U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

MSc

2º
CICLO

FCUP
2020

U.PORTO

On assembling bacterial genomes from long reads: a case study in walnut-associated *Xanthomonas* spp.

Miguel Teixeira

FC

On assembling bacterial genomes from long reads: a case study in walnut-associated *Xanthomonas* spp.

Miguel Teixeira

Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto em
Bioinformática e Biologia Computacional

2020

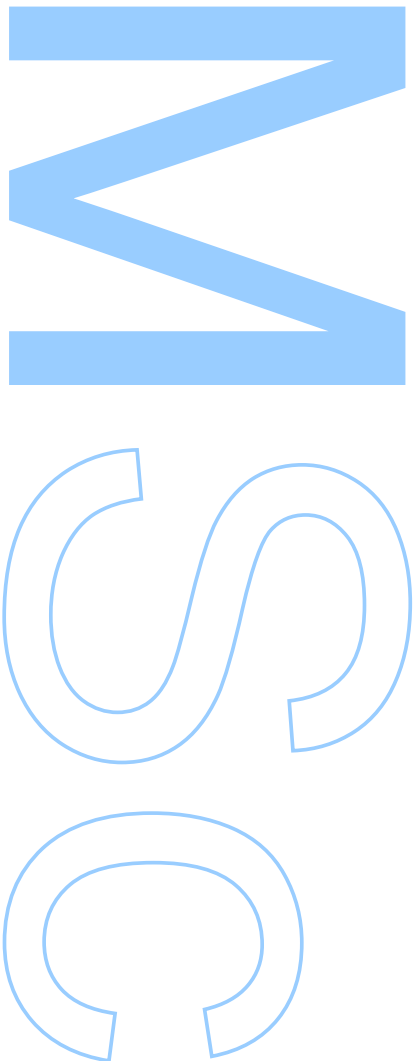
MSc

On assembling bacterial genomes from long reads: a case study in walnut-associated *Xanthomonas* spp.

Miguel Teixeira
Mestrado em Bioinformática e Biologia Computacional
Departamento de Biologia
2020

Orientador
Nuno A. Fonseca,
Investigador Coordenador, CIBIO-INBIO

Orientador
Fernando Tavares,
Professor Auxiliar, Faculdade de Ciências, Universidade do Porto



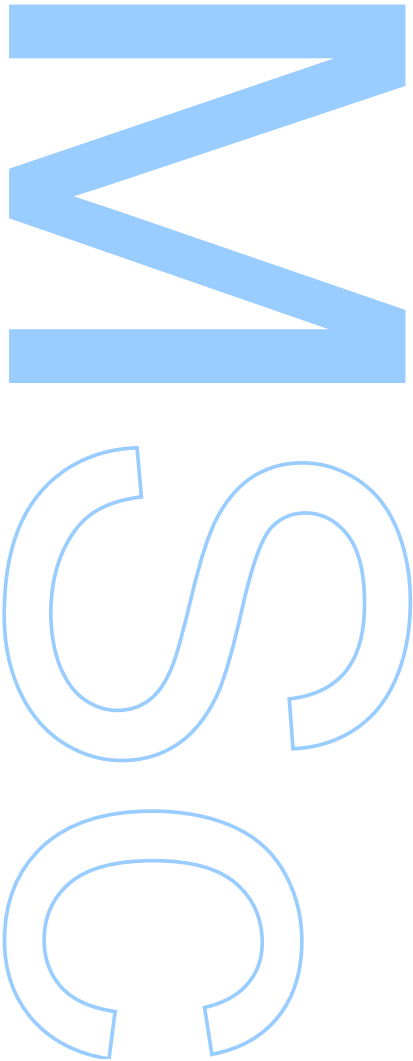
U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____



UNIVERSIDADE DO PORTO

MASTERS THESIS

On assembling bacterial genomes from long reads: a case study in walnut-associated *Xanthomonas* spp.

Author:
Miguel Teixeira

Supervisor:
Prof. Dr. Nuno A. Fonseca

Supervisor:
Prof. Dr. Fernando Tavares

A thesis submitted in fulfilment of the requirements
for the degree of MSc. in Bioinformatics and Computational Biology

at the

Faculdade de Ciências da Universidade do Porto
Departamento de Biologia

January 25, 2021

" See that bird? It's a Spencer's warbler. Well, in Italian, it's a Chutto Lapittida, in German it's called a halzenfugel, in Chinese, it's a Chung-long-tah, and in Japanese, it's a Katano Tekeda. You can know the name of that bird in all the languages of the world, but when you're finished, you'll still know nothing about the bird, absolutely nothing about the bird. So let's look at the bird and see what it's doing, that's what counts. "

Richard P. Feynman

Acknowledgements

I would like to acknowledge my thesis supervisors, professor Nuno Fonseca and professor Fernando Tavares. For the mentorship and for being always available to answer my questions, making this a pleasant and grateful journey.

To my family, for everything.

To my friends. Those 104, SRSB, the Gang and the Bioinformaticians from The association.

UNIVERSIDADE DO PORTO

Abstract

Faculdade de Ciências da Universidade do Porto
Departamento de Biologia

MSc. in Bioinformatics and Computational Biology

On assembling bacterial genomes from long reads: a case study in walnut-associated *Xanthomonas* spp.

by Miguel Teixeira

Xanthomonas spp. associated with walnut trees include pathogens responsible for several diseases that decrease orchards productivity and cause severe damages in the long term. Despite the efforts that are being made, practices to eradicate or control these pathogens remain ineffective, mainly due to the lack of detection methods capable to identify particularly virulent *Xanthomonas* strains, and the limited knowledge concerning the patrimony of *Xanthomonas* genetic determinants putatively involved in bacteria adaptation to the host. To address these issues, it is particularly important to acquire a set of genomic data to build a robust pangenome of walnut-associated *Xanthomonas*, that may translate the adaptations to the host walnut trees, but also assembling the genome sequences of each *Xanthomonas* isolates into a single contig, which is most useful to carry out accurate comparative genomic studies. Starting with datasets of genomic short- and long-reads, obtained by Illumina and ONT sequencing platforms, from sixteen *Xanthomonas* spp. isolated from disease plants hosts, this work aimed to disclose the most suitable bioinformatics pipeline leading to the complete genome sequences for each of the sixteen strains sequenced. In this regard, we define pipelines to assemble isolated bacterial genomes from ONT data, and to perform hybrid assemblies with ONT and Illumina data. For ONT data, we define a pipeline supported by statistical analysis, concerning the preponderance of basecalling models, preprocessing raw reads (correcting and trimming), different assemblers (Canu, Flye, Miniasm, Wtdbg2) and polishing strategies (with Racon and Medaka). These assemblies were shown to be highly contiguous and could recover most of the genomic repertoire as evaluated by assessing the core-genome and the

BUSCO genes. From the methodology for hybrid assembling, we achieve contiguous, complete and accurate assemblies for eleven isolates of *X. arboricola* and five of *X. euroxanthea*, isolated from *Juglans regia* or *Carya illinoinensis*. A comparative analysis comprising the sixteen genomes and focused on four T3E (*avrBs2*, *xopF1*, *xopN* and *xopR*) which are acknowledged to be involved in *Xanthomonas* phytopathogenicity by conditioning the plant host defense mechanisms, revealed striking differences that suggest a need to revise their importance for infection. Furthermore, this case-study on T3E emphasizes the utility of assembling the genome sequences of each member of a bacterial consortium to a couple or single contig to allow assertive comparative genomics analysis that may ultimately contribute to profile the full set of putative genetic determinants of pathogenicity and virulence. This data will certainly enhance the capacity to design more efficient surveillance practices of epidemic bacteria, such as walnut-associated *Xanthomonas*, and discover target-specific phytosanitary treatments, therefore contributing to a tight control of epidemic outbreaks.

UNIVERSIDADE DO PORTO

Resumo

Faculdade de Ciências da Universidade do Porto
Departamento de Biologia

Mestrado em Bioinformática e Biologia Computacional

On assembling bacterial genomes from long reads: a case study in walnut-associated *Xanthomonas* spp.

por Miguel Teixeira

O género *Xanthomonas* inclui espécies patogénicas responsáveis por diminuições na produtividade e danos severos a longo prazo no cultivo de nogueira. Práticas para controlar e erradicar estes agentes resultam em esforços infrutíferos devido à falta de métodos de deteção capazes de identificar estirpes virulentas e ao desconhecimento do património genético responsável pela adaptação da bactéria ao hospedeiro. Para abordar este problema é importante adquirir informação genómica para identificar o pangenoma de *Xanthomonas* associadas a nogueira, de forma a identificar os determinantes genéticos da adaptação ao hospedeiro, e, obter *assemblies* de isolados em *contigs* únicos, de forma a permitir estudos de genómica comparativa mais completos. O objetivo deste trabalho foi alcançar *assemblies* completos de dezasseis genomas de *Xanthomonas* isolados de hospedeiros sintomáticos. Para tal, definimos *pipelines* para efetuar *assemblies* de *long reads* respetivas a genomas sequenciados por ONT e *assemblies* híbridos a partir de *long reads* e *short reads* provenientes de sequenciação por ONT e Illumina. A pipeline destinada a *assemblies* de *long reads* foi definida com base numa análise estatística considerando preponderância dos modelos de *basecalling*, do pré-processamento de *reads*, da performance de diferentes *assemblers* e de diferentes estratégias de correção para obter *assemblies* completos. Esta pipeline demonstrou ser capaz de efetuar *assemblies* contíguos e completos, recuperando a maior parte do reportório genético. No caso de *assemblies* híbridos, obtivemos sequências contíguas, completas e precisas para onze genomas de *X. arboricola* e cinco de *X.euroxanthea*, isolados a partir de *Juglans regia* ou *Carya illinoinensis*.

Uma análise comparativa dos dezasseis genomas verificou diferenças consideráveis em quatro genes efetores relacionados com patogenicidade, questionando a sua preponderância no processo de infecção. Além disso, esta análise evidencía a imporância de obter *assemblies* completos para avaliar reportórios genéticos presumivelmente relacionados com fatores de patogenicidade e virulência. Os resultados deste trabalho são relevantes para definir métodos de resposta eficientes a epidemias de fitopatogénios através de tratamentos específicos e da vigilância de agentes etiológicos.

Contents

List of Tables

List of Figures

List of Abbreviations

ANI	Average Nucleotide Identity
BLAST	Basic Local Alignment Search Tool
BUSCO	Benchmarking Universal Single-Copy Orthologs
DBG	De Bruijn graphs
ddNTP	Di-deoxynucleotidetriphosphates
DNA	Deoxyribonucleic Acid
dsDNA	double stranded Deoxyribonucleic Acid
EMBOSS	European Molecular Biology Open Software Suite
ENA	European Nucleotide Archive
FCUP	Faculdade de Ciências da Universidade do Porto
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
OLC	Overlap Layout Consensus
PCR	Polymerase Chain Reaction
pv.	pathovar
ssDNA	single stranded Deoxyribonucleic Acid
SGS	Second Generation Sequencing
T3E	Type III Effectors
TGS	Third Generation Sequencing
UniProt	Universal Protein Resource
Xaj	*Xanthomonas arboricola* pv. *juglandis*
Xe	*Xanthomonas euroxanthea*
Xca	*Xanthomonas campestris*
Xci	*Xanthomonas citri*

Chapter 1

Introduction

1.1 Context

Over the last two centuries, the seven-fold increase in the human population required the development of novel agriculture techniques to satisfy the world demand for food [1, 2]. The cultivation of a single or few species, supported by fertilizers, pesticides and resistance genes, are regular practices that aim to reduce plant infections and increase the yield. Therefore, agro-ecosystems tend to be genetically uniform environments under high selective pressure, which promotes the emergence of highly adapted and aggressive pathogens that affect crops and cause social-economic problems [3, 4].

Xanthomonas is a wide genus of plant-associated bacteria, comprising members mostly known for its pathogenic role in various crops, and, for the high specificity regarding host and tissue colonization[5, 6]. In particular, walnut-associated *Xanthomonas* are subject of increasing concern for recurrent outbreaks of several diseases that remain to be controlled. The cause of these diseases have confidently been attributed to *X. arboricola*, yet, reports of individual hosts colonized by *Xanthomonas* populations with close related non-pathogenic and pathogenic strains (including a novel *Xanthomonas* non-arboricola species) raise questions regarding the population role in the epidemics [7, 8]. Thus, to address this problem, there is an urge to improve surveillance practices with fine detection methods that effectively discriminate close related strains, and, to further comprehend the diseases epidemiology beyond the known hosts and pathogens.

Pathogen surveillance is fundamental to act upon an invading epidemic before it gets out of control. However, surveillance is *per se* limited by the sampling process, regarding frequency, environmental conditions and the number of individuals/structures/tissues

sampled [9]. It relies on the knowledge of growth requirements, laborious work in successive cultivations to isolate a pure lineage, and reliable DNA markers for taxonomic identification[10–13]. Thus, surveillance initiatives are mainly implemented when disease symptoms appear, which may only occur when the epidemic prevalence is already high, and rely on *a priori* knowledge of suspicious targets to select target-specific DNA markers for its identification. The outcome of this paradigm is that unculturable or unknown co-colonizing pathogens are likely unnoticed and, among the culturable bacteria, close related ones may pass indistinguishable due to the markers limited resolution to discriminate infrasubspecific taxonomic levels [12, 14–16].

To further comprehend a disease epidemiology one must assess biotic factors beyond the host and pathogen populations[4]. This is notedly relevant when dealing with bacterial species due to the abundance of mobile genetic mechanisms that drive evolutionary processes and shape interactions among populations. This is remarked in the mentioned paradigm of walnut-associated *Xanthomonas*, where particular genetic determinants for host specificity and pathogenicity emerge within a microbial community[5–7]. Nevertheless, population genomics is strongly limited by the ability to culture bacteria and the use of typing methods, either DNA or phenotype-based, that regard few genomic regions [10, 11, 17]. Analyses of wide genomic domains are possible with whole genome sequencing, yet, fragmented assemblies from short-read sequencing miss to unveil the chromosomal structure and tend to keep comparative genomics focused in well defined features. Single molecule sequencing technologies, as MinION from Oxford Nanopore Technologies (ONT), produce long reads that promote contiguous assemblies of isolated genomes and the reconstruction of wide genomic regions from unculturable bacteria [18–20] . Such sequences provide further genomic information, relevant to mine novel infrasubspecific DNA markers, to assess the evolutive history of cohabiting bacteria and the emergence of functional traits related to pathogenicity or host adaptation. ONT MinION also stands for its portability and the possibility of real time sequencing, which facilitates field practices in epidemiology. As a counterpart, those technologies present higher sequencing errors that made them less suitable for finer analysis[21, 22].

1.2 Thesis objectives

The objective of this thesis was to *de novo* assembly bacterial genomes from ONT long reads, considering, as a case study, the paradigm of walnut-associated Xanthomonads.

First, we aimed to define a pipeline to assemble complete and accurate genomes using just ONT data. Second, we aimed to perform hybrid assemblies, from ONT and Illumina data, of sixteen *Xanthomonas* spp. isolated from walnut trees. These assemblies should be contiguous, complete and accurate, reflecting the advantages, and overcoming the disadvantages, of both long and short-read sequencing approaches.

Fulfilling these objectives, must result in high quality *Xanthomonas* spp. assemblies to support future epidemiological studies. In addition, it should demonstrate the potential of ONT sequencing to assemble contiguous genomes and allow us to make considerations regarding its use to reconstruct wide regions of environmental genomes.

1.3 Thesis outline

Here we describe the five chapters of this thesis:

- **Introduction**: Contextualization of the work and objectives.
- **Background**: Topics on *Xanthomonas* spp. epidemiology, bacterial detection and identification, sequencing technologies, and whole genome assemblies.
- **On assembling bacterial genomes with long reads**: Addresses the selection of a pipeline to assemble long reads and to perform hybrid assemblies.
- **On assembling *de novo* the genomes of 16 *Xanthomonas* spp. isolates**: Addresses the whole genome assembling and a comparative analysis of 16 *Xanthomonas* spp. isolates.
- **Conclusion**: Provides the final remarks of this work and future perspectives.

Chapter 2

Background

2.1 Plant-pathogen epidemiology: Xanthomonads in walnuts

Xanthomonas is a large bacteria genus belonging to the gamma subdivision of proteobacteria, which includes numerous plant pathogens that cause several diseases in economically important crops distributed worldwide. This genus is composed strictly by aerobic bacteria, with an optimal growth temperature between 25 and 30 °C, morphologically described as rod shaped with a single polar flagellum cells, forming yellow pigmented colonies in culture media[15, 23].

The species *Xanthomonas arboricola* became a pathogen of serious concern in Europe due to its constant threat to stone fruits and nut trees[24–26]. In particular, *X. arboricola* pathovar *juglandis* (Xaj) is the etiological agent of Walnut Bacterial Blight [27], Brown apical necrosis [28] and Vertical Oozing Canker[29] in walnut trees, namely *Juglans regia*[30] and *Carya illinoinensis*[31].

These plant diseases, which remain poorly controlled, are responsible for major crop losses and mainly mitigated by the phytosanitary application of ecosystem-harmful copper compounds. The ecology and epidemiology of walnut-associated *Xanthomonas*, which include both pathogenic and non-pathogenic strains and epithytic and endophytic behaviors, are still badly understood due to difficulties regarding bacterial detection in asymptomatic hosts, accurate identification of strains particularly virulent, and a scarce understanding of the bacteria population genetics. In fact, Xaj detection still require bacteria isolation in culture media and reliable DNA-specific markers that are often insufficient to discriminate closely related strains. Regardless these constraints, advances are being

made, as for instance the recent characterization and announcement of a new walnut-associated species named *Xanthomonas euroxanthea* (Xe) that includes pathogenic and non-pathogenic strains and has been shown to co-colonize together with Xaj walnut host trees [8, 32]. This example highlights the importance to optimize infrasubspecific detection methods capable to accurately discriminate closely related strains sharing the same niche, which is absolutely needed to determine populations dynamics and comprehend how its members co-evolve and interact with a host and the environment.

2.2 Bacterial identification and detection

Traditional bacterial detection and identification methods rely on the isolation of bacteria in selective culture media followed by phenotypic characterization of the cells, and colonies. These procedures are indispensable to describe a bacteria isolate and are still routinely used, however they are unsuitable to distinguish bacterial lineages within the same species and depend on bacteria culturability, which is time-consuming, laborious and incompatible to study non-culturable bacteria [11, 33]. Bacterial detection and identification methods based in biochemical or serological techniques, aim to detect biomolecules and consequently to characterize the chemical composition of bacterial cell structures, detect the presence of antigens epitopes or profile metabolic pathways. Although some of these methods are culture-independent and can be used to detect bacteria in complex environment matrices, they require a comprehensive knowledge of the metabolic and molecular properties of the target bacteria, and frequently do not allow to assess important bacterial traits such as presence of virulence factors or resistance to chemical treatments [33].

DNA-based methods for bacterial characterization rely on the analysis of genomic fingerprints, or patterns, due to DNA polymorphisms of small genomic regions [11, 33]. Although these methodologies have contributed to important breakthroughs in the last decades, they are still strongly dependent on the culturability, among other disadvantages. For instance, highly conserved genes, as the 16S rRNA gene resulted in the remarkable resolution of the phylogenetic structure of the prokaryotic domain and become widely used as a molecular marker in microbial ecology [34]. However, its utility for lower taxonomic discrimination, i.e., species and infrasubspecific taxa, is debatable, particularly due to the polymorphisms between the multiple gene copies found in a single chromosome [35]. Alternatives, as tandem repeats, have also been extensively used as

molecular markers to discriminate closely related bacteria, although with contrasting results due to unstable number of repetitive elements [36]. More recently, sequence-based methods targeting bacteria housekeeping genes, such as the multilocus sequence typing (MLST) or multilocus sequence analysis (MLSA), have been successfully used for genotyping of numerous bacterial species [10], although they are inappropriate as detection methods[16].

Overall, characterizing a bacteria based on few, well conserved regions of its genome, requires a previous laborious analysis to select markers for the species in the given context, and analysis with different combinations of markers can only hardly imply ecological, evolutionary or phylogenetic relations. These features, together with the plasticity of bacterial genomes, often change established phylogenetic relations or lead to equivocated taxonomic attributions[37–40]. Furthermore, these molecular and detection-based approaches, strongly oriented to specific targets, often miss relevant data for epidemiological studies, such as antibiotic-resistance and virulence traits, that could provide valuable information to implement the most suitable sanitary measures [41–43]. Not surprisingly, with the major developments of sequencing technologies observed in the last two decades, coupled with the abrupt decrease of sequencing costs, whole bacterial genome analysis is foresee as the future gold standard for bacterial genotyping capable to provide an unprecedented capacity to discriminate strains, unveil complex phylogenies, and disclose genetic determinants of specific adaptations to distinct environments. This knowledge, will certainly contribute to improve epidemiological surveys ultimately capable to make risk assessment analysis[44–46].

2.3 Sequencing technologies

First generation sequencing: Sanger chain termination method

The first generation of sequencing technologies was born in 1970s when Maxam and Gilbert introduced a chemical method that compared the electrophoretic patterns from DNA fragments cleaved in purines, pyrimidines, adenines and cytosines to determine its nucleotide order [47]. This method was the first sequencing method widely used, however, due to the technical complexity, hazardous chemicals and difficulties to scale the method for sequences longer than a few hundred bases, lost popularity with the appearance of Sanger chain termination method [48].

Sanger method uses radiolabeled dideoxynucleotides (analogous to usual nucleotides present in DNA, but without the hydroxyl group in the sugar third carbon that enables the bond to the phosphate group in the next deoxynucleotide to extend the chain) to randomly stop DNA extension reactions in order to obtain fragments of all sizes with a terminal nucleotide labeled. Running the results of four parallel reactions (one for each deoxynucleotide) side by side on a polyacrylamide gel enables the inference of the original nucleotide order. This method, with technical improvements, outstanded for its robustness, accuracy and usability, and became the most used sequencing approach until the 2000s. In fact, Sanger method remains the most accurate method for sequencing and, in modern versions, results in reads up to several hundreds base pairs, nevertheless, its very low throughput per run makes it costly, laborious and inappropriate to sequence entire genomes [49].

Second generation sequencing: Illumina sequencing-by-synthesis and paired-end reads

In 2005, the second generation of sequencing technologies started to emerge with remarkable advantages over the first generation, namely, the very high throughput, reduced time, labor and cost of the process. Among the various SGS technologies, such as Roche 454, Illumina, ABI/SOLiD and Ion Torrent, the most popular approach is, arguably, the paired-end sequencing-by-synthesis from Illumina platforms that, in a simplified explanation, consist in three steps:

1. ssDNA is randomly fragmented and oligonucleotide adapters are bound to both ends;
2. DNA fragments are attached to a surface by complementarity with the adapters and amplified;
3. The complementary strand of the fragments is synthesized using color-labeled nucleotides to identify the sequence (sequencing-by-synthesis).

In the third step just the ends of the fragments are sequenced resulting in paired-end reads, usually around 150 nucleotides. Filtering the fragments by length in the first step allows to control the insert size between reads, therefore, it is possible to have overlapped paired reads or distant reads to flank unknown regions with specific length. Overlapped reads are easy to assemble, distant reads may help to define the boundaries of regions

impossible to assemble, as long repetitive regions. In fact, the main disadvantage of this technology is the short length of reads that is insufficient to resolve highly repetitive regions and hardly enables assemblies of whole genomes longer than 1-2 Mbp or complex genomes from environmental samples. Other possible disadvantages are the computational challenges to store and process massive amounts of data, requiring specialized manipulation and bioinformatic tools, nevertheless, with a throughput of more than 600 Gbp per run with an overall error rate around 1%, it results in extensive sequencing coverage with statistical confidence to be readily interpreted by their alignment to a reference genome [49–51].

Third generation sequencing: MinION single molecule sequencing and long reads

Third generation sequencing (TGS) technologies, as the ones by Oxford Nanopore Technologies and Pacific Biosciences, are characterized by its ability to sequencing single molecules in real time, producing long reads without the traditionally mandatory amplification process. Among TGS technologies, MinION from Oxford Nanopore Technologies is a device that measures a few centimeters in length and can be connected to a laptop using a standard USB port, standing out by its portability and the lowest sequence cost [21, 50]. MinION operating principle can be described, in a simplified manner, as follows:

1. dsDNA is sheared and adapters are bound to both ends;
2. An electric current is applied and measured in a membrane with protein nanopores embedded;
3. dsDNA is loaded into the membrane and, as it approaches a pore, it starts to unzip through the interaction of the adapter and pore;
4. As a single strand passes through the pore causes a variation in the electric current measured.

The signal from measuring the electric current is immediately available and therefore it is not necessary to wait until the end of the run to access the data. The raw signal must be converted to the desired nucleotide sequencing in a basecalling process using a neural network model. Several models are available, usually trading accuracy for speed [52]. MinION outputs long reads up to 900 kbp [53], likely enabling the resolution of

repetitive regions which leds to contiguous assemblies of whole genomes, without reference, and improves the reconstruction of environmental genomes. In comparison to SGS technologies MinION as lower throughput and high error rates (around 10%), however, it is claimed that errors can be computationally corrected to achieve an accuracy over 99%[21, 49–51].

2.4 *De novo* assembling of bacterial genomes

Assembling genomes is the process of reconstructing genomes from sequencing reads [54]. It can be done by mapping reads to a reference assembly (reference based) or from scratch (*de novo*). The first approach is faster and computationally less demanding, yet, its reliability depends on the quality and similarity of a previous assembled genome. That is, homologous regions can be successfully reconstructed by mapping reads, but structural variants, as insertions, inversions or translocations may be missed. Thus, assembling by mapping to a reference genome is suitable for methodologies that target particular genomic regions (e.g., inferring allelic variants for clinical diagnostics) but, if the whole genome and its structure are object of concern (e.g. population genomics or phylogeny inference), *de novo* assemblies are more likely to reproduce the real genomic structure.

There are several computational approaches to assemble a genome *de novo*. Two popular approaches rely on Overlap Layout Consensus (OLC) or De Bruijn Graphs (DBG) to build a graph representing overlaps between all sequenced reads [55, 56]. These, are computationally demanding processes, with major scalability issues and high sensitivity to the length and quality of the reads. For these reasons, assemblers are designed to address particular genomic features and sequencing methods, with concern for the computational resources.

Popular long read assemblers, as Canu, Miniasm and Wtdbg2 follow an OLC approach [57–60]. In OLC, each read is a vertex and edges between vertices are established if the reads share a common prefix or suffix. These graphs are more informative and less sensitive to errors, allowing the interpretation of sequences formed by the overlaps consensus. However, repetitive regions longer than the reads can not be resolved, and, insertions in those repetitive regions are easily missed. Flye assembler is based on DBG: the reads are splitted in k-mers (a read substring of length k, where the k value is adjusted); Each k-mer is a vertex, and edges are established if vertices share a prefix or suffix of length k-1; Strict k-1 long overlaps in sequences of length k, are strongly compromised

by the reads errors which often causes dead-end paths in graphs. Overall, DBG are computationally more efficient to build and store than OLC graphs [55]. Both approaches are found, for instance, in pipelines for hybrid assemblers. As an example, Unicycler first uses SPAdes to assemble short reads in a DBG, and then uses the long reads to build bridges following an OLC approach with Miniasm [61, 62].

Ideally, the assembly graphs should result in one unique path that represents a single continuous sequence, i.e., a contig. However, depending on the assemblers approach (OLC or DBG), the reads quality, and the presence of repetitive regions in the genome, the graphs are susceptible to artifacts such as tips (short dead-end paths diverging from the main path), bubbles (parallel short paths that start and end in the same two vertices), and bulges (low-coverage paths that create alternate paths between two nodes) [55, 56].

The assembly quality can be evaluated according to its accuracy, contiguity and completeness. Contiguity can be evaluated by assessing metrics as the number of contigs or N50. The N50 value indicates that 50% of the genome is assembled in contigs of length N50 or longer. Variations of this metric consider major fractions of the assembly (N70, N90) or, indicate the number of contigs that comprise a fraction of the genome (L50, L70, L90) [54, 63, 64]. An additional indication of contiguity can be obtained by assessing the existence of reads that overlap both extremities of a contig, suggesting an assembly of a circular structure, as bacterial chromosomes and plasmids usually are. Completeness concerns the genetic patrimony recovered by an assembly. It can be evaluated by assessing the core genome content, or the presence of highly conserved genetic features [65, 66]. Accuracy regards the confidence per base, an approach to assess it, involves mapping the reads against the assembly to evaluate the consensus of the overlapped reads. It reflects, for instance, the reliability of the allelic variants that one observes [63, 64].

Chapter 3

On assembling bacterial genomes with long reads

Abstract

Pathogen surveillance and populational genomic studies are crucial to comprehend disease emergence and to predict outbreaks. These, rely on bacterial detection and identification methods that generally consist in DNA-based typing of single lineages isolated in cultures. As most bacteria remain unculturable and typing methods concern few genomic aspects, environmental or non-pathogenic strains are overlooked along genomic features with relevant functional meaning.

Long reads from single-molecule sequencing technologies, as ONT, promote assemblies contiguity that may provide further genomic information and potentially recover wide genomic regions from unculturable bacteria. Here, we explore the limitations, due to high error rates, of ONT sequencing data, and the advantages of its long reads to assemble *Xanthomonas* spp. genomes. We achieved assemblies with improved contiguity from ONT data, and, contiguous and complete assemblies from hybrid approaches that include Illumina data.

3.1 Introduction

In recent times, outbreaks of walnut diseases caused by *Xanthomonas* are becoming more recurrent and causing progressive major damages [3, 9]. Comprehending its emergence and surveil the respective etiological agents is essential for rapid and effective actions to control outbreaks, nevertheless, efforts to do so, are challenged from sampling decision-making till the identification of infrasubspecific taxa.

Sampling processes are laborious and narrow by default. Beyond that, particular conditions must be satisfied to grow a bacteria, successive cultivations must be performed to isolate pure lineages from single cells, and then, selective DNA markers must be available to identify an isolate. Thus, bacterial detection and identification methods are oriented to suspicious targets. These must set appropriate growth conditions, that are unknown for most bacteria, and select taxonomic DNA markers, that are often unreliable to discriminate close related strains. As a consequence, pathogen monitoring is often driven by the appearance of disease symptoms, asymptomatic hosts are unsighted, and diseases epidemiology does not take into account the population genomics of environmental or non-pathogen strains [14]. In the case of walnut-associated *Xanthomonas*, populations of close strains, comprising pathogens and non-pathogens, have been isolated within the same individual. This sympatric occurrence of close-related commensal and pathogenic strains, remarks the call for highly discriminatory detection methods and extended disclosure of population genomic structure, to enlighten the emergence of pathogen traits and the role that cohabiting populations have on it. The mentioned challenges are, in part, tied to the characteristics of short-read sequencing technologies, and might be relieved with technologies as the portable long-read sequencer from Oxford Nanopore Technologies (ONT), MinION.

Short-read technologies, namely Illumina and Ion Torrent, sequence by synthesizing the target molecules, which requires a considerable initial amount of molecules obtained through amplification processes. It produces reads up to a few hundred base-pairs that are insufficient to assemble contiguous regions with problematic domains, as frequent repetitive patterns. Assemblies of isolated genomes are, therefore, very fragmented (typically with several dozen contigs) and the reconstruction of wide genomic regions from environmental samples is virtually impossible. As a main advantage, its runs have a massive yield of reads (sequencing depth from few hundred to thousand times) with low error rates (up to 1% of mostly mismatches [49]) that allow an high confidence per base in

the assembled sequences. Contrary, technologies such as ONT, can sequence long regions in few fragments without amplification processes. Its very long-reads (with mean length above a few thousand bases) easily cover extensive regions beyond possible problematic domains, enhancing contiguous assemblies of isolated genomes and the reconstruction of wide regions from environmental samples. ONT main disadvantage is the high error rate (above 10% of mismatches and indels [49]) that blur the assembly resolution. Nevertheless, the errors effect can be mitigated. In particular, complementing long-reads with short-reads, likely results in contiguous and complete assemblies with high confidence per base call.

Altogether, these possibilities may provide new insights into unculturable bacteria genomes and extend populational genomics beyond small target-regions, as the 16S rRNA gene and well conserved genes [18, 67, 68]. In addition, complete and contiguous assemblies are valuable to mine novel DNA markers, and disclose whole genetic repertoires and structural rearrangements. This would contribute to epidemics management through improved surveillance methodologies and further understanding of pathogenicity determinants.

In this chapter, we evaluate several tools and strategies to attenuate the high errors of long reads, in order to define a pipeline to assemble *Xanthomonas* spp. genomes from ONT data. We also tried hybrid approaches to assemble the same genomes from ONT and Illumina data. This way, we access some potential advantages and limitations of ONT sequencing for assembling bacterial genomes and support epidemiological studies.

3.2 Material and methods

Bacterial isolation and sequencing

Three *Xanthomonas* spp. strains (CPBF367, CPBF426 and CPBF427) were isolated in April 2016, in Loures, Portugal, from asymptomatic dormant buds of a isolated walnut tree (*Juglans regia*), known to develop symptoms of walnut bacterial blight during the growing season. The strains were grown on bacterial culture medium M2 (yeast extract, $2gL^{-1}$; Bacto peptone, $5g^{-1}$; $NaCl$, $5gL^{-1}$; KH_2PO_4, $0.45gL^{-1}$; Na_2HPO_4 $12H_2O$, $2.39gL^{-1}$) at 28 °C and $100rpm$ for $48h$. DNA was extracted using the E.Z.N.A. bacterial DNA purification kit (Omega Bio-tek, Norcross, GA) and sequenced with Illumina and ONT MinION platforms. Illumina sequencing was outsourced to GATC Biotech,

AG (Konstanz, Germany) using a HiSeq Illumina instrument with a standard 2 × 150 bp paired-end library protocol. Nanopore sequencing libraries were prepared with the SQK-LSK109 kit and multiplexed using the EXP-NBD104 barcoding kit. Sequencing was performed on a MinION sequencer using a R9.4.1 flow cell. Reads were basecalled and demultiplexed using Guppy* (version 3.4.1) with high accuracy mode and fast mode. Reads statistics were calculated with FastQC (version 0.11.5 [69]) and NanoStat (version 1.2.1 [70]).

De novo assemblies from ONT data

A pipeline to assemble a bacterial genome from long reads involves choices in 3 main steps:

1. **preprocessing raw reads**: we sequentially corrected and trimmed raw reads using the Canu (version 1.9[59]) modules;

2. **assembling**: we tried Canu (version 1.9), Flye (version2.7-b1585, with the option to rescue short unassembled plasmids [71]), Miniasm (version 0.3-r179 [60], with Minimap version 2.17-r974-dirty [72]) and Wtdbg2 (version 2.5, with the consenser wtpoa [58]);

3. **improving draft assemblies**: we polish with Racon (version 1.4.3 [73]), up to four iterations, followed by one correction with Medaka† (version 0.11.5, with model r941_min_high_g330 for reads basecalled using the accurate mode or the model r941_min_fast_g303 for reads basecalled in fast mode).

The tools were selected to cover different characteristics regarding accuracy, computational performance and input requirements. Its possible combinations result in ninety nine different pipelines. Default parameters were used for all software unless otherwise noted. A scheme for the above process is represented in Figure 3.1.

*Only available to Oxford Nanopore customers through their community site (http://community.nanoporetech.com)

†Sequence correction provided by ONT Research: https://github.com/nanoporetech/medaka

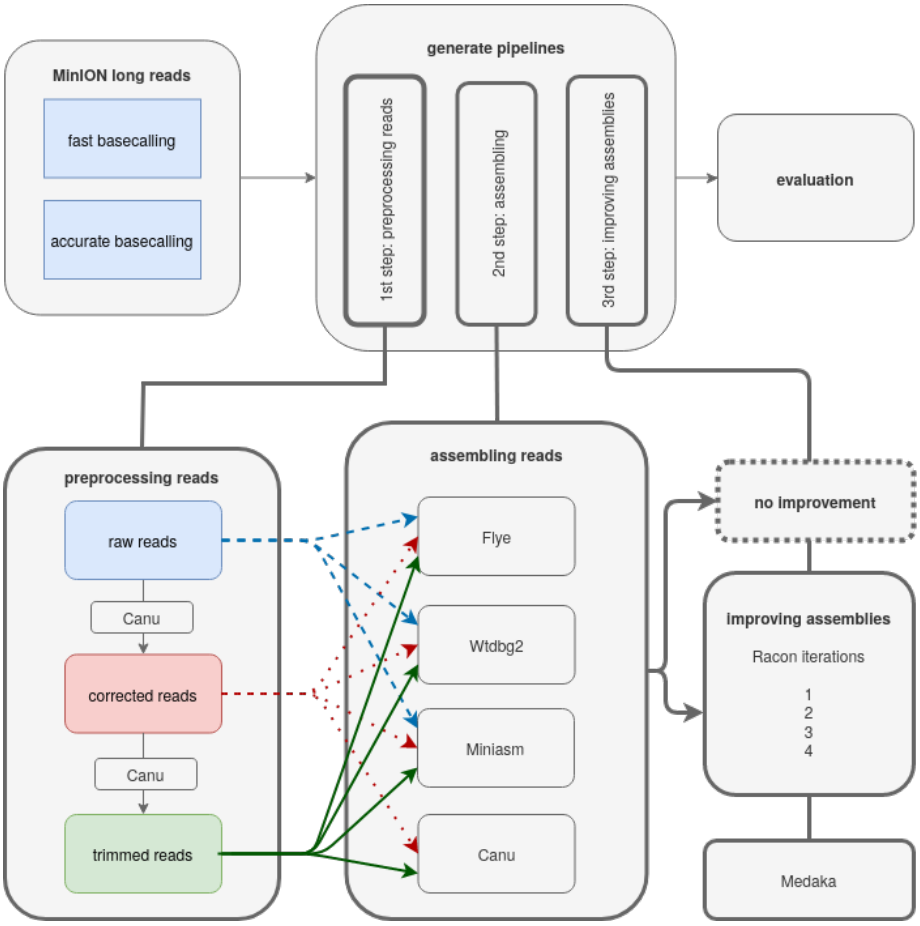


FIGURE 3.1: Methodology scheme, steps and tools to select a pipeline to assemble long reads.

The ninety nine pipelines to assemble long reads were evaluated considering 6 variables that mostly concern the assembly completeness and accuracy per base:

- Number of single nucleotide differences regarding a reference assembly: It consists in the sum of single nucleotide polymorphisms and single nucleotide insertions/deletions determined with the *dnadiff* program from Mummer (version 4.0.0 [74]). The reference assemblies for strains CPBF367, CPBF426 and CPBF427 were generated with Illumina data and are available in GenBank under the accession numbers NZ_UNRN00000000.1, NZ_UNRM00000000.1 and NZ_UNRO00000000.1, respectively.

- Number of complete core genes: The core genome was defined using EDGAR (version 2.3 [75]) for the *Xanthomonas* strains CPBF367, CPBF426, CPBF427, CPBF1521 and CPBF424 and resulted in a set of 3423 genes [7, 32, 76]. Strain-specific gene sequences were mapped with Minimap2. We consider just the best alignment for each

gene (by excluding secondary alignments with the parameter –secondary=no), that way, more alignments than queried genes meant that some genes were spliced for alignment. Counting the number of extra alignments and considering very unlikely that a gene was spliced more than twice, we calculate the n°of complete genes as the difference between the n°of queried genes and the n°of extra alignments;

- Number of differences in the core genes: Sum of mismatches and gaps when mapping the core genes (indicated by the NM flag in minimap2 alignment, secondary alignments were not considered);
- Number of housekeeping genes present: Considering a set of 7 informative genes commonly used for Multi Locus Sequence Analysis in *Xanthomonas* spp. [77–82]. The genes selected were *atpD*, *DnaK*, *efp*, *fyuA*, *glnA*, *gyrB* and *rpoD*, and were sequenced by Sanger. The sequences were mapped with Minimap2. Secondary, supplementary and chimeric alignments were not considered.
- Number of differences in the housekeeping genes: Sum of mismatches and gaps when mapping the housekeeping genes, assessed as mentioned for core genes.
- BUSCO-score: Reflects the genome completeness based on 1152 orthologous genes expected to be present in a species from the order *Xanthomonadales*. It was defined as $C - M - F/2$, where C, M and F represent the percentage of complete, missed and fragmented genes retrieved from an analysis with BUSCO (version 4.0.6, database xanthomonadales_odb10 [65, 66]).

Variables regarding computational performance of each pipeline, as the total elapsed time and maximum memory used, were registered.

An overall score was computed based on the pipelines relative performance regarding the six variables mentioned above, in the following manner: i) the pipelines were ranked based on the value obtained for each variable; ii) a pipeline score for a given strain was calculated by summing its ranks on all variables; iii) an overall pipeline score was calculated by summing its scores for all strains. A low overall score implies that a given pipeline had the first ranks (i.e., performed better) for the variables considered, therefore, the pipeline with lowest overall score was elected as the best one. We assess if pipelines comprising particular tools/steps tend to perform better than the others through its overall score. To do so, we perform Mann-Whitney U tests to test if the average overall score

was significantly ($\alpha < 0.01$) different depending on the choice made on each step [83]. For step 3 (improving draft assemblies) we also perform Wilcoxon signed-rank tests to test if the pipelines using Medaka were significantly different than the equivalent ones without Medaka (e.g., r_flye_m1 vs r_flye_r1; t_canu_m2 vs t_canu_r2) [84].

The procedure described above was also applied to reads basecalled in fast mode. Moreover, to compare the basecallings we pair same-pipeline assemblies from both basecalling modes and perform a Wilcoxon signed-rank test for the BUSCO-score. To verify if pipelines' performance was consistent for both basecalling modes we calculated a Pearson's correlation coefficient for the overall score ranks.

A pipeline unique name indicates how it is composed as follows:
{reads set}-{assembler}-{improvement}

- **reads set**: {r (raw), c (corrected), t (trimmed)}
- **assembler**: {canu, miniasm, flye, wtdbg2}
- **improvement step**: {no improvement (N), improved just with Racon (r), improved with Medaka after Racon (m)} {1,2,3,4 (number of Racon iterations)}

Hybrid *De novo* assemblies

We considered two approaches to perform hybrid assemblies using long ONT reads (basecalled in accurate mode) and short Illumina reads. In the first approach, we took long read assemblies from some the pipelines previously described, and polished them with short reads using Pilon (version 1.23 [85]). The pipelines were chosen, among the 99 mentioned in the anterior topic, considering the best overall score, the best score for individual samples, and, an apparent good structure, i.e., less contigs. For the second approach we used the Unicycler pipeline (version 0.4.8 [61]). We compared the assemblies from these two approaches with assemblies from long reads only (from the best pipeline in the anterior topic) and from short reads only (the public available assemblies used as reference in the anterior topic) considering the length, N50, number of contigs and BUSCO-score.

As a preliminary study to infer the sequencing depth needed to achieve contiguous and complete assemblies, we performed Unicycler assemblies varying the coverage from both technologies. We varied the ONT coverage and applied a linear regression model to estimate the minimum needed to achieve circular contiguous genomes. In the same

fashion, we varied the Illumina coverage to estimate the minimum needed to achieve complete genomes, according to the BUSCO-score. Based on the linear models information, we varied the sequencing depth of both technologies as an attempt to define the relation between the amount of ONT and Illumina data necessary to achieve contiguous and complete assemblies.

3.3 Results and discussion

From now one, for practical reasons concerning results writing and labels for data visualization, we replaced the CPBF in strains identifiers for an X.

The statistics for the ONT and Illumina reads are presented in Table 3.1 and Table 3.2. The mean quality is indicated in a phred-scale [86], and the coverage was estimated considering that the total genome size was 5M bp. The ONT reads basecalled with the accurate mode have a quality score around 11.8 which corresponds to an error probability of 6.6%. The N50 for all ONT datasets indicates that half the bases sequenced are in reads longer than 10,000 bases. Illumina datasets coverage ranges from 392x to 508x with pair-end reads of 2 x 151 bp and a mean quality score around 29, which implies an error probability near 0.1%. ONT and Illumina data are available under the accession numbers ERX4296808 and ERX2780809, for X367, ERX4296809 and ERX2780811, for X426 and, ERX4296810 and ERX2780812 for X427.

TABLE 3.1: ONT sequencing data statistics.

strain	dataset	n°of reads	mean length	mean quality	N50	coverage
X367	ONT (accurate)	18,857	6,386	11.8	14,257	24
X426	ONT (accurate)	19,997	5,895	11.9	10,553	23
X427	ONT (accurate)	14,302	6,358	11.8	13,332	18

TABLE 3.2: Illumina sequencing data statistics.

strain	dataset	n°of reads	mean length	mean quality	coverage
X367	Illumina	8,413,466	2 x 151	29.3	508
X426	Illumina	6,494,807	2 x 151	28.8	392
X427	Illumina	7,562,695	2 x 151	29.3	456

The number of reads, mean read length and mean read quality are represented in Figure 3.2, emphasizing the higher yield from Illumina and the trade off between quality and length for Illumina and ONT.

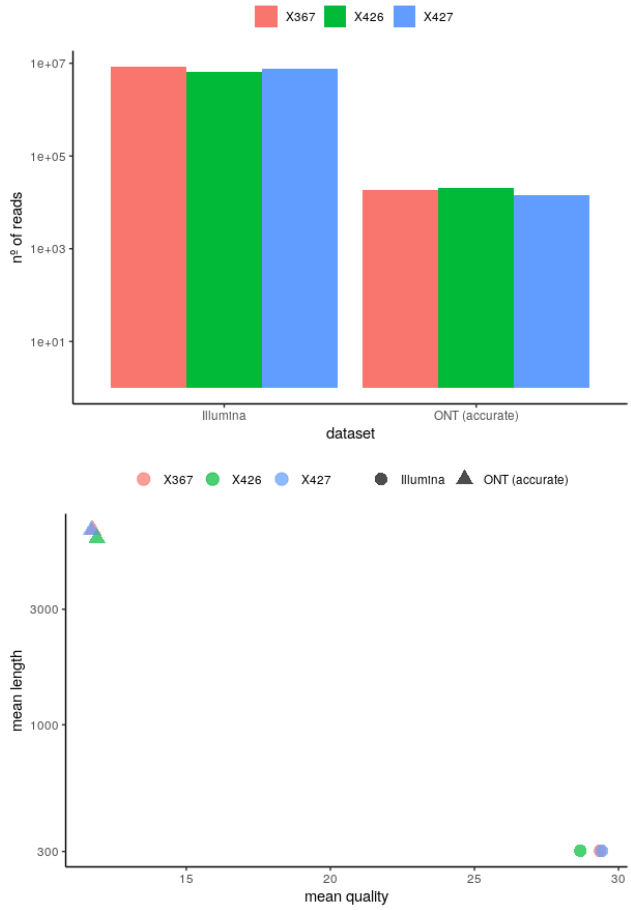


FIGURE 3.2: Yield, reads length and quality for ONT and Illumina data.

De novo assemblies from ONT data

The values distribution for the different variables considered are presented in Figure 3.3. BUSCO-score, single nucleotide differences regarding a reference assembly and number of differences in core genes, have similar, slightly skewed distributions for the three samples with worst performances for X427. Most pipelines produced assemblies containing all core genes and most BUSCO's, yet, there is a notorious skewness for X427. For the variables related to the housekeeping genes, all assemblies contained the seven housekeeping genes (*atpD*, *DnaK*, *efp*, *fyuA*, *glnA*, *gyrB* and *rpoD*) with few differences regarding the reference sequences for X367 and X427.

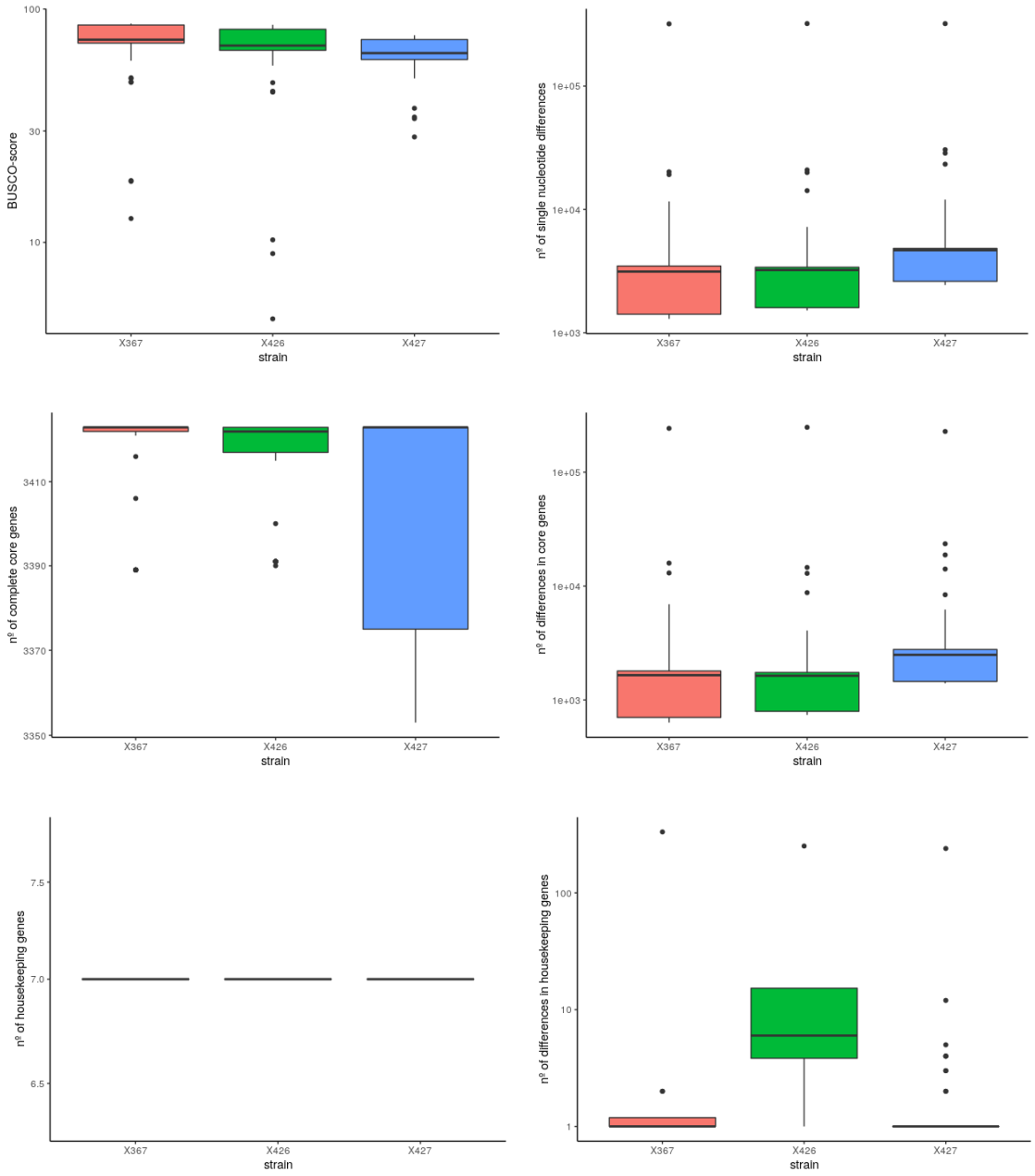


FIGURE 3.3: Boxplots diplaying the distribution of the variables used to assess assemblies quality.

→ Impact of the preprocessing choices in the assembled genomes

As demonstrated in Figure 3.4a, correcting and trimming excludes a considerable amount (up to a quarter) of reads, yet those are mostly short reads. That is noted by the superior mean length (around 2000 bases more) of corrected and trimmed sets.

Overall score values have similar distributions according to the preprocessing choice, as shown in Figure 3.4b, suggesting that preprocessing reads do not impact significantly the pipelines performance (Mann-Whitney U test: $\alpha < 0.01$, with a $p = 0.36$)

→ Which assembler performs better?

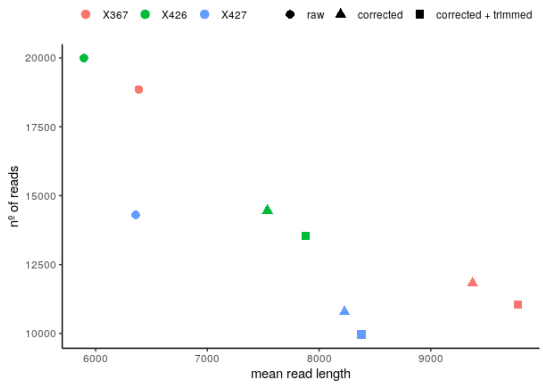
The overall score of the pipelines with Canu, Miniasm and Wtdbg2 were not significantly different (Mann-Whitney U test: $\alpha < 0.01$, with values of $p \geq 0.28$). The overall score of the pipelines with Flye is significantly lower (One-sided Mann-Whitney U test: $\alpha < 0.01$, $p = 0.0015$), as evidenced in Figure 3.4c.

→ Impact of improving draft assemblies

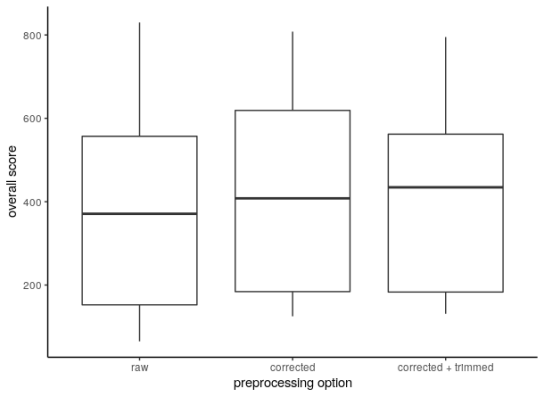
Improving assemblies (with Racon and/or Medaka) led to overall scores significantly inferior (Mann-Whitney U test: $\alpha < 0.01$, with a $p \approx 0$) as evidenced in Figure 3.4d. Furthermore, among the pipelines that improved the draft assemblies, running Medaka after Racon also results in significantly lower overall scores (One-sided Wilcoxon signed rank test: $\alpha < 0.01$, with a $p \approx 0$), it is evidenced in Figure 3.4e, where is represented the overall score for pairs of pipelines with and without Medaka.

There is also no significant difference in the overall score (Mann-Whitney U test: $\alpha < 0.01$, with values of $p \geq 0.1$) due to the number of Racon iterations before Medaka. This is represented in Figure 3.4f.

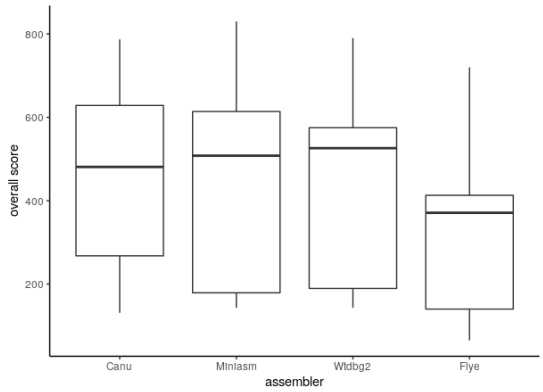
Further details on the statistical tests are available in Appendix A.



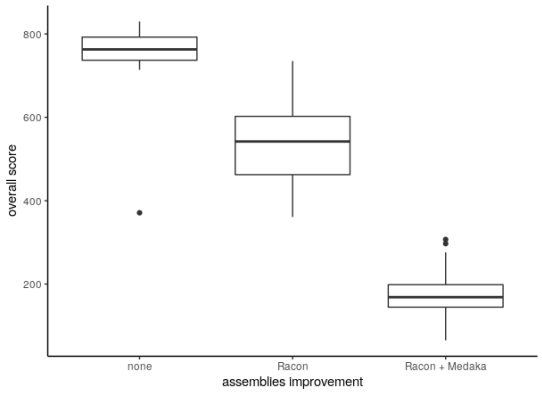
(A) The impact of preprocessing reads in the number of reads and its mean length.



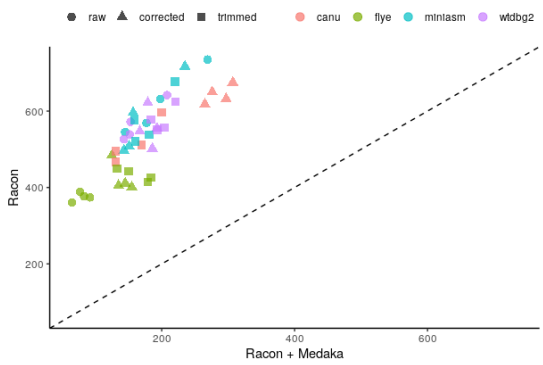
(B) Overall score distribution according to the preprocessing option.



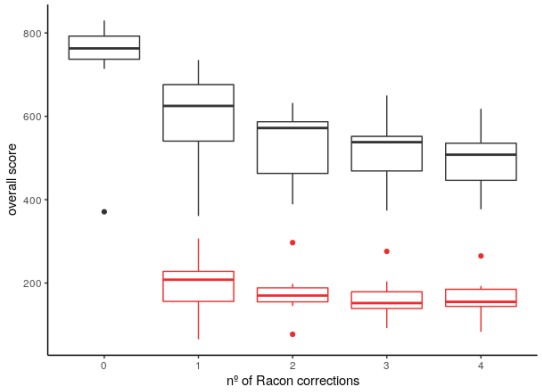
(C) Overall score distribution according to the assembler.



(D) Overall score distribution according to the improvement choice.



(E) Overall score pairs for pipelines with and without Medaka.



(F) The impact of Racon iterations. Pipelines without Medaka in black, with Medaka in red.

FIGURE 3.4: How different choices impacted assemblies quality.

What is the best pipeline?

In Table 3.3 are listed the overall scores of the 10 best pipelines. Unsurprisingly, pipelines that assembled raw reads with Flye and improve the draft assemblies with Racon and Medaka performed better. The best one was r-flye-m1. The individual scores per strain are indicated in Appendix A.

TABLE 3.3: Top 10 pipelines.

pipelines	overall score
r-flye-m1	65
r-flye-m2	77
r-flye-m4	83
r-flye-m3	92
c-flye-m1	125
t-canu-m4	131
t-flye-m1	133
r-wtdbg2-m4	143
c-flye-m2	145
c-flye-m4	155

In Table 3.4 we present some statistics for r-flye-m1 assemblies. The small number of contigs and an N50 close to the total length, demonstrate the advantage of ONT reads to achieve contiguous assemblies. The QUAL represents the mean error probability per base in a phred-scale, C,F and M respectively represent the percentage of complete, fragmented and missed BUSCO genes.

Similar results were obtained with an equivalent analysis for the fast basecalling, it is available in Appendix B, along with a comparison between basecallings that, as expected, shows accurate basecalled reads result in better assemblies.

TABLE 3.4: Statistics for assemblies from r-flye-m1.

assembly	length	n °contigs	N50	QUAL	n°genes	C	F	M
X367	4971547	4	4924476	56	4,210	91.4	4.9	3.7
X426	4901420	2	4884048	56	4,175	91.2	5.5	3.3
X427	5224511	3	4915292	50	4,520	86.4	6.0	7.6

As final considerations it seems that preprocessing reads is an option tied to the chosen assembler. As an example, Canu requires corrected reads and its pipelines only appear in the top 10 when using the set of corrected and trimmed reads, on the contrary, Flye and Wtdbg2 have the best scores in pipelines using the raw set of reads. The number of reads may also impact the assemblers relative performance. Strain X427 has considerably less reads than the remaining ones and Wtdbg2 performed better for it, but do not appear once in the top 10 for X367 and X426, contrary to Canu and Miniasm that have appearances in the top 10 for X367 and X426 but are not mentioned for X427. In the last step, using Racon followed by Medaka clearly results in improved assemblies [*] . After all, we consider that the choices of assembling raw reads with Flye and run one iteration of Racon and Medaka compose a quality pipeline to assemble bacterial genomes from ONT data, as already reported[†].

Hybrid assemblies

Hybrid approaches result in more complete and more contiguous assemblies than the ones from ONT and Illumina, respectively. The statistics in Table 3.5 evidence how assemblies from Illumina are less contiguous (with a minimum of 11 contigs and a maximum of 57) than assemblies from ONT data that have, at maximum, 4 contigs. On the other hand, assemblies completeness, accessed with BUSCO, is equal to 99.8% for all assemblies containing Illumina data, whilst for assemblies from ONT is inferior to 92%. These results are represented in Figure 3.5.

Among the hybrid assemblies, the approach of just polishing with Pilon is simpler than assembling with Unicycler. That approach does not assemble reads, instead, it takes an assembly from ONT and polish its contigs with Illumina reads, as consequence it is capable of improving the accuracy to the level of Illumina assemblies but the assembly contiguity remains the same from the ONT assembly.

The Unicycler pipeline assembled Illumina and ONT reads, and, also polished the assembly with Pilon. This approach results in contiguous and circular genomes, as complete as the assemblies from Illumina. It returned a circular chromosome for all strains and, for X367 and X426, it also returned a closed plasmid.

[*]Medaka benchmarks: https://nanoporetech.github.io/medaka/benchmarks.html

[†]Oxford Nanopore Technologies: Assembling microbial genomes using long nanopore sequencing reads. 2020

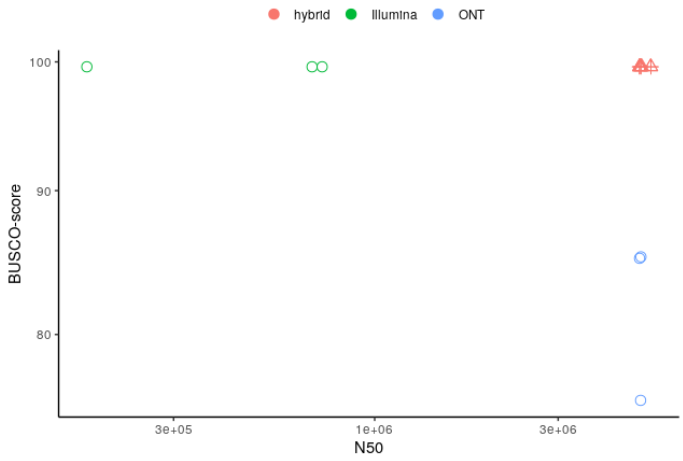


FIGURE 3.5: Contiguity and completeness of Illumina, ONT and hybrid assemblies.

TABLE 3.5: Statistics for Illumina, ONT and hybrid assemblies.

reads	assembly	length	n°contigs	N50	C	F	M
Illumina	X367-reference	4956382	22	687415	99.8	0.0	0.2
	X426-reference	4894012	11	730188	99.8	0.0	0.2
	X427-reference	5190560	57	178455	99.8	0.0	0.2
ONT	X367-r-flye-m1	4971547	4	4924476	91.4	4.9	3.7
	X426-r-flye-m1	4901420	2	4884048	91.2	5.5	3.3
	X427-r-flye-m1	5224511	3	4915292	86.4	6.0	7.6
hybrid	X367-rflye-m1-pilon	4971497	4	4924433	99.8	0.0	0.2
	X367-rflye-m2-pilon	4969660	2	4924431	99.8	0.0	0.2
	X367-rflye-m4-pilon	4969660	2	4924431	99.8	0.0	0.2
	X367-tcanu-m4-pilon	4951462	1	4951462	99.8	0.0	0.2
	X367-wtdbg2-N-pilon	4931145	2	4886483	99.8	0.0	0.2
	X426-rflye-m1-pilon	4901303	2	4883917	99.8	0.0	0.2
	X426-rflye-m4-pilon	4901272	2	4883886	99.8	0.0	0.2
	X427-rflye-m1-pilon	5224296	3	4915062	99.8	0.0	0.2
	X427-wtdbg2-N-pilon	5222448	1	5222448	99.8	0.0	0.2
	X427-wtdbg2-m4-pilon	5222497	1	5222497	99.8	0.0	0.2
	X367-unicycler	4968459	2	4923218	99.8	0.0	0.2
	X426-unicycler	4900648	2	4883254	99.8	0.0	0.2
	X427-unicycler	5228174	1	5228174	99.8	0.0	0.2

A preliminary study to estimate the minimum sequencing depth needed to achieve contiguous and complete genomes is available in Appendix C.

3.4 Conclusion

We verify that long reads from ONT enable more contiguous assemblies than short reads from Illumina. The superior error rate of long reads results in blurred regions, as supported by the analysis of BUSCO that indicates some fragmented and missed genes. Nevertheless, ONT assemblies have a completeness above 85% and are remarkably contiguous, which is presumably enough to assess structural rearrangements and mine novel DNA markers for bacterial detection/identification. The results obtained suggest that similar methodologies developed for environmental samples will likely provide new insights into unculturable bacteria genomes, and with mean read lengths superior to 10000 bp, single unassembled reads may be sufficient for detection methods.

Hybrid assemblies, in particular the ones achieved with Unicycler, reunite the advantages of both technologies and resulted in high quality genomes regarding accuracy per base, completeness and contiguity. These genomes are valuable information for comparative studies once they comprise whole chromosomes and plasmids in circular structures, which enables the inference of structural rearrangements as regions from horizontal gene transfer events and the respective functional meaning and impact in bacteria lifestyle. In addition, the disclosure of whole genetic repertoires enables the identification of gene clusters related to specialized traits as pathogenicity determinants. The Unicycler assemblies are accessible in the European Nucleotide Archive (ENA) under the numbers GCA_903989455, GCA_903989465 and GCA_903989465 for X367, X426 and X427, respectively. The genomes of X367 and X426 are also announced in the following publication: "Teixeira, M., Martins, L., Fernandes, C., Chaves, C., Pinto, J., Tavares, F., & Fonseca, N. A. (2020). Complete Genome Sequences of Walnut-Associated Xanthomonas euroxanthea Strains CPBF 367 and CPBF 426 Obtained by Illumina/Nanopore Hybrid Assembly. Microbiology Resource Announcements, 9(45)".

Chapter 4

On assembling *de novo* the genomes of 16 *Xanthomonas* spp. isolates

Abstract

Walnut associated *Xanthomonas* spp. are object of concern due to the increased recurrence of outbreaks. Recent reports highlight the co-colonization of pathogen and non-pathogen *Xanthomonas* spp. within the same walnut tree, providing an important paradigm to study the evolution of genomic determinants for host specificity and pathogenic traits.

Here, an hybrid approach was followed to assemble the genomes of sixteen *Xanthomonas* spp. isolated from symptomatic individuals of *Juglans regia* and *Carya illinoinensis*, it resulted in finished assemblies in accurate, complete and contiguous sequences. A preliminary study regarding its evolutive history demarcated two species based on the average nucleotide identity, and the observed patters of putative genetic determinants of pathogenicity were inconsistent with pathogenicity assays.

4.1 Introduction

Agricultural ecosystems form homogeneous genetic environments under permanent selection that promote the emergence of highly adapted, specialized and virulent pathogens in co-evolution with the host [3]. These traits are well noted in the genus *Xanthomonas* that, in spite of being associated with more than 400 plant species around the world, has an high host and tissue specificity that allow to discriminate in pathovars for species

causing the same disease in the same host range [6]. Additionally, the occurrence of *Xanthomonas* populations containing pathogenic and non-pathogenic strains in the same host suggest a sympatric evolution, which raises questions regarding how benign and commensal strains are involved in pathogen evolution and what conditions promote its emergence [5].

Genomic studies are the means to reconstruct this shared evolutionary history and have in fact identified genetic exchanges that shift host ranges, or cause host jumps[87, 88], and, gene clusters directly linked with virulent traits, as the Type III Effectors, whose pattern enable an *in silico* prediction of pathogenicity [89–91]. Nevertheless, we have strongly relied on incomplete genomic information that influence analysis for genetic content related with phenotypic traits and well known content as housekeeping genes[77–82]. It is therefore from utter importance, the disclosure of structurally complete and accurate genomes to couple the whole genetic content with information regarding its location and context in the genome, as well the definition and delimitation of chromosomes and secondary genetic elements. The identification of conserved genomic regions that define a genus, as well as the more dynamic regions involved in specialization, as host adaptation and pathogenic traits would help to understand the molecular basis of pathogenicity and allow us to infer how susceptible a crop is to an outbreak and how likely pathogens are to prevail from a given microbiota.

Here, we *de novo* assembled sixteen strains of walnut-associated *Xanthomonas* spp. strains following an hybrid approach combining Illumina short reads and ONT long reads. The sixteen strains were isolated from *Juglans regia* and *Carya illinoiensis* along 2 years in different regions of Portugal, composing a genetic diverse group from different environment conditions. We aimed to achieve high quality and complete genome assemblies that would increase the available genomic collection of *Xanthomonas* spp. and lay the groundwork for extensive comparative genomic studies.

4.2 Material and methods

Bacterial isolation and sequencing

Sixteen *Xanthomonas* spp. strains were isolated along 2 years from walnut trees (*Juglans regia* and *Carya illinoinensis*) in different geographic regions of Portugal as referenced in Table 4.1. The strains were grown on bacterial culture medium M2 (yeast extract, $2gL^{-1}$;

Bacto peptone, $5gL^{-1}$; $NaCl$, $5gL^{-1}$; KH_2PO_4, $0.45gL^{-1}$; Na_2HPO_4 $12H_2O$, $2.39gL^{-1}$) at 28 °C and $100rpm$ for $48h$. DNA was extracted using the E.Z.N.A. bacterial DNA purification kit (Omega Bio-tek, Norcross, GA) and sequenced with Illumina and ONT MinION platforms. Illumina sequencing for strains X367, X426, X424 and X427 was outsourced to GATC Biotech, AG (Konstanz, Germany) using a HiSeq Illumina instrument with a standard 2×150 bp paired-end library protocol while for the remaining strains were outsourced to Macrogen Inc. (Seoul, South Korea) using a TruSeq DNA kit for a 2×150 bp paired end library protocol. Nanopore sequencing libraries were prepared with the SQK-LSK109 kit and multiplexed using the EXP-NBD104 barcoding kit. Sequencing was performed on a MinION sequencer using a R9.4.1 flow cell for 36 hours. Reads were basecalled and demultiplexed using Guppy* (version 3.4.1 and version 4.0.14 with high accuracy basecalling mode). Reads statistics were calculated with FastQC (version 0.11.5 [69]) and NanoStat (version 1.2.1 [70]).

TABLE 4.1: *Xanthomonas* spp. isolates metadata.

strain	host	tree	plant samples	year	location
X122	*Juglans regia*	Jr#39	leaves	2015	Ponte de Barca
X237	*Juglans regia*	Jr#41	leaves	2015	Ponte de Lima
X367	*Juglans regia*	Jr#18	buds	2016	Loures
X424	*Juglans regia*	Jr#18	buds	2016	Loures
X426	*Juglans regia*	Jr#18	buds	2016	Loures
X427	*Juglans regia*	Jr#18	buds	2016	Loures
X554	*Juglans regia*	Jr#45	branches	2016	Carrazeda de Ansiães
X761	*Carya illinoinensis*	Cr#02	leaves	2016	Alcobaça
X765	*Carya illinoinensis*	Cr#03	leaves	2016	Alcobaça
X766	*Carya illinoinensis*	Cr#03	leaves	2016	Alcobaça
X796	*Juglans regia*	Jr#61	leaves	2016	Alcobaça
X1483	*Juglans regia*	Jr#02	branches	2014	Alcobaça
X1494	*Carya illinoinensis*	Cr#02	leaves	2014	Alcobaça
X1514	*Juglans regia*	Jr#15	leaves	2014	Estremoz
X1567	*Juglans regia*	Jr#30	fruit	2015	Bombarral
X1586	*Juglans regia*	Jr#18	leaves	2015	Loures

*Only available to Oxford Nanopore customers through their community site (http://community.nanoporetech.com)

Complete *De novo* hybrid assemblies

Illumina and ONT data was combined with the aim to achieve structurally complete assemblies with low errors at base level. The assemblies were generated with the Unicycler pipeline (version 0.4.8[61]) for all strains except X1567, where Unicycler was unable to solve the genomic structure. Therefore, for X1567, the genome was assembled *de novo* using an alternate approach: ONT reads were assembled with Flye (version 2.7-b1585[71]) and the resulting assembly was improved using one iteration of Racon (version 1.4.3[73]) and Medaka* (version0.11.5); Circlator (version 1.5.5 [92]) was used to verify its circularity and to rotate it, in order to start the sequence with the dnaA gene (as the remaining assemblies from Unicycler); Finally, the assembly was polished with the Illumina reads using Pilon (version 1.23 [85]). All assemblies were annotated with PGAP (version 2020-03-30.build4489[93]) and its metrics calculated with QUAST (version 5.0.2 [64]). The confidence of each base was estimated by realigning Illumina reads in the assembly with bwa (version 0.7.17-r1188[94]) and samtools (version 1.8[95]). From this realignment, bcftools functions mpileup and call (version 1.9[96]), calculate, respectively, the possible variants and their likelihood in a phred quality scale. The phred quality values per base were converted to probabilities and calculate its mean as an indicator of the expected error per base for the whole genome. It is presented again in a phred scale.

Contigs circularity is indicated by Unicycler and Circlator. The assembly graphs from Unicycler and Flye were generated with Bandage (version 0.8.1[97]), that was also used to map contigs from the X1567 Unicycler assembly in the respective alternative assembly. Genome completeness was assessed with BUSCO (version 4.0.6, database xanthomonadales_odb10 [65, 66]). For the missing genes, we align the corresponding protein sequence against the assembly using tblastn (version 2.10.1 [98]) and compare the best hit to the PGAP annotation for that region.

Comparative genomics

Average nucleotide identity (ANI)[99], an *in silico* alternative for DNA-DNA hybridization, was calculated using OrthoANI (version 1.40[100]) with the usual threshold of 95% identity for isolates from the same species[101].

*Sequence correction provided by ONT Research: https://github.com/nanoporetech/medaka

The presence of four type III effector genes (T3E) has been reported as an indicator of pathogenicity among *Xanthomonas* spp.[89–91]. Correspondent protein sequences are accessible in the database UniProt as indicated in Table 4.2.

TABLE 4.2: T3E accession numbers

T3E gene	accession number	n°of residues
avrBs2	Q3BZN0_XANC5	714
xopF1	Q3BYL8_XANC5	670
xopN	K4LAY3_XANOO	621
xopR	H6UWS2_XANOO	437

To evaluate the presence of these T3E in our isolates, the protein sequences were aligned against our isolates transcriptomes. For that, we use Bedtools (version 2.26.0[102]) to get the individual coding regions indicated in PGAP annotations from the assemblies FASTA files. The alignment was performed with tblastn. The E-value, Bit-score, alignment identity and length, as well as the functional annotation of PGAP was taken into account to infer the presence of a gene. In addition, for each protein, we use transeq from EMBOSS (version 6.6.0.0 [103]) to translate the nucleotide sequences from the best hit in each assembly and perform a multiple alignment with TCoffee (version 11.00.8cbe486[104]).

Phylogenetic relations were inferred using three approaches:

- A distance matrix was calculated based on the ANI data;
- Alfpy (calc_word version 1.0.6[105]) was used to compare the genomic sequences based on the frequency of k-length words, (considered k=3,5,7,9,11,15,20);
- The genes used by BUSCO (highly conserved, single-copy orthologous genes). Considering that our isolates all belong to *Xanthomonas* spp. associated with walnut trees, we presume that they all have been under the similar selective pressure and therefore compose an adequate record to evaluate divergences among our isolates. We aligned the 1152 BUSCO proteins for *Xanthomonadales* in our assemblies using tblastn and took the nucleotide sequence corresponding to the best hit. Next, we individually align the genes from all species using TCoffee and concatenate those alignments to build a phylogenetic tree with RAxML (version 8.2.11 [106]) using the model $GTR + \gamma$ and a bootstrap value of 500.

We include two complete genome sequences of *Xanthomonas* spp. as outgroups in this phylogenetic analysis. Those sequences, from a *X. campestris* and a *X. citri*, are accessible in the ATCC collection with the identifiers ATCC33913 and ATCC4918, respectively. A visual representation of the approaches was achieved with Figtree (version 1.4.4 [*]).

The R package ctc[†] (version 1.62.0.) was used to convert the distance matrices to phylogenetic trees. Consensus trees were generated with the Phylip consense module[107].

4.3 Results and discussion

The results from ONT and Illumina sequencing are available in Appendix E.

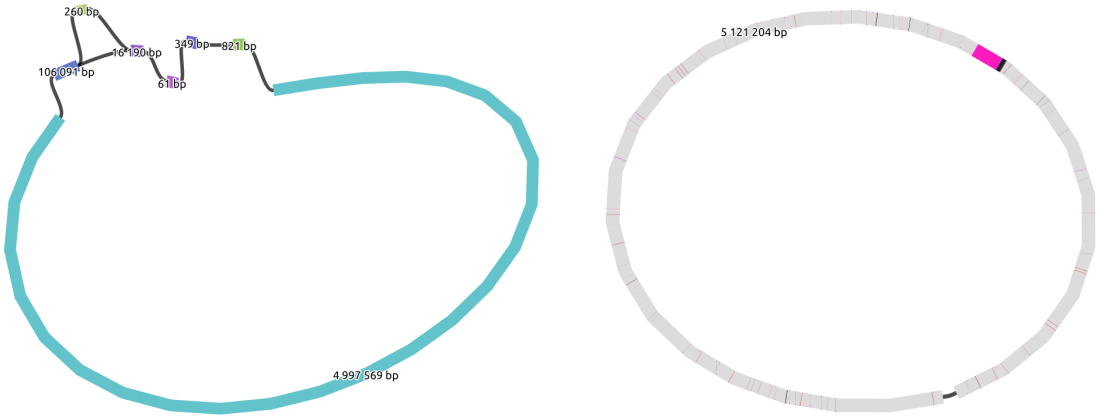
De novo complete assemblies

The Unicycler pipeline enabled assemblies in circular contigs for all stains, except X1567. Samples X237, X367 and X426 have one plasmid each. Representations of assemblies graphs, obtained with Bandage, are accessible in Figure D.1 and Figure D.2 in Appendix D. Unicycler was incapable to resolve the genomic structure of X1567, as represented in Figure 4.1, resulting in 6 contigs with lengths indicated in Table 4.3. The assembly graph representation in Figure 4.1a shows that most of genome sequence is in a single contig, and that the second and third longer contigs can be directly linked or have an insertion of 260 bp. The other contigs are inferior to 821 bp. Since the fragmented region is composed of contigs smaller than an average ONT read we speculated that the Unicycler inability to resolve the whole structure was related to short reads. Unicycler produces an Illumina assembly and than uses its contigs and the long reads to produce the complete hybrid assembly, therefore, the final result may be conditioned by errors in the Illumina assembly, i.e., the missassemble of short reads can result in artifacts that roughly overlap with long reads making impossible to establish the bridges that should link them. A contiguous circular assembly for the genome of X1567 was achieved from the ONT data assembled with the pipeline r-flye-m1 (with Flye, Racon and Medaka, see chapter 3 for more details) and polished with Pilon using Illumina data. The graph representation of the assembly is presented in Figure 4.1b along the matches with contigs from the Unicycler assembly. Concerning completeness, both assemblies have very similar annotations as indicated in Table 4.4.

[*]FigTree: graphical viewer of phylogenetic trees. Rambaut A. 2010
[†]ctc: Cluster and Tree Conversion. Lucas A, Gautier L. 2020

TABLE 4.3: Contigs length for X1567 assembly from Unicycler.

contig n°	1	2	3	4	5	6
length (bp)	4997569	106091	16190	821	349	260



(A) Unicycler assembly

(B) r-flye-m1-pilon assembly

FIGURE 4.1: Assembly graphs highlighting the contiguity of X1567 assemblies.

TABLE 4.4: PGAP annotation for X1567 assemblies.

Features	Unicycler	r-flye-m1-pilon
Genes (total)	4,361	4,366
CDSs (total)	4,258	4,263
Genes (coding)	4,128	4,128
CDSs (with protein)	4,128	4,128
Genes (RNA)	103	103
rRNAs (5S, 16S, 23S)	2, 2, 2	2, 2, 2
complete rRNAs (5S, 16S, 23S)	2, 2, 2	2, 2, 2
tRNAs	53	53
ncRNAs	44	44
Pseudo Genes (total)	130	135
CDSs (without protein)	130	135

Missing BUSCO genes

The analysis with BUSCO to assess the genetic completeness indicates that some strains miss the following genes:

- **28445at135614**: BFD-like
- **14127at135614**: Band 7, C-terminal extension
- **10621at135614**: tRNA 2-thiocytidine(32) synthetase TtcA
- **17339at135614**: Methyltransferase
- **17191at135614**: Short-chain dehydrogenase/reductase, conserved site

All the genes missed by BUSCO were detected by PGAP. The PGAP annotation, location in the genome and the BUSCO id are indicated in Table 4.5. A possible explanation for BUSCO not detecting these genes is the divergence between the assembled sequences and the BUSCO sequences that, for the order *Xanthomonadales*, are consensus sequences of two families (*Xanthomonadaeceae* and *Rhodanobacteraceae*).

Statistics regarding the structure and genetic content of the assemblies are present in Table 4.6.

TABLE 4.5: PGAP annotation for the missing BUSCO genes.

strain	BUSCO id	location	annotation
X122	28445at135614	661032..661241	(2Fe-2S)-binding protein
X237	14127at135614	1545764..1546903	FtsH protease activity modulator HflK
	28445at135614	4594436..4594645	(2Fe-2S)-binding protein
X367	14127at135614	1328508..1329647	FtsH protease activity modulator HflK
	28445at135614	581829..582038	(2Fe-2S)-binding protein
X424	14127at135614	1314525..1315664	FtsH protease activity modulator HflK
	28445at135614	623421..623630	(2Fe-2S)-binding protein
X426	14127at135614	1313144..1314283	FtsH protease activity modulator HflK
	28445at135614	591484..591693	(2Fe-2S)-binding protein
X427	10621at135614	4963579..4964504	tRNA 2-thiocytidine(32) synthetase TtcA
	28445at135614	4507993..4508202	(2Fe-2S)-binding protein
x554	28445at135614	3215955..3216200	(2Fe-2S)-binding protein
X761	14127at135614	1283101..1284240	FtsH protease activity modulator HflK
	28445at135614	591491..591700	(2Fe-2S)-binding protein
X765	-	-	-
X766	14127at135614	1308899..1310038	FtsH protease activity modulator HflK
	28445at135614	617140..617349	(2Fe-2S)-binding protein
X796	14127at135614	3660414..3661553	FtsH protease activity modulator HflK
	17339at135614	4258366..4259433	Methyltransferase
	28445at135614	642392..642601	(2Fe-2S)-binding protein
X1483	10621at135614	203387..204312	tRNA 2-thiocytidine(32) synthetase TtcA
	28445at135614	658449..658658	(2Fe-2S)-binding protein
X1494	28445at135614	68435..68926	(2Fe-2S)-binding protein
X1514	-	-	-
X1567	14127at135614	3689195..3690334	FtsH protease activity modulator HflK
	17339at135614	4296617..4297684	Methyltransferase
	28445at135614	637485..637694	(2Fe-2S)-binding protein
X1586	17191at135614	2714089..2714592	SDR family NAD(P)-dependent oxidoreductase
	17339at135614	4363064..4364131	Methyltransferase
	28445at135614	667517..667726	(2Fe-2S)-binding protein

TABLE 4.6: Assemblies statistics.

strain	contigs	length (bp)	GC content (%)	N50	QUAL	CDSs	Genes (coding)	RNAs	rRNAs (5S, 16S, 23S)	tRNAs	ncRNAs	Pseudo Genes
X122	1	4987644	65.62	4987644	98	4.175	4.271 (4.116)	96	2, 2, 2	53	37	59
X237	2	5276251	65.19	5256807	80	4.430	4.534 (4.307)	104	2, 2, 2	53	45	123
X367	2	4968459	65.81	4923218	69	4.077	4.157 (4.012)	80	2, 2, 2	56	18	65
X424	1	4900930	65.88	4900930	82	4.040	4.119 (3.993)	79	2, 2, 2	56	17	47
X426	2	4900648	65.85	4883254	68	4.066	4.143 (4.003)	77	2, 2, 2	53	18	63
X427	1	5228174	65.38	5228174	79	4.367	4.465 (4.237)	98	2, 2, 2	54	38	130
X554	1	5299179	65.4	5299179	83	4.440	4.540 (4.295)	100	2, 2, 2	53	41	145
X761	1	4853125	65.92	4853125	91	4.016	4.093 (3.953)	77	2, 2, 2	53	18	63
X765	1	5114374	65.54	5114374	98	4.291	4.385 (4.229)	94	2, 2, 2	53	35	62
X766	1	4806525	66.08	4806525	82	4.005	4.084 (3.956)	79	2, 2, 2	53	20	49
X796	1	5082336	65.45	5082336	76	4.214	4.317 (4.094)	103	2, 2, 2	53	44	120
X1483	1	5238196	65.39	5238196	73	4.389	4.487 (4.260)	98	2, 2, 2	54	38	129
X1494	1	4997863	65.69	4997863	71	4.204	4.295 (4.154)	91	2, 2, 2	54	31	50
X1514	1	5110073	65.55	5110073	94	4.287	4.381 (4.225)	94	2, 2, 2	53	35	62
X1567	1	5122148	65.38	5122148	67	4.263	4.366 (4.128)	103	2, 2, 2	53	44	135
X1586	1	5101599	65.56	5101599	67	4.237	4.327 (4.167)	90	2, 2, 2	53	31	70

Comparative genomics

The samples can be clustered in two groups considering the usual threshold of 95% of ANI for a species (the average nucleotide identities are provided in Appendix F, Table F.1). One group contains X122, X1483, X1494, X1567, X1586, X237, X427, X554, X765 and X796, and the group contains X367, X424, X426, X761 and X766. Interestingly, the groups do not reflect the hosts and both include strains isolated from *Juglans regia* and *Carya illinoinensis*. From previous studies it is known that X427 is a *X. arboricola* pv. *juglandis* and X424 is from the novel *X. euroxanthea* [7, 8, 32]. Therefore, as all strains share an ANI superior to 95% with one of these, an identification could be attributed to all. Henceforth, the acronyms Xaj and Xe will be used to refer to samples from *X. arboricola* and *X. euroxanthea*.

Annotation and alignment statistics for *avrBs2* are provided in Table 4.7, the multiple alignment is provided in Appendix F, section F.2. The whole *avrBs2* sequence (714 residues long) is aligned with more than 70% identities in Xaj strains. Contrary, only 8% of the *avrBs2* sequence was aligned in Xe strains. This result is explained by the length of Xe sequences that have approximately half the residues (around 300) than the reference and Xajs sequences. Nevertheless, PGAP annotation for all Xe strains indicates a glycerophosphodiester phosphodiesterase family protein, which is consistent with the annotation for most Xaj strains. The multiple alignment shows conserved domains among all strains and major missing regions for all Xe. Thus, the gene *avrBs2* should be present in all strains, but incomplete for Xe.

TABLE 4.7: Annotation and alignment statistics for *avrBs2* best hit.

strain	Bit-score	E-value	length (%)	identities (%)	annotation
Xaj122	995	<1e-10	100	77	avirulence protein
Xaj237	978	<1e-10	100	75	glycerophosphodiester phosphodiesterase family protein
Xe367	57.4	2.8e-9	8	48	glycerophosphodiester phosphodiesterase family protein
Xe424	57.4	2.8e-9	8	48	glycerophosphodiester phosphodiesterase family protein
Xe426	57.8	2.8e-9	8	48	glycerophosphodiester phosphodiesterase family protein
Xaj427	978	<1e-10	100	75	glycerophosphodiester phosphodiesterase family protein
Xaj554	978	<1e-10	100	75	glycerophosphodiester phosphodiesterase family protein
Xe761	57.8	2.8e-9	8	48	glycerophosphodiester phosphodiesterase family protein
Xaj765	998	<1e-10	100	76	glycerophosphodiester phosphodiesterase family protein
Xe766	57.4	2.8e-9	8	48	glycerophosphodiester phosphodiesterase family protein
Xaj796	978	<1e-10	100	75	glycerophosphodiester phosphodiesterase family protein
Xaj1483	978	<1e-10	100	75	glycerophosphodiester phosphodiesterase family protein
Xaj1494	994	<1e-10	100	76	avirulence protein
Xaj1514	998	<1e-10	100	76	glycerophosphodiester phosphodiesterase family protein
Xaj1567	978	<1e-10	100	75	glycerophosphodiester phosphodiesterase family protein
Xaj1586	1000	<1e-10	100	77	avirulence protein

xopF1 was not found in Xe367, Xe426 and Xe761 but detected in the remaining strains. Annotation and alignment statistics for *xopF1* best hit are present in Table 4.8. For the multiple alignment see Appendix F, section F.3. The reference sequence (670 residue long) was completely aligned for most strains with similarities around 67%. The exceptions are Xe367, that result from not finding a significant alignment, and the strains Xe426 and Xe761. For these two strains, the best hit aligns 6% in a sequence of around 840 residues long (annotated as a phosphotransferase). Nevertheless, further investigation is required, since PGAP annotation is not conclusive and the multiple sequence alignment suggests the existence of rearrangements among strains, in particular for Xaj237, Xaj427 and Xaj554.

TABLE 4.8: Annotation and alignment statistics for *xopF1* best hit.

strain	Bit-score	E-value	length (%)	identities (%)	annotation
Xaj122	704	<1e-10	100	67	hypothetical protein
Xaj237	684	<1e-10	99	67	hypothetical protein
Xe367	NA	NA	NA	NA	NA
Xe424	693	<1e-10	100	67	hypothetical protein
Xe426	27.7	7.5	6	43	phosphoenolpyruvate–protein phosphotransferase
Xaj427	686	<1e-10	99	67	hypothetical protein
Xaj554	686	<1e-10	99	67	hypothetical protein
Xe761	27.7	7.4	6	43	phosphoenolpyruvate–protein phosphotransferase
Xaj765	708	<1e-10	100	67	hypothetical protein
Xe766	684	<1e-10	100	67	hypothetical protein
Xaj796	686	<1e-10	99	67	hypothetical protein
Xaj1483	686	<1e-10	99	67	hypothetical protein
Xaj1494	703	<1e-10	100	68	hypothetical protein
Xaj1514	708	<1e-10	100	67	hypothetical protein
Xaj1567	686	<1e-10	99	67	hypothetical protein
Xaj1586	701	<1e-10	100	68	hypothetical protein

xopN protein was only detected in Xaj237, Xaj427, Xaj554, Xaj796, Xaj1483 and Xaj1567 (statistics from the best alignment are present in Table 4.9), where a significant alignment was found with almost 70% identity to sequences annotated as T3E (by PGAP). These sequences have inserted regions that make the alignment longer than the reference sequence, therefore the alignment length, calculated as percentage of reference sequence length, is greater than 100%. Among these six strains there seems to be two lineages, one for Xaj237, Xaj427 and Xaj554 and another for Xaj796, Xaj1483 and Xaj1567 that are more similar to the reference as noted in the multiple alignment (see Appendix F, section F.4). For the remaining strains, the best hit is annotated as an enzyme involved in the biosynthesis of histidine. These sequences are shorter (around 200 residues instead of

the expected 437) and the respective alignments only comprises 25% of the reference in sequences.

TABLE 4.9: Annotation and alignment statistics for *xopN* best hit.

strain	Bit-score	E-value	length (%)	identities (%)	annotation
Xaj122	30	1.1	25	25	imidazole glycerol phosphate synthase subunit HisH
Xaj237	723	<1e-10	132	68	type III secretion system effector protein
Xe367	30.8	0.57	25	25	imidazole glycerol phosphate synthase subunit HisH
Xe424	30.4	0.66	25	25	imidazole glycerol phosphate synthase subunit HisH
Xe426	30.4	0.66	25	25	imidazole glycerol phosphate synthase subunit HisH
Xaj427	723	<1e-10	132	68	type III secretion system effector protein
Xaj554	723	<1e-10	132	68	type III secretion system effector protein
Xe761	30.4	0.65	25	25	imidazole glycerol phosphate synthase subunit HisH
Xaj765	30	1.1	25	25	imidazole glycerol phosphate synthase subunit HisH
Xe766	30.8	0.55	25	25	imidazole glycerol phosphate synthase subunit HisH
Xaj796	723	<1e-10	132	68	type III secretion system effector protein
Xaj1483	723	<1e-10	132	68	type III secretion system effector protein
Xaj1494	30	1.1	25	25	imidazole glycerol phosphate synthase subunit HisH
Xaj1514	30	1.1	25	25	imidazole glycerol phosphate synthase subunit HisH
Xaj1567	723	<1e-10	132	68	type III secretion system effector protein
Xaj1586	27.7	5	25	25	imidazole glycerol phosphate synthase subunit HisH

xopR is present and well conserved among all strains, but diverges considerably from the reference sequence (see Table 4.10 for alignment statistics and Appendix F,section F.5 for the multiple sequence alignment). Xaj and Xe sequences have less 300 residues than the reference sequence (720 residues), therefore, alignments length correspond to just 21% and the identities are inferior than 50%. Nevertheless, PGAP annotates all sequences as a type III secretion protein.

TABLE 4.10: Annotation and alignment statistics for *xopR* best hit.

strain	Bit-score	E-value	length (%)	identities (%)	annotation
Xaj122	138	<1e-10	21	45	type III secretion protein
Xaj237	139	<1e-10	21	45	type III secretion protein
Xe367	144	<1e-10	21	46	type III secretion protein
Xe424	141	<1e-10	21	45	type III secretion protein
Xe426	139	<1e-10	21	45	type III secretion protein
Xaj427	137	<1e-10	21	44	type III secretion protein
Xaj554	137	<1e-10	21	44	type III secretion protein
Xe761	139	<1e-10	21	45	type III secretion protein
Xaj765	140	<1e-10	21	45	type III secretion protein
Xe766	145	<1e-10	21	46	type III secretion protein
Xaj796	139	<1e-10	21	45	type III secretion protein
Xaj1483	137	<1e-10	21	44	type III secretion protein
Xaj1494	138	<1e-10	21	45	type III secretion protein
Xaj1514	140	<1e-10	21	45	type III secretion protein
Xaj1567	139	<1e-10	21	45	type III secretion protein
Xaj1586	139	<1e-10	21	45	type III secretion protein

A complete set of T3E suggests that a given strain is pathogenic. As Xe424 has confirmed pathogenicity hence other strains sharing the T3E pattern are presumably pathogens too. Contrary, Xe367 is confirmed as non-pathogen and thus, all strains with its pattern should be commensal[7].

Figure 4.2 depicts the consensus tree of the three approaches. Individual trees based on the ANI, the BUSCO genes and the the k-mer frequencies are represented in Figure G.1, Figure G.2 and Figure G.3 in Appendix G.

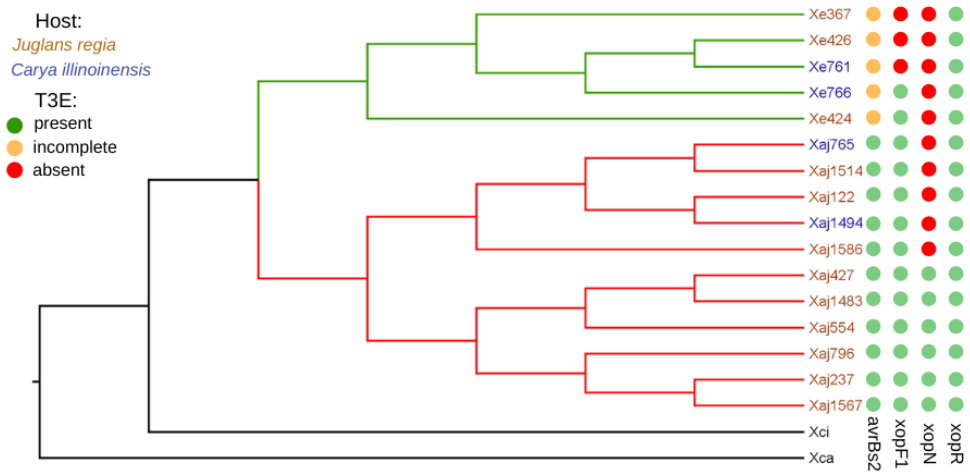


FIGURE 4.2: Species, host, presence of T3E and consensus phylogeny tree for the sixteen isolates.

4.4 Conclusion

Numerous whole-genome sequences of walnut-associated *Xanthomonas arboricola* are currently available at the NCBI Genome Resources, providing a valuable pangenome patrimony to scrutinize putative pathoadaptations. However, most of these genomes are composed of variable numbers of contigs or scaffolds, making difficult to link their genomic landscape with specific pathoadaptations using comparative genomics analysis. Therefore, it is not surprising the efforts that are being made to obtain complete circular genome sequences in a single contig through hybrid assembly of Illumina short-reads and ONT long-reads. In fact, hybrid assemblies led to high quality genomes, regarding genome contiguity, completeness and accuracy.

From previous results (see chapter 3) we determined that the Unicycler approach to assemble long and short reads was most likely to result in contiguous assemblies. On the contrary, in this chapter, we observed the inability of Unicycler pipeline to obtain the circular chromosomal sequence in a single contig using both Illumina short-reads an ONT long-reads for one (X1567) of the sixteen strains analysed in this study. The hybrid assembling obtained with Unicycler for strain X1567, resulted in six contigs, that we hypothesize to be due to misassemblies in the short reads leading to assembling artifacts.

In order to assess the genomic contents, i.e. completeness, each of the sixteen *Xanthomonas* genomes sequenced were analyzed using BUSCO. Although BUSCO revealed the absence of five genes, these genes were retrieved by PGAP, suggesting that the no gene losses occurred during the assembling.

The clustering genomics analysis of the sixteen xanthomonads genomes confirmed the existence of two distinct species, with five strains belonging to the recently proposed species *Xanthomonas euroxanthea* [8], while the remaining eleven strains belong to *Xanthomonas arboricola* pv. *juglandis*. The three putative non-pathogenic strains (Xe367, Xe426, Xe761) are all *X. euroxanthea*, but two were isolated in *Juglans regia* (Xe367 and Xe426) and one in *Carya illinoinensis* (Xe761). This data further suggests that the dendogram clusters do not reflect host specificity neither putative pathogenicity, indicating that these traits are likely linked to genetic determinants transversal to both *Xanthomonas* species.

The screening of four type III effectors (T3Es) acknowledged as important determinants of virulence and pathogenicity in *Xanthomonas*, revealed interesting differences. In fact, while *xopR* was observed to be present in all the sixteen genomes, which include both pathogenic and non-pathogenic strains [8] , the other three T3E revealed a heterogeneous pattern. *xopN* was absent from the five Xe strains and some of the Xaj strains, suggesting that is not essential for pathogenicity. On the contrary, *xopF1* was present in the eleven Xaj and in two Xe strains, including strain Xe424 which was shown to be pathogenic, but absent in the remaining three Xe which include the non-pathogenic strain Xe367. Regarding the T3E *avrBs2*, the results reveal that although gene homologs have been retrieved in the 16 genomes, there are striking differences between Xe and Xaj characterized by major and consistent deletions in the *avrBs2* gene sequences in Xe strains comparatively to Xaj. Taken together, the distinct profiles of these T3E in Xe and Xaj, call for functional genetic studies that may definitively address their importance for pathogenicity and virulence.

Chapter 5

Conclusion

An objective of this work was to optimize a pipeline to assemble bacterial genomes from long reads. The pipeline defined assemble reads with Flye and correct the draft assemblies with one iteration of Racon and Medaka. It results in contiguous assemblies, with few (and often one) contig for each chromosome/plasmid. Although these assemblies are less accurate and less complete than assemblies from short reads, the completeness achieved from the accurate basecalling is around 90%. Thus, we can conclude that long read assemblies can recover most the genomic repertoire and enable insights in structural rearrangements involved in bacterial adaptation processes.

The use of Illumina reads to polish ONT assemblies with Pilon, effectively improve accuracy and completeness, yet, that approach has no impact in the genome contiguity, that remains the same originated by the long reads assembly. This could be enhanced using the Unicycler pipeline to assemble ONT long reads and Illumina short reads, which resulted in high quality assemblies, regarding accuracy, completeness and contiguity. For Xaj1567, Unicycler was unable to resolve the whole genome, presumably due to a misassemble of short reads, hence, for that strain we assembled just the long reads with the pipeline mentioned above and polished the draft assembly with Illumina reads. In the end, sixteen assemblies with circular chromosomes and plasmids were obtained. This data will enrich the public collection of *Xanthomonas* spp. genomes, with complete assemblies of *X. arboricola* and the first complete assemblies of *X. euroxanthea*.

The phylogenies of these strains did not reflect host specialization, and the genomic determinants of pathogenicity and virulence remain an open question. As the epiphytic and endophytic bacteria consortia colonizing a host may include populations of pathogenic and non-pathogenic bacteria, genomic approaches supported by these assemblies, may

help to unveil genetic trade-offs and clarify the role played by the different members of the bacteria consortia during a disease process. Well known virulence determinants, such as the type III effector, coded by *avrBs2*, has been observed to vary considerably between Xe and Xaj, which biological meaning is still unknown. Lastly, plasmids are not usual in walnut-associated *Xanthomonas*, thus its existence in Xaj237, Xe367 and Xe426 should be experimentally assessed, and its content ascertained.

In sum, this work comprises a basis for further genomic studies addressing the paradigm of walnut diseases emerging from *Xanthomonas* populations. We conclude that long read sequencing methodologies can recover most of a bacterial genetic patrimony and promote whole genome assemblies in contiguous sequences. Furthermore, due to the portability of MinION device, long reads are a promising asset to detect and identify bacteria on the field, in real time and without the traditional succession of cultures in the laboratory.

Appendix A

Analysis details on assembling reads from the accurate basecalling model

→ **Impact of the preprocessing choices in the assembled genomes**

Mann-Whitney U tests ($\alpha < 0.01$):

- **raw reads**

  H0: $overall\ score\{raw\} = overall\ score\{corrected, corrected + trimmed\}$

  H1: $overall\ score\{raw\} \neq overall\ score\{corrected, corrected + trimmed\}$

  $$W = 856.5, p = 0.36$$

- **corrected reads**

  H0: $overall\ score\{corrected\} = overall\ score\{raw, corrected + trimmed\}$

  H1: $overall\ score\{corrected\} \neq overall\ score\{raw, corrected + trimmed\}$

  $$W = 1226.5, p = 0.50$$

- **corrected+trimmed reads**

  H0: $overall\ score\{corrected + trimmed\} = overall\ score\{raw, corrected\}$

  H1: $overall\ score\{corrected + trimmed\} \neq overall\ score\{raw, corrected\}$

  $$W = 1157, p = 0.87$$

→ **Which assembler performs better?**

Mann-Whitney U tests ($\alpha < 0.01$):

- **Canu**

  H0: $overall\ score\{Canu\} = overall\ score\{Miniasm, Wtdbg2, Flye\}$

  H1: $overall\ score\{Canu\} \neq overall\ score\{Miniasm, Wtdbg2, Flye\}$

  $$W = 844.5, p = 0.30$$

- **Miniasm**

  H0: $overall\ score\{Miniasm\} = overall\ score\{Canu, Wtdbg2, Flye\}$

  H1: $overall\ score\{Miniasm\} \neq overall\ score\{Canu, Wtdbg2, Flye\}$

  $$W = 1109.5, p = 0.28$$

- **Wtdbg2**

  H0: $overall\ score\{Wtdbg2\} = overall\ score\{Canu, Miniasm, Flye\}$

  H1: $overall\ score\{Wtdbg2\} \neq overall\ score\{Canu, Miniasm, Flye\}$

  $$W = 1096.5, p = 0.33$$

- **Flye**

  H0: $overall\ score\{Flye\} = overall\ score\{Canu, Miniasm, Wtdbg2\}$

  H1: $overall\ score\{Flye\} \neq overall\ score\{Canu, Miniasm, Wtdbg2\}$

  $$W = 594.5, p = 0.0030$$

  As pipelines with Flye had a different average overall score, we repeated a Mann-Whitney U test for one side and verified, above the 99% confidence level, that its overall score is inferior, i.e., Flye performed significantly better than the other assemblies.

  H0: $overall\ score\{Flye\} \geq overall\ score\{Canu, Miniasm, Wtdbg2\}$

  H1: $overall\ score\{Flye\} < overall\ score\{Canu, Miniasm, Wtdbg2\}$

  $$W = 594.5, p = 0.0015$$

→ **Impact of improving draft assemblies**

Mann-Whitney U test ($\alpha < 0.01$):

H0: $overall\ score\{with\ improvement\} \geq overall\ score\{without\ improvement\}$

H1: $overall\ score\{with\ improvement\} < overall\ score\{without\ improvement\}$

$$W = 922, p = 5.5 \times 10^{-7}$$

Performing a Wilcoxon signed rank test ($\alpha < 0.01$) we verify that pipelines with Medaka had lower scores than the equivalent ones with just Racon, this evidence is represented in Figure 3.4e.

H0: $overall\ score\{Racon + Medaka\} \geq overall\ score\{Racon\}$

H1: $overall\ score\{Racon + Medaka\} < overall\ score\{Racon\}$

$$V = 990, p = 3.9 \times 10^{-9}$$

→ **How many Racon corrections should we perform?**

There was no significant difference in running Racon more than once before Medaka.

Mann-Whitney U tests ($\alpha < 0.01$):

- **1 iteration**

  H0: $overall\ score\{1iteration\} = overall\ score\{1, 2, 4iterations\}$

  H1: $overall\ score\{1iteration\} \neq overall\ score\{1, 2, 4iterations\}$

  $$W = 242.5, p = 0.10$$

- **2 iterations**

  H0: $overall\ score\{2iterations\} = overall\ score\{1, 3, 4iterations\}$

  H1: $overall\ score\{2iterations\} \neq overall\ score\{1, 3, 4iterations\}$

  $$W = 189, p = 0.85$$

- **3 iterations**

  H0: $overall\ score\{3iterations\} = overall\ score\{1, 2, 4iterations\}$

  H1: $overall\ score\{3iterations\} \neq overall\ score\{1, 2, 4iterations\}$

  $$W = 144, p = 0.32$$

- **4 iterations**

  H0: $overall\ score\{4iterations\} = overall\ score\{1, 2, 3iterations\}$

  H1: $overall\ score\{4iterations\} \neq overall\ score\{1, 2, 3iterations\}$

$$W = 150.5, p = 0.41$$

What is the best pipeline?

In Table A.1 are listed the individual scores of the 10 best pipelines for each strain. In general, pipelines that assembled raw reads with Flye combined with Racon and Medaka performed better.

TABLE A.1: Top 10 pipelines per strain.

X367	score	X426	score	X427	score
t-canu-m4	14	r-flye-m1	13	r-wtdbg2-m4	17
t-flye-m1	19	r-flye-m2	13	r-wtdbg2-m2	20
t-canu-m3	21	r-flye-m4	14	r-wtdbg2-m3	25
t-canu-m2	23	r-flye-m3	23	r-flye-m1	28
r-flye-m1	24	c-flye-m3	27	r-flye-m4	32
t-flye-m3	24	c-flye-m1	31	c-wtdbg2-m1	33
c-flye-m1	28	c-flye-m2	31	r-flye-m3	34
r-flye-m2	29	c-miniasm-m4	38	r-flye-m2	35
r-miniasm-m4	30	t-miniasm-m2	39	c-wtdbg2-m4	36
c-flye-m4	31	c-miniasm-m3	43	c-wtdbg2-m3	38

Appendix B

Assembling reads from the fast basecalling model and basecalling models comparison

The fast basecalling mode resulted in more reads and bases than the accurate mode, yet, its reads are shorter and have an inferior quality (a phred quality score around 11.1, corresponding to an error probability of 7.8%) than the accurate basecalled reads (a quality score around 11.8, corresponding to an error probability of 6.6%). The N50 for all ONT datasets indicates that half the bases sequenced are in reads longer than 10,000 bases (see Table B.1).

TABLE B.1: ONT sequencing data statistics.

strain	dataset	n°of reads	mean length	mean quality	N50	coverage
X367	ONT (fast)	24,826	5,571	11.0	13,082	27
	ONT (acc)	18,857	6,386	11.8	14,257	24
X426	ONT (fast)	25,187	5,364	11.1	10,015	27
	ONT (acc)	19,997	5,895	11.9	10,553	23
X427	ONT (fast)	16,589	5,980	11.1	12,894	19
	ONT (acc)	14,302	6,358	11.8	13,332	18

The number of reads, mean read length and mean read quality are represented for both ONT and the Illumina datasets in Figure B.1.

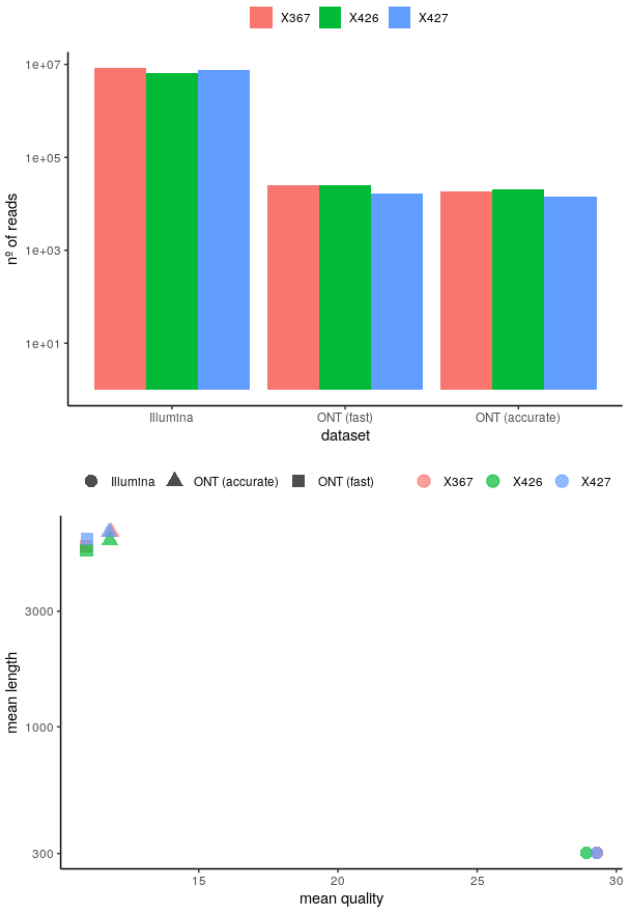


FIGURE B.1: Yield, reads length and quality for ONT and Illumina data.

The values distribution for the different variables considered are presented in Figure B.2, and are similar to accurate basecalling ones. BUSCO-score, single nucleotide differences regarding a reference assembly and number of differences in core genes, have similar, slightly skewed distributions, for the three samples, with worst performances for X427. Most pipelines produced assemblies containing all core genes and most BUSCO's, yet, there is a notorious skewness for X427. For the variables related to the housekeeping genes, all assemblies contained the seven housekeeping genes.

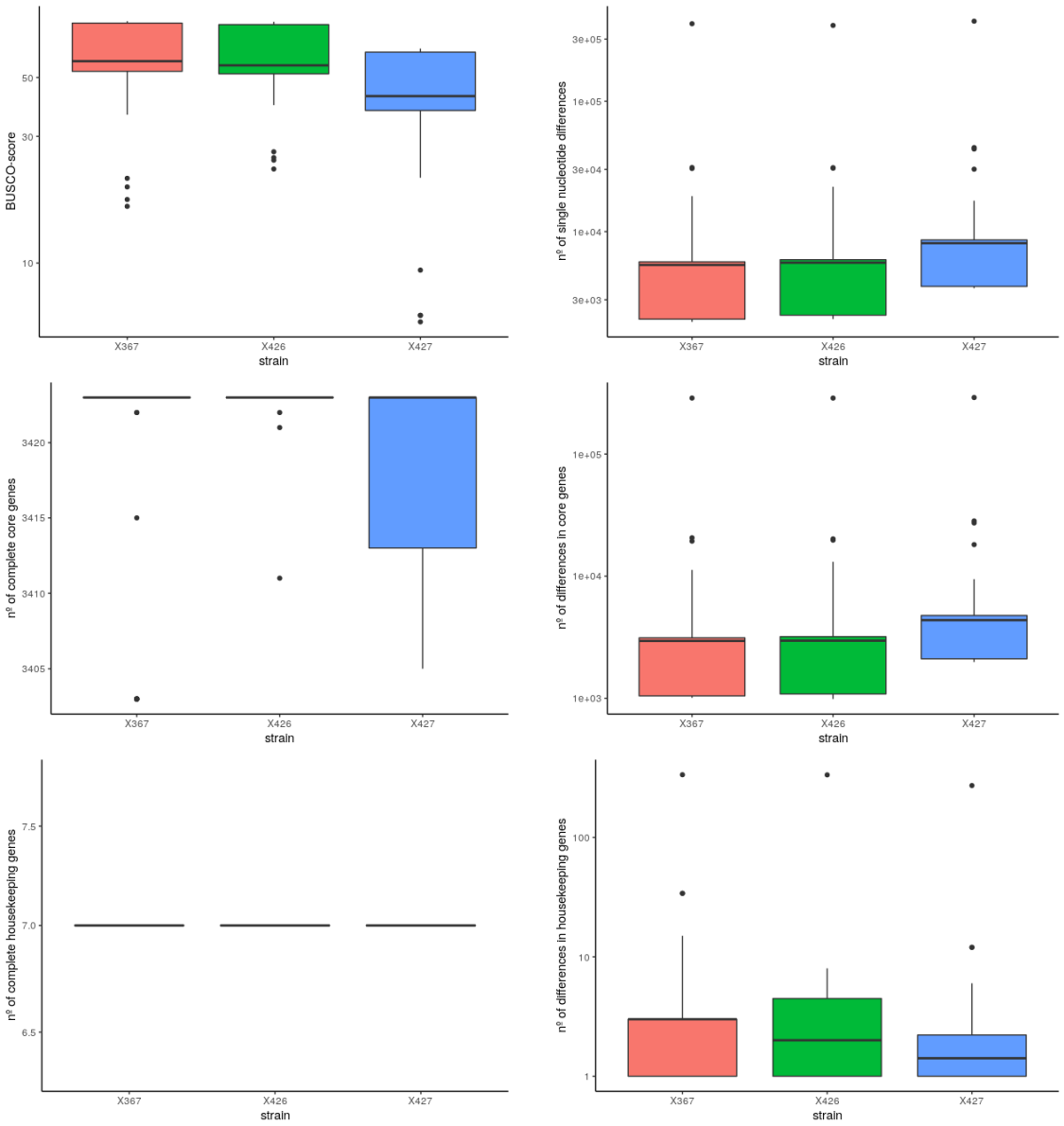


FIGURE B.2: Boxplots diplaying the distribution of the variables used to assess assemblies quality.

→ **Impact of the preprocessing choices in the assembled genomes**

Preprocessing reads do not significantly impact the pipelines performance (see Figure B.3b)

Mann-Whitney U test ($\alpha < 0.01$):

- **raw reads**

  H0: $overall\ score\{raw\} = overall\ score\{corrected, corrected + trimmed\}$

  H1: $overall\ score\{raw\} \neq overall\ score\{corrected, corrected + trimmed\}$

  $$W = 906.5, p = 0.61$$

- **corrected reads**

  H0: $overall\ score\{corrected\} = overall\ score\{raw, corrected + trimmed\}$

  H1: $overall\ score\{corrected\} \neq overall\ score\{raw, corrected + trimmed\}$

  $$W = 1209, p = 0.59$$

- **corrected+trimmed reads**

  H0: $overall\ score\{corrected + trimmed\} = overall\ score\{raw, corrected\}$

  H1: $overall\ score\{corrected + trimmed\} \neq overall\ score\{raw, corrected\}$

  $$W = 1124.5, p = 0.95$$

→ **Which assembler performs better?**

Pipelines with Flye have significantly inferior overall score (see B.3c).

Mann-Whitney U tests ($\alpha < 0.01$):

- **Canu**

  H0: $overall\ score\{Canu\} = overall\ score\{Miniasm, Wtdbg2, Flye\}$

  H1: $overall\ score\{Canu\} \neq overall\ score\{Miniasm, Wtdbg2, Flye\}$

  $$W = 572, p = 0.52$$

- **Miniasm**

  H0: $overall\ score\{Miniasm\} = overall\ score\{Canu, Wtdbg2, Flye\}$

H1: $overall\ score\{Miniasm\} \neq overall\ score\{Canu, Wtdbg2, Flye\}$

$$W = 1304, p = 0.009$$

- **Wtdbg2**

  H0: $overall\ score\{Wtdbg2\} = overall\ score\{Canu, Miniasm, Flye\}$

  H1: $overall\ score\{Wtdbg2\} \neq overall\ score\{Canu, Miniasm, Flye\}$

$$W = 800, p = 0.98$$

- **Flye**

  H0: $overall\ score\{Flye\} = overall\ score\{Canu, Miniasm, Wtdbg2\}$

  H1: $overall\ score\{Flye\} \neq overall\ score\{Canu, Miniasm, Wtdbg2\}$

$$W = 969, p = 0.002$$

As pipelines with Flye and Miniasm had a different average overall score, we repeated a Mann-Whitney U test for one side.

H0: $overallscore\{Flye\} \geq overallscore\{Canu, Miniasm, Wtdbg2\}$

H1: $overallscore\{Flye\} < overallscore\{Canu, Miniasm, Wtdbg2\}$

$$W = 572, p = 0.0008$$

H0: $overallscore\{Miniasm\} \geq overallscore\{Canu, Wtdbg2, Flye\}$

H1: $overallscore\{Miniasm\} < overallscore\{Canu, Wtdbg2, Flye\}$

$$W = 1304, p = 0.99$$

→ **Impact of improving draft assemblies**

Improving assemblies (with Racon and/or Medaka) led to overall scores signficantly inferior (see Figure B.3d).

Mann-Whitney U test ($\alpha < 0.01$):

H0: $overallscore\{withoutimprovement\} \geq overallscore\{withimprovement\}$

H1: $overallscore\{withimprovement\} < overallscore\{withoutimprovement\}$

$$W = 922, p = 5.5 \times 10^{-7}$$

Running Medaka after Racon also results in significantly lower overall scores (see Figure B.3e).

Wilcoxon signed rank test: $\alpha < 0.01$

H0: $overall\ score\{Racon + Medaka\} \geq overall\ score\{Racon\}$

H1: $overall\ score\{Racon + Medaka\} < overall\ score\{Racon\}$

$$V = 990, p = 3.9 \times 10^{-9}$$

There is also no significant difference in the overall score due to the number of Racon iterations before Medaka. This is represented in Figure B.3f.

Mann-Whitney U tests ($\alpha < 0.01$):

- **1 iteration**

  H0: $overall\ score\{1iteration\} = overall\ score\{1, 2, 4iterations\}$

  H1: $overall\ score\{1iteration\} \neq overall\ score\{1, 2, 4iterations\}$

  $$W = 240.5, p = 0.11$$

- **2 iterations**

  H0: $overall\ score\{2iterations\} = overall\ score\{1, 3, 4iterations\}$

  H1: $overall\ score\{2iterations\} \neq overall\ score\{1, 3, 4iterations\}$

  $$W = 181, p = 1$$

- **3 iterations**

  H0: $overall\ score\{3iterations\} = overall\ score\{1, 2, 4iterations\}$

  H1: $overall\ score\{3iterations\} \neq overall\ score\{1, 2, 4iterations\}$

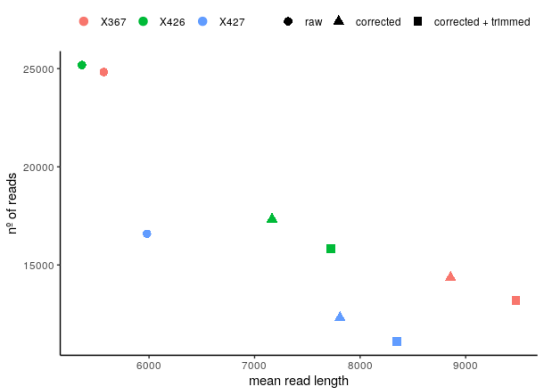
  $$W = 172.5, p = 0.82$$

- **4 iterations**

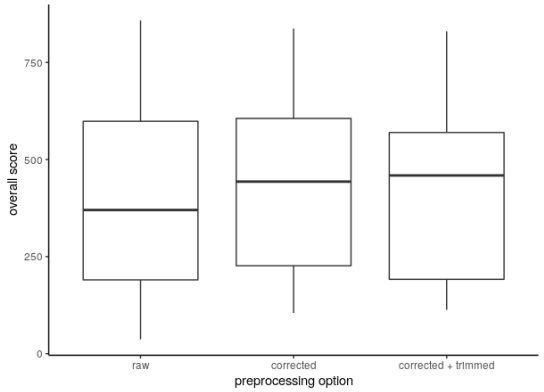
  H0: $overall\ score\{4iterations\} = overall\ score\{1, 2, 3iterations\}$

  H1: $overall\ score\{4iterations\} \neq overall\ score\{1, 2, 3iterations\}$

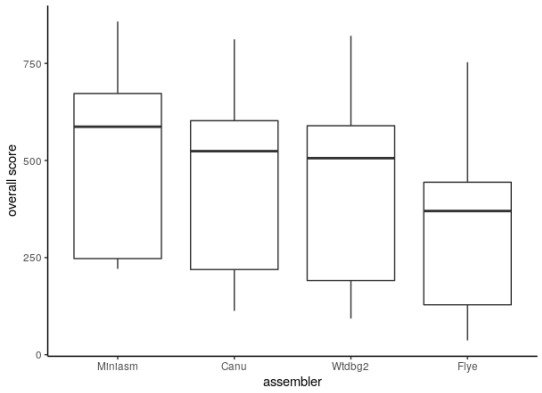
  $$W = 132, p = 0.18$$



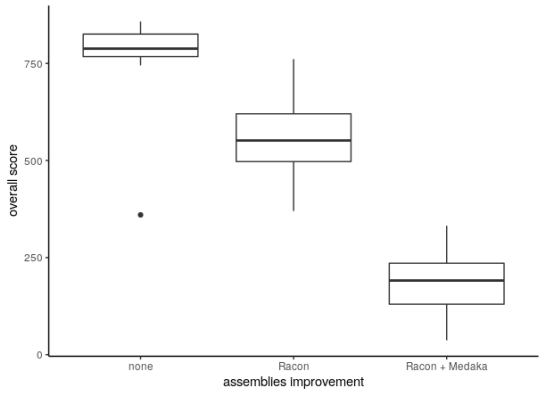
(A) The impact of preprocessing reads in the n°of reads and its mean length.



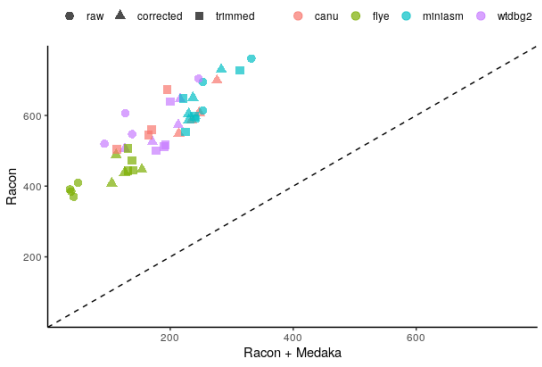
(B) Overall score distribution according to the set of reads.



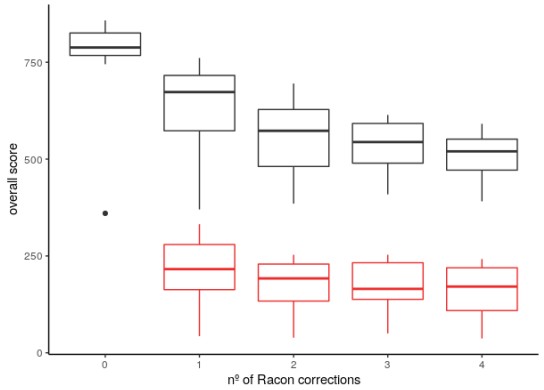
(C) Overall score distribution according to the assembler.



(D) Overall score distribution according to the assembly improvement method.



(E) Overall score with and without Medaka.



(F) Overall score progression as the Racon iterations increase. Pipelines without Medaka in black, with Medaka in red.

FIGURE B.3: How different choices impact assemblies quality.

Basecallings comparison

Considering the BUSCO-score, assemblies from reads basecalled with the accurate model resulted in genomes with higher completeness, as represented in Figure B.4, where is evident that all pipelines had an higher BUSCO-score for the accurate mode. It is supported by the following Wilcoxon signed rank test ($\alpha < 0.01$):

H0: $BUSCO - score\{fast\} \geq BUSCO - score\{accurate\}$

H1: $BUSCO - score\{fast\} < BUSCO - score\{accurate\}$

$$V = 44252, p < 2.2 \times 10^{-16}$$

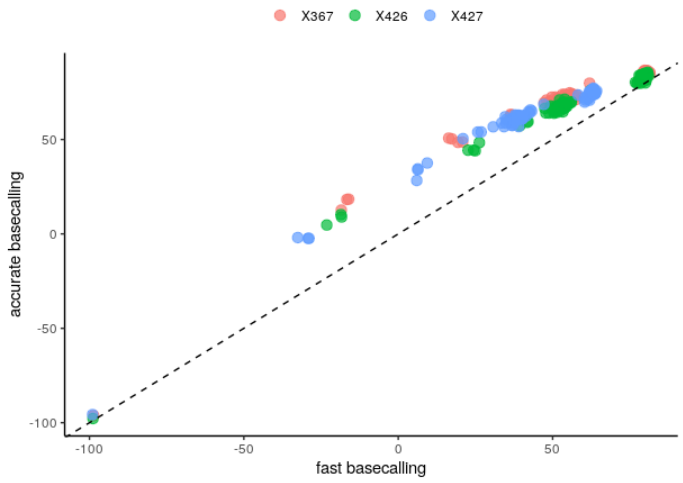


FIGURE B.4: BUSCO-score according to the basecalling mode.

Furthermore, beside the evidence that an accurate basecalling led to completer assemblies, an high and positive Pearson's correlation coefficient ($\rho = 0.96$) indicates that, for BUSCO-score, the relative performance of the pipelines is similar for both basecallings. To generalize this last assumption, we calculated the Pearson's correlation coefficient for the ranks of the overall score and obtained a $\rho = 0.97$. This result proves that the pipelines are ranked in similar order, regarding the overall score, and thus, we can have confidence that a pipeline that performed well (or poorly) for a dataset from one basecalling will have a similar performance for a dataset from the other basecalling too.

What is the best pipeline?

Over all samples, for accurate basecalled reads the best pipeline was r-flye-m1, and for fast basecalled reads was r-flye-m4. When considering the performances for both basecallings, as listed on Table B.2, r-flye-m1 is elected as the best pipeline.

TABLE B.2: Top 10 pipelines for both basecallings.

pipelines	overall score (fast)	overall score (accurate)	combined score
r-flye-m1	43	65	108
r-flye-m2	39	77	116
r-flye-m4	37	83	120
r-flye-m3	50	92	142
r-wtdbg2-m4	93	143	236
c-flye-m1	112	125	237
t-canu-m4	113	131	244
c-flye-m4	105	155	260
t-flye-m1	131	133	264
c-flye-m2	126	145	271

In Table B.3 we present some statistics for r-flye-m1 assemblies. The number of contigs is small and equal in each strain, and the assembly length and N50 are very similar, which demonstrates the advantage of ONT reads to achieve contiguous assemblies. Interestingly, a few more genes are detected for assemblies from fast basecalled reads, in spite of its inferior accuracy per base, as the QUAL indicates. Statistics from BUSCO show that accurate basecalled reads result in completer assemblies with more complete and less fragmented genes than fast basecalled reads.

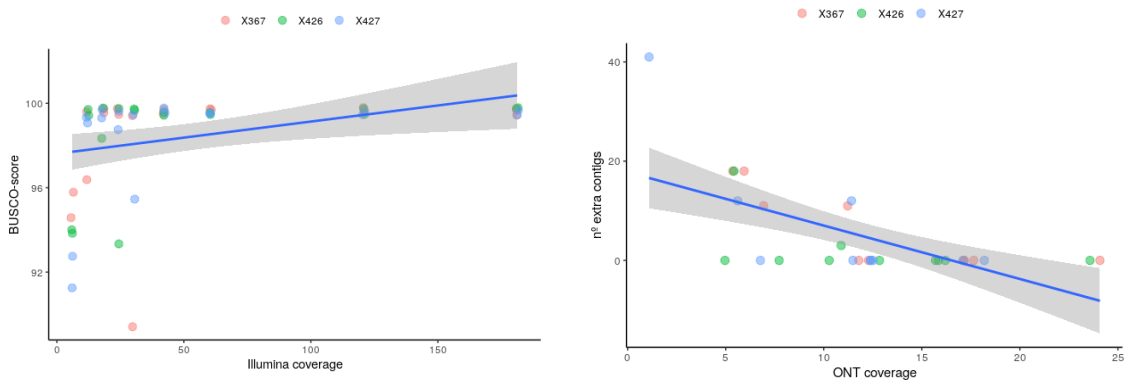
TABLE B.3: Statistics for assemblies from r-flye-m1.

assembly	Guppy model	length	n°contigs	N50	QUAL	n°genes	C	F	M
X367	fast	4969962	4	4924060	54	4,216	89.1	6.6	4.3
X367	accurate	4971547	4	4924476	56	4,210	91.4	4.9	3.7
X426	fast	4900888	2	4883520	53	4,188	88.2	7.2	4.6
X426	accurate	4901420	2	4884048	56	4,175	91.2	5.5	3.3
X427	fast	5228245	3	5228245	49	4,521	79.5	10.5	10.0
X427	accurate	5224511	3	4915292	50	4,520	86.4	6.0	7.6

Appendix C

On sequencing depth for hybrid assemblies

The simulations to estimate the minimum coverage from Illumina necessary to achieve a complete genome, i.e., a BUSCO-score of 100 are represented in Figure C.1a and the correspondent linear regression in equation (C.1) where B denotes the BUSCO-score and I the Illumina coverage. We estimate that to achieve a BUSCO-score of 100 it is required a 159x coverage of Illumina data, when looking into the data, all assemblies with more than 30x coverage resulted in the maximum Busco-score observed. The estimations for the minimum coverage necessary from ONT to achieve the ideal structure is represented in Figure C.1b and the correspondent linear regression in equation (C.2), where E denotes the number of extra-contigs and M the ONT coverage. We estimate that the minimum ONT coverage to achieve a contiguous assembly should be 16x, when looking into the data, all assemblies with more than 11x coverage result in contiguous structures.



(A) The impact of Illumina sequencing depth.

(B) The impact of MinION sequencing depth.

FIGURE C.1: Sequencing depths to achieve contiguous and complete assemblies.

$$B = 97.61 + 0.015I \Leftrightarrow B = 100 \implies C \approx 159 \tag{C.1}$$

$$E = 17.79 - 1.08M \Leftrightarrow E = 0 \implies C \approx 16 \tag{C.2}$$

Considering the limits determined above we performed assemblies varying both the Illumina and ONT coverages. As mentioned, a coverage of Illumina superior to 30x results in the maximum BUSCO-score observed, yet, we could not define a region where all assemblies have a complete structure with circular contigs. This is represented in Figure C.2 where is suggested that optimal results are just achieved for coverages superior to 18x for MinION and around 400x for Illumina and, we can observe that increasing coverage do not always lead to a better structure. Therefore, this study is a preliminary, but insufficient, step to reasonably propose a limit of coverage.

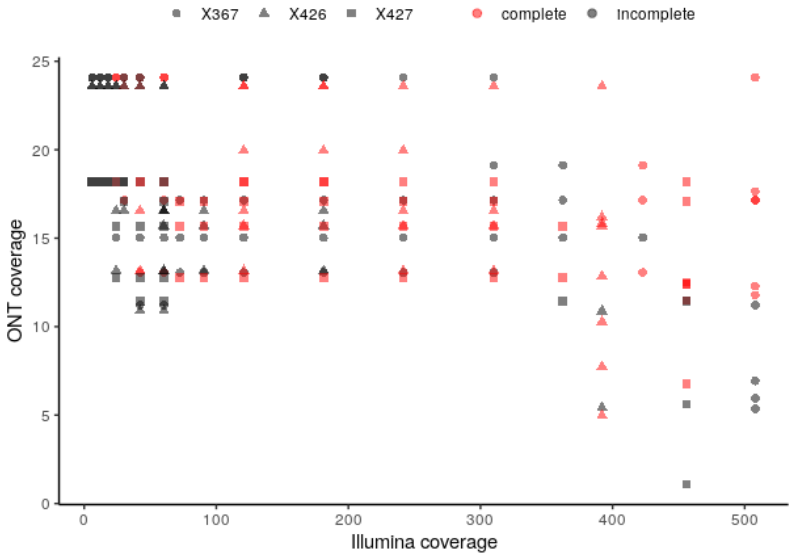
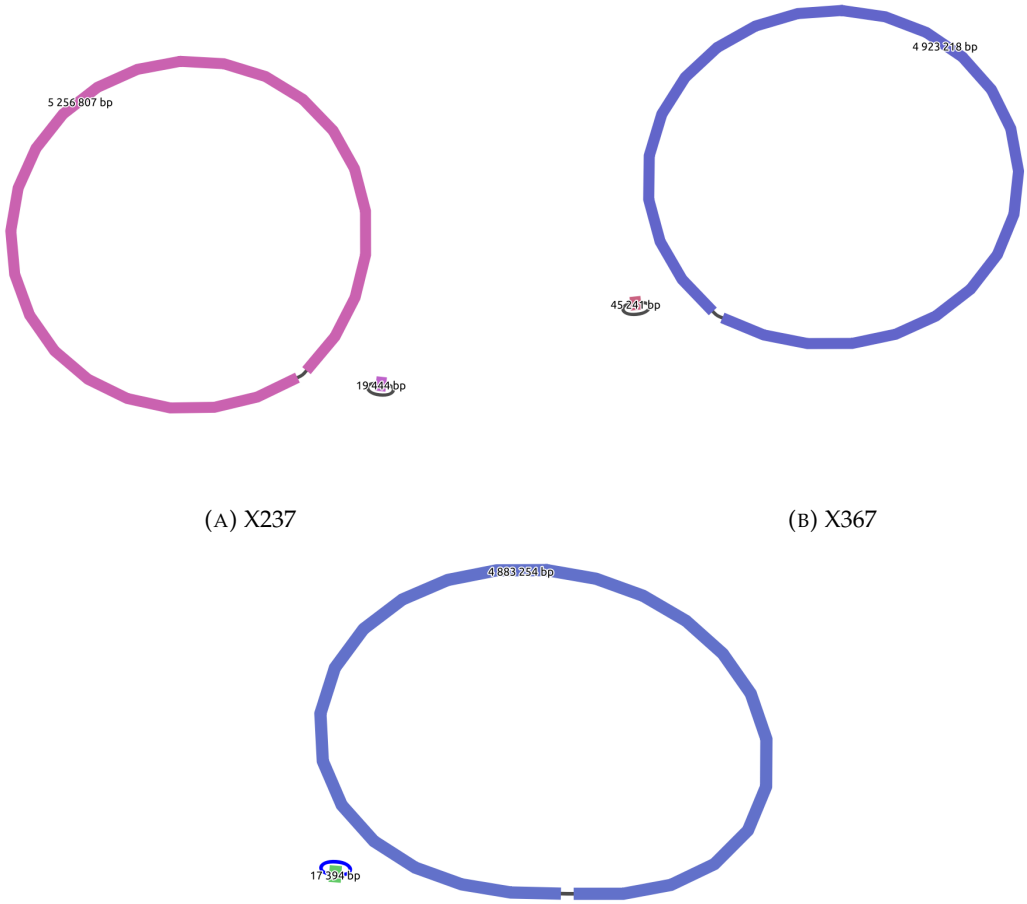


FIGURE C.2: Assemblies contiguity for different coverage of Illumina and ONT.

Appendix D

Assembly graphs

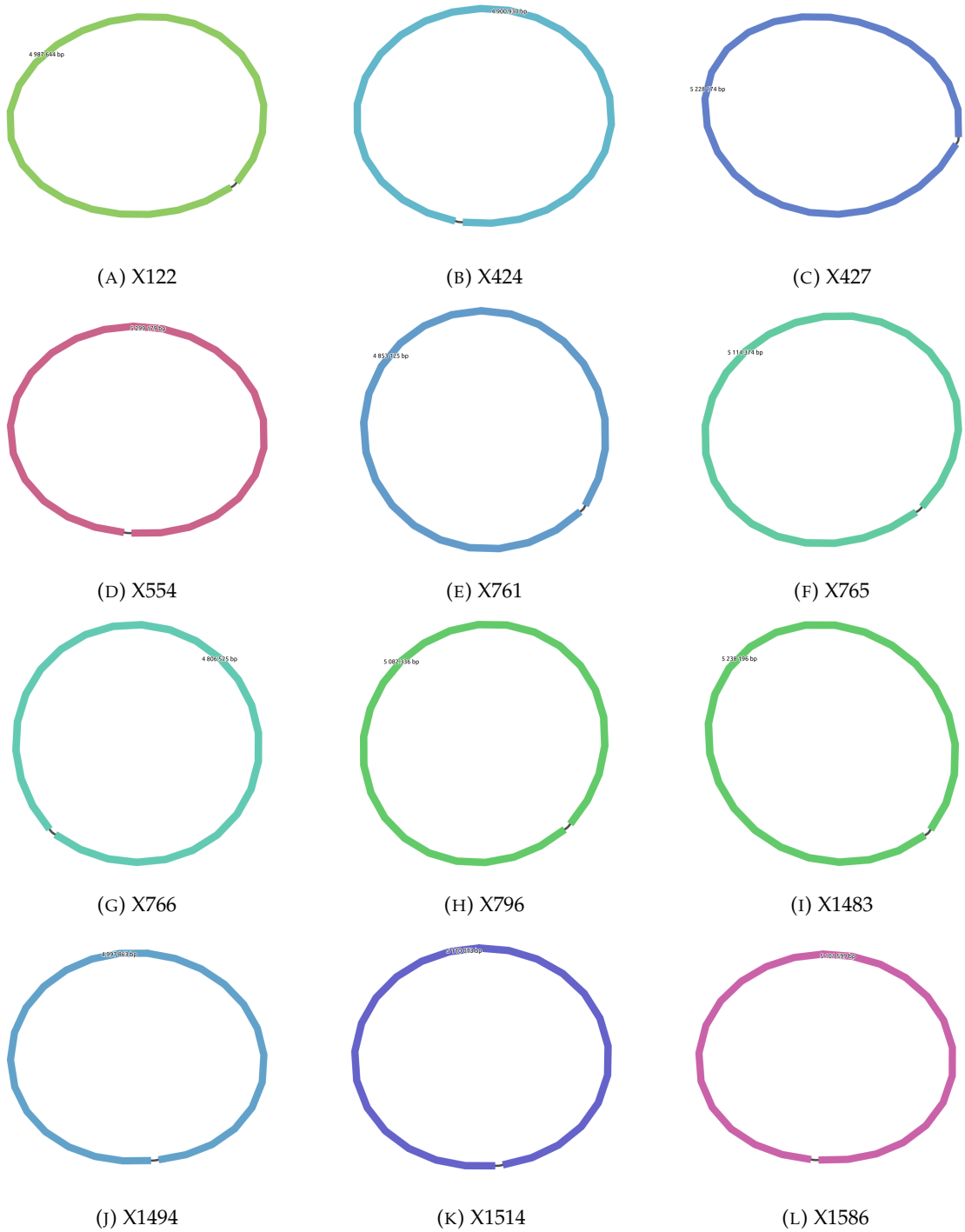


(A) X237

(B) X367

(C) X426

FIGURE D.1: Assembly graphs representation for strains with plasmids.



(A) X122 (B) X424 (C) X427

(D) X554 (E) X761 (F) X765

(G) X766 (H) X796 (I) X1483

(J) X1494 (K) X1514 (L) X1586

FIGURE D.2: Assembly graphs representation.

Appendix E

Sequencing data for *Xanthomonas* spp. isolates

The statistics for the ONT and Illumina datasets are present in Table E.1 and Table E.2. The mean quality is indicated in a phred-scale and the coverage was estimated considering a total genome size of 5M bp. Regarding the ONT data, the dataset name identifies the run and the Guppy version (v3.4.1 or v4.0.14). ONT datasets quality ranges from 11.3 to 13.2, in a phred scale, that correspond to error rates between 7.4% and 4.8%. Illumina datasets minimum coverage is 392x with pair-end reads of 2 x 151 bp and a mean quality score superior than 29, which implies an error probability below 0.1%.

TABLE E.1: Illumina sequencing data statistics.

strain	n°of reads	mean length	mean quality	coverage
X122	18,149,031	2 x 151	33.0	1096
X237	10,003,841	2 x 151	32.9	604
X367	8,413,466	2 x 151	29.3	508
X424	8,701,514	2 x 151	29.0	525
X426	6,494,807	2 x 151	28.8	392
X427	7,562,695	2 x 151	29.3	456
X554	17,448,616	2 x 151	32.6	1053
X761	10,085,534	2 x 151	33.0	609
X765	18,456,463	2 x 151	33.2	1114
X766	10,080,212	2 x 151	33.0	608
X796	10,714,224	2 x 151	33.2	647
X1483	18,295,142	2 x 151	33.3	1105
X1494	10,303,786	2 x 151	32.9	622
X1514	10,684,250	2 x 151	33.0	645
X1567	10,937,504	2 x 151	32.9	660
X1586	18,310,688	2 x 151	33.4	1105

TABLE E.2: ONT sequencing data statistics.

strain	dataset	n°of reads	mean length	mean quality	N50	coverage
X122	ONT r21v3	11,619	5,999	12.2	13,563	13
	ONT r22v3	6,389	4,815	11.4	10,716	6
X237	ONT r21v3	7,443	6,672	12.1	14,550	9
	ONT r22v3	4,139	5,541	11.4	12,067	4
X367	ONT r16v3	18,857	6,386	11.8	14,257	24
X424	ONT r31v4	15,263	6,401	11.7	12,236	19
X426	ONT r16v3	19,997	5,895	11.9	10,553	23
X427	ONT r16v3	14,302	6,358	11.8	13,332	18
X554	ONT r21v4	31,604	6,599	13.2	13,472	41
	ONT r22v4	31,534	4,963	12.6	12,308	31
	ONT r31v4	14,265	6,151	11.5	10,896	17
X761	ONT r21v3	13,979	5,832	12.2	13,451	16
	ONT r22v3	7,772	4,796	11.4	10,818	7
X765	ONT r21v4	31,310	7,272	13.1	14,086	45
	ONT r22v4	29,845	5,531	12.5	12,979	33
	ONT r31v4	12,555	3,647	11.6	6,968	9
X766	ONT r21v3	18,652	5,529	12.2	12,600	20
	ONT r22v3	10,002	4,514	11.4	9,983	9
X796	ONT r21v3	14,032	5,617	12.3	12,511	15
	ONT r22v3	7,308	4,735	11.5	10,429	6
X1483	ONT r21v4	32,689	8,490	13.2	16,401	55
	ONT r22v4	28,846	6,479	12.7	15,130	37
	ONT r31v4	13,998	6,707	11.8	12,920	18
X1494	ONT r21v3	4,092	7,014	12.2	15,268	5
	ONT r22v3	2,281	5,745	11.3	13,516	3
x1514	ONT r31v4	32,533	5,523	11.8	10,416	35
x1567	ONT r31v4	13,670	6,867	11.7	13,127	18
X1586	ONT r21v4	12,783	7,434	13.2	14,370	19
	ONT r22v4	13,586	5,676	12.5	13,644	15
	ONT r31v4	16,048	5,944	11.7	10,664	19

Appendix F

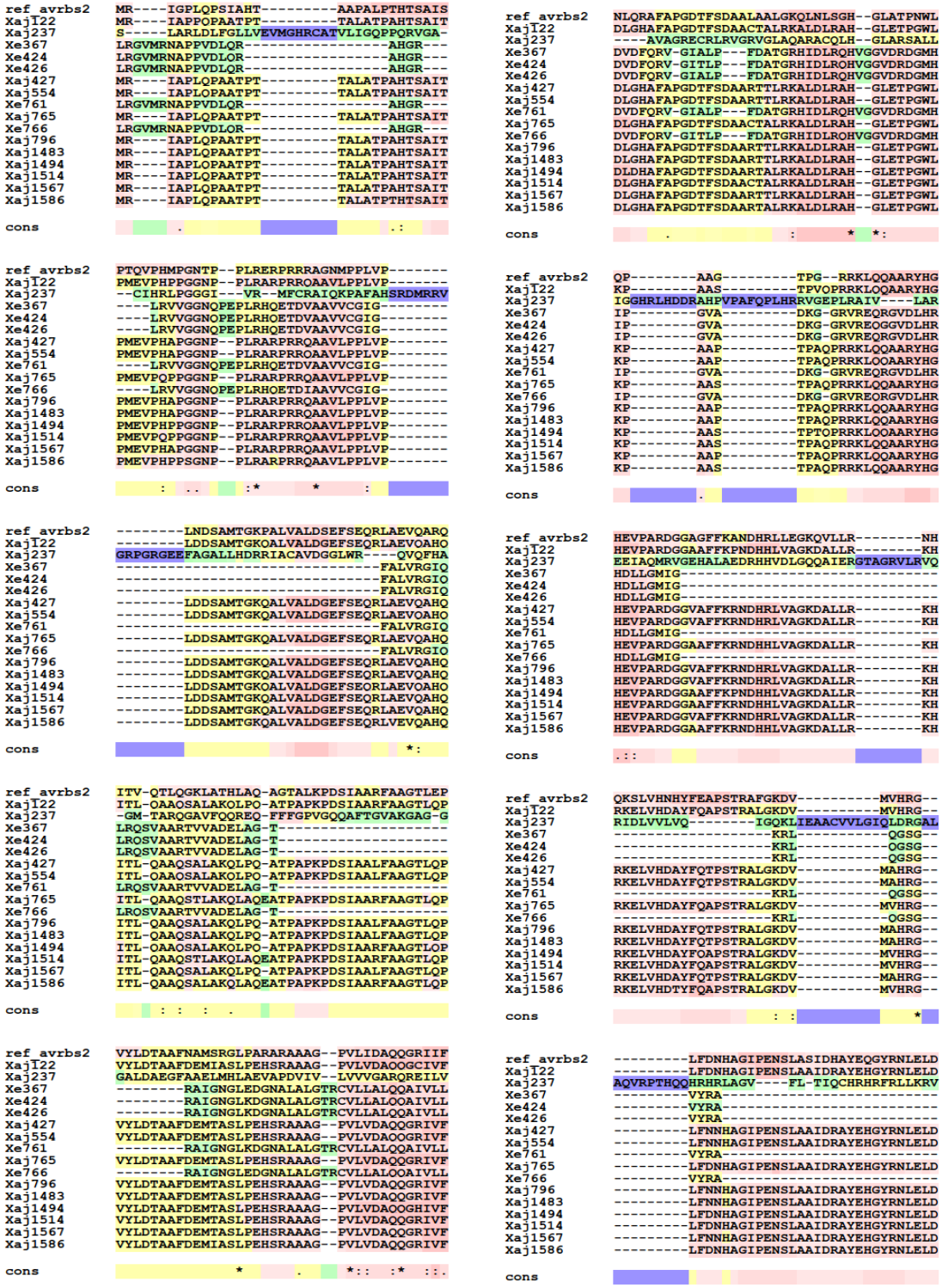
Average nucleotide identities and T3E alignments

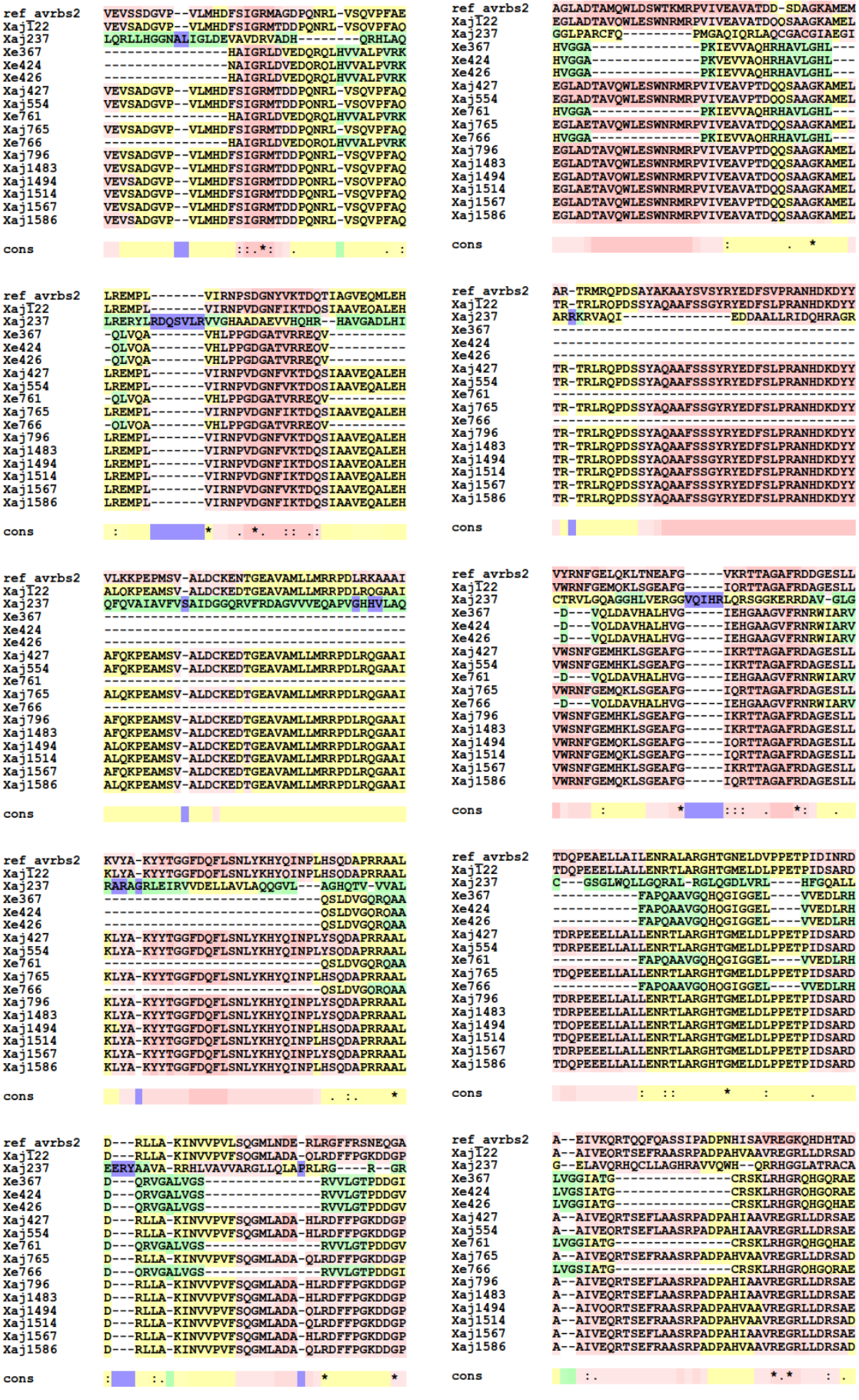
F.1 Average nucleotide identity

TABLE F.1: Average nucleotide identities.

ANI(%)	X122	X1483	X1494	X1514	X1567	X1586	X237	X367	X424	X426	X427	X554	X761	X765	X766	X796	Xca	Xci
X122	100	96.6	97.1	97.0	96.7	96.7	96.7	93	92.9	92.9	96.6	96.6	92.8	97.0	92.9	96.8	86.4	87.1
X1483	96.6	100	96.6	96.4	98.2	96.3	98.1	93.1	93	93	100	98.9	93	96.4	93.1	98.3	86.5	87.1
X1494	97.1	96.6	100	96.9	96.6	96.8	96.5	92.8	92.9	92.8	96.6	96.6	92.7	96.9	92.8	96.6	86.2	87.2
X1514	97	96.4	96.9	100	96.5	96.7	96.3	93.1	93.1	93	96.3	96.3	93	100	93.1	96.5	86.3	87.1
X1567	96.7	98.2	96.6	96.5	100	96.3	99.4	93.1	93.1	93.1	98.2	98.4	93.1	96.4	93.1	99.4	86.2	87.1
X1586	96.7	96.3	96.8	96.7	96.3	100	96.3	93.2	93.2	93.3	96.3	96.3	93.2	96.7	93.1	96.4	86.2	87.1
X237	96.7	98.1	96.5	96.3	99.4	96.3	100	93	93	93.1	98.1	98.1	93.1	96.2	93.1	99.3	86.4	87.0
X367	93	93.1	92.8	93.1	93.1	93.2	93	100	97.9	98	93	93.1	98	93	97.9	93	86.1	87.1
X424	92.9	93	92.9	93.1	93.1	93.2	93	97.9	100	98	93	93	98	93.1	98.2	93	86.3	87.1
X426	92.9	93	92.8	93	93.1	93.3	93.1	98	98	100	93.1	93	100	93	98.2	93.1	86.3	87.3
X427	96.6	100	96.6	96.3	98.2	96.3	98.1	93	93	93.1	100	98.9	93.1	96.3	93.1	98.3	86.3	86.9
X554	96.6	98.9	96.6	96.3	98.4	96.3	98.1	93.1	93	93	98.9	100	93.1	96.3	93.1	98.6	86.4	87.1
X761	92.8	93	92.7	93	93.1	93.2	93.1	98	98	100	93.1	93.1	100	93	98.2	93	86.3	87.1
X765	97	96.4	96.9	100	96.4	96.7	96.2	93	93.1	93	96.3	96.3	93	100	93	96.4	86.3	87.1
X766	92.9	93.1	92.8	93.1	93.1	93.1	93.1	97.9	98.2	98.2	93.1	93.1	98.2	93	100	93.1	86.3	87.1
X796	96.8	98.3	96.6	96.5	99.4	96.4	99.3	93	93	93.1	98.3	98.6	93	96.4	93.1	100	86.4	87.2
Xca	86.4	86.5	86.2	86.3	86.2	86.2	86.4	86.1	86.3	86.3	86.3	86.4	86.3	86.3	86.3	86.4	100	85.1
Xci	87.1	87.1	87.2	87.1	87.1	87.1	87	87.1	87.1	87.3	86.9	87.1	87.1	87.1	87.1	87.2	85.1	100

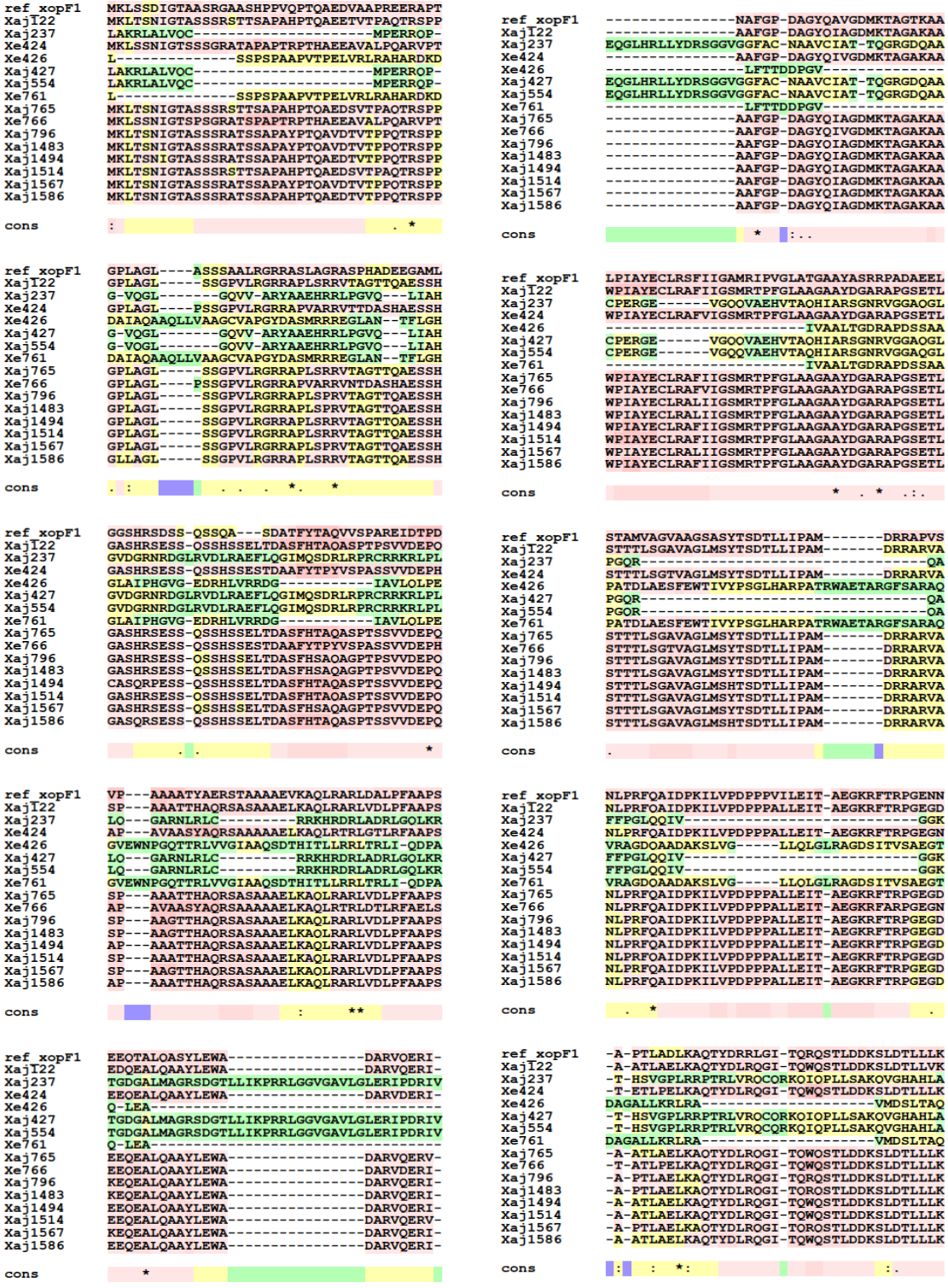
F.2 *avrBs2* multiple alignment

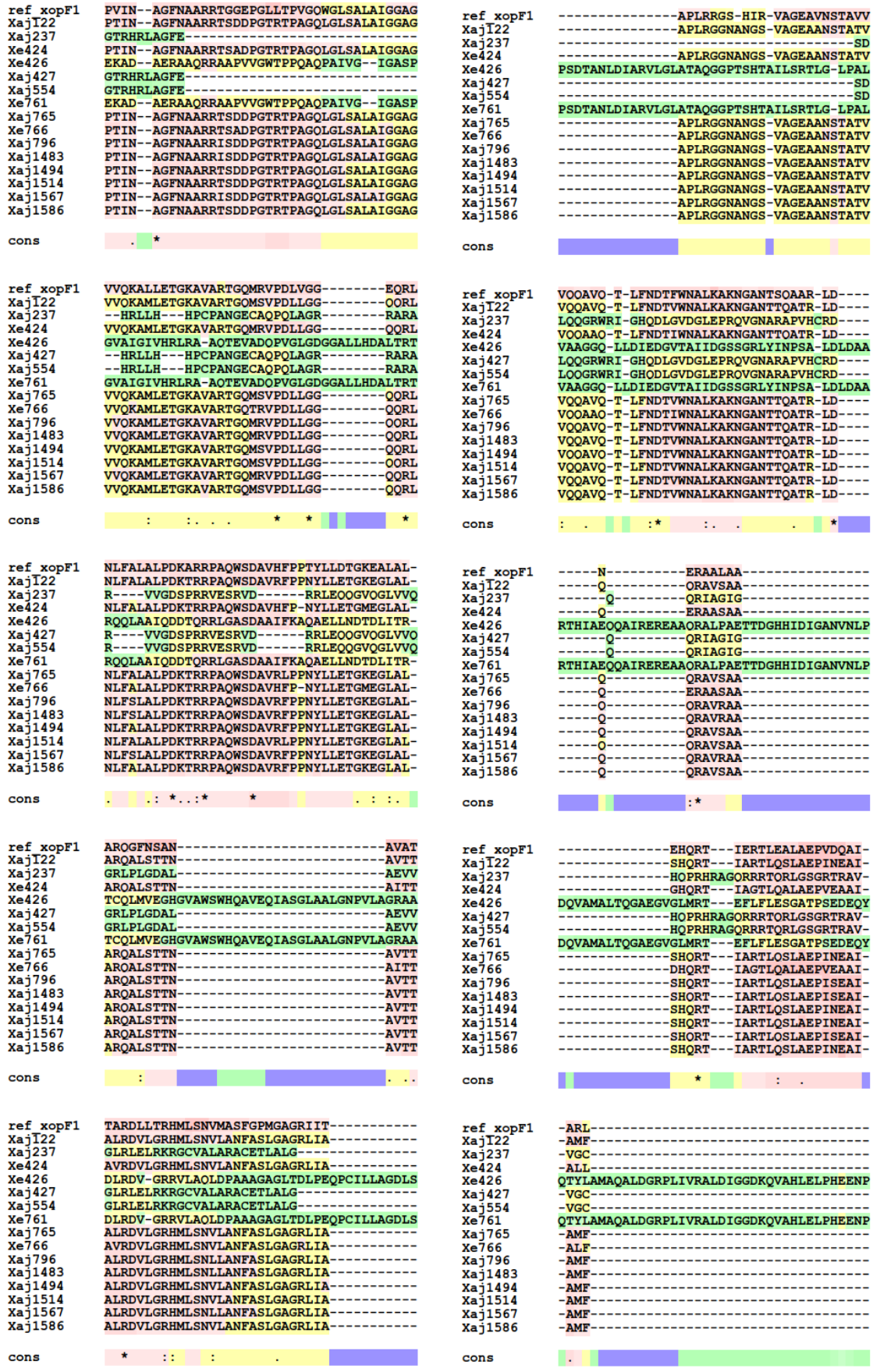


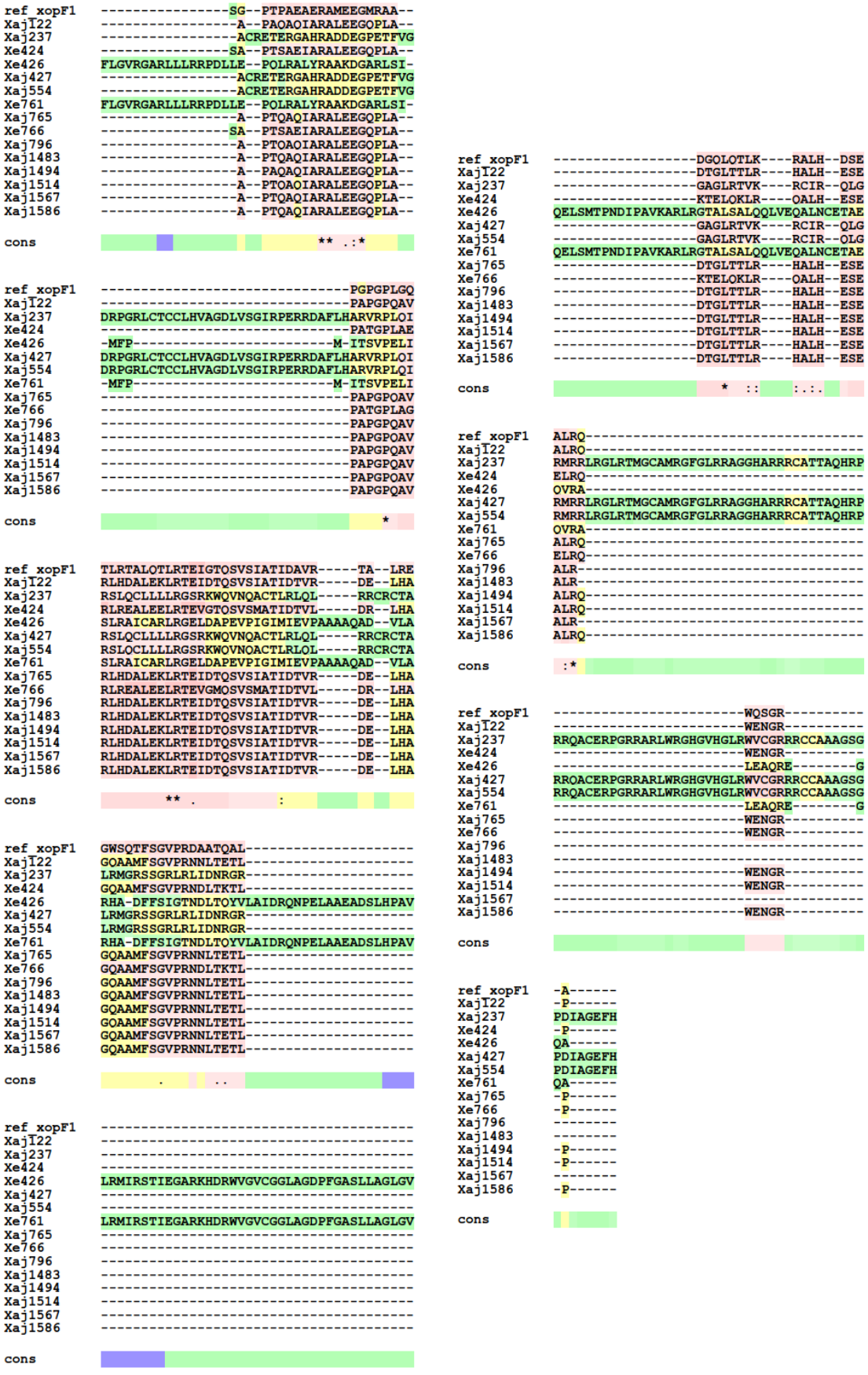

ref_avrbs2 VEVSSDGVP--VLMHDFSIGRMAGDPQNRL-VSQVPFAE
XajT22 VEVSADGVP--VLMHDFSIGRMTDDPQNRL-VSQVPFAQ
Xaj237 LQRLLHGGNALIGLDEVAVDRVADH--------QRHLAQ
Xe367 ----------HAIGRLDVEDQRQLHVVALPVRK
Xe424 ----------NAIGRLDVEDQRQLHVVALPVRK
Xe426 ----------HAIGRLDVEDQRQLHVVALPVRK
Xaj427 VEVSADGVP--VLMHDFSIGRMTDDPQNRL-VSQVPFAQ
Xaj554 VEVSADGVP--VLMHDFSIGRMTDDPQNRL-VSQVPFAQ
Xe761 ----------HAIGRLDVEDQRQLHVVALPVRK
Xaj765 VEVSADGVP--VLMHDFSIGRMTDDPQNRL-VSQVPFAQ
Xe766 ----------HAIGRLDVEDQRQLHVVALPVRK
Xaj796 VEVSADGVP--VLMHDFSIGRMTDDPQNRL-VSQVPFAQ
Xaj1483 VEVSADGVP--VLMHDFSIGRMTDDPQNRL-VSQVPFAQ
Xaj1494 VEVSADGVP--VLMHDFSIGRMTDDPQNRL-VSQVPFAQ
Xaj1514 VEVSADGVP--VLMHDFSIGRMTDDPQNRL-VSQVPFAQ
Xaj1567 VEVSADGVP--VLMHDFSIGRMTDDPQNRL-VSQVPFAQ
Xaj1586 VEVSADGVP--VLMHDFSIGRMTDDPQNRL-VSQVPFAQ
cons ::.*: . . :
ref_avrbs2 AGLADTAMQWLDSWTKMRPVIVEAVATDD-SDAGKAMEM
XajT22 EGLADTAVQWLESWNRMRPVIVEAVATDQQSAAGKAMEL
Xaj237 GGLPARCFQ---------PMGAQIQRLAQCGACGIAEGI
Xe367 HVGGA----------PKIEVVAQHRHAVLGHL---
Xe424 HVGGA----------PKVEVVAQHRHAVLGHL---
Xe426 HVGGA----------PKIEVVAQHRHAVLGHL---
Xaj427 EGLADTAVQWLESWNRMRPVIVEAVPTDQQSAAGKAMEL
Xaj554 EGLADTAVQWLESWNRMRPVIVEAVPTDQQSAAGKAMEL
Xe761 HVGGA----------PKIEVVAQHRHAVLGHL---
Xaj765 EGLAETAVQWLESWNRMRPVIVEAVATDQQSAAGKAMEL
Xe766 HVGGA----------PKIEVVAQHRHAVLGHL---
Xaj796 EGLADTAVQWLESWNRMRPVIVEAVPTDQQSAAGKAMEL
Xaj1483 EGLADTAVQWLESWNRMRPVIVEAVPTDQQSAAGKAMEL
Xaj1494 EGLADTAVQWLESWNRMRPVIVEAVATDQQSAAGKAMEL
Xaj1514 EGLAETAVQWLESWNRMRPVIVEAVATDQQSAAGKAMEL
Xaj1567 EGLADTAVQWLESWNRMRPVIVEAVPTDQQSAAGKAMEL
Xaj1586 EGLADTAVQWLESWNRMRPVIVEAVATDQQSAAGKAMEL
cons : .
ref_avrbs2 LREMPL-------VIRNPSDGNYVKTDQTIAGVEQMLEH
XajT22 LREMPL-------VIRNPVDGNFIKTDQSIAAVEQALEH
Xaj237 LRERYLRDQSVLRVVGHAADAEVVHQHR--HAVGADLHI
Xe367 -QLVQA-------VHLPPGDGATVRREQV----------
Xe424 -QLVQA-------VHLPPGDGATVRREQV----------
Xe426 -QLVQA-------VHLPPGDGATVRREQV----------
Xaj427 LREMPL-------VIRNPVDGNFVKTDQSIAAVEQALEH
Xaj554 LREMPL-------VIRNPVDGNFVKTDQSIAAVEQALEH
Xe761 -QLVQA-------VHLPPGDGATVRREQV----------
Xaj765 LREMPL-------VIRNPVDGNFIKTDQSIAAVEQALEH
Xe766 -QLVQA-------VHLPPGDGATVRREQV----------
Xaj796 LREMPL-------VIRNPVDGNFVKTDQSIAAVEQALEH
Xaj1483 LREMPL-------VIRNPVDGNFVKTDQSIAAVEQALEH
Xaj1494 LREMPL-------VIRNPVDGNFIKTDQSIAAVEQALEH
Xaj1514 LREMPL-------VIRNPVDGNFIKTDQSIAAVEQALEH
Xaj1567 LREMPL-------VIRNPVDGNFVKTDQSIAAVEQALEH
Xaj1586 LREMPL-------VIRNPVDGNFIKTDQSIAAVEQALEH
cons : * . *. :: .:
ref_avrbs2 AR-TRMRQPDSAYAKAAYSVSYRYEDFSVPRANHDKDYY
XajT22 TR-TRLRQPDSSYAQAAFSSGYRYEDFSLPRANHDKDYY
Xaj237 ARRKRVAQI----------EDDAALLRIDQHRAGR
Xe367 ----------
Xe424 ----------
Xe426 ----------
Xaj427 TR-TRLRQPDSSYAQAAFSSSYRYEDFSLPRANHDKDYY
Xaj554 TR-TRLRQPDSSYAQAAFSSSYRYEDFSLPRANHDKDYY
Xe761 ----------
Xaj765 TR-TRLRQPDSSYAQAAFSSGYRYEDFSLPRANHDKDYY
Xe766 ----------
Xaj796 TR-TRLRQPDSSYAQAAFSSSYRYEDFSLPRANHDKDYY
Xaj1483 TR-TRLRQPDSSYAQAAFSSSYRYEDFSLPRANHDKDYY
Xaj1494 TR-TRLRQPDSSYAQAAFSSGYRYEDFSLPRANHDKDYY
Xaj1514 TR-TRLRQPDSSYAQAAFSSGYRYEDFSLPRANHDKDYY
Xaj1567 TR-TRLRQPDSSYAQAAFSSSYRYEDFSLPRANHDKDYY
Xaj1586 TR-TRLRQPDSSYAQAAFSSGYRYEDFSLPRANHDKDYY
cons
ref_avrbs2 VLKKPEPMSV-ALDCKENTGEAVAMLLMRRPDLRKAAAI
XajT22 ALQKPEAMSV-ALDCKEDTGEAVAMLLMRRPDLRQGAAI
Xaj237 QFQVAIAVFVSAIDGGQRVFRDAGVVVEQAPVGHHVLAQ
Xe367 ----------
Xe424 ----------
Xe426 ----------
Xaj427 AFQKPEAMSV-ALDCKEDTGEAVAMLLMRRPDLRQGAAI
Xaj554 AFQKPEAMSV-ALDCKEDTGEAVAMLLMRRPDLRQGAAI
Xe761 ----------
Xaj765 ALQKPEAMSV-ALDCKEDTGEAVAMLLMRRPDLRQGAAI
Xe766 ----------
Xaj796 AFQKPEAMSV-ALDCKEDTGEAVAMLLMRRPDLRQGAAI
Xaj1483 AFQKPEAMSV-ALDCKEDTGEAVAMLLMRRPDLRQGAAI
Xaj1494 ALQKPEAMSV-ALDCKEDTGEAVAMLLMRRPDLRQGAAI
Xaj1514 ALQKPEAMSV-ALDCKEDTGEAVAMLLMRRPDLRQGAAI
Xaj1567 AFQKPEAMSV-ALDCKEDTGEAVAMLLMRRPDLRQGAAI
Xaj1586 ALQKPEAMSV-ALDCKEDTGEAVAMLLMRRPDLRQGAAI
cons
ref_avrbs2 VYRNFGELQKLTNEAFG-----VKRTTAGAFRDDGESLL
XajT22 VWRNFGEMQKLSGEAFG-----IQRTTAGAFRDAGESLL
Xaj237 CTRVLGQAGGHLVERGGVQIHRLQRSGGKERRDAV-GLG
Xe367 -D---VQLDAVHALHVG-----IEHGAAGVFRNRWIARV
Xe424 -D---VQLDAVHALHVG-----IEHGAAGVFRNRWIARV
Xe426 -D---VQLDAVHALHVG-----IEHGAAGVFRNRWIARV
Xaj427 VWSNFGEMHKLSGEAFG-----IKRTTAGAFRDAGESLL
Xaj554 VWSNFGEMHKLSGEAFG-----IKRTTAGAFRDAGESLL
Xe761 -D---VQLDAVHALHVG-----IEHGAAGVFRNRWIARV
Xaj765 VWRNFGEMQKLSGEAFG-----IQRTTAGAFRDAGESLL
Xe766 -D---VQLDAVHALHVG-----IEHGAAGVFRNRWIARV
Xaj796 VWSNFGEMHKLSGEAFG-----IKRTTAGAFRDAGESLL
Xaj1483 VWSNFGEMHKLSGEAFG-----IKRTTAGAFRDAGESLL
Xaj1494 VWRNFGEMQKLSGEAFG-----IQRTTAGAFRDAGESLL
Xaj1514 VWRNFGEMQKLSGEAFG-----IQRTTAGAFRDAGESLL
Xaj1567 VWSNFGEMHKLSGEAFG-----IKRTTAGAFRDAGESLL
Xaj1586 VWRNFGEMQKLSGEAFG-----IQRTTAGAFRDAGESLL
cons : * ::: . *: .
ref_avrbs2 KVYA-KYYTGGFDQFLSNLYKHYQINPLHSQDAPRRAAL
XajT22 KLYA-KYYTGGFDQFLSNLYKHYQINPLHSQDAPRRAAL
Xaj237 RARAGRLEIRVVDELLAVLAQQGVL---AGHQTV-VVAL
Xe367 ----------QSLDVGQRQAA
Xe424 ----------QSLDVGQRQAA
Xe426 ----------QSLDVGQRQAA
Xaj427 KLYA-KYYTGGFDQFLSNLYKHYQINPLYSQDAPRRAAL
Xaj554 KLYA-KYYTGGFDQFLSNLYKHYQINPLYSQDAPRRAAL
Xe761 ----------QSLDVGQRQAA
Xaj765 KLYA-KYYTGGFDQFLSNLYKHYQINPLHSQDAPRRAAL
Xe766 ----------QSLDVGQRQAA
Xaj796 KLYA-KYYTGGFDQFLSNLYKHYQINPLYSQDAPRRAAL
Xaj1483 KLYA-KYYTGGFDQFLSNLYKHYQINPLYSQDAPRRAAL
Xaj1494 KLYA-KYYTGGFDQFLSNLYKHYQINPLHSQDAPRRAAL
Xaj1514 KLYA-KYYTGGFDQFLSNLYKHYQINPLHSQDAPRRAAL
Xaj1567 KLYA-KYYTGGFDQFLSNLYKHYQINPLYSQDAPRRAAL
Xaj1586 KLYA-KYYTGGFDQFLSNLYKHYQINPLHSQDAPRRAAL
cons . :. *
ref_avrbs2 TDQPEAELLAILENRALARGHTGNELDVPPETPIDINRD
XajT22 TDQPEEELLALLENRTLARGHTGMELDLPPETPIDSARD
Xaj237 C---GSGLWQLLGQRAL-RGLQGDLVRL----HFGQALL
Xe367 ----------FAPQAAVGQHQGIGGEL----VVEDLRH
Xe424 ----------FAPQAAVGQHQGIGGEL----VVEDLRH
Xe426 ----------FAPQAAVGQHQGIGGEL----VVEDLRH
Xaj427 TDRPEEELLALLENRTLARGHTGMELDLPPETPIDSARD
Xaj554 TDRPEEELLALLENRTLARGHTGMELDLPPETPIDSARD
Xe761 ----------FAPQAAVGQHQGIGGEL----VVEDLRH
Xaj765 TDQPEEELLALLENRTLARGHTGMELDLPPETPIDSARD
Xe766 ----------FAPQAAVGQHQGIGGEL----VVEDLRH
Xaj796 TDRPEEELLALLENRTLARGHTGMELDLPPETPIDSARD
Xaj1483 TDRPEEELLALLENRTLARGHTGMELDLPPETPIDSARD
Xaj1494 TDQPEEELLALLENRTLARGHTGMELDLPPETPIDSARD
Xaj1514 TDQPEEELLALLENRTLARGHTGMELDLPPETPIDSARD
Xaj1567 TDRPEEELLALLENRTLARGHTGMELDLPPETPIDSARD
Xaj1586 TDQPEEELLALLENRTLARGHTGMELDLPPETPIDSARD
cons : :: * : .
ref_avrbs2 D---RLLA-KINVVPVLSQGMLNDE-RLRGFFRSNEQGA
XajT22 D---RLLA-KINVVPVFSQGMLADA-QLRDFFPGKDDGP
Xaj237 EERYAAVA-RRHLVAVVARGLLQLAPRLRG----R--GR
Xe367 D---QRVGALVGS----------RVVLGTPDDGI
Xe424 D---QRVGALVGS----------RVVLGTPDDGV
Xe426 D---QRVGALVGS----------RVVLGTPDDGV
Xaj427 D---RLLA-KINVVPVFSQGMLADA-HLRDFFPGKDDGP
Xaj554 D---RLLA-KINVVPVFSQGMLADA-HLRDFFPGKDDGP
Xe761 D---QRVGALVGS----------RVVLGTPDDGV
Xaj765 D---RLLA-KINVVPVFSQGMLADA-QLRDFFPGKDDGP
Xe766 D---QRVGALVGS----------RVVLGTPDDGI
Xaj796 D---RLLA-KINVVPVFSQGMLADA-HLRDFFPGKDDGP
Xaj1483 D---RLLA-KINVVPVFSQGMLADA-HLRDFFPGKDDGP
Xaj1494 D---RLLA-KINVVPVFSQGMLADA-QLRDFFPGKDDGP
Xaj1514 D---RLLA-KINVVPVFSQGMLADA-QLRDFFPGKDDGP
Xaj1567 D---RLLA-KINVVPVFSQGMLADA-HLRDFFPGKDDGP
Xaj1586 D---RLLA-KINVVPVFSQGMLADA-QLRDFFPGKDDGP
cons : :. * *
ref_avrbs2 A--EIVKQRTQQFQASSIPADPNHISAVREGKQHDHTAD
XajT22 A--AIVEQRTSEFRAASRPADPAHVAAVREGRLLDRSAD
Xaj237 G--ELAVQRHQCLLAGHRAVVQWH--QRRHGGLATRACA
Xe367 LVGGIATG----------CRSKLRHGRQHGQRAE
Xe424 LVGSIATG----------CRSKLRHGRQHGQRAE
Xe426 LVGGIATG----------CRSKLRHGRQHGQHAE
Xaj427 A--AIVEQRTSEFLAASRPADPAHIAAVREGRLLDRSAE
Xaj554 A--AIVEQRTSEFLAASRPADPAHIAAVREGRLLDRSAE
Xe761 LVGGIATG----------CRSKLRHGRQHGQHAE
Xaj765 A--AIVEQRTSEFRAASRPADPAHVAAVREGRLLDRSAD
Xe766 LVGSIATG----------CRSKLRHGRQHGQRAE
Xaj796 A--AIVEQRTSEFLAASRPADPAHIAAVREGRLLDRSAE
Xaj1483 A--AIVEQRTSEFLAASRPADPAHIAAVREGRLLDRSAE
Xaj1494 A--AIVQQRTSEFRAASRPADPAHVAAVREGRLLDRSAE
Xaj1514 A--AIVEQRTSEFRAASRPADPAHVAAVREGRLLDRSAD
Xaj1567 A--AIVEQRTSEFLAASRPADPAHIAAVREGRLLDRSAE
Xaj1586 A--AIVEQRTSEFRAASRPADPAHVAAVREGRLLDRSAD
cons :. *.* : .

```
ref_avrbs2    MVNDPAATRALDKRAKALGLLTDKYRGAPVTHYLNEQAR
XajI22        HSHDAAARQAVDARADALGRLTDQYRGAPVTHYLNEQAR
Xaj237        QWRVAAWRVRHLHRGDRAGVRRCGQRGGGRGGGLQGSNT
Xe367         QAA----------------------GTG-QGMWNEGGR
Xe424         QAA----------------------GTG-QGMWNEGGR
Xe426         QAA----------------------GTG-QGMWNEGGR
Xaj427        HPHDAAARQAVDARADTLGRLTDQYRGAPVTHYLNEQAK
Xaj554        HPHDAAARQAVDARADTLGRLTDQYRGAPVTHYLNEQAK
Xe761         QAA----------------------GTG-QGMWNEGGR
Xaj765        HSHDAAARQAVDARADALGRLTDQYRGAPVTHYLNEQAK
Xe766         QAA----------------------GTG-QGMWNEGGR
Xaj796        HPHDAAARQAVDARADTLGRLTDQYRGAPVTHYLNEQAK
Xaj1483       HPHDAAARQAVDARADTLGRLTDQYRGAPVTHYLNEQAK
Xaj1494       HPHDAAARQAVDARADALGRLTDQYRGAPVTQYLNEQAK
Xaj1514       HSHDAAARQAVDARADALGRLTDQYRGAPVTHYLNEQAK
Xaj1567       HPHDAAARQAVDARADTLGRLTDQYRGAPVTHYLNEQAK
Xaj1586       HSHDAAARQAVDARADALGLLTDQYRGAPVTHYLNEQAR

cons                                *      :

ref_avrbs2    QTET-D
XajI22        QIEPGE
Xaj237        H-----
Xe367         H-----
Xe424         H-----
Xe426         H-----
Xaj427        QIEPGE
Xaj554        QIEPGE
Xe761         H-----
Xaj765        QIEPGE
Xe766         H-----
Xaj796        QIEPGE
Xaj1483       QIEPGE
Xaj1494       QIEPGE
Xaj1514       QIEPGE
Xaj1567       QIEPGE
Xaj1586       QIEPGE

cons          :
```

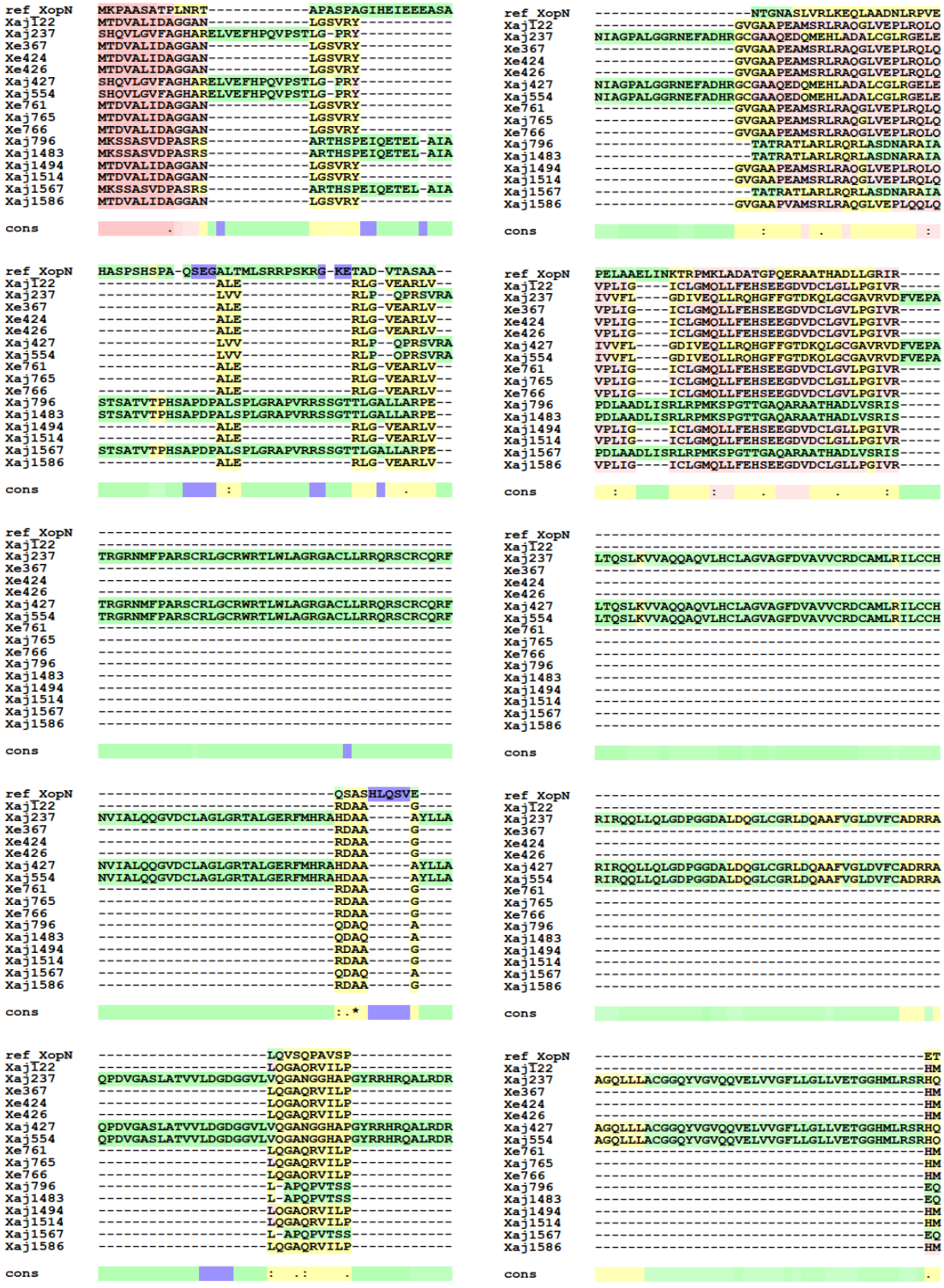
F.3 xopF1 multiple alignment

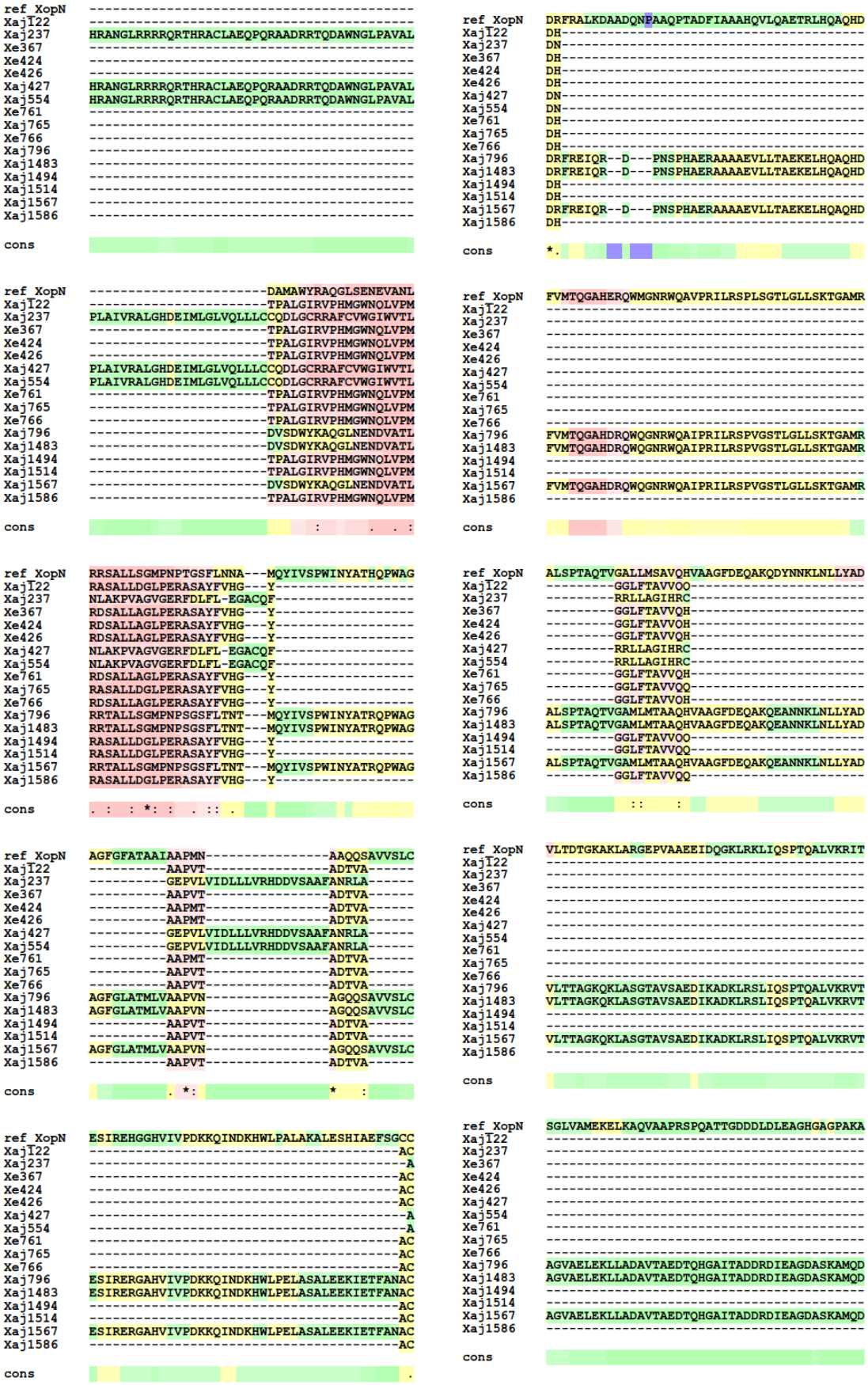


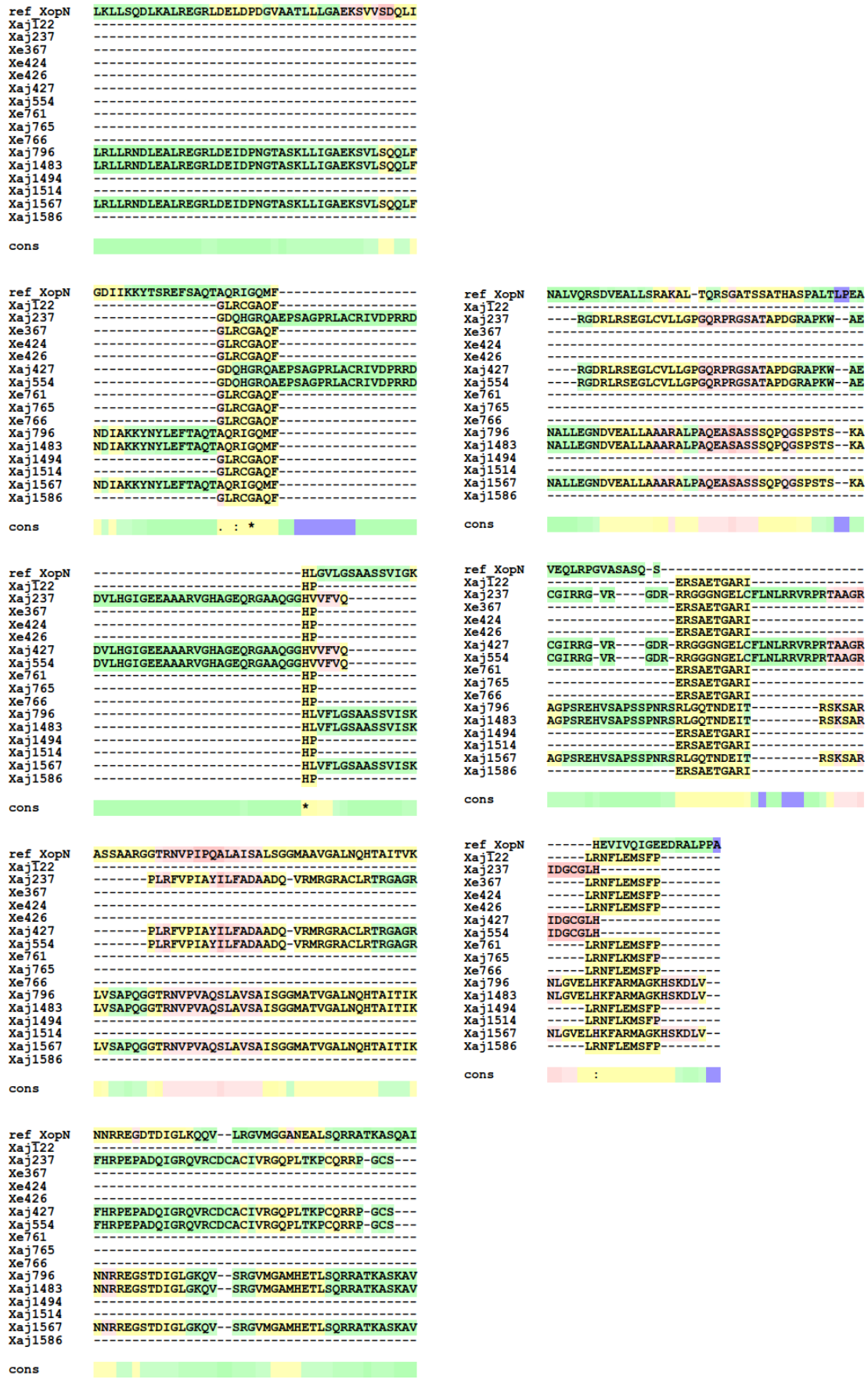

ref_xopF1 PVIN--AGFNAARRTGGEPGLLTPVGQWGLSALAIGGAG
XajI22 PTIN--AGFNAARRTSDDPGTRTPAGQLGLSALAIGGAG
Xaj237 GTRHRLAGFE----------
Xe424 PTIN--AGFNAARRTSADPGTRTPAGQLGLSALAIGGAG
Xe426 EKAD--AERAAQRRAAPVVGWTPPQAQPAIVG--IGASP
Xaj427 GTRHRLAGFE----------
Xaj554 GTRHRLAGFE----------
Xe761 EKAD--AERAAQRRAAPVVGWTPPQAQPAIVG--IGASP
Xaj765 PTIN--AGFNAARRTSDDPGTRTPAGQLGLSALAIGGAG
Xe766 PTIN--AGFNAARRTSADPGTRTPAGQLGLSALAIGGAG
Xaj796 PTIN--AGFNAARRISDDPGTRTPAGQLGLSALAIGGAG
Xaj1483 PTIN--AGFNAARRTSDDPGTRTPAGQLGLSALAIGGAG
Xaj1494 PTIN--AGFNAARRTSDDPGTRTPAGQLGLSALAIGGAG
Xaj1514 PTIN--AGFNAARRTSDDPGTRTPAGQLGLSALAIGGAG
Xaj1567 PTIN--AGFNAARRTSDDPGTRTPAGQLGLSALAIGGAG
Xaj1586 PTIN--AGFNAARRTSDDPGTRTPAGQLGLSALAIGGAG
cons
ref_xopF1 ----------APLRRGS-HIR-VAGEAVNSTAVV
XajI22 ----------APLRGGNANGS-VAGEAANSTATV
Xaj237 ----------SD
Xe424 ----------APLRGGNANGS-VAGEAANSTATV
Xe426 PSDTANLDIARVLGLATAQGGPTSHTAILSRTLG-LPAL
Xaj427 ----------SD
Xaj554 ----------SD
Xe761 PSDTANLDIARVLGLATAQGGPTSHTAILSRTLG-LPAL
Xaj765 ----------APLRGGNANGS-VAGEAANSTATV
Xe766 ----------APLRGGNANGS-VAGEAANSTATV
Xaj796 ----------APLRGGNANGS-VAGEAANSTATV
Xaj1483 ----------APLRGGNANGS-VAGEAANSTATV
Xaj1494 ----------APLRGGNANGS-VAGEAANSTATV
Xaj1514 ----------APLRGGNANGS-VAGEAANSTATV
Xaj1567 ----------APLRGGNANGS-VAGEAANSTATV
Xaj1586 ----------APLRGGNANGS-VAGEAANSTATV
cons
ref_xopF1 VVQKALLETGKAVARTGQMRVPDLVGG--------EQRL
XajI22 VVQKAMLETGKAVARTGQMSVPDLLGG--------QQRL
Xaj237 --HRLLH---HPCPANGECAQPQLAGR--------RARA
Xe424 VVQKAMLETGKAVARTGQMRVPDLLGG--------QQRL
Xe426 GVAIGIVHRLRA-AQTEVADQPVGLGDGGALLHDALTRT
Xaj427 --HRLLH---HPCPANGECAQPQLAGR--------RARA
Xaj554 --HRLLH---HPCPANGECAQPQLAGR--------RARA
Xe761 GVAIGIVHRLRA-AQTEVADQPVGLGDGGALLHDALTRT
Xaj765 VVQKAMLETGKAVARTGQMSVPDLLGG--------QQRL
Xe766 VVQKAMLETGKAVARTGQTRVPDLLGG--------QQRL
Xaj796 VVQKAMLETGKAVARTGQMRVPDLLGG--------QQRL
Xaj1483 VVQKAMLETGKAVARTGQMRVPDLLGG--------QQRL
Xaj1494 VVQKAMLETGKAVARTGQMSVPDLLGG--------QQRL
Xaj1514 VVQKAMLETGKAVARTGQMSVPDLLGG--------QQRL
Xaj1567 VVQKAMLETGKAVARTGQMRVPDLLGG--------QQRL
Xaj1586 VVQKAMLETGKAVARTGQMSVPDLLGG--------QQRL
cons
ref_xopF1 VQQAVQ-T-LFNDTFWNALKAKNGANTSQAAR-LD----
XajI22 VQQAVQ-T-LFNDTVWNALKAKNGANTTQATR-LD----
Xaj237 LQQGRWRI-GHQDLGVDGLEPRQVGNARAPVHCRD----
Xe424 VQQAAQ-T-LFNDTIWNALKAKNGANTTQATR-LD----
Xe426 VAAGGQ-LLDIEDGVTAIIDGSSGRLYINPSA-LDLDAA
Xaj427 LQQGRWRI-GHQDLGVDGLEPRQVGNARAPVHCRD----
Xaj554 LQQGRWRI-GHQDLGVDGLEPRQVGNARAPVHCRD----
Xe761 VAAGGQ-LLDIEDGVTAIIDGSSGRLYINPSA-LDLDAA
Xaj765 VQQAVQ-T-LFNDTVWNALKAKNGANTTQATR-LD----
Xe766 VQQAAQ-T-LFNDTIWNALKAKNGANTTQATR-LD----
Xaj796 VQQAVQ-T-LFNDTVWNALKAKNGANTTQATR-LD----
Xaj1483 VQQAVQ-T-LFNDTVWNALKAKNGANTTQATR-LD----
Xaj1494 VQQAVQ-T-LFNDTVWNALKAKNGANTTQATR-LD----
Xaj1514 VQQAVQ-T-LFNDTVWNALKAKNGANTTQATR-LD----
Xaj1567 VQQAVQ-T-LFNDTVWNALKAKNGANTTQATR-LD----
Xaj1586 VQQAVQ-T-LFNDTVWNALKAKNGANTTQATR-LD----
cons
ref_xopF1 NLFALALPDKARRPAQWSDAVHFPPTYLLDTGKEALAL-
XajI22 NLFALALPDKTRRPAQWSDAVRFPPNYLLETGKEGLAL-
Xaj237 R----VVGDSPRRVESRVD------RRLEQQGVQGLVVQ
Xe424 NLFALALPDKTRRPAQWSDAVHFP-NYLLETGMEGLAL-
Xe426 RQQLAAIQDDTQRRLGASDAAIFKAQAELLNDTDLITR-
Xaj427 R----VVGDSPRRVESRVD------RRLEQQGVQGLVVQ
Xaj554 R----VVGDSPRRVESRVD------RRLEQQGVQGLVVQ
Xe761 RQQLAAIQDDTQRRLGASDAAIFKAQAELLNDTDLITR-
Xaj765 NLFALALPDKTRRPAQWSDAVRLPPNYLLETGKEGLAL-
Xe766 NLFALALPDKTRRPAQWSDAVHFP-NYLLETGMEGLAL-
Xaj796 NLFSLALPDKTRRPAQWSDAVRFPPNYLLETGKEGLAL-
Xaj1483 NLFSLALPDKTRRPAQWSDAVRFPPNYLLETGKEGLAL-
Xaj1494 NLFALALPDKTRRPAQWSDAVRFPPNYLLETGKEGLAL-
Xaj1514 NLFALALPDKTRRPAQWSDAVRLPPNYLLETGKEGLAL-
Xaj1567 NLFSLALPDKTRRPAQWSDAVRFPPNYLLETGKEGLAL-
Xaj1586 NLFALALPDKTRRPAQWSDAVRFPPNYLLETGKEGLAL-
cons
ref_xopF1 -----N----------ERAALAA----------
XajI22 -----Q----------QRAVSAA----------
Xaj237 ------Q---------QRIAGIG----------
Xe424 -----Q----------ERAASAA----------
Xe426 RTHIAEQQAIRERERAAQRALPAETTDGHHIDIGANVNLP
Xaj427 ------Q---------QRIAGIG----------
Xaj554 ------Q---------QRIAGIG----------
Xe761 RTHIAEQQAIRERERAAQRALPAETTDGHHIDIGANVNLP
Xaj765 -----Q----------QRAVSAA----------
Xe766 -----Q----------ERAASAA----------
Xaj796 -----Q----------QRAVRAA----------
Xaj1483 -----Q----------QRAVRAA----------
Xaj1494 -----Q----------QRAVSAA----------
Xaj1514 -----Q----------QRAVSAA----------
Xaj1567 -----Q----------QRAVRAA----------
Xaj1586 -----Q----------QRAVSAA----------
cons
ref_xopF1 AROGFNSAN----------AVAT
XajI22 ARQALSTTN----------AVTT
Xaj237 GRLPLGDAL----------AEVV
Xe424 ARQALSTTN----------AITT
Xe426 TCQLMVEGHGVAWSWHQAVEQIASGLAALGNPVLAGRAA
Xaj427 GRLPLGDAL----------AEVV
Xaj554 GRLPLGDAL----------AEVV
Xe761 TCQLMVEGHGVAWSWHQAVEQIASGLAALGNPVLAGRAA
Xaj765 ARQALSTTN----------AVTT
Xe766 ARQALSTTN----------AITT
Xaj796 ARQALSTTN----------AVTT
Xaj1483 ARQALSTTN----------AVTT
Xaj1494 ARQALSTTN----------AVTT
Xaj1514 ARQALSTTN----------AVTT
Xaj1567 ARQALSTTN----------AVTT
Xaj1586 ARQALSTTN----------AVTT
cons
ref_xopF1 ----------EHQRT---IERTLEALAEPVDQAI-
XajI22 ----------SHQRT---IARTLQSLAEPINEAI-
Xaj237 ----------HQPRHRAGQRRRTQRLGSGRTRAV-
Xe424 ----------GHQRT---IAGTLQALAEPVEAAI-
Xe426 DQVAMALTQGAEGVGLMRT---EFLFLESGATPSEDEQY
Xaj427 ----------HQPRHRAGQRRRTQRLGSGRTRAV-
Xaj554 ----------HQPRHRAGQRRRTQRLGSGRTRAV-
Xe761 DQVAMALTQGAEGVGLMRT---EFLFLESGATPSEDEQY
Xaj765 ----------SHQRT---IARTLQSLAEPINEAI-
Xe766 ----------DHQRT---IAGTLQALAEPVEAAI-
Xaj796 ----------SHQRT---IARTLQSLAEPISEAI-
Xaj1483 ----------SHQRT---IARTLQSLAEPISEAI-
Xaj1494 ----------SHQRT---IARTLQSLAEPINEAI-
Xaj1514 ----------SHQRT---IARTLQSLAEPINEAI-
Xaj1567 ----------SHQRT---IARTLQSLAEPISEAI-
Xaj1586 ----------SHQRT---IARTLQSLAEPINEAI-
cons
ref_xopF1 TARDLLTRHMLSNVMASFGPMGAGRIIT----------
XajI22 ALRDVLGRHMLSNVLANFASLGAGRLIA----------
Xaj237 GLRLELRKRGCVALARACETLALG----------
Xe424 AVRDVLGRHMLSNVLANFASLGAGRLIA----------
Xe426 DLRDV-GRRVLAQLDPAAAGAGLTDLPEQPCILLAGDLS
Xaj427 GLRLELRKRGCVALARACETLALG----------
Xaj554 GLRLELRKRGCVALARACETLALG----------
Xe761 DLRDV-GRRVLAQLDPAAAGAGLTDLPEQPCILLAGDLS
Xaj765 ALRDVLGRHMLSNVLANFASLGAGRLIA----------
Xe766 AVRDVLGRHMLSNVLANFASLGAGRLIA----------
Xaj796 ALRDVLGRHMLSNLLANFASLGAGRLIA----------
Xaj1483 ALRDVLGRHMLSNLLANFASLGAGRLIA----------
Xaj1494 ALRDVLGRHMLSNVLANFASLGAGRLIA----------
Xaj1514 ALRDVLGRHMLSNVLANFASLGAGRLIA----------
Xaj1567 ALRDVLGRHMLSNLLANFASLGAGRLIA----------
Xaj1586 ALRDVLGRHMLSNVLANFASLGAGRLIA----------
cons
ref_xopF1 -ARL----------
XajI22 -AMF----------
Xaj237 -VGC----------
Xe424 -ALL----------
Xe426 QTYLAMAQALDGRPLIVRALDIGGDKQVAHLELPHEENP
Xaj427 -VGC----------
Xaj554 -VGC----------
Xe761 QTYLAMAQALDGRPLIVRALDIGGDKQVAHLELPHEENP
Xaj765 -AMF----------
Xe766 -ALF----------
Xaj796 -AMF----------
Xaj1483 -AMF----------
Xaj1494 -AMF----------
Xaj1514 -AMF----------
Xaj1567 -AMF----------
Xaj1586 -AMF----------
cons


ref_xopF1
XajI22
Xaj237
Xe424
Xe426
Xaj427
Xaj554
Xe761
Xaj765
Xe766
Xaj796
Xaj1483
Xaj1494
Xaj1514
Xaj1567
Xaj1586
cons

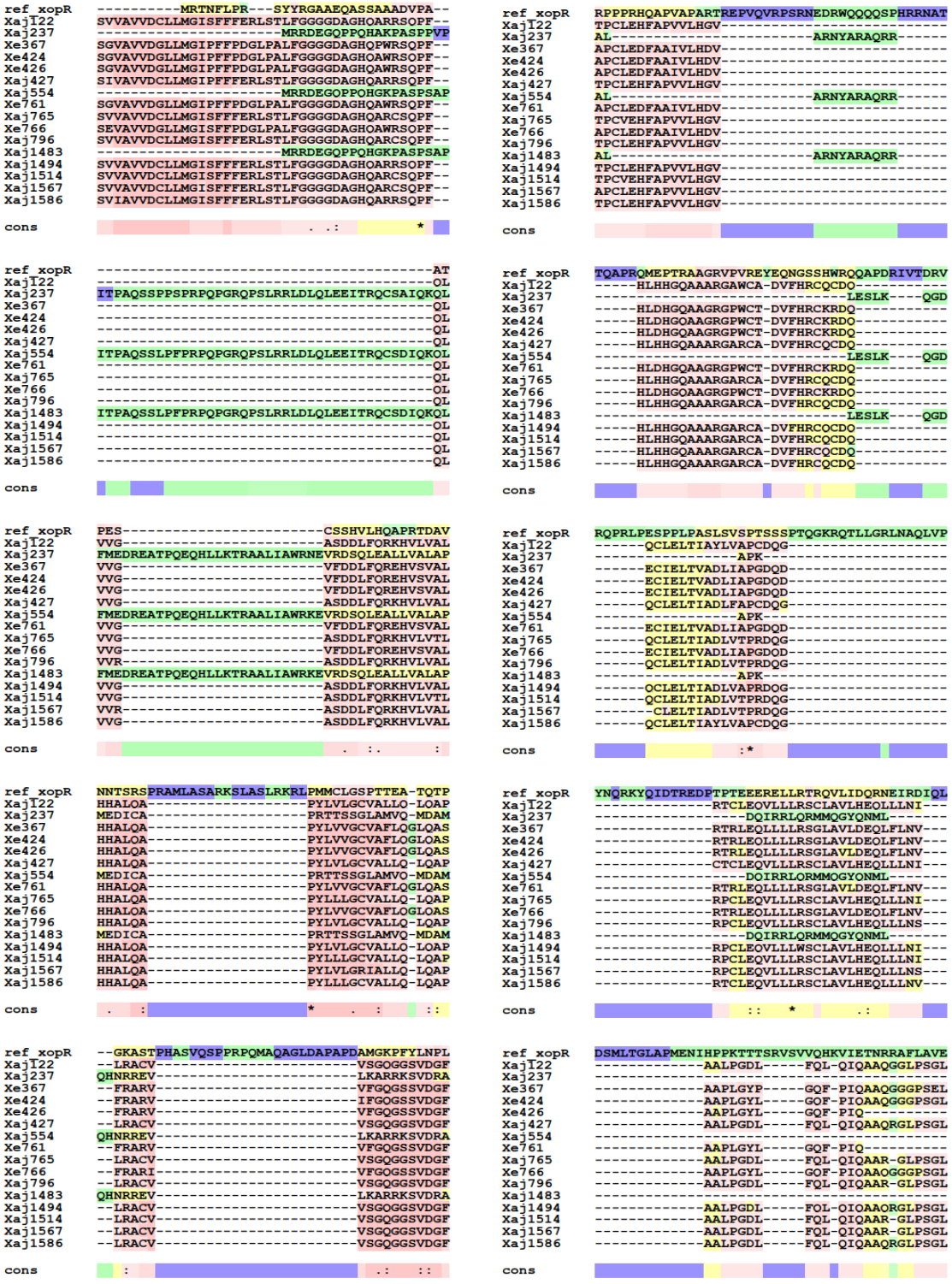
F.4 xopN multiple alignment

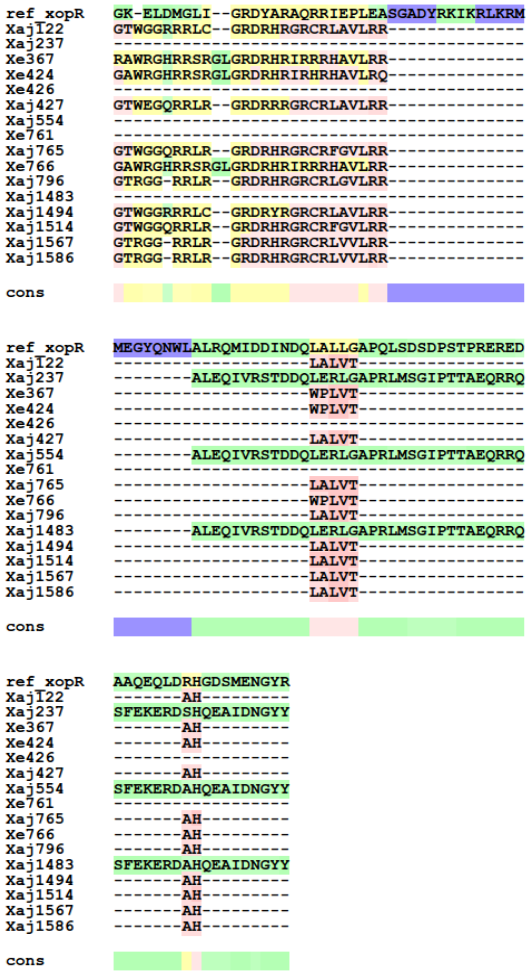



ref_XopN
XajI22
Xaj237
Xe367
Xe424
Xe426
Xaj427
Xaj554
Xe761
Xaj765
Xe766
Xaj796
Xaj1483
Xaj1494
Xaj1514
Xaj1567
Xaj1586
cons
HRANGLRRRRQRTHRACLAEQPQRAADRRTQDAWNGLPAVAL
DRFRALKDAADQNPAAQPTADFIAAAHQVLQAETRLHQAQHD
DAMAWYRAQGLSENEVANL
TPALGIRVPHMGWNQLVPM
PLAIVRALGHDEIMLGLVQLLLCCQDLGCRRAFCVWGIWVTL
FVMTQGAHERQWMGNRWQAVPRILRSPLSGTLGLLSKTGAMR
RRSALLSGMPNPTGSFLNNA---MQYIVSPWINYATHQPWAG
ALSPTAQTVGALLMSAVQHVAAGFDEQAKQDYNNKLNLLYAD
AGFGFATAAIAAPMN----------AAQQSAVVSLC
VLTDTGKAKLARGEPVAAEEIDQGKLRKLIQSPTQALVKRIT
ESIREHGGHVIVPDKKQINDKHWLPALAKALESHIAEFSGCC
SGLVAMEKELKAQVAAPRSPQATTGDDDLDLEAGHGAGPAKA
FVMTQGAHDRQWQGNRWQAIPRILRSPVGSTLGLLSKTGAMR
ALSPTAQTVGAMLMTAAQHVAAGFDEQAKQEANNKLNLLYAD
VLTTAGKQKLASGTAVSAEDIKADKLRSLIQSPTQALVKRVT
AGVAELEKLLADAVTAEDTQHGAITADDRDIEAGDASKAMQD


ref_XopN
XajI22
Xaj237
Xe367
Xe424
Xe426
Xaj427
Xaj554
Xe761
Xaj765
Xe766
Xaj796
Xaj1483
Xaj1494
Xaj1514
Xaj1567
Xaj1586
cons

F.5 xopR multiple alignment




ref_xopR GK-ELDMGLI--GRDYARAQRRIEPLEASGADYRKIKRLKRM
XajI22 GTWGGRRRLC--GRDRHRGRCRLAVLRR--------------
Xaj237 --
Xe367 RAWRGHRRSRGLGRDRHRIRRRHAVLRR--------------
Xe424 GAWRGHRRSRGLGRDRHRIRHRHAVLRQ--------------
Xe426 --
Xaj427 GTWEGQRRLR--GRDRRRGRCRLAVLRR--------------
Xaj554 --
Xe761 --
Xaj765 GTWGGQRRLR--GRDRHRGRCRFGVLRR--------------
Xe766 GAWRGHRRSRGLGRDRHRIRRRHAVLRR--------------
Xaj796 GTRGG-RRLR--GRDRHRGRCRLGVLRR--------------
Xaj1483 --
Xaj1494 GTWGGRRRLC--GRDRYRGRCRLAVLRR--------------
Xaj1514 GTWGGQRRLR--GRDRHRGRCRFGVLRR--------------
Xaj1567 GTRGG-RRLR--GRDRHRGRCRLVVLRR--------------
Xaj1586 GTRGG-RRLR--GRDRHRGRCRLVVLRR--------------
cons
ref_xopR MEGYQNWLALRQMIDDINDQLALLGAPQLSDSDPSTPRERED
XajI22 --------------------LALVT-----------------
Xaj237 --------ALEQIVRSTDDQLERLGAPRLMSGIPTTAEQRRQ
Xe367 --------------------WPLVT-----------------
Xe424 --------------------WPLVT-----------------
Xe426 --
Xaj427 --------------------LALVT-----------------
Xaj554 --------ALEQIVRSTDDQLERLGAPRLMSGIPTTAEQRRQ
Xe761 --
Xaj765 --------------------LALVT-----------------
Xe766 --------------------WPLVT-----------------
Xaj796 --------------------LALVT-----------------
Xaj1483 --------ALEQIVRSTDDQLERLGAPRLMSGIPTTAEQRRQ
Xaj1494 --------------------LALVT-----------------
Xaj1514 --------------------LALVT-----------------
Xaj1567 --------------------LALVT-----------------
Xaj1586 --------------------LALVT-----------------
cons
ref_xopR AAQEQLDRHGDSMENGYR
XajI22 -------AH---------
Xaj237 SFEKERDSHQEAIDNGYY
Xe367 -------AH---------
Xe424 -------AH---------
Xe426 ------------------
Xaj427 -------AH---------
Xaj554 SFEKERDAHQEAIDNGYY
Xe761 ------------------
Xaj765 -------AH---------
Xe766 -------AH---------
Xaj796 -------AH---------
Xaj1483 SFEKERDAHQEAIDNGYY
Xaj1494 -------AH---------
Xaj1514 -------AH---------
Xaj1567 -------AH---------
Xaj1586 -------AH---------
cons

Appendix G

Phylogenies of *Xanthomonas* spp. isolates

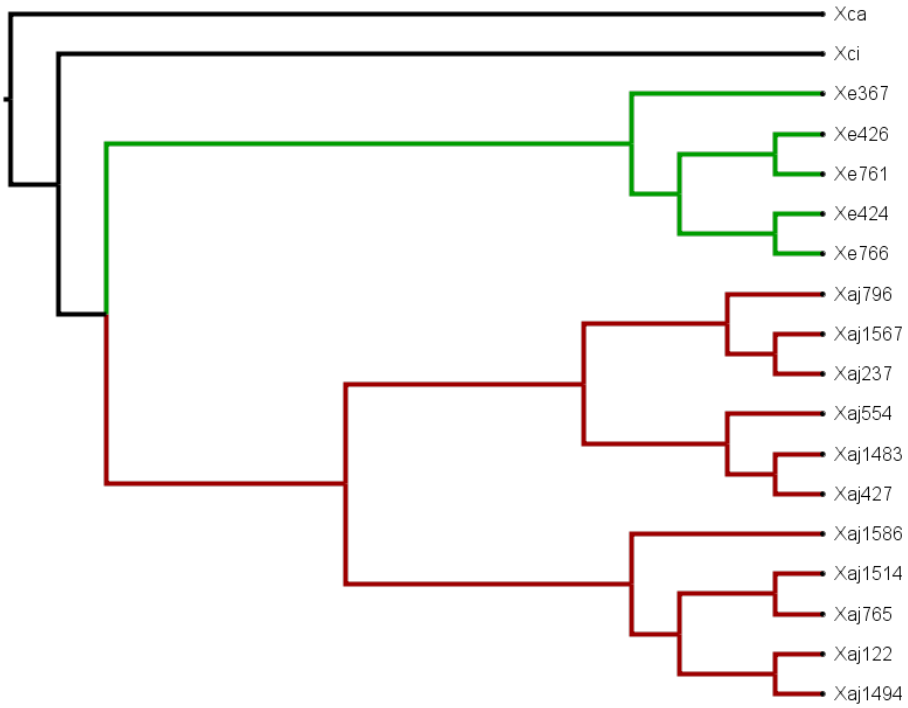


FIGURE G.1: Phylogeny based on average nucleotide identities.

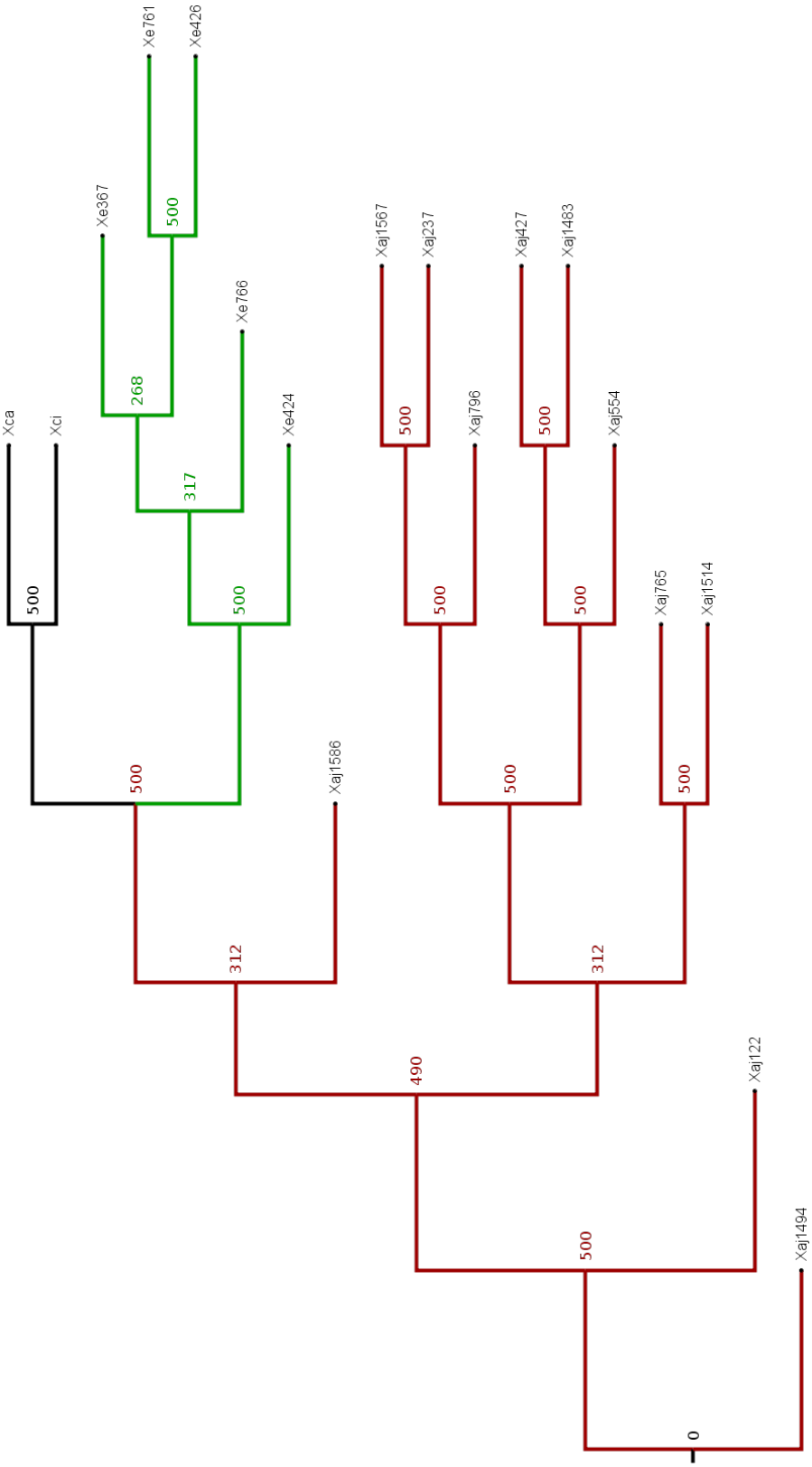


FIGURE G.2: Phylogeny based on BUSCO genes.

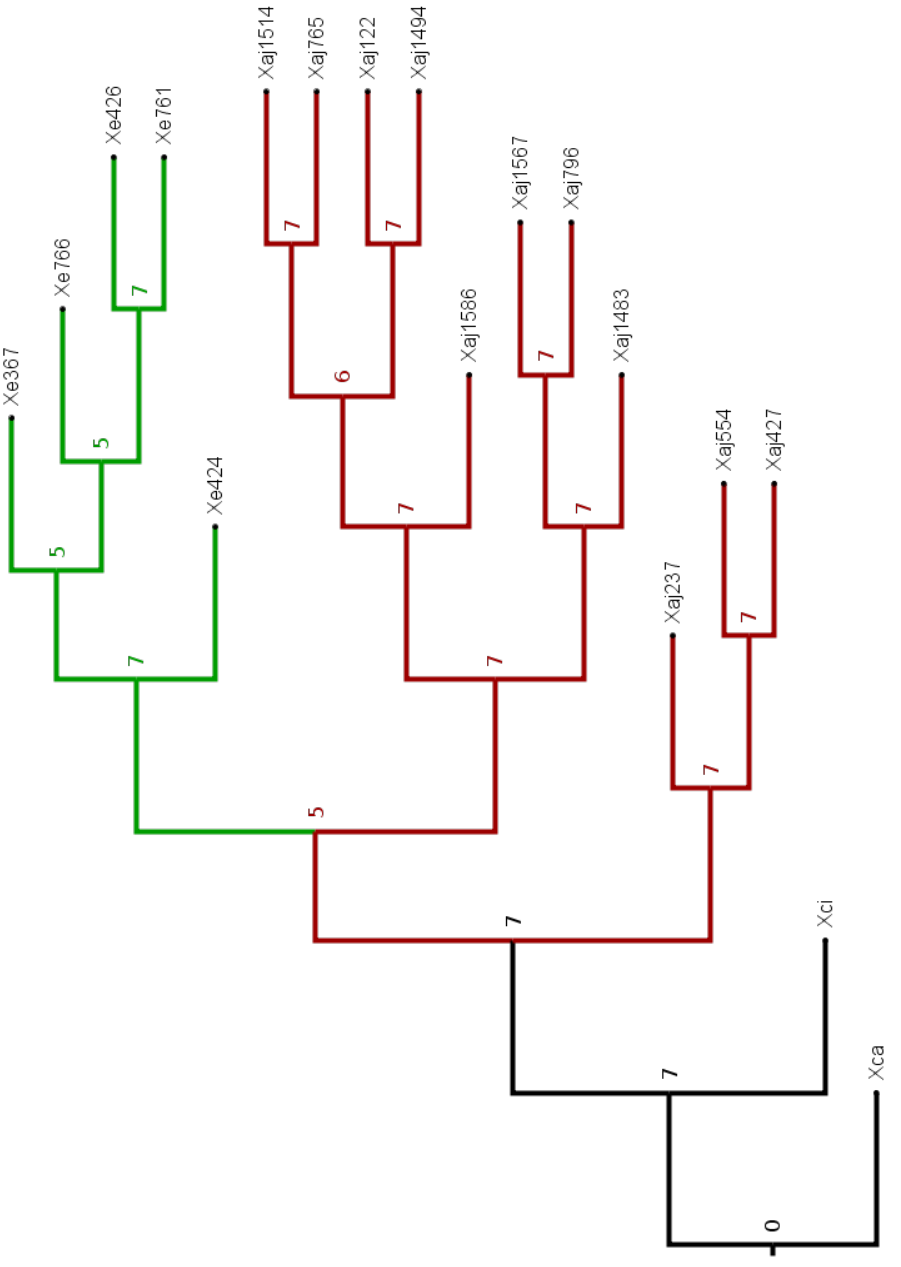


FIGURE G.3: Phylogeny based on k-mer frequencies.

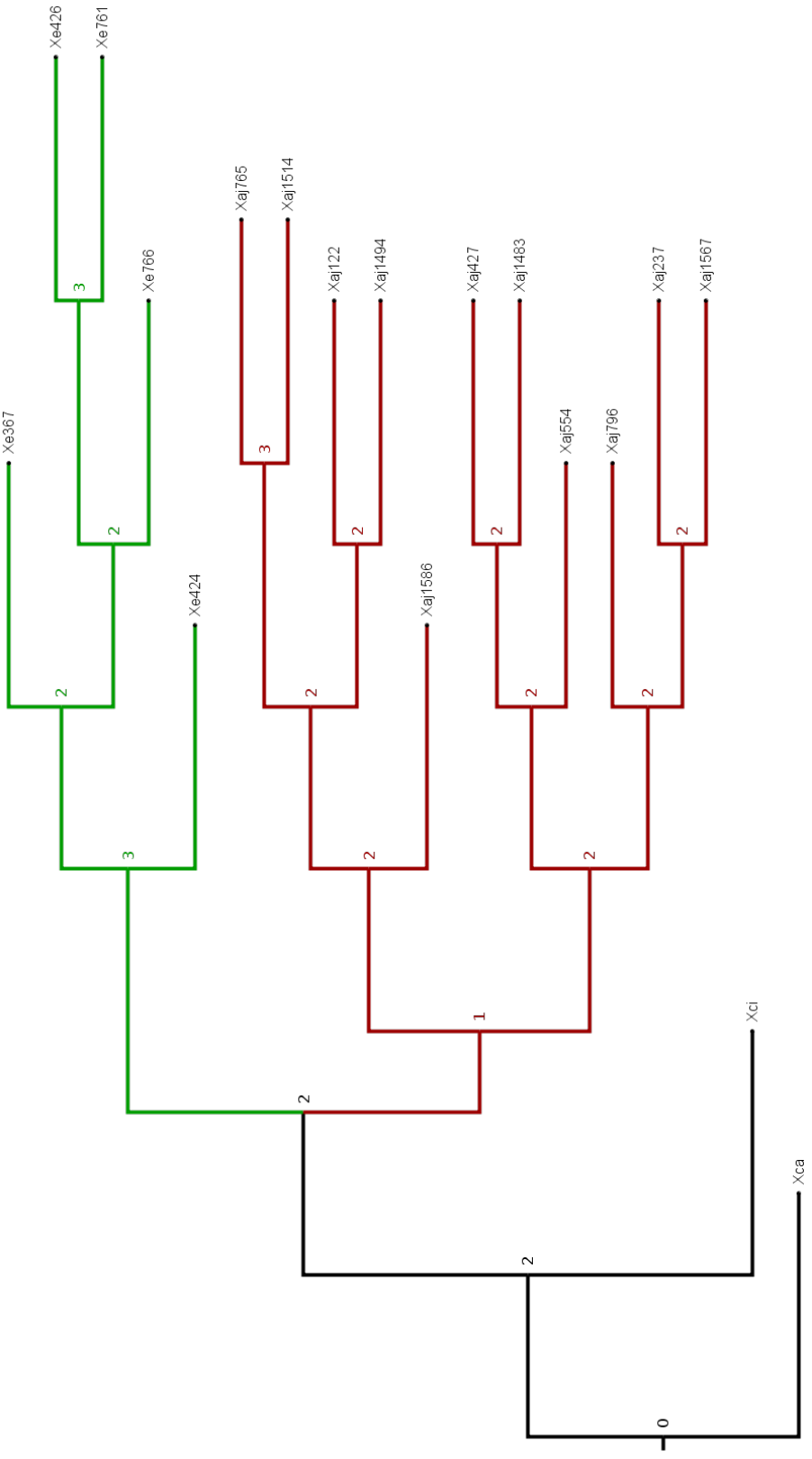


FIGURE G.4: Consensus phylogeny.

Bibliography

[1] Our world in data: World population growth. Accessed: 2020-07-30. [Online]. Available: https://ourworldindata.org/world-population-growth [Cited on page 1.]

[2] R. M. Welch and R. D. Graham, "A new paradigm for world agriculture: meeting human needs: productive, sustainable, nutritious," *Field crops research*, vol. 60, no. 1-2, pp. 1–10, 1999. [Cited on page 1.]

[3] E. H. Stukenbrock and B. A. McDonald, "The origins of plant pathogens in agro-ecosystems," *Annual Review of Phytopathology*, vol. 46, pp. 75–100, 2008. [Cited on pages 1, 14, and 29.]

[4] M. G. Milgroom and T. L. Peever, "Population biology of plant pathogens," *Plant Disease*, vol. 87, no. 6, pp. 608–617, 2003. [Cited on pages 1 and 2.]

[5] M.-A. Jacques, M. Arlat, A. Boulanger, T. Boureau, S. Carrère, S. Cesbron, N. W. Chen, S. Cociancich, A. Darrasse, N. Denancé, M. Fischer-Le Saux, L. Gagnevin, R. Koebnik, E. Lauber, L. D. Noël, I. Pieretti, P. Portier, O. Pruvost, A. Rieux, I. Robène, M. Royer, B. Szurek, V. Verdier, and C. Vernière, " Using Ecology, Physiology, and Genomics to Understand Host Specificity in Xanthomonas ," *Annual Review of Phytopathology*, vol. 54, no. 1, pp. 163–187, 2016. [Cited on pages 1, 2, and 30.]

[6] R. P. Ryan, F. J. Vorhölter, N. Potnis, J. B. Jones, M. A. Van Sluys, A. J. Bogdanove, and J. M. Dow, "Pathogenomics of Xanthomonas: Understanding bacterium-plant interactions," *Nature Reviews Microbiology*, vol. 9, no. 5, pp. 344–355, 2011. [Cited on pages 1 and 30.]

[7] C. Fernandes, P. Albuquerque, L. Cruz, and F. Tavares, "Genotyping and epidemiological metadata provides new insights into population structure of xanthomonas

isolated from walnut trees," *bioRxiv*, p. 397703, 2018. [Cited on pages 1, 2, 17, 39, and 42.]

[8] L. Martins, C. Fernandes, J. Blom, N. C. Dia, J. F. Pothier, and F. Tavares, "Xanthomonas euroxanthea sp. nov., a new xanthomonad species including pathogenic and non-pathogenic strains of walnut," *International Journal of Systematic and Evolutionary Microbiology*, p. ijsem004386, 2020. [Cited on pages 1, 6, 39, and 44.]

[9] S. Parnell, F. Van Den Bosch, T. Gottwald, and C. A. Gilligan, "Surveillance to Inform Control of Emerging Plant Diseases: An Epidemiological Perspective," *Annual Review of Phytopathology*, vol. 55, no. January 2019, pp. 591–610, 2017. [Cited on pages 2 and 14.]

[10] P. Albuquerque, M. V. Mendes, C. L. Santos, P. Moradas-Ferreira, and F. Tavares, "DNA signature-based approaches for bacterial detection and identification," *Science of the Total Environment*, vol. 407, no. 12, pp. 3641–3651, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.scitotenv.2008.10.054 [Cited on pages 2 and 7.]

[11] D. Emerson, L. Agulto, H. Liu, and L. Liu, "Identifying and Characterizing Bacteria in an Era of Genomics and Proteomics," *BioScience*, vol. 58, no. 10, pp. 925–936, 2008. [Cited on pages 2 and 6.]

[12] A. M. Phillippy, K. Ayanbule, N. J. Edwards, and S. L. Salzberg, "Insignia: a dna signature search web server for diagnostic assay development," *Nucleic acids research*, vol. 37, no. suppl_2, pp. W229–W234, 2009. [Cited on page 2.]

[13] P. Albuquerque, C. M. Caridade, A. S. Rodrigues, A. R. Marcal, J. Cruz, L. Cruz, C. L. Santos, M. V. Mendes, and F. Tavares, "Evolutionary and experimental assessment of novel markers for detection of xanthomonas euvesicatoria in plant samples," *PLoS ONE*, vol. 7, no. 5, 2012. [Cited on page 2.]

[14] H. C. McCann, "Skirmish or war: the emergence of agricultural plant pathogens," *Current Opinion in Plant Biology*, vol. 56, pp. 147–152, 2020. [Online]. Available: https://doi.org/10.1016/j.pbi.2020.06.003 [Cited on pages 2 and 14.]

[15] L. Vauterin, J. Rademaker, and J. Swings, "Synopsis on the taxonomy of the genus Xanthomonas," *Phytopathology*, vol. 90, no. 7, pp. 677–682, 2000. [Cited on page 5.]

[16] A. K. Tsang, H. H. Lee, S. M. Yiu, S. K. Lau, and P. C. Woo, "Failure of phylogeny inferred from multilocus sequence typing to represent bacterial phylogeny," *Scientific Reports*, vol. 7, no. 1, pp. 1–12, 2017. [Cited on pages 2 and 7.]

[17] S. R. Vartoukian, R. M. Palmer, and W. G. Wade, "Strategies for culture of 'unculturable' bacteria," *FEMS microbiology letters*, 2010. [Cited on page 2.]

[18] B. L. Brown, M. Watson, S. S. Minot, M. C. Rivera, and R. B. Franklin, "MinIONTM nanopore sequencing of environmental metagenomes: A synthetic approach," *GigaScience*, vol. 6, no. 3, pp. 1–10, 2017. [Cited on pages 2 and 15.]

[19] A. D. Tyler, L. Mataseje, C. J. Urfano, L. Schmidt, K. S. Antonation, M. R. Mulvey, and C. R. Corbett, "Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications," *Scientific Reports*, vol. 8, no. 1, pp. 1–12, 2018. [Online]. Available: http://dx.doi.org/10.1038/s41598-018-29334-5

[20] N. Kono and K. Arakawa, "Nanopore sequencing: Review of potential applications in functional genomics," *Development Growth and Differentiation*, vol. 61, no. 5, pp. 316–326, 2019. [Cited on page 2.]

[21] H. Lu, F. Giordano, and Z. Ning, "Oxford Nanopore MinION Sequencing and Genome Assembly," *Genomics, Proteomics and Bioinformatics*, vol. 14, no. 5, pp. 265–279, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.gpb.2016.05.004 [Cited on pages 2, 9, and 10.]

[22] S. Goldstein, L. Beka, J. Graf, and J. L. Klassen, "Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing," *BMC Genomics*, vol. 20, no. 1, pp. 1–17, 2019. [Cited on page 2.]

[23] R. P. Ryan, F. J. Vorhölter, N. Potnis, J. B. Jones, M. A. Van Sluys, A. J. Bogdanove, and J. M. Dow, "Pathogenomics of Xanthomonas: Understanding bacterium-plant interactions," *Nature Reviews Microbiology*, vol. 9, no. 5, pp. 344–355, 2011. [Cited on page 5.]

[24] J. R. Lamichhane, "Xanthomonas arboricola diseases of tone fruit, Almond, and walnut trees: Progress toward understanding and management," *Plant Disease*, vol. 98, no. 12, pp. 1600–1610, 2014. [Cited on page 5.]

[25] J. R. Lamichhane and L. Varvaro, "Xanthomonas arboricola disease of hazelnut : current status and future perspectives for its management," *Plant pathology*, pp. 243–254, 2014.

[26] D. Frutos, "Bacterial diseases of walnut and hazelnut and genetic resources," *Journal of Plant Pathology*, vol. 92, no. 1 SUPPL., 2010. [Cited on page 5.]

[27] R. E. Smith, *Walnut culture in California: walnut blight*. Agricultural Experiment Station, 1912, no. 231. [Cited on page 5.]

[28] C. Moragrega, J. Matias, N. Aletà, E. Montesinos, and M. Rovira, "Apical necrosis and premature drop of Persian (English) walnut fruit caused by xanthomonas arboricola pv. juglandis," *Plant Disease*, vol. 95, no. 12, pp. 1565–1570, 2011. [Cited on page 5.]

[29] A. Hajri, D. Meyer, F. Delort, J. Guillaumès, C. Brin, and C. Manceau, "Identification of a genetic lineage within Xanthomonas arboricola pv. juglandis as the causal agent of vertical oozing canker of Persian (English) walnut in France," *Plant Pathology*, vol. 59, no. 6, pp. 1014–1022, 2010. [Cited on page 5.]

[30] M. L. Martínez, D. O. Labuckas, A. L. Lamarque, and D. M. Maestri, "Walnut (Juglans regia L.): Genetic resources, chemistry, by-products," *Journal of the Science of Food and Agriculture*, vol. 90, no. 12, pp. 1959–1967, 2010. [Cited on page 5.]

[31] H. Y. Wetzstein, A. P. M. Rodriguez, J. A. Burns, and H. N. Magner, *Carya illinoensis (Pecan)*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996, pp. 50–75. [Online]. Available: https://doi.org/10.1007/978-3-662-10617-4{_}4 [Cited on page 5.]

[32] C. Fernandes, J. Blom, J. F. Pothier, and F. Tavares, "high-quality draft genome sequence of xanthomonas sp. strain cpbf 424, a walnut-pathogenic strain with atypical features," *Microbiology Resource Announcements*, vol. 7, no. 15, 2018. [Cited on pages 6, 17, and 39.]

[33] R. Franco-Duarte, L. Černáková, S. Kadam, K. S. Kaushik, B. Salehi, A. Bevilacqua, M. R. Corbo, H. Antolak, K. Dybka-Stępień, M. Leszczewicz, S. R. Tintino, V. C. A. de Souza, J. Sharifi-Rad, H. D. M. Coutinho, N. Martins, and C. F. Rodrigues, "Advances in chemical and biological methods to identify microorganismsfrom past to present," *Microorganisms*, vol. 7, no. 5, 2019. [Cited on page 6.]

[34] C. R. Woese and G. E. Fox, "Phylogenetic structure of the prokaryotic domain: The primary kingdoms," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 11, pp. 5088–5090, 1977. [Cited on page 6.]

[35] R. T. Espejo and N. Plaza, "Multiple Ribosomal RNA operons in bacteria; Their concerted evolution and potential consequences on the rate of evolution of their 16S rRNA," *Frontiers in Microbiology*, vol. 9, no. JUN, pp. 1–6, 2018. [Cited on page 6.]

[36] C. Ahlstrom, H. W. Barkema, K. Stevenson, R. N. Zadoks, R. Biek, R. Kao, H. Trewby, D. Haupstein, D. F. Kelton, G. Fecteau, O. Labrecque, G. P. Keefe, S. L. B. Mckenna, and J. D. Buck, "Limitations of variable number of tandem repeat typing identified through whole genome sequencing of Mycobacterium avium subsp . paratuberculosis on a national and herd level," *BMC genomics*, pp. 1–9, 2015. [Cited on page 7.]

[37] J. M. Janda, "Clinical Decisions: How Relevant is Modern Bacterial Taxonomy for Clinical Microbiologists?" *Clinical Microbiology Newsletter*, vol. 40, no. 7, pp. 51–57, 2018. [Online]. Available: http://dx.doi.org/10.1016/j.clinmicnews.2018.03.005 [Cited on page 7.]

[38] F. Hayashi, S. Anna, E. Bach, R. Z. Porto, F. Guella, E. Hayashi, S. Anna, L. M. P. Passaglia, F. Hayashi, S. Anna, E. Bach, R. Z. Porto, F. Guella, F. Hayashi, S. Anna, and E. B. Ã, "Critical Reviews in Microbiology Genomic metrics made easy : what to do and where to go in the new era of bacterial taxonomy bacterial taxonomy," *Critical Reviews in Microbiology*, vol. 45, no. 2, pp. 182–200, 2019. [Online]. Available: https://doi.org/10.1080/1040841X.2019.1569587

[39] F. M. Cohan, "What are Bacterial Species?" *Annual Review of Microbiology*, vol. 56, no. 1, pp. 457–487, 2002.

[40] ——, "Bacterial Species and Speciation," *Systematic Biology*, vol. 50, no. 4, pp. 513–524, 2001. [Cited on page 7.]

[41] A. C. Schürch, R. J. L. Willems, and R. V. Goering, "Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene e based approaches," *Clinical Microbiology and Infection*, vol. 24, no. 4, pp. 350–354, 2018. [Online]. Available: https://doi.org/10.1016/j.cmi.2017.12.016 [Cited on page 7.]

[42] S. Quainoo, J. P. M. Coolen, S. A. F. T. van Hijum, W. J. G. M. Martijn A. Huynen, W. van Schaik, and H. F. L. Wertheim, "Whole-Genome Sequencing of Bacterial Pathogens : the Future of Nosocomial," *Clinical microbiology reviews*, vol. 30, no. 4, pp. 1015–1064, 2017.

[43] S. R. Leopold, R. V. Goering, A. Witten, D. Harmsen, and A. Mellmann, "Standardized Analysis for Typing and Detection of Virulence and Antibiotic Resistance Genes," *Journal of clinical microbiology*, vol. 52, no. 7, pp. 2365–2370, 2014. [Cited on page 7.]

[44] J. Besser, H. A. Carleton, P. Gerner-Smidt, R. L. Lindsey, and E. Trees, "Next-generation sequencing technologies and their application to the study and control of bacterial infections," *Clinical Microbiology and Infection*, vol. 24, no. 4, pp. 335–341, 2018. [Online]. Available: https://doi.org/10.1016/j.cmi.2017.10.013 [Cited on page 7.]

[45] N. Van Goethem, T. Descamps, S. De Keersmaecker, D. Jamine, N. Roosens, H. Van Oyen, and A. Robert, "Status and potential of pathogen genomics for public health practice: a scoping review," *European Journal of Public Health*, vol. 28, no. suppl_4, pp. 1–16, 2018.

[46] P. Tang, M. A. Croxen, M. R. Hasan, W. W. L. Hsiao, and L. M. Hoang, "American Journal of Infection Control Infection control in the new age of genomic epidemiology," *AJIC: American Journal of Infection Control*, vol. 45, no. 2, pp. 170–179, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.ajic.2016.05.015 [Cited on page 7.]

[47] A. M. Maxam and W. Gilbert, "A new method for sequencing DNA. 1977." *Biotechnology (Reading, Mass.)*, vol. 24, no. 2, pp. 99–103, 1992. [Cited on page 7.]

[48] F. Sanger, S. Nicklen, and A. R. Coulson, "Dna sequencing with chain-terminating inhibitors," *Proceedings of the national academy of sciences*, vol. 74, no. 12, pp. 5463–5467, 1977. [Cited on page 7.]

[49] M. Kchouk, J. F. Gibrat, and M. Elloumi, "Generations of Sequencing Technologies: From First to Next Generation," *Biology and Medicine*, vol. 09, no. 03, 2017. [Cited on pages 8, 9, 10, 14, and 15.]

[50] K. C. Wong, J. Zhang, S. Yan, X. Li, Q. Lin, S. Kwong, and C. Liang, "DNA sequencing technologies: Sequencing data protocols and bioinformatics tools," *ACM Computing Surveys*, vol. 52, no. 5, 2019. [Cited on page 9.]

[51] W. R. McCombie, J. D. McPherson, and E. R. Mardis, "Next-generation sequencing technologies," *Cold Spring Harbor Perspectives in Medicine*, vol. 9, no. 11, 2019. [Cited on pages 9 and 10.]

[52] R. R. Wick, L. M. Judd, and K. E. Holt, "Performance of neural network basecalling tools for Oxford Nanopore sequencing," *Genome Biology*, vol. 20, no. 1, pp. 1–10, 2019. [Cited on page 9.]

[53] S. Morganti, P. Tarantino, E. Ferraro, P. DAmico, G. Viale, D. Trapani, B. A. Duso, and G. Curigliano, "Complexity of genome sequencing and reporting: Next generation sequencing (ngs) technologies and implementation of precision medicine in real life," *Critical reviews in oncology/hematology*, vol. 133, pp. 171–182, 2019. [Cited on page 9.]

[54] M. Baker, "De novo genome assembly: What every biologist should know," *Nature Methods*, vol. 9, no. 4, pp. 333–337, 2012. [Cited on pages 10 and 11.]

[55] R. Rizzi, S. Beretta, M. Patterson, Y. Pirola, M. Previtali, G. Della Vedova, and P. Bonizzoni, "Overlap graphs and de Bruijn graphs: data structures for de novo genome assembly in the big data era," *Quantitative Biology*, vol. 7, no. 4, pp. 278–292, 2019. [Cited on pages 10 and 11.]

[56] Z. Li, Y. Chen, D. Mu, J. Yuan, Y. Shi, H. Zhang, J. Gan, N. Li, X. Hu, B. Liu, B. Yang, and W. Fan, "Comparison of the two major classes of assembly algorithms: Overlap-layout-consensus and de-bruijn-graph," *Briefings in Functional Genomics*, vol. 11, no. 1, pp. 25–37, 2012. [Cited on pages 10 and 11.]

[57] N. De Maio, L. P. Shaw, A. Hubbard, S. George, N. D. Sanderson, J. Swann, R. Wick, M. AbuOun, E. Stubberfield, S. J. Hoosdally *et al.*, "Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes," *Microbial Genomics*, vol. 5, no. 9, 2019. [Cited on page 10.]

[58] J. Ruan and H. Li, "Fast and accurate long-read assembly with wtdbg2," *Nature methods*, vol. 17, no. 2, pp. 155–158, 2020. [Cited on page 16.]

[59] 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 research*, vol. 27, no. 5, pp. 722–736, 2017. [Cited on page 16.]

[60] H. Li, "Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences," *Bioinformatics*, vol. 32, no. 14, pp. 2103–2110, 2016. [Cited on pages 10 and 16.]

[61] 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 computational biology*, vol. 13, no. 6, p. e1005595, 2017. [Cited on pages 11, 19, and 32.]

[62] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski *et al.*, "Spades: a new genome assembly algorithm and its applications to single-cell sequencing," *Journal of computational biology*, vol. 19, no. 5, pp. 455–477, 2012. [Cited on page 11.]

[63] L. A. Yang, Y. J. Chang, S. H. Chen, C. Y. Lin, and J. M. Ho, "SQUAT: A Sequencing Quality Assessment Tool for data quality assessments of genome assemblies," *BMC Genomics*, vol. 19, no. Suppl 9, pp. 1–12, 2019. [Cited on page 11.]

[64] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler, "Quast: quality assessment tool for genome assemblies," *Bioinformatics*, vol. 29, no. 8, pp. 1072–1075, 2013. [Cited on pages 11 and 32.]

[65] F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, "BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs," *Bioinformatics*, vol. 31, no. 19, pp. 3210–3212, 2015. [Cited on pages 11, 18, and 32.]

[66] R. M. Waterhouse, M. Seppey, F. A. Simao, M. Manni, P. Ioannidis, G. Klioutchnikov, E. V. Kriventseva, and E. M. Zdobnov, "BUSCO applications from quality assessments to gene prediction and phylogenomics," *Molecular Biology and Evolution*, vol. 35, no. 3, pp. 543–548, 2018. [Cited on pages 11, 18, and 32.]

[67] E. L. Moss and A. S. Bhatt, "Generating closed bacterial genomes from long-read nanopore sequencing of microbiomes," *bioRxiv*, p. 489641, 2018. [Cited on page 15.]

[68] T. Charalampous, G. L. Kay, H. Richardson, A. Aydin, R. Baldan, C. Jeanes, D. Rae, S. Grundy, D. J. Turner, J. Wain, R. M. Leggett, D. M. Livermore, and J. O'Grady, "Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection," *Nature Biotechnology*, vol. 37, no. 7, pp. 783–792, 2019. [Online]. Available: http://dx.doi.org/10.1038/s41587-019-0156-5 [Cited on page 15.]

[69] S. Andrews, F. Krueger, A. Segonds-Pichon, L. Biggins, C. Krueger, and S. Wingett, "FastQC," Babraham Institute, Babraham, UK, Jan. 2012. [Cited on pages 16 and 31.]

[70] W. De Coster, S. DHert, 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, 03 2018. [Online]. Available: https://doi.org/10.1093/bioinformatics/bty149 [Cited on pages 16 and 31.]

[71] M. Kolmogorov, J. Yuan, Y. Lin, and P. A. Pevzner, "Assembly of long, error-prone reads using repeat graphs," *Nature biotechnology*, vol. 37, no. 5, pp. 540–546, 2019. [Cited on pages 16 and 32.]

[72] H. Li, "Minimap2: pairwise alignment for nucleotide sequences," *Bioinformatics*, vol. 34, no. 18, pp. 3094–3100, 2018. [Cited on page 16.]

[73] R. Vaser, I. Sović, N. Nagarajan, and M. Šikić, "Fast and accurate de novo genome assembly from long uncorrected reads," *Genome research*, vol. 27, no. 5, pp. 737–746, 2017. [Cited on pages 16 and 32.]

[74] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg, "Versatile and open software for comparing large genomes," *Genome biology*, vol. 5, no. 2, p. R12, 2004. [Cited on page 17.]

[75] J. Blom, S. P. Albaum, D. Doppmeier, A. Pühler, F. J. Vorhölter, M. Zakrzewski, and A. Goesmann, "EDGAR: A software framework for the comparative analysis of prokaryotic genomes," *BMC Bioinformatics*, vol. 10, 2009. [Cited on page 17.]

[76] C. Fernandes, J. Blom, J. F. Pothier, and F. Tavares, "High-quality draft genome sequence of xanthomonas arboricola pv. juglandis cpbf 1521, isolated from leaves of a symptomatic walnut tree in portugal without a past of phytosanitary treatment," *Microbiology Resource Announcements*, vol. 7, no. 16, 2018. [Cited on page 17.]

[77] N. Parkinson, V. Aritua, J. Heeney, C. Cowie, J. Bew, and D. Stead, "Phylogenetic analysis of xanthomonas species by comparison of partial gyrase b gene sequences," *International Journal of Systematic and Evolutionary Microbiology*, vol. 57, no. 12, pp. 2881–2887, 2007. [Cited on pages 18 and 30.]

[78] J. Young, D.-C. Park, H. Shearman, and E. Fargier, "A multilocus sequence analysis of the genus xanthomonas," *Systematic and applied microbiology*, vol. 31, no. 5, pp. 366–377, 2008.

[79] S. Marcelletti, P. Ferrante, and M. Scortichini, "Multilocus sequence typing reveals relevant genetic variation and different evolutionary dynamics among strains of xanthomonas arboricola pv. juglandis," *Diversity*, vol. 2, no. 11, pp. 1205–1222, 2010.

[80] Ž. Ivanović, T. Popović, J. Janse, M. Kojić, S. Stanković, V. Gavrilović, and D. Fira, "Molecular assessment of genetic diversity of xanthomonas arboricola pv. juglandis strains from serbia by various dna fingerprinting techniques," *European journal of plant pathology*, vol. 141, no. 1, pp. 133–145, 2015.

[81] D. Giovanardi, S. Bonneau, S. Gironde, M. Fischer-Le Saux, C. Manceau, and E. Stefani, "Morphological and genotypic features of xanthomonas arboricola pv. juglandis populations from walnut groves in romagna region, italy," *European Journal of Plant Pathology*, vol. 145, no. 1, pp. 1–16, 2016.

[82] M. Kałużna, J. Pulawska, M. Waleron, and P. Sobiczewski, "The genetic characterization of xanthomonas arboricola pv. juglandis, the causal agent of walnut blight in poland," *Plant pathology*, vol. 63, no. 6, pp. 1404–1416, 2014. [Cited on pages 18 and 30.]

[83] P. E. McKnight and J. Najab, "Mann-whitney u test," *The Corsini encyclopedia of psychology*, pp. 1–1, 2010. [Cited on page 19.]

[84] F. Wilcoxon, S. Katti, and R. A. Wilcox, "Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test," *Selected tables in mathematical statistics*, vol. 1, pp. 171–259, 1970. [Cited on page 19.]

[85] B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young *et al.*, "Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement," *PloS one*, vol. 9, no. 11, p. e112963, 2014. [Cited on pages 19 and 32.]

[86] B. Ewing and P. Green, "Base-calling of automated sequencer traces using phred. II. Error probabilities," *Genome Research*, vol. 8, no. 3, pp. 186–194, 1998. [Cited on page 20.]

[87] T. Coutinho, L. Van der Westhuizen, J. Roux, S. McFarlane, and S. Venter, "Significant host jump of xanthomonas vasicola from sugarcane to a eucalyptus grandis clone in south africa," *Plant Pathology*, vol. 64, no. 3, pp. 576–581, 2015. [Cited on page 30.]

[88] C.-H. Huang, G. E. Vallad, H. Adkison, C. Summers, E. Margenthaler, C. Schneider, J. Hong, J. B. Jones, K. Ong, and D. J. Norman, "A novel xanthomonas sp. causes bacterial spot of rose (rosa spp.)," *Plant disease*, vol. 97, no. 10, pp. 1301–1307, 2013. [Cited on page 30.]

[89] S. Essakhi, S. Cesbron, M. Fischer-Le Saux, S. Bonneau, M. A. Jacques, and C. Manceau, "Phylogenetic and variable-number tandem-repeat analyses identify nonpathogenic Xanthomonas arboricola lineages lacking the canonical type III secretion system," *Applied and Environmental Microbiology*, vol. 81, no. 16, pp. 5395–5410, 2015. [Cited on pages 30 and 33.]

[90] V. Meline, W. Delage, C. Brin, C. Li-Marchetti, D. Sochard, M. Arlat, C. Rousseau, A. Darrasse, M. Briand, G. Lebreton *et al.*, "Role of the acquisition of a type 3 secretion system in the emergence of novel pathogenic strains of Xanthomonas," *Molecular Plant Pathology*, vol. 20, no. 1, pp. 33–50, 2019.

[91] A. Hajri, J. F. Pothier, M. Fischer-Le Saux, S. Bonneau, S. Poussier, T. Boureau, B. Duffy, and C. Manceau, "Type three effector gene distribution and sequence analysis provide new insights into the pathogenicity of plant-pathogenic xanthomonas arboricola," *Applied and Environmental Microbiology*, vol. 78, no. 2, pp. 371–384, 2012. [Online]. Available: https://aem.asm.org/content/78/2/371 [Cited on pages 30 and 33.]

[92] M. Hunt, N. De Silva, T. D. Otto, J. Parkhill, J. A. Keane, and S. R. Harris, "Circlator: automated circularization of genome assemblies using long sequencing reads," *Genome biology*, vol. 16, no. 1, pp. 1–10, 2015. [Cited on page 32.]

[93] T. Tatusova, M. DiCuccio, A. Badretdin, V. Chetvernin, E. P. Nawrocki, L. Zaslavsky, A. Lomsadze, K. D. Pruitt, M. Borodovsky, and J. Ostell, "Ncbi prokaryotic genome

annotation pipeline," *Nucleic acids research*, vol. 44, no. 14, pp. 6614–6624, 2016. [Cited on page 32.]

[94] H. Li and R. Durbin, "Fast and accurate short read alignment with BurrowsWheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 05 2009. [Online]. Available: https://doi.org/10.1093/bioinformatics/btp324 [Cited on page 32.]

[95] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The sequence alignment/map format and samtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009. [Cited on page 32.]

[96] H. Li, "A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data," *Bioinformatics*, vol. 27, no. 21, pp. 2987–2993, 2011. [Cited on page 32.]

[97] R. R. Wick, M. B. Schultz, J. Zobel, and K. E. Holt, "Bandage: interactive visualization of de novo genome assemblies," *Bioinformatics*, vol. 31, no. 20, pp. 3350–3352, 2015. [Cited on page 32.]

[98] E. M. Gertz, Y.-K. Yu, R. Agarwala, A. A. Schäffer, and S. F. Altschul, "Composition-based statistics and translated nucleotide searches: improving the tblastn module of blast," *BMC biology*, vol. 4, no. 1, pp. 1–14, 2006. [Cited on page 32.]

[99] S.-H. Yoon, S.-m. Ha, J. Lim, S. Kwon, and J. Chun, "A large-scale evaluation of algorithms to calculate average nucleotide identity," *Antonie van Leeuwenhoek*, vol. 110, no. 10, pp. 1281–1286, 2017. [Cited on page 32.]

[100] I. Lee, Y. O. Kim, S.-C. Park, and J. Chun, "Orthoani: an improved algorithm and software for calculating average nucleotide identity," *International journal of systematic and evolutionary microbiology*, vol. 66, no. 2, pp. 1100–1103, 2016. [Cited on page 32.]

[101] Konstantinidis, "Genomic insights that advance the species definition for prokaryotes," *Tetrahedron*, vol. 41, no. 19, pp. 4147–4156, 1985. [Cited on page 32.]

[102] A. R. Quinlan and I. M. Hall, "Bedtools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–842, 2010. [Cited on page 33.]

[103] P. Rice, I. Longden, and A. Bleasby, "Emboss: the european molecular biology open software suite," 2000. [Cited on page 33.]

[104] C. Notredame, D. G. Higgins, and J. Heringa, "T-coffee: A novel method for fast and accurate multiple sequence alignment," *Journal of molecular biology*, vol. 302, no. 1, pp. 205–217, 2000. [Cited on page 33.]

[105] A. Zielezinski, S. Vinga, J. Almeida, and W. M. Karlowski, "Alignment-free sequence comparison: benefits, applications, and tools," *Genome biology*, vol. 18, no. 1, p. 186, 2017. [Cited on page 33.]

[106] A. Stamatakis, "Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models," *Bioinformatics*, vol. 22, no. 21, pp. 2688–2690, 2006. [Cited on page 33.]

[107] J. Felsenstein, *PHYLIP (phylogeny inference package), version 3.5 c*. Joseph Felsenstein., 1993. [Cited on page 34.]