UNIVERSITÀ DEGLI STUDI DI UDINE
hic sunt futura

www.uniud.it

Ph.D. Course in
Agricultural Science and Biotechnology

In convenzione con Fondazione Edmund Mach

XXX Cycle

Title of the thesis

"Next-Generation-Sequencing Genomic and Metagenomic Analysis of Phytopathogenic Prokaryotes"

Ph.D. Candidate
Cesare POLANO

Supervisor
Prof. Giuseppe FIRRAO

Year 2018

Table of Contents

Abstract

Phytopatology, as a discipline that deals with the complexity of living communities, needs methods to screen what could be considered ‘useful’ data from ‘background noise’. Until the Nineties, this was achieved by simplifications that were deemed adequate enough: from Koch’s postulates that require microorganisms to be culturable, to DNA barcoding that assumed genetic markers to be universal and precise enough to distinguish minute differences, to the disease triangle model that mostly downplayed the role of the micro-community context the pathogens find themselves in.

With the introduction of whole genome sequencing (WGS) technologies in the last thirty years, we started to realise that those simplifications, while not wrong, constituted sufficient but not necessary conditions: some pathogens (*e.g.* phytoplasmas) are remarkably difficult to cultivate *in vitro*, DNA structural variations can produce diverse strains without changing markers, and the micro-community can significantly impact on a pathogen’s ability to spread.

This work shows, from different perspectives all tied by the use of WGS data and analysis, how a deeper understanding of these complex dynamics can prompt new practical concepts to manage economically impactful plant diseases:

- The characterisation of *Pseudomonas* sp. strain Pf4 shows how the most fit strains, both from pathogens and biocontrol agents, derive their qualities from sizable sets of ‘secondary’ – but in fact crucial, as we are now aware – metabolites (SM) gene clusters;
- The comparison of the biocontrol activity of Pf4 and Pf11 shows that while a wide set of SM clusters is important, the inclusion of such set doesn’t necessarily translate into a ‘stronger’ control activity, but points to a better adaptability to changing environmental conditions;
- The use of third-generation WGS, which produces longer (~10,000 nts) reads, was essential to characterise the CRAFRU 12.29 and 14.08 strains (one produ-

cing hypersensitive response (HR) on leaves, the other not), as their difference lies in a transposon-mediated structural variation that would not have been possible to identify with older sequencing methods;

- Developing the *Phytoassembly* pipeline contributed to a novel method of obtaining phytoplasma (and other non-culturable organisms) genome, which circumvent the laborious *in vitro* protocols employed so far to obtain similar results;
- The *Phytoassembly* pipeline showed its potentiality by not only isolating a Chicory Phyllody (ChP) phytoplasma, but allowing to detect the presence of a companion spiroplasma, later shown to frequently occur together in mixed infections of chicory;
- *Phytoassembly* also helped characterising a Cassava Frogskin Disease (CFSD) phytoplasma, which showed some differences from other representatives in the group;
- The spatialisation of the genomic samples from the kiwifruit endophyte populations allows to correlate their spatial and temporal variation to the severity of the symptoms displayed by the plants and the time of *Pseudomonas syringae* pt. *actinidiae* (Psa) infection.

On the whole, the research projects presented in this work give insights into the greater complexity of microbial genome structure and variation, the dynamics between pathogens and the wider microbial community, the necessity for research methodologies based on more complex data, and the essential role that WGS technologies plays and will play in plant protection research and development.

Research activities

List of publications

- Polano, C., Ermacora, P., Martini, M., Musetti, R., Loi, N., & Firrao, G. (2016, September). A revised and effective pipeline based on relative coverage for the genome reconstruction of phytoplasmas and other fastidious prokaryotes. Poster presented at the XXII national conference of the Italian Phytopathological Society, Rome.
- Martini, M., Moruzzi, S., Polano, C., Musetti, R., Loi, N., Firrao, & G., Ermacora, P. (2016, September). Genome drafts of fluorescent pseudomonas biocontrol strains isolated from hydroponic cultures. Poster presented at the XXII national conference of the Italian Phytopathological Society, Rome.
- Polano, C., & Firrao, G. (2017). Next-Generation-Sequencing Metagenomic Analysis of Phytopathogenic Prokaryotes. Poster presented at the Ph.D. Expo 2017, University of Udine.
- Moruzzi, S., Firrao, G., Polano, C., Borselli, S., Loschi, A., Ermacora, P., Loi, N., & Martini, M. (2017). Genomic-assisted characterisation of *Pseudomonas* sp. strain Pf4, a potential biocontrol agent in hydroponics. *Biocontrol Science and Technology,* **27(8)**, 969–991. doi:10.1080/09583157.2017.1368454
- Polano, P., & Firrao, G. (n.d.). An Effective Pipeline Based on Relative Coverage for the Genome Assembly of Phytoplasmas and Other Fastidious Prokaryotes. *Current Genomics* [*submitted*]
- Firrao, G., Torelli, E., Polano, C., Ferrante, P., Ferrini, F., Martini, M., Scortichini, M., & Ermacora, P. (n.d.). Genomic structural variations during clonal expansion of *Pseudomonas syringae* pv. *actinidiae* biovar 3 in Europe. *Molecular Plant Pathology.* [*submitted*]
- Polano, C., Martini, M., Savian, F., Moruzzi, S., Ermacora, P., & Firrao, G. (n.d.). Genome sequence and antifungal activity in two niche-sharing *Pseudomonas protegens* strains isolated from hydroponics. [*to be submitted in 2017*]
- Polano, C., Moruzzi, S., Ermacora, P., Ferrini, F., Martini, M., & Firrao, G. (n.d.). Metagenomics highlighted mixed infection of spiroplasma and phytoplasma in chicory. [*to be submitted in 2017*]

- Polano, C., & Firrao, G. (n.d.). Multivariate analysis of endophytes diversity in kiwifruit in relation with *Pseudomonas syringae* pv. *actinidiae. [to be submitted in 2017]*

- Polano, C., Martini, M., & Firrao, G. (n.d.) “Obtaining high quality phytoplasma genome drafts with Illumina and Phytoassembly.” In: Phytoplasma – Methods and Protocols. *Springer Nature,* London [*to be submitted in 2018*]

- Neves de Souza, A., Polano, C., Martini, M., Firrao, G., & Carvalho, C. (n.d.). Molecular characterization of organisms associated with cassava plants showing cassava frogskin disease. [*to be submitted in 2018*]

Seminars and conferences attended

- Monalisa’s Quidproquo IV (Early) Midsummer Festival, Udine, June 4th–5th, 2015

- Galileo Festival 2016, Padua, May 5th–7th, 2016

- Workshop “Bioinformatica per tutti e per tutto: genomica, epigenomica, trascrittomica”, Italian Society of Agricultural Genetics, Udine, June 28th–July 7th, 2016

- XXII SIPaV Conference, Italian Phytopathological Society, Rome, September 19th–21st, 2016

- Summer School 2017, University of Udine, Paluzza (UD), September 6th–9st, 2017

Collaborations with other institutions

- Internship with the developing team of the UNITE/PlutoF database, University of Tartu, Estonia, August 13th–October 20th, 2017

1 Introduction

1.1 Whole Genome Sequencing

Genome sequencing, the process of determining the complete DNA sequence of an organism's chromosomal, mitochondrial and (for plants) chloroplast DNA, has come to fulfil an essential role in biological research, as a detailed map of an organism's genetic assets allows a greater understanding of the mechanisms of its adaptation to an environment (or in case of pathogens, its host), its potential weaknesses and its phylogenetic position relative to its closest species. Strengths and weaknesses can be suggested by the sets of proteins available, such as enzymes, regulators and transporters, while transposable elements can potentially deactivate genes by inserting in the middle of their sequences.

Initially, sequencing was a laborious, manual procedure; one of the earliest sequencings was done for the bacteriophage MS2 coat protein (Jou *et al.*, 1972), about 3500 nts long, and was carried out using 2D gel electrophoresis, an adaptation of a technique developed by Sanger and colleagues (Adams *et al.*, 1969). The Sanger method, formalised in 1975 (*see below*) quickly became the golden standard for sequencing methods for the reliability and (at the time) relative speed of its output. During the Nineties however its limits became too tight, and with the introduction of new techniques sequencing became faster, more economical, and allowed for longer uninterrupted sequences to be produced (Shendure and Ji, 2008). Sequencing technologies can be roughly divided in three 'generations', each improving on the size and speed of the output.

1.1.1 First Generation: the Sanger method

The 'first generation', exemplified by the Sanger and the Maxam–Gilbert methods, have in common an electrophoretic run as their last step, and as such are limited in the length of the sequence that can be determined, but are on the other hand quite accurate. The Sanger method was developed by Frederick Sanger in 1975 (Sanger and Coulson, 1975), is still used *e.g.* for sequencing individual fragments generated through polymerase chain reaction (PCR) and is still regarded as the benchmark against which other methods are calibrated and compared.

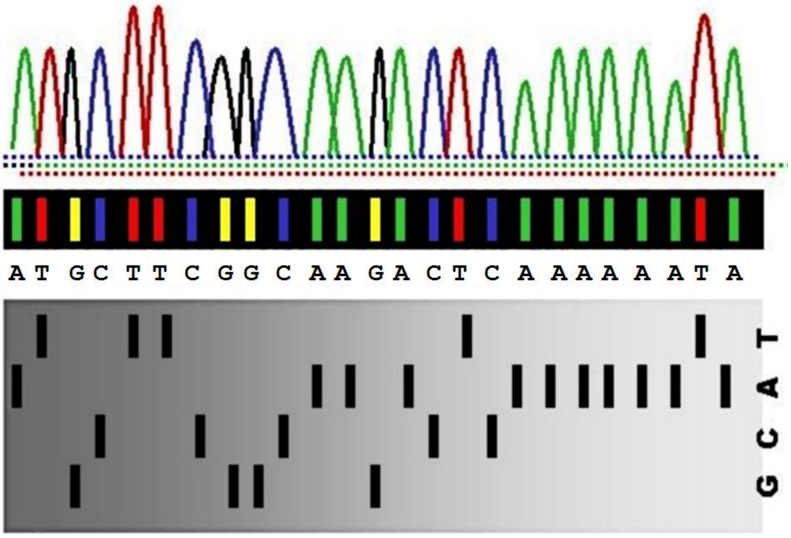


Figure 1.1 – An example of Sanger sequencing. Each band correspond to a fragment of *n* bases, and each lane is marked for one base. Optical readings of radio-labelled bands can be translated in intensity peaks for automatic transcription. [Source: https://dodona.ugent.be/en/exercises/144497797/]

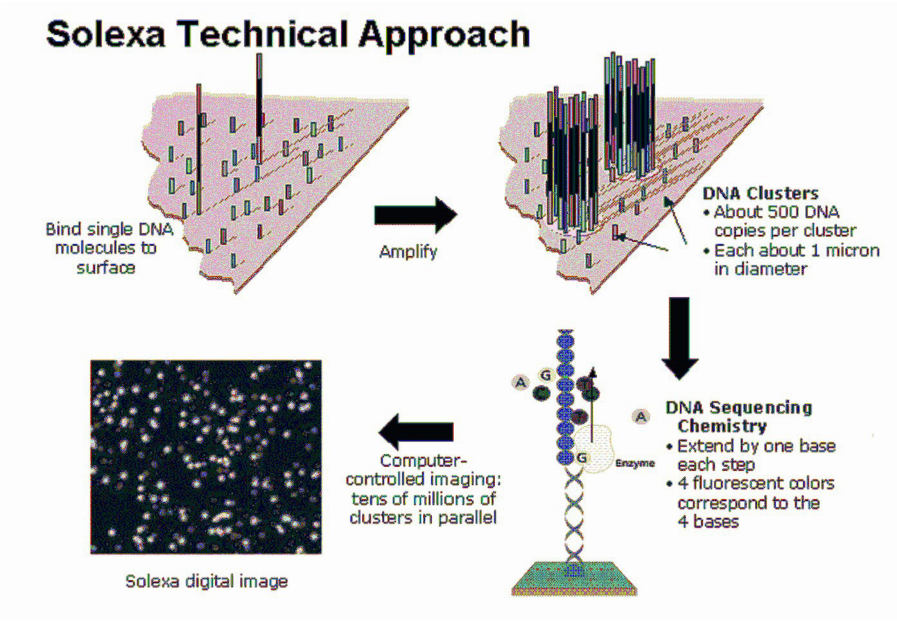

Solexa Technical Approach
Bind single DNA molecules to surface
Amplify
DNA Clusters
• About 500 DNA copies per cluster
• Each about 1 micron in diameter
DNA Sequencing Chemistry
• Extend by one base each step
• 4 fluorescent colors correspond to the 4 bases
A
G
A
G
Enzyme
Computer-controlled imaging: tens of millions of clusters in parallel
Solexa digital image

Figure 1.2 – Illumina (Solexa) sequencing method. 100–500 nt long fragments are binded to the surface of a flow cell, then amplified to obtain optically-detectable clusters; the complementary strand is then synthesised one nucleotide at a time, producing a flash which is captured by an optical sensor. [Source: https://www.sec.gov/Archives/edgar/data/913275/000095013407000492/f26433a1e425.htm#010]

This method can sequence up to around 900 nts and takes advantage of the property of modified di-deoxynucleotidetriphosphates (ddNTPs) of interrupting the extension of DNA, due to the lack of a 3′-OH group, preventing the formation of a phosphodiester bond with the successive nucleotide (Sanger *et al.*, 1977). In a PCR amplification, this ideally produces as many groups of fragments as the number of the specific ddNTP base. In sequencing machines the ddNTPs can be fluorescently labelled for automatic detection.

A typical Sanger reaction employs four parallel PCR reactions, each containing all of the standard dNTPs and one type of ddNTP. The resulting fragments are denatured and separated with gel electrophoresis according to size. Aligning the four gel lanes, the relative positions of the bands correspond to the DNA sequence (Figure 1.1). In automated Sanger sequencers, up to 380 reactions can be run in parallel and optically read, producing intensity curves whose peaks translate to individual bases (Hutchison, 2007).

The main limitations of the Sanger method are that the amplification quality for the first 15–40 bases is rather poor due to primer binding, and the quality of sequencing traces deteriorates again after 700–900 bases, as beyond that length it becomes difficult to separate fragments that differ in length by one nucleotide.

1.1.2 Second Generation: the Illumina method

As mentioned, the main problem of early sequencing methods like Sanger is that it can only sequence about 1000 bases at a time, while a small bacterial genome comprises millions of bases. A second problem is that at least initially it was a manual task and not a trivial one to complete; while the latest Sanger sequencers raised the output to about 380,000 bases per run, they still required significant economical resources that only major research centers could afford. The need for faster, more accessible acquisition of larger portions of genome led to the development of automated Whole Genome Sequencing (WGS) technologies during the Nineties.

This ‘second generation’ sequencing achieved full-genome length by mean of various strategies; of those still in use, the most common method involves splitting the genome into numerous reads, to be later assembled by software. Many methods have been developed between the ’90s and the early 2000s, to name a few: Massively Parallel Signature Sequencing, 454 Pyrosequencing (phased out in 2016), Illumina (formerly Solexa) Sequencing and Ion Torrent Semiconductor Sequencing. One of the most common methods – and the one used for most of the papers included in this thesis – is that by Illumina, which uses DNA polymerase fluorescent substrates with reversible 3′-terminators (Canard and Sarfati, 1994). A typical procedure includes (Figure 1.2):

1. The DNA is randomly fragmented and adapters are ligated to the 5′ and 3′ extremities ("tagmentation"). The ligated fragments are PCR-amplified and gel-purified.
2. The amplified fragments are bounded to the surface of an acrylamide-coated flow cell, where each 'lane' is a cluster of fragment duplicates (usually around 1000 per lane) generated by bridge amplification. The reverse sequences are then removed.
3. The lanes are complemented with fluorescent-tagged nucleotides. The 3′-terminators prevent the polymerase from joining more than one base at a time, allowing to image each lane in one shot. The fluorescent labels are then removed and the process is repeated until the end of the sequence.
4. In paired-end sequencing, the lanes undergo a second bridge amplification to invert the sequences, then the bases are read one by one a second time.

The consensus sequences from each lane are individual reads. It is not possible to determine *a priori* where each read belong into the whole sequence; shotgun sequencing relies on the likelihood that, with enough coverage, any point in a genome is represented by at least one read. In practice, this is often not the case: the major downside of splitting the genome into reads is that the subsequent assembling (in some cases, based on heuristic techniques) is highly reliant on their quality, can be misled by repetitive sequences and some portions of the genome might not be covered altogether.

1.1.3 Third Generation: the PacBio method

The current ‘third generation’ or ‘long-read’ sequencing (since around 2008) methods attempt to circumvent the assembling issues by transcribing the sequences on a single-molecule level, therefore obtaining much longer reads, and can potentially allow for direct detection of epigenetic markers (Simpson *et al.*, 2017). The most known methods are those by Pacific Biosciences (PacBio) and Oxford Nanopore Technology.

The PacBio method, also known as Single Molecule Real-Time (SMRT) sequencing, utilised in one of the papers of this thesis, is based on zero-mode waveguides (ZMWs), structures that can guide optical waves into picolitre wells, and phospholinked nucleotides (Levene, 2003; Eid *et al.*, 2009).

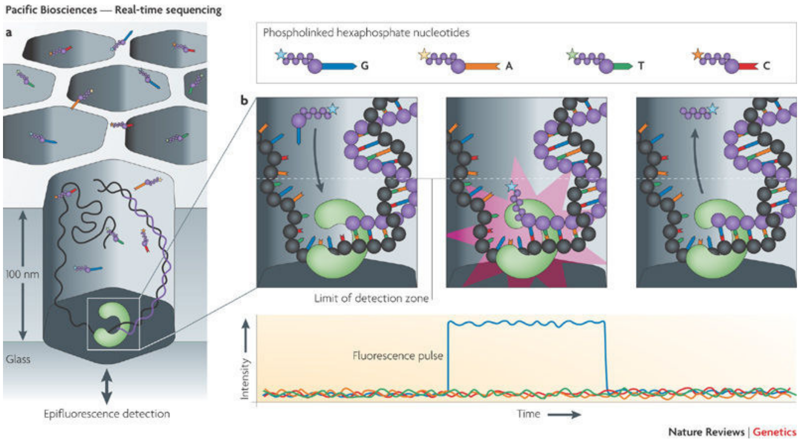

Pacific Biosciences — Real-time sequencing
a
b
Phospholinked hexaphosphate nucleotides
G
A
T
C
100 nm
Glass
Epifluorescence detection
Limit of detection zone
Fluorescence pulse
Intensity
Time
Nature Reviews | Genetics

Figure 1.3 – Pacific Biosciences SMRT sequencing. **a**: the zero-mode waveguide (ZMW) reduces the observation volume and the number of stray fluorescently labelled molecules that enter the detection layer for a given period; **b**: The residence time of phospholinked nucleotides is usually on the millisecond scale. The released, dye-labelled pentaphosphate by-product quickly diffuses away, dropping the fluorescence signal to background levels. The template is then translocated before binding and incorporating the next incoming phospholinked nucleotide. [Source: https://www.nature.com/nrg/journal/v11/n1/fig_tab/nrg2626_F4.html]

1. A SMRT cell is comprised of tens of thousands of ZMWs wells, about 50×100 nm in size; a DNA template-polymerase complex sits at the bottom, which is illuminated from below (Figure 1.3).

2. Phospholinked nucleotides, labeled with coloured fluorophores, are released on the SMRT cell.
3. Ligation of the nucleotides releases the fluorophore and emits a light pulse, with little background noise because of the small size of the well.

Third generation methods attempt not only to produce longer reads, but also to reduce the background noise from the fluorophore flashes occurring in nearby wells or clusters. The drawback of many of these methods is that sequencing errors are often unrecoverable, so they can be less suitable for *e.g. de novo* assembling (see below); but in applications like metagenomics or large structural variant calling, which are more tolerant to errors, these newer technologies can often outperform their predecessors.

The ideal sequencing tool would of course be able to sequence the whole genome from start to end, without interruptions and with negligible error rates. While the current technologies are still far from that ideal, in the last few years many strategies have been proposed that come reasonably close to it, and increasingly more sophisticated post-sequencing tools can help 'fill the gap' with current sequencers. One notable class of such tools is that of *de novo* assemblers.

1.2 *De novo* sequence assembly

Whole genome ('shotgun') sequencing produces fragments of various length (100–500 nts with second generation, 10,000–60,000 with third generation), which need to be aligned and merged to reconstruct the original sequence by forming contigs (Figure 1.4) (Johnson *et al.*, 2012). This, of course, is not a trivial procedure, as it has to deal with repetitive sequences, reading errors, and fragments not belonging to the same organism. Also, while some early algorithms were devised to combine *e.g.* a few Sanger sequences, WGS requires assembling many millions of reads, which requires more sophisticated strategies to complete the assembly in a reasonable time frame.

Assembling can be divided in two main types: *mapping*, or aligning reads against an already existing, similar but not necessarily identical, reference sequence; and *de novo*, in which full-length sequences are generated without previous knowledge.

Mapping algorithms compare each read to a reference and are relatively straightforward, although they need to accommodate for the possible presence of insertions, deletions, transpositions and other sources of variability. A more in-depth discussion of sequence aligners is present in Chapter 1.4 “Comparative genomics”.

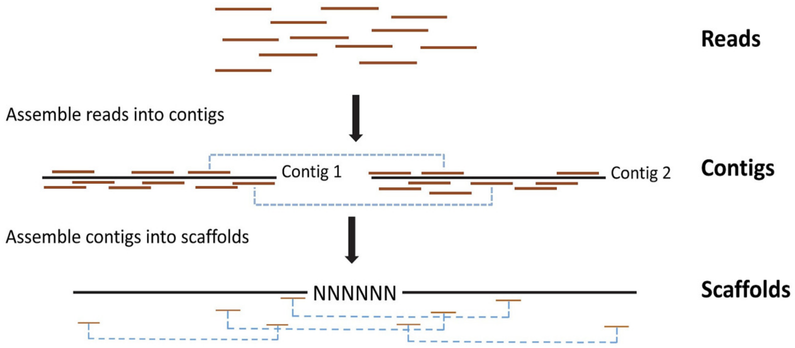


Figure 1.4 – *De novo* assembling of reads involves joining their overlapping extremities to form *contigs* (from “*contig*uous”); additional information can be used to *scaffold* the contigs into a single sequence. [Source: (Johnson *et al.*, 2012)]

De novo algorithms have an $O(n^2)$ complexity, as they need to compare every read with every other read. The speed of assembling depends on various contrasting factors, *e.g.* shorter reads align faster, but overlaps are less univocal (Henson *et al.*, 2012). Early assemblers employed ‘greedy’ algorithms: first they calculate pairwise distances, clustering reads with greatest overlap, then assembling these reads into contigs. These algorithms are optimised for local *optima* and are less suitable for larger sets (Bang-Jensen *et al.*, 2004). Once commonly used greedy assemblers were *SEQAID* (Peltola *et al.*, 1984) and *Phrap* (Machado *et al.*, 2011), part of the Phred-Phrap-Consed package that introduced the Phred quality score, later adopted for the FASTQ format (Cock *et al.*, 2009).

Later assemblers have been programmed with WGS in mind, adopting De Bruijn graph methods that search global optima: reads are broken into smaller fragments of specified size (*k*-mers), which are then used as nodes in a graph; nodes that overlap by some amount are then connected and sequences are constructed based on the graph (Myers, 1995). Commonly used assemblers of this type are *SPAdes* (Bankevich *et al.*, 2012) and the *A5 pipeline* (Tritt *et al.*, 2012), along with many others (Bradnam *et al.*, 2013). It might be interesting to note that while early assemblers, and other types of genomic software, were for the most commercial products, many of the more recent ones are being developed under open source licenses (Koboldt, 2015).

1.3 Genome annotation

Once the raw, full sequence has been obtained from the organism of interest, in-deep analysis of its content becomes possible. One of the first desirable steps is often annotating the coding regions and their functions (Figure 1.5), by identifying protein-coding portions, transposons, predicting genes and gene clusters, delimiting pathogenicity islands and secondary metabolite production-associated clusters, verifying repetitive sequences, and separating plasmid genomes from the main sequence (Stein, 2001).

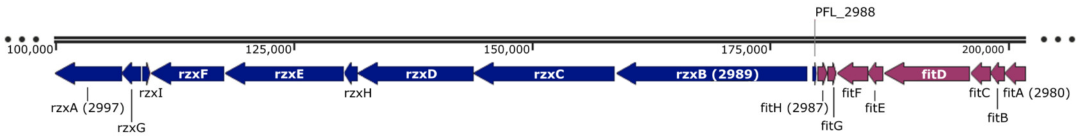
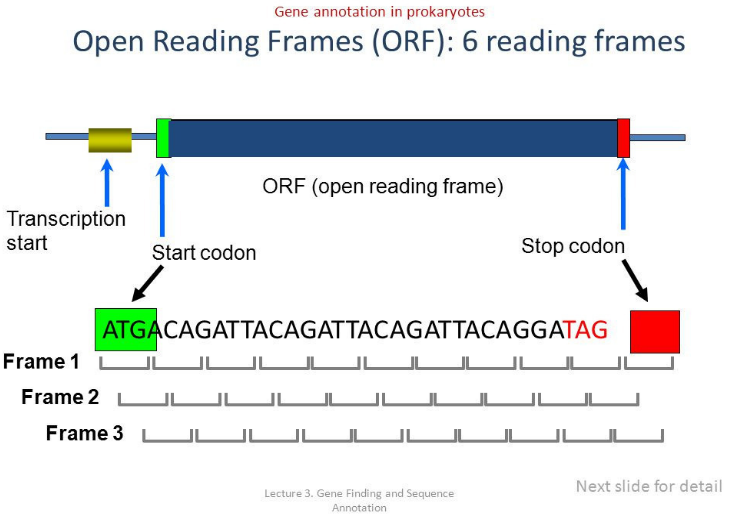


Figure 1.5 – An example of the representation of a cluster of genes; the arrows indicate the direction of transcription. [Source: (Moruzzi *et al.*, 2017)]

One of the first major problems of genome annotation is identifying the correct Open Reading Frames (ORFs) of the sequence, the translation from triplets of nucleotides to aminoacids between and including a start and a stop codon (Figure 1.6). As DNA has two antiparallel strands, there are 6 possible ORFs for any given sequence. There is rarely, if ever, a single ORF for a whole sequence: interruptions such as mutations and Single-Nucleotide Polymorphisms (SNPs) can shift the reading frame, and the reading direction can switch (Sharma *et al.*, 2011).

Initially, for short sequences, annotation was a lengthy procedure done entirely manually by experienced annotators, often using search tools such as *BLAST* (NCBI Resource Coordinators, 2013) to find homologous genes in specific or multipurpose databases. Manual annotation of whole genomes is of course unfeasible, so pipelines have been developed to automatise the process; manual annotation is however still necessary, as annotators' output can be unreliable, depends on databases that can be incomplete, and sequences might have different attributions depending on the species analysed (Koonin and Galperin, 2003). Some of the most common annotation softwares are the NCBI Prokaryotic Genome Annotation Pipeline [https://www.ncbi.nlm.nih.gov/genome/annotation_prok/], the RAST server (Aziz *et al.*, 2008) and the MG-RAST server (Glass and Meyer, 2011).

Annotation can be *structural*, identifying the genomic components like ORFs, coding regions and gene structures, or *functional*, attributing roles at the genomic components, *e.g.* regulative or expressive; often both types are done sequentially. The main challenge however remains predicting and attributing functions to proteins, tasks that still require long computational times even with the aid of computers, although mass-spectrometry can help improve the speed and quality of annotation (Gupta *et al.*, 2007).



Gene annotation in prokaryotes

Open Reading Frames (ORF): 6 reading frames

Lecture 3. Gene Finding and Sequence Annotation

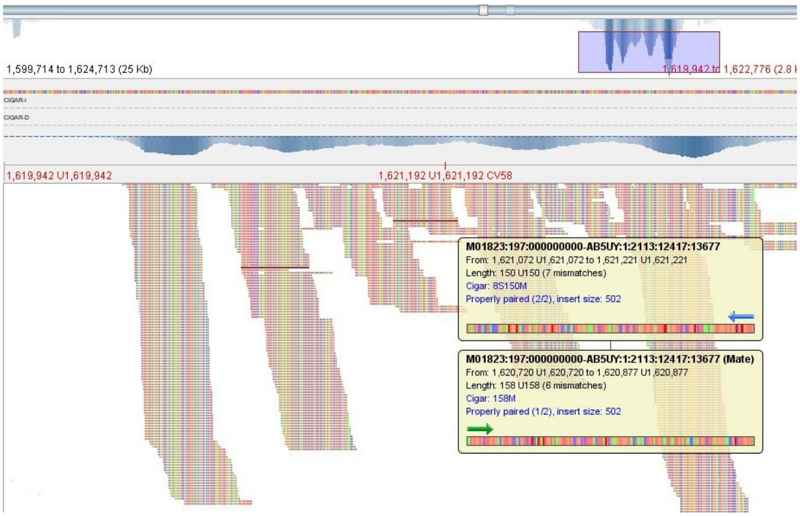
Next slide for detail

Figure 1.6 – The three possible open reading frames (ORFs) of a single DNA strand; with the antiparallel strand, the possible ORFs are up to six. The start codon is most commonly ATG, while the stop codon is usually TAA, TAG or TGA. ORFs enclose both exons and introns (see the next Figure) [Source: http://slideplayer.com/slide/5750865/]

1.4 Comparative genomics

Once a genome has been sequenced, and preferably annotated, it is often interesting to compare it to pre-existent sequences of close relatives, to *e.g.* find functional or structural variances between strains or species. Features that are investigated include genes, clusters, SNPs and introns. The comparison can be simply between two sequences (*pairwise*, Figure 1.7) or across many (Figure 1.8).

The simplest case is that of aligning short sequences to a longer reference, or comparing very short sequences; complexity increases with longer sequences, multiple alignments and with greater divergence between sequences. Pairwise alignments can be *global*, which assumes that the sequences have similar length, or *local*, where a smaller query is aligned more precisely over a portion of the longer reference, although the exact location of the alignment can be ambiguous (Polyanovsky *et al.*, 2011). A smaller query aligned globally over the reference would result in wide gaps inserted in the query, likely making the alignment nonsensical.


1,599,714 to 1,624,713 (25 Kb)
1,619,942 to 1,622,776 (2.8 k
CIGAR-I
CIGAR-D
1,619,942 U1,619,942
1,621,192 U1,621,192 CV58

M01823:197:000000000-AB5UY:1:2113:12417:13677
From: 1,621,072 U1,621,072 to 1,621,221 U1,621,221
Length: 150 U150 (7 mismatches)
Cigar: 8S150M
Properly paired (2/2), insert size: 502

M01823:197:000000000-AB5UY:1:2113:12417:13677 (Mate)
From: 1,620,720 U1,620,720 to 1,620,877 U1,620,877
Length: 158 U158 (6 mismatches)
Cigar: 158M
Properly paired (1/2), insert size: 502

Figure 1.7 – Alignment of reads against a reference sequence, using the software *Tablet*.
[Source: http://2014.igem.org/Team:Imperial/Gluconacetobacter]

Some of the most commonly used mapping tools of this type are *Bowtie* (Langmead *et al.*, 2009), *BWA* (Li and Durbin, 2010) and *SOAPdenovo* (Luo *et al.*, 2012); for visualisation, two commonly used programs are *Tablet* (Milne *et al.*, 2013) and the *Integrative Genome Viewer* (Thorvaldsdottir *et al.*, 2013).

Extending the comparison to multiple genomes is a step in identifying evolutionary relationships between samples, by clustering sequences on the basis of their similiarity, a method that is also utilised to produce phylogenetic trees (Figure 1.9). Commonly used programs for multiple sequence alignment are *MUMmer* (Kurtz *et al.*, 2004), *Clustal* (Larkin *et al.*, 2007) and *Mauve* (Darling, 2004). An important step in mapping similiarities and differences in related genomes is the identification of orthologs, genes with equivalent functions shared between related species that can shift position due to changes such as insertions or transpositions (Kuzniar *et al.*, 2008); specific software is being developed to help identifying orthologs, such as the OMA browser (Altenhoff *et al.*, 2015).

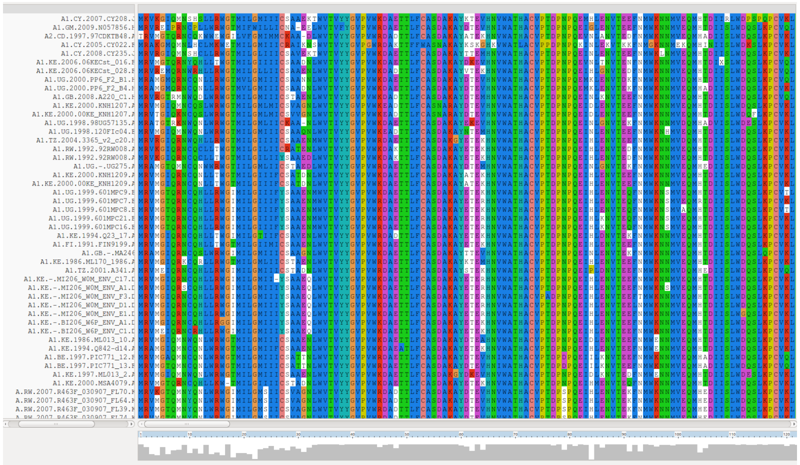


Figure 1.8 – Multiple sequence alignment, using the software *ClustalX*. [Source: http://bioinfopoint.com/index.php/code/3-multiple-sequence-alignment-with-bioperl-and-muscle]

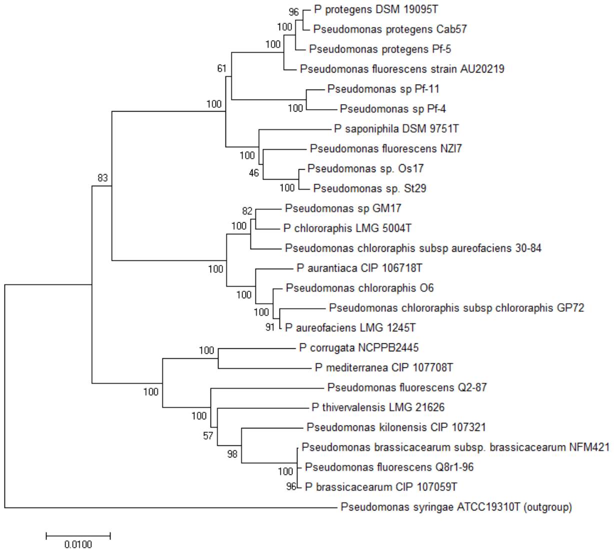

96
100
100
61
100
100
46
100
83
82
100
100
100
100
91
100
100
100
57
98
100
96
P protegens DSM 19095T
Pseudomonas protegens Cab57
Pseudomonas protegens Pf-5
Pseudomonas fluorescens strain AU20219
Pseudomonas sp Pf-11
Pseudomonas sp Pf-4
P saponiphila DSM 9751T
Pseudomonas fluorescens NZI7
Pseudomonas sp. Os17
Pseudomonas sp. St29
Pseudomonas sp GM17
P chlororaphis LMG 5004T
Pseudomonas chlororaphis subsp aureofaciens 30-84
P aurantiaca CIP 106718T
Pseudomonas chlororaphis O6
Pseudomonas chlororaphis subsp chlororaphis GP72
P aureofaciens LMG 1245T
P corrugata NCPPB2445
P mediterranea CIP 107708T
Pseudomonas fluorescens Q2-87
P thivervalensis LMG 21626
Pseudomonas kilonensis CIP 107321
Pseudomonas brassicacearum subsp. brassicacearum NFM421
Pseudomonas fluorescens Q8r1-96
P brassicacearum CIP 107059T
Pseudomonas syringae ATCC19310T (outgroup)
0.0100

Figure 1.9 – Example of a phylogenetic tree. The lengths of the ‘branches’ are proportional to the calculated distances between samples.

1.5 Whole genome sequencing and plant pathology

Previous chapters gave a panoramic of the various tools currently available to genomic-based plant pathology research. The main problem in phytopatology, in common with other disciplines like ecology, is that the subject of its research is very complex and dynamic (Mazzocchi, 2008). To borrow signal processing terminology, it therefore needs to improve the ‘signal-to-noise ratio’ by excluding what could be considered ‘negligible’ information. Until recent times, this implied simplifications that were deemed reasonable: Robert H. Koch postulated that to derive a causative relationship between a microbe and a disease, the microorganism needs to be isolated and grown in pure culture (Koch, 1876). The classic disease triangle model, attributed to R.B. Stevens, characterise a disease as a relation between the pathogen, the host and the environment (Francl, 2001), the last essentially being ‘everything else’. DNA barcoding is used on the assumption that genetic markers such as 16S rRNA or ITS are universally applicable and precise enough to distinguish minute differences between pathogen strains (Hajibabaei *et al.*, 2007).

The introduction of WGS tecnologies in the last thirty years provided the obvious advantage of allowing access to whole genomes within reasonable times even to small research labs, but they also led to the realisation that many of those simplifications, while not entirely wrong, constituted sufficient but not necessary conditions: defining a causative relationship between microbes and diseases required to include those pathogens that are remarkably difficult to cultivate *in vitro* (such as phytoplasmas or bacteria like *Xylella fastidiosa,* see chapter 4), who would otherwise be excluded. More refined disease models should include the microbial community context in which pathogens live, since the interactions within and between bacterial species can profoundly impact the outcome of their competition (Hibbing *et al.*, 2010). DNA structural variations, more complex and involving longer sections of the genome (DePristo *et al.*, 2011), can produce diverse strains without modifying the barcode markers, and as noted the microbial community can influence these variations.

The deeper understanding that the wealth of now available data required, in turn led to the implementation of more sophisticated analytical and synthetic tools (Green *et al.*, 2005) and methodologies outlined in the *Introduction*, but also suggests new, more sophisticated strategies to contain or contrast plant disease causative agents, especially those of more recent introduction, which are challenging to control with traditional methods and occasionally (as in the case of the kiwifruit canker agent, *Pseudomonas syringae* pv. *actinidiae*) can produce the opposite effect, *e.g.* unexpectedly selecting more resistant strains.

1.6 Aims and objective of the thesis

Objective of this thesis is to illustrate how rigorous bioinformatic analyses, backed by cutting-edge computing techniques, are essential to understand the data provided by Whole Genome Sequencing, and how they can help answer new and more complex questions.

In the following chapters relevant applications of these techniques will be presented, with an increasingly wide perspective: from the use of genomics to understand bacterial interactions with the environment, to a metagenomic approach for the characterisation of fastidious prokaryotes, to the metagenomic characterisation of whole communities; at each level, be it single bacterial genomes, types of microorganisms or communities in their entirety, WGS provided a better insights and suggested new strategies. An introduction to each topic will be given in each chapter.

1.7 Bibliography

Adams, J. M. *et al.* (1969) 'Nucleotide Sequence from the Coat Protein Cistron of R17 Bacteriophage RNA', *Nature*, 223(1009).

Altenhoff, A. M. *et al.* (2015) 'The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements', *Nucleic Acids Research*, 43(D1), pp. D240–D249. doi: 10.1093/nar/gku1158.

Aziz, R. K. *et al.* (2008) 'The RAST Server: Rapid Annotations using Subsystems Technology', *BMC Genomics*, 9(1), p. 75. doi: 10.1186/1471-2164-9-75.

Bang-Jensen, J., Gutin, G. and Yeo, A. (2004) 'When the greedy algorithm fails', *Discrete Optimization*, 1(2), pp. 121–127. doi: 10.1016/j.disopt.2004.03.007.

Bankevich, A. *et al.* (2012) 'SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing', *Journal of Computational Biology*, 19(5), pp. 455–477. doi: 10.1089/cmb.2012.0021.

Bradnam, K. R. *et al.* (2013) 'Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species', *GigaScience*, 2(1), p. 10. doi: 10.1186/2047-217X-2-10.

Canard, B. and Sarfati, R. S. (1994) 'DNA polymerase fluorescent substrates with reversible 3′-tags', *Gene*, 148(1), pp. 1–6. doi: 10.1016/0378-1119(94)90226-7.

Cock, P. J. A. *et al.* (2009) 'The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants', *Nucleic Acids Research*, 38(6), pp. 1767–1771. doi: 10.1093/nar/gkp1137.

Darling, A. C. E. (2004) 'Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements', *Genome Research*, 14(7), pp. 1394–1403. doi: 10.1101/gr.2289704.

DePristo, M. A. *et al.* (2011) 'A framework for variation discovery and genotyping using next-generation DNA sequencing data', *Nature Genetics*, 43(5), pp. 491–498. doi: 10.1038/ng.806.

Eid, J. *et al.* (2009) 'Real-Time DNA Sequencing from Single Polymerase Molecules', *Science*, 323(5910), pp. 133–138. doi: 10.1126/science.1162986.

Francl, L. J. (2001) 'The Disease Triangle: A Plant Pathological Paradigm Revisited', *The Plant Health Instructor*. doi: 10.1094/PHI-T-2001-0517-01.

Glass, E. M. and Meyer, F. (2011) 'The Metagenomics RAST Server: A Public Resource for the Automatic Phylogenetic and Functional Analysis of Metagenomes', *Handbook of Mo-*

lecular Microbial Ecology I: Metagenomics and Complementary Approaches, 8, pp. 325–331. doi: 10.1002/9781118010518.ch37.

Green, J. L. *et al.* (2005) ‘Complexity in Ecology and Conservation: Mathematical, Statistical, and Computational Challenges’, *BioScience*, 55(6), pp. 501–510. doi: 10.1641/0006-3568(2005)055[0501:CIEACM]2.0.CO;2.

Gupta, N. *et al.* (2007) ‘Whole proteome analysis of post-translational modifications: Applications of mass-spectrometry for proteogenomic annotation’, *Genome Research*, 17(9), pp. 1362–1377. doi: 10.1101/gr.6427907.

Hajibabaei, M. *et al.* (2007) ‘DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics’, *Trends in Genetics*, 23(4), pp. 167–172. doi: 10.1016/j.tig.2007.02.001.

Henson, J., Tischler, G. and Ning, Z. (2012) ‘Next-generation sequencing and large genome assemblies’, *Pharmacogenomics*, 13(8), pp. 901–915. doi: 10.2217/pgs.12.72.

Hibbing, M. E. *et al.* (2010) ‘Bacterial competition: surviving and thriving in the microbial jungle.’, *Nature reviews. Microbiology*. NIH Public Access, 8(1), pp. 15–25. doi: 10.1038/nrmicro2259.

Hutchison, C. A. (2007) ‘DNA sequencing: bench to bedside and beyond’, *Nucleic Acids Research*, 35(18), pp. 6227–6237. doi: 10.1093/nar/gkm688.

Johnson, M. T. J. *et al.* (2012) ‘Evaluating Methods for Isolating Total RNA and Predicting the Success of Sequencing Phylogenetically Diverse Plant Transcriptomes’, *PLoS ONE*. Edited by C. Quince, 7(11), p. e50226. doi: 10.1371/journal.pone.0050226.

Jou, W. M. *et al.* (1972) ‘Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein’, *Nature*, 237(5350), pp. 82–88. doi: 10.1038/237082a0.

Koboldt, D. (2015) *The Open Source Software Debate in NGS Bioinformatics*. Available at: http://massgenomics.org/2015/11/open-source-ngs-bioinformatics.html.

Koch, R. (1876) ‘Untersuchungen ueber Bakterien V. Die Aetiologie der Milzbrand-Krankheit, begruendent auf die Entwicklungsgeschichte des Bacillus Anthracis’, *Beitrage zur Biologie der Pflanzen*, pp. 277–310. doi: http://edoc.rki.de/documents/rk/508-5-26/PDF/5-26.pdf.

Koonin, E. V. and Galperin, M. Y. (2003) ‘Genome Annotation and Analysis’, in *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics*. Boston: Kluwer Academic. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21089240.

Kurtz, S. *et al.* (2004) ‘Versatile and open software for comparing large genomes.’, *Genome biology*, 5(2), p. R12. doi: 10.1186/gb-2004-5-2-r12.

Kuzniar, A. *et al.* (2008) ‘The quest for orthologs: finding the corresponding gene across genomes’, *Trends in Genetics*, 24(11), pp. 539–551. doi: 10.1016/j.tig.2008.08.009.

Langmead, B. *et al.* (2009) ‘Ultrafast and memory-efficient alignment of short DNA sequences to the human genome’, *Genome Biology*, 10(3), p. R25. doi: 10.1186/gb-2009-10-3-r25.

Larkin, M. A. *et al.* (2007) ‘Clustal W and Clustal X version 2.0’, *Bioinformatics*, 23(21), pp. 2947–2948. doi: 10.1093/bioinformatics/btm404.

Levene, M. J. (2003) ‘Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations’, *Science*. Nature Publishing Group, 299(5607), pp. 682–686. doi: 10.1126/science.1079700.

Li, H. and Durbin, R. (2010) ‘Fast and accurate long-read alignment with Burrows–Wheeler transform’, *Bioinformatics*, 26(5), pp. 589–595. doi: 10.1093/bioinformatics/btp698.

Luo, R. *et al.* (2012) ‘SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler’, *GigaScience*, 1(1), p. 18. doi: 10.1186/2047-217X-1-18.

Machado, M. *et al.* (2011) ‘Phred-Phrap package to analyses tools: a pipeline to facilitate population genetics re-sequencing studies’, *Investigative Genetics*, 2(1), p. 3. doi: 10.1186/2041-2223-2-3.

Mazzocchi, F. (2008) ‘Complexity in biology. Exceeding the limits of reductionism and determinism using complexity theory’, *EMBO reports*, 9(1), pp. 10–14. doi: 10.1038/sj.embor.7401147.

Milne, I. *et al.* (2013) ‘Using Tablet for visual exploration of second-generation sequencing data’, *Briefings in Bioinformatics*, 14(2), pp. 193–202. doi: 10.1093/bib/bbs012.

Moruzzi, S. *et al.* (2017) ‘Genomic-assisted characterisation of Pseudomonas sp. strain Pf4, a potential biocontrol agent in hydroponics’, *Biocontrol Science and Technology*, 27(8), pp. 969–991. doi: 10.1080/09583157.2017.1368454.

Myers, E. W. (1995) ‘Toward Simplifying and Accurately Formulating Fragment Assembly’, *Journal of Computational Biology*, 2(2), pp. 275–290. doi: 10.1089/cmb.1995.2.275.

NCBI Resource Coordinators (2013) ‘Database resources of the National Center for Biotechnology Information’, *Nucleic Acids Research*, 41(D1), pp. D8–D20. doi: 10.1093/nar/gks1189.

Peltola, H., Söderlund, H. and Ukkonen, E. (1984) ‘SEQAID: a DNA sequence assembling program based on a mathematical model’, *Nucleic Acids Research*, 12(1Part1), pp. 307–321. doi: 10.1093/nar/12.1Part1.307.

Polyanovsky, V. O., Roytberg, M. A. and Tumanyan, V. G. (2011) ‘Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences’, *Algorithms for Molecular Biology*, 6(1), p. 25. doi: 10.1186/1748-7188-6-25.

Sanger, F. and Coulson, A. R. (1975) ‘A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase’, *Journal of Molecular Biology*, 94(3), pp. 441–448. doi: 10.1016/0022-2836(75)90213-2.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) ‘DNA sequencing with chain-terminating inhibitors.’, *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp. 5463–7. Available at: http://www.pubmedcentral.nih.gov/articlerender-.fcgi?artid=431765&tool=pmcentrez&rendertype=abstract (Accessed: 29 October 2012).

Sharma, V. *et al.* (2011) ‘A pilot study of bacterial genes with disrupted ORFs reveals a surprising profusion of protein sequence recoding mediated by ribosomal frameshifting and transcriptional realignment’, *Molecular Biology and Evolution*, 28(11), pp. 3195–3211. doi: 10.1093/molbev/msr155.

Shendure, J. and Ji, H. (2008) ‘Next-generation DNA sequencing’, *Nature Biotechnology*, 26(10), pp. 1135–1145. doi: 10.1038/nbt1486.

Simpson, J. T. *et al.* (2017) ‘Detecting DNA cytosine methylation using nanopore sequencing’, *Nature Methods*, 14(4), pp. 407–410. doi: 10.1038/nmeth.4184.

Stein, L. (2001) ‘Genome annotation: from sequence to biology’, *Nature Reviews Genetics*, 2(7), pp. 493–503. doi: 10.1038/35080529.

Thorvaldsdottir, H., Robinson, J. T. and Mesirov, J. P. (2013) ‘Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration’, *Briefings in Bioinformatics*, 14(2), pp. 178–192. doi: 10.1093/bib/bbs017.

Tritt, A. *et al.* (2012) ‘An Integrated Pipeline for de Novo Assembly of Microbial Genomes’, *PLoS ONE*. Edited by D. Zhu, 7(9), p. e42304. doi: 10.1371/journal.pone.0042304.

2 Genomics as a tool to understand bacterial interactions with the environment that are relevant for plant pathology

A typical example of a pathogen that can effectively be analysed using WGS tools, due to the number of studies published and the availability of genome sequences of most of its pathovars, is *Pseudomonas syringae*, a Gram-negative, rod-shaped, flagellated bacterium characterised and named by C.J.J. van Hall in 1904 (Prasanth *et al.*, 2015). *P. syringae* is a causal agent of diseases in a wide range of plant species (including *Arabidopsis thaliana*, tobacco and tomato, species often used as model plants in experimental research for their ease of cultivation) and comprises more than 50 pathovars (infraspecies taxon that are distinguished by pathogenic abilities, without clearly established phylogenetic relationships with the others) with variable specificity.

P. syringae is of interest because of its various mechanisms of pathogenicity. The bacterium is not able to create an opening in the plant by itself, but can use chemical signals to find its way to natural openings (Ichinose *et al.*, 2013). It can produce a biofilm to adhere to the host cell wall and protect itself from other bacteria, using a *quorum sensing* strategy, by which genes are expressed when enough pseudomonad cells are present to produce effective quantities of said molecules. It can facilitate frost injury by acting as ice nucleation, raising the temperature at which water inside plant tissues, normally in a supercooled state, frosts (Lindow *et al.*, 1978). *P. syringae* is also interesting because it has one of the best characterised effector repertoire.

In summary, the plant-pathogen relation can be outlined by a 4-step model: initially, plants recognise *pathogen* (or *microbe*)*-associated molecular patterns* (PAMPs/MAMPs) (Zhang *et al.*, 2007) – which include many different molecules, like lipopolysaccharides, flagellin, lipoteichoic acid and peptidoglycan – and react with *MAMP-triggered immunity* (MTI, Figure 2.1); in turn pathogens can develop strategies to hide their MAMPs, or suppress the host's ability to recognise them by delivering effectors through the type 3 secretion system (T3SS) encoded in the *hrp* gene cluster; plants in turn recognise T3SS effectors through resistance (R) genes and employ *effector-triggered immunity* (ETI) responses, which most often translates to a *hypersen-*

sitive response (HR), where the affected cells die in an attempt to immobilise and stop the spread of the pathogen. Bacteria in turn can develop strategies to overcome ETI responses, *e.g.* by losing or changing their T3SS effectors (Bent and Mackey, 2007; Newman *et al.*, 2013).

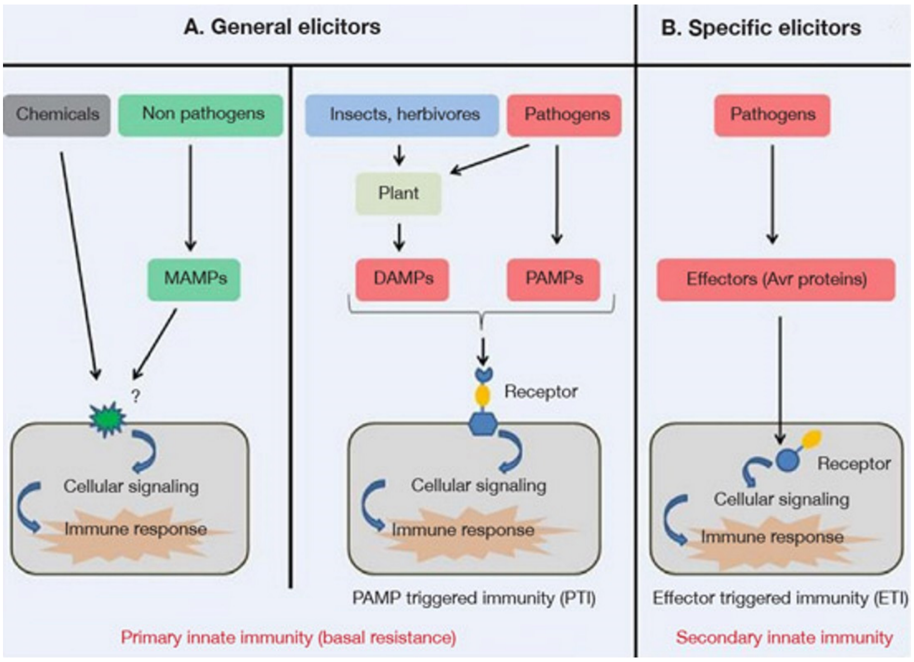


Figure 2.1 – M/PAMP-triggered immunity (MTI), constituting the primary form of plant resistance, *vs.* effector-triggered immunity (ETI), which is the plant's response to the pathogen's adaptation to MTI. [Source: (Henry *et al.*, 2012)]

This model, in which both the pathogen and the plant change in response to new strategies deployed by the other, has an interesting ramification, in that closely-related but less aggressive bacterial strains could potentially thrive thanks to this 'arms race', because of their closeness, instead of being cornered due to having a less optimised gene set, as it was generally assumed in the past. Genomic research can intervene by helping plants develop new T3SS effectors detectors (*e.g.* by breeding resistant cultivars), but also by altering the bacterial community in the proximity of the plant, to create a less hospitable environment to pathogens by promoting a stable presence of non-pathogenic antagonists or closely related strains. Considering the classic relationship between pathogen, host and environment, influencing the microbial community (besides looking for plant resistances) could help managing pathogens by leveraging on various tactics to slow its spread (Arneson, 2001): the initial inoculum could be reduced through dilution with closely-related non-pathogenic strains; the infection rate could be slowed by keeping the pathogen busy with competing with other species or multiple biocontrol agents (Guetsky *et al.*, 2002); the duration of the epidemic could be shortened by inducing the need for more time-consuming strategies to reach the plants.

The papers that follow illustrate a few applications of genomic analysis on *Pseudomonas* strains. In the first one, the biocontrol activity of *Pseudomonas* sp. strain Pf-4 isolated from hydroponics was assessed and compared to the set of secondary metabolites-producing gene clusters of the well-known soil biocontrol *P. protegens* strain Pf-5. Supplementary data is listed in the *Appendix*, Chapter 6.1.

In the second paper, the inhibition strength of *Pseudomonas* sp. strain Pf-4 and its close relative *Pseudomonas* sp. strain Pf-11 against fungal species were compared, to evaluate differences in their genomic and biological (in terms of fungal inihibition capabilities) features.

In the third paper, two very close strains of *P. s.* pv. *actinidiae* (PSA) produced a markedly different response when inoculated in tobacco leaves; using a Third Generation sequencing, the analysis showed that an insertion disrupted the functionality of the T3SS, suggesting that control strategies that do not promote recombination might have a lesser chance of favouring more virulent variant strains.

Bibilography

Arneson, P. A. (2001) ‘Plant Disease Epidemiology’, *The Plant Health Instructor*. doi: 10.1094/PHI-A-2001-0524-01.

Bent, A. F. and Mackey, D. (2007) ‘Elicitors, Effectors, and R Genes: The New Paradigm and a Lifetime Supply of Questions’, *Annual Review of Phytopathology*, 45(1), pp. 399–436. doi: 10.1146/annurev.phyto.45.062806.094427.

Guetsky, R. *et al.* (2002) ‘Improving Biological Control by Combining Biocontrol Agents Each with Several Mechanisms of Disease Suppression’, *Phytopathology*, 92(9), pp. 976–985. doi: 10.1094/PHYTO.2002.92.9.976.

Henry, G., Thonart, P. and Ongena, M. (2012) ‘PAMPs, MAMPs, DAMPs and others: an update on the diversity of plant immunity elicitors’, *Biotechnologie, Agronomie, Société et ...*, 16(2), p. 12. Available at: http://www.doaj.org/doaj?func=fulltext&aId=1044522.

Ichinose, Y., Taguchi, F. and Mukaihara, T. (2013) ‘Pathogenicity and virulence factors of Pseudomonas syringae’, *Journal of General Plant Pathology*, 79(5), pp. 285–296. doi: 10.1007/s10327-013-0452-8.

Lindow, S. E., Arny, D. C. and Upper, C. D. (1978) ‘Distribution of ice nucleation-active bacteria on plants in nature’, *Applied and Environmental Microbiology*, 36(6), pp. 831–838.

Newman, M.-A. *et al.* (2013) ‘MAMP (microbe-associated molecular pattern) triggered immunity in plants’, *Frontiers in Plant Science*. Frontiers Media SA, 4, p. 139. doi: 10.3389/fpls.2013.00139.

Prasanth, M. *et al.* (2015) ‘Pseudomonas Syringae : An Overview and its future as a “Rain Making Bacteria”’, *International Research Journal of Biological Sciences*, 4(2), pp. 70–77.

Zhang, J. *et al.* (2007) ‘A Pseudomonas syringae Effector Inactivates MAPKs to Suppress PAMP-Induced Immunity in Plants’, *Cell Host and Microbe*, 1(3), pp. 175–185. doi: 10.1016/j.chom.2007.03.006.

BIOCONTROL SCIENCE AND TECHNOLOGY, 2017
VOL. 27, NO. 8, 969–991
https://doi.org/10.1080/09583157.2017.1368454

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Genomic-assisted characterisation of *Pseudomonas* sp. strain Pf4, a potential biocontrol agent in hydroponics

Serena Moruzzi, Giuseppe Firrao, Cesare Polano, Stefano Borselli, Alberto Loschi, Paolo Ermacora, Nazia Loi and Marta Martini

Department of Agricultural, Food, Environmental and Animal Sciences (DI4A), University of Udine, Udine, Italy

ABSTRACT

In an attempt to select potential biocontrol agents against *Pythium* spp. and *Rhizoctonia* spp. root pathogens for use in soilless systems, 12 promising bacteria were selected for further investigations. Sequence analysis of the 16S rRNA gene revealed that three strains belonged to the genus *Enterobacter*, whereas nine strains belonged to the genus *Pseudomonas*. In *in vitro* assays, one strain of *Pseudomonas* sp., Pf4, closely related to *Pseudomonas protegens* (formerly *Pseudomonas fluorescens*), showed noteworthy antagonistic activity against two strains of *Pythium aphanidermatum* and two strains of *Rhizoctonia solani* AG 1-IB, with average inhibition of mycelial growth >80%. Strain Pf4 was used for *in vivo* treatments on lamb's lettuce against *R. solani* root rot in small-scale hydroponics. Pf4-treated and untreated plants were daily monitored for symptom development and after two weeks of infection, a significant protective effect of Pf4 against root rot was recorded. The survival and population density of Pf4 on roots were also checked, demonstrating a density above the threshold value of 10^5 CFU g^{-1} of root required for disease suppression. Known loci for the synthesis of antifungal metabolites, detected using PCR, and draft-genome sequencing of Pf4 demonstrated that *Pseudomonas* sp. Pf4 has the potential to produce an arsenal of secondary metabolites (*plt*, *phl*, *ofa* and *fit-rzx* gene clusters) very similar to that of the well-known biocontrol *P. protegens* strain Pf-5.

ARTICLE HISTORY
Received 18 January 2017
Accepted 14 August 2017

KEYWORDS
Biological control; *Rhizoctonia solani*; *Pythium* spp; population dynamic; secondary metabolites; draft-genome sequencing

1. Introduction

Soilless, hydroponic systems are well suited for the cultivation of many crops, including leafy vegetables. Their main feature is the possibility to control all environmental factors, that is, nutrient solution supply, temperature, pH, dissolved oxygen concentration, electrical conductivity, light radiation, that translates into higher production, energy conservation, better control of growth, independence from soil quality (Van Os, 1999).

Although soilless cultures have been reported as successful alternatives to the use of methyl bromide and other fumigants to avoid root diseases caused by soil-borne pathogen

CONTACT Marta Martini marta.martini@uniud.it Department of Agricultural, Food, Environmental and Animal Sciences (DI4A), University of Udine, Via delle Scienze 206, 33100 Udine, Italy

Supplemental data for this article can be accessed at https://doi.org/10.1080/09583157.2017.1368454

microorganisms (Van Os, 1999), root diseases still occur in these systems. Sometimes disease outbreaks are even greater than in soil (McPherson, Harriman, & Pattison, 1995), promoted by suitable environmental conditions, and rapid dispersal of root-colonising agents through the cultural system (Vallance et al., 2011). The most harmful pathogenic microorganisms in hydroponic cultures are those producing zoospores, that is, *Pythium* spp. and *Phytophthora* spp., particularly adapted to wet environment, but also *Fusarium* spp. and *Rhizoctonia solani* are of major concern (Paulitz & Bélanger, 2001; Schnitzler, 2004). In particular, *R. solani* was recently detected in Italy on many leafy vegetables (Colla, Gilardi, & Gullino, 2012), including lamb's lettuce [*Valerianella locusta* (L.) Laterr.] (Garibaldi, Gilardi, & Gullino, 2006).

Prevention of pathogen infections, particularly in closed hydroponic systems, has become a major challenge in recent years, particularly in the light of the increasing public concern regarding the use of chemical pesticides and subsequent legislative issues (e.g. Directive 2009/128/EC). Biological control is regarded as a potentially solid alternative to the use of chemical pesticides, and can be effective also in soilless systems especially against *Pythium* spp., *Phytophthora* spp. and *Fusarium* spp. (Postma, 2010; Vallance et al., 2011). Since studies on suppressiveness demonstrated the potential of indigenous microflora to inhibit root diseases in hydroponic cultures (McPherson, 1998), one of the main strategies is the addition of antagonistic microorganisms to increase the level of suppressiveness (Vallance et al., 2011).

Rhizobacteria are the most efficient microorganisms against soil-borne pathogens, which occur in the environment at the interface of root and soil (Handelsman & Stabb, 1996). In particular, fluorescent pseudomonads can persistently colonise the rhizosphere (Couillerot, Prigent-Combaret, Caballero-Mellado, & Moënne-Loccoz, 2009), compete with root pathogens for micronutrients (especially for iron and carbon) and root surface colonisation (Haas & Défago, 2005; Raaijmakers, Paulitz, Steinberg, Alabouvette, & Moënne-Loccoz, 2009), trigger induced systemic resistance (ISR) response in plants (Bakker, Pieterse, & Van Loon, 2007). A major component of biocontrol potential appears to be connected with secretion: fluorescent pseudomonads that are active biocontrol agents produce secondary metabolites that act as antimicrobial compounds, that is, 2,4-diacetylphloroglucinol (2,4-DAPG), phenazines, pyrrolnitrin, pyoluteorin, hydrogen cyanide (HCN) (Handelsman & Stabb, 1996; Raaijmakers, Vlami, & De Souza, 2002), but also siderophores such as pyoverdin, biosurfactants and extracellular lytic enzymes (Compant, Duffy, Nowak, Clément, & Barka, 2005).

Only a limited number of studies on biological control by rhizobacteria have been carried out in soilless systems and consequently a limited number of biocontrol agents have been isolated and characterised from soilless systems. Yet it is important to understand to what extent the growing system is a relevant component in determining the potential of the biological control agent. Are rhizobacteria with biological control potential isolated from hydroponics different from those isolated from soil? Are they relying on different mechanisms for the control of pathogens?

In this work we selected a biocontrol agent from endogenous source, the hydroponics, characterised it for both its biocontrol performances and its genomic features, with particular reference to secondary metabolites, and compared it with other known biological agents isolated from soil. Surprisingly, the strain was not dramatically different from other previously known pseudomonads biocontrol agents, indicating that the hydroponic conditions do not significantly change the mechanisms involved in biocontrol.

2. Materials and methods

2.1. Plant pathogen strains

Fungal and oomycete pathogens were obtained from culture collection and by isolation from diseased plants. Specifically, *Pythium aphanidermatum* strain CBS 118745 and strain CBS 116664 were obtained from the Centraal Bureau voor de Schimmelcultures (CBS) culture collection, and were grown on oatmeal agar (OA, 30 g l^{-1} oatmeal flakes boiled and filtered, 15 g l^{-1} bacteriological agar). Whereas fungal isolations were made in 2009 from diseased plants showing symptoms of root rot and wilting in a hydroponic farm in Friuli Venezia Giulia (FVG) region, north-eastern Italy. Sixty portions of lamb's lettuce or chicory roots and seedlings were washed in sterile distilled water, placed on water agar (WA, 20 g l^{-1} bacteriological agar) plates and incubated at 24°C for 48 h. The isolates were transferred to Petri dishes containing OA. Fungal isolates with the morphological characters of *R. solani* were consistently recovered and their identity confirmed by internal transcribed spacer (ITS) analysis. DNA extraction and PCR amplification of the ITS region using the universal primers ITS1/ITS4 (White, Bruns, Lee, & Taylor, 1990) and GoTaq Flexi DNA Polymerase (Promega, Madison, WI, USA) from 12 isolates of *R. solani* were carried out as previously described by Martini et al. (2009). PCR products were then digested with endonuclease *Tru*1I and visualised on a 2% agarose gel, stained with GelRed™ (Biotium Inc., Hayward, CA, USA). The subsequent restriction profiles were compared, and found to be identical to each other. Two strains of *R. solani*, TR15 and TP20, were selected for sequencing and analysis of the ITS region as described by Martini et al. (2009), and successively used in this work. ITS sequences (652 bp) of *R. solani* strains TR15 and TP20 were submitted to GenBank under accessions KM589032 and KM589033 respectively. BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) analysis allowed confirmation of their morphological identification as *R. solani* and their assignment to the anastomosis group AG 1-IB (Sharon, Kuninaga, Hyakumachi, & Sneh, 2006) with 100% similarity with the GenBank sequence AJ868450 of *R. solani* (*Thanatephorus cucumeris*) strain AG1 (CBS 522.96).

2.2. Isolation of potential bacterial biocontrol agents and preliminary screening

Bacteria strains were isolated from the rhizosphere of healthy hydroponic lamb's lettuce plants grown in the same hydroponic farm as before. Thirty root samples were collected from healthy plants, cut into 1–1.5 cm pieces, washed in sterile distilled water and transferred on WA; plates were incubated at 24°C for 48–72 h. Each colony was re-streaked three times, and grown in pure culture on nutrient agar medium (NA, 1 g l^{-1} beef extract, 2 g l^{-1} yeast extract, 5 g l^{-1} peptone, 5 g l^{-1} sodium chloride, 15 g l^{-1} bacteriological agar) at 24°C for 48 h.

Fifty-one bacterial strains were preliminarily tested by a dual culture method according to Gravel, Martinez, Antoun, and Tweddell (2005) with *P. aphanidermatum* strains CBS

118745 and CBS 116664, on potato dextrose agar medium (PDA, 38 g l^{-1}). Bacteria were inoculated at one side of a Petri dish and, after 48-h incubation, a mycelium plug was placed on the opposite site of the Petri dish, approximately 5 cm apart from the bacterial inoculation point. At the same time, positive controls of fungal pathogens were prepared by placing a mycelium plug in a Petri dish. After incubation for 7 days at room temperature (about 24°C), the presence/absence of an inhibition zone between the pathogen and each bacterium was recorded. Twelve bacterial strains that proved to inhibit the tested pathogens were selected for further investigations.

2.3. Bacteria identification

DNAs from the 12 selected bacterial strains were extracted according to the procedure reported on *Current protocols in Molecular Biology* (Wilson, 1997). PCR amplification of the 16S rRNA gene was performed with universal primers fD1/rP1 (Weisburg, Barns, Pelletier, & Lane, 1991). Amplifications were performed with the automated One Advanced thermocycler (EuroClone, Celbio, Milan, Italy) in 25 μl reactions containing 200 μM of each of the four dNTPs, 0.4 μM of each primer, 1.5 mM $MgCl_2$, 0.625 units of GoTaq Flexi DNA Polymerase (Promega, Madison, WI, USA) and 1 μl of diluted bacterial DNA (5 ng μl^{-1}). The PCR programme consisted of initial denaturation for 2 min at 94°C; 36 cycles of 1 min at 94°C, 1 min at 58°C, 2 min at 72°C; and a final extension for 8 min at 72°C.

PCR products were purified using the Wizard® SV Gel and PCR Clean-Up System Kit (Promega, Madison, WI, USA) and sent to Genechron laboratory, (ENEA Casaccia, Rome, Italy) for sequencing. The sequences were determined with forward and reverse primers and assembled with BioEdit (Hall, 1999). For bacteria identification, 16S rRNA gene sequences 1303–1409 bp long were compared with those present in GenBank using BLASTN analysis. The nucleotide sequences were deposited in GenBank.

2.4. In vitro antagonistic activity

The antagonistic activity of the 12 preliminarily selected bacterial strains against *P. aphanidermatum* strains CBS 118745 and CBS 116664 and *R. solani* strains TR15 and TP20 was further characterised as follows. Bacterial strains were inoculated on Petri dishes containing PDA supplemented with 3 g l^{-1} peptone and 2 g l^{-1} yeast extract, in four diametrically opposite sites, approximately 3 cm from the centre. After 48-h incubation at 24°C, plugs of mycelium (about 5 mm in diameter) were placed in the centre of the Petri dishes. At the same time, mycelium plugs were also inoculated on Petri dishes containing only growth medium as the control reference. The plates were further incubated for 9 days, and the mycelial growth was measured daily. The assays were repeated twice, and each combination of bacterial antagonist–plant pathogen was replicated at least three times. The average inhibitory effect of each strain against the two pathogens was estimated based on the per cent inhibition of radial growth, calculated using the following formula (Fokkema, 1976): % inhibition = $[(C - T)\ C^{-1}] \times 100$, where C is the radial growth of the pathogen without antagonist and T is the radial growth of the pathogen in the presence of the antagonist.

2.5. *In vivo activity of* Pseudomonas *sp. strain Pf4 against* R. solani

The bacterial strain that showed the best *in vitro* antagonistic activity (% inhibition of fungal growth ≥ 82, Figure 1), that is, *Pseudomonas* sp. strain Pf4, was chosen for *in vivo* application with the aim to evaluate its protective effect against *R. solani* root rot and its persistence and concentration on the rhizosphere of lamb's lettuce plants growing in a soilless system. Pf4 was cultured in flasks with 50 ml of nutrient broth (NB, 1 g l^{-1} beef extract, 2 g l^{-1} yeast extract, 5 g l^{-1} peptone, 5 g l^{-1} sodium chloride) at 24°C for 36 h, pelleted with centrifugation at 6500 rpm for 10 min at 4°C and suspended in sterile distilled water to a final concentration of 10^9 CFU ml^{-1}. *R. solani* was cultured in flasks with 200 ml malt extract broth (MEB, malt extract 6 g l^{-1}, maltose 1.8 g l^{-1}, dextrose 6 g l^{-1}, yeast extract 1.2 g l^{-1}) at 24°C for 14–18 d; the mycelium was rinsed with sterile distilled water and thoroughly grinded to obtain a homogeneous suspension. Lamb's lettuce plants were grown in a plant growth room, with the following conditions:

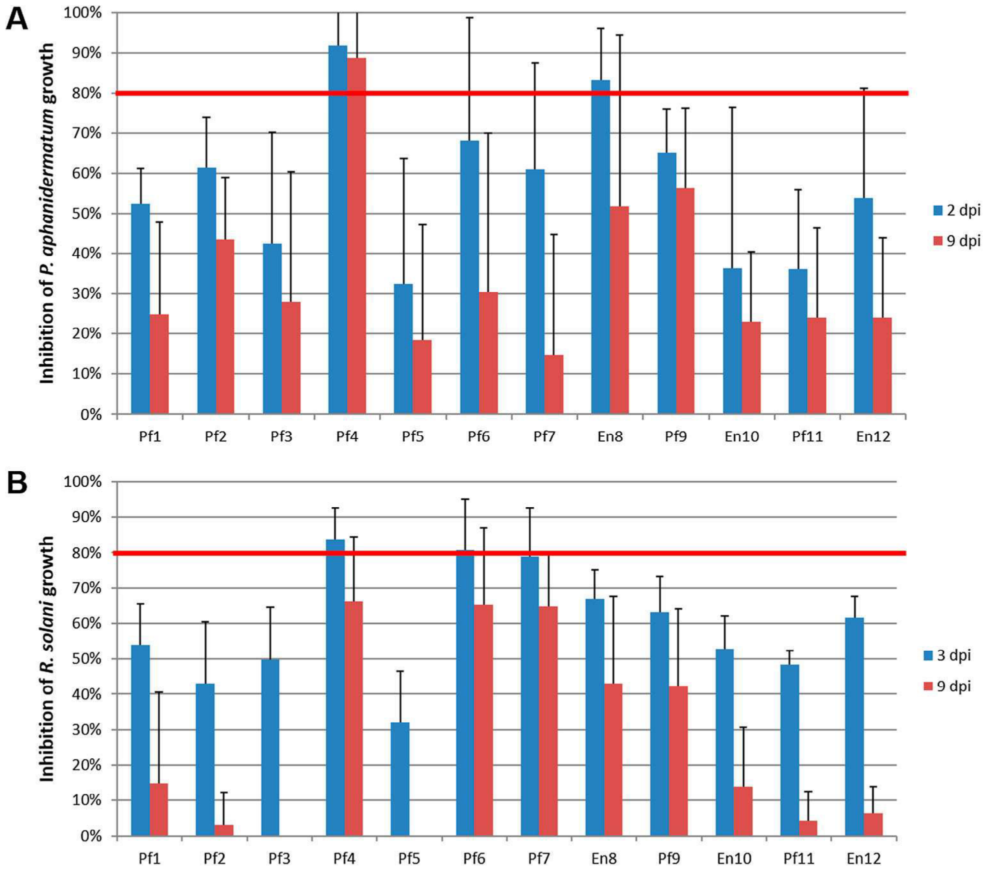


Figure 1. Antagonistic activity (% inhibition of fungal growth, *y* axis) of 12 potential antagonistic bacterial strains (*x* axis) against *P. aphanidermatum* CBS 118745 and CBS 116664 (A), and *R. solani* TR15 and TP20 (B), under *in vitro* conditions (on PDA medium) after 2 or 3 days of incubation respectively, and at the end of the experiments (9 days of incubation). The data shown are the average antagonistic activity of all bacterial strains derived from two experiments with at least three replicates each. Error bars indicate standard deviations.

temperature 26°C, photoperiod of 11 h light/13 h dark, in small-scale floating systems (15 l tanks) with a standard solution widely used by horticultural farms in north-eastern Italy, as reported by Iacuzzo et al. (2011). Specifically, 8 tanks were prepared, in each tank about 50 lamb's lettuce plants were grown. Bacterial treatments were carried out on four of the eight tanks (four replicates for Pf4 treatment) and successively infected with the pathogen; the other four tanks were only infected with the pathogen (four replicates for untreated plants). Eight additional tanks, prepared as above and not inoculated with the pathogen, served as negative controls.

Pf4 bacterial suspensions were used for three treatments: the first was applied on seeds by immersion in the bacterial suspension for 10 min, the second was applied on seedlings at the root level (approximately 10^7 CFU/seedling) about 7 days after seeding; whereas the third one was applied 18 days after seeding directly into the nutrient solution at a final concentration of 10^6 CFU ml^{-1}. Successively, Pf4-treated and untreated plants were artificially infected with the fungal pathogen. For fungal infection, a bunch of lamb's lettuce plants growing in a miniaturised floating system were infected through root immersion for 2 h in the suspension of *R. solani* mycelium. Three days after the third bacterial treatment, six infected plants were put in each of the eight tanks, and used as source of inoculum. Disease development was scored daily as wilted or not wilted for up to three weeks. The number of plants with *R. solani* symptoms (limping, wilting and/or complete withering) was scored.

The experiment was repeated twice (trial I and trial II). Statistical analysis was performed separately on data obtained from each experiment. The data of disease incidence in percentage were subjected to arcsine transformation and to unpaired *T*-test with Welch correction using the software GraphPad InStat version 3.00 (GraphPad Software Inc., San Diego, CA, USA).

2.5.1. Survival and population density of Pseudomonas *sp. strain Pf4 on lamb's lettuce roots in hydroponics*

In order to determine the survival and population density of the inoculated bacteria, root samples (30–300 mg) were weekly collected from two plants randomly selected from each negative control tank of trial I for a period of four weeks, starting 18 days after seeding, just before the application of bacterial suspension into the nutrient solution. Roots from Pf4-treated and untreated plants were weighed, placed in sterile distilled water (1 ml 10 mg^{-1} root tissue) and kept on a rotary shaker for 2 h. Aliquots (100 µl) of the obtained suspensions and of 10-fold serial dilutions were plated in duplicate, using a spreader, onto King's B medium (20 g l^{-1} proteose peptone, 10 ml l^{-1} glycerol, 1.5 g l^{-1} K_2HPO_4, 1.5 g l^{-1} $MgSO_4{\cdot}7\ H_2O$, 15 g l^{-1} agar, pH 7.2) (King, Ward, & Raney, 1954) plates. Colonies were counted (the CFU counting method) after 48 h incubation at 25°C using UV-light.

Molecular identity of 15 colonies from each of the 4 weekly samplings, for a total of 60 colonies from treated plants and 60 colonies from untreated plants, was assessed by a strain-specific EvaGreen® real-time PCR method, the development of which will be described in a separate paper (Martini & Moruzzi, 2017). Bacterial suspensions were prepared with 100 µl of sterile PCR water and bacteria scraped from the agar surface with a sterile plastic loop, successively boiled for 10 min at 99°C. One micro litre of boiled bacterial suspensions was used as a template in 20 µl-PCR reactions including 0.3 µM each primer Pfluor4GyrBF3 and Pfluor4GyrBR2 (5′-CTGTTCAAGTACGAAGGTGGCT-3′/5′-TAAGGTTACGCGT-CAGAGCA-3′) (Martini & Moruzzi, 2017), 1× Sso Fast EvaGreen SuperMix (Bio-Rad Inc., Hercules, CA, USA), and sterile H_2O. Diluted total genomic DNA (2 ng μl^{-1}) of Pf4 was used as positive control in real-time PCRs. Cycling conditions in a 96-well Bio-Rad CFX96 RealTime PCR System (Bio-Rad Inc., Hercules, CA, USA) were as follows: initial denaturation at 98°C for 2 min; 45 cycles of 5 s at 98°C; 5 s at 64°C. A low-resolution melting curve (ramp from 65°C to 95°C with 0.5°C increments and holding times of 5 s) was programmed at the end of the cycling reaction.

2.6. In vitro screening for genes associated with antibiotic production in *Pseudomonas* sp. Pf4

Bacterial strain Pf4 was examined by PCR for the presence of genes involved in antibiotic production using gene-specific primers. Table 1 lists the target genes and PCR primer sets used for the detection of genes encoding the selected antibiotics: 2,4-diacetylphloroglucinol (2,4-DAPG), phenazine-1-carboxylic acid, pyrrolnitrin, pyoluteorin and hydrogen cyanide. All primer sets were used in PCR mixtures with a total volume of 25 μl containing dNTPs 200 μM each, $MgCl_2$ 1.5 mM, each primer 0.4 μM, 0.625U GoTaq Flexi (Promega, Madison, WI, USA). The PCR cycling conditions were: initial denaturation for 2 min at 94°C; 34 cycles of 1 min at 94°C, 40 s at 68°C (or 62/64°C) (Table 1), 1 min at 72°C; and a final extension for 8 min at 72°C. PCR products were separated by electrophoresis in a 1% agarose gel, stained with ethidium bromide, and captured with a DigiDoc-It imaging system (UVP, Cambridge, United Kingdom).

Table 1. Target genes encoding enzymes involved in the biosynthesis of several antibiotics and primer sets used for their amplification in *Pseudomonas* sp. Pf4 strain from this study.

Target gene (antibiotic)	Primer	Sequence (5′-3′)	Annealing T°	Expected size of PCR product	Reference
phlD (2,4-DAPG)	Phl2a Phl2b	GAGGACGTCGAAGACCACCA ACCGCAGCATCGTGTATGAG	62°C	745	Raaijmakers, Weller, and Thomashow (1997)
phzCD (phenazine-1-carboxylic acid)	PCA2a PCA3b	TTGCCAAGCCTCGCTCCAAC CCGCGTTGTTCCTCGTTCAT	68°C	1150	Raaijmakers et al. (1997)
prnD (pyrrolnitrin)	PRND1 PRND2	GGGGCGGGCCGTGGTGATGGA YCCCGCSGCCTGYCTGGTCTG	68°C	786	de Souza and Raaijmakers (2003)
prnC (pyrrolnitrin)	PrnCf PrnCr	CCACAAGCCCGGCCAGGAGC GAGAAGAGCGGGTCGATGAAGCC	64°C	720	Mavrodi et al. (2001)
pltC (pyoluteorin)	PLTC1 PLTC2	AACAGATCGCCCCGGTACAGAACG AGGCCCGGACACTCAAGAAACTCG	68°C	438	de Souza and Raaijmakers (2003)
pltB (pyoluteorin)	PltBf PltBr	CGGAGCATGGACCCCCAGC GTGCCCGATATTGGTCTTGACC	68°C	791	Mavrodi et al. (2001)
hcnBC (hydrogen cyanide)	Aca Acb	ACTGCCAGGGGCGGATGTGC ACGATGTGCTCGGCGTAC	62°C	587	Ramette, Frapolli, Défago, and Moënne-Loccoz (2003)
hcnAB (hydrogen cyanide)	PM2 PM7-26R	TGCGGCATGGGCGTGTGCCATTGCTGCCTGG CCGCTCTTGATCTGCAATTGCAGGCC	68°C	570	Svercel, Duffy, and Défago (2007)

2.7. Library preparation, draft-genome sequencing, assembly and annotation.

Genomic DNA was prepared for sequencing by the Nextera DNA sample preparation kit (Illumina), according to the manufacturer's instructions. Sequencing was performed on an Illumina MiSeq platform using indexed paired-end 300-nucleotide v2 chemistry at the Istituto di Genomica Applicata (Udine, Italy). Paired reads were assembled into contigs using the A5-miseq pipeline (Tritt, Eisen, Facciotti, & Darling, 2012).

Automated annotation of *Pseudomonas* sp. Pf4 draft-genome sequence was performed using the RAST server (Aziz et al., 2008) and the NCBI Prokaryotic Genome Annotation Pipeline (http://www.ncbi.nlm.nih.gov/genome/annotation_prok/). Orthologs inference and comparison with *Pseudomonas protegens* Pf-5 were achieved with the standalone OMA program (http://omabrowser.org/standalone/).

Secondary metabolite production clusters were examined using the antiSMASH program (Medema et al., 2011). Sequence (BLAST) analysis of gene clusters for the synthesis of antibiotics, exoenzyme, cyclic lipopeptide, siderophores, toxin, and of Gac/Rsm homologues in *Pseudomonas* sp. Pf4 was conducted and similarities to those in *P. protegens* and other closely related *Pseudomonas* spp. strains were recorded (Flury et al., 2016; Garrido-Sanz et al., 2016; Loper et al., 2012; Takeuchi et al., 2015).

Contig 8 sequence of *Pseudomonas* sp. Pf4 containing the *fit-rzx* cluster was scanned for regions of genomic islands, putative signatures of HGT, using the IslandViewer3 website (Dhillon et al., 2015) with the algorithms IslandPick (Langille, Hsiao, & Brinkman, 2008), SIGI-HMM (Waack et al., 2006) and IslandPath-DIMOB (Hsiao, Wan, Jones, & Brinkman, 2003).

2.8. Phylogenetic analysis based on multilocus sequence analysis (MLSA)

For the MLSA-based phylogenies, a total of 28 *Pseudomonas* strains of *P. chlororaphis* (including *P. protegens*- and *P. saponiphila*-related strains) and *P. corrugata* subgroups in the *Pseudomonas fluorescens* group according to Mulet, Lalucat, and García-Valdés (2010) and Mulet et al. (2012) were analysed, comprising Pf4, 10 type strains (Gomila, Peña, Mulet, Lalucat, & García-Valdés, 2015) and 17 *Pseudomonas* strains whose complete or draft genome is available in the databases. The sequences of *gyrB*, *rpoD* and *rpoB* housekeeping genes along with the 16S rDNA gene sequence were retrieved from the genomic annotation, if available, and by performing BLASTN on the genomic sequence if otherwise. Genes for the type strains were retrieved from the PseudoMLSA database (http://www.uib.es/microbiologiaBD/Welcome.php).

The sequences of four genes were cut and concatenated as described by Mulet et al. (2010), and successively aligned with CLUSTAL W from the Molecular Evolutionary Genetics Analysis program-MEGA7 (Kumar, Stecher, & Tamura, 2016). The maximum parsimony (MP) tree was obtained using the Tree-Bisection-Regrafting (TBR) algorithm, implemented in the MEGA7, with search level 3 in which the initial trees were obtained by the random addition of sequences (10 replicates). *P. syringae* ATCC19310 type strain was used as an outgroup taxon to root the tree. Bootstrapping (500 replicates) was performed to estimate the stability and support for the inferred clades.

3. Results

3.1. Isolations and preliminary screenings

Bacterial colonies isolated from 30 lamb's lettuce root samples were used in preliminary dual culture tests with two *P. aphanidermatum* strains (CBS 118745 and 116664). Among the 51 bacterial strains tested, 12 strains showed growth limiting activity, as summarised in Table 2. After 4 days of incubation, 3 of the 12 bacteria showed an inhibition zone of more than 10 mm, while 4 showed an inhibition zone ranging from 1 to 10 mm. The remaining five bacteria showed a reduced inhibition zone, although no physical contact was observed between the bacterial and the oomycete growth.

The identification of the 12 bacterial strains was preliminarily carried out by sequence analysis using BLASTN of PCR amplified ribosomal DNAs, that resulted in about 1303–1409 bp in length (accession numbers listed in Table 2). According to the sequence analysis, three bacterial strains (En8, En10, En12) with 16S rDNA gene sequence similarities of 99.2–99.3% among them belonged to *Enterobacter* spp., showing sequence identities of about 99% with three different *Enterobacter* sp. strain sequences deposited in GenBank, while the other nine strains belonged to *P. fluorescens* group. Specifically, six strains (Pf1, Pf2, Pf3, Pf4, Pf5, Pf11) were closely related to *P. protegens* showing a 99–100% sequence similarity with strain CHA0^T (=DSM 19095^T) (AJ278812), two strains (Pf6 and Pf7) to *P. fluorescens* with 99% similarity with strain ATCC 13525^T (AF094725) and one strain (Pf9) to *Pseudomonas poae* with 99% similarity with strain DSM 14936^T (AJ492829).

3.2. In vitro antagonistic activity

The results of *in vitro* antagonism tests of each of the 12 bacterial strains towards the plant pathogens *P. aphanidermatum* and *R. solani* are shown in Figure 1(A,B) respectively. Since *P. aphanidermatum* strains CBS 118745 and CBS 116664, and the *R. solani* strains TR15 and TP20 showed a nearly identical behaviour, combined data for each species are shown. The data from all replicates of the two experiments were also combined (Figure 1). Examples

Table 2. Preliminary data of antagonistic activity against *P. aphanidermatum* after 4 days of incubation on PDA plates, and molecular identification based on BLASTn analysis of 16S rRNA gene sequences with corresponding GenBank accession numbers of 12 selected bacterial strains.

Bacterial strain ID	Antagonistic activity[a]	Accession No.	GenBank closest relative (accession no.)	% similarity
Pf1	++	KM589020	*P. protegens* CHA0^T (AJ278812)	99%
Pf2	+++	KM589021	*P. protegens* CHA0^T (AJ278812)	100%
Pf3	+	KM589022	*P. protegens* CHA0^T (AJ278812)	99%
Pf4	+++	KM589023	*P. protegens* CHA0^T (AJ278812)	100%
Pf5	+	KM589024	*P. protegens* CHA0^T (AJ278812)	99%
Pf6	++	KM589027	*P. fluorescens* ATCC[b] 13525^T (AF094725)	99%
Pf7	+	KM589028	*P. fluorescens* ATCC[b] 13525^T (AF094725)	99%
En8	+++	KM589029	*Enterobacter* sp. TM 1.3 (DQ279307)	99%
Pf9	++	KM589026	*P. poae* DSM[c] 14936^T (AJ492829)	99%
En10	+	KM589030	*Enterobacter* sp. 638 (CP000653)	99%
Pf11	++	KM589025	*P. protegens* CHA0^T (AJ278812)	99%
En12	+	KM589031	*Enterobacter aerogenes* KNUC5012 (JQ682638)	99%

Notes: Pf: bacteria belonging to *P. fluorescens* group; En: bacteria belonging to *Enterobacter* spp.

[a] +: <1 mm inhibition zone; ++: 1 to 10 mm inhibition zone; +++: >10 mm inhibition zone.

[b] ATCC: American Type Culture Collection.

[c] DSM: Deutsche Sammlung von Mikroorganismen.

of the recorded bacterial antagonisms are given in Figure 2. All bacterial strains demonstrated the ability to inhibit the growth of both fungal pathogens, at least in the first 2–3 days of incubation, however bacterial strain Pf4 was the most effective in exhibiting an inhibitory activity of ≥ 82% against both pathogens *P. aphanidermatum* and *R. solani*, after 2 and 3 days of incubation, respectively. After 9 days of incubation, its inhibitory activity was still very high showing an inhibition of fungal growth of > 66% against both pathogens (Figure 1). Interestingly, *P. aphanidermatum* could not be recovered from plates where it was incubated together with Pf4, suggesting that Pf4 had a fungicidal activity against it. After 2–3 days of incubation, moderate effective strains with an inhibitory activity against both pathogens, ranging from 62% to 82%, were Pf6, Pf7, En8 and Pf9. Finally, the least effective strains with a percentage of inhibition ≤ 62 were Pf1, Pf2, Pf3, Pf5, En10, Pf11 and En12.

3.3. In vivo activity of Pseudomonas *sp. strain Pf4 against* R. solani

Pf4-treated and untreated lamb's lettuce plants were artificially infected with the fungal pathogen *R. solani* in order to test the protective effect of Pf4. In both groups of plants,

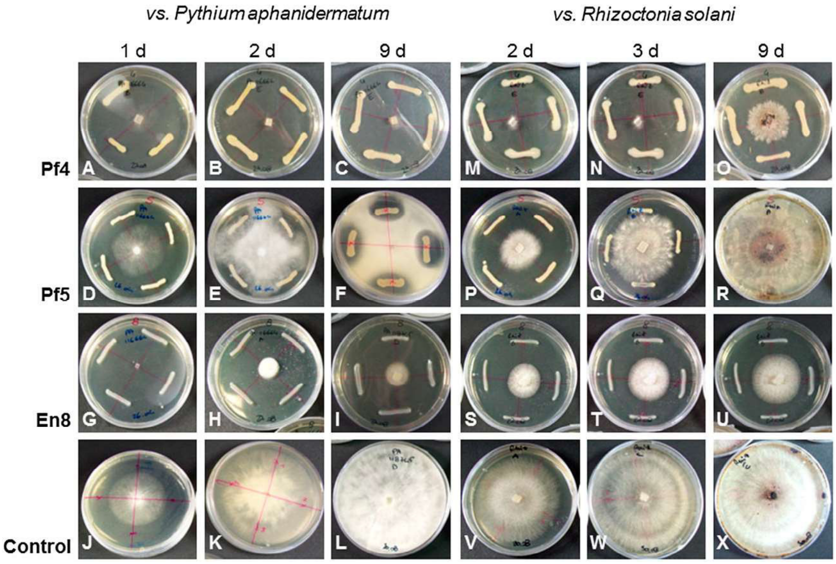


Figure 2. (A–L) Growth of *P. aphanidermatum* cultures at 1, 2 and 9 days of incubation with different bacterial antagonists: A–C, Pf4 (strain with maximum antagonistic activity); D–F, Pf5 (strain with minimum antagonistic activity); G–I, En8 (strain with strong antagonistic activity); J–L, pure culture of *P. aphanidermatum*. Control colony reached the maximum diameter in 2 days (K); at that time even the less-efficient strains showed a quite high inhibition activity, ranging between 32.41% and 68.13% (E). No physical contact was observed for the entire duration of the assay between all the bacteria tested, including those showing low inhibition activity (F), and the mycelium of *P. aphanidermatum*. (M–X) Growth of *R. solani* cultures at 2, 3 and 9 days of incubation with different bacterial antagonists: M–O, Pf4; P–R, Pf5; S–U, En8; V–X, pure culture of *R. solani*. Control colony reached the maximum diameter in 3 days (W), and even the less-efficient strains showed at that time a significant inhibition, ranging between 31.94% and 61.67% (Q). In some cases, a change in *R. solani* mycelium colour becoming darker brown (R), or a change in the shape of the colony edges becoming uneven and jagged (O), were observed.

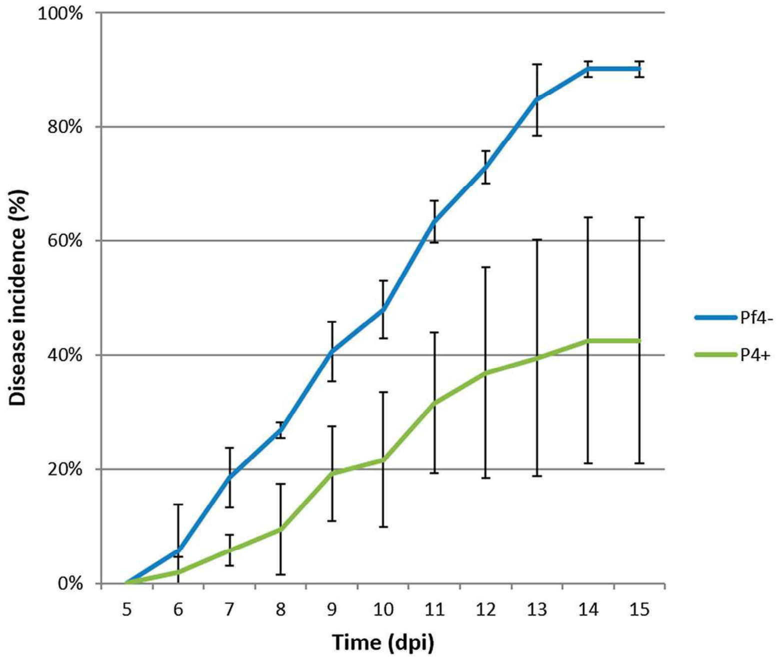


Figure 3. Disease incidence (average % of symptomatic plants per total number of plants observed in the two trials) dynamics of root rot caused by *R. solani* on lamb's lettuce plants, Pf4-treated (Pf4+) or untreated (Pf4−), from 5 to 16 dpi. Error bars indicate standard deviations. In the graph of Pf4-treated plants, the error bar is rather broad because of some inconsistency in the degree of suppressive activity shown in the two trials.

the first symptoms of disease appeared at 6 days after fungal infection (dpi) and developed very fast, especially on untreated plants (Figure 3). In fact, on untreated plants there was a sudden rise at 7 dpi, and then the number of symptomatic plants increased constantly; on Pf4-treated plants, there was a sudden rise at 8–9 dpi, and a slow progression of the disease until 14 dpi. After 14 days, no new infections were observed, neither on untreated or treated plants. In any case, plants infected by *R. solani* showed a sudden shrivelling of leaves, and withered completely in 1–2 days; roots and crown became yellowish-brown and rotted.

Figure 4 with data of disease incidence from the two trials (four replicates each) shows the effects of Pf4 inoculation on lamb's lettuce plants infected with *R. solani* at 14 dpi, when the maximum number of wilted plants was reported. Untreated plants showed a very high disease incidence in both trials with an average disease incidence equal to 91.10 ± 7.59% (mean of four replicates ± SD) in trial I and 89.23 ± 15.05% in trial II, whereas plants treated with Pf4 showed a much lower disease incidence, even though the protection effect in the two trials showed some statistically significant difference. Namely, Pf4-treated plants exhibited a very high protection against *R. solani* in the first trial with an average disease incidence equal to 25.17 ± 5.78% and a lower degree of protection in the second trial with an average disease incidence of 55.60 ± 6.97%. Nevertheless, statistical analysis showed that Pf4 displayed an extremely significant ($P \leq 0.001$, Welch's approximate $t = 9.757$ with 4 degrees of freedom) and significant ($P \leq 0.05$,

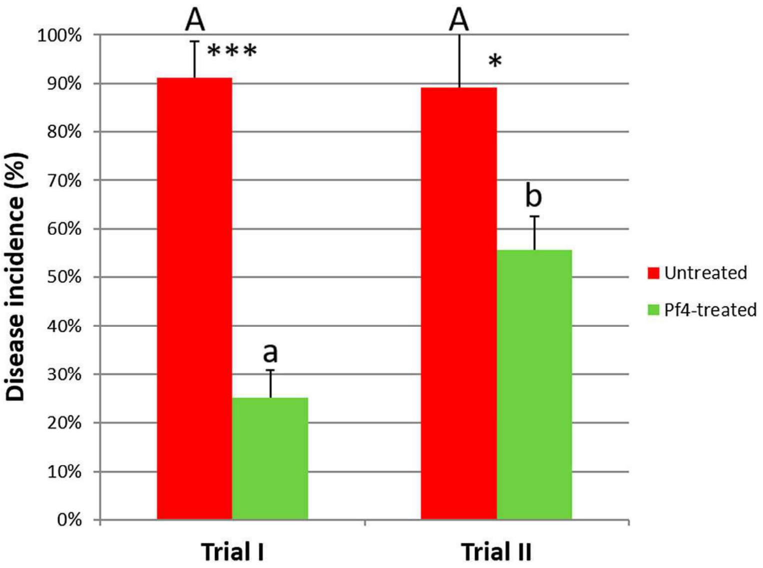


Figure 4. Data of disease incidence (% of symptomatic plants per total number of plants observed) of root rot caused by *R. solani* in the two trials at 14 dpi on Pf4-treated or untreated lamb's lettuce plants. Error bars indicate standard deviations. For each trial, four replicates were carried out. Differences between treated and untreated conditions in each experiment were calculated using unpaired *T*-test with Welch correction. Family-wise significance and confidence level: .05. $^{*}P \leq .05$, $^{***}P \leq .001$. For comparisons among the two different experiments, the same test was used. Different letters indicate different significance levels. Untreated plants: A, no significant difference. Pf4-treated plants: a, no significant difference; b, $P \leq .01$.

Welch's approximate $t = 3.832$ with 3 degrees of freedom) biocontrol activity in trial I and II respectively, against the unprotected control with pathogen alone.

3.3.1. Survival and population density of Pseudomonas *sp. strain Pf4 on lamb's lettuce roots in hydroponics*

The survival and population density of Pf4 on the rhizosphere of lamb's lettuce plants growing in small-scale floating systems, as determined by CFU counting method, is reported in Figure 5. Lines A and C show the overall CFU counts on King's B agar of fluorescent pseudomonads on the roots of Pf4-treated and untreated plants, respectively.

On treated plants, CFU counts ranged from 2×10^5 to 1.5×10^7, and on untreated plants from 0 to 1×10^5. Data obtained from colony counting were then adjusted on the basis of the results of molecular analysis (Figure 5; lines B and D) carried out on randomly sampled fluorescent colonies. In each sample taken from treated roots, 80–100% of the colonies gave a positive reaction (Figure 5, line B) with specific primers Pfluor4gyrB F3/R2, displaying a Ct range between 9 and 17 and a unique melting peak at 86.0°C; whilst in samples collected from untreated roots none of the fluorescent colonies gave a positive reaction (Figure 5, line D). CFU counts of Pf4, over a time span longer than the average growing cycle of lamb's lettuce in hydroponics, ranged between 1.60×10^5 and 1.29×10^7 CFU g^{-1} of root tissue. In particular, Pf4 went across a quick increase in the first week after its inoculation in the tanks, rising the initial concentration of 5.00×10^5 to a maximum of 1.29×10^7 CFU g^{-1}

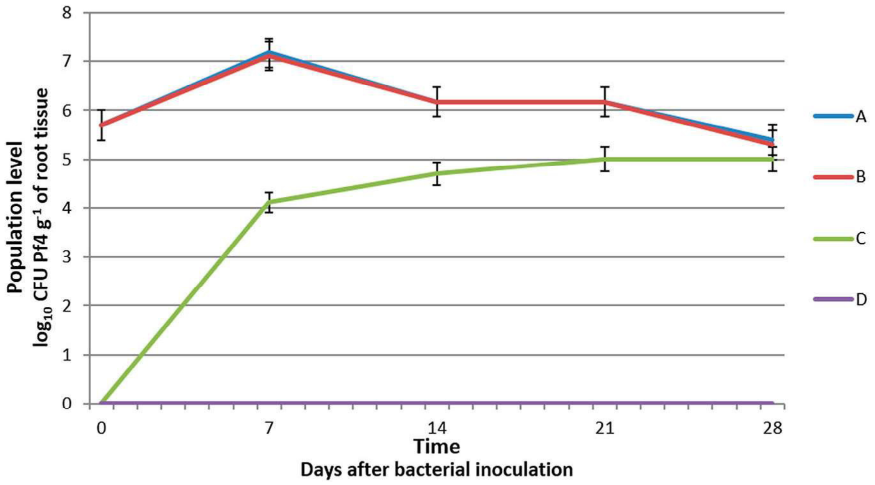


Figure 5. Population density of Pf4 (log10 CFU g^{-1} of root tissue) on lamb's lettuce roots in hydroponics determined by CFU counting method. Lines A: CFU of fluorescent pseudomonads g^{-1} of treated roots; B: CFU of Pf4 g^{-1} of treated roots; C: CFU of fluorescent pseudomonads g^{-1} of untreated roots; D: CFU of Pf4 g^{-1} of untreated roots. Error bars indicate standard deviations.

of root tissue; then Pf4 slowly decreased in the following weeks reaching the minimum concentration of 1.60 × 10^5 CFU g^{-1} of root tissue after four weeks.

3.4. In vitro screening for genes associated with antibiotic production in Pseudomonas strain Pf4

PCR primer sets for conserved sequences of genes involved in the biosynthesis of five antibiotics were targeted against the Pf4 strain. Of the five genes investigated, those involved in the synthesis of 2,4 DAPG (*phlD*), pyrrolnitrin (in both loci *prnD* and *prnC*), pyoluteorin (in both loci *pltC* and *pltB*) and in cyanide production (in both loci *hcnBC* and *hcnAB*) were detected in *Pseudomonas* sp. Pf4, although in locus *hcnAB* a faint PCR signal was obtained even with less stringent PCR conditions. Whereas the gene sequence for phenazine-1-carboxylic acid was not detected in Pf4. In all cases where a positive signal was obtained, the PCR products were of the expected size.

3.5. Genome-wide sequence data

We conducted draft-genome sequencing to obtain information on strain Pf4. The Illumina sequencing provided 1,149,353,940 nts of 300 nts reads that passed the quality check. Sequencing of the Pf4 library provided 3,828,938 reads which were assembled into 36 contigs (N50 = 688,889; largest contig: 1,018,138) for a total of 6,832,152 nts (a coverage of 100.9X). The G + C content was 62.5%, which is similar to that of other sequenced *Pseudomonas* sp. genomes.

Automated annotation of the *Pseudomonas* sp. Pf4 draft-genome sequence using the NCBI pipeline assigned a total of 5907 candidate protein coding-genes, with 1324 (22.41%) annotated as hypothetical proteins. The assembly predicted a total of 62 tRNA and 11 (6 5S, 3 16S, 2 23S) rRNA sequences. The draft-genome sequence of *Pseudomonas* sp. Pf4 has been deposited in the DDBJ/EMBL/GenBank database under the accession no. LUUD00000000. The BioProject designation for this project is PRJNA315258 and the BioSample accession no. is SAMN04554942.

Four-gene clusters (*hcn*, *plt*, *prn*, and *phl*) encoding the enzymes for the synthesis of the typical antibiotics of *P. protegens* were found in the genomic sequence of strain Pf4 (Table 3 and Table S1), which supported the results obtained by PCR analyses for all four antibiotic biosynthetic genes described above. The *hcn* and *phl* gene clusters showed high homology (91–99% and 92–99% respectively) with those of *P. protegens* strains (CHA0^T, Pf-5 and Cab57) (Gross & Loper, 2009; Takeuchi, Noda, & Someya, 2014) and closely related *Pseudomonas* sp. Os17 and St29 (Takeuchi et al., 2015). The *plt* gene cluster showed very high homology (98–100%) only with that of *P. protegens* strains; and the *prn* gene cluster showed high homology (92–98%) with those of *P. protegens* strains and *P. chlororaphis* strains (Table S1).

Other typical gene clusters encoding factors associated with biocontrol found in the Pf4 genome and highly similar to their homologues in *P. protegens* and/or *Pseudomonas* sp. Os17 and St29 (Table 3 and Table S1) include the *aprA* gene cluster (for the major extracellular protease AprA); the genes associated with the Gac/Rsm signal transduction pathway; the gene clusters for pyoverdine, found in the Pf4 genome at four different loci (Gene ID 17855-17860, 29340–29435, 04660–04610, and 04555–04545) as reported in Pf-5 (Gross & Loper, 2009) and Cab57 (Takeuchi et al., 2014); and the genes associated with the synthesis of other siderophores (i.e. enantio-pyochelin, hemophore biosynthesis and ferric-enterobactin receptor) (Table 3 and Table S1).

Among more uncommon genes encoded in the Pf4 genome, we found the gene cluster for orfamides (82–85% similar to that of *P. protegens*), and the complete *rzx* gene cluster (approximately 79 kb, with the highest homology 98–99% to that of Pf-5) encoding analogues of the antimitotic macrolide rhizoxin in *P. protegens* Pf-5 (Loper, Henkels, Shaffer, Valeriote, & Gross, 2008), just upstream the *fit* cluster (with the highest homology 89–97% to that of *P. protegens* strains) (Figure 6, Table S1) encoding a functional insect toxin reported in *P. protegens* Pf-5 (Péchy-Tarr et al., 2008).

The homology search of the gene cluster over the entire genome suggested that the known pathways for the synthesis of phenazine may not be present in the Pf4 strain, confirming PCR results described above.

3.6. Phylogenetic analysis based on MLSA

A phylogenetic tree (Figure 7) was generated based on the concatenated sequences with a total length of 3712 nucleotides in the following order: 16S rRNA (1288 nt), gyrB (798 nt), rpoD (711 nt), and rpoB (915 nt).

In the phylogenetic tree, three well-supported clades can be distinguished, two of them including *P. protegens*-/*P. saponiphila*-related strains (*P. protegens* clade) and *P. chlororaphis*-related strains (*P. chlororaphis* clade) respectively, both belonging to the *P. chlororaphis* subgroup according to Mulet et al. (2010, 2012), and the third clade (*P. corrugata* clade) corresponding to *P. corrugata* subgroup (Mulet et al., 2010, 2012).

Table 3. Overview on presence (+)/absence (−) of secondary metabolites biosynthetic gene clusters in *P. protegens* and closely related *Pseudomonas* spp. strains.

		Gene cluster											
Species	Strain	*hcn*[a]	*plt*[a]	*prn*[a]	*phl*[a]	*aprA*[a]	*pvd*[a]	*pch*[a]	*has*[a]	*pfe*[a]	*ofa*[a]	*fit*[a]	*rzx*[a]
P. protegens	CHA0^T	+	+	+	+	+	+	+	+	+	+	+	−
	Cab57	+	+	+	+	+	+	+	+	+	+	+	−
	Wayne1	+	+	+	+	+	+	+	+	+	+	+	−
	Pf-5	+	+	+	+	+	+	+	+	+	+	+	+
	PF	+	+	+	+	+	+	+	+	+	+	+	+
Pseudomonas spp.	Pf4	+	+	+	+	+	+	+	+	+	+	+	+
	Os17	+	−	−	+	+	+	+	+	+	−	+	+
	St29	+	−	−	+	+	+	+	+	+	−	+	−
	NZI7	+	−	−	+	+	+	+	+	+	−	−	−
	PH1b	+	+	−	−	+	+	+	+	+	+	+	−
	CMR5c	+	+	+	+	+	+	+	+	+	+	+	−
	CMAA1215	−	+	−	+	+	+	+	+	+	+	+	−

Notes: Except Pf4 isolated in the present work from roots in hydroponics, all the other strains were isolated mostly from roots of plants grown in soil.

[a] *hcn*: for hydrogen cyanide; *plt*: for pyoluteorin; *prn*: for pyrrolnitrin; *phl*: for 2,4-diacetylphloroglucinol; *aprA*: for major extracellular protease AprA; *pvd*: for pyoverdine; *pch*: for enantio-pyochelin; *has*: for hemophore biosynthesis; *pfe*: for ferric-enterobactin receptor; *ofa*: for orfamide; *fit*: for FitD toxin; *rzx*: for rhizoxin.

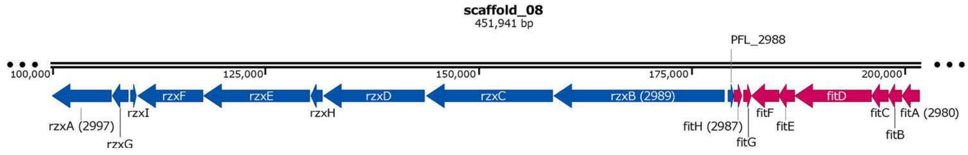


Figure 6. Genetic organisation of the *fit* (for FitD toxin) and *rzx* (for rhizoxin analogues could you please eliminate these two references to the arrows from the text? In the black and white version of the figure that we newly provided, it is no longer possible to perceive the color difference of the arrows.) gene clusters in the genome of Pf4, obtained using SnapGene software (from GSL Biotech; available at snapgene.com).

Phl$^+$ Plt$^+$ *Pseudomonas* strain Pf4 represents a separate branch in the well-supported *P. protegens* clade, which includes Phl$^+$ Plt$^+$ *Pseudomonas* strains closely related to *P. protegens* species (Ramette et al., 2011) (Figure 7, Table 3) and Phl$^+$ Plt$^-$ *Pseudomonas* strains closely related to *P. saponiphila* (Takeuchi et al., 2015; Wu et al., 2016).

In the MLSA of these four genes, sequence similarity of Pf4 was 97.28% with *P. protegens* CHA0^T and 96.8% with *P. saponiphila* DSM 9751^T, demonstrating that Pf4 is a member of the *P. chlororaphis* subgroup, most closely related to *P. protegens* strains.

4. Discussion

A pool of bacterial microorganisms was isolated from roots of healthy lamb's lettuce plants growing in the floating system in a farm in which a *R. solani* root rot outbreak occurred in

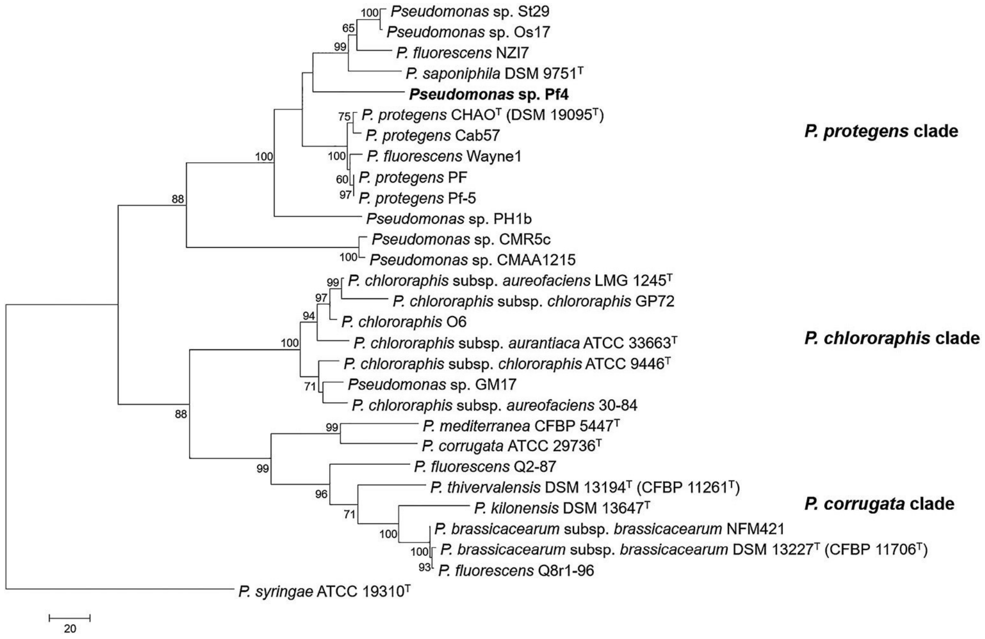


Figure 7. MP phylogenetic tree of strains belonging to *P. chlororaphis* and *P. corrugata* subgroups based on four-gene (16S rRNA, *gyrB*, *rpoD* and *rpoB*) MLSA scheme of Mulet et al. (2010, 2012). Bootstrap values over 50% are indicated in the tree.

2009, with the aim to select microorganisms well adapted to soilless environment and synchronised with the pathogen in time and space (Postma, 2010). Molecular identification based on 16S rRNA gene sequences revealed that 9 of the 12 selected bacteria belonged to genus *Pseudomonas* (6 strains most closely related to *P. protegens*, 2 to *P. fluorescens* and 1 to *P. poae*), and 3 to *Enterobacter*. Bacteria from these genera are common inhabitants of rhizosphere, both in soil and in soilless system, and are well known as biocontrol agents against diseases caused by soil-borne fungal pathogens (Couillerot et al., 2009; Haas & Défago, 2005; Pliego, Ramos, de Vicente, & Cazorla, 2011).

Pf4, the isolate showing the strongest antagonistic *in vitro* activity, was further characterised. It was able to clearly inhibit the growth of both pathogens *P. aphanidermatum* and *R. solani in vitro*; it was then shown in *in vivo* tests with pre-treatment of lamb's lettuce plants growing in hydroponics to reduce significantly *R. solani* disease incidence, despite some inconsistency in the degree of the suppressive activity in the two trials. Whether the variability in the efficacy could be ascribed to the growing system (soilless) or due to factors not associated with the growing system, such as poor host colonisation by the biocontrol agent or variable expression of genes involved in disease suppression, as reported for experiments carried out in soil (Raaijmakers et al., 2002) could not be ascertained and deserves further investigations.

During *in vivo* test (trial I), the persistence and concentration of Pf4 on the rhizosphere were monitored by a conventional culturing method and molecular analysis, which demonstrated that the totality or majority of the fluorescent pseudomonads from treated roots corresponded to Pf4, while in the case of untreated ones none of the fluorescent pseudomonads resembled Pf4. Hence, Pf4 was capable of surviving at high level of population in the rhizosphere for a period of 4 weeks starting 18 days after seeding, therefore exceeding the entire lamb's lettuce growing cycle in the floating system. The population dynamics were consistent with those reported in the literature for soil (Haas & Défago, 2005), that is, artificially inoculated biocontrol agent initially colonise roots at 10^7–10^8 CFU g^{-1}, then decline within few weeks. The lowest colonisation level shown by Pf4 was 1.60×10^5 CFU g^{-1} of lamb's lettuce root, corresponding to the threshold population density (10^5–10^6 CFU g^{-1} of root) that must be reached by *Pseudomonas* spp. strains for effective disease suppression in soil (Haas & Défago, 2005).

Since the fluorescent pseudomonads population level of untreated plants was quite similar at the end of the monitoring period, we could confirm previous works (Vallance et al., 2011) indicating that also in soilless cultures a bacterial population could naturally and quickly develop without artificial inoculation, even though starting with a 'microbiological vacuum' (Postma, 2010).

In order to shed light on the mechanisms underlying the biocontrol properties of *Pseudomonas* sp. Pf4, PCRs detecting known loci for the synthesis of antifungal metabolites, and draft-genome sequencing were undertaken. Indeed, both methods showed the presence in Pf4 of genes involved in the biosynthesis of typical *P. protegens* secondary metabolites, such as gene clusters *hcn*, *plt*, *prn*, and *phl*, involved in the production of hydrogen cyanide, pyoluteorin, pyrrolnitrin and 2,4-DAPG, respectively. The biosynthesis of pyoluteorin was claimed (Garrido-Sanz et al., 2016) to be specific of *P. protegens* group strains within the *P. fluorescens* complex; the results of this study and of that of Flury et al. (2016) demonstrated indeed that in the *P. chlororaphis* subgroup defined according to Mulet et al. (2010, 2012), also other *Pseudomonas* spp. strains (i.e. Pf4, PH1b, CMR5c and CMAA1215, Table 3 and Figure 7) besides *P. protegens* species strains harbour the *plt* gene cluster.

In addition to the above, also other gene clusters coding for extracellular enzymes such as *apr* gene cluster and siderophores such as *pch*, *has* and *pfe* gene clusters, besides Gac/Rsm homologues and small regulatory RNAs, showed high homology with *P. protegens* strains, as well as with *Pseudomonas* sp. Os17 and St29, supporting the notion of a close relatedness of Pf4 to both groups of fluorescent pseudomonads. Interestingly, Pf4 also has the biosynthetic potential for metabolites that are less universally spread among the fluorescent pseudomonads; in particular, with our genomic drafting we discovered in Pf4 the gene clusters for the cyclic lipopeptide orfamide A, for the insect toxin FitD and for rhizoxin analogues, recently identified natural products discovered through genomics-guided approaches. Orfamide A, a biosurfactant influencing swarming motility of Pf-5, was shown to function as an antifungal agent, to lyse oomycete zoospores, and to act as an insecticidal agent (Gross & Loper, 2009; Ma et al., 2016). The gene cluster for orfamides, which has been identified in strain Pf-5 mining *Pseudomonas* genomes (Gross et al., 2007), was also found in the genomes of other *P. protegens* strains, CHA0^T and Cab57 (Takeuchi et al., 2014), and of *P. protegens*-related strains (i.e. *Pseudomonas* spp. CMR5c, CMR12a, CMAA1215, PH1b) (Ma et al., 2016). The Fit insect toxin cluster was first identified in *P. protegens* Pf-5, in which the production of this toxin has been associated with the lethality of this strain for the tobacco hornworm *Manduca sexta* (Péchy-Tarr et al., 2008). The complete gene cluster has also been identified in *P. protegens* CHA0^T and several other *P. protegens* strains, in closely related *Pseudomonas* spp. Os17, St29 and CMR5c, in *P. chlororaphis* strains O6, 30–84 and many others, suggesting that the Fit toxin is consistently and exclusively shared by strains belonging to the *P. chlororaphis* subgroup [corresponding to sub-clade 1 after Loper et al. (2012)] (Flury et al., 2016; Garrido-Sanz et al., 2016; Loper et al., 2012; Péchy-Tarr et al., 2013; Takeuchi et al., 2015).

Rhizoxins are 16-membered polyketide macrolides that exhibit significant phytotoxic, antifungal and antitumoral properties by binding to b-tubulin, thereby interfering with microtubule dynamics during mitosis. The complete *rxz* cluster has been initially reported in *P. protegens* Pf-5 (Loper et al., 2008). This cluster has been found to be absent from two other fully sequenced *P. protegens* strains, CHA0^T and Cab57 (Takeuchi et al., 2014), but present in *P. protegens* PF and closely related *Pseudomonas* sp. Os17 (Loper et al., 2016; Takeuchi et al., 2015) in the *P. fluorescens* group.

In Pf4 the rhizoxin biosynthesis gene cluster is adjacent to the gene cluster encoding for the production of the FitD insect toxin. To date only few other closely related *Pseudomonas* spp. strains, *P. protegens* strains Pf-5 and PF and the related strain *Pseudomonas* sp. Os17, are known to have the Fit and rhizoxin gene clusters linked (i.e. the *fit-rzx* cluster) in their genomes. As in *P. protegens* Pf-5 and *Pseudomonas* sp. Os17, the genomic region with the *fit-rzx* gene clusters of Pf4 did not show the characteristics of a genomic island, although Loper et al. (2016) suggested that the *fit-rzx* clusters of Pf-5 and closely related strains have a complex evolutionary history that includes HGT. Loper et al. (2016) demonstrated that the *fit-rxz* cluster confers oral and injectable toxicity to a broader set of insects than either the *fit* or *rzx* clusters alone, therefore Pf4 represents a potential bacteria that may exhibit oral toxicity towards agriculturally relevant insect pests such as Pf-5. Testing *in vivo* insecticidal activity would be an interesting address for future research on Pf4. After the recent discovery that certain pseudomonads can not only suppress fungal plant diseases but also have the potential to control insect pests, the results of this work further widen the application targets of the so called *P. chlororaphis* subgroup, adding value to their use as biocontrol agents and opening up new industrial opportunities toward the development of unique biopesticides for biological control of plant diseases and pests using the same product in different growth environments.

Draft genome of Pf4 allowed also to obtain the sequence of the housekeeping *rpoD*, *gyrB* and *rpoB* genes, which represent the three genes besides the 16S rRNA gene used in the MLSA developed by Mulet et al. (2010) and proved to be a useful tool for *Pseudomonas* spp. identification at the species level (Gomila et al., 2015). MLSA is a major contribution to accurate identification, needed since a large number of strains with disease suppression potential are reported as *P. fluorescens*, but only some of them are presently retained within this species (Bossis, Lemanceau, Latour, & Gardan, 2000; Mulet et al., 2010). Mulet et al. (2010) established a similarity of 97.0% in the MLSA of these four genes as the threshold value for strains in the same species in the genus *Pseudomonas*. The sequence similarity obtained between Pf4 and *P. protegens* CHA0^T or *P. saponiphila* DSM 9751^T (97.28 and 96.80% respectively) and the phylogenetic analysis indicated that Pf4 potentially belong to a novel *Pseudomonas* species, as it forms a clearly distinct lineage within the *P. protegens* clade (Figure 7) in the *P. chlororaphis* subgroup defined according to Mulet et al. (2010, 2012).

Despite the fact that it was isolated from the roots of plants in hydroponic culture, Pf4 was not only at the genomic level, but also at the taxonomic level, rather similar to other strains of *Pseudomonas* spp. that have been isolated from soil and shown to be active biocontrol agent in soil.

5. Conclusions

Pf4 displayed the ability to inhibit the growth of *R. solani* and *P. aphanidermatum in vitro*, and the capacity to suppress root rot caused by *R. solani in vivo*, on lamb's lettuce plants grown in hydroponics. It could be inferred from the drafted genome sequence that Pf4 has the potential to produce an arsenal of secondary metabolites very similar to that of the well-known biocontrol *P. protegens* strain Pf-5. Actually, Pf4 is the only *P. protegens*-related strain among those analysed, which is more like Pf-5 in the type of secondary metabolites produced. Moreover, Pf4 can colonise lamb's lettuce roots for the entire growth cycle of this crop in the floating system at a density of 10^5–10^7 CFU g^{-1} of root, therefore above the threshold required for suppression of root diseases in soil. This work supports the notion that key factors conferring the ability to suppress root diseases in soil are also of paramount relevance in hydroponics.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by a grant (L. R. 26) from Friuli Venezia Giulia Region Administration (Italy) and by "Ager - Agroalimentare e Ricerca" Foundation, project "Novel strategies meeting the needs of the fresh-cut vegetable sector - STAYFRESH", under Grant number 2010 2370.

ORCID

Giuseppe Firrao http://orcid.org/0000-0002-7890-0899
Paolo Ermacora http://orcid.org/0000-0003-0757-7956
Nazia Loi http://orcid.org/0000-0002-9738-9248
Marta Martini http://orcid.org/0000-0002-7271-5297

References

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., … Meyer, F. (2008). The RAST server: Rapid annotations using subsystems technology. *BMC Genomics*, *9*, 75. doi:10.1186/1471-2164-9-75

Bakker, P. A., Pieterse, C. M., & Van Loon, L. C. (2007). Induced systemic resistance by fluorescent *Pseudomonas* spp. *Phytopathology*, *97*, 239–243. doi:10.1094/PHYTO-97-2-0239

Bossis, E., Lemanceau, P., Latour, X., & Gardan, L. (2000). The taxonomy of *Pseudomonas fluorescens* and *Pseudomonas putida*: Current status and need for revision. *Agronomie*, *20*, 51–63. doi:10.1051/agro:2000112

Colla, P., Gilardi, G., & Gullino, M. L. (2012). A review and critical analysis of the European situation of soilborne disease management in the vegetable sector. *Phytoparasitica*, *40*, 515–523. doi:10.1007/s12600-012-0252-2

Compant, S., Duffy, B., Nowak, J., Clément, C., & Barka, E. A. (2005). Use of plant growth-promoting bacteria for biocontrol of plant diseases: Principles, mechanisms of action, and future prospects. *Applied and Environmental Microbiology*, *71*, 4951–4959. doi:10.1128/AEM.71.9.4951-4959.2005

Couillerot, O., Prigent-Combaret, C., Caballero-Mellado, J., & Moënne-Loccoz, Y. (2009). *Pseudomonas fluorescens* and closely-related fluorescent pseudomonads as biocontrol agents of soil-borne phytopathogens. *Letters in Applied Microbiology*, *48*, 505–512. doi:10.1111/j.1472-765X.2009.02566.x

de Souza, J. T., & Raaijmakers, J. M. (2003). Polymorphisms within the *prnD* and *pltC* genes from pyrrolnitrin and pyoluteorin-producing *Pseudomonas* and *Burkholderia* spp. *FEMS Microbiology Ecology*, *43*, 21–34. doi:10.1111/j.1574-6941.2003.tb01042.x

Dhillon, B. K., Laird, M. R., Shay, J. A., Winsor, G. L., Lo, R., Nizam, F., … Brinkman, F. S. (2015). Islandviewer 3: More flexible, interactive genomic island discovery, visualization and analysis. *Nucleic Acids Research*, *43*, W104–W108. doi:10.1093/nar/gkv401

Flury, P., Aellen, N., Ruffner, B., Péchy-Tarr, M., Fataar, S., Metla, Z., … Maurhofer, M. (2016). Insect pathogenicity in plant-beneficial pseudomonads: Phylogenetic distribution and comparative genomics. *The ISME Journal*, *10*, 2527–2542. doi:10.1038/ismej.2016.5

Fokkema, N. J. (1976). Antagonism between fungal saprophytes and pathogens on aerial plant surface. In C. H. Dickinson & T. F. Preece (Eds.), *Microbiology of aerial plant surfaces* (pp. 487–506). London: Academic Press Ltd.

Garibaldi, A., Gilardi, G., & Gullino, M. L. (2006). First report of *Rhizoctonia solani* AG 4 on lamb's lettuce in Italy. *Plant Disease*, *90*, 1109. doi:10.1094/PD-90-1109C

Garrido-Sanz, D., Meier-Kolthoff, J. P., Göker, M., Martín, M., Rivilla, R., & Redondo-Nieto, M. (2016). Genomic and genetic diversity within the *Pseudomonas fluorescens* complex. *PloS one*, *11*, e0150183. doi:10.1371/journal.pone.0150183

Gomila, M., Peña, A., Mulet, M., Lalucat, J., & García-Valdés, E. (2015). Phylogenomics and systematics in *Pseudomonas*. *Frontiers in Microbiology*, 6, 214. doi:10.3389/fmicb.2015.00214

Gravel, V., Martinez, C., Antoun, H., & Tweddell, R. J. (2005). Antagonist microorganisms with the ability to control *Pythium* damping-off of tomato seeds in rockwool. *BioControl*, *50*, 771–786. doi:10.1007/s10526-005-1312-z

Gross, H., & Loper, J. E. (2009). Genomics of secondary metabolite production by *Pseudomonas* spp. *Natural Product Reports*, *26*, 1408–1446. doi:10.1039/b817075b

Gross, H., Stockwell, V. O., Henkels, M. D., Nowak-Thompson, B., Loper, J. E., & Gerwick, W. H. (2007). The genomisotopic approach: A systematic method to isolate products of orphan biosynthetic gene clusters. *Chemistry & Biology*, *14*, 53–63. doi:10.1016/j.chembiol.2006.11.007

Haas, D., & Défago, G. (2005). Biological control of soil-borne pathogens by fluorescent pseudomonads. *Nature Reviews Microbiology*, *3*, 307–319. doi:10.1038/nrmicro1129

Hall, T. A. (1999). Bioedit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, *41*, 95–98.

Handelsman, J., & Stabb, E. V. (1996). Biocontrol of soilborne plant pathogens. *The Plant Cell*, *8*, 1855–1869. doi:10.1105/tpc.8.10.1855

Hsiao, W., Wan, I., Jones, S. J., & Brinkman, F. S. (2003). Islandpath: Aiding detection of genomic islands in prokaryotes. *Bioinformatics (Oxford, England)*, *19*, 418–420. doi:10.1093/bioinformatics/btg004

Iacuzzo, F., Gottardi, S., Tomasi, N., Savoia, E., Tommasi, R., Cortella, G., … Cesco, S. (2011). Corn salad (*Valerianella locusta* (L.) Laterr.) growth in a water-saving floating system as affected by iron and sulfate availability. *Journal of the Science of Food and Agriculture*, *91*, 344–354. doi:10.1002/jsfa.4192

King, E. O., Ward, M. K., & Raney, D. E. (1954). Two simple media for the demonstration of pyocyanin and fluorescin. *The Journal of Laboratory and Clinical Medicine*, *44*, 301–307.

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, *33*, 1870–1874, msw054. doi:10.1093/molbev/msw054

Langille, M. G., Hsiao, W. W., & Brinkman, F. S. (2008). Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinformatics*, *9*, 329. doi:10.1186/1471-2105-9-329

Loper, J. E., Hassan, K. A., Mavrodi, D. V., Davis II, E. W., Lim, C. K., Shaffer, B. T., … Paulsen, I. T. (2012). Comparative genomics of plant-associated *Pseudomonas* spp.: Insights into diversity and inheritance of traits involved in multitrophic interactions. *PLoS Genetics*, *8*, e1002784. doi:10.1371/journal.pgen.1002784

Loper, J. E., Henkels, M. D., Rangel, L. I., Olcott, M. H., Walker, F. L., Bond, K. L., … Taylor, B. J. (2016). Rhizoxin, orfamide a, and chitinase production contribute to the toxicity of *Pseudomonas protegens* strain Pf-5 to *Drosophila melanogaster*. *Environmental Microbiology*, *18*, 3509–3521. doi:10.1111/1462-2920.13369

Loper, J. E., Henkels, M. D., Shaffer, B. T., Valeriote, F. A., & Gross, H. (2008). Isolation and identification of rhizoxin analogs from *Pseudomonas fluorescens* Pf-5 by using a genomic mining strategy. *Applied and Environmental Microbiology*, *74*, 3085–3093. doi:10.1128/AEM.02848-07

Ma, Z., Geudens, N., Kieu, N. P., Sinnaeve, D., Ongena, M., Martins, J. C., & Höfte, M. (2016). Biosynthesis, chemical structure, and structure-activity relationship of orfamide lipopeptides produced by *Pseudomonas protegens* and related species. *Frontiers in Microbiology*, *7*, 382. doi:10.3389/fmicb.2016.00382

Martini, M., & Moruzzi, S. (2017). *Specific detection and quantification of the biocontrol agent Pseudomonas sp. strain Pf4 by real-time PCR and high-resolution melting (HRM) analysis*. Unpublished manuscript.

Martini, M., Musetti, R., Grisan, S., Polizzotto, R., Borselli, S., Pavan, F., & Osler, R. (2009). DNA-dependent detection of the grapevine fungal endophytes *Aureobasidium pullulans* and *Epicoccum nigrum*. *Plant Disease*, 93, 993–998.

Mavrodi, O. V., McSpadden Gardener, B. B., Mavrodi, D. V., Bonsall, R. F., Weller, D. M., & Thomashow, L. S. (2001). Genetic diversity of phlD from 2, 4-diacetylphloroglucinol-producing fluorescent *Pseudomonas* spp. *Phytopathology*, *91*, 35–43. doi:10.1094/PHYTO.2001.91.1.35

McPherson, G. M. (1998). *Root diseases in hydroponics – their control by disinfection and evidence for suppression in closed systems*. Paper presented at Abstract 7th International congress on plant pathology, Edinburgh, UK.

McPherson, G. M., Harriman, M. R., & Pattison, D. (1995). The potential for spread of root diseases in recirculating hydroponic systems and their control with disinfection. *Mededelingen-Faculteit Landbouwkundige en Toegepaste Biologische Wetenschappen Universiteit Gent (Belgium)*, *60*, 371–379.

Medema, M. H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M. A., … Breitling, R. (2011). antiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research*, 39(suppl 2), W339–W346. doi:10.1093/nar/gkr466

Mulet, M., Gomila, M., Scotta, C., Sánchez, D., Lalucat, J., & García-Valdés, E. (2012). Concordance between whole-cell matrix-assisted laser-desorption/ionization time-of-flight mass spectrometry and multilocus sequence analysis approaches in species discrimination within the genus *Pseudomonas*. *Systematic and Applied Microbiology*, *35*, 455–464. doi:10.1016/j.syapm.2012.08.007

Mulet, M., Lalucat, J., & García-Valdés, E. (2010). DNA sequence-based analysis of the *Pseudomonas* species. *Environmental Microbiology*, *12*, 1513–1530. doi:10.1111/j.1462-2920.2010.02181.x

Paulitz, T. C., & Bélanger, R. R. (2001). Biological control in greenhouse systems. *Annual Review of Phytopathology*, *39*, 103–133. doi:10.1146/annurev.phyto.39.1.103

Péchy-Tarr, M., Borel, N., Kupferschmied, P., Turner, V., Binggeli, O., Radovanovic, D., … Keel, C. (2013). Control and host-dependent activation of insect toxin expression in a root-associated biocontrol pseudomonad. *Environmental Microbiology*, *15*, 736–750. doi:10.1111/1462-2920.12050

Péchy-Tarr, M., Bruck, D. J., Maurhofer, M., Fischer, E., Vogne, C., Henkels, M. D., … Keel, C. (2008). Molecular analysis of a novel gene cluster encoding an insect toxin in plant-associated strains of *Pseudomonas fluorescens*. *Environmental Microbiology*, *10*, 2368–2386. doi:10.1111/j.1462-2920.2008.01662.x

Pliego, C., Ramos, C., de Vicente, A., & Cazorla, F. M. (2011). Screening for candidate bacterial biocontrol agents against soilborne fungal plant pathogens. *Plant and Soil*, *340*, 505–520. doi:10.1007/s11104-010-0615-8

Postma, J. (2010). The status of biological control of plant diseases in soilless cultivation. In U. Gis, I. Chet, & M. L. Gullino (Eds.), *Recent developments in management of plant diseases* (pp. 133–146). Dordrecht: Springer.

Raaijmakers, J. M., Paulitz, T. C., Steinberg, C., Alabouvette, C., & Moënne-Loccoz, Y. (2009). The rhizosphere: A playground and battlefield for soilborne pathogens and beneficial microorganisms. *Plant and Soil*, *321*, 341–361. doi:10.1007/s11104-008-9568-6

Raaijmakers, J. M., Vlami, M., & De Souza, J. T. (2002). Antibiotic production by bacterial biocontrol agents. *Antonie van Leeuwenhoek*, *81*, 537–547. doi:10.1023/A:1020501420831

Raaijmakers, J. M., Weller, D. M., & Thomashow, L. S. (1997). Frequency of antibiotic-producing *Pseudomonas* spp. In natural environments. *Applied and Environmental Microbiology*, *63*, 881–887.

Ramette, A., Frapolli, M., Défago, G., & Moënne-Loccoz, Y. (2003). Phylogeny of HCN synthase-encoding *hcnBC* genes in biocontrol fluorescent pseudomonads and its relationship with host plant species and HCN synthesis ability. *Molecular Plant-Microbe Interactions*, *16*, 525–535. doi:10.1094/MPMI.2003.16.6.525

Ramette, A., Frapolli, M., Fischer-Le Saux, M., Gruffaz, C., Meyer, J. M., Défago, G., … Moënne-Loccoz, Y. (2011). *Pseudomonas protegens* sp. nov., widespread plant-protecting bacteria producing the biocontrol compounds 2, 4-diacetylphloroglucinol and pyoluteorin. *Systematic and Applied Microbiology*, *34*, 180–188. doi:10.1016/j.syapm.2010.10.005

Schnitzler, W. H. (2004). Pest and disease management of soilless culture. *South Pacific Soilless Culture Conference-SPSCC. Acta Horticulturae*, *648*, 191–203. doi:10.17660/ActaHortic.2004.648.23

Sharon, M., Kuninaga, S., Hyakumachi, M., & Sneh, B. (2006). The advancing identification and classification of *Rhizoctonia* spp. Using molecular and biotechnological methods compared with the classical anastomosis grouping. *Mycoscience*, *47*, 299–316. doi:10.1007/s10267-006-0320-x

Svercel, M., Duffy, B., & Défago, G. (2007). PCR amplification of hydrogen cyanide biosynthetic locus *hcnAB* in *Pseudomonas* spp. *Journal of Microbiological Methods*, *70*, 209–213. doi:0.1016/j.mimet.2007.03.018

Takeuchi, K., Noda, N., Katayose, Y., Mukai, Y., Numa, H., Yamada, K., & Someya, N. (2015). Rhizoxin analogs contribute to the biocontrol activity of a newly isolated *Pseudomonas* strain. *Molecular Plant-Microbe Interactions*, *28*, 333–342. doi:10.1094/MPMI-09-14-0294-FI

Takeuchi, K., Noda, N., & Someya, N. (2014). Complete genome sequence of the biocontrol strain *Pseudomonas protegens* Cab57 discovered in Japan reveals strain-specific diversity of this species. *PloS one*, *9*, e93683. doi:10.1371/journal.pone.0093683

Tritt, A., Eisen, J. A., Facciotti, M. T., & Darling, A. E. (2012). An integrated pipeline for de novo assembly of microbial genomes. *PloS one*, *7*, e42304. doi:10.1371/journal.pone.0042304

Vallance, J., Déniel, F., Le Floch, G., Guérin-Dubrana, L., Blancard, D., & Rey, P. (2011). Pathogenic and beneficial microorganisms in soilless cultures. *Agronomy for Sustainable Development*, *31*, 191–203. doi:10.1051/agro/2010018

Van Os, E. A. (1999). Closed soilless growing systems: A sustainable solution for Dutch greenhouse horticulture. *Water Science and Technology*, *39*, 105–112.

Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W. F., … Merkl, R. (2006). Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics*, *7*, 142. doi:10.1186/1471-2105-7-142

Weisburg, W. G., Barns, S. M., Pelletier, D. A., & Lane, D. J. (1991). 16S ribosomal DNA amplification for phylogenetic study. *Journal of Bacteriology*, *173*, 697–703. doi:10.1128/jb.173.2.697-703.1991

White, T. J., Bruns, T., Lee, S. J. W. T., & Taylor, J. W. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *PCR Protocols: a Guide to Methods and Applications*, *18*, 315–322.

Wilson, K. (1997). Preparation of genomic DNA from bacteria. *Current Protocols in Molecular Biology*, 2.4.1–2.4.5. doi:10.1002/0471142727.mb0204s56

Wu, L., Shang, H., Wang, Q., Gu, H., Liu, G., & Yang, S. (2016). Isolation and characterization of antagonistic endophytes from *Dendrobium candidum* Wall ex Lindl., and the biofertilizing potential of a novel *Pseudomonas saponiphila* strain. *Applied Soil Ecology*, *105*, 101–108. doi:10.1016/j.apsoil.2016.04.008

2.2 Genome sequence and antifungal activity in two niche-sharing *Pseudomonas protegens* strains isolated from hydroponics

Authors: Cesare Polano, Marta Martini, Francesco Savian, Serena Moruzzi, Paolo Ermacora, Giuseppe Firrao.

Manuscript submitted to "Biological control".

2.2.1 Introduction

The rhizosphere environment hosts a complex of microorganisms that interact with plants in a multitude of ways, such as the mutually beneficial relationship between nitrogen-fixing bacteria and the radical apparatus; adverse interactions, such as the continuously evolving conflict between pathogens and their more-or-less specific hosts; and neutral interactions, in which neither the microorganisms nor the plants derive any particular benefit, but also no significant detriment (Raaijmakers *et al.*, 2009). Vegetable crops are particularly sensitive to adverse interactions for various reasons, to name a few: the intensity of cultivation can often push their physiological capabilities, requiring a surge in the uptake of resources during the productive season, which can leave the plants more susceptible to pathogen aggression; the use of a limited set of commercially sought-after cultivars, which are often on the lower side of disease resistance, when compared to more recently developed cultivars and breeds; and also the need for more environmentally sound methods of pest and disease management (Colla *et al.*, 2012), motivated both by scientific and social reasons. The impact of pathogens can be reduced, sometimes significantly, by employing various strategies, including the competition and antibiosis between microorganisms, the induction of local or systemic resistance in hosts, and by influencing the chemical characteristics of the soil itself (Mazzola, 2002).

While not all root diseases can be avoided (Vallance *et al.*, 2011), their impact can be decisively limited by employing microorganisms as biocontrol agents (Handelsman and Stabb, 1996). A particularly suited genus is *Pseudomonas*, common in all agricultural soils (Paulitz and Bélanger, 2001; Weller, 2007); *P. fluorescens* strains have been intensely studied as models for rhizosphere ecological studies and analysis of secondary metabolism (Couillerot *et al.*, 2009), and *Pseudomonas* species have shown the ability of inhibiting the growth of fungal pathogens in e.g. hydroponic cultures of *Valerianella locusta* (L.) Laterr., *Pythium aphanidermatum* and *Rhizoctonia solani* in particular (Rankin, 1994) (Van Os, 1999)

Bacterial communities constitute a complex relation of interactions, driven by the necessity of controlling limited resources (Hibbing *et al.*, 2010; Stubbendieck and Straight, 2016). These interactions can be competitive or mutualistic, and can be between different species, different strains in the same species, or between members of the same species and strain. As a general rule, interactions become more and more competitive the more distant the individuals are on a genetic level.

Cooperation often involves quorum sensing (Hense *et al.*, 2007): as small quantities of antimicrobial compounds can induce a physiologically tolerant state (bacteria that produce these compounds are most likely tolerant themselves because of this mechanism), a common strategy is to delay production of antimicrobials until enough individuals are present (Lyon and Novick, 2004), so that the release of these compounds will reach full inhibitory level of concentration. Such antimicrobial compounds can have various level of specificity; organism with narrower range of habitats will often have more specific targets.

Some degree of cooperation can however be present also between different species, such as between *Pseudomonas putida* strain R1 and *Acinetobacter* strain C6, where the former depends on the benzoate produced by latter to grow on benzyl alcohol (Christensen *et al.*, 2002).

Competition has been often compared to an arms race, with broadly two categories of strategies: one involves direct interaction, where two or more individuals (or colonies) attempt to displace the competition and get access to most of the resources; the other is indirect, and consists e.g. in a faster uptake of limited resources, outgrowing the competition (Nicholson, 1954). In large enough communities, some individuals can take advantage of metabolites produced by other organisms and shift the cost of producing them, e.g. using heterologous siderophores (Khan *et al.*, 2006). This phenomenon is more frequent against different species, but has also been observed intra-species.

Access to favourable locations involves colonising new niches or displacing existing competition. Long term persistence requires mechanisms for keeping hold of the position; some species produce receptors that bind to specific surfaces (Johnson-Henry *et al.*, 2007), while others produce antimicrobials or molecules that facilitate dispersal of competing organisms (Xavier and Foster, 2007)

The diversity and composition of the microbial community can consequently change in time, with some species prevailing over others while they compete for the same resources. If the 'weaker' species happen to be those that are inoculated for biocontrol purposes, it becomes necessary to reintroduce them periodically to keep them at an efficacious population level. Alternatively, a better strategy would be attempting to facilitate a stable presence, or at least a slower decline, of the species of economical interest to plant protection.

While the simplest form of biocontrol is one metabolite *vs* one antagonist, more complex patterns can be identified: the same metabolites can affect more than one antagonist, while two or more metabolites can act synergically on the same one, either acting on the same biochemical mechanism or on multiple fronts (Kannan and Sureendar, 2009).

In a previous work, a group of *Pseudomonas protegens* related strains were isolated from hydroponic cultures of lamb's lettuce, for their ability to inhibit selected fungal pathogens (Moruzzi *et al.*, 2017). The aim of this work was to investigate the biological activity of two of those strains against a larger number of fungal strains, and correlate it with their genomic features, especially those related to secondary metabolism.

2.2.2 Materials and methods

***Pseudomonas* genomes**: the *Pseudomonas* sp. Pf-4 and Pf-11 strains were isolated in 2009 from the roots of healthy *Valerianella locusta* (L.) Laterr. plants grown in a hydroponic farm in Friuli Venezia Giulia (FVG) region, north-eastern Italy. Genomic DNA was extracted from 1 ml of 24 hrs old cultures grown in Nutrient Broth with agitation using a Wizard DNA purification kit (Promega Italia, Padova, Italy) following the manufacturer's instructions. DNA was measured and checked for quality using a NanoDrop (NanoDrop products, Wilmington, DE, USA). Illumina libraries were prepared as described previously (Scortichini *et al.*, 2013) and sent to the Istituto di Genomica Applicata (Udine, Italy) for sequencing on a Illumina Myseq with a 2x300 Reagent kit. Genomes were assembled using the A5-miseq pipeline (Tritt *et al.*, 2012) and annotated using the NCBI Prokaryotic Genome Annotation Pipeline (http://www.ncbi.nlm.nih.gov/genome/annotation_prok/) and the RAST server (Aziz *et al.*, 2008). The BioSample accession for Pf-4 is SAMN04554942; the BioSample accession for Pf-11 is SAMN04554943. A preliminary genome draft of Pf-4 has been published (Moruzzi *et al.*, 2017), while no genomic information about Pf-11 has been previously reported.

Orthologs and metabolic pathways: the NCBI-annotated sequences of Pf-11 and Pf-4 were compared for orthologs using the standalone version of the *Orthologous Matrix* tool (OMA; http://omabrowser.org/standalone/). The output was converted into a comparison table, using a custom Perl/Bash script.

Secondary metabolic pathways were investigated by verifying the presence of a selection of genes and gene clusters typical of *P. protegens* strains (Mavrodi *et al.*, 2010; Loper *et al.*, 2012; Blankenfeldt and Parsons, 2014) in the NCBI annotation. The two genomes were also submitted to the *antiSMASH* 3.0 tool for rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences (Weber *et al.*, 2015), for comparison with the hand-selected cluster findings. Structural and functional features of the two genomes were compared using Mauve (Darling, 2004), Busco (Simão *et al.*, 2015), and some *ad hoc* Perl scripts.

Fungal growth inhibition tests: to test the ability of both *Pseudomonas* strains to inhibit fungal growth, a total of 18 fungal strains were assayed in four separate inhibition tests. Most strains were freshly isolated identified by rDNA sequencing as belonging to the following species: *Alternaria alternata, Aspergillus flavus, A. niger, Fusarium oxysporum* f. sp. *niveum, F. oxysporum* f. sp. *vasinfectum, F. verticillioides, Penicillium chrysogenum, P. griseofulvum, P. verrucosum; Ilyonectria europaea, I. robusta, Epicoccum nigrum, Neopestalotiopsis rosae; Phoma betae, Botritis cinerea, Colletotrichum* sp. In addition, *Pythium aphanidermatum* strain CBS 118745 and strain CBS 116664, obtained from the Centraal Bureau voor de Schimmelcultures (CBS) and a two *Rhizoctonia solani* strains TR15 and TP20 isolated in 2009 from symptomatic plants in the same hydroponic farm as the *Pseudomonas* strains, were used.

Inhibition tests: nine repetitions of each fungus were placed in the center of standard Petri dishes (internal diameter 85 mm) containing PDA supplemented with 3 g/l peptone and 2 g/l yeast estract, three of which were streaked at the sides, in a square shape, with Pf-4 and another three with Pf-11; the last three were the control samples, with the fungus alone. The Petri dishes were incubated at room temperature in a dim-lit environment. The dishes were photographed about every 24 hours, for at least 10 days, and the diameter of each fungal colony was recorded (for the early square shapes, an average diameter was noted). Statistical analysis was carried out by comparing modeled growth curves using the “nlme” package (Pinheiro *et al.*, 2017) of R (R Development Core Team, 2007).

2.2.3 Results

Genome sequencing: the ILLUMINA Sequencing of strain Pf-11 DNA produced 3,727,137 reads, 300 nts each, for a total of $1.1 \cdot 10^9$ nucleotides, while sequencing of Pf-4 DNA produced 1,914,469 reads, 300 nts each, for a total of $0.6 \cdot 10^9$ bp. The assembled Pf-11 draft genome sequence is 7,053,517 bp long in total, with a 62.0% G+C content, and consists of 125 contigs ranging from 605 to 1,372,031 bp in size (N50: 1,036,338), with a coverage of 205.3×. The assembled Pf-4 draft genome sequence is 6,832,152 bp long in total, with a 62.5% G+C content, and consists of 36 contigs ranging from 605 to 1,018,138 bp in size (N50: 688,889), with a coverage of 100.9×.

The genome sequence draft of Pf-11 contains 6154 predicted protein-coding sequences and 135 pseudogenes. In addition, 63 tRNA genes and 11 rRNA genes were identified. The genome sequence of Pf-4 contains 5907 predicted protein-coding sequences and 61 pseudogenes. In addition, 62 tRNA genes and 11 rRNA genes were identified.

Comparison of the Pf-4 and Pf-11 genomes: strains Pf-4 and Pf-11 are very similar to each other. Their 16S rRNA gene sequences are almost identical (1 nt difference in the entire sequence). They are also very similar at the genome level, sharing a large number of orthologous genes. OMA found 5534 orthologs, representing 89.9% of the Pf-11 genome and 93.7% of the Pf-4 genome. We selected the predicted protein sequences of 437 orthologs that are highly conserved among the gammaproteobacteria according to the BUSCO data-set (Simão *et al.*, 2015), and estimated that the identical sequences were 261, while 100 had a single amino acid difference. In the total 153,633 amino acid residues resulting from concatenation of the 437 core protein sequences only 403 (0.26%) were different between Pf-4 and Pf-11. Differences among the two genomes were mostly found in the accessory genome; the OMA program listed 427 additional genes exclusive to Pf-4 (Table 2.2.1 and 6.1.1) and 741 exclusive to Pf-11 (Table 2.2.2 and 6.1.2).

Table 2.2.1 – Highlights of OMA-isolated genes exclusive to Pf-4. An exhaustive list is present in the supplemental data, Table 6.1.1.

Gene code	Description
A1348_00295	iron dicitrate transport regulator FecR
A1348_01100, A1348_12715	antitoxin
A1348_01105	plasmid maintenance protein
A1348_05325	immunity protein
A1348_06835	organic radical-activating protein
A1348_07960	nitrilotriacetate monooxygenase
A1348_08840	RTX toxin
A1348_11565	ferredoxin
A1348_12720	addiction module toxin RelE
A1348_12950	pesticin immunity protein
A1348_15975	p-hydroxycinnamoyl CoA hydratase/lyase
A1348_16040	Vanillate O-demethylase oxidoreductase
A1348_16045	Rieske (2Fe-2S) protein
A1348_16275	nucleoid-associated protein YejK
A1348_17340	antibiotic ABC transporter permease
A1348_21675	nitric oxide synthase
A1348_22195	agmatine deiminase
A1348_25120	Holliday junction resolvase
A1348_25380	plasmid stabilization protein ParE
A1348_27055	flavin reductase
A1348_27075	monodechloroaminopyrrolnitrin synthase PrnB
A1348_29440	chemotaxis protein
A1348_30105, A1348_12615	large adhesive protein

Table 2.2.2 – Highlights of OMA-isolated genes exclusive to Pf-11. An exhaustive list is present in the supplemental data, Table 6.1.2.

Gene code	Description
A1395_07930	methanobactin biosynthesis cassette protein MbnB
A1395_08725	acriflavin resistance protein
A1395_08815, A1395_21000	DNA repair protein RadC
A1395_09195	multidrug transporter
A1395_09200	multidrug efflux RND transporter permease subunit
A1395_16665	phenol degradation protein meta
A1395_17490, A1395_29460	addiction module toxin RelE
A1395_17495	toxin-antitoxin system protein
A1395_17640	cytotoxic translational repressor of toxin-antitoxin stability system
A1395_20785	coproporphyrinogen III oxidase
A1395_21095	nitrilase
A1395_22460	metal-chelation protein CHAD
A1395_25370	prevent-host-death protein
A1395_27990	host specificity protein
A1395_28685, A1395_28745	lysozyme
A1395_29380	large adhesive protein
A1395_29465	antitoxin of toxin-antitoxin stability system
A1395_31465	spermidine/putrescine ABC transporter substrate-binding protein
A1395_31485	nitrate reductase
A1395_31570	SfnB family sulfur acquisition oxidoreductase
nusA	transcription termination/antitermination protein NusA

By using the complete sequence of Pf-5 as a reference (CP0000076.1; Paulsen *et al.*, 2005), we carried out the scaffolding of 16 contigs of Pf-4 (accounting for 6,802,786 nts, which correspond to >99.5% of the Pf-4 estimated genome size) on one hand and, on the other, scaffolding of 15 contigs of Pf-11 (accounting for 6,934,975 nts, which correspond to >98.3% of the Pf-11 estimated genome size). According to the genome alignment carried out with Mauve (shown in Figure 2.2.1), the two genomes have a strong colinearity and conservation. However, the alignment highlighted 96 sequence regions (larger than 1,000 nts) in the genome of Pf-4 that were missing in Pf-11 (460,862 nts total), and 68 in the genome of Pf-11 that were missing in Pf-4 (600,403 nts total). The 8 unique regions in Pf-4 and the 11 regions in Pf-11 that are larger than 10,000 nts are marked with a red dot in the genome alignment of Figure 2.2.1. Noticeably, in the

Pf-4 genome, region 4 in scaffold 1 and region 5 in scaffold 2, as well as the region of poor similarity and rearrangements located around and including region 3 in scaffold 4, contain many genes involved in secondary metabolism. Moreover, several strain specific polyketide synthase (PKS) could be located around and in region 1 in scaffold 8. Conversely, in the Pf-11 genome, only the large region 1 in scaffold 1 was rich in genes involved in secondary metabolism.

On the whole, the genome of Pf-11 was about 200 kb (3%) larger than the genome of Pf-4, the difference being related with a larger accessory genome. The count of genes annotated as conjugative protein and transposase sums 43 in Pf-11 and only 4 in Pf-4, suggests that the presence of mobile elements is more extensive in Pf-11.

Genes involved in the production of secondary metabolites: in a comparison for their potential in the production of secondary metabolites, the two genomes resulted similar, yet with some significant differences. A summary of the comparison of the secondary metabolite gene content of the two strains is given in Table 2.2.3, and reported in full detail in Table 6.1.3.

Nine gene clusters for antibiotic metabolites typical of *P. protegens* were found in both Pf-4 and Pf-11 strains, along with *gac*/*rsm* homologues and small regulatory RNAs: hydrogen cyanide (*hcn*), 2,4-diacetylphloroglucinol (*phl*), AprX protease (*apr*), pyoverdine (*pvd*), enantio-pyochelin (*pch*), hemophore biosynthesis (*has*), ferric-enterobactin receptor (*pfe*), orfamide A (*ofa*), and FitD toxin (*fit*). Three clusters, *i.e.* pyoluteorin (*plt*), pyrrolnitrin (*prn*), and rhizoxin (*rzx*), were present only in Pf-4.

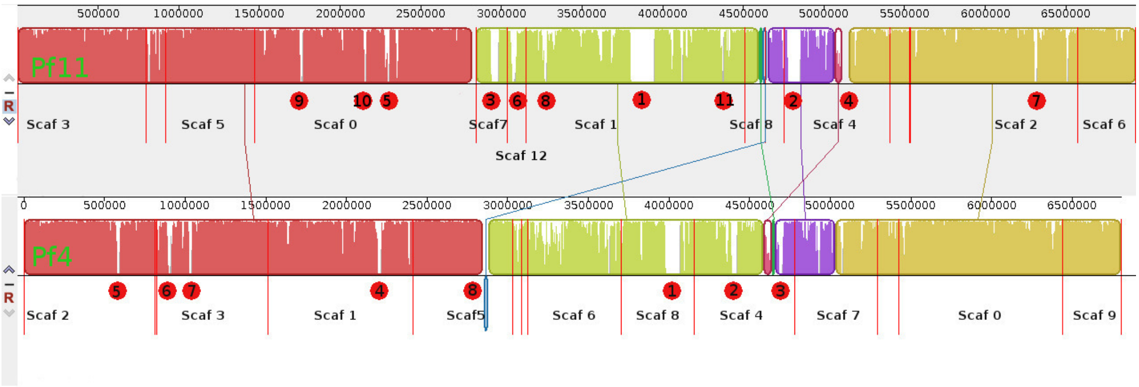


Figure 2.2.1 – Mauve alignment of Pf-11 and Pf-4. The red dots mark unique regions larger than 10,000 nts.

Table 2.2.3 – Summary of the sequence analysis of gene clusters for the synthesis of antibiotics, exoenzyme, cyclic lipopeptide, siderophores, and toxin, and of Gac/Rsm homologues in P. protegens Pf-11 and Pf-4 and similarities to those in P. protegens strains (Pf-5, PH1b) and other closely related Pseudomonas sp. (CMAA1215, NFPP17, Os17). A more detailed list of genes is present in Table 6.1.3.

Pf-11 NCBI Gene ID (A1395_)	**Pf-4 NCBI Gene ID (A1348_)**	**Gene name (Pf5 equiv. PFL ID)**	**Pf-11 Locus**	**Pf-4 Locus**
hcn **gene cluster (for hydrogen cyanide) – present in both**				
10425–10415	23065–23075	*hcnA* (2577)–*hcnC* (2579)	1: 991726–994695 (–)	6: 391003–393972 (+)
plt **gene cluster (for pyoluteorin) – only in Pf-4**				
–	17270–17350	*pltM* (2784)–*pltP* (2800)	–	4: 360091–390616
prn **gene cluster (for pyrrolnitrin) – only in Pf-4**				
–	27080–27065	*prnA* (3604)–*prnD* (3607)	–	8: 326813–332375
phl **gene cluster (for 2,4-diacetylphloroglucinol) – present in both**				
18635–18670	10485–10520	*phlH* (5951)–*phlE* (5958)	3: 364619–372851	2: 363678–371910
apr **gene cluster (for AprX protease) – present in both**				
08470–08450	26990–26970	*aprA* (3210)–*aprF* (3206)	1: 533253–539877 (–)	8: 303649–310279 (–)
Gac/Rsm homologues – present in both				
13645	03275	*gacS* (4451)	2: 326117–328870 (+)	0: 690217–692970 (–)
21170	25980	*gacA* (3563)	4: 104938–105522 (–)	7: 486282–486866 (+)
13900	03020	*rsmA* (4504)	2: 377278–377466 (–)	0: 641626–641814 (+)
17930	09780	*rsmE* (2095)	3: 220271–220990 (+)	2: 219078–219797 (+)
24025	15270	*retS* (0664)	5: 78482–81268 (+)	3: 607391–610177 (–)
26950	28385	*ladS* (5426)	6: 187267–189633 (–)	9: 172345–174711 (+)
Small regulatory RNAs – present in both				
N.A.	N.A.	*rsmZ* (6285)	0: 514076–513951 (–)	1: 506535–506661 (+)
N.A.	N.A.	*rsmY* (6291)	3: 74313–74197 (–)	2: 73788–73906 (+)
N.A.	N.A.	*rsmX* (6289)	10: 33390–33506 (+)	10:86797–86915 (+)

***pvd* gene cluster (for pyoverdine) – present in both**				
07080–07085	17855–17860	*pvdQ* (2902)–*fpvR* (2903)	1: 189376–192763	4: 506592–509979
30155–30060	29340–29435	*pvdA* (4079)–PFL_4097	10: 46820–92493	10: 26184–71830
12240–12290	04660–04610	PFL_4169–*pvdH* (4179)	2: 17612–26974	10: 56263–59334 (–) 0: 990920–999310
12360–12370	04555–04545	*pvdL* (4189)–*pvdY* (4191)	2: 41461–55794	0: 962639–976972
***pch* cluster (for enantio-pyochelin) – present in both**				
30475–30520	15840–15885	*pchR* (3497)–*pchA* (3488)	11: 53981–72965	4: 49492–68476
***has* gene cluster (for hemophore biosynthesis) – present in both**				
26720–26690	28615–28645	*hasI* (5380)–*hasF* (5374)	6: 128190–138010 (–)	9: 223960–233779 (+)
***pfe* gene cluster (for ferric-enterobactin receptor) – present in both**				
10085–10095	23430–23420	*pfeR* (2665)–*pfeA* (2663)	1: 916810–921183 (+)	6: 470135–474508 (–)
***ofa* gene cluster (for orfamide A) – present in both**				
27845–27835	18430–18420	*ofaA* (2145)–*ofaC* (2147)	7: 7700–42217 (–)	5: 7709–42188 (–)
***fit* gene cluster (for FitD toxin) – present in both**				
08015–07980	26560–26525	*fitA* (2980)–*fitH* (2987)	1: 402656–424286	8: 180030–201661
***rzx* gene cluster (for rhizoxin) – only in Pf-4**				
–	26520–26475	PFL_2988–*rzxA* (2997)	–	8: 99945–179906

Genes in the *hcn* cluster showed high similarity (between 91% and 98%) to those of *P. protegens* strain Pf-5 in the case of Pf-11, while in the case of Pf-4 the best matches were to those of strains *P.* sp. Os17 and St29 (95%–99%); similarly, the genes in the Pf-11 *phl* cluster have high similarity (92%–98%) to those of *P. protegens* strain Pf-5 and to those of *P.* sp. Os17, and St29 and *P. protegens* strains in the case of Pf-4 (90%–99%).

In both Pf-11 and Pf-4, high similarity to the PH1b, CMAA1215 and Os17 strains was found for the *apr* cluster (92–99%) and to Pf-5 for the cluster associated with the *gac*/*rsm* signal transduction (91–100%).The *pvd* cluster for pyoverdine, whose product has been reported in Pf-5 (Gross and Loper, 2009), is divided in four different loci, with varying levels of similarity; the largest locus spans genes A1395_30060–A1395_30155, with similarity ranging between 31% and 96%, in Pf-11, and (NCBI ID) A1348_29340–A1348_29435, with similarity ranging between 35% and 100% in Pf-4.

Clusters for enantio-pyochelin, fully conserved in Pf-5 (Youard *et al.*, 2007), hemophore biosynthesis, ferric-enterobactin receptor and orfamide A were also found in both strains; the *fit* cluster (Péchy-Tarr *et al.*, 2008), in the downstream region of the *rzx* cluster in Pf-5, has 88–97% identity in both cases to *P. protegens* Pf-5 homologous, and appears inverted compared to *P. protegens* strain Pf-5 and *P.* sp. Os17.

Differently from Pf-4, the *plt* cluster for pyoluteorin and the *prn* cluster for pyrrolnitrin, typical antibiotic metabolites in *P. protegens*, as well as the *rzx* gene cluster encoding analogs of the antimitotic macrolide rhizoxin, are not present in Pf-11.

For comparison, the *antiSMASH* tool found 6 metabolic pathways common to Pf-11 and Pf-4 (amychelin, arylpolyene, 2,4-diacetylphloroglucinol, mitomycin, orfamide, pyoverdine), 1 exclusive to Pf-11 (alginate) and 3 exclusive to Pf-4 (pyoluteorin, pyrrolnitrin, rhizoxins), as reported in Table 2.2.4 and Table 2.2.5. The enantio-pyochelin biosynthesis cluster, however, was not detected.

Fungal growth inhibition tests: strains of all fungal species but *Pythium* and *Rhizoctonia* took at least 9 days to reach the plate border, therefore growth curves were constructed with data of 8 days of growth. The diameter of the fungal colonies grown for 10 days in the presence of strain Pf-11 ranged from 22 mm (*P. verrucosum*) to 57 mm (*E. nigrum)*, while those grown in the presence of Pf-4 ranged from 15 mm (*N. rosae*) to 53 mm (*A. niger*).

Table 2.2.4 – Gene clusters in Pf-11 determined by *antiSMASH* 3.0 web tool. Under the "Most similar known cluster" column, the percentage is the proportion of genes that show similarity.

Cluster	**Type**	**From**	**To**	**Most similar known cluster**	**MIBiG BGC-ID**
scaffold 0					
Cluster 1	T1pks	269602	315724	Alginate biosynthetic g.c.* (100%)	BGC0000726 c1
Cluster 2	Bacteriocin	604419	615309	-	-
Cluster 3	Bacteriocin	643914	655860	-	-
scaffold 1					
Cluster 4	Thiopeptide-Lantipeptide-Bacteriocin	60239	100470	-	-
Cluster 5	Bacteriocin	387369	398181	-	-
Cluster 6	Nrps	1101588	1153056	Mitomycin biosynthetic g.c.* (5%)	BGC0000915 c1
scaffold 10					

Cluster 7	Nrps	28880	92959	Pyoverdine biosynthetic g.c.* (31%)	BGC0000413 c1
scaffold 11					
Cluster 8	Nrps	35259	89304	Amychelin biosynthetic g.c.* (12%)	BGC0000300 c1
scaffold 2					
Cluster 9	Nrps	21461	74477	Pyoverdine biosynthetic g.c.* (16%)	BGC0000413 c1
Cluster 10	Nrps	503749	554778	-	-
scaffold 3					
Cluster 11	T3pks	350414	391463	2,4-Diacetylphloroglucinol biosynthetic g.c.* (87%)	BGC0000281 c1
Cluster 12	Bacteriocin	629699	640544	-	-
scaffold 5					
Cluster 13	Arylpolyene	294021	337638	APE Vf biosynthetic g.c.* (40%)	BGC0000837 c1
scaffold 7					
Cluster 14	Nrps	1	62217	Orfamide biosynthetic g.c.* (70%)	BGC0000399 c1

*g.c.: "gene cluster".

For all fungi except *A. niger* and *A. flavus*, a statistically significant inhibition effect was observed for both Pf4 and Pf11, starting from the 4th day post inoculum (DPI) (Figure 2.2.2).

In a first group of fungi (*E. nigrum, Colletotrichum* sp.*, A. alternata, I. robusta, P. betae, P. verrucosum*), the inhibition effect from Pf4 was more intense than Pf11 and both were more intense than the controls. Significant difference from the controls was determined for all these fungi at the 2nd DPI, while the effect difference between Pf4 and Pf11 ranged from the 2nd (*E. nigrum*) to the 3rd (*Collectricum* sp., *A. alternata*), 5th (*I. robusta*) or 9th DPI (*P. verrucosum*).

Table 2.2.5 – Gene clusters in Pf-4 determined by *antiSMASH* 3.0 web tool. Under the "Most similar known cluster" column, the percentage is the proportion of genes that show similarity.

Cluster	**Type**	**From**	**To**	**Most similar known cluster**	**MIBiG BGC-ID**
scaffold 0					
Cluster 1	Nrps	464199	515228	-	-
Cluster 2	Nrps	943956	996972	Pyoverdine biosynthetic g.c.* (17%)	BGC0000413 c1
scaffold 1					
Cluster 3	Bacteriocin	597423	608313	-	-
Cluster 4	Bacteriocin	636903	648849	-	-
scaffold 10					
Cluster 5	Nrps	25701	89768	Pyoverdine biosynthetic g.c.* (27%)	BGC0000413 c1
scaffold 2					
Cluster 6	T3pks	349473	390522	2,4-Diacetylphloroglucinol biosynthetic g.c.* (87%)	BGC0000281 c1

Cluster 7	Bacteriocin	644287	655132	-	-
scaffold 3					
Cluster 8	Arylpolyene	350533	394150	APE Vf biosynthetic g.c.* (40%)	BGC0000837 c1
scaffold 4					
Cluster 9	Nrps	30770	84815	Amychelin biosynthetic g.c.* (12%)	BGC0000300 c1
Cluster 10	T1pks	344776	397525	Pyoluteorin biosynthetic g.c.* (100%)	BGC0000127 c1
Cluster 11	Lantipeptide-Bacteriocin	396010	420165	-	-
scaffold 5					
Cluster 12	Nrps	1	62188	Orfamide biosynthetic g.c.* (70%)	BGC0000399 c1
scaffold 6					
Cluster 13	Nrps	230297	281765	Mitomycin biosynthetic g.c.* (3%)	BGC0000915 c1
scaffold 8					
Cluster 14	Transatpks	79945	198849	Rhizoxins biosynthetic g.c.* (12%)	BGC0001112 c1
Cluster 15	Other	309674	350759	Pyrrolnitrin biosynthetic g.c.* (100%)	BGC0000924 c1

*g.c.: “gene cluster”.

In a second group of fungi (*F. verticillioides, F. oxysporum* f. sp. *vasinfectum, F. oxysporum* f. sp. *niveum, N. rosae, B.cinerea, P. chrysogenum and I.europaea*) Pf4 and Pf11 affect the growth rate of the fungi if compared to the controls, but there was no difference in effect between Pf4 and Pf11 themselves.

In *A. niger,* the presence of Pf4 or Pf11 enhance, rather than inhibit, the fungus growth rate, the difference becoming statistically different from the controls at the 7th DPI. No significant difference was observed between the Pf4 and Pf11 theses.

Finally, in *A. flavus,* no statistical difference was found between Pf4 and Pf11, and between these and the controls.

2.2.4 Discussion

A full understanding of the dynamics and composition of the microbial communities in soil is of paramount relevance for the establishment of biological control strategies against fungal pathogens. The conventional approach relies on *in vitro* screening of potential biocontrol agents by evaluation of their ability to inhibit pathogen growth. As shown in this report, the bacteria selected for the strongest ability to inhibit pathogens are characterized by the production of the

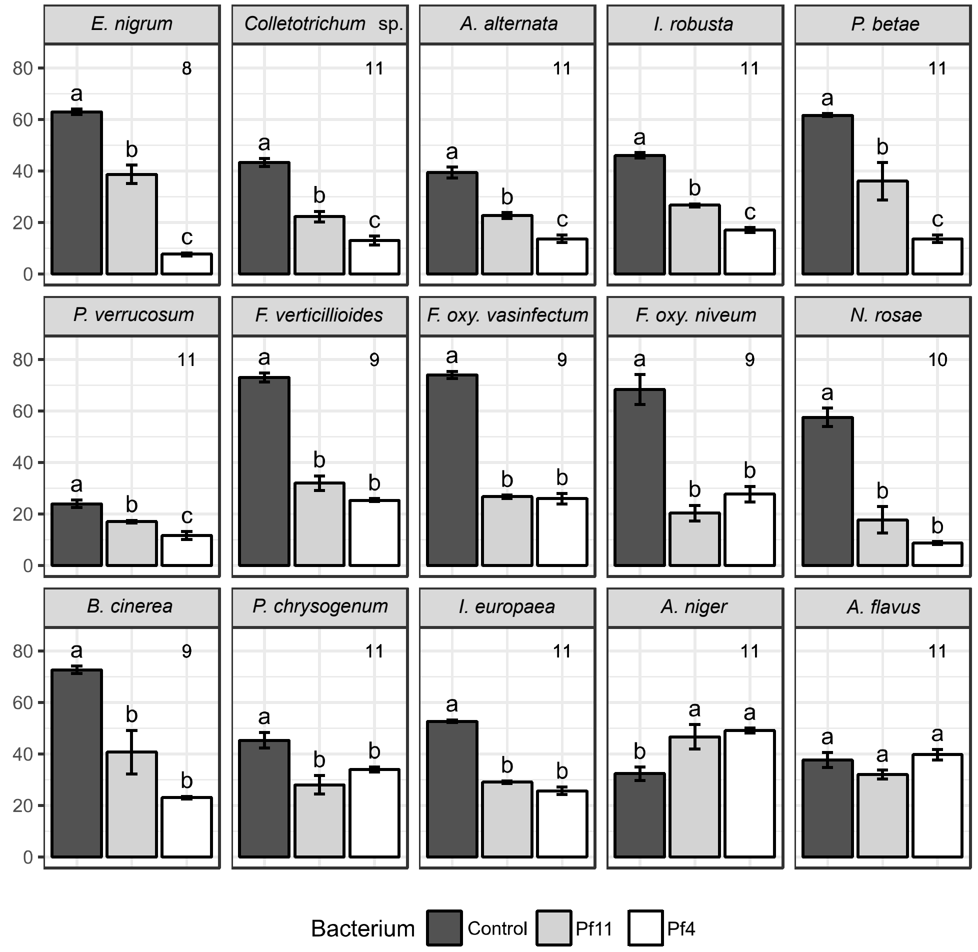


Figure 2.2.2 – Results of the statistical tests of the inhibition assays. The bars represent the diameters means for Control, Pf11, Pf4 replicas observed at the last date relevant for statistical analysis (*DPIend*). Error bars are the standard deviation, while different letter indicate different means based on post hoc Tukey test at 0.01 level of significance. The number at the top right of each graph specify the *DPIend*.

largest array of metabolite and a wider (broader) activity against a variety of fungal species: a larger set of metabolites can allow a wider spectrum of biocontrol activity, and/or a stronger control towards the same competitor, by exploiting different strategies simultaneously, possibly resulting in a synergistic effect.

Although the strains producing a wider range of secondary metabolites may result the most effective in restricting pathogen growth, it remain to be established whether or not their use is the most profitable choice when aiming at a durable protection. The isolation from the same environment of strains that are taxonomically strictly related, yet significantly different in their interaction with other microorganisms, suggested that environmental demand for within species diversity may grant to seemingly less fit bacterial strains, with narrower metabolite production patterns, an opportunity to survive along with more competitive ones. It is tempting to speculate that while the strong fungal growth inhibitory activity of Pf4 on one hand advocates a role of deployment against fungal pathogens, on the other it has the potential of significantly alter the equilibrium in the microorganism community, causing a comparably strong response that might associate with a simplification of the community and eventual decline of the Pf4 population itself due to a lower ability to adapt to changing conditions. Under this view, a less impactful bacterium like Pf11, producing a more limited array of metabolites, might provide the conditions for a more diverse microbial community.

The results of this work highlight the contrast between a classification based on taxonomic markers and one based on ecological roles; species that may appear homogeneous on a taxonomical level might on the contrary present a high level of heterogeneity in terms of interactions with other microorganisms. A dynamic equilibrium among different strains comprised in the same species, *i.e.* those that allow the maximum exploitation of competitive feature based on secondary metabolites and those that preserve a more complex microbial community, may be functional to the evolutionary success of the species.

An effective analysis of microbial diversity in ecological complex system needs to take into account the concepts outlined above. Although barcoding using taxonomically informative genes such as ribosomal DNA is presently the most widely used approach to characterize complex microbial communities, it severely underestimates the diversity of the communities. Indeed, it would interesting to verify whether bacteria that are indistinguishable using rDNA and other gene markers, but differ significantly in the genetic features that control interaction with other microorganisms, can coexist in the same environment. When referring to *P. protegens*, a species comprising strains that have a significant production of bio-active secondary metabolites, the intra-species diversity and variability may play a major role in determining the composition of the microbial community.

The cohabitation of different strains that are strictly taxonomically related and share a prevalent fraction of their genomes, yet with significantly different secondary metabolite profiles, is functional to their continued coevolution. A continuous trade of horizontally transferred genetic material needs to be fueled with new genetic information and, to this end, the intra-species diversity plays an instrumental role. According to the results presented in this paper, the production of pyoluteorin, pyrrolnitrin, and rhizoxin, typical antibiotic metabolites in *P. protegens,* is relevant in determining the fungal growth inhibition pattern of competitive *P. protegens* strains. Clusters *prn, rzx* and *plt* were found only in Pf4 scaffolds 8 and 4, along with two regions on scaffolds 1 and 5 coding for secondary metabolites; therefore the difference between the two strains can't be attributed to a simple insertion event, but implies a relatively complex differentiation focused on the accessory genomes. In contrast to the core genome, the evolution in the accessory genomes progresses exploiting primarily horizontal gene transfer consequent to multiple invasion of foreign DNA, that could be more or less stably integrated in the genome. *P. protegens* have the largest genomes among the bacteria of the fluorescens group and in general among *Pseudomonas* (whose genomes range from 4.17 to 8.6 Mb). Conceivably, larger genomes allow to accommodate metabolic gene clusters conferring environmental fitness advantages that compensate for the relax of an otherwise strict genome size constrains. In particular, with 7.05 Mb, Pf11 stands at the high range of the genomes size allowed for the species. It can be speculated that genome expansion (with horizontal acquisition of genes) and contraction is a dynamic process that lead to a more stable genome. In this view, the Pf11 genome, with its larger size and the larger number of transposable elements appears to be in more dynamic evolutionary stage as compared to Pf4 genome that already gained a richer pattern of secondary metabolite production associated gene clusters.

2.2.5 Acknowledgments

This research was supported by a grant (L. R. 26) from Friuli Venezia Giulia Region Administration (Italy) and by "Ager – Agroalimentare e Ricerca" Foundation, project "Novel strategies meeting the needs of the fresh-cut vegetable sector – STAYFRESH", nº 2370.

2.2.6 References

Aziz, R. K. *et al.* (2008) ‘The RAST Server: Rapid Annotations using Subsystems Technology’, *BMC Genomics*, 9(1), p. 75. doi: 10.1186/1471-2164-9-75.

Blankenfeldt, W. and Parsons, J. F. (2014) ‘The structural biology of phenazine biosynthesis’, *Current Opinion in Structural Biology*. Elsevier Ltd, 29, pp. 26–33. doi: 10.1016/j.sbi.2014.08.013.

Christensen, B. B. *et al.* (2002) ‘Metabolic Commensalism and Competition in a Two-Species Microbial Consortium’, *Applied and Environmental Microbiology*, 68(5), pp. 2495–2502. doi: 10.1128/AEM.68.5.2495-2502.2002.

Colla, P., Gilardi, G. and Gullino, M. L. (2012) ‘A review and critical analysis of the European situation of soilborne disease management in the vegetable sector’, *Phytoparasitica*, 40(5), pp. 515–523. doi: 10.1007/s12600-012-0252-2.

Couillerot, O. *et al.* (2009) ‘Pseudomonas fluorescens and closely-related fluorescent pseudomonads as biocontrol agents of soil-borne phytopathogens’, *Letters in Applied Microbiology*, 48(5), pp. 505–512. doi: 10.1111/j.1472-765X.2009.02566.x.

Darling, A. C. E. (2004) ‘Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements’, *Genome Research*, 14(7), pp. 1394–1403. doi: 10.1101/gr.2289704.

Gross, H. and Loper, J. E. (2009) ‘Genomics of secondary metabolite production by Pseudomonas spp.’, *Natural Product Reports*, 26(11), p. 1408. doi: 10.1039/b817075b.

Handelsman, J. and Stabb, E. V (1996) ‘Biocontrol of soilborne plant pathogens’, *Plant Cell*, 8, pp. 1855–1869.

Hense, B. A. *et al.* (2007) ‘Does efficiency sensing unify diffusion and quorum sensing?’, *Nature reviews. Microbiology*, 5(3), pp. 230–9. doi: 10.1038/nrmicro1600.

Hibbing, M. E. *et al.* (2010) ‘Bacterial competition: surviving and thriving in the microbial jungle.’, *Nature reviews. Microbiology*. NIH Public Access, 8(1), pp. 15–25. doi: 10.1038/nrmicro2259.

Johnson-Henry, K. C. *et al.* (2007) ‘Surface-layer protein extracts from Lactobacillus helveticus inhibit enterohaemorrhagic Escherichia coli O157:H7 adhesion to epithelial cells.’, *Cellular microbiology*, 9(2), pp. 356–67. doi: 10.1111/j.1462-5822.2006.00791.x.

Kannan, V. and Sureendar, R. (2009) ‘Synergistic effect of beneficial rhizosphere microflora in biocontrol and plant growth promotion’, *Journal of Basic Microbiology*. WILEY-VCH Verlag, 49(2), pp. 158–164. doi: 10.1002/jobm.200800011.

Khan, A. *et al.* (2006) ‘Differential cross-utilization of heterologous siderophores by nodule bacteria of Cajanus cajan and its possible role in growth under iron-limited conditions’, *Applied Soil Ecology*, 34(1), pp. 19–26. doi: 10.1016/j.apsoil.2005.12.001.

Loper, J. E. *et al.* (2012) ‘Comparative Genomics of Plant-Associated Pseudomonas spp.: Insights into Diversity and Inheritance of Traits Involved in Multitrophic Interactions’, *PLoS Genetics*. Edited by D. S. Guttman, 8(7), p. e1002784. doi: 10.1371/journal.pgen.1002784.

Lyon, G. J. and Novick, R. P. (2004) ‘Peptide signaling in Staphylococcus aureus and other Gram-positive bacteria.’, *Peptides*, 25(9), pp. 1389–403. doi: 10.1016/j.peptides.2003.11.026.

Mavrodi, D. V. *et al.* (2010) ‘Diversity and Evolution of the Phenazine Biosynthesis Pathway’, *Applied and Environmental Microbiology*, 76(3), pp. 866–879. doi: 10.1128/AEM.02009-09.

Mazzola, M. (2002) ‘Mechanisms of natural soil suppressiveness to soilborne diseases’, *Antonie van Leeuwenhoek*, 81(1/4), pp. 557–564. doi: 10.1023/A:1020557523557.

Moruzzi, S. *et al.* (2017) ‘Genomic-assisted characterisation of Pseudomonas sp. strain Pf4, a potential biocontrol agent in hydroponics’, *Biocontrol Science and Technology*, 27(8), pp. 969–991. doi: 10.1080/09583157.2017.1368454.

Nicholson, A. (1954) ‘An outline of the dynamics of animal populations.’, *Australian Journal of Zoology*, 2(1), p. 9. doi: 10.1071/ZO9540009.

Van Os, E. A. (1999) ‘Closed soilless growing systems: A sustainable solution for Dutch greenhouse horticulture’, *Water Science and Technology*, 39(5), pp. 105–112. doi: 10.1016/S0273-1223(99)00091-8.

Paulitz, T. C. and Bélanger, R. R. (2001) ‘Biological control in greenhouse systems’, *Annual Review of Phytopathology*, 39(1), pp. 103–133. doi: 10.1146/annurev.phyto.39.1.103.

Paulsen, I. T. *et al.* (2005) ‘Complete genome sequence of the plant commensal Pseudomonas fluorescens Pf-5’, *Nature Biotechnology*, 23(7), pp. 873–878. doi: 10.1038/nbt1110.

Péchy-Tarr, M. *et al.* (2008) ‘Molecular analysis of a novel gene cluster encoding an insect toxin in plant-associated strains of Pseudomonas fluorescens’, *Environmental Microbiology*, 10(9), pp. 2368–2386. doi: 10.1111/j.1462-2920.2008.01662.x.

Pinheiro, J. *et al.* (2017) ‘nlme: Linear and Nonlinear Mixed Effects Models’. Available at: https://cran.r-project.org/package=nlme.

Raaijmakers, J. M. *et al.* (2009) 'The rhizosphere: a playground and battlefield for soilborne pathogens and beneficial microorganisms', *Plant and Soil*, 321(1–2), pp. 341–361. doi: 10.1007/s11104-008-9568-6.

R Development Core Team (2007) 'R: a Language and Environment for Statistical Computing'. Vienna, Austria: R Foundation for Statistical Computing. Available at: http://www.r-project.org.

Rankin, L. (1994) 'Evaluation of Rhizosphere Bacteria for Biological Control of Pythium Root Rot of Greenhouse Cucumbers in Hydroponic Culture', *Plant Disease*, 78(5), p. 447. doi: 10.1094/PD-78-0447.

Scortichini, M. *et al.* (2013) 'A Genomic Redefinition of Pseudomonas avellanae species', *PLoS ONE*. Edited by D. Arnold, 8(9), p. e75794. doi: 10.1371/journal.pone.0075794.

Simão, F. A. *et al.* (2015) 'BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs', *Bioinformatics*, 31(19), pp. 3210–3212. doi: 10.1093/bioinformatics/btv351.

Stubbendieck, R. M. and Straight, P. D. (2016) 'Multifaceted Interfaces of Bacterial Competition', *Journal of Bacteriology*. Edited by W. Margolin. American Society for Microbiology, 198(16), pp. 2145–2155. doi: 10.1128/JB.00275-16.

Tritt, A. *et al.* (2012) 'An Integrated Pipeline for de Novo Assembly of Microbial Genomes', *PLoS ONE*. Edited by D. Zhu, 7(9), p. e42304. doi: 10.1371/journal.pone.0042304.

Vallance, J. *et al.* (2011) 'Pathogenic and beneficial microorganisms in soilless culture', *Agronomy for Sustainable Development*, 31, pp. 191–203.

Weber, T. *et al.* (2015) 'antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters', *Nucleic Acids Research*. Oxford University Press, 43(W1), pp. W237–W243. doi: 10.1093/nar/gkv437.

Weller, D. M. (2007) 'Pseudomonas Biocontrol Agents of Soilborne Pathogens: Looking Back Over 30 Years', *Phytopathology*, 97(2), pp. 250–256. doi: 10.1094/PHYTO-97-2-0250.

Xavier, J. B. and Foster, K. R. (2007) 'Cooperation and conflict in microbial biofilms.', *Proceedings of the National Academy of Sciences of the United States of America*, 104(3), pp. 876–81. doi: 10.1073/pnas.0607651104.

Youard, Z. A. *et al.* (2007) 'Pseudomonas fluorescens CHA0 Produces Enantio-pyochelin, the Optical Antipode of the Pseudomonas aeruginosa Siderophore Pyochelin', *Journal of Biological Chemistry*, 282(49), pp. 35546–35553. doi: 10.1074/jbc.M707039200.

2.3 Genomic structural variations during clonal expansion of *Pseudomonas syringae* pv. *actinidiae* biovar 3 in Europe

Authors: Giuseppe Firrao[1-2], Emanuela Torelli[1], Cesare Polano[1], Patrizia Ferrante[3], Francesca Ferrini[1], Marta Martini[1], Marco Scortichini[3], Paolo Ermacora[1].

Manuscript submitted to "Frontiers in Microbiology".

[1] Dipartimento di Scienze Agroalimentari, Ambientali e Animali – Università di Udine – 33100 Udine, Italy

[2] Istituto Nazionale Biostrutture e Biosistemi, Italy

[3] Council for Agricultural Research and Analysis of Agricultural Economics (CREA), Research Centre for Olive, Fruit Trees and Citrus; Via di Fioranello 52, I-00134 Roma, Italy.

2.3.1 Summary

Pseudomonas syringae pv. *actinidiae* (Psa) biovar 3 has caused worldwide pandemic bacterial canker of *Actinidia chinensis* and *A. deliciosa* since 2008. In Europe, the disease spread rapidly in the kiwifruit cultivation areas from a single introduction from China. In this study, we investigated the genomic diversity of Psa biovar 3 strains during the primary clonal expansion in Europe using single molecule real-time (SMRT), Illumina and Sanger sequencing technologies. By comparing the genome sequences obtained from strains isolated from symptomatic kiwifruit tissue, we showed that despite the modest changes in terms of nucleotide polymorphysms, structural variations based on rearrangements of genetic elements occur frequently in the population of Psa biovar 3 undergoing clonal expansion in Europe. We recorded evidences of frequent mobilization and loss of transposon Tn6212, large chromosome inversions, and ectopic integration of IS sequences (remarkably ISPsy31 and ISPsy37) that, at least in one case, interrupted a pathogenesis related gene cluster and caused the loss in the ability to cause hypersensitivity reaction (HR) on tobacco and eggplant leaves. The evidence of gene loss in variant strains with reduced virulence in Europe is in streaking contrast with the emergence in New Zealand of copper resistant variant strains characterized by gene gain.

This divergence may be due to different environmental conditions or to the adoption of different strategies in the management of the epidemics.

2.3.2 Introduction

Pseudomonas syringae pv. actinidiae (Psa) is the causal agent of bacterial canker of green-fleshed (*Actinidia deliciosa*) and yellow-fleshed kiwifruit (*A. chinensis*) (Scortichini *et al.*, 2012). The pathogen was first isolated in Japan (Takikawa et al., 1989), where the disease was reported since 1984 and, subsequently, in Italy (Scortichini, 1994) and South Korea (Koh *et al.*, 1994). In the years 2008–2011, sudden and repeated epidemics of bacterial canker developed firstly in central Italy (Balestra *et al.*, 2009; Ferrante and Scortichini, 2009; Ferrante and Scortichini, 2010), and, subsequently, in all the other major areas of kiwifruit cultivation such as New Zealand (Everett *et al.*, 2011), and Chile (EPPO, 2016). In Europe, the epidemics spread to Portugal, France, Spain, Switzerland, Germany, Slovenia and Greece (Abelleira *et al.*, 2011; Cunty *et al.*, 2015b; Dreo *et al.*, 2014; EPPO, 2016; Holeva *et al.*, 2015).

Genomic and genetic analyses have soon revealed that the Psa strains causing the 2008–2011 epidemics differed significantly from those previously found in Italy (Marcelletti *et al.*, 2011) and that the first outbreaks of kiwifruit bacterial canker in Italy were caused by a rapid and clonal expansion of the pathogen in the cultivated areas (Marcelletti and Scortichini, 2011). Then, the availability of strains isolated in China, the area of origin of many *Actinidia* spp., and the intensive use of Illumina sequencing of bacterial genomes (Butler *et al.*, 2013; Mazzaglia *et al.*, 2012; McCann *et al.*, 2013; McCann *et al.*, 2017) and VNTR analysis (Ciarroni *et al.*, 2015; Cunty *et al.*, 2015a, Cesbron, *et al.*, 2015) paved the way to the understanding of the epidemiology of this important disease.

At present, Psa is subdivided into four biovars, three of which with distinct phylogeographic structure. Biovar 1 produces phaseolotoxin and has been isolated in Japan and Italy before 2008. Biovar 2 produces coronatine instead of phaseolotoxin and has been isolated only in South Korea. Biovar 3 produces neither phaseolotoxin nor coronatine and is responsible for the global outbreak of bacterial canker of kiwifruit in recent years. Biovar 5 does not produce phaseolotoxin nor coronatine, but unlike biovar 3 it is found only in the Saga Prefecture of Japan

(Fujikawa and Sawada, 2016). A fifth clade, initially identified as Psa biovar 4, has been recently described as a new pathovar, *P.s.* pv. *actinidifoliorum* (Cunty *et al.*, 2015b; Ferrante and Scortichini, 2015). Genome analysis performed so far is consistent with the hypothesis that all Psa biovars originated independently from a single natural source population and established subsequent outbreaks on cultivated kiwifruit. McCann *et al.* (2013) highlighted the overall clonal population structure with signatures of within-pathovar, intra-biovar recombination.

Psa biovar 3 is distinct from other biovars for the virulence and the sudden world-wide epidemic spread, that has unveiled major weakness of our kiwifruit cultivation system, while calling for efforts in the clarification of its dynamics in view of future prevention. Several genome-wide diversity studies revealed that epidemics in Europe, New Zealand and Chile of Psa biovar 3 originated from independent introductions of a single founder variant from China (Butler *et al.*, 2013; Mazzaglia *et al.*, 2012; McCann *et al.*, 2013) which, however, is not considered the center of origin of the biovar 3 (McCann et al., 2016).

In this work, we examined a sampling of the Psa population originated in Europe from the putative single introduction occurred in 2008. Through the analysis of Illumina sequence data-sets of 11 European and one non-European Psa genomes, and through the reconstruction and comparison of two complete genomes, a picture emerged that accounts for the significant differences in the pathways of genome evolution of this bacterium before and after the clonal expansion associated with the pandemic. DNA mobilization due to transposable elements was a major cause of structural differences and, at least in one case, resulted in the disruption of genes relevant in pathogen-host interaction, with a factual reduction of strain virulence on kiwifruit.

2.3.3 Results and discussion

Differential HR response of Psa CRAFRU 12.29 is due to insertional inactivation of the *hrp* gene cluster

Psa biovar 3 strains isolated in different regions of Europe were investigated to assess their phytopathogenic and genomic diversity. While most strains, as expected, induced HR in eggplant and tobacco leaves when infiltrated at concentrations of $1–2 \cdot 10^8$ cfu/ml, strains CRAFRU 12.50 and CRAFRU 12.29 failed in eliciting HR (not shown). Strains CRAFRU 12.50 and

CRAFRU 12.29 were also compared with the reference strain CRAFRU 8.43 in their ability to colonize *A. chinensis* leaves. Visual observations clearly revealed differences between CRAFRU 8.43 (HR+), that caused leaf spots, on one hand, and CRAFRU 12.29 (HR–) and CRAFRU 12.50 (HR–), on the other, that failed in inciting foliar symptoms. The estimate of bacterial concentration in leaves 22 days after inoculation, reported in Figure 2.3.1, showed that the population sizes of strain CRAFRU 12.29 and CRAFRU 12.50 did not increase during the assay time, while those of the virulent strain CRAFRU 8.43 peaked up to 100 times the inoculum. Thus, although the bacterial populations of HR– strains did not increase as much as the wild type, the bacteria remained detectable after 22 days. Further experiments carried out on micropropagated plantlets inoculated by dipping, revealed that the CRAFRU 12.29 cells move within the stem and were detectable by PCR in the stem segments above the point of inoculation 10 days after the dipping (results not shown).

Furthermore, a preliminary SNPs analysis, based on Illumina sequencing data, suggested that one of the HR– strains, namely CRAFRU 12.29, was highly similar, if not identical, to a HR+ strain, namely CRAFRU 14.08. Hence, the genome sequences of strains CRAFRU 14.08 and CRAFRU 12.29 were completed by SMRT (Single Molecule, Real Time) and Sanger sequencing. The resulting finished chromosomes, as shown in the alignment of Figure 2.3.2, differ for several structural features.

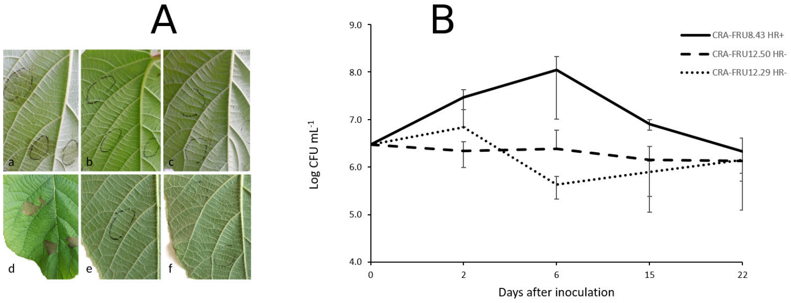


Figure 2.3.1 – A. Symptoms on kiwifruit leaves 2 days (a, b, c) and 15 days (d, e, f) after inoculation with CRAFRU 8.43 (a, d), CRAFRU 12.29 (b, e) and CRAFRU 12.50 (c, f). B. Population dynamics of Psa strains CRAFRU 8.43 (HR+), CRAFRU 12.29 (HR–) and CRAFRU 12.50 (HR–) after inoculation of kiwifruit leaves.

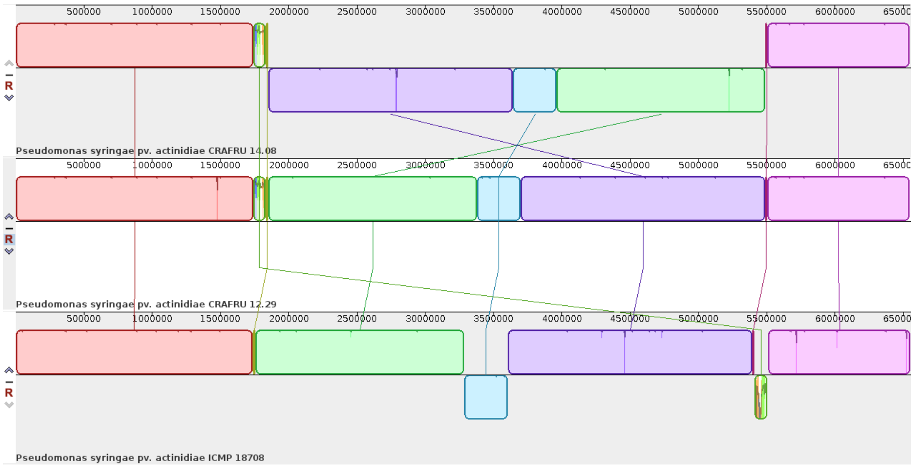

Pseudomonas syringae pv. actinidiae CRAFRU 14.08
Pseudomonas syringae pv. actinidiae CRAFRU 12.29
Pseudomonas syringae pv. actinidiae ICMP 18708
500000
1000000
1500000
2000000
2500000
3000000
3500000
4000000
4500000
5000000
5500000
6000000
6500
R

Figure 2.3.2 – Mauve alignment of the chromosomes of strains ICMP 18884, CRAFRU 14.08, and CRAFRU 12.29.

First of all, CRAFRU 14.08 displays a large inversion of about half of the chromosome (3,637,997 nts) as compared to CRAFRU 12.29. The inversion occurred by recombination of the two identical copies of the gene encoding an integrating conjugative element protein of the PFL_4705 family that are located, together with some other complete and incomplete copies, at position 1850000-1858000 and 5488000-5500000 in the chromosome of CRAFRU 12.29. Chromosome inversions have been reported to affect gene expression and occasionally the phenotype (Cui *et al.*, 2012). However, whether or not the large genome inversion in CRAFRU 14.08 is associated with phenotype could not be determined in the present study.

The second major difference in strain CRAFRU 12.29 concerns a 1700 bp integrative sequence, encoding an integrase and an IS3/IS911 transposase. This small integrative unit was inserted in the *hrpS* gene, within a transcriptional unit that spans several components of the type III secretion system, including the gene encoding harpin, *hrpZ* (Figure 2.3.3). Since, according to annotation and Blast searches, there are no other copies of *hrpZ* in the genome of Psa CRAFRU 12.29, the lack of expression of *hrpZ* may conceivably be the reason for the reduced virulence on kiwifruit and inability to elicit HR on eggplant and tobacco leaves. The phenotype is indeed reminiscent of previously characterized *hrpZ* deletion mutants (He *et al.*, 1993).

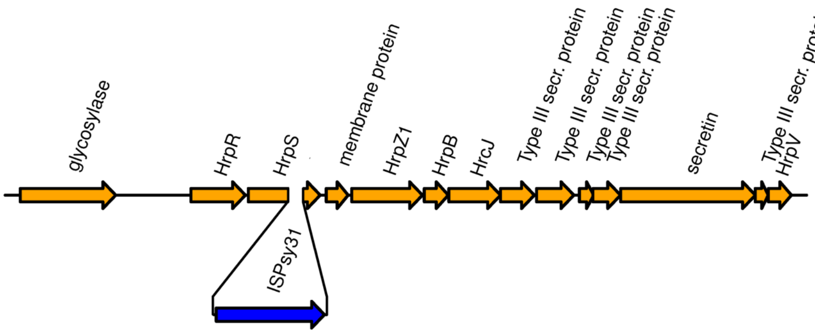


Figure 2.3.3 – Drawing of part of the *Hrp* cluster of Psa, with the location of the insertion of ISPsy31 in strain CRAFRU 12.29.

The mobilization of IS3/IS911 elements has been already reported by Butler *et al.* (2013), who found that in the comparison of Pac_ICE1 from four New Zealand strains (ICMP 18708, ICMP18800, TP1 and 6.1) the presence of an IS element of the type IS3/IS911 in strain 6.1 was the only difference. They designated this small transposable element ISPsy31 at the IS Finder database (Siguier *et al.*, 2016) and we will follow this nomenclature. Also, as remarked by Butler *et al.* (2013), ISPsy31 is predicted to have two, partially overlapping reading frames associated with a 21 frame shift (the typical pattern found in IS3/IS911 type elements). While this shift encodes no functions other than those involved in its mobility, yet it may still significantly impact the behavior of the pathogen in its interaction with the host.

There are many copies of ISPsy31 in the Psa genome. In strain CRAFRU 12.29 we counted 52 completed and five incomplete copies in the chromosome, and two complete copies in the plasmid. With the notable exception of the one interrupting *hrpS*, all other ISPsy31 copies are in corresponding positions in the chromosomes of strains CRAFRU 12.29 and CRAFRU 14.08.

On the other hand, strain CRAFRU 14.08 genome displays (position 5223542-5224799) the insertion of another IS element of the IS3/IS911 family, related to but well distinct from ISPsy31, and designated as ISPsy37 at the ISFinder database (Siguier *et al.*, 2016). There are two copies of this transposon in CRAFRU 14.08, and only a single occurrence in CRAFRU 12.29.

Finally, one variation associated with variable number tandem repeats (VNTR) was also scored at positions 2787533-2786633 in CRAFRU 14.08, in additions to two unique SNPs (*see below*).

Differences between the chromosomes of CRAFRU 14.08 and CRAFRU 12.29 are summarized in supplementary Table 6.2.2.

Structural diversity in the chromosomes of the European population of Psa biovar 3

The availability of finished genomes of European Psa isolates allowed to precisely map SNPs in additional 10 genomes (Table 2.3.1) of strains isolated in Europe, using Illumina data, as summarized in Table 2.3.2. Accordingly, a single SNP between the chromosomes of strains CRAFRU 14.08 and CRAFRU 12.29 was scored, at position 39328331 in CRAFRU 14.08. Comparison of the two finished chromosome sequences using MUMmer (Delcher *et al.*, 2002) revealed an additional SNP at position 2736260; that position corresponds to a transposase gene that is present in several copies in the genome and therefore was not detectable by reads mapping (supplementary Table 6.2.2). In summary, the SNP comparison of the 12 European Psa genomes revealed that they differ from each other in 0 to 8 sites, on a total of 19 polymorphisms detected.

The SNP analysis reported here supports the assertion of Butler *et al.* (2013) that the clonal populations in New Zealand and Chile are undergoing divergence, but as yet the frequency of idiosyncratic SNPs is less than one per Mb. A similar rate was determined in this work for European strains, as it was also anticipated by Mazzaglia et al. (2012). However, these figures are significantly lower than those reported by McCann et al. (2013) who identified 28-70 polymorphisms among the four Italian strains included in their study. The explanation of this inconsistency may lay in the fact that for three out of the four strains compared by those Authors, they used the data from *de novo* draft assemblies deposited in the database by Marcelletti (Marcelletti *et al.*, 2011), Butler (Butler *et al.*, 2013), and Mazzaglia (Mazzaglia *et al.*, 2012), respectively, and *de novo* assembly is much more error prone than the conservative read mapping method used in this work.

Mazzaglia and co-workers (2012) identified the presence, in the chromosome of Psa, of a divergent genomic island ~100 kb long, similar to PPHGI-1, an integrative conjugative elements (ICE) described earlier in *P. syringae* pv. *phaseolicola* (Pitman *et al.*, 2005), and also similar to PsyrGI-6, an ICE of *P. syringae* pv. *syringae* B728a (Feil *et al.*, 2005). The genomic island was analyzed in more detail by Butler *et al.* (2013), who named Pac_ICE2 the type shared by

European strains of Psa (in contrast with Pac_ICE1, for New Zealand strains, Pac_ICE3, for Chilean).

Butler *et al.* (2013) reported that the islands in ICMP 18708 (New Zealand), ICMP 18744 (Italy) and ICMP 19455 (Chile) were broadly synthenic, although the sequences shared by the ICEs were significantly divergent (~85% identical). Two regions with high conservation were detected, corresponding to transposons named Tn6211 and Tn6212. While Tn6211 occupies distinct positions in each of the three ICE types, the second conserved region (bases 55201–71516 in Pac_ICE1 from ICMP 18708), designated Tn6212 and almost identical in all ICEs, was synthenic in the three ICE types.

Mapping of Illumina reads examined in this work revealed two distinct types of Pac_ICE2 among the 12 European Psa genomes. The Illumina reads from five strains (namely CRAFRU 12.50, CRAFRU 12.29, CRAFRU 14.21, CRAFRU 14.08, and CRAFRU 13.27) did not cover the about 16.3 kbp of Tn6212 (Figure 2.3.4). “Split reads” containing Tn6212 flanking sequences were also found suggesting that the transposon was excised.

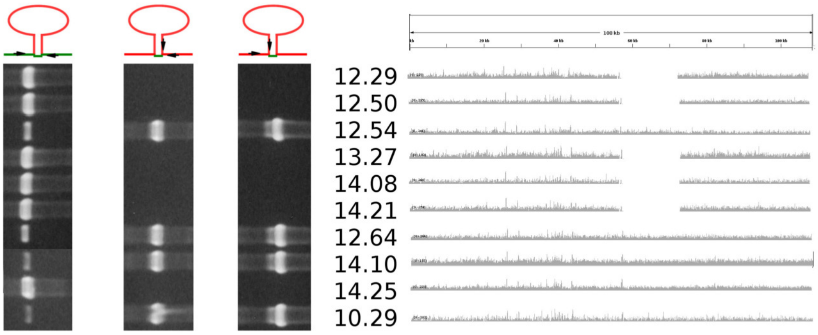


Figure 2.3.4 – Evidence of integration/excision of Tn6212. Left: Agarose gels of PCR amplification products with primers that amplify the upstream transposon junction, the downstream transposon junction, and the chromosome region resulting from excision (from left to right, as indicated in the top scheme of PCR primers positions). Right: density of reads mapping on Tn6212 and flanking regions. The numbers indicate the CRAFRU strains.

Table 2.3.1 – Strains and DNA sequences used in this work.

Stain name	Received as	Origin	Isolation year	Host plant	DNA sequence reference	HR on tobacco	HR on eggplant	Tn6212 integration	SRA database accession	GenBank accession
CRAFRU 14.08	Psa 354	Portugal	2010	A. deliciosa Summer	This work	+	+	-	SAMN06349005	CP019730
CRAFRU 12.29	23b	Italy (Piemonte)	2011	A. deliciosa Hayward	This work	-	-	-	SAMN06349003	CP019732
CRAFRU 14.25	our isolate	Italy (Latium)	2012	A. chinensis Hort16A	This work	+	+	+	SAMN06348997	n.a.
CRAFRU 12.54	1616-291a	Italy (Piemonte)	2011	A. deliciosa Hayward	This work	+	+	+	SAMN06348998	n.a.
CRAFRU 14.10	Psa 490	Italy (Calabria)	2010	A. chinensis Jintao	This work	+	+	+	SAMN06348999	n.a.
CRAFRU 12.64	1616-231Aa	Italy (Piemonte)	2010	A. chinensis Jintao	This work	+	+	+	SAMN06349000	n.a.
CRAFRU 10.29	4252 A,1	Italy (Emilia Romagna)	2009	A. chinensis Jintao	This work	+	+	+	SAMN06349001	n.a.
CRAFRU 12.50	our isolate	Italy (Campania)	2011	A. chinensis Jintao	This work	-	-	-	SAMN06349002	n.a.
CRAFRU 14.21	37.51	France	2011	A. chinensis Jintao	This work	+	+	-	SAMN06349004	n.a.
CRAFRU 13.27	IVIA 3729.2	Spain	2011	A. deliciosa Hayward	This work	+	+	-	SAMN06349006	n.a.
CRAFRU 8.43	our isolate	Italy (Latium)	2008	A. chinensis Hort16A	Marcelletti et al., 2011	n.i. (1)	n.i.	+		AFTG00000000
CRAFRU 13.04	ICMP 18884	New Zealand	2010	A. deliciosa Hayward	Templeton et al., 2016	n.i.	n.i.	n.i.	SAMN06349007	CP011972
Additional sequences used in this work										
7286		Italy			Mazzaglia et al., 2012				SRX105337	
ICMP 18708, V13		New Zealand			Poulter et al., unpublished (2)					CP012179

(1) Not Investigated

(2) deposited as Poulter,R.T.M., Poulter,G.T.M., Stockwell,P.A., Lamont,I.L. and Butler,M.I. (unpublished)

Table 2.3.2 – SNPs identified among the strains used in this work by Illumina reads mapping. Position relative to the chromosome of CRAFRU 14.08.

Strain Position	CRAFRU 12.64	CRAFRU 10.29	CRAFRU 12.50	CRAFRU 12.29	CRAFRU 14.21	CRAFRU 14.08	#7286	CRAFRU 13.27	CRAFRU 8.43	CRAFRU 14.25	CRAFRU 12.54	CRAFRU 14.10
32022	G	G	G	G	G	G	G	G	C	G	G	G
1537885	C	T	C	C	C	C	C	C	C	T	C	C
1791521	G	G	G	G	G	G	C	G	G	G	G	G
1791522	G	G	G	G	G	G	C	G	G	G	G	G
2109838	C	C	C	C	C	C	C	C	T	C	C	C
2554115	G	G	A	G	G	G	G	G	G	G	G	G
3540152	A	T	A	A	A	A	A	A	A	A	A	A
3540154	C	T	C	C	C	C	C	C	C	C	C	C
3932833	C	C	C	T	C	C	C	C	C	C	C	C
4207959	C	C	C	C	C	C	C	C	C	C	T	C
4262863	G	G	T	G	G	G	G	G	G	G	G	G
5267844	C	A	C	C	C	C	C	C	C	C	C	C
5268734	C	C	C	C	C	C	A	C	C	C	C	C
5346399	A	T	T	T	T	T	T	T	T	T	T	T
5379834	C	C	C	C	A	C	C	C	C	C	C	C
5719829	G	G	G	G	G	G	G	G	T	G	G	G
5803673	C	C	C	C	C	C	G	C	C	C	C	C
6189845	C	C	T	C	C	C	C	C	C	C	C	C
6357274	C	C	T	C	C	C	C	C	C	C	C	C

PCR carried out with primers placed on the borders of Tn6212 (Figure 2.3.4) provided confirmation of the excision and loss of Tn6212 in the named five strains: with their DNA extracts as templates, both the PCRs with primers located on left end of Tn6212 and flanking region, and the PCRs with primers located on right end of Tn6212 and flanking region, failed to amplify a DNA fragment of the expected size. Conversely, PCRs with primers specific for the left and right flanking regions amplified a DNA fragment that was 686 bp in length, i.e. lacking the Tn6212 sequence. Unexpectedly, the DNA samples from the other strains were positive not only to PCRs designed to amplify the ends of Tn6212 and flanking regions, but also primed amplification of the 686 bp DNA fragment with primers specific for the left and right flanking regions.

Since the DNA samples were prepared from 24 hrs old liquid cultures started from single colonies, we hypothesize that Tn6212 may occur with high frequency *in vitro*, so that at the time of DNA extraction the sample contained a mixture of genomes with and without Tn6212 integration. A similar hypothesis may explain the incongruity of the results concerning strain CRAFRU 14.25, that showed reads coverage of the Tn6212 region but no amplification products with primers located on its ends. Since the sequencing was carried out more than one year before PCRs, we hypothesize that subculturing ultimately selected genomes missing Tn6212.

The evidence of optional and frequent excision of Tn6212 raised the question of its potential role in the interaction with the plant host, that could warrant its maintenance in the pathogen population over time and its detection in fresh isolates.

Tn6212 has been reported to be the Psa specific part that distinguished ICEs of Psa and *Ps. syringae* pv. *phaseolicola* (Psp). McCann *et al.* (2013) pointed out the presence within the Tn6212 region of genes that may be implicated in the interaction with the plant host, such as those encoding a predicted enolase and various transporters, including an ortholog of DctT (a putative di- carboxylic acid transporter with N-terminus predicted to be targeted to the Type III Secretion System) and a methyl-accepting chemotaxis protein predicted to be involved in taxis toward malate.

In an attempt to detect differences in virulence and within-plant movement of strains, we inoculated plantlets with strain CRAFRU 8.43 and CRAFRU 14.08 and, after 10 days incubation, estimated by qPCR the bacterial population in the point of inoculation (“bottom” in Figure 2.3.5) and in the stem segment 3 cm above ("Top" in Figure 2.3.5). Although the bacterial cell number estimated of CRAFRU 8.43 were higher, the detected difference was not statistically significant. The optional excision of Tn6212 is the only significant variation in ICE2 among the 12 genomes examined. In fact, ICE2 resulted identical in all strains except for a single polymorphism in strain CRAFRU 10.29 at position 51525.

Furthermore, we examined the results of Illumina re-sequencing of all Psa strains with the aim of discovering new genes possibly acquired during clonal expansion. Following read mapping on the complete genome of strain CRAFRU 12.29, we selected and assembled the Illumina reads that were not mapped. After filtering for Tn6212 (missing in the reference) sequences, we obtained in total 175 contigs for a total of 105,000 nts. The encoded aminoacids sequences whose function could be recognized according to RAST annotation were exclusively phage associated proteins (Table 2.3.S4, see *Supporting Information*). Hence, we could find no evidence of gene gain in our sample of 12 European genomes, reveling a picture divergent from that described by Colombi *et al.* (2017) who showed the acquisition by strains isolated in New Zealand of exogenous integrative conjugative elements carrying copper resistance genes during clonal expansion.

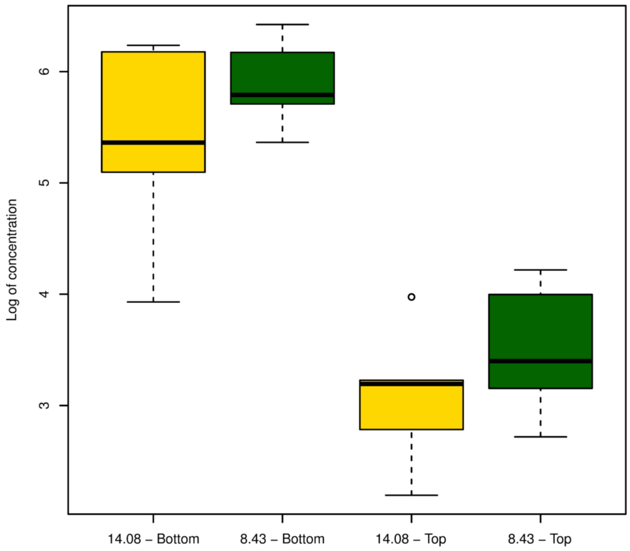


Figure 2.3.5 – Boxplot of the estimated bacterial population in the upper ("Top") and lower ("Bottom") part of the stem 10 days after inoculation with strains CRAFRU 8.43 and CRAFRU 14.08.

To confirm that genome diversity in the European strains is mostly due to rearrangement of genetic elements, the Illumina dataset was used to investigate structural changes in the chromosomes of the collections of 10 European strains. The detection of structural changes, in most cases associated with repetitive sequences, is challenging when using short Illumina reads datasets, hence we used different approaches to highlight clues of rearrangement events.

We mapped the Illumina reads from all strains on both the CRAFRU 12.29 and CRAFRU 14.08 chromosomes and visualized the alignments in the regions covering the structural changes that differentiate those chromosomes among themselves. As a result, we found that the ISPsy31 insertion in CRAFRU 12.29, as well as the ISPsy37 insertion and the large inversion in CRAFRU 14.08 were unique in the respective strain chromosomes and not shared by any other of the remaining European strains. We therefore focused on the detection of specific structural changes in the chromosomes of the other strains.

To this end, we prepared an inventory of the mobile elements that can be detected in the two complete chromosomes of European Psa, CRAFRU 12.29 and CRAFRU 14.08 (supplementary Table 6.2.5), then mapped their ends on the assemblies of other strains to detect traces of transposon mobilization. By using this approach, we found contigs ending with sequences associated with mobile element borders that were not present in the reference chromosome. In particular, we found IS3 related sequences in unique positions in CRAFRU 12.64 and CRAFRU 8.43, and an IS3 related sequence present in the same position in both CRAFRU 13.27 and CRAFRU 10.29.

The assemblies were also scaffolded using CRAFRU 12.29 genome as a reference and visualized, allowing the detection of an inversion around position 5508000 (CRAFRU 12.29 numbering) in strain CRAFRU 8.43.

Comparison of chromosomes of European vs. New Zealand Psa biovar 3 strains

The comparison of the European strain CRAFRU 12.29 and the two complete genomes of New Zealand isolates that were available from NCBI in October 2016, i.e. strains ICMP 18708 and ICMP 18884, showed substantial syntheny of the chromosomes (Figure 2.3.2).

As previously noticed the sequences diverged largely in the ICE region, and much less in the rest of the genome. As it has already been reported for other strains (Butler *et al.*, 2013) the ICE is inserted in a different lysine tRNA site in the genomes of European Psa strain CRAFRU 12.29 and in the New Zealand strain ICMP 18708.

Apart from the ICE region, the chromosomes of the two New Zealand strains were identical except for seven SNPs (including single nucleotide indels), according to the results of direct comparison using MUMmer (Delcher *et al.*, 2002) and Mauve (Darling *et al.*, 2004). Two of the indels occurred in homopolynucleotide stretches and were not confirmed by our Illumina sequencing and reads mapping of strain ICMP 18884. Thus, the number of single nucleotide variations between the two New Zealand strains were similar to that occurring among the European strains. Conversely, 27 SNPs (including indels) and three sequence variations affecting multiple nucleotides were detected between the European Psa strain CRAFRU 12.29 and the New Zealand strain ICMP 18884 in the remaining (after exclusion of ICE) about 6 Mb of the chromosome (pos 1-5410820 and 5511674-6555571, strain ICMP 18708 numbering). This finding is in substantial agreement with the hypothesis that Psa strains originating the epidemics in Chile, New Zealand and Europe were independently invaded by Pac_ICE1/3, supporting the notion that this ICE may contain genetic elements that significantly affect the virulence of the pathogen.

In addition to SNPs, several genome rearrangement events distinguished the genome of the European Psa strain CRAFRU 12.29 and the New Zealand strains ICMP 18708/18884, as presented in supplementary Table 6.2.3. Major events include the insertion of a copy of a mobile selfish genetic element of the group named bacterial group II intron reverse transcriptase/maturase in CRAFRU 12.29 at positions 1023375-1025252. Proteins in this group have an N-terminal reverse transcriptase (RNA-directed DNA polymerase) domain (pfam00078) followed by an RNA-binding maturase domain (pfam08388). This mobile element is present in 14 copies in CRAFRU 12.29 and 13 copies in ICMP 18708/18884 genomes.

On the other hand, ICMP 18708 and ICMP 18884 are characterized by a similar event, the insertion of another distinct bacterial group II intron reverse transcriptase/maturase starting at position 5715260 and ending at position 5717133. Also this transcriptase/maturase is present in several identical copies in the Psa genomes, namely 14 copies in ICMP 18708 and 13 copies in CRAFRU 12.29, respectively. There are, in total, 54 proteins annotated as bacterial group II intron reverse transcriptase/maturase in each of the two genomes in comparison. Another major difference between the two genomes concerns an insertion of two transposase genes at positions 3287490-3288700 in a DNA region that includes sequences encoding IS630 transposases, a phage invertase and related proteins that are associated with a 316 kb inversion in ICMP 18708/18884. Another IS630 insertion that is specific of ICMP 18708/18884 occurs in those genomes at position 6522179 – 6523356 (ICMP 18708 numbering). In ICMP 18708/18884 there are 61 complete and five incomplete IS630 transposases, while CRAFRU 12.29 displayed 59 complete and five incomplete copies of this gene. Two minor variation associated to repeats of variable lengths were also scored, one of which corresponding to the same repeat region that differentiated CRAFRU 14.08 from CRAFRU 12.29.

2.3.4 Conclusions

Mobile DNA elements contribute to bacterial evolution, as their ability to mobilize themselves and unrelated DNA in their proximity can lead to genome rearrangements that affect the microorganism phenotype (Bardaji *et al.*, 2011). Their role in improving fitness and, potentially, pathogenicity and virulence of phytopathogenic bacteria is well established (Jackson *et al.*, 2011). Many studies stressed the role of mobile DNA dependent gene gain in pathogen populations during epidemics, leading to the differentiation and development of more adapted clones (Holden *et al.*, 2009; Mutreja *et al.*, 2011; Petrovska *et al.*, 2016; Santagati *et al.,* 2012). Psa biovar 3 represents a relevant example of such a process, considering the primary role of mobile DNA mediated horizontal genetic transfer (particularly the gain of ICE) in its emergence as a pandemic pathogen of kiwifruit, according to several studies (Butler *et al.*, 2013; Marcelletti *et al.*, 2011; McCann *et al.*, 2013; McCann *et al.*, 2017).

However, Mobile DNA-induced mutations are often deleterious (Wu et al., 2015), and transposable elements have been regarded as a sort of genomic disease (Wagner, 2009). Loss of fitness due to the accumulation of deleterious mutations has been reported for small, obligate asexual populations, as these are incapable of reconstituting highly fit genotypes by recombination or back mutation (Lynch *et al.*, 1993; Moran, 1996).

According to the results of a pangenomic study by Bolotin and Hershberg (2015), while non-clonal species diversify through a combination of changes to gene sequences (gene loss and gene gain), gene loss completely dominates as a source of genetic variation among clonal species, for which it needs to be taken into account as a potential dominant source of phenotypic variation. In the case of Psa biovar 3, we report here a relevant number (considering the small sample) of transposon mediated structural variations, occasionally impairing relevant phenotypic aspects of the interaction with the host, as occurred in the genome of strain CRAFRU 12.29 where a ISPsy31 insertion in the *hrpS* gene disrupted the functionality of the TTSS. In all cases, structural variations implied rearrangement of genetic elements and not incorporation of external DNA.

There is a growing body of evidence supporting the hypothesis of two phases in the recent evolution of Psa biovar 3, with a landmark in the initiation of the worldwide pandemic in 2008. The SNP based comparisons (this work, McCann *et al.*, 2017), as well as the evidence of independent invasions of ICE (Butler *et al.*, 2013), suggest the conservation of within-biovar diversity in the natural environment of the region of origin and during initial spread in China, before pandemic initiation. In this phase, acquisition of exogenous DNA through mobile DNA and selection for increased fitness were drivers of the evolution, promoting the emergence of adapted individuals. Also in this phase, recombination (intra- and inter-pathovar; McCann *et al.*, 2013; McCann *et al.*, 2017) and selection limited the proliferation of transposons and the deleterious mutations associated to DNA mobilization.

A new phase begun with the introduction of adapted highly virulent strains from China into the kiwifruit cultivated areas in Europe, Chile and New Zealand. In Europe, Psa biovar 3 established and spread clonally in an ecological niche lacking competitive selection, such as that represented by the highly sensitive *A. chinensis* cv. Hort 16A. The results of this study show that the new phase was associated to an increase in the number of small transposons in the bacterial genome, with rearrangements leading to gene loss rather than to gain of functions by horizontal transfer. The data collected herein would suggest that clonal spread of the pathogen in a free ecological niche occurred with no access to the environmental gene pool, with diversification through rearrangement of genetic elements, and in the absence of the recombination-selection process that mitigates genome degeneration associated with transposon mobilization (Bast *et al.*, 2016).

While evidence of gene gain associated with the emergence of copper-resistant strains was recently reported by Colombi *et al.* (2017) for Psa in New Zealand, in this study we report evidence of gene loss and the isolation of some low virulence variant Psa biovar 3 strains. The different outcomes of the surveys may be related with differences in the environmental conditions or in epidemic dynamics or disease management, such as timing of the disease spread on the territory, introduction of tolerant cultivars, use of containment measures directed to the reduction of the inoculum size (particularly copper treatments) or to the reduction of pathogen dispersal and the establishment of conducive conditions for the epidemics (pruning, girdling, cultivation under cover), prevalence of the crop in the region (Vanneste, 2017).

Modern strategies for the management of destructive epidemics, such as that caused by Psa biovar 3 on kiwifruit, may benefit from the awareness of their effect on short-term genome evolution and population structure of the pathogen. The results presented in this paper would suggest that strategies that do not promote recombination and preserve the clonal structure of the invasive microorganism may be associated with lower risk of developing variant strains with enhanced fitness or virulence.

2.3.5 Experimental procedures

Strains and sequencing

The strains investigated in this work and their genome data accessions are listed in Table 2.3.1.

Genomic DNA was extracted from 1 ml of 24 hrs old cultures grown in Nutrient Broth with agitation using a Wizard DNA purification kit (Promega Italia, Padova, Italy) following the manufacturer's instructions. DNA was measured and checked for quality using a NanoDrop spectrophotometer (NanoDrop products, Wilmington, DE, USA). Illumina libraries were prepared as described previously (Scortichini et al., 2013) and sent to the Istituto di Genomica Applicata (Udine, Italy) for sequencing on a Illumina Genome Analyser IIx (Illumina, USA). An average of 14 million single (50 nts) reads were obtained, filtered for quality using Prinseq (Schmieder and Edwards, 2011) and further processed. The sequence reads of strain 7286, obtained by Mazzaglia et al. (2012) were downloaded from the Sequence Read Archive (SRA accession SRX105337; https://www.ncbi.nlm.nih.gov/sra). The complete genome sequence of strain ICMP ICMP 18884 (Templeton *et al.*, 2015) and ICMP 18708 (yet unpublished but made available by Poulter, R.T.M., Poulter, G.T.M., Stockwell, P.A., Lamont, I.L. and Butler, M.I.) were obtained from the NCBI nucleotide database and used as comparative reference for non-European strains.

Genomic DNA extracted from strains CRAFRU 12.29 and CRAFRU 14.08 was also sent for single molecule real-time (SMRT) sequencing to the University of Washington PacBio Sequencing Services. The genomes were then completed with Sanger sequencing using a primer walking approach on PCR fragments amplified from putatively adjacent contigs ends, as resulted by scaffolding using ICMP 18708 as a reference; fragments were sent for sequencing to Genelab, Casaccia, Italy. Sequences were edited and manipulated using Seaview (Gouy *et al.*, 2010) and Ugene (Okonechnikov *et al.*, 2012).

Sequence analysis

Preliminary reads alignments and alignments manipulation were carried out using widely used tools such as BWA 0.5.5 (Li and Durbin, 2009), SAMtools 0.1.16 (Li *et al.*, 2009) and PICARD tools (http://picard.sourceforge.net). SNP calling was carried out with the GATK package (McKenna *et al.*, 2010); SNPs call was supported by a depth of coverage of at least 5 and a consensus of at least 95% of the aligned reads. Tablet (Milne *et al.*, 2010) and IGV (Robinson *et al.*, 2011) were used for the visualization of the alignments.

Assemblies of small DNA regions were carried out with Edena (Hernandez *et al.*, 2008). Genome reconstructed from Illumina reads were assembled with SPAdes (Bankevich *et al.*, 2012) and scaffolded with Ragout (Kolmogorov *et al.*, 2014). Alignments were carried out with Mauve (Darling *et al.*, 2004) and MUMmer (Delcher *et al.*, 2002). The above listed tools were integrated with several *ad-hoc* Perl scripts into Bash scripts and run on Linux instances launched on the infrastructures of the DIAG (http://www.igs.umaryland.edu/resources/irc/) and CyVerse (Merchant *et al.*, 2016) projects.

Annotation of Insertion Sequences (IS) in the complete genomes was carried out at the ISsaga (Insertion Sequence semi-automatic genome annotation) engine (Varani *et al.*, 2011).

Plantlet inoculations

To investigate whether or not Psa strains were impaired in their within plant spread capabilities, micropropagated kiwifruit plantlets *Actinidia chinensis* (cv. Soreli) at the stage of 6 leaves, provided by Az. Agr. Fanna Giampaolo (Moimacco, Italy) were used for plantlet inoculation. Bacterial strains grown for 24 hrs in Nutrient Broth with agitation were washed twice and resuspended in 0.9% saline solution in concentration of 1–2·10^9 cfu/ml. Plantlets were cut from callus, dipped in the inoculum and transferred to a fresh medium. Control plants were dipped in sterile saline. After 10 days the plantlets were collected, cut into two halves (about 3 cm from inoculation point), and DNA was extracted from each subsample according to standard protocols (Doyle and Doyle, 1990). The bacterial populations were quantified by qPCR according to published protocol (Gallelli *et al.*, 2014). For statistical analysis, carried out with R (R Core Team, 2013), the medians of three PCR reactions for each of five repetitions per strain was used.

Leaves inoculations

To compare the capability of strains to induce disease symptoms and to determine their growth *in planta*, *Actinidia chinensis* (cv. Dorì®) leaves were inoculated with the method described previously (Marcelletti *et al.*, 2011). Leaf areas of approximately 1 cm in diameter were inoculated at the concentrations of 1–2·10^6 cfu/ml. For each thesis, 10 leaves were inoculated in four sites. Control plants were treated using solely sterile solution (0.85 % NaCl). Two, 6, 15 and 22 days after inoculation, leaf disks of about 0.5 cm of diameter were sampled and ground in 1 ml of sterile saline, then serial ten-fold dilutions were counted by colony growth onto nutrient agar supplemented with 3% of sucrose (NSA).

Hypersensitive reactions were tested by infiltrating aqueous bacterial suspensions at 1–2 x 10E8 cfu/ml on fully expanded tobacco and eggplant leaves using a needless syringe. The development of typical hypersensitivity reaction was checked within 48 hrs after infiltration. Assays were repeated three times.

Other wet lab methods

To determine the excised/integrated state of Tn6212, primers (supplementary Table 6.2.1) were designed on the inner and outer borders of the transposon. PCRs with primer pairs fX1/rX2; fX1/rX4 and fX3/rX4 were performed with the automated One Advanced thermocycler (EuroClone, Celbio, Milan, Italy) in 25 µl reactions containing 200 µM of each of the four dNTPs, 0.4 µM of each primer, 1.5 mM MgCl2, 0.625 units of GoTaq Flexi DNA Polymerase (Promega, Madison, WI, USA) and 1 µl of diluted bacterial DNA (5 ng/µl). The PCR program consisted of initial denaturation for 2 min at 94 °C; 35 cycles of 1 min at 94 °C, 45 sec at 58 °C, 1 min at 72 °C; and a final extension for 8 min at 72 °C.

PCR products were separated by electrophoresis in a 1% agarose gel, stained with ethidium bromide, and captured with a DigiDoc-It imaging system (UVP, Cambridge, United Kingdom).

2.3.6 Acknowledgements

The following colleagues are gratefully acknowledged for sharing their isolates with us: G.M. Balestra, Tuscia University, Viterbo, Italy; C. Morone, Regione Piemonte, Servizio Fitosanitario regionale, Torino, Italy; M.M. Lopez, Instituto Valenciano de Investigaciones Agrarias, Moncada-Valencia, Spain; S. Poliakoff, ANSES, Angers, France. CyVerse, supported by the National Science Foundation under Award Numbers DBI-0735191 and DBI-1265383, and the Data Intensive Academic Grid (DIAG, sadly retired on Jan 2017), supported by the National Science Foundation under Grant No. 0959894 are acknowledged for providing computer resources. Financial support of this work came from the Region Friuli Venezia Giulia, Italy.

2.3.7 References

Abelleira, A., Lopez, M.M., Penalver, J., Aguin, O., Mansilla, J.P., Picoaga, A. and Garcia, M.J. (2011) First Report of Bacterial Canker of Kiwifruit Caused by *Pseudomonas syringae* pv. *actinidiae* in Spain. *Plant Dis.* **95**, 1583–1583.

Balestra, G.M., Mazzaglia, A., Quattrucci, A., Renzi, M. and Rossetti, A. (2009) Occurrence of *Pseudomonas syringae* pv. *actinidiae* in Jintao kiwi plants in Italy. *Phytopathol. Mediterr.* **48**, 299–301.

Bankevich, A., Nurk, S., Antipov, D., et al. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **19**, 455–477.

Bardaji, L., Añorga, M., Jackson, R.W., Martínez-Bilbao, A., Yanguas-Casás, N. and Murillo, J. (2011) Miniature transposable sequences are frequently mobilized in the bacterial plant pathogen *Pseudomonas syringae* pv. *phaseolicola*. *PLoS ONE* **6**, e25773**.**

Bast, J., Schaefer, I., Schwander, T., Maraun, M., Scheu, S. and Kraaijeveld, K. (2016) No accumulation of transposable elements in asexual arthropods. *Mol. Biol. Evol.* **33**, 697–706.

Bolotin, E. and Hershberg, R. (2015) Gene loss dominates as a source of genetic variation within clonal pathogenic bacterial species. *Genome Biology and Evolution* **7**, 2173-2187.

Butler, M.I., Stockwell, P.A., Black, M.A., Day, R.C., Lamont, I.L. and Poulter, R.T.M. (2013) *Pseudomonas syringae* pv. *actinidiae* from recent outbreaks of kiwifruit bacterial canker belong to different clones that originated in China. *PloS One* **8**, e57464.

Ciarroni, S., Gallipoli, L., Taratufolo, M.C., Butler, M.I., Poulter, R.T.M., Pourcel, C., Vergnaud, G., Balestra, G.M. and Mazzaglia, A. (2015) Development of a Multiple Loci Variable Number of Tandem Repeats Analysis (MLVA) to Unravel the Intra-Pathovar Structure of *Pseudomonas syringae* pv. *actinidiae* Populations Worldwide. *PloS One* **10**, e0135310.

Colombi, E., Straub, C., Künzel, S., Templeton, M.D., McCann, H.C. and Rainey, P.B. (2017) Evolution of copper resistance in the kiwifruit pathogen *Pseudomonas syringae* pv. *actinidiae* through acquisition of integrative conjugative elements and plasmids. *Environ. Microbiol.***19**, 819-832

Cui, L., Neoh, H., Iwamoto, A. and Hiramatsu, K. (2012) Coordinated phenotype switching with large-scale chromosome flip-flop inversion observed in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E1647-1656.

Cunty, A., Cesbron, S., Poliakoff, F., Jacques, M.-A. and Manceau, C. (2015a) Origin of the outbreak in France of *Pseudomonas syringae* pv. *actinidiae* biovar 3, the causal agent of bacterial canker of kiwifruit, revealed by a multilocus variable-number tandem-repeat analysis. *Appl. Environ. Microbiol.* **81**, 6773–6789.

Cunty, A., Poliakoff, F., Rivoal, C., Cesbron, S., Fischer-Le Saux, M., Lemaire, C., Jacques, M.A., Manceau, C. and Vanneste, J.L. (2015b) Characterization of *Pseudomonas syringae* pv. *actinidiae* (Psa) isolated from France and assignment of Psa biovar 4 to a de novo pathovar: *Pseudomonas syringae* pv. *actinidifoliorum* pv. nov. *Plant Pathol.* **64**, 582–596.

Darling, A.C.E., Mau, B., Blattner, F.R. and Perna, N.T. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403.

Delcher, A.L., Phillippy, A., Carlton, J. and Salzberg, S.L. (2002) Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30**, 2478–2483.

Dolgin, E.S. and Charlesworth, B. (2006) The fate of transposable elements in asexual populations. *Genetics* **174**, 817–827.

Doyle, J.J. and Doyle, J.L. (1990) Isolation of plant DNA from fresh tissue. *Focus* **12**, 13–15.

Dreo, T., Pirc, M., Ravnikar, M., Zezlina, I., Poliakoff, E., Rivoal, C., Nice, F. and Cunty, A. (2014) First Report of *Pseudomonas syringae* pv. *actinidiae*, the Causal Agent of Bacterial Canker of Kiwifruit in Slovenia. *Plant Dis.* **98**, 1578–1578.

EPPO (2016) EPPO Global Database (available online, url:https://gd.eppo.int/).

Everett, K.R., Taylor, R.K., Romberg, M.K., Rees-George, J., Fullerton, R.A., Vanneste, J.L. and Manning, M.A. (2011) First report of *Pseudomonas syringae* pv. *actinidiae* causing kiwifruit bacterial canker in New Zealand. *Australas. Plant Dis. Notes* **6**, 67–71.

Feil, H., Feil, W.S., Chain, P., et al. (2005) Comparison of the complete genome sequences of *Pseudomonas syringae* pv. *syringae* B728a and pv. *tomato* DC3000. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 11064–11069.

Ferrante, P. and Scortichini, M. (2009) Identification of *Pseudomonas syringae* pv. *actinidiae* as Causal Agent of Bacterial Canker of Yellow Kiwifruit (*Actinidia chinensis* Planchon) in Central Italy. *J. Phytopathol.* **157**, 768–770.

Ferrante, P. and Scortichini, M. (2010) Molecular and phenotypic features of *Pseudomonas syringae* pv. *actinidiae* isolated during recent epidemics of bacterial canker on yellow kiwifruit (*Actinidia chinensis*) in central Italy. *Plant Pathol.* **59**, 954–962.

Ferrante, P. and Scortichini, M. (2015) Redefining the global populations of *Pseudomonas syringae* pv. *actinidiae* based on pathogenic, molecular and phenotypic characteristics. *Plant Pathol.* **64**, 51–62.

Ferrante, P., Takikawa, Y. and Scortichini, M. (2015) *Pseudomonas syringae* pv. *actinidiae* strains isolated from past and current epidemics to *Actinidia* spp. reveal a diverse population structure of the pathogen. *Eur. J. Plant Pathol.* **142**, 677–689.

Fujikawa, T. and Sawada, H. (2016) Genome analysis of the kiwifruit canker pathogen *Pseudomonas syringae* pv. *actinidiae* biovar 5. *Sci. Rep.* **6**, 21399.

Gallelli, A., Talocci, S., Pilotti, M. and Loreti, S. (2014) Real- time and qualitative PCR for detecting *Pseudomonas syringae* pv. *actinidiae* isolates causing recent outbreaks of kiwifruit bacterial canker. *Plant Pathol.* **63**, 264–276.

Gouy, M., Guindon, S. and Gascuel, O. (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **27**, 221–224.

He, S.Y., Huang, H.C. and Collmer, A. (1993) *Pseudomonas syringae* pv. *syringae* harpinPss: a protein that is secreted via the Hrp pathway and elicits the hypersensitive response in plants. *Cell* **73**, 1255–1266.

Hernandez, D., François, P., Farinelli, L., Osterås, M. and Schrenzel, J. (2008) De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res.* **18**, 802–809.

Holden, M.T.G., Hauser, H., Sanders, M., et al. (2009) Rapid evolution of virulence and drug resistance in the emerging zoonotic pathogen *Streptococcus suis*. *PloS One* **4**, e6072.

Holeva, M.C., Glynos, P.E. and Karafla, C.D. (2015) First Report of Bacterial Canker of Kiwifruit Caused by *Pseudomonas syringae* pv. *actinidiae* in Greece. *Plant Dis.* **99**, 723–723.

Jackson, R.W., Vinatzer, B., Arnold, D.L., Dorus, S. and Murillo, J. (2011) The influence of the accessory genome on bacterial pathogen evolution. *Mob. Genet. Elem.* **1**, 55–65.

Koh, J.K., Cha, B.J., Chung, H.J. and Lee, D.H. (1994) Outbreak and spread of bacterial canker in kiwifruit. *Korean J. Plant Pathol.* **10**, 68–72.

Kolmogorov, M., Raney, B., Paten, B. and Pham, S. (2014) Ragout-a reference-assisted assembly tool for bacterial genomes. *Bioinforma. Oxf. Engl.* **30**, i302-309.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079.

Lynch, M., Bürger, R., Butcher, D. and Gabriel, W. (1993) The mutational meltdown in asexual populations. *J. Hered.* **84**, 339–344.

Marcelletti, S., Ferrante, P., Petriccione, M., Firrao, G. and Scortichini, M. (2011) *Pseudomonas syringae* pv. *actinidiae* draft genomes comparison reveal strain-specific features involved in adaptation and virulence to *Actinidia* species. *PloS One* **6**, e27297.

Marcelletti, S and Scortichini, M. (2011). Clonal outbreaks of bacterial canker of kiwifruit caused by *Pseudomonas syringae* pv. *actinidiae* on *Actinidia chinensis* and *A. deliciosa* in Italy. *J. Plant Pathol.* **93**, 479-483.

Mazzaglia, A., Studholme, D.J., Taratufolo, M.C., Cai, R., Almeida, N.F., Goodman, T., Guttman, D.S., Vinatzer, B.A. and Balestra, G.M. (2012) *Pseudomonas syringae* pv. *actinidiae* (PSA) isolates from recent bacterial canker of kiwifruit outbreaks belong to the same genetic lineage. *PloS One* **7**, e36518.

McCann, H.C., Li, L., Liu, Y., et al. (2017) Origin and evolution of a pandemic lineage of the kiwifruit pathogen *Pseudomonas syringae* pv. *actinidiae*. *Genome Biol Evol*. **9**, 932-944.

McCann, H.C., Rikkerink, E.H.A., Bertels, F., et al. (2013) Genomic analysis of the Kiwifruit pathogen *Pseudomonas syringae* pv. *actinidiae* provides insight into the origins of an emergent plant disease. *PLoS Pathog.* **9**, e1003503.

McKenna, A., Hanna, M., Banks, E., et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303.

Merchant, N., Lyons, E., Goff, S., Vaughn, M., Ware, D., Micklos, D. and Antin, P. (2016) The iPlant Collaborative: Cyberinfrastructure for enabling data to discovery for the life sciences. *PLOS Biol.* **14**, e1002342.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F. and Marshall, D. (2010) Tablet—next generation sequence assembly visualization. *Bioinformatics* **26**, 401–402.

Moran, N.A. (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 2873–2878.

Mutreja, A., Kim, D.W., Thomson, N.R., et al. (2011) Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477**, 462–465.

Okonechnikov, K., Golosova, O. and Fursov, M. (2012) Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* **28**, 1166–1167.

Petrovska, L., Mather, A. E., Abuoun, M., Branchu, P., Harris, S. R., Connor, T., and Kingsley, R. A. (2016). Microevolution of monophasic Salmonella Typhimurium during epidemic, United Kingdom, 2005-2010. *Emerging Infectious Diseases* **22**, 617-624.

Pitman, A.R., Jackson, R.W., Mansfield, J.W., Kaitell, V., Thwaites, R. and Arnold, D.L. (2005) Exposure to host resistance mechanisms drives evolution of bacterial virulence in plants. *Curr. Biol. CB* **15**, 2230–2235.

R Core Team (2013) R: A language and environment for statistical computing. Vienna, Austria: the R Foundation for Statistical Computing.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. and Mesirov, J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26.

Schmieder, R. and Edwards, R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864.

Santagati, M., Campanile, F. and Stefani, S. (2012). Genomic diversification of enterococci in hosts: The role of the mobilome. *Frontiers in Microbiology* **3**, 95.

Scortichini, M. (1994) Occurrence of *Pseudomonas syringae* pv. *actinidiae* on Kiwifruit in Italy. *Plant Pathol.* **43**, 1035–1038.

Scortichini, M., Marcelletti, S., Ferrante, P. and Firrao, G. (2013) A Genomic redefinition of *Pseudomonas avellanae* species. *PloS One* **8**, e75794.

Scortichini, M., Marcelletti, S., Ferrante, P., Petriccione, M. and Firrao, G. (2012) *Pseudomonas syringae* pv. *actinidiae*: a re-emerging, multi-faceted, pandemic pathogen. *Mol. Plant Pathol.* **13**, 631–640.

Siguier, P., Mahillon, J. and Chandler, M. (2016) IS Finder Database. url: www-is.biotoul.fr

Templeton, M.D., Warren, B.A., Andersen, M.T., Rikkerink, E.H.A. and Fineran, P.C. (2015) Complete DNA Sequence of *Pseudomonas syringae* pv. *actinidiae*, the causal agent of kiwifruit canker disease. *Genome Announc.* **3**, e01054-15.

Vanneste, J.L. (2017) The scientific, economic, and social impacts of the New Zealand outbreak of bacterial canker of kiwifruit (*Pseudomonas syringae* pv. *actinidiae*). *Ann. Rev. Phytopathol.* **55**, in press.

Varani, A.M., Siguier, P., Gourbeyre, E., Charneau, V. and Chandler, M. (2011) ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biol.* **12**, R30.

Wagner, A. (2009) Transposable elements as genomic diseases. *Mol BioSyst.* **5**, 32-35.

Wu, Y., Aandahl, R.Z. and Tanaka, M.M. (2015) Dynamics of bacterial insertion sequences: can transposition bursts help the elements persist? *BMC Evol. Biol.* **15,** 288

3 Metagenomic approaches for the characterization of fastidious prokaryotes

In an ideal situation, a pathogen is isolated and cultivated for further studies and for sequencing. Unfortunately, some microorganisms cannot be reliably cultivated *in vitro*, which in turns makes it difficult to amplify them to the amount required for analyses. A typical example of fastidious pathogens are phytoplasmas, wall-less obligate parasites of phloematic tissues, transmitted by insect vectors; belonging to the class *Mollicutes* (thus related to mycoplasmas and spiroplasmas), they were originally identified in 1967 and named mycoplasma-like organisms (MLOs) (Doi *et al.*, 1967). Because of the inability to cultivate them, a step that is required by the International Code for the Nomenclature of the Bacteria, phytoplasmas are presently classified under the '*Candidatus* Phytoplasma' genus (IRPCM Phytoplasma/Spiroplasma Working Team–Phytoplasma Taxonomy Group, 2004)

Currently, only for four phytoplasmas the finished genome sequence is available: the onion yellows (Oshima *et al.*, 2004) and the aster yellows witches' broom phytoplasmas (both belonging to '*Candidatus* Phytoplasma asteris') (Bai *et al.*, 2006), '*Ca.* P. australiense' (Tran-Nguyen *et al.*, 2008) and '*Ca.* P. mali' (Kube *et al.*, 2008). Phytoplasmas genomes are among the smallest known genomes, 600–1300 kbp long, and often lack genes encoding essential metabolic functions, and cannot therefore survive outside vectors or plant phloem. Phytoplasmas have resisted most attempts to culture them in vitro, although there are reports of successful cultivation (Contaldo *et al.*, 2012).

In the following papers, a new WGS-based *in vitro* pipeline to derive a *Phytoplasma* sequence directly from infected samples is presented, followed by two works that employ this strategy: in one not only was the phytoplasma isolated, but a spiroplasma was additionally identified, while in the other it is shown that the pipeline can employ a combined reference genome to screen the pathogen.

Bibliography

Bai, X. *et al.* (2006) ‘Living with genome instability: the adaptation of phytoplasmas to diverse environments of their insect and plant hosts.’, *Journal of bacteriology*. American Society for Microbiology, 188(10), pp. 3682–96. doi: 10.1128/JB.188.10.3682-3696.2006.

Contaldo, N. *et al.* (2012) ‘Axenic culture of plant pathogenic phytoplasmas’, *Phytopathologia Mediterranea*, 51(3), pp. 607–617.

Doi, Y. *et al.* (1967) ‘Mycoplasma- or PLT Group-like Microorganisms Found in the Phloem Elements of Plants Infected with Mulberry Dwarf, Potato Witches’ Broom, Aster Yellows, or Paulownia Witches’ Broom’, *Japanese Journal of Phytopathology*, 33(4), pp. 259–266. doi: 10.3186/jjphytopath.33.259.

IRPCM Phytoplasma/Spiroplasma Working Team–Phytoplasma Taxonomy Group (2004) ‘“Candidatus Phytoplasma”, a taxon for the wall-less, non-helical prokaryotes that colonize plant phloem and insects’, *International Journal of Systematic and Evolutionary Microbiology*, 54(4), pp. 1243–1255. doi: 10.1099/ijs.0.02854-0.

Kube, M. *et al.* (2008) ‘The linear chromosome of the plant-pathogenic mycoplasma “Candidatus Phytoplasma mali”’, *BMC Genomics*. BioMed Central, 9(1), p. 306. doi: 10.1186/1471-2164-9-306.

Oshima, K. *et al.* (2004) ‘Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma’, *Nature Genetics*. Nature Publishing Group, 36(1), pp. 27–29. doi: 10.1038/ng1277.

Tran-Nguyen, L. T. T. *et al.* (2008) ‘Comparative genome analysis of “Candidatus Phytoplasma australiense” (subgroup tuf-Australia I; rp-A) and “Ca. Phytoplasma asteris” Strains OY-M and AY-WB.’, *Journal of bacteriology*. American Society for Microbiology, 190(11), pp. 3979–91. doi: 10.1128/JB.01301-07.

3.1 An Effective Pipeline Based on Relative Coverage for the Genome Assembly of Phytoplasmas and Other Fastidious Prokaryotes

Authors: Cesare Polano[1], Giuseppe Firrao[1]. ***Manuscript submitted to "Current Genomics".***

[1] Dipartimento di Scienze Agroalimentari, Ambientali e Animali – Università di Udine – 33100 Udine, Italy

3.1.1 Abstract

A pipeline for the genome assembly of pathogens that cannot be axenically cultivated, with particular reference to the plant pathogenic phytoplasmas, is presented. The *Phytoassembly* pipeline uses ILLUMINA sequencing data produced from DNA isolated from an infected plant, using a healthy host genome reference as a filter and exploiting the difference in coverage between the sequences of the pathogen and those of the host. For phytoplasma infected samples containing >2-4% of pathogen DNA and an isogenic reference sequence the resulting assemblies can be next to complete. The pipeline has been benchmarked using simulated and real ILLUMINA runs.

Using this pipeline, high quality draft assemblies were obtained for '*Ca.* Phytoplasma aurantifolia' strain 2034 causing Lime Witches' Broom of Lime, the phytoplasma strain associated to Cassava Frogskin Disease (CFSD) and that associated to Chicory Phyllody (ChiP).

3.1.2 Introduction

Phytoplasmas are bacterial plant pathogens that cause disease in over 100 plant families (Lee *et al.*, 2000); they belong to the class *Mollicutes,* bacteria characterized by the absence of a cell wall, and are typically about 200–300 nm in size, with a genome of $0.5–1.2 \cdot 10^6$ nts (Zhao *et al.*, 2005). They live in the host phloem cells and propagate by vectors such as insects (mainly *Cicadellidae, Fulgoroidea* and *Psyllidae;* (Weintraub and Beanland, 2006)) and parasitic plants (Marcone *et al.*, 1997).

Genomics of fastidious prokaryotes is made challenging by the fact that they are difficult to cultivate *in vitro* (Tran-Nguyen and Gibb, 2007). For the phytoplasmas, protocols typically involve time consuming isolation and purification of DNA from plant or insect infected tissue using CsCl equilibrium buoyant density gradient in the presence of bisbenzimide (Saeed *et al.*, 1994), or physical isolation by pulsed-field gel electrophoresis

(PFGE) of entire chromosomes (Oshima *et al.*, 2004). Currently, only for four phytoplasmas the genomes have been sequenced to completion: '*Ca.* Phytoplasma asteris' Onion Yellows phytoplasma strain M (Oshima *et al.*, 2004), '*Ca.* P. asteris' Aster Yellows phytoplasma strain Witches' Broom ph. (Bai *et al.*, 2006), '*Ca.* P. mali' strain AT (Kube *et al.*, 2008) and '*Ca.* P. australiense' strains Paa and SLY (Tran-Nguyen *et al.*, 2008; Andersen *et al.*, 2013).

Genomic surveys have also been published for multiple phytoplasmas (Liefting and Kirkpatrick, 2003; Garcia-Chapa *et al.*, 2004; Cimerman, Arnaud and Foissac, 2006; Kawar *et al.*, 2010). With the introduction of New Generation Sequencing (NGS) methods, an emerging alternative, made possible by informatics tools, is to random sequence a large library of DNA extracted from diseased plants and then select the sequences of the pathogen. However, the pathogen sequence selection is not trivial and therefore many genome drafts obtained with this approach so far are incomplete (Casati *et al.*, 2011; Saccardo *et al.*, 2012; Chung *et al.*, 2013; Davis *et al.*, 2013; Quaglino *et al.*, 2013, 2015; Chen *et al.*, 2014).

The pipeline developed here, named *Phytoassembly*, is an evolution of the procedure described in (Saccardo *et al.*, 2012) and exploits on one hand the differential coverage of sequences originating from the pathogen and those from the host, due to the relative abundance of pathogen genome units even in samples with less than 10% pathogen DNA, and on the other hand the filtration of reads that map on a reference healthy plant genome assembly.

3.1.3 Materials and methods

Design and implementation of the pipeline

A major point in the procedure presented here is that plant sequences are separated first by setting a cutoff point based on the differential coverage of the plant (host) and the phytoplasma (pathogen) sequence contigs resulting from a pre-assembly. Indeed, in samples collected from phytoplasma infected plants, despite the prevalence of host DNA, the number of phytoplasma genome copies exceed the number of host genome copies. Phytoplasma genomes sizes range around 10^6 bp, while plant genomes are about 3 orders of magnitude larger (Zonneveld, Leitch and Bennett, 2005); therefore when counting the reads in an ILLUMINA data-set obtained from a diseased plant sample containing 1% phytoplasma DNA, the coverage of phytoplasma DNA is expected to be 10 times greater than the coverage of the plant DNA. As the sequence of phytoplasma DNA are over-represented, it would be possible to select phytoplasma reads in an ILLUMINA data-set from infected samples assuming a cutoff in a coverage graph; with data obtained by phytoplasma enriched samples from well infected plants the peaks are distinct, but in many other cases there is overlap between the phytoplasma and the host peaks, hence determining the optimal cutoff requires an estimation, that is carried out by the program, to ensure that all phytoplasma reads are retained during the selection.

Thus, the first steps of the pipeline consist in a preassembly, the estimation of pre-contigs coverage and calculation of the optimal cutoff. Then the ILLUMINA reads belonging to contigs above the cutoff are selected and aligned against the healthy plant genome reference, so that those pertaining to the plant can be discarded and the non-plant reads can be assembled in preliminary phytoplasma assembly. Further polishing is carried out to filter out ambiguous contigs, originating from low-quality reads from the plant. This is based on the percentage of indentity of BLAST matches against the healthy plant reference, the threshold being any match greater than 95%.

The standard procedure requires a reference genome from an uninfected plant in FASTA format and the sequence reads from an ILLUMINA MiSeq in FASTQ format. If necessary, the pipeline can also assemble reference genome reads in FASTQ format, and it is possible to also input the already assembled sequence reads in FASTA format. For best results, the healthy plant should be isogenic to, and grown in the same environment as the diseased specimen, so as to match the plant genome and include the same contaminants. The aforementioned BLAST verification becomes a necessity if the reference does not meet these qualities. On the other hand, it is possible to input a collection of reference genomes (simply by joining the relative FASTQ files), e.g. to filter out known pathogens.

The pipeline is written in the Bash and Perl languages and requires a working installation of *BioPerl* (http://bioperl.org/), *NCBI Blast+* (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and the *A5 pipeline* (Tritt *et al.*, 2012). *Phytoassembly* has been tested on Linux Ubuntu 16.04 LTS and Mac OS X 10.11.6.

In detail, the pipeline includes the following steps:

Stage 0: data preparation. Phytoassembly calls the A5 pipeline to assemble the healthy plant sequence reads (producing the file *Healthy.contigs.fasta*), unless an already assembled sequence is provided. Next, the diseased plant reads are assembled (producing the file *Diseased.contigs.fasta*). A step in the A5 pipeline produces error corrected reads (*Diseased.ec.fastq*), which are used in all the subsequent steps. The assembled reference sequence file is then indexed and aligned with the error corrected reads using the *BWA* tool (Li and Durbin, 2009). The resulting file is converted to the *bam* format (*Diseased.mapped.bam*) and, using *samtools* (http://www.htslib.org/doc/samtools.html), a summary of statics is produced (*Diseased.sorted.csv*), consisting of the reference sequence name, sequence length, number of mapped and unmapped reads.

Stage 1: cutoff. The pipeline estimates the optimal cutoff value by running once with cutoff 0, then using a fraction of the ratio between the sum of the lengths of the non-mapping reads at cutoff 0 (*Stage2.0.nonmatch.fastq,* see below) and the sum of the lengths of the error corrected reads (*Diseased.ec.fastq*) of the diseased plant, multiplied by 100. Alternatively, if the user wants to supply a range of specifies fixed cutoff values, then the pipeline repeats the following steps from the lowest to the highest values provided (represented here as *$cutoffval*). From the summary of statistical data (*Diseased.sorted.csv*), per-contig coverages are calculated (as the ratio between the sum of the lengths of the mapped reads and the length of the contig, multiplied by 100), and saved in a text file (*Diseased.sorted.cov.csv*). The contigs with a coverage higher than *$cutoffval* are exported to a FASTA file (*Diseased.cutoff.$cutoffval.fasta*, where *$cutoffval* is *e.g.* “10”). The error-corrected reads from the diseased plant (*Assembly.ec.fastq*) are then aligned to the contigs in that last file using BWA. From the alignment file (*Stage1.$cutoffval.match.sam*) the reads above the cutoff are extracted and exported in a FASTQ file (*Stage1.$cutoffval.match.fastq*).

Stage 2: re-alignment and filtering. The reads from the cutoff (*Stage1.$cutoffval.match.fastq*) are now aligned with *BWA* against the healthy plant reference (*Healthy.contigs.fasta*) and a FASTQ file with the reads that do not align is exported (*Stage2.$cutoffval.nonmatch.fastq*). These non-aligned reads are assembled with the A5 pipeline (*Stage3.$cutoffval.contigs.fasta*).

Stage 3: Blast. A Blast nucleotide database is created from the reference healthy plant file (*Healthy.contigs.fasta,* which could also be a combination of different references) and used to query the contigs outputted by the previous stage (*Stage3.$cutoffval.contigs.fasta*) using *tblastx* (translated nucleotide query *vs.* translated nucleotide database Blast). The results are saved in a text file (*Stage3.$cutoffval.contigs.csv*), which is then filtered according to the identity percentage (IP): entries with an IP greater than 95% are attributed to the plant (*Stage3.$cutoffval.contigs.plant.csv*), while those with a lower IP are attributed to the phytoplasma (*Stage3.$cutoffval.contigs.phyto.csv*). Using this last file the contigs pertaining to the phytoplasma are extracted from the query and saved in a FASTA file (*Stage3.$cutoffval.phyto.fasta*).

Stage 4: clean-up. Lastly, the main outputs are compressed in the *gzip* format, moved to a folder (*Results_$timestamp*), statistical data such as contigs size and number are calculated, while the intermediate files are moved to a sub-folder (*Other_files*), which also contains the assembly of the reference and/or the diseased plant reads, unless skipped in Stage 0. If the user did not input a cutoff value, the *Results* folder will contain files for cutoff 0, the calculated maximum value and half of the maximum.

A flow chart of the *Phytoassembly* pipeline is provided as supplementary Figure 6.3.1.

Source of data

Genome assemblies of '*Ca.* Phytoplasma asteris', strain Aster Yellows Witches'-Broom (AYWB; Bai *et al.,* 2006; accession number CP000061), Milkweed Yellows phytoplasma (MW1; (Saccardo *et al.*, 2012); accession number AKIL00000000), Italian Clover Phyllody phytoplasma (MA1; (Saccardo *et al.*, 2012); accession number AKIM00000000), Vaccinium Witches' Broom phytoplasma (VAC; (Saccardo *et al.*, 2012); accession number AKIN00000000) and Poinsettia branch-inducing phytoplasma strain JR1 (JR1; (Saccardo *et al.*, 2012); accession number AKIK00000000) were downloaded from the NCBI database. The ILLUMINA reads data-sets of MW1 and MA1, and from '*Ca.* Phytoplasma aurantifolia' strain Witches' Broom of Lime 2034 (WBDL; Siqueira Alves *et al.*, submitted), Cassava Frogskin Disease associated phytoplasma (CFSD; Neves *et al.*, manuscript in preparation) and Chicory

Phyllody associated phytoplasma (ChiP2; Martini *et al.*, in preparation) were provided by the authors of the cited papers.

Simulations and further data analysis

Comparisons of the assemblies were carried out using BUSCO (Simão *et al.*, 2015), *MUMmer* (Delcher *et al.*, 2002), and OMA (Altenhoff *et al.*, 2015). To benchmark the pipeline, a sequencing experiment was simulated from an existing complete phytoplasma genome. Artificial sequence reads were generated from a complete sequencing of AYWB, using an ad-hoc Perl script that introduces reading errors and combines the phytoplasma and the plant reads. Reads obtained from a healthy periwinkle in a previous work ((Saccardo *et al.*, 2012); SRA accession number SRS356159) were combined with the artificially generated reads, so that phytoplasma reads resulted in adding 5%, 10% and 15% proportions to the plant reads.

3.1.4 Results

Validation

As presented in the introduction, the procedure described here exploits the different coverages of pathogen and host contigs resulting for a preliminary assembly of the ILLUMINA reads. Figure 3.1.1 shows a coverage graphs of the contigs resulting from a preassembly of an ‘artificial’ dataset generated from the genome of AYWB, and mixed in proportion of 15% to real ILLUMINA reads from a healthy periwinkle. Although the two peaks corresponding to the host and pathogen contigs are clearly distinguishable in the graph, maximizing the recovery of the pathogen data in order to obtain the most complete genome reconstruction requires the estimation and use of an inclusive, cautious cutoff value. We found that an optimal cutoff value can be estimated as 0.3 times the ratio between the sum of the lengths of the non-mapping reads at cutoff 0 and the sum of the lengths of the error corrected reads, multiplied by 100. To test the robustness of the pipeline with this estimate, we performed a number of tests using artificial and real datasets.

First, the pipeline was run for cutoff values between 0 and 15 with various simulated datasets and the size of the resulting final assemblies evaluated (Figure 3.1.2). With optimized cutoff the pipeline recovered 88.1% (with 5% of phytoplasma reads and cutoff 2), 94.2% (with 10% of phytoplasma reads and cutoff 4) and 93.9% (with 15% of phytoplasma reads and cutoff 5) of the original AYWB sequence. The number of reconstructed genes (including partials) was 711, 666 and 666, respectively, compared to 534 in the actual AYWB genome. The higher value of the gene number in the assemblies generated by the pipeline was due to the fragmentation of genes located at contigs ends.

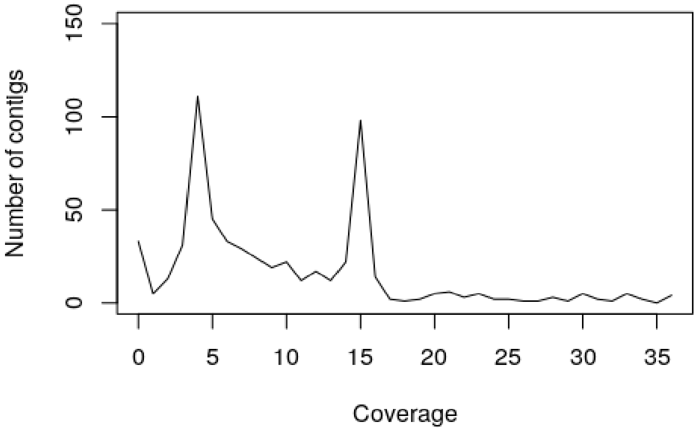


Figure 3.1.1 – Coverage graph of the artificial aster yellows phytoplasma strain witches' broom sample pre-assembly. The graph, from a dataset with 15% of phytoplasma reads, illustrates the position of the plant (left) and the phytoplasma (right) peaks. The optimal cutoff site determined by Phytoassembly falls between the two peaks. On the x-axis is the per-contig coverage, calculated as the ratio between the sum of the lengths of the reads aligned on the contig and the length of the contig, multiplied by 100. On the y-axis is the number of contigs with similar coverage. The plant peak has 111 contigs at coverage 4, the phytoplasma peak has 98 contigs at coverage 15.

As a quality evaluation, we compared the genes found in the complete AYWB genome with those in the assembly generated by the pipeline from the dataset with 10% of phytoplasma reads and cutoff 4 using OMA. According to the results, 59 genes of AYWB did not have an identical counterpart in the *Phytoassembly* reconstructed genome. However, 20 of those genes showed >95% identity with a gene in the AYWB genome, the differences being due to misassembly of genes that are present in multiple, non identical, copies. The remaining 39 genes (7%) were all annotated as hypothetical proteins or phage associated proteins, and were characterized by low complexity in sequence. In conclusion the pipeline provided suitable data for the complete reconstruction of the genetic features of the AYWB phytoplasma, failing only in areas of the genome with low complexity likely associated with phage integrations.

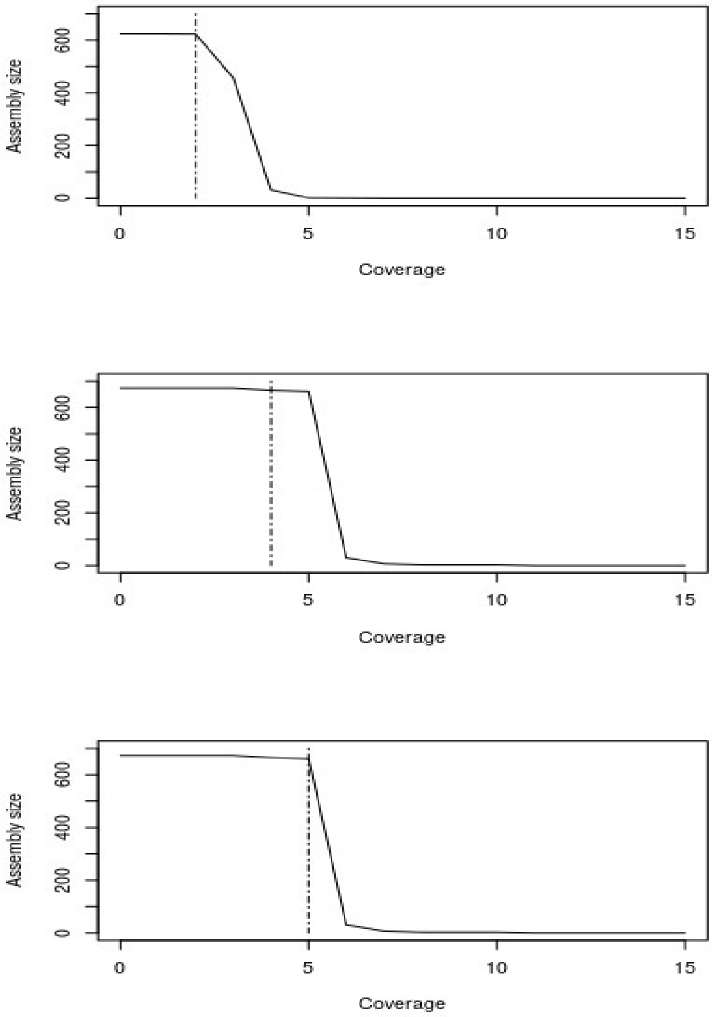


Figure 3.1.2 – Size (in knts) of the artificial aster yellows phytoplasma strain witches' broom (AYWB) sequences resulting from the use of different cutoff values. Datasets have phytoplasma/plant reads ratio of 5% (top), 10% (middle), or 15% (bottom). The vertical line shows the optimal cutoff determined by Pythoassembly. blast filtering did not remove any sequence from the output.

A second test used actual ILLUMINA reads of MW1 and MA1, and the results were compared with the previously obtained assemblies (Saccardo *et al.*, 2012). The reference genome used was a *Velvet* (https://www.ebi.ac.uk/~zerbino/velvet/) assembly from ILLUMINA reads of periwinkle. The reconstructed assembly of MW1 was 632,844 nts long without cutoff and 631,878 nts long with a 10 cutoff (222 contigs), while the 2012 assembly comprised 583,806 nts

(158 contigs) (Table 3.1.1); the minimum size of the contigs in the *Phytoassembly* reconstructions is 307 nts (N50 6,099 nts), while in the 2012 one is 231 nts (N50 7,972 nts). The reconstructed assembly of MA1 was 710,075 nts long without cutoff and 708,886 nts long with a 10 cutoff (299 contigs), while the previously obtained assembly comprised 597,245 nts (197 contigs); the minimum size of the contigs in the *Phytoassembly* assembly is 188 nts and 184 nts (N50 10,390 nts and 10,407 nts), while in the 2012 one is 230 nts (N50 12,309 nts). The MW1 assemblies differ on 128 contigs, 308–5477 nts in size; MA1 assemblies differ on 35 contigs, 299–1227 nts in size.

Table 3.1.1 – Data relative to draft phytoplasma assemblies obtained with *Phytoassembly*.

	Nucleotides	Contigs	Min. size	Max. size	N50 size	N50 contigs	G+C
AYWB reference	706,569	1	706,569	706,569	706,569	1	27%
AYWB 5% cutoff 0	624,492	242	398	21,808	3,987	47	27%
AYWB 5% cutoff 2	622,737	243	398	21,808	3,845	47	27%
AYWB 10% cutoff 0	673,019	111	407	137,058	30,483	7	27%
AYWB 10% cutoff 4	665,375	95	559	137,058	30,472	7	27%
AYWB 15% cutoff 0	664,899	95	512	90,316	28,048	8	27%
AYWB 15% cutoff 5	663,628	97	500	87,545	25,058	9	27%
Milkweed Yellows ph. (MW1) reference	583,806	158	231	22,485	7,972	26	27%
Phytoassembly MW1, cutoff 0	632,844	224	308	22,483	6,099	32	28%
Phytoassembly MW1, cutoff 10	631,878	222	307	22,483	6,099	32	28%
Italian Clover Phyllody ph. (MA1) reference	597,245	197	230	40,778	12,309	16	27%
Phytoassembly MA1, cutoff 0	710,075	296	188	39,685	10,390	20	27%
Phytoassembly MA1, cutoff 10	708,886	299	184	39,685	10,407	20	27%
Cassava Frogskin Disease (CFSD)	818,980	293	311	35,791	7,796	28	29%
'*Ca.* Phytoplasma Aurantifolia' (WBDL)	794,372	182	602	56,244	13,769	17	28%
Chicory Phyllody (ChiP2) raw	1,931,149	370	605	83,360	11,391	35	26%
Chicory Phyllody (ChiP2)	547,918	138	605	25,180	4,832	30	25%

To assess the completeness of the MA1 and MW1 genome reconstructions by *Phytoassembly*, the assemblies were checked for missing conserved genes, using BUSCO. Running the program with the set of 14 phytoplasma genome drafts used in (Firrao *et al.*, 2013), we generated an *ad hoc* list comprising a subset of 77 BUSCOs (conserved genes) that are common to all phytoplasma genomes. As shown in Table 3.1.2, one gene was missing in the assembly of MW1 and two genes were missing in the assembly of MA1. It was therefore estimated that *Phytoassembly* can recover >95% of the coding information of the sampled genomes.

Table 3.1.2 – Conserved genes missing from new genome drafts built by *Phytoassembly*.

Assembly	Missing BUSCOs	Description
MA1	POG090A00A0 POG090A001V	tRNA uridine 5-carboxymethylaminomethyl modification protein ribosomal protein S15
MW1	POG090A019O	signal recognition particle protein Srp54
CSFD	None	
CHIP2	POG090A00VB POG090A012Q	transcription termination/antitermination factor NusG ribosomal protein L35
WBDL	POG090A00FL	Elongation factor G

Novel drafts

Using this pipeline, high quality draft assemblies of the WBDL, CFSD, and ChiP2 were obtained. The size of the assemblies varied from about 550,000 to about 800,000 nts (Table 3.1.1).

Each of the phytoplasma genomes reconstructed by *Phytoassembly* was analyzed along with the four complete phytoplasma genomes available (Oshima *et al.*, 2004; Bai *et al.*, 2006; Kube *et al.*, 2008; Tran-Nguyen *et al.*, 2008) using standalone OMA, in order to identify shared orthologs. 274 'shared' orthologs are present in all of the four phytoplasma genomes.

The CFSD sample was processed using a healthy cassava sample, obtaining a phytoplasma genome assembly of 818,980 nts in 293 contigs, ranging from 311 to 35,791 nts in length (see Table 3.1.1 for a full comparison between the samples). This sample shares 457 orthologs with at least one of the four phytoplasmas, and 247 with all of them.

The WBDL sample was processed with an ensemble of *Citrus sinensis* and *Citrus clementina,* because an isogenic reference was not available. After annotation, the phytoplasma genome assembly was 794,372 nts long divided in 182 contigs, ranging from 602 to 56,244 nts. This sample shares 479 orthologs with at least one of the four phytoplasmas, and 220 with all of them. An additional about 1,000,000 nts long set of small contigs could not be attributed to the phytoplasma nor to the plant, as they were not represented in the available *Citrus* genomes, but are assumed to be specific lime repeated sequences.

The ChiP2 sample was processed using the healthy periwinkle specimen (see MW1 and MA1 above), obtaining an assembly of 1,931,149 nts. The output of the pipeline was consistently oversized for a phytoplasma, which rarely exceeds 10^6 nts. It was therefore annotated using RAST (Aziz *et al.*, 2008), and the result showed that 1,338,982 nts (69.3%) actually belonged to a spiroplasma, while the true phytoplasma genome was 547,918 nts (28.4%), assembled in 138 contigs, ranging from 605 to 25,180 nts.

The check for draft completeness, carried out with BUSCO and the *ad hoc* conserved gene list revealed, as shown in Table 3.1.2, that no conserved genes were missing in the CSFD assembly, one gene was missing in the assembly of WBDL, and two genes were missing in the assembly of Chip2.

3.1.5 Discussion

The *Phytoassembly* pipeline successfully addresses the problem of obtaining the genomic sequences of phytoplasmas, by selectively excluding the reads of the host plant from a infected plant ILLUMINA sequence data-set. It does so by first by filtering out reads with low coverage, which can be assumed to belong to the plant, because of the vast disparity in coverage between the plant and the pathogen genome; then by removing the reads that can be aligned on the healthy plant genome.

As an improvement of the procedure developed in (Saccardo *et al.*, 2012), which required *ad hoc* tuning and various manual or external steps for the *de novo* assembly, *Phytoassembly* can carry out autonomously the complete analysis, and relies on an assembler (the A5 pipeline) which doesn't require additional input from the user. The assembler is tailored for ILLUMINA reads, and works with paired-ends.

The sequences that pass the re-alignment step are those that do not map on the healthy plant reference, therefore they can only belong to genes not attributable to the plant host. While the main aim of the *Phytoassembly* procedure is the isolation of phytoplasma genes, by virtue of the mechanism employed it can also isolate other non-culturable pathogens, or mask specific pathogens by adding their genomes to the healthy plant reference.

The pipeline attempts to determine a cutoff value using the ratio between the total length of the non-mapping reads at cutoff 0 and the error corrected reads of the diseased plant. This ratio was chosen because the error corrected reads exclude any ambiguous or unreliable data from the estimation, and the non-mapping reads represent a fraction roughly proportional to the pathogen quota in the sequencing. Using the value as is, however, leads to an excessive cutoff. Plotting the nucleotide count of the phytoplasma reconstructions at various cutoffs (Figure 3.1.2), a common feature is a significant drop after a value that appears correlated to the percentage of pathogen genome in the diseased plant specimen. Based on the results of the artificial reads test, a more conservative estimation is obtained by using 1/3 of the aforementioned ratio.

An alternative method to determine the optimal value would be to run the pipeline at cutoff 0, increasing the value until the last estimation has a significant drop (in the order of more than 1000 nts) in the reconstructed genome size. This however can increase the computation time significantly, while the chosen method repeats the procedure only once. Testing different values is still allowed, simply by inputting the minimum and maximum values and the distance between the cutoffs (*e.g.* from 3 to 12, with step 3, produces cutoffs 3, 6, 9, 12).

In conclusion, *Phytoassembly* is a focused tools that allows a user-frendly and performant processing of ILLUMINA sequence data from a pair of samples, a phytoplasma infected plant sample and its uninfected reference sample, outputting a high quality genome draft of the pathogen. Given the increasing availability of access to ILLUMINA technology, *Phytoassembly* is expected to be a valuable help in the characterization of the genomes of the large, diverse and economically relevant group of plant pathogens that belong to the genus '*Ca*. Phytoplasma'.

The *Phytoassembly* source code is available on GitHub at https://github.com/cpolano/ /phytoassembly .

3.1.6 References

Altenhoff, A. M. *et al.* (2015) 'The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements', *Nucleic Acids Research*, 43(D1), pp. D240–D249. doi: 10.1093/nar/gku1158.

Andersen, M. T. *et al.* (2013) 'Comparison of the complete genome sequence of two closely related isolates of "Candidatus Phytoplasma australiense" reveals genome plasticity', *BMC Genomics*. BMC Genomics, 14(1), p. 529. doi: 10.1186/1471-2164-14-529.

Aziz, R. K. *et al.* (2008) 'The RAST Server: Rapid Annotations using Subsystems Technology', *BMC Genomics*, 9(1), p. 75. doi: 10.1186/1471-2164-9-75.

Bai, X. *et al.* (2006) 'Living with genome instability: the adaptation of phytoplasmas to diverse environments of their insect and plant hosts.', *Journal of bacteriology*. American Society for Microbiology, 188(10), pp. 3682–96. doi: 10.1128/JB.188.10.3682-3696.2006.

Casati, P. *et al.* (2011) 'Multiple gene analyses reveal extensive genetic diversity among "Candidatus Phytoplasma mali" populations', *Annals of Applied Biology*. Blackwell Publishing Ltd, 158(3), pp. 257–266. doi: 10.1111/j.1744-7348.2011.00461.x.

Chen, W. *et al.* (2014) 'Comparative Genome Analysis of Wheat Blue Dwarf Phytoplasma, an Obligate Pathogen That Causes Wheat Blue Dwarf Disease in China', *PLoS ONE*. Edited by M. Gijzen. Public Library of Science, 9(5), p. e96436. doi: 10.1371/journal.pone.0096436.

Chung, W.-C. *et al.* (2013) 'Comparative Analysis of the Peanut Witches'-Broom Phytoplasma Genome Reveals Horizontal Transfer of Potential Mobile Units and Effectors', *PLoS ONE*. Edited by M. Robinson-Rechavi. Public Library of Science, 8(4), p. e62770. doi: 10.1371/journal.pone.0062770.

Cimerman, A., Arnaud, G. and Foissac, X. (2006) 'Stolbur phytoplasma genome survey achieved using a suppression subtractive hybridization approach with high specificity.', *Applied and environmental microbiology*. American Society for Microbiology, 72(5), pp. 3274–83. doi: 10.1128/AEM.72.5.3274-3283.2006.

Davis, R. E. *et al.* (2013) '"Candidatus Phytoplasma pruni", a novel taxon associated with X-disease of stone fruits, Prunus spp.: multilocus characterization based on 16S rRNA, secY, and ribosomal protein genes', *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY*. Microbiology Society, 63(Pt 2), pp. 766–776. doi: 10.1099/ijs.0.041202-0.

Delcher, A. L. *et al.* (2002) 'Fast algorithms for large-scale genome alignment and comparison.', *Nucleic acids research*, 30(11), pp. 2478–83.

Firrao, G. *et al.* (2013) ‘Genome wide sequence analysis grants unbiased definition of species boundaries in “Candidatus Phytoplasma”’, *Systematic and Applied Microbiology*. Elsevier GmbH., 36(8), pp. 539–548. doi: 10.1016/j.syapm.2013.07.003.

Garcia-Chapa, M. *et al.* (2004) ‘PCR-mediated whole genome amplification of phytoplasmas’, *Journal of Microbiological Methods*, 56(2), pp. 231–242. doi: 10.1016/j.mimet.2003.10.010.

Kawar, P. G. *et al.* (2010) ‘Identification and Isolation of SCGS Phytoplasma-specific Fragments by Riboprofiling and Development of Specific Diagnostic Tool’, *Journal of Plant Biochemistry and Biotechnology*. Springer India, 19(2), pp. 185–194. doi: 10.1007/BF03263339.

Kube, M. *et al.* (2008) ‘The linear chromosome of the plant-pathogenic mycoplasma “Candidatus Phytoplasma mali”’, *BMC Genomics*. BioMed Central, 9(1), p. 306. doi: 10.1186/1471-2164-9-306.

Lee, I.-M., Davis, R. E. and Gundersen-Rindal, D. E. (2000) ‘Phytoplasma: Phytopathogenic Mollicutes’, *Annual Review of Microbiology*. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA, 54(1), pp. 221–255. doi: 10.1146/annurev.micro.54.1.221.

Li, H. and Durbin, R. (2009) ‘Fast and accurate short read alignment with Burrows-Wheeler transform’, *Bioinformatics*. Oxford University Press, 25(14), pp. 1754–1760. doi: 10.1093/bioinformatics/btp324.

Liefting, L. W. and Kirkpatrick, B. C. (2003) ‘Cosmid cloning and sample sequencing of the genome of the uncultivable mollicute, Western X-disease phytoplasma, using DNA purified by pulsed-field gel electrophoresis’, *FEMS Microbiology Letters*. Oxford University Press, 221(2), pp. 203–211. doi: 10.1016/S0378-1097(03)00183-6.

Marcone, C., Ragozzino, A. and Seemuller, E. (1997) ‘Dodder transmission of alder yellows phytoplasma to the experimental host Catharanthus roseus (periwinkle)’, *Forest Pathology*, 27(6), pp. 347–350. doi: 10.1111/j.1439-0329.1997.tb01449.x.

Oshima, K. *et al.* (2004) ‘Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma’, *Nature Genetics*. Nature Publishing Group, 36(1), pp. 27–29. doi: 10.1038/ng1277.

Quaglino, F. *et al.* (2013) ‘“Candidatus Phytoplasma solani”, a novel taxon associated with stolbur- and bois noir-related diseases of plants’, *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY*. Microbiology Society, 63(Pt 8), pp. 2879–2894. doi: 10.1099/ijs.0.044750-0.

Quaglino, F. *et al.* (2015) ‘“Candidatus Phytoplasma phoenicium” associated with almond witches’-broom disease: from draft genome to genetic diversity among strain populations', *BMC Microbiology*. BioMed Central, 15(1), p. 148. doi: 10.1186/s12866-015-0487-4.

Saccardo, F. *et al.* (2012) 'Genome drafts of four phytoplasma strains of the ribosomal group 16SrIII', *Microbiology*. Microbiology Society, 158(Pt_11), pp. 2805–2814. doi: 10.1099/mic.0.061432-0.

Saeed, E. *et al.* (1994) 'Molecular Cloning, Detection of Chromosomal DNA of the Mycoplasmalike Organism (MLO) Associated with Faba Bean (Vicia faba L.) Phyllody by Southern Blot Hybridization and the Polymerase Chain Reaction (PCR)', *Journal of Phytopathology*. Blackwell Publishing Ltd, 142(2), pp. 97–106. doi: 10.1111/j.1439-0434.1994.tb04519.x.

Simão, F. A. *et al.* (2015) 'BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs', *Bioinformatics*, 31(19), pp. 3210–3212. doi: 10.1093/bioinformatics/btv351.

Tran-Nguyen, L. T. T. *et al.* (2008) 'Comparative genome analysis of "Candidatus Phytoplasma australiense" (subgroup tuf-Australia I; rp-A) and "Ca. Phytoplasma asteris" Strains OY-M and AY-WB.', *Journal of bacteriology*. American Society for Microbiology, 190(11), pp. 3979–91. doi: 10.1128/JB.01301-07.

Tran-Nguyen, L. T. T. and Gibb, K. S. (2007) 'Optimizing Phytoplasma DNA purification for genome analysis.', *Journal of biomolecular techniques : JBT*, 18(2), pp. 104–12. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17496222.

Tritt, A. *et al.* (2012) 'An Integrated Pipeline for de Novo Assembly of Microbial Genomes', *PLoS ONE*. Edited by D. Zhu, 7(9), p. e42304. doi: 10.1371/journal.pone.0042304.

Weintraub, P. G. and Beanland, L. (2006) 'Insect Vectors of Phytoplasmas', *Annual Review of Entomology*. Annual Reviews, 51(1), pp. 91–111. doi: 10.1146/annurev.ento.51.110104.151039.

Zhao, Y., Davis, R. E. and Lee, I.-M. (2005) 'Phylogenetic positions of "Candidatus Phytoplasma asteris" and Spiroplasma kunkelii as inferred from multiple sets of concatenated core housekeeping proteins', *International Journal of Systematic and Evolutionary Microbiology*. Microbiology Society, 55(5), pp. 2131–2141. doi: 10.1099/ijs.0.63655-0.

Zonneveld, B. J. M., Leitch, I. J. and Bennett, M. D. (2005) 'First Nuclear DNA Amounts in more than 300 Angiosperms', *Annals of Botany*. Oxford University Press, 96(2), pp. 229–244. doi: 10.1093/aob/mci170.

3.2 Metagenomics highlighted mixed infection of spiroplasma and phytoplasma in chicory

Authors: Polano C., Moruzzi S., Ermacora P., Ferrini F., Martini M., Firrao G.

Manuscript in preparation

3.2.1 Summary

Phytoplasma disease symptoms were observed on chicory in a restricted area near Carlino (North East Italy). Preliminary analyses demonstrated the presence of a phytoplasma belonging to pigeon pea witches' broom (16SrIX) group. Using ILLUMINA sequencing, we obtained a high quality draft of the genome of the phytoplasma associated with chicory phyllody (ChiP), consisting in an assembly of 126 contigs for a total length of 547,918 nucleotides. The assembly allowed a clearer look to the genome-wide phylogeny of the 16SrIX group and to the secreted protein potential and diversity of the genome. While carrying out the assembly of the phytoplasma it became evident that the sample used for the analysis was mixed infected by a phytoplasma and a spiroplasma. Preliminary field sampling actually confirmed that the two organisms occur frequently in mixed natural infections of chicory.

3.2.2 Introduction

In 2011 a severe outbreak of chicory phyllody has been reported in the Carlino area (Udine province) in Friuli Venezia Giulia region (FVG, North-eastern Italy). Plants of *Cichorium intybus* L., (chicory, family *Asteraceae*) showed symptoms of phyllody, virescence and proliferation of axillary buds. Molecular characterization of chicory phyllody (ChiP) phytoplasma strains (Martini et al., 2012; Ermacora et al., 2013) based on the three genes 16S rDNA, ribosomal protein (*rp, rpl22* and *rps3*) and *secY* showed that all the strains were nearly identical and were closely related to strains PEY (*Picris echioides* yellows) and *NaxY* (*Naxos* periwinkle virescence), belonging to 16SrIX-C, rp(IX)-C1 and secY(IX) C1 subgroups (Lee et al., 2012).

Since the genome of the 16SrIX group phytoplasma is poorly known, in this paper we report on our use of a metagenomic approach to obtain a genome draft of the phytoplasma causing chicory phyllody (ChiP).

3.2.3 Materials and methods

PCR amplification. Chicory phyllody phytoplasma (16SrIX-C) specific primers based on rp (rpl22-rps3) and *secY* gene sequence alignments of 16SrIX phytoplasma strains have been used for the diagnostic direct and nested PCR amplification of chicory phyllody phytoplasma DNA as reported by (Martini *et al.*, 2012), according to the published method. Spiroplasmas infection in field chicory plants was assessed using a set of primers for spiral in gene PCR amplification developed by Martini *et al.* (manuscript in preparation).

Illumina sequencing. A total of 10 mg DNA from each sample was fragmented by incubation for 70 min with 5 µl dsDNA Fragmentase (New England Biolabs). The following steps in library preparation were carried out as described elsewhere (Marcelletti *et al.*, 2011). The samples were run on an ILLUMINA MySeq that provided paired reads of 300 nt in length, at the Istituto di Genomica Applicata (Udine, Italy).

Phylogenetic analysis. In order to provide a solid alignment of DNA sequence a multistep procedure was set up with the development of a set of ad hoc PERL scripts. To assess the completeness of the MA1 and MW1 genome reconstructions by *Phytoassembly*, the assemblies were checked for missing conserved genes using BUSCO (Simão *et al.*, 2015), which was run with a set of 8 phytoplasma genome drafts used to generate an *ad hoc* list comprising a subset BUSCOs (conserved genes) that are common to all phytoplasma genomes investigated. Orthologous groups that contained more than one protein for at least one genome (paralogs) were not discarded. Then the alignments in each othologous group were split, sorted and re-merged in order to identify and exclude alignment regions that contained a number of gaps higher than a cutoff (10 gaps/50 aa. positions) and that could therefore be of uncertain alignment. The protein alignments were analyzed individually and as a concatenated sequence. Alignment inspection and preliminary analyses were carried out with SEAVIEW (Gouy et al., 2010).

Maximum likelihood analysis was carried out with PHYML (Guindon and Gascuel, 2003), using LC as a substitution model for protein sequence analysis, respectively. Tree topologies were estimated using the better topology obtained using Nearest Neighbor Interchange (NNI) or Subtree Pruning and Regrafting (SPR). A most parsimonious tree was used as input tree. The support of the data for each internal branch of the phylogeny was estimated using non-parametric bootstrap with 100 replicates.

Concatenated gene sequence data were also analyzed using split networks with the aid of the software SPLITTREE4 (Huson and Bryant, 2006). Split networks are used to represent incompatible and ambiguous signals in a data set. The median network of all most parsimonious trees used here, is depicted as a tree with additional edges, so that the distance between two taxa is equal to the length of the shortest path connecting them (Bandelt *et al.*, 1995). It is therefore capable of highlighting taxa relationships that are not tree-like, taking in account polytomy at branching points, i.e. the fact that one sequence may share identities with a sequence that is more distant in the tree in positions where its neighbour sequence(s) differ(s).

For the construction of a consensus network, trees from individual protein sequence alignments were obtained by recursively running PHYML using NNI, then processed with SPLITTREE4 using a median network construction (Holland *et al.*, 2004). In these split networks, the lengths of the edges are proportional to the number of gene trees in which a particular edge occurs. Thus, the presence of boxes in the networks indicates contradictory evidence for grouping.

3.2.4 Results and discussion

DNA sequence analysis of the Chicory with Phyllody symptoms

Field collected samples displaying phyllody symptoms were preliminary analyzed by PCR and qPCR using phytoplasma specific primers as reported elsewhere (Martini *et al.*, 2012). One sample that, according to qPCR, resulted to contain >2% phytoplasma DNA was further processed, as described in methods, for ILLUMINA genome sequencing. As a results, 3,360,210 paired reads, for 1,009,810,635 nucleotides overall, were obtained.

The reads were assembled with the A5 pipeline (Tritt *et al.*, 2012) and the sequences belonging to the plant (SRA accession number SRS356159) were separated with the *Phytoassembly* pipeline (Polano and Firrao, submitted). The pipeline produced a genome draft fragmented into 390 contigs, accounting for 1,931,149 bp. overall, that was anomalously large for the expected phytoplasma genome. RAST annotation (Aziz *et al.*, 2008) clarified that the pipeline selected the sequences of two mollicutes, due to the presence in the annotation of a large number of sequences with similarity to Spiroplasma spp. genes, in addition to sequences encoding phytoplasmal typical proteins. Using an *ad hoc* Perl script, the RAST annotation result was used to sort the contigs in three batches: one with contigs that are unambiguously assigned to phytoplasmas, one with contigs that are unambiguously assigned to spiroplasmas, and one with contigs that were spurious or did not allow to differentiate the sequences as belonging to the phytoplasma (Table 3.2.1). After sorting the assemblies were separately re-submitted to RAST annotation, since the spiroplasma have a different genetic code.

Table 3.2.1 – Assemblies data for the ChiP sample, as it was obtained from *Phytoassembly* and further separated in phytoplasma and other microorganisms. Row 1 is not the sum of rows 2 and 3 because the Illumina reads were re-mapped and re-assembled after the first split.

	Nucleotides	**Contigs**	**Min. size**	**Max. size**	**G+C**
Chicory Phyllody, raw	1,931,149	370	605	83,360	26%
Chicory Phyllody, phytoplasma	541,091	134	605	25,180	25%
Chicory Phyllody, spiroplasma	1,560,885	334	621	83,360	27%

Characterization of the phytoplasma genome

The phytoplasma genome resulting after the exclusion of non-phytoplasma sequences from the assembly resulted 541,091 nucleotides, with the predicted encoding potential of 583 proteins. This size is significantly larger than the recently reported genome of '*Ca.* P. phoenicium' strain AlmWB (Almond Witches' Broom), another phytoplasma of the 16SrIX group. The Venn diagram in Figure 3.2.2 shows a comparison. However, comparing the predicted proteins of strains AlmWB and ChiP (Table 3.2.2) it became apparent that the AlmWB genome draft misses a relatively large number of proteins that are common to ChiP and all other phytoplasmas, hence the difference in size between ChiP and AlmWB genome draft sequences is due to the large incompleteness of the latter. Therefore the AlmWB genome draft was not used for further analyses.

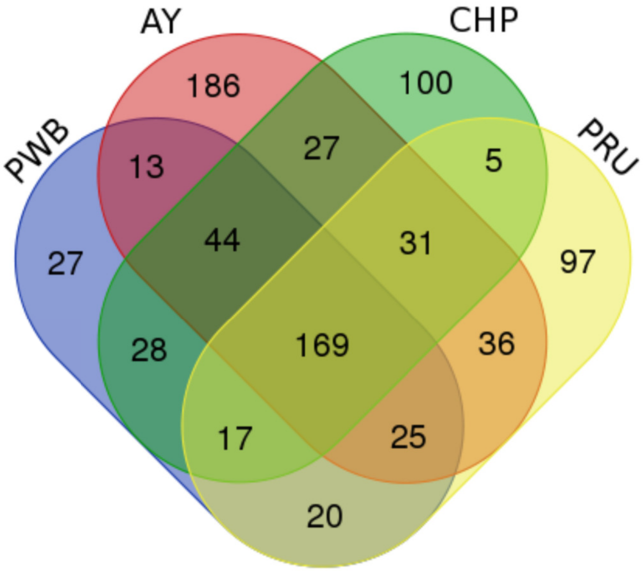


Figure 3.2.1 – Venn diagram of the orthologous gene number of four phytoplasma representative of major clades. CHP: '*Ca.* P. phoenicium'-related strain ChiP; AY: '*Ca.* P. asteris' strain Aster Yellows Witches' Broom; PWB: Peanut Witches' Broom ph.; PRU: '*Ca.* P. pruni'.

Table 3.2.2 – Protein coding potential of the genome drafts of strains AlmWB and their comparison with that of other '*Ca.* Phytoplasma' species. Values with an asterisk include duplicates.

Total number of predicted proteins for ChiP	583*
Total number of predicted proteins for AlmWB	286*
Proteins shared among AlmWB and ChiP	167
Proteins shared among AlmWB, ChiP, '*Ca.* P. asteris', '*Ca.* P. mali' and '*Ca.* P. pruni'	104
Proteins shared among ChiP, PWB, '*Ca.* P. asteris', and '*Ca.* P. pruni'	65
Proteins shared among AlmWB, PWB, '*Ca.* P. asteris', and '*Ca.* P. pruni'	10

The comparison of the orthologous gene content of ChiP with three phytoplasmas as representatives of other major clades, that is presented in the Venn diagram of Figure 3.2.2, shows that there is a well conserved set of 169 core genes shared by this set of four very diverse phytoplasmas. The similar number of genes shared between three out of four genomes is likely a balance between the process of individual reductive evolution of the phytoplasma from a centre of radiation and the result of the intense horizontal gene trafficking. It can also be observed a relatively high number of strain specific genes in all genomes but PWB.

For the analysis of the genome wide phylogeny of the strain ChiP, we selected 23 gene fragments from 16 genes of the core genome that were present and well conserved among the "*Candidatus* Phytoplasma" species compared (5285 aa with less than 100 indels), in order to construct a robust alignment despite the relevant differences in the gene sequence of the phytoplasmas. According to the analyses carried out on single or a few genes that have been reported widely in the literature, in the Maximum Likelihood phylogram of core genome concatenated gene sequences (Error: Reference source not found) ChiP phytoplasma resulted well distinct from other phytoplasmas. Strain ChiP, and thus '*Ca.* P. phoenicium', belongs to a branch of the phytoplasma evolutionary tree that includes '*Ca.* P. pruni' and PWB, although is not related to those species (strain MA1, that is related to '*Ca.* P. pruni' is included in the trees for reference and comparison). The evolution of the core genome of ChiP has been limitedly influenced by exchanges and genome hybridisation with other phytoplasmas, as shown by the phylogenetic split network of Figure 3.2.3, that has substantially a tree/like structure. The consensus analysis of 16 trees presented in Figure 3.2.4 shows the presence of some contrasting phylogenetic information at the basis of the major branch separating PRU-PHE-PWB from other phytoplasmas, but a substantial independent evolution.

A different picture emerges from the analysis of accessory gene content. In particular, a closer look to the secreted proteins as predicted by SignalP provides some hints about the extensive gene exchange among ChiP phytoplasma and other phytoplasmas. Table 3.2.3 reports the 35 proteins predicted as secreted in the ChiP phytoplasma genome and their similarities and identities with the secreted proteins that have been found in the genomes in other phytoplasmas. As

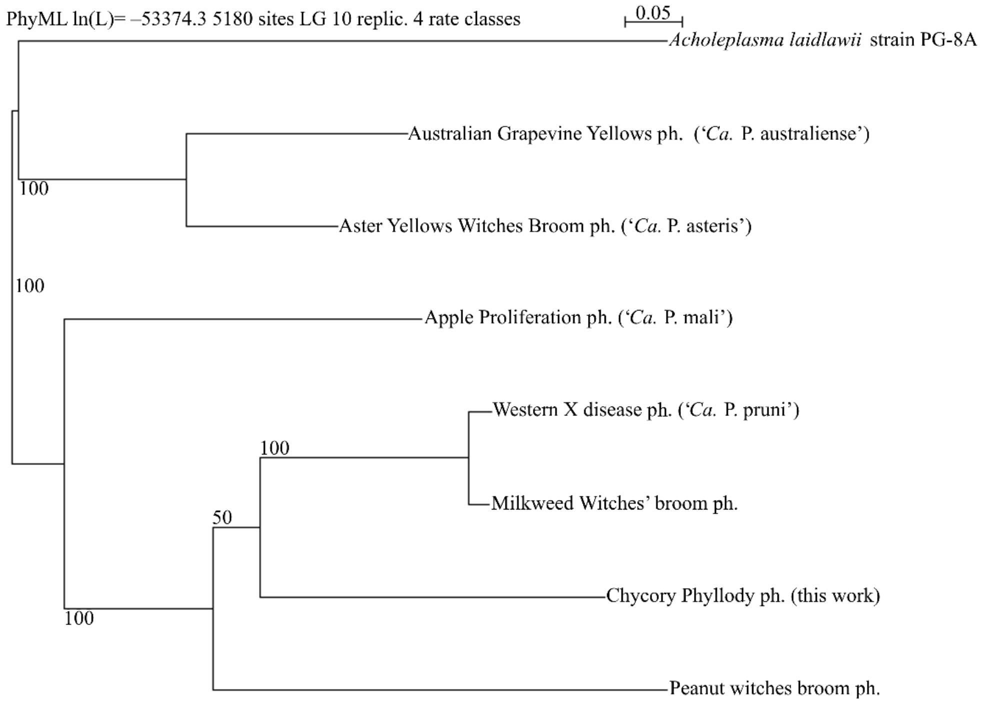

PhyML ln(L)= –53374.3 5180 sites LG 10 replic. 4 rate classes
0.05
Acholeplasma laidlawii strain PG-8A
Australian Grapevine Yellows ph. (‘*Ca.* P. australiense’)
Aster Yellows Witches Broom ph. (‘*Ca.* P. asteris’)
Apple Proliferation ph. (‘*Ca.* P. mali’)
Western X disease ph. (‘*Ca.* P. pruni’)
Milkweed Witches’ broom ph.
Chycory Phyllody ph. (this work)
Peanut witches broom ph.
100
100
100
50
100

Figure 3.2.2 – Maximum likelihood phylogram of concatenated gene sequences.

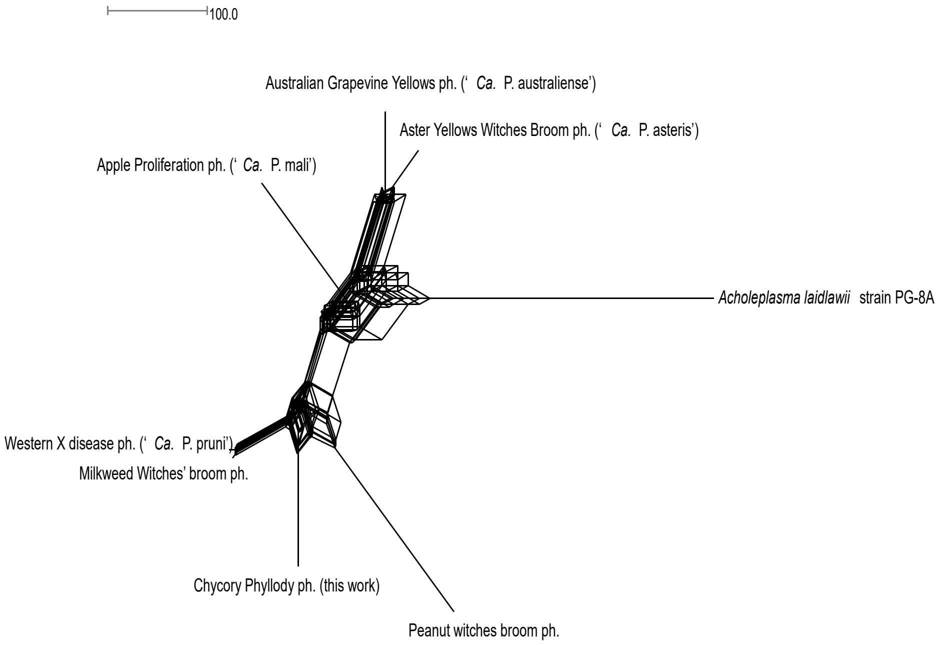

100.0
Australian Grapevine Yellows ph. (‘ *Ca.* P. australiense’)
Aster Yellows Witches Broom ph. (‘ *Ca.* P. asteris’)
Apple Proliferation ph. (‘ *Ca.* P. mali’)
Acholeplasma laidlawii strain PG-8A
Western X disease ph. (‘ *Ca.* P. pruni’)
Milkweed Witches’ broom ph.
Chycory Phyllody ph. (this work)
Peanut witches broom ph.

Figure 3.2.3 – Neighbor phylogenetic network calculated on gene concatenation alignment

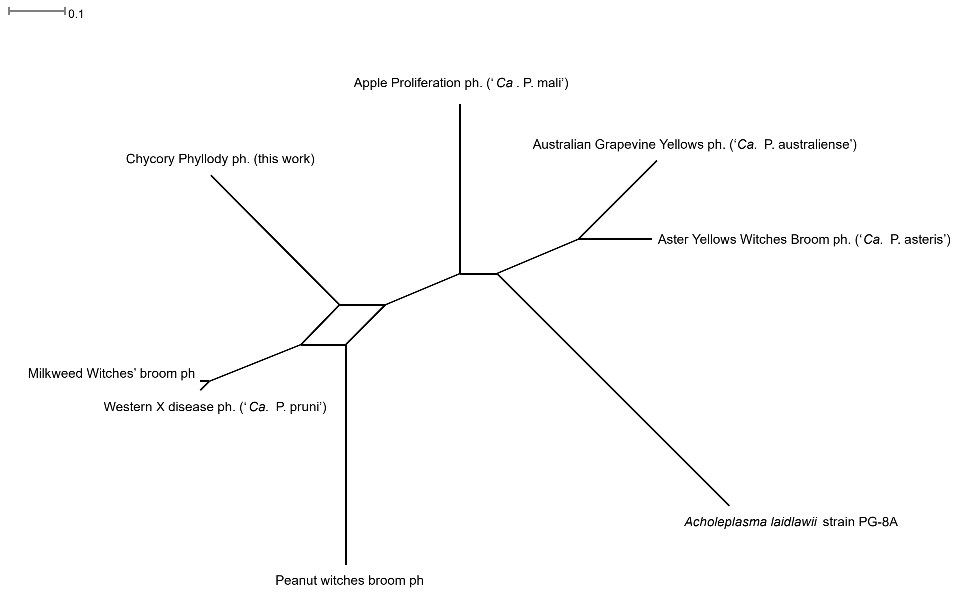

0.1
Apple Proliferation ph. ('*Ca* . P. mali')
Australian Grapevine Yellows ph. ('*Ca.* P. australiense')
Chycory Phyllody ph. (this work)
Aster Yellows Witches Broom ph. ('*Ca.* P. asteris')
Milkweed Witches' broom ph
Western X disease ph. ('*Ca.* P. pruni')
Acholeplasma laidlawii strain PG-8A
Peanut witches broom ph

Figure 3.2.4 – Consensus of 16 trees.

Table 3.2.3 – Results of blast searches of ChiP proteins predicted as secreted by SignalP against a database of phytoplasma putatively secreted proteins.

Chicory Phyllody Phytoplasma	Best matching phytoplasma	%similarity	%identity
orf00012CHP1contig19	orf04485AP1contig1	44.26	27.87
orf00010CHP1contig19	orf00010MA1contig108	51.47	23.53
orf00001CHP1contig40	orf98840AUScontig1	56.25	28.12
orf00013CHP1contig59	orf00001VACcontig1461	94.12	88.24
orf00005CHP1contig69	orf56218AY1contig1	64.84	43.96
orf00005CHP1contig78	orf98965AUScontig1	58.33	33.33
orf00002CHP1contig80	orf04091AP1contig1	80.63	66.14
orf00005CHP1contig86	orf56572AY1contig1	83.72	71.51
orf00002CHP1contig90	orf99298AUScontig1	95.71	87.14
orf00007CHP1contig90	orf50623OYMcontig1	96.15	91.03
orf00009CHP1contig90	orf50621OYMcontig1	96.05	88.16
orf00001CHP1contig101	orf00009JR1contig524	60.98	41.46
orf00007CHP1contig118	orf98797AUScontig1	45.00	32.50
orf00001CHP1contig139	orf00002MA1contig512	97.14	91.43
orf00006CHP1contig139	orf50770OYMcontig1	94.17	87.38
orf00011CHP1contig139	orf00002VACcontig1552	79.38	61.48

orf00009CHP1contig179	orf00002MA1contig512	97.14	91.43
orf00003CHP1contig185	orf00001MA1contig172	66.28	46.51
orf00001CHP1contig202	orf00002MA1contig512	97.14	91.43
orf00003CHP1contig223	orf00001JR1contig3782	51.06	38.30
orf00002CHP1contig223	orf98938AUScontig1	80.87	67.83
orf00002CHP1contig247	orf04489AP1contig1	50.00	31.48
orf00008CHP1contig265	orf00001VACcontig1461	94.12	88.24
orf00003CHP1contig321	orf99065AUScontig1	84.25	70.08
orf00002CHP1contig349	orf50301OYMcontig1	95.60	93.41
orf00002CHP1contig350	orf50645OYMcontig1	93.51	89.23
orf00008CHP1contig354	orf00002VACcontig1552	93.33	76.19
orf00004CHP1contig445	orf00002JR1contig142	45.14	29.86
orf00001CHP1contig509	orf00001JR1contig2836	47.25	28.57
orf00003CHP1contig635	orf00001VACcontig1461	97.69	94.10
orf00003CHP1contig851	orf50608OYMcontig1	48.84	37.21
orf00001CHP1contig891	orf00002VACcontig1552	87.85	73.83
orf00002CHP1contig1867	orf00002VACcontig1552	81.01	69.38
orf00003CHP1contig2350	orf50443OYMcontig1	64.71	35.29
orf00001CHP1contig5435	orf00002VACcontig1552	85.98	71.96

shown, similarities are very high, and in some cases the proteins are nearly identical to those of phytoplasmas of the 16SrIII or 16SrII clades that, as shown above, are very distantly related phylogenetically and genomically.

Characterization of the spiroplasma genome

The second genome draft obtained and annotated in this work resulted 1,560,885 bp in size, assembled into 334 contigs, with an N50 of 13807 bp. According to a preliminary analysis based on 16S rDNA sequence (not shown), the genome belonged to a *Spiroplasma* sp. that is part of the Citri-Chrysopicola-Mirum phylogentic clade, as defined by Lo and coworkers (Lo *et al.*, 2013). Most of the plant pathogenic Spiroplasma species were assigned to this clade, including *S. citri*, that causes the Citrus Stubborn Disease (Saglio et al., 1973) and *S. kunkelii*, that causes the Corn Stunt Disease (Whitcomb *et al.*, 1986).

The drafted genome was compared with the 8 genomes presently available for the Citri-Chrysopicola-Mirum phylogentic clade, in order to precisely determine its taxonomic position and highlight any peculiarity in the genome content of of the strain in comparation with its relatives. Orthologous search using OMA resulted in the identification of 442 genes shared by all 8 strains; a concatenated alignment, construced with 261 partial protein sequences including a total of 73,856 aa with less than 365 prot indels, was used to build a Maximum Likelihood phylogram and a Neighbor phylogenetic network (Figure 3.2.5 and 3.2.6). According to the phylogram based on the conserved gene sequences, the Chicory hosted Spiroplasma sp. strain whose genome was drafted in this work (ChiSsp) is a close relative of *Spiroplasma citri*. The phylogenetic network has a tree-like structure, indicating indipendent evolution of the genomes, with minor, if any, within clade gene exchanges.

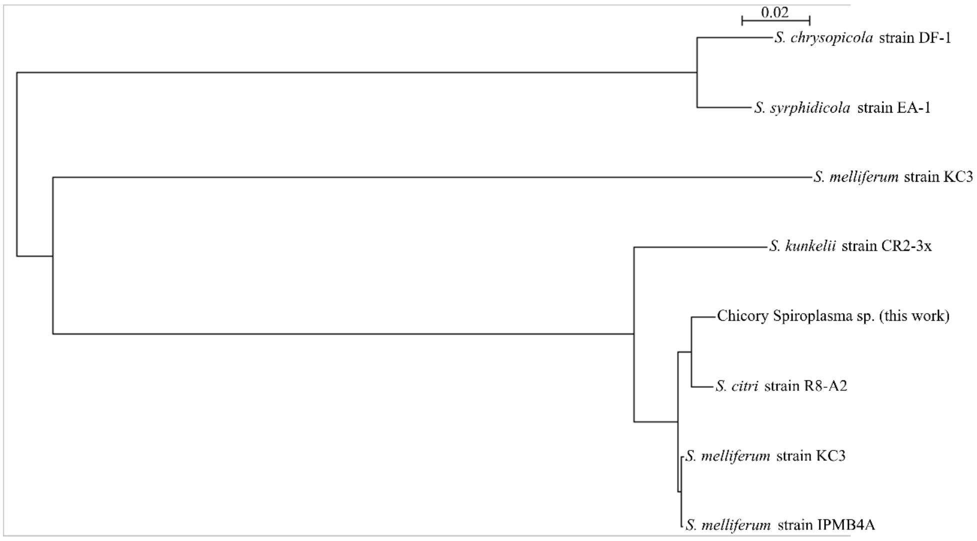

0.02
S. chrysopicola strain DF-1
S. syrphidicola strain EA-1
S. melliferum strain KC3
S. kunkelii strain CR2-3x
Chicory Spiroplasma sp. (this work)
S. citri strain R8-A2
S. melliferum strain KC3
S. melliferum strain IPMB4A

Figure 3.2.5 – Maximum Likelihood phylogram of concatenated sequences from genomes belonging to the Citri-Chrysopicola-Mirum phylogentic clade.

The consensus network (Figure 3.2.7), shows the complete congruence of the phylogenetic signal and absence of any contradictory evidence for grouping.

According to the phylogenetic analysis presented above, the genomes of our ChiSsp and recently sequenced (Davis *et al.*, 2017) genome of *Spiroplasma citri* strain R8-A2 are closely related. The alignment of the genomes (Figure 3.2.8), obtained with Mauve (Darling *et al.*, 2004) shows several genome rearrangements, despite the fragmentation of the ChiS draft into

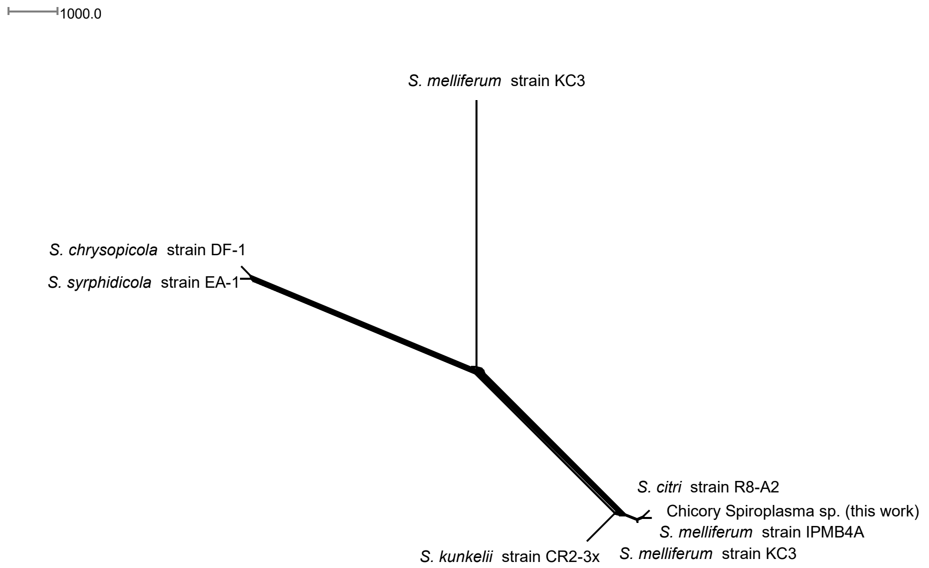

1000.0
S. melliferum strain KC3
S. chrysopicola strain DF-1
S. syrphidicola strain EA-1
S. citri strain R8-A2
Chicory Spiroplasma sp. (this work)
S. melliferum strain IPMB4A
S. kunkelii strain CR2-3x
S. melliferum strain KC3

Figure 3.2.6 – Neighbor phylogenetic network of concatenated sequences from genomes belonging to the Citri-Chrysopicola-Mirum phylogentic clade.

345 contigs, that is expected to hide rearrangements occurring at contig ends, the most numerous as most often the assembly stops when a repetitive sequence occours.

Moreover, in striking contraddiction with the the results from OMA just presented, a relevant fraction of the genomes does not align: coincevaibly, an unusually large fraction of the genomes consists in virus associated sequences that are distinct in the two genomes. Indeed, the unusually, extremely abundant presence of sequences of viruses, particularly plectoviruses, in the genome of *Spiroplasma citri* has been reported and it is a well known obstacle that delayed the completion of the organism genome sequence determination for a long time (Carle *et al.*, 2010) until the availability of the SMRT sequencing technology (Davis *et al.*, 2017).

To understand the functional differences in the two genomes we compared the predicted proteome of the two strains. The *S. citri* R8-A2 genome encodes 1812 proteins, among which 1061 are annotated as hypothetical proteins. In the remaining 751 proteins, 118 are annotated as preudogenes. Conversely, the ChiSsp genome encodes 1994 proteins, among which 1027 are

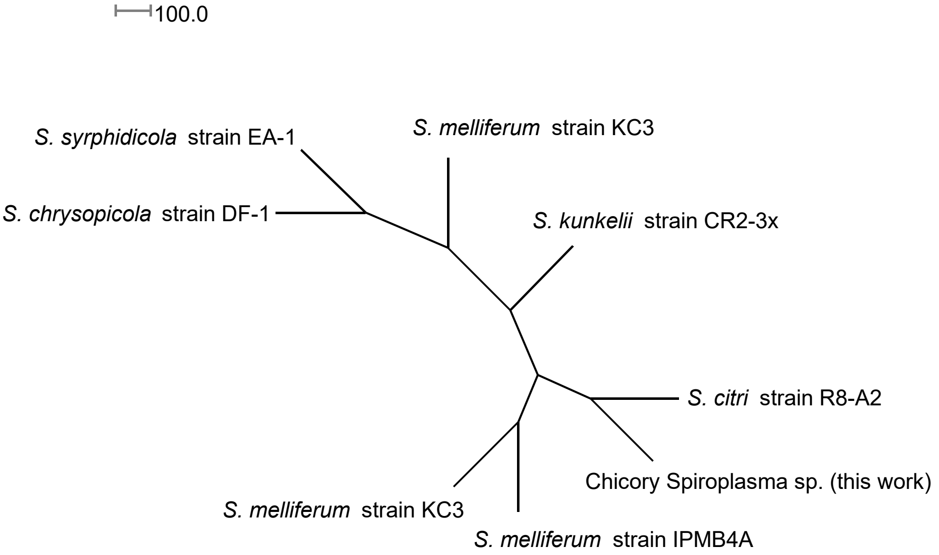

100.0
S. syrphidicola strain EA-1
S. melliferum strain KC3
S. chrysopicola strain DF-1
S. kunkelii strain CR2-3x
S. citri strain R8-A2
Chicory Spiroplasma sp. (this work)
S. melliferum strain KC3
S. melliferum strain IPMB4A

Figure 3.2.7 – Consensus network, showing the complete congruence of the phylogenetic signal and absence of any contradictory evidence for grouping.

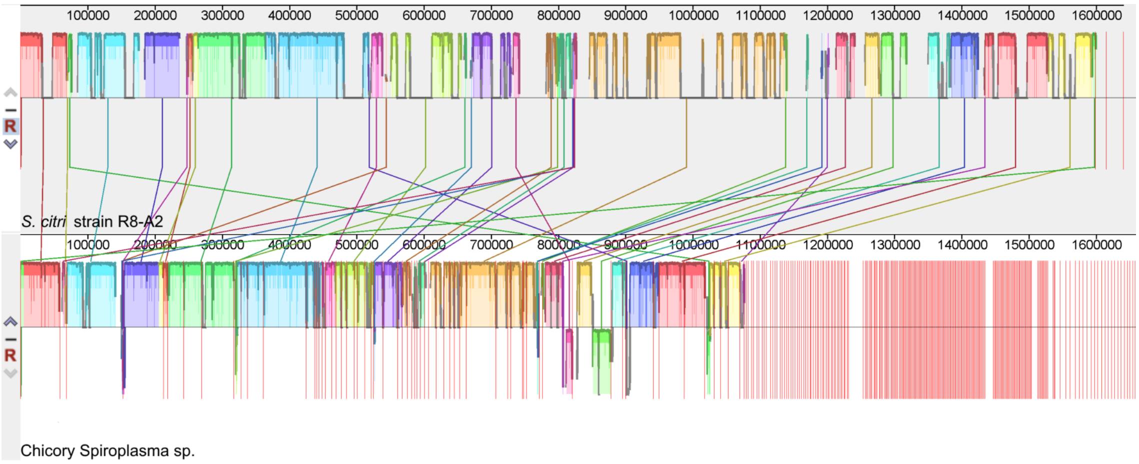


Figure 3.2.8 – Mauve alignment between *Spiroplasma citri* strain R8-A2 and Chicory Spiroplasma sp. (ChiS).

annotated as hypothetical proteins. In the remaining 967 proteins, we estimated that at least 159 are preudogenes.

A preliminary orthologous search using OMA of annotated proteins found 485 pairs of orthologous proteins, suggesting that the number of functions that are not shared by the two strains could be relevant. However, a closer look to the putatively strain-specific gene sequences revealed that they were mostly *composed* by complete and (most often) incomplete copies of genes already used by OMA. We therefore used a collection of perl scripts (that we named "Comparator"; Firrao & Marcelletti, unpublished) developed for this task in previous works (Saccardo *et al.*, 2012; Scortichini *et al.*, 2013; Torelli *et al.*, 2015) to identify gene families by homology. While OMA identifies pairs of orthologous genes, one in each genome in comparison, *Comparator* identifies groups of homologous genes that includes orthologs and paralogs, *i.e.* including one or more genes from each genome.

Comparator found 1031 homologous gene families containing at least one gene in each genome in comparison, comprehensive of 1258 genes in *S. citri* R8-A2 and 1445 in ChiSsp. The number of genes not included in common families was 454 in *S. citri* R8-A2 and 549 in ChiSsp.

Among the 454 *S. citri* R8-A2 specific putative proteins 9 were phage/plasmid associated proteins, 369 were unannotated as hypothetical proteins (most likely of viral origin), 21 were mobile element associated proteins. The remaining only 4 *S. citri* R8-A2 specific putative proteins included 2 methyltransferases and 1 HAD family hydrolase and one incomplete copy of the same gene.

Among the 549 ChiSsp specific putative proteins 114 were phage/plasmid associated proteins, 364 were unannotated hypothetical proteins, 31 were mobile elements associated proteins. The remaining 36 ChiSsp specific putative proteins include genes and gene fragments that are annotates as methylases/methyltransferases (21 among genes and fragments), as adhesins (3 genes and 1 fragment), or as genes implicated in sugar transport and metabolism (8 genes), and two other additional genes as detailed in Table 3.2.4. Adhesins and transporters may play a role in insect host specificity (Dénes *et al.*, 2003; Boutareaud *et al.*, 2004; Bové *et al.*, 2003).

In summary, the second Mollicute genome characterized from our chicory samples resulted to belong to a strain of Spiroplasma citri that have a functional content very similar to *S. citri* R8-A2, despite the divergent structure, the different viral content and the divergent distribution of gene fragments and repetitive sequences. Similarly to the other *S. citri* genomes investigated so

Table 3.2.4 – ChiSsp putative proteins found by *Comparator* (Firrao & Marcelletti, unpublished), classified by their functions.

Context	Name	Number of genes/fragments	Similar to:
Adhesion	orf00596CHScontig16	1	Putative adhesin P89
	orf00633CHScontig169	1	Putative adhesin P89
	orf02110CHScontig82	1	Putative adhesin P89
	orf00663CHScontig171	1	Streptococcal hemagglutinin protein
Methilases	orf00400CHScontig127	1	Site-specific DNA methylase
	orf00630CHScontig168	9	Adenine-specific methyltransferase
	orf00502CHScontig141	8	DNA-cytosine methyltransferase
	orf00156CHScontig1	3	tRNA:m(5)U-54 MTase gid
Sugar metabolism and transport	orf00708CHScontig18	1	PTS system, diacetylchitobiose-specific IIC component
	orf02096CHScontig8	1	PTS system, fructose-specific IIA/IIB/IIC component
	orf02094CHScontig8	1	PTS system, galactose-inducible IIA component
	orf00921CHScontig22	1	PTS system, IIA component
	orf01804CHScontig58	1	PTS system, N-acetylglucosamine-specific IIB/IIC component
	orf00145CHScontig1	1	ABC transporter ATP-binding protein
	orf00697CHScontig18	1	Outer surface protein of unknown function - cellobiose operon
	orf02170CHScontig9	1	D-lactate dehydrogenase (EC 1.1.1.28)
Other genes	orf02100CHScontig8	1	Bona fide RidA/YjgF/TdcF/RutC subgroup
	orf00162CHScontig1	1	Nucleoside-diphosphate-sugar epimerases

far, a large fraction of the genome of the chicory strain is the result of plectovirus invasion. Comparative analysis of *S. melliferum* IPMB4A (Lo *et al.*, 2013) showed that these phages have facilitated extensive genome rearrangements in these bacteria, and our result provide confirming evidence for this notion as far as *S. citri* is concerned. The Authors also suggest that this feature contribute to horizontal gene transfers that led to species-specific adaptation to different eukaryotic hosts, a contribution that we did not evidence through the analysis of our samples.

Presence of the phytoplasma and the spiroplasma in the environment

The evidence of double infection in the sample used for the ILLUMINA sequencing prompted us to the investigation of field samples in order to ascertain whether or not mixed infections were a common occurrence and may have epidemiological significance.

As reported in Table 3.2.5, we found out that as much as two third of spiroplasma positive samples and nearly the same for phytoplasma positive samples were in mixed infections, a strong indication of vector preferential behaviour.

Table 3.2.5 – Result of the field sample analysis by direct (d:) and nested (n:) PCR.

Total samples	Phytoplasma	Spiroplasma	Mixed infections
45	31 (direct:24 + nested:7)	27 (direct:11 + nested:16)	18

It has been reported in the literature that phytoplasma may manipulate the host gene expression making plants more actractive for the insects, by altering volatile profiles (Bertaccini *et al.*, 2011; Tan *et al.*, 2016; Janik *et al.*, 2017), hormonal patterns and other physiological traits (Cettul and Firrao, 2011; MacLean *et al.*, 2011; Sugio *et al.*, 2011; Sugio and Hogenhout, 2012). A similar effector-based interaction, that results in a more favorable environment for the insect vector on one hand, and in the typical symptoms such as phyllody and witches' broom on the other, has not been reported for the spiroplasmas. On the basis of mutant analysis, spiroplasma associated symptoms, such as yellowing, have been related with selective sugar uptake of the pathogen cells in the plant host phloem (Gaurivaud *et al.*, 2000). It is therefore conceivable that phytoplasma infection results in plants that are more attractive for several insect species, including those vectoring spiroplasmas. Epidemiological studies are presently ongoing in our laboratory to further elucidate this complex interactive network.

3.2.5 References

Aziz, R. K. *et al.* (2008) 'The RAST Server: Rapid Annotations using Subsystems Technology', *BMC Genomics*, 9(1), p. 75. doi: 10.1186/1471-2164-9-75.

Bandelt, H. J. *et al.* (1995) 'Mitochondrial portraits of human populations using median networks.', *Genetics*. Genetics Society of America, 141(2), pp. 743–53.

Bertaccini, A. *et al.* (2011) 'Effects of "Candidatus Phytoplasma asteris" on the Volatile Chemical Content and Composition of Grindelia robusta Nutt.', *Journal of Phytopathology*, 159(2), pp. 124–126.

Boutareaud, A. *et al.* (2004) 'Disruption of a gene predicted to encode a solute binding protein of an ABC transporter reduces transmission of Spiroplasma citri by the leafhopper Circulifer haematoceps.', *Applied and environmental microbiology*. American Society for Microbiology (ASM), 70(7), pp. 3960–7. doi: 10.1128/AEM.70.7.3960-3967.2004.

Bové, J. M. *et al.* (2003) ' *S PIROPLASMA CITRI* , A P LANT P ATHOGENIC M OLLICUTE : Relationships with Its Two Hosts, the Plant and the Leafhopper Vector', *Annual Review of Phytopathology*, 41(1), pp. 483–500. doi: 10.1146/annurev.phyto.41.052102.104034.

Carle, P. *et al.* (2010) 'Partial chromosome sequence of Spiroplasma citri reveals extensive viral invasion and important gene decay.', *Applied and environmental microbiology*. American Society for Microbiology, 76(11), pp. 3420–6. doi: 10.1128/AEM.02954-09.

Cettul, E. and Firrao, G. (2011) 'Development of phytoplasma-induced flower symptoms in Arabidopsis Thaliana', *Physiological and Molecular Plant Pathology*, 76(3–4), pp. 204–211. doi: 10.1016/j.pmpp.2011.09.001.

Darling, A. C. E. C. E. *et al.* (2004) 'Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements', *Genome Research*, 14(7), pp. 1394–1403. doi: 10.1101/gr.2289704.

Davis, R. E. *et al.* (2017) ‘Complete Genome Sequence of Spiroplasma citri Strain R8-A2T, Causal Agent of Stubborn Disease in Citrus Species.’, *Genome announcements*. American Society for Microbiology (ASM), 5(16). doi: 10.1128/genomeA.00206-17.

Dénes, B. *et al.* (2003) ‘Recognition of multiple Mycoplasma bovis antigens by monoclonal antibodies’, *Hybridoma and Hybridomics*, 22(1), pp. 11–16.

Gaurivaud, P. *et al.* (2000) ‘Fructose Utilization and Phytopathogenicity of Spiroplasma citri’, *Molecular Plant-Microbe Interactions*. The American Phytopathological Society, 13(10), pp. 1145–1155. doi: 10.1094/MPMI.2000.13.10.1145.

Gouy, M., Guindon, S. and Gascuel, O. (2010) ‘SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building.’, *Molecular biology and evolution*, 27(2), pp. 221–4. doi: 10.1093/molbev/msp259.

Guindon, S. and Gascuel, O. (2003) ‘A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.’, *Systematic biology*, 52(5), pp. 696–704.

Holland, B. R. *et al.* (2004) ‘Using consensus networks to visualize contradictory evidence for species phylogeny.’, *Molecular biology and evolution*, 21(7), pp. 1459–61. doi: 10.1093/molbev/msh145.

Huson, D. H. and Bryant, D. (2006) ‘Application of phylogenetic networks in evolutionary studies.’, *Molecular biology and evolution*, 23(2), pp. 254–67. doi: 10.1093/molbev/msj030.

Janik, K. *et al.* (2017) ‘An effector of apple proliferation phytoplasma targets TCP transcription factors-a generalized virulence strategy of phytoplasma?’, *Molecular Plant Pathology*, 18(3), pp. 435–442. doi: 10.1111/mpp.12409.

Lo, W.-S. *et al.* (2013) ‘Comparative genome analysis of Spiroplasma melliferum IPMB4A, a honeybee-associated bacterium’, *BMC Genomics*, 14(1), p. 22. doi: 10.1186/1471-2164-14-22.

MacLean, A. M. *et al.* (2011) ‘Phytoplasma effector SAP54 induces indeterminate leaf-like flower development in arabidopsis plants’, *Plant Physiology*, 157(2), pp. 831–841.

Marcelletti, S. *et al.* (2011) ‘Pseudomonas syringae pv. actinidiae Draft Genomes Comparison Reveal Strain-Specific Features Involved in Adaptation and Virulence to Actinidia Species’, *PLoS ONE*. Edited by M. A. Webber. 10 [doi, 6(11), p. e27297. doi: 10.1371/journal.pone.0027297.

Martini, M. *et al.* (2012) ‘Molecular characterization of phytoplasma strains associated with epidemics of chicory phyllody’, *Journal of Plant Pathology*, 94(4, Supplement), p. S4.49.

Saccardo, F. *et al.* (2012) ‘Genome drafts of four phytoplasma strains of the ribosomal group 16SrIII’, *Microbiology*. Microbiology Society, 158(Pt_11), pp. 2805–2814. doi: 10.1099/mic.0.061432-0.

Saglio, P., Lhospital, M. and Lafleche, D. (1973) ‘Spiroplasma citri gen. and sp. n.: a mycoplasma like organism associated with “Stubborn” disease of citrus’, *International Journal of Systematic Bacteriology*, 23(3), pp. 191–204.

Scortichini, M. *et al.* (2013) ‘A Genomic Redefinition of Pseudomonas avellanae species’, *PLoS ONE*. Edited by D. Arnold, 8(9), p. e75794. doi: 10.1371/journal.pone.0075794.

Simão, F. A. *et al.* (2015) ‘BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs’, *Bioinformatics*, 31(19), pp. 3210–3212. doi: 10.1093/bioinformatics/btv351.

Sugio, A. *et al.* (2011) ‘Phytoplasma protein effector SAP11 enhances insect vector reproduction by manipulating plant development and defense hormone biosynthesis’, *Proceedings of the National Academy of Sciences*, 108(48), pp. E1254-63. doi: 10.1073/pnas.1105664108.

Sugio, A. and Hogenhout, S. A. (2012) ‘The genome biology of phytoplasma: modulators of plants and insects.’, *Current Opinion in Microbiology*, 15(3), pp. 247–54. doi: 10.1016/j.mib.2012.04.002.

Tan, C. M. *et al.* (2016) ‘Phytoplasma SAP11 alters 3-isobutyl-2-methoxypyrazine biosynthesis in Nicotiana benthamiana by suppressing NbOMT1’, *Journal of Experimental Botany*, 67(14), pp. 4415–4425. doi: 10.1093/jxb/erw225.

Torelli, E. *et al.* (2015) ‘Draft genome of a Xanthomonas perforans strain associated with pith necrosis’, *FEMS Microbiology Letters*, 362(4). doi: 10.1093/femsle/fnv001.

Tritt, A. *et al.* (2012) ‘An Integrated Pipeline for de Novo Assembly of Microbial Genomes’, *PLoS ONE*. Edited by D. Zhu, 7(9), p. e42304. doi: 10.1371/journal.pone.0042304.

Whitcomb, R. F. *et al.* (1986) ‘Spiroplasma kunkelii sp. nov.: Characterization of the Etiological Agent of Corn Stunt Disease’, *International Journal of Systematic Bacteriology*. Microbiology Society, 36(2), pp. 170–178. doi: 10.1099/00207713-36-2-170.

3.3 Molecular characterization of organisms associated with cassava plants showing cassava frogskin disease

Abstract; manuscript draft to be prepared by the first Author.

Authors: Neves de Souza A.[1], Polano C.[2], Martini M.[2], Firrao G.[2], Carvalho C.[1]

[1] Department of Plant Pathology, Universitad Federal de Viçosa, Brazil

[2] Dipartimento di Scienze AgroAlimentari, Ambientali e Animali, Università di Udine

Cassava Frogskin Disease (CFSD) is a disease of great concern to the cultivation of cassava, mainly affecting their primary product, the tubers. Efforts have been made with the aim of better understanding the infectious process and the appearance of CFSD symptoms, but due to its etiology still controversial, these studies are challenging.

The identification and complete characterization of organisms associated with cassava plants showing CFSD are important to allow a better understanding of the disease, a more detailed studies on its etiology and on host-pathogen interaction. Therefore, the main aim of this study was the utilization of next-generation sequencing to identify and characterize potential organisms involved on the development of this disease. A deep sequencing of DNA and RNA from cassava plants showing symptoms of CFSD was performed.

In the DNA sequencing, the emphasis was on sequencing of a phytoplasma previously associated with this disease. The phytoplasma belonging to the ribosomal group 16SrIII had its genome sequenced, and we obtained its draft genome. This phytoplasma was compared with other phytoplasma from the same ribosomal group and it seemed to be slightly different from the other representatives of the group.

In the RNA deep sequencing, a new RNA subviral agent of 1228 bp in length was identified, and it shows two putative ORFs in its genome. One of the ORFs shows 156 aa in length and a common conserved domain from *Potexvirus* coat protein, and the second ORF, a putative 90 aa protein of unknown function. This was the first report of an RNA subviral agent associated with cassava plants. The presence of this subviral agent does not appear to be related to the occurrence of CFSD in cassava plants.

4 Metagenomic characterisation of communities

Until the 2000s, microbiology, microbial genome sequencing and genomics were conducted using cultivated clonal cultures and specific genes to simulate natural diversity, with the limitation of losing a large part of microbial biodiversity (Hugenholtz, Goebel and Pace, 1998). Used first in 1998, the term *metagenomics* indicates a set of research techniques, mainly including the use of shotgun sequencing, and a research field, the study of genetic material obtained directly from environmental samples (Handelsman *et al.*, 1998; Board on Life Sciences, 2007). By analysing microorganisms as an aggregate and focusing on how genes might influence each other's activities in serving collective functions, metagenomics tries to circumvents the unculturability and genomic diversity of most microbes, and employs computational methods designed to interpret the genetic composition and activities of communities that cannot be characterized at the individual level. With the advent of next-generation sequencing technologies, it became possible to obtain datasets where the sequences belong to a wide range and number of individuals, as opposed to one or few. The data used in metagenomic derives from high-throughput sequencing, but unlike genomic sequencing the coverage is generally low, as each read may come from a different individual. These reads can be attributed to the respective taxons using metabarcoding, which uses genetic markers (typically, 16S rRNA for bacteria and ITS for fungi) for identification.

Most of the tools and methods used to characterise isolates can be used in metabarcoding, but additional tools have become necessary to further characterise and perform statistical analyses on the datasets (Thomas, Gilbert and Meyer, 2012). First, low-quality sequences are filtered; duplicates are noted and represented only once to lower the computational requirements. Then the sequences are *binned*, grouped according to similarity, either by aligning them to a database of known sequences, attributing to each a taxonomic position, or by determining operational taxonomic unit (OTUs) that not necessarily correspond to known taxons (Blaxter *et al.*, 2005). OTUs can be used with distance-matrix methods to determine phylogenetic trees. While an in-depth exposition of the methods for calculating phylogenetic trees is beyond the scope of this thesis, it is worth mentioning a few methods, like UPGMA and neighour-joining, along with maximum parsimony, maximum likelihood, and Bayesian methods (Lemey, Salemi and Vandamme, 2009).

Additional tools from statistics can be employed to further explore metagenomic datasets, most notably multidimensional scaling, in which the distances between the OTUs are represented in an *n*-dimensional space (usually 2 dimensional *scatterplots*, Figure 4.1) with arbitrary *x* and *y* axes (Borg, and Groenen, 2005). The most commonly used are Principal Coordinate Analysis (PCoA) and non-metric multidimensional scaling (nMDS) (Ramette, 2007).

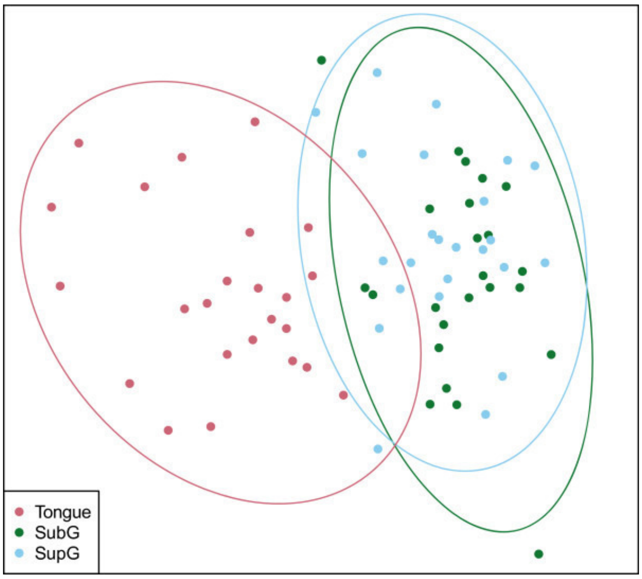


Figure 4.1 – An example of a multidimensional scaling (MDS) scatterplot. [Source: (Galimanas *et al.*, 2014)]

Statistical indices such as the Simpson and the Shannon estimators, and tools such as the analysis of molecular variance (AMOVA) (Excoffier, Smouse and Quattro, 1992), can of course be applied to assess the significativity of the elaborated data results. Commonly used tools to perform all of these analyses are *MEGAN* (Huson *et al.*, 2007), *Mothur* (Schloss *et al.*, 2009) and *QIIME* (Caporaso *et al.*, 2010).

In the following paper, a metagenomic approach is used in attempt to understand, and possibly take advantage of, the relations between the kiwifruit endophytes and *Pseudomonas syringae* pv. *actinidiae*, causal agent of the kiwifruit canker.

Bibliography

Blaxter, M. *et al.* (2005) 'Defining operational taxonomic units using DNA barcode data', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1462), pp. 1935–1943. doi: 10.1098/rstb.2005.1725.

Board on Life Sciences (2007) *The New Science of Metagenomics*. Washington, D.C.: National Academies Press. doi: 10.17226/11902.

Borg, I. and Groenen, P. (2005) *Modern Multidimensional Scaling: theory and applications*. 2nd edn. New York: Springer-Verlag.

Caporaso, J. G. *et al.* (2010) 'QIIME allows analysis of high-throughput community sequencing data', *Nature Methods*, 7(5), pp. 335–336. doi: 10.1038/nmeth.f.303.

Excoffier, L., Smouse, P. E. and Quattro, J. M. (1992) 'Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes: Application', 491, pp. 479–491. doi: 10.1038/nmeth.3176.

Galimanas, V. *et al.* (2014) 'Bacterial community composition of chronic periodontitis and novel oral sampling sites for detecting disease indicators', *Microbiome*, 2(1), p. 32. doi: 10.1186/2049-2618-2-32.

Handelsman, J. *et al.* (1998) 'Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products', *Chemistry & Biology*, 5(10), pp. R245–R249. doi: 10.1016/S1074-5521(98)90108-9.

Hugenholtz, P., Goebel, B. M. and Pace, N. R. (1998) 'Impact of culture independent studies on the emerging phylogenetic view of bacterial diversity', *Journal of Bacteriology*, v(18), p. 180p4765-4774. doi: 0021-9193/98/$04.00+0.

Huson, D. H. *et al.* (2007) 'MEGAN analysis of metagenomic data', *Genome Research*, 17(3), pp. 377–386. doi: 10.1101/gr.5969107.

Lemey, P., Salemi, M. and Vandamme, A.-M. (eds) (2009) *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. 2nd edn. Cambridge University Press.

Ramette, A. (2007) ‘Multivariate analyses in microbial ecology’, *FEMS Microbiology Ecology*, 62(2), pp. 142–160. doi: 10.1111/j.1574-6941.2007.00375.x.

Schloss, P. D. *et al.* (2009) ‘Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities’, *Applied and Environmental Microbiology*, 75(23), pp. 7537–7541. doi: 10.1128/AEM.01541-09.

Thomas, T., Gilbert, J. and Meyer, F. (2012) ‘Metagenomics - a guide from sampling to data analysis’, *Microbial Informatics and Experimentation*, 2(1), p. 3. doi: 10.1186/2042-5783-2-3.

4.1 Multivariate analysis of endophytes diversity in kiwifruit in relation with *Pseudomonas syringae* pv. *actinidiae*

Authors: Cesare Polano, Marta Martini, Paolo Ermacora, Francesca Ferrini, Nazia Loi, Giuseppe Firrao. ***Manuscript in preparation.***

4.1.1 Introduction

Pseudomonas syringae is a Gram-negative bacterium whose pathovars can infect a variety of plant species; some pathovars target specific species, while others can infect a spectrum of hosts; In particular, *P. syringae pv. actinidiae* (Psa) targets *Actinidia deliciosa* and *A. chinensis* plants as the causal agent of bacterial canker. This disease was identified as early as 1989 in Japan (Takikawa *et al.*, 1989) and 1994 in Italy (Scortichini, 1994). Yellow and red fleshed kiwifruits are generally more susceptible than green ones; Psa3 was in fact first detected on yellow cultivars (Koh *et al.*, 2012). Symptoms include leaf spots and necrosis, extensive twig dieback, reddening of the lenticels, bleeding cankers on the trunk and leader with whitish to orange ooze, and in the worse cases can lead to the death of the host (Agrios, 2005).

However, the impact of the disease worldwide has not been as severe as in Japan and Korea until an outbreak occurred a decade later, which has caused grave damage to kiwifruit culture particularly in southern Europe, New Zealand and Korea (Vanneste, 2012; Kim *et al.*, 2016). Psa from these outbreaks has been divided into 3 biovars: Psa1 and Psa2 caused the cankers in Japan and Korea in 1980s and produce phaseolotoxin (Psa1) and coronatine (Psa2), while Psa3 caused the cankers in Italy and worldwide in 2008 (Scortichini *et al.*, 2012). It was determined that the new strain emerged as a result of horizontal transfers events from asian strains (Marcelletti *et al.*, 2011), including the incorporation of Integrative Conjugative Elements (ICEs) (Marcelletti *et al.*, 2011; Butler *et al.*, 2013).

Strategies for controlling the spread of Psa are limited and consist in eliminating the affected plants and protecting the healthy ones in spring and fall using copper formulations (Vanneste *et* *al.*, 2011) and avoiding excess humidity, *e.g.* by covering the plant with nets. Excessive recourse of copper treatments however led to the differentiation of resistant subclones, by acquisition of another ICE (Colombi *et al.*, 2017). It was recently observed that some plants do not seem to get infected, even after a few years of exposure. The hypothesis tested in this work is whether this resilience could be influenced by the interaction between the pathogen and the endophyte population (Reinhold-Hurek and Hurek, 2011); the testing was conducted using a metagenomic analysis of 16S and ITS sequences sampled in 3 consecutive years.

4.1.2 Materials and methods

Two sets of data were sampled: a preliminary set ("alpha") of 16S sequences included data from 12 samplings of kiwifruit bark and leaves obtained between 2014 and 2015 (Table 4.1.1), from an experimental field in Dandolo di Sopra, Friuli-Venezia Giulia, Italy, using the 926F and 1392R primers. Samples 16S-1/2, 16S-3/4/5/6/7 and 16S-8/9/10/11/12 were sampled each from a single plant.

A second, more complete set ("beta") of 16S (V6–V8) and ITS2 sequences included data from 24 samplings of kiwifruit bark obtained in July 2016, from the same field, using the 926F, 1392R (Engelbrektson *et al.*, 2010), ITS3_KYO2, ITS4 (Toju *et al.*, 2012) primers. The ITS2 primers are not specific, in order to amplify as many fungi as possible, though they can fail to pick up a few Orders (Asemaninejad *et al.*, 2016). Symptoms from the plants used in the beta set were recorded during three years, from 2015 to 2017.

Genomic DNA was extracted from 24 bark samples with various levels of PSA symptoms, using modified version of the Doyle and Doyle method (Doyle and Doyle, 1990). Amplicons were amplified using a PCR method described in (Martini *et al.*, 2009), with a KAPA HiFi HotStart PCR kit; primers used were 799F/1492R for bacteria (Goodfellow and Stackebrandt, 1991) and ITS1F-KYO1/ITS4 for fungi (Toju *et al.*, 2012). To avoid possible plant contaminations, the amplicons were run in an electrophoretic apparatus, and the excised bands purified with an RBC Real Biotech kit. The purified amplicons were then amplified with the aforementioned 926F/1392R primers for bacteria and ITS3_KYO2/ITS4 for fungi, and purified with an RBC

Real Biotech kit. The samples were then sent to were sent to the Institute of Applied Genomics (IGA, Udine, Italy) to be sequenced with Illumina MiSeq 2×300.

Because of the limitations of phylotype-based methods, such as ambiguously-defined taxons (particularly below the order level) and, consequently, the limited completeness of available taxonomical databases (Schloss and Westcott, 2011), an Operational Taxonomic Unit (OTU)-based strategy for sequence clustering was chosen, in order to verify whether the microbial community in the samples can be correlated to the severity of the symptoms and time of infection. The correlation between the samples was assessed using a multivariate analysis, which has increasingly become an essential tool in exploring and understanding large data sets (Ramette, 2007). Non-metric multidimensional scaling analysis (nMDS) was chosen over principal component analysis (PCoA) because it does not require data preprocessing and is generally regarded as a more robust method (Taguchi and Oono, 2005), while the distances were calculated using the Canberra method, as it is known to perform especially well for detecting clusters (Kuczynski *et al.*, 2010).

The OTUs were determined and clustered using the *LotuS* processing pipeline (Hildebrand *et al.*, 2014); the nMDS spatialisation was carried out using the software suite *Mothur* (Schloss *et al.*, 2009) and graphed using the R statistical language (R Core Team, 2017).

4.1.3 Results

In the alpha set each sample had on average 48,729 reads, ranging from 14,994 to 87,988, for a total of 584,752 reads (Table 4.1.1); in the beta set, each 16S sample had on average 519,800 reads, ranging from 199,492 to 1,085,258, for a total of 24,950,422 reads (Figure 4.1.4), while each ITS sample had on average 291,075 reads, ranging from 121,877 to 582,370, for a total of 13,971,610 reads (Figure 4.1.5).

The preliminary analysis on the alpha set was done to assess the variability resulting from the sampling method. Comparing replicas showed that there is some variability between replicas: in Figure 4.1.1, for example, *Enterobacteriales* are present in less than 5% of sample 16S-1 reads and over 25% of sample 16S-2, while *Burkholderiales* are present in over 15% of sample 16S-1 reads and slightly over 5% of sample 16S-2 reads.

There are also some differences between the populations in the bark and the leaves (Figure 4.1.2): *Burkholderiales* are twice as numerous in bark as they are in leaves, while in half of the leaves samples *Rhizobiales* and *Sphingomonadales* represent 25% of the reads, as opposed to 15% of the rest of the samples. As for the changes over time (Figure 4.1.3), *Sphingobacteriales* and *Actimomycetales* are almost absent in June but have a significant presence in September, then decline again, while *Bacillales* are absent in September and October but moderately present in June and March.

Table 4.1.1 – Preliminary samples ("alpha set") used to calibrate the analysis. The sampling was done to include samples from autumn and spring and from bark and leaves.

Sample ID	Source	Date of sampling	Reads		Sample ID	Source	Date of sampling	Reads
16S-3	bark	June 2014	29.648		16S-8	bark	June 2014	76.100
16S-4	bark	September 2014	34.893		16S-9	bark	September 2014	44.493
16S-5	bark	October 2014	77.521		16S-10	bark	October 2014	69.072
16S-6	leaf	October 2014	87.899		16S-11	leaf	October 2014	19.985
16S-7	bark	March 2015	34.777		16S-12	bark	March 2015	14.994
16S-1	bark	April 2015	26.041		16S-2	bark	April 2015	69.329

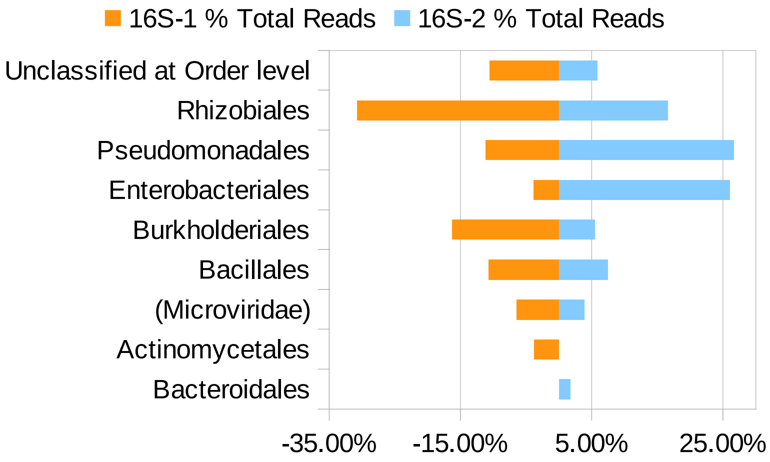


Figure 4.1.1 – Orders distribution for samples 16S-1 and 16S-2 of the alpha set, repetitions made at the same time (April 2015) from the same location (bark) of the same plant.

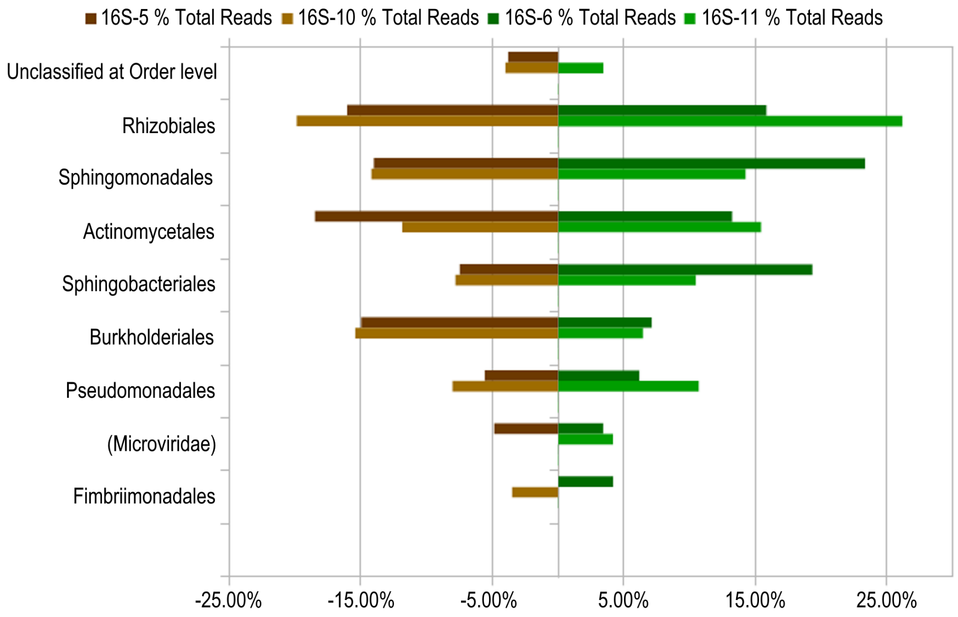


Figure 4.1.2 – Orders distribution differences between bark (samples #5 and #10 of the alpha set) and leaves (samples #6 and #11); pairs come from different plants.

For the complete analysis on the beta set, the samples were divided in 4 groups, based on the time of the first occurrence of the symptoms (Table 4.1.2): group A has no observed symptoms, group B had symptoms since 2017, group C since 2016 and group D since 2015. For the 16S clusters, the spatialisation (Figure 4.1.6) shows a fairly good segregation, as the samples from group C or D are completely on the left side, while the samples from group A or B are on the right side, although not separated. Similarly, most of the ITS samples from group C or D are segregated to the lower-left side (Figure 4.1.7) with a few outliers on the upper-right, while samples from group A and B are on the right side, in this case without intermixing.

From Table 4.1.3 it can be noted that 16S24 and 16S03 differ by the proportion of *Cytophagaceae, Oxalobacteriaceae, Microbacteriaceae*, and of most notably, *Pseudomonadaceae*. Similarly, ITS22 and ITS03 differ by the proportion of *Phaeosphaeriaceae, Leptosphaeriaceae, Taphrinaceae, Montagnulaceae* and *Mycosphaerellaceae*.

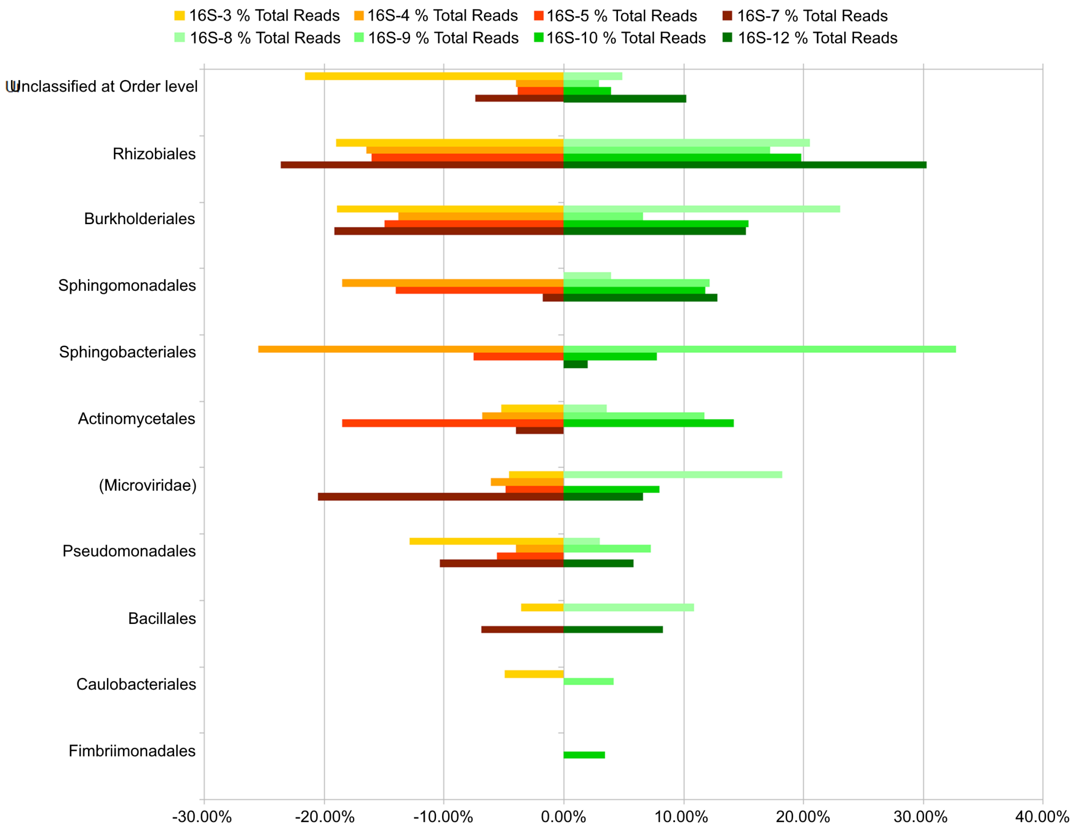


Figure 4.1.3 – Orders distribution differences between months: June (samples #3 and #8), September (samples #4 and #9), October (samples #5 and #10), March (samples #7 and #12); pairs come from different plants.

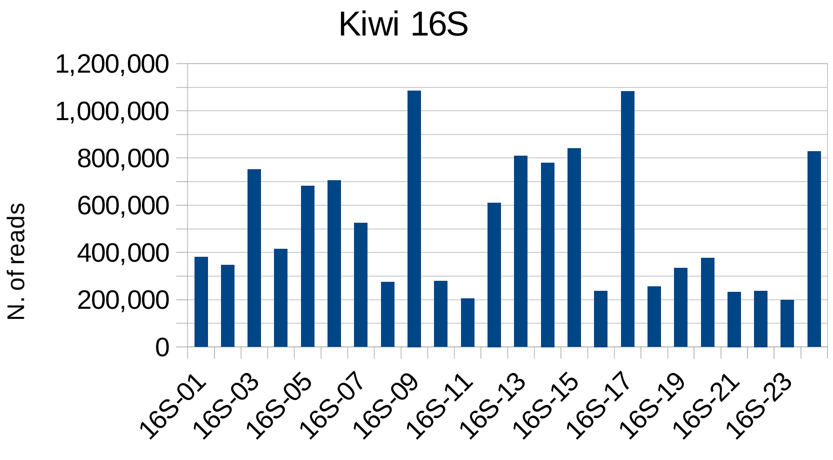


Figure 4.1.4 – Read size and distribution of the kiwifruit samples in the 16S region.

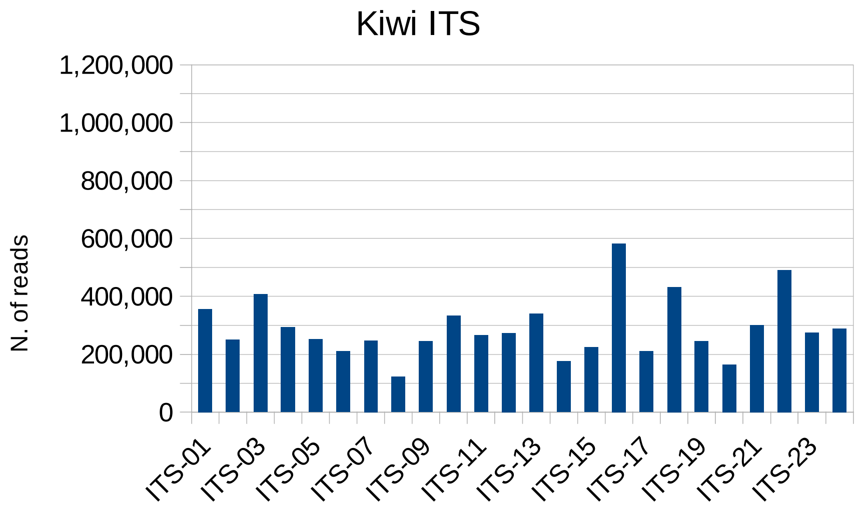


Figure 4.1.5 – Read size and distribution of the kiwifruit samples in the ITS region.

4.1.4 Discussion

In the last decade, *Pseudomonas syrigae* pt. *syringae* (Psa) has become a gravely damaging causal agent of disease in kiwifruits. Attempts to contain it with copper formulations produced the undesired effect of selecting resistant strains. Psa has proven to have a rather complex set of elicitors and T3SS effectors, and a dynamic system of transposable elements.

With the introduction of metagenomic analyses made possible by Whole Genome Sequencing, a more comprehensive scope of investigating the relation between pathogens, plants and the rest of the microbial community has been made possible. The complexity of the pathogenic mechanisms of Psa, along with the experimental observation that some plants appear less affected than others while not being more resistant themselves, suggested to investigate the microbial population as a whole, in order to verify whether this form of 'resistance' is influenced by differences in the diversity of the endophyte population of kiwifruit samples. At the time of this writing, it was not possible to include data from unculturable microorganisms, therefore their role in this analysis could not be assessed.

With this ongoing 4-year project, we have shown that such a metagenomic analysis has found a correlation with the severity of the symptoms and the time of infection of the plants. While further analyses will be required to elucidate whether the differences in Order distributions in 16S and ITS samples are linked, and whether these differences could be used as predictors for Psa spread, the graphical spatialisation of the OTUs did reflect fairly well the preexistent data relative to the symptomatological classes derived from visual and real-time PCR observations.

Table 4.1.2 – Samples taken in three consecutive years ("beta set"), with real-time PCR results on the same DNA extracted from kiwi vines, used for categorising the samples of the microbial community. Symptomatological classes are: 0 – no exudates; 1 – few exudates without cankers or dryings; 2 – exudates, cankers and dryings of young parts of the plant; 3 – abundance of exudates, wide presence of cankers and dryings of older parts of the plant. Real-time PCR was done on PSA using DNA extracted from vines in July 2016.

Label	Group	Symptomatological classes			Real-time PCR
		2015-04-24	2016-03-21	2017-04-20	
#1	B	0	0	1	–
#2	A	0	0	0	–
#3	A	0	0	0	–
#4	B	0	0	2	–
#5	B	0	0	1	–
#6	A	0	0	0	–
#7	B	0	0	2	–
#8	B	0	0	1	–
#9	C	0	1	0	+ (ct 19.40)
#10	C	0	1	3	–
#11	C	0	1	2	–
#12	C	0	1	0	–
#13	C	0	1	1	–
#14	C	0	1	1	–
#15	C	0	1 (pollarded)	0	–
#16	C	0	1	2	–
#17	D	2 (pollarded)	0	3	+ (ct 23.73)
#18	D	2	1	1	–
#19	D	2	1	2	+ (ct 23.33)
#20	D	2	1 (pollarded)	1	+ (ct 23.89)
#21	D	2	2	1	+ (ct 23.14)
#22	D	2	1	2	+ (ct 23.57)
#23	D	2	1	2	+ (ct 23.13)
#24	D	1	1 (pollarded)	1	+ (ct 22.74)

Table 4.1.3 – Distribution of the most represented families in the most distant samples (group D 16S24 *vs.* group A 16S03 and group D ITS22 *vs.* group A ITS03) from the 16S and the ITS clusterings.

Families	Reads
16S24	
Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Comamonadaceae	55,570
Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Oxalobacteraceae	22,011
Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae	19,997
Bacteria;Actinobacteria;Actinobacteria;Actinomycetales;Propionibacteriaceae	15,261
Bacteria;Firmicutes;Bacilli;Bacillales;Staphylococcaceae	10,905
Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales_Incertae Sedis XI	8,166
Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Methylobacteriaceae	2,883
16S03	
Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Comamonadaceae	46,557
Bacteria;Bacteroidetes;Cytophagia;Cytophagales;Cytophagaceae	13,790
Bacteria;Actinobacteria;Actinobacteria;Actinomycetales;Microbacteriaceae	9,443
Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Oxalobacteraceae	8,954
Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Methylobacteriaceae	8,103
Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae	7,483
Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae	4,402
ITS22	
Fungi;Ascomycota;Dothideomycetes;Pleosporales;Pleosporaceae	54,069
Fungi;Ascomycota;Dothideomycetes;Pleosporales;Phaeosphaeriaceae	15,046
Fungi;Basidiomycota;Tremellomycetes;Tremellales;Incertae sedis	10,669
Fungi;Ascomycota;Dothideomycetes;Pleosporales;Leptosphaeriaceae	10,133
Fungi;Ascomycota;Dothideomycetes;Pleosporales;Incertae sedis	4,068
Fungi;Ascomycota;Taphrinomycetes;Taphrinales;Taphrinaceae	2,923
Fungi;Ascomycota;Dothideomycetes;Incertae sedis;?	2,215
ITS03	
Fungi;Ascomycota;Dothideomycetes;Pleosporales;Pleosporaceae	33,200
Fungi;Basidiomycota;Tremellomycetes;Tremellales;Incertae sedis	9,683
Fungi;Ascomycota;Dothideomycetes;Pleosporales;Montagnulaceae	8,844
Fungi;Ascomycota;Dothideomycetes;Capnodiales;Mycosphaerellaceae	6,946
Fungi;Ascomycota;Dothideomycetes;Pleosporales;Incertae sedis	5,155
Fungi;Basidiomycota;Tremellomycetes;Tremellales;unidentified	3,184
Fungi;Ascomycota;Dothideomycetes;Incertae sedis;Incertae sedis	2,455

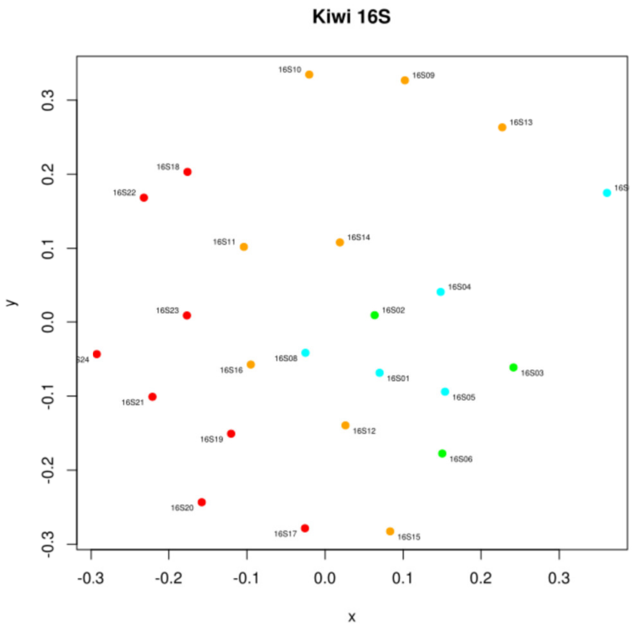


Figure 4.1.6 – Non-metric multidimensional scaling spatialisation of the kiwifruit samples, based on the 16S OTUs distances calculated using the Canberra method. Colours correspond to the symptomatic grouping: green – no observed symptoms, cyan – symptoms since 2017, orange – symptoms since 2016, red – symptoms since 2015.

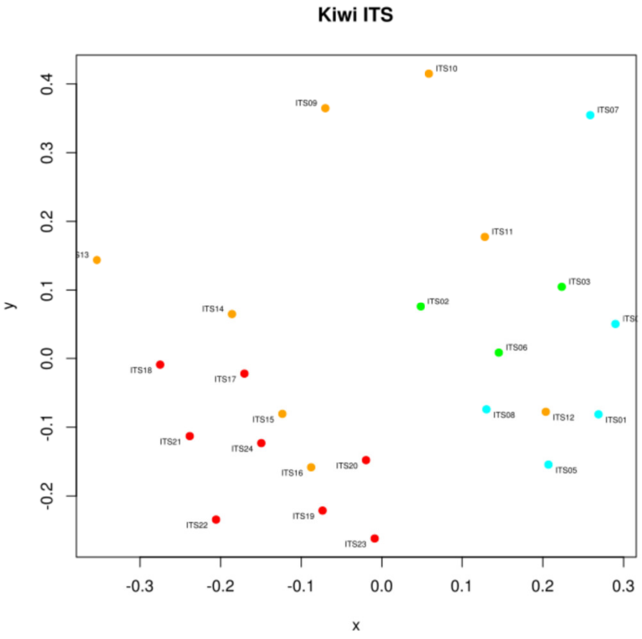


Figure 4.1.7 – Non-metric multidimensional scaling spatialisation of the kiwifruit samples, based on the ITS OTUs distances calculated using the Canberra method. Colours correspond to the symptomatic grouping: green – no observed symptoms, cyan – symptoms since 2017, orange – symptoms since 2016, red – symptoms since 2015.

4.1.5 References

Agrios, G. N. (2005) *Plant Pathology*. 5th edn. Amsterdam: Elsevier Academic Press.

Asemaninejad, A. *et al.* (2016) 'New Primers for Discovering Fungal Diversity Using Nuclear Large Ribosomal DNA', *PLOS ONE*. Edited by S. Pöggeler, 11(7), p. e0159043. doi: 10.1371/journal.pone.0159043.

Butler, M. I. *et al.* (2013) 'Pseudomonas syringae pv. actinidiae from Recent Outbreaks of Kiwifruit Bacterial Canker Belong to Different Clones That Originated in China', *PLoS ONE*, 8(2). doi: 10.1371/journal.pone.0057464.

Colombi, E. *et al.* (2017) 'Evolution of copper resistance in the kiwifruit pathogen Pseudomonas syringae pv. actinidiae through acquisition of integrative conjugative elements and plasmids', *Environmental Microbiology*, 19(2), pp. 819–832. doi: 10.1111/1462-2920.13662.

Doyle, J. J. and Doyle, J. L. (1990) 'Isolation of plant DNA from fresh tissue', *Focus*, 12, pp. 13–15.

Engelbrektson, A. *et al.* (2010) 'Experimental factors affecting PCR-based estimates of microbial species richness and evenness', *The ISME Journal*, 4(5), pp. 642–647. doi: 10.1038/ismej.2009.153.

Goodfellow, M. and Stackebrandt, E. (1991) *Nucleic acid techniques in bacterial systematics. Vol. 5*. Wiley.

Hildebrand, F. *et al.* (2014) 'LotuS: an efficient and user-friendly OTU processing pipeline', *Microbiome*, 2(1), p. 30. doi: 10.1186/2049-2618-2-30.

Kim, G. H. *et al.* (2016) 'Outbreak and Spread of Bacterial Canker of Kiwifruit Caused by Pseudomonas syringae pv. actinidiae Biovar 3 in Korea', *The Plant Pathology Journal*, 32(6), pp. 545–551. doi: 10.5423/PPJ.OA.05.2016.0122.

Koh, Y. J. *et al.* (2012) 'Occurrence of a New Type of Pseudomonas syringae pv. actinidiae Strain of Bacterial Canker on Kiwifruit in Korea', *The Plant Pathology Journal*, 28(4), pp. 423–427. doi: 10.5423/PPJ.NT.05.2012.0061.

Kuczynski, J. *et al.* (2010) 'Microbial community resemblance methods differ in their ability to detect biologically relevant patterns', *Nature Methods*, 7(10), pp. 813–819. doi: 10.1038/nmeth.1499.

Marcelletti, S. *et al.* (2011) 'Pseudomonas syringae pv. actinidiae Draft Genomes Comparison Reveal Strain-Specific Features Involved in Adaptation and Virulence to Actinidia Species', *PLoS ONE*. Edited by M. A. Webber. 10 [doi, 6(11), p. e27297. doi: 10.1371/journal.pone.0027297.

Martini, M. *et al.* (2009) 'DNA-Dependent Detection of the Grapevine Fungal Endophytes Aureobasidium pullulans and Epicoccum nigrum', *Plant Disease*. The American Phytopathological Society, 93(10), pp. 993–998. doi: 10.1094/PDIS-93-10-0993.

R Core Team (2017) 'R: a Language and Environment for Statistical Computing'. Vienna, Austria: R Foundation for Statistical Computing. Available at: http://www.r-project.org.

Ramette, A. (2007) 'Multivariate analyses in microbial ecology', *FEMS Microbiology Ecology*, 62(2), pp. 142–160. doi: 10.1111/j.1574-6941.2007.00375.x.

Reinhold-Hurek, B. and Hurek, T. (2011) 'Living inside plants: bacterial endophytes', *Current Opinion in Plant Biology*, 14(4), pp. 435–443. doi: 10.1016/j.pbi.2011.04.004.

Schloss, P. D. *et al.* (2009) 'Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities', *Applied and Environmental Microbiology*, 75(23), pp. 7537–7541. doi: 10.1128/AEM.01541-09.

Schloss, P. D. and Westcott, S. L. (2011) 'Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis', *Applied and Environmental Microbiology*, 77(10), pp. 3219–3226. doi: 10.1128/AEM.02810-10.

Scortichini, M. (1994) 'Occurrence of Pseudomonas syringae pv. actinidiae on kiwifruit in Italy', *Plant Pathology*, 43(6), pp. 1035–1038. doi: 10.1111/j.1365-3059.1994.tb01654.x.

Scortichini, M. *et al.* (2012) 'Pseudomonas syringae pv. actinidiae: a re-emerging, multi-faceted, pandemic pathogen', *Molecular Plant Pathology*, 13(7), pp. 631–640. doi: 10.1111/j.1364-3703.2012.00788.x.

Taguchi, Y. -h. and Oono, Y. (2005) 'Relational patterns of gene expression via non-metric multidimensional scaling analysis', *Bioinformatics*, 21(6), pp. 730–740. doi: 10.1093/bioinformatics/bti067.

Takikawa, Y. *et al.* (1989) 'Pseudomonas syringae pv. actinidiae pv. nov.: The causal bacterium of canker of kiwifruit in Japan.', *Japanese Journal of Phytopathology*, 55(4), pp. 437–444. doi: 10.3186/jjphytopath.55.437.

Toju, H. *et al.* (2012) 'High-Coverage ITS Primers for the DNA-Based Identification of Ascomycetes and Basidiomycetes in Environmental Samples', *PLoS ONE*. Edited by O. Lespinet. Public Library of Science, 7(7), p. e40863. doi: 10.1371/journal.pone.0040863.

Vanneste, J. (2012) ‘Pseudomonas syringae pv. actinidiae (Psa): a threat to the New Zealand and global kiwifruit industry’, *New Zealand Journal of Crop and Horticultural Science*, 40(4), pp. 265–267. doi: 10.1080/01140671.2012.736084.

Vanneste, J. L. *et al.* (2011) ‘Recent advances in the characterisation and control of Pseudomonas syringae pv. actinidiae, the causal agent of bacterial canker on kiwifruit’, *Acta Horticulturae*, (913), pp. 443–455. doi: 10.17660/ActaHortic.2011.913.59.

5 General conclusions and perspectives

Whole genome sequencing in the last decade has seen a sharp increase in the amount and quality of data that can be made available, at a fraction of the cost and the labour it used to have. This promped the need for more complex strategies and tools aimed at 'making sense' of such data, but also opened the doors to an unprecedented scope of enquiries. On a phytopathological perspective, this technological improvement has the potential for a deeper understanding of the relation between pathogens and their hosts and between pathogens and the rest of the microbial community (Stubbendieck and Straight, 2016), and ultimately the potential for better (less-impacting) preventive defence strategies against current, but also future plant diseases.

WGS has great potential in many aspects of phytopathology; in characterising bacterial strains, it allows a comparison between known (and occasionally unknown) availabilities of secondary metabolites (Gross and Loper, 2009): in the case of the *Pseudomonas* sp. strain Pf-4, comparing sequences of clusters pertaining to previously-identified secondary metabolite production, it was possible to prove that the genome of Pf-4 includes an 'arsenal' of metabolites quite similar to that of already well-characterised biocontrol agents, *P. protegens* strain Pf-5 in particular (Takeuchi *et al.*, 2014). This is significant, because Pf-5 is a biocontrol agent employed in soil cultures, while Pf-4 was isolated from a hydroponic system, suggesting that the mechanisms involved in the biocontrol activity of these strains are very similar, regardless of the environmental conditions in which they developed. Comparing clusters has become a common strategy to characterise and understand the internal relations of strains (Takeuchi *et al.*, 2015; Garrido-Sanz *et al.*, 2016; Loper *et al.*, 2016)

A more comprehensive approach also suggests that a more careful understanding of the dynamics and composition of the microbial communities is necessary, in order to formulate more attentive biological control strategies against fungal pathogens (Colla *et al.*, 2012). While selecting biocontrol agents for the strongest ability to inhibit pathogens through a wider range of secondary metabolites may be the most effective, it might not be the preferable choice for a durable protection.

It could be hypothesised that a too strong inhibitory activity might potentially alter the equilibrium in the microorganism community, leading to a comparable response and a simplification of the community itself, eventually causing a decline of the strong biocontrol agent population itself, if it is less capable of adapting to changing conditions. If that is the case, a less impactful bacterium (such as Pf-11), with a more limited array of metabolites, might allow for a more diverse microbial community. For sure, it has become more and more evident that plant disease management cannot overlook the impact of the complex relations between pathogen strains, other competing microorganisms, the various types on environment (soil, hydroponics, root, etc.) and the plants themselves (Hibbing *et al.*, 2010; Stubbendieck and Straight, 2016; Tollenaere *et al.*, 2016).

Sequencing Pf-11 allowed me to prove that its genome includes a large array of secondary metabolite clusters with a broader activity against a variety of fungal species, even larger than Pf-4, yet lacking many of those available to the latter. A larger set of metabolites allows for a wider spectrum of biocontrol activity, or a stronger control towards the same competitor, by exploiting different strategies, possibly in a synergistic combination (Kannan and Sureendar, 2009).

Comparing Pf-4 and Pf-11 also pointed out an issue related to metabarcoding, *i.e.* the use of conserved sequences such as the 16S rRNA: while being the most commonly used method to characterise complex communities, it potentially underestimates the diversity of said communities (Hengstmann *et al.*, 1999; Vrålstad, 2011); in fact, Pf-4 and Pf-11 are indistinguishable using these markers, yet they significantly differ in their inhibitory activity. Nevertheless, they coexist in the same environment, suggesting the necessity of a within-species diversity, which allows for seemingly less fit bacterial strains to survive along with more competitive ones.

A possible explanation might lie in the role of the horizontal transfer of genetic material that is favoured by intra-specific diversity. In the simplest cases, this is caused by insertion events, but the more relevant situation is that of a complex differentiation of accessory genomes (Jackson *et al.*, 2011), due to multiple invasions of foreign DNA that are then integrated in the genome. It could be argued that an expanding/contracting genome is in a more dynamic evolutionary stage, and might eventually result in a stabler genome.

Genomic diversity is often caused by rearrangement of genetic elements; it has become increasingly evident that detecting structural changes, especially those associated with repetitive sequences, can require Third Generation sequencing strategies that produce longer reads than those usually provided by Second Generation sequencers (Stapley *et al.*, 2010). Mobile DNA elements contribute to bacterial evolution, as they can lead to genome rearrangements that can influence their fitness, and possibly their pathogenicity and virulence, in some cases suggesting 'two speed' mechanisms that help pathogens adapting to quickly changing environmental conditions (Faino *et al.*, 2016; Seidl and Thomma, 2017)

Psa biovar 3 is a typical example of this process, as its emergence as a pandemic pathogen of kiwifruit was influenced by horizontal transfer (ICE sequences, in particular). By comparing PacBio sequencings of the CRAFRU 12.29, CRAFRU 14.08 and ICMP 18708 strains, whose longer uninterrupted reads allowed to pinpoint the structural variations between them, using multiple alignment tools, it was possible to note that those structural variations (an insertion in the *hrpS* gene that disrupted the functionality of the T3SS) were caused by a rearrangement of genetic elements, and not by incorporating external DNA, without the recombination-selection process that mitigates genome degeneration associated with transposon mobilization.

In turn, this suggest that more attentive strategies for managing destructive epidemics might want to keep in mind their effect on the short-term genome evolution and population structure of the pathogen, as strategies that do not promote recombination might be at a lesser risk of developing variant, more virulent strains.

Considering the metagenomic approach to pathogen control, another significant difficulty in drawing a complete picture of the relations between pathogens and the larger microbial community is caused by unculturable pathogens, such as phytoplasmas (Lee *et al.*, 2000). As in the case of using barcodes, the difficulties in obtaining reliable *in vitro* cultures of these organisms can lead to an underestimation of the diversity in the community they live in.

While various isolation and purification protocols have been developed, they are generally time-consuming and occasionally specific to single species or even strains. With *Phytoassembly*, it was possible to develop a pipeline that employ a WGS-based strategy to indirectly derive fairly complete genomes from infected plant samples, by using a (preferably isogenic) reference genome of a healthy plant as a filter and exploiting the differential coverage of sequences from the host and the (more abundant) ones from pathogen; this *in silico* approach is an additional evidence of the growing importance of WGS and its potentials in overcoming the difficulties in understanding the more elusive pathogens (Barba *et al.*, 2014; Kakizawa and Yoneda, 2015).

Phytoassembly not only worked well to derive single phytoplasma genomes, but as in the case of the Chicory Phyllody-associated phytoplasma (Martini *et al.*, 2012), it allowed the discovery of a double infection of phytoplasma and spiroplasma, due to the unusal size of the output. By filtering the RAST-annotated genome set with a custom Perl script, it was possible to reconstruct both the phytoplasma and the spiroplasma sequences. In the light of the results of recent investigations it may be speculated that the manipulation of the host gene expression by the phytoplasma is advantageous for and affect the epidemic behaviour of other pathogens, another example of a diveristy-influencing factor that might be underestimated with traditional analysis methods.

As mentioned, the reference genome used with *Phytoassembly* should preferably be isogenic with the infected sample, but if not available a combination of reference genomes can also be used, as with the CFSD-associated phytoplasma (Alvarez *et al.*, 2009).

Expanding the scope to a fully metagenomic approach, WGS is the technology that made possible such perspective, which as mentioned elsewhere can provide a better understanding of the relations between pathogens and the other microorganisms present in the environment, by including all the species (Flynn *et al.*, 2015). While possibly not covering the differences deriving from intra-specific diversity, metabarcoding is still a comprehensive method of surveying a microbial population.

In the case of the kiwifruit endophyte populations, the attempt was to correlate their spatial and temporal variation to the physiological state of the plant related to the severity of the symptoms and time of Psa infection. While further elaborations will be required to fully explore the differences between the samples as resulting from the spatialisation of the OTUs, it was still possible to infer a relation between the observed state of the plants, the composition of the communities from each sample and how these could interact with the pathogen (Brader *et al.*, 2017).

The results from each paper included in this thesis, each from a slightly different perspective, showed how a deep understanding of the data provided by Whole Genome Sequencing requires the use of rigorous bioinformatic analyses and modern computing techniques. Complex, virulent pathogens like Psa requires a level of understanding of their dynamics that, as suggested by the works presented here, benefits from including their relation with the other microorganisms (Tringe *et al.*, 2005), and designing better strategies to contain them could rely on non-pathogenic strains, or on apparently less efficient biocontrol agents, which in the past would have otherwise been overlooked.

5.1 Bibliography

Alvarez, E. *et al.* (2009) 'Characterization of a Phytoplasma Associated with Frogskin Disease in Cassava', *Plant Disease*, 93(11), pp. 1139–1145. doi: 10.1094/PDIS-93-11-1139.

Barba, M., Czosnek, H. and Hadidi, A. (2014) 'Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology', *Viruses*, 6(1), pp. 106–136. doi: 10.3390/v6010106.

Brader, G. *et al.* (2017) 'Ecology and Genomic Insights into Plant-Pathogenic and Plant-Nonpathogenic Endophytes', *Annual Review of Phytopathology*, 55(1), pp. 61–83. doi: 10.1146/annurev-phyto-080516-035641.

Colla, P., Gilardi, G. and Gullino, M. L. (2012) 'A review and critical analysis of the European situation of soilborne disease management in the vegetable sector', *Phytoparasitica*, 40(5), pp. 515–523. doi: 10.1007/s12600-012-0252-2.

Faino, L. *et al.* (2016) ‘Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen’, *Genome Research*, 26(8), pp. 1091–1100. doi: 10.1101/gr.204974.116.

Flynn, J. M. *et al.* (2015) ‘Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods’, *Ecology and Evolution*, 5(11), pp. 2252–2266. doi: 10.1002/ece3.1497.

Garrido-Sanz, D. *et al.* (2016) ‘Genomic and Genetic Diversity within the Pseudomonas fluorescens Complex’, *PLOS ONE*. Edited by B. A. Vinatzer, 11(2), p. e0150183. doi: 10.1371/journal.pone.0150183.

Gasparich, G. E. (2010) ‘Spiroplasmas and phytoplasmas: Microbes associated with plant hosts’, *Biologicals*, 38(2), pp. 193–203. doi: 10.1016/j.biologicals.2009.11.007.

Gross, H. and Loper, J. E. (2009) ‘Genomics of secondary metabolite production by Pseudomonas spp.’, *Natural Product Reports*, 26(11), p. 1408. doi: 10.1039/b817075b.

Hengstmann, U. *et al.* (1999) ‘Comparative phylogenetic assignment of environmental sequences of genes encoding 16S rRNA and numerically abundant culturable bacteria from an anoxic rice paddy soil.’, *Applied and environmental microbiology*, 65(11), pp. 5050–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10543822.

Hibbing, M. E. *et al.* (2010) ‘Bacterial competition: surviving and thriving in the microbial jungle.’, *Nature reviews. Microbiology*. NIH Public Access, 8(1), pp. 15–25. doi: 10.1038/nrmicro2259.

Jackson, R. W. *et al.* (2011) ‘The influence of the accessory genome on bacterial pathogen evolution’, *Mobile Genetic Elements*, 1(1), pp. 55–65. doi: 10.4161/mge.1.1.16432.

Kakizawa, S. and Yoneda, Y. (2015) ‘The role of genome sequencing in phytoplasma research’, *Phytopathogenic Mollicutes*, 5(1), p. 19. doi: 10.5958/2249-4677.2015.00058.4.

Kannan, V. and Sureendar, R. (2009) ‘Synergistic effect of beneficial rhizosphere microflora in biocontrol and plant growth promotion’, *Journal of Basic Microbiology*. WILEY-VCH Verlag, 49(2), pp. 158–164. doi: 10.1002/jobm.200800011.

Lee, I.-M., Davis, R. E. and Gundersen-Rindal, D. E. (2000) ‘Phytoplasma: Phytopathogenic Mollicutes’, *Annual Review of Microbiology*. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA, 54(1), pp. 221–255. doi: 10.1146/annurev.micro.54.1.221.

Loper, J. E. *et al.* (2016) ‘Rhizoxin, orfamide a, and chitinase production contribute to the toxicity of pseudomonas protegens strain pf-5 to drosophila melanogaster’, *Environmental Microbiology*. doi: 10.1111/1462-2920.13369.

Martini, M. *et al.* (2012) ‘Molecular characterization of phytoplasma strains associated with epidemics of chicory phyllody’, *Journal of Plant Pathology*, 94(4, Supplement), p. S4.49.

Seidl, M. F. and Thomma, B. P. H. J. (2017) ‘Transposable Elements Direct The Coevolution between Plants and Microbes’, *Trends in Genetics*, 33(11), pp. 842–851. doi: 10.1016/j.tig.2017.07.003.

Stapley, J. *et al.* (2010) ‘Adaptation genomics: the next generation’, *Trends in Ecology & Evolution*, 25(12), pp. 705–712. doi: 10.1016/j.tree.2010.09.002.

Stubbendieck, R. M. and Straight, P. D. (2016) ‘Multifaceted Interfaces of Bacterial Competition’, *Journal of Bacteriology*. Edited by W. Margolin. American Society for Microbiology, 198(16), pp. 2145–2155. doi: 10.1128/JB.00275-16.

Takeuchi, K. *et al.* (2015) ‘Rhizoxin Analogs Contribute to the Biocontrol Activity of a Newly Isolated Pseudomonas Strain’, *Molecular Plant-Microbe Interactions*, 28(3), pp. 333–342. doi: 10.1094/MPMI-09-14-0294-FI.

Takeuchi, K., Noda, N. and Someya, N. (2014) ‘Complete Genome Sequence of the Biocontrol Strain Pseudomonas protegens Cab57 Discovered in Japan Reveals Strain-Specific Diversity of This Species’, *PLoS ONE*. Edited by P. J. Janssen, 9(4), p. e93683. doi: 10.1371/journal.pone.0093683.

Tollenaere, C., Susi, H. and Laine, A.-L. (2016) ‘Evolutionary and Epidemiological Implications of Multiple Infection in Plants’, *Trends in Plant Science*. Elsevier Ltd, 21(1), pp. 80–90. doi: 10.1016/j.tplants.2015.10.014.

Tringe, S. G. *et al.* (2005) ‘Comparative Metagenomics of Microbial Communities’, *Science*, 308(5721), pp. 554–557. doi: 10.1126/science.1107851.

Vrålstad, T. (2011) ‘ITS, OTUs and beyond-fungal hyperdiversity calls for supplementary solutions’, *Molecular Ecology*. Blackwell Publishing Ltd, 20(14), pp. 2873–2875. doi: 10.1111/j.1365-294X.2011.05149.x.

6 Appendix: Supplementary data

6.1 Genome sequence and antifungal activity in two niche-sharing *Pseudomonas protegens* strains isolated from hydroponics

Table 6.1.1 – OMA-isolated genes exclusive to Pf-4.

Gene code	Description	Position	
A1348_00125	hypothetical protein	29948:30160	F
A1348_00215	lysine transporter LysE	48045:48686	F
A1348_00270	transcriptional regulator	55172:55489	R
A1348_00290	RNA polymerase subunit sigma	59893:60405	F
A1348_00295	iron dicitrate transport regulator FecR	60381:61349	F
A1348_00300	ligand-gated channel	61469:63949	F
A1348_00305	acid phosphatase	64010:64849	R
A1348_00465	lipid A 3-O-deacylase	97975:98493	F
A1348_00985	7-cyano-7-deazaguanine synthase	203207:203905	F
A1348_01010	hypothetical protein	208353:208850	R
A1348_01100	antitoxin	225842:226075	F
A1348_01105	plasmid maintenance protein	226075:226473	F
A1348_01575	cupin	330307:330603	R
A1348_01980	hypothetical protein	415988:416218	R
A1348_02040	hypothetical protein	426717:427472	F
A1348_03470	cupin	733169:733480	F
A1348_03835	hypothetical protein	804067:804255	R
A1348_04005	GDP-fucose synthetase	846573:847550	F
A1348_04010	transferase	847681:848172	F
A1348_04460	mannose-1-phosphate guanylyltransferase/mannose-6-phosphate isomerase	943809:945266	R
A1348_04580	hypothetical protein	981628:984981	R
A1348_05010	hypothetical protein	54372:54707	F
A1348_05320	hypothetical protein	123702:123884	R
A1348_05325	immunity protein	126358:126621	F
A1348_05875	SAM-dependent methyltransferase	243409:244230	R

A1348_06830	AAA family ATPase	462393:464003	R
A1348_06835	organic radical-activating protein	464000:464557	R
A1348_06840	hypothetical protein	464558:465190	R
A1348_06845	hypothetical protein	465192:466175	R
A1348_07190	hypothetical protein	535974:537785	F
A1348_07815	hypothetical protein	669710:670660	F
A1348_07835	transporter	673474:674388	R
A1348_07875	sulfurtransferase	681625:683208	R
A1348_07880	cysteine dioxygenase	683205:683834	R
A1348_07885	LysR family transcriptional regulator	683941:684828	F
A1348_07890	ABC transporter permease	685097:685894	F
A1348_07895	sulfonate ABC transporter ATP-binding protein	685897:686676	F
A1348_07900	acyl-CoA dehydrogenase	686673:687878	F
A1348_07905	dihydrofolate reductase	687875:688813	F
A1348_07910	hypothetical protein	689149:689958	F
A1348_07915	aliphatic sulfonate ABC transporter substrate-binding protein	689971:690933	F
A1348_07920	peptidase M19	691062:692273	F
A1348_07925	monoamine oxidase	692488:694065	F
A1348_07930	hypothetical protein	694154:694651	F
A1348_07935	TonB-dependent receptor	694677:697145	F
A1348_07940	ABC transporter substrate-binding protein	697221:698135	F
A1348_07945	proline hydroxylase	698429:699205	F
A1348_07950	tRNA-dependent cyclodipeptide synthase	699210:699959	F
A1348_07955	MFS transporter	700072:701499	F
A1348_07960	nitrilotriacetate monooxygenase	701492:702844	F
A1348_08035	restriction endonuclease	714963:716048	F
A1348_08735	transposase	875906:876886	F
A1348_08795	hypothetical protein	887576:887758	R
A1348_08840	RTX toxin	898144:900932	F
A1348_29255	type IV secretion protein Rhs	1:572	F
A1348_29440	chemotaxis protein	71986:73602	R
A1348_29515	hypothetical protein	88414:89601	F
A1348_29755	hypothetical protein	21764:22006	R
A1348_29760	hypothetical protein	22632:22916	F
A1348_29780	hypothetical protein	29349:29705	F
A1348_29785	hypothetical protein	29754:30008	R
A1348_29790	AraC family transcriptional regulator	30077:30946	R

A1348_29805	hypothetical protein	32224:32499	R
A1348_30105	large adhesive protein	7799:12114	R
A1348_30115	hypothetical protein	1989:2201	F
A1348_30120	hypothetical protein	1:713	R
A1348_30125	hypothetical protein	710:1210	R
A1348_10605	hypothetical protein	388292:388588	R
A1348_10725	hypothetical protein	414063:414335	F
A1348_10770	hypothetical protein	421772:422425	F
A1348_11280	hypothetical protein	535747:536175	F
A1348_11290	hypothetical protein	536466:537725	R
A1348_11495	ABC transporter ATP-binding protein	577020:577856	F
A1348_11500	nitrate ABC transporter substrate-binding protein	577894:578895	F
A1348_11505	ABC transporter permease	578925:579701	F
A1348_11515	cupin	580478:581014	F
A1348_11520	hydrolase	581019:581897	F
A1348_11535	hypothetical protein	583519:585201	F
A1348_11540	aspartate dehydrogenase	585212:586015	F
A1348_11545	aldehyde dehydrogenase	586129:587625	F
A1348_11550	glyoxalase	587622:588587	F
A1348_11555	hypothetical protein	588615:588842	F
A1348_11560	3-phenylpropionate dioxygenase	588864:589898	F
A1348_11565	ferredoxin	589957:590916	F
A1348_11570	hypothetical protein	591241:592470	F
A1348_11585	hypothetical protein	595537:596082	R
A1348_11695	transporter	624286:624726	F
A1348_12040	histidine kinase	696144:696548	F
A1348_12045	hypothetical protein	696571:696867	F
A1348_30155	type IV secretion protein Rhs	915:1621	F
A1348_30165	type I secretion protein	1:899	F
A1348_30170	hypothetical protein	1:899	F
A1348_30175	hypothetical protein	1:812	R
A1348_12615	large adhesive protein	1:4801	R
A1348_12635	taurine dioxygenase	10395:11282	F
A1348_12640	nitrate ABC transporter substrate-binding protein	11309:12346	F
A1348_12645	sulfonate ABC transporter ATP-binding protein	12352:13212	F
A1348_12650	ABC transporter permease	13246:14112	F

A1348_12715	antitoxin	29384:29629	F
A1348_12720	addiction module toxin RelE	29619:29900	F
A1348_12930	phosphatidylinositol kinase	66585:67805	R
A1348_12935	transcriptional regulator	67798:68046	R
A1348_12945	fatty acid desaturase	69321:70394	F
A1348_12950	pesticin immunity protein	70547:70954	F
A1348_12955	arylsulfatase	71015:72625	R
A1348_12960	ABC transporter ATP-binding protein	72640:73455	R
A1348_12965	ABC transporter permease	73452:75047	R
A1348_12970	nitrate ABC transporter substrate-binding protein	75434:76456	F
A1348_12975	transcriptional regulator	76723:77643	R
A1348_12980	taurine dioxygenase	77827:78732	F
A1348_12985	phosphonate monoester hydrolase	78974:80590	R
A1348_12990	transcriptional regulator	80865:81809	R
A1348_12995	TonB-dependent receptor	81964:84336	F
A1348_13000	ArsR family transcriptional regulator	84403:85752	F
A1348_13030	transcriptional regulator	88576:88827	F
A1348_13035	phosphatidylinositol kinase	88827:89759	F
A1348_13065	hypothetical protein	94041:94364	F
A1348_13070	energy transducer TonB	94784:95599	F
A1348_13075	biopolymer transporter ExbB	95655:96380	F
A1348_13080	biopolymer transporter ExbD	96382:96783	F
A1348_13180	hypothetical protein	119210:119596	R
A1348_13540	hypothetical protein	201586:203268	F
A1348_13545	hypothetical protein	203432:205084	R
A1348_13550	hypothetical protein	205483:206022	F
A1348_13555	transcriptional regulator	206345:206680	F
A1348_13560	hypothetical protein	208517:208879	F
A1348_13565	hypothetical protein	211448:211723	R
A1348_13570	integrase	211716:212700	R
A1348_13585	amidase	215393:216037	F
A1348_13710	hypothetical protein	244712:245041	R
A1348_13730	hypothetical protein	246892:247968	R
A1348_14105	hypothetical protein	331346:331726	R
flgK	flagellar biosynthesis protein FlgK	1:735	F
A1348_30190	terminase	1:727	R

A1348_15630	hypothetical protein	1:642	F
A1348_15635	hypothetical protein	1434:2237	F
A1348_15910	MarR family transcriptional regulator	73975:74934	F
A1348_15920	GntR family transcriptional regulator	76426:77094	F
A1348_15925	ABC transporter substrate-binding protein	77170:77976	F
A1348_15930	polar amino acid ABC transporter permease	77989:78762	F
A1348_15935	ectoine/hydroxyectoine ABC transporter ATP-binding protein EhuA	78765:79544	F
A1348_15940	gamma-glutamyltransferase	79581:81191	F
A1348_15950	sugar ABC transporter substrate-binding protein	81884:82831	F
A1348_15955	TonB-dependent receptor	82899:85298	F
A1348_15960	3-(3-hydroxy-phenyl)propionate transporter MhpT	85381:86592	R
A1348_15965	porin	86925:88244	F
A1348_15970	MarR family transcriptional regulator	88266:88787	R
A1348_15975	p-hydroxycinnamoyl CoA hydratase/lyase	88995:89825	F
A1348_15980	salicylaldehyde dehydrogenase	89931:91379	F
A1348_15985	feruloyl-CoA synthase	91606:93489	F
A1348_15990	acetyl-CoA acetyltransferase	93486:94727	F
A1348_15995	acyl-CoA dehydrogenase	94787:96541	F
A1348_16000	MFS transporter	97060:98385	F
A1348_16035	GntR family transcriptional regulator	106692:107405	F
A1348_16040	Vanillate O-demethylase oxidoreductase	107582:108532	R
A1348_16045	Rieske (2Fe-2S) protein	108592:109650	R
A1348_16095	hypothetical protein	121097:121291	F
A1348_16110	hypothetical protein	123365:123565	F
A1348_16215	hypothetical protein	138378:138758	R
A1348_16230	hypothetical protein	141682:141933	R
A1348_16235	hypothetical protein	142040:142429	F
A1348_16240	hypothetical protein	142430:142720	R
A1348_16245	hypothetical protein	142999:143271	R
A1348_16250	hypothetical protein	143940:144680	R
A1348_16255	hypothetical protein	144791:144997	F
A1348_16260	hypothetical protein	145009:145411	R
A1348_16265	hypothetical protein	145510:145737	F
A1348_16270	hypothetical protein	145877:146554	F
A1348_16275	nucleoid-associated protein YejK	146687:147694	R
A1348_16280	hypothetical protein	148111:148437	R

A1348_16300	hypothetical protein	150519:150725	F
A1348_16320	hypothetical protein	152288:152560	R
A1348_16325	hypothetical protein	152618:152836	R
A1348_16330	hypothetical protein	152913:153113	R
A1348_16335	hypothetical protein	153865:154740	R
A1348_16340	transposase	154980:155420	F
A1348_16345	hypothetical protein	155706:155999	F
A1348_16350	hypothetical protein	156217:156426	R
A1348_16355	hypothetical protein	156452:157222	R
A1348_16360	hypothetical protein	157322:157534	F
A1348_16365	hypothetical protein	157588:158202	F
A1348_16375	hypothetical protein	159125:159469	F
A1348_16380	phage replication protein	159471:160436	F
A1348_16390	hypothetical protein	161205:161711	F
A1348_16395	hypothetical protein	161708:161995	F
A1348_16405	hypothetical protein	162572:162838	F
A1348_16410	hypothetical protein	162835:163065	F
A1348_16425	hypothetical protein	165241:165558	F
A1348_16435	hypothetical protein	166334:166693	R
A1348_16450	terminase	168161:169483	F
A1348_16465	hypothetical protein	172478:173206	F
A1348_16475	hypothetical protein	174228:174812	F
A1348_16580	integrase	193933:194325	R
A1348_16595	hypothetical protein	195955:196605	R
A1348_16600	hypothetical protein	197225:197467	R
A1348_16605	hypothetical protein	198425:198790	F
A1348_16610	hypothetical protein	199046:199399	R
A1348_16615	acetyltransferase	200642:201172	R
A1348_16635	hypothetical protein	205076:205375	F
A1348_16640	aminoglycoside phosphotransferase	205405:206181	F
A1348_16890	hypothetical protein	253936:255027	F
A1348_16950	hypothetical protein	267554:268399	F
A1348_16955	hypothetical protein	268483:268833	F
A1348_17170	alkaline phosphatase	324812:325972	F
A1348_17270	halogenase	360091:361599	R
A1348_17275	transcriptional regulator	361596:362627	R

A1348_17280	peptidyl carrier protein PltL	363114:363380	F
A1348_17285	FADH2-dependent halogenase PltA	363394:364743	F
A1348_17290	polyketide synthase	364776:372152	F
A1348_17300	halogenase	377576:379210	F
A1348_17305	acyl-CoA dehydrogenase	379212:380354	F
A1348_17310	D-alanine--poly(phosphoribitol) ligase	380351:381844	F
A1348_17315	thioesterase	381848:382630	F
A1348_17320	transcriptional regulator	382636:383307	R
A1348_17325	hypothetical protein	383383:384396	F
A1348_17330	ABC transporter ATP-binding protein	384393:386162	F
A1348_17335	hypothetical protein	386172:387314	F
A1348_17340	antibiotic ABC transporter permease	387331:388437	F
A1348_17345	RND transporter	388449:389945	F
A1348_17350	transporter	390011:390616	F
A1348_17365	alpha/beta hydrolase	392399:393268	R
A1348_17375	alkene reductase	394187:395287	R
A1348_17475	hypothetical protein	414065:414451	R
A1348_17480	hypothetical protein	414521:414805	R
A1348_17485	hypothetical protein	414923:415360	F
A1348_17495	amidase	416352:417638	F
A1348_17510	hypothetical protein	419490:420233	F
A1348_17515	hypothetical protein	422227:423993	F
A1348_17525	hypothetical protein	425271:425900	F
A1348_17750	glycosyl transferase	481588:482241	R
A1348_17755	methyltransferase	482238:482837	R
A1348_17760	acetylglucosaminylphosphatidylinositol deacetylase	482834:483592	R
A1348_17765	acyl-CoA dehydrogenase	483589:484602	R
A1348_18380	hypothetical protein	622032:622726	F
A1348_30205	hypothetical protein	1:189	R
A1348_18495	hypothetical protein	57784:58044	F
A1348_18665	type B chloramphenicol O-acetyltransferase	85320:85955	F
A1348_18950	hypothetical protein	142049:143482	F
A1348_18955	hypothetical protein	143872:144651	F
A1348_18960	DNA polymerase V subunit UmuC	144768:146045	R
A1348_18965	peptidase S24	146038:146463	R
A1348_18970	hypothetical protein	146560:146769	R

A1348_18975	hypothetical protein	147004:148581	R
A1348_19010	phage tail protein	152627:155185	R
A1348_19015	hypothetical protein	155327:155632	R
A1348_19020	phage tail protein	155645:156154	R
A1348_19025	phage tail protein	156167:157333	R
A1348_19035	phage tail protein	157869:159773	R
A1348_19040	phage tail protein	159770:160375	R
A1348_19045	baseplate assembly protein	160377:161258	R
A1348_19050	phage baseplate protein	161255:161581	R
A1348_19055	hypothetical protein	161586:161789	R
A1348_19060	phage baseplate protein	161854:162447	R
A1348_19065	hypothetical protein	162444:162956	R
A1348_19070	hypothetical protein	162949:163602	R
A1348_19075	hypothetical protein	163599:163913	R
A1348_19080	major capsid protein E	163916:164911	R
A1348_19085	hypothetical protein	164975:165319	R
A1348_19090	Clp protease ClpP	165316:166458	R
A1348_19095	portal protein	166455:167936	R
A1348_19100	hypothetical protein	167936:168142	R
A1348_19105	terminase	168144:170156	R
A1348_19110	terminase small subunit	170161:170733	R
A1348_19140	peptidase S24	175807:176520	F
A1348_19170	hypothetical protein	178601:178822	F
A1348_19175	hypothetical protein	178864:179160	F
A1348_19180	DNA methyltransferase	179150:180883	F
A1348_19185	hypothetical protein	181407:181880	F
A1348_19950	hypothetical protein	361445:361741	F
A1348_30220	conjugal transfer protein	1:615	R
A1348_30225	mammalian cell entry protein	1:608	R
A1348_21245	permease	9187:10101	R
A1348_21250	transcriptional regulator	10202:10978	R
A1348_21420	IclR family transcriptional regulator	44127:44903	R
A1348_21425	ABC transporter substrate-binding protein	45074:45838	F
A1348_21430	amino acid ABC transporter permease	45899:46555	F
glnQ	glutamine ABC transporter ATP-binding protein	46552:47298	F
A1348_21440	FAD-dependent oxidoreductase	47295:48584	F

A1348_21480	LysR family transcriptional regulator	57715:58622	R
A1348_21485	short-chain dehydrogenase	58821:59678	F
A1348_21505	hypothetical protein	63792:64385	F
A1348_21605	hypothetical protein	87826:88128	F
A1348_21610	hypothetical protein	88354:89010	R
A1348_21615	hypothetical protein	89122:89529	R
A1348_21645	AraC family transcriptional regulator	94388:95353	R
A1348_21650	MFS transporter	95513:96724	F
A1348_21665	hypothetical protein	101116:101817	R
A1348_21670	thiamine biosynthesis protein ApbE	101873:102847	R
A1348_21675	nitric oxide synthase	102837:105026	R
A1348_21680	Tat pathway signal protein	105053:105523	R
A1348_21685	DNA-binding response regulator	105758:106417	F
A1348_21690	two-component sensor histidine kinase	106414:107757	F
A1348_21755	hypothetical protein	121340:122269	F
A1348_21760	hypothetical protein	122419:122943	F
A1348_21765	hypothetical protein	123174:123527	F
A1348_21775	hypothetical protein	125735:126490	R
A1348_21780	colicin transporter	126737:126991	R
A1348_21805	hypothetical protein	131316:132215	R
A1348_21815	MerR family transcriptional regulator	133791:134180	F
A1348_21825	hypothetical protein	134871:135530	R
A1348_21910	hypothetical protein	152325:152837	R
A1348_21930	GNAT family acetyltransferase	156912:157487	F
A1348_21935	hypothetical protein	157524:157958	R
A1348_21955	hypothetical protein	160070:161089	F
A1348_21960	hypothetical protein	161086:161421	F
A1348_21965	hypothetical protein	161411:161599	F
A1348_22020	ABC transporter permease	168325:169284	R
A1348_22025	transcriptional regulator	169438:170334	F
A1348_22035	riboflavin-specific deaminase	170816:171253	F
A1348_22040	hypothetical protein	171316:171846	F
A1348_22090	hypothetical protein	181168:181653	R
A1348_22100	hypothetical protein	182448:182999	R
A1348_22105	hypothetical protein	183068:183829	R
A1348_22125	ligand-gated channel protein	186693:189131	R

A1348_22135	RNA polymerase subunit sigma	190196:190702	R
A1348_22175	LysR family transcriptional regulator	199277:200152	R
A1348_22180	peptidyl-arginine deiminase	200290:201348	F
A1348_22185	N-carbamoylputrescine amidase	201345:202253	F
A1348_22190	ABC transporter substrate-binding protein	202255:203376	F
A1348_22195	agmatine deiminase	203524:204651	F
A1348_22265	hypothetical protein	216898:219108	R
A1348_22270	hypothetical protein	219099:219299	R
A1348_22275	hypothetical protein	219706:220326	R
A1348_22370	hypothetical protein	238369:238836	R
A1348_22820	hypothetical protein	335526:336224	F
A1348_22870	diaminopimelate epimerase	347954:348748	R
A1348_22940	ABC transporter permease	365055:366068	F
A1348_22960	hypothetical protein	369107:369514	R
A1348_23055	hypothetical protein	389106:389492	F
A1348_23060	hypothetical protein	389682:390623	F
A1348_23265	acyl-CoA synthetase	435248:436237	F
A1348_23270	ABC transporter ATP-binding protein	436234:437010	F
A1348_23275	ABC transporter permease	437010:437888	F
A1348_23280	ABC transporter permease	437893:438942	F
A1348_23285	ABC transporter permease	438971:440305	F
A1348_23290	ABC transporter ATP-binding protein	440309:441076	F
A1348_23405	hypothetical protein	467911:468312	F
A1348_23445	hypothetical protein	476273:476920	F
A1348_23450	hypothetical protein	476917:477387	F
A1348_23480	hypothetical protein	480392:481999	F
A1348_23485	hypothetical protein	482058:482744	R
A1348_23500	RNA polymerase subunit sigma-24	484024:484623	F
A1348_23580	ATP-dependent endonuclease	499198:500982	F
A1348_23585	DNA/RNA helicase	500964:502865	F
A1348_23590	hypothetical protein	503249:503974	F
A1348_23745	hypothetical protein	542601:544184	R
A1348_23750	hypothetical protein	544181:546673	R
A1348_23875	hypothetical protein	573885:574388	R
A1348_23965	DNA-binding protein	17387:17962	R
A1348_24010	transcriptional regulator	27036:27233	R

A1348_24015	hypothetical protein	27230:27604	R
A1348_24285	cytochrome B	85173:85727	R
A1348_24330	amine oxidase	95481:96866	F
A1348_24335	cupin	96882:97226	F
A1348_24340	regulator	97216:98721	R
A1348_24345	spermidine/putrescine ABC transporter substrate-binding protein	98930:99946	F
A1348_24895	CopG family transcriptional regulator	215279:215551	F
A1348_24905	restriction endonuclease	216530:217441	R
A1348_24910	restriction endonuclease subunit R	217451:220648	R
A1348_24915	hypothetical protein	220648:221868	R
A1348_24920	restriction endonuclease subunit M	222045:224027	R
A1348_24925	fructose-bisphosphate aldolase	224343:224624	R
A1348_25115	hypothetical protein	262102:262341	F
A1348_25120	Holliday junction resolvase	262462:262785	R
A1348_25125	hypothetical protein	262903:263616	R
A1348_25135	hypothetical protein	264171:264452	F
A1348_25140	hypothetical protein	264459:264728	F
A1348_25225	hypothetical protein	280126:281694	R
A1348_25230	hypothetical protein	281875:283428	R
A1348_25380	plasmid stabilization protein ParE	318168:318524	R
A1348_25440	hypothetical protein	335315:337831	R
A1348_25445	hypothetical protein	337898:339991	R
A1348_25485	hypothetical protein	359473:359823	F
A1348_25490	fimbrial protein	360488:361036	F
A1348_25500	fimbrial protein	361911:364373	F
A1348_25505	hypothetical protein	364461:368051	R
A1348_25510	DNA-binding response regulator	368062:368685	R
A1348_25515	hypothetical protein	369021:369980	F
A1348_25560	hypothetical protein	379276:379551	R
A1348_25565	glucosidase	379748:382381	F
A1348_25725	hypothetical protein	424620:425150	F
A1348_25775	hypothetical protein	435725:435970	R
A1348_25780	serine hydrolase	436222:437841	F
A1348_25795	hypothetical protein	444938:445366	R
A1348_25800	hypothetical protein	445348:446046	R
A1348_25855	hypothetical protein	457880:458704	F

A1348_25860	hypothetical protein	458856:459623	F
A1348_25875	transcriptional regulator	462042:462467	R
A1348_25880	mRNA interferase	462518:462697	R
A1348_25885	hypothetical protein	462787:463053	R
A1348_26080	type IV secretion protein Rhs	511853:512559	F
A1348_26085	hypothetical protein	1:651	F
A1348_26230	hypothetical protein	30283:31623	F
A1348_26240	hypothetical protein	32325:33269	R
A1348_26245	hypothetical protein	33256:34881	R
A1348_26320	hypothetical protein	60100:60516	R
A1348_26325	hypothetical protein	61124:62077	F
A1348_26330	ATP-dependent exonuclease	62151:63221	R
A1348_26335	ATP-dependent endonuclease	63214:64818	R
A1348_26380	NAD-dependent dehydratase	79396:80252	R
A1348_26445	hypothetical protein	92020:93603	R
A1348_26450	hypothetical protein	94157:94720	F
A1348_26455	hypothetical protein	94743:96380	R
A1348_26460	hypothetical protein	96380:97372	R
A1348_26465	hypothetical protein	97378:97959	R
A1348_26470	hypothetical protein	98131:99558	R
A1348_26475	polyketide synthase	99945:107012	R
A1348_26480	acyl transferase	106937:108964	R
A1348_26485	SAM-dependent methyltransferase	109125:109991	F
A1348_26490	polyketide synthase	110029:117654	R
A1348_26495	polyketide synthase	117720:130220	R
A1348_26500	cytochrome P450	130286:131695	R
A1348_26505	polyketide synthase	131692:143814	R
A1348_26510	polyketide synthase	143811:158636	R
A1348_26515	polyketide synthase	158807:178849	R
A1348_26520	hypothetical protein	179502:179906	F
A1348_27055	flavin reductase	324880:325440	R
A1348_27060	potassium transporter	325465:326700	R
A1348_27065	2Fe-2S ferredoxin	326813:327904	R
A1348_27070	FAD-dependent oxidoreductase	327929:329632	R
A1348_27075	monodechloroaminopyrrolnitrin synthase PrnB	329674:330759	R
A1348_27080	tryptophan halogenase	330759:332375	R

A1348_27135	hypothetical protein	345390:345572	R
A1348_27540	hypothetical protein	447277:449721	F
A1348_28115	hypothetical protein	114675:115046	R
A1348_28605	hypothetical protein	222974:223395	F
A1348_29000	diguanylate phosphodiesterase	307568:308742	R

Table 6.1.2 – OMA-isolated genes exclusive to Pf-11.

Gene code	Description	Position	
A1395_00130	hypothetical protein	31786:32013	F
A1395_00295	hypothetical protein	63051:63428	R
A1395_00300	GCN5 family acetyltransferase	64070:64507	R
A1395_00305	GNAT family acetyltransferase	64805:65353	R
A1395_00310	AraC family transcriptional regulator	65511:66359	F
A1395_00460	hypothetical protein	99323:99508	R
A1395_00575	hypothetical protein	125921:126262	F
A1395_00680	hypothetical protein	153783:154220	R
A1395_01315	hypothetical protein	283126:284442	R
A1395_01320	hypothetical protein	284476:285237	R
A1395_01325	hypothetical protein	285284:286510	R
A1395_01330	hypothetical protein	286498:287235	R
A1395_01335	hypothetical protein	287228:288253	R
A1395_01340	hypothetical protein	288250:289605	R
A1395_02460	phage tail protein	543394:544479	F
A1395_02465	hypothetical protein	544476:544856	F
A1395_02645	hypothetical protein	578866:579078	F
A1395_03120	GntR family transcriptional regulator	681101:682102	R
A1395_03125	amidase	682469:683827	F
A1395_03130	polysaccharide deacetylase	683875:684756	F
A1395_03135	MFS transporter	684785:686125	F
A1395_03150	hypothetical protein	687791:688090	F
A1395_03155	cobalamin biosynthesis protein CobW	688087:689094	F
A1395_03160	signal peptide prediction	689094:690302	F
A1395_03190	hypothetical protein	695707:696003	F

A1395_03470	hypothetical protein	751292:751510	F
A1395_03520	hypothetical protein	758156:758347	F
A1395_03780	hypothetical protein	811187:812647	F
A1395_03785	hypothetical protein	812664:813638	R
A1395_03790	hypothetical protein	813681:813905	R
A1395_03795	hypothetical protein	813902:814240	R
A1395_03800	hypothetical protein	814345:814644	R
A1395_03805	hypothetical protein	814659:815354	R
A1395_03810	hypothetical protein	815503:815898	R
A1395_03815	serine recombinase	816281:816916	R
A1395_03820	hypothetical protein	817443:818366	F
A1395_03850	LysR family transcriptional regulator	823835:824779	R
A1395_03855	tricarboxylate transporter	824850:826379	R
A1395_03860	tripartite tricarboxylate transporter TctB	826376:826909	R
A1395_03865	tricarboxylate transporter	826985:827983	R
A1395_03870	4-hydroxythreonine-4-phosphate dehydrogenase	828481:829485	F
A1395_03875	MFS transporter	829829:831154	F
A1395_03880	L-rhamnonate dehydratase	831188:832363	F
A1395_03885	FAH family protein	832416:833411	F
A1395_03890	ketoglutarate semialdehyde dehydrogenase	833454:835034	F
A1395_04065	transposase	887179:887571	F
A1395_04070	transposase	887663:888322	F
A1395_04075	hypothetical protein	888858:889109	F
A1395_04080	hypothetical protein	889288:889503	F
A1395_04085	hypothetical protein	889555:890160	F
A1395_05130	hypothetical protein	1109019:1109246	R
A1395_05430	hypothetical protein	1167530:1167827	R
A1395_05475	hypothetical protein	1177038:1177223	F
A1395_05770	hypothetical protein	1253202:1254143	F
A1395_06200	hypothetical protein	1349559:1350614	F
A1395_06205	DNA cytosine methyltransferase	1350967:1352031	R
A1395_06215	hypothetical protein	1352605:1353651	R
A1395_06220	hypothetical protein	1353656:1354189	R
A1395_06225	hypothetical protein	1354186:1354803	R
A1395_06230	hypothetical protein	1354880:1355692	R
A1395_06235	hypothetical protein	1355732:1356082	R

A1395_06240	transcriptional regulator	1356133:1356375	R
A1395_06245	hypothetical protein	1356372:1356755	R
A1395_06250	hypothetical protein	1356752:1357027	R
A1395_06255	deoxynucleotide monophosphate kinase	1357038:1357613	R
A1395_06260	hypothetical protein	1357610:1357795	R
A1395_06265	hypothetical protein	1357943:1358665	R
A1395_06270	hypothetical protein	1358782:1358994	F
A1395_06275	hypothetical protein	1359133:1359621	F
A1395_06285	hypothetical protein	1359817:1360854	F
A1395_06290	hypothetical protein	1360855:1362348	F
A1395_06295	hypothetical protein	1362596:1362979	F
A1395_06300	hypothetical protein	1362976:1363458	F
A1395_06310	hypothetical protein	1364112:1364855	R
A1395_06315	holin	1365147:1365464	F
A1395_06320	terminase	1365639:1366169	F
A1395_06325	terminase	1366132:1367967	F
A1395_06330	hypothetical protein	1367964:1368152	F
A1395_06335	portal protein	1368155:1369426	F
A1395_06340	capsid protein	1369423:1370406	F
A1395_06345	phage capsid protein	1370488:1371807	F
A1395_06350	hypothetical protein	1:244	R
A1395_06485	hypothetical protein	33006:33248	F
A1395_06725	hypothetical protein	98802:99857	R
A1395_06790	hypothetical protein	112442:114394	F
A1395_06885	lipase	136790:139096	R
A1395_06890	hypothetical protein	139093:139887	R
A1395_06895	hypothetical protein	139928:140722	R
A1395_06900	hypothetical protein	140719:141567	R
A1395_06905	type IV secretion protein Rhs	141564:143638	R
A1395_06925	hypothetical protein	147677:148468	R
A1395_06935	hypothetical protein	149385:150185	R
A1395_07270	hypothetical protein	232630:232917	F
A1395_07740	hypothetical protein	332845:334165	F
A1395_07745	hypothetical protein	334565:335089	F
A1395_07825	hypothetical protein	360614:361339	R
A1395_07830	hypothetical protein	361750:365682	R

A1395_07910	TonB-dependent receptor	388024:390006	F
A1395_07915	hypothetical protein	390084:390458	F
A1395_07920	hypothetical protein	390452:391807	R
A1395_07925	hypothetical protein	391794:392315	R
A1395_07930	methanobactin biosynthesis cassette protein MbnB	392369:393181	R
A1395_07935	cytochrome-c peroxidase	393485:394621	R
A1395_07940	hypothetical protein	394618:395514	R
A1395_07945	membrane receptor protein	395531:397666	R
A1395_07950	hypothetical protein	397739:398164	R
A1395_08120	hypothetical protein	449154:449438	R
A1395_08125	hypothetical protein	449605:451452	F
A1395_08130	hypothetical protein	451449:453092	F
A1395_08135	hypothetical protein	453093:454982	F
A1395_08140	hypothetical protein	455032:456114	F
A1395_08590	zinc-binding alcohol dehydrogenase	568197:568754	R
A1395_08595	NADPH dehydrogenase	568890:569954	R
A1395_08600	HxlR family transcriptional regulator	570095:570517	R
A1395_08605	NmrA family transcriptional regulator	570753:571616	F
A1395_08610	alcohol dehydrogenase	571694:572683	F
A1395_08615	branched-chain amino acid ABC transporter substrate-binding protein	573081:574361	F
A1395_08620	ABC transporter permease	574358:575260	F
A1395_08625	branched-chain amino acid ABC transporter permease	575260:576201	F
A1395_08630	ABC transporter ATP-binding protein	576198:576938	F
A1395_08635	ABC transporter ATP-binding protein	576950:577669	F
A1395_08640	aldehyde dehydrogenase	577666:579099	F
fabG	3-ketoacyl-ACP reductase	579144:579896	F
A1395_08650	alcohol dehydrogenase	580164:581261	F
A1395_08655	aldehyde dehydrogenase	581304:582761	F
A1395_08660	hypothetical protein	582777:583748	R
A1395_08665	aminomethyltransferase	584187:585845	F
A1395_08670	sarcosine oxidase	585847:586452	F
A1395_08680	cation transporter	589461:589790	R
A1395_08685	cation transporter	589793:590134	R
A1395_08690	mammalian cell entry protein	590800:592491	R
A1395_08695	paraquat-inducible protein A	592484:593104	R
A1395_08700	paraquat-inducible protein A	593101:593706	R

A1395_08705	two-component sensor histidine kinase	593970:595154	R
A1395_08710	DNA-binding response regulator	595151:595852	R
A1395_08715	hypothetical protein	596057:597445	F
A1395_08720	hypothetical protein	597442:598275	F
A1395_08725	acriflavin resistance protein	599430:602516	F
A1395_08730	hypothetical protein	603883:604071	R
A1395_08735	hypothetical protein	605103:605381	F
A1395_08740	hypothetical protein	605666:606421	F
A1395_08745	AlpA family transcriptional regulator	606531:606761	F
A1395_08750	cobyrinic acid a,c-diamide synthase	606766:607605	F
A1395_08755	hypothetical protein	607691:607900	F
A1395_08760	chromosome partitioning protein ParB	607915:609495	F
A1395_08765	hypothetical protein	609492:610040	F
A1395_08770	hypothetical protein	610037:611227	F
A1395_08775	conjugal transfer protein	611373:612128	F
A1395_08780	integrase	612125:612643	F
A1395_08785	single-stranded DNA-binding protein	612640:613110	F
A1395_08790	DNA topoisomerase III	613333:615300	F
A1395_08795	hypothetical protein	615342:615596	R
A1395_08800	transcriptional regulator	615805:616146	F
A1395_08805	hypothetical protein	617059:617640	F
A1395_08810	hypothetical protein	617711:617968	F
A1395_08815	DNA repair protein RadC	618069:618563	F
A1395_08820	ABC transporter substrate-binding protein	618616:619371	F
A1395_08825	pili assembly chaperone	619460:620002	F
A1395_08830	chemotaxis protein	619999:620667	F
A1395_08835	hypothetical protein	620677:621390	F
A1395_08840	lytic transglycosylase	621372:621920	F
A1395_08845	conjugal transfer protein	621917:622432	F
A1395_08850	conjugal transfer protein TraG	622442:624547	F
A1395_08855	hypothetical protein	624637:625368	F
A1395_08860	hypothetical protein	625389:625970	R
A1395_08865	conjugal transfer protein	626270:626605	F
A1395_08870	conjugal transfer protein	626602:626841	F
A1395_08875	conjugal transfer protein	626859:627230	F
A1395_08880	conjugal transfer protein	627238:627651	F

A1395_08885	hypothetical protein	627648:628289	F
A1395_08890	conjugal transfer protein	628286:629101	F
A1395_08895	conjugal transfer protein	629091:630608	F
A1395_08900	conjugal transfer protein	630571:630987	F
A1395_08905	conjugal transfer protein	630987:633725	F
A1395_08910	hypothetical protein	634115:634690	F
A1395_08915	hypothetical protein	634928:635185	R
A1395_08920	transcriptional regulator	635357:635611	F
A1395_08925	conjugal transfer protein	635717:636673	F
A1395_08930	conjugal transfer protein	636670:638052	F
A1395_08935	conjugal transfer protein TraG	638375:639916	F
A1395_08940	hypothetical protein	639905:640303	R
A1395_08945	hypothetical protein	640761:641090	F
A1395_08950	DNA primase	641226:642185	F
A1395_08955	hypothetical protein	642323:642520	R
A1395_08960	relaxase	642742:644466	F
A1395_08965	hypothetical protein	644784:648611	F
A1395_08970	DNA helicase UvrD	648635:650521	R
A1395_08975	ATP-dependent endonuclease	650524:652494	R
A1395_08980	hypothetical protein	656186:656383	F
A1395_08985	hypothetical protein	657752:658132	F
A1395_08990	hypothetical protein	658179:659084	R
A1395_08995	hypothetical protein	659406:660026	F
A1395_09000	DNA-binding protein	660019:661137	F
A1395_09010	hypothetical protein	662506:663186	R
A1395_09015	hypothetical protein	663173:664132	R
A1395_09020	hypothetical protein	664147:664704	R
A1395_09025	hypothetical protein	664841:665104	F
A1395_09030	hypothetical protein	665341:666342	R
A1395_09035	hypothetical protein	667043:668242	R
A1395_09040	phytanoyl-CoA dioxygenase	668534:669403	F
A1395_09045	glycosyl hydrolase	669520:670482	R
A1395_09050	RND transporter	670493:672907	R
A1395_09055	arylsulfatase	672916:674502	R
A1395_09060	hypothetical protein	674587:675768	R
A1395_09065	hypothetical protein	676483:677178	R

A1395_09070	hypothetical protein	677308:677523	R
A1395_09075	phytanoyl-CoA dioxygenase	677595:678482	R
A1395_09080	TetR family transcriptional regulator	678865:679443	F
A1395_09085	hypothetical protein	679437:679652	R
A1395_09090	FAD-containing monooxygenase EthA	679745:681271	F
A1395_09095	alpha/beta hydrolase	681268:682143	F
A1395_09100	1,3-propanediol dehydrogenase	682337:683494	F
A1395_09105	amino acid permease	683507:684838	F
A1395_09110	gamma-aminobutyraldehyde dehydrogenase	684918:686342	F
A1395_09115	diaminobutyrate--2-oxoglutarate transaminase	686388:687641	F
A1395_09120	hypothetical protein	687917:688174	F
A1395_09125	oxidoreductase	688326:689126	F
A1395_09130	amino acid permease	689194:689346	F
A1395_09135	cation acetate symporter	689397:691055	R
A1395_09140	hypothetical protein	691052:691375	R
A1395_09145	acyl-CoA synthetase	691432:693084	R
A1395_09155	acyl-CoA dehydrogenase	693935:695107	R
A1395_09160	acyl-CoA dehydrogenase	695161:695517	R
A1395_09165	aminoglycoside phosphotransferase	695514:696572	R
A1395_09170	propionate catabolism operon regulatory protein PrpR	696772:698724	F
A1395_09175	transcriptional regulator	698910:699470	F
A1395_09180	transcriptional regulator	699667:700188	R
A1395_09185	mammalian cell entry protein	700230:702542	R
A1395_09190	paraquat-inducible protein A	702535:703155	R
A1395_09195	multidrug transporter	703156:704592	R
A1395_09200	multidrug efflux RND transporter permease subunit	704592:707723	R
A1395_09205	efflux transporter periplasmic adaptor subunit	707751:708902	R
A1395_09210	transcriptional regulator	709027:709632	F
A1395_09215	paraquat-inducible protein A	709658:710296	F
A1395_09220	hypothetical protein	710906:711655	R
A1395_09670	ATPase	824172:828644	R
A1395_09675	diguanylate cyclase	828673:830373	R
A1395_09880	hypothetical protein	878131:878385	R
A1395_09930	Cro/Cl family transcriptional regulator	890638:890889	F
A1395_09935	hypothetical protein	890940:891221	R
A1395_09940	hypothetical protein	891442:894669	R

A1395_09945	hypothetical protein	894859:895572	R
A1395_09975	hypothetical protein	899701:900729	R
A1395_10010	alpha/beta hydrolase	908065:908679	F
A1395_10035	hypothetical protein	911077:911289	R
A1395_10195	hypothetical protein	943968:944363	F
A1395_10225	hypothetical protein	950446:950694	R
A1395_10465	hypothetical protein	1001825:1002019	F
A1395_10540	ABC transporter permease	1017547:1019631	R
A1395_10610	hypothetical protein	1035794:1036242	F
A1395_11195	hypothetical protein	1162351:1162548	F
A1395_11295	hypothetical protein	1183129:1183890	F
A1395_11300	hypothetical protein	1183887:1184528	F
A1395_11355	hypothetical protein	1193655:1193972	F
A1395_11420	hypothetical protein	1204360:1204701	F
A1395_11475	hypothetical protein	1213964:1214329	F
A1395_11510	hypothetical protein	1218427:1218633	F
A1395_11515	GCN5 family acetyltransferase	1218715:1218951	F
A1395_11570	hypothetical protein	1231649:1232203	F
A1395_11575	fimbrial assembly protein	1232896:1235331	F
A1395_11585	hypothetical protein	1236144:1236635	F
A1395_11590	hypothetical protein	1236632:1237129	F
A1395_11595	hypothetical protein	1237126:1237629	F
A1395_11600	hypothetical protein	1237638:1238168	F
A1395_11605	hypothetical protein	1238690:1239010	F
A1395_11610	hypothetical protein	1239531:1240047	R
A1395_11615	DNA-binding response regulator	1240178:1240807	R
A1395_11620	diguanylate phosphodiesterase	1241030:1242229	F
A1395_11625	hybrid sensor histidine kinase/response regulator	1242327:1245974	F
A1395_11635	hypothetical protein	1247131:1247337	F
A1395_11665	hypothetical protein	1253113:1253742	F
A1395_11785	hypothetical protein	1281902:1282573	F
A1395_11790	hypothetical protein	1282586:1283764	R
A1395_30045	hypothetical protein	44393:44761	R
A1395_30240	Rhs element Vgr protein	118205:118801	R
A1395_31585	flagellar M-ring protein FliF	94:774	F
A1395_31590	sigma-54-dependent Fis family transcriptional regulator	1:434	R

A1395_31595	hypothetical protein	442:756	R
A1395_30260	hypothetical protein	4380:5180	F
A1395_30525	hypothetical protein	73155:73391	R
A1395_30530	hypothetical protein	73463:73933	F
A1395_30630	hypothetical protein	98586:100481	F
A1395_30635	integrase	103147:104211	R
A1395_30640	hypothetical protein	104215:104457	R
A1395_30645	hypothetical protein	104481:105113	F
A1395_30675	hypothetical protein	107690:108856	F
A1395_30695	pyocin R2, holin	114067:114408	F
A1395_30700	holin	114469:114786	F
A1395_30705	terminase	114961:115491	F
A1395_31605	hypothetical protein	1:214	R
A1395_31615	hypothetical protein	1:260	R
A1395_31620	hypothetical protein	257:613	R
A1395_31625	glutamine synthetase	1:741	R
A1395_31630	hypothetical protein	1:739	F
A1395_31635	ribosome maturation factor	1:198	F
nusA	transcription termination/antitermination protein NusA	246:737	F
A1395_30810	hypothetical protein	22280:22522	F
A1395_30835	hypothetical protein	26864:27322	R
A1395_30840	antibiotic biosynthesis monooxygenase	27355:27723	R
A1395_30845	MarR family transcriptional regulator	27856:28284	F
A1395_30850	hypothetical protein	28478:28909	F
A1395_30860	hypothetical protein	29672:29854	R
A1395_30865	hypothetical protein	29935:30177	F
A1395_30870	transposase	30241:30501	R
A1395_30875	hypothetical protein	30634:31209	R
A1395_30880	hypothetical protein	31808:32395	F
A1395_30885	acetylornithine aminotransferase	33761:34999	F
A1395_30890	hypothetical protein	35052:36764	F
A1395_30895	short-chain dehydrogenase	36861:37706	F
A1395_30900	hypothetical protein	37699:38322	F
A1395_30905	NTD biosynthesis hydrolase NtdB	38329:39114	F
A1395_30910	acetylserotonin O-methyltransferase	39133:40158	F
A1395_30915	hypothetical protein	40205:41458	F

A1395_30920	CopG family transcriptional regulator	41929:42286	R
A1395_30925	hypothetical protein	42452:43045	R
A1395_30930	hypothetical protein	43042:43515	R
A1395_30935	CopG family transcriptional regulator	45501:45902	R
A1395_30940	hypothetical protein	46384:46626	R
A1395_30945	His-Xaa-Ser system radical SAM maturase HxsC	47222:48451	R
A1395_30950	His-Xaa-Ser system radical SAM maturase HxsB	48457:49944	R
A1395_30960	hypothetical protein	51891:52589	F
A1395_31055	hypothetical protein	75592:75848	R
A1395_31645	leucyl aminopeptidase	1:453	F
A1395_31650	DNA polymerase III subunit chi	511:733	F
A1395_31660	type IV secretion protein Rhs	1:731	R
A1395_31215	phage head-tail joining protein	55:417	R
A1395_31220	hypothetical protein	419:1006	R
A1395_31225	hypothetical protein	1009:1356	R
A1395_31230	phage capsid protein	1411:2730	R
A1395_31235	capsid protein	2812:3795	R
A1395_31240	portal protein	3792:5063	R
A1395_31245	hypothetical protein	5066:5254	R
A1395_31250	terminase	5251:5913	R
A1395_31665	transposase	70:730	R
A1395_31670	FAD-dependent oxidoreductase	1:442	R
A1395_31675	phage tail protein	1:720	R
A1395_31680	hypothetical protein	1:717	F
A1395_31685	SAM-dependent methyltransferase	89:493	F
A1395_31690	hypothetical protein	25:701	F
A1395_31695	hypothetical protein	1:699	R
A1395_31705	hypothetical protein	494:693	F
A1395_31710	phage tail protein	1:504	R
A1395_31715	bifunctional ADP-dependent (S)-NAD(P)H-hydrate dehydratase/NAD(P)H-hydrate epimerase	1:501	R
A1395_31720	hypothetical protein	1:675	F
A1395_31725	3-oxoadipate enol-lactonase	80:672	F
A1395_31280	hypothetical protein	1:564	F
A1395_31285	hydroxyacid dehydrogenase	569:1819	F
A1395_31290	baseplate protein	1823:2938	F
A1395_31295	hypothetical protein	2935:3441	F

A1395_31300	hypothetical protein	3444:3842	F
A1395_31730	hypothetical protein	1:329	F
nudF	ADP-ribose pyrophosphatase	1:441	R
purH	bifunctional phosphoribosylaminoimidazolecarboxamide formyltransferase/inosine monophosphate cyclohydrolase	1:276	F
A1395_31745	hypothetical protein	476:658	F
A1395_31750	glycosyl transferase family 2	1:280	F
A1395_31755	polysaccharide deacetylase	294:655	F
A1395_31760	sulfate adenylyltransferase	80:655	F
A1395_31765	hypothetical protein	1:193	F
A1395_31770	threonine synthase	299:654	F
A1395_12330	TonB-dependent receptor	34512:36953	R
A1395_12355	hypothetical protein	40912:41250	R
A1395_12455	mannose-1-phosphate guanylyltransferase/mannose-6-phosphate isomerase	73239:73983	F
A1395_12905	hypothetical protein	168536:169027	R
A1395_12910	hypothetical protein	170792:171199	F
A1395_12920	integration host factor subunit beta	173045:173335	R
A1395_13445	MFS transporter	284483:285688	R
A1395_13450	TetR family transcriptional regulator	285781:286326	F
A1395_14880	hypothetical protein	591485:591763	R
A1395_15255	fatty acid desaturase	669195:670061	R
A1395_15710	integrase	770635:771876	F
A1395_15715	hypothetical protein	772123:773061	F
A1395_15720	hypothetical protein	773395:773700	F
A1395_15725	hypothetical protein	774231:774515	F
A1395_15730	hypothetical protein	774512:774781	F
A1395_15735	DNA/RNA helicase, superfamily II protein	774778:777672	F
A1395_15740	chromosome segregation protein SMC	777899:778741	F
A1395_15745	hypothetical protein	778805:778990	F
A1395_15750	transcriptional regulator	779060:779425	F
A1395_15755	hypothetical protein	782247:783194	R
A1395_15760	hypothetical protein	784016:784897	F
A1395_15835	hypothetical protein	801542:801751	R
A1395_15870	ArsR family transcriptional regulator	808925:809590	R
A1395_15875	hypothetical protein	809700:810209	F
A1395_16655	AraC family transcriptional regulator	968805:969740	R

A1395_16660	phenylacetaldoxime dehydratase	970296:971354	F
A1395_16665	phenol degradation protein meta	971446:972312	F
A1395_16670	amidase	972504:974018	F
A1395_16675	nitrile hydratase subunit alpha	974095:974688	F
A1395_16680	nitrile hydratase subunit beta	974730:975392	F
A1395_16685	hypothetical protein	975389:976669	F
A1395_16690	chemotaxis protein	976666:977958	F
A1395_16845	hypothetical protein	1006137:1006355	F
A1395_31305	phage head-tail joining protein	1:245	F
A1395_31310	hypothetical protein	238:759	F
A1395_31315	hypothetical protein	821:1315	F
A1395_31325	hypothetical protein	2061:2651	F
A1395_31775	pyridoxalphosphate dependent aminotransferase	1:654	F
A1395_31785	hypothetical protein	251:651	R
A1395_31330	hypothetical protein	21:611	R
A1395_31335	hypothetical protein	1058:1393	F
A1395_31345	hypothetical protein	2183:2419	R
A1395_31790	4-hydroxy-tetrahydrodipicolinate synthase	3:646	F
A1395_31795	hypothetical protein	1:644	R
A1395_31800	signal recognition particle-docking protein FtsY	1:629	F
A1395_31350	hypothetical protein	1665:2337	F
A1395_31805	hypothetical protein	1:524	R
A1395_31355	phage tail protein	1:1173	F
A1395_31360	phage tail protein	1231:1578	F
A1395_31365	ArsR family transcriptional regulator	1575:1871	F
A1395_31810	gamma-glutamylputrescine oxidoreductase	1:621	R
A1395_31815	hypothetical protein	1:573	R
A1395_31375	phage tail tape measure protein	1:1790	R
A1395_31380	hypothetical protein	1790:1980	R
A1395_31820	aminodeoxychorismate lyase	89:613	F
A1395_31825	carbamoyltransferase	1:611	R
A1395_31835	hypothetical protein	1:605	F
A1395_31385	phage tail tape measure protein	1:1790	R
A1395_31390	hypothetical protein	1790:1980	R
A1395_31395	hypothetical protein	41:241	R
A1395_31400	hypothetical protein	1:1655	F

A1395_17490	addiction module toxin RelE	117211:117525	R
A1395_17495	toxin-antitoxin system protein	117491:117781	R
A1395_17640	cytotoxic translational repressor of toxin-antitoxin stability system	155044:155298	F
A1395_17645	transcriptional regulator	155282:155605	F
A1395_18870	hypothetical protein	415076:416449	F
A1395_19425	hypothetical protein	538177:538509	R
A1395_19430	hypothetical protein	538910:539713	R
A1395_20090	hypothetical protein	681599:682114	F
A1395_20095	hypothetical protein	682303:682815	F
A1395_20100	hypothetical protein	682812:683195	F
A1395_20310	hypothetical protein	720727:720930	F
A1395_31410	hypothetical protein	491:898	R
A1395_31415	hypothetical protein	1:1200	F
A1395_20755	hypothetical protein	23916:24176	F
A1395_20760	hypothetical protein	25221:25973	F
A1395_20765	transcriptional regulator	26092:26304	F
A1395_20770	cobyrinic acid a,c-diamide synthase	26347:27222	F
A1395_20775	hypothetical protein	27206:27454	F
A1395_20780	hypothetical protein	27447:29096	F
A1395_20785	coproporphyrinogen III oxidase	29112:29672	F
A1395_20790	hypothetical protein	29676:30902	F
A1395_20795	conjugal transfer protein	31221:32000	F
A1395_20800	integrase	31997:32524	F
A1395_20805	single-stranded DNA-binding protein	32597:33037	F
A1395_20810	DNA topoisomerase III	33316:35328	F
A1395_20820	hypothetical protein	37698:38255	F
A1395_20825	hypothetical protein	38587:38799	F
A1395_20830	hypothetical protein	38821:39213	F
A1395_20835	hypothetical protein	39387:40073	F
A1395_20840	ABC transporter substrate-binding protein	40154:40891	F
A1395_20845	uridylate kinase	40989:41267	F
A1395_20850	GTPase	41556:42368	F
A1395_20855	hypothetical protein	42493:42846	F
A1395_20860	hypothetical protein	43215:44132	F
A1395_20865	hypothetical protein	44190:44879	F
A1395_20870	conserved plasmid protein	44974:45297	F

A1395_20875	hypothetical protein	45400:45807	F
A1395_20880	hypothetical protein	45824:46084	F
A1395_20885	hypothetical protein	46161:46808	F
A1395_20890	methyltransferase	46873:47982	F
A1395_20895	methyltransferase	48034:48354	F
A1395_20900	hypothetical protein	48445:48750	F
A1395_20905	DEAD/DEAH box helicase	48886:51165	F
A1395_20910	hypothetical protein	51281:51487	F
A1395_20915	integrating conjugative element protein pill, pfgi-1	51635:52234	F
A1395_20920	hypothetical protein	52231:52872	F
A1395_20925	hypothetical protein	52887:53612	F
A1395_20930	lytic transglycosylase	53594:54199	F
A1395_20935	hypothetical protein	54196:54735	F
A1395_20940	conjugal transfer protein TraG	54740:56929	F
A1395_20945	hypothetical protein	56926:57675	F
A1395_20950	RAQPRD family plasmid	57774:58157	F
A1395_20955	hypothetical protein	58154:58387	F
A1395_20960	conjugal transfer protein	58404:58763	F
A1395_20965	hypothetical protein	58775:59173	F
A1395_20970	hypothetical protein	59170:59859	F
A1395_20975	hypothetical protein	59856:60761	F
A1395_20980	conjugal transfer protein	60751:62169	F
A1395_20985	conjugal transfer protein	62150:62590	F
A1395_20990	conjugal transfer protein	62590:65457	F
A1395_20995	disulfide bond formation protein DsbA	65471:66235	F
A1395_21000	DNA repair protein RadC	66411:66905	F
A1395_21005	hypothetical protein	67067:67513	F
A1395_21010	conjugal transfer protein	67510:68457	F
A1395_21015	conjugal transfer protein	68467:69861	F
A1395_21020	hypothetical protein	69858:70214	F
A1395_21025	conjugal transfer protein TraG	70230:71750	F
A1395_21030	hypothetical protein	71762:72127	R
A1395_21035	hypothetical protein	72212:72844	R
A1395_21040	integrase	72857:73483	R
A1395_21045	relaxase	73801:75639	F
A1395_21050	excinuclease ABC subunit A	75720:78371	R

A1395_21055	sodium:proton antiporter	78397:80262	R
A1395_21060	recombination factor protein RarA	80323:81636	R
A1395_21065	NADH dehydrogenase	81688:82995	R
A1395_21070	S-formylglutathione hydrolase	83014:83844	R
A1395_21075	Egg lysin	83903:85237	R
A1395_21080	glyoxalase	85420:85818	R
A1395_21085	S-(hydroxymethyl)glutathione dehydrogenase/class III alcohol dehydrogenase	85862:86971	R
A1395_21090	regulator	87031:87306	R
A1395_21095	nitrilase	87589:88479	R
A1395_21100	GNAT family acetyltransferase	88476:89084	R
A1395_21105	isopropylmalate/homocitrate/citramalate synthase	89086:89493	R
A1395_21110	LysR family transcriptional regulator	89648:90553	F
A1395_21115	excinuclease ABC subunit B	90704:92680	R
A1395_21120	hypothetical protein	93055:93486	R
A1395_21125	hypothetical protein	94342:94596	R
A1395_21130	formaldehyde dehydrogenase, glutathione-independent	95387:96601	F
A1395_21135	D-alanyl-D-alanine endopeptidase	97105:97995	F
A1395_21140	LysR family transcriptional regulator	98431:99084	F
A1395_21145	peroxidase	99203:100171	F
A1395_21150	integrase	100209:102152	R
A1395_21190	hypothetical protein	108981:109181	R
A1395_21330	hypothetical protein	143709:144818	F
A1395_21345	hypothetical protein	152027:152341	F
A1395_21350	hypothetical protein	152313:152576	F
A1395_21555	hypothetical protein	204959:205849	F
A1395_21560	hypothetical protein	205985:206746	F
A1395_21605	hypothetical protein	216743:217222	R
A1395_21645	type IV secretion protein Rhs	236722:241398	F
A1395_21650	hypothetical protein	242512:242862	F
A1395_21720	AraC family transcriptional regulator	260277:261314	R
A1395_21725	hypothetical protein	261575:262525	F
A1395_21730	alanine racemase	262890:264164	R
A1395_21735	FAD-linked oxidoreductase	264181:265575	R
A1395_21740	cytochrome C	265595:266017	R
A1395_21955	hypothetical protein	313262:313555	R
A1395_21965	transcriptional regulator	314840:315220	F

A1395_21970	hypothetical protein	315204:315701	F
A1395_21985	hypothetical protein	317310:318137	R
A1395_21990	hypothetical protein	318137:318427	R
A1395_22015	hypothetical protein	323655:323834	R
A1395_22020	hypothetical protein	323831:324046	R
A1395_22040	hypothetical protein	329641:330837	R
A1395_22045	nucleotide pyrophosphohydrolase	330848:331198	R
A1395_22080	hypothetical protein	334278:334631	F
A1395_22090	hypothetical protein	335135:335443	R
A1395_22100	hypothetical protein	336454:337170	R
A1395_22105	hypothetical protein	337253:337576	R
A1395_22120	terminase	340133:341434	R
A1395_22135	hypothetical protein	342605:343057	F
A1395_22140	hypothetical protein	343127:343471	R
A1395_22145	MFS transporter	343468:343797	R
A1395_22170	hypothetical protein	347093:348025	R
A1395_22180	Rha family transcriptional regulator	348979:349401	R
A1395_22185	repressor	349484:349687	R
A1395_22190	XRE family transcriptional regulator	349804:350592	F
A1395_22195	hypothetical protein	350649:351194	F
A1395_22200	hypothetical protein	351238:351471	F
A1395_22205	hypothetical protein	351540:352436	F
A1395_22210	hypothetical protein	353231:353824	F
A1395_22215	hypothetical protein	353990:354463	F
A1395_22220	hypothetical protein	355263:355694	F
A1395_22255	hypothetical protein	358913:359226	F
A1395_22260	hypothetical protein	360105:361214	F
A1395_22265	hypothetical protein	361698:361928	R
A1395_22270	DNA methyltransferase	362013:364058	F
A1395_22275	hypothetical protein	364071:365123	R
A1395_22285	hypothetical protein	366960:367427	F
A1395_22290	hypothetical protein	367464:367706	F
A1395_22305	hypothetical protein	370262:370966	R
A1395_22310	hypothetical protein	371003:371311	F
A1395_22315	hypothetical protein	371668:371862	F
A1395_22320	hypothetical protein	372076:372927	R

A1395_22325	hypothetical protein	373346:373537	R
A1395_22330	hypothetical protein	373728:374240	R
A1395_22335	hypothetical protein	374590:377007	R
A1395_22340	hypothetical protein	377922:378350	F
A1395_22345	hypothetical protein	378531:379022	F
A1395_22350	hypothetical protein	379057:380211	F
A1395_22355	hypothetical protein	380208:380804	F
A1395_22360	hypothetical protein	380816:381574	F
A1395_22365	hypothetical protein	381587:381994	F
A1395_22370	toll-Interleukin receptor	382350:382931	R
A1395_22375	hypothetical protein	382973:383812	R
A1395_22380	hypothetical protein	384179:384589	F
A1395_22385	hypothetical protein	385017:385727	R
A1395_22390	transposase	386457:386735	F
A1395_22395	transposase	386726:387607	F
A1395_22400	serine/threonine protein phosphatase	387604:388326	R
A1395_22405	Cro/Cl family transcriptional regulator	388655:389089	F
A1395_22410	hypothetical protein	389254:389664	R
A1395_22415	DNA helicase UvrD	389710:391692	R
A1395_22420	ATP-dependent endonuclease	391689:393752	R
A1395_22425	transposase	394336:394968	R
A1395_22430	transposase	394910:395289	F
A1395_22435	transposase	395429:396121	F
A1395_22440	hypothetical protein	398352:399188	R
A1395_22445	hypothetical protein	399420:399605	F
A1395_22450	hypothetical protein	399909:401375	R
A1395_22455	hypothetical protein	401344:402396	R
A1395_22460	metal-chelation protein CHAD	403252:404037	R
A1395_22465	tautomerase	404227:404676	R
A1395_22470	AraC family transcriptional regulator	404773:405732	F
A1395_22475	diguanylate cyclase	406144:407076	F
A1395_22660	hypothetical protein	442275:443483	F
A1395_22665	nucleotide pyrophosphohydrolase	443829:444212	F
A1395_22670	ATP-binding protein	444237:446111	F
A1395_22680	HNH endonuclease	447689:448012	F
A1395_23230	hypothetical protein	563038:563412	F

A1395_23580	hypothetical protein	639622:640011	R
A1395_23660	type IV secretion protein Rhs	658803:660354	F
A1395_31420	hypothetical protein	1:1017	R
A1395_31425	phage tail protein	2:1174	R
A1395_31430	hypothetical protein	1:639	R
A1395_31435	phage tail protein	1:623	R
A1395_31440	hypothetical protein	623:823	R
A1395_31445	hypothetical protein	820:1098	R
A1395_23730	hypothetical protein	15421:15603	R
A1395_23825	hypothetical protein	35919:36125	R
A1395_25190	poly(3-hydroxyalkanoate) granule-associated protein PhaI	356247:356669	R
A1395_25370	prevent-host-death protein	394005:394277	F
A1395_25375	plasmid stabilization protein	394274:394606	F
A1395_25575	cell filamentation protein Fic	440614:441768	F
A1395_25715	amidase	472403:472711	R
A1395_26040	3-ketoacyl-ACP reductase	542830:543183	F
A1395_31460	aminotransferase	1:448	F
A1395_31465	spermidine/putrescine ABC transporter substrate-binding protein	621:1023	F
A1395_31470	hypothetical protein	440:1019	F
A1395_26330	hypothetical protein	53031:53210	R
A1395_26335	diguanylate phosphodiesterase	53351:54526	F
A1395_26730	hypothetical protein	138575:138997	R
A1395_27220	hypothetical protein	246137:246745	F
A1395_27225	hypothetical protein	246742:247392	F
A1395_27325	cell filamentation protein Fic	274924:276087	F
A1395_31475	autotransporter outer membrane beta-barrel domain-containing protein	1:949	F
A1395_31480	isoleucine--tRNA ligase	1:939	R
A1395_31485	nitrate reductase	1:917	F
A1395_31490	transcriptional regulator	1:905	F
A1395_31495	hybrid sensor histidine kinase/response regulator	1:900	F
A1395_31500	TfdA	170:896	R
A1395_27905	hypothetical protein	56315:56497	F
A1395_27915	NAD-dependent deacylase	58451:59227	R
A1395_27920	hypothetical protein	59452:59841	R
A1395_27925	transposase	60441:61322	R

A1395_27930	transposase	61313:61591	R
A1395_27935	hypothetical protein	62187:62468	F
A1395_27940	hypothetical protein	62623:62861	F
A1395_27945	hypothetical protein	62895:63926	R
A1395_27955	hypothetical protein	65471:65728	R
A1395_27960	hypothetical protein	65725:66099	R
A1395_27965	structural protein P5	66096:66539	R
A1395_27975	hypothetical protein	67050:68558	R
A1395_27990	host specificity protein	69790:73524	R
A1395_27995	phage tail protein	73577:74161	R
A1395_28000	hydrolase Nlp/P60	74158:74940	R
A1395_28005	phage tail protein	74943:75644	R
A1395_28010	phage tail protein	75681:76028	R
A1395_28015	phage tail tape measure protein	76028:79423	R
A1395_28020	hypothetical protein	79480:79815	R
A1395_28025	hypothetical protein	79828:80082	R
A1395_28030	hypothetical protein	80112:80450	R
A1395_28035	phage tail protein	80460:81182	R
A1395_28040	hypothetical protein	81235:81603	R
A1395_28045	hypothetical protein	81611:82075	R
A1395_28050	head-tail adaptor protein	82068:82409	R
A1395_28055	hypothetical protein	82409:82885	R
A1395_28060	hypothetical protein	82889:83278	R
A1395_28065	capsid protein	83322:84536	R
A1395_28070	peptidase	84551:85426	R
A1395_28075	portal protein	85440:86837	R
A1395_28080	terminase	86837:88528	R
A1395_28085	hypothetical protein	88531:88749	R
A1395_28090	HNH endonuclease	89093:89431	R
A1395_28095	hypothetical protein	89431:89796	R
A1395_28105	hypothetical protein	90628:91179	R
A1395_28110	antitermination protein Q	91516:91890	R
A1395_28115	hypothetical protein	91887:92174	R
A1395_28120	helicase DnaB	92167:93555	R
A1395_28125	ATP-binding protein	93552:94352	R
A1395_28130	hypothetical protein	94402:95094	R

A1395_28135	hypothetical protein	95091:95318	R
A1395_28140	hypothetical protein	95315:95692	R
A1395_28145	hypothetical protein	95689:96420	R
A1395_28150	hypothetical protein	96417:96677	R
A1395_28155	hypothetical protein	96670:96942	R
A1395_28160	hypothetical protein	96939:97271	R
A1395_28165	XRE family transcriptional regulator	98111:98878	F
A1395_28170	helix-turn-helix transcriptional regulator	99006:99242	F
A1395_28175	carbon storage regulator	99397:99699	F
A1395_28180	hypothetical protein	99686:100051	F
A1395_28185	hypothetical protein	100048:100809	F
A1395_28190	hypothetical protein	100806:101030	F
A1395_28195	hypothetical protein	101117:101329	F
A1395_28200	integrase	101329:102441	F
A1395_28370	type B chloramphenicol O-acetyltransferase	129607:130240	F
A1395_28555	hypothetical protein	162602:162793	F
A1395_28660	RelA/SpoT family protein	186651:187541	R
A1395_28665	hypothetical protein	187541:189664	R
A1395_28670	hypothetical protein	190222:190911	F
A1395_28675	hypothetical protein	191051:191389	F
A1395_28680	hypothetical protein	191663:192586	F
A1395_28685	lysozyme	192649:193164	R
A1395_28695	phage tail protein	193764:194189	R
A1395_28700	hypothetical protein	194617:195462	R
A1395_28705	phage tail protein	195463:196071	R
A1395_28710	baseplate J protein	196059:197099	R
A1395_28715	hypothetical protein	197089:197302	R
A1395_31505	hypothetical protein	562:747	F
A1395_31510	diguanylate cyclase	1:454	F
A1395_31515	hybrid sensor histidine kinase/response regulator	416:881	R
A1395_31520	ribonuclease R	1:876	F
A1395_31525	peptide transporter	1:160	F
dppD	peptide ABC transporter ATP-binding protein	171:870	F
A1395_31535	hypothetical protein	178:378	F
A1395_28720	hypothetical protein	1:214	F
A1395_28725	baseplate J protein	204:1244	F

A1395_28730	phage tail protein	1232:1840	F
A1395_28735	hypothetical protein	1841:2686	F
A1395_28745	lysozyme	3591:4142	F
A1395_28755	hypothetical protein	6618:7547	F
A1395_28780	hypothetical protein	12281:13408	R
A1395_28785	hypothetical protein	13647:14726	R
A1395_28790	transposase	14816:15091	R
A1395_28795	hypothetical protein	15800:16138	R
A1395_28820	glyoxalase	20599:20991	F
A1395_28920	integrase	34843:35840	R
A1395_28930	hypothetical protein	36350:36622	R
A1395_28935	hypothetical protein	36667:36897	F
A1395_28945	hypothetical protein	37367:37681	R
A1395_28950	hypothetical protein	37766:37963	F
A1395_28955	hypothetical protein	40672:40977	R
A1395_28975	GNAT family acetyltransferase	45039:45638	F
A1395_31540	histidine kinase	1:624	R
A1395_31545	DNA-binding response regulator	617:859	R
A1395_31550	ABC transporter ATP-binding protein	1:840	F
A1395_31555	helicase	1:839	R
A1395_31560	AraC family transcriptional regulator	261:835	F
A1395_29380	large adhesive protein	7637:23593	R
A1395_29460	addiction module toxin RelE	44112:44393	R
A1395_29465	antitoxin of toxin-antitoxin stability system	44390:44668	R
A1395_29555	hydroxyacid dehydrogenase	59845:60849	R
A1395_29560	phosphoserine phosphatase	60851:61687	R
A1395_29565	hypothetical protein	62334:62681	R
A1395_29660	cytochrome C	80509:81828	F
A1395_29665	sorbitol dehydrogenase	81825:82337	F
A1395_29670	dehydrogenase	82334:84571	F
A1395_29710	hypothetical protein	90401:90658	F
A1395_31565	hypothetical protein	1:813	F
A1395_31570	SfnB family sulfur acquisition oxidoreductase	1:577	R
A1395_31575	zinc metallopeptidase RseP	1:777	F

Table 6.1.3 – Sequence analysis of gene clusters for the synthesis of antibiotics, exoenzyme, cyclic lipopeptide, siderophores, and toxin, and of Gac/Rsm homologues in *P. protegens* Pf11 and Pf-4 and similarities to those in *P. protegens* strains (Pf-5, PH1b) and other closely related *Pseudomonas* sp. (CMAA1215, NFPP17, Os17).

Gene clusters present only in Pf-4 are: pyoluteorin (*plt*), pyrrolnitrin (*prn*), rhizoxin (*rzx*). Gene clusters present in both are: hydrogen cyanide (*hcn*), 2,4-diacetylphloroglucinol (*phl*), AprX protease (*apr*), *gac*/*rsm* homologues, small regulatory RNAs, pyoverdine (*pvd*), enantio-pyochelin (*pch*), hemophore biosynthesis (*has*), ferric-enterobactin receptor (*pfe*), orfamide A (*ofa*), and FitD toxin (*fit*).

Pf-11 Gene ID (A1395_)	**Pf-4 Gene ID (A1348_)**	**Gene name (Pf5 equiv. PFL ID)**	**Pf-11 Locus**	**Pf-4 Locus**	**Length (a.a.)**	**Pf-11 % a.a. indentity**	**Pf-4 % a.a. identity**
***hcn* gene cluster (for hydrogen cyanide) – present in both**							
10425	23065	*hcnA* (2577)	1: 994378–994695 (–)	6: 391003–391320 (+)	105	98.1 *P. protegens* Pf-5	98 *P. protegens* 97 *P.* sp. Os17, St29
10420	23070	*hcnB*	1: 992972–994381 (–)	6: 391317–392726 (+)	469	91.5 *P. protegens* Pf-5	95 *P.* sp. Os17, St29 91 *P. protegens*
10415	23075	*hcnC* (2579)	1: 991726–992979 (–)	6: 392719–393972 (+)	417	95.7 *P. protegens* Pf-5	99 *P.* sp. Os17, St29 96 *P. protegens*
***plt* gene cluster (for pyoluteorin) – only in Pf-4**							
–	17270	*pltM* (2784)	–	4: 360091–361599 (–)	502	–	99 *P. protegens*
–	17275	*pltR*	–	4: 361596–362627 (–)	343	–	98 *P. protegens*

–	17280	*pltL*	–	4: 363114–363380 (+)	88	–	100 *P. protegens*
–	17285	*pltA*	–	4: 363394–364743 (+)	449	–	100 *P. protegens*
–	17290	*pltB*	–	4: 364776–372152 (+)	2458	–	98 *P. protegens*
–	17295	*pltC*	–	4: 372201–377525 (+)	1774	–	99 *P. protegens*
–	17300	*pltD*	–	4: 377576–379210 (+)	544	–	99 *P. protegens*
–	17305	*pltE*	–	4: 379212–380354 (+)	380	–	99 *P. protegens*
–	17310	*pltF*	–	4: 380351–381844 (+)	497	–	99 *P. protegens*
–	17315	*pltG*	–	4: 381848–382630 (+)	260	–	99 *P. protegens*
–	17320	*pltZ*	–	4: 382636–383307 (–)	223	–	99 *P. protegens*
–	17325	*pltI*	–	4: 383383–384396 (+)	337	–	99 *P. protegens*
–	17330	*pltJ*	–	4: 384393–386162 (+)	589	–	99 *P. protegens*
–	17335	*pltK*	–	4: 386172–387314 (+)	380	–	99 *P. protegens*
–	17340	*pltN*	–	4: 387331–388437 (+)	368	–	99 *P. protegens*
–	17345	*pltO*	–	4: 388449–389945 (+)	498	–	98 *P. protegens*
–	17350	*pltP* (2800)	–	4: 390011–390616 (+)	201	–	99 *P. protegens*
prn **gene cluster (for pyrrolnitrin) – only in Pf-4**							
–	27080	*prnA* (3604)	–	8: 330759–332375 (–)	538	–	96 *P. chlororaphis* PA23, O6, subsp. *aurantiaca*

							JD37, PB-St2; *P. protegens*
–	27075	*prnB*	–	8: 329674–330759 (–)	361	–	93 *P. chlororaphis* 92 *P. protegens*
–	27070	*prnC*	–	8: 327929–329632 (–)	567	–	97 *P. protegens* 96 *P. chlororaphis*
–	27065	*prnD* (3607)	–	8: 326813–327904 (–)	363	–	95 *P. chlororaphis* 94 *P. protegens*
phl gene cluster (for 2,4-diacetylphloroglucinol) – present in both							
18635	10485	*phlH* (5951)	3: 364619–365293 (–)	2: 363678–364352 (–)	224	93.2 *P. protegens* Pf-5	93 *P. protegens* 90 *P.* sp. Os17, St29
18640	10490	*phlG*	3: 365436–366320 (+)	2: 364495–365379 (+)	294	92.5 *P. protegens* Pf-5	96 *P.* sp. Os17, St29 93 *P. protegens*
18645	10495	*phlF*	3: 366373–366975 (–)	2: 365432–366034 (–)	200	96.5 *P. protegens* Pf-5	97 *P. protegens*; *P.* sp. Os17, St29
18650	10500	*phlA*	3: 367438–368526 (+)	2: 366497–367585 (+)	362	93.9 *P. protegens* Pf-5	96 *P.* sp. Os17, St29 94 *P. protegens*
18655	10505	*phlC*	3: 368556–369752 (+)	2: 367615–368811 (+)	398	98.7 *P. protegens* Pf-5	99 *P.* sp. Os17, St29; *P. protegens*
18660	10510	*phlB*	3: 369765–370205 (+)	2: 368824–369264 (+)	146	95.9 *P. protegens* Pf-5	99 *P.* sp. Os17, St29 97 *P. protegens*

18665	10515	*phlD*	3: 370414–371463 (+)	2: 369473–370522 (+)	349	98.6 *P. protegens* Pf-5	99 *P. protegens* 98 *P.* sp. Os17, St29
18670	10520	*phlE* (5958)	3: 371574–372851 (+)	2: 370633–371910 (+)	425	92.5 *P. protegens* Pf-5	92 *P.* sp. Os17, St29; *P. protegens*
			***apr* gene cluster (for AprX protease) – present in both**				
08470	26990	*aprA* (3210)	1: 538429–539877 (–)	8: 308831-310279 (–)	482	96.1 *P. protegens* Pf-5 95 *P. protegens* PH1b 92 *P.* sp. CMAA1215	96 *P. protegens* 93 *P.* sp. Os17, St29
08465	26985	*Inh* (3209)	1: 537952–538335 (–)	8:308354..308737 (–)	128	99 *P. protegens* 98 *P.* sp. Os17, St29, CMAA1215	84 *P. protegens* 96 *P.* sp. Os17, St29
08460	26980	*aprD*	1: 535948–537741 (–)	8: 306344–308137 (–)	597	95.3 *P. protegens* Pf-5 94.0 *P.* sp. Os17 95.0 *P.* sp. NFPP17	95 *P. protegens* 94 *P.* sp. Os17, St29
08455	26975	*aprE*	1: 534617–535951 (–)	8: 305013–306347 (–)	444	97.5 *P. protegens* Pf-5 97.0 *P.* sp. Os17 94.0 *P.* sp. CMAA1215	97 *P. protegens* 96 *P.* sp. Os17
08450	26970	*aprF* (3206)	1: 533253–534614 (–)	8: 303649–305010 (–)	453	94.7 *P. protegens* Pf-5 99.0 *P.* sp. Os17 94.0 *P.* sp. PH1b	98 *P.* sp. Os17, St29 94 *P. protegens*

			Gac/Rsm homologues – present in both				
13645	03275	*gacS* (4451)	2: 326117–328870 (+)	0: 690217–692970 (–)	917	96.9 *P. protegens* Pf-5	97 *P.* sp. Os17, St29; *P. protegens*
21170	25980	*gacA* (3563)	4: 104938–105522 (–)	7: 486282–486866 (+)	194	100.0 *P. protegens* Pf-5	100 *P.* sp. Os17, St29; *P. protegens*
13900	03020	*rsmA* (4504)	2: 377278–377466 (–)	0: 641626–641814 (+)	62	100.0 *P. protegens* Pf-5	100 *P.* sp.
17930	09780	*rsmE* (2095)	3: 220271–220990 (+)	2: 219078–219797 (+)	239	93.3 *P. protegens* Pf-5	96 *P.* sp. Os17, St29 92 *P. protegens*
24025	15270	*retS* (0664)	5: 78482–81268 (+)	3: 607391–610177 (–)	928	96.7 *P. protegens* Pf-5	97 *P.* sp. Os17, St29; *P. protegens*
26950	28385	*ladS* (5426)	6: 187267–189633 (–)	9: 172345–174711 (+)	788	91.1 *P. protegens* Pf-5	93 *P.* sp. Os17, St29 91 *P. protegens*
			Small regulatory RNAs – present in both				
N.A.	N.A.	*rsmZ* (6285)	0: 514076–513951 (–)	1: 506535–506661 (+)	127 nt	99 *P. protegens* 98 *P.* sp. Os17, St29	99 *P. protegens* 98 *P.* sp. Os17, St29
N.A.	N.A.	*rsmY* (6291)	3: 74313–74197 (–)	2: 73788–73906 (+)	118 nt	100 *P.* sp. Os17, St29 99 *P. protegens*	100 *P.* sp. Os17, St29 99 *P. protegens*
N.A.	N.A.	*rsmX* (6289)	10: 33390–33506 (+)	10:86797–86915 (+)	119 nt	98 *P.* sp. Os17, St29 97 *P. protegens*	98 *P.* sp. Os17, St29 97 *P. protegens*
			***pvd* gene cluster (for pyoverdine) – present in both**				
07080	17855	*pvdQ* (2902)	1: 189376–191709 (+)	4: 506592–508925 (+)	777	90.5 *P. protegens* Pf-5	91 *P. protegens*

						90.0 *P. protegens* PH1b 87.0 *P.* sp. CMAA1215	85 *P.* sp. Os17, St29
07085	17860	*fpvR* (2903)	1: 191762–192763 (–)	4: 508978–509979 (–)	333	91.4 *P. protegens* Pf-5 92.0 *P.* sp. CMAA1215	91 *P. protegens* 90 *P.* sp. Os17, St29
30155	29340	*pvdA* (4079)	10: 91156–92493 (+)	10: 26184–27521 (–)	445	88.1 *P. protegens* Pf-5	88 *P. protegens*; *P.* sp. Os17, St29
30150	29345	*fpvI*	10: 90476–90958 (+)	10: 27719–28201 (–)	160	85.5 *P. protegens* Pf-5	85 *P.* sp. Os17, St29 84 *P. protegens*
30145	29350	RND efflux Transporter (4081)	10: 88981–90105 (–)	10: 28524–29696 (+)	390	95.6 *P. protegens* Pf-5	96 *P. protegens*; *P.* sp. Os17, St29
30140	29355	ABC efflux Transporter (4082)	10: 87007–88980 (–)	10: 29697–31670 (+)	657	91.5 *P. protegens* Pf-5	97 *P.* sp. Os17, St29 91 *P. protegens*
30135	29360	RND efflux Transporter (4083)	10: 85608–86999 (–)	10: 31678–33069 (+)	463	77.3 *P. protegens* Pf-5	95 *P.* sp. Os17, St29 76 *P. protegens*
30130	29365	PFL_4084	10: 85191–85490 (–)	10: 33186–33485 (+)	99	50.0 *P. protegens* Pf-5	94 *P.* sp. St29 90 *P.* sp. Os17 48 *P. protegens*

30125	29370	PFL_4085	10: 84725–85162 (–)	10: 33514–33951 (+)	145	61.8 *P. protegens* Pf-5	62 *P. protegens*
30120	29375	*pvdP* (4086)	10: 83068–84672 (+)	10: 34004–35632 (–)	542	59.0 *P. protegens* Pf-5	95 *P.* sp. Os17, St29 59 *P. protegens*
30115	29380	*pvdM*	10: 81521..82870 (–)	10: 35806–37155 (+)	449	73.6 *P. protegens* Pf-5	99 *P.* sp. Os17 95 *P.* sp. St29 72 *P. protegens*
30110	29385	*pvdN*	10: 80202–81488 (–)	10: 37188–38474 (+)	428	69.0 *P. protegens* Pf-5	99 *P.* sp. Os17 91 *P.* sp. St29 68 *P. protegens*
30105	29390	*pvdO*	10: 79264–80157 (–)	10: 38522–39412 (+)	296	65.5 *P. protegens* Pf-5	100 *P.* sp. Os17 76 *P.* sp. St29 66 *P. protegens*
30100	29395	*pvdF*	10: 78212–79231 (–)	10: 39445–40464 (+)	339	30.6 *P. protegens* Pf-5	100 *P.* sp. Os17
30095	29400	*pvdE*	10: 76227–77882 (–)	10: 40789–42444 (+)	551	75.0 *P. protegens* Pf-5	100 *P.* sp. Os17 79 *P.* sp. St29 74 *P. protegens*
30090	29405	*fpvA*	10: 73636–76068 (–)	10: 42552–45035 (+)	810 (Pf-11) 827 (Pf-4)	39.9 *P. protegens* Pf-5	100 *P.* sp. Os17 42 *P.* sp. St29 40 *P. protegens*

30085	29410	*pvdD*	10: 62418–72959 (+)	10: 45701–56242 (–)	3513	52.8 *P. protegens* Pf-5	99 *P.* sp. Os17 53 *P. protegens* 45 *P.* sp. St29
30080	29415	*pvdJ* (4094)	10: 59326–62397 (+)	10: 56263–59334 (–)	1023	35.1 *P. protegens* Pf-5	99 *P.* sp. Os17 37 *P.* sp. St29 35 *P. protegens*
30075		???	10: 58257–59315 (+)		352	28.8 *P. protegens* Pf-5	
30070	29425	*pvdI* (4095)	10: 48880–58188 (+)	10: 60472–69768 (–)	3102 (Pf-11) 3098 (Pf-4)	47.7 *P. protegens* Pf-5	97 *P.* sp. Os17 63 *P.* sp. St29; *P. protegens*
30065	29430	Siderophore-Interacting protein (4096)	10: 47737–48705 (–)	10: 69943–70911 (+)	322	86.0 *P. protegens* Pf-5	91 *P.* sp. Os17, St29 85 *P. protegens*
30060	29435	PFL_4097	10: 46820..47560 (+)	10: 71090–71830 (–)	246	91.4 *P. protegens* Pf-5	98 *P.* sp. St29 97 *P.* sp. Os17 91 *P. protegens*
12240	04660	PFL_4169	2: 17612–18835 (+)	10: 56263–59334 (–)	407	93.2 *P. protegens* Pf-5	99 *P.* sp. Os17 93 *P. protegens* 90 *P.* sp. St29
12245	04655	PFL_4170	2: 18832–19371 (+)	0: 998771–999310 (–)	179	93.8 *P. protegens* Pf-5	99 *P.* sp. Os17

							95 *P. protegens*
12250	04650	PFL_4171	2: 19371–19709 (+)	0: 998433–998771 (–)	112	92.0 *P. protegens* Pf-5	97 *P.* sp. Os17 95 *P. protegens*
12255	04645	PFL_4172	2: 19706–20278 (+)	0: 997864–998436 (–)	190	84.7 *P. protegens* Pf-5	100 *P.* sp. St29 98 *P.* sp. Os17 84 *P. protegens*
12260	04640	PFL_4173	2: 20314–21243 (+)	0: 996899–997828 (–)	309	97.7 *P. protegens* Pf-5	98 *P. protegens* 98 *P.* sp. St29 96 *P.* sp. Os17
12265	04635	PFL_4174	2: 21240–21983 (+)	0: 996159–996902 (–)	247	98.0 *P. protegens* Pf-5	98 *P. protegens* 98 *P.* sp. St29 97 *P.* sp. Os17
12270	04630	PFL_4175	2: 21997–22896 (+)	0: 995246–996145 (–)	299	99.3 *P. protegens* Pf-5	99 *P. protegens* 99 *P.* sp. Os17, St29
12275	04625	PFL_4176	2: 22897–23874 (+)	0: 994262–995245 (–)	325 (Pf-11) 327 (Pf-4)	91.1 *P. protegens* Pf-5	97 *P.* sp. Os17, St29 93 *P. protegens*
12280	04620	PFL_4177	2: 24262–25089 (+)	0: 993202–994029 (–)	275	87.2 *P. protegens* Pf-5	94 *P.* sp. Os17, St29 88 *P. protegens*
12285	04615	PFL_4178	2: 25255–25479 (–)	0: 992415–992639 (+)	74	98.6 *P. protegens* Pf-5	99 *P. protegens*

							99 *P.* sp. Os17, St29
12290	04610	*pvdH* (4179)	2: 25562–26974 (–)	0: 990920–992332 (+)	470	94.9 *P. protegens* Pf-5	97 *P.* sp. Os17, St29 95 *P. protegens*
12360	04555	*pvdL* (4189)	2: 41461–54477 (–)	0: 963956–976972 (+)	4338	95.9 *P. protegens* Pf-5	97 *P.* sp. Os17, St29 95 *P. protegens*
12365	04550	*pvdS*	2: 54852–55400 (+)	0: 963033–963581 (–)	182	100.0 *P. protegens* Pf-5	100 *P. protegens* 99 *P.* sp. Os17, St29
12370	04545	*pvdY* (4191)	2: 55441–55794 (–)	0: 962639–962992 (+)	117	70.6 *P. protegens* Pf-5	70 *P. protegens* 67 *P.* sp. Os17, St29
			***pch* cluster (for enantio-pyochelin) – present in both**				
30475	15840	*pchR* (3497)	11: 53981–54883 (–)	4: 49492–50394 (–)	300	95.3 *P. protegens* Pf-5	97 *P.* sp. Os17, St29 95 *P. protegens*
30480	15845	*pchD*	11: 55259–56926 (+)	4: 50770–52437 (+)	555	90.1 *P. protegens* Pf-5	90 *P. protegens* 88 *P.* sp. Os17, St29
30485	15850	*pchH*	11: 56910–58664 (+)	4: 52421–54175 (+)	584	89.1 *P. protegens* Pf-5	90 *P.* sp. Os17, St29 89 *P. protegens*
30490	15855	*pchI*	11: 58661–60424 (+)	4: 54172–55935 (+)	587	86.2 *P. protegens* Pf-5	87 *P.* sp. Os17, St29 86 *P. protegens*

30495	15860	*pchE*	11: 60417–63887 (+)	4: 55928–59398 (+)	1156	88.4 *P. protegens* Pf-5	88 *P.* sp. Os17 88 *P. protegens* 87 *P.* sp. St29
30500	15865	*pchF*	11: 63884–69304 (+)	4: 59395–64815 (+)	1806	93.9 *P. protegens* Pf-5	94 *P. protegens* 93 *P.* sp. Os17, St29
30505	15870	*pchK*	11: 69316–70416 (+)	4: 64827–65927 (+)	366	84.1 *P. protegens* Pf-5	85 *P. protegens* 84 *P.* sp. Os17, St29
30510	15875	*pchC*	11: 70413–71192 (+)	4: 65924–66703 (+)	259	90.7 *P. protegens* Pf-5	93 *P.* sp. Os17, St29 90 *P. protegens*
30515	15880	*pchB*	11: 71216–71539 (+)	4: 66727–67050 (+)	107	85.0 *P. protegens* Pf-5	85 *P.* sp. Os17, St29 84 *P. protegens*
30520	15885	*pchA* (3488)	11: 71532–72965 (+)	4: 67043–68476 (+)	477	88.8 *P. protegens* Pf-5	89 *P. protegens* 86 *P.* sp. Os17, St29
			has gene cluster (for hemophore biosynthesis) – present in both				
26720	28615	*hasI* (5380)	6: 137489–138010 (–)	9: 223960–224481 (+)	173	95.9 *P. protegens* Pf-5	96 *P. protegens* 95 *P.* sp. Os17, St29
26715	28620	*hasS*	6: 136412–137425 (–)	9: 224545–225558 (+)	337	94.1 *P. protegens* Pf-5	93 *P. protegens* 87 *P.* sp. Os17, St29
26710	28625	*hasR*	6: 133574–136279 (–)	9: 225690–228395 (+)	901	96.2 *P. protegens* Pf-5	95 *P. protegens*

							95 *P.* sp. Os17, St29
26705	28630	*hasA*	6: 132873–133490 (–)	9: 228479–229096 (+)	205	96.6 *P. protegens* Pf-5	97 *P. protegens* 92 *P.* sp. Os17, St29
26700	28635	*hasD*	6: 130870–132654 (–)	9: 229315–231099 (+)	594	97.2 *P. protegens* Pf-5	97 *P. protegens*
26695	28640	*hasE*	6: 129524–130873 (–)	9: 231096–232445 (+)	449	96.9 *P. protegens* Pf-5	96 *P. protegens*
26690	28645	*hasF* (5374)	6: 128190–129527 (–)	9: 232442–233779 (+)	445	95.1 *P. protegens* Pf-5	94 *P. protegens*
pfe gene cluster (for ferric-enterobactin receptor) – present in both							
10085	23430	*pfeR* (2665)	1: 916810–917502 (+)	6: 473816–474508 (–)	230	94.8 *P. protegens* Pf-5 95.0 *P. protegens* PH1b 92.0 *P.* sp. CMAA1215	93 *P.* sp. Os17, St29 92 *P. protegens*
10090	23425	*pfeS*	1: 917502–918839 (+)	6: 472479–473816 (–)	445	94.4 *P. protegens* Pf-5 94.0 *P. protegens* PH1b 95.0 *P. protegens* CHA0	96 *P.* sp. Os17, St29 94 *P. protegens*
10095	23420	*pfeA* (2663)	1: 918943–921183 (+)	6: 470135–472375 (–)	746	96.1 *P. protegens* Pf-5 97.0 *P. protegens* Cab57 97.0 *P.* sp. St29	96 *P. protegens*; *P.* sp. Os17, St29
ofa gene cluster (for orfamide A) – present in both							
27845	18430	*ofaA* (2145)	7: 35837–42217 (–)	5: 35808–42188 (–)	2126	82.1 *P. protegens* Pf-5	82 *P. protegens*

27840	18425	*ofaB*	7: 22420–35535 (–)	5: 22429–35544 (–)	4371	85.1 *P. protegens* Pf-5	85 *P. protegens*
27835	18420	*ofaC* (2147)	7: 7700–22423 (–)	5: 7709–22432 (–)	4907	84.2 *P. protegens* Pf-5	84 *P. protegens*
			***fit* gene cluster (for FitD toxin) – present in both**				
08015	26560	*fitA* (2980)	1: 422145–424286 (–)	8: 199520–201661 (–)	713	96.3 *P. protegens* Pf-5 96.0 *P.* sp. NFPP17 93.0 *P.* sp. Os17	96 *P. protegens* 93 *P.* sp. Os17, St29 91 *P. chlororaphis*
08010	26555	*fitB*	1: 420760–422148 (–)	8: 198135–199523 (–)	462	97.6 *P. protegens* Pf-5 97.0 *P.* sp. NFPP17 94.0 *P.* sp. PH1b	96 *P. protegens* 93 *P.* sp. Os17, St29 91 *P. chlororaphis*
08005	26550	*fitC*	1: 418598–420757 (–)	8: 195973–198132 (–)	719	96.9 *P. protegens* Pf-5 97.0 *P.* sp. NFPP17 92.0 *P.* sp. Cab57	97 *P. protegens* 91 *P. chlororaphis* 90 *P.* sp. Os17, St29
08000	26545	*fitD*	1: 409471–418482 (–)	8: 186846–195857 (–)	3003	94.8 *P. protegens* Pf-5 95.0 *P.* sp. CHA0 84.0 *P.* sp. PH1b	93 *P. protegens* 80 *P. chlororaphis* 80 *P.* sp. Os17, St29
07995	26540	*fitE*	1: 407887–409248 (–)	8: 185262–186767 (–)	453 (Pf-11) 501 (Pf-4)	97.3 *P. protegens* Pf-5 97.0 *P.* sp. NFPP17 94.0 *P.* sp. PH1b	94 *P. protegens* 88 *P. chlororaphis* 86 *P.* sp. Os17, St29
07990	26535	*fitF*	1: 404571–407807 (–)	8: 181945–185181 (–)	1078	88.6 *P. protegens* Pf-5	89 *P. protegens*

						89.0 *P.* sp. NFPP17, Cab57	
07985	26530	*fitG*	1: 403657–404574 (+)	8: 181031–181948 (+)	305	96.4 *P. protegens* Pf-5 91.0 *P.* sp. PH1b 89.0 *P.* sp. GM17	95 *P. protegens* 88 *P.* sp. Os17, St29 87 *P. chlororaphis*
07980	26525	*fitH* (2987)	1: 402656–403636 (+)	8: 180030–181010 (+)	326	90.5 *P. protegens* Pf-5 80.0 *P.* sp. PH1b	90 *P. protegens* 80 *P. chlororaphis* 80 *P.* sp. Os17, St29
			***rzx* gene cluster (for rhizoxin) – only in Pf-4**				
–	26520	PFL_2988	–	8: 179502–179906 (+)	134	–	98 *P. protegens* Pf-5 84 *P.* sp. Os17
–	26515	*rzxB* (2989)	–	8: 158807–178849 (–)	6680	–	98 *P. protegens* Pf-5 79 *P.* sp. Os17
–	26510	*rzxC*	–	8: 143811–158636 (–)	4941	–	98 *P. protegens* Pf-5 81 *P.* sp. Os17
–	26505	*rzxD*	–	8: 131692–143814 (–)	4040	–	98 *P. protegens* Pf-5 80 *P.* sp. Os17
–	26500	*rzxH*	–	8: 130286–131695 (–)	469	–	99 *P. protegens* Pf-5 90 *P.* sp. Os17
–	26495	*rzxE*	–	8: 117720–130220 (–)	4166	–	98 *P. protegens* Pf-5

							80 *P.* sp. Os17
–	26490	*rzxF*	–	8: 110029–117654 (–)	2541	–	98 *P. protegens* Pf-5 78 *P.* sp. Os17
–	26485	*rzxI*	–	8: 109125–109991 (+)	288	–	99 *P. protegens* Pf-5 88 *P.* sp. Os17
–	26480	*rzxG*	–	8: 106937–108964 (–)	675	–	98 *P. protegens* Pf-5 84 *P.* sp. Os17
–	26475	*rzxA* (2997)	–	8: 99945–107012 (–)	2355	–	98 *P. protegens* Pf-5 74 *P.* sp. Os17

6.2 Genomic structural variations during clonal expansion of *Pseudomonas syringae* pv. *actinidiae* biovar 3 in Europe

Table 6.2.1 – Primer used in this work. Expected PCR products were fX1/rX2 (933 bp);fX1/rX4 (686 bp); fX3/rX4 (739 bp).

Primer name	Primer sequence 5′-3′
fX1	TAGCCACGGTTTTCTTTGCT
rX2	GACGTTTTACCCCATGCACT
fX3	TTCACGGCCAAGAACAACTG
rX4	CCGCTGACTCGTCTTCTCTC

Table 6.2.2 – Summary of the differences between the chromosomes of strains CRAFRU 12.29 and CRAFRU 14.08

	Position			
Event	**12.29_left**	**12.29_right**	**14.08_left**	**14.08_right**
ISPsy31 insertion	1.474.713	1.476.386	1.474.713	1.474.714
Chromosomal inversion	1.852.631	5.490.147	5.490.627	1.850.961
ISPsy32 insertion	2.118.457	2.118.458	5.224.801	5.223.546
SNP	3.409.171	/	3.932.833	/
VNTR	4.554.472	4.554.473	2.787.533	2.786.633
SNP	4.604.846	/	2.736.260	/

seq0_leftend	seq0_rightend	seq1_leftend	seq1_rightend
1474711	1476386	1474710	1474715
1852630	–4554477	1850959	1852272
–5488836	–2118462	2786629	2787533

–4554472	–1852633	5223542	5224801
–2118457	5490148	5490625	5490628
6548978	6552959	6549454	6553426
6620009	1474712	6620488	0
1476385	5488837	0	0
5490147	6548979	0	0
6552958	0	0	1850960
0	0	1852271	2786630
0	0	2787532	5223543
0	0	5224800	6549455
0	0	6553425	0

Table 6.2.3 – Summary of major structural differences between the chromosomes of strains CRAFRU 12.29 and ICMP 18884.

	Position			
Event	**12.29_left**	**12.29_right**	**18807_left**	**18808_right**
Reverse transcriptase/maturase insertion	1023375	1025252	1023375	/
Reverse transcriptase/maturase insertion	5708293	/	5715260	5717133
Transposase insertion	3694419	/	3287490	3288700
Chromosomal inversion	3380078	3697838	3284077	3603006
IS631 transposase insertion	6513331	/	6522179	6523356
VNTR	4554392		4459558	4460460
VNTR	4824720		4730787	4730844

Table 6.2.4 – Gene finding and annotation in filtered assemblies of reads not mapping on the CRAFRU 12.29 chromosome.

Strain	Contig	Pos. Start	Pos. End	Strand	Annotation
CRAFRU 14.25	3	1694	1236	–	hypothetical protein
CRAFRU 14.25	3	1951	1691	–	Phage single stranded DNA synthesis
CRAFRU 14.25	3	3543	1948	–	Phage DNA replication protein
CRAFRU 14.25	3	4539	3553	–	Phage minor capsid protein - DNA pilot protein
CRAFRU 14.25	3	5120	4548	–	Phage major spike protein
CRAFRU 14.25	3	5332	5186	–	Phage major capsid protein
CRAFRU 12.54	1	535	275	–	Phage single stranded DNA synthesis
CRAFRU 12.54	1	2127	532	–	Phage DNA replication protein
CRAFRU 12.54	1	3123	2137	–	Phage minor capsid protein - DNA pilot protein
CRAFRU 12.54	1	3704	3132	–	Phage major spike protein
CRAFRU 12.54	31	943	29	–	Phage major capsid protein
CRAFRU 12.54	43	487	642	+	Error-prone, lesion bypass DNA polymerase V (UmuC)
CRAFRU 12.54	46	620	453	–	hypothetical protein
CRAFRU 12.54	48	410	120	–	Lyzozyme M1 (1,4-beta-N-acetylmuramidase)
CRAFRU 12.54	49	441	298	–	hypothetical protein
CRAFRU 12.54	49	440	553	+	hypothetical protein
CRAFRU 12.29	2	31	465	+	Phage major capsid protein
CRAFRU 12.29	2	531	1103	+	Phage major spike protein
CRAFRU 12.29	2	1112	2098	+	Phage minor capsid protein - DNA pilot protein
CRAFRU 12.29	2	2108	3703	+	Phage DNA replication protein
CRAFRU 12.29	2	3700	3960	+	Phage single stranded DNA synthesis
CRAFRU 12.29	2	3957	4415	+	hypothetical protein
CRAFRU 14.21	1	7	654	+	Phage major capsid protein
CRAFRU 14.21	1	720	1292	+	Phage major spike protein
CRAFRU 14.21	1	1301	2287	+	Phage minor capsid protein - DNA pilot protein
CRAFRU 14.21	16	620	453	–	hypothetical protein
CRAFRU 14.21	3	1	579	+	Phage DNA replication protein

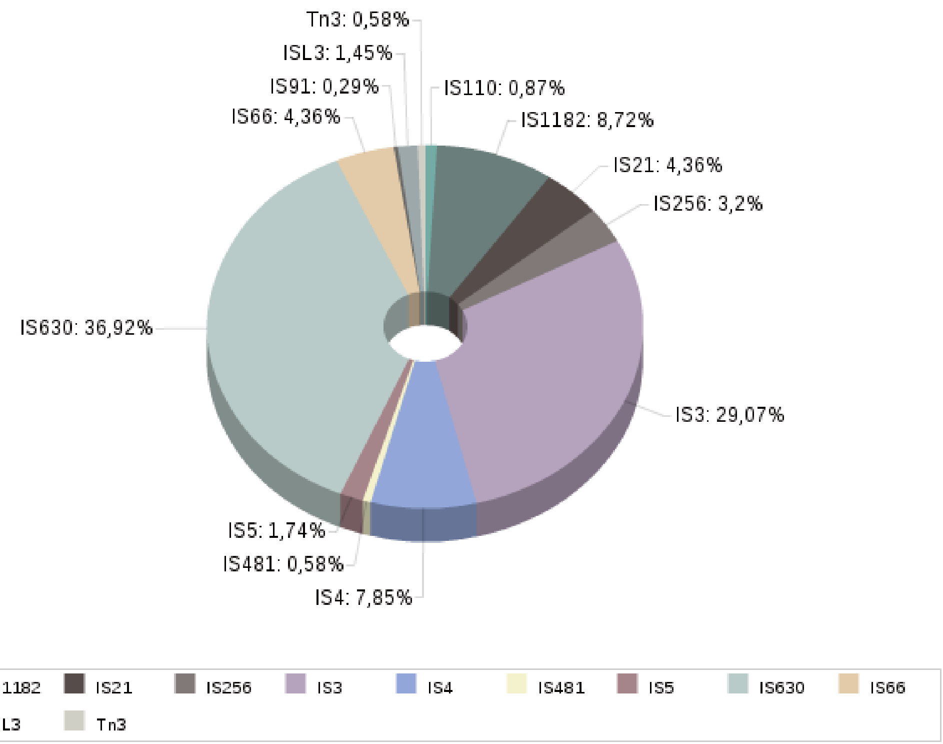
CRAFRU 14.21	3	576	836	+	Phage single stranded DNA synthesis
CRAFRU 14.21	40	192	79	–	hypothetical protein
CRAFRU 14.21	40	191	334	+	hypothetical protein
CRAFRU 14.08	1	535	275	–	Phage single stranded DNA synthesis
CRAFRU 14.08	1	2127	532	–	Phage DNA replication protein
CRAFRU 14.08	1	3123	2137	–	Phage minor capsid protein - DNA pilot protein
CRAFRU 14.08	1	3704	3132	–	Phage major spike protein
CRAFRU 14.08	1	4684	3770	–	Phage major capsid protein
CRAFRU 14.08	14	750	127	–	Lyzozyme M1 (1,4-beta-N-acetylmuramidase)
CRAFRU 14.08	22	169	56	–	hypothetical protein
CRAFRU 14.08	22	168	311	+	hypothetical protein
CRAFRU 14.08	29	620	453	–	hypothetical protein
CRAFRU 14.08	63	191	316	+	hypothetical protein
CRAFRU 14.08	73	168	37	–	hypothetical protein
CRAFRU 13.27	4	1399	941	–	hypothetical protein
CRAFRU 13.27	4	1656	1396	–	Phage single stranded DNA synthesis
CRAFRU 13.27	4	3248	1653	–	Phage DNA replication protein
CRAFRU 13.27	4	4244	3258	–	Phage minor capsid protein - DNA pilot protein
CRAFRU 13.27	4	4825	4253	–	Phage major spike protein
CRAFRU 13.27	4	5358	4891	–	Phage major capsid protein
CRAFRU 14.10	4	905	447	–	Phage external scaffolding protein #Protein D
CRAFRU 14.10	4	1162	902	–	Phage single stranded DNA synthesis
CRAFRU 14.10	4	2754	1159	–	Phage DNA replication protein
CRAFRU 14.10	4	3750	2764	–	Phage minor capsid protein - DNA pilot protein
CRAFRU 14.10	4	4331	3759	–	Phage major spike protein
CRAFRU 14.10	4	5311	4397	–	Phage major capsid protein
CRAFRU 12.50	14	577	317	–	Phage single stranded DNA synthesis
CRAFRU 12.50	14	1152	574	–	Phage DNA replication protein
CRAFRU 12.50	18	1	684	+	Phage minor capsid protein - DNA pilot protein
CRAFRU 12.50	36	60	227	+	hypothetical protein
CRAFRU 12.50	39	320	30	–	Lyzozyme M1 (1,4-beta-N-acetylmurami-

					dase)
CRAFRU 12.50	8	7	654	+	Phage major capsid protein
CRAFRU 12.50	8	720	1292	+	Phage major spike protein
CRAFRU 12.64	1	1749	1066	–	Phage minor capsid protein - DNA pilot protein
CRAFRU 12.64	10	19	549	+	Phage DNA replication protein
CRAFRU 12.64	14	576	433	–	hypothetical protein
CRAFRU 12.64	14	575	688	+	hypothetical protein
CRAFRU 12.64	17	620	453	–	hypothetical protein
CRAFRU 12.64	4	79	651	+	Phage major spike protein
CRAFRU 12.64	43	143	18	–	hypothetical protein
CRAFRU 10.29	1	102	449	+	Phage minor capsid protein - DNA pilot protein
CRAFRU 10.29	10	30	224	+	Phage major spike protein
CRAFRU 10.29	121	198	55	–	hypothetical protein
CRAFRU 10.29	16	535	275	–	Phage single stranded DNA synthesis
CRAFRU 10.29	18	564	397	–	Phage external scaffolding protein #Protein D
CRAFRU 10.29	25	422	132	–	Lyzozyme M1 (1,4-beta-N-acetylmuramidase)
CRAFRU 10.29	6	7	654	+	Phage major capsid protein
CRAFRU 10.29	89	138	269	+	hypothetical protein

Table 6.2.5 – Inventory of the Insertion Sequences (IS) in the chromosome of strain CRAFRU 12.29.

IS Family	IS elements	Copies
IS4 (subgroup IS4)	1	27
ISL3	5	5
IS5 (subgroup IS427)	4	4
IS5 (subgroup IS5)	1	1
IS3 (subgroup IS3)	5	97
IS256	1	11

IS481	2	2
IS66	3	15
IS91	1	1
Tn3	1	2
IS1182	2	30
IS21	1	15
IS5 (subgroup IS1031)	1	1
IS630	4	127
IS3 (subgroup IS51)	2	3
IS110	1	3
Total	**35**	**344**



6.3 An Effective Pipeline Based on Relative Coverage for the Genome Assembly of Phytoplasmas and Other Fastidious Prokaryotes

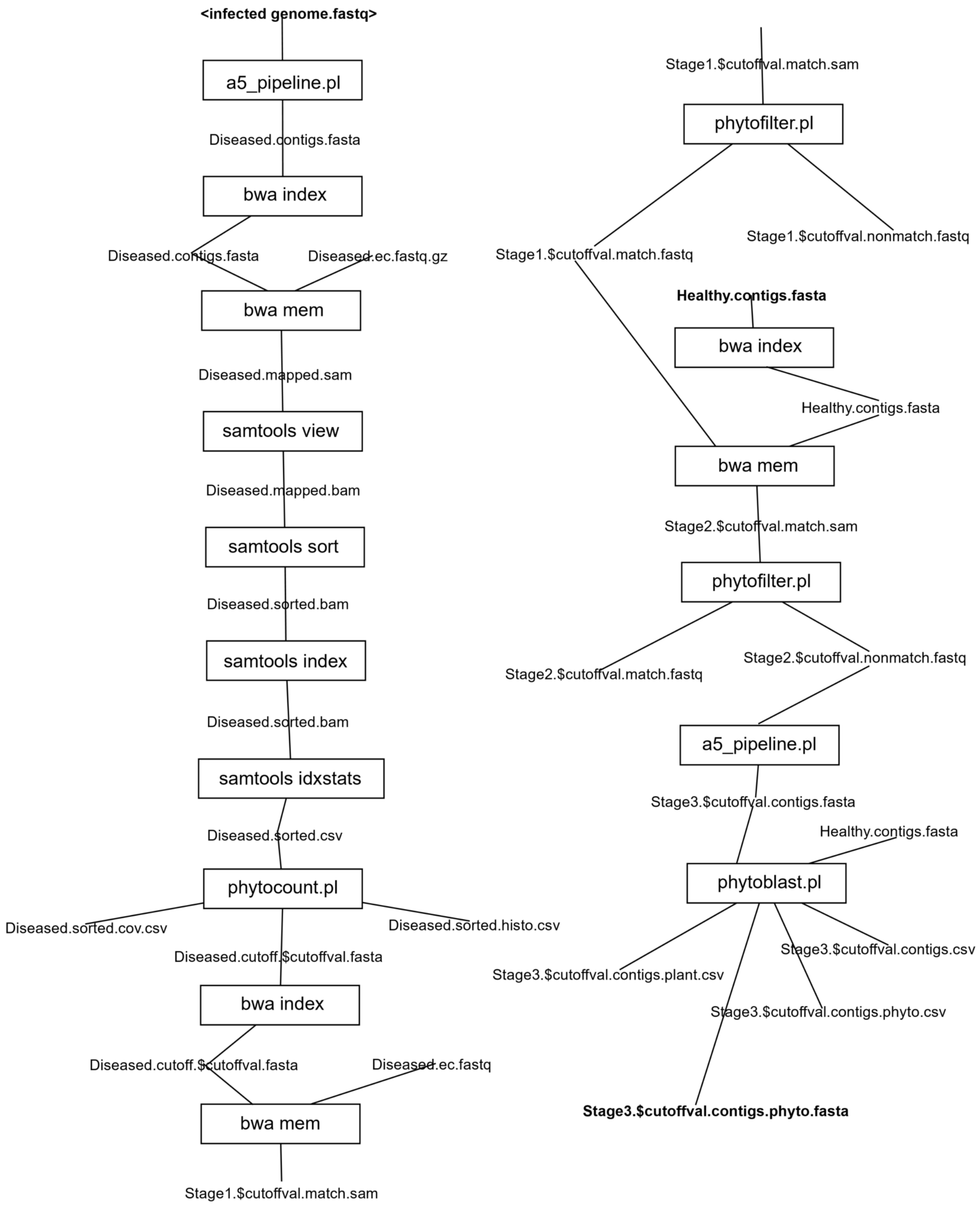


Figure 6.3.1 – Flowchart of the *Phytoassembly* pipeline.