališ, F. Van Gijsegem, J. Pédron, M. Ravnikar, and T. Dreo, "Diversity within the novel *Dickeya fangzhongdai* sp., isolated from infected orchids, water and pears," *Plant Pathol.*, vol. 67, no. 7, pp. 1612–1620, Sep. 2018.

[162] H. Steele, B. E. Laue, G. A. MacAskill, S. J. Hendry, and S. Green, "Analysis of the natural infection of European horse chestnut (*Aesculus hippocastanum*) by *Pseudomonas syringae* pv. *aesculi*," *Plant Pathol.*, vol. 59, no. 6, pp. 1005–1013, Dec. 2010.

[163] J. R. Lamichhane and C. Bartoli, "Plant pathogenic bacteria in open irrigation systems: what risk for crop health?," *Plant Pathol.*, vol. 64, no. 4, pp. 757–766, Aug. 2015.

[164] A. Chai *et al.*, "Aerosol transmission of *Pseudomonas amygdali* pv. *lachrymans* in greenhouses," *Sci. Total Environ.*, vol. 748, p. 141433, Dec. 2020.

[165] C. L. Monteil, M. Bardin, and C. E. Morris, "Features of air masses associated with the deposition of *Pseudomonas syringae* and *Botrytis cinerea* by rain and snowfall," *ISME J.*, vol. 8, no. 11, pp. 2290–2304, Nov. 2014.

[166] C. Digonnet *et al.*, "Deciphering the route of *Ralstonia solanacearum* colonization in *Arabidopsis thaliana* roots during a compatible interaction: Focus at the plant cell wall," *Planta*, vol. 236, no. 5, pp. 1419–1431, Oct. 2012.

[167] D. A. Duong, R. V. Jensen, and A. M. Stevens, "Discovery of *Pantoea stewartii* ssp. *stewartii* genes important for survival in corn xylem through a Tn-Seq analysis," *Mol. Plant Pathol.*, vol. 19, no. 8, p. 1929, Aug. 2018.

[168] H. J. Choi, Y. J. Kim, Y. J. Lim, and D. H. Park, "Survival of *Erwinia amylovora* on Surfaces of Materials Used in Orchards," *Res. Plant Dis.*, vol. 25, no. 2, pp. 89–93, Jun. 2019.

[169] V. Trkulja, A. Tomić, R. Iličić, M. Nožinić, and T. P. Milovanović, "*Xylella fastidiosa* in Europe: From the Introduction to the Current Status," *Plant Pathol. J.*, vol. 38, no. 6, pp. 551–571, Dec. 2022.

[170] A. Degrave, S. Siamer, T. Boureau, and M. A. Barny, "The AvrE superfamily: ancestral type III effectors involved in suppression of pathogen-associated molecular pattern-triggered immunity," *Mol. Plant Pathol.*, vol. 16, no. 8, p. 899, Oct. 2015.

[171] L. Andresen, V. Kõiv, T. Alamäe, and A. Mäe, "The Rcs phosphorelay modulates the expression of plant cell wall degrading enzymes and virulence in *Pectobacterium carotovorum* ssp. *carotovorum*," *FEMS Microbiol. Lett.*, vol. 273, no. 2, pp. 229–238, Aug. 2007.

[172] M. Hu *et al.*, "Genomic and Functional Dissections of Dickeya zeae Shed Light on the Role of Type III Secretion System and Cell Wall-Degrading Enzymes to Host Range and Virulence," *Microbiol. Spectr.*, vol. 10, no. 1, Feb. 2022.

[173] C. L. Bender, F. Alarcón-Chaidez, and D. C. Gross, "*Pseudomonas syringae*

Phytotoxins: Mode of Action, Regulation, and Biosynthesis by Peptide and Polyketide Synthetases," *Microbiol. Mol. Biol. Rev.*, vol. 63, no. 2, p. 266, Jun. 1999.

[174] T. Komatsu and N. Kondo, "Winter habitat of *Xylophilus ampelinus*, the cause of bacterial blight of grapevine, in Japan," *J Gen Plant Pathol*, vol. 81, pp. 237–242, 2015.

[175] R. A. Redak, A. H. Purcell, J. R. S. Lopes, M. J. Blua, R. F. Mizell, and P. C. Andersen, "The Biology of Xylem Fluid-Feeding Insect Vectors Of *Xylella fastidiosa* and Their Relation to Disease Epidemiology," *Annu. Rev. Entomol.*, vol. 49, no. Volume 49, 2004, pp. 243–270, Jan. 2004.

[176] M. D. Koutsoudis, D. Tsaltas, T. D. Minogue, and S. B. Von Bodman, "Quorum-sensing regulation governs bacterial adhesion, biofilm development, and host colonization in *Pantoea stewartii* subspecies *stewartii*," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 15, pp. 5983–5988, Apr. 2006.

[177] S. Timilsina *et al.*, "Xanthomonas diversity, virulence and plant–pathogen interactions," *Nat. Rev. Microbiol. 2020 188*, vol. 18, no. 8, pp. 415–427, Apr. 2020.

[178] M. Nandi, J. Macdonald, P. Liu, B. Weselowski, and Z. C. Yuan, "*Clavibacter michiganensis* ssp. *michiganensis*: bacterial canker of tomato, molecular interactions and disease management," *Mol. Plant Pathol.*, vol. 19, no. 8, pp. 2036–2050, Aug. 2018.

[179] B. A. Williamson-Benavides and A. Dhingra, "Understanding Root Rot Disease in Agricultural Crops," *Hortic. 2021, Vol. 7, Page 33*, vol. 7, no. 2, p. 33, Feb. 2021.

[180] C. I. Kado, "Historical account on gaining insights on the mechanism of crown gall tumorigenesis induced by *Agrobacterium tumefaciens*," *Front. Microbiol.*, vol. 5, no. AUG, p. 64946, Aug. 2014.

[181] R. Czajkowski, M. C. M. Pérombelon, J. A. Van Veen, and J. M. Van der Wolf, "Control of blackleg and tuber soft rot of potato caused by *Pectobacterium* and *Dickeya* species: a review," *Plant Pathol.*, vol. 60, no. 6, pp. 999–1013, Dec. 2011.

[182] G. W. Sundin, L. F. Castiblanco, X. Yuan, Q. Zeng, and C. H. Yang, "Bacterial disease management: challenges, experience, innovation and future prospects," *Mol. Plant Pathol.*, vol. 17, no. 9, pp. 1506–1518, Dec. 2016.

[183] "EPPO activities on plant quarantine." [Online]. Available: https://www.eppo.int/ACTIVITIES/quarantine_activities. [Accessed: 21-Mar-2025].

[184] A. Bonaterra, E. Badosa, N. Daranas, J. Francés, G. Roselló, and E. Montesinos, "Bacteria as Biological Control Agents of Plant Diseases," *Microorg. 2022, Vol. 10, Page 1759*, vol. 10, no. 9, p. 1759, Aug. 2022.

[185] A. Nawaz *et al.*, "Bacteriophages: an overview of the control strategies against phytopathogens," *Egypt. J. Biol. Pest Control*, vol. 33, no. 1, pp. 1–10, Dec. 2023.

[186] M. Verhaegen *et al.*, "On the use of antibiotics to control plant pathogenic bacteria: a genetic and genomic perspective," *Front. Microbiol.*, vol. 14, p. 1221478, Jun. 2023.

[187] J. Mergaert, L. Verdonck, and K. Kersters, "Transfer of *Erwinia ananas* (synonym, *Erwinia uredovora*) and *Erwinia stewartii* to the genus *Pantoea* emend. as *Pantoea ananas* (Serrano 1928) comb. nov. and *Pantoea stewartii* (Smith 1898) comb. nov.,

respectively, and description of *Pantoea stewartii* subsp. *indologenes* subsp. nov.," *Int. J. Syst. Bacteriol.*, vol. 43, no. 1, pp. 162–173, Jan. 1993.

[188] M. C. Roper, "*Pantoea stewartii* subsp. *stewartii*: Lessons learned from a xylem-dwelling pathogen of sweet corn," *Mol. Plant Pathol.*, vol. 12, no. 7, pp. 628–637, Sep. 2011.

[189] M. Jeger *et al.*, "Pest categorisation of *Pantoea stewartii* subsp. *stewartii*," *EFSA J.*, vol. 16, no. 7, p. e05356, Jul. 2018.

[190] P. Doblas-Lbéfiez *et al.*, "Dominant, Heritable Resistance to Stewart's Wilt in Maize Is Associated with an Enhanced Vascular Defense Response to Infection with Pantoea stewartii," *Mol. Plant. Microbe. Interact.*, vol. 32, no. 12, pp. 1581–1597, 2019.

[191] D. Cui *et al.*, "First Report of *Pantoea stewartii* subsp. *stewartii* Causing Bacterial Leaf Wilt of Sugarcane in China," *Plant Disease*, vol. 105, no. 4, p. 1190, Feb. 2021.

[192] S. Koirala *et al.*, "Identification of Two Novel Pathovars of *Pantoea stewartii* subsp. *indologenes* Affecting *Allium* sp. and Millets," *Phytopathology*, vol. 111, no. 9, pp. 1509–1519, Sep. 2021.

[193] F. Mangeli, F. Sistani, Z. Lori, M. Azadvar, A. Hosseinipour, and N. Shahriyari, "First report of maize leaf blight disease caused by *Pantoea stewartii* subsp *indologenes* in Kerman Province, Iran," *Crop Prot.*, vol. 184, p. 106836, Oct. 2024.

[194] M. Ahmad, D. R. Majerczak, S. Pike, M. E. Hoyos, A. Novacky, and D. L. Coplin, "Biological Activity of Harpin Produced by *Pantoea stewartii* subsp. *stewartii*," *MPMI*, vol. 14, no. 10, pp. 1223–1234, Feb. 2007.

[195] V. R. Correa *et al.*, "the bacterium *Pantoea stewartii* uses two different type III secretion systems to colonize its plant host and insect vector," *Appl. Environ. Microbiol.*, vol. 78, no. 17, pp. 6327–6336, Sep. 2012.

[196] J. H. Ham, D. R. Majerczak, A. S. Arroyo-Rodriguez, D. M. Mackey, and D. L. Coplin, "WtsE, an AvrE-Family Effector Protein from *Pantoea stewartii* subsp. *stewartii*, Causes Disease-Associated Cell Death in Corn and Requires a Chaperone Protein for Stability," *MPMI*, vol. 19, no. 10, pp. 1092–1102, Feb. 2007.

[197] C. M. Herrera, M. D. Koutsoudis, X. Wang, and S. B. Von Bodman, "*Pantoea stewartii* subsp. *stewartii* Exhibits Surface Motility, Which is a Critical Aspect of Stewart's Wilt Disease Development on Maize," *MPMI*, vol. 21, no. 10, pp. 1359–1370, 2008.

[198] M. Mohammadi, L. Burbank, and M. C. Roper, "*Pantoea stewartii* subsp. *stewartii* Produces an Endoglucanase That Is Required for Full Virulence in Sweet Corn," *MPMI*, vol. 25, no. 4, pp. 463–470, Mar. 2012.

[199] D. A. Duong, A. M. Stevens, and R. V. Jensen, "Expanded Analysis of the *Pantoea stewartii* subsp. *stewartii* DC283 Complete Genome Reveals Plasmid-borne Virulence Factors," *bioRxiv*, p. 407825, Sep. 2018.

[200] P. De Maayer *et al.*, "The large universal *Pantoea* plasmid LPP-1 plays a major role in biological and ecological diversification," *BMC Genomics*, vol. 13, no. 1, pp. 1–12, Nov. 2012.

[201] D. Y. Shyntum, J. Theron, S. N. Venter, L. N. Moleleki, I. K. Toth, and T. A. Coutinho, "*Pantoea ananatis* utilizes a type VI secretion system for pathogenesis

and bacterial competition," *Mol. Plant-Microbe Interact.*, vol. 28, no. 4, pp. 420–431, Apr. 2015.

[202] C. Panagopoulos, "The disease tsilik marasi of grapevine: its description and identification of the causal agent (*Xanthomonas ampelina* sp. nov.)," *Ann. Inst. Phytopathol. Beanki*, no. 9, pp. 59–81, 1969.

[203] A. Willems, M. Gillis, K. Kersters, L. Van Den Broecke, and J. De Ley, "Transfer of *Xanthomonas ampelina* Panagopoulos 1969 to a New Genus, *Xylophilus* gen. nov., as *Xylophilus ampelinus* (Panagopoulos 1969) comb. nov.," *Int. J. Syst. Evol. Microbiol.*, vol. 37, no. 4, pp. 422–430, Oct. 1987.

[204] S. A. Lee *et al.*, "*Xylophilus rhododendri* sp. nov., Isolated from Flower of Royal Azalea, *Rhododendron schlippenbachii*," *Curr. Microbiol.*, vol. 77, no. 12, pp. 4160–4166, Dec. 2020.

[205] F. I. Chacón *et al.*, "Native Cultivable Bacteria from the Blueberry Microbiome as Novel Potential Biocontrol Agents," *Microorg. 2022, Vol. 10, Page 969*, vol. 10, no. 5, p. 969, May 2022.

[206] E. Laanto and H. M. Oksanen, "Three Phages from a Boreal Lake during Ice Cover Infecting *Xylophilus*, *Caulobacter*, and *Polaromonas* Species," *Viruses*, vol. 15, no. 2, Feb. 2023.

[207] P. Portier *et al.*, "Analysis of the Diversity of *Xylophilus ampelinus* Strains Held in CIRM-CFBP Reveals a Strongly Homogenous Species," *Microorganisms*, vol. 10, no. 8, p. 1531, Aug. 2022.

[208] "*Xylophilus ampelinus*," *EPPO Bull.*, vol. 39, no. 3, pp. 403–412, Dec. 2009.

[209] H. Khlaif, I. Abu-Obeid, and B. Werikat, "African Journal of Agricultural Research Occurrence of plant bacterial diseases in Jordan," vol. 13, no. 40, pp. 2104–2117, 2018.

[210] H. D. Erasmus, F. N. Matthee, and H. A. Louw, "A comparison between plant pathogenic species of *Pseudomonas*, *Xanthomonas* and *Erwinia* with special reference to the bacterium responsible for bacterial blight of vines," *Phytophylactica*, vol. 6, pp. 11–18, 1974.

[211] J. L. S. S. Serfontein, J. J. Serfontein, W. J. Botha, "The isolation and characterisation of *Xylophilus ampelinus*," *Vitis*, vol. 36, no. 4, pp. 209–209, Jan. 1997.

[212] C. Manceau, S. Grall, C. Brin, and J. Guillaumes, "Bacterial extraction from grapevine and detection of *Xylophilus ampelinus* by a PCR and Microwell plate detection system," *EPPO Bull.*, vol. 35, no. 1, pp. 55–60, Apr. 2005.

[213] S. Sapkota, M. Mergoum, and Z. Liu, "The translucens group of *Xanthomonas translucens*: Complicated and important pathogens causing bacterial leaf streak on cereals," *Mol. Plant Pathol.*, vol. 21, no. 3, pp. 291–302, Mar. 2020.

[214] H. Alvandi *et al.*, "Pathovar-Specific PCR Method for Detection and Identification of *Xanthomonas translucens* pv. *undulosa*," *Plant Dis.*, vol. 107, no. 8, pp. 2279–2287, Aug. 2023.

[215] K. E. Ledman, V. Roman-Reyna, R. D. Curland, N. Heiden, J. M. Jacobs, and R. Dill-Macky, "Comparative Genomics of *Xanthomonas translucens* pv. *undulosa* Strains Isolated fromWeedy Grasses and Cultivated Wild Rice," *Phytopathology*,

vol. 113, no. 11, pp. 2083–2090, Nov. 2023.

[216] N. F. Charkhabi *et al.*, "Complete genome sequencing and targeted mutagenesis reveal virulence contributions of Tal2 and Tal4b of *Xanthomonas translucens* pv. *undulosa* ICMP11055 in bacterial leaf streak of wheat," *Front. Microbiol.*, vol. 8, no. AUG, p. 279604, Aug. 2017.

[217] D. Willner, J. Daly, D. Whiley, K. Grimwood, C. E. Wainwright, and P. Hugenholtz, "Comparison of DNA extraction methods for microbial community profiling with an application to pediatric bronchoalveolar lavage samples," *PLoS One*, vol. 7, no. 4, 2012.

[218] L. Abusleme, B. Y. Hong, A. K. Dupuy, L. D. Strausbaugh, and P. I. Diaz, "Influence of DNA extraction on oral microbial profiles obtained via 16S rRNA gene sequencing," *Journal of Oral Microbiology*, vol. 6, no. 1. 2014.

[219] M. Panek *et al.*, "Methodology challenges in studying human gut microbiota-effects of collection, storage, DNA extraction and next generation sequencing technologies OPEN."

[220] R. Eisenhofer, J. J. Minich, C. Marotz, A. Cooper, R. Knight, and L. S. Weyrich, "Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations," *Trends Microbiol.*, vol. 27, no. 2, pp. 105–117, 2019.

[221] N. A. Kennedy *et al.*, "The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing," *PLoS One*, vol. 9, no. 2, pp. 1–9, 2014.

[222] J. Sinclair, N. P. West, and A. J. Cox, "Comparison of four DNA extraction methods for 16s rRNA microbiota profiling of human faecal samples," *BMC Res. Notes*, vol. 16, no. 1, pp. 1–5, Dec. 2023.

[223] L. Terranova *et al.*, "How to process sputum samples and extract bacterial DNA for microbiota analysis," *Int. J. Mol. Sci.*, vol. 19, no. 10, pp. 1–12, 2018.

[224] G. Agarwal, R. D. Gitaitis, and B. Dutta, "Pan-genome of novel *Pantoea stewartii* subsp. *indologenes* reveals genes involved in onion pathogenicity and evidence of lateral gene transfer," *Microorganisms*, vol. 9, no. 8, p. 1761, Aug. 2021.

[225] K. C. Crosby *et al.*, "Genomic delineation and description of species and within-species lineages in the genus *Pantoea*," *Front. Microbiol.*, vol. 14, p. 1254999, Nov. 2023.

[226] R. Ibrahim, S. Izera Ismail, M. Yasin Ina-Salwany, and D. Zulperi, "Biology, Diagnostics, Pathogenomics and Mitigation Strategies of Jackfruit-Bronzing Bacterium *Pantoea stewartii* subspecies *stewartii*: What Do We Know So Far about This Culprit?," *Horticulturae*, vol. 8, no. 8, p. 702, 2022.

[227] F. Kerff *et al.*, "Crystal structure and activity of *Bacillus subtilis* YoaJ (EXLX1), a bacterial expansin that promotes root colonization," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 44, pp. 16876–16881, Nov. 2008.

[228] M. A. Jacques *et al.*, "Using Ecology, Physiology, and Genomics to Understand Host Specificity in *Xanthomonas*," *Annu. Rev. Phytopathol.*, vol. 54, no. Volume 54, 2016, pp. 163–187, Aug. 2016.

[229] M. Bergsma-Vlami *et al.*, "Exploring the genetic diversity of recent *Ralstonia pseudosolanacearum* phylotype I findings in Europe," Oct. 2025.

[230] Š. Alič *et al.*, "Genome-Informed Design of a LAMP Assay for the Specific Detection of the Strain of '*Candidatus* Phytoplasma asteris' Phytoplasma Occurring in Grapevines in South Africa," *Plant Dis.*, vol. 106, no. 11, pp. 2927–2939, Nov. 2022.

[231] T. Dreo, K. Gruden, C. Manceau, J. D. Janse, and M. Ravnikar, "Development of a real-time PCR-based method for detection of *Xylophilus ampelinus*," *Plant Pathol.*, vol. 56, no. 1, pp. 9–16, Feb. 2007.

Bibliography

Publications Related to the Thesis

Clavijo F., Barrera C., Benčič A., Croce V., Jacobs J. M., A. J. Bernal A. J., Koebnik R., Roman-Reyna V. (2022). Complete genome sequence resource for Xanthomonas translucens pv. undulosa MAI5034, a wheat pathogen from Uruguay. Phytopathology, 112(9):2036-2039. doi: 10.1094/PHYTO-01-22-0025-A

Benčič, A., Pothier, J., Dreo, T. (2025). Comparative genomics and pathogenicity of *Pantoea stewartii* subsp. *stewartii* reveal multiple introductions and limited distribution in Europe. bioRxiv, 2025.11.07.687269. https://doi.org/10.1101/2025.11.07.687269

Benčič, A., Toplak, N., Koren, S., Bogožalec Košir, A., Milavec, M., Tomič, V., Lužnik, D., Dreo, T. (2024). Metrological evaluation of DNA extraction method effects on the bacterial microbiome and resistome in sputum. mSystems, 2024, Vol. 9, Iss. 9, Art. No. e00735-24. https://doi.org/10.1128/msystems.00735-24

Benčič, A., Bogožalec Košir, A., Matičič, J., Pirc, M., Turnšek, N., Dreo, T. (2025). Development of a multi-targeted real-time PCR assay for the detection of the grapevine pathogen *Xylophilus ampelinus*. Plant Methods, 2025, Vol. 21, Art. No. 99, https://doi.org/10.1186/s13007-025-01422-4

Anteljević, M., Rosić, I., Medić, O., Ranković, T., Sunjog, K., Kračun-Kolarević, M., Kolarević, S., Dreo, T., Benčič, A., Berić, T., Stanković, S., Nikolić, I. (2025). 2.4 Irrigation systems as reservoirs of diverse and pathogenic *Pseudomonas syringae strains* endangering crop health. Water Research X, 2025, Vol. 28, Art. No. 100380. https://doi.org/10.1016/j.wroa.2025.100380

Biography

Aleksander Benčič was born in Ljubljana, Slovenia, on 12 October 1992. He completed his primary and secondary education in Koper, Slovenia. He obtained his Bachelor of Science (BSc) and Master of Science (MSc) degrees in Biochemistry from the Faculty of Chemistry and Chemical Technology (UL FKKT) at the University of Ljubljana in 2014 and 2018, respectively. He conducted his BSc thesis research at the Department of Molecular Biology and Nanobiotechnology of the National Institute of Chemistry under the supervision of Prof. Dr. Gregor Anderluh. He completed his MSc thesis at the UL FKKT Chair of Biochemistry under the supervision of Assistant Prof. Gregor Gunčar. Both theses were in the fields of protein biotechnology and nanobiotechnology. In 2017, he undertook an internship at the German Institute of Human Nutrition Potsdam-Rehbruecke, working in nutritional epidemiology. Following graduation, he briefly worked at the Marine Biology Station Piran, focusing on using molecular methods to detect fraud in fish products. In 2019, he started a position as a young researcher at the National Institute of Biology, working under the supervision of Dr Tanja Dreo. He enrolled in the Nanosciences and Nanotechnologies doctoral programme at the Jožef Stefan International Postgraduate School, where he was a member of the student council (2021–2022) and played an active role in organising the annual student conference. His research at NIB primarily focuses on various applications of high-throughput sequencing (HTS) in bacterial genomics. He utilises genomic and metagenomic data for microbiome research, comparative genomic analysis, and developing new molecular methods to detect plant pathogenic bacteria. His research has applications in areas ranging from clinical microbiology to plant protection and disease management. During his PhD studies, he undertook visits to the Zurich University of Applied Sciences (in 2022) and the Flanders Research Institute for Agriculture, Fisheries and Food (in 2023), where he gained expertise in HTS, bioinformatics, and comparative genomics. He has also collaborated with the University Clinic Golnik to facilitate the introduction of targeted HTS in clinical diagnostics and the detection of genes associated with antibiotic resistance. He has attended several workshops on whole genome sequencing, bioinformatics, and data analysis. He has been involved in multiple international projects organised within the scope of the Euphresco network. The candidate is involved in the promotion and popularisation of science among the general population. He has participated in the production of promotional videos showcasing NIB activities and has been active in the institute's promotion group.