

**Genome analyses of resistance-breaking
isolates of the spinach downy mildew
*Peronospora effusa***

Genoomanalyse van resistentie-doorbrekende isolaten van *Peronospora effusa*,
de valse meeldauw van spinazie

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. H. R. B. M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op
woensdag 18 december 2019 des middags te 2.30 uur

door

Joël Klein
geboren op 31 mei 1988
te Rotterdam

Promotor: Prof. dr. A.F.J.M. van den Ackerveken

Table of contents

| | | |
|-----------|---|-----|
| Chapter 1 | General introduction | 9 |
| Chapter 2 | Genome reconstruction of the non-culturable spinach downy mildew <i>Peronospora effusa</i> | 29 |
| Chapter 3 | Phylogenetic analysis of resistance-breaking isolates of the spinach downy mildew <i>Peronospora effusa</i> | 85 |
| Chapter 4 | <i>Peronospora effusa</i> effector polymorphisms associated with the breaking of resistance loci in spinach | 119 |
| Chapter 5 | Summarizing Discussion | 157 |
| | Summary | 174 |
| | Samenvatting | 176 |
| | Dankwoord | 178 |
| | Curriculum vitae | 183 |
| | List of publications | 184 |





Chapter 1

General introduction

General introduction

1 Oomycetes, also known as “water molds”, are eukaryotic microbial organisms that belong to the Stramenopile taxon [1]. They encompass a wide range of species many of which can actively penetrate their eukaryotic host and cause disease on mammals (e.g. *Pythium insidiosum*), fish (e.g. *Saprolegnia parasitica*) and many plant species (e.g. *Phytophthora infestans*) [2]. Oomycete phytopathogens are often host-specific; meaning that they can only infect a single plant species, while others are broad host pathogens that can infect multiple different host species. The majority of plant pathogenic oomycetes belong to the Peronosporales (e.g. *Peronospora effusa*) and Pythiales (e.g. the damping-off pathogen *Pythium ultimum*), and cause destructive plant diseases in crops, but also in natural ecosystems [3-5]. Although their evolutionary history is largely unknown, fossil evidence indicates that oomycetes first appeared as endophytes of land plants around the Carboniferous period (300-350 million years ago) [6].

Phytopathogenic oomycetes are capable of decimating or even destroying entire crop yields in agriculture. For example, the potato late blight pathogen *Phytophthora infestans* was responsible for the Irish potato famine in the mid-nineteenth century [7]. The genus *Phytophthora* (meaning plant destroyer) includes many well-known species, such as *Phytophthora ramorum* (sudden oak death), *Phytophthora sojae* (stem and root rot of soybean), *Phytophthora capsici* (fruit rot of peppers and other important commercial crops), *Phytophthora palmivora* (bud-rot of palms, fruit-rot or kole-roga of coconut and areca nut), and many more. Other well-known members of the Peronosporales include important downy mildew pathogens of crops e.g. *Bremia lactucae* (lettuce), *Plasmopara viticola* (grape), *Peronospora belbahrii* (basil), *Peronospora effusa* (spinach) and many more.

Phytopathogenic oomycetes have different lifestyles, some are necrotrophic (e.g. *Pythium* species) and actively kill the plant to acquire its nutrients, while others are hemibiotrophic (e.g. *Phytophthora* species) and initially keep the host alive. However, at sufficient colonization levels, they start to secrete hydrolytic enzymes and toxins that kill plant tissues causing the typical blight and rot symptoms. A third lifestyle is obligate biotrophy (e.g. in downy mildews) characterized by an intimate association with living plant tissue and full dependency on the host [8-10].

Oomycete can reproduce both sexually and asexually and are either homothallic or heterothallic. Sexual reproduction starts with the formation of gametangia (female-oogonium and male-antheridium) in which meiosis takes place to produce the gametes. After fertilization the diploid thick-walled oospores are produced that serve as survival structures as they can

remain dormant for many years in the soil and can infect through the root to start a new infection cycle. Asexual spores are produced on leaves and other infected parts of the plant on sporangiophores that are visible with the naked eye. The sporangiospores can be transported by wind or rain and can germinate directly or in certain species further develop to produce motile zoospores [11-13].

Downy mildews

Downy mildews are obligate biotrophic species that generally have a narrow host range. Until now, 20 downy mildew genera consisting of more than 700 different species have been identified and described [12; 14; 15]. Their obligate biotrophic lifestyle is remarkable, as taxonomically downy mildews are related to the hemi-biotrophic *Phytophthora* species, together forming the *Peronosporaceae* clade [12]. Downy mildews have long been considered a sister clade to the *Phytophthora* species based on morphological traits and single gene phylogenies [9; 16]. However, recent studies based on multiple genes and full genome phylogenies suggest that the downy mildew species are not monophyletic but have evolved within different *Phytophthora* clades [17-19].

The majority of downy mildew species are clustered in one clade that includes graminicolous grass-infecting downy mildews (GDM) and two groups that were previously distinguished based on the morphology of their sporangia; the downy mildews with colored conidia (DMCC) and the brassicolous downy mildews (BDM). The *Peronospora* species, e.g. *P. effusa*, belong to the DMCC, whereas the Arabidopsis infecting downy mildew *Hyaloperonospora arabidopsidis* is a BDM species, and *Sclerospora graminicola* is an example of a grass-infecting downy mildew (GDM) [12]. The downy mildews with pyriform haustoria (DMPH), like the species in the genus of *Plasmopara* and *Bremia*, have evolved independently and are actually more closely related to *Phytophthora* clade 1 than to the other downy mildews [17-19].

Downy mildews infect a wide range of agronomically-important crops. These include grapes (*Plasmopara viticola*), cucurbits (*Pseudoperonospora cubensis*), basil (*Peronospora belbahrii*), sunflowers (*Plasmopara halstedii*), lettuce (*Bremia lactucae*), spinach (*Peronospora effusa*), Pearl millet (*Sclerospora graminicola*) and many more [11; 20]. Despite their importance to agriculture, many questions regarding downy mildew biology, phylogeny and their infection mechanisms remain unanswered to date [21; 22].

Structure of oomycete genomes

Most oomycetes are diploid, with varying genome sizes ranging from 18 to 250 Mb [23]. Larger oomycete genomes such as that of *P. infestans*, and to some extent *H. arabidopsidis*, typically have a high repeat content (74%, 42% respectively) [24; 25]. Repeats in these genomes mostly consist of Gypsy and Copia-like elements, but also numerous small and large DNA transposable elements [26]. However, not all oomycetes are characterized by a high repeat content, for example *Phytophthora capsici* (19%) [27], *P. ultimum* (7%) [28] and *Peronospora tabacina* (24%) [29]. These oomycetes species have smaller genome sizes (43 Mb, 64 Mb, and 68 Mb respectively) compared to *P. infestans* and *H. arabidopsidis* (240 Mb and 82 Mb respectively) consistent with the repeat-driven genome expansion model [26].

In the genome of *P. infestans*, certain parts were found to be relatively gene dense and repeat-poor, while other regions were gene-sparse and repeat-rich. Repeat-rich regions likely contribute to changes in gene copy number, family size and sequence [30]. This led to the two-speed genome hypothesis, where the evolutionary adaptations of genes in gene-sparse regions were found to occur at a higher rate than in gene dense parts of the genome. This would allow the proteins encoded in these gene-sparse regions to adapt faster to the host resistance [31].

Oomycete effectors: beating host resistance

To establish a successful infection and use to the host for nutrition, oomycete pathogens must prevent or circumvent plant immune responses. For this purpose, oomycetes deploy an array of secreted virulence molecules (aka effectors) that can manipulate biochemical and physiological processes in their hosts. Best known are protein effectors that have an N-terminal signal peptide (SP) to allow their secretion from the pathogen. We can distinguish apoplastic and host-translocated effectors based on the location of their activity *in planta*. Apoplastic or extracellular effectors are secreted into the apoplastic space between the pathogen and host cell. In the apoplast a diverse range of activities is performed by effectors, e.g. inhibiting host enzymes, degrading host molecules and cell walls, killing plant cells using toxins, etc. [32; 33].

Host-translocated effectors, also known as intracellular or cytoplasmic effectors, also contain a SP but are translocated into host cells. Two major classes of host-translocated effector types have been identified in oomycete genomes, the RxLR and Crinkler (CRN) proteins, respectively, that each have a distinct translocation motif [34; 35]. Many host-translocated effector proteins suppress plant defense responses to enhance disease susceptibility [36]. To be transported into plant cells, a host-translocation

motif is required [37]. *P. infestans* transformants that express a variant of the AVR3a effector, where the RxLR-EER domain was replaced, no longer translocated the effector into plant cells. However, replacement of the translocation domain did not prevent secretion of the effector from haustoria [35], feeding structures that invaginate plant cells for nutrient acquisition and manipulation of the host [38].

Although an SP is sufficient for the secretion into the apoplast, many effectors are host-translocated into host cells [37]. The SP directs proteins to the ER from which proteins travel through the endomembrane system where they are packaged into vesicles that then migrate to the cell membrane where the cargo is released via exocytosis into the apoplast [39]. In a recent study, Wawra *et al.*, observed the cleavage of effector AVR3a directly after the RxLR motif and acetylation of the remaining N-terminus of the mature effector [40]. The processing and modification of the AVR3a effector resembles the cleavage during transport occurring with the so-called PEXEL export system that is found in *Plasmodium* species [41]. In oomycetes the RxLR motif might therefore function as a signal for processing and unconventional protein secretion UPS [42]. This UPS pathway provides an alternative route for SP-containing cargo through a direct transport from the ER to the plasma membrane and thus bypassing the Golgi, also referred to as Golgi-bypass or type IV pathway, as was recently shown for effector protein Pi04314 of *P. infestans* [43; 44].

Effectors can also be recognized by plant resistance (R) proteins encoded by resistance genes (R-genes) [4], resulting in effective defense that is often associated with the hypersensitive response (HR), and referred to as effector-triggered immunity (ETI) [36]. This creates a strong selection pressure on these recognized proteins and drives natural selection to avoid detection by the host or to suppress the host disease response.

Effectors can avoid recognition in the host in several ways [45; 46], for example, through specific amino acids substitutions in the C-terminal effector domains [47]. Besides, recognition can also be avoided when the offending effector is deleted from the genome or when it is not expressed. Alternatively, ETI can be suppressed by new effectors that have evolved through duplication and diversification [48].

Agriculture has accelerated the adaptation of effectors by the repeated introduction of R-genes in crop species, creating a strong selection pressure on the oomycete pathogen to overcome this resistance. This has led to a so-called 'molecular arms-race' [49] which often results in a 'boom-bust' cycle [50; 51]. In these cases, crop varieties are resistant for some time until the introduced resistance is broken, leading to a severe pathogen outbreak. Overall the selection pressures inflicted by R-genes have resulted

in fast-evolving oomycete pathogens, giving rise to new races and isolates that can evade host resistance [12; 52; 53].

Before the advent of genome sequencing and bioinformatics, effectors were identified and studied by biochemical methods or by analysis of individual genes [8; 54]. However, wet-lab research is often tedious, slow and experiments usually depend on prior knowledge. The availability of a genome or transcriptome can be used for many different bioinformatic analyses, and thus aid in the identification of interesting gene candidates for further study [55].

1

Oomycete genome analysis and effector mining

The advancement of low-cost high-throughput DNA sequencing technologies has led to the availability of genome sequences for many important oomycete pathogens, including species such as *Phytophthora sojae* [56] *Phytophthora infestans* [57] and *Hyaloperonospora arabidopsidis* [25]. These oomycete genomes have proven to be a valuable resource to address questions on oomycete biology, lifestyle, and phylogeny [23; 26; 58; 59]. These genome resources have provided rapid and novel insights into the infection mechanisms that are employed by oomycete pathogens.

To study an oomycete pathogen *in silico*, first, the genome or transcriptome needs to be sequenced and assembled (**Box 1**). Gene prediction software is then used to derive a predicted proteome. The existence of genes in the genome can be further supported or discovered using transcriptome sequencing, e.g. by RNA sequencing (RNA-seq), that is also very valuable to improve the accuracy of gene model prediction [60]. In addition, coding sequences can also be obtained from assembled RNA-seq data [61].

Putative de-novo annotations for gene models residing on the genome can be obtained based on homology with genes in other species, and the identification of conserved functional domains. However, annotations for a genome assembly are often not complete as species-specific genes are absent in databases. Despite this, functional domains present in these genes or identified homologs often yields predictions of gene function [62].

As mentioned before, effectors play a major role in the infection of the host by oomycete pathogens and are therefore key to understand the infection mechanism. Oomycete effectors are modular proteins, as opposed to fungal effectors [63], meaning they contain conserved translocation and effector domains that can be identified *in silico* [64]. This allows for the mining and cataloging of effectors present in assembled oomycete genomes. Candidate genes encoding for host-translocated effectors can be identified by a signal peptide for secretion at the N-terminal side, follow by a conserved RxLR-EER translocation motif in RxLR effectors, and LxLFLAK

motif in Crinkler (CRN) effectors. The conserved RxLR or CRN translocation domains can be identified by using regular expressions or an effector specific Hidden Markov Model (HMM) [47]. Some effectors also contain additional C-terminal WY and LWY domains often in tandem repeats [55; 65] that provide structural folds of the effector domains. These folds consist of three or four α -helices with one hydrophobic pocket (WY) [64] or five α -helices with two hydrophobic pockets (LWY) [66].

Although the identification of putative host-translocated oomycete effectors is relatively straightforward, given that most contain conserved motifs in the protein sequence, it does not disclose their exact function in the infection process. To accurately determine the functionality and localization of effector proteins in the host cell often involves wet-lab experiments [34]. However, comparing the gene models to the genes of well-characterized oomycetes could give insights into the functionality of the effectors and other genes in the pathogen under study [67-69]. For example, a comparative functional characterization of two homologous RxLR effector sharing the same functional domains found in *H. arabidopsidis* (*HaRxL96*) and *P. sojae* (*PsAvh163*) showed that they are both capable of suppressing ETI and/or PTI responses in soybean [70].

Besides the comparison of effectors between distinct oomycete species the direct comparisons between effectors of different isolates within the same species can reveal how they have adapted to host immunity [64]. Moreover, effectors that evolve at a faster pace compared to the rest of the genome likely play an important role in the adaptation to host immunity, and their identification can thus reveal core components underlying host-pathogen interaction. Combined with the determination of the phylogenetic relationship between isolates, effector adaptations can reveal how new races have emerged [71-73] or even have adapted to a different host (host jump) [9; 74]. The phylogenetic relationship between different races of the same species is often determined using ITS (internal transcribed spacers), rDNA (ribosomal DNA), mitochondrial genome sequences, or conserved coding sequences in the genome [72]. However, using SNPs identified on whole genome data provides more robust data to determine phylogenetic relationships [75].

Overall, bioinformatic analysis of oomycete genomes provides candidate genes and hypothesis that can be further tested experimentally to reveal the biological functions and activities of genes and proteins. Effector mining, based on conserved motifs and domains, provides predictions, but effector candidates need experimental confirmation to elucidate their localization, target protein in the plant host, and ultimately their function [37; 55; 76]. Overall, bioinformatics can provide the basic groundwork on which lab experiments can be designed. This increases the effectiveness of many

studies on plant pathogens and contributes to finding effective weapons against the 'boom-bust cycle'.

Spinach downy mildew

The downy mildew of spinach (*P. effusa*) is the biggest threat to this popular crop, in particular in California where most of the production of this leafy green occurs [77]. Over the last decades, worldwide spinach production has significantly increased, making it an important agronomical crop. Before 1990, only three races of *Pfs* were known. Now, more than 17 races are denominated for this fast-evolving pathogen.

Pfs mostly spreads through airborne sporangiospores that are primarily formed on the abaxial side of the host leaves. When released sporangiospores land on the surface of a suitable host they germinate, form a germ tubes, and penetrate the leaf by forming appressoria (**Figure 1A**). After entering the host tissue hyphae are formed that grow intercellularly. Similar to many other oomycetes, *Pfs* forms specialized feeding structures, called haustoria, as side branches that cross the plant cell wall and invaginate the plant plasma membrane (**Figure 1B**). Besides a presumed role in nutrient uptake, haustoria also function as specialized organs for the secretion of virulence proteins and effectors to suppress plant defense and enhance susceptibility of the host [78]. After having colonized infected tissues, the pathogen emerges by forming sporangiophores on the leaf when suitable conditions for sporulation (cold temperatures and high humidity) are met (**Figure 1G**). Sporangiophores are branched structures bearing multiple sporangiospores, which are often dispersed by wind currents and splashing (**Figure 1H**) [79].

Besides asexual reproduction through sporangiospores, *Pfs* can also reproduce sexually by forming oospores (**Figure 1D-E**). Oospores serve as survival structures which can remain dormant for long periods in the soil and potentially enter the host through the root to start a new infection cycle (**Figure 1F**) [11]. To effectively combat this pathogen, breeding for resistance to spinach downy mildew is achieved by introgression of different R-genes in cultivated spinach lines. The R-genes are located on RPF loci (Resistance against *Peronospora farinosa*). These race-specific resistance loci to *Pfs* have been identified and obtained from spinach cultivars and wild relatives. Currently, eight R-loci have been identified and characterized in spinach, and are available for the breeding of resistant spinach cultivars [77; 81; 82]. Although monoecious spinach plants can be found, spinach is largely a dioecious crop, this means that there are separate male and female plants [83]. This allows for the breeding of hybrid spinach lines that contain two R-loci and combine unique resistances to the downy mildew pathogen [84]. In a previous study, the analysis of segregation data indicated that resistance was controlled by a single dominant gene, designated

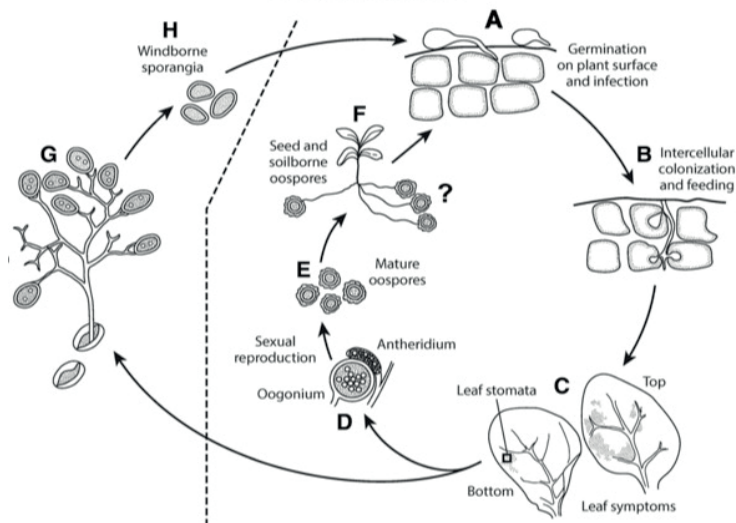


Figure 1. Disease cycle of *P. effusa*. **A.** Germination of spores or oospores and the subsequent penetration of the leaf surface through the formation of appressoria. **B.** Intercellular growth of hyphae and haustoria formation. **C.** Formation of spore-bearing sporangiophores, disease symptoms become visible to the naked eye. **D.** Sexual reproduction through contact between the oogonium with antheridium. **E.** Formation of mature oospores which are located in leaves and seeds. **F.** Infection of the root by germinating oospores. **G.** Sporangiophores formation bearing sporangia (spores); **H.** Sporangia become wind-borne and are disseminated. The sexual and asexual cycles are depicted in dashed squares. Adapted from Kandel *et al.* (2019) [80].

as Pfs-1 [85]. As stated before, the use of R-loci poses a strong selection pressure on the downy mildew pathogen to adapt and overcome the newly bred host resistance. As a consequence, within the last 20 years, more than 16 new races have been identified, and the pathogen is increasingly difficult to control [81]. This underlines the need for elaborate studies on this plant-pathogen interaction, and to bootstrap durable resistance breeding against this downy mildew pathogen.

Box 1.

Sequencing of complex genomes

Next-generation sequencing (NGS), enables a rapid and affordable way to achieve whole-genome de-novo sequencing. Over time most NGS technologies improved significantly in terms of lower cost and higher read output [86]. Although millions of reads are produced with short read technologies e.g. Illumina, the sequenced fragments are relatively small (75~350 bp) and often contain a few miscalled bases. To obtain a genome assembly for analysis, short reads need to be assembled in larger contigs and scaffolds [87]. The major drawback of using short reads for whole-genome assembly is that they often fail to assemble the repetitive areas in a genome. This poses a problem as many oomycete genomes have been shaped by repeat-driven expansions resulting in a repeat-rich genome [88]. Longer sequencing reads can span these repeats, and thus allow for gapless scaffolding leading to more contiguity [89]. Illumina mate-pair sequencing [90] uses a larger insert size between the read pairs which can extend over larger distances and thus potentially span repetitive regions [91]. However, this often leads to gaps in the assembly, where the length of the gap is known but not its nucleotide content. This limitation is largely overcome by using long-read DNA sequencing methods such as PacBio [92]. With this technology, it is possible to generate long reads up to 60-80 kb. However, the error rate in PacBio reads is relatively high (around 11%-15%). These errors are randomly distributed and can be corrected after sequencing by aligning the shorter sequencing reads to longer reads. Recently, Nanopore sequencing was introduced that can also produce long reads at a lower cost than PacBio. This single molecule method uses a relatively small sequencing device that can produce the results in real-time [93]. Furthermore, each sequencing technology has its own biases, for example, the GC bias that is often observed in Illumina sequencing [94]. Including (long) reads from other sequencing technologies can offset the biases in different sequencing technologies [95] and allows for more contiguous genome assemblies.

Thesis outline

This study aims to identify effector sequences in the genome of *Peronospora effusa* (*Pfs*), to study their adaptation in different *Pfs* isolates to overcome resistance in the spinach host and establish the evolutionary relationship between *Pfs* isolates.

Chapter 2 describes the sequencing and genome assembly of *Pfs* race 1. *Pfs* is an obligate biotroph and only grows axenically on living spinach plants. Consequently, the isolated DNA contains many DNA fragments derived from other microbes, which must be removed to obtain a clean genome assembly. For this, a pipeline was used that can taxonomically classify sequences from a pre-assembly constructed from unfiltered reads. Sequences that do not belong to the Stramenopiles were used as a library to filter the sequence reads. Reads that do not map to this library were used to obtain a relatively clean assembly. Genome completeness of the resulting assembly was confirmed by the identification of 191 out of 215 (88.9%) conserved core orthologous genes. Also, RNA-seq reads were employed to guide gene model prediction. These gene models were used to mine for putative effectors, which resulted in the identification of 99 RxLR(-like), and 14 CRN (Crinkler) effectors. Furthermore, we found that a significant number of effectors reside in gene-sparse repeat-rich regions compared to other genes in the genome. This has also been previously observed in *P. infestans*, where these regions are regarded as a cradle for adaptive evolution. In conclusion, the *Pfs* genome is highly gene dense and remarkably small in comparison to other sequenced oomycetes to date.

Chapter 3 describes the phylogenetic relationships between 16 *Pfs* races and 8 additional field isolates. DNA from these isolates was sequenced and aligned to the assembled reference genome of *Pfs1*. These alignments were used to call variants between the reference and the isolates. This provided the input to employ different methods of phylogenetic analysis. The analyses on the genetic structure of the isolates showed three distinct genetic groups and one admixed group consisting of one or more combinations of the aforementioned groups. This is an indication that repeated sexual crosses have occurred and yielded novel *Pfs* races and isolates. Also, a phylogenetic tree was constructed that showed the relationship between the different isolates. Besides, the nuclear genome, also the mitochondrial genome (mt) of *Pfs* race 1 was assembled from the reads. Similar to the whole genome analyses the nucleotide variations between the isolates and the reference mt genome were used to create an mt phylogenetic tree. This tree revealed that there are two distinct mitochondrial haplogroups present in the isolates under study. These mt haplogroups can be found in different branches of the phylogenetic tree based on vari-

ants identified in the whole genome that shows the relationship between the isolates. Combined with the data on the genetic structure of the *Pfs* isolates we conclude that sexual recombination has likely played a pivotal role in the emergence of new resistance breaking *Pfs* races and isolates.

Chapter 4 describes the analyses on *Pfs* effector adaptation to resistant spinach cultivars. Variants between the genomes of the isolates and the reference (that were identified in **Chapter 3**) were annotated to determine the effect on the coding sequences. Also, we identified (effector) genes that are missing in one or more of the isolates. Data on the non-synonymous and synonymous variants in the coding sequences were used to calculate the selection pressure (dN/dS ratio) over the isolates for each gene. The mean dN/dS ratios of the RxLR, RxLR-like and CRN effectors were compared to the other genes in the genome. Here we found that the RxLR effectors had significantly higher dN/dS values compared to other genes in the genome, suggesting that RxLR effectors have been under positive selection. More importantly, we were able to correlate the breaking of specific R-loci to missing effectors and changes in the coding sequences of effectors. In conclusion, this study showed that RxLR effectors undergo more rapid evolution to adapt to the host and revealed interesting changes in effector candidates that are associated with the breaking of R-loci in the spinach host.

In **Chapter 5** the results of the previous chapters are summarized and discussed in a broader perspective.

References

1. Patterson DJ. 1989. *Stramenopiles : chromophytes from a protistan perspective*. Clarendon Press. 357-79 pp.
2. Fry WE, Grünwald NJ. 2010. *Introduction to oomycetes*. <https://www.apsnet.org/edcenter/disandpath/oomycete/introduction/Pages/IntroOomycetes.aspx>
3. Soanes DM, Richards TA, Talbot NJ. 2007. Insights from sequencing fungal and oomycete genomes: what can we learn about plant disease and the evolution of pathogenicity? *The Plant Cell* 19:3318-26
4. Schornack S, Huitema E, Cano LM, Bozkurt TO, Oliva R, et al. 2009. Ten things to know about oomycete effectors. *Mol. Plant Pathol.* 10:795-803
5. van West P, Appiah AA, Gow NA. 2003. Advances in research on oomycete root pathogens. *Physiol. Mol. Plant Pathol.* 62:99-113
6. Krings M, Taylor TN, Dotzler N. 2011. The fossil record of the Peronosporomycetes (Oomycota). *Mycologia* 103:455-7
7. Fry WE, Goodwin SB. 1997. Resurgence of the Irish potato famine fungus. *Bioscience* 47:363-71
8. Singh S, Wilson J, Navi S, Talukdar B, Hess D, Reddy K. 1997. *Screening techniques and sources of resistance to downy mildew and rust in pearl millet*. 108 pp.
9. Thines M. 2014. Phylogeny and evolution of plant pathogenic oomycetes – a global overview. *Eur. J. Plant Pathol.* 138:431-47
10. Restrepo S, Tabima JF, Mideros MF, Grunwald NJ, Matute DR. 2014. Speciation in fungal and oomycete plant pathogens. *Annu. Rev. Phytopathol.* 52:289-316
11. Kamoun S, Furzer O, Jones JD, Judelson HS, Ali GS, et al. 2015. The Top 10 oomycete pathogens in molecular plant pathology. *Mol. Plant Pathol.* 16:413-34
12. Thines M, Choi YJ. 2016. Evolution, diversity, and taxonomy of the Peronosporaceae, with focus on the genus *Peronospora*. *Phytopathology* 106:6-18
13. Judelson HS. 2007. *Sexual reproduction in plant pathogenic oomycetes: biology and impact on disease*. pp 445-458. ASM Press
14. Thines M, Telle S, Ploch S, Runge F. 2009. Identity of the downy mildew pathogens of basil, coleus, and sage with implications for quarantine measures. *Mycol. Res.* 113:532-40
15. Kobayashi M, Hiraka Y, Abe A, Yaegashi H, Natsume S, et al. 2017. Genome analysis of the foxtail millet pathogen *Sclerospora graminicola* reveals the complex effector repertoire of graminicolous downy mildews. *BMC Genomics* 18:897
16. Beakes GW, Glockling SL, Sekimoto S. 2012. The evolutionary phylogeny of the oomycete “fungi”. *Protoplasma* 249:3-19
17. Bourret TB, Choudhury RA, Mehl HK, Blomquist CL, McRoberts N, Rizzo DM. 2018. Multiple origins of downy mildews and mito-nuclear discordance within the paraphyletic genus *Phytophthora*. *PLoS ONE* 13:e0192502

18. Dussert Y, Mazet ID, Couture C, Gouzy J, Piron MC, *et al.* 2019. A high-quality grapevine downy mildew genome assembly reveals rapidly evolving and lineage-specific putative host adaptation genes. *Genome Biol. Evol.* 11:954-69
19. McCarthy CGP, Fitzpatrick DA. 2017. Phylogenomic reconstruction of the oomycete phylogeny derived from 37 genomes. *mSphere* 2:e00095-17
20. Telle S, Shivas RG, Ryley MJ, Thines M. 2011. Molecular phylogenetic analysis of *Peronosclerospora* (Oomycetes) reveals cryptic species and genetically distinct species parasitic to maize. *Eur. J. Plant Pathol.* 130:521-8
21. Thines M, Kamoun S. 2010. Oomycete-plant coevolution: recent advances and future prospects. *Curr. Opin. Plant Biol.* 13:427-33
22. Kemen E, Jones JDG. 2012. Obligate biotroph parasitism: Can we link genomes to lifestyles? *Trends Plant Sci.* 17:448-57
23. Kamoun S. 2003. Molecular genetics of pathogenic oomycetes. *Eukaryot. Cell* 2:191-9
24. Schoina C, Govers F. 2015. *The oomycete Phytophthora infestans, the irish potato famine pathogen.* pp 371-378. Springer. XXIV, 448 pp.
25. Baxter L, Tripathy S, Ishaque N, Boot N, Cabral A, *et al.* 2010. Signatures of adaptation to obligate biotrophy in the *Hyaloperonospora arabidopsidis* genome. *Science* 330:1549-51
26. Judelson HS. 2012. Dynamics and innovations within oomycete genomes: Insights into biology, pathology, and evolution. *Eukaryot. Cell* 11:1304-12
27. Lamour KH, Mudge J, Gobena D, Hurtado-Gonzales OP, Schmutz J, *et al.* 2012. Genome sequencing and mapping reveal loss of heterozygosity as a mechanism for rapid adaptation in the vegetable pathogen *Phytophthora capsici*. *Mol. Plant-Microbe Interact.* 25:1350-60
28. Levesque CA, Brouwer H, Cano L, Hamilton JP, Holt C, *et al.* 2010. Genome sequence of the necrotrophic plant pathogen *Pythium ultimum* reveals original pathogenicity mechanisms and effector repertoire. *Genome Biol.* 11:R73
29. Derevnina L, Chin-Wo-Reyes S, Martin F, Wood K, Froenicke L, *et al.* 2015. Genome sequence and architecture of the tobacco downy mildew pathogen *Peronospora tabacina*. *Mol. Plant-Microbe Interact.* 28:1198-215
30. Dong S, Yu D, Cui L, Qutob D, Tedman-Jones J, *et al.* 2011. Sequence variants of the *Phytophthora sojae* RXLR effector Avr3a/5 are differentially recognized by *Rps3a* and *Rps5* in soybean. *PLoS ONE* 6:e20172
31. Dong S, Raffaele S, Kamoun S. 2015. The two-speed genomes of filamentous pathogens: waltz with plants. *Curr. Opin. Genet. Dev.* 35:57-65
32. McGowan J, Fitzpatrick DA. 2017. Genomic, network, and phylogenetic analysis of the oomycete effector arsenal. *mSphere* 2:e00408-17
33. Kamoun S. 2006. A catalogue of the effector secretome of plant pathogenic oomycetes. *Annu. Rev. Phytopathol.* 44:41-60
34. Kamoun S. 2007. Groovy times: filamentous pathogen effectors revealed. *Curr. Opin. Plant Biol.* 10:358-65

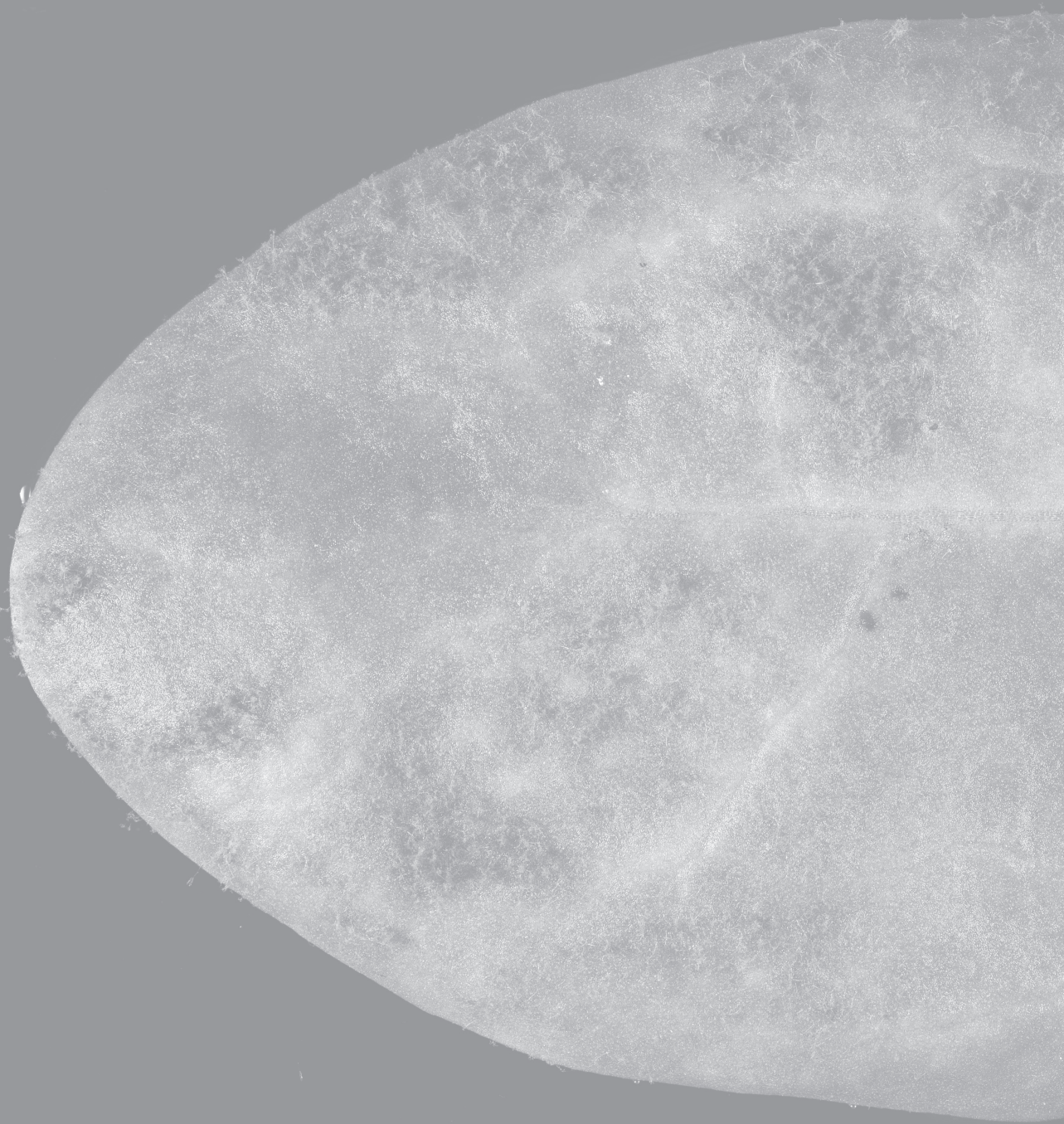
35. Whisson SC, Boevink PC, Moleleki L, Avrova AO, Morales JG, *et al.* 2007. A translocation signal for delivery of oomycete effector proteins into host plant cells. *Nature* 450:115-8
36. Stassen JH, Van den Ackerveken G. 2011. How do oomycete effectors interfere with plant life? *Curr. Opin. Plant Biol.* 14:407-14
37. Birch PR, Rehmany AP, Pritchard L, Kamoun S, Beynon JL. 2006. Trafficking arms: oomycete effectors enter host plant cells. *Trends Microbiol.* 14:8-11
38. Latijnhouwers M, de Wit PJ, Govers F. 2003. Oomycetes and fungi: Similar weaponry to attack plants. *Trends Microbiol.* 11:462-9
39. Birch PR, Boevink PC, Gilroy EM, Hein I, Pritchard L, Whisson SC. 2008. Oomycete RXLR effectors: delivery, functional redundancy and durable disease resistance. *Curr. Opin. Plant Biol.* 11:373-9
40. Wawra S, Trusch F, Matena A, Apostolakis K, Linne U, *et al.* 2017. The RxLR motif of the host targeting effector AVR3a of *Phytophthora infestans* is cleaved before secretion. *Plant Cell* 29:1184-95
41. de Koning-Ward TF, Dixon MW, Tilley L, Gilson PR. 2016. Plasmodium species: master renovators of their host cells. *Nat. Rev. Microbiol.* 14:494
42. Van den Ackerveken G. 2017. Seeing is believing: imaging the delivery of pathogen effectors during plant infection. *New Phytol.* 216:8-10
43. Rabouille C. 2017. Pathways of unconventional protein secretion. *Trends Cell Biol.* 27:230-40
44. Wang S, Boevink PC, Welsh L, Zhang R, Whisson SC, Birch PRJ. 2017. Delivery of cytoplasmic and apoplastic effectors from *Phytophthora infestans* haustoria by distinct secretion pathways. *New Phytol.* 216:205-15
45. Frantzeskakis L, Di Pietro A, Rep M, Schirawski J, Wu CH, Panstruga R. 2019. Rapid evolution in plant-microbe interactions – a molecular genomics perspective. *New Phytol.*
46. Petit-Houdenet Y, Fudal I. 2017. Complex interactions between fungal avirulence genes and their corresponding plant resistance genes and consequences for disease resistance management. *Front. Plant Sci.* 8:1072
47. Win J, Morgan W, Bos J, Krasileva KV, Cano LM, *et al.* 2007. Adaptive evolution has targeted the C-terminal domain of the RXLR effectors of plant pathogenic oomycetes. *Plant Cell* 19:2349-69
48. Dou D, Kale SD, Wang X, Chen Y, Wang Q, *et al.* 2008. Conserved C-terminal motifs required for avirulence and suppression of cell death by *Phytophthora sojae* effector Avr1b. *The Plant Cell* 20:1118-33
49. Anderson JP, Gleason CA, Foley RC, Thrall PH, Burdon JB, Singh KB. 2010. Plants versus pathogens: an evolutionary arms race. *Funct. Plant Biol.* 37:499-512
50. Brown JK, Tellier A. 2011. Plant-parasite coevolution: bridging the gap between genetics and ecology. *Annu. Rev. Phytopathol.* 49:345-67
51. Milgroom MG, Peever TL. 2003. Population biology of plant pathogens: The synthesis of plant disease epidemiology and population genetics. *Plant Dis.* 87:608-17

52. Derevnina L, Petre B, Kellner R, Dagdas YF, Sarowar MN, *et al.* 2016. Emerging oomycete threats to plants and animals. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 371:20150459
53. Raffaele S, Farrer RA, Cano LM, Studholme DJ, MacLean D, *et al.* 2010. Genome evolution following host jumps in the Irish potato famine pathogen lineage. *Science* 330:1540-3
54. St Clair DA. 2010. Quantitative disease resistance and quantitative resistance loci in breeding. *Annu. Rev. Phytopathol.* 48:247-68
55. Ellis JG, Rafiqi M, Gan P, Chakrabarti A, Dodds PN. 2009. Recent progress in discovery and functional analysis of effector proteins of fungal and oomycete plant pathogens. *Curr. Opin. Plant Biol.* 12:399-405
56. Tyler BM. 2007. *Phytophthora sojae*: root rot pathogen of soybean and model oomycete. *Mol. Plant Pathol.* 8:1-8
57. Haas BJ, Kamoun S, Zody MC, Jiang RH, Handsaker RE, *et al.* 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* 461:393-8
58. Lévesque CA. 2011. Fifty years of oomycetes—from consolidation to evolutionary and genomic exploration. *Fungal Divers.* 50:35
59. Fang Y, Cui L, Gu B, Arredondo F, Tyler BM. 2017. Efficient genome editing in the oomycete *Phytophthora sojae* using CRISPR/Cas9. *Curr. Protoc. Microbiol.* 44:21A. 1.1-A.1.6
60. Mathé C, Sagot MF, Schiex T, Rouzé P. 2002. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* 30:4103-17
61. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, *et al.* 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8:1494-512
62. Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* 13:329-42
63. Rafiqi M, Ellis JG, Ludowici VA, Hardham AR, Dodds PN. 2012. Challenges and progress towards understanding the role of effectors in plant-fungal interactions. *Curr. Opin. Plant Biol.* 15:477-82
64. Boutemy LS, King SRF, Win J, Hughes RK, Clarke TA, *et al.* 2011. Structures of *Phytophthora* RXLR effector proteins: A conserved but adaptable fold underpins functional diversity. *J. Biol. Chem.* 286:35834-42
65. He J, Ye W, Choi DS, Wu B, Zhai Y, *et al.* 2019. Structural analysis of *Phytophthora* suppressor of RNA silencing 2 (PSR2) reveals a conserved modular fold contributing to virulence. *Proc. Natl. Acad. Sci. U. S. A.* 116:8054-9
66. Win J, Krasileva KV, Kamoun S, Shirasu K, Staskawicz BJ, Banfield MJ. 2012. Sequence divergent RXLR effectors share a structural fold conserved across plant pathogenic oomycete species. *PLoS Pathog.* 8:e1002400

67. Adhikari BN, Hamilton JP, Zerillo MM, Tisserat N, Levesque CA, Buell CR. 2013. Comparative genomics reveals insight into virulence strategies of plant pathogenic oomycetes. *PLoS ONE* 8:e75072
68. Huang S, van der Vossen EA, Kuang H, Vleeshouwers VG, Zhang N, *et al.* 2005. Comparative genomics enabled the isolation of the *R3a* late blight resistance gene in potato. *Plant J.* 42:251-61
69. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 96:4285-8
70. Anderson RG, Casady MS, Fee RA, Vaughan MM, Deb D, *et al.* 2012. Homologous RXLR effectors from *Hyaloperonospora arabidopsidis* and *Phytophthora sojae* suppress immunity in distantly related plants. *Plant J.* 72:882-93
71. Rouxel M, Mestre P, Comont G, Lehman BL, Schilder A, Delmotte F. 2013. Phylogenetic and experimental evidence for host-specialized cryptic species in a biotrophic oomycete. *New Phytol.* 197:251-63
72. Cooke DE, Drenth A, Duncan JM, Wagels G, Brasier CM. 2000. A molecular phylogeny of *Phytophthora* and related oomycetes. *Fungal Genet. Biol.* 30:17-32
73. Goker M, Voglmayr H, Riethmuller A, Oberwinkler F. 2007. How do obligate parasites evolve? A multi-gene phylogenetic analysis of downy mildews. *Fungal Genet. Biol.* 44:105-22
74. Lamour KH, Stam R, Jupe J, Huitema E. 2012. The oomycete broad-host-range pathogen *Phytophthora capsici*. *Mol. Plant Pathol.* 13:329-37
75. Brocchieri L. 2001. Phylogenetic inferences from molecular sequences: review and critique. *Theor. Popul. Biol.* 59:27-40
76. Deb D, Anderson RG, How-Yew-Kin T, Tyler BM, McDowell JM. 2018. Conserved RxLR effectors from oomycetes *Hyaloperonospora arabidopsidis* and *Phytophthora sojae* suppress PAMP – and Effector-Triggered Immunity in diverse plants. *Mol. Plant-Microbe Interact.* 31:374-85
77. Correll JC, Bluhm BH, Feng C, Lamour K, du Toit LJ, Koike ST. 2011. Spinach: better management of downy mildew and white rust through genomics. *Eur. J. Plant Pathol.* 129:193-205
78. Lo Presti L, Kahmann R. 2017. How filamentous plant pathogen effectors are translocated to host cells. *Curr. Opin. Plant Biol.* 38:19-24
79. Rumbolz J, Wirtz S, Kassemeyer HH, Guggenheim R, Schafer E, Buche C. 2002. Sporulation of *Plasmopara viticola*: Differentiation and light regulation. *Plant Biology* 4:413-22
80. Kandel SL, Mou B, Shishkoff N, Shi A, Subbarao KV, Klosterman SJ. 2019. Spinach downy mildew: advances in our understanding of the disease cycle and prospects for disease management. *Plant Dis.* 103:791-803

81. Feng C, Saito K, Liu B, Manley A, Kammeijer K, *et al.* 2018. New races and novel strains of the spinach downy mildew pathogen *Peronospora effusa*. *Plant Dis.* 102:613-8
82. Kock VLA, Feitsma JG, Frijters RJJM. 2018. *Peronospora* Resistance In *Spinacia oleracea*. Rijk Zwaan Zaadteelt En Zaadhandel Bv. *Patent No. WO 2018/060445 A1*
83. Morelock TE, Correll JC. 2008. *Vegetables I. Handbook of Plant Breeding; Spinach.* pp 189-218. New York, NY: Springer. 430 pp.
84. Barchenger DW, Lamour KH, Bosland PW. 2018. Challenges and strategies for breeding resistance in *Capsicum annuum* to the multifarious pathogen, *Phytophthora capsici*. *Front. Plant Sci.* 9:628
85. Irish BM, Correll JC, Feng C, Bentley T, de Los Reyes BG. 2008. Characterization of a resistance locus (Pfs-1) to the spinach downy mildew pathogen (*Peronospora farinosa f. sp. spinaciae*) and development of a molecular marker linked to Pfs-1. *Phytopathology* 98:894-900
86. Levy SE, Myers RM. 2016. Advancements in Next-Generation Sequencing. *Annu. Rev. Genomics Hum. Genet.* 17:95-115
87. Miller JR, Koren S, Sutton G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics* 95:315-27
88. Raffaele S, Kamoun S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat. Rev. Microbiol.* 10:417-30
89. Koren S, Phillippy AM. 2015. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr. Opin. Microbiol.* 23:110-20
90. Wetzell J, Kingsford C, Pop M. 2011. Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies. *BMC Bioinformatics* 12:95
91. Zhang HB, Wu CC. 2001. BAC as tools for genome sequencing. *Plant Physiol. Bioc.* 39:195-209
92. Rhoads A, Au KF. 2015. PacBio Sequencing and Its Applications. *Genom. Proteom. Bioinf.* 13:278-89
93. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, *et al.* 2010. *The potential and challenges of nanopore sequencing.* pp 261-268. World Scientific. 368 pp.
94. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, *et al.* 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13:341
95. English AC, Richards S, Han Y, Wang M, Vee V, *et al.* 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* 7:e47768





Chapter 2

Genome reconstruction of the non-culturable spinach downy mildew *Peronospora effusa* by metagenome filtering

Joël Klein¹, Manon Neilen¹, Marcel van Verk^{1,2}, Bas E. Dutilh³,
Guido Van den Ackerveken¹

¹ Plant-Microbe Interactions, Department of Biology,
Utrecht University, Utrecht, The Netherlands

² Crop Data Science, KeyGene, Wageningen, The Netherlands

³ Theoretical Biology and Bioinformatics, Department of Biology,
Utrecht University, Utrecht, The Netherlands

A slightly adapted version of this chapter is published as a preprint in BioRxiv.
Available at: www.biorxiv.org/content/10.1101/842658v1

Abstract

Peronospora effusa (previously known as *P. farinosa* f. sp. *spinaciae*, and here referred to as *Pfs*) is an obligate biotrophic oomycete that causes downy mildew on spinach (*Spinacia oleracea*). To combat this destructive disease resistant cultivars are continually bred. However, new *Pfs* races rapidly break the employed resistance genes. To get insight into the gene repertoire of *Pfs* and identify infection-related genes, the genome of the first reference race, *Pfs1*, was sequenced, assembled, and annotated. Due to the obligate biotrophic nature of this pathogen, material for DNA isolation can only be collected from infected spinach leaves that, however, also contain many other microorganisms. The obtained sequences are, therefore, considered a metagenome. To filter and obtain *Pfs* sequences we utilized the CAT tool to taxonomically annotate ORFs residing on long sequences of a genome pre-assembly. This study is the first to show that CAT filtering performs well on eukaryotic contigs. Based on the taxonomy, determined on multiple ORFs, contaminating long sequences and corresponding reads were removed. Filtered reads were re-assembled to provide a clean and improved *Pfs* genome sequence of 32.40 Mbp consisting of 8,635 scaffolds. Transcript sequencing of a range of infection time points aided the prediction of a total of 13,277 gene models, including 99 RxLR(-like) effector, and 14 putative Crinkler genes. Comparative analysis identified common features in the secretomes of different obligate biotrophic oomycetes, regardless of their phylogenetic distance. Their secretomes are generally smaller, compared to hemi-biotrophic and necrotrophic oomycete species. We observe a reduction in proteins involved cell wall degradation, in Nep1-like proteins (NLPs), proteins with PAN/apple domains, and host translocated effectors. The genome of *Pfs1* will be instrumental in studying downy mildew virulence and for understanding the molecular adaptations by which new isolates break spinach resistance.

Introduction

Phytopathogenic oomycetes are eukaryotic microbes that infect a large range of plant species. Due to their hyphal infection structures, they appear fungal-like, however, taxonomically they are not related to fungi but belong to the Stramenopiles [1]. The most devastating phytopathogenic oomycetes are found within the orders of *Albuginales*, *Peronosporales*, and *Pythiales*. The highly radiated *Peronosporales* order contains species with different lifestyles. The most infamous species of this order are in the hemibiotrophic *Phytophthora* genus. Other species within the *Peronosporales* are the obligate biotrophic downy mildews that exploit the plant while keeping host cells alive.

The initial phase of infection is very similar between *Phytophthora* and downy mildew species. Hyphae grow intercellularly through the tissue and locally breach through cell walls to allow the formation of haustoria [2]. These invaginating feeding structures form a platform for the intimate interaction between plant and pathogen cells, and function as a site for the exchange of nutrients, signals, and proteins. Several of the delivered proteins are known to alter host immunity [3] to avoid pathogen detection, and subsequently escape the immune response [4]. These and other molecules secreted by pathogens to promote the establishment and maintenance of an infection in the host are called effectors. Effectors act at two different sites within the host. Apoplastic effectors are secreted into the extracellular space and interact with extracellular plant proteins and cell surface proteins. Host translocated effectors, such as the RxLR and Crinkler (CRN) effectors, are translocated into the host cells and interact with proteins in different compartments of the host cell [5].

The downy mildew *Peronospora effusa* (previously known as *P. farinosa forma specialis spinaciae*, and here referred to as *Pfs*), is the most important pathogen of spinach. *Pfs* is an obligate biotroph that affects the leaves, severely damaging the harvestable parts of the spinach crop. Under favorable environmental conditions, *Pfs* infection can progress rapidly resulting in abundant sporulation within a week post-inoculation that is visible as a thick grey ‘furry layer’ of sporangiophores producing abundant asexual spores [6]. Preventing spread of this pathogen is difficult, since only a few fungicides are effective in chemical control [7]. As a result, the disease can cause severe losses in this popular crop, and infected fields often completely lose their market value. Before 1990, only three races of *Pfs* were known, and the disease could be well controlled. However, as a result of intensified agriculture and, in California year-round production, more than 16 new races have been identified within the last 20 years, and the pathogen is increasingly difficult to control [8]. New *Pfs* races rapidly break resistance

genes that are deployed in newly introduced spinach varieties [8]. It is clear that new resistance sources or alternative methods to protect this crop against this pathogen are needed [9; 10].

Here we describe the sequencing of genomic DNA obtained from spores collected from infected spinach plants using a combination of Illumina and PacBio sequencing. Since the spore washes contained many other microorganisms besides *Pfs*, bioinformatic filtering on taxonomy using CAT [11] was deployed to be able to obtain a clean assembly that is highly enriched for *Pfs* sequences. The obtained assembly of race *Pfs1* was used to predict genes and compare its proteome, in particular, its secretome, with that of other oomycete taxa. We show that the secretomes of obligate biotrophic oomycetes are functionally more similar to each other than to that of more closely related species with a different lifestyle.

2

Materials and Method

Downy mildew infection

Peronospora effusa race 1 (*Pfs1*) was provided to us by the Dutch breeding company Rijk Zwaan Breeding BV in 2014. As *Pfs1* is an obligate biotrophic pathogen maintenance was done on *Spinacia oleracea* Viroflay plants. Seeds were sown on soil, stratified for two days at 4°C and grown under the long-day condition for two weeks (16 h light, 70% humidity, 21°C). sporangiophores were washed off infected plant material in 50 ml falcon tubes. The solution is filtered through miracloth and the spore concentration was checked under the microscope. Four-day-old *Spinacia oleracea* Viroflay plants were infected with *Pfs* by spraying a spore solution (70 spores/μl) in tap water. Seven days post-inoculation, *Pfs* sporangiospores were collected from heavily-infected spinach leaves with tap water, using a soft brush to prevent plant and soil contamination and used for DNA isolation and genome sequencing.

DNA isolation and genome sequencing

The sporangiospores were freeze-dried, ground and dissolved in CTAB (Cetyltrimethyl ammonium bromide) extraction buffer, lysed for 30 minutes at 65°C, followed by a phenol-chloroform/isoamyl-alcohol, and chloroform/isoamyl-alcohol extraction. DNA was precipitated from the aqueous phase with NaOAc and ice-cold isopropanol. The precipitate was collected by centrifugation, and the resulting pellet washed with ice-cold 70% ethanol. DNA was further purified using a QIAGEN Genomic-tip 20/G, following the standard protocol provided by the manufacturer. DNA was quantified using a Qubit HS dsDNA assay (Thermo Fisher Scientific) and

sheared using the Covaris S220 ultrasonicator set to 550 bp. The sequencing library was constructed with the Illumina TruSeq DNA PCR-Free kit. Fragment size distribution in the library was determined before and after the library preparation using the Agilent Bioanalyzer 2100 with HS-DNA chip (Agilent Technologies). The library was sequenced on an Illumina Nextseq machine in high output mode with a 550 bp genomic insert paired-end 150 bp reads at the Utrecht Sequencing Facility (USEQ). Illumina reads with low-quality ends were trimmed ($Q < 36$) using prinseq-lite [12].

For PacBio sequencing, the input DNA was amplified by WGA (Whole Genome Amplification) using the Illustra GenomiPhi V2 DNA Amplification (GE Healthcare). The sequencing library for PacBio was constructed according to the manufacturer protocol. The resulting library was sequenced on 24 SMRT cells (P6 polymerase and C4 chemistry) using the RSII sequencer (Keygene, Wageningen the Netherlands). The obtained PacBio reads were error-corrected using the Falcon pipeline [13] with the standard settings using the SMRT Portal that is part of the SMRT analysis software package version 2.3.0 from PacBio [14]. The analysis software package was installed according to the installation instructions on an Amazon WebService (AWS) cloud-based computer and operated via its build-in GUI.

Taxonomic classification of long reads and contigs

The taxonomic origin of each error-corrected PacBio read was determined using the CAT (Contig Annotation Tool) pipeline [11]. The pipeline identifies open reading frames (ORFs) on long sequences or contigs using Prodigal [15] with default parameters. The protein translations of the predicted ORFs are then queried against the NCBI non-redundant (nr) protein database (retrieved November 2016) using DIAMOND [16] and the taxonomic classification for the resulting hits are derived from the NCBI taxonomy database. Based on the bit-score of the top hit for each ORF a cutoff was set at 90% of this value so that only hits with a bit score above this threshold are maintained. For the high scoring hits of each ORF, a mean bit score is calculated. The maximum achievable bit score (ΣB_{max}) for a long sequence or contig is the summation of mean bit scores of all its ORFs. Then the bit score values of all hits for each ORFs assigned to the same taxon are summed (ΣB_{taxon}), where the taxonomic annotations of each taxon and all its parents in the taxonomic lineage are considered. Based on the maximum achievable bit score (ΣB_{max}) a threshold is set at 50% ΣB_{max} . The highest scoring taxon (ΣB_{taxon}) above the threshold (0.5 times ΣB_{max}) is used to assign the taxonomy for the entire sequence. If ΣB_{taxon} is below the threshold no taxonomy is assigned.

Genome assembly and identification of repeats

A pre-assembly was made using taxonomically filtered and error-corrected PacBio sequences and 60% of the Illumina reads using SPAdes version 3.5.0 [17]. The error-corrected PacBio reads were used as long reads in the assembly, SPAdes was set to use k -mer lengths 21, 33, 55, 77, 99, 127 for the assembly and the `-careful` option was used to minimize the number of mismatches in the final contigs. The contigs derived from the pre-assembly were filtered using the CAT tool (see above), and sequences that were designated as bacterial or non-stramenopile eukaryotes were collected. The entire set of Illumina sequencing reads were aligned to the collection of removed sequences (annotated as bacterial and non-stramenopile) with Bowtie version 2.2.7 using default settings [18]. Illumina reads that aligned to these sequences were removed from the Illumina data set. The remaining Illumina reads (Illumina filtered), and PacBio sequences were re-assembled with SPAdes (same settings as the preassembly), which resulted in a final *Pfs1* genome assembly. A custom repeat library for the *Pfs1* genome assembly was generated with RepeatModeler [19]. Repeat regions in the assembled *Pfs1* genome were predicted using RepeatMasker 4.0.7 [19].

Quality evaluation of the assembly

k -mers of length 21 in the filtered Illumina data set were counted with Jellyfish count version 2.0 [20] with settings `-C -m 21 -s 1000000000` followed by Jellyfish histo. The histogram was plotted with GenomeScope [21] to produce a graphical output and an estimate of the genome size. The coverage of the genome by PacBio sequences was determined by aligning the unfiltered error-corrected PacBio reads to the *Pfs1* genome assembly using BWA-mem [22] and selected `-x pacbio` option. The BBmap pileup [23] script was used to determine the percentage covered bases by PacBio reads in the final assembly of the *Pfs1*.

The GC-content per contig larger than 1kb was calculated using a Perl script [24]. GC density plots were generated in Rstudio version 1.0.143 using GGplot version 3.1 [25]. For comparison, the same analysis was done on a selection of other publicly available oomycete assemblies; *Hyaloperonospora arabidopsidis* [26], *Peronospora belbahrii* [27], *Phytophthora infestans* [28], *Bremia lactucae* [29], *Phytophthora parasitica* [10], *Phytophthora ramorum* (Pr102) [10], *Phytophthora sojae* [30], *Peronospora tabacina* (968-S26) [31] and *Plasmopara viticola* [32].

Kaiju [33] was used to analyze the taxonomic origin by mapping reads to the NCBI nr nucleotide database (November 2017). The input for Kaiju was generated using ART [34] set at 20x coverage with 150 bp Illumi-

na to create artificial sequencing reads from the various FASTA assembly files of the genomes of different oomycetes.

Genome completeness and gene duplications were analyzed with BUSCO version 3 [35] with default settings using the protists Ensembl database (May 2018).

RNA sequencing and gene model prediction

RNA of *Pfs1* at different stages during the infection was isolated and sequenced to aid gene model prediction. Infected leaves and cotyledons were harvested every day from three days post-infection (dpi) until sporulation (7 dpi). Besides these infected leaves, spores were harvested, and a subset of these spores was placed in a petri dish with water and incubated overnight at 16°C to allow them to germinate. RNA was isolated using the RNeasy Plant Mini Kit from Qiagen, and the RNA was analyzed using the Agilent 2100 bioanalyzer to determine the RNA quality and integrity. The RNA-sequencing libraries were made with the Illumina TruSeq Stranded mRNA LT kit. Paired-end 150 bp reads were obtained from the different samples with the Illumina Nextseq 500 machine on high output mode. RNA-seq reads from all the samples were pooled, aligned to the *Pfs1* assembly using Tophat [36], and used as input for gene model prediction using BRAKER1 [37]. The obtained gene models for the *Pfs1* genome together with the RNA-seq alignment result, the repeat models, and results obtained from a BLAST-p search to the nr NCBI database (January 2017) were loaded into a custom made WebApollo [38] instance. Gene models on the 100 largest contigs of the genome were manually curated and all gene models were exported from the WebApollo instance for further use.

Gene annotation and the identification of functional domains

Bedtools intersect version 2.27 was used to determine the overlap between *Pfs1* gene models and annotated repeat elements in the genome. Gene models that had more than 20% overlap with a region marked as a repeat-containing gene. ANNIE [39] was used to annotate proteins on the *Pfs1* genome based on Pfam domains [40] and homologous sequences in the NCBI-Swissprot database (accessed Augustus 2017). Sequences that were annotated as transposons by ANNIE were removed from the gene set. SignalP 4.1 [41] was used to predict the presence and location of a signal peptide, the D-cutoff for noTM and TM networks were set at 0.34 to increase sensitivity [42]. TMHMM version 2 [43] was used to predict the presence of transmembrane helices in the proteins of *Pfs1*. To identify proteins that possess one or more WY domains an HMM model made by Win *et al.* [44] was used. Protein sequences that possessed a WY domain were extracted and realigned. This alignment was used to construct a new

HMM model using HMMER version 3.2.1 [45] and queried again against all protein models in the *Pfs* genome to obtain the full set of WY domains containing proteins.

Effector identification

Putative effectors residing on the genome of *Pfs1* were identified with a custom made pipeline [46] constructed using the Perl [47] scripting language. Secreted proteins were screened for the occurrence of known translocation domains within the first 100 amino acids after the signal peptide. Proteins with a canonical RxLR or a degenerative RxLR (xxLR or RxLx) combined with either an EER-like or a WY domain or both were considered putative RxLR effectors. A degenerative EER domain was allowed to vary from the canonical EER by at most one position.

Proteins with a canonical LFLAK motif or a degenerative LFLAK and HVL motif in the first 100 amino acids of the protein sequence. An HMMer profile was constructed based on the LFLAK or HVL containing proteins. This HMMer profile was used to identify Crinkler effector candidates lacking the LFLAK or HVL motif.

Proteins with an additional transmembrane domain or a C-terminal ER retention signal (H/KDEL) were removed. WY domains were identified using *hmmsearch* version 3.1b2 [48] with the published *Phytophthora* HMM model (see above) [49]. *Pfs* WY-motif containing protein sequences were realigned and used to construct a *Pfs* specific WY HMM model using *hmmbuild* version 3.1b2 [50]. Based on the *Pfs* specific HMM model WY-motif containing *Pfs* proteins were determined.

The effector prediction for the comparative analysis was done in a similar fashion, except for the published *Phytophthora* HMM model for RxLR prediction and a published model for CRN prediction was used. The prediction of effectors using the same model in each species enabled the comparison [51].

Comparative gene distance analysis

Based on the gene locations encoded in the GFF file the 3' and 5' intergenic distances between genes on contigs were calculated as a measure of local gene density. When a gene is located next to the beginning or end of a contig, the distance was taken from the start or end of the gene to the end of the contig. Putative high confidence RxLR effector sequences that encode for proteins with either an exact canonical RxLR motif or an RxLR-like motif in combination with one or more WY-motifs were selected for the comparison (66 in total). Distances were visualized using a heat map constructed with the GGPlot *geom_hex* function [25]. Statistical significance was determined using the Wilcoxon signed-rank test [52].

Comparative secretomics

The predicted proteomes of eighteen plant pathogenic oomycetes were obtained from Ensembl and NCBI (**Supplemental Table 1**). Proteins in the collected proteomes that have a predicted secretion signal [41] (SignalP v.4.1, D-cutoff for SignalP-noTM and TM networks = 0.34 [42]), no additional transmembrane domain (TMHMM 2.0 [43]) or C-terminal K/HDEL domain were considered secreted. Functional annotations of the secreted proteins were predicted using InterProScan [53] and the CAZymes database [54] using the dbCAN2 meta server [55].

Phylogenetic analysis

The phylogenetic relationships between the proteomes of the studied species were inferred using Orthofinder [56]. Orthofinder first identifies 'orthogroups' of proteins that descended from a single ancestral protein. Next, it determines pairwise orthologs between each pair of species. Orthogroups with only one protein of each species were used to make gene trees using MAFFT [57]. The species tree was inferred from the gene trees using the distance algorithms of FastMe [58] and visualized using EvolView v2 [59].

Principal Component Analysis (PCA)

The total number of InterPro and CAZymes domain per species was summarized in a counts table. For each domain, the number was divided by the total number of domains for that species. The normalized matrix has been loaded into Phyloseq version 1.22.3 [60] with R version 3.4.4 [61] in RStudio [62]. A PCA plot has been made with the Phyloseq ordinate function on euclidean distance. The PCA plot has been made with the GGPlot R package [25]. The biplot has been generated with the standard prcomp function in R with the same normalized matrix. Figures were optimized using Adobe Photoshop 2017.01.1.

Permutational analysis of variance (PERMANOVA)

A PERMANOVA using distance matrices was used to statistically test whether there is a difference between the clades based on their CAZymes and InterPro domains. PERMANOVA is a non-parametric method for multivariate analysis of variance using permutations. The data has been double root transformed with the vegdist function from the R-package vegan version 2.5-3 [63]. After the transformation, the PERMANOVA has been calculated with the adonis function from the Vegan package. A total number of 999 permutations have been made to retrieve a representative permutation result.

Enrichment analysis

A chi-square test with Bonferroni correction was used to identify under –and over-represented InterPro domains in each group (*Hyaloperonospora/Peronospora*, *Plasmopara*, *Albugo*) compared to *Phytophthora*. The actual range was the sum of the proteins that have a given domain. The expected range was the fraction of proteins with a given domain that is expected to belong to a species cluster giving the overall ratio of InterPro domains between species clusters.

Results

2 An early isolate of *Peronospora effusa*, race 1 (*Pfs1*), was used as a reference as it predates resistance breeding in spinach and its infection is effectively stopped by all spinach resistance genes known to date. Since downy mildews cannot be grown axenically we isolated asexual sporangiospores by carefully washing highly-infected leaves of the universally susceptible cultivar Viroflay. Genomic DNA was isolated from freeze-dried spores and used to construct libraries for PacBio and Illumina sequencing, resulting in 1.09 million PacBio reads with an N50 of 9,253 bp, and 535 million Illumina reads of 150 bp. The paired-end Illumina reads were used for a trial assembly using Velvet. Inspection of the draft assembly showed that many contigs were of bacterial origin, and not of *Pfs*. This is likely caused by contamination of the isolated *Pfs* spores with other microorganisms that reside on infected leaves and that are collected in the wash-offs. We, therefore, decided to treat the sequences as a metagenome and bioinformatically filter the sequences and corresponding reads.

Taxonomic filtering

To filter out the sequences that could be classified as contaminants we deployed the Contig Annotation Tool (CAT) [11] on long reads and contigs derived from assemblies (outlined in the flow diagram in **Figure 1**). Details on the CAT method are described in the materials and methods section. In short, CAT utilizes the combined taxonomic annotations of multiple individual ORFs found on each sequence to determine its likely taxonomic origin. This allows for a more robust taxon classification that is based on multiple hits, rather than a single best hit. An example of the CAT taxonomic classification for two of our sequences is visualized in **Figure 2**.

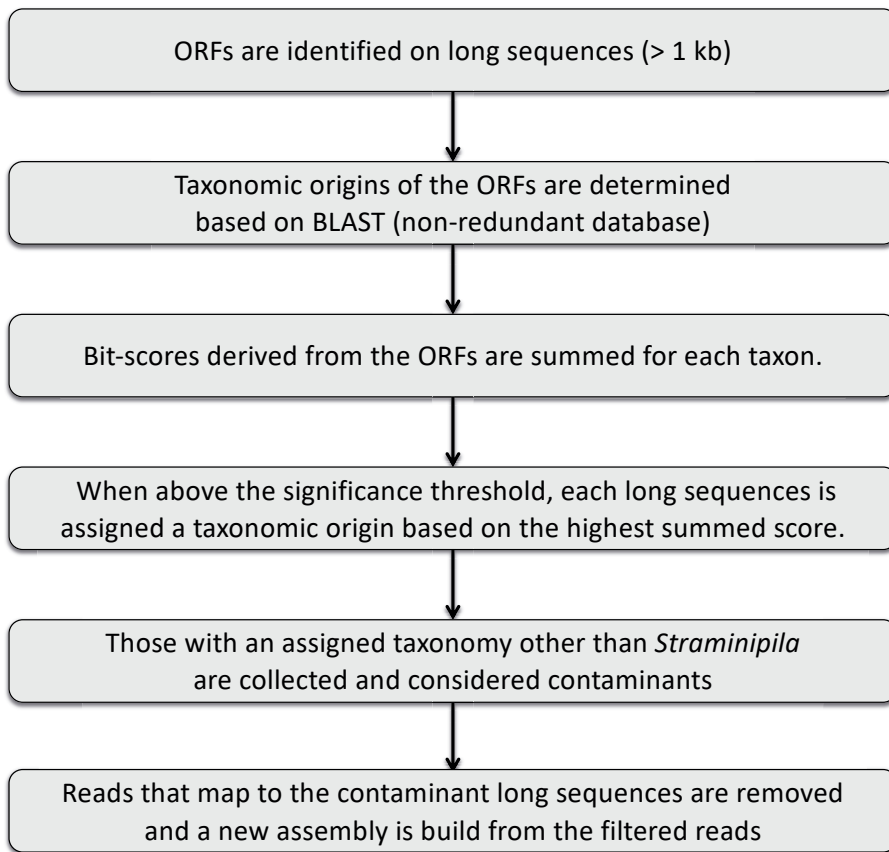
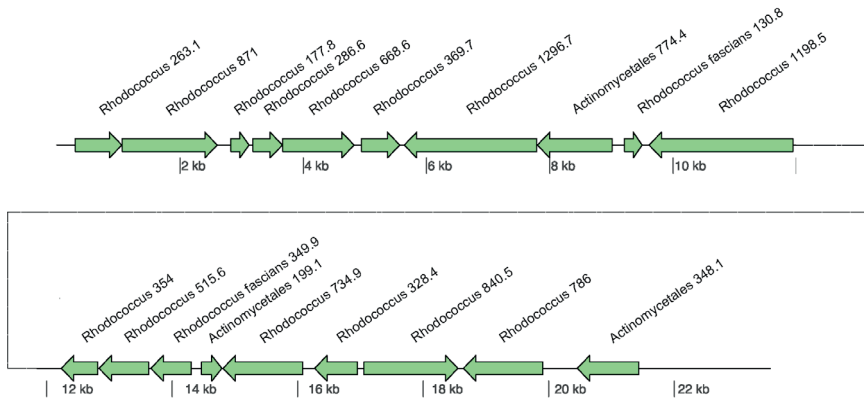


Figure 1. Flow diagram depicting the steps in the taxonomic filtering of long sequences (PacBio reads or assembled contigs).

CAT was first used on the long PacBio reads. As these reads contain on average around 15% base call errors, they were first error-corrected using the FALCON pipeline that fixes long PacBio reads by mapping short reads obtained in the same runs. The resulting 466,225 PacBio reads had a total length of 1,003 Mb with an N50 of 3,325 bp and were subsequently assigned a taxonomic classification using CAT. PacBio reads that were classified as prokaryotic, or non-stramenopile eukaryotic (e.g. Fungi) were removed, whereas reads with the assigned taxonomy stramenopiles or unknown were retained. This resulted in a cleaned set of 232,846 PacBio reads with a total length of 522 Mb with an N50 of 3,458 bp that was used for a hybrid pre-assembly.

A



B

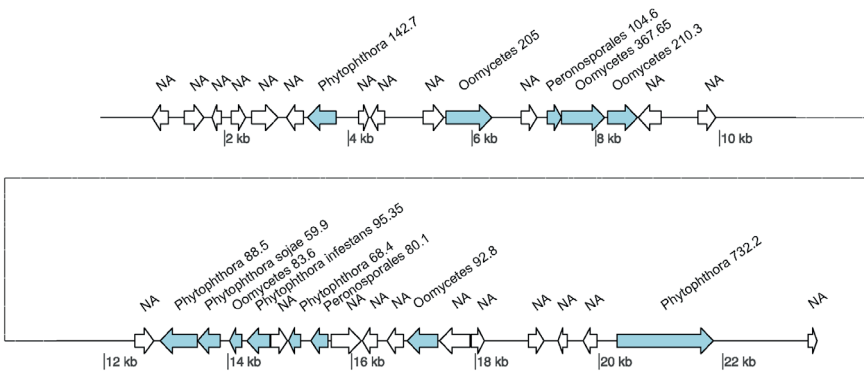


Figure 2. Taxonomic classification by CAT. Two contigs are depicted and per ORF a single top hit is shown. **A.** Contig from the pre-assembly assigned by the CAT tool as bacterial, ORFs of bacterial origin are colored green, and ORF with no hits to the database are colored white. On this contig, most ORFs had the highest blast hit with *Rhodococcus* species. The ΣB_{max} for this contig is 10982, and the highest ΣB_{taxon} is for the *Rhodococcus* genus at 9660, which is well above the cutoff of 5491 ($\Sigma B_{max} * 0.5$). The taxonomic origin of this contig was therefore assigned to the genus *Rhodococcus*, and as a consequence, this contig was regarded as non-*Pfs* and removed. **B.** Contig from the pre-assembly assigned by the CAT tool as an oomycete contig. On this contig, all ORFs have the best hit to an oomycete species, and the ΣB_{max} is 2328. In fact, most ORFs have the best hit to species in the *Phytophthora* genus (ΣB_{taxon} : 1184), or the Peronosporales family (ΣB_{taxon} : 184). The ΣB_{taxon} for the *Phytophthora* genus is above the cutoff at 1164 ($\Sigma B_{max} * 0.5$) thus assigning this contig to the *Phytophthora* genus, and consequently, this contig is maintained for the *Pfs* genome assembly.

In order to evaluate the effectiveness of the CAT tool in removing contaminating genomic sequences, we analyzed the GC-content of the reads. The corrected PacBio reads showed two distinct peaks (**Figure 3A**), whereas oomycete genomes tend to have a rather narrow GC bandwidth around 50%, e.g. as shown in **Supplemental Figure 1** for the contigs of the *Phytophthora infestans* genome [28]. After CAT filtering a single peak remained with a narrow GC-content distribution around ~48%, demonstrating that the tool, that does not take into account GC-content, was effective in removing contaminating sequences (**Figure 3B**).

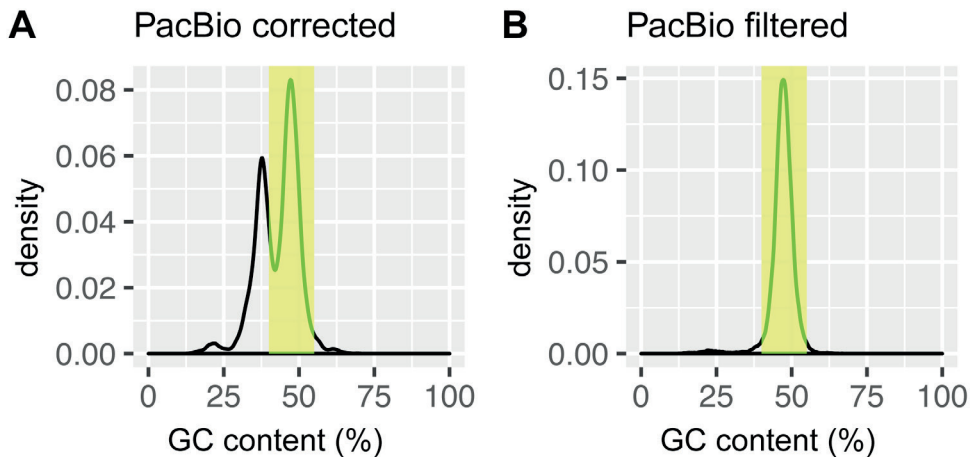


Figure 3. Density plot of the GC values of assembled contigs from the *Pfs1* assembly, and of the PacBio reads before and after CAT filtering sequences. The green bar indicates the region between 40 and 55% GC, based on reads > 1 kb. **A.** PacBio reads before CAT-filtering shows a bimodal distribution with a presumed peak of contaminating sequences with a GC content of ~40%. **B.** PacBio reads after CAT-filtering show a relatively clean distribution with a GC content around ~46%.

Hybrid assembly

A hybrid pre-assembly was generated using the genome assembler SPAdes that can combine long PacBio with short Illumina reads. The input consisted of all corrected and filtered PacBio reads together with 60% randomly extracted Illumina reads (321 million read, 96.3 Gb, to decrease assembly run time and memory requirements). The pre-assembly consisted of 170,143 contigs with a total length of 176 Mb and an N50 of 6,446 bp, of which only 21,690 contigs were larger than 1 kb. CAT filtering was applied to the contigs of the pre-assembly, CAT marked 16,518 contigs consisting of 91.5 Mb (52% of total assembled bases) as contaminant sequences. Next, Illumina reads were aligned to these and Illumina read-pairs of which at least one end aligned were removed from the data set. A final assembly

was generated with the CAT-filtered PacBio and the remaining 77.6 million Illumina reads, resulting in 8,635 scaffolds with a total length of 32.4 Mb. The assembly size corresponds with the estimated genome size of 36.18 Mb that was determined based on *k*-mer count frequency (**Table 1**) in the filtered Illumina reads.

Filtering results

The effect of filtering with CAT on the pre-assembly is well visualized by plotting the GC-content of the contigs, similar as for the PacBio reads. As is visible in **Figure 4A** many contigs with a GC-percentage deviating from the 40-55% range are present in the pre-assembly, indicating that it contains many contaminating sequences. After filtering, the final assembly (H7) shows one peak of the expected GC-content of 50%, although there is a small shoulder remaining of slightly higher GC-content.

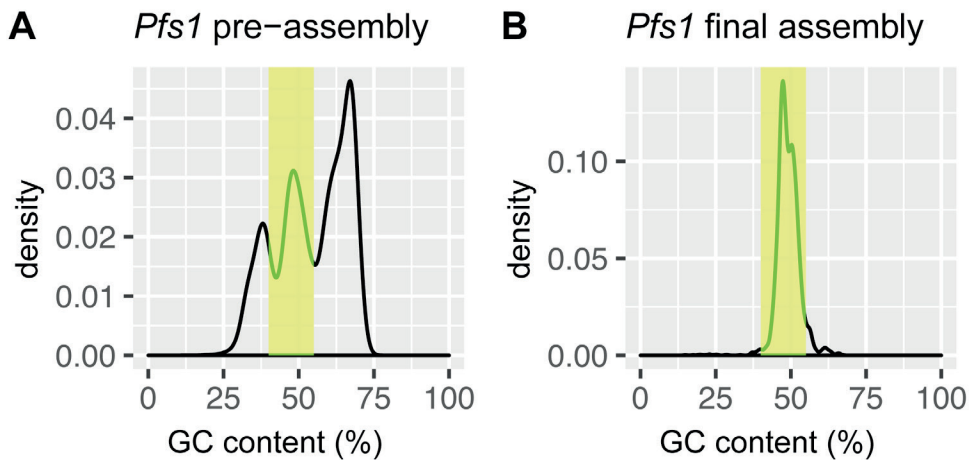


Figure 4. Density plot of the GC values of assembled contigs larger than 1 kb before and after the assembly with filtered reads. The green bar indicates the region between 45 and 49% GC **A.** GC content of the *Pfs1* contigs from the pre-assembly before filtering shows additional peaks at around 30 and 60 GC%, indicating that there are many contaminant contigs. **B.** GC content of the *Pfs1* contigs after filtering of the reads with the CAT tool shows that the additional peaks are no longer present and have thus been successfully filtered out.

In a different approach, Kaiju [33] was used to assess the effectiveness of our taxonomic filtering. Kaiju is typically used for the taxonomic classification of sequencing reads in metagenome analysis but here we used it to determine the effect of taxonomic filtering by CAT. For this, genome assemblies of *Pfs1* and other oomycetes were divided into artificial short reads. The taxonomic distributions generated by Kaiju provide a clear

picture of the removal of contaminating sequences from the *Pfs1* genome data (**Figure 5**). Whereas the pre-assembly mostly contained artificial reads with an assigned bacterial taxonomy, this was reduced to 14% in the final assembly. The percentage of > 80% of oomycete-assigned reads in the *Pfs1* final assembly is similar to what we observe for the high-quality genome assemblies of *P. infestans* and *P. sojae*, pathogens that can be grown axenically, i.e. free of contaminating other microbes (**Figure 5**).

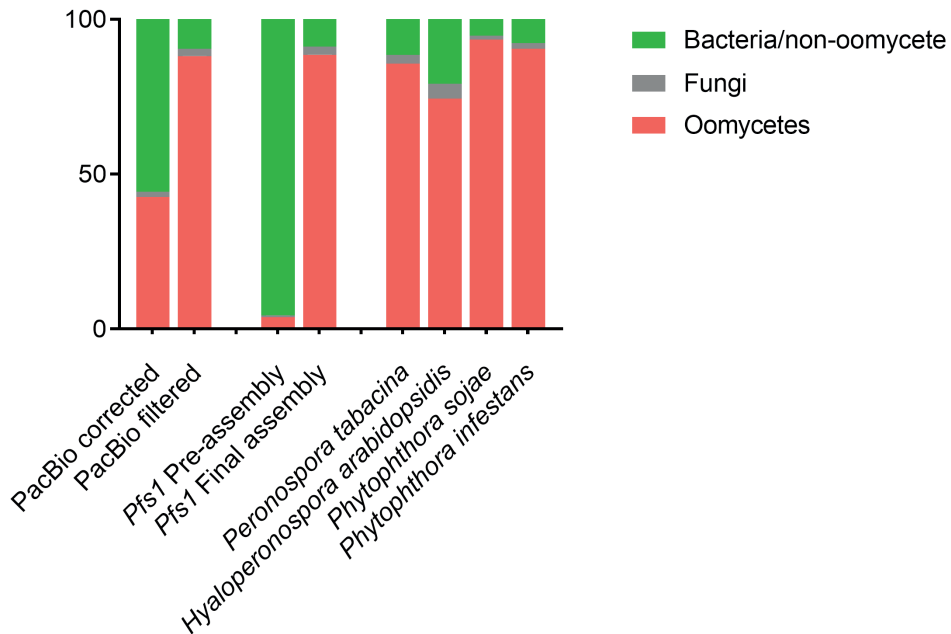


Figure 5. Kaiju bar plot showing the percentage of reads assigned to certain taxonomy and summarized for three classes: Bacteria and other non-oomycetes, Fungi and Oomycetes. In error-corrected PacBio reads, 42.64% are assigned to oomycetes, after filtering with CAT 88.09% of the reads are assigned to oomycetes. For the pre-assembly, only 5% of the reads are assigned to oomycetes. For the *Pfs1* final assembly (H7), 88.6% of the reads are assigned to oomycetes. This is comparable to other oomycetes that can be sterile grown on plates, indicating that the remaining non-oomycetes are most likely a result of an incorrect classification in the database.

Genome statistics

To assess the quality of the assembly we re-aligned the Illumina reads to the contigs and found a large variation in coverage between the contigs smaller than 1 kb compared to larger contigs, suggesting that these small contigs contain a high number of repeats or assembly errors. In addition, the CAT filtering pipeline is only reliable on relatively large contigs where enough ORFs can be called to determine taxonomy in an accurate

way. Therefore, several small contigs could possibly be derived from other microbes besides *Pfs*. Removing contigs smaller than 1 kb (5027 contigs) resulted in a small reduction of 1.9 Mb in genome length, slightly reducing the assembly size to 30.5 Mb, but resulting in a 58% reduction in the number of contigs. The remaining 3608 contigs, larger than 1 kb, had an N50 of 36,273 bp. The statistics of the filtered assembly are further detailed in **Table 1**.

Table 1. Summary of statistics for the hybrid assembly of the *Pfs1* genome. Data is provided for the final assembly (*Pfs1* final) and size-filtered assembly omitting the scaffold smaller than 1 kb (*Pfs1* filtered). In addition, genome information based on k-mer counting of the Illumina reads is providing an estimate for the predicted genome size and repeat content.

| | <i>Pfs1</i> final | <i>Pfs1</i> filtered | <i>k-mer</i> estimation | |
|------------------|-------------------|----------------------|-------------------------|-----------|
| Assembly size | 32.40 Mbp | 30.48 Mbp | Assembly size | 36.18 Mbp |
| GC content | 47.75% | 47.80% | Repeat size | 8.76 Mbp |
| Longest scaffold | 310.10 kbp | 310.10 kbp | Read Error Rate | 1.04% |
| Repeat size | 6.93 Mbp | 6.38 Mbp | Model Fit | 94.01% |
| # Contigs | 8,635 | 3,608 | | |
| N50 | 32,837 bp | 36,273 bp | | |
| # Gene models | 13,227 | 9,007 | | |

To assess the gene space completeness of our assembly in comparison to other oomycete genomes we used BUSCO which identifies single-core orthologs that are conserved in a certain lineage. Here, we used the protist Ensembl database as the protist lineage encompasses the oomycetes and other Stramenipila. According to the BUSCO analysis, the gene space in our assembly is 88.9% complete with only 0.5% fragmented genes and 0.5% duplicates. This gene space completeness score is similar to that of other downy mildew genomes but slightly lower than of genomes of *Phytophthora* species (**Supplemental Table 2**). Furthermore, the low number of duplicates suggests that there is a low incidence of erroneous assembly of haplotypes, therefore we concluded that the obtained *Pfs* assembly represents most of the single-copy gene portion of the *Pfs* genome [35].

Repeat content

In addition to a genome size estimate, the *k-mer* analysis estimated a repeat content of ~8.8 Mb. This is slightly higher than the observed repeat content in the final assembly of ~6.9 Mb (~6.4 Mb in the filtered assembly) (**Table 1**). The difference between the estimated repeat size and the repeat content in the assembly (1.87 Mb) is most likely caused by long repetitive

elements that are hard to assemble. Repeatmasker [19] identified a total of 13,089 of which most are part of the Gypsy and Copia superfamily. We also identified 562 LINEs (Long interspersed nuclear elements) and only 16 SINE (short interspersed nuclear elements), these repeat types belong to the class I transposon (retrotransposons). Other repeat elements consisted of 2297 simple repeats, 298 Low complexity regions, 391 different types of DNA transposons (**Table 2**), and several (278) other minor repeat types; full details can be found in **Supplemental Table 3**.

Table 2. The total number of major repeat types identified in the *Pfs1* genome assembly.

| Repeat type | Count | Percentage | Length (bp) |
|-------------------|-------|------------|-------------|
| LTR | 9,247 | 70.65% | 6,532,069 |
| LINE | 562 | 4.29% | 201,127 |
| Simple repeat | 2,297 | 17.55% | 97,983 |
| DNA repeats/TE | 391 | 2.99% | 46,677 |
| Rolling Circle TE | 97 | 0.74% | 26,123 |

When we compare the genome assembly size of *Pfs* (30.5 Mb) to other sequenced oomycete genomes such as those of *P. infestans* (240 Mb), *H. arabidopsidis* (100 Mb), *Pl. halstedii* (75.3 Mb) or the relatively small genome of *P. tabacina* (63.1 Mb), *Pfs* has a strikingly compact genome (**Supplemental Table 4**). The repeat content (21%) is also low compared to that of other oomycetes, e.g. *P. infestans* (74%), *H. arabidopsidis* (43%), *Pl. halstedii* (39% Mbp) and more comparable to *P. tabacina* (24%).

***Pfs* gene prediction**

Gene prediction is greatly aided by transcript sequence information. We, therefore, isolated and sequenced mRNA from *Pfs* spores and *Pfs*-infected spinach leaves at several time points during the infection. For this, leaves were harvested daily starting from 3 days post-inoculation (dpi) until 7 dpi when sporulation was observed. In addition, mRNA was also isolated from sporangiospores and germlings grown from spores that were incubated in water overnight. The 7 different samples would ensure a broad sampling of transcripts that will facilitate gene identification. Illumina transcript sequences (659 million) were aligned to the assembled *Pfs* genome which resulted in 100 million aligned read pairs. Most of the other reads map to the spinach genome but were no further analyzed.

Predicted proteins

The aligned transcript read pairs served as input for the BRAKER1 [37] pipeline to generate a *Pfs* specific training set for gene model predic-

tion. The gene prediction model was used to predict 13,227 gene models on the *Pfs* genome final assembly. The resulting protein models identified in the assembly of *Pfs1* were annotated using ANNIE [39]. ANNIE provided a putative annotation for 7297 *Pfs* proteins (**Supplemental Table 5**). We found that 12630 protein models reside on a contig larger than 1 kb and are thus contained in the filtered assembly. In addition, we found that 2983 gene models had 20% or more overlap with a repeat that was identified by RepeatMasker [19], another 952 protein models were annotated by ANNIE as transposable elements (**Figure 6**).

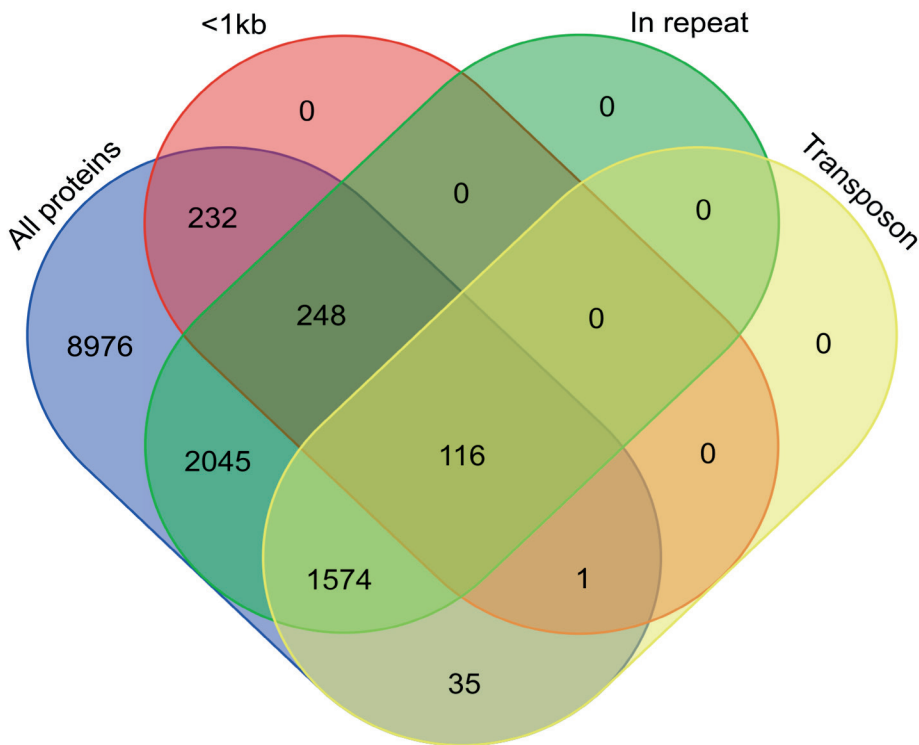


Figure 6. Venn diagram of the *Pfs* protein models. In the < 1kb category consists of proteins located on contigs smaller than 1kb. In repeat contains protein models that had more than 20% overlap with an identified repeat region in the genome. The transposons category contains the proteins that were annotated by ANNIE as transposons.

Based on the numbers presented in **Figure 6** we can conclude that most of the protein models residing on contigs smaller than 1 kb which have a significant overlap with a repeat region or are marked by ANNIE as transposons. The number of gene models found in the assembly of *Pfs1* is strikingly low in comparison to *P. infestans* (17,792), *H. arabidopsidis* (14,321), *Pl. halstedii* (15,469) and *P. tabacina* (11,310).

Secretome and host-translocated effectors

For the identification of the *Pfs* secretome as well as of candidate host-translocated RxLR and Crinkler effectors we choose to start with the proteins encoded by the initial 13,227 gene set. This reduces the risk of missing effectors that are encoded on smaller contigs (< 1 kb). SignalP [41] prediction identified 783 proteins with an N-terminal signal peptide. Of these, 231 were found to have an additional transmembrane domain (as determined by TMHMM [43] analysis) leaving 557 proteins. In addition, five of these carried a C-terminal KDEL motif that functions as an ER retention signal. The resulting set of 552 secreted proteins was used for secretome comparison, which is ~4% of the *Pfs1* proteome.

Previous research showed that some effectors of the lettuce downy mildew species *Bremia* have a single transmembrane domain in addition to the signal peptide [64]. Therefore, we chose to predict the host-translocated effectors not only from the secretome but also from the set of proteins with a signal peptide and an additional transmembrane domain. From these, a total of 99 putative RxLR or RxLR-like proteins and 14 putative Crinkler effectors were identified. Ten putative RxLR effector proteins were found to have a single transmembrane domain. Also, five putative RxLR effectors are located on contigs smaller than 1 kb (**Figure 7, Supplemental Table 6**). Of the 99 RxLR effectors, 64 had a canonical RxLR domain, while 35 had a degenerative RxLR domain combined with an EER-like and/or WY domain (**Figure 7**). The number of host-translocated effectors in *Pfs* is significantly smaller compared to that of *Phytophthora* species. Five of the identified putative Crinkler effectors had a canonical LFLAK domain. The others had a degenerative LFLAK combined with an HVL domain or were identified using the Crinkler HMM (**Figure 8**).

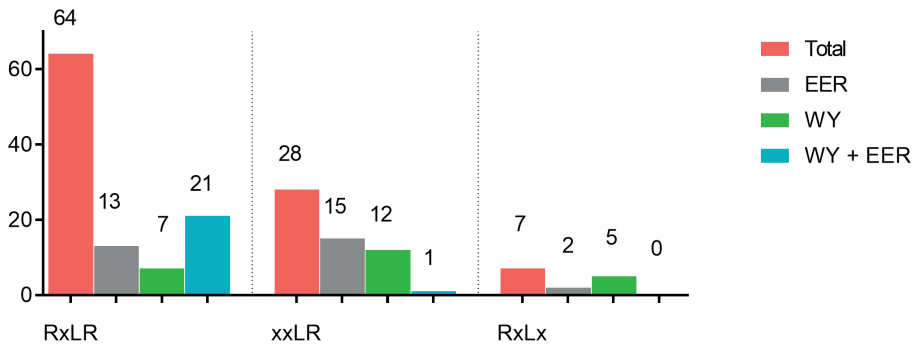


Figure 7. Overview of RxLR (-like) motifs observed in the putative RxLR effectors identified in the genome of *Pfs1*. For each (degenerate) RxLR motif the presence of a WY domain (orange), EER-like (green) domain or both (purple) is shown.

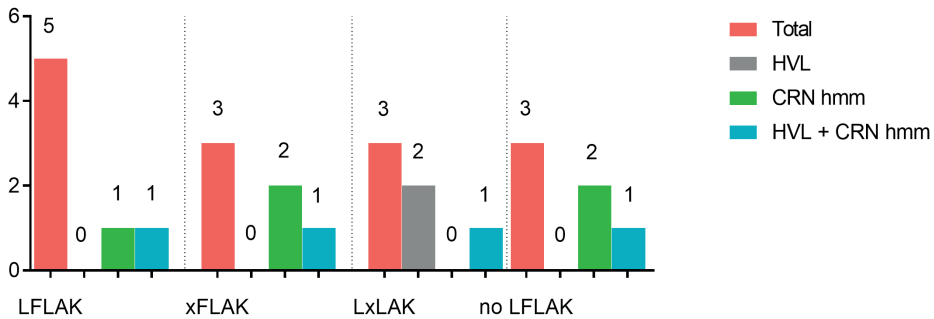
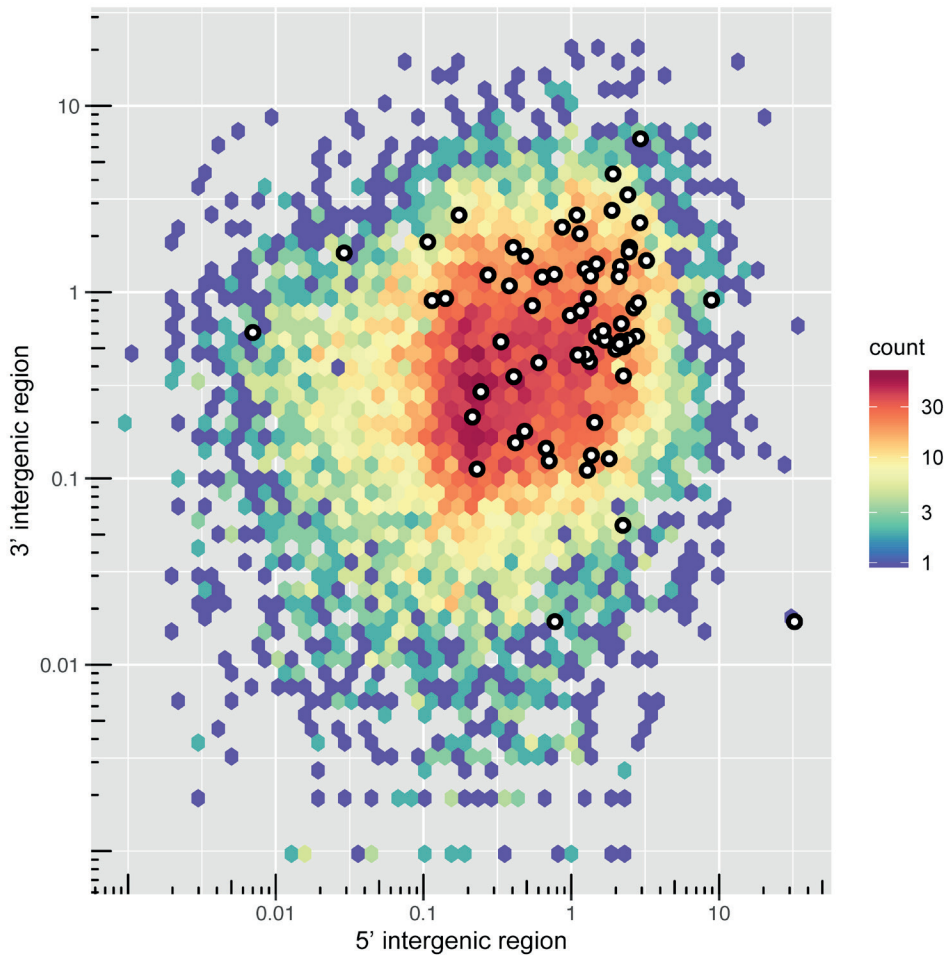


Figure 8. Overview of CRN (-like) motifs observed in the putative CRN effectors identified in the genome of *Pfs1*. For each (degenerate) CRN protein the presence of an HVL domain (orange), identified with a CRN HMM model (red) or both (green).

Genomic distribution of effectors

It has been previously described for the potato late blight pathogen *P. infestans* that effectors often reside on contigs with relatively large repeat content compared to the rest of the genome [65]. To gain insight into the genomic context of the *Pfs1* genome with regards to all 13,277 genes in the genome, and 66 selected RxLR effector genes we calculated the distance between neighboring genes as a measure of local gene density. To get a good overview of the gene intergenic distances in the genome we plotted the 3' and 5' values for all the genes in the *Pfs1* genome on a log₁₀ scaled heat map (**Figure 9**).



2

Figure 9. The genome spacing of each of 13,227 predicted genes of *Pfs1*, depicted by plotting the 5' and 3' intergenic distances (on a log₁₀ scale). The scale bar represents the number of genes in each bin, shown as a color-coded hexagonal heat map in which red indicates a gene dense and blue a gene-poor region. The locations of effector genes are indicated with white dots.

The genome of *Pfs1* is highly gene dense and effectors show a modest but significant (Wilcoxon rank-sum test resulting in P-value of $1.914e^{-11}$) enrichment in the gene-spare regions of the genome (**Figure 9**). The median 3' and 5' spacing for all genes is 925 bp, while for the selected effector genes it is 2,976 bp. However, the difference in gene density between the effector and core genes is not as strong as in *P. infestans* two-speed genome [28].

Comparative analysis of orthologs

Eighteen phytopathogenic oomycete species, that represent a diverse taxonomic range and different lifestyles, were chosen for a comparative analysis with *Pfs* (Table 3). The objective of the comparison is to see whether the biotrophic lifestyle of downy mildew species, like *Pfs*, is reflected in the secretome. For the analysis, the secretome of *Pfs* was compared to that of closely related *Phytophthora* (hemibiotrophic), *Plasmopara* (biotrophic) and more distantly related *Pythium* (necrotrophic) and *Albugo* (biotrophic) species. First, the predicted proteins of each species were used to create a multigene phylogenetic tree to infer their taxonomic relationships using Orthofinder. In total, 86.9% (267,813) of all proteins were assigned to 14,484 orthogroups. Of those, 2,383 had proteins from all species in the dataset of which 152 groups contained only one copy of a gene from each species. These single-copy orthologous genes of each species were used to infer a Maximum-likelihood species tree (Figure 10).

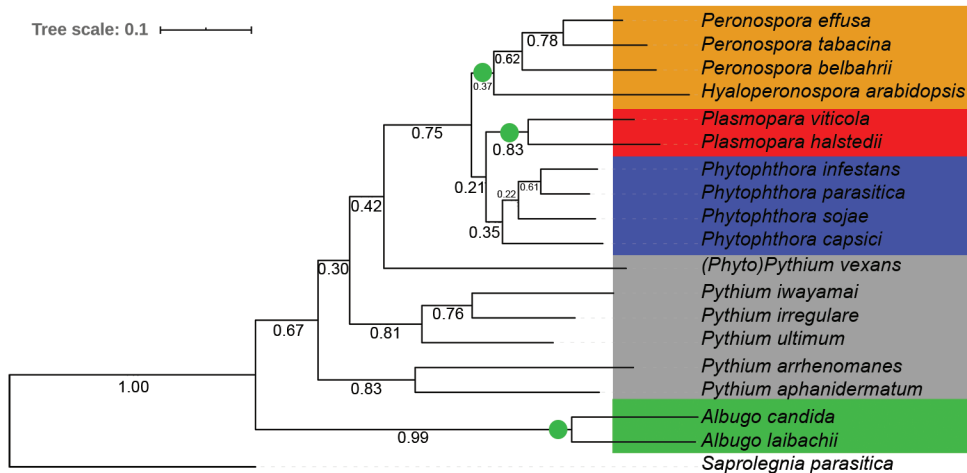


Figure 10. Maximum likelihood tree inferred from 152 single-copy orthogroups of proteins derived from the proteome of 18 plant infecting oomycete species. Branch numbers represent bootstrap values of N=12171 trees. Five phylogenetic clusters were defined for further analysis; *Hyaloperonospora/Peronospora* (green), *Plasmopara* (red), *Phytophthora* (blue), *Pythium* (grey) and *Albugo* (green). The lifestyle of the species in each cluster are indicated with symbols. The fish infecting species *Saprolegnia parasitica* was used as an outgroup.

The resulting tree shows that *Pfs* clusters with *H. arabidopsidis* (*Hpa*), *P. tabacina* (*Pta*) and *P. belbahrii* (*Pbe*). The closest relative of *Pfs* based on single-copy orthologs is the downy mildew of tobacco *Pta*, followed by the basil-infecting *Pbe*. Based on the tree, *Hpa* is more divergent from the former three downy mildew species within the *Hyaloperonospora/*

Peronospora clade. The *Plasmopara* downy mildew species are in a different clade that is more closely related to the *Phytophthora* species used in this study. The separation between the *Peronospora* lineage and the *Phytophthora/Plasmopara* lineages is well supported with a bootstrap value of 0.75. This clustering pattern is in line with the recent studies that suggest that the downy mildew species are not monophyletic within the Peronosporales [66; 67]. The *Phytophthora* species, although belonging to three different *Phytophthora* clades, are more closely related to each other than to the other species in this study. (*Phytophythium vexans* appears as a sister group to the *Phytophthora/Peronospora* lineage, which is in line with a recently published multi-gene phylogeny [68]. The other five species of *Pythium* form two clusters, as previously observed [68]. The two *Albugo* species form a cluster that is separated from the other clades with maximum bootstrap support.

Based on the core ortholog protein tree, we grouped the species into five phylogenetically-related clades; *Hyaloperonospora/Peronospora*, *Plasmopara*, *Phytophthora*, *Pythium*, and *Albugo* for further analysis of the secretome. Three of these clades have obligate biotrophic species (*Hyaloperonospora/Peronospora*, *Plasmopara* and *Albugo*), the *Phytophthora* cluster is hemibiotrophic and the species in the *Pythium* cluster have a necrotrophic lifestyle. (*Phyto*)*Pythium vexans* was included in the *Pythium* cluster. The fish-infecting *Saprolegnia parasitica* served as an outgroup for the phylogenetic tree and is not used for further comparison.

Secretome comparison

For each species, the total number of proteins and the subset that is considered secreted (signal peptide, no additional transmembrane domains, no ER retention signal) is shown in **Table 3**. *Phytophthora* species generally have a larger proteome than downy mildew species and secrete a larger percentage of the predicted proteins. The *Phytophthora* species in this study are predicted to secrete 1,976 proteins on average, whereas the *Plasmopara* and *Peronospora* species secrete an average of 1,461 and 703 proteins, respectively.



Table 3. Predicted secretomes of 18 oomycete species used in this study. The total number of predicted proteins, those with a signal peptide (SP), proteins with SP but without additional transmembrane domains (TM), and the number of proteins with SP, no TM, and no C-terminal KDEL sequence are shown. In the final column, the percentage of the proteome that is predicted to be secreted is highlighted.

| Species | Predicted proteins | Secretome | % secreted |
|---------------------------|--------------------|-----------|------------|
| <i>P. effusa</i> | 13,227 | 552 | 4.2 |
| <i>P. belbahrii</i> | 9,049 | 494 | 4.7 |
| <i>H. arabidopsidis</i> | 14,321 | 999 | 7.0 |
| <i>P. tabacina</i> | 18,447 | 798 | 4.3 |
| <i>Pl. halstedii</i> | 15,498 | 1,071 | 6.9 |
| <i>Pl. viticola</i> | 12,201 | 1,850 | 15.2 |
| <i>Ph. infestans</i> | 18,138 | 1,885 | 10.4 |
| <i>Ph. parasitica</i> | 27,942 | 2,250 | 8.1 |
| <i>Ph. sojae</i> | 26,584 | 2,337 | 8.8 |
| <i>Ph. capsici</i> | 19,805 | 14,33 | 7.2 |
| <i>Py. arrhenomanes</i> | 13,805 | 913 | 6.6 |
| <i>Py. aphanidermatum</i> | 12,312 | 928 | 7.5 |
| <i>Py. irregulare</i> | 13,805 | 961 | 7.0 |
| <i>Py. iawyamai</i> | 15,249 | 1,067 | 7.0 |
| <i>Py. vexans</i> | 11,958 | 863 | 7.2 |
| <i>Py. ultimum</i> | 15,322 | 1,071 | 7.0 |
| <i>A. candida</i> | 13,310 | 888 | 6.8 |
| <i>A. laibachii</i> | 13,804 | 679 | 4.9 |

Carbohydrate active enzymes

To functionally compare the secretomes between the five defined clades of species we first investigated the carbohydrate-active enzymes (CAZymes). These enzymes, among others, are involved in degrading and modifying plant cell walls, which is an important part of the infection process. Degradation and modification of cell walls are mediated by large numbers of enzymes that have complex interactions. A total of 95 unique CAZyme domains were found, using the CAZymes database (via db-CAN2.0 [55]), in the combined secretomes of the 18 oomycete species. The total number of CAZymes per species ranges from 35 in *A. laibachii* to 336 in *P. sojae* and was lower in obligate biotrophic species (35-193) compared to *Phytophthora* species (197-336) (**Supplemental Table 7**).

The presence and numbers of CAZyme domains were compared between species using a Principal Component Analysis (PCA), a statistical reduction technique that determines what variables contribute most to the variation observed in a data set. We report the relative abundance of each

CAZyme domain to the total number of secreted CAZyme domains per species to account for the large variation in absolute numbers of proteins between the species (**Figure 11**). The species clusters as depicted in **Figure 11** are supported by PERMANOVA (P-value = $1e^{-3}$).

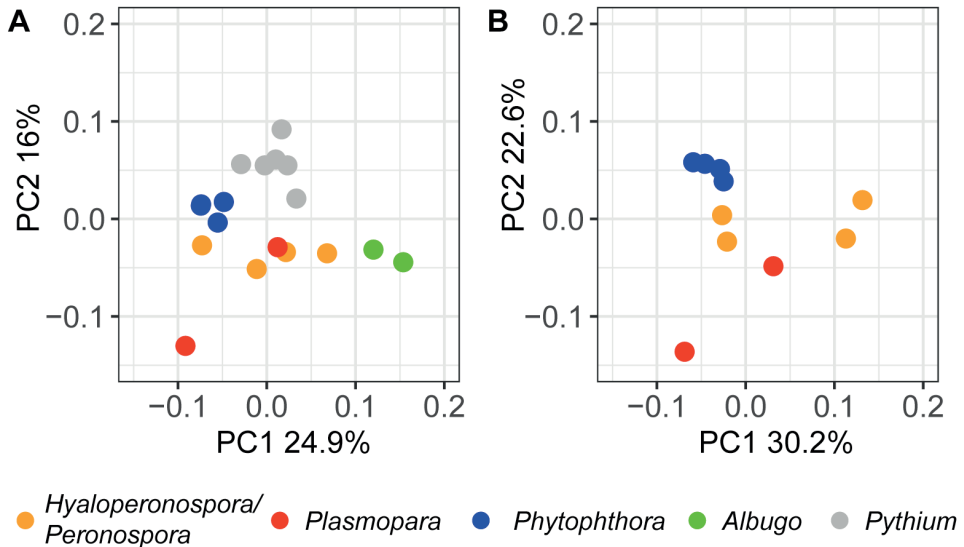


Figure 11. Principal component analysis (PCA) of the variation between species based on the predicted CAZymes in their secretome using the relative abundance of the identified CAZyme domains for each of the 18 species **A.** or for the *Peronospora*, *Plasmopara* and *Phytophthora* species only. **B.** The PERMANOVA test shows that the grouping based on the CAZyme domains is significant ($Pr > F = 1e^{-3}$). Species are grouped by color based on the classes that were defined in the phylogenetic tree (**Figure 10**). *Phytophthora* (blue), *Peronospora* (yellow), *Plasmopara* (red), *Albugo* (green) and *Pythium* (grey).

In accordance with the core ortholog protein tree (**Figure 10**), the *Albugo*, *Phytophthora* and *Pythium* species form separate clusters in the PCA. Remarkably, the *Hyaloperonospora/Peronospora* species do not cluster together, indicating that the secreted CAZyme domains vary largely between these species, despite their close phylogenetic relationship and same lifestyle. Similarly, the two *Plasmopara* species do not cluster together and are not closer to *Phytophthora* than *Peronospora* as could be expected based on the core ortholog protein tree (**Figure 10**). Instead, the *Plasmopara* and *Hyaloperonospora/Peronospora* species are not separated in the PCA based on the secreted CAZyme domains. The repertoire of CAZymes in all these downy mildew species is more similar than would be expected based on their taxonomic relationship. This could be the result of convergent evolution towards the obligate biotrophic lifestyle. However, the

Plasmopara and *Hyaloperonospora/Peronospora* species are in no way clustered as the position of the individual species which is highly variable among component one. The *Albugo* species do not cluster with the other biotrophic species.

To exclude the effect of the more distantly-related species on the separation between the downy mildew and *Phytophthora* species, the PCA was also performed on the set without the *Pythium* and *Albugo* species (**Figure 11B**). The pattern, as observed in the total set, is maintained when the more distantly related species are excluded from the analysis.

We conclude that a different composition and abundance in secreted CAZyme domains may contribute to obligate biotrophy as these species are clearly separated from the necrotrophic and hemibiotrophic ones. However, the variation within obligate biotrophic species is large, making it impossible to find a clear carbohydrate-active enzyme signature that defines the obligate biotrophs.

To look further into the properties of the secreted CAZymes we highlighted literature-curated domains of phytopathogenic oomycetes that are known to modify the main plant cell wall components; lignin, cellulose, and hemicellulose [69] (**Supplemental Figure 2**). We found that the secretomes differ more in terms of the absolute number of plant cell wall-degrading enzymes than in the relative occurrence of the different corresponding CAZyme catalytic activities per species. Secretomes of obligate biotrophic and hemibiotrophic/necrotrophic oomycetes have secreted proteins with similar functions (like the breakdown of cellulose, pectin, hemicellulose, etc.) but the numbers and diversity of those proteins in obligate biotrophic species are reduced.

Pfam domains

In a different approach, the secretomes were compared based on their Pfam annotation. InterPro was used to identify known protein domains and the numbers and presence of domains were compared between the species. A total of 1,354 unique domains were found in the combined secretomes of the oomycetes analyzed. The number of domains identified ranged from 304 in *Al. candida* to 1,710 in *Ph. parasitica*. The total number, as well as the relative number of Pfam domains in secretomes of obligate biotrophic species, was lower in obligate biotrophic species compared to *Phytophthora* and *Pythium* (**Supplemental Table 8**). PCA analysis was used to compare the Pfam annotations of the secretomes based on the relative number of secreted Pfam domains per species (**Figure 12**).

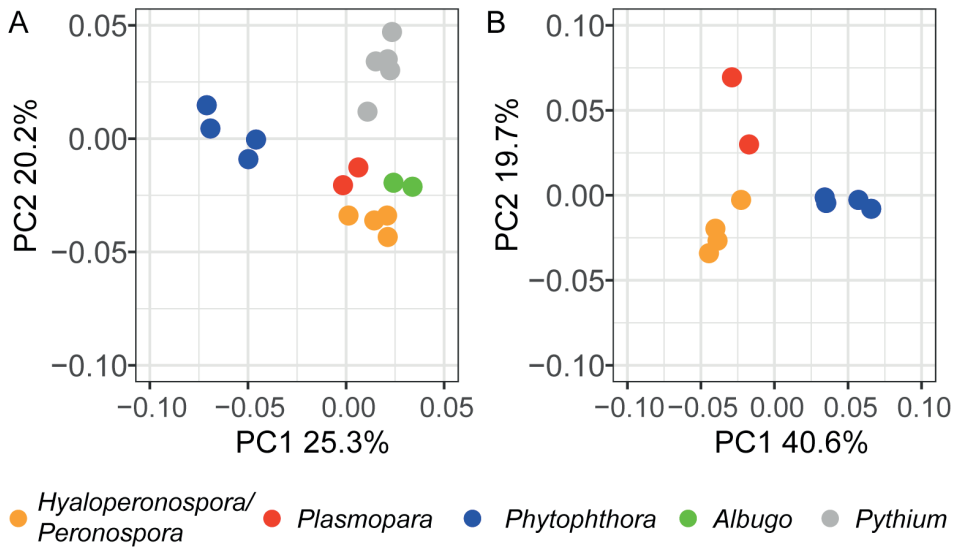


Figure 12. Principal component analysis (PCA) of the variation between species based on the predicted Pfam domains in their secretome using the relative abundance of the identified Pfam domains for each species. In **A** all species are included, while **B** shows the variation between downy mildew, *Plasmopara* and *Phytophthora* species only. Species are grouped by color based on the classes that were defined in the phylogenetic tree (Figure 1). The PERMANOVA test shows that the grouping based on the Pfam domains is significant (P -value $1e^{-3}$). *Phytophthora* (blue), downy mildew (yellow), *Plasmopara* (red), *Albugo* (green) and *Pythium* (grey).

The PCA in **Figure 12** shows a clear separation between lifestyles. The *Phytophthora* species cluster together and separate from all other species along PC1 (25.3%) (**Figure 12A**). The *Pythium* species form a cluster that separates clearly from the other species along PC2 (20.2%) (**Figure 12A**). Remarkably, all biotrophic species, including both groups of downy mildews and the *Albugo* species cluster together. Within the obligate biotrophic cluster, the phylogenetic groups (*Hyaloperonospora/Peronospora*, *Plasmopara*, *Albugo*) as found in the core ortholog tree are still present but the differences are minor. The species clusters as depicted in **Figure 12** are supported by PERMANOVA (P -value = $1e^{-3}$). To exclude the effect of the more distantly related species on the separation between the obligate biotrophs, the PCA was also performed without *Pythium* and *Albugo* species (**Figure 12B**). The pattern observed in **Figure 12A** is maintained when the more distantly related species are excluded from the analysis (**Figure 12B**). The Pfam domains that contribute to the variance in PC1 and PC2 were identified using a biplot. In a biplot, the variables are presented as vectors, with their length reflecting their contribution. Many of the domains contribute

to the differences between the biological groups, but five of them stand out (**Figure 13** and **Table 4**).

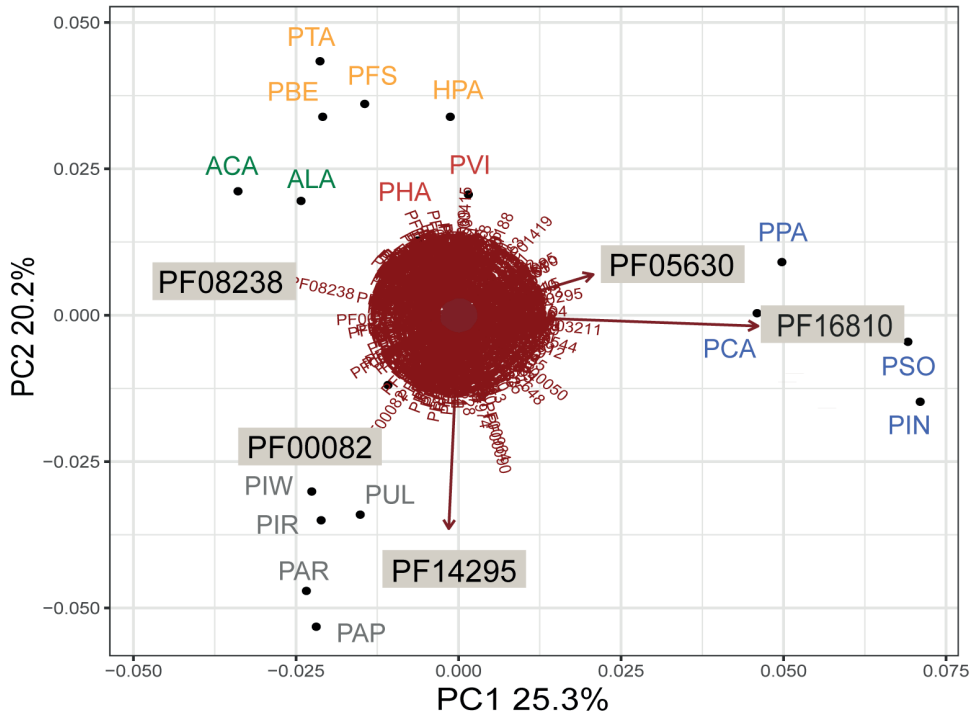


Figure 13. Biplot that shows the Pfam domains that provide to the variation between species based on the relative abundance of Pfam domains in their secretomes. Although many domains contribute to the variation, PF16810, PF05630, PF08238, PF14295, and PF00082 are the domains that contribute most, as evidenced by the length of their vectors. Two Pfam domains that have a higher relative abundance in *Phytophthora* contribute strongly to the separation between *Phytophthora* and the other species. The first, PF16810, represents an RxLR protein family with a conserved core α -helical fold (WY-fold). Some of the proteins that this domain was based on having a known avirulence activity [49]. On average, 82 PF16810 domains were identified in *Phytophthora* compared to 1.3 in *Peronospora*, 1.0 in *Plasmopara* and none in the *Albugo* clade. Using HMMer searches, many more WY domain candidates can be identified in these species, however, they apparently do not fit the Pfam domain that is based on a larger part of the proteins as the HMM. Abbr. PFS; *Peronospora* (*P.*) *effusa*, PBE; *P. belbahrii*, PTA; *P. tabacina*, HPA, *Hyaloperonospora arabidopsidis*, PHA; *Plasmopara* (*Pl.*) *halstedii*, PVI; *Pl. viticola*, PIN; *Phytophthora* (*Ph.*) *infestans*, PSO; *Ph. sojae*, PCA; *Ph. capsici*, PPA; *Ph. parasitica*, ACA; *Albugo candida*, ALA; *A. laibachii*, PUL; *Pythium* (*Py.*) *ultimum*, PAR; *Py. arrhenomanes*, PAP; *Py. Aphanidermatum*, PIR; *Py. irregulare*, PIW; *Py. lawyamai*, PVE; *Phytophthium vexans*

Table 4. Pfam domains that contribute most to the variation between species in the PCA. Numbers represent the number of domains per secretome per species. Domains that are relatively less abundant are blue, domains that occur in relatively high numbers are yellow.

| | | PF16810 RxLR | PF05630 NPP1 | PF08238 Sel1 repeat | PF14295 PAN/Apple | PF00082 Subtilase |
|------------------------|------------|-----------------|-----------------|------------------------|----------------------|----------------------|
| <i>Hpa/Peronospora</i> | <i>Pfs</i> | 3 | 7 | 16 | 1 | 1 |
| | <i>Pta</i> | 0 | 18 | 39 | 1 | 1 |
| | <i>Pbe</i> | 0 | 2 | 14 | 2 | 1 |
| | <i>Hpa</i> | 2 | 10 | 7 | 0 | 1 |
| <i>Plasmopara</i> | <i>Pha</i> | 2 | 15 | 10 | 3 | 5 |
| | <i>Pvi</i> | 0 | 10 | 23 | 0 | 20 |
| <i>Phytophthora</i> | <i>Pin</i> | 92 | 31 | 14 | 39 | 5 |
| | <i>Ppa</i> | 90 | 54 | 20 | 36 | 4 |
| | <i>Pso</i> | 104 | 59 | 22 | 31 | 2 |
| | <i>Pca</i> | 41 | 42 | 16 | 22 | 0 |
| <i>Pythium</i> | <i>Par</i> | 0 | 4 | 27 | 64 | 10 |
| | <i>Pap</i> | 0 | 3 | 30 | 60 | 21 |
| | <i>Pir</i> | 0 | 3 | 27 | 35 | 19 |
| | <i>Piw</i> | 0 | 2 | 24 | 33 | 17 |
| | <i>Pve</i> | 0 | 5 | 10 | 21 | 9 |
| | <i>Pul</i> | 0 | 7 | 25 | 33 | 26 |
| <i>Albugo</i> | <i>Aca</i> | 0 | 0 | 6 | 1 | 2 |
| | <i>Ala</i> | 0 | 0 | 13 | 5 | 2 |

Two Pfam domains that have a higher relative abundance in *Phytophthora* contribute strongly to the separation between *Phytophthora* and the other species. The first, PF16810, represents a RxLR protein family with a conserved core α -helical fold (WY-fold). Some of the proteins that this domain was based on have a known avirulence activity [50]. On average, 82 PF16810 domains were identified in *Phytophthora* compared to 1.3 in *Peronospora*, 1.0 in *Plasmopara* and none in the *Albugo* clade. Using HMMer searches, many more WY domain candidates can be identified in these species, however they apparently do not fit the Pfam domain that is based on a larger part of the proteins as the HMM.

The second domain, PF05630, is a necrosis-inducing protein domain (NPP1) that is based on a protein of *Ph. parasitica* [70]. This domain is conserved in proteins belonging to the family of Nep1-like proteins (NLPs) that occur in bacteria, fungi, and oomycetes [71]. Infiltration of cytotoxic NLPs in eudicot plant species results in cytolysis and cell death, visible as necrosis [72]. *Phytophthora* species are known to have high numbers of recently expanded NLP genes in their genomes, encoding both cytotoxic and

2

non-cytotoxic NLPs [71]. *H. arabidopsidis* and other obligate biotrophs tend to have lower numbers and only encode non-cytotoxic NLPs [71; 73].

Domain PF08238 contributes to the distance between the *Phytophthora* and obligate biotrophic species and is relatively more abundant in the biotrophs (PC1). PF08238 is a Sel1 repeat domain that is found in bacterial as well as eukaryotic species. Proteins with Sel1 repeats are suggested to be involved in protein or carbohydrate recognition and ER-associated protein degradation in eukaryotes [74]. No function of proteins with a PF08238 domain is known for oomycete or fungal pathogens.

The distance between *Pythium* and the obligate biotrophic species along PC2 is largely caused by differences in two domains that are commonly reported in oomycete secretomes [67]. The first, PF14295, a PAN/Apple domain, is known to be associated with a carbohydrate-binding module (CBM)-containing proteins that recognize and bind saccharide ligands in *Ph. parasitica*. Loss of these genes, as in the biotrophs, may facilitate the evasion of host recognition as they can induce plant defenses [75]. Second, PF00082 is a subtilase domain, which is found in a family of serine proteases. Secreted serine proteases are ubiquitous in secretomes of plant pathogens and play a role in nutrient acquisition through the breakdown of host proteins [76]. Secreted proteases from fungal species have been shown to enhance infection success by degrading plant-derived antimicrobial compounds [77].

Over and under-representation of Pfam domains in obligate biotrophic species

The enrichment analysis confirmed the pattern that was shown in the biplot. In total, 60 Pfam domains were found to be differentially abundant in obligate biotrophic species clusters compared to *Phytophthora* (Table 5). Three out of the five Pfam domains that contributed most to the separation between phylogenetic groups in the PCA (Figure 12 and Table 4) were also found to be differentially abundant in at least one obligate biotrophic cluster compared to *Phytophthora* in the enrichment analysis. The RxLR domain (PF16810) was under-represented in all three obligate biotrophic clusters. The PAN domain (PF14295) was under-represented in *Plasmopara* and *Peronospora* and the NPP1 domain (PF05630) in *Albugo*. Upon inspection a trend towards under-representation of NPP1 (Bonferroni corrected $P = 0.159$) in *Hyaloperonospora/Peronospora* is evident as well, although not significantly. The Sel1 domain (PF08238) significantly over represented between *Phytophthora* and *Hyaloperonospora/Peronospora*. The fifth domain that was found to contribute to the variation in the PCA was not found in the enrichment analysis, as it accounts for separation between *Pythium* and the other clusters. No additional domains were found to

be differentially abundant in more than one obligate biotrophic species cluster compared to *Phytophthora*. Previous studies identified Pfam domains that are associated with virulence in other phytopathogenic oomycete species like *Pythium*, *Plasmopara*, *Peronospora* and *Phytophthora* [78]. The occurrence of known virulence-associated domains of *Pfs* is summarized in **Supplemental Figure 3**. We found that obligate biotrophic species have a lower total number of secreted virulence-associated domains and that these domains make up a smaller portion of the total number of secreted proteins compared to the other species.

Table 5. Over and under-representation of Pfam domains in the secretomes of *Hyaloperonospora/Peronospora*, *Plasmopara* and *Albugo* compared to *Phytophthora* species. Over (green) and under (blue)-representation was tested relative to the expected distribution of each Pfam domain. The abundance of each domain was compared between the species clusters using a Chi-square test with Bonferroni correction. Bonferroni corrected P-values are shown in the table.

| Pfam | Name | Interpro | P | PI | AI |
|---------|--------------------------------------|-----------|---------------------|---------------------|---------------------|
| PF16810 | RxLR protein, Avirulence activity | IPR031825 | 8.3e ⁻²⁴ | 2.7e ⁻¹⁶ | 3.9e ⁻⁰⁷ |
| PF14295 | PAN domain | IPR003609 | 1.8e ⁻⁰⁷ | 1.8e ⁻⁴ | 16.913 |
| PF00090 | Thrombospondin type 1 | IPR000884 | 4.5e ⁻⁰⁵ | 77.4987 | 1.98225 |
| PF05630 | Necrosis inducing protein (NPP1) | IPR008701 | 1.6e ⁻¹ | 1.36788 | 2.1e ⁻³ |
| PF08238 | Sel1 repeat | IPR006597 | 1.7e ⁻⁰⁷ | 4.61559 | 1.07208 |
| PF00254 | FKBP-type cis-trans isomerase | IPR001179 | 1.8e ⁻⁴ | | |
| PF00050 | Kazal-type serine protease inhibitor | IPR002350 | 2.0e ⁻³ | | |
| PF07974 | eGF-like domain | IPR013111 | 1.2e ⁻² | | |
| PF13456 | Reverse transcriptase-like | IPR002156 | 2.6e ⁻¹⁰ | | |
| PF00300 | Histidine phosphatase superfamily | IPR013078 | 7.7e ⁻³ | | |
| PF00665 | Integrase core domain | IPR001584 | 1e ⁻² | | |
| PF00571 | CBS domain | IPR000644 | 1.7e ⁻² | | |
| PF00089 | Trypsin | IPR001254 | | 1.1e ⁻¹² | |
| PF01833 | IPT/TIG domain | IPR002909 | | 8e ⁻¹² | |
| PF00082 | Subtilase family | IPR000209 | | 3.1e ⁻¹⁰ | |
| PF01341 | Glycosyl hydrolases family 6 | IPR016288 | | 3.3e ⁻⁰⁵ | |
| PF00182 | Chitinase class I | IPR000726 | | 2.6e ⁻⁴ | |
| PF01670 | Glycosyl hydrolase family 12 | IPR002594 | | 1e ⁻³ | |
| PF03184 | DDe superfamily endonuclease | IPR004875 | | | 2.4e ⁻⁰⁶ |
| PF09818 | Predicted ATPase of the ABC class | IPR019195 | | | 4.6e ⁻⁰⁶ |
| PF00169 | PH domain | IPR001849 | | | 8.1e ⁻⁰⁶ |
| PF01764 | Lipase (class 3) | IPR002921 | | | 1.3e ⁻⁴ |
| PF00026 | eukaryotic aspartyl protease | IPR033121 | | | 2.3e ⁻⁴ |
| PF13405 | eF-hand domain | IPR002048 | | | 3.7e ⁻⁴ |

Table 5. continued.

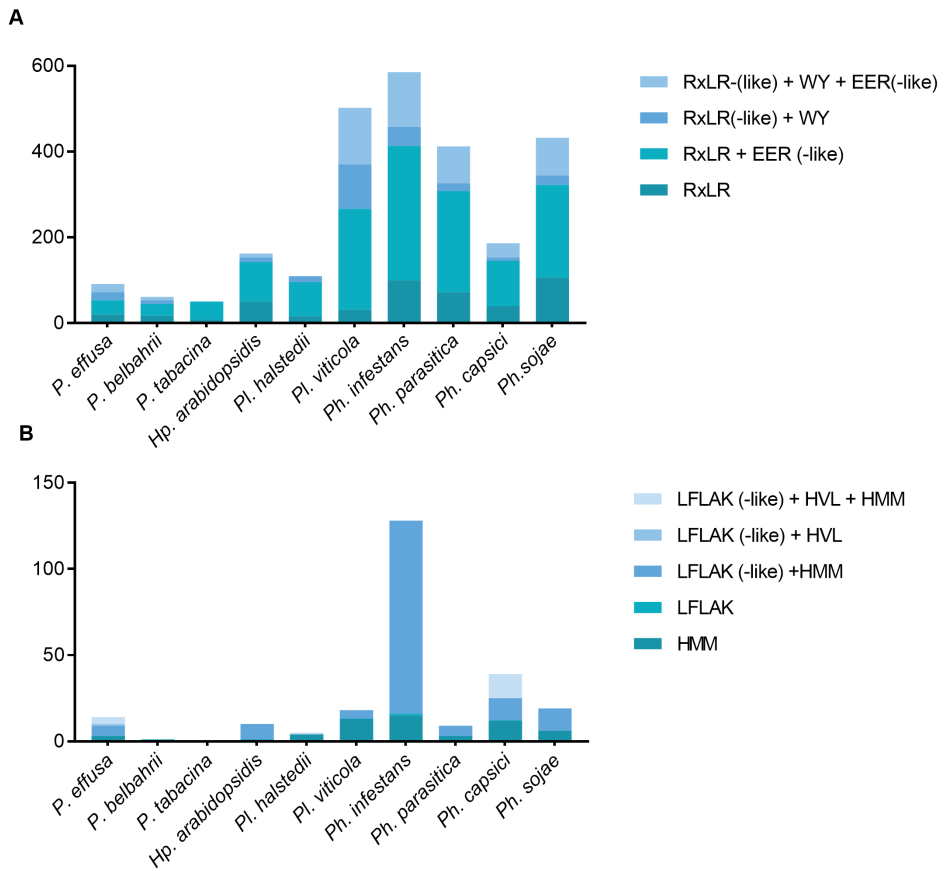
| Pfam | Name | Interpro | P | PI | AI |
|---------|---------------------------------------|-----------|---|----|--------------------|
| PF15924 | ALG11 mannosyltransferase | IPR031814 | | | 3.7e ⁻⁴ |
| PF01546 | Peptidase family M20/M25/M40 | IPR002933 | | | 3.7e ⁻⁴ |
| PF07687 | Peptidase dimerisation domain | IPR011650 | | | 3.7e ⁻⁴ |
| PF03870 | RNA polymerase Rpb8 | IPR005570 | | | 3.7e ⁻⁴ |
| PF13041 | PPR repeat family | IPR002885 | | | 3.7e ⁻⁴ |
| PF00443 | Ubiquitin carboxyl-terminal hydrolase | IPR001394 | | | 1.4e ⁻³ |
| PF10152 | Subunit CCDC53 of WASH complex | IPR019309 | | | 1.6e ⁻³ |
| PF00041 | Fibronectin type III domain | IPR003961 | | | 2.3e ⁻³ |
| PF07727 | Reverse transcriptase | IPR013103 | | | 6e ⁻³ |
| PF04130 | Spc97/Spc98 family | IPR007259 | | | 2.2e ⁻² |
| PF01753 | MYND finger | IPR002893 | | | 2.2e ⁻² |
| PF03577 | Peptidase family C69 | IPR005322 | | | 2.2e ⁻² |
| PF03388 | Legume-like lectin family | IPR005052 | | | 3e ⁻² |
| PF03133 | Tubulin-tyrosine ligase family | IPR004344 | | | 3e ⁻² |
| PF13181 | Tetratricopeptide repeat | IPR019734 | | | 3e ⁻² |
| PF01156 | Nucleoside hydrolase | IPR001910 | | | 3e ⁻² |
| PF06367 | Diaphanous FH3 Domain | IPR010472 | | | 3e ⁻² |
| PF04910 | Transcriptional repressor TCF25 | IPR006994 | | | 3e ⁻² |
| PF00044 | Glyceraldehyde 3-ph.dehydrogenase | IPR020828 | | | 3e ⁻² |
| PF02800 | Glyceraldehyde 3-ph.dehydrogenase | IPR020829 | | | 3e ⁻² |
| PF01428 | AN1-like Zinc finger | IPR000058 | | | 3e ⁻² |
| PF00766 | Electron transfer FAD-binding domain | IPR014731 | | | 3e ⁻² |
| PF01012 | Electron transfer flavoprotein domain | IPR014730 | | | 3e ⁻² |
| PF03690 | UPF0160 (uncharacterized) | IPR003226 | | | 3e ⁻² |
| PF13307 | Helicase C-terminal domain | IPR006555 | | | 3e ⁻² |
| PF08683 | Microtubule-binding calmodulin-reg | IPR014797 | | | 3e ⁻² |
| PF01846 | FF domain | IPR002713 | | | 3e ⁻² |
| PF13418 | Galactose oxidase | | | | 3e ⁻² |
| PF03776 | MinE | IPR005527 | | | 3e ⁻² |
| PF13815 | Iguana/Dzip1-like DAZ-interacting | IPR032714 | | | 3e ⁻² |
| PF04851 | Type III restriction enzyme | IPR006935 | | | 3e ⁻² |
| PF13831 | PHD-finger | | | | 3e ⁻² |
| PF04045 | Arp2/3 complex.p34-Arc | IPR007188 | | | 3e ⁻² |
| PF08144 | CPL (NUC119) domain | IPR012959 | | | 3e ⁻² |
| PF00659 | POLO box duplicated region | IPR000959 | | | 3e ⁻² |
| PF08450 | SMP-30/Gluconolactonase/LRE-like | IPR013658 | | | 4.8e ⁻² |

Host-translocated effectors

The RxLR effector models in the Pfam database (PF16810 and PF16829) mentioned above cover only a small fraction of the predicted RxLR effectors in secretomes of phytopathogenic oomycetes. We predicted the total number of host translocated effectors for each secretome using a Perl regex script and HMM searches (see methods) to get a more comprehensive view of the total numbers of host translocated effectors. This includes RxLR effectors without WY domains and CRN effectors.

The predicted host translocated effectors are classified into groups based on the identified domains and shown in **Figure 14**. Please note that the number of *Pfs* effectors is slightly different from the numbers reported before. The numbers in **Figure 14** are derived from predictions based on published HMM models instead of models trained for *Pfs* (**Figure 7**), for the purpose of comparison between species. The same criteria for effector prediction were used for all species in this comparison.

RxLR effector proteins are more abundant in *Phytophthora* compared to the obligate biotrophic species in this study. On average 399 RxLR effector proteins were found in *Phytophthora* whereas *Plasmopara* and *Hyaloperonospora/Peronospora* had 79 and 90 on average. The same pattern is evident for CRN effectors. The average number of CRN proteins in *Hyaloperonospora/Peronospora* is 11, while *Plasmopara* has 12 and *Phytophthora* 56 on average. We conclude that downy mildew species (*Hyaloperonospora/Peronospora* and *Plasmopara*) have fewer host translocated effectors compared to *Phytophthora* species.



Discussion

The ability to sequence full genomes at a high pace and relatively low cost has aided research in phytopathology dramatically. Over the past few years, the genomes of many phytopathogenic oomycetes have been sequenced and their genomes revealed an arsenal of protein-coding genes that were found to be important in virulence. However, technical difficulties restricted the sequencing and assembly of genomes of obligate biotrophic oomycetes that cannot be cultured axenically. Obligate biotrophic species can only grow on living host tissue so when collecting spores for DNA isolation bacterial and plant DNA will inevitably contaminate the sample, which complicates the genome assembly. In this paper we describe the assembly of a relatively clean genome sequence of the obligate biotrophic downy mildew of spinach, *Peronospora effusa*.

To get a clean assembly, sequence that are derived from different species needed to be filtered out and removed. Several methods were considered to identify and filter contigs or reads that were likely contaminants in our data. Initially we considered to filter contigs or reads based on their GC content, since this differs between oomycete genomes [79] and most other microbes [80]. However, this is not the case for all bacteria, for example *E. coli* has a GC content of 51.7% [80], which is close to that of most oomycetes. In addition, the GC bases are not evenly distributed over the genome, so filtering based on GC content could therefore potentially remove valuable parts of the genome.

Alternatively, reads of non-oomycete origin could be identified by mapping them to databases with sequences of known taxonomy. For example, a database containing only oomycete or bacterial genomes. This is not ideal as the databases are incomplete and are likely to contain annotation errors. In addition, it could lead to the removal of novel parts of the downy mildew genome that are not present in other oomycetes, eventually, this will hamper the study on the valuable species-specific parts of your genome.

The filtering we applied with the CAT tool does not classify a contig based on a single hit [11]. Instead, it determines the taxonomic origin of each ORF on an assembled contig or corrected PacBio read. When the accumulated bit-scores of each ORF are above the set threshold it will assign the full contig to a specific taxon. After filtering, 50% of the error-corrected PacBio reads remained and were used in the assembly and 56% of the Illumina sequencing reads aligned with the assembled genome. This indicates that roughly half of our sequenced data originated from another microbial source besides *Pfs*. Notably, while the classifications in the original CAT paper were only benchmarked on prokaryotic sequences [11], our study shows that the tool also performs well for classifying eukaryotic contigs. Thus, the

CAT tool may also be promising for classification of eukaryotes including oomycetes in metagenomic datasets, provided that long contigs, or corrected PacBio or Nanopore sequencing reads are available. It should be noted that sequences of unknown taxonomy were maintained for the assembly, making it possible that these are still contaminants. When we compare the taxonomic distributions generated by Kaiju of the pre-assembly and final assembly, we see a dramatic reduction of sequences of bacterial origin (**Figure 5**). The oomycete content according to Kaiju and the overall GC content of the final assembly is similar to that of genome assemblies of axenically grown oomycetes. We can, therefore, conclude that with the CAT filtering method, we successfully filtered out most of the genomic reads that did not belong to the oomycete genus.

Most oomycete genomes sequenced to date were found to contain long repeat regions [81] that cannot be resolved using only a short read technology such as Illumina. Long reads can potentially sequence over these long stretches of repeat sequences, and contribute to the contiguity of the genome assembly [82]. Therefore, the data was complemented with long-read PacBio sequences in an attempt to close the gaps between the contigs. The resulting assembled *Pfs1* genome consists of a large number of contigs. This suggests that even with PacBio reads we were unable to span many repeat regions. Besides a biological reason for the large number of contigs, there could also be a technical reason. Prior to PacBio sequencing whole genome amplification with random primers was performed as the initial sequencing attempt with non-amplified DNA barely yield sequencing reads. This creates a bias, where some parts of the oomycete the genome may be differentially represented in the PacBio data set by chance. Nevertheless, CAT effectively removed the bacterial contamination providing a clean input for the assembly.

The assembled *Pfs1* genome size is 32.40 Mb divided over of 8,635 contigs. The genome is highly gene dense and contains in total 13,227 genes. Overall, the BUSCO analysis showed that this assembly contains most of the gene-space. Many of the 8,635 contigs were smaller than 1 kb. However, the CAT filtering method performs best on relatively large contigs containing multiple ORFs. Therefore, small contigs could still contain sequences derived from other microbes. The removal of these small contigs results in a small genome size reduction (1.9 Mb), but significantly reduces the number of contigs (by 5,027). When we account for genes that have significant overlap (> 20%) with repeats, or that were annotated as transposable elements we come to 9,007 high-confidence gene models.

The genomes of *Pfs* R13 and R14 (race *Pfs13* and *14*) have recently been published [83], with similar genome sizes (32.1 Mb, and 30.8 Mb respectively) and gene content (~8,000 gene models) compared to the assembly of *Pfs1*. Contrary to our assembly method, the input data for

these genome assemblies were filtered against an oomycete and bacterial database to discard reads that do not belong to the oomycete genus. This filtering method is prone to the incorporation of bacterial sequences that are not in the public databases. Besides, the positive filtering for oomycete scaffolds against the NCBI nt database could have resulted in the loss of *Pfs* specific genome sequences. In addition, the assembly of R13 and R14 were based on the merging of several assemblies of different *k*-mer characteristics. Although this improves the contiguity of the assembly, this method might yield collapsed and chimeric contigs.

Recent sequencing of *Peronospora* species shows that they have remarkably small and compact genomes (32.3–63.1 Mb) compared to *Phytophthora* (82–240 Mb) species [28; 31; 83; 84]. The *k*-mer analysis predicts the *Pfs1* genome to be 36.18 Mb containing 8.78 Mb of repeats (24%). The predicted genome size of *Pfs R13* and *R14* based on *k*-mer prediction is 44.1–41.2 Mb (repeats; 24–22%) [83].

The increased genome size of *Phytophthora* is attributed to an ancestral whole genome duplication in the lineage leading to *Phytophthora* and to an increase in the proportion of repetitive non-coding DNA [28; 85]. The duplication event has been proposed to have taken place after the speciation of *Hp. arabidopsidis* [86]. However new multigene phylogenies show that the *Peronospora* lineage has speciated after the divergence of *Phytophthora* clade 7 from clade 1 and 2. Notably, these three clades all contain species with duplicated genomes [32; 66; 87; 88]. This would suggest that an ancestral whole genome duplication before this speciation point would also apply to *Peronospora*, so that duplication cannot account for the difference in genome size. The availability of more *Peronospora* genomes for comparisons asks for a reevaluation of the timing of the duplication and subsequent speciation events.

Biologically, the question of how *Peronospora* species can be host-specific and obligate biotrophic while maintaining only a small and compact genome is interesting. It is argued that the trend in filamentous phytopathogens is towards large genomes with repetitive stretches to enhance genome plasticity [85]. Plasticity may enable host jumps and adaptations that favor the species for survival over species with small, less flexible genomes [85]. The reduced genomes of *Peronospora* species show an opposing trend that cannot be attributed to their obligate biotrophic lifestyle alone, as it is not evident in *Plasmopara* species (75–92 Mb) [32; 89]. Sequencing of multiple races of the same *Peronospora* species may shed light on genome plasticity at the species level.

The biotrophic lifestyle has emerged on several independent occasions in filamentous plant pathogens, in several branches of the tree of life. Convergent evolution is thought to be the main driving factor behind the development of biotrophy in such distantly related organisms [90]. However, it

was shown that horizontal gene transfer can also occur between fungi and oomycetes, resulting in 21 fungal proteins in the secretome of *Hp. arabidopsidis*. Out of these 21 proteins, 13 were predicted to be secreted, indicating that horizontal gene transfer may affect species pathogenicity and interaction with the host [91; 92].

It was proposed that the critical step for adopting biotrophy in filamentous phytopathogens is the ability to create and maintain functional haustoria [93]. To do so, a species needs to be able to avoid host recognition or suppress the host defense response. A proposed mechanism for avoidance of host recognition is the loss of proteins involved in cell wall degradation, as evidenced by the reduction of cell wall degrading enzymes in mutualistic species compared to biotrophs [94]. In this and other studies, we find a reduction of the number of cell wall degrading enzymes in obligate biotrophic species compared to hemibiotrophic *Phytophthora* species (**Supplemental Figure 2**) [26]. This is true for all three obligate biotrophic groups in this study (*Hyaloperonospora/Peronospora*, *Plasmopara* and *Albugo*) although the difference is less clear in *Plasmopara*. Possibly this reduction is the result of a similar selection pressure to reduce recognition by the host plant in the biotrophic species, where the hemibiotrophic nature of the interaction between host and *Phytophthora* allows for slightly less caution in recognition avoidance.

The other mechanism of establishing a strong interaction is suppression or avoidance of the host defense response. Biotrophic infections are often accompanied by co-infection of species that are unable to infect the plant in the absence of the biotroph, indicating efficient defense suppression [93; 95]. We found enhanced numbers of secreted serine proteases (PF00082) (suppression) and reduced numbers of proteins with PAN/Apple domains that are known to be recognized by the plant immune system.

While the expansion of host translocated RxLR effectors is evident in both hemibiotrophic and biotrophic species, their numbers are smaller in secretomes of obligate biotrophs. Crinkler effectors are especially reduced in secretomes of biotrophic species. As opposed to RxLR effectors, CRNs are an ancient class of effectors that are known to induce cell death. Obligate biotrophic species presumably lost them as they are not beneficial for their survival.

In this study we first showed that the CAT tool performs well for taxonomic filtering of eukaryotic contigs. We provided a clean reference genome of the oldest known isolate of the spinach infecting downy mildew, *Pfs1*. In a comparative approach, we found that the secretomes of the obligate biotrophic oomycetes are more similar to each other than to more closely related hemibiotrophic species when comparing the presence and absence of functional domains, including the host translocated effectors.

We conclude that adaptation to biotrophy is reflected in the secretome of oomycete species.

Acknowledgments

We wish to thank Topsector Horticulture and Starting Materials (TKI) for funding the project. B. Dutilh was supported by the Netherlands Organisation for Scientific Research (NWO) Vidi grant 864.14.004. We thank Ronnie de Jonge (Utrecht University) for useful input for the orthology analysis, Bjorn Wouterse for helping out with the comparative and statistical analysis, and the Utrecht Sequencing Facility for providing the Illumina sequencing service and data. Utrecht Sequencing Facility is subsidized by the University Medical Center Utrecht, Hubrecht Institute, Utrecht University and The Netherlands X-omics Initiative (NWO). We thank Keygene for technical assistance on the whole genome amplification, preparation and subsequent sequencing of the PacBio libraries.

References

1. Lee SC, Ristaino JB, Heitman J. 2012. Parallels in intercellular communication in oomycete and fungal pathogens of plants and humans. *PLoS Pathog.* 8:e1003028
2. Wang S, Welsh L, Thorpe P, Whisson SC, Boevink PC, Birch PR. 2018. The *Phytophthora infestans* haustorium is a site for secretion of diverse classes of infection-associated proteins. *MBio* 9:e01216-18
3. Ellis JG, Rafiqi M, Gan P, Chakrabarti A, Dodds PN. 2009. Recent progress in discovery and functional analysis of effector proteins of fungal and oomycete plant pathogens. *Curr. Opin. Plant Biol.* 12:399-405
4. Deb D, Anderson RG, How-Yew-Kin T, Tyler BM, McDowell JM. 2018. Conserved RxLR effectors from oomycetes *Hyaloperonospora arabidopsidis* and *Phytophthora sojae* suppress PAMP – and Effector-Triggered Immunity in diverse plants. *Mol. Plant-Microbe Interact.* 31:374-85
5. Birch PR, Rehmany AP, Pritchard L, Kamoun S, Beynon JL. 2006. Trafficking arms: oomycete effectors enter host plant cells. *Trends Microbiol.* 14:8-11
6. Kandel SL, Mou B, Shishkoff N, Shi A, Subbarao KV, Klosterman SJ. 2019. Spinach downy mildew: advances in our understanding of the disease cycle and prospects for disease management. *Plant Dis.* 103:791-803
7. Koike S, Smith R, Schulbach K. 1992. Resistant cultivars, fungicides combat downy mildew of spinach. *Calif. Agric.* 46:29-30
8. Feng C, Saito K, Liu B, Manley A, Kammeijer K, *et al.* 2018. New races and novel strains of the spinach downy mildew pathogen *Peronospora effusa*. *Plant Dis.* 102:613-8
9. Xu C, Jiao C, Sun H, Cai X, Wang X, *et al.* 2017. Draft genome of spinach and transcriptome diversity of 120 *Spinacia* accessions. *Nat. Commun.* 8:15275
10. Thines M, Choi YJ. 2016. Evolution, diversity, and taxonomy of the Peronosporaceae, with focus on the genus *Peronospora*. *Phytopathology* 106:6-18
11. von Meijenfeldt FAB, Arkhipova K, Cambuy DD, Coutinho FH, Dutilh BE. 2019. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.* 20:217
12. Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27:863-4
13. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, *et al.* 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* 13:1050-4
14. 2019. SMRT Analysis Software. PacBio
15. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119

16. Buchfink B, Xie C, Huson DH. 2014. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12:59
17. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, *et al.* 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19:455-77
18. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25
19. Smit A, Hubley R, Green P. 2015. RepeatMasker Open-4. 0. 2013-2015. <http://www.repeatmasker.org>
20. Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27:764-70
21. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, *et al.* 2017. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33:2202-4
22. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*
23. Bushnell B. 2014. BMAP: a fast, accurate, splice-aware aligner, Lawrence Berkeley National Lab. (LBNL), United States
24. Meneghin J. 2009. Get GC Content (Perl script). https://github.com/CostaLab/practical_SS2015/
25. Wickham H. 2016. *ggplot2: elegant graphics for data analysis*. Springer. VIII, 213 pp.
26. Baxter L, Tripathy S, Ishaque N, Boot N, Cabral A, *et al.* 2010. Signatures of adaptation to obligate biotrophy in the *Hyaloperonospora arabidopsidis* genome. *Science* 330:1549-51
27. Thines M, Sharma R, Rodenburg SYA, Gogleva A, Judelson HS, *et al.* 2019. The genome of *Peronospora belbahrii* reveals high heterozygosity, a low number of canonical effectors and CT-rich promoters. *bioRxiv:721027*
28. Haas BJ, Kamoun S, Zody MC, Jiang RH, Handsaker RE, *et al.* 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* 461:393-8
29. Fletcher K, Gil J, Bertier LD, Kenefick A, Wood KJ, *et al.* 2019. Genomic signatures of somatic hybrid vigor due to heterokaryosis in the oomycete pathogen, *Bremia lactucae*. *bioRxiv:516526*
30. Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RH, *et al.* 2006. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* 313:1261-6
31. Derevnina L, Chin-Wo-Reyes S, Martin F, Wood K, Froenicke L, *et al.* 2015. Genome sequence and architecture of the tobacco downy mildew pathogen *Peronospora tabacina*. *Mol. Plant-Microbe Interact.* 28:1198-215
32. Dussert Y, Mazet ID, Couture C, Gouzy J, Piron MC, *et al.* 2019. A high-quality grapevine downy mildew genome assembly reveals rapidly evolving and lineage-specific putative host adaptation genes. *Genome Biol. Evol.* 11:954-69

33. Menzel P, Ng KL, Krogh A. 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* 7:11257
34. Huang W, Li L, Myers JR, Marth GT. 2012. ART: a next-generation sequencing read simulator. *Bioinformatics* 28:593-4
35. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210-2
36. Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25:1105-11
37. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* 32:767-9
38. Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, *et al.* 2013. Web Apollo: a web-based genomic annotation editing platform. *Genome Biol.* 14:R93
39. Ooi HS, Kwo CY, Wildpaner M, Sirota FL, Eisenhaber B, *et al.* 2009. ANNIE: integrated de novo protein sequence annotation. *Nucleic Acids Res.* 37:W435-40
40. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, *et al.* 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44:D279-85
41. Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4. 0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8:785-6
42. Sperschneider J, Williams AH, Hane JK, Singh KB, Taylor JM. 2015. Evaluation of secretion prediction highlights differing approaches needed for oomycete and fungal effectors. *Front. Plant Sci.* 6:1168
43. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. 2001. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.* 305:567-80
44. Win J, Krasileva KV, Kamoun S, Shirasu K, Staskawicz BJ, Banfield MJ. 2012. Sequence divergent RXLR effectors share a structural fold conserved across plant pathogenic oomycete species. *PLoS Pathog.* 8:e1002400
45. Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics (Oxford, England)* 14:755-63
46. Klein J. 2018. *GitHub repository, <https://github.com/kleinjoell/bioscripts/>.*
47. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, *et al.* 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* 12:1611-8
48. Johnson LS, Eddy SR, Portugaly E. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* 11:431
49. Boutemy LS, King SRF, Win J, Hughes RK, Clarke TA, *et al.* 2011. Structures of *Phytophthora* RXLR effector proteins: A conserved but adaptable fold underpins functional diversity. *J. Biol. Chem.* 286:35834-42
50. Eddy S, Wheeler T. 2007. *HMMER-biosequence analysis using profile hidden Markov models.*

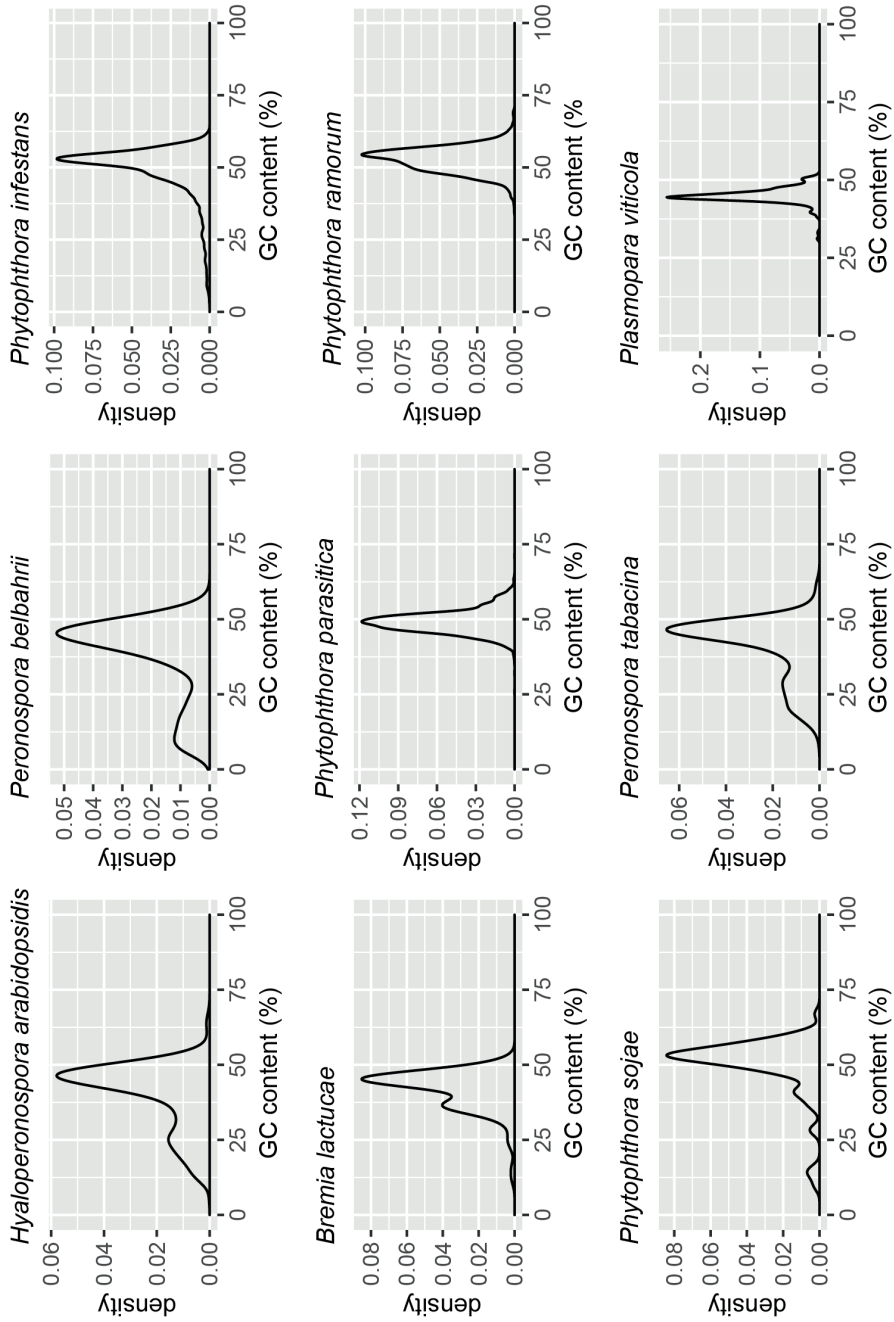
51. Armitage AD, Lysøe E, Nellist CF, Lewis LA, Cano LM, *et al.* 2018. Bioinformatic characterisation of the effector repertoire of the strawberry pathogen *Phytophthora cactorum*. *PLoS ONE* 13:e0202305
52. Wilcoxon F, Katti S, Wilcox RA. 1970. *Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test.* . 171-259 pp.
53. Jones P, Binns D, Chang H-Y, Fraser M, Li W, *et al.* 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30:1236-40
54. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* 37:D233-D8
55. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, *et al.* 2018. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* 46:W95-W101
56. Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16:157
57. Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30:772-80
58. Lefort V, Desper R, Gascuel O. 2015. FastME 2. 0: A comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.* 32:2798-800
59. He Z, Zhang H, Gao S, Lercher MJ, Chen WH, Hu S. 2016. Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. *Nucleic Acids Res.* 44:W236-W41
60. McMurdie PJ, Holmes S. 2013. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* 8:e61217
61. Team R. 2013. R: A language and environment for statistical computing.
62. Team R. 2015. RStudio: integrated development for R. RStudio, Inc., Boston, MA <http://www.rstudio.com>
63. Dixon P. 2003. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* 14:927-30
64. Meisrimler CN, Pelgrom AJE, Oud B, Out S, van den Ackerveken G. 2019. Multiple downy mildew effectors target the stress-related NAC transcription factor LsNAC069 in lettuce. *Plant J.* 0
65. Dong S, Raffaele S, Kamoun S. 2015. The two-speed genomes of filamentous pathogens: waltz with plants. *Curr. Opin. Genet. Dev.* 35:57-65
66. Bourret TB, Choudhury RA, Mehl HK, Blomquist CL, McRoberts N, Rizzo DM. 2018. Multiple origins of downy mildews and mito-nuclear discordance within the paraphyletic genus *Phytophthora*. *PLoS ONE* 13:e0192502
67. McGowan J, Fitzpatrick DA. 2017. Genomic, network, and phylogenetic analysis of the oomycete effector arsenal. *mSphere* 2:e00408-17



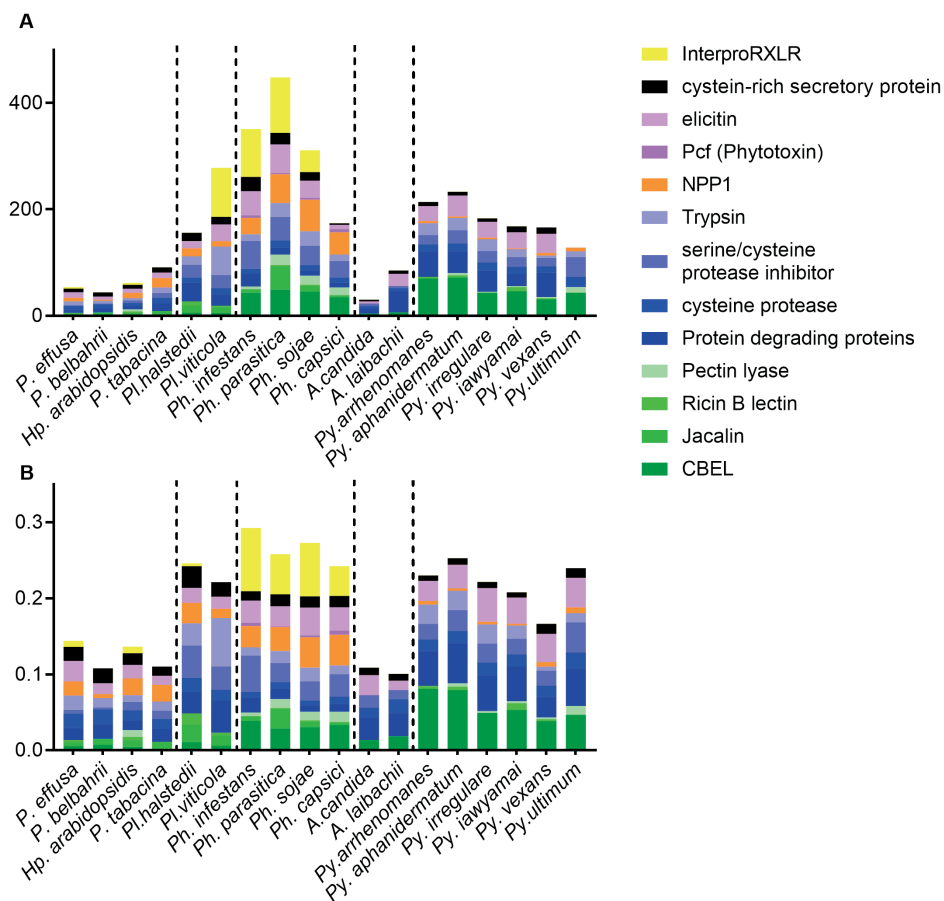
68. Ascunce MS, Huguet-Tapia JC, Ortiz-Urquiza A, Keyhani NO, Braun EL, Goss EM. 2017. Phylogenomic analysis supports multiple instances of polyphyly in the oomycete peronosporalean lineage. *Mol. Phylogenet. Evol.* 114:199-211
69. Blackman LM, Cullerne DP, Hardham AR. 2014. Bioinformatic characterisation of genes encoding cell wall degrading enzymes in the *Phytophthora parasitica* genome. *BMC Genomics* 15:785
70. Fellbrich G, Romanski A, Varet A, Blume B, Brunner F, et al. 2002. NPP1, a *Phytophthora*-associated trigger of plant defense in parsley and Arabidopsis. *Plant J.* 32:375-90
71. Seidl MF, Van den Ackerveken G. 2019. Activity and phylogenetics of the broadly occurring family of microbial Nep1-Like proteins. *Annu. Rev. Phytopathol.* 57
72. Ottmann C, Luberacki B, Kűfner I, Koch W, Brunner F, et al. 2009. A common toxin fold mediates microbial attack and plant defense. *Proc. Natl. Acad. Sci. U. S. A.* 106:10359-64
73. Cabral A, Oome S, Sander N, Kűfner I, Nűrnberger T, Van den Ackerveken G. 2012. Nontoxic Nep1-like proteins of the downy mildew pathogen *Hyaloperonospora arabidopsidis*: repression of necrosis-inducing activity by a surface-exposed region. *Mol. Plant-Microbe Interact.* 25:697-708
74. Mittl PR, Schneider-Brachert W. 2007. Sel1-like repeat proteins in signal transduction. *Cell. Signal.* 19:20-31
75. Larroque M, Barriot R, Bottin A, Barre A, Rougė P, et al. 2012. The unique architecture and function of cellulose-interacting proteins in oomycetes revealed by genomic and structural analyses. *BMC Genomics* 13:605
76. Hu G, Leger RJS. 2004. A phylogenomic approach to reconstructing the diversification of serine proteases in fungi. *J. Evol. Biol.* 17:1204-14
77. Jashni MK, Dols IHM, Iida Y, Boeren S, Beenen HG, et al. 2015. Synergistic action of a metalloprotease and a serine protease from *Fusarium oxysporum f. sp. Lycopersici* cleaves chitin-binding tomato chitinases, reduces their antifungal activity, and enhances fungal virulence. *Mol. Plant-Microbe Interact.* 28:996-1008
78. Adhikari BN, Hamilton JP, Zerillo MM, Tisserat N, Levesque CA, Buell CR. 2013. Comparative genomics reveals insight into virulence strategies of plant pathogenic oomycetes. *PLoS ONE* 8:e75072
79. McGowan J, Byrne KP, Fitzpatrick DA. 2018. Comparative analysis of oomycete genome evolution using the oomycete gene order browser (OGOB). *Genome Biol. Evol.* 11:189-206
80. Bohlin J, Eldholm V, Pettersson JH, Brynildsrud O, Snipen L. 2017. The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. *BMC Genomics* 18:151
81. Lamour K, Kamoun S. 2009. *Oomycete genetics and genomics: diversity, interactions and research tools*. John Wiley & Sons. 592 pp.
82. De Bustos A, Cuadrado A, Jouve N. 2016. Sequencing of long stretches of repetitive DNA. *Scientific reports* 6:36665

83. Fletcher K, Klosterman SJ, Derevnina L, Martin F, Bertier LD, *et al.* 2018. Comparative genomics of downy mildews reveals potential adaptations to biotrophy. *BMC Genomics* 19:851-84
84. Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF, Jiang RHY, *et al.* 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* 290:972-7
85. Raffaele S, Kamoun S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat. Rev. Microbiol.* 10:417-30
86. Seidl MF, van den Ackerveken G, Govers F, Snel B. 2012. Reconstruction of oomycete genome evolution identifies differences in evolutionary trajectories leading to present-day large gene families. *Genome Biol. Evol.* 4:199-211
87. McCarthy CGP, Fitzpatrick DA. 2017. Phylogenomic reconstruction of the oomycete phylogeny derived from 37 genomes. *mSphere* 2:e00095-17
88. Cui C, Herlihy J, Bombarely A, McDowell JM, Haak DC. 2019. Draft assembly of *Phytophthora capsici* from long-read sequencing uncovers complexity. *Mol. Plant-Microbe Interact.*
89. Sharma R, Xia X, Cano LM, Evangelisti E, Kemen E, *et al.* 2015. Genome analyses of the sunflower pathogen *Plasmopara halstedii* provide insights into effector evolution in downy mildews and *Phytophthora*. *BMC Genomics* 16:741
90. Latijnhouwers M, de Wit PJ, Govers F. 2003. Oomycetes and fungi: Similar weaponry to attack plants. *Trends Microbiol.* 11:462-9
91. Richards TA, Dacks JB, Jenkinson JM, Thornton CR, Talbot NJ. 2006. Evolution of filamentous plant pathogens: gene exchange across eukaryotic kingdoms. *Curr. Biol.* 16:1857-64
92. Richards TA, Soanes DM, Jones MDM, Vasieva O, Leonard G, *et al.* 2011. Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. *Proc. Natl. Acad. Sci. U. S. A.* 108:15258-63
93. Kemen E, Gardiner A, Schultz-Larsen T, Kemen AC, Balmuth AL, *et al.* 2011. Gene gain and loss during evolution of obligate parasitism in the white rust pathogen of *Arabidopsis thaliana*. *PLoS Biol.* 9:e1001094
94. Kemen E, Jones JDG. 2012. Obligate biotroph parasitism: Can we link genomes to lifestyles? *Trends Plant Sci.* 17:448-57
95. Cooper AJ, Latunde-Dada AO, Woods-Tör A, Lynn J, Lucas JA, *et al.* 2008. Basic compatibility of *Albugo candida* in *Arabidopsis thaliana* and *Brassica juncea* causes broad-spectrum suppression of innate immunity. *Mol. Plant-Microbe Interact.* 21:745-56

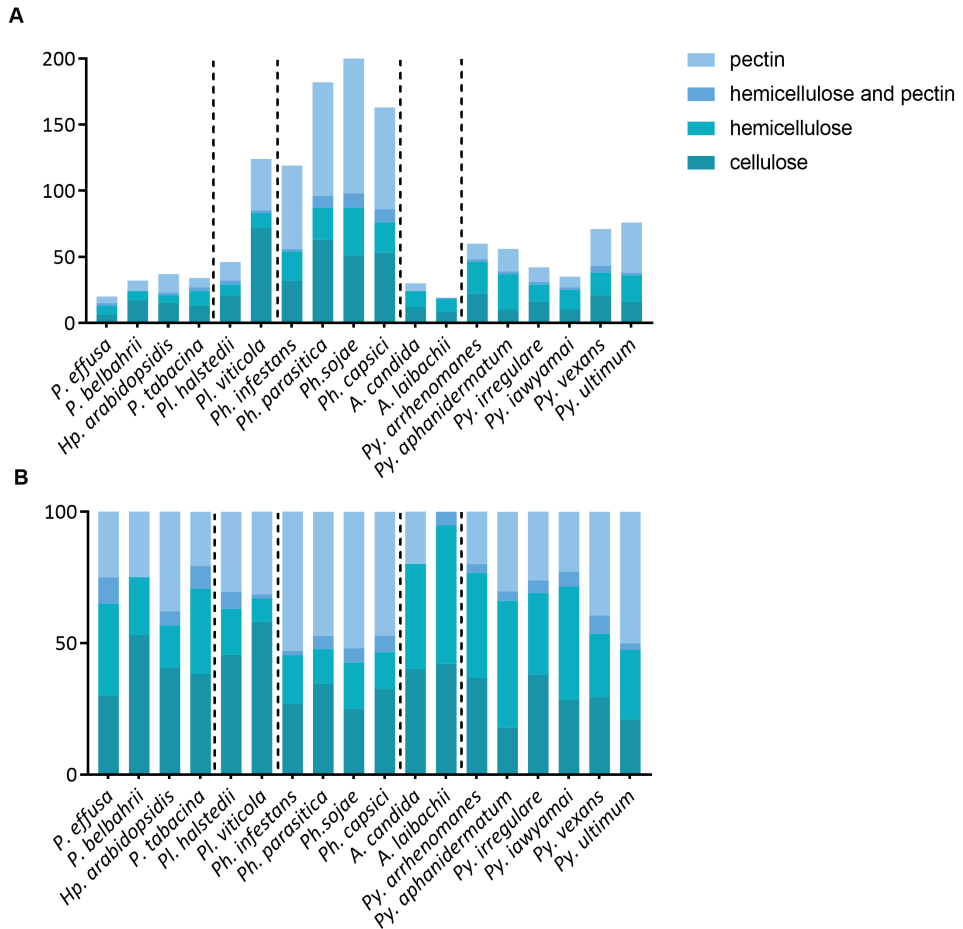
Supplemental data



Supplemental Figure 1. GC plot of various oomycete assemblies on contigs larger than 1kb.



Supplemental Figure 2. Occurrence of Pfam domains known to be involved in pathogenicity within the secretome of each species. **A.** shows the absolute number of Pfam domains, while **B.** shows the number relative to the total number of Pfam domains per species.



Supplemental Figure 3. A. Numbers of literature curated plant cell wall degrading (CA-Zymes) enzymes per species. **B.** The same data represented as fraction of the total number cell wall degrading protein domains per species.

Supplemental Table 1. Species used for comparative secretomics.

| Species | lifestyle | source | assembly | assembly size (mb) |
|---|---------------------|---------------------------------------|--------------------|--------------------|
| <i>Albugo laibachii</i> | obligate biotrophic | The Sainsbury laboratory | ENA1 | 37.0 |
| <i>Albugo candida</i> | obligate biotrophic | The Sainsbury laboratory | ASM107853v1 | 34.5 |
| <i>Pythium aphanidermatum</i> | necrotrophic | Michigan State University | AKXX000000000 | 35.9 |
| <i>Pythium arrhenomanes</i> | necrotrophic | Michigan State University | AKXY000000000 | 44.7 |
| <i>Pythium irregulare</i> | necrotrophic | Michigan State University | AKXZ000000000 | 42.9 |
| <i>Pythium iwayamai</i> | necrotrophic | Michigan State University | AKYA000000000 | 43.3 |
| <i>Pythium vexans</i> | necrotrophic | Michigan State University | AKYC000000000 | 33.9 |
| <i>Pythium ultimum</i> (var. <i>ultimum</i>) | necrotrophic | Michigan State University | DAOM BR144 | 42.8 |
| <i>Plasmopara viticola</i> | obligate biotrophic | INRA | INRA-PV221 isolate | 92.9 |
| <i>Plasmopara halstedii</i> | obligate biotrophic | Goethe-Universität, Frankfurt am Main | GCA_900000015.1 | 75.3 |
| <i>Peronospora belbahrii</i> | obligate biotrophic | Goethe-Universität, Frankfurt am Main | PRJEB15119 | 35.4 |
| <i>Hyaloperonospora arabidopsidis</i> | obligate biotrophic | Virginia Bio Informatics Institute | HyaAraEmoy2_2.0 | 81.6 |
| <i>Phytophthora infestans</i> | hemibiotrophic | Broad Institute | ASM14294v1 | 229.0 |
| <i>Phytophthora capsici</i> | hemibiotrophic | Joint Genome Institute, California | GCA_000325885.1 | 56.1 |
| <i>Phytophthora sojae</i> | hemibiotrophic | Joint Genome Institute, California | P. sojae V3.0 | 95.0 |
| <i>Phytophthora parasitica</i> | hemibiotrophic | Broad Institute | PP_INRA-310_V2 | 82.0 |
| <i>Peronospora effusa</i> (Race 1) | obligate biotrophic | Utrecht University | Hybrid7 | 32.4 |
| <i>Peronospora tabacina</i> (968-J2) | obligate biotrophic | University of California, Davis | PRJNA285243 | 63.1 |

Supplemental Table 2. Comparison of conserved eukaryotic genes for different oomycetes and the *Pfs1* assembly using BUSCO.

| Species | Complete | Single | Duplicated | Fragmented | Missing | #predicted genes | Assembly version |
|---------------------------------------|----------|--------|------------|------------|---------|------------------|------------------|
| <i>Phytophthora sojae</i> | 99.5% | 95.8% | 3.7% | 0.5% | 0.0% | 28,142 | V3.0 |
| <i>Phytophthora ramorum</i> | 98.1% | 94.4% | 3.7% | 0.5% | 1.4% | 15,955 | pr102_1 |
| <i>Hyaloperonospora arabidopsidis</i> | 96.8% | 92.1% | 4.7% | 0.5% | 2.7% | 14,606 | HyaAraEmoy2 |
| <i>Phytophthora infestans</i> | 95.9% | 92.6% | 3.3% | 0.9% | 3.2% | 25,638 | ASM14294v1 |
| <i>Plasmopara halstedii</i> | 95.8% | 95.8% | 0.0% | 0.5% | 3.7% | 15,469 | V2 |
| <i>Pseudoperonospora cubensis</i> | 93.1% | 91.2% | 1.9% | 1.4% | 5.5% | 27,591 | MSU-1 |
| <i>Peronospora tabacina</i> | 91.6% | 62.8% | 28.8% | 1.9% | 6.5% | 18,000 | 968-s26 |
| <i>Peronospora effusa</i> | 88.9% | 88.4% | 0.5% | 0.5% | 10.6% | 13,223 | H7 |
| <i>Saprolegnia parasitica</i> | 87.9% | 81.4% | 6.5% | 2.3% | 9.8% | 20,121 | cbs_223.65_2 |
| <i>Phytophthora capsici</i> | 31.6% | 31.6% | 0.0% | 0.5% | 67.9% | 19,805 | 11 |

Supplemental Table 3. Repeat elements identified in the *Pfs1* genome, for each repeat type the total numbers and percentage are shown. In addition, also a detailed annotation for each repeat element is provided.

| Repeat type | Count | Percentage | Length (bp) |
|-------------------|--------|------------|-------------|
| LTR | 9,247 | 70.65% | 6,532,069 |
| LINE | 562 | 4.29% | 201,127 |
| Simple repeat | 2,297 | 17.55% | 97,983 |
| DNA repeats/TE | 391 | 2.99% | 46,677 |
| Rolling Circle TE | 97 | 0.74% | 26,123 |
| Low complexity | 298 | 2.28% | 15,907 |
| tRNA | 132 | 1.01% | 9,413 |
| rRNA | 24 | 0.18% | 9,169 |
| SINE | 16 | 0.12% | 1,140 |
| Satellite | 12 | 0.09% | 964 |
| snRNA | 9 | 0.07% | 742 |
| Unknown | 4 | 0.03% | 272 |
| Total | 13,089 | 100.00% | 6,941,586 |

Supplemental Table 4. Genome sizes and repeat content of different assembled oomycete genomes

| Species | Assembly size (Mb) | Number of genes | GC content | Repeat content |
|---------------------------------------|--------------------|-----------------|------------|----------------|
| <i>Phytophthora infestans</i> | 240.00 | 17,797 | 50.96% | 74% |
| <i>Phytophthora sojae</i> | 95.00 | 19,027 | 54.60% | 39% |
| <i>Phytophthora ramorum</i> | 65.00 | 15,743 | 53.85% | 28% |
| <i>Peronospora tabacina</i> | 63.10 | 11,310 | 41.17% | 24% |
| <i>Hyaloperonospora arabidopsidis</i> | 100.00 | 14,321 | 47.21% | 43% |
| <i>Phytophthora capsici</i> | 64.00 | 19,805 | 50.42% | 19% |
| <i>Albugo laibachii</i> | 32.70 | 13,032 | 44.34% | 22% |
| <i>Pythium ultimum</i> | 42.80 | 15,290 | 48.27% | 7% |
| <i>Peronospora effusa</i> (Pfs1) | 30.48 | 8,927 | 48.25% | 21% |
| <i>Peronospora effusa</i> (Pfs13) | 32.10 | 8,607 | 47.63% | 33% |
| <i>Plasmopara viticola</i> | 101.30 | 17,014 | 45.00% | 26% |
| <i>Plasmopara halstedii</i> | 75.30 | 15,469 | 45.29% | 39% |
| <i>Albugo candida</i> | 34.50 | 15,824 | 43.00% | 17% |

Supplemental Table 5. Putative annotations of the *Pfs* proteins as obtained with ANNIE. In addition, the presence of a N-terminal signal peptide for secretion, WY motif, TM motif and overlap with a repeat region are listed for each protein coding gene.

This supplemental table can be found in the online repository located at:
<https://doi.org/10.17026/dans-zzg-mjgb>



Supplemental Table 6. Overview of the different effectors (RxLR and CRN) that were identified in the genome of *Pfs1*. Also, their respective functional domains and locations are listed per effector. Selected effectors that were used in the gene intergenic distance analysis are listed in the second tab.

| Gene name | Gene length (bp) | located on contig < 1kb | In repeat region (> 20%) | secreted (SignalP) | TargetP localization | WY-motif | #WY-domains | Trans membrane helix | #TM-helix | RxLR-motif | Motif | RxLR position | EER-motif | EER position |
|--------------------------|------------------|-------------------------|--------------------------|--------------------|----------------------|----------|-------------|----------------------|-----------|------------|-------|---------------|-----------|--------------|
| <i>pfs111739-00001</i> | 495 | NA | NA | secreted | S | no WY | 0 | Tm | 1 | AHLR | AXLR | 49 | DER | 106 |
| <i>pfs113210-00001</i> | 190 | NA | NA | secreted | S | no WY | 0 | no Tm | | ANLR | AXLR | 27 | EER | 116 |
| <i>pfs1112722-00003</i> | 265 | NA | NA | secreted | S | WY | 3 | no Tm | | DALR | DXLR | 36 | | |
| <i>pfs1113024-00001</i> | 407 | NA | NA | secreted | S | WY | 5 | no Tm | | DALR | DXLR | 36 | | |
| <i>pfs119526aa-00001</i> | 425 | NA | NA | secreted | S | WY | 2 | no Tm | | DALR | DXLR | 36 | | |
| <i>pfs118068-00001</i> | 510 | NA | NA | secreted | S | WY | 5 | no Tm | | DTLR | DXLR | 45 | | |
| <i>pfs111709-00001</i> | 1074 | NA | NA | secreted | C | WY | 11 | no Tm | | EFLR | EXLR | 83 | | |
| <i>pfs113069-00001</i> | 850 | NA | NA | secreted | S | WY | 7 | no Tm | | EGLR | EXLR | 94 | | |
| <i>pfs11842-00001</i> | 371 | NA | NA | secreted | S | WY | 3 | no Tm | | EKLR | EXLR | 85 | EDK | 91 |
| <i>pfs111034-00001</i> | 128 | NA | NA | secreted | S | no WY | 0 | Tm | 1 | FRLR | FXLR | 30 | EER | 55 |
| <i>pfs116706-00001</i> | 634 | NA | NA | secreted | S | no WY | 0 | no Tm | | GPLR | GXLR | 27 | EDR | 114 |
| <i>pfs115871-00001</i> | 159 | NA | NA | secreted | S | no WY | 0 | Tm | 1 | HHLR | HXLR | 40 | EER | 50 |
| <i>pfs1110165-00001</i> | 216 | NA | NA | secreted | S | no WY | 0 | no Tm | | IALR | IXLR | 55 | EDR | 108 |
| <i>pfs115773-00001</i> | 507 | NA | NA | secreted | S | WY | 5 | no Tm | | KFLR | KXLR | 41 | ENR | 76 |
| <i>pfs117229-00001</i> | 191 | NA | NA | secreted | S | no WY | 0 | no Tm | | KHLR | KXLR | 43 | EER | 56 |
| <i>pfs113188-00001</i> | 442 | NA | NA | secreted | S | WY | 2 | no Tm | | KNLR | KXLR | 43 | DDK | 53 |
| <i>pfs1110418-00001</i> | 206 | NA | NA | secreted | S | no WY | 0 | no Tm | | KTLR | KXLR | 68 | DEK | 76 |
| <i>pfs115909-00003</i> | 202 | NA | NA | secreted | S | WY | 1 | no Tm | | NALR | NXLR | 36 | | |
| <i>pfs119196-00001</i> | 738 | NA | NA | secreted | M | no WY | 0 | no Tm | | NGLR | NXLR | 41 | DER | 113 |
| <i>pfs118598-00001</i> | 640 | NA | NA | secreted | M | no WY | 0 | no Tm | | NQLR | NXLR | 21 | EER | 106 |
| <i>pfs111028-00001</i> | 129 | NA | NA | secreted | S | no WY | 0 | Tm | 1 | PRLR | PXLR | 30 | EER | 55 |
| <i>pfs111298-00001</i> | 350 | NA | NA | secreted | C | WY | 4 | no Tm | | PVLR | PXLR | 43 | | |
| <i>pfs117655-00001</i> | 127 | NA | NA | secreted | S | no WY | 0 | no Tm | | QALR | QXLR | 46 | EER | 58 |
| <i>pfs117659-00001</i> | 127 | NA | NA | secreted | S | no WY | 0 | no Tm | | QALR | QXLR | 46 | EER | 58 |
| <i>pfs112360-00001</i> | 158 | NA | NA | secreted | S | no WY | 0 | no Tm | | QMLR | QXLR | 55 | EER | 73 |
| <i>pfs117306-00001</i> | 498 | NA | NA | secreted | S | no WY | 0 | no Tm | | RQLR | RXLR | 29 | EDR | 52 |
| <i>pfs111601-00001</i> | 196 | NA | NA | secreted | S | no WY | 0 | no Tm | | RALE | RXLR | 89 | EDR | 94 |
| <i>pfs1110453-00001</i> | 327 | NA | NA | secreted | S | no WY | 0 | no Tm | | RGLK | RXLR | 30 | EER | 64 |
| <i>pfs112474-00001</i> | 136 | NA | NA | secreted | S | no WY | 0 | Tm | 1 | RYLK | RXLR | 49 | EER | 62 |
| <i>pfs1110123-00001</i> | 135 | NA | NA | secreted | S | no WY | 0 | no Tm | | RDLL | RXLR | 71 | EER | 75 |
| <i>pfs1110014-00001</i> | 86 | < 1kb | NA | secreted | S | no WY | 0 | no Tm | | RALR | RXLR | 23 | DOK | 49 |
| <i>pfs118000-00001</i> | 192 | NA | NA | secreted | S | no WY | 0 | no Tm | | RALR | RXLR | 40 | EER | 52 |
| <i>pfs113214-00001</i> | 363 | NA | NA | secreted | M | no WY | 0 | no Tm | | RALR | RXLR | 67 | | |
| <i>pfs118288a-00001</i> | 389 | NA | NA | secreted | S | WY | 4 | no Tm | | RFLR | RXLR | 49 | DDK | 73 |
| <i>pfs114980-00001</i> | 135 | NA | NA | secreted | S | no WY | 0 | no Tm | | RFLR | RXLR | 52 | EER | 64 |

Supplemental Table 6. continued.

| Gene name | Gene length (bp) | located on contig < 1kb | In repeat region (> 20%) | secreted (SignalP) | TargetP localization | WY -motif | #WY-domains | Trans membrane helix | #TM-helix | RxLR-motif | Motif | RxLR position | EER-motif | EER position |
|-------------------------|------------------|-------------------------|--------------------------|--------------------|----------------------|-----------|-------------|----------------------|-----------|------------|-------|---------------|-----------|--------------|
| <i>pfs110015-00001</i> | 252 | NA | NA | secreted | M | WY | 2 | no Tm | | RFLR | RXLR | 42 | EER | 53 |
| <i>pfs11524-00001</i> | 161 | NA | NA | secreted | M | WY | 1 | no Tm | | RFLR | RXLR | 42 | EER | 53 |
| <i>pfs118208a-00001</i> | 188 | NA | NA | secreted | S | WY | 1 | no Tm | | RFLR | RXLR | 57 | EER | 75 |
| <i>pfs112936-00001</i> | 495 | NA | inside repeat | secreted | | no WY | 0 | no Tm | | RFLR | RXLR | 113 | | |
| <i>pfs110814-00001</i> | 76 | NA | NA | secreted | S | no WY | 0 | Tm | 1 | RFLR | RXLR | 25 | | |
| <i>pfs116786-00001</i> | 492 | NA | | secreted | S | no WY | 0 | no Tm | | RGLR | RXLR | 94 | | |
| <i>pfs11792-00002</i> | 350 | NA | NA | secreted | S | WY | 3 | no Tm | | RHLR | RXLR | 32 | DER | 46 |
| <i>pfs11792-00002</i> | 179 | NA | NA | secreted | S | no WY | 0 | no Tm | | RHLR | RXLR | 35 | EER | 49 |
| <i>pfs114249-00001</i> | 240 | NA | NA | secreted | C | no WY | 0 | no Tm | | RHLR | RXLR | 43 | EER | 55 |
| <i>pfs11707-00001</i> | 334 | NA | NA | secreted | S | WY | 2 | no Tm | | RHLR | RXLR | 39 | EER | 53 |
| <i>pfs112732-00001</i> | 418 | NA | NA | secreted | S | WY | 3 | no Tm | | RHLR | RXLR | 50 | EER | 63 |
| <i>pfs119818-00001</i> | 339 | NA | NA | secreted | M | WY | 3 | no Tm | | RHLR | RXLR | 39 | EER | 53 |
| <i>pfs117215-00001</i> | 174 | NA | NA | secreted | S | no WY | 0 | no Tm | | RHLR | RXLR | 31 | | |
| <i>pfs112056-00001</i> | 422 | NA | NA | secreted | S | WY | 3 | no Tm | | RHLR | RXLR | 60 | | |
| <i>pfs112059-00001</i> | 395 | NA | NA | secreted | S | WY | 3 | no Tm | | RHLR | RXLR | 60 | | |
| <i>pfs112759-00001</i> | 308 | NA | NA | secreted | S | WY | 2 | no Tm | | RKLR | RXLR | 40 | DNK | 50 |
| <i>pfs11848-00001</i> | 220 | NA | NA | secreted | S | no WY | 0 | no Tm | | RKLR | RXLR | 47 | EER | 62 |
| <i>pfs115786-00001</i> | 266 | NA | NA | secreted | S | WY | 1 | no Tm | | RKLR | RXLR | 40 | EER | 53 |
| <i>pfs111520-00001</i> | 381 | NA | NA | secreted | S | no WY | 0 | no Tm | | RKLR | RXLR | 47 | | |
| <i>pfs113089-00001</i> | 492 | NA | NA | secreted | S | WY | 5 | no Tm | | RLLR | RXLR | 58 | EER | 71 |
| <i>pfs118804-00001</i> | 599 | NA | NA | secreted | M | no WY | 0 | no Tm | | RLLR | RXLR | 20 | EOR | 55 |
| <i>pfs119474-00002</i> | 323 | NA | NA | secreted | S | WY | 1 | no Tm | | RMLR | RXLR | 57 | EER | 75 |
| <i>pfs11509-00001</i> | 941 | NA | NA | secreted | M | WY | 8 | no Tm | | RMLR | RXLR | 42 | EER | 53 |
| <i>pfs110282-00001</i> | 99 | <1kb | inside repeat | secreted | M | no WY | 0 | no Tm | | RMLR | RXLR | 17 | | |
| <i>pfs1110648-00001</i> | 99 | <1kb | NA | secreted | M | no WY | 0 | no Tm | | RMLR | RXLR | 17 | | |
| <i>pfs113989-00001</i> | 506 | NA | NA | secreted | C | no WY | 0 | Tm | 1 | RMLR | RXLR | 55 | | |
| <i>pfs114650-00001</i> | 160 | NA | NA | secreted | S | no WY | 0 | Tm | 1 | RMLR | RXLR | 35 | | |
| <i>pfs116386-00001</i> | 127 | NA | NA | secreted | M | no WY | 0 | no Tm | | RMLR | RXLR | 99 | | |
| <i>pfs113683-00001</i> | 160 | NA | NA | secreted | S | no WY | 0 | no Tm | | RMLR | RXLR | 43 | | |
| <i>pfs113869-00001</i> | 190 | NA | NA | secreted | S | no WY | 0 | no Tm | | RMLR | RXLR | 43 | | |
| <i>pfs118634-00001</i> | 346 | NA | NA | secreted | S | WY | 2 | no Tm | | RMLR | RXLR | 48 | EER | 65 |
| <i>pfs118777-00001</i> | 499 | NA | NA | secreted | S | WY | 1 | no Tm | | RMLR | RXLR | 51 | EER | 68 |
| <i>pfs119985-00001</i> | 387 | NA | NA | secreted | S | WY | 3 | no Tm | | RMLR | RXLR | 38 | EER | 52 |
| <i>pfs117438-00001</i> | 255 | NA | NA | secreted | M | no WY | 0 | no Tm | | RMLR | RXLR | 51 | DER | 72 |
| <i>pfs110789-00001</i> | 123 | NA | NA | secreted | S | no WY | 0 | no Tm | | RMLR | RXLR | 52 | EER | 65 |
| <i>pfs113172-00001</i> | 165 | NA | NA | secreted | S | no WY | 0 | Tm | 1 | RMLR | RXLR | 40 | EER | 50 |
| <i>pfs114214-00001</i> | 215 | NA | NA | secreted | S | no WY | 0 | Tm | 1 | RMLR | RXLR | 41 | EER | 51 |
| <i>pfs116054-00001</i> | 125 | NA | NA | secreted | S | no WY | 0 | no Tm | | RMLR | RXLR | 52 | EER | 65 |

Supplemental Table 6. continued.

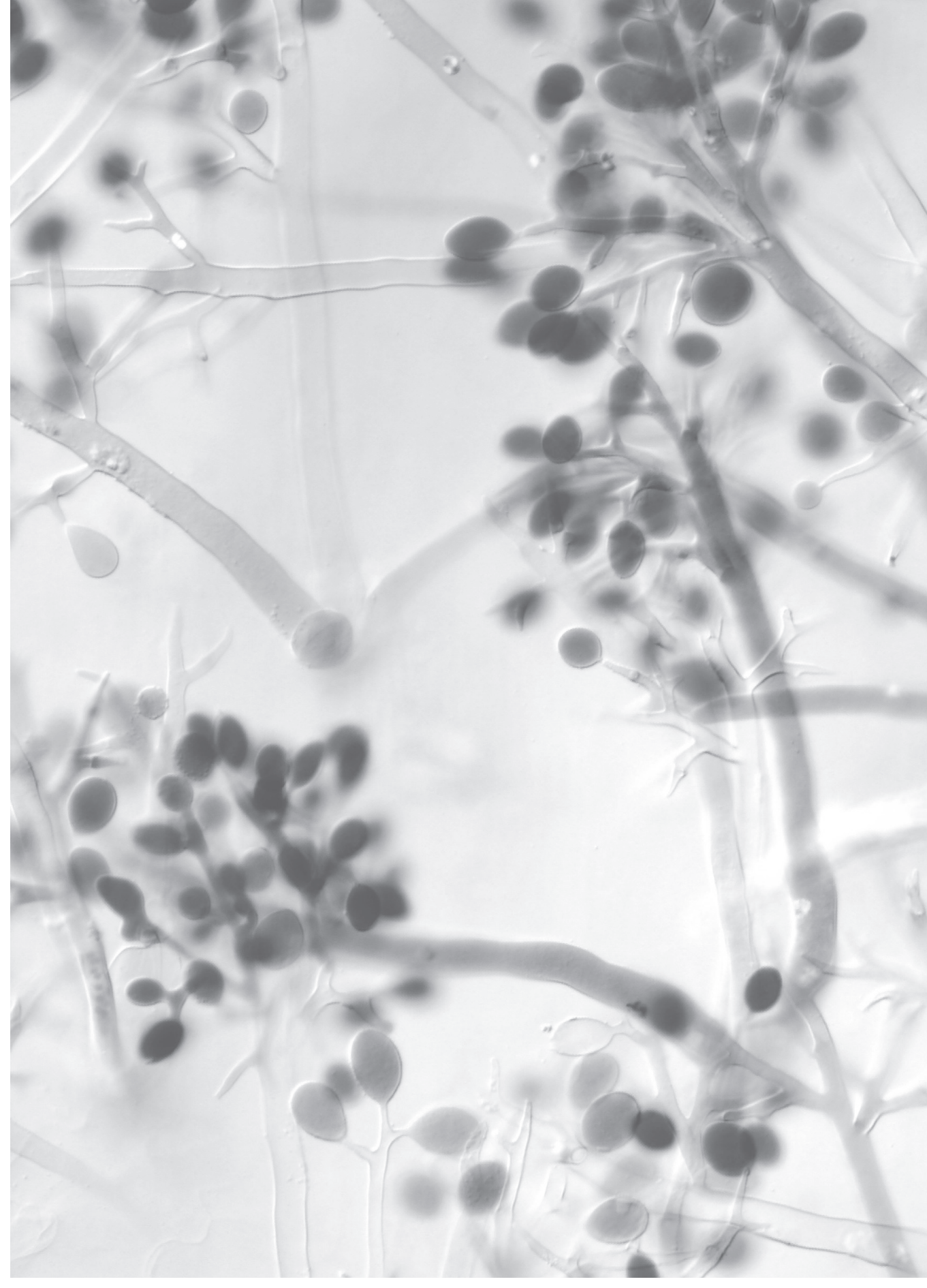
| Gene name | Gene length | located on contig < 1kb | In repeat region (> 20%) | secreted (SignalP) | TargetP localization | WY-motif domains | #WY-domains | Trans membrane helix | #TM-helix | RxLR-motif | Motif | RxLR position | EER-motif | EER position |
|-------------------------|-------------|-------------------------|--------------------------|--------------------|----------------------|------------------|-------------|----------------------|-----------|------------|-------|---------------|-----------|--------------|
| <i>pfs11192-00001</i> | 684 | NA | NA | secreted | S | WY | 6 | no Tm | | RRLR | RXLR | 50 | EER | 65 |
| <i>pfs119365-00001</i> | 377 | NA | NA | secreted | | WY | 3 | no Tm | | RRLR | RXLR | 39 | EER | 53 |
| <i>pfs114-00001</i> | 539 | NA | NA | secreted | S | WY | 4 | no Tm | | RRLR | RXLR | 39 | ENR | 100 |
| <i>pfs1117910-00001</i> | 113 | NA | NA | secreted | S | no WY | 0 | no Tm | | RRLR | RXLR | 106 | | |
| <i>pfs11556-00001</i> | 119 | NA | NA | secreted | S | no WY | 0 | no Tm | | RRLR | RXLR | 114 | | |
| <i>pfs113286-00001</i> | 873 | NA | NA | secreted | S | WY | 8 | no Tm | | RRLR | RXLR | 50 | | |
| <i>pfs111883-00001</i> | 177 | NA | NA | secreted | S | no WY | 0 | no Tm | | RSLR | RXLR | 44 | EER | 61 |
| <i>pfs111982-00001</i> | 232 | NA | NA | secreted | C | WY | 1 | no Tm | | RSLR | RXLR | 33 | EER | 54 |
| <i>pfs1113378-00001</i> | 522 | NA | NA | secreted | M | no WY | 0 | no Tm | | RSLR | RXLR | 36 | | |
| <i>pfs119324-00001</i> | 295 | NA | NA | secreted | M | no WY | 0 | no Tm | | RSLR | RXLR | 56 | | |
| <i>pfs111735-00001</i> | 765 | NA | NA | secreted | S | no WY | 0 | no Tm | | RTL | RXLR | 88 | | |
| <i>pfs11257-00001</i> | 1101 | NA | NA | secreted | S | no WY | 0 | no Tm | | RTL | RXLR | 70 | | |
| <i>pfs11439-00001</i> | 241 | NA | NA | secreted | S | no WY | 0 | no Tm | | RTL | RXLR | 41 | | |
| <i>pfs1123175-00001</i> | 1170 | NA | NA | secreted | S | no WY | 0 | no Tm | | RYLR | RXLR | 46 | DDK | 52 |
| <i>pfs114263-00001</i> | 353 | NA | NA | secreted | S | WY | 3 | no Tm | | RYLR | RXLR | 41 | DER | 55 |
| <i>pfs1110230-00001</i> | 165 | NA | NA | secreted | M | no WY | 0 | no Tm | | RYLR | RXLR | 60 | EDR | 77 |
| <i>pfs1110492-00001</i> | 699 | NA | NA | secreted | M | WY | 5 | no Tm | | RYLR | RXLR | 58 | EDR | 75 |
| <i>pfs118028-00001</i> | 755 | NA | NA | secreted | M | WY | 5 | no Tm | | RYLR | RXLR | 60 | EDR | 77 |
| <i>pfs111950-00001</i> | 366 | NA | NA | secreted | S | WY | 2 | no Tm | | RYLR | RXLR | 32 | | |
| <i>pfs11809-00001</i> | 681 | NA | NA | secreted | M | WY | 5 | no Tm | | RYLR | RXLR | 60 | | |
| <i>pfs11668-00001</i> | 367 | NA | NA | secreted | S | no WY | 0 | no Tm | | RKLS | RXLS | 46 | DER | 80 |
| <i>pfs112002-00001</i> | 164 | NA | NA | secreted | S | no WY | 0 | no Tm | | RKLT | RXLT | 39 | EER | 59 |
| <i>pfs1110382-00001</i> | 133 | NA | NA | secreted | S | no WY | 0 | no Tm | | SALR | SXLR | 27 | EER | 73 |
| <i>pfs111878-00001</i> | 291 | NA | NA | secreted | S | WY | 1 | no Tm | | SRLR | SXLR | 101 | | |
| <i>pfs1117975-00001</i> | 810 | NA | NA | secreted | S | WY | 2 | no Tm | | SYLR | SXLR | 38 | DDK | 52 |
| <i>pfs119033-00001</i> | 843 | NA | NA | secreted | | no WY | 0 | no Tm | | VRLR | VXLR | 73 | EEK | 118 |

Supplemental Table 7. Secreted CAZyme domains per species.

This supplemental table can be found in the online repository located at:
<https://doi.org/10.17026/dans-zzg-mjgb>

Supplemental Table 8. Secreted Pfam domains per species.

This supplemental table can be found in the online repository located at:
<https://doi.org/10.17026/dans-zzg-mjgb>



A grayscale micrograph showing a dense network of thin, branching hyphae. Several dark, oval-shaped spores are visible, some attached to the hyphae and others free. The background is light and slightly textured.

Chapter 3

Phylogenetic analysis of resistance-breaking isolates of the spinach downy mildew *Peronospora effusa*

Joël Klein¹, Guido Van den Ackerveken¹

**¹ Plant-Microbe Interactions, Department of Biology,
Utrecht University, Utrecht, The Netherlands**

Abstract

The downy mildew *Peronospora effusa* (*Pfs*) causes a destructive disease on cultivated spinach. There are currently eight resistance loci available to breed spinach lines that confer resistance to different *Pfs* races. However, the extensive use of these resistance loci in cultivars on a large scale has posed a strong selection pressure on the pathogen. Before 1990 there were only three races of the pathogen known. The use of resistance loci accelerated the appearance of resistance-breaking isolates, resulting in the denomination of 14 new races within the last 20 years. To study the genetic relationship between *Pfs* races and isolates, we sequenced and phylogenetically analyzed the genomes of 16 denominated *Pfs* races and 8 field isolates. This resulted in the identification of three distinct genetic *Pfs* groups and one additional admixed group. Furthermore, mitochondrial DNA analysis revealed two main haplotypes in the sampled isolates. The presence of admixed genotypes and closely related *Pfs* isolates possessing different mitochondrial haplotypes strongly suggests that sexual recombination has played a role in the evolution of recent *Pfs* isolates. Other isolates show clear signs of clonal expansion allowing these successful genotypes to proliferate asexually.

Introduction

Peronospora effusa (*Pfs*, previously known as *P. farinose* f. sp. *spinaciae*) is one of the most economically important diseases of cultivated spinach [1]. Spinach is often cultivated in high-density fields year-round, e.g. in California, creating ideal conditions for continuous disease development [2; 3]. Downy mildew infection can already start at the cotyledon stage and eventually results in discolored and distorted plant development. In some cases, the disease development is so severe that entire fields are abandoned all together [4].

The downy mildew of spinach was first identified and reported in Britain in 1824 [5]. More than a century later race 2 (*Pfs*2) was described in 1958 [2]. Before 1990 there were only three races of the pathogen known, and the disease could be well controlled [6]. After 1990, because of the use of resistance loci (R-loci) in spinach varieties, the number of identified races increased exponentially [7], and more than 14 new races were discovered within the last 20 years [4]. To date, there are eight disease R-loci known, and available for breeding resistance *Pfs* spinach cultivars [2; 8]. However, the use of resistant cultivars on a large scale in agriculture has posed a strong selection pressure on the pathogen to adapt and overcome the employed R-loci. This has led to the emergence of new races and severe outbreaks of downy mildew infections in spinach cultivars that were previously resistant to this disease [9].

In addition to the use of R-loci for resistance, there are also chemical crop protection methods available that are effective against *Pfs* [10]. However, only a few fungicides are effective against this pathogen and with severe inoculum pressure combined with favorable environmental conditions the use of these chemicals is only partially effective [10]. Besides, an increasing number of growers are switching to organic farming, which prohibits the use of synthetic fungicides [11]. These untreated fields often serve as a reservoir for the pathogen from which the disease can spread providing huge amounts of inoculum.

Pfs isolates are assigned to different races based on their virulence spectrum on a standardized set of differential spinach lines (the so-called differential set) developed by the International Seed Federation (ISF) [12]. These spinach differentials contain one, or a combination of two known spinach R-loci effective against different *Pfs* races. The differential set was improved by the inclusion of Near Isogenic Lines (NIL) containing single R-loci [13]. Also, cultivars that contain a combination of two R-loci were included for the accurate determination of recent races. (**Supplemental Table 1**) [12]. The International Working Group on *Peronospora* (IWGP) performs the determination of new races and annually evaluates newly

identified isolates that break previously known resistance traits [14]. When the disease response of novel *Pfs* isolates on the differential set show a consistent unique virulence spectrum a new race number can be assigned, and a chosen isolate denominated as the reference race [7].

A major factor that contributes to the rapid proliferation of this pathogen is its ability to spread spores over large areas in a short time through wind currents and dispersal by rain. Like many oomycetes, *Pfs* can reproduce both asexually (clonal propagation) and sexually. In the asexual cycle, which is most relevant for spread during the spinach growing season, the pathogen sporulates profusely on infected plant tissues releasing a vast amount of sporangiospores from emerging sporangiophores [15]. In general, limiting sexual reproduction is considered a common virulence strategy in oomycetes that allows the proliferation of clonal populations that are well adapted to their host [15-17].

Sexual reproduction occurs when two *Pfs* isolates of the opposite mating type (P1 and P2) grow in the same infected spinach tissue [18]. When hypha of opposite mating type meet, they develop antheridia and oogonia in which meiosis occurs followed by fertilization and formation of a single diploid oospore. These thick-walled spores can survive for prolonged periods in soil and thus serve as survival structures for the pathogen [19]. Sexual reproduction results in new combinations of parental chromosomes and recombination in the diploid *Pfs* genome, resulting in a unique mix of virulence and effector genes. This enlarged genotypic diversity aids the emergence of new *Pfs* isolates over time and allows the pathogen in adapting to selection pressure from the host [17]. Once a new successful genotype has emerged it can then reproduce asexually leading to rapid population expansion. The occurrence of sexual and clonal propagation events in historic isolates of the same species can be assessed using phylogenetic methods. This has been done previously for several plant pathogenic fungi and oomycetes. For example, the study on the phylogenetic relationship between 50 *Magnaporthe oryzae* isolates based on whole-genome sequencing, resulted in the identification of six distinct lineages within this rice blast species. The major lineage containing *M. oryzae* isolates, endemic to continental Southeast Asia, showed signatures of sexual recombination while other lineages were found to be largely clonal [20]. A similar result was found in the potato late blight pathogen *P. infestans* where the genotypes of 652 isolates collected from various regions in the Netherlands were found to be clustered in three genetic groups. The first cluster consists of a single clonal lineage in which isolates all contain the Ia mitochondrial haplotype. The second and third clusters have a more complex substructure containing many unique genotypes and genetic variation which indicates that sexual reproduction has occurred in these lineages. In the third cluster, several distinct clonal lineages were also identified [21].

A genomic population study on the relationship between several *Pfs* isolates and races has previously been published in which they used SNP markers. This revealed that historical *Pfs* isolates and races showed greater genotypic diversity than recent field isolates. Although most of the genotypic variation found within recent field isolates were produced during asexual development, the overall genetic diversity in historical races and isolates gathered from different geographical locations was likely caused by sexual recombination [7].

Also, the mitochondrial genes *cox2* and *nad1* in *Pfs* have been utilized to shed light on the relationship between different *Pfs* isolates. These mitochondrial genes were sequenced in 33 *Pfs* isolates derived from different locations in Asia, Europe and America. This revealed two distinct mitochondrial haplotype groups in *Pfs* that could be linked to their geographical origin (Group I for Asia and Oceania and Group II for American, European, and two Japanese samples) [22].

Here we describe the phylogenetic analysis of 16 denominated races and eight recent field isolates of *Pfs* based on whole-genome sequencing data. The sequencing reads of different isolates were compared with the assembled nuclear and mitochondrial genome of *Pfs1*. We assess the genetic variation between races and isolates to determine their phylogenetic relationship and predict the possible involvement of sexual recombination in their evolution.

3

Materials and methods

Pfs isolates and DNA isolation

Spore material for 16 denominated *Pfs* races (1-16) and eight field isolates were obtained from two different Dutch breeding companies. These races are designated by the IWGP (International Working Group on *Peronospora* in spinach) and maintained by the Dutch organization Naktuinbouw (the Netherlands Inspection Service for Horticulture). The exact geographical origin is not known in many cases, especially for older races. The virulence spectrum of the field isolates was determined on the current standard set of international spinach differentials (**Supplemental Table 1**). *Pfs* sporangiospores were harvested from infected spinach plants (approximately 3 days after the start of sporulation) using a soft brush and 10 ml of tap water to brush off and collect the spores. The resulting spore suspensions were filtered through Miracloth to remove debris and larger particles, collected in a 50 ml Greiner tube, and centrifuged at 4,000 RPM (1,878 RCF) for 20 min in a table-top centrifuge. The cleared supernatant was removed until in total 1 ml of pellet and liquid remained. The pellet was

then re-suspended and transferred to an Eppendorf tube and centrifuged for 10 minutes at 20200 RCF after which the supernatant was discarded. The resulting spore pellet was flash-frozen in liquid nitrogen and the remaining moisture was removed by freeze-drying. Dried spores were stored at -80°C . Spore material of the downy mildew of beetroot (*P. farinosa* f. *sp. betae*) was isolated in the same way. For the isolation of genomic DNA, freeze-dried spores were disrupted using the Qiagen tissue lyser II using glass beads at a frequency of 30/s twice for 1 minute. DNA was extracted from the pulverized spores with the Qiagen DNeasy Plant Kit using the manufacturers' standard protocol. The DNA was quantified with the Qubit 2.0 fluorometer using the Qubit dsDNA HS (High Sensitivity) assay kit (Life Technologies).

DNA sequencing

Sequencing libraries were prepared from 1 μg of total genomic DNA per race/isolate. DNA was mechanically sheared in the Covaris M220 ultrasonicator (Covaris, Inc., Woburn, MA) into 550 bp fragments. Paired-end libraries for sequencing were prepared using the Illumina NeoPrep liquid handling robot. Library quality assessments and quantification were verified using the 2100 Bioanalyzer using high sensitivity DNA chip (Agilent Genomics) and the Qubit 2.0 fluorometer using the Qubit dsDNA HS kit (Life Technologies). The resulting libraries were paired-end sequenced (2 x 150 bp) using the Illumina NextSeq500 machine in high output mode.

Mitochondrial genome assembly

The mitochondrial (mt) genome of *Pfs1* was assembled with MitoMaker version 1.14 [23] using genomic Illumina reads and the mitochondrial (mt) sequence of *Peronospora tabacina* 968-S-25 as a backbone. Error-corrected PacBio reads were aligned to the incomplete mt-contigs and PacBio reads that had an overlap with these contigs were extracted. The selected PacBio reads together with the partially assembled mt-contigs were assembled using Newbler version 2.7 [24] using default parameters. The resulting mt-contig was used as an improved backbone in MitoMaker. Gene models for the *Pfs* mt-genome were automatically annotated using GeSeq [25], based on sequence similarity with genes located on the backbone sequence (*P. tabacina* 968-S-25) and common mt-genes located in the database of the GeSeq program. Mitochondrial tRNAs were identified with tRNAscan-SE 2.0 [26] using the "mitochondrial other" database. Repeat regions in the mt-genome were identified using the repeat finder in Geneious Prime version 2019.2 [27]. The identified gene models located on the mt-genome gene were visually inspected using Geneious Prime. A homology search to the NCBI RefSeq database (July 2019) was performed with the mt-sequences of ORFs lacking a functional annotation. The

sequence annotations on the mt-genome of *Pfs* were compared to those present in the mt-genome of *P. tabacina* [28]. An alignment visualization between the mt-sequences of *Pfs1* and *P. tabacina* was created with Easyfig version 2.2.2 [29]. The circular mt-genome was visualized using OGDRAW version 1.3.1 [30].

Mitochondrial genome phylogeny

A subset of 80 million genomic reads was randomly selected for each isolate and aligned to the assembled mt-genome of *Pfs1* with BWA version 0.7.17 using default settings. mt-variants were called using the GATK haplotype variant caller with default settings. The coverage at each base position was recorded and the mean coverage was calculated over the entire mt-genome for each isolate. A minimal coverage value was determined by multiplying the mean coverage per isolate with 0.2 (20% of mean read coverage). Variants called in regions where the base coverage was below this minimal coverage value were removed with Vcfintersect (part of vcflib [31]). GATK FastaAlternateReferenceMaker was used to create isolate specific mt-sequences by incorporation of the variants identified between the reference (*Pfs1*) and each isolate. The resulting isolate-specific mt-sequences were aligned using MUSCLE [27] with default settings. An mt-genome tree was constructed using the Geneious Prime tree builder with Tamura-Nei [32] as a distance model and UPGMA [33] for tree building. The tree was visualized in an unrooted manner with ITOL [34]. Haplogroup names for the mt-genomes were inferred from Choi *et al.* [22]. Variants in the mt-genome that are shared between all isolates of haplogroup I were combined using vcf-isec part of VCFtools [35]. Bedtools intersect [36] was used to remove variants in the combined VCF file that are also present in one of the *Pfs* isolates belonging to mt-haplogroup II. Variants that are unique to the mt-genome of haplogroup I were annotated based on the GFF file containing the location of the coding sequences in the mt-genome of *Pfs1* using Bedtools intersect.

Read mapping to the *Pfs1* genome and quality filtering

The sequencing reads for each isolate were aligned to the *Pfs1* reference genome using BWA mem version 0.7.13 using default settings [37]. To optimize BWA runtime, 40 million paired-end reads (80 million reads in total) for each isolate were randomly extracted from the total read set. The resulting SAM file was converted to BAM format using Samtools view version 1.3.1 [38], only reads with a mapping quality (MAPQ) higher than 45 were retained, and unpaired reads were removed using the is-Paired option. The hard clipped (H-tag) and soft clipped (S-tag) were used to filter out clipped reads that only partially align to the reference sequence. Picard MarkDuplicates version 2.10.5 [39] with settings REMOVE_DUPLI-

CATES=true was used to remove secondary read alignments with a lower base-quality scores. Samtools AddOrReplaceReadGroups option was used to add the sample name (*Pfs* race or isolate name) to each of the BAM alignments. Samtools fixmate was used to update and correct the insert size as well as the mate related flags that were lost after filtering. After filtering the remaining number of mapped reads was recorded, and the number of covered bases compared to the reference sequence for each isolate was calculated using Samtools mpileup.

Variant calling

SNP and INDEL variants were identified for each isolate using GATK HaplotypeCaller version 3.8 [40] with settings `-ploidy 2 -minReadsPerAlignStart 7` with gVCF using a masked *Pfs1* reference FASTA file (details on the genome masking can be found in **Chapter 2**). The individual isolate gVCF files were merged using GATK CombineGVCFs and GenotypeGVCFs with standard settings to obtain a VCF file containing variant and genotype data for each variant in all isolates. Called variants at positions where one of the other isolates did not have read coverage were removed from the data set. Variants were further quality filtered with GATK VariantFiltration with filtering options: “QD < 20.0 || FS > 60.0 || MQ < 40.0 || HaplotypeScore > 13.0 || MappingQualityRankSum < -12.5 || MQRankSum < -12.5 || ReadPosRankSum < -8.0” [40]. Variants that did not pass these filtering criteria were discarded. The resulting VCF file was split per *Pfs* isolate and analyzed with BCFtools [41] using the stats options to count the number and nature of the identified variants per isolate.

Nuclear genome phylogeny

For the phylogenetic analysis of the isolates, the previously obtained VCF files were further filtered. INDELS longer than 1 base pair were removed. Positions, where a reliable variant call could not be made, were filtered using GATK SelectVariants program with settings `-maxNOCALL-number 0 -selectType SNP`. Gdsfmt [42] was used to convert the isolates VCFs into a genofile using snpgds VCF2GDS. A PCA was done based on the genofile with snpgdsPCA with a missing rate set to 0.10, maf to 0.05, otherwise standard settings were used. PGDspider2 [43] was used to convert the combined VCF containing all variants from all isolates understudy to PHYLIP, STRUCTURE and NEXUS format for further use.

A 500-times bootstrapped phylogenetic tree was constructed with the previously obtained PHYLIP formatted file using RAxML [44] within the raxmlHPC-HYBRID program with settings `-f a -m GTRGAMMA -N 500 -x 12345 -p 54321`. First, a tree was constructed of the *Pfs* isolates together with *ES-15* (*Peronospora farinosa* f. sp. *betae*) as an outgroup. A second tree without *ES-15* was generated to produce a dendrogram that

was visualized in ITOL. The PHYLIP formatted file was also used to create a Neighbor-Net of phylogenetic multifurcations in SplitsTree4 [45] version 4.14.4 using the NJ with uncorrected-P clustering method. The presence of recombination was statistically verified using the pairwise homoplasy test (PHI) implemented in SplitsTree v.4 using a window size of 100 with $k = 71$ [46]. Genetic groups within the sequenced isolates were analyzed in STRUCTURE [47] version 2.3.2 that was run at K-values ranging from 1 till 7 and 6 runs with 500 burn-ins and 1,000 repeats assuming an admixed model. The STRUCTURE output was imported in the R package PopHelper [48] version 2.2.6 to visualize the result.

Results

A selection of spinach downy mildew isolates was made consisting of 16 denominated *Pfs* races and 8 *Pfs* field isolates. These isolates were collected over the last 20 years from spinach fields in the USA and Europe (**Supplemental Table 1**). The denominated races and isolates have each broken one or a combination of resistance genes residing in R-loci that have been introduced into commercial spinach cultivars over the last decades. Genomic DNA was isolated from harvested sporangiospores and used for paired-end sequencing. Details on the number of obtained reads for each isolate and other NGS statistics can be found in **Supplemental Table 2**. The reads from each isolate were mapped to the assembled nuclear and mitochondrial (mt) genome of reference race *Pfs1* and used to identify variants present in the *Pfs* isolates. Before describing the analysis of nuclear polymorphisms, we first detail the comparison of mitochondrial sequences.

Mitochondrial genome assembly

Polymorphisms in mt-DNA sequences and genes are useful to study phylogenetic relationships between and within species [49; 50]. To study the mt-phylogeny of *Pfs* we first generated a reference mt-genome of race *Pfs1*. Illumina and error-corrected PacBio reads were assembled on the mt-genome of the related downy mildew *Peronospora tabacina* (968-S-26). The *Pfs1* assembly resulted in a complete circular mt-genome with a size of 40.801 bp. To assess whether the mt-genome was correctly scaffolded we aligned the Illumina and PacBio reads to the assembled contig (**Supplemental Figure 1**). Although we see that the error-corrected PacBio reads cover the entire *Pfs* mt-genome which indicates correct scaffolding, we observe two distinct regions with coverage gaps in the Illumina read alignment. These regions correspond with two AT-rich repeats of approximately 600 bp (**Supplemental Figure 1**). In general, the GC-poor regions in the

genome are problematic to sequence with the Illumina sequencing technologies [51] which explains the lack of Illumina read coverage for these regions.

The *Pfs* mt-genome contains 76 features, consisting of 42 genes of which three are undefined ORFs. Other features are two repeat regions with a low GC-content, 25 tRNAs, one small subunit ribosomal RNA (SSU rRNA), ribosomal RNA subunits *rnl* and *rns*, and one *rns* and three *rnl* fragments (**Figure 1**). The three undefined mt-ORFs were queried against the NCBI nr database (July 2019), but no functional annotation for these ORFs could be established.

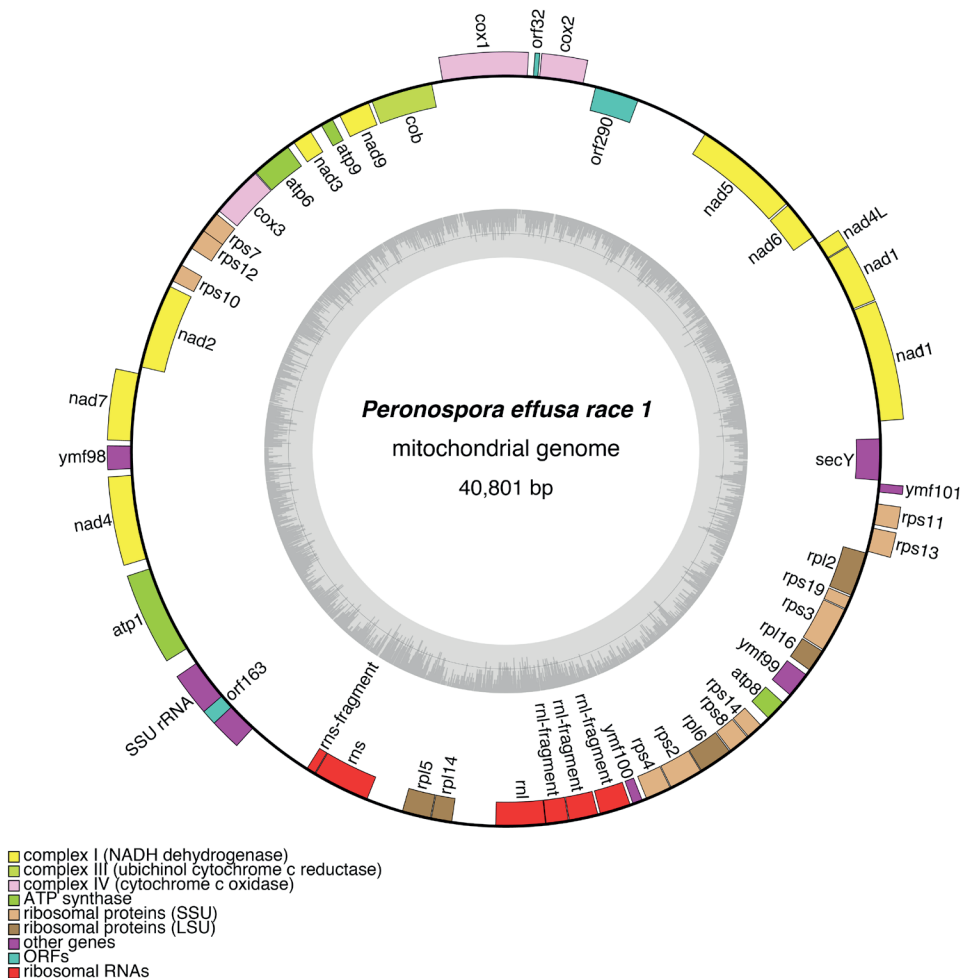


Figure 1. Annotated mitochondrial genome of *P. effusa* reference race *Pfs1*. The inner-circle depicts the GC-content, the grey arrows indicate the predicted origin of replication. The tRNA sequences present in the mt-genome of *Pfs1* are not depicted in this figure.

The mt-genome of *P. tabacina* (968-S-26) [28], that was used as a backbone in the assembly has 94.14% sequence identity to the mt-genome of *Pfs*. The overall gene order between both mt-genomes is conserved, and they share 62 features. However, the mt-genome of *P. tabacina* contains 5 ORFs and one repeat region that are not present on the mt-genome of *Pfs*. Likewise, the mt-genome of *Pfs* contains two repeats, and ribosomal RNA sequences *rnl* (including three additional fragments) and *rns* (including one additional fragment) that are not present on the mt-genome of *P. tabacina*. The most notable difference is a single large inverted region in the mt-genome of *Pfs* compared to *P. tabacina* (**Supplemental Figure 2**).

Mitochondrial genome phylogeny of *Pfs* isolates

Sequencing reads derived from the different *Pfs* isolates were aligned to the assembled mt-genome of *Pfs1* and used to identify SNP and INDEL polymorphisms. In general, the *Pfs* mt-genomes for all isolates were well covered by reads (**Supplemental Table 3**). However, as mentioned before, two repeat regions low in GC-content were poorly covered by Illumina reads. Variants in a region with low coverage (20% of mean read coverage) were therefore filtered out to reduce the risk of incorporating false-positives. The high-confidence variants identified between the *Pfs1* mt-reference genome and the different *Pfs* isolates under study were used to generate isolate-specific mt-sequences. The resulting mt-sequences were aligned and used to construct a phylogenetic tree (**Figure 2A**) that revealed two distinct haplogroups.

The nucleotide variants in the mt-genes *nad1* and *cox2*, previously described by Choi *et al.* [22], could be matched to the two haplogroups (**Figure 2B, Supplemental Table 3**). Comparing all variants in the full mt-sequences of the *Pfs* isolates under study, two distinctive barcode-like patterns clearly emerge that match to the two haplogroups (**Figure 2C**). The mt-genome of all isolates that possess mt-haplogroup I contain 35 variants unique variants that are not present in the mt-genomes of isolates in haplogroup II. Of these unique variants in haplogroup I, 16 were present in coding sequences, 7 in putative ORFs, 3 in the SSU rRNA, and 3 in tRNA features of which tRNAscan [26] could not determine the anticodon. Only six variants were located outside coding regions (**Supplemental Table 4**).

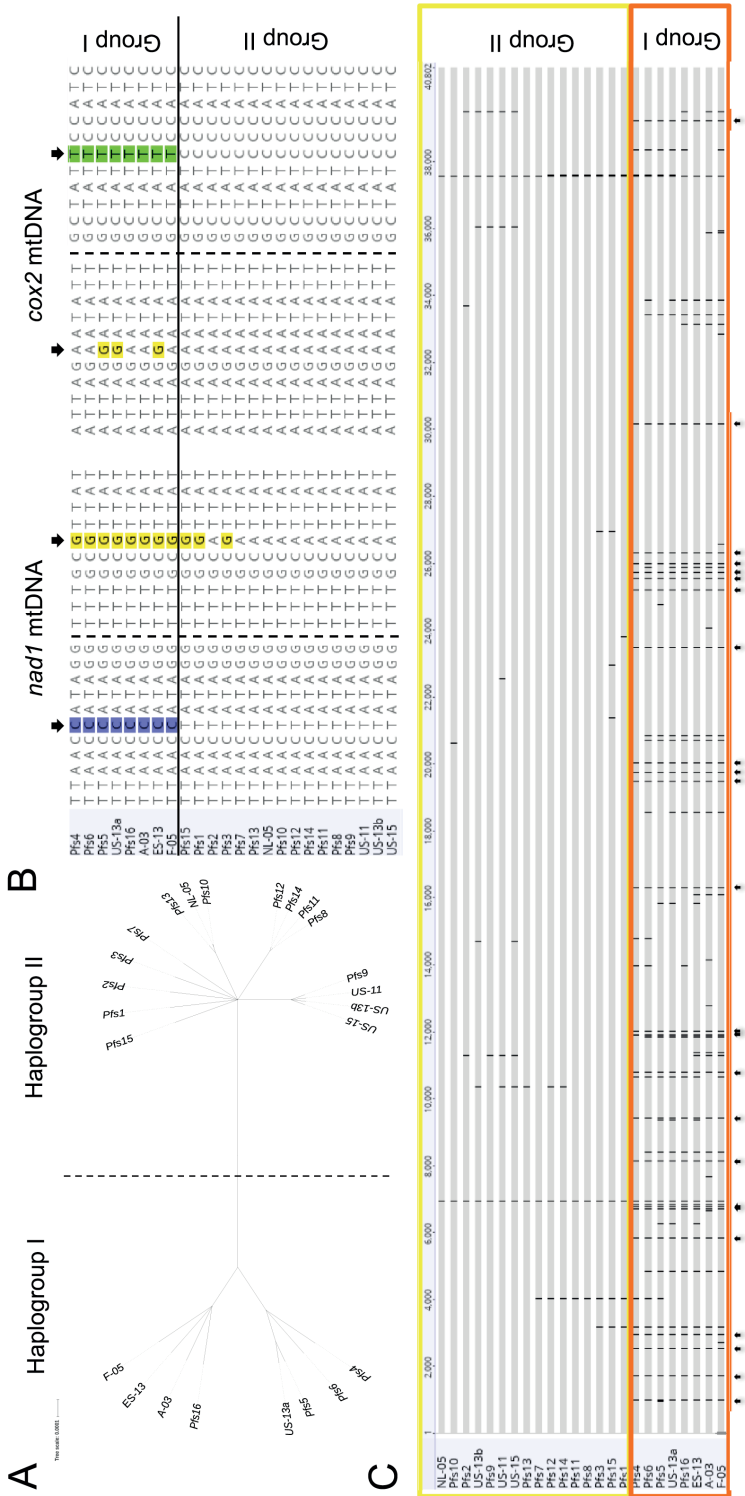


Figure 2 Phylogenetic comparison of mt-variants between *PfIs* isolates. **A**. Unrooted phylogenetic tree based on the mt-sequences of the different *PfIs* isolates. The tree was constructed with a Jukes Cantor substitution matrix and UP-GMA clustering. The mt-phylogenetic tree depicts two clear haplogroups (I and II). **B**. Partial sequence alignment of the *nad1* and *cox2* gene in the mitochondrial genome of *PfIs*; black arrows indicates the sites of substitution as previously found by Choi *et al.* [22], colors are assigned based on the nucleotides that differ from the consensus. **C**. Alignment of *PfIs* mt sequences as previously compared to the *PfIs1* reference are indicated with black lines, variants that are unique to haplogroup I are indicated with black arrows. From the partial and full alignment (**B**, **C**) the difference between the two-mitochondrial haplogroups I (orange box) and II (yellow box) are clearly visible.

Nuclear genome phylogeny

The phylogeny of *Pfs* isolates and races was studied based on whole-genome polymorphisms. Variants between the nuclear genome of the isolates and the reference genome were identified simultaneously using a joint variant calling approach. This allowed for the filtering of variants that were lacking read coverage at the same position in other isolates.

In total 398,645 heterozygous and homozygous variants were identified of which 139,575 remained after removing variants with a low quality score. Only single-nucleotide INDELS and SNPs that were reliably genotyped in all *Pfs* isolates were maintained, resulting in 103,720 homozygous and heterozygous variants, both shared and unique, that were represented in an SNP-matrix that was used for further phylogenetic analysis. The similarities and differences between *Pfs* isolates were analyzed using a principal component analysis (PCA) on the SNP matrix which visualizes the genetic relatedness between *Pfs* isolates (Figure 3).

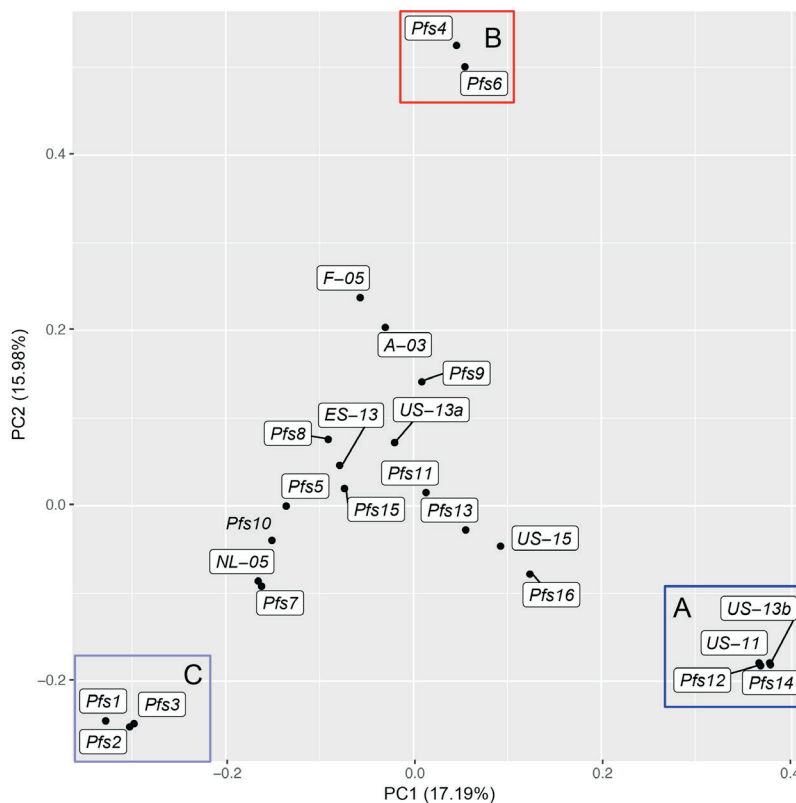


Figure 3. Principle component analysis (PCA) based on the isolate SNP matrix visualizing the genetic relatedness between the isolates. Isolates that are genetically close are grouped together in the PCA. Distinct groups are indicated by colored squares and numbered (A, B and C).

Three distinct groups, consisting of isolates *Pfs1*, 2, and 3 (A); *Pfs4*, and 6 (B) and *US-13b*, *US-11*, *Pfs12*, and 14 (C), could be distinguished (**Figure 3**). This implies that isolates in these groups have similar genetic content and are likely clonal variants. The remaining isolates cluster in the middle of the PCA, suggesting that they have a mixed genetic make-up, possibly because of sexual recombination between parents belonging to the three distinct groups (A, B, and C).

To further examine to what extent sexual recombination has contributed to the genetic content in these isolates, we analyzed our data using STRUCTURE. This analysis makes use of allele frequency information and performs a Bayesian clustering which assigns isolates to a pre-defined number of populations (K). Detection of the best number of clusters (K) in a sample of individuals is difficult when patterns of dispersal among populations are not homogeneous [52], as is likely the case for our sampled *Pfs* isolates. The observed grouping in the PCA plot (**Figure 3**) provided support for 3 distinct genetic groups (A, B and C) and 1 possible admixed group. Based on this we performed a STRUCTURE analysis assuming three clusters (K = 3) (**Figure 4**).

The same three groups (A, B and C) observed in the PCA plot (**Figure 3**) could also be distinguished in **Figure 4**. Other isolates that reside in the middle of the PCA show signs of mixtures between groups A and B (for 3 isolates) between all three groups (the 13 remaining isolates). This genetic mixture between the different groups could have occurred through sexual recombination. For example, we see that *Pfs7* and *NL-05* appear to be a genetic combination of variants found in group A and B. This could be an indication that *Pfs7* is the result of sexual reproduction between an isolate of group A and one of group B. We also see genetic combinations of group A, B, and C in certain isolates. For example, the STRUCTURE analysis shows that *Pfs13* appears to be a mix of the 3 defined groups, as shown by the equal proportions of the three colors. This pattern could have emerged from multiple sexual recombination events between isolates of all three defined genetic groups. However, after a new isolate has emerged through sexual reproduction, its subsequent asexual progeny could further evolve to new isolates and races that appear mixed in our analysis.

Phylogenetic analysis

A phylogenetic tree provides, in contrast to a PCA or STRUCTURE analysis, information on the possible evolutionary relatedness between the individual isolates. To analyze this, the same SNP matrix was also used to generate a maximum likelihood tree of all the sequenced isolates (**Figure 5**).

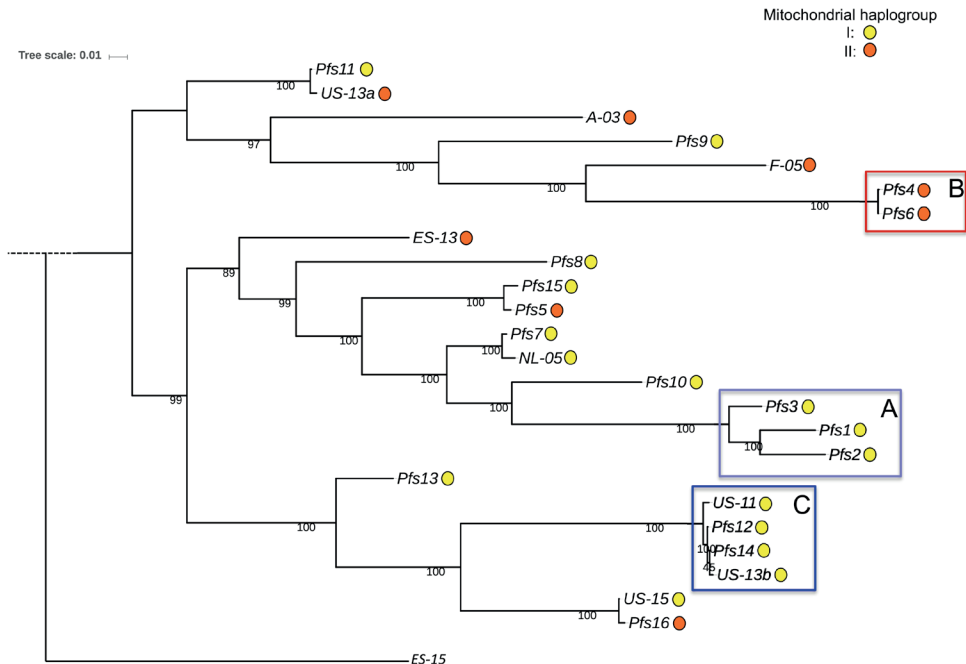


Figure 5. Maximum likelihood phylogenetic tree constructed on variants identified in *Pfs* isolates based on sequencing read alignment to the *Pfs1* reference genome. The tree was constructed with 500 bootstrap replicates to determine the likelihood of each split. The root was determined using the *P. farinosa f. sp. betae* isolate *ES-15*, which is connected to the tree with a dashed line to indicate the longer branch length. The colored circle indicates the mitochondrial genome haplogroup (I: yellow, II: orange). The colored boxes indicate distinct groups that were observed in the PCA (**Figure 3**) and STRUCTURE plot (**Figure 4**).

Most branches of the tree are well resolved, and clear groups (BS-score > 80) can be observed (**Figure 5**). The races are divided over three main clades, where the bottom one contains group C and the higher race numbers (the more recent races *Pfs12* to *Pfs16*), the middle clade contains group A (*Pfs1*, 2 and 3) but also many potentially admixed isolates, and the top one group B and a mix of relatively old and more recent isolates. An interesting observation is that *Pfs5* (1996) and *Pfs15* (2012) belong to the same monophyletic group, which was not expected as *Pfs15* was

isolated much later. Interestingly, they also possess different mitochondrial haplogroups. Races belonging to one monophyletic clade but of different mt-haplogroup are suggested to be the result of sexual recombination.

Split network analysis

The genetic make-up of many *Pfs* isolates is suggested to be admixed, hinting to sexual recombination, as we have observed in the STRUCTURE and PCA plot (**Figure 4**). A Neighbor-Net network can illustrate the evolutionary relationships among the different isolates in the presence of sexual recombination. Possible events of sexual recombination can be inferred from variants that possess conflicting phylogenetic signals in the distance matrix (SNP matrix). Potential sources of phylogenetic conflict among *Pfs* isolates may arise but are not limited to, hybridization, recombination, and horizontal gene transfer [53]. In the Neighbor-Net these conflicting signals are represented by reticulation patterns resulting in parallel lines.

The Neighbor-net phylogenetic network (**Figure 6**) shows clear evidence of phylogenetic conflicts indicated by reticulations in the network between the isolates. A reticulate pattern (lines that connect clades) with parallel lines indicates possible sexual recombination events between different *Pfs* genotypes. Isolates located close to the center of the net suggest that they have more recently emerged from sexual recombination, e.g. *Pfs11*, *13*, *16*, and *US-13a*. While the identified *Pfs* groups (A, B and C) are connected to the main 'net' with relatively long branches, suggesting that they have accumulated variants in the absence of sexual reproduction and are likely clonal variants of each other.

In the split network, *Pfs5* and *Pfs15* are located at a short distance from each other and are separated by reticulation patterns (**Figure 6**). In **Figure 5** we have already observed that these races are genetically closely related but contain different mitochondrial haplotypes. This suggests that sexual recombination between similar unknown common parental lines has probably resulted in the emergence of both *Pfs5* and *Pfs15*. The maternal parent is likely of opposite mt-haplotype to explain the difference between the two otherwise genetically-related races. A PHI (Pairwise Homoplasly Index) test [46] was performed to statistically confirm the occurrence of sexual recombination between the *Pfs* isolates under study. Homoplasies are shared similarities found in different branches of the phylogenetic tree which are not inherited directly from an ancestor. The test is significant when an excess of homoplasies is found in the dataset, compared to the number of homoplasies expected (~5%) by a mutation in the absence of recombination. The PHI analysis between all *Pfs* isolates and races provides strong support for the occurrence of recombination (P-value < 1e⁻⁶). The combination of the information derived from the different phylogenetic anal-

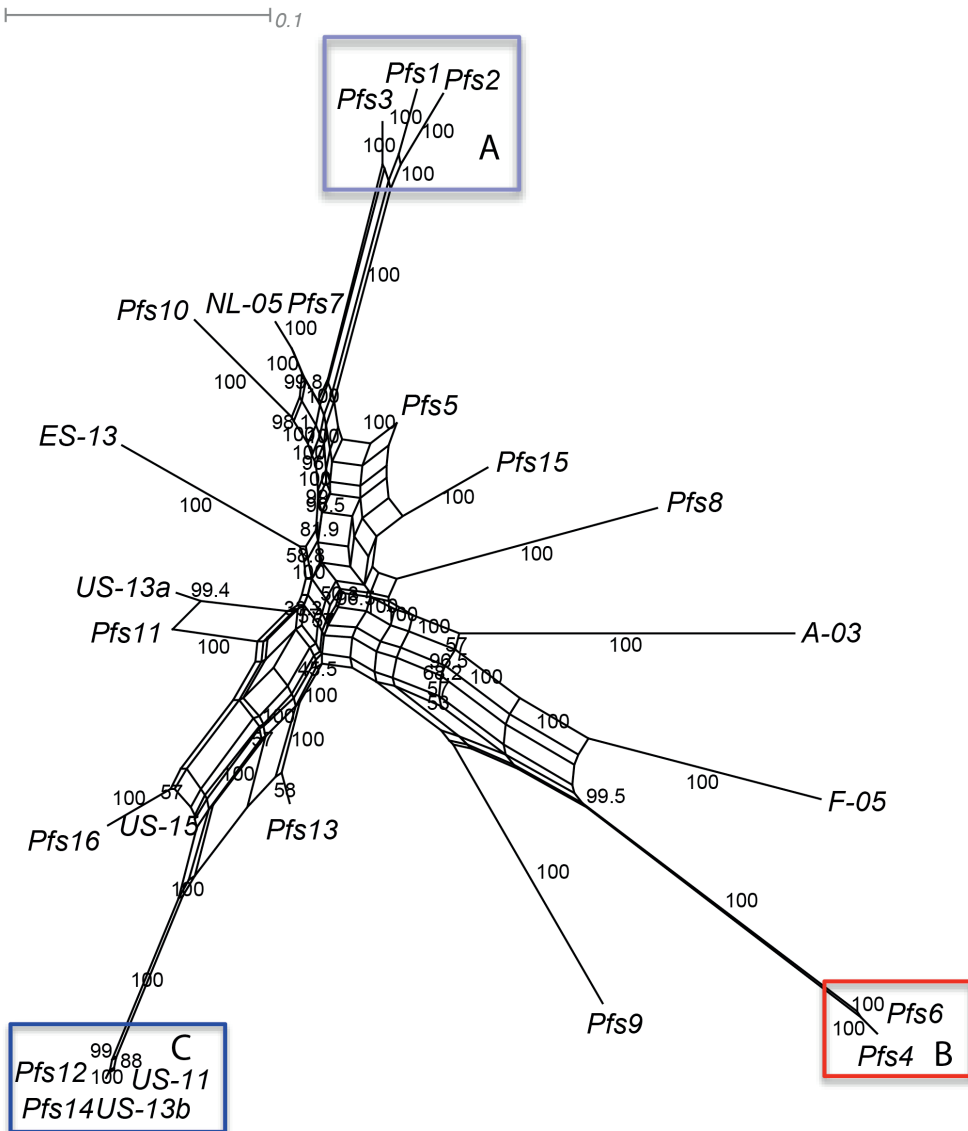


Figure 6. Split decomposition network constructed using the Neighbor-Net algorithm on the *Pfs* isolate SNPmatrix build with SplitsTree [45], branch confidence was calculated using 1,000 bootstraps. Branch lengths are proportional to the expected number of substitutions per site. The parallel edges connecting different isolates indicate conflicting phylogenetic signals. Colored and numbered boxes indicate distinct groups that were observed in the PCA plot (Figure 3) and the STRUCTURE plot (Figure 4).

yses and the PHI-test result provides a convincing indication that sexual recombination between isolates played a role in the emergence of several *Pfs* isolates.

Comparison between denominated *Pfs* races versus field isolates

As mentioned before, the designation of races *is based on* their virulence spectrum on a standard set of spinach differentials (**Supplemental Table 1**). Also, the *Pfs* field isolates used in this study were tested on the differential set (**Supplemental Table 2**). The virulence spectrums of four of the field isolates (*ES-13*, *NL-05*, *US-11*, and *US-15*) were identical to a race type. However, two isolates (*A-03* and *F-05*) were found to be more virulent compared to their race type (indicated with + in the table), and two isolates had a slightly deviating virulence spectrum (*US-13a* and *US13b*).

When we compare the virulence spectrum of the field isolates to their position in the phylogenetic tree (**Figure 5, 6**), we see that almost all are in the same branch as the reference race that best matches their virulence spectrum, indicating they are genetically closely related (**Supplemental Table 1**). This is the case for isolate *US-15* and race *Pfs16* and for isolate *NL-05* and race *Pfs7*. Furthermore, isolates *US-11* and *US13b* are located close to *Pfs14* in the tree to which they are similar but not identical in their virulence spectra. In contrast, *US-13a* shows a virulence spectrum resembling *Pfs14* but in the phylogenetic tree and Neighbor-net, it is located next to *Pfs11* (**Figure 5**). It is striking that *US-13a* and *Pfs11* are so similar while both are capable of breaking different R-loci (**Supplemental Table 1**).

Other notable differences between the phylogenetic tree (**Figure 5**) and the Neighbor-Net network (**Figure 6**) are the locations of *Pfs* isolates *A-03*, *F-05*, and *Pfs8*. Isolates *A-03* and *F-05* are more virulent compared to *Pfs8*. These isolates are in one clade in **Figure 5**, while *Pfs8* resides in a different clade. However, in the Neighbor-Net the reticulation pattern observed between isolates *A-03* (+), *F-05* (++) and *Pfs8* suggest that multiple events of sexual recombination have taken place and perhaps have contributed to their difference in virulence.

Discussion

Over the last two decades, many new *Pfs* races have emerged that have overcome introduced resistance traits in the spinach host. In 1990 there were only three *Pfs* races (1 to 3) known, but because of resistance breeding and breaking of traits, this number rapidly increased to 17 identified races to date [14; 54]. Although the phylogenetic relationship between *Pfs* races and isolates have been studied before based on a limited number of SNP markers [7], a phylogenetic study based on whole-genome sequencing that includes all denominated *Pfs* races except *Pfs17*, was not available yet.

A comparison of the mitochondrial genomes of *Pfs* and *P. tabacina* [28] revealed high levels of similarity both in the overall sequence and gene present. A few notable differences were observed, the mt-genome of *Pfs* is slightly smaller than that of *P. tabacina*. Furthermore, an alignment between the mt-sequences showed a large sequence inversion between the mt-genomes of both pathogens. Certain ORFs and repeats are present on the mt-genome of *P. tabacina* and not of *Pfs*, suggesting that these mt-features have recently evolved or were lost in the mt-genome of *Pfs1*. Notably, the mt-genome of *Pfs* also contains ribosomal sequences *rns* and *rnl* that are not present in the mt-genome of *P. tabacina*. These two ribosomal RNA subunits (*rnl* and *rns*) also have fragmented parts present on the mt-genome of *Pfs*.

Fragmentation and occasional dispersion of ribosomal subunits in the mt-genome have also been observed in other species, such as in several green algae [55] and protists [56]. In oyster species, fragmented ribosomal subunits are also present in the mt-genomes. RT-PCR on these ribosomal subunits revealed that these segments are separately transcribed but still form functional rRNAs and ribosomes [57]. This makes it possible that the fragmented ribosomal sequences in the mt-genome of *Pfs* might form functional ribosomal subunits.

Based on the variants between *Pfs1* and the mt-genomes of the other isolates under study, we can conclude that there are two distinct haplogroups present. This observation has been done before by Choi *et al.* [22] for the *cox2* and *nad1* derived from several *Pfs* isolates and was linked to the geographical location where these isolates were found (**Figure 2**). Although most *Pfs* isolates that were isolated outside the USA can be found in haplogroup I, a clear correlation between the mt-haplogroup and the geographical locations where they originated from could not be established for our samples (**Supplemental Table 2**). For this study, we focused primarily on denominated races of which all were first identified, but not necessarily, isolated in the USA. Similar resistance-breaking isolates are often found at different geographical locations around the same time [8; 58]. This is likely due to the globalized spinach market, in which spinach seeds get transported and sold worldwide. These seeds could contain oospores that can survive for long periods under harsh conditions and likely start an infection via the spinach root [59]. Moreover, a resistance breaking isolate is denominated as a race when it occurs in different geographical locations and has a consistent novel virulence spectrum on the differential set [8, 56]. Therefore, it is difficult to determine where a novel resistance breaking isolates first occurred, and in our case, to know the exact continental origin of our *Pfs* races.

We also identified 35 nucleotide variants that are present in all isolates belonging to haplogroup I and are not present in haplogroup II. These

variants can serve as potential SNP markers of which only a few SNPs have to be determined to identify the mt-haplogroup of new field isolates.

The nuclear phylogeny of the *Pfs* races and isolates under study was also determined. A PCA revealed the presence of 3 distinct groups and one admixed group of *Pfs* isolates in the isolates under study (**Figure 3**). The STRUCTRE analysis (assuming $K = 3$) (**Figure 4**), also showed the presence of 3 clear genetic groups and 1 admixed group. The same groups can also be observed in the phylogenetic tree in **Figure 5** where *Pfs* reference races 1 to 3 (genetic group A) belong to the same monophyletic group and share mitochondrial haplogroup I. This suggests that these races belong to a clonal lineage with sub-clonal variation. The clonality of these isolates is also supported by long parallel lines connecting them to the main net in the Neighbor-net network (**Figure 6**). Races *Pfs4* (isolated in 1990) and *Pfs6* belong to a different monophyletic group in another branch in the phylogenetic tree in **Figure 5** (genetic group B) and share mitochondrial haplogroup II. This suggests that after 1990 *Pfs* isolates with a different genetic background emerged that did not directly evolve from races *Pfs1*, 2, and 3. Also, *Pfs13*, *Pfs14*, *US-11*, and *US-13b* belong to a monophyletic group (C) that share mitochondrial haplogroup I. Interestingly, all isolates in these three groups possess the same mt-haplogroup, however, isolates in the admixed group possess a mixture of the two haplogroups. This suggests that at least some of these isolates have emerged through sexual recombination.

When we look at the genetic composition of the admixed group, we see that *Pfs* race 7 and isolate *NL-05* both show almost equal amounts of genetic background present in groups A and B. As they both possess the same mitochondrial haplogroup (I) it is likely that race *Pfs7* and isolate *NL-05* are clonal variants. *Pfs* races 5 until 11 contain a mixture of genetic backgrounds from all genetic groups (A, B and C) in varying proportions. This is likely the result of repeated sexual recombination between parental lines of different groups.

No linear evolutionary relationship, where each incremental race number has directly evolved from the preceding race, can be found based on the phylogeny. Based on the joint results of the PCA, STRUCTURE analysis, Neighbor-Net network and the phylogenetic tree we see convincing evidence that sexual reproduction events have occurred in the evolutionary history of *Pfs* to evolve races that have overcome consecutive resistance traits.

A PHI-test to assess whether sexual recombination has occurred within the *Pfs* isolates under study provided significant evidence (P -value $< 1e^{-6}$) that sexual recombination has played a role in the evolution and emergence of resistance-breaking *Pfs* races. However, the exact parental lines that have contributed to sexual recombination in the isolates cannot be

determined based on this data. Once a new isolate has emerged through sexual recombination, its clonal offspring would also contain the same recombined loci which would be observed as mixed genotypes in our analysis. Therefore, it is possible that some *Pfs* isolates in the admixed group are clonal variants of isolates that have emerged through sexual recombination. Besides, the sampling of races and isolates used in this study is limited and spread out over time. This means that we are missing nodes which makes interpretation of the exact phylogeny between races and isolates and the parental lines they have emerged from difficult.

The presence of sexual recombination within *Pfs* isolates is in accordance with an earlier study on the phylogeny and the genetic composition of several *Pfs* isolates [7]. This study was based on a limited number of SNP markers and showed that the genotypic variation found within recent field isolates was primarily produced during asexual development. However, the authors also found that overall genetic diversity in historical isolates collected from different geographical locations is likely influenced by sexual recombination on broader geographical and temporal scales. This suggests that both sexual and asexual reproduction has contributed to the diversity within *Pfs*.

The phylogeny of *Pfs* isolates in this study was based on whole-genome data that yields a high number of informative sites. This provides a higher resolution leading to a more robust phylogeny compared to the use of a limited number of SNP markers. Analyzing the phylogenetic relationships between isolates combined with their mitochondrial genome phylogeny has proven to be a powerful method to determine the relatedness of the *Pfs* isolates. However, variants in the isolates can only be determined on parts of the genome that are already present in the reference. A more contiguous and complete reference genome would, therefore, yield additional informative sites. Longer contigs would also allow for the determination of structural variants between the isolates, and recombined regions in the genomes of different isolates.

Although whole-genome sequencing is getting less expensive over time, it is still costly to do this for many field isolates. The identification of regions in the *Pfs* genome with high levels of variation can be used to identify SNP markers. This enables the study of phylogenetic relationships between isolates based on high-throughput genotyping. The results can be incorporated with the phylogenetic data presented in this study. This enables the rapid determination in which parental lines have likely contributed to the emergence of novel field isolates.

In summary, we can conclude that sexual recombination has contributed significantly to the genetic diversity within the studied *Pfs* population. Although the sexual progeny might have a lower fitness it has potentially given rise to novel resistance-breaking isolates. Once a new suc-

successful *Pfs* genotype emerged the asexual cycle ensures the propagation and acts as the dispersal mechanism.

The data and methods presented in this study can be used to assess the genetic composition of emerging and future resistance-breaking isolates, that are highly relevant for steering disease resistance breeding programs and selecting and combining desired resistance traits in spinach cultivars.

Acknowledgments

We wish to thank the breeding companies Enza Zaden, Pop Vriend Seeds, Rijk Zwaan Breeding, and Syngenta Seeds for sharing *Pfs isolate* material and data used in this project. We thank TKI T&U (Topsector Tuinbouw & Uitgangsmaterialen) along with the companies for financing this project.

References

1. Choi Y-J, Hong S-B, Shin H-D. 2007. Re-consideration of *Peronospora farinosa* infecting *Spinacia oleracea* as distinct species, *Peronospora effusa*. *Mycol. Res.* 111:381-91
2. Correll JC, Bluhm BH, Feng C, Lamour K, du Toit LJ, Koike ST. 2011. Spinach: better management of downy mildew and white rust through genomics. *Eur. J. Plant Pathol.* 129:193-205
3. Klosterman SJ, Anchieta A, McRoberts N, Koike ST, Subbarao KV, *et al.* 2014. Coupling spore traps and quantitative PCR assays for detection of the downy mildew pathogens of spinach (*Peronospora effusa*) and beet (*P. schachtii*). *Phytopathology* 104:1349-59
4. Feng C, Saito K, Liu B, Manley A, Kammeijer K, *et al.* 2018. New races and novel strains of the spinach downy mildew pathogen *Peronospora effusa*. *Plant Dis.* 102:613-8
5. Greville RK. 1824. *Flora Edinensis* pp 1-478.
6. Brandenberger L, Morelock T, Correll J. 1991. Reactions of spinach accessions to a new race (race 4) of downy mildew. *Hortscience* 26:712-
7. Lyon R, Correll J, Feng C, Bluhm B, Shrestha S, *et al.* 2016. Population structure of *Peronospora effusa* in the southwestern United States. *PLoS ONE* 11:e0148385
8. Satou M, Nishi K, Kubota M, Fukami M, Tsuji H, Van Ettekovén K. 2006. Appearance of race Pfs: 5 of spinach downy mildew fungus, *Peronospora farinosa* f. sp. *spinaciae*, in Japan. *J. Gen. Plant Pathol.* 72:193-4
9. Correll J, Feng C, Irish B, Koike S, Morelock T, *et al.* 2007. *Spinach downy mildew: overview of races and the development of molecular markers linked to major resistance genes.* pp 135-142. 265 pp.
10. Koike S, Smith R, Schulbach K. 1992. Resistant cultivars, fungicides combat downy mildew of spinach. *Calif. Agric.* 46:29-30
11. Ching L. 2017. *Spinach growers see increased demand.* <http://www.agalert.com/story/?id=10725>
12. Correll J, du Toit L, Koike S, van Ettekovén K. 2010. *Guidelines for spinach downy mildew: Peronospora farinosa f. sp. spinaciae (Pfs).* http://www.worldseed.org/cms/medias/file/TradelIssues/DiseasesResistance/StrainIdentification/Spinach_downy_mildew_Differentials_29102010.pdf
13. Irish BM, Correll JC, Koike ST, Morelock TE. 2007. Three new races of the spinach downy mildew pathogen identified by a modified set of spinach differentials. *Plant Dis.* 91:1392-6
14. 2018. *Denomination of Pfs: 17, a new race of downy mildew in spinach.* <https://www.naktuinbouw.com/about-naktuinbouw/news/denomination-pfs-17-new-race-downy-mildew-spinach>
15. Kamoun S. 2003. Molecular genetics of pathogenic oomycetes. *Eukaryot. Cell* 2:191-9

16. Goss EM, Carbone I, Grunwald NJ. 2009. Ancient isolation and independent evolution of the three clonal lineages of the exotic sudden oak death pathogen *Phytophthora ramorum*. *Mol. Ecol.* 18:1161-74
17. Heitman J. 2006. Sexual reproduction and the evolution of microbial pathogens. *Curr. Biol.* 16:R711-25
18. Inaba T, Takahashi K, Morinaka T. 1983. Seed Transmission of Spinach Downy Mildew. *Plant Dis.* 67:1139-41
19. Fry WE, Grünwald NJ. 2010. *Introduction to oomycetes*. <https://www.apsnet.org/edcenter/disandpath/oomycete/introduction/Pages/IntroOomycetes.aspx>
20. Gladieux P, Ravel S, Rieux A, Cros-Arteil S, Adreit H, *et al.* 2018. Coexistence of multiple endemic and pandemic lineages of the rice blast pathogen. *MBio* 9:e01806-17
21. Li Y, van der Lee TA, Evenhuis A, van den Bosch GB, van Bekkum PJ, *et al.* 2012. Population dynamics of *Phytophthora infestans* in the Netherlands reveals expansion and spread of dominant clonal lineages and virulence in sexual offspring. *G3 (Bethesda)* 2:1529-40
22. Choi YJ, Thines M, Han JG, Shin HD. 2011. Mitochondrial phylogeny reveals intraspecific variation in *Peronospora effusa*, the spinach downy mildew pathogen. *J. Microbiol. (Seoul)* 49:1039-43
23. Schomaker-Bastos A, Prosdocimi F. 2018. mitoMaker: A pipeline for automatic assembly and annotation of animal mitochondria using raw NGS data. *Preprints:2018080423*
24. Silva GG, Dutilh BE, Matthews TD, Elkins K, Schmieder R, *et al.* 2013. Combining de novo and reference-guided assembly with scaffold_builder. *Source Code Biol. Med.* 8:23
25. Tillich M, Lehwerk P, Pellizzer T, Ulbricht-Jones ES, Fischer A, *et al.* 2017. GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 45:W6-W11
26. Chan PP, Lowe TM. 2019. tRNAscan-SE: searching for tRNA genes in genomic sequences. In *Gene Prediction:1-14*: Springer. Number of 1-14 pp.
27. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792-7
28. Derevnina L, Chin-Wo-Reyes S, Martin F, Wood K, Froenicke L, *et al.* 2015. Genome sequence and architecture of the tobacco downy mildew pathogen *Peronospora tabacina*. *Mol. Plant-Microbe Interact.* 28:1198-215
29. Sullivan MJ, Petty NK, Beatson SA. 2011. Easyfig: a genome comparison visualizer. *Bioinformatics* 27:1009-10
30. Lohse M, Drechsel O, Bock R. 2007. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr. Genet.* 52:267-74
31. Garrison E. 2012. Vcfliib: A C++ library for parsing and manipulating VCF files. *GitHub* <https://github.com/lekg/vcfliib>

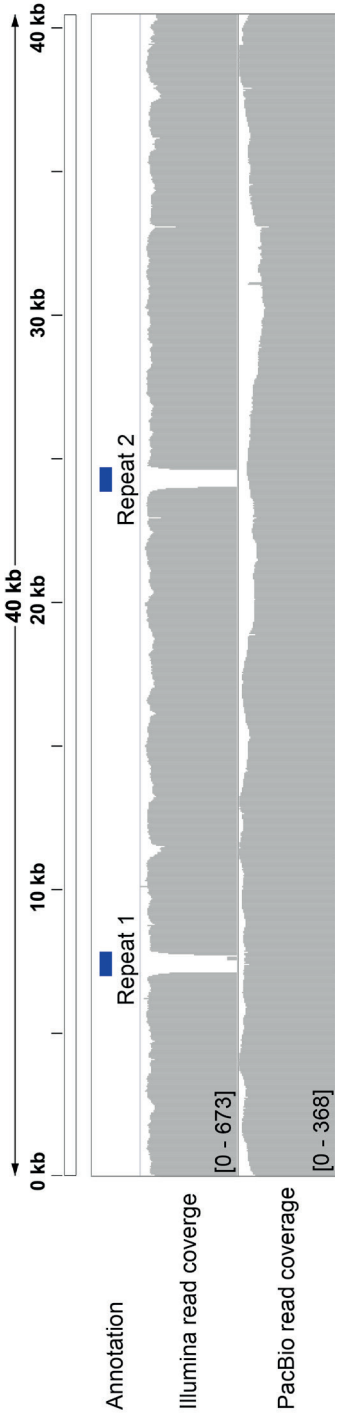
32. Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10:512-26
33. Dawyndt P, De Meyer H, De Baets B. 2006. UPGMA clustering revisited: A weight-driven approach to transitive approximation. *Int. J. Approx. Reason.* 42:174-91
34. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* 44:W242-5
35. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, *et al.* 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156-8
36. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841-2
37. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754-60
38. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al.* 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078-9
39. 2018. Picard Toolkit Broad Institute
40. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, *et al.* 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* 43:110 1-33
41. Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27:2987-93
42. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28:3326-8
43. Lischer HE, Excoffier L. 2012. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* 28:298-9
44. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312-3
45. Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23:254-67
46. Bruen TC, Philippe H, Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172:2665-81
47. Boutemy LS, King SRF, Win J, Hughes RK, Clarke TA, *et al.* 2011. Structures of *Phytophthora* RXLR effector proteins: A conserved but adaptable fold underpins functional diversity. *J. Biol. Chem.* 286:35834-42
48. Francis RM. 2017. pophelper: an R package and web app to analyse and visualize population structure. *Mol. Ecol. Resour.* 17:27-32
49. Martin FN. 2008. Mitochondrial haplotype determination in the oomycete plant pathogen *Phytophthora ramorum*. *Curr. Genet.* 54:23-34

50. Cooke DE, Drenth A, Duncan JM, Wagels G, Brasier CM. 2000. A molecular phylogeny of *Phytophthora* and related oomycetes. *Fungal Genet. Biol.* 30:17-32
51. Chen Y-C, Liu T, Yu C-H, Chiang T-Y, Hwang C-C. 2013. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS ONE* 8:e62856
52. Verity R, Nichols RA. 2016. Estimating the number of subpopulations (K) in structured populations. *Genetics* 203:1827-39
53. Galtier N, Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 363:4023-9
54. Feng C, Lamour KH, Bluhm BH, Sharma S, Shrestha S, *et al.* 2018. Genome sequences of three races of *Peronospora effusa*: a resource for studying the evolution of the spinach downy mildew pathogen. *Mol. Plant-Microbe Interact.* 31:1230-1
55. Nedelcu AM, Lee RW, Lemieux C, Gray MW, Burger G. 2000. The complete mitochondrial DNA sequence of *Scenedesmus obliquus* reflects an intermediate stage in the evolution of the green algal mitochondrial genome. *Genome Res.* 10:819-31
56. Gray MW, Lang BF, Cedergren R, Golding GB, Lemieux C, *et al.* 1998. Genome structure and gene content in protist mitochondrial DNAs. *Nucleic Acids Res.* 26:865-78
57. Milbury CA, Lee JC, Cannone JJ, Gaffney PM, Gutell RR. 2010. Fragmentation of the large subunit ribosomal RNA gene in oyster mitochondrial genomes. *BMC Genomics* 11:485
58. Irish BM, Correll JC, Koike ST, Schafer J, Morelock TE. 2003. Identification and cultivar reaction to three new races of the spinach downy mildew pathogen from the United States and Europe. *Plant Dis.* 87:567-72
59. Kunjeti SG, Anchieta A, Subbarao KV, Koike ST, Klosterman SJ. 2016. Plasmolysis and vital staining reveal viable oospores of *Peronospora effusa* in spinach seed lots. *Plant Dis.* 100:59-65

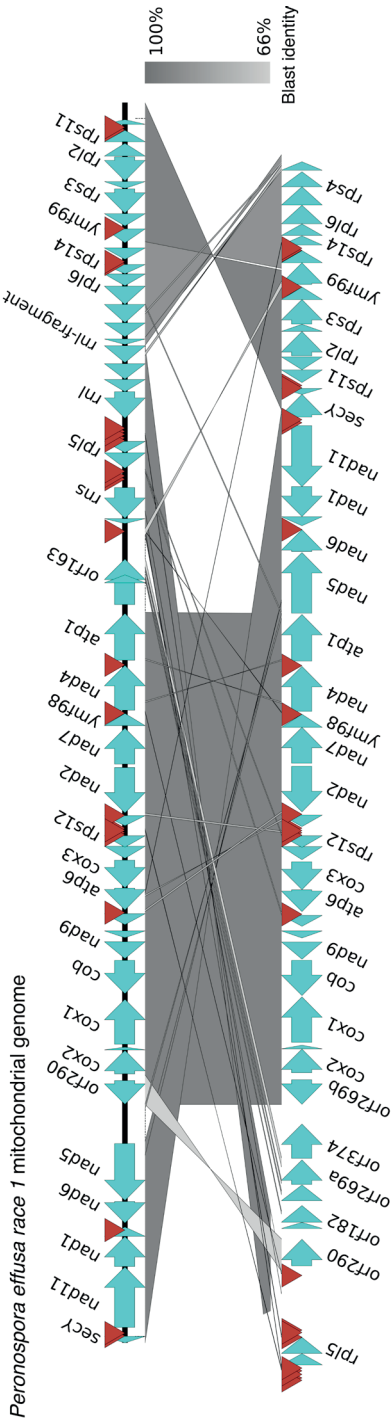
Supplemental data



3



Supplemental figure 1. Visual representation of non-quality filtered Illumina and error-corrected PacBio read coverage to the mitochondrial genome of *Pfs1*. The two gaps in the Illumina read coverage are found at GC-poor repeat regions (indicated in the annotation track) in the mitochondrial genome.



Peronospora tabacina 968-S-26 mitochondrial genome

Supplemental Figure 2. Alignment between the mt-sequence of *Pfs* (Race 1) and *P. tabacina* (968-S-26). The arrows indicate the location of that gene on the mt-sequence, tRNAs are depicted with a brown arrow. Gene annotations for coding sequences are depicted above the arrows.

Supplemental Table 1. Differential set of spinach cultivars (April 2018) currently used in the evaluation and denomination of *P. effusa* races and isolates. R-loci present in the parental lines are shown in the second and third columns. OP = Open pollinated and ?? = Unknown R-loci. The virulence spectrum for each race is indicated using + for susceptible and – for resistant or (-) for intermediate resistance.

| Differential | Parental resistance | | Race/Pfs | | | | | | | | | | | | | | | |
|--------------|---------------------|--------|----------|---|---|-----|---|-----|-----|---|---|-----|----|----|-----|----|-----|----|
| | Male | Female | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| Viroflay | OP | none | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| NIL 5 | OP | RPF5 | - | - | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| NIL 3 | OP | RPF3 | - | + | - | + | - | + | + | - | - | + | - | - | + | - | + | - |
| NIL 4 | OP | RPF4 | - | - | - | - | + | + | + | + | + | + | + | + | + | + | - | + |
| NIL 6 | OP | RPF6 | - | + | - | - | - | + | - | + | + | + | - | + | (-) | + | - | - |
| NIL 1 | OP | RPF1 | - | - | - | - | - | - | - | + | - | + | - | + | - | + | - | - |
| NIL 2 | OP | RPF2 | - | - | - | - | - | - | - | - | - | - | + | + | + | + | - | + |
| Whale | ?? | RPF3 | - | - | - | (-) | - | (-) | (-) | - | - | (-) | - | - | + | - | (-) | - |
| Pigeon | RPF2 | RPF9 | - | - | - | - | - | - | - | - | - | - | - | - | - | + | - | + |
| Caladonia | RPF3 | RPF9 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | + | - |
| Meerkat | RPF2 | RPF10 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | + |
| Hydrus | ?? | RPF11 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |

Supplemental Table 2. Overview of the location where the different *Pfs* races and isolates used in this study were first identified. The number of sequencing reads obtained and the percentage of reads that aligned to the reference assembly (*Pfs1*) as well as the nucleotides covered in the reference are indicated in the columns. Also, the total number of polymorphisms and type (homozygous or heterozygous) identified in each isolate is listed.

| Race/ Isolate | First identified | Virulence spectrum | sequenced reads | Input reads | Mapped reads | % mapped | SNPs | Hetrozygous SNP | Homozygous SNP | Indels | % bases in reference covered |
|------------------|------------------|-----------------------|--------------------|----------------|-----------------|-------------|--------|--------------------|-------------------|--------|---------------------------------|
| <i>Pfs1</i> | USA | | 535,471,470 | 40,083,766 | 22,383,214 | 55.84% | 8,896 | 7,291 | 1,605 | 87 | 100.00% |
| <i>Pfs2</i> | USA | | 57,784,612 | 40,083,766 | 24,574,442 | 61.31% | 15,656 | 5,822 | 9,834 | 2,986 | 100.00% |
| <i>Pfs3</i> | USA | | 47,236,596 | 40,083,766 | 22,974,323 | 57.32% | 15,956 | 5,588 | 10,368 | 3,064 | 99.98% |
| <i>Pfs4</i> | USA | | 68,315,382 | 40,083,766 | 22,609,933 | 56.41% | 68,071 | 27,080 | 40,991 | 8,438 | 99.62% |
| <i>Pfs5</i> | USA | | 54,701,916 | 40,083,766 | 24,030,053 | 59.95% | 55,642 | 44,425 | 11,217 | 7,162 | 99.97% |
| <i>Pfs6</i> | USA | | 55,053,740 | 40,083,766 | 23,843,641 | 59.48% | 69,666 | 30,555 | 39,111 | 8,767 | 99.12% |
| <i>Pfs7</i> | USA | | 61,908,142 | 40,083,766 | 24,752,024 | 61.75% | 46,564 | 30,954 | 15,610 | 6,177 | 99.89% |
| <i>Pfs8</i> | USA | | 66,263,460 | 40,083,766 | 20,243,716 | 50.50% | 54,135 | 28,424 | 25,711 | 7,078 | 99.42% |
| <i>Pfs9</i> | USA | | 52,501,840 | 40,083,766 | 23,476,452 | 58.57% | 60,055 | 28,077 | 31,978 | 7,606 | 99.63% |
| <i>Pfs10</i> | USA | | 62,278,520 | 40,083,766 | 24,343,863 | 60.73% | 26,992 | 7,291 | 19,701 | 6,307 | 99.69% |
| <i>Pfs11</i> | USA | | 54,266,776 | 40,083,766 | 23,921,377 | 59.68% | 59,596 | 40,234 | 19,362 | 7,613 | 100.00% |
| <i>Pfs12</i> | USA | | 50,792,422 | 40,083,766 | 24,381,532 | 60.83% | 53,940 | 16,745 | 37,195 | 7,127 | 99.45% |
| <i>Pfs13</i> | USA | | 58,193,096 | 40,083,766 | 21,417,710 | 53.43% | 56,961 | 36,525 | 20,436 | 7,406 | 99.70% |
| <i>Pfs14</i> | USA | | 51,920,130 | 40,083,766 | 24,278,534 | 60.57% | 53,725 | 14,527 | 39,198 | 7,120 | 99.42% |
| <i>Pfs15</i> | USA | | 30,137,642 | 40,083,766 | 14,735,927 | 48.90% | 54,784 | 33,703 | 21,081 | 6,960 | 99.20% |
| <i>Pfs16</i> | USA | | 59,725,890 | 40,083,766 | 21,125,948 | 52.70% | 56,770 | 31,226 | 25,544 | 7,248 | 99.63% |
| A-03 | Australia | <i>Pfs8+</i> | 64,698,358 | 40,083,766 | 24,402,378 | 60.88% | 53,973 | 21,296 | 32,677 | 6,969 | 99.70% |
| ES-13 | Spain | <i>Pfs15</i> variant | 57,929,332 | 40,083,766 | 23,822,587 | 59.43% | 49,726 | 27,351 | 22,375 | 6,601 | 99.89% |
| F-05 | France | <i>Pfs8++</i> | 82,411,908 | 40,083,766 | 23,801,468 | 59.38% | 61,593 | 30,445 | 31,148 | 7,828 | 99.67% |
| NL-05 | the Netherlands | <i>Pfs7</i> variant | 52,547,074 | 40,083,766 | 23,397,326 | 58.37% | 43,340 | 27,221 | 16,119 | 5,858 | 100.00% |
| US-11 | USA | <i>Pfs14</i> variant | 55,196,784 | 40,083,766 | 23,433,149 | 58.46% | 52,793 | 14,551 | 38,242 | 6,926 | 99.58% |
| US-13a | USA | <i>Pfs14</i> ? | 75,837,594 | 40,083,766 | 24,782,694 | 61.83% | 63,764 | 47,066 | 16,698 | 7,971 | 99.76% |
| US-13b | USA | <i>Pfs14</i> ? | 63,732,754 | 40,083,766 | 23,457,539 | 58.52% | 52,410 | 12,445 | 39,965 | 6,851 | 99.63% |
| US-15 | USA | <i>Pfs16</i> variant | 68,333,652 | 40,083,766 | 22,975,640 | 57.32% | 63,101 | 42,340 | 20,761 | 7,958 | 100.00% |

Supplemental Table 3. Overview of the locations where the different *Pfs* races and isolates used in this study were first identified combined with their mt-haplogroup. The mean number of reads covering the mt-genome per isolate is indicated in the last column.

| Race/isolate | First discovered | Haplogroup | Mean coverage (number of reads) |
|---------------|------------------|------------|---------------------------------|
| <i>Pfs1</i> | USA | II | 6,400.59 |
| <i>Pfs2</i> | USA | II | 7,756.92 |
| <i>Pfs3</i> | USA | II | 4,286.77 |
| <i>Pfs4</i> | USA | I | 2,855.20 |
| <i>Pfs5</i> | USA | I | 4,206.79 |
| <i>Pfs6</i> | USA | I | 4,022.38 |
| <i>Pfs7</i> | USA | II | 1,781.90 |
| <i>Pfs8</i> | USA | II | 3,268.98 |
| <i>Pfs9</i> | USA | II | 9,278.19 |
| <i>Pfs10</i> | USA | II | 4,824.29 |
| <i>Pfs11</i> | USA | II | 4,356.95 |
| <i>Pfs12</i> | USA | II | 17,227.20 |
| <i>Pfs13</i> | USA | II | 4,911.03 |
| <i>Pfs14</i> | USA | II | 4,627.26 |
| <i>Pfs15</i> | USA | II | 821.51 |
| <i>Pfs16</i> | USA | I | 4,325.34 |
| <i>US-11</i> | USA | II | 15,172.30 |
| <i>US-13a</i> | USA | I | 33,42.56 |
| <i>US-13b</i> | USA | II | 15,349.60 |
| <i>US-15</i> | USA | II | 9,288.10 |
| <i>A-03</i> | Australia | I | 9,405.90 |
| <i>ES-13</i> | Spain | I | 5,809.73 |
| <i>F-05</i> | France | I | 13,289.30 |
| <i>NL-05</i> | the Netherlands | II | 5,958.51 |

Supplemental Table 4. The position of each variant present in all isolates containing mt-haplogroup I that are not present in any isolate that possesses mt-haplogroup II. Also, the reference base (Ref) and base identified in the isolate (Alt) for each variant as well as the mt-genome feature the variant resides in (if applicable). The last three columns indicate the start and end of the mt-feature as well as the coding strand.

| Position | Ref | Alt | Feature name | Feature start | Feature stop | Strand |
|----------|-----|-----|-------------------------------|---------------|--------------|--------|
| 1,013 | G | A | nad11 gene | 515 | 2,522 | + |
| 1,732 | A | C | nad11 gene | 515 | 2,522 | + |
| 2,546 | C | A | tRNA-Undet15 gene | 2,350 | 2,551 | + |
| 2,546 | C | A | tRNA-Undet45 gene | 2,362 | 2,562 | - |
| 2,546 | C | A | nad1 gene | 2,518 | 3,499 | + |
| 2,965 | T | C | nad1 gene | 2,518 | 3,499 | + |
| 5,836 | C | G | nad5 gene | 4,683 | 6,183 | - |
| 6,716 | T | G | NA | NA | NA | NA |
| 6,784 | A | T | NA | NA | NA | NA |
| 6,837 | C | T | NA | NA | NA | NA |
| 6,943 | C | A | NA | NA | NA | NA |
| 8,145 | G | A | orf290 gene | 7,854 | 8,634 | - |
| 8,145 | G | A | orf269b gene | 7,854 | 8,601 | - |
| 8,148 | A | T | orf290 gene | 7,854 | 8,634 | - |
| 8,148 | A | T | orf269b gene | 7,854 | 8,601 | - |
| 9,422 | C | T | cox2 gene | 8,870 | 9,629 | + |
| 10,793 | T | C | cox1 gene | 9,842 | 11,306 | + |
| 11,921 | A | G | cob gene | 11,461 | 12,583 | - |
| 12,017 | G | A | cob gene | 11,461 | 12,583 | - |
| 16,311 | T | C | rps7 gene | 16,123 | 16,336 | - |
| 16,311 | T | C | rps12 gene | 16,310 | 16,694 | - |
| 19,490 | T | C | nad7 gene | 19,046 | 20,225 | + |
| 19,748 | T | G | nad7 gene | 19,046 | 20,225 | + |
| 20,027 | T | G | nad7 gene | 19,046 | 20,225 | + |
| 23,475 | A | C | atp1 gene | 22,983 | 23,997 | + |
| 25,207 | T | A | smallsubunitribosomalRNA gene | 24,292 | 25,792 | + |
| 25,207 | T | A | orf172 gene | 25,030 | 25,274 | + |
| 25,534 | A | G | smallsubunitribosomalRNA gene | 24,292 | 25,792 | + |
| 25,534 | A | G | orf269a 2 gene | 25,434 | 25,546 | + |
| 25,534 | A | G | orf269a gene | 25,530 | 25,681 | + |
| 25,709 | T | C | smallsubunitribosomalRNA gene | 24,292 | 25,792 | + |
| 25,990 | C | T | NA | NA | NA | NA |
| 26,304 | G | T | NA | NA | NA | NA |
| 30,150 | A | T | tRNA-Undet60 gene | 30,093 | 30,165 | - |
| 39,216 | C | T | rps13 gene | 39,046 | 39,368 | + |



Chapter 4

Peronospora effusa effector polymorphisms associated with the breaking of resistance loci in spinach

Joël Klein¹, Sjoerd Gremmen¹, Guido Van den Ackerveken¹

¹ Plant-Microbe Interactions, Department of Biology,
Utrecht University, Utrecht, The Netherlands

Abstract

The downy mildew *Peronospora effusa* is a major threat to spinach production worldwide. *P. effusa*, like other oomycete pathogens, uses host-translocated effector proteins to suppress immune responses and to facilitate disease in its plant host. In turn, plants have evolved resistance (R) genes that are capable of recognizing effectors or their activity and triggering an effective immune response. Spinach varieties with resistance to *P. effusa* have been bred by introgressing R-loci derived from wild germplasm. The wide-spread use of these varieties, however, has created a strong selection pressure on the pathogen to overcome resistance. This has resulted in the evolution and selection of resistance-breaking *P. effusa* isolates. We found that many effectors in the genome of *P. effusa* are part of paralogous gene groups that have evolved from a common ancestral effector gene through duplication and diversification. By comparing effector sequences from 16 reference races and 8 field isolates we established that *P. effusa* genes encoding effectors of the RxLR class accumulate non-synonymous polymorphism at a higher rate compared to other genes in the genome. Two effector genes were found to be absent in several isolates under study, and amino acid substitutions in three effectors associate to the breaking of single resistance loci in spinach.

Introduction

Peronospora effusa (here referred to as *Pfs*) [1], previously known as *Peronospora farinosa* f. sp. *spinaciae*, is economically the most important disease of cultivated spinach [1]. This obligate biotrophic downy mildew belongs to the oomycetes, a group of filamentous microbes that are a part of the Stramenopile taxon. The oomycetes include many plant pathogens that form a significant threat to global food production [2; 3].

Plants possess an innate defense response against oomycetes and other microbial pathogens [4], which is activated upon the detection of pathogen-associated molecular patterns (PAMPs). Pattern-recognition receptors detect PAMPs and trigger a defense response called pattern-triggered immunity (PTI) [5]. However, oomycetes are known to secrete proteins (commonly referred to as effectors) that can counteract plant immune responses.

Two major classes of effectors can be distinguished based on the plant compartment they are active in; apoplastic effectors and host-translocated effectors [6]. Apoplastic effectors act in the extracellular space between pathogen and host cells (the apoplast), while host-translocated effectors act inside host cells. In *Phytophthora* and downy mildew species, the major classes of host-translocated effectors can be further divided into RxLR and Crinkler (CRN) effectors. This division is made based on their conserved translocation motif (RxLR-EER for RxLR, and LFLAK-HVL for CRN). Host-translocated effector molecules are known for their activity as suppressors of PTI and other defense responses. They can also serve as disease-enhancing factors in the host, which subsequently enables the progression of the infection [7; 8]. In turn, plants have evolved intracellular receptor proteins that are encoded by resistance (R) genes and recognize effectors or their activity [9]. Most intracellular R proteins belong to the family of nucleotide-binding leucine-rich repeat (NB-LRR) protein, also known as NLRs [10]. Upon recognition of an effector, the plant mounts a defense response, known as effector-triggered immunity (ETI) [9; 11]. ETI often involves a hypersensitive response (HR) encompassing programmed cell death and associated arrest of pathogen invasion. A good understanding of host resistance against downy mildews is found for the model plant species *Arabidopsis thaliana*. Several of the *RPP*-genes have been cloned and characterized, as well as some of the matching avirulence genes of the downy mildew *Hyaloperonospora arabidopsidis* [12; 13].

The presence of R-genes in the host imposes a strong selection pressure on the oomycete effector genes to adapt and overcome resistance. A subset of effectors in oomycetes has been described to be located in repeat-rich regions. These regions are associated with a higher rate of

point variants, sequence rearrangements, duplications and deletions [14]. Effector proteins in *Phytophthora infestans*, *Phytophthora ramorum* and *Hyaloperonospora peronospora* located near repeat-rich regions showed an uneven rate of compared to other genes in the genome [15].

Pathogens can evade recognition of their effectors by immune receptors through several genetic mechanisms; (I) the pathogen can supplement its effector repertoire through duplication and diversification to suppress ETI, (II) recognition can be avoided through amino acid changes in the recognized effector protein, (III) through partial or complete deletion of the effector gene from the genome, or (IV) through silencing of the effector gene [16].

There are many examples of effector adaptation in oomycetes to avoid host recognition [16]. For instance, in *H. arabidopsidis* a single amino acid substitution in the effector domain of RxLR protein ATR13 was found to be sufficient to impair the recognition by resistance protein RPP13 [17]. In *P. infestans* effector *PiAVR2* was found to indirectly trigger resistance protein R2 [18]. Variants of these effectors were either not transcribed or contained polymorphisms resulting in amino acid changes yielding allelic variants which circumvented recognition by the potato host [19].

Besides, many effectors of plant pathogenic oomycetes have undergone recent expansions through gene duplication and divergence [20-24]. Duplicated and diverged effectors often share C-terminal conserved motifs forming tandem repeats with the residues W, Y and L [25; 26]. These conserved domains form a structural fold consisting of three or four α -helices with one hydrophobic pocket (WY) [27] or five α -helices with two hydrophobic pockets (LWY) [28]. Resistance breeding to *Pfs* started almost a century ago with the identification of two resistant spinach accessions from Iran that were effective against *Pfs* race 1.

Shortly after *Pfs* race 2 was identified (1958), a trait was discovered that conveyed resistance against both *Pfs1* and *Pfs2* [29]. Resistance against race 3 (1976) was facilitated by the simultaneous incorporation of two R-genes into hybrids that were introduced in 1978 [30]. *Pfs4*, isolated in 1990, was able to break this resistance. Resistance against this race was found by screening 707 spinach accession originating from 41 countries [31]. This resulted in the identification of nine accessions that were partially resistant of which two accessions exhibited a high level of resistance to *Pfs4*. However, after the rapid emergence and identification of races *Pfs5* to 11, which were identified in a time-span of 12 years, it was clear that new strategies and better sources for resistance were needed to broaden protection against this pathogen.

To aid in the resistance breeding against *Pfs*, a set of spinach differentials that contain one or two R-loci [32] (denoted as RPF for Resistance to *Peronospora farinosa*) effective against several *Pfs* isolates was devel-

oped (**Chapter 3, Supplemental Table 1**) [33]. Currently, this set includes near-isogenic lines containing a single R-loci in the genetic background of the susceptible Viroflay line [34; 35]. To date, there are 6 NIL lines available that hold 6 resistance loci (*RPF1* to *6*). Also, *RPF9* and *RPF10* are available and currently used in hybrid spinach lines conveying resistance against *Pfs* races that have broken the six previously used R-loci [29; 36; 37]. However, a complete understanding of the molecular mechanisms by which R-loci, and in extent R-genes, convey resistance to the spinach host is not available yet [38].

Overall, previous studies on the relation between host R-genes and oomycete effectors have broadened the understanding of how these pathogens can adapt and overcome host resistance. To expand on this knowledge and determine how effectors in *Pfs* isolates adapt to known R-loci, we acquired sequencing data of 16 *Pfs* races and 8 field isolates. Differential interactions were used to associate changes effector genes of the different *Pfs* races to the breaking of known R-loci in spinach. Moreover, this data was used to evaluate the selection pressure imposed on putative effector sequences in the *Pfs* genome.

Materials and methods

Identification of conserved effectors, LWY domains and clustering of *Pfs* effectors sequences

Conserved *Pfs* effectors with orthologs present in other oomycete species were identified in the *Pfs1* reference genome (**Chapter 2**) determined based on bi-directional best BLAST hit (BBH) approach, where the protein sequences of *Pfs* effectors were queried against the NCBI-RefSeq protist database. Protein sequences of the top hit (E-value cutoff = $1e^{10}$) for each sequence were queried against the entire proteome of *Pfs*. An effector is considered to be conserved when the top hit to the database also produced a top hit with the sequence it was originally identified with.

Pfs effectors belonging to paralogous gene groups (PGGs) were identified for each effector by performing all-versus-all BlastP searches using mature (without a signal peptide) *Pfs* effector peptide sequences at an E-value cutoff set at $5e^{-11}$. The tabular BLAST output table was used as input for the Markov Cluster Algorithm (MCL) version 14-137 [39] using MCL's mclxload function. User-selected options for mclxload were --stream-mirror --stream-neg-log10 --stream-tf 'ceil(200)', otherwise default settings were used. The clustering was performed using an inflation value of 2. Clustering in Cytoscape was done based on the number of edges that are shared between the nodes within a cluster. The backgrounds of the nodes

in Cytoscape are colored based on the presence or absence of a WY domain, and the connecting edge thickness is based on $-\text{Log}(E\text{-value})$. This E-value was obtained with the pairwise blast of *Pfs* effectors (see above).

CLC Genomic Workbench version 11 was used to align the peptide sequences of eight homologous RxLR effectors and the construction of a phylogenetic tree. A visual representation of the alignment was created using ESPrit 3.0 [40]. The phylogenetic tree was built in CLC using a Jukes-Cantor substitution model in combination with the UPGMA clustering method and 100 bootstraps. The conserved domains RxLR, EER/DER and (L)WY and their relative locations were annotated at the branches of the phylogram using Adobe InDesign.

Presence, absence and paralogous *Pfs* genes

The read coverage depth for all genes in the different *Pfs* isolates was determined using the Samtools depth functionality [41]. The coverage value for each individual gene was obtained by multiplying the number of reads mapped to a gene by the read length and divided over the total length of the gene (**Supplemental Table 4**). These values were normalized per isolate based on the mean coverage over the genome. A gene is considered to be present when the complete coding sequence is covered by sequencing reads. A gene is considered to be paralogous when the coverage score is higher than two standard deviations above the mean average coverage score of an isolate (**Supplemental Table 5**). Paralogous genes that were likely not correctly assembled in the *Pfs1* reference genome, or recently duplicated in a specific *Pfs* isolate, were identified based on their relatively higher coverage score compared to mean coverage over the genome.

Evolutionary selection pressure on *Pfs* genes in different isolates

The VCF files for each *Pfs* isolate were processed with SNPeff version 4.3i [42]. User-selected options were `-v -ud 0 -canon -no-downstream -no-intergenic -no-intron -no-upstream -no-utr -onlyProtein`, otherwise standard settings were used. Based on the tabular output of SNPeff the dN/dS value for each gene in each isolate was calculated by dividing the non-synonymous (variants_effect_missense_variant) over the synonymous variants (variants_effect_synonymous_variant). Genes that are likely paralogous (172 in total), (**Supplemental Table 5**), located on contigs smaller than 1 kb or with a significant overlap with a repeat sequence (> 20%), or annotated as a transposable element by ANNIE (**Chapter 2, Supplemental Table 5**) were left out from the analysis of selection pressure.

The remaining gene models (8,937) with corresponding dN/dS values were partitioned into four sets: RxLR genes that encode proteins with either an exact canonical RxLR motif or an RxLR-like motif in combination with one or more WY motifs, (66 in total). RxLR-like genes encoding proteins with an RxLR-like motif without WY domains (28 in total), CRN genes encoding proteins with an LFLAK or HVL-like (8 in total), all containing a signal peptide for secretion (SP). Also a reference group consisting of conserved single core-orthologous genes identified with BUSCO (BUSCO genes, 187 in total) was defined. Remaining genes (8,648) not belonging to one of the aforementioned groups are designated in this study as “other genes”.

***Pfs* effector-based phylogenetic tree**

Based on the VCF files containing the variant data between the *Pfs* isolates and the assembled reference genome (*Pfs1*) new VCF files were created that only contained variants for each isolate residing in coding sequences of genes in the RxLR set. This set consists of 66 RxLR effectors, however, effector genes *Pfs1|1792* and *Pfs1|1509* were found absent in some of the isolates under study. The missing genes were removed from the analysis. GATK [43] FastaAlternateReferenceMaker program was used to generate isolate-specific coding sequences for the 64 remaining effector genes in which heterozygous variants were encoded using IUPAC codes. The resulting 64 isolate-specific effector nucleotide sequences were individually aligned using MUSCLE [44] with default settings, and the resulting alignments concatenated. The concatenated alignment served as an input for RAxML to construct a maximum likelihood phylogenetic tree using a GTR-GAMMA nucleotide substitution model and 500 bootstraps to estimate branch likelihood. The resulting phylogenetic tree was combined in a tanglegram for comparison with the phylogenetic tree that was previously constructed on genome-wide variants (**Chapter 3**). The SPR distance [45] between the trees was calculated using phylo. io [46].

Identification of homozygous non-synonymous variants in *Pfs* isolates

Based on the variant annotations provided by SNPeff [42], homozygous non-synonymous (HNS) variants were identified and extracted for all putative effector sequences and all other proteins in the secretome (**Chapter 2**). To compare phylogenetically close *Pfs* isolates, the VCF files containing the variant annotations were intersected using bedtools intersect –a VCF1 –b VCF2 –v –header [47].

Results

Effector families in *P. effusa*

The evolutionary relationships between the *Pfs* host-translocated effectors were analyzed as it is known that, in oomycetes, effectors evolve by rapid duplication and divergence [24]. To examine this for *Pfs* we determined the sequence similarity between the 113 *Pfs* candidate host-translocated effectors (RxLR, RxLR-like, and CRN) (**Chapter 2**) and grouped them in paralogous gene groups (PGGs) with MCL Markov clustering. This resulted in the identification of nine PGGs with each more than two effector sequences, encompassing in a total of 46 effector proteins (**Supplemental Table 1**). A complementary method to group gene families is by clustering based on their $-\log(\text{E-value})$ and visualizing this in a similarity network in Cytoscape (**Figure 1**). In the protein sequence similarity network different PGGs identified by MCL can be clearly distinguished (**Figure 1**). Effectors that contain one or more WY motifs (indicated with red background) are mostly clustered together. This is expected as an effector duplication would maintain the protein domains within an expanded family. Some larger groups of effector proteins in the network are connected to each other by a thin edge (low significance).

4 A bi-directional best BLAST hit (BBH) approach revealed that 36 RxLR and one CRN sequence have likely orthologs in other species and are therefore regarded as conserved effectors (**Supplemental Table 2**). Most of the conserved effectors in *Pfs* showed similarity with the effectors of *Phytophthora* species. Notably, all conserved RxLR sequences in *Pfs* contain one or more WY motif. Three PGGs groups with more than two effectors contained one or more conserved effectors, while the other six groups (containing in total 30 effectors) did not contain any conserved effectors. Of these three groups, PGG2 contains three conserved RxLR effectors while PGG5 and PGG7 each contain one conserved effector.

Signatures of selection in RxLR effector genes

Over the last decades, the introduction of new resistant spinach cultivars has resulted in strong selection pressures on the *Pfs* genome to adapt and overcome host resistance. Therefore, it is expected that corresponding host-recognized effector genes (also known as avirulence genes) show signs of positive selection. Scenario's that can be envisioned to overcome the newly introduced resistance in the spinach host are adaptive amino acid substitutions in effector proteins. Also, the complete deletion of the recognized effector gene and disruptive mutations, e.g. leading to premature start or stop codons, frameshift resulting from insertions/dele-

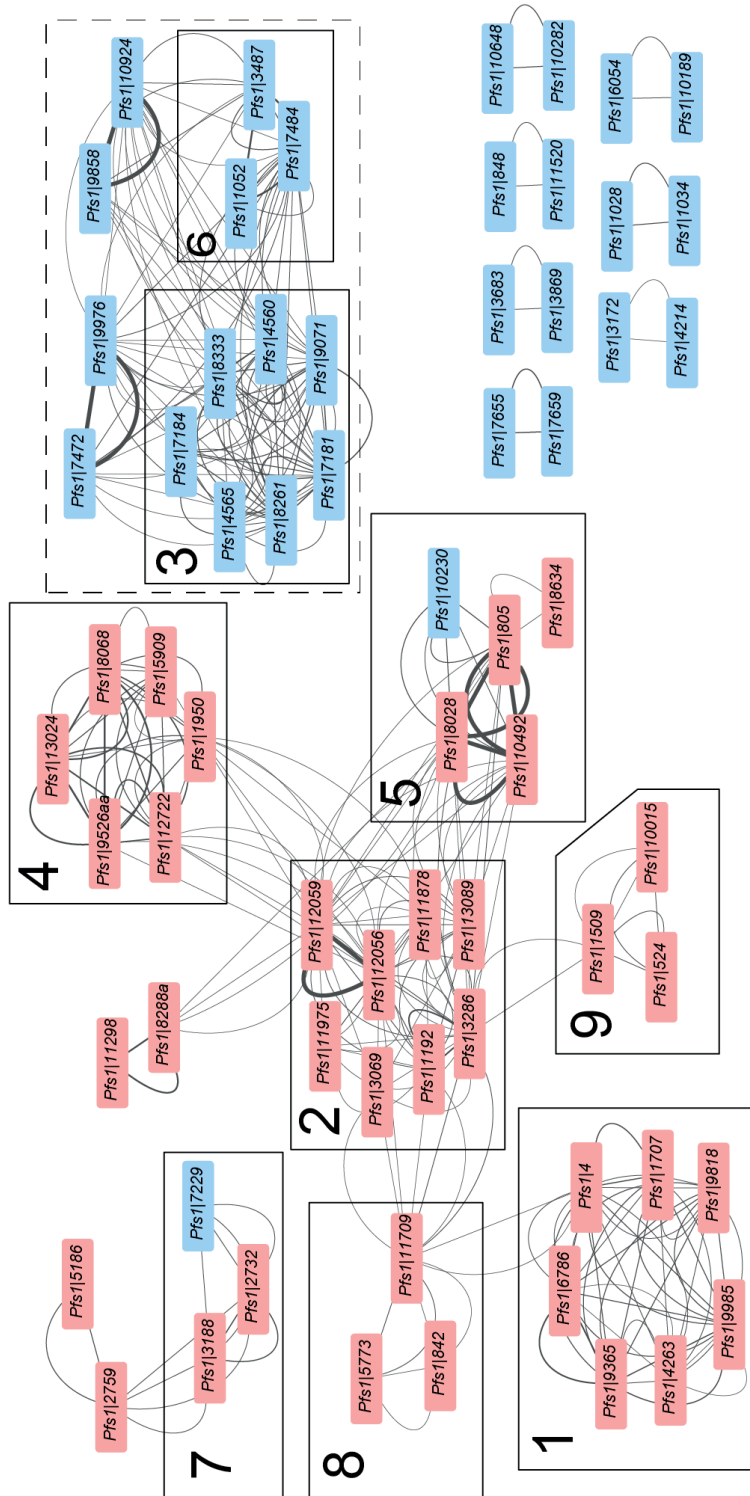


Figure 1. Protein sequence similarity network of effectors in the *Pfs* genome. Effectors residing in PGGs identified with MCL larger than 2 sequences are indicated with a box with their respective group numbers. Node labels with a red background indicate the presence of one or more WY motif, a blue background indicates that no WY domain was found in these effector proteins. Edge thickness that connects the labels are based on the $-\log(E\text{-value})$, a thicker edge indicates a more significant E-value. Effectors belonging to the CRN class are located in the dashed square.

tions. This would result in aberrant proteins and loss of effector function or effector recognition. A second important driver of positive selection is the adaptation of effectors to act optimally on plant host target molecules.

The mean of the dN/dS values for each gene over the isolates was used as a measure of selection pressure and compared between each group. The *Pfs* effectors (94 RxLRs and 8 CRN) were divided into three groups: RxLR, RxLR-like, and CRN, based on their conserved motifs. A fourth additional group consists of conserved single-core orthologs identified by BUSCO. Genes that do not belong to these aforementioned groups were labeled “other genes” (**Figure 2**).

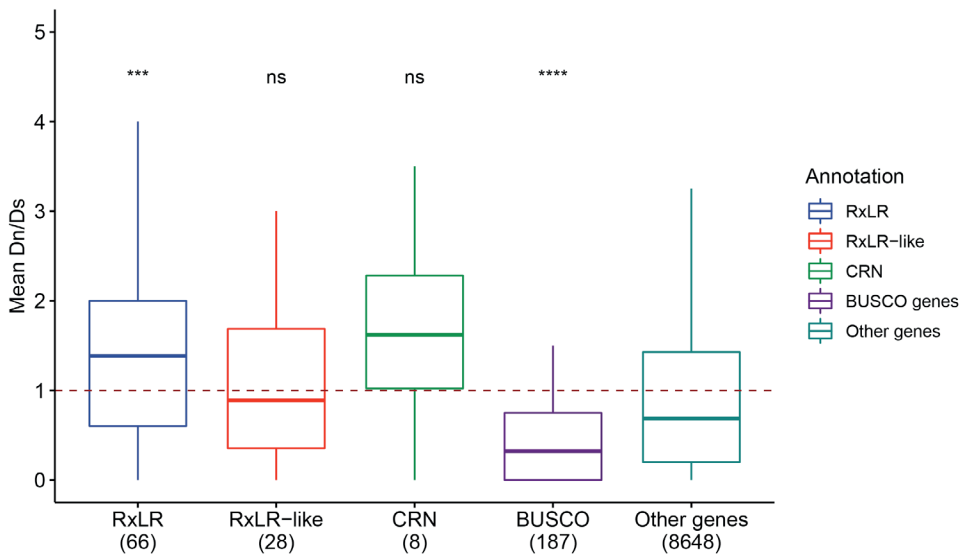


Figure 2. Comparison between the mean dN/dS values over the isolates for the RxLR, RxLR-like CRN, BUSCO, and other genes. The boxplot shows the distribution of dN/dS values for each group; the dotted line indicates neutral selection (dN/dS = 1). The number of sequences in each group is denoted between brackets. Statistical significance is determined using the Wilcoxon signed-rank test where all groups were compared to the “other genes” group, the level of significance is denoted with one or more asterisks (** = P-value $\leq 1.0e^{-2}$ and *** = P-value $\leq 1.0e^{-3}$) and “ns” for non-significant.

There is evidence for positive selection in the RxLR effector group that shows a significantly higher median dN/dS value of 1.39 (P-value = $3.1e^{-4}$) compared to the “other genes” group with a median of 0.69. Contrary to this, the RxLR-like effector group has a median dN/dS value of 0.88, which is not significantly different compared to the “other genes” group. Although the Crinkler (CRN) effectors have a median dN/dS of 1.62 which is higher than that of “other genes”, the difference was not significant in the

Wilcoxon test. This is likely due to the small number of CRN genes (8) used in this analysis. Negative selection was observed for the BUSCO genes group that has a significantly (P -value = $3.2e^{-8}$) lower median dN/dS value of 0.32 compared to the “other genes” group. That BUSCO genes are under negative selection is expected as these genes are core orthologous sequences in the protist lineage involved in essential processes in the cell. The group with the highest dN/dS values is clearly the RxLR effector group. The observed signs of positive selection are possibly caused by a combination of adaption to host targets and to overcome host resistance.

The RxLR effector genes, that showed signs of positive selection, were further analyzed by splitting them into two groups; conserved and *Pfs*-specific RxLR effector genes. When we analyzed the two effector groups separately, we find that the 19 conserved RxLR effectors have a median dN/dS value of 1.00 over the isolates, while the 47 *Pfs*-specific RxLR effectors have a mean dN/dS value of 1.83 (**Figure 3**).

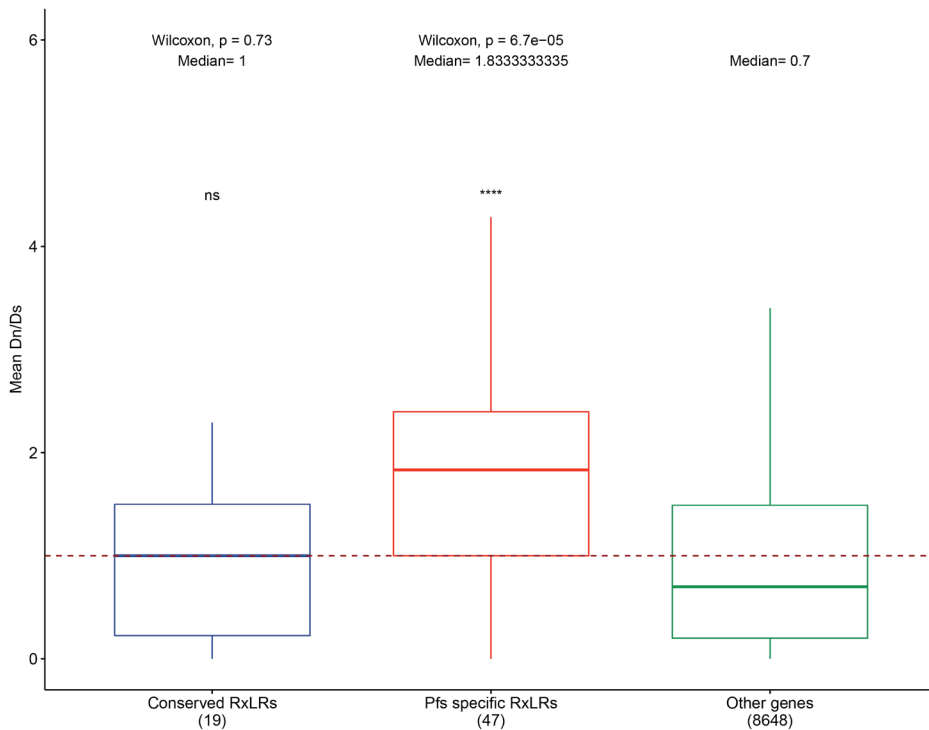


Figure 3. Comparison between the mean dN/dS values over the isolates in conserved and *Pfs* specific RxLR effectors. The boxplot shows the distribution of dN/dS values for each group; the dotted line indicates neutral selection (dN/dS = 1). The number of genes in each group is denoted between the brackets. Statistical significance is determined using the Wilcoxon signed-rank test where each group is compared to the “other genes” group (*** = $P \leq 1.0e^{-3}$, and “ns” for non-significant).

4

There is a strong signal of positive selection in the *Pfs*-specific RxLR effector genes compared to “other genes” (P -value = $3.85e^{-5}$), while the conserved RxLR effector genes do not differ significantly. The lower dN/dS of the conserved effectors could be explained by the fact that they are more ancient, occurring in different oomycete taxa. These proteins have possibly evolved an important activity in the different host plant species that requires a neutral or negative selection ($dN/dS \leq 1$) to maintain their activity.

***Pfs* isolate phylogeny based on effector genes**

To determine if effectors evolve co-linearly or at an uneven rate compared to the rest of the nuclear genome, we constructed a phylogenetic tree based on the effector nucleotide sequences in different isolates and compared the result with the phylogenetic tree based on whole-genome variants. As input for a phylogenetic tree, we used the 66 RxLR effector candidates that, as a group, showed significantly higher dN/dS ratios compared to “other genes” within the *Pfs* isolates analyzed (**Figure 2**). Based on read coverage, we found that two effector genes, *Pfs1|1792* and *Pfs1|1509*, were absent in several isolates (discussed below) and were therefore not included in the phylogenetic analysis.

For each isolate, the 64 effector coding sequences were concatenated, aligned, and used as input to construct a maximum likelihood phylogenetic tree. The reference race *Pfs1* contained 54 heterozygous variants in the 64 effector genes. In total, the alignment contained 915 homozygous and 1,414 heterozygous SNP and INDEL variants. The vast majority of variants were shared between several isolates, as only nine variants were unique to single isolates.

The virulence spectrum on the spinach differential set was also determined for the non-denominated *Pfs* isolates included in this study (**Chapter 3, Supplemental Table 1**). This information can be used to associate the virulence spectrum to differences and similarities in the genomes and effectors of different *Pfs* races and isolates.

To compare the effector-based to the genome-wide phylogeny we combined them in a tanglegram. To minimize the number of crossings, the tree based on genome-wide variants was ordered differently in the tanglegram (**Figure 4**) than the original tree based on genome-wide variants (**Chapter 3, Figure 5**). The two trees contain notable similarities. For example, in the effector-based tree, we see that distinctive isolate groups, that were previously identified in **Chapter 3**, are also present in identical monophyletic branches in the effector tree (indicated with colored boxes with labels A, B and C). Also, *US-15* (a *Pfs16* variant) clusters with *Pfs16* in one monophyletic clade in both trees as well. Moreover, *Pfs7* and isolate *NL-05* (a *Pfs7* variant, isolated in the Netherlands) are both present in one monophyletic clade in both trees.

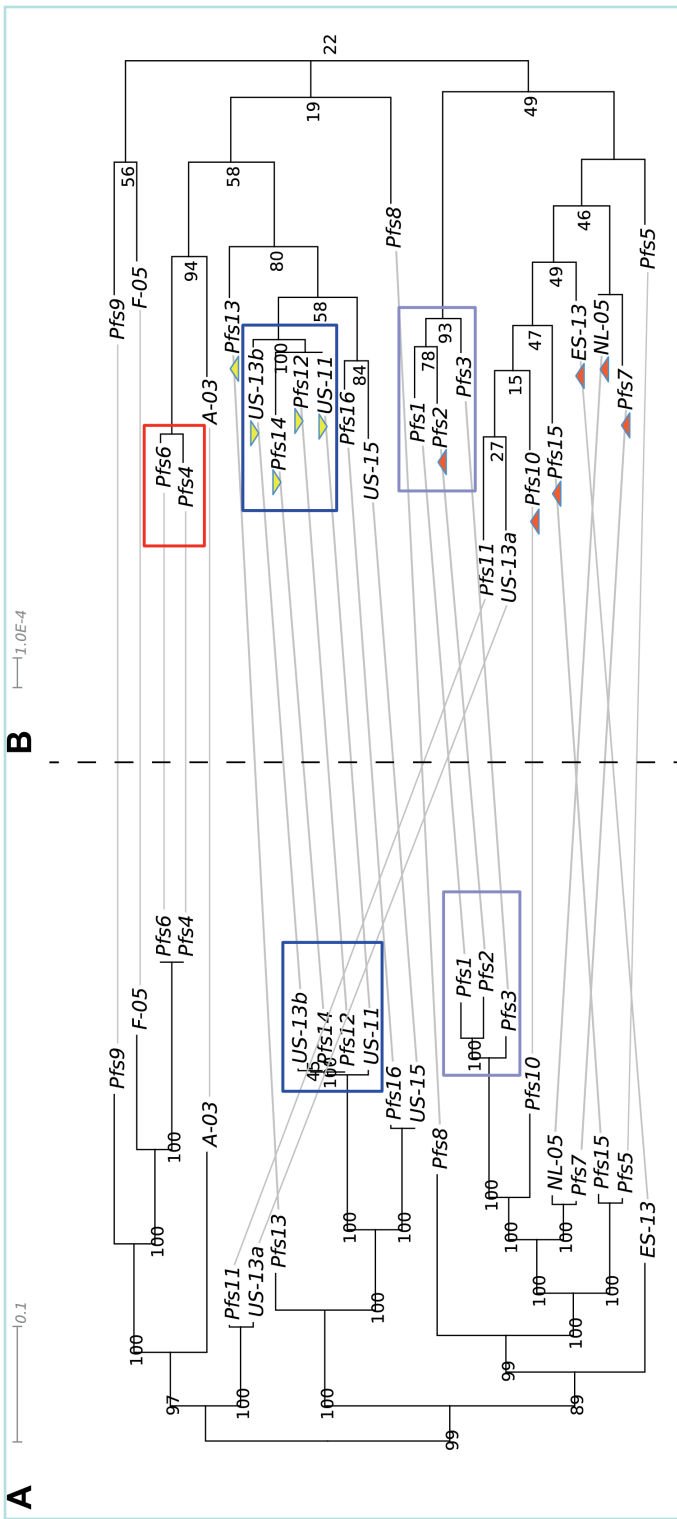


Figure 4. Tanglegram comparing the phylogenetic tree based on genome-wide variants and 64 RxLR effector genes **A**. Maximum likelihood phylogenetic tree based on genome-wide variants identified in the different *Pfs* isolates (modified from the original tree in **Chapter 3**). **B**. Maximum likelihood phylogenetic tree based on the concatenated alignment of 64 RxLR effector genes. Both trees were constructed using the RAxML nucleotide substitution model and 500 bootstraps to estimate branch confidence. The colored boxes indicate the three distinct groups that were observed in the phylogenetic analysis of the different *Pfs* isolates (**Chapter 3**). The triangles indicate *Pfs* isolates in which either gene *Pfs*1792 (red) or *Pfs*11509 (yellow) are missing. The scale bars (different for both trees) indicates the number of nucleotide substitutions per site.

As an estimate of similarity between the trees in **Figure 4**, we determined the number of instances where nodes or clades of the phylogram that differ must be repositioned to obtain the same tree structure. This measure is also denoted as the Subtree Prune and Regraft (SPR) distance [48]. The two trees could be aligned with an SPR distance of 6, while the highest possible SPR distance between these trees is 18. This indicates that the overall structure in both trees is similar but differs in a few clades. Especially the position of the clade containing *Pfs11* and *US-13a* differs significantly between the trees. In the genome-wide-based tree, this clade is more closely related to *A-03*, whereas in the effector-based tree *Pfs11* and *US-13a* are grouped closer to *Pfs10* and *Pfs15*. This latter grouping fits with the fact that the virulence spectrum of *Pfs11* is most similar to that of *Pfs10*. Although several other noticeable differences between both trees can be observed, the bootstrap confidence is relatively low for several clades in the effector-based tree, not allowing confident conclusions. Overall, the effector-based tree mostly follows the whole genome-based tree suggesting that *Pfs* effector genes do not evolve very differently from the rest of the genome.

Effector gene deletions associated with the breaking of R-loci

The absence of an effector gene in resistance-breaking isolates provides a strong hint that the encoded effector could be recognized through a matching spinach resistance protein. Two effector genes, *Pfs1|1792* and *Pfs1|1509*, had besides a few spurious reads, no coverage of sequence reads in certain *Pfs* isolates and was therefore scored as absent (**Figure 5**).

Effector gene *Pfs1|1792* is absent in *Pfs* races 2, 7, 10, 13, and 15 (**Table 1**). We experimentally confirmed, by PCR amplification, the presence of *Pfs1|1792* in races 1, 3, 5, and 8, and its absence in races 10 and 13 (**Supplemental Figure 1**).

The presence or absence of the *Pfs1|1792* gene shows a strong but incomplete association with the breaking of *RPF3*. Only the two *RPF3*-breaking races *Pfs4* and *Pfs6* still possess the *Pfs1|1792* gene. Therefore, we looked in *Pfs4* and *Pfs6* if there are allelic variants, with non-synonymous mutations, of *Pfs1|1792* present. As *Pfs* is a diploid organism, variants that overcome R-loci are expected to be homozygous as recognition is a dominant trait. However, no homozygous non-synonymous (HNS) variants were found in effector gene *Pfs1|1792* in *Pfs4* and 6. It could be that these races suppress the immune response to effector *Pfs1|1792* by another effector, or that *Pfs1|1792* is not expressed in these races (not analyzed).

Pfs effector gene *Pfs1|1509* was found to be absent in *Pfs12*, *Pfs14* and isolates *US-11* and *US-13b* (**Supplemental Figure 2, Table 1**) For

Peronospora effusa effector polymorphisms associated with the breaking of resistance loci in spinach

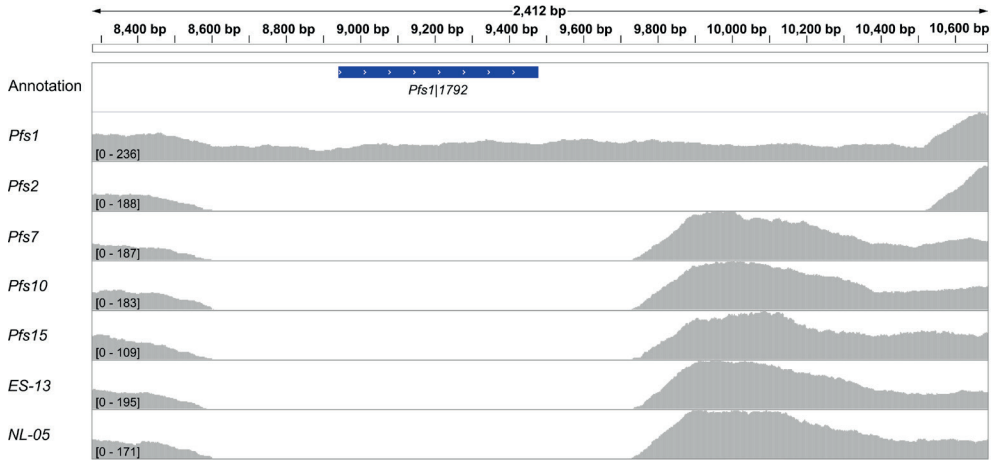


Figure 5. Read coverage on a 2.4 kb region on contig Pb_140 for different *Pfs* races that are missing the *Pfs1|1792* gene, gene location is indicated by the blue bar. As a reference, the coverage for this segment in *Pfs1* is included. The coverage range is denoted on the left side between brackets, coverage is not to scale. The observed deletion in *Pfs2* is considerably larger than in the other isolates that are missing effector gene *Pfs1|1792*. In other *Pfs* races, the coverage profile is similar to that observed for *Pfs1*.

Table 1. Overview of the isolates in which the *Pfs* genes *Pfs1|1792* and *Pfs1|1509* are present or absent. Light red indicates that the gene is scored as absent, and cadetblue indicates it is present in the genome based on coverage (per base coverage over the gene and normalized) in the different races. The virulence spectrum of the *Pfs* races on the differential spinach lines with the resistance locus *RPF3* is indicated in the last column for gene *Pfs1|1792*.

| <i>Pfs1 1792</i> | | | | <i>Pfs1 1509</i> | | |
|------------------|------------------|----------|-------------|------------------|------------------|----------|
| Race | Presence/absence | Coverage | <i>RPF3</i> | Race | Presence/absence | Coverage |
| <i>Pfs1</i> | Present | 0.644 | Resistant | <i>Pfs1</i> | Present | 0.576 |
| <i>Pfs2</i> | Absent | 0.004 | Susceptible | <i>Pfs2</i> | Present | 0.647 |
| <i>Pfs3</i> | Present | 0.592 | Resistant | <i>Pfs3</i> | Present | 0.303 |
| <i>Pfs4</i> | Present | 0.713 | Susceptible | <i>Pfs4</i> | Present | 0.328 |
| <i>Pfs5</i> | Present | 0.568 | Resistant | <i>Pfs5</i> | Present | 0.522 |
| <i>Pfs6</i> | Present | 0.768 | Susceptible | <i>Pfs6</i> | Present | 0.292 |
| <i>Pfs7</i> | Absent | 0.003 | Susceptible | <i>Pfs7</i> | Present | 0.496 |
| <i>Pfs8</i> | Present | 0.567 | Resistant | <i>Pfs8</i> | Present | 0.539 |
| <i>Pfs9</i> | Present | 0.554 | Resistant | <i>Pfs9</i> | Present | 0.321 |
| <i>Pfs10</i> | Absent | 0.005 | Susceptible | <i>Pfs10</i> | Present | 0.612 |
| <i>Pfs11</i> | Present | 0.526 | Resistant | <i>Pfs11</i> | Present | 0.306 |
| <i>Pfs12</i> | Present | 0.564 | Resistant | <i>Pfs12</i> | Absent | 0.001 |
| <i>Pfs13</i> | Absent | 0.006 | Susceptible | <i>Pfs13</i> | Present | 0.649 |
| <i>Pfs14</i> | Present | 0.732 | Resistant | <i>Pfs14</i> | Absent | 0.000 |
| <i>Pfs15</i> | Absent | 0.008 | Susceptible | <i>Pfs15</i> | Present | 0.495 |
| <i>Pfs16</i> | Present | 0.686 | Resistant | <i>Pfs16</i> | Present | 0.257 |

these races and isolates there were no reads present giving coverage of the complete but small contig (3,416 bp). We also found that *Pfs* races *Pfs8*, 9, 10, 11, 15 and 16 contain one or more HNS variants in effector gene *Pfs1|1509*. However, a clear association with the differential set could not be made based on the presence or absence of this gene or HNS variant in other *Pfs* races. We also found this gene to be absent in the field isolates *US-11* and *US-13b*. Together with *Pfs12* and 14, *Pfs* isolates *US-11* and *US-13b* form a monophyletic clade (defined as genetic group C in **Chapter 3**) (**Figure 4**). These isolates and races were all isolated from the same geographical location (USA). This suggests that effector *Pfs1|1509* is not essential to establish a successful infection and a single deletion event in a common ancestral isolate resulted in the loss of this gene in these 4 isolates.

Contrary to this, *Pfs* isolates where effector gene *Pfs1|1792* is missing are more dispersed throughout the tree in **Figure 4B**. This suggests that the loss of *Pfs1|1792* might have occurred more than once during the evolution of new *Pfs* races. The independent loss is supported by different deletion sizes in the region where effector *Pfs1|1792* resides. All isolates that are missing *Pfs1|1792* except for *Pfs2* share the same deletion size of approximately 1,150 bp (**Figure 5**). The deletion in *Pfs2* starts almost at the same position, corresponding to position 8.6 kb on *Pfs1* contig Pb_140, but ends at 10.5 kb, whereas in the other isolates it ends at 9.7 kb, making the deletion in *Pfs2* almost 0.8 kb longer.

4

Polymorphisms in effectors of closely-related *Pfs* races

Based on the genome polymorphisms in the analyzed *Pfs* isolates and denominated races in **Chapter 3**, we observed three distinct genetic groups (**Chapter 3, Figure 4**). Group A consists of *Pfs1*, 2, and 3, group B consists of *Pfs4* and 6 and group C consist of *Pfs12* and 14 as well as isolates *US-11* and *US-13b*. Although the isolates and races within a group are genetically closely related, they are capable of breaking different spinach *RPF* resistance loci. We examined the effector polymorphisms between two closely-related races in each group and their possible link to the breaking of one or more *RPF* genes. A comparison of HNS variants in 113 identified candidate host-translocated effector genes (RxLR, RxLR-like, and CRN) was obtained by subtracting the HNS variants identified in the breaking race from those of the non-breaking race. As a result, only HNS variants that differ between the two selected *Pfs* isolates remain (**Table 2**).

Table 2. Summary of the number of HNS variants, identified after comparison of two genetically related *Pfs* races, in the whole genome, in coding sequences (genes), secreted protein genes (secretome) and effector genes. Note that coding sequences can contain more than one HNS variant.

| | Total HNS variants | In <i>Pfs</i> Genes | In secretome | In effectors |
|------------------------------|--------------------|---------------------|--------------|--------------|
| <i>Pfs2</i> vs <i>Pfs3</i> | 1,288 | 703 | 41 | 8 |
| <i>Pfs3</i> vs <i>Pfs2</i> | 1,517 | 832 | 55 | 9 |
| <i>Pfs6</i> vs <i>Pfs4</i> | 540 | 274 | 21 | 4 |
| <i>Pfs12</i> vs <i>Pfs14</i> | 870 | 351 | 20 | 3 |

Group A: *Pfs3* vs *Pfs2*

Pfs3 (as well as all other races and isolates in this study except *Pfs1* and 2) has overcome resistance mediated by *RPF5*. On the other hand, *Pfs2* has broken resistance loci *RPF3* and *RPF6*. Reciprocal comparison of *Pfs2* and 3, can provide candidates effectors with HNS variants that are associated with the breaking of *RPF5* (*Pfs3*) or a combination of *RPF3* and *RPF6* (*Pfs2*). We identified nine effectors in *Pfs3* that contain HNS variants that were not present in the effector genes in the genome of *Pfs2* (**Table 3**).

Besides *Pfs3*, most other races have also broken *RPF5* except for *Pfs1* and *Pfs2*. We see that *Pfs* gene *Pfs1*|3839 contains an HNS variant (leading to substitution R103H) that occurs in all *Pfs* isolates capable of breaking *RPF5*. Consequently, the protein encoded by effector gene *Pfs1*|3989, more precisely the variant occurring in *Pfs1* and *Pfs2*, is a good candidate to be recognized by the *RPF5* resistance protein.

The other way around, *Pfs2* is capable of breaking resistance genes *RPF3* and *RPF6*. *RPF3* is also broken by *Pfs4*, 6, 7, 10, 13 and 15. *RPF6* is broken by a different set of races, namely *Pfs8*, 9, 10, 12, 14 and to some extent by *Pfs5*, 6 and 13 (these races show an intermediate disease phenotype on the differential set). The comparison revealed eight effector sequences that contained one or more HNS variants in *Pfs2* that are not present in *Pfs3* (**Table 4**).

Table 3. Differential HNS variants and their effect on the protein sequence of the effectors containing these variants. The *RPF* loci (*RPF5*) that *Pfs3* has broken compared to *Pfs2* and corresponding virulence spectrum (– = resistant and + = susceptible) are indicated in the last column.

| Race | <i>Pfs1</i> 1257 | <i>Pfs1</i> 6706 | <i>Pfs1</i> 8777 | <i>Pfs1</i> 9526aa | <i>Pfs1</i> 11883 | <i>Pfs1</i> 11975 | <i>Pfs1</i> 1052 | <i>Pfs1</i> 1192 | <i>Pfs1</i> 3989 | <i>RPF5</i> |
|--------------|--------------------------|-----------------------------|-----------------------------|---------------------|--------------------|------------------------|--|--|--------------------------|-------------|
| <i>Pfs1</i> | | | | | | | | | | - |
| <i>Pfs2</i> | G88E, C672G, 11070V | S44P, P91L, C213G, Q217E | | | | | I5T | | | - |
| <i>Pfs3</i> | F964L, 11070V | N595D | T393P | M15L, Q58H | P175S | | I5T, G764S, N767S, Trp772C | N153K | R103H | + |
| <i>Pfs4</i> | F964L | V353A, N595D | T393P | | D66N, Q157E | | Y620N, N752D, K756Q, G768C, P792S, E798Q, R201S | N153K, H400L, D404H, G408fs, E409fs, M639L, Y643D, V646M, K647R, K648A | R103H | + |
| <i>Pfs5</i> | P1002S, 11070V | | | Q58H | | | | N153K | R103H | + |
| <i>Pfs6</i> | F964L | V353A, N595D | T393P | | | | | N153K, H400L, D404H, G408fs, E409fs, N627S, I629S, M639L, K648A | R103H | + |
| <i>Pfs7</i> | 11070V | N595D | | | P175S | | | N153K | R103H | + |
| <i>Pfs8</i> | F964L, P1002S, 11070V | V353A, N595D | T393P | | P175S | | T354fs, N767S, R201S | R3K, N153K, N358D | N47K, R103H, Q435R | + |
| <i>Pfs9</i> | 11070V | V353A, N595D | S44P, P91L, C213G, Q217E | | P175S | | V81A | N153K | R103H | + |
| <i>Pfs10</i> | C672G | N595D | | Q58H | | | | N153K | R103H | + |
| <i>Pfs11</i> | G88E, C672G, 11070V | N595D | | | P175S | | V81A | | R103H | + |
| <i>Pfs12</i> | G88E, C672G, 11070V | N595D | T393P | | P175S | | | N153K | R103H | + |
| <i>Pfs13</i> | G88E, C672G, 11070V | N595D | T393P | | P175S | V394fs, E91K, S578P | | R3K, N153K | R103H | + |
| <i>Pfs14</i> | G88E, C672G, 11070V | N595D | T393P | | P175S | S722R | | N153K | R103H | + |
| <i>Pfs15</i> | C672G | N595D | | M15L, Q58H | | | | N153K | R103H | + |
| <i>Pfs16</i> | | N595D | T393P | Q58H | P175S | | R201S, Y620N, N752D, K756Q, G768C, P792S, E798Q | N153K | R103H | + |

Table 4. Differential HNS variants and their effect on the protein sequence of the effectors containing these variants. The *RPF* loci (*RPF3* and *RPF6*) that *Pfs2* has broken compared to *Pfs3* and corresponding virulence spectrum (– = resistant and + = susceptible) are indicated in the last column.

| Race | <i>Pfs11257</i> | <i>Pfs112002</i> | <i>Pfs113069</i> | <i>Pfs118777</i> | <i>Pfs11805</i> | <i>Pfs115186</i> | <i>Pfs117484</i> | <i>Pfs111709</i> | <i>RPF3</i> | <i>RPF6</i> |
|--------------|-----------------------|---------------------------|---------------------------|--------------------------|-----------------|------------------|------------------|---|-------------|-------------|
| <i>Pfs1</i> | | | | | | | | | - | - |
| <i>Pfs2</i> | G88E, C672G, I1070V | S82T | T633fs | S44P, P91L, C213G, Q217E | A43S | P182L | E201K | S27L | + | + |
| <i>Pfs3</i> | F964L, I1070V | | | T393P | | | | | - | - |
| <i>Pfs4</i> | F964L | S83G, K234I, T633fs | S83G, K234I, T633fs | T393P | S304T | E201K | | G619fs, G619E, E817D, S973A, H977Q, R1043Q, Q1069H, K1072fs, Ter1075Qext*?, D1039G | + | - |
| <i>Pfs5</i> | P1002S, I1070V | | | | | | | | - | (-) |
| <i>Pfs6</i> | F964L | S83G, K234I, T633fs | S83G, K234I, T633fs | T393P | A43S, S304T | | | G619fs, G619E, E817D, S973A, H977Q, R1043Q, Ter1075Qext*? | + | - |
| <i>Pfs7</i> | I1070V | | | | A43S | | | | + | - |
| <i>Pfs8</i> | F964L, P1002S, I1070V | | T633fs | T393P | A43S | | | | - | + |
| <i>Pfs9</i> | I1070V | S83G, S75P, K234I, T633fs | S83G, S75P, K234I, T633fs | S44P, P91L, C213G, Q217E | A43S | | | | - | + |
| <i>Pfs10</i> | C672G | | | | A43S | | | | + | + |
| <i>Pfs11</i> | G88E, C672G, I1070V | | S83G, K234I, T633fs | | | | | | - | - |
| <i>Pfs12</i> | G88E, C672G, I1070V | K36N | S83G, K234I, T633fs | T393P | A43S | | | G619fs, G619E, E817D, S973A, H977Q, R1043Q, Q1069H, K1072fs, Ter1075Qext*? | - | + |
| <i>Pfs13</i> | G88E, C672G, I1070V | | S83G, K234I, T633fs | T393P | A43S | E201K | | | + | (-) |
| <i>Pfs14</i> | G88E, C672G, I1070V | K36N | S83G, K234I, T633fs | T393P | A43S | | | G619fs, G619E, E817D, S973A, H977Q, R1043Q, Ter1075Qext*? | - | + |
| <i>Pfs15</i> | C672G | | | | A43S | | | | + | - |
| <i>Pfs16</i> | | | | T393P | A43S | | | G619fs, G619E, E817D, S973A, H977Q, R1043Q, Q1069H, K1072fs, Ter1075Qext*?, V914F, A918, N919insV, N919K, A920G, D1039G, S27L | - | - |

As mentioned before, some *Pfs* races are also capable of breaking either the *RPF3* or *RPF6* resistance loci. However, we did not identify differential HNS variants present in all these specific *Pfs* races that can be directly associated with the breaking of the *RPF3* or *RPF6* resistance loci. Notably, effector gene *Pfs1|11709* has a stop lost variant in several races, resulting in much longer effector protein.

Group B: *Pfs6* vs *Pfs4*

Pfs6 has overcome resistance mediated by *RPF3*, *RPF4*, and *RPF5*, while *Pfs4* has overcome two of these: *RPF3* and *RPF5*. This means that *Pfs6* has broken *RPF4*. Almost all *Pfs* races, from *Pfs5* until 16 except for *Pfs15*, have also broken *RPF4* resistance. We identified two effectors (*Pfs1|7655* and *Pfs1|805*) in *Pfs4* that contain one HNS variant that is not present in *Pfs6*, and one effector (*Pfs1|4*) that contain two HNS variants in *Pfs6* that are not present in *Pfs4* (**Table 5**).

Table 5. Differential HNS variants and their effect on the protein sequence of the effectors containing these variants. The *RPF* loci (*RPF4*) that *Pfs6* has broken compared to *Pfs4* and corresponding virulence spectrum (– = resistant and + = susceptible) is indicated in the last column.

| Race | <i>Pfs1 7655</i> | <i>Pfs1 4</i> | <i>Pfs1 805</i> | <i>Pfs1 1192</i> | <i>RPF4</i> |
|--------------|------------------|---------------------|-----------------|--|-------------|
| <i>Pfs1</i> | | | | | - |
| <i>Pfs2</i> | F64L | E344K, D346A | A43S | | - |
| <i>Pfs3</i> | F64L | E344K, D346A | | N153K | - |
| <i>Pfs4</i> | F64L | | S304T | N153K, H400L, D404H, G408fs, E409fs, M639L, Y643D, V646M, K647R, K648A | - |
| <i>Pfs5</i> | | E344K, D346A | | N153K | + |
| <i>Pfs6</i> | | E344K, D346A | S304T, A43S | N153K, H400L, D404H, G408fs, E409fs, N627S, I629S, M639L, K648A | + |
| <i>Pfs7</i> | | E344K, D346A | A43S | N153K | + |
| <i>Pfs8</i> | | E344K, D346A | A43S | R3K, N153K, N358D | + |
| <i>Pfs9</i> | F64L | E344K, D346A | A43S | N153K | + |
| <i>Pfs10</i> | F64L | E344K, D346A | A43S | N153K | + |
| <i>Pfs11</i> | F64L | E344K, D346A | | | + |
| <i>Pfs12</i> | F64L | E286K, E344K, D346A | A43S | N153K | + |
| <i>Pfs13</i> | F64L | E286K, E344K, D346A | A43S | R3K, N153K | + |
| <i>Pfs14</i> | F64L | E286K, E344K, D346A | A43S | N153K | + |
| <i>Pfs15</i> | | E344K, D346A | A43S | N153K | - |
| <i>Pfs16</i> | F64L | E286K, E344K, D346A | A43S | N153K | + |

None of the four effectors show HNS variants that are associated with the breaking of *RPF4* resistance. In addition, we also determined if a direct link with the breaking of *RPF4* could be found with other secreted proteins. HNS variants that differ between *Pfs6* compared to *Pfs4* were found in 17 other proteins. However, none of these HNS variants showed association with the breaking of *RPF4* over all races.

Group C: *Pfs14* vs *Pfs12*

Pfs14 and *12* are both in a monophyletic clade (group C) together with field isolates *US-11* and *US-13b*, which both have the same differential interaction as *Pfs14* and can thus be considered race 14 isolates (**Chapter 3 Supplemental Table 2**). Compared to *Pfs12*, *Pfs14* has overcome the resistance of spinach cultivar Pigeon that contains both the *RPF2* and *RPF9* resistant loci. However, *Pfs12* has also overcome the single *RPF2* locus in NIL2, so *Pfs14* has specifically broken *RPF9*. The comparison revealed eight effector sequences that contained one or more HNS variants in *Pfs14* that are not present in *Pfs12* (**Table 6**).

Table 6. Differential HNS variants and their effect on the protein sequence of the effectors containing these variants. The *RPF* loci (*RPF9*) that *Pfs14* has broken compared to *Pfs12* and corresponding virulence spectrum (– = resistant and + = susceptible) is indicated in the last column.

| Race | <i>Pfs1</i> 3188 | <i>Pfs1</i> 3210 | <i>Pfs1</i> 11975 | <i>RPF9</i> |
|--------------|--|------------------------|-------------------------------|-------------|
| <i>Pfs1</i> | | | | - |
| <i>Pfs2</i> | | | | - |
| <i>Pfs3</i> | | | | - |
| <i>Pfs4</i> | | E117K, T131N, G139S | | - |
| <i>Pfs5</i> | | | | - |
| <i>Pfs6</i> | | E117K, T131N, G139S | | - |
| <i>Pfs7</i> | A47T, A136_ E137insGPSQPTLAEPSQPTLA | | | - |
| <i>Pfs8</i> | A47T | | | - |
| <i>Pfs9</i> | A47T | | E91K, V394fs, S578P, N660D | - |
| <i>Pfs10</i> | | | | - |
| <i>Pfs11</i> | | | | - |
| <i>Pfs12</i> | A47T | | | - |
| <i>Pfs13</i> | A47T | | E91K, V394fs, S578P | - |
| <i>Pfs14</i> | A47T, A136_ E137insGPSQPTLAEPSQPTLA | E117K, T131N, G139S | S722R | + |
| <i>Pfs15</i> | | | | + |
| <i>Pfs16</i> | | | | + |

We found that effector gene *Pfs1|3188* in *Pfs14* contains an insertion of 16 amino acids (large INDEL variant) that is not present in *Pfs12*. This insertion (including 4 prolines) is likely to have a large effect on the structure of this protein. Strikingly, the same insertion of codons is also found in *Pfs7*, a race that is not genetically close to *Pfs14*. None of the HNS variants found in the three effectors of *Pfs14* is present in *Pfs16* that can also break the *RPF9* resistance. Therefore, we cannot clearly associate HNS variants in these effectors with the breaking of *RPF9*.

The relationship between R-loci and HNS variants in effector alleles of all races

Of the effectors that were identified in the genome of *Pfs1* (99 RxLR and 14 CRN), 41 effectors did not contain any DNA polymorphisms and are thus fully conserved in all races and isolates tested (**Supplemental Table 7**). Of these non-polymorphic effectors, ten were also found to be conserved in other species. The other 72 effectors were found to be polymorphic within *Pfs* and contained a total of 941 homozygous variants in all *Pfs* races under study. These variants can be further divided based on their predicted impact on the coding sequences. In the polymorphic effectors, we found 871 HNS variants of which SNPeffect annotated 70 as high impact variants (e.g. frameshift variants, premature stop codons, or stop codons lost), and 468 as moderate impact variants (e.g. missense variants or in-frame deletions). Furthermore, 3613 heterozygous non-synonymous variants were found in all effectors of which 139 were high impact and 1889 moderate impact variants (**Supplemental Table 8**).

Besides the comparison of HNS variants in effector sequences of closely related races; we also compared the presence of HNS variant(s) in all races in relation to the differential set of spinach (**Supplemental Table 9**). With this comparison, we were able to couple three *Pfs* effectors directly to the breaking of a single R-locus. HNS variants in effector gene *Pfs1|3839* in the different races could be directly linked to *RPF5*. This association was already found in the direct comparison between *Pfs3* and 2 (**Table 3**). We also found that HNS variants in effector gene *Pfs1|6786* and *Pfs1|9985* could also be directly associated with the breaking of *RPF2* resistance. Multiple amino acid substitutions were found in the effectors *Pfs1|6786* and *Pfs1|9985* that are fully associated with the breaking of *RPF2* resistance (**Table 7**).

We also looked if HNS variants in effectors could be directly associated with differential spinach lines that contain more than one R-locus (Pigeon, Caladonia, and Meerkat). We found that effector gene *Pfs1|11709*, *Pfs1|13089*, *Pfs1|11982* and *Pfs1|8634* contain HNS variants in *Pfs16*, the only race that is virulent on cultivar Meerkat, containing *RPF2* and *RPF10* resistance (**Chapter 3 Supplemental Table 1**). However, any HNS vari-

ant in an effector protein of *Pfs16* could be responsible for the breaking of resistant loci *RPF2* and *RPF10*.

Effector genes *Pfs1|6786* and *Pfs1|9985*, that are associated to the breaking of *RPF2* resistance, are homologous to each other and found together with five other effectors in PGG1 (**Figure 1**). Also, *Pfs* gene *Pfs1|1057* is part of PGG1 but was previously not identified as an effector (**Chapter 1**) as an SP was missing. Manual re-annotation revealed an SP making *Pfs1|1057* an additional RxLR effector.

Table 7. *Pfs* effectors with HNS variants (indicated by light red rows) in isolates that can be directly linked to the *RPF2* loci. The amino acid substitution(s) and location are indicated in the second column, and corresponding disease phenotype (– = resistant and + = susceptible) linked to a specific R-loci are indicated in the third column.

| Race | <i>Pfs1 6786</i> | <i>Pfs1 9985</i> | <i>RPF2</i> |
|--------------|------------------------------|------------------|-------------|
| <i>Pfs1</i> | | | - |
| <i>Pfs2</i> | | | - |
| <i>Pfs3</i> | | | - |
| <i>Pfs4</i> | | | - |
| <i>Pfs5</i> | | | - |
| <i>Pfs6</i> | | | - |
| <i>Pfs7</i> | | | - |
| <i>Pfs8</i> | | | - |
| <i>Pfs9</i> | | | - |
| <i>Pfs10</i> | | | - |
| <i>Pfs11</i> | D26G, G55S, M60I | L115V, F273I | + |
| <i>Pfs12</i> | D26G, S51N, F52L, G55S, M60I | L115V, F273I | + |
| <i>Pfs13</i> | D26G, S51N, F52L, G55S, M60I | L115V, F273I | + |
| <i>Pfs14</i> | D26G, S51N, F52L, G55S, M60I | L115V, F273I | + |
| <i>Pfs15</i> | | | - |
| <i>Pfs16</i> | D26G, S51N, F52L, G55S, M60I | L115V, F273I | + |

All eight effectors belonging to family PGG1 share two or more conserved domains after the RxLR motif. Overall these proteins show high levels of homology over the entire sequence, and in particular in the RxLR and EER/DER motifs, and (L)WY repeats (**Supplemental Figure 3**). Interestingly, the F273I variant in effector *Pfs1|9985*, that associated with the breaking of *RPF2*, is located in an LWY motif and could potentially affect the tertiary structure of the LWY motif, and consequently affect effector function and recognition.

Although these effectors are homologous to each other, an ortholog in another species identified was not found with the BBH approach (**Sup-**

plemental Table 2). However, a distant ortholog was found in *Phytophthora cactorum* (PC110_g16699) [49] and used as an outgroup in our phylogenetic tree to show the relatedness between these effectors (Figure 6). The order of the motifs/domains is conserved in all family members, except for *Pfs1|1707* that lacks an LWY motif. Notably, the *Pfs1|4* protein contains an extra copy of the WY and LWY domain and is almost twice as long in the effector domain.

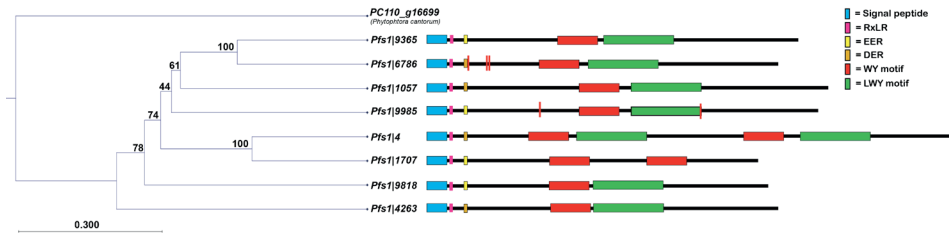


Figure 6. Phylogenetic relationships and domain organization of eight different RxLR effectors in *Pfs* belonging to the family PGG1. The tree was constructed in CLC, using the UPGMA algorithm. An ortholog in *Phytophthora cactorum* was used as an out-group. Their respective domains and approximate location within the proteins are indicated at the right-hand side of the node label. The red lines in RxLR effectors *Pfs1|6786* and *Pfs1|9985* are polymorphisms that correspond directly with the breaking of *RPF2* (Table 2).

Discussion

The downy mildew pathogen *Pfs* is well known for its rapid adaptation to resistance in the spinach host. Typically, newly introduced resistance in spinach is overcome in a short period [35], as exemplified by the development of 14 new races over the last 28 years (1990 to 2018), the denominated races *Pfs4* to *17*. Due to the lack of genomic resources for this pathogen, it was not possible to determine how genes in the *Pfs* genome adapt to overcome the introduced resistance traits in spinach. Current sequencing technology now made it feasible to sequence multiple different isolates. We chose to sequence 16 denominated races (*Pfs1* to *16*) and eight additional field isolates. Of these eight field isolates, four had an identical virulence spectrum as one of the denominated races, whereas four showed a deviating virulence spectrum on the differential set.

Since host-translocated effectors play a crucial role in the successful infection of the host, we determined the dN/dS ratio of candidate effectors over the isolates as a measure of evolutionary selection pressure. We observed that host-translocated effector genes display significantly more non-synonymous polymorphisms compared to the “other genes” group in the *Pfs* genome and are under positive selection. The uneven selection pressure on effector genes compared to other genes in the genome is also found in other oomycetes. Most proteins in the oomycete genomes

are under neutral or purifying selection. Neutral or purifying selection conserves the amino-acid sequence of proteins, and in extent their functions. However, in several oomycete species, it was shown that effectors accumulate non-synonymous variants at an uneven rate, especially those that reside in repeat-rich gene-poor regions of the genome [21; 23; 24]. In contrast, *Pfs* effectors belonging to the RxLR-like group had a median dN/dS value of 0.88. This suggests that these RxLR-like sequences have not been under selection pressure to change, implying that they are not recognized in the host, or are not evolving to adapt to host target proteins.

When we split the RxLR group into conserved effectors containing orthologs in other oomycete species and *Pfs*-specific effectors we see that the conserved effectors have a lower dN/dS ratio than the *Pfs*-specific effectors. A study in *Phytophthora infestans* recently selected 18 conserved RxLR effector genes that were highly expressed during the early stages of pathogen infection and observed that many of them showed defense suppressive activities [50]. This indicates that conserved effectors are likely involved in important activities in the oomycete, which requires negative or neutral selection to maintain their function. Conserved effectors are useful to explore further, as breeding for recognition of these proteins could provide long-term durable resistance against the pathogen [51].

The dN/dS values below 1 for the “other genes” group and the BUSCO genes group in the *Pfs* genome are expected as these genes are under purifying selection. These results also validate our approach in which we take the median dN/dS values for each gene over the isolates. Although the comparison of median dN/dS values shows a clear difference between the different groups, we also observed several outliers with a very high dN/dS value, especially in the “other genes” group. This is mainly the result of frameshift variants caused by an out of frame insertion or deletion. Frameshifts result in a completely different translation from the original in a large part of the protein and often the premature termination of translation caused by stop codons. However, these outliers have a relatively small influence on the median value which is therefore preferred over the mean in this analysis.

To determine the duplication and diversification between effector sequences we grouped the *Pfs* effectors by Markov Clustering and also clustered the effector based on $-\log(E\text{-values})$. Both analyses clearly showed five distinct effector clusters that contained more than two effector sequences each. Each effector in a PGG is likely derived from a common ancestral effector gene that has duplicated and diversified during the evolution of the *Pfs* pathogen. Effector proteins present in these PGGs mostly possessed the conserved WY motif.

An extremely large superfamily of effectors was identified in the *Phytophthora* species *P. sojae* and *P. ramorum* consisting of > 370 can-

didate effector genes with sequence similarity to *P. sojae* RxLR effector Avr1b in each of the species. The identified avirulence homolog (*Avh*) genes likely evolved from a common ancestor by rapid duplication and divergence. Many *Avh* proteins contain conserved C-terminal motifs (termed W, Y, and L) [24], similar to what we observed in *Pfs*.

The loss of an effector that is recognized by an R-gene in the host could prevent the recognition of the pathogen and ensure the successful progression of the infection. We identified two effectors (*Pfs1|1509* and *Pfs1|1792*) that were missing in several isolates. *Pfs* effector *Pfs1|1792* shows a strong correlation with the breaking of *RPF3* resistance in spinach. A strong association with *RPF3*-breaking could not be established as the virulent races *Pfs4* and *6* both still have a copy of the *Pfs1|1792* gene. However, in **Chapter 3** we found that *Pfs4* and *6* belong to a different genetic clade than reference race *Pfs1*. Perhaps a different effector capable of suppressing ETI could be responsible for the breaking of *RPF3* loci in these races.

We also looked at differential amino acid substitutions in effector genes of different *Pfs* races for an association with the breaking of R-loci. We chose to only screen for HNS variants since we did not measure the allele-specific expression of effector genes. As avirulence is mostly a dominant trait, heterozygosity would often not be sufficient to avoid recognition by the resistant host [52; 53].

Initially, we compared the difference between HNS variants between two closely related *Pfs* isolates of which one has broken an R-gene. Although we identified many effectors that contain HNS variants in the resistance-breaking race, only one effector contains HNS variants that are also present in other races that break the same R-gene. This effector gene, *Pfs1|9985*, contains HNS variants that can be directly associated with the breaking of *RPF2* and was found to be part of a larger diverging effector family that was identified in the Markov clustering analysis. When we compared HNS variants in all effector genes of different *Pfs* races we found three effectors with a strong association with overcoming *RPF*-mediated resistance in spinach; *Pfs* effector *Pfs1|3989* with *RPF5* and effectors *Pfs1|6786* and *Pfs1|9985* with *RPF2*. Functional assays are now being deployed to test the recognition of these effectors in resistant spinach lines to experimentally determine if they are truly responsible for pathogen recognition by the matching R-loci. For the resistance sources for which we did not find associated effector variants, it is likely that other mechanisms play a role in breaking resistance. Heterozygous non-synonymous polymorphisms could perhaps prevent the recognition by R-loci in spinach if only the allele that contains the variant is expressed. The allele-specific expression could be determined using RNA-seq data. The implementation of such results could reveal additional resistance breaking effector variants in *Pfs*.

Oomycete effectors are often located near repeat regions rich in transposable elements [54], which are notoriously hard to assemble [55]. This could mean that we are still missing, or incorrectly annotated putative effector genes in the *Pfs1* genome. Based on the relatively higher coverage of certain genes in the genome of *Pfs* we identified the presence of some paralogous sequences. These sequences are highly similar to each other in terms of nucleotide composition and are likely merged and not assembled correctly as separate genes in the genome of *Pfs1*. A complete and more contiguous *Pfs* genome assembly is expected to yield additional effector sequences that we are currently unaware of. In addition, genome assemblies of other *Pfs* races would also allow assessing potential effector gene loss in the *Pfs1* race.

Including future isolates (and races) would strengthen the study of resistance breaking *Pfs* effectors, as the continuous effector adaptation can be further monitored and studied. This would broaden our understanding of how the *Pfs* pathogen and which (emerging) effectors play a role in the evasion of the immune response in the spinach host.

Overall, this study provided insight into the evolution and selection pressures of effector- and other genes in the *Pfs* genome. Importantly, several *Pfs* effector candidates were found to be associated with the breaking of spinach resistance. The *Pfs1* variants of these candidates are now being tested experimentally to validate their recognition in spinach NILs to determine if their recognition is R-gene specific. *In planta* recognition assays with conserved effectors can be potentially used to identify novel R-genes in spinach to strengthen the effort to breed for durable resistance [56].

Acknowledgments

We wish to thank the breeding companies Enza Zaden, Pop Vriend Seeds, Rijk Zwaan Breeding, and Syngenta Seeds for sharing *Pfs isolate* material and data used in this project. We thank TKI T&U (Topsector Tuinbouw & Uitgangsmaterialen) along with the companies for financing this project.

References

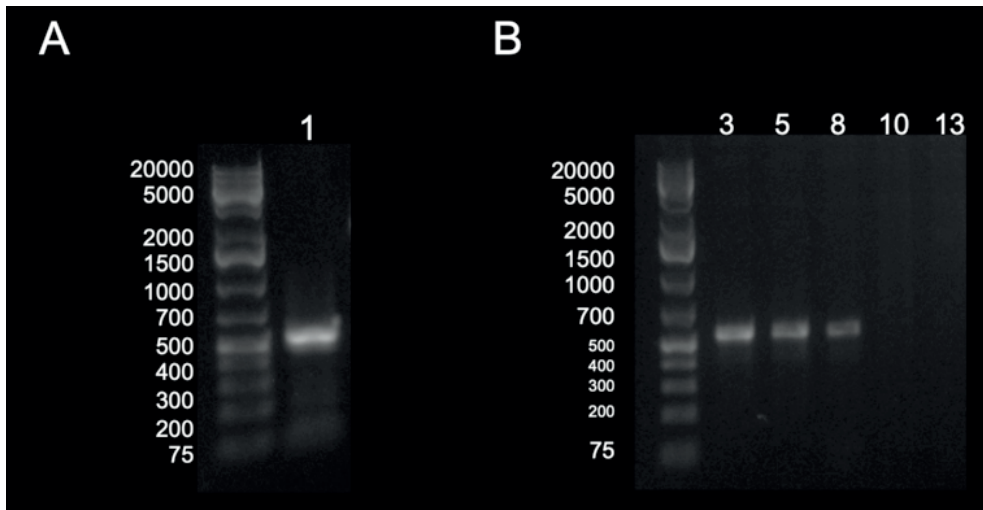
1. Choi YJ, Hong SB, Shin HD. 2007. Re-consideration of *Peronospora farinosa* infecting *Spinacia oleracea* as distinct species, *Peronospora effusa*. *Mycol. Res.* 111:381-91
2. McCarthy CGP, Fitzpatrick DA. 2017. Phylogenomic reconstruction of the oomycete phylogeny derived from 37 genomes. *mSphere* 2:e00095-17
3. Kamoun S, Furzer O, Jones JD, Judelson HS, Ali GS, *et al.* 2015. The Top 10 oomycete pathogens in molecular plant pathology. *Mol. Plant Pathol.* 16:413-34
4. Raaymakers TM, Van den Ackerveken G. 2016. Extracellular recognition of oomycetes during biotrophic infection of plants. *Front. Plant Sci.* 7:906 – 17
5. Schwessinger B, Zipfel C. 2008. News from the frontline: recent insights into PAMP-triggered immunity in plants. *Curr. Opin. Plant Biol.* 11:389-95
6. Asai S, Shirasu K. 2015. Plant cells under siege: plant immune system versus pathogen effectors. *Curr. Opin. Plant Biol.* 28:1-8
7. Stassen JH, Van den Ackerveken G. 2011. How do oomycete effectors interfere with plant life? *Curr. Opin. Plant Biol.* 14:407-14
8. Fabro G, Steinbrenner J, Coates M, Ishaque N, Baxter L, *et al.* 2011. Multiple candidate effectors from the oomycete pathogen *Hyaloperonospora arabidopsidis* suppress host plant immunity. *PLoS Pathog.* 7:e1002348
9. Cui H, Tsuda K, Parker JE. 2015. Effector-triggered immunity: from pathogen perception to robust defense. *Annu. Rev. Plant Biol.* 66:487-511
10. Cesari S. 2018. Multiple strategies for pathogen perception by plant immune receptors. *New Phytol.* 219:17-24
11. Jones JD, Dangl JL. 2006. The plant immune system. *Nature* 444:323-9
12. Coates ME, Beynon JL. 2010. *Hyaloperonospora arabidopsidis* as a Pathogen Model. *Annu. Rev. Phytopathol.* 48:329-45
13. Schlaich NL, Slusarenko A. 2009. *Downy mildew of arabidopsis caused by Hyaloperonospora arabidopsidis (Formerly Hyaloperonospora parasitica)*. pp 263-285. John Wiley & Sons. 574 pp.
14. Judelson HS. 2012. Dynamics and innovations within oomycete genomes: Insights into biology, pathology, and evolution. *Eukaryot. Cell* 11:1304-12
15. Haas BJ, Kamoun S, Zody MC, Jiang RH, Handsaker RE, *et al.* 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* 461:393-8
16. Wang Y, Tyler BM, Wang Y. 2019. Defense and counterdefense during plant-pathogenic oomycete infection. *Annu. Rev. Microbiol.* 73:667-96
17. Allen RL, Meitz JC, Baumber RE, Hall SA, Lee SC, *et al.* 2008. Natural variation reveals key amino acids in a downy mildew effector that alters recognition specificity by an Arabidopsis resistance gene. *Mol. Plant Pathol.* 9:511-23

18. Saunders DG, Breen S, Win J, Schornack S, Hein I, *et al.* 2012. Host protein BSL1 associates with *Phytophthora infestans* RXLR effector AVR2 and the *Solanum demissum* Immune receptor R2 to mediate disease resistance. *Plant Cell* 24:3420-34
19. Gilroy EM, Breen S, Whisson SC, Squires J, Hein I, *et al.* 2011. Presence/absence, differential expression and sequence polymorphisms between *PiAVR2* and *PiAVR2*-like in *Phytophthora infestans* determine virulence on R2 plants. *New Phytol.* 191:763-76
20. Raffaele S, Farrer RA, Cano LM, Studholme DJ, MacLean D, *et al.* 2010. Genome evolution following host jumps in the Irish potato famine pathogen lineage. *Science* 330:1540-3
21. Goss EM, Press CM, Grunwald NJ. 2013. Evolution of RXLR-class effectors in the oomycete plant pathogen *Phytophthora ramorum*. *PLoS ONE* 8:e79347
22. Qutob D, Tedman-Jones J, Dong S, Kuflu K, Pham H, *et al.* 2009. Copy number variation and transcriptional polymorphisms of *Phytophthora sojae* RXLR effector genes *Avr1a* and *Avr3a*. *PLoS ONE* 4:e5066
23. Win J, Morgan W, Bos J, Krasileva KV, Cano LM, *et al.* 2007. Adaptive evolution has targeted the C-terminal domain of the RXLR effectors of plant pathogenic oomycetes. *Plant Cell* 19:2349-69
24. Jiang RHY, Tripathy S, Govers F, Tyler BM. 2008. RXLR effector reservoir in two *Phytophthora* species is dominated by a single rapidly evolving superfamily with more than 700 members. *Proc. Natl. Acad. Sci. U. S. A.* 105:4874-9
25. Ellis JG, Rafiqi M, Gan P, Chakrabarti A, Dodds PN. 2009. Recent progress in discovery and functional analysis of effector proteins of fungal and oomycete plant pathogens. *Curr. Opin. Plant Biol.* 12:399-405
26. He J, Ye W, Choi DS, Wu B, Zhai Y, *et al.* 2019. Structural analysis of *Phytophthora* suppressor of RNA silencing 2 (PSR2) reveals a conserved modular fold contributing to virulence. *Proc. Natl. Acad. Sci. U. S. A.* 116:8054-9
27. Boutemy LS, King SRF, Win J, Hughes RK, Clarke TA, *et al.* 2011. Structures of *Phytophthora* RXLR effector proteins: A conserved but adaptable fold underpins functional diversity. *J. Biol. Chem.* 286:35834-42
28. Win J, Krasileva KV, Kamoun S, Shirasu K, Staskawicz BJ, Banfield MJ. 2012. Sequence divergent RXLR effectors share a structural fold conserved across plant pathogenic oomycete species. *PLoS Pathog.* 8:e1002400
29. Correll JC, Bluhm BH, Feng C, Lamour K, du Toit LJ, Koike ST. 2011. Spinach: better management of downy mildew and white rust through genomics. *Eur. J. Plant Pathol.* 129:193-205
30. Morelock T. 1999. Spinach: variety test and description. *Hortscience* 34:987-8
31. Brandenberger LP, Correll JC, Morelock TE. 1991. Identification of and cultivar reactions to a new race (Race-4) of *Peronospora farinosa* f. sp. *Spinaciae* on spinach in the United-States. *Plant Dis.* 75:630-4

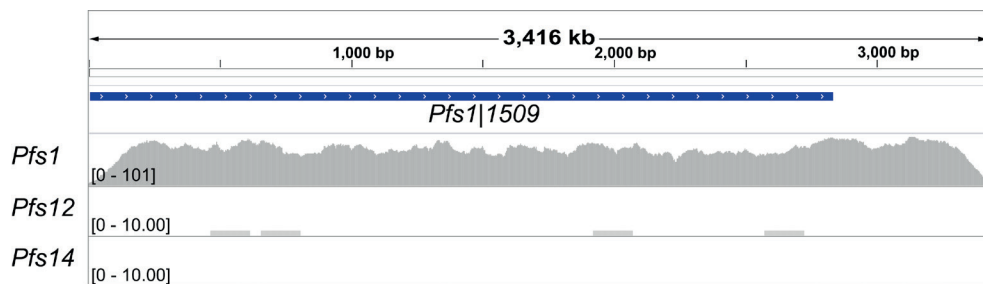
32. Brugmans BW, Meeuwsen J, Nooyen C. 2015. Compositions and methods for peronospora resistance in spinach. Seminis Vegetable Seeds Inc. *United States Patent No. US20150240256A1*
33. Correll J, du Toit L, Koike S, van Ettehoven K. 2010. *Guidelines for spinach downy mildew: Peronospora farinosa f. sp. spinaciae (Pfs)*. http://www.worldseed.org/cms/medias/file/Tradelssues/DiseasesResistance/StrainIdentification/Spinach_downy_mildew_Differentials_29102010.pdf
34. Irish BM, Correll JC, Feng C, Bentley T, de Los Reyes BG. 2008. Characterization of a resistance locus (Pfs-1) to the spinach downy mildew pathogen (*Peronospora farinosa f. sp. spinaciae*) and development of a molecular marker linked to Pfs-1. *Phytopathology* 98:894-900
35. Feng C, Saito K, Liu B, Manley A, Kammeijer K, et al. 2018. New races and novel strains of the spinach downy mildew pathogen *Peronospora effusa*. *Plant Dis.* 102:613-8
36. Jansen JPA. 2016. Hybrid spinach variety 51-518 RZ. Rijk Zwaan Zaadteelt en Zaadhandel BV *United States Patent No. US9295220B2*
37. Dijkstra J. 2013. Hybrid spinach variety Andromeda. Nunhems BV *United States Patent No. US8563807B2*
38. She H, Qian W, Zhang H, Liu Z, Wang X, et al. 2018. Fine mapping and candidate gene screening of the downy mildew resistance gene RPF1 in Spinach. *Theor. Appl. Genet.* 131:2529-41
39. Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30:1575-84
40. Robert X, Gouet P. 2014. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* 42:W320-4
41. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078-9
42. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, et al. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6:80-92
43. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, et al. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* 43:110 1-33
44. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792-7
45. Hickey G, Dehne F, Rau-Chaplin A, Blouin C. 2008. SPR distance computation for unrooted trees. *Evol. Bioinform. Online* 4:17-27
46. Robinson O, Dylus D, Dessimoz C. 2016. Phylo. io: Interactive viewing and comparison of large phylogenetic trees on the web. *Mol. Biol. Evol.* 33:2163-6
47. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841-2

48. Whidden C, Matsen FA. 2019. Calculating the unrooted subtree prune-and-regraft distance. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16:898-911
49. Armitage AD, Lysoe E, Nellist CF, Lewis LA, Cano LM, *et al.* 2018. Bioinformatic characterisation of the effector repertoire of the strawberry pathogen *Phytophthora cactorum*. *PLoS One* 13:e0202305
50. Yin JL, Gu B, Huang GY, Tian Y, Quan JL, *et al.* 2017. Conserved RXLR effector genes of *Phytophthora infestans* expressed at the early stage of potato infection are suppressive to host defense. *Front. Plant Sci.* 8:2155-66
51. Vleeshouwers VGAA, Oliver RP. 2014. Effectors as tools in disease resistance breeding against biotrophic, hemibiotrophic, and necrotrophic plant pathogens. *Mol. Plant-Microbe Interact.* 27:196-206
52. Woods-Tor A, Studholme DJ, Cevik V, Telli O, Holub EB, Tor M. 2018. A suppressor/avirulence gene combination in *Hyaloperonospora arabidopsidis* determines race specificity in *Arabidopsis thaliana*. *Front. Plant Sci.* 9:265
53. Huang J, Chen L, Lu X, Peng Q, Zhang Y, *et al.* 2019. Natural allelic variations provide insights into host adaptation of *Phytophthora* avirulence effector PsAvr3c. *New Phytol.* 221:1010-22
54. Dong S, Raffaele S, Kamoun S. 2015. The two-speed genomes of filamentous pathogens: waltz with plants. *Curr. Opin. Genet. Dev.* 35:57-65
55. Schmid M, Frei D, Patrignani A, Schlapbach R, Frey JE, *et al.* 2018. Pushing the limits of de novo genome assembly for complex prokaryotic genomes harboring very long, near identical repeats. *Nucleic Acids Res.* 46:8953-65
56. Lokossou AA, Park TH, van Arkel G, Arens M, Ruyter-Spira C, *et al.* 2009. Exploiting knowledge of *R/Avr* genes to rapidly clone a new LZ-NBS-LRR family of late blight resistance genes from potato linkage group IV. *Mol. Plant-Microbe Interact.* 22:630-41

Supplemental data

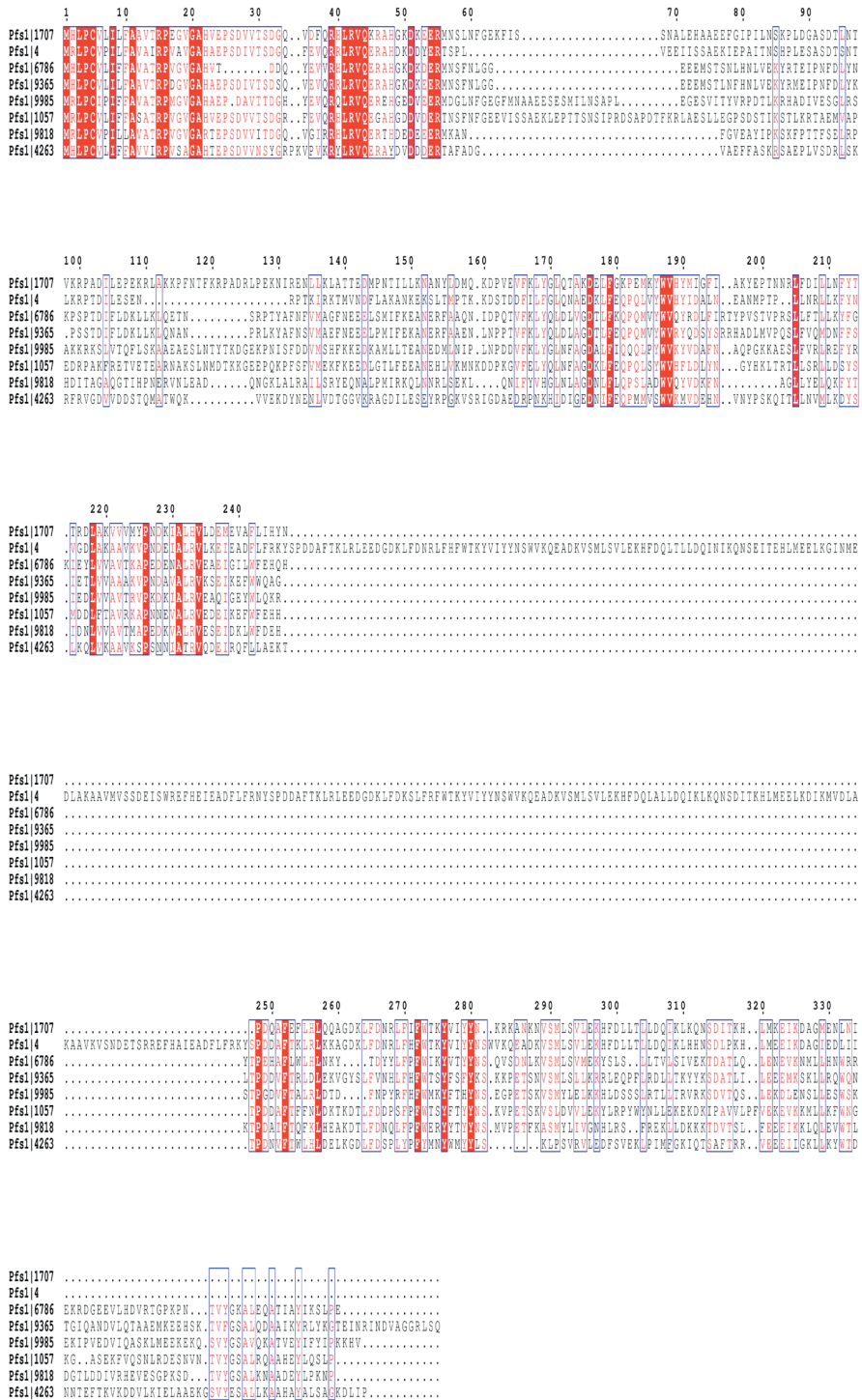


Supplemental Figure 1. Confirmation of the presence or absence of effector gene *Pfs1|1792*. Ethidium bromide stained agarose gel (1%) showing DNA fragments produced by PCR amplification of the *Pfs1|1792* gene. The *Pfs* race number where the DNA for PCR analysis was obtained from is marked at the top of each lane. The unmarked lane on the left of each gel contains a 1 kb plus DNA ladder (Invitrogen), fragment sizes are denoted on the left side of the marker lane. **A.** PCR results from *Pfs1*, a clear band can be seen between 500 and 700 bp which corresponds with the length of *Pfs1|1792* (541 bp). **B.** PCR results for *Pfs3*, 5, 8, 10, 13 with the band for the *Pfs1|1792* gene clearly missing for race 10 and 13.



Supplemental Figure 2. Read coverage over the whole region of contig Pb_1342 for different *Pfs* races that are missing the *Pfs1|1509* gene that is indicated by the blue bar. As a reference, the coverage for this contig in *Pfs1* is included. The coverage range (not to scale) is denoted on the left-hand side between brackets. In races where this gene is missing the entire contig is not present, In all other *Pfs* races, the coverage profile is similar to that observed for *Pfs1*.

Peronospora effusa effector polymorphisms associated with the breaking of resistance loci in spinach



Supplemental Figure 3. CLUSTALW alignment of eight paralogous *Pf*s RxLR effectors that belong to PGG1.

Supplemental Table 1. MCL clustering results of homologous *Pfs* effector sequences.

| PGG number | Effectors | | | | | | | | | | Effector type |
|--|--------------------|---------------------|--------------------|--------------------|--------------------|--------------------|--------------------|-------------------|--|--|---------------|
| PGG1 | <i>Pfs</i> 1 19985 | <i>Pfs</i> 1 9365 | <i>Pfs</i> 1 6786 | <i>Pfs</i> 1 9818 | <i>Pfs</i> 1 4 | <i>Pfs</i> 1 1707 | <i>Pfs</i> 1 4263 | | | | RxLR |
| PGG2 | <i>Pfs</i> 1 12056 | <i>Pfs</i> 1 1192 | <i>Pfs</i> 1 3286 | <i>Pfs</i> 1 12059 | <i>Pfs</i> 1 13089 | <i>Pfs</i> 1 11975 | <i>Pfs</i> 1 11878 | <i>Pfs</i> 1 3069 | | | RxLR |
| PGG3 | <i>Pfs</i> 1 8261 | <i>Pfs</i> 1 9071 | <i>Pfs</i> 1 7181 | <i>Pfs</i> 1 4560 | <i>Pfs</i> 1 8333 | <i>Pfs</i> 1 7184 | <i>Pfs</i> 1 4565 | | | | CRN |
| PGG4 | <i>Pfs</i> 1 8068 | <i>Pfs</i> 1 9526aa | <i>Pfs</i> 1 13024 | <i>Pfs</i> 1 12722 | <i>Pfs</i> 1 1950 | <i>Pfs</i> 1 5909 | | | | | RxLR |
| PGG5 | <i>Pfs</i> 1 10492 | <i>Pfs</i> 1 805 | <i>Pfs</i> 1 8028 | <i>Pfs</i> 1 10230 | <i>Pfs</i> 1 8634 | | | | | | RxLR |
| PGG6 | <i>Pfs</i> 1 7484 | <i>Pfs</i> 1 3487 | <i>Pfs</i> 1 1052 | | | | | | | | CRN |
| PGG7 | <i>Pfs</i> 1 7229 | <i>Pfs</i> 1 2732 | <i>Pfs</i> 1 3188 | | | | | | | | RxLR |
| PGG8 | <i>Pfs</i> 1 11709 | <i>Pfs</i> 1 842 | <i>Pfs</i> 1 5773 | | | | | | | | RxLR |
| PGG9 | <i>Pfs</i> 1 1509 | <i>Pfs</i> 1 524 | <i>Pfs</i> 1 10015 | | | | | | | | RxLR |
| PGGs containing less than three effectors sequences | | | | | | | | | | | |
| PGG10 | <i>Pfs</i> 1 10648 | <i>Pfs</i> 1 10282 | | | | | | | | | RxLR |
| PGG11 | <i>Pfs</i> 1 9976 | <i>Pfs</i> 1 7472 | | | | | | | | | CRN |
| PGG12 | <i>Pfs</i> 1 10924 | <i>Pfs</i> 1 9858 | | | | | | | | | CRN |
| PGG13 | <i>Pfs</i> 1 7655 | <i>Pfs</i> 1 7659 | | | | | | | | | RxLR |
| PGG14 | <i>Pfs</i> 1 2759 | <i>Pfs</i> 1 5186 | | | | | | | | | RxLR |
| PGG15 | <i>Pfs</i> 1 8288a | <i>Pfs</i> 1 11298 | | | | | | | | | RxLR |
| PGG16 | <i>Pfs</i> 1 3172 | <i>Pfs</i> 1 4214 | | | | | | | | | RxLR |
| PGG17 | <i>Pfs</i> 1 1028 | <i>Pfs</i> 1 1034 | | | | | | | | | RxLR |
| PGG18 | <i>Pfs</i> 1 6054 | <i>Pfs</i> 1 10189 | | | | | | | | | RxLR |
| PGG19 | <i>Pfs</i> 1 3683 | <i>Pfs</i> 1 3869 | | | | | | | | | RxLR |
| PGG20 | <i>Pfs</i> 1 848 | <i>Pfs</i> 1 11520 | | | | | | | | | RxLR |

Contains an ortholog in another oomycete species

Supplemental Table 2. Conserved Pfs effectors, of which an ortholog was found in another species based on the BBH approach.

| Gene name | Species | Hit gene name | WY-motif | # WY-motifs | Transmembrane | # helix | RxLR motif | EER motif | CRN-motif |
|------------------|---|---------------|----------|-------------|---------------|---------|------------|-----------|-----------|
| Pfs11739-00001 | <i>Phytophthora sojae</i> | gnt1_3 | no WY | 0 | TMHMM | 1 | AHLR | DER | |
| Pfs113210-00001 | <i>Phytophthora infestans</i> T30-4 | NA | no WY | 0 | no TMHMM | | ANLR | EEL | |
| Pfs116706-00001 | <i>Phytophthora parasitica</i> INRA-310 | NA | no WY | 0 | no TMHMM | | GPLR | EDR | |
| Pfs113188-00001 | <i>Ichthyophthirius multifiliis</i> | Spr11a | WYfound | 2 | no TMHMM | | KNLR | DDK | |
| Pfs110418-00001 | <i>Phytophthora parasitica</i> INRA-310 | NA | no WY | 0 | no TMHMM | | KTLLR | DEK | |
| Pfs119196-00001 | <i>Phytophthora parasitica</i> INRA-310 | At1g74750 | no WY | 0 | no TMHMM | | NGLR | DER | |
| Pfs118598-00001 | <i>Phytophthora sojae</i> | smurf2_1 | no WY | 0 | no TMHMM | | NQLR | EER | |
| Pfs112360-00001 | <i>Phytophthora parasitica</i> INRA-310 | NA | no WY | 0 | no TMHMM | | QMLR | EEL | GFSAK |
| Pfs117306-00001 | <i>Phytophthora sojae</i> | NA | no WY | 0 | no TMHMM | | RQLA | EDR | |
| Pfs111601-00001 | <i>Phytophthora parasitica</i> INRA-310 | SELENOF | no WY | 0 | no TMHMM | | RALE | EDR | |
| Pfs110453-00001 | <i>Phytophthora parasitica</i> INRA-310 | NA | no WY | 0 | no TMHMM | | RGLK | EEL | |
| Pfs112474-00001 | <i>Phytophthora infestans</i> T30-4 | NA | no WY | 0 | TMHMM | 1 | RYLK | EER | |
| Pfs110123-00001 | <i>Phytophthora parasitica</i> INRA-310 | NA | no WY | 0 | no TMHMM | | RDLL | EEL | HVL |
| Pfs113214-00001 | <i>Phytophthora parasitica</i> INRA-310 | NA | no WY | 0 | no TMHMM | | RALL | EER | |
| Pfs118288a-00001 | <i>Phytophthora parasitica</i> INRA-310 | NA | WYfound | 4 | no TMHMM | | RFLR | DDK | |
| Pfs118208a-00001 | <i>Phytophthora parasitica</i> INRA-310 | NA | WYfound | 1 | no TMHMM | | RFLR | EER | |
| Pfs114249-00001 | <i>Winogradskya epiphytica</i> | NA | no WY | 0 | no TMHMM | | RHLR | EER | |
| Pfs117215-00001 | <i>Plasmopara halstedii</i> | ykwD_2 | no WY | 0 | no TMHMM | | RHLR | | |
| Pfs112056-00001 | <i>Phytophthora parasitica</i> INRA-310 | NA | WYfound | 3 | no TMHMM | | RHLR | | |
| Pfs112059-00001 | <i>Phytophthora sojae</i> | NA | WYfound | 3 | no TMHMM | | RHLR | | |
| Pfs118804-00001 | <i>Phytophthora parasitica</i> INRA-310 | infB_1 | no WY | 0 | no TMHMM | | RLLR | EOR | |
| Pfs113989-00001 | <i>Phytophthora infestans</i> T30-4 | BROM1 | no WY | 0 | TMHMM | 1 | RMLR | | |
| Pfs114650-00001 | <i>Plasmopara halstedii</i> | NA | no WY | 0 | TMHMM | 1 | RMLR | | |
| Pfs116386-00001 | <i>Phytophthora infestans</i> T30-4 | LSM10 | no WY | 0 | no TMHMM | | RMLR | | |
| Pfs117438-00001 | <i>Phytophthora sojae</i> | rply | no WY | 0 | no TMHMM | | RRLR | DER | |
| Pfs11192-00001 | <i>Phytophthora sojae</i> | NA | WYfound | 6 | no TMHMM | | RRLR | EER | |
| Pfs113378-00001 | <i>Phytophthora parasitica</i> INRA-310 | Adck1 | no WY | 0 | no TMHMM | | RSLR | | |
| Pfs119324-00001 | <i>Phytophthora parasitica</i> INRA-310 | NA | no WY | 0 | no TMHMM | | RSLR | | |
| Pfs111135-00001 | <i>Phytophthora parasitica</i> INRA-310 | NA | no WY | 0 | no TMHMM | | RTL | | |
| Pfs11257-00001 | <i>Phytophthora sojae</i> | pep1 | no WY | 0 | no TMHMM | | RTL | | |
| Pfs11439-00001 | <i>Phytophthora sojae</i> | NA | no WY | 0 | no TMHMM | | RTL | | |
| Pfs112315-00001 | <i>Phytophthora parasitica</i> INRA-310 | TUBGCP6_0 | no WY | 0 | no TMHMM | | RYLR | DDK | |
| Pfs11805-00001 | <i>Phytophthora infestans</i> T30-4 | NA | WYfound | 5 | no TMHMM | | RYLR | | |
| Pfs11668-00001 | <i>Phytophthora sojae</i> | DNAJA1_1 | no WY | 0 | no TMHMM | | RKLS | DER | |
| Pfs1110382-00001 | <i>Phytophthora infestans</i> T30-4 | NA | no WY | 0 | no TMHMM | | SALR | EER | |
| Pfs119033-00001 | <i>Phytophthora sojae</i> | kif11 | no WY | 0 | no TMHMM | | VRLR | EEL | |

Supplemental Table 3. Overview of LWY containing (effector) proteins in the *Pfs1* genome.

This supplemental table can be found in the online repository located at:
<https://doi.org/10.17026/dans-zzg-mjgb>

Supplemental Table 4. Mean per base coverage per gene for each *Pfs* isolate sequenced.

This supplemental table can be found in the online repository located at:
<https://doi.org/10.17026/dans-zzg-mjgb>

Supplemental Table 5. Possible paralogous genes identified based on the read alignment to coding sequences in the different isolates. A gene is considered to be paralogous when the coverage score is higher than two standard deviations above the mean average coverage score of an isolate. Paralogs are denoted as “2” non paralogous genes as “0”.

This supplemental table can be found in the online repository located at:
<https://doi.org/10.17026/dans-zzg-mjgb>

Supplemental Table 6. Putative annotation of effector genes, dN/dS values of genes in different isolates, mean dN/dS values per gene over the isolates, and identified orthologs in other oomycetes species.

This supplemental table can be found in the online repository located at:
<https://doi.org/10.17026/dans-zzg-mjgb>

Supplemental Table 7. List of non-polymorphic *Pfs* effectors

| Gene name | Under 1 kbp | In repeat | # WY motifs | TMHMM | #helix | RxLR motif | EER motif | CRN motif | HVL motif |
|--------------------------|-------------|-----------|-------------|-------------|--------|------------|-----------|-----------|-----------|
| <i>Pfs1</i> 12722-00003 | NA | NA | 3 | no TMHMM | | DALR | | | |
| <i>Pfs1</i> 5871-00001 | NA | NA | 1 | TMHMM found | 1 | HHLR | EER | | |
| <i>Pfs1</i> 10418-00001 | NA | NA | 2 | no TMHMM | | KTLR | DEK | | |
| <i>Pfs1</i> 1028-00001 | NA | NA | 2 | TMHMM found | 1 | PRLR | EER | | |
| <i>Pfs1</i> 7659-00001 | NA | NA | 3 | no TMHMM | | QALR | EER | | HVL |
| <i>Pfs1</i> 12474-00001 | NA | NA | 1 | TMHMM found | 1 | RYLK | EER | | |
| <i>Pfs1</i> 10123-00001 | NA | NA | 3 | no TMHMM | | RDLL | EEK | | HVL |
| <i>Pfs1</i> 10014-00001 | < 1Kbp | NA | 1 | no TMHMM | | RALR | DQK | | |
| <i>Pfs1</i> 8000-00001 | NA | NA | 2 | no TMHMM | | RALR | EER | | |
| <i>Pfs1</i> 4980-00001 | NA | NA | 1 | no TMHMM | | RFLR | EER | | |
| <i>Pfs1</i> 8208a-00001 | NA | NA | 3 | no TMHMM | | RFLR | EER | | |
| <i>Pfs1</i> 3927-00001 | NA | NA | 3 | TMHMM found | 1 | RFLR | | | |
| <i>Pfs1</i> 10814-00001 | NA | in repeat | 1 | no TMHMM | | RGLR | | | |
| <i>Pfs1</i> 4249-00001 | NA | NA | 4 | no TMHMM | | RHLR | EER | | |
| <i>Pfs1</i> 1707-00001 | NA | NA | 4 | no TMHMM | | RHLR | EER | | |
| <i>Pfs1</i> 2732-00001 | NA | NA | 1 | no TMHMM | | RHLR | EER | | |
| <i>Pfs1</i> 9414-00002 | NA | NA | 2 | no TMHMM | | RMLR | EER | | |
| <i>Pfs1</i> 10282-00001 | < 1Kbp | in repeat | 1 | no TMHMM | | RMLR | | | |
| <i>Pfs1</i> 10648-00001 | < 1Kbp | NA | 1 | no TMHMM | | RMLR | | | |

Supplemental Table 7. continued.

| Gene name | Under 1kbp | In repeat | # WY motifs | TMHMM | #helix | RxLR motif | EER motif | CRN motif | HVL motif |
|-------------------------|------------|-----------|-------------|-------------|--------|------------|-----------|-----------|-----------|
| <i>Pfs1 4650-00001</i> | NA | NA | 1 | TMHMM found | 1 | RMLR | | | |
| <i>Pfs1 3869-00001</i> | NA | NA | 3 | no TMHMM | | RNLR | | | |
| <i>Pfs1 10189-00001</i> | NA | NA | 1 | no TMHMM | | RRLR | EER | | |
| <i>Pfs1 3172-00001</i> | NA | NA | 1 | TMHMM found | 1 | RRLR | EER | AFTAK | |
| <i>Pfs1 4214-00001</i> | NA | NA | 1 | TMHMM found | 1 | RRLR | EER | | |
| <i>Pfs1 6054-00001</i> | NA | NA | 1 | no TMHMM | | RRLR | EER | | |
| <i>Pfs1 9365-00001</i> | NA | NA | 5 | no TMHMM | | RRLR | EER | | |
| <i>Pfs1 556-00001</i> | NA | NA | 2 | no TMHMM | | RRLR | | | |
| <i>Pfs1 13378-00001</i> | NA | NA | 3 | no TMHMM | | RSLR | | | |
| <i>Pfs1 439-00001</i> | NA | NA | 1 | no TMHMM | | RTLr | | | |
| <i>Pfs1 1950-00001</i> | NA | NA | 4 | no TMHMM | | RYLR | | | |
| <i>Pfs1 10382-00001</i> | NA | NA | 3 | no TMHMM | | SALR | EER | | |
| <i>Pfs1 9033-00001</i> | NA | NA | 3 | no TMHMM | | VRLR | EEK | | |
| <i>Pfs1 3487-00001</i> | NA | NA | 4 | no TMHMM | | RKLR | DNK | LFLAK | HVL |
| <i>Pfs1 8261-00001</i> | < 1Kbp | in repeat | 5 | no TMHMM | | RNLQ | | LFLAK | |
| <i>Pfs1 9071-00001</i> | < 1Kbp | in repeat | 5 | no TMHMM | | RNLQ | | LFLAK | |
| <i>Pfs1 7181-00001</i> | NA | in repeat | 5 | no TMHMM | | | | LFLAK | |
| <i>Pfs1 9976-00001</i> | NA | NA | 5 | no TMHMM | | RDLQ | | LFLAK | HVL |
| <i>Pfs1 7184-00001</i> | NA | in repeat | 4 | no TMHMM | | | | TFLAK | |
| <i>Pfs1 4565-00001</i> | NA | NA | 5 | no TMHMM | | LSLR | | TFLAK | |
| <i>Pfs1 4560-00001</i> | NA | in repeat | 5 | no TMHMM | | RSLQ | | | |
| <i>Pfs1 8333-00001</i> | < 1Kbp | in repeat | 5 | no TMHMM | | RSLQ | | | |

Supplemental Table 8. Homozygous and heterozygous SNP variants with impact annotation in the protein-coding sequence of *Pfs* effectors.

This supplemental table can be found in the online repository located at:
<https://doi.org/10.17026/dans-zzg-mjgb>

Supplemental Table 9. Summary of SNP variants and their impact on the coding sequence of *Pfs* effectors.

This supplemental table can be found in the online repository located at:
<https://doi.org/10.17026/dans-zzg-mjgb>





Chapter 5

Summarizing discussion

Summarizing discussion

The characterization of genome sequences of isolates of the spinach downy mildew *Peronospora effusa* (*Pfs*), described in this thesis, has been facilitated by using various bioinformatics methods, e.g. taxonomic filtering of contigs and comparative genome analysis. The findings have provided insights into possible *Pfs* infection strategies, and importantly, adaptations in effector sequences to overcome spinach resistance. Bioinformatics is a fast-evolving specialization that can be used to study the large amounts of experimental data obtained from organisms and other biological samples. In particular in the field of genomics has bioinformatics been a powerful enabling technology to complement experimental research and thereby accelerated studies on many species significantly.

In this thesis, I describe the sequencing and assembly of the genome of *Pfs* race 1 (*Pfs1*), the prediction of gene and protein models, the selection of the secretome, and the subsequent identification of candidate effectors. Proteins with a predicted function in pathogenicity were compared to those of other oomycete species to provide insights into potential infection strategies employed by *Pfs* and other downy mildew pathogens (**Chapter 2**).

Besides *Pfs1*, 15 other races (2-16), and 8 field isolates were sequenced. Comparison of Illumina sequencing reads with the *Pfs1* reference genome revealed possible phylogenetic and evolutionary relationships between denominated *Pfs* races and the sequenced field isolates (**Chapter 3**). This comparison was also used to assess the selection pressures on effector genes compared to other genes in the *Pfs* genome. This revealed that two effectors were absent in several isolates. The loss of one effector gene could be associated with the breaking of *RPF3* resistance. Other associations to the breaking of resistance loci were found for amino acid substitutions in several effector genes in a selection of *Pfs* isolates (**Chapter 4**). Here I reflect on these bioinformatic analyses, place the main results in a broader perspective and discuss the strategies to further increase the understanding of the *Pfs*-spinach interaction for the benefit of resistance breeding.

Challenges in genome assembly of obligate biotrophic plant pathogens

Over the last two decades, we have witnessed the publication of genome analyses of a broad range of oomycetes species [1]. This started with the sequencing and assembly of the oomycete genomes of *Phytophthora sojae* and *ramorum* [2] in 2006 and was followed by that of *Phytophthora infestans* [3] and *Hyaloperonospora arabidopsidis* [4] a few years later. In

the following years, the genomes of several *Pythium* (e.g. *Pythium ultimum* [5]) and *Albugo* (e.g. *Albugo laibachii* [6]) species were also published. In recent years, several downy mildew genomes have been sequenced as well (**Table 1**). These genome assemblies have provided an interesting insight into the mechanism of the downy mildew infection.

Table 1. Published downy mildew genomes.

| Downy mildew species | Year published | Host | Reference |
|---------------------------------------|----------------|--------------|-----------|
| <i>Hyaloperonospora arabidopsidis</i> | 2010 | Arabidopsis | [4] |
| <i>Pseudoperonospora cubensis</i> | 2011 | Cucumber | [7] |
| <i>Plasmopara halstedii</i> | 2015 | Sunflower | [8] |
| <i>Peronospora tabacina</i> | 2015 | Tobacco | [9] |
| <i>Sclerospora graminicola</i> | 2017 | Pearl millet | [10] |
| <i>Peronospora belbahrii</i> | 2019 | Basil | [11] |
| <i>Pseudoperonospora humuli</i> | 2019 | Hop | [12] |
| <i>Plasmopara viticola</i> | 2019 | Grapevine | [13] |
| <i>Bremia lactucae</i> | 2019 | Lettuce | [14] |

Downy mildews are by definition obligate biotrophic organisms, which entails a different challenge for the sequencing and assembly of the genomes of these pathogens. These microbial pathogens cannot be cultivated on media and only grow on living plants. Consequentially, when DNA is isolated from sporangia harvested from living plants, the DNA of other microbes present on these plants is also collected. Thus, the resulting assembly of the unfiltered reads results in a meta-genome rather than a single pathogen genome. From a meta-genome assembly, it is difficult to distinguish which sequences are derived from the pathogen of interest and which originated from other microbes. This problem has not been addressed adequately in many previous downy mildew assemblies.

We encountered this problem when we sequenced and assembled a draft genome based on sequencing genomic DNA derived from isolated *Pfs* sporangia. In the initial draft assembly, we found that more than half of the assembled contigs belonged to bacteria and other microbial organisms. This emphasizes the need for an adequate filtering strategy to avoid the unintended assembly of sequences originating from bacteria, fungi, and other microbes into the *Pfs* genome. Several methods were considered to determine which reads or sequences are not part of *Pfs* and should be excluded from the assembly. Initially, we investigated the GC content of each contig, since this is quite conserved within oomycete genomes and different from that of many other microbes. However, this is not the case for all bacteria, for example, *E. coli* has a GC content of 51.7% [15], which is close to that of most oomycetes. Also, the GC bases are not evenly distributed over the genome, for example in regions with many AT or GC-rich repeats. Thus, filtering based on this criterium will likely result in the elimination of parts of

the oomycete genome that deviate in GC-content while still including sequences from other microbes.

Another method considered was to align the sequencing reads to a database containing previously assembled oomycete genomes. However, genome databases are not complete, and each oomycete genome contains unique parts that would be lost with this approach. Besides, these databases contain assembled downy mildew genomes that are not filtered and could, therefore, contain bacterial and fungal sequences that are annotated as being of downy mildew origin. In a different approach, reads that align with sequences in microbial databases could be discarded prior to assembly [16]. However, also bacterial databases are not complete, and likely contain taxonomic annotation errors [17].

For these reasons, we chose to determine the taxonomic origin of the assembled contigs and long reads (PacBio) using the Contig Annotation Tool (CAT) pipeline [18]. This pipeline identifies putative open reading frames (ORFs) on sequences and taxonomically classifies them. For each ORF present on an (assembled) sequence, a likely taxonomic origin is determined. CAT utilizes the combined taxonomic annotations of multiple individual ORFs to provide a taxonomic annotation for the entire sequence. This allows for a more robust taxon classification that is based on multiple hits, rather than a single best hit. Based on the CAT-method we taxonomically annotated the error-corrected PacBio reads and assembled contigs.

In total, we filtered out 50% of the error-corrected PacBio reads, and 44% of the Illumina reads did not align with the filtered contigs. This indicates that roughly half of the sequencing reads were derived from other species. Assembly with the filtered data resulted in a *Pfs* genome sequence consisting of 32.4 Mb, which is close to the *k*-mer based estimated genome size of 36.2 Mb. Based on RNA-seq guided gene model prediction we identified in total 13,227 gene models present in this assembly. Of these gene models, 9007 were deemed high confidence as these are present on contigs larger than 1 kb and do not have significant overlap (> 20%) with a repeat sequence.

According to the BUSCO analysis, the gene space in our assembly is 88.9% complete with only 0.5% fragmented genes. The gene space completeness score is similar to that of other downy mildew genomes but lower than that of the assembled genomes of different *Phytophthora* species. Furthermore, the assembled genomes of *Pfs* races R13 and R14 (32.2 and 30.8 Mb, respectively) presented by Fletcher *et al.* showed similar genome sizes compared to our *Pfs1* genome assembly [19]. In total, we were able to assemble 6.93 Mb of repeats for the *Pfs1* genome, which is slightly less than the assembled repeat sizes for *Pfs* races 13 and 14 (10.6 Mb and 9.14 Mb respectively) [19]. Compared to the assemblies of R13 and R14, the *Pfs1* genome assembly contains more gene coding (non-repeat) parts. This

also explains why we identified more high-confidence gene models in the genome assembly of *Pfs1* (9007) compared to R13 and R14 (~8000 genes each) [19].

Although the gene-space confined in our *Pfs* assembly is ~90% complete with only 0.5% duplicates, the *Pfs1* genome assembly is fragmented, consisting of 8,635 contigs, possibly due to large repeats residing in the *Pfs* genome. The repeat content assembled for *Pfs* (21%) is much lower than that of *H. arabidopsidis* (43%) and *P. infestans* (74%), which explains in part its relatively small genome size [6]. However, despite the use of longer PacBio sequences, we were unable to span many of the repeat regions in the *Pfs1* genome. The use of whole-genome amplification before PacBio sequencing could have resulted in a bias for certain genomic regions. Moreover, DNA obtained from the isolated spores contained a large amount of contaminating microbial sequences making that only half of the PacBio sequences could be used for genome assembly.

Effector genes in the *Pfs* genome

Effectors play a major role in immune suppression and other virulence activities in the host, and thereby enable the successful infection by pathogens [20; 21]. We can distinguish apoplastic [22] and host-translocated effectors based on the location of their activity *in planta*. Host-translocated effectors, also known as intracellular or cytoplasmic effectors, have an N-terminal signal peptide but are translocated into host cells. In my thesis, I primarily focus on host-translocated effectors of *Pfs1*. Host-translocated effectors of oomycete pathogens belonging to the Peronosporales (*Phytophthora* species and downy mildews) can be further divided into two major classes, namely the RxLR and Crinkler (CRN) effectors. Especially RxLR effectors have been studied extensively over the last 10-15 years [23]. This effector type carries an N-terminal SP followed by a conserved RxLR motif which is often accompanied by a conserved EER motif. The RxLR motif was found to be necessary for the translocation of effectors into plant cells [24; 25].

CRN effectors were also found to be host-translocated and can be identified by a conserved LFLAK motif, which is, similar to the RxLR motif, responsible for translocation [26]. Consequentially, the presence of these conserved N-terminal motifs makes oomycete host-translocated effectors modular proteins [21]. This also makes their bioinformatic identification possible, enabling the cataloging of RxLR and CRN in *Pfs* [2; 27].

In *Pfs*, we identified 99 RxLR(-like) effectors based on the presence RxLR, or a degenerative RxLR (xxLR or RxLx) combined with either an EER-like or a WY domain. Also, 14 CRN effectors were found in *Pfs* based on the presence of a canonical LFLAK motif or a degenerative LFLAK and HVL motif in the first 100 amino acids of the protein sequence. The num-

ber of RxLR and CRN effector sequences identified in the genome of *Pfs* is significantly smaller than that in the genomes of *Phytophthora* species that each encode for hundreds of potential effector proteins [2; 24; 28]. The number of *Pfs* effectors is comparable to other downy mildews such as *H. arabidopsidis* (134 RxLR, 20 CRN genes) [4] and *P. tabacina* (120 RxLR, 6 CRN genes) [9]. An exception is the genome of sunflower downy mildew, *Pl. halstedii*, that encodes more effectors (274 RxLR, 77 CRN genes) [29]. However, assembling downy mildew genomes without filtering method could lead to the unintended incorporation of sequences derived from other microbes, some of which are known to possess effector sequences that show sequence similarity with oomycete effectors [30]. This could greatly hinder the correct identification of oomycete effector sequences.

Oomycete genome size and adaptability

The relatively small genome size and effector repertoire of *Pfs* is in contrast with the hypothesis that a larger oomycete genome can adapt faster to host immunity [31; 32]. But it fits well with the hypothesis of genome reduction, which states that gene loss can provide a selective advantage by conserving an organism limiting resources [33]. An organism that is depended on another organism for its survival can obtain nutrients from its host or the surrounding community, and thus no longer needs complex metabolic pathways to produce these compounds. As a result, the evolution of symbiosis and parasitism is typically associated with genome reduction and has been observed in many microbial communities and parasites [34-36].

In some oomycete genomes e.g. *P. infestans* and to some extent *H. arabidopsidis*, the relatively larger genome size is linked to the high content of repeats. Besides, it was found that oomycete effectors often reside in gene-sparse repeat-rich regions [31]. These gene-sparse areas in the genome are associated with an accelerated rate of evolution through repeat-driven expansions, deletions and gene fusions. Analysis of the genomic context of putative effector genes compared to non-effector genes in the *Pfs1* genome revealed that effector genes are often located in less gene dense regions. In *Pfs* this might contribute to the accelerated adaptation of effectors in a similar way as occurs in *Phytophthora* species, e.g. in *P. infestans* [31].

Phylogenetic relationship between *Pfs* races and isolates

Many resistance-breaking races and isolates of *Pfs* have emerged in recent decades. However, it was not known how these isolates are related, how they have emerged, and what the driving force is in their adaptation to overcome host resistance. Therefore, we sequenced genomic DNA and aligned reads of 15 additional races (*Pfs2-16*) and 8 *Pfs* field isolates to

the assembled *Pfs1* mitochondrial (mt) genome. Contrary to a comparison between other oomycete lineages [37], the number of variants in the mt-genomes of different *Pfs* isolates alone is not enough to calculate a reliable distance for each individual isolate.

However, we did find that there are two distinct haplogroups present. These two groups have been previously proposed by Choi *et al.* based on the analysis of the mitochondrial *cox2* and *nad1* genes [38] and were associated with the geographical origin of the isolates: an Asian/Oceanian clade and a clade that includes American/European *Pfs* isolates. A similar observation was done in the late blight pathogen *P. infestans*. In a study on this oomycete, 37 different mt-haplogroups were found that could be divided over 2 main lineages. mt-haplotypes from US and Mexican isolates form a loose group but are distinct from isolates found in Europe, South America and other regions [39]. Contrary to these studies a clear association between the geographical locations and the mt-haplogroup could not be established for our isolates. The data that we have for the location of the *Pfs* races is based on the location they were first discovered and could be different from the location they originated.

In order to resolve the nuclear genome phylogeny of the *Pfs* races and isolates used in this study, we aligned the sequencing reads obtained from genomic DNA to the assembled *Pfs1* reference genome. A principal component analysis (PCA) based on this SNP data revealed four groups present in our collection of isolates. This observation was confirmed by population structure analyses that revealed three distinct groups (A, B and C), and a fourth admixed group. The presence of an admixed group suggests that sexual recombination contributed to the evolution and emergence of the different *Pfs* isolates [40].

The Neighbor-Net analysis of the *Pfs* isolates [41] can accommodate evolutionary processes such as recombination and hybridization. These processes result in conflicting phylogenetic signals in the distance matrix (SNP matrix) which are visualized in a split network as reticulate patterns. Isolates belonging to the admixed group are centered in the middle of the phylogenetic split network and are connected with rectangular patterns, that suggest the involvement of sexual recombination. Isolates present in the three distinct groups are connected to the main 'net' with relatively long branches. The distance between these groups and the main 'net' suggests that these groups are distinct clonal lineages and have accumulated variants in the absence of sexual reproduction. Also, the phylogenetic tree constructed with RaXML shows the three distinct groups, observed in the structure analysis, as monophyletic groups in each of the three main branches of the tree.

From these analyses, we can conclude that *Pfs1* till 3 (discovered before 1990) belong to the same monophyletic group and therefore share

a common ancestor. However, *Pfs4* (discovered in 1990) and *Pfs6* belong to a different monophyletic group and possess a distinctly different genetic background than *Pfs1*, 2 and 3. *Pfs12* and 14 are present in a third monophyletic clade in the phylogram together with *Pfs* isolates *US-11* and *US-13b*. Notably, each distinct genetic group in our data contained only one mt-haplogroup, whereas the admixed group contained mixed haplogroups. The presence of different mt-haplogroups in the closely-related isolates in the admixed group strongly suggests that sexual recombination has played a role in the evolution of recent *Pfs* isolates.

A study using SNP markers on the genetic relationship between several *Pfs* isolates and races has also been reported by Lyon *et al.* [40]. Their data analysis showed that historical *Pfs* isolates and races showed greater genetic diversity, which indicated that sexual recombination has likely occurred. However, recent isolates (2014–2019) were genetically closely related suggesting that asexual development was responsible for the genotypic variation between these isolates [40]. We have observed similar results in our study, where many *Pfs* isolates showed signs of sexual reproduction whereas other isolates are genetically more similar to each other.

Effector adaption to introduced resistance in the host

Resistance breeding against *Pfs* in spinach has resulted in a strong selection pressure on the pathogen to adapt to the introduced R-loci. A major challenge until now has been to determine how and which effector has adapted to a specific R-locus and how one can leverage this information.

Since effectors play a significant role in the infection of the host, we initially looked at the dN/dS ratio of effectors over the isolates as a measure for evolutionary selection/adaptation. We observed that, on average, effector genes acquire more polymorphisms over time compared to non-effector genes of *Pfs*. This has previously been reported for other oomycetes, where the authors found that effectors accumulate polymorphisms at an accelerated rate compared to the rest of the genome [42; 43].

In contrast, non-effector genes in the oomycete genome are under purifying or negative selection to conserve amino acid sequences, and thereby protein function [42; 44]. Consequentially, the quantification of positive selection pressure on effectors can be used as a criterion for the identification of key effectors that are candidates for breaking resistance or for improved virulence function, e.g. by enhanced interaction with an effector target [45-47].

However, the evolutionary selection analysis alone does not answer the question of how specific effectors have adapted to the introduced spinach resistance loci. *Pfs* races and isolates are capable of breaking different spinach R-loci (*RPF* for Resistance against *Peronospora farinosa*) [48]. Based on this premise, a differential set of spinach lines was used to

assign race designations to *Pfs* isolates based on their virulence spectrum. A recent version of the differential set includes six Near Isogenic Lines (NIL) containing a single R-locus or cultivars that contain a combination of two R-loci for the accurate determination of recent races [49; 50]. With this publicly available differential set we were able to determine which changes in effector proteins in the different races are potentially associated with the breakage of R-loci in spinach.

Effectors in the *Pfs* genome can adapt in various ways to these R-loci, which include but are not limited to the substitution of specific amino acid to avoid recognition, to discard or silence the offending effector all together. Also an the introduction of premature stop codon in the effector sequence would yield a truncated effector that is likely no longer recognized [51]. New effectors could also emerge through duplication and diversification to actively suppress effector-triggered immune responses mediated by host R-loci [52]. However, gene diversification resulting in novel protein functions occurs over longer time periods. Considering the relatively short time span in which most *Pfs* races have appeared, between 1 and 2 years on average, this would not be very likely.

Co-evolutionary patterns in relation to effectors have been described for bacterial [53], fungal [54] and several oomycete pathogens [45; 55]. The evolution between the pathogen and host is driven by the fixation of allelic variants that either prevent the infection of the pathogen or overcome the host resistance [51]. There are many examples of effectors in other oomycetes that have adapted to host resistance genes to avoid recognition [56]. For example, in *H. arabidopsidis* amino acid substitutions in the effector domain of RxLR effector ATR13 were sufficient to impair the recognition of the effector by Arabidopsis RPP13 [57]. Moreover, in *P. infestans* presence/absence, amino acid polymorphism and differential transcription determine whether avirulence gene *PiAVR2* is recognized by the *R2* resistance gene in potato [58].

Absence of effector gene *Pfs1|1792*, based on sequencing depth, was observed in races: *Pfs2*, 7, 10, 13, 15. Effector gene *Pfs1|1509* was missing in *Pfs12* and 14. The presence/absence profile of effector *Pfs1|1792* associates with the breaking of R-locus *RPF3* by 5 races, with the exception of *Pfs4* and 6. *Pfs4* and *Pfs6*, however, belong to a different clade based on the phylogenetic and structure analysis. These two races might have evolved another mechanism to break the resistance mediate by *RPF3* or do not transcribe effector gene *Pfs1|1792*. Effector gene *Pfs1|1509* could not be associated with the breaking of known spinach R-loci. *Pfs* races that are lacking effector gene *Pfs1|1509* are together in genetic group C. This indicates that the deletion *Pfs1|1509* has likely occurred in a shared ancestral *Pfs* lineage. The absence of this effector did not reduce the virulence of these races as they have persisted over time.

To determine if an amino acid substitution in an effector might have led to the loss of recognition, we looked at homozygous non-synonymous (HNS) variants that alter the amino acid sequence of effectors in different *Pfs* races. Comparison of HNS variants in effector genes of different *Pfs* races with the differential set of spinach cultivars yielded three effector substitutions with a strong association with the breaking of resistance loci in spinach. We found HNS variants in effector gene *Pfs1|3989* that associate with the breaking of the *RPF5* resistance locus. In the same way, we found that HNS variants in effectors *Pfs1|6786* and *Pfs1|9985* are associated with the breaking of *RPF2*.

In this study, we only evaluated a linear association of effector changes to the breaking of R-loci in the differential set. The more complex relationships of multiple effectors combined and other *Pfs* proteins in relation to host immunity could be further studied by advanced computational approaches, e.g. machine learning. Also, the inclusion of more resistance-breaking isolates and differential spinach lines with novel combinations of *RPF* loci can reveal additional effectors that contain polymorphism linked to the breaking of a specific R-loci.

Overall this method of comparing R-loci to variations in effector genes of resistance breaking races has yielded several *Pfs* effectors that are likely recognized by R proteins. These candidates can now be validated to determine if recognition is R-gene specific.

***Pfs* genomics: what can we still expect?**

Although the assembled genome of *Pfs1* has proven to be a good resource to identify effectors and to study the evolution of different *Pfs* races and isolates, an improved assembly would greatly expand on the results obtained so far. A major point that can be improved is the contiguity of the genome assembly. This can be done with the inclusion of additional long reads with for example the newer version of the PacBio sequencing platform, or with Oxford Nanopore [59]. The large number of long sequencing reads (> 10 kb) obtained with these technologies have great potential to span the repeats regions in the genome. However, to do this a large number of long reads must be generated since many sequences will belong to other microbial sources due to the obligate biotrophic nature of the downy mildew *Pfs*. Alternatively, various strategies could be employed to reduce the bacterial growth on the host plant, or on the separation of bacteria from the spores after their collection. Furthermore, optical mapping could be used to correctly rearrange and connect scaffolds and contigs in the assembly.

In order to determine the full complement of genes and effectors present in the isolates within the *Pfs* species, a pan-genome (or supra-genome) should be generated. However, the construction of a *Pfs*

pan-genome is challenging and requires a large amount of long-read data from various races and isolates to include as much of the genotypic diversity in this species [60]. Nonetheless, a *Pfs* pan-genome would provide an excellent basis to determine which genes are present or absent in each isolate to elucidate isolate-specific effectors. The alignment of reads from individual isolates to the pan-genome will allow studying nucleotide variations in effectors between *Pfs* isolates [61].

Overall, these improved assemblies would lead to significantly longer contigs, possibly chromosomal level contigs. This allows better filtering on taxonomy leading to a clean genome and providing a more comprehensive set of genes and effectors for further study.

Final remarks and outlook

The results presented in this thesis have increased the understanding of the biology of *Pfs*, the downy mildew pathogen of spinach and its adaptation to resistant spinach cultivars. In addition, we gained knowledge of how the different *Pfs* races and isolates are related and how they are possibly evolutionary linked. Moreover, this study has revealed effector candidates with variations in different races that are associated with the breaking of a specific R-loci in the host. However, an important step that still remains is the application of the knowledge gained from this study to provide strategies for improved resistance in spinach.

The use of single R-loci or the combination of two R-loci in spinach lines has been successfully used in breeding for resistance against *Pfs*. However, this strategy has until now often been a short-lived success as new *Pfs* variants emerged that could overcome the introduced resistance. To combat this, new breeding strategies are necessary to ensure the future availability of *Pfs* resistant spinach crops.

In the assembled genome of *Pfs*, we have identified many *Pfs* effectors, that can be used to identify novel or match known R-genes in spinach. These *Pfs* effectors could potentially serve as markers that can be employed in *in planta* expression assays [62]. Recognition assays can be used to identify spinach lines or wild germplasm that recognize multiple effectors and thus potentially confer broad-spectrum resistance [63; 64]. Besides, this could reveal R-genes that can recognize more or ideally all allelic variants (except deletions) of resistance-breaking effectors, that were found in the different *Pfs* races [65]. The use of effectors in recognition assays also has the potential to accelerate resistance breeding against this pathogen as the recognition of a single resistance-breaking effector can be tested on (hybrid) spinach lines that are still under development [65].

In this study, we identified effectors with homologs in different oomycete species. R-genes that are known to recognize the effector homolog in other plant species can likely also recognize the homologous *Pfs*

effector [66]. In turn, this information can be used to identify an R-gene analog (RGA) in the (wild-)spinach pool based on sequence similarity [67]. Once such an R-gene analog is found it can be introgressed into cultivated spinach lines [68; 69].

Alternatively, *Pfs* effectors can be used to identify susceptibility genes (S-genes) in the spinach host. S-genes encode plant proteins that are the target of effectors employed by the pathogen to promote disease development [70]. Consequently, removal or mutation of the S-gene would render the host resistant to the pathogen [71]. Effector mediated S-gene identification can be done with for example a yeast-two-hybrid (Y2H) method. This method has been usefully used in *P. infestans* where the effector target of effector AVR3a interacts with the host U-box E3 ligase CMPG1 protein. The E3 ligase CMPG1 is required for (INF1)-triggered cell death, and the targeting by AVR3a prevents this response [72]. Also, the silencing of six orthologs of Arabidopsis S-genes in potato was shown to be effective in providing resistance to *P. infestans* [73]. Effectors identified in *Pfs* have been used to identify Y2H interactors from the plant (unpublished results Neilen *et al.*). Since stable transformation of spinach is possible [74], one can attempt to silence or knock-out an effector-identified candidate S-gene to test for loss of susceptibility to *Pfs*.

The research described in this thesis has allowed a first view on the genome of *Pfs* and has provided important insights into the differences between genomes of races that have appeared in the last 20 years. The selection of races has enabled the discovery of candidate avirulence effectors that greatly enhance our knowledge of how this downy mildew pathogen overcomes spinach resistance traits. The generation of fundamental knowledge on this pathosystem now forms the basis for advancing resistance breeding strategies. Knowing your enemy, understanding its weapons and tactics, and how this pathogen responds to host defense, forms a strong fundament for future in-depth studies. The close collaboration with breeders on this project has inspired them to think more about “the enemy”. Moreover, it challenges them to envision using these pathogen insights for resistance breeding and the future development of more durable forms of resistance. Genetic resistance is crucial for the sustainable production of this nutritious leafy green vegetable so that the spinach production is not all eaten by downy mildew but can be enjoyed by us all.

References

1. Judelson HS. 2012. Dynamics and innovations within oomycete genomes: Insights into biology, pathology, and evolution. *Eukaryot. Cell* 11:1304-12
2. Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RH, *et al.* 2006. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* 313:1261-6
3. Haas BJ, Kamoun S, Zody MC, Jiang RH, Handsaker RE, *et al.* 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* 461:393-8
4. Baxter L, Tripathy S, Ishaque N, Boot N, Cabral A, *et al.* 2010. Signatures of adaptation to obligate biotrophy in the *Hyaloperonospora arabidopsidis* genome. *Science* 330:1549-51
5. Levesque CA, Brouwer H, Cano L, Hamilton JP, Holt C, *et al.* 2010. Genome sequence of the necrotrophic plant pathogen *Pythium ultimum* reveals original pathogenicity mechanisms and effector repertoire. *Genome Biol.* 11:R73
6. Kemen E, Gardiner A, Schultz-Larsen T, Kemen AC, Balmuth AL, *et al.* 2011. Gene gain and loss during evolution of obligate parasitism in the white rust pathogen of *Arabidopsis thaliana*. *PLoS Biol.* 9:e1001094
7. Tian M, Win J, Savory E, Burkhardt A, Held M, *et al.* 2011. 454 Genome sequencing of *Pseudoperonospora cubensis* reveals effector proteins with a QXLR translocation motif. *Mol. Plant-Microbe Interact.* 24:543-53
8. Sharma R, Xia X, Cano LM, Evangelisti E, Kemen E, *et al.* 2015. Genome analyses of the sunflower pathogen *Plasmopara halstedii* provide insights into effector evolution in downy mildews and *Phytophthora*. *BMC Genomics* 16:741
9. Derevnina L, Chin-Wo-Reyes S, Martin F, Wood K, Froenicke L, *et al.* 2015. Genome sequence and architecture of the tobacco downy mildew pathogen *Peronospora tabacina*. *Mol. Plant-Microbe Interact.* 28:1198-215
10. Kobayashi M, Hiraka Y, Abe A, Yaegashi H, Natsume S, *et al.* 2017. Genome analysis of the foxtail millet pathogen *Sclerospora graminicola* reveals the complex effector repertoire of graminicolous downy mildews. *BMC Genomics* 18:897
11. Thines M, Sharma R, Rodenburg SYA, Gogleva A, Judelson HS, *et al.* 2019. The genome of *Peronospora belbahrii* reveals high heterozygosity, a low number of canonical effectors and CT-rich promoters. *bioRxiv:721027*
12. Rahman A, Gongora-Castillo E, Bowman MJ, Childs KL, Gent DH, *et al.* 2019. Genome sequencing and transcriptome analysis of the hop downy mildew pathogen *Pseudoperonospora humuli* reveal species-specific genes for molecular detection. *Phytopathology* 109:1354-66
13. Dussert Y, Mazet ID, Couture C, Gouzy J, Piron MC, *et al.* 2019. A high-quality grapevine downy mildew genome assembly reveals rapidly evolving and lineage-specific putative host adaptation genes. *Genome Biol. Evol.* 11:954-69

14. Fletcher K, Gil J, Bertier LD, Kenefick A, Wood KJ, *et al.* 2019. Genomic signatures of somatic hybrid vigor due to heterokaryosis in the oomycete pathogen, *Bremia lactucae*. *bioRxiv*:516526
15. Bohlin J, Eldholm V, Pettersson JH, Brynildsrud O, Snipen L. 2017. The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. *BMC Genomics* 18:151
16. O'Leary NA, Wright MW, Brister JR, Ciufu S, Haddad D, *et al.* 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44:D733-45
17. Promponas VJ, Iliopoulos I, Ouzounis CA. 2015. Annotation inconsistencies beyond sequence similarity-based function prediction – phylogeny and genome structure. *Stand. Genomic Sci.* 10:108
18. von Meijenfeldt FAB, Arkhipova K, Cambuy DD, Coutinho FH, Dutilh BE. 2019. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.* 20:217
19. Fletcher K, Klosterman SJ, Derevnina L, Martin F, Bertier LD, *et al.* 2018. Comparative genomics of downy mildews reveals potential adaptations to biotrophy. *BMC Genomics* 19:851-84
20. Wawra S, Belmonte R, Lobach L, Saraiva M, Willems A, van West P. 2012. Secretion, delivery and function of oomycete effector proteins. *Curr. Opin. Microbiol.* 15:685-91
21. Schornack S, Huitema E, Cano LM, Bozkurt TO, Oliva R, *et al.* 2009. Ten things to know about oomycete effectors. *Mol. Plant Pathol.* 10:795-803
22. Kamoun S. 2009. *The secretome of plant-associated fungi and oomycetes.* pp 173-180. Springer. 398 pp.
23. Anderson RG, Deb D, Fedkenheuer K, McDowell JM. 2015. Recent Progress in RXLR Effector Research. *Mol. Plant-Microbe Interact.* 28:1063-72
24. Whisson SC, Boevink PC, Moleleki L, Avrova AO, Morales JG, *et al.* 2007. A translocation signal for delivery of oomycete effector proteins into host plant cells. *Nature* 450:115-8
25. Wang S, Boevink PC, Welsh L, Zhang R, Whisson SC, Birch PRJ. 2017. Delivery of cytoplasmic and apoplastic effectors from *Phytophthora infestans* haustoria by distinct secretion pathways. *New Phytol.* 216:205-15
26. Schornack S, van Damme M, Bozkurt TO, Cano LM, Smoker M, *et al.* 2010. Ancient class of translocated oomycete effectors targets the host nucleus. *Proc. Natl. Acad. Sci. U. S. A.* 107:17421-6
27. Kamoun S. 2006. A catalogue of the effector secretome of plant pathogenic oomycetes. *Annu. Rev. Phytopathol.* 44:41-60
28. Dou D, Kale SD, Wang X, Jiang RH, Bruce NA, *et al.* 2008. RXLR-mediated entry of *Phytophthora sojae* effector Avr1b into soybean cells does not require pathogen-encoded machinery. *The Plant Cell* 20:1930-47

29. Sharma R, Xia X, Cano LM, Evangelisti E, Kemen E, *et al.* 2015. Genome analyses of the sunflower pathogen *Plasmopara halstedii* provide insights into effector evolution in downy mildews and *Phytophthora*. *BMC Genomics* 16:741
30. Rovenich H, Boshoven JC, Thomma BP. 2014. Filamentous pathogen effector functions: of pathogens, hosts and microbiomes. *Curr. Opin. Plant Biol.* 20:96-103
31. Raffaele S, Kamoun S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat. Rev. Microbiol.* 10:417-30
32. Dong S, Raffaele S, Kamoun S. 2015. The two-speed genomes of filamentous pathogens: waltz with plants. *Curr. Opin. Genet. Dev.* 35:57-65
33. Bosi E, Mascagni F. 2019. Less is more: Genome reduction and the emergence of cooperation—implications into the coevolution of microbial communities. *Int. J. Genomics* 2019
34. Weinert LA, Welch JJ. 2017. Why might bacterial pathogens have small genomes? *Trends Ecol. Evol.* 32:936-47
35. Gardner MJ, Hall N, Fung E, White O, Berriman M, *et al.* 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419:498-511
36. Morrison HG, McArthur AG, Gillin FD, Aley SB, Adam RD, *et al.* 2007. Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. *Science* 317:1921-6
37. Hudspeth DSS, Nadler SA, Hudspeth MES. 2000. A COX2 molecular phylogeny of the peronosporomycetes. *Mycologia* 92:674-84
38. Choi YJ, Thines M, Han JG, Shin HD. 2011. Mitochondrial phylogeny reveals intraspecific variation in *Peronospora effusa*, the spinach downy mildew pathogen. *J. Microbiol. (Seoul)* 49:1039-43
39. Martin FN, Zhang YH, Cooke DEL, Coffey MD, Grunwald NJ, Fry WE. 2019. Insights into evolving global populations of *Phytophthora infestans* via new complementary mtDNA haplotype markers and nuclear SSRs. *PLoS ONE* 14:e0208606
40. Lyon R, Correll J, Feng C, Bluhm B, Shrestha S, *et al.* 2016. Population structure of *Peronospora effusa* in the southwestern United States. *PLoS ONE* 11:e0148385
41. Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406-25
42. Win J, Morgan W, Bos J, Krasileva KV, Cano LM, *et al.* 2007. Adaptive evolution has targeted the C-terminal domain of the RXLR effectors of plant pathogenic oomycetes. *Plant Cell* 19:2349-69
43. Goss EM, Press CM, Grunwald NJ. 2013. Evolution of RXLR-class effectors in the oomycete plant pathogen *Phytophthora ramorum*. *PLoS ONE* 8:e79347
44. Jiang RHY, Tripathy S, Govers F, Tyler BM. 2008. RXLR effector reservoir in two *Phytophthora* species is dominated by a single rapidly evolving superfamily with more than 700 members. *Proc. Natl. Acad. Sci. U. S. A.* 105:4874-9

45. Thines M, Kamoun S. 2010. Oomycete-plant coevolution: recent advances and future prospects. *Curr. Opin. Plant Biol.* 13:427-33
46. Jiang RH, Tyler BM. 2012. Mechanisms and evolution of virulence in oomycetes. *Annu. Rev. Phytopathol.* 50:295-318
47. Stassen JH, Van den Ackerveken G. 2011. How do oomycete effectors interfere with plant life? *Curr. Opin. Plant Biol.* 14:407-14
48. Correll JC, Bluhm BH, Feng C, Lamour K, du Toit LJ, Koike ST. 2011. Spinach: better management of downy mildew and white rust through genomics. *Eur. J. Plant Pathol.* 129:193-205
49. Correll J, du Toit L, Koike S, van Ettehoven K. 2010. Guidelines for spinach downy mildew: *Peronospora farinosa* f. sp. *spinaciae* (Pfs).
50. Irish BM, Correll JC, Koike ST, Morelock TE. 2007. Three new races of the spinach downy mildew pathogen identified by a modified set of spinach differentials. *Plant Dis.* 91:1392-6
51. Frantzeskakis L, Di Pietro A, Rep M, Schirawski J, Wu CH, Panstruga R. 2019. Rapid evolution in plant-microbe interactions – a molecular genomics perspective. *New Phytol.*
52. Petit-Houdenet Y, Fudal I. 2017. Complex interactions between fungal avirulence genes and their corresponding plant resistance genes and consequences for disease resistance management. *Front. Plant Sci.* 8:1072
53. Stavrinides J, McCann HC, Guttman DS. 2008. Host-pathogen interplay and the evolution of bacterial effectors. *Cell. Microbiol.* 10:285-92
54. Stukenbrock EH, McDonald BA. 2009. Population genetics of fungal and oomycete effectors involved in gene-for-gene interactions. *Mol. Plant-Microbe Interact.* 22:371-80
55. Aguilera G, Refregier G, Yockteng R, Fournier E, Giraud T. 2009. Rapidly evolving genes in pathogens: methods for detecting positive selection and examples among fungi, bacteria, viruses and protists. *Infect. Genet. Evol.* 9:656-70
56. Wang Y, Tyler BM, Wang Y. 2019. Defense and counterdefense during plant-pathogenic oomycete infection. *Annu. Rev. Microbiol.* 73:667-96
57. Allen RL, Meitz JC, Baumber RE, Hall SA, Lee SC, *et al.* 2008. Natural variation reveals key amino acids in a downy mildew effector that alters recognition specificity by an Arabidopsis resistance gene. *Mol. Plant Pathol.* 9:511-23
58. Gilroy EM, Breen S, Whisson SC, Squires J, Hein I, *et al.* 2011. Presence/absence, differential expression and sequence polymorphisms between PiAVR2 and PiAVR2-like in *Phytophthora infestans* determine virulence on R2 plants. *New Phytol.* 191:763-76
59. Kono N, Arakawa K. 2019. Nanopore sequencing: Review of potential applications in functional genomics. *Development, growth & differentiation* 61:316-26
60. Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, *et al.* 2015. High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat. Commun.* 6:6914

61. McCarthy CGP, Fitzpatrick DA. 2019. Pan-genome analyses of model fungal species. *Microb Genom* 5:e000243
62. Vleeshouwers VG, Rietman H, Krenek P, Champouret N, Young C, *et al.* 2008. Effector genomics accelerates discovery and functional profiling of potato disease resistance and *Phytophthora infestans* avirulence genes. *PLoS ONE* 3:e2875
63. Lokossou AA, Rietman H, Wang M, Krenek P, van der Schoot H, *et al.* 2010. Diversity, distribution, and evolution of *Solanum bulbocastanum* late blight resistance genes. *Mol. Plant-Microbe Interact.* 23:1206-16
64. Van Der Vossen E, Sikkema A, Hekkert BtL, Gros J, Stevens P, *et al.* 2003. An ancient R gene from the wild potato species *Solanum bulbocastanum* confers broad-spectrum resistance to *Phytophthora infestans* in cultivated potato and tomato. *The plant journal* 36:867-82
65. Vleeshouwers VGAA, Oliver RP. 2014. Effectors as tools in disease resistance breeding against biotrophic, hemibiotrophic, and necrotrophic plant pathogens. *Mol. Plant-Microbe Interact.* 27:196-206
66. McDowell JM, Woffenden BJ. 2003. Plant disease resistance genes: recent insights and potential applications. *Trends Biotechnol.* 21:178-83
67. Huang S, van der Vossen EA, Kuang H, Vleeshouwers VG, Zhang N, *et al.* 2005. Comparative genomics enabled the isolation of the *R3a* late blight resistance gene in potato. *Plant J.* 42:251-61
68. Brugmans BW, Meeuwssen J, Nooyen C. 2015. Compositions and methods for *peronospora* resistance in spinach. Seminis Vegetable Seeds Inc. *United States Patent No. US20150240256A1*
69. Xu C, Jiao C, Sun H, Cai X, Wang X, *et al.* 2017. Draft genome of spinach and transcriptome diversity of 120 *Spinacia* accessions. *Nat. Commun.* 8:15275
70. Lapin D, Van den Ackerveken G. 2013. Susceptibility to plant disease: more than a failure of host immunity. *Trends Plant Sci.* 18:546-54
71. Gust AA, Brunner F, Nurnberger T. 2010. Biotechnological concepts for improving plant innate immunity. *Curr. Opin. Biotechnol.* 21:204-10
72. Bos JI, Armstrong MR, Gilroy EM, Boevink PC, Hein I, *et al.* 2010. *Phytophthora infestans* effector AVR3a is essential for virulence and manipulates plant immunity by stabilizing host E3 ligase CMPG1. *Proc. Natl. Acad. Sci. U. S. A.* 107:9909-14
73. Sun K, Wolters AM, Vossen JH, Rouwet ME, Loonen AE, *et al.* 2016. Silencing of six susceptibility genes results in potato late blight resistance. *Transgenic Res.* 25:731-42
74. Bao JH, Chin DP, Fukami M, Ugaki M, Nomura M, Mii M. 2009. Agrobacterium-mediated transformation of spinach (*Spinacia oleracea*) with *Bacillus thuringiensis cry1Ac* gene for resistance against two common vegetable pests. *Plant Biotechnology* 26:249-54

Summary

The consumption of spinach (*Spinacia oleracea*) has increased substantially during recent decades. To meet this growing demand, spinach growers must limit crop loss due to pathogens. The biggest threat to spinach cultivation is the downy mildew *Peronospora effusa* (*Pfs*), a pathogen belonging to the oomycetes. *Pfs* is an obligate biotroph, meaning that it only grows on living host tissue. Although some fungicides are effective against this pathogen, the appearance of fungicide-resistance and reduction in use of agro-chemicals, makes this form of protection less suitable. A more durable form of protection against this pathogen is by making use of genetic resistance mediated by resistance (R)-genes. These *RPF* genes mediate the recognition of the intruding pathogen resulting in an effective resistance response. To sustain an infection, the pathogen must, therefore, prevent or suppress immune responses of the host. *Pfs*, like other oomycetes, produces host-translocated effectors that can suppress defense responses or otherwise enhance susceptibility of the host. However, host-translocated effectors can also be recognized by R-proteins, and trigger plant immunity. This creates a strong selection pressure on the genome of the pathogen to adapt its recognized effector genes and thus avoid recognition. Alternatively, new effectors can evolve that can block the immune response.

Overall, the use of R-genes on a wide scale has resulted in the rapid increase of *Pfs* isolates capable of breaking the employed resistances. This has led to the identification of more than 16 resistance-breaking *Pfs* races and numerous isolates within the last 20 years. A better understanding of the tools that this pathogen uses for infection, and how it can adapt to resistance in the host is needed to better control this pathogen and prevent crop loss. A description of the *Pfs* pathogen, its infection mechanisms and effectors, and how to study them using bioinformatic approaches is provided in **Chapter 1**.

To study the *Pfs* genes and effectors, the genome of race *Pfs1* was sequenced and assembled (**Chapter 2**). Due to the obligate biotrophic nature of this pathogen it cannot be cultured axenically, which led to the unintended sequencing of many other microbes present on the spinach leaves. To effectively filter sequences belonging to other microbes and remove them from the assembly, we used the CAT tool capable of determining the taxonomic origin of long error-corrected PacBio sequences and contig sequences resulting from a pre-assembly. This resulted in a clean *Pfs1* genome assembly of 32.4 Mb. In addition *Pfs1* mRNA was sequenced to aid gene prediction, resulting in 13,227 gene models. From the corresponding peptide sequences 99 RxLR(-like), and 14 Crinkler(-like) effector candidates were identified.

Besides *Pfs1*, we also sequenced 16 denominated *Pfs* races and 8 field isolates. In **Chapter 3** we describe the identification of genomic variants between the *Pfs* isolates and the reference sequence (*Pfs1*) and their use for determining the phylogenetic relationship between the isolates. The results indicate the presence of three distinct genetic groups, and one admixed group of *Pfs* isolates. The *Pfs1* mitochondrial reference genome was assembled and used to identify two distinct mitochondrial haplotypes in the *Pfs* population. The genome and mitochondrial analyses showed that there is not a linear evolutionary relationship between the denominated *Pfs* races, and that sexual recombination has likely played an important role in the emergence of new *Pfs* races.

In **Chapter 4** we analyzed the effectors found in *Pfs1* further revealing that many show sequence similarities and can be grouped, indicating that they have evolved by duplication and diversification. We found that effectors accumulate non-synonymous polymorphism at a higher rate compared to other genes in the genome. We also determined that two effector genes were absent in certain isolates under study, of which the loss of one effector could be partially correlated to the breaking the *RPF3*-locus. Besides, amino acid substitutions in two effector proteins could be directly correlated to the breaking of *RPF2*, and substitutions in another effector were correlated to the breaking of *RPF5*. These candidates can now be used to determine if recognition is R-gene specific. Overall, the results presented in this thesis have increased our understanding of the biology of *Pfs*, the downy mildew pathogen of spinach and its adaptation to resistant spinach cultivars.

In **Chapter 5** the main findings of this thesis are summarized and discussed in a broader perspective. In addition, this chapter gives future challenges and provides an outlook on the application of the findings reported in this thesis for resistance breeding in spinach.

Samenvatting

De consumptie van spinazie (*Spinacia oleracea*) is de afgelopen decennia aanzienlijk toegenomen. Om aan deze groeiende vraag te voldoen, moeten spinazietelers gewasverlies door ziekteverwekkers beperken. De grootste bedreiging voor de teelt van spinazie is de valse meeldauw *Peronospora effusa* (*Pfs*), een pathogeen die behoort tot de oomycete klasse. *Pfs* is een obligaat biotroof organisme, wat betekent dat het alleen kan groeien op een levende gastheer. Hoewel sommige fungiciden effectief zijn tegen deze ziekteverwekker, leidt het herhaald gebruik tot fungicide resistentie. Daarnaast zijn steeds minder gewasbeschermingsmiddelen toegestaan wat deze vorm van bescherming minder geschikt maakt. Een duurzamere methode tegen deze ziekteverwekker kan worden verkregen door gebruik te maken van genetische resistentie doormiddel van resistentie-loci (R-loci). De R-loci, genaamd *RPF* (Resistance to *Peronospora farinosa*) coderen voor eiwitten die de binnendringende pathogeen kunnen detecteren en zorgen zo in voor een effectieve afweerreactie.

Voor een succesvolle infectie moet de ziekteverwekker dus deze immunoreactie voorkomen of onderdrukken in de gastheer. De *Pfs*-ziekteverwekker produceert net als andere oomyceten “gastheer getransloceerde effectoren” die in staat zijn om de afweerreactie te onderdrukken of de vatbaarheid van de gastheer vergroten. Deze effectoren kunnen echter ook potentiële doelwitten zijn van R-genen en een immunorespons veroorzaken. Dit creëert een sterke selectiedruk op het genoom van de ziekteverwekker om de eiwitsequentie van de herkende effector aan te passen of herkenning te voorkomen door nieuwe effectoren te ontwikkelen.

Deze selectiedruk resulteerde in een snelle toename van het aantal resistentie brekende *Pfs*-isolaten en heeft geleid tot de identificatie van meer dan 16 nieuwe rassen en talloze *Pfs*-isolaten in de afgelopen 20 jaar. Een beter begrip van de infectie mechanismes die deze pathogeen gebruikt is daarom hard nodig voor een effectieve bestrijding. Een gedetailleerd overzicht van de infectiemechanismen van oomyceten en *Pfs* in het bijzonder, maar ook hoe deze zijn te bestuderen met behulp van de bioinformatica is te vinden in **hoofdstuk 1**.

Om het *Pfs*-genoom en de effectoren te bestuderen hebben we allereerst de genoomsequentie van *Peronospora effusa*-ras: 1 gesequenced en geassembleerd (**hoofdstuk 2**).

Doordat *Pfs* een obligaat biotroof organisme is wat niet gekweekt kan worden op dood materiaal onder steriele condities zijn er onbedoeld sequentie) geïsoleerd die behoren tot andere microben die van nature aanwezig zijn op de spinaziebladeren. Om deze sequenties er effectief uit te filteren en te verwijderen uit de genoomsequenties, hebben wij gebruik

gemaakt van de CAT-methode, die in staat is om van lange sequenties de taxonomisch oorsprong te bepalen. Dit resulteerde in een schoon *Pfs1* genoom met een lengte 32,40 Mb. Naast de nucleaire genoom sequenties van *Pfs1* werden ook mRNA-sequenties bepaald door middel van korte reads. Deze mRNA reads werden gebruikt om de nauwkeurigheid van de gen-modelvoorspelling te verbeteren, wat resulteerde in 13.227 gen-modellen. Op basis deze gen-modellen hebben we 99 R_xLR(-achtige) effectoren en 14 Crinkler(-achtige) effectoren geïdentificeerd.

Naast het genoom van *Pfs1* hebben we ook sequenties verkregen (reads) van 16 andere gedomineerde *Pfs*-rassen (fysios) en 8 recente isolaten. In **hoofdstuk 3** beschrijven de varianten tussen de *Pfs*-isolaten en de referentiesequentie. Op basis van deze varianten hebben we verschillende fylogenetische analyses toegepast om de relatie tussen de isolaten te bepalen. Dit resulteerde in de identificatie van drie verschillende genetische groepen en een gemengde groep in de *Pfs*-isolaten. Ook hebben we ook een mitochondriaal genoom samengesteld voor *Pfs1* met behulp van de genomische reads en deze gebruikt voor de identificatie van twee verschillende mitochondriale haplotypen in de *Pfs*-populatie. In het algemeen lieten de resultaten zien dat er geen lineair evolutionair verband bestaat tussen de *Pfs*-rassen en dat seksuele recombinatie een rol heeft gespeeld bij het ontstaan van een aantal van de *Pfs*-rassen en isolaten.

In **hoofdstuk 4** hebben we effector genen, geïdentificeerd in het genoom van *Pfs1* verder geanalyseerd. De analyse liet zien dat veel effector sequenties vergelijkbaar zijn en kunnen worden gegroepeerd, wat aangeeft dat ze zijn geëvolueerd door duplicatie en diversificatie. Ook hebben we gevonden dat de peptide sequentie van effectoren versneld veranderde ten opzichte van andere eiwitten in het genoom. We hebben ook vastgesteld dat twee effectoren niet aanwezig waren in sommige isolaten, waarvan een het verlies van een effector nagenoeg overeenkwam met het doorbreken van het *RPF3*-locus. Daarnaast konden wij veranderingen in de aminozuursequenties van twee effector genen direct correleren aan het breken van een enkele *RPF2*-locus en substituties in een andere effector aan het doorbreken van *RPF5*. Deze effector kandidaten kunnen nu worden gebruikt om te bepalen of herkenning R-gen specifiek is.

Samenvattend hebben de resultaten beschreven in dit proefschrift bijgedragen aan ons begrip van de biologie van *Pfs*, de valse meeldauw pathogeen van spinazie en hoe deze pathogeen zich aanpast aan resistente spinazie gewassen.

In **hoofdstuk 5** worden de belangrijkste bevindingen van dit proefschrift samengevat en besproken in een breder perspectief. Bovendien geeft dit hoofdstuk een gedetailleerd overzicht van toekomstige uitdagingen en een overzicht van de toepassing van de bevindingen in dit proefschrift voor resistentie veredeling in spinazie.

Dankwoord

Dat was het dan. Na jaren werk voeg ik het laatste stuk toe aan mijn thesis. Mijn promotieonderzoek is afgerond, het vele werk gecondenseerd tot dit bescheiden boekje. Nu mijn promotie bijna achter de rug is en de rust langzaam weerkeert, kijk ik terug op de afgelopen jaren. Jaren die voorbij zijn vlogen, maar waar ook veel in is gebeurd. Waar ik als net afgestudeerde Wageningse student de vakgroep Plant-microbe interacties binnenliep, ging wonen in de grote stad Utrecht en nu dan eindelijk de deuren van de universiteit achter mij kan sluiten met het beoogde resultaat. Een tijd waarin veel is gebeurd, ik veel heb meegemaakt en ook als persoon ben veranderd en gegroeid. Er waren hoogtepunten waarop we successen mochten vieren maar zeker ook dieptepunten waarin ik me soms alleen voelde. Maar alleen wás ik zeker niet. Ik ben gezegend met veel mensen om mij heen, vrienden, familie en collega's die elke keer daar waren wanneer je ze nodig had. Die interesse toonden, betrokken waren en mij mede vormden tot wie ik nu ben.

Allereerst wil ik Guido bedanken, als promotor stond je altijd voor mij klaar. Je wist mijn betere kanten naar boven te halen en mij te begeleiden op een manier die bij mij past. Je eindeloze geduld, hulp en bemoediging waren onmisbaar op de momenten waar ik door de bomen soms het bos niet meer zag. Maar ook dank voor de leuke en interessante gesprekken tijdens congressen en tal van andere momenten die we hebben mogen delen. In de laatste maanden heb ik ook vaak gedacht: "Guido begeleidt niet alleen promovendi, hij maakt ze ook!"

But I also obtained more than just a title from my time at the PMI group. From the first moment I saw Yang I was already impressed by her. Although it took a while before I got her attentions well, I can now call myself a fortunate man with her by my side. I can vividly remember the first time I really spoke with her at the PMI Christmas dinner about three years ago. Since then we have been inseparable, and she has enriched my life in many ways. I never had to regret my choice, Yang was the one who motivated me in difficult times, was inextricably by my side and was always there for me. My promotion would never have happened without you and I hope we can experience many magnificent moments in our lives together.

Daarnaast mocht ik ook altijd rekenen op mijn familie, mijn ouders en broertjes die betrokken waren, mij steunden en waar ik altijd het vertrouwen heb om op terug kon vallen. Lieve papa en mama dank dat jullie er altijd voor mij zijn. Ruben en Chris, ook al zijn we belast met dezelfde genen, als de wil er is, dan lukt het altijd wel. Ik wil ook in het bijzonder mijn opa en oma Kok bedanken waar ik bij mijn start van mij promotie mocht wonen. Ik denk nog wel eens terug aan de tijd dat wij bij ome Huub en tante

Emma langs gingen en met tante Riek als date naar hun jubileumfeest. Jullie waren altijd betrokken, geïnteresseerd en altijd bereid om te helpen en stonden altijd voor mij klaar. Ook dank aan oma Klein voor uw betrokkenheid en interesse al die jaren en de vele wijze woorden die u met mij deelde. Veel dank ook aan mij ooms en tantes, neefjes en nichtjes voor jullie interesse in de malle neef die zijn tijd vulde met spinazie-onderzoek.

Naast familie waren er ook veel andere mensen die ik praktisch mijn hele leven al heb mogen kennen. In het bijzonder wil ik de familie Wesdorp bedanken. Met veel plezier hebben we de afgelopen jaren elkaar nog regelmatig mogen zien. Ik denk nog vaak terug aan de tijd dat ik als kind samen met mijn broertjes, vader, Leen, Hanna, Ella en Mirjam naar het strand gingen. Ook de familie Schot heb ik nog regelmatig kunnen zien, zeker nu Johan werkt bij de Universiteit Utrecht. Regelmatig mocht ik ook met veel plezier op jullie kat Maxime/Mifan passen die mij de nodige afleiding gaf.

Na mijn verhuizing naar Utrecht viel ik in een warm bad. Al vrij snel voelde ik me hier thuis met oude Wageningse vrienden. Max, met jou heb ik al veel mogen meemaken maar ook in Utrecht hebben we elkaar vaak kunnen zien om weer even bij te praten. Piet, dank voor alle gezellig avonden waarin we vele diepzinnige discussies mochten hebben. Als de appels het toelieten was Pierre er ook bij om de avond wat meer luister bij te zetten. Wat fijn dat jullie daar al waren en dankjewel voor de mooie tijd die we hebben gehad in Utrecht, vaak ook samen met andere leden van de Heeren van Dodona. Het was goed om af en toe een vertrouwd gezicht te zien en even bij te kunnen praten.

Maar naast oude bekenden heb ik ook veel nieuwe mensen mogen leren kennen in Utrecht. Andries sinds de eerste keer dat we samen met Eefje naar het strand gingen was er al meteen een klik. Dank voor alle avonturen en goede gesprekken die we de afgelopen jaren mochten hebben. Eefje, met jou heb ik veel lief en leed over het PhD-traject mogen delen. Hoe we vaak tegen dezelfde dingen aanliepen en er gelijkgestemd over dachten. Romee, het was altijd een aangename verrassing als je onder mijn raam stond te fluiten of op de deur klopte voor een drankje en een goed gesprek bij Orloff. Dank voor het organiseren van de vele burenborels samen met Arnold en Mirjam, jullie interesse en spontane gesprekken in en om het portiek. Maar natuurlijk wil ik ook graag Esther en Freek bedanken. Onze diners samen en spontane borels waren altijd vol van gezelligheid en door jullie open blik op de wereld voelde ik me al gauw bij jullie thuis.

Ook via de ICF heb ik veel mensen leren kennen. Bedankt Joan dat jij de afgelopen jaren mijn geestelijk begeleider wilde zijn en dat we openhartig over dingen konden praten. Ruita, de dingen zijn niet altijd zo gelopen als we zelf hadden gedacht, maar bedankt voor de fijne tijd samen. Ook ben ik Willy dankbaar. Jouw tomeloze inzet waarmee je iedereen met

open armen ontving zal ik nooit vergeten. Marc-Etienne, jij bent al wat verder in de academische wereld en mede daardoor heb ik veel van je kunnen leren. Dank dat je ook als buurman altijd klaarstond om te helpen mij van advies te voorzien en de goede gesprekken. Pierre en Anke, het was altijd gezellig om jullie te bezoeken in Houten en te ontmoeten in de kerk en met jullie van gedachte te wisselen.

In het hoge Noorden werden mijn avonden nog regelmatig verrijkt met het zien van mensen die ik al praktisch vanaf de basisschool kende. Hillebrant, Gerlach en Wimer, Friesland voelde voor mij warmer aan als we de afgelopen jaren nog wel eens een spelletje speelden of elkaar zagen.

Maar de meeste tijd heb ik natuurlijk doorgebracht met mijn collega's. Ik mocht mij welkom voelen in deze zeer diverse vakgroep. Uiteraard op de eerste plaats de 'Guidogroep'. Toen ik vrijwel op dag één Alexandra ontmoette was het al meteen duidelijk, wij komen van twee verschillende planeten. Ondanks dat, heb ik je leren kennen als een zeer gewaardeerde collega waarmee ik een ontzettende leuke tijd mee heb gehad. Het was ook een eer om als paranimf op jouw promotie aanwezig te mogen zijn.

Manon, met jou heb ik toch wel echt de meeste tijd doorgebracht. Ik had je al gewaarschuwd, je gaat mij meer zien dan je man! Samen hebben wij het spinazie-project tot grotere hoogte mogen brengen. Dank voor al je inzet, vertrouwen en gezelligheid. Annemiek, je was al die jaren als een moeder voor mij. Bij jou kon ik altijd terecht om lief en leed te delen en heb je leren kennen als een warm en fijn persoon. Ook aan de andere technicus in de "Guidogroep" heb ik veel gehad. Kim dank voor alle gezelligheid en je bereidheid om mij af toe met experimenten uit de brand te helpen. Joyce dank voor al je inzichten, suggesties en goede gesprekken. Pim, ondanks dat je snel na mijn komst vertrok uit de groep heb ik veel aan je gehad. Jij hebt mij de klappen van de zweep leren kennen en mij in korte tijd geïntroduceerd bij de vakgroep.

Tijmen, wat hebben we veel gelachen de afgelopen jaren. Jij was altijd wel in voor een verzetje en alles kon gezegd worden. Ook dank aan alle andere "Guido's" voor de goede tijd. Nora ik zal je nooit vergeten, hoewel we soms tegenpolen waren, waardeerde ik onze openhartig gesprekken zeer. Marjolein, Pim, Marrit, Sebastian, Tom, Claudia, Dongping, Nicole dank voor alle gezelligheid en bemoedigingen. Tijdens mijn promotie heb ik ook met veel genoegen een aantal studenten mogen begeleiden tijdens hun onderzoeksstage. Joris, Lotte, Rolf, Edwin, Bjorn en Sjoerd dank voor jullie inzet en bijdrage aan dit werk en jullie interesse in de valse meeldauw. Ik hoop dat ik wat liefde voor dit vak heb kunnen overdragen en jullie een stap verder op weg heb kunnen helpen in jullie verdere carrière.

Dank ook aan alle PI's uit de groep. Corné van jouw sterke verhalen stonden mijn oren altijd te klapperen maar ik heb er met genoeg naar geluisterd. Dank ook voor je inzet voor de groep als geheel en mij persoonlijk,

je betrokkenheid en hulp. Ronnie, het was fijn om de afgelopen jaren weer een ervaren bio-informaticus in de groep te mogen hebben. Dank voor al je suggesties en tips die me elke keer weer een stap verder hielpen. Saskia, fijn dat we regelmatig van gedachten konden wisselen over allerlei onderwerpen. Dank ook voor al je bemoedigende woorden de afgelopen jaren. Roeland, het heeft even geduurd voordat jij mijn relatie met Yang wilde goedkeuren. Dank dat je wilde ontfermen over Yang en fijn dat je uiteindelijk mocht concluderen dat we allebei het beste met haar voor hebben. Samen met deze mensen hebben ook vele collega's en oud-collega's uit binnen – en buitenland deze vakgroep vormgegeven. Hans bedankt voor alle prachtige foto's. Anja dank voor je inzet en de goede gesprekken. Dankjewel, Gilles, Sietske, Niels, Sanne, Merel, Paul, Irene, Lotte, Eline en Colette. Also special thanks to my many international colleagues Giannes, Juan, Dharani, Shu-Hua, Miguel, Ke, Erqin, Richard, Marciel, Hao, Yeling, Gonzalo, Marco, Alberto, Pauline and Silvia. With all your experiences from all over the world, you made this group a very interesting and versatile place to work. Ook ben ik Johanna Westerdijk veel verschuldigd. Ik heb niet de eer gehad om u in persoon te mogen ontmoeten maar dankzij u heb ik wel vele internationale congressen kunnen bijwonen.

Dank ook de mensen van de bedrijven die betrokken waren bij dit project. Mathieu Pel, Peter Paul Damen, Beatrice Lindhout, Annoma Lokossou, Michel de Lange, David Courand, Jan de Visser, Johan Rijk, John Groenendijk en Suxian Zhu. Dank voor jullie suggesties steun en betrokkenheid. Het was altijd een genoegen om mijn voortgang met jullie te bespreken, hier heb ik ontzettend veel van geleerd.

Uiteindelijk was ik nooit aan mijn PhD begonnen zonder alle hulp en adviezen van een aantal mensen die ik heb mogen leren kennen in Wageningen. Dank je wel Ans, dat ik regelmatig met jou kon spreken je bemoedigingen, steun en begeleiding hebben zeer bijgedragen aan de voorspoedige voortgang van mijn studie biotechnologie. Ook zijn er veel docenten geweest in Wageningen die mij hebben geïnspireerd, kansen hebben geboden en geholpen. Dank je wel Huub Haaker, Sonja Isken en Marian Vermue voor alle adviezen en hulp op verschillende cruciale momenten in mijn studie. Sacco de Vries, naast docent biochemie heb ik je ook goed persoonlijk leren kennen en hebben we veel mooie momenten gehad, van kopstootjes drinken in de kroeg tot op excursie in Barcelona. Dit gaf mij een unieke inkijk op het academische leven. René Geurts, dank dat jij mijn masterthesis wilde begeleiden en dat jij mijn talent voor bioinformatica al snel opmerkte. De vrijheid die ik bij moleculair biologie heb gehad om mijn talenten te ontwikkelen en jouw persoonlijk advies zijn doorslaggevend geweest voor mijn keuze om te gaan promoveren. Henk Franssen, het was een eer om te weten dat na al die jaren jij nog steeds vertrouwen had in buffer die ik op je verjaardag heb gemaakt. Dank voor alle goede gesprekken.

ken en grappen die mij ook het plezier van het onderzoek hebben laten zien.

Hartelijk dank allemaal voor de mooie tijd en betrokkenheid van iedereen en die mij op verschillende manieren hebben geholpen en mijn leven hebben verrijkt. Ik besef dat zonder de hulp van al deze mensen ik nooit geworden was tot wat ik nu ben, een doctor in de biologie.

Curriculum vitae

Joël Klein is geboren op 31 mei 1988 te Rotterdam. Na het behalen van zijn middelbareschooldiploma begon hij met de opleiding Biotechnologie aan het Van Hall Larenstein in Leeuwarden. Overtuigd van zijn interesse in dit vakgebied begon hij, na het behalen van zijn propedeuse, met de bachelor Biotechnologie aan Wageningen University & Research (WUR). Tijdens zijn opleiding volgde Joël met succes een semester aan Cornell University in Ithaca New York. Na zijn bachelor begon hij met de masteropleiding Biotechnologie aan de WUR. Daar liep hij zijn onderzoeksstage bij de vakgroep Moleculair Biologie onder leiding van dr. ir. Rene Geurts. Tijdens deze stage onderzocht hij de droogteresistentie van wortelnodulerende bacteriën (*Rhizobium*). Hij raakte geïnteresseerd in de bioinformatica en de toepassing in het onderzoek. Zijn tweede stage voerde Joël uit in Beijing aan de Chinese Academy Of Agricultural Sciences waar hij werkte aan de assemblage van het aardappelgenoom. Na het behalen van zijn masterdiploma Biotechnology in 2014 begon hij aan een promotietraject aan de Universiteit Utrecht bij de Plant-Microbe Interacties groep onder leiding van prof. dr. Guido van den Ackerveken. Hij onderzocht de genomische basis van resistentie doorbreking in spinazie door de valse meeldauw *Peronospora effusa*. De resultaten van zijn promotieonderzoek zijn in dit proefschrift beschreven.



List of publications

Sharma R, Xia X, Cano LM, Evangelisti E, Kemen E, Judelson H, Oome S, Sambles C, van den Hoogen DJ, Kitner M, **Klein J**, Meijer HJG, Spring O, Win J, Zipper R, Bode HB, Govers F, Kamoun S, Schornack S, Studholme DJ, Van den Ackerveken G, Thines M. 2015. Genome analyses of the sunflower pathogen *Plasmopara halstedii* provide insights into effector evolution in downy mildews and *Phytophthora*. *BMC Genomics* 16:741

Thines M, Sharma R, Rodenburg SYA, Gogleva A, Judelson HS, Xia X, van den Hoogen J, Kitner M, **Klein J**, Neilen M, de Ridder D, Seidl MF, Van den Ackerveken G, Govers F, Schornack S, Studholme DJ. 2019. The genome of *Peronospora belbahrii* reveals high heterozygosity, a low number of canonical effectors and CT-rich promoters. *bioRxiv:721027*

Pelgrom AJE, Eikelhof J, Elberse J, Meisrimler C-N, Raedts R, **Klein J**, Van den Ackerveken G. 2019. Recognition of lettuce downy mildew effector BLR38 in *Lactuca serriola* LS102 requires two unlinked loci. *Mol. Plant Pathol.* 20:240-53

Klein J, Neilen M, van Verk M, Dutilh BE, Van den Ackerveken G. 2019. Genome reconstruction of the non-culturable spinach downy mildew *Peronospora effusa* by metagenome filtering. *bioRxiv:842658*