



# Propositions

1. Whole-genome comparative genomics is not a suitable approach to reveal the basis of pathogenicity within the *Verticillium* genus.  
(this thesis)
2. Gene loss has been a main driver for *Verticillium* speciation.  
(this thesis)
3. Strictly asexually reproducing organisms are not evolutionary dead ends
4. All scientists must receive ethics training to distinguish between “having the right to do” and “what is right to do”.
5. Every biologist will be a bioinformatician in the near future.
6. Preventing food waste is a more efficient and sustainable manner for combating food shortage than increasing crop production.
7. Agricultural chemical sprays are a hazard that GMOs can prevent.

Propositions belonging to the thesis, entitled  
'Comparative genomics of the fungal genus *Verticillium*'

Xiaoqian Shi  
Wageningen, 3 May 2019

**Comparative genomics  
of the fungal genus *Verticillium***

**Xiaoqian Shi**

## **Thesis committee**

### **Promotor**

Prof. Dr B.P.H.J. Thomma  
Professor of Phytopathology  
Wageningen University & Research

### **Co-promotor**

Dr M.F. Seidl  
Assistant professor, Laboratory of Phytopathology  
Wageningen University & Research

### **Other members**

Prof. Dr B.J. Zwaan, Wageningen University & Research  
Prof. Dr A.F.J.M. van den Ackerveken, Utrecht University  
Dr S. Smit, Wageningen University & Research  
Dr A. Wittenberg, Keygene N.V., Wageningen

This research was conducted under the auspices of the Graduate School Experimental Plant Sciences.



# **Comparative genomics of the fungal genus *Verticillium***

**Xiaoqian Shi**

## **Thesis**

submitted in fulfilment of the requirements for the degree of doctor  
at Wageningen University  
by the authority of the Rector Magnificus,  
Prof. Dr A.P.J. Mol,  
in the presence of the  
Thesis Committee appointed by the Academic Board  
to be defended in public  
on Friday 3 May 2019  
at 1:30 p.m. in the Aula.

Xiaoqian Shi

Comparative genomics of the fungal genus *Verticillium*,  
196 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2019)  
With references, with summary in English

DOI: <https://doi.org/10.18174/469317>

ISBN: 978-94-6343-421-8

# Table of Contents

<b>Chapter 1</b>	General introduction and thesis outline	<b>7</b>
<b>Chapter 2</b>	The genome of the saprophytic fungus <i>Verticillium tricorpus</i> reveals a complex effector repertoire resembling that of its pathogenic relatives	<b>17</b>
<b>Chapter 3</b>	Genus-wide comparative genomics reveals similar features between pathogenic and non-pathogenic <i>Verticillium</i> species	<b>47</b>
<b>Chapter 4</b>	Evolution within the fungal genus <i>Verticillium</i> is characterized by chromosomal rearrangement and gene loss	<b>67</b>
<b>Chapter 5</b>	Dynamic virulence-related regions of the fungal plant pathogen <i>Verticillium dahliae</i> display remarkably enhanced sequence conservation	<b>95</b>
<b>Chapter 6</b>	The genome of the fungal pathogen <i>Verticillium dahliae</i> reveals extensive bacterial to fungal gene transfer	<b>117</b>
<b>Chapter 7</b>	<i>In silico</i> prediction and characterisation of secondary metabolite clusters in the plant pathogen <i>Verticillium dahliae</i>	<b>139</b>
<b>Chapter 8</b>	General discussion	<b>155</b>
	<b>Summary</b>	<b>169</b>
	<b>References</b>	<b>171</b>
	<b>Acknowledgements</b>	<b>187</b>
	<b>Curriculum vitae</b>	<b>189</b>
	<b>Publication list</b>	<b>190</b>
	<b>Education statement</b>	<b>191</b>





Chapter

1

**General introduction**

## Abstract

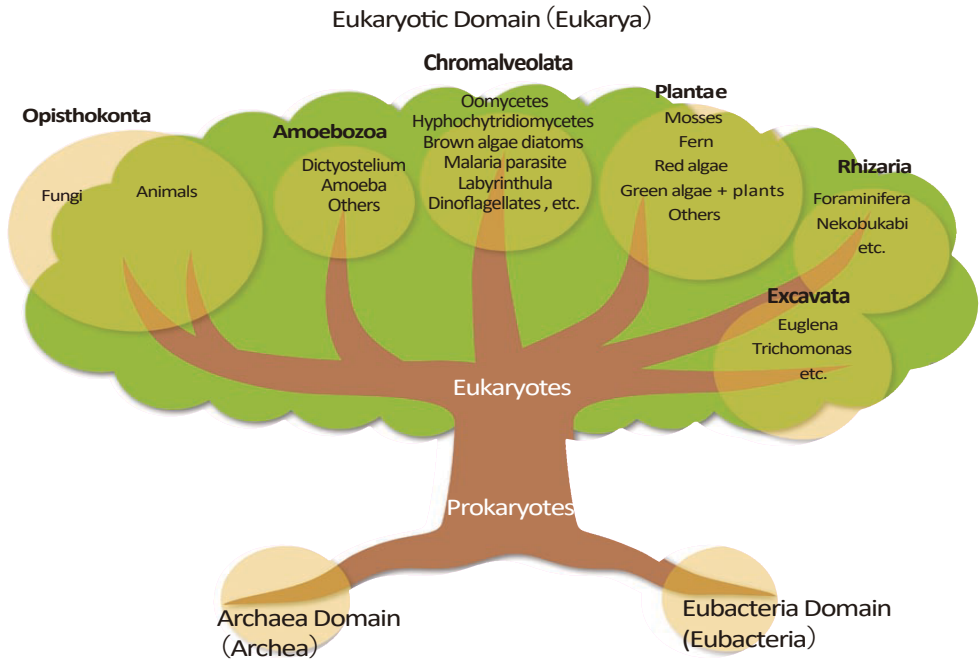
Fungi are organisms that are grouped in a distinct kingdom within eukaryotes. Like animals and most bacteria, fungi are heterotrophs, i.e. they obtain their carbon and energy from other organisms. The fungal kingdom includes many species that are able to cause considerable yield losses in crop production systems worldwide. The recent advances in genome sequencing and assembly facilitated the availability of high-quality genome assemblies of many fungal species. Consequently, comparative genomics provided insights in the evolution of pathogen genomes.

## Introduction

### Fungi and taxonomy

Species from the fungal kingdom have played important roles for human civilization. Several fungal species are consumed as food or used as active agent in processing or fermentation of food products. For example, edible mushrooms have been used directly as food sources for thousands of years. Humans have exploited the natural abilities of fungi to ferment fruits and grains to produce alcoholic beverages and bread since as early as 6000 BC (McGovern et al., 1996; Samuel, 1996; McGovern et al., 2004) and cheese since 7500 BC (Salque et al., 2013). Yeasts are used for the fermentation of bread, wine, and beer, while filamentous fungi are used for the maturation of cheeses and the production of alcohol from rice (e.g. sake) and soy sauce from soy beans (Salque et al., 2013). Fungi are also used to ferment and preserve meat from food spoilage, in particular *Penicillium nalgiovense*, which forms the white crust on salamis (Bernáldez et al., 2013). Moreover, in the 1930s the antibiotic penicillin was discovered in the fungal species *Penicillium chrysogenum*, which was subsequently developed into a life-saving medicine against bacterial infections (Ligon, 2004). Since then, a considerable number of medical drugs have been developed from fungal products including anti-infective, anti-cancer, anti-cholesterolemic, anti-parasitic and immunosuppressive drugs (Saravanamuthu, 2010).

Fungi were once considered to be primitive members of the plant kingdom because they grow out of the soil and have rigid cell walls (Webster and Weber, 2007). More recently, taxonomic classification projects revealed that fungi should be placed independently in an own kingdom of equal rank as animals and plants (Webster and Weber, 2007). The kingdoms of fungi and animals both belong to the group Opisthokonta, while plants, in contrast, belong to group Plantae (Figure 1) (Simpson and Roger, 2004). Thus, although fungi may not be our next of kin, they are more closely related to animals than they are to plants. The taxonomy of fungal species is subject to constant flux, especially due to the limited abilities to separate species based on morphological characteristics and the advent of research based on DNA comparisons. Thus, modern-day phylogenetic analyses often overturn historical classifications that were solely based on morphological methods. Currently, the kingdom of Fungi consists of eight phyla, Glomeromycota, Zygomycota, Basidiomycota, Ascomycota, Chytridiomycota, Neocallimatiomycota, Blastocladiomycota, and Microsporidia (Choi and Kim, 2017).



**FIGURE 1. Diagrammatic tree depicting the organisation of eukaryotes into six major phyla.** The tree is modified from Simpson and Roger(2004).

## Fungi and diverse life styles

Fungal species can be found in a multitude of environments. Particular species even adapted to the most extreme and inhospitable conditions, such as high salt content, low oxygen levels and extreme temperatures (Dix and Webster, 1995). For example, several fungal species (including *Penicillium crustosum*, *Penicillium svalbardense*, and *Aureobasidium pullulans*) have been isolated from Arctic glaciers and from sea water (Gunde-Cimerman et al., 2003; Sonjak et al., 2007; Zalar et al., 2008). Recently, *Exophiala dermatitidis* was found in the extreme environment of a dishwasher, which has a high temperature, high salinity and varied pH (Zupančič et al., 2016). Collectively, fungi are one of the most versatile and ecologically successful groups of organisms. However, most fungal species thrive as saprophytes that decompose organic matter. Fewer fungal species evolved, presumably from saprophytic progenitors, to establish symbiotic relationships with other living organisms, in a continuum ranging from pathogenic through endophytic to mutualistic (Little et al., 2012). Collectively, these symbionts are engaged in interactions with a broad range of hosts, including plants, vertebrate and invertebrate animals and even other microbes. Some of the plant pathogenic fungal species are causal agents of devastating plant diseases that have huge impact on agriculture. It is estimated



that approximately 10% of the worldwide agricultural production is lost due to fungal infections annually (Savary et al., 2012). Various plant pathogenic fungi can infect plant hosts at diverse developmental stages, including seeds, seedlings and adult plants, and different parts of the plant, such as roots, leaves, shoots, stems, fruits and flowers. Moreover, the lifestyles of plant pathogenic fungi are highly diverse, which leads to distinct strategies to interact with host plants. For instance, necrotrophic fungi infect and kill host tissue to extract nutrients from dead host cells while, in contrast, biotrophic fungi can only retrieve nutrients and complete their life cycle on living host tissues (Lo Presti et al., 2015). In between these extremes, there is a wide array of hemi-biotrophic fungi that display a switch from an initial biotrophic phase to a subsequent necrotrophic stage at a particular stage in their infection process (Lo Presti et al., 2015).

Irrespective of their life style, plant symbiotic fungi are typically recognized by the plant immune system and elicit host defenses (Dodds and Rathjen, 2010; Thomma et al., 2011; Cook et al., 2015). In order to establish their infection, adapted pathogens secrete molecules, termed “effectors”, that enable them to positively influence the outcome of the interaction (Jones and Dangl, 2006; de Jonge et al., 2011; Rovenich et al., 2014; Cook et al., 2015). Typical effectors are known as small, secreted, cysteine-rich proteins that are produced during host invasion (Stergiopoulos and de Wit, 2009). These effectors also comprise carbohydrate-active enzymes (CAZyme) that include a plethora of enzymes that are involved in degrading host cell walls. Therefore, plant pathogens often encode larger CAZyme repertoires than saprophytes that thrive on decaying organic matter (Zhao et al., 2013). Additionally, it is increasingly recognized that also other type of molecules should be seen as effectors, such as secondary metabolites and small RNAs, since they can serve a primary function to aid in the establishment of host interactions (Cook et al., 2015). Secondary metabolites (SMs) are small bioactive molecules that often play crucial roles in the establishment of specific ecological niches, but are not essential for normal growth (Fox and Howlett, 2008; Ponts, 2015; Derntl et al., 2017). Many pathogenic fungi employ SMs to promote virulence (Ponts, 2015; Pusztahelyi et al., 2015). These SMs are generally classified as either host-specific toxins (HSTs) that have specific targets in the host or non-HSTs that are generally toxic to a wide-range of organisms (Wolpert et al., 2002). Other SMs that do not directly damage plants through toxic activity could play important roles in protecting the fungus against environmental stresses, interfering with host hormone signaling and suppressing plant defense (Collemare and Lebrun, 2011; Stergiopoulos et al., 2013).

A comprehensive characterization of effector repertoires and determination of the mode of action of individual effectors is key to decipher pathogenicity mechanisms of plant pathogens to potentially design new, effective and durable disease management strategies with minimal environmental impact (Gibriel et al., 2016).

## Genome sequencing and comparative genomics

The first whole genome sequence of the yeast *Saccharomyces cerevisiae* is a landmark in fungal genomics (Goffeau et al., 1996). This genome sequencing project was a worldwide effort of over 600 researchers and took years to complete. By 2017, the genomes of more than 900 fungal species have been published, owing to the technology breakthrough of next-generation sequencing (NGS) that has reduced the effort, duration and costs of genome projects (Aylward et al., 2017). NGS encompasses second- and third- generation sequencing. Second-generation sequencing technologies can produce millions of reads in a single run in only a few days (Kchouk et al., 2017). However, they are not suitable to accurately assemble whole genome sequences, as they typically produce small DNA fragments, up to a few hundred base pairs (bp), that are not able to span repetitive regions (Kchouk et al., 2017). Third-generation sequencing uses single-molecule real-time (SMRT) sequencing technologies that produce long reads. These Reads can exceed 30 kb and may therefore span repeats (Braslavsky et al., 2003). Pacific Biosciences (PacBio) developed the first third-generation genomic sequencer. However, in addition to the relatively high sequencing cost, the major drawback of PacBio sequencing is its high error rate of ~13% (Kulski, 2016) that is dominated by insertion and deletion errors (Koren et al., 2012). Consequently, a considerable sequence depth is required to generate an error-free genome assembly based on SMRT sequencing only. Many researchers developed correction algorithms and assembly strategies that use short, high-fidelity sequences to correct the errors in long reads (Koren et al., 2012; Faino and Thomma, 2014; Seidl et al., 2015). Subsequent advances of bioinformatics tools and drops in the cost of PacBio sequencing have made assemblies based on PacBio sequencing affordable to many research laboratories. High-quality and even gapless genome assemblies of several fungal species were obtained (Faino et al., 2015; Dallery et al., 2017; Depotter et al., 2017; van Kan et al., 2017; Depotter et al., 2018). More recently, a pocket-sized, portable sequencing device, was released as the MinION sequencer that utilizes nanopore sequencing technology (Ip et al., 2015). The MinION can provide long reads that improve the contiguity of a *de novo* assembly. However, like PacBio sequencing it also produces a high error rate of ~12% (Ip et al., 2015), but the sequencing cost is much more affordable.

The availability of a high quality genome assembly of a fungal species facilitates in-depth genome analysis (Thomma et al., 2016). Once the genomes of multiple organisms have been sequenced and assembled, comparative genomics can be exploited to find shared and unique genes, traits and other genomic features that might explain their commonality or differences in morphology, life style or niche adaption. Comparative genomics of plant pathogens highlighted that many pathogens contain a so-called “two-speed” genome where effector genes reside in genomic compartments that are considerable more plastic than the core genome, facilitating the swift evolution of effector catalogs (Raffaele and Kamoun, 2012; Dong et al., 2015). Furthermore, comparative genomics approaches have

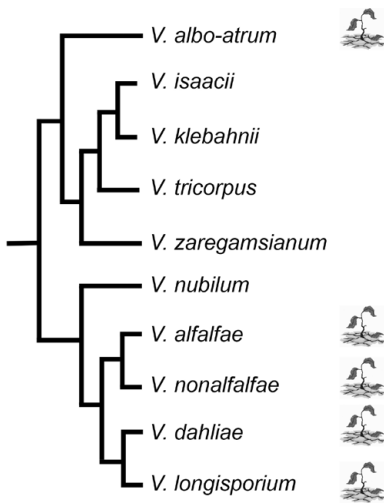
also been used for reconstructing the evolutionary history of species in order to identify the genomic divergence that may be associated to adaptation (Gordon et al., 2009; Shi-Kunne et al., 2018). For example, comparative genomics was applied to study genetic diversity between and within powdery mildew *formae speciales* to explain adaptation to new hosts (Wicker et al., 2013).

## The *Verticillium* genus

Verticillium wilt is a devastating vascular disease that is caused by several *Verticillium* spp. (Fradin and Thomma, 2006). Pathogenic *Verticillium* spp. infect their host plants through the roots, after which they traverse the root cortex and enter xylem vessels in which they proliferate and sporulate. Spores of the fungus are carried upwards to the above-ground plant tissues with the water flow and are trapped in pits or at vessel end walls, causing blockage of the water and mineral flow that can result in wilting of the plant (Agrios, 2005). In the absence of a suitable plant host, *Verticillium* spp. can remain dormant in the soil as microsclerotia for years. Microsclerotia are small, globular, melanized resting structures that are extremely durable and will only germinate in the proximity of a suitable host.

The *Verticillium* genus has a long and complicated taxonomic history. The first *Verticillium* species was found in 1816 (von Esenbeck, 1817) and approximately 190 species have since been described with similar morphological features (Rasoul et al., 2004). Molecular approaches subsequently revealed that the *Verticillium* genus comprised various distantly related species, which were subsequently removed from the genus (Rehner and Samuels, 1995). The reduced *Verticillium* genus then consisted of only five species, *V. dahliae*, *V. albo-atrum*, *V. tricorpus*, *V. nubilum* and *V. longisporum* (Karapapa et al., 1997; Zare et al., 2007). A more recent morphological and multi-genetic dataset was used to revise the *Verticillium* genus such that it presently contains ten species (Inderbitzin et al., 2011a). This taxonomic revision includes the designation of five new species, *V. alfalfae*, *V. nonalfalfae*, *V. isaacii*, *V. klebahnii* and *V. zaregamsianum* (Figure 2). *V. longisporum* is a near diploid hybrid, whereas all other species are haploids (Inderbitzin et al., 2011b; Depotter et al., 2017). Among these, *Verticillium dahliae* is the most notorious plant pathogen since it can cause vascular wilt disease on hundreds of plant species, resulting in large economic losses every year (Fradin and Thomma, 2006; Klimes et al., 2015). Furthermore, also *V. longisporum*, *V. albo-atrum*, *V. alfalfae* and *V. nonalfalfae* are plant pathogens, albeit with narrower host ranges than *V. dahliae* (Inderbitzin et al., 2011a). Although the remaining species *V. tricorpus*, *V. zaregamsianum*, *V. nubilum*, *V. isaacii* and *V. klebahnii* have been incidentally reported as plant pathogens, they are mostly considered as saprophytes (Ebihara et al., 2003; Inderbitzin et al., 2011a; Gurung et al., 2015).

Genomic studies of *Verticillium* species have mainly been focused on the pathogenic species, especially on *V. dahliae*. Various strains of *V. dahliae* have been sequenced (de Jonge et al., 2012; Faino et al., 2015; Kombrink et al., 2017; Depotter et al., 2018), and for two strains a gapless genome assembly has been generated (Faino et al., 2015). Comparative genomics revealed the occurrence of extensive large-scale genomic rearrangements between these strains (de Jonge et al., 2013), likely mediated by erroneous double-stranded break repair (Faino et al., 2015). These rearrangements gave rise to lineage-specific genomic regions that are enriched for *in planta*-expressed effector genes that are important for host colonization (de Jonge et al., 2013; Faino et al., 2016).



**FIGURE 2. Phylogenetic tree of the *Verticillium* genus.** Pathogenic *Verticillium* species are indicated with symbols of wilted plants.

## Main objective and research questions

The main objective of this research was to investigate the whole *Verticillium* genus, including non-pathogenic species. By using comparative genomics I aimed to address the following general questions:

1. How did the *Verticillium* genus evolve?
2. Can genomic traits that are commonly associated with pathogenicity of filamentous microbes also be implicated in *Verticillium* pathogenicity?

## Thesis outline

In **Chapter 2**, a strain of the saprophytic species *V. tricorpus* was sequenced and assembled using a hybrid approach that combines second and third generation sequencing. With comparative genomics, we aimed to identify genomic features that confer *V. dahliae* the ability to cause vascular wilt disease. Unexpectedly, both species were found to encode similar effector repertoires and share a genomic structure with genes encoding secreted proteins clustered in genomic islands. In conclusion, we highlight the technical advances of a hybrid sequencing and assembly approach and reveal that the saprophyte *V. tricorpus* shares many hallmark features with the pathogen *V. dahliae*.

In **Chapter 3**, we sequenced the genomes of all haploid *Verticillium* spp. with the aim to identify genomic features that can be associated with pathogenicity. We demonstrate that all species display similar genomic features, including the occurrence of extensive genomic rearrangements and the presence of extensive effector catalogs. Overall, this study reveals that no particular genomic feature can be associated with pathogenicity in the genus *Verticillium*.

**Chapter 4** is a genus-wide genomic study that was performed in order to infer genome organizations and gene contents of ancestral *Verticillium* species. Interestingly, inferring the *Verticillium* ancestral genome revealed that frequent rearrangements and gene losses occurred during evolution of this fungal genus. The findings of this study provide a basis for investigating the mechanism underlying species divergence as well as niche adaptation in the *Verticillium* genus.

**Chapter 5** describes the evolution of lineage specific regions in haploid *Verticillium* spp. by investigating sequence divergence within core and lineage-specific regions in *V. dahliae* and comparisons to the other haploid species. These lineage-specific regions are characterized by their dynamic evolution with respect to presence-absence polymorphisms, yet their conserved evolution with respect to sequence divergence.

**Chapter 6** describes a systematic search for evidence of inter-kingdom HGT events in the genome of *V. dahliae*. We provide evidence for extensive ancient horizontal gene acquisition from bacterial donors.

In **Chapter 7**, we performed *in silico* prediction and in-depth analyses of SM clusters (SMC) in the *V. dahliae* genome. We identified 25 predicted SMCs, among which are loci that may be implicated in DHN-melanin ferricrocin, triacetyl fusarinine and fujikurin production.

Finally, **Chapter 8** discusses the most important findings of this research in the broader context of genome evolution of filamentous plants pathogens, and highlights the importance of determining high-quality genome assemblies.



# Chapter **2**

## **The genome of the saprophytic fungus *Verticillium tricorpus* reveals a complex effector repertoire resembling that of its pathogenic relatives**

Michael F Seidl\*, Luigi Faino\*, Xiaoqian Shi-Kunne,  
Grady CM van den Berg, Melvin D Bolton and Bart PHJ Thomma

\*These authors contributed equally

This chapter has been published as:

Seidl MF\*, Faino L\*, Shi-Kunne X, van den Berg GCM, Bolton MD, Thomma BPHJ (2015)  
The genome of the saprophytic fungus *Verticillium tricorpus* reveals a complex effector  
repertoire resembling that of its pathogenic relatives. Mol Plant-Microbe Interact 28: 362–  
373 (\*equal contribution)

## Abstract

Vascular wilts caused by *Verticillium* spp. are destructive plant diseases, affecting hundreds of hosts. Only few *Verticillium* spp. are causal agents of vascular wilt diseases, of which *V. dahliae* is the most notorious pathogen, and several *V. dahliae* genomes are available. In contrast, *V. tricorpus* is mainly known as saprophyte and causal agent of opportunistic infections. Based on a hybrid approach that combines second and third generation sequencing, a near-gapless *V. tricorpus* genome assembly was obtained. With comparative genomics, we aimed to identify genomic features in *V. dahliae* that confer the ability to cause vascular wilt disease. Unexpectedly, both species encode similar effector repertoires and share a genomic structure with genes encoding secreted proteins clustered in genomic islands. Intriguingly, *V. tricorpus* contains significantly less repetitive elements and an extended spectrum of secreted carbohydrate-active enzymes when compared with *V. dahliae*. In conclusion, we highlight the technical advances of a hybrid sequencing and assembly approach and reveal that the saprophyte *V. tricorpus* shares many hallmark features with the pathogen *V. dahliae*.



## Introduction

Vascular wilt diseases caused by members of the fungal genus *Verticillium* collectively affect hundreds of plant hosts (Fradin and Thomma, 2006; Klosterman et al., 2009). The *Verticillium* genus was recently revised to encompass only ten species of soil-borne fungi that differ in morphological features, such as their resting structures, as well as in their ability to cause plant diseases (Inderbitzin et al., 2011a). Within this genus, *V. dahliae* is the most notorious plant pathogen due to the ability to cause vascular wilt on hundreds of dicotyledonous plant species, including ecologically important plant hosts and many high-value crops (Fradin and Thomma, 2006; Klosterman et al., 2009). Furthermore, *V. albo-atrum*, *V. alfalfae*, *V. nonalfalfae* and *V. longisporum* are also plant pathogens, albeit with narrower host ranges and more restricted distribution than *V. dahliae* (Inderbitzin and Subbarao, 2014). However, not all *Verticillium* species are genuine plant pathogens (Inderbitzin and Subbarao, 2014). For instance, although occasionally reported as an opportunistic pathogen, *V. tricorpus* is mainly known as a soil-borne saprophyte that thrives on decaying organic matter (Isaac, 1967; Qin et al., 2008; Klosterman et al., 2009; Powell et al., 2013). Comparative genomics between pathogenic members of the *Verticillium* genus, in particular strains of *V. dahliae*, recently facilitated the identification of key factors that mediate successful establishment of host infection and mechanisms that foster adaptation in the evolutionary arms race with plant hosts (Klosterman et al., 2011; de Jonge et al., 2012; de Jonge et al., 2013; Seidl and Thomma, 2014). To further identify genomic differences within the *Verticillium* genus and identify genetic components that enable particular *Verticillium* species to successfully infect plant hosts, determination of the genome sequence of non-pathogenic members of the *Verticillium* genus, such as *V. tricorpus*, is instrumental.

In the past, *de novo* sequencing of complex eukaryotic genomes was expensive and time consuming. However, recent advances in next generation sequencing (NGS) technologies, which can now be divided in second and third generation, facilitate affordable, rapid and high-quality genome sequencing (Metzker, 2010; Faino and Thomma, 2014). Second generation sequencing generates billions of short strings of nucleotides (called reads) of high quality and can be used for genome re-sequencing, where genome assemblies are based on mapping of reads to a reference genome, as well as for *de novo* genome sequencing. However, *de novo* assemblies that are based on short reads are obstructed by repetitive elements, GC- or AT-rich stretches, and homo-nucleotide stretches. Consequently, genome assemblies based only on second generation sequencing generally contain many gaps, leading to fragmented draft genomes (Faino and Thomma, 2014). Third generation technology, such as single molecule real-time (SMRT) sequencing, generates long continuous sequences (up to 35 kb) that overcome assembly problems by increasing the probability of having unique overlaps between reads (Schadt et al. 2010;

Faino and Thomma 2014). However, SMRT sequencing has a high per-base error rate that can only be corrected by large sequence depths, which results in considerable sequencing costs. To overcome the specific limitations of second and third generation sequencing platforms, hybrid approaches have been developed (Gnerre et al., 2011; Koren et al., 2012; Faino and Thomma, 2014). Such approaches combine reads generated by second and third generation sequencing platforms to maximize continuity and minimize the per base error rate. So far, however, hybrid approaches have only been used for relatively small microbial (bacterial) genomes (Koren et al., 2012).

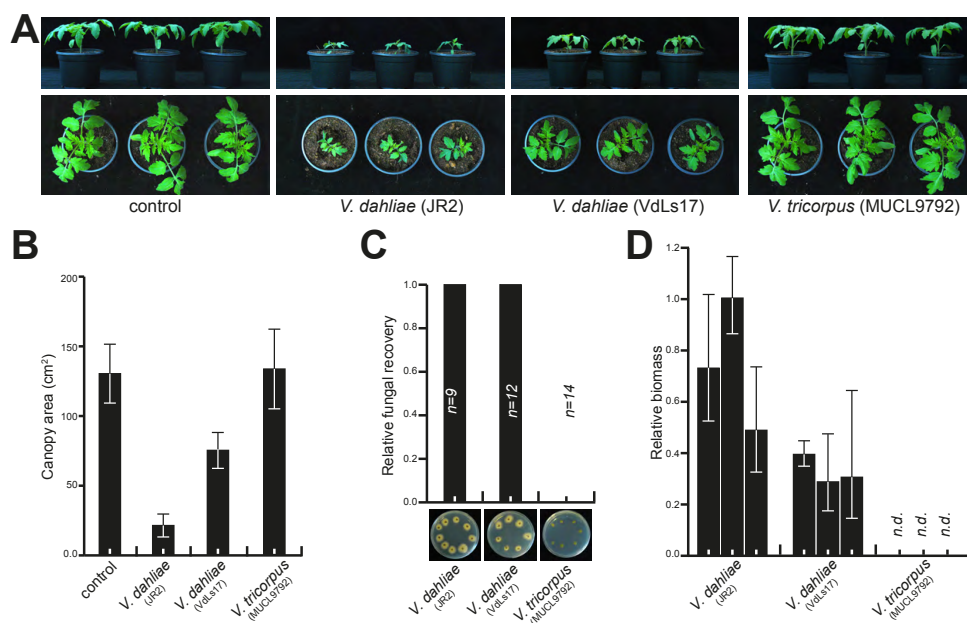
Here, we report on the genome of the saprophyte *V. tricorpus*. We applied a recently described hybrid assembly pipeline (Faino and Thomma, 2014) to establish a draft genome of *V. tricorpus*. This 36 Mb genome assembly encompasses 58 nearly gap-less scaffolds, of which four represent complete chromosomes. Comparative genomics between *V. tricorpus* and *V. dahliae* was performed in an attempt to identify genomic features that foster pathogenicity within the *Verticillium* genus.

## Results & Discussion

### ***V. tricorpus* strain MUCL9792 does not cause wilt disease on tomato**

In order to confirm the inability of *V. tricorpus* to cause Verticillium wilt disease on tomato (*Solanum lycopersicum*), inoculations were carried out with *V. tricorpus* strain MUCL9792 on MoneyMaker tomato plants that lack any characterized source of Verticillium wilt resistance (Fradin et al., 2009). As controls, inoculations were performed with two *V. dahliae* strains, VdLs17 and JR2, that have previously been characterized as relatively weak and strong tomato pathogens, respectively (Fradin et al., 2009; Klosterman et al., 2009; de Jonge et al., 2013). As expected, inoculation of tomato plants with *V. dahliae* strain JR2 resulted in clear stunting of the inoculated plants as a typical symptom of Verticillium wilt disease, whereas inoculation with strain VdLs17 resulted in considerably less disease symptoms as only mild stunting was observed (Figure 1A). In contrast, inoculation of tomato plants with *V. tricorpus* did not lead to visible stunting of tomato plants (Figure 1A). This was quantitatively corroborated as plants inoculated with *V. dahliae* strains JR2 or VdLs17 show considerably reduced foliage area when compared with mock-inoculated plants, while plants inoculated with *V. tricorpus* do not show any reduction (Figure 1B). Moreover, fungal recovery assays using stem sections of the inoculated plants showed that all sections of *V. dahliae*-inoculated plants were colonized by fungus, while no *Verticillium* was recovered from stem sections of *V. tricorpus*-inoculated plants (Figure 1C). Finally, real-time PCR quantification of fungal biomass revealed that no *V. tricorpus* biomass could be recorded in inoculated tomato plants, while ample *V. dahliae* biomass could be detected in VdLs17- and JR2-inoculated plants (Figure 1D). These data confirm

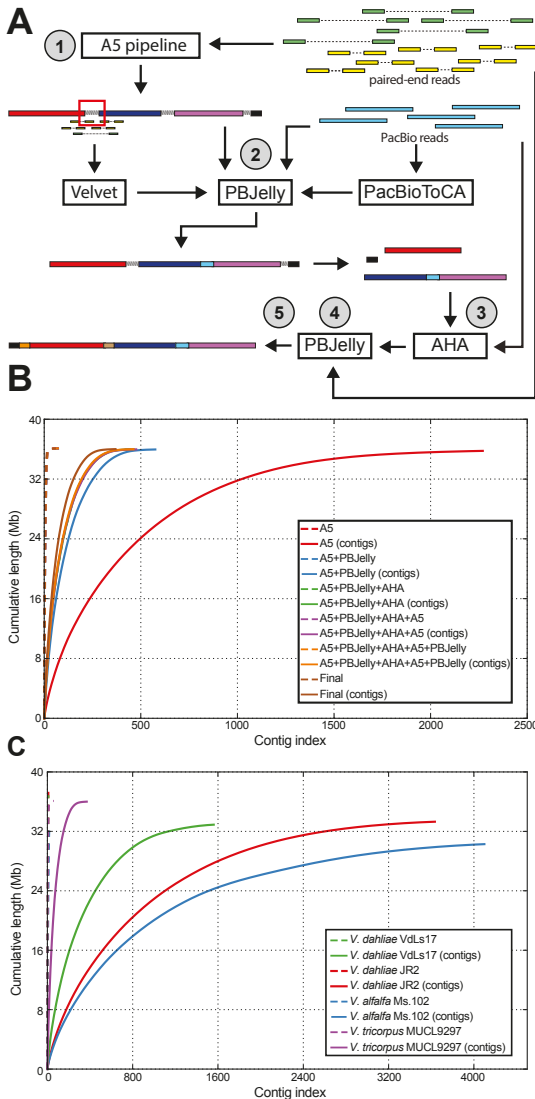
previous observations that *V. tricorpus* is not a genuine tomato pathogen (Isaac, 1967). *V. tricorpus* has only incidentally been reported as a plant pathogen on other hosts (Qin et al., 2008; Powell et al., 2013). For instance, inoculation of lettuce with *V. tricorpus* resulted in mild disease symptoms on only 2 out of 400 assessed lettuce plants (Qin et al. 2008). Thus, *V. tricorpus* should be considered an opportunistic pathogen that causes occasional infections on weakened plant hosts. Indeed, *V. tricorpus* is able to induce wilt symptoms when plants have been challenged with additional abiotic stresses, such as high soil nitrogen levels or waterlogging conditions (Isaac, 1956).



**FIGURE 1. *V. tricorpus* does not cause wilt disease on tomato.** Pathogenicity of two *V. dahliae* strains (highly pathogenic JR2, mildly pathogenic VdLs17) and of *V. tricorpus* was assessed on tomato. (A) Side and top view of tomato plants at 13 days after mock inoculation (control) or inoculation with *V. dahliae* strain JR2, *V. dahliae* strain VdLs17 and *V. tricorpus* MUCL9792, respectively. (B) Average canopy area of six tomato plants at 13 days after mock inoculation or inoculation with *V. dahliae* strain JR2, *V. dahliae* strain VdLs17 and *V. tricorpus* MUCL9792, respectively. An independently performed inoculation resulted in highly similar results (squared Pearson correlation coefficient = 0.99) (C) Fungal outgrowth at 10 days after plating tomato stem sections harvested at 15 days post-inoculation with *V. dahliae* strain JR2, *V. dahliae* strain VdLs17 and *V. tricorpus* MUCL9792, respectively. Quantification shows the percentage of stem section slices that showed fungal outgrowth. (D) Real-time PCR quantification of fungal biomass based on tomato plants harvested at 15 days post inoculation with *V. dahliae* strain JR2, *V. dahliae* strain VdLs17 and *V. tricorpus* MUCL9792, respectively.

## A high-quality *V. tricorpus* genome assembly

The genome sequence of *V. tricorpus* strain MUCL9792 was determined using sequencing-by-synthesis on the Illumina HiSeq2000 and Single Molecule Real-Time (SMRT) sequencing on the PacBio platform. To this end, 1.2 Gb of paired-end (PE) library-derived reads (500 bp insert size; 100 bp read length) was produced in combination with 1 Gb of mate-pair (MP) library-derived reads (5 kb insert size; 50 bp read length), together with ~200 Mb of long reads produced with PacBio sequencing (10 kb insert size) (Table S1). In order to assemble the Illumina-derived short reads and the PacBio-derived long reads, we applied a recently described hybrid assembly approach (Figure 2A; Faino and Thomma, 2014) that increases contiguity of the assembly in several steps (Figure 2B). In the first step, the PE and MP sequences were assembled using the A5 pipeline (Tritt et al., 2012) (Figure 2A). This initial assembly resulted in 195 scaffolds with an N50 of ~4 Mb, with the longest contig of 240 kb and the longest scaffold of 5.3 Mb (Table 1). This assembly has ~2,500 gaps and ~600 Ns (unknown nucleotides) per 100 kb. In the second step, gaps were filled using three types of sequences: (i) corrected long reads, (ii) uncorrected long reads, and (iii) sequences generated upon local re-assembly of short reads at gap positions (Figure 2A). This step improved the assembly statistics considerably, reducing the number of contigs from ~2,700 to ~750 with the longest contig of 535 kb, and reducing the number of gaps from ~2,500 to ~550. At this stage, gaps that are not closed may actually represent wrongly scaffolded contigs. Therefore, uncorrected long reads were used to scaffold contigs in the third step. Subsequently, the obtained scaffolds were subjected to another scaffolding step using MP reads in the fourth step, resulting in 604 contigs and 166 scaffolds (Table 1). More importantly, the total number of gaps was reduced to 438, while the number of Ns per 100 kb dropped to ~240. Manual curation included the removal of about 100 extremely short contigs (<500 bp) and further scaffolding. The A5 software will not merge contigs into scaffolds in case of insufficient support by paired Illumina reads. However, in about 10 cases where MP reads could unambiguously be mapped at the edges of two contigs, these were joined manually. Subsequently, we globally evaluated the scaffolding by remapping the MP reads to the scaffolds. As this did not lead to the identification of any discordantly mapped MP reads that would suggest erroneous scaffolding, this assembly was considered final (Figure S1). This manual curation yielded a final assembly of 399 contigs that composed 61 scaffolds (58 genomic and 3 mitochondrial), ranging in size between 9 Mb and 500 bp (scaffold N50 of 4.7 Mb; contig N50 of 250 kb) with 338 gaps and 229 Ns per 100 kb on average (Table 1). The ten largest scaffolds encompass 95% of the genome (Figure 2C, Figure 3A).



**FIGURE 2. A hybrid sequencing and assembly approach yields a nearly gap-less draft genome of *V. tricorpus*.** (A) A five-step hybrid assembly approach applied to assemble the 36 Mb draft genome of *V. tricorpus*. (B) Sequence length accumulation curve showing the successive reduction in the number of contigs and scaffolds during the different steps of the hybrid assembly. The curve displays the number of bases in the assembly as a function of the (sorted) number of sequences (contigs or scaffolds) (C) Sequence length accumulation curve showing the comparison between the draft genome assembly of *V. tricorpus* and the previously assembled genomes of *V. dahliae* strains JR2 and VdLs17 as well as *V. alfalfae*.

To determine its quality, the genome assembly was compared to that of previously sequenced *Verticillium* genomes. The genome of *V. dahliae* strain VdLs17 as well as that of *V. alfalfae* (formerly known as *V. albo-atrum*) strain VaMs102 were sequenced using whole-genome shotgun Sanger sequencing (7.5X and 4X coverage, respectively; (Klosterman et al., 2011)), while *V. dahliae* strain JR2 was sequenced with the Illumina HiSeq2000 (30X coverage; (de Jonge et al., 2012)). All these assemblies yielded a higher number of contigs (1,562 for *V. dahliae* strain VdLs17, 4,098 for *V. alfalfae* and 4,515 for *V. dahliae* strain JR2) when compared with the *V. tricorpus* assembly (Figure 2C), leading to a considerable higher amount of Ns per 100 kb. Consequently, the scaffolded genome assembly of *V. alfalfae* and

*V. dahliae* strain VdLs17 contained 2.5 Mb and 0.9 Mb of Ns, respectively. Although optical mapping further refined the *V. dahliae* assemblies by combining individually assembled scaffolds into eight chromosomes (Klosterman et al., 2011; de Jonge et al., 2013), this process introduced a considerable amount of additional gaps.

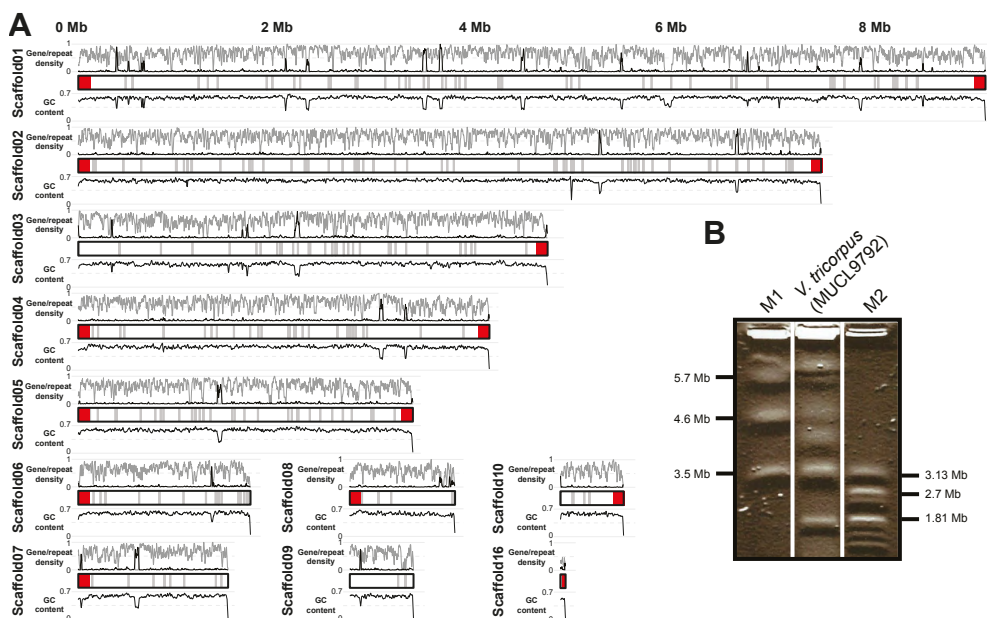
**TABLE 1. Genome assembly statistics for the A5 assembly, the hybrid assembly and the final assembly of the draft genome of *V. tricornutus*.**

	A5	Hybrid assembly	Final assembly
<b>Final statistics</b>			
Read depth (fold)	62.1X	67.5X	67.5X
Estimate genome size (Mb)	36.16	36.10	36.09
Number of Ns	217,012	87,959	82,887
Number of Ns (per 100 kb)	600.20	242.65	229.6
<b>Contig statistics</b>			
Number	2,789	604	399
Total length (Mb)	35.94	36.04	36.01
Largest (kb)	241.5	677.2	677.2
N <sub>50</sub> (kb)	35.9	180.7	249.2
Number of genes (partial)	10,417 (1,032)	11,291 (167)	11,338 (120)
<b>Scaffold statistics</b>			
Number	195	166	61
Largest (kb)	5,284	6,960	9,084
N <sub>50</sub> (kb)	4,123	4,117	4,670
Number of Ns (per 100 kb)	600.71	243.65	229.64
Number of genes (partial)	11,107 (345)	11,398 (60)	11,406 (52)

To assess whether any of the 58 scaffolds of the *V. tricornutus* assembly represent full-length chromosomes, we queried the scaffolds for characteristic fungal telomeric sequence repeats (Fulnečková et al., 2013). We identified fourteen such repeats, all of which were located at the ends of scaffolds (Figure 3A). Moreover, eight telomeric repeats were found on the ends of four of the ten largest scaffolds that are thus likely to represent full-length chromosomes (Figure 3A). Pulse-field gel electrophoresis was performed to determine the karyotype of *V. tricornutus* strain MUCL9792. This analysis revealed the presence of seven chromosomes, suggesting that we sequenced and assembled all telomeres present in the genome (Figure 3B). Combining the information from the karyotyping and the assembly, we estimate the sizes of the chromosomes to be 9 Mb (Scaffold 1; not entered the gel), 7.5 Mb (Scaffold 2), 5.7 Mb (not completely assembled), 4.8 Mb (not completely assembled), 4.1 Mb (Scaffold 4), 3.3 Mb (Scaffold 5) and 1.8 Mb (not completely assembled) in size (Figure 3B), summing up to 36.2 Mb and thus corresponding to the calculated genome

size. Thus, whereas previously sequenced *V. dahliae* strains carry eight chromosomes, this *V. tricorpus* strain only carries seven.

To further assess the completeness of the assembled gene space we queried for orthologs of 248 core eukaryotic gene families using the CEGMA pipeline (Parra et al. 2007; 2009). The assembled *V. tricorpus* genome assembly was found to contain 237 members (95.56%) of these eukaryotic core families, comparable to previously assembled *V. dahliae* draft genomes and higher when compared with *V. alfalfae* (81%) (Klosterman et al., 2011; de Jonge et al., 2013), indicative of a highly complete gene space in the *V. tricorpus* genome assembly.



**FIGURE 3. Overview of the draft genome assembly of *V. tricorpus*.** (A) Schematic representation of the ten largest scaffolds in the genome assembly of *V. tricorpus*. Characteristic fungal telomeric repeats are displayed on the ends of the scaffolds (indicated in red; an additional single small scaffold with distal telomeric repeat is displayed as well) and assembly gaps are displayed on each scaffold (grey). Gene (grey) and repeat (black) density as well as GC content is displayed in a sliding window of 10 kb with a slide of 2 kb. (B) Chromosome karyotype by pulsed-field gel electrophoresis of the sequenced *V. tricorpus* strain MUCL9792. Chromosomal DNA of *Schizosaccharomyces pombe* (M1) and *Hansenula wingei* (M2) were loaded as size markers.

### *V. tricorpus* genome annotation

We classified different types of repetitive elements in the genome of *V. tricorpus* with Repeatmasker, utilizing a combination of known repetitive elements and *de novo* repeat identification (Smit et al., 1996; Jurka et al., 2005). In total, 2.5% (800 kb) of the assembled *V. tricorpus* genome accounts for repetitive DNA and included traces of long terminal



2

repeats (110 kb; 0.3%), long interspersed nuclear elements (35 kb; 0.1%) and short interspersed nuclear elements (1 kb; 0.002%) that are not uniformly distributed along the scaffolds (Figure 3A). Surprisingly, their quantity and relative contribution is considerably smaller compared to that of previously sequenced *Verticillium* genomes, where 4.2% of the assembled genome account for repetitive elements (4.05% for *V. dahliae* strain VdLs17 and 4.33% for *V. dahliae* strain JR2) (Klosterman et al., 2011; de Jonge et al., 2013). Repetitive elements are strong drivers of genome evolution since they promote large-scale genomic variations (Seidl and Thomma, 2014). In *V. dahliae*, repetitive elements have been shown to foster genomic rearrangements that drive the evolution of pathogenicity (de Jonge et al., 2013; Seidl and Thomma, 2014). It is thus tempting to speculate that the increase of repetitive elements in the genome of *V. dahliae* is associated with its pathogenic capacity.

We inferred a reference gene annotation by integrating *de novo*, homology-based and transcription data using the Maker2 pipeline (Holt and Yandell, 2011). To this end, we generated 1 Gb RNA-Seq data of *V. tricorpus* growing in a suspension of tomato cells (Table S1), to mimic a plant-associated environment, and about 97% of the reads could be aligned to the genome assembly using TopHat2 (Kim et al., 2013). To further improve gene structure annotation by Maker2, we assembled the RNA-Seq reads into transcripts using Oases (Schulz et al., 2012) and Trinity (Grabherr et al., 2011), which were subsequently aligned to the genome assembly using PASA (Haas et al., 2003). Additionally, we provided Maker2 (Holt and Yandell, 2011) with 35 predicted fungal proteomes that represent a broad phylogenetic diversity (Table S2) to guide gene structure annotation, and applied three independent *de novo* gene predictors; SNAP (Korf 2004), GeneMark-HMM (Lukashin and Borodovsky, 1998), and Augustus (Stanke et al., 2004). Maker2 combined these data and produced a consolidated set of gene models. Subsequently, we refined these gene models by manually performing RNA-Seq- and homology-guided gene structure annotation. If necessary, gene models were corrected leading to the identification of 418 previously undetected genes, mainly due to erroneous gene fusions, and the correction of ~4,000 gene models, thus yielding a total of 11,454 protein-coding genes. Of these, 89% have RNA-Seq support ( $\geq 1$  read mapped; on average 464 reads per gene model) and 91% have homology-based support in other fungal species. The 11,454 protein-coding genes were functionally annotated using InterProScan (Zdobnov and Apweiler, 2001) and Blast2GO (Conesa et al., 2005) yielding ~65% of the predicted genes with a predicted functional annotation (Figure S2, Table S3).

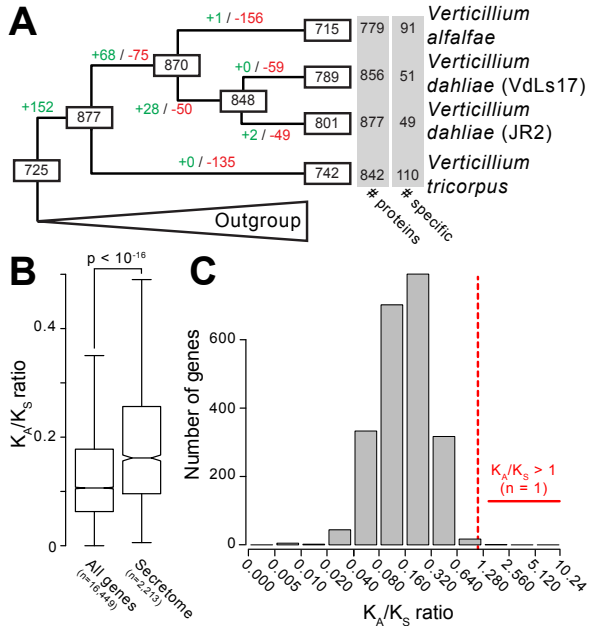
### The secretome of *V. tricorpus*

The vast majority of fungal species comprise saprophytes that thrive on decaying organic matter. These species deploy a plethora of secreted proteins to decompose complex organic material in their proximity to acquire nutrients and to establish their niche.



Independently emerging from saprophytic ancestors, particular fungal species evolved to establish symbioses with various hosts, including plants, requiring to successfully overcome host immunity. Irrespective whether the symbiosis involves parasitism, commensalism or mutualism, secreted fungal proteins play an important role as they comprise so-called effectors that are required for establishment of the interaction (de Jonge et al., 2011; Thomma et al., 2011; Rovenich et al., 2014). Therefore, pathogenic species often encode larger secretomes (i.e. the total set of secreted proteins) compared to their saprophytic relatives (Choi et al., 2010), potentially reflecting the acquisition of additional factors, e.g. effector genes, that facilitate their pathogenic lifestyle. To predict the secretome of *V. tricorpus*, we queried the predicted proteome for secretion signals, leading to the identification of 952 secreted proteins, which accounts for ~8.3% of its proteome. Using the same approach, the genomes of the pathogenic *V. dahliae* strains JR2 and VdLs17 encode 926 and 907 secreted proteins accounting for ~8.5% of their predicted proteomes, respectively. The genome of *V. alfalfae* encodes fewer (870) secreted proteins, which may be the consequence of the relatively low sequence coverage that hampered whole genome assembly and annotation (Klosterman et al., 2011), and thus should be taken with caution.

By tracing the gain and losses of gene families encoding secreted proteins in the *Verticillium* genus using Dollo parsimony, we observed comparable absolute numbers of these gene families in *V. dahliae*, *V. alfalfae* and *V. tricorpus* ranging from 715 to 801, corresponding to 779 to 877 proteins (Figure 4). Since pathogenic *Verticillium* species likely evolved from a non-pathogenic saprophytic ancestor, the 667 families (2,691 proteins) that are shared between *V. tricorpus*, *V. dahliae* and *V. alfalfae* could represent a 'core' saprophytic secretome. In *V. dahliae* strain JR2, only 56 of the 668 genes belonging to these families are *in planta* induced ( $>1.5 \log_{10}$  (fold-change)), and thus the majority has other functions that are likely not associated with its pathogenic lifestyle. Only few gene families (99), encompassing 254 genes encoding secreted proteins, have been specifically gained in the genomes of *V. dahliae* and *V. alfalfae*, of which the vast majority encode proteins of unknown function. Moreover, a considerable amount of secreted proteins do not display similarity to any other sequence and thus are considered lineage-specific. For example, *V. dahliae* strain JR2 contains 49 lineage-specific genes encoding secreted proteins, including the experimentally verified virulence effector *XLOC009059* and the *in planta*-induced candidate effector *XLOC008951* (de Jonge et al., 2013). Similarly, *V. tricorpus* encodes a considerable amount of lineage-specific genes, mainly encoding secreted proteins with unknown function. The vast majority (70%) of these genes are expressed during *in vitro* growth.



**FIGURE 4. The secretome of *V. tricorpus* encodes a secretome of similar size as pathogenic *Verticillium* spp.**

(A) Evolutionary history of the gene families constituting the predicted secretome of the analyzed *Verticillium* spp. according to DOLLOP. At each branch, the number of gene family gains (green) and losses (red) is displayed and the inferred ancestral gene family set is shown. (B) Distribution of non-synonymous substitutions per non-synonymous site values ( $K_A$ ) to synonymous substitutions per synonymous site ( $K_S$ ) ratios for *V. tricorpus* genes encoding secreted proteins with orthologs in any of the analyzed *Verticillium* secretomes, and for all other genes and with orthologs. The difference between the distributions was assessed using the non-parametric Wilcoxon rank sum test and significance is highlighted. Whiskers indicate the 1.5 interquartile range and outlier were not displayed for clarity. (C) Distribution of  $K_A/K_S$  ratios for *V. tricorpus* genes encoding secreted proteins with orthologs in the other analyzed *Verticillium* secretomes. Genes with  $K_A/K_S$  ratios  $> 1$ , indicative of positive selection, are highlighted.

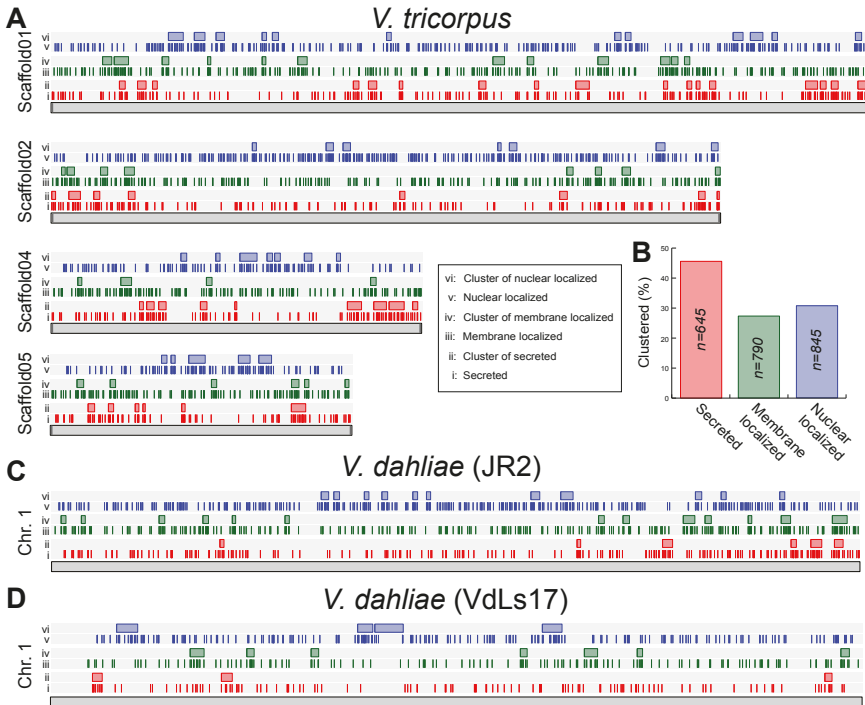
Many genes encoding secreted proteins, but in particular pathogen effectors, are expected to rapidly diversify due to selection pressure imposed by the host during the evolutionary arms-race between pathogens and their hosts (Raffaele et al. 2010; Guyon et al. 2014; de Jonge et al. 2013; Klosterman et al. 2011; Wicker et al. 2013). To assess the speed of gene evolution and positive selection within the secretome, we used the ratio of non-synonymous substitutions per non-synonymous site ( $K_A$ ) to synonymous substitutions per synonymous site ( $K_S$ ) for *V. tricorpus* genes with orthologs in the other analyzed *Verticillium* spp. as a proxy (Figure 4B). While the vast majority of *V. tricorpus* genes were found to be under purifying selection (i.e.  $K_A/K_S \ll 1$ ), genes encoding secreted proteins were found to display significantly higher  $K_A/K_S$  ratios when compared with pairs of genes derived from the full set of protein coding genes (Figure 4B, Figure 4C), thus suggesting accelerated evolution within the secretome. However, only a single highly

expressed gene, encoding a secreted protein with similarity to chitinases, displays signs of positive selection, i.e.  $K_A/K_S > 1$  (Figure 4C). In contrast, *V. dahliae* strain JR2 contains 17 genes encoding secreted proteins with signs of positive selection (Figure S3), of which the majority is of unknown function, but that also includes genes encoding secreted proteins with predicted functions in cellulose binding or cutinase activity. In many pathogenic microbes, an extended repertoire of genes encoding secreted proteins has been described (Raffaele et al. 2010; Wicker et al. 2013). However, the small number of genes under positive selection corroborates previous analysis that suggested that diversification at nucleotide levels does not play a major role in *V. dahliae* evolution (de Jonge et al. 2013). Notably, the majority of effector genes in *V. dahliae* with roles in virulence are lineage-specific, hampering the assessment of positive selection, and suggesting that presence/absence polymorphisms are crucial to drive the evolution of pathogenicity in *V. dahliae*.

We observed a distinct quantitative difference in repeat content between *V. dahliae* strain JR2 and *V. tricorpus*. To further assess the hypothesis that the expansion of repeats in *V. dahliae* strain JR2 is associated with its pathogenicity (de Jonge et al. 2013; Seidl and Thomma 2014), we analyzed whether genes, and in particular those encoding secreted proteins, are directly flanked by repetitive elements. In *V. dahliae* strain JR2, 25% of all genes, and 27% of those encoding secreted proteins, are directly flanked by at least a single repetitive element (excl. simple nucleotide repeats). In contrast, only 15% of all predicted genes, and 18% of those encoding secreted proteins, are flanked by repeats in *V. tricorpus*. Interestingly, when considering *in planta* expressed genes ( $>1.5 \log_{10}$  (fold-change)), 20 out of 79 genes that encode secreted proteins are directly flanked by repeats in *V. dahliae* strain JR2, including two previously functionally characterized JR2-specific effectors proteins (de Jonge et al. 2013). In contrast, only three highly expressed *V. tricorpus* genes ( $>2 \log_{10}$  (fold-change) when compared with the average genome-wide expression) that encode secreted proteins are flanked by repeats. Thus, our data suggests that the presence and expansion of repetitive elements is associated with pathogenicity in the *Verticillium* genus.

### Genes encoding secreted proteins are clustered in *Verticillium*

The ten largest scaffolds encode on average 9.5% secreted proteins. On some of the smaller scaffolds considerably more genes encode secreted proteins, e.g. up to 18% of the genes on scaffold 10, thus indicating that many of these genes localize in close physical proximity to each other. Moreover, these small scaffolds contained one of the characteristic telomeric repeats and thus many proteins encoding secreted protein cluster in proximity of the chromosomal ends. Therefore, we analyzed the global position of genes encoding secreted proteins on the complete chromosomes (scaffolds 1, 2, 4 and 5) (Figure 5A). Interestingly, genes encoding secreted proteins are not randomly distributed



**FIGURE 5. Genes encoding distinct functional classes cluster in the genome of *V. tricoloris*.** (A) Genomic position of genes encoding secreted proteins (i; red vertical bar), genes encoding transmembrane localized protein (iii; green vertical bar) and genes encoding nuclear localized proteins (v; blue vertical bars) are indicated along the four scaffolds that represent complete *V. tricoloris* chromosomes. Clusters of genes for each of the functional classes are shown (ii, iv and vi; window of 20 genes, min of five genes, consecutive windows were merged). (B) Fraction of genes in each of the functional classes that is clustered on the four complete chromosomes. Total number of genes per class (n) is indicated. Clustering of genes for a chromosome of *V. dahliae* strain JR2 (C) and VdLs17 (D) is shown for comparison.

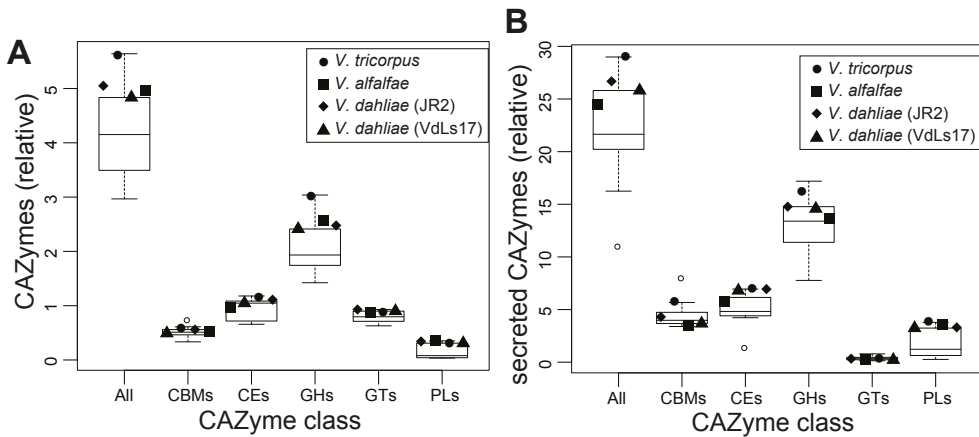
along the chromosomes of *V. tricoloris* but rather cluster locally, both in areas within the chromosomes and also in some cases close to telomeres (Figure 5A). Intriguingly, also genes encoding predicted membrane- or nuclear localized proteins cluster (Figure 5A). However, compared to genes encoding secreted proteins, where ~45% of the genes are clustered, only ~27% of the genes encoding proteins predicted to localize in membranes or in the nucleus are located within clusters (Figure 5B). Similarly, this local clustering, even though less pronounced, is also observed in both *V. dahliae* strains, e.g. on chromosome 1 for strain JR2 and chromosome 2 for VdLs17 (Figure 5C and D). Clustering of genes encoding secreted proteins has been reported before in different strains of *V. dahliae*, where *in planta* expressed effector genes are enriched in highly dynamic lineage-specific regions (de Jonge et al., 2013). Similarly, many filamentous pathogens contain a genome architecture that is often characterized by the local co-occurrence of effector genes

(Raffaele and Kamoun, 2012). For example, genes encoding secreted proteins in *Fusarium graminearum* are clustered, preferentially in regions close to the telomeres or in regions with increased recombination frequency (Brown et al., 2012). Clustering of genes, and in particular genes encoding secreted proteins, in or in the proximity of unstable genomic regions can foster the swift alterations of these genes, thus enable the rapidly adaption to changing environments (Seidl and Thomma, 2014).

### **The genome of *V. tricorpus* encodes an expanded repertoire of CAZymes**

Saprophytes and plant pathogens secrete a plethora of enzymes that are involved in the breakdown of complex polysaccharides that compose plant cell walls. Whereas saprophytes degrade these complex polysaccharides solely for nutritional purposes, pathogens also need to overcome these physical barriers to establish infection. Therefore, plant pathogens often encode larger repertoires of carbohydrate-active enzymes (CAZymes) when compared with saprophytes (Zhao et al., 2013). To explore the diversity of CAZymes in *Verticillium*, we queried the predicted proteomes of *V. tricorpus*, *V. alfalfae* and the two *V. dahliae* strains (JR2 and VdLs17) for CAZyme signature domains (Yin et al., 2012) that belong to the four catalytic classes of glycoside hydrolases (GHs), polysaccharide lyases (PLs), carbohydrate esterases (CEs), and glycosyltransferases (GTs) and carbohydrate-binding modules (CBMs), leading to the identification of 646, 508, 531 and 532 CAZymes (~5% of their proteomes), respectively. When compared to six filamentous pathogens (Figure S4) and three saprophytes, *Verticillium* spp., together with the vascular wilt pathogen *Fusarium oxysporum*, encode the highest absolute number of CAZymes in their genomes (Figure S5, Table S4) (Klosterman et al., 2011). Moreover, *Verticillium* spp. devote the largest proportion of their proteome to CAZymes (Figure 6A). Notably, the saprophytic dung-decomposing fungus *Podospira anserina*, the white rot fungus *Phanerochaete chrysosporium* and *Neurospora crassa* encode considerably lower CAZyme numbers (Table S4). Moreover, *Verticillium* spp. devote a large proportion (~25%) of their secretome to CAZymes (Figure 6B), which is higher compared to most other fungi with the exception of the saprophyte *P. chrysosporium* (Figure 6B). Contrary to previous observations that saprophytes encode fewer CAZymes when compared to pathogens (Zhao et al., 2013), *V. tricorpus* contains more CAZymes than the two pathogenic *V. dahliae* strains. The high number of CAZymes in *V. tricorpus* is mainly due to a higher abundance of GH families (in particular families GH3, GH43 and GH78) and secreted CBM1-containing proteins. CBM1 is generally considered as a fungal cellulose-binding domain that is often identified in proteins containing glycoside hydrolases (Klosterman et al., 2011). Saprophytic fungi such as *P. anserina* and *P. chrysosporium* also encode an expanded repertoire of CBM1 containing genes (Table S4). The expanded CAZyme repertoire in *V. tricorpus* therefore likely reflects its saprophytic lifestyle. In contrast, only six CAZyme families are expanded in the pathogenic *Verticillium* strains when compared with *V.*

tricornus (Table S4). Moreover, a single member belonging to CAZyme family GH54 is present in all pathogenic *Verticillium* strains while it is absent from *V. tricornus*. This gene is predicted to encode a secreted arabinofuranosidase with a C-terminal carbohydrate-binding module CBM42. Interestingly, the gene that encodes this member is highly induced during in planta infection in *V. dahliae* strain JR2 (up to 3 log<sub>10</sub> (fold-change) at 12 dpi), suggesting a possible role of this gene during pathogenicity. However, only two other genes that belong to the six expanded CAZyme families show induced in planta expression (>1.5 log<sub>10</sub> (fold-change)), suggesting that expansion or presence/absence of particular CAZymes is not associated with pathogenicity.



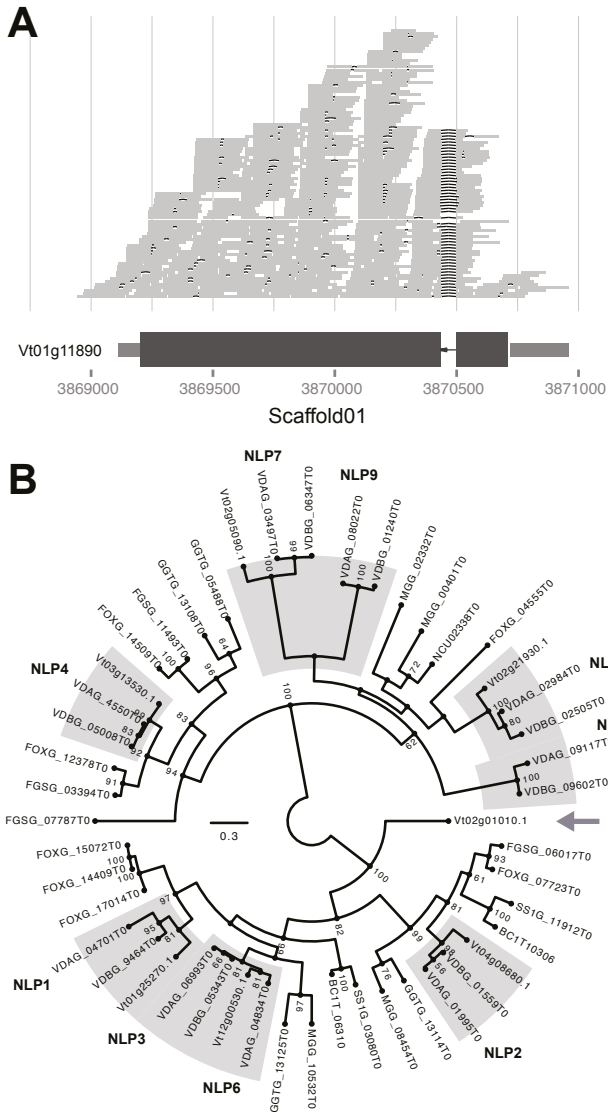
**FIGURE 6. The genomes of *Verticillium* spp. encode a large repertoire of carbohydrate active enzymes.** Boxplots display the relative number of carbohydrate active enzymes (CAZymes) in the (A) full proteomes, and (B) secretome of *Verticillium* spp. and related fungi (Figure S4). The relative contribution belonging to the five CAZyme classes - glycoside hydrolases (GHs), polysaccharide lyases (PLs), carbohydrate esterases (CEs), and glycosyltransferases (GTs), and carbohydrate-binding modules (CBMs) - are shown as boxplots, whiskers indicate the 1.5 interquartile range, and outliers are shown as open circles. The position of each of the *Verticillium* spp. is indicated.

### The *V. tricornus* effector repertoire largely resembles that of other *Verticillium* spp.

The surprising finding that the secretome of the saprophyte *V. tricornus* is even larger than the secretome of the *V. dahliae* strains urged us to further explore the presence and abundance of genes that have been implicated in fungal pathogenicity, and thus could reflect the adaptation of pathogenic *Verticillium* spp. to their particular lifestyle. We queried the genomes for genes encoding small secreted cysteine-rich proteins (SSCPs) that are often characterized as lineage-specific pathogen effector molecules (de Jonge et al., 2011; Thomma et al., 2011). When querying secreted proteins for size and cysteine content ( $\leq 150$  aa and contain  $\geq 3\%$  cysteine-residues; (Saunders et al., 2012)), we identified 46 SSCP in *V. tricornus*, of which >75% do not have any functional annotation. Similarly, 67

SSCPs were identified in *V. dahliae* strain JR2, 43 SSCP in VdLs17 and 36 SSCP in *V. alfalfae*, of which >85%, >70% and >70% do not have any functional annotation, respectively (Table S5). *V. tricorpus* and *V. dahliae* considerably overlap in the functionally annotated proteins present in their respective SSCP catalogue, suggesting that *V. tricorpus* shares effectors with *V. dahliae*. However, the *V. tricorpus* catalogue of SSCP contains proteins that cannot unambiguously be implicated in pathogenicity, such as hydrophobins. Even though particular hydrophobins play roles in pathogenicity (Talbot et al., 1993), they are also present in non-pathogenic fungi and act in the formation of resting structures in *Verticillium* (Klimes et al., 2008). Similarly, several proteins with a cerato-platanin domain were identified of which homologs occur in filamentous fungi with all types of lifestyles and that may act in in fungal growth and development (Gaderer et al., 2014).

We further mined the secretomes for the presence of LysM effectors, a class of secreted proteins that do not contain any other recognizable features except LysM domains (Bolton et al. 2008; de Jonge and Thomma 2009; Kombrink and Thomma 2013; de Jonge et al. 2010). Even though several LysM effectors have been implicated in fungal pathogenicity (de Jonge et al., 2010; Marshall et al., 2011; Mentlak et al., 2012; Sánchez-Vallet et al., 2013), they are commonly observed in saprophytes as well, and may thus play alternative roles in fungal biology besides promoting pathogenicity (Kombrink and Thomma, 2013; Rovenich et al., 2014). The genome of *V. tricorpus* encodes five LysM effectors, containing 3-5 LysM domains, of which one (Vt01g11890) is expressed when *V. tricorpus* is grown in a tomato cell suspension (Figure 7A). Initially, seven LysM effector candidates were identified in the genome of *V. dahliae* strain VdLs17 (Klosterman et al., 2011), of which only five qualify as *bona fide* LysM effectors (Kombrink, 2014). Notably, one of these five is a strain-specific LysM effector that is expressed during infection and acts as a virulence factor on tomato (de Jonge et al., 2013). So far, the role of the remaining four core LysM effectors during plant infection could not be demonstrated and their function remains enigmatic (Kombrink and Thomma, 2013). Rather than a role in pathogenicity, *Verticillium* LysM effectors could be involved in protection from competing microbes in the soil (de Jonge and Thomma, 2009; Kombrink and Thomma, 2013; Rovenich et al., 2014).



**FIGURE 7. The genome of *V. tricolorpus* encodes similar effector candidates as pathogenic *Verticillium* spp. (A) RNA-Seq expression of a gene encoding the LysM effector Vt01g11890. (B) Phylogenetic relationship of necrosis and ethylene inducing-like (NLPs) in related filamentous fungi (Figure S4). Orthologous groups of previously described NLP classes in *Verticillium* are highlighted in grey (Santhanam et al. 2013) and the single highly expressed NLP (Vt02g01010) in *V. tricolorpus* is indicated by an arrow.**

Similar to LysM effectors, also necrosis and ethylene inducing-like proteins (NLPs) are widespread and occur in organisms with diverse life styles (Gijzen and Nürnberger, 2006). Even though they have been observed in both pathogenic and saprophytic species, e.g. *Neurospora crassa* (Figure 7B), they have gained considerable attention in pathogenic



species (Gijzen and Nürnberger, 2006). Many fungi contain few (1-2) NLPs, while the NLP family is expanded in several pathogenic oomycetes and particular fungi (Gijzen and Nürnberger, 2006; Santhanam et al., 2013), including *V. dahliae* and *V. alfalfae* that, depending on the strain, contain either seven or eight members (Santhanam et al., 2013). Recent analyses revealed that only two of the *V. dahliae* NLPs, NLP1 and NLP2, induce cell death and are required for virulence on tomato and Arabidopsis (Santhanam et al., 2013), while the role of the remaining, non-cytolytic NLPs remains unclear. Interestingly, VdNLP1 also plays a role in vegetative growth and asexual production, hinting to alternative functions of NLPs besides their necrotic activity (Santhanam et al., 2013). An initial survey of *V. tricorpus* prior to the completion of the genome assembly identified only four NLP homologs (Santhanam et al., 2013). However, the presently assembled genome of *V. tricorpus* contains seven NLP genes, of which six are clear orthologs of previously described *V. dahliae* NLPs (Figure 7B). Interestingly, *V. tricorpus* encodes orthologs of NLP-1 and NLP-2. Both of these orthologs were expressed at low levels (7.8 and 15 read counts per 1 kb, respectively), in contrast to a *V. tricorpus*-specific NLP homolog (Vt02g01010) that was highly expressed in the tomato cell suspension (368.7 read counts per 1 kb). Thus, presently the function of *V. tricorpus* NLPs remains enigmatic.

In summary, we revealed an effector repertoire in *V. tricorpus* that is surprisingly similarly composed when compared to its pathogenic relatives. Members of some of the identified effector families are commonly observed in non-pathogenic species such as saprophytes, and thus do not necessarily represent genuine pathogenicity factors. Moreover, homologs of *V. dahliae* effector genes, such as *NLP2*, are not always identical in *V. tricorpus* and may thus have diverged to different functions. Furthermore, a considerable amount of SSCPs do not display significant similarity to any known (effector) family and several of these uncharacterized proteins might play important roles in niche establishment and competition with other microbes. For example, *V. tricorpus* has recently been shown to be an antagonist of *V. dahliae* on lettuce and artichoke (Qin et al., 2008), and SSCPs of unknown function might be involved in this interaction. Thus, the genome assembly of *V. tricorpus* presents the foundation for the identification and initial characterization of effector families. Realistically, however, the functions of most effector families and their members remain unknown and need experimental characterization.

## Conclusions

In the last decade, rapid technological advances made genomic sequencing a commodity that provides a plethora of genomic sequences. Many of these genomes are from closely related species that differ in their ecological niches. We anticipated that a high-quality genomic sequence of the saprophyte *V. tricorpus* would shed light on the origin of

2

pathogenicity within the genus *Verticillium*. However, contrary to our expectations, the genome of *V. tricorpus* contains many of the hallmarks that also characterize pathogenic *Verticillium* spp. In particular, *V. tricorpus* encodes an effector repertoire that largely resembles that of its pathogenic relatives. Assuming that pathogenicity in fungi evolved from saprophytes, alternative roles for effectors, e.g. during interactions with other microorganisms that are encountered in the same niche, seem likely (Rovenich et al., 2014). During the evolution of pathogenicity, these proteins might have acquired altered functionalities that facilitate the pathogenic lifestyle. However, a distinct difference in the repetitive element content between *V. tricorpus* and its pathogenic relatives may suggest that repeats are involved in the evolution of pathogenicity in the *Verticillium* genus. We anticipate that further comparative genomics together with functional characterization of effector proteins in species with different lifestyles will enhance our understanding of their role in facilitating a pathogenic lifestyle.

## Materials & Methods

### Plant inoculations

All plant inoculations were performed on ten-day-old tomato seedlings (cv. MoneyMaker) using root-dip inoculation as described previously (Fradin et al., 2009). Briefly, seedlings were uprooted and inoculated by dipping the roots for 2 minutes in a conidial suspension ( $10^6$  conidia/ml) in water. After replanting in soil, plants were incubated at standard greenhouse conditions of a 16-h/8-h light/dark regime and 60% RH. Disease progression was monitored until 15 days after inoculation. To determine *in planta* colonization, stem sections at the height of the first internode were taken, surface sterilized and sliced. The slices were placed on PDA supplemented with chloramphenicol at 50 µg/ml and incubated at 22°C for 10 days. For *in planta* biomass quantification, stem of three inoculated plants were harvested at 15 days after inoculation. The samples were ground to powder and genomic DNA was isolated. Real-time PCR on genomic DNA was carried and analyzed using the 7300 SDS software (Applied Biosystems). A *Verticillium*-specific primer pair was used to quantify fungal colonization (5'-CATTGCCCAAGTTTACCTCC-3' combined with 5'-GCCAGCGTGTCTATCTTCTC-3').

### DNA extraction

DNA was isolated from conidia and mycelium fragments that were harvested from 10-day-old cultures grown in liquid potato dextrose broth (PDB) at 28°C. DNA isolation was performed according to the following protocol. Firstly, mycelium was scraped directly from the surface of an agar plate, suspended in water and transferred to a 50 ml tube. The tube was vortexed briefly and centrifuged at maximum speed for 10 minutes.

After centrifugation, 0.5 ml of 0.1 M Tris HCl (pH 8) was added to the pellet, which was mixed and subsequently transferred to a 2 ml screw cap tube. Three metal beads were added and the tube was shaken in a bead beater at the highest speed for 45 seconds. After shaking, 1 ml phenol:chloroform:isoamylalcohol (24:24:1) was added. The tube was mixed and centrifuged at maximum speed for 7 minutes, and then the water phase (400  $\mu$ l) was transferred to a new tube. In the new tube, 800  $\mu$ l isopropanol was added and the tube was centrifuged after mixing at maximum speed for 20 minutes in the cold room. After centrifugation, the isopropanol was removed and the DNA pellet was air-dried. Subsequently, the pellet was dissolved in 50  $\mu$ l of deionized H<sub>2</sub>O.

### Genome sequencing and assembly

Two sequence libraries (500 bp and 5 kb insert size) of *V. tricorpus* were prepared and paired-ends were sequenced at the Beijing Genome Institute (BGI) using the Illumina high-throughput sequencing platform. In addition, a long insert library (10 kb) was generated and sequenced with SMRT sequencing technology on a PacBio platform at BaseClear (Leiden, The Netherlands). Genome assembly was performed as described previously (Faino and Thomma, 2014). Contiguity of scaffolds and the correct placement of contigs were manually assessed (Figure S1). To this end, we manually inspected paired-end reads for concordant mapping and coverage of gapped regions between contigs (Figure S1). Genome statistics were calculated using QUAST (Gurevich et al., 2013). Repeats were identified and characterized using the *de novo* repeat identification with RepeatModeler (Smit and Hubley, 2008) (1.0.7; default setting). Subsequently, *de novo* identified repeats were combined with the repeat library from RepBase (Jurka et al., 2005) (release 31-04-2014) to identify and annotate families of repetitive elements using RepeatMasker (Smit et al., 1996) (open-4.0.5) in sensitive mode.

### Karyotyping

Karyotyping was performed as outlined by de Jonge et al (2013) (de Jonge et al., 2013). Briefly, *V. tricorpus* protoplasts were prepared as described by Mehrabi and colleagues (Mehrabi et al., 2012). Karyotyping was performed using a CHEF Mapper XA pulsed field electrophoresis system (Bio-Rad) using the auto algorithm function with low- and high-molecular weight setting at 2 and 6 Mb, respectively. Chromosomes were separated in 0.8% low EEO agarose (US Biologicals) gel. Two chromosome size markers (Bio-Rad) were loaded as a reference marker. After electrophoresis, the gel was stained with ethidium bromide (1  $\mu$ g/ml) in water for 1 h and de-stained in water for 2 h.

## Gene prediction and annotation

For RNA sequencing, *V. tricolorpus* was cultured in a ten-day-old tomato cell suspension for two days. RNA was isolated based on Trizol RNA extraction (Simms et al., 1993). An RNA-Seq library (180 bp insert size) was prepared and sequenced (100-bp reads) at BGI. The RNA-Seq reads were mapped onto the draft *V. tricolorpus* genome using Tophat (Trapnell et al., 2009) and PASA (Haas et al., 2003). Oases (Schulz et al., 2012) and Trinity (Grabherr et al., 2011) were used to assemble the RNA-Seq data independently from the genome assembly. The assembled transcripts derived from Oases, Trinity and PASA were combined and used as cDNA/ESTs evidence in Maker2 (Holt and Yandell, 2011). Moreover, assembled RNA-Seq reads from *V. dahliae* were used as evidence for gene prediction as well as protein sequences from 35 different fungal proteomes (Table S2). The annotation of the predicted genes was done by BLASTp (Altschul et al., 1990) searches to the NCBI NR database and by InterProScan (Jones et al., 2014). The BLASTp analysis and the InterProScan were merged together using Blast2GO (Conesa et al., 2005).

To estimate gene expression levels in *V. tricolorpus*, RNA-Seq reads were mapped to the draft genome of *V. tricolorpus* using TopHat (Trapnell et al., 2009). Subsequently, the number of reads that were mapped to each gene were determined and normalized for the length of the transcript (number of reads per 1 kb transcript). Similarly, gene expression in *V. dahliae* strain JR2 was estimated using previously generated RNA-Seq data (de Jonge et al., 2012). For each condition, the reads were mapped to the draft genome assembly of *V. dahliae* strain JR2 using TopHat (Trapnell et al., 2009) and summarized as reads per kilobase of transcript per million of mapped reads.

Secreted proteins were predicted using a combination of SignalP3 (Dyrløv Bendtsen et al. 2004), SignalP4 (Petersen et al., 2011), TargetP (Emanuelsson et al., 2000) and TMHMM (Krogh et al., 2001). Proteins showing positive predictions from (i) both predictors (HMM and neural network, default settings) of SignalP3, (ii) SignalP4 (default settings), and (iii) TargetP (predicted localization “secreted”; default settings) were considered as candidate effectors. Moreover, we filtered this list by removing proteins that contained (i) more than a single TMHMM predicted transmembrane domain, or (ii) proteins with a single TMHMM predicted transmembrane domain, if this transmembrane domain starts outside of the first 35 amino acids or overlaps less than 10 amino acids with the predicted signal peptide. Protein domains observed in candidate effectors proteins, e.g. NLPs and LysMs, were identified by mining the InterProScan annotations.

## Comparative genomics

The species phylogeny was based on the amino acid sequences of 130 single copy gene families that were identified in the predicted proteomes of *V. tricorpus* and nine related fungi using the CEGMA pipeline (Parra et al., 2007). Individual families were aligned using mafft (LINSi; v7.047b) (Kato et al., 2002) and subsequently concatenated. Maximum likelihood phylogeny was inferred using RAxML (v7.6.3) (Stamatakis, 2006; Stamatakis et al., 2008) with the GAMMA model of rate heterogeneity and the Whelan and Goldman (WAG) model of amino acid substitutions. The robustness of the inferred phylogeny was assessed by 1,000 rapid bootstrap approximations.

The phylogeny of the NLPs was based on the identified NLP domain (PF05630, excluding short fragments) using InterProScan (Jones et al., 2014) of ten filamentous fungi (Figure S4). Subsequently, the NLP domain was excised, aligned with mafft (LINSi; v7.047b) (Kato et al., 2002). Maximum likelihood phylogeny was inferred using RAxML (v7.6.3) (Stamatakis, 2006; Stamatakis et al., 2008) with the GAMMA model of rate heterogeneity and the WAG model of amino acid substitutions. The robustness of the inferred phylogeny was assessed by 1,000 rapid bootstrap approximations.

We built families of gene encoding secreted proteins in the analyzed *Verticillium* and seven related fungi using OrthoMCL (default setting) (Li et al., 2003). Beforehand, similarity between proteins was established by an all-vs-all analyses using BLASTp (e-value cutoff  $1e-5$ , soft filtering) (Altschul et al., 1990; Altschul et al., 1997). The ancestral set of families and the gains and losses were inferred by reconciliation of the binary presence/absence profiles of member constituting an orthologous group with the species phylogeny (Figure S4) using DOLLOP (Felsenstein, 2002) assuming the irreversibility of the loss an orthologous group, i.e. an orthologous group can only be gained, however individual members can be lost multiple times independently. To calculate the number of synonymous substitutions per synonymous site ( $K_S$ ), the number of non-synonymous substitutions per non-synonymous site ( $K_A$ ), and the  $K_A/K_S$  ratio, we extracted sequences of the four analyzed *Verticillium* strains from the previously determined OrthoMCL families. Subsequently, we aligned the coding sequences per gene family guided by protein alignment and determined the pairwise  $K_S$ ,  $K_A$  and  $K_A/K_S$  ratios using the 'kaks' function of the 'seqinr' package (<http://cran.r-project.org/web/packages/seqinr/index.html>). For further analysis, only genes with  $K_A$  or  $K_S$  value  $\leq 5$  were considered to avoid saturation effects. Similarly, we built OrthoMCL families for all genes of the four *Verticillium* and determined  $K_A$ ,  $K_S$  and  $K_A/K_S$  ratios as outlined above.

To identify clustering of genes that encode secreted proteins, we scanned each scaffold with CROC (Pignatelli et al., 2009). Briefly, CROC scans each scaffold, represented as a string of genes, with a sliding window (number of genes per window = 20; slide per window = 1; min 5 genes of interest per window). For each window, the probability to

observe that many (or more) genes encoding secreted genes by chance is assessed using a hypergeometric distribution test (Pignatelli et al., 2009). Subsequently, multiple-testing correction using Benjamini-Hochberg was applied. Transmembrane localized proteins were predicted using TMHMM (Krogh et al., 2001) and nuclear localized proteins were predicted using a combination of WoLF PSORT (Horton et al., 2007) and NLStradamus (Ba et al., 2009), and subsequently clustering was assessed as described above.

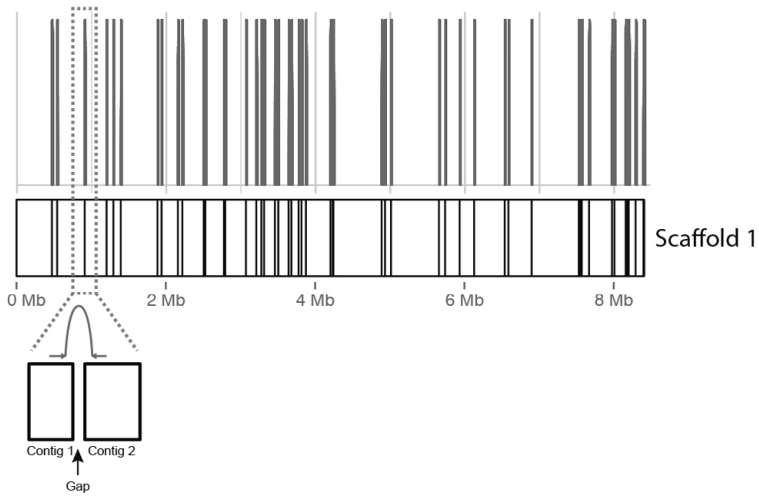
### Data access

The genome assembly has been associated with the NCBI GenBank BioProject PRJNA229139 and the BioSample SAMN02415140. The assembly is deposited at DDBJ/EMBL/GenBank under the accession JPET00000000. The raw RNA-Seq sequence reads have been associated with the NCBI Sequence Read Archive under the same BioProject and BioSample numbers and is deposited under the accession SRX658508.

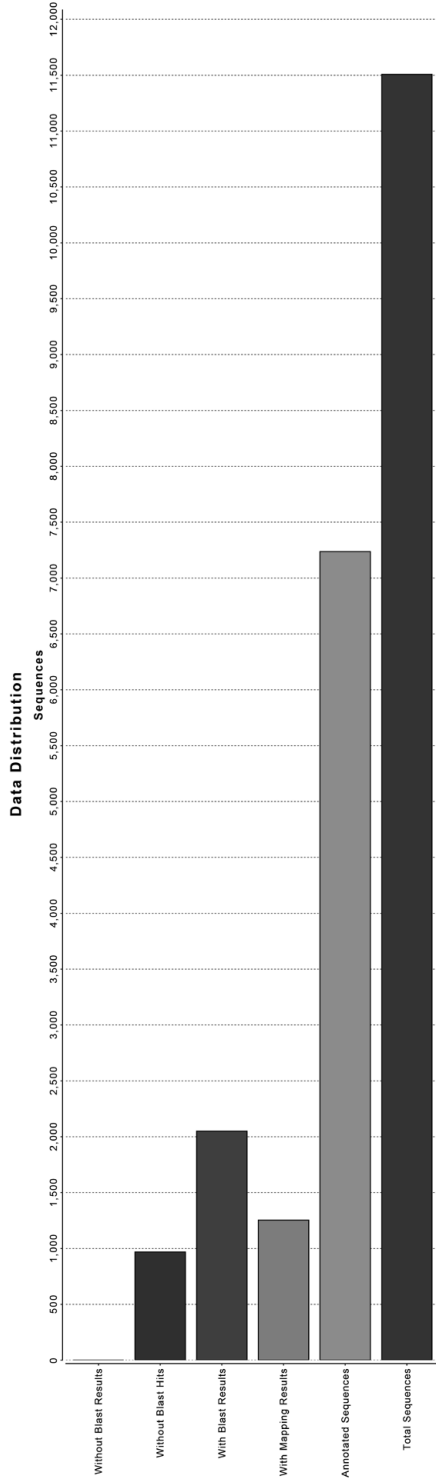
### Acknowledgements

The authors acknowledge support by the Research Council Earth and Life Sciences (ALW) of the Netherlands Organization of Scientific Research (NWO). Moreover, we thank B. Essenstam for excellent plant care and X. Wang (USDA-ARS) for excellent technical assistance.

## Supplementary material

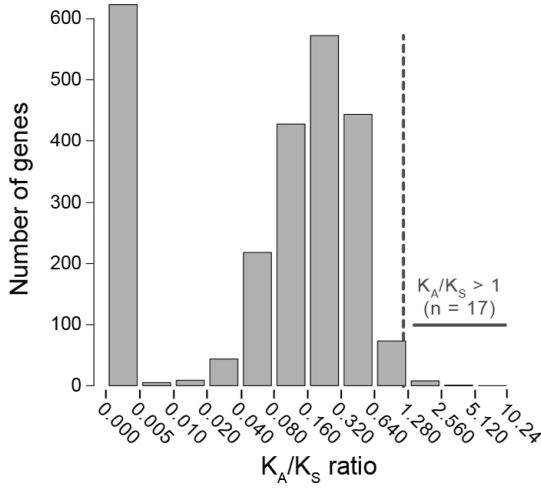


**FIGURE S1. Manual inspection of contig placement using mate-pair reads.** Assembly gaps were manually inspected by confirming the mapping of mate-pair reads (blue arches) that connect the edges of two contigs, as exemplified for the gaps (black lines) on scaffold 1 (illustrated by the close-up). This analysis confirmed that all contigs were correctly placed.

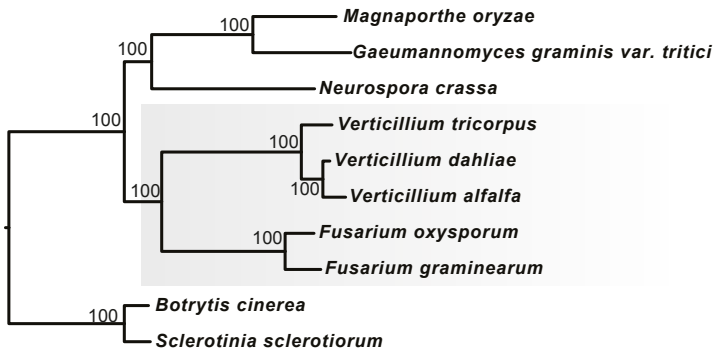


**FIGURE S2. Distribution of annotated sequences.** Bar chart displays the distribution of all predicted proteins and their respective fractions that have been annotated using Blast2GO and InterProScan.

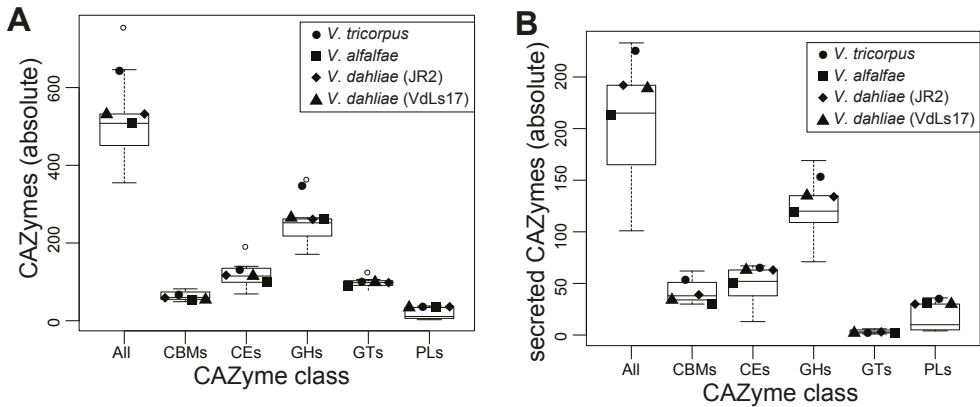




**FIGURE S3. Distribution of  $K_A/K_S$  ratios for genes encoding secreted proteins of *V. dahliae* strain JR2 with orthologs in any of the other analyzed *Verticillium* secretomes.** Genes with  $K_A/K_S$  ratios > 1, indicative of positive selection, are highlighted.



**FIGURE S4. Phylogenetic relationship of *V. tricorpus*, *V. dahliae*, *V. alfalfae* and related fungal species.** Phylogenetic relationship between these fungi was based on a concatenated marker (130 families), and the robustness of the tree was assessed using 1,000 bootstrap replicates.



**FIGURE S5. The genomes of *Verticillium* spp. encode a large repertoire of carbohydrate active enzymes.**

Boxplots display the absolute number of carbohydrate active enzymes (CAZymes) in the (A) full proteomes, and (B) secretome of *Verticillium* spp. and related fungi (Figure S4). The abundance of proteins belonging to the five CAZyme classes - glycoside hydrolases (GHs), polysaccharide lyases (PLs), carbohydrate esterases (CEs), and glycosyltransferases (GTs), and carbohydrate-binding modules (CBMs) - are shown. The position of each of the *Verticillium* is explicitly highlighted, and outliers of the data are shown as open circles.

**TABLE S1. Read statistics for the different DNA and RNA sequencing data sets used in the genome assembly and RNA-Seq analyses.**

	DNA-seq			RNA-seq
	paired-end (Illumina)	mate-pair (Illumina)	long reads (PacBio)	paired-end (Illumina)
Number reads	6,001,030	10,362,806	97,870	7,931,365
Bases (Mb)	1,200,206	1,036,281	195,114	1,586,273
Read length (average)	100	50	1,993.6	100

**TABLE S2. Predicted proteomes used to aid the genome annotation of *V. tricorpus*.**

Species	Source	Date accessed	Citation
<i>Allomyces macrogynus</i>	BROAD Institute	30-6-2013	[1]
<i>Aspergillus niger</i>	BROAD Institute	30-6-2013	Andersen et al. (2011)
<i>Aspergillus oryzae</i>	BROAD Institute	30-6-2013	Machida et al. (2005)
<i>Batrachochytrium dendrobatidis</i>	BROAD Institute	30-6-2013	[2]
<i>Blastomyces dermatitidis</i>	BROAD Institute	30-6-2013	[7]
<i>Botrytis cinerea</i>	BROAD Institute	30-6-2013	Amselem et al. (2011)
<i>Candida albicans</i>	BROAD Institute	30-6-2013	Jones et al. (2004)
<i>Chaetomium globosum</i>	BROAD Institute	30-6-2013	[6]
<i>Coccidioides immitis</i>	BROAD Institute	30-6-2013	Sharpton et al. (2009)
<i>Coprinopsis cinerea</i>	BROAD Institute	30-6-2013	Stajich et al. (2010)
<i>Cryptococcus neoformans grubii</i>	BROAD Institute	30-6-2013	Janbon et al. (2014)
<i>Fusarium graminearum</i>	BROAD Institute	30-6-2013	Cuomo et al. (2007)
<i>Fusarium oxysporum</i>	BROAD Institute	30-6-2013	Ma et al. (2010)
<i>Gaeumannomyces graminis var. tritici</i>	BROAD Institute	30-6-2013	[8]
<i>Histoplasma capsulatum</i>	BROAD Institute	30-6-2013	[3]
<i>Laccaria bicolor</i>	BROAD Institute	30-6-2013	Martin et al. (2008)
<i>Lodderomyces elongisporus</i>	BROAD Institute	30-6-2013	[5]
<i>Magnaporthe oryzae</i>	BROAD Institute	30-6-2013	Dean et al. (2005)
<i>Microsporum canis</i>	BROAD Institute	30-6-2013	Martinez et al. (2012)
<i>Mortierella verticillata</i>	BROAD Institute	30-6-2013	[1]
<i>Neurospora crassa</i>	BROAD Institute	30-6-2013	Galagan et al. (2003)
<i>Paracoccidioides brasiliensis</i>	BROAD Institute	30-6-2013	Desjardins et al. (2011)
<i>Phaeosphaeria nodorum</i>	BROAD Institute	30-6-2013	Hane et al. (2007)
<i>Puccinia graminis f. sp. tritici</i>	BROAD Institute	30-6-2013	Duplessis et al. (2011)
<i>Pyrenophora tritici-repentis</i>	BROAD Institute	30-6-2013	Manning et al. (2013)
<i>Rhizopus delemar</i>	BROAD Institute	30-6-2013	Ma et al. (2009)
<i>Schizosaccharomyces pombe</i>	BROAD Institute	30-6-2013	Wood et al. (2002)
<i>Sclerotinia sclerotiorum</i>	BROAD Institute	30-6-2013	Amselem et al. (2011)
<i>Spizellomyces punctatus</i>	BROAD Institute	30-6-2013	[1]
<i>Trichophyton rubrum</i>	BROAD Institute	30-6-2013	Martinez et al. (2012)
<i>Uncinocarpus reesii</i>	BROAD Institute	30-6-2013	[4]
<i>Ustilago maydis</i>	BROAD Institute	30-6-2013	Kämper et al. (2006)
<i>Verticillium alfalfae</i>	BROAD Institute	30-6-2013	Klosterman et al. (2011)
<i>Verticillium dahliae</i>	BROAD Institute	30-6-2013	Klosterman et al. (2011)
<i>Saccharomyces cerevisiae</i>	Saccharomyces Genome Database	30-6-2013	Goffeau et al. (1996)

[1] Origins of Multicellularity Sequencing Project, Broad Institute of Harvard and MIT (<http://www.broadinstitute.org/>)

[2] "*Batrachochytrium dendrobatidis* Sequencing Project, Broad Institute of Harvard and MIT (<http://www.broadinstitute.org/>)

[3] "*Histoplasma capsulatum* Sequencing Project, Broad Institute of Harvard and MIT (<http://www.broadinstitute.org/>)"

[4] "*Uncinocarpus reesii* Sequencing Project, Broad Institute of Harvard and MIT (<http://www.broadinstitute.org/>)

[5] "*Candida* Sequencing Project, Broad Institute of Harvard and MIT (<http://www.broadinstitute.org/>)"

[6] "*Chaetomium globosum* Sequencing Project, Broad Institute of Harvard and MIT (<http://www.broadinstitute.org/>)"

[7] "*Blastomyces dermatitidis* Sequencing Project, Broad Institute of Harvard and MIT (<http://www.broadinstitute.org/>)"

[8] "*Magnaporthe comparative* Sequencing Project, Broad Institute of Harvard and MIT (<http://www.broadinstitute.org/>)"

## Chapter 2

**TABLE S3. Functional annotation of the predicted proteins in *V. tricorpus* extracted from Blast2GO.**

[http://www.scientificsocieties.org/MPMIXtras/2014/MPMI-06-14-0173-R\\_TableS3.xls](http://www.scientificsocieties.org/MPMIXtras/2014/MPMI-06-14-0173-R_TableS3.xls)

**TABLE S4. Number of members of different families of carbohydrate active enzymes in the predicted proteomes of several analyzed fungi.**

[http://www.scientificsocieties.org/MPMIXtras/2014/MPMI-06-14-0173-R\\_TableS4.xls](http://www.scientificsocieties.org/MPMIXtras/2014/MPMI-06-14-0173-R_TableS4.xls)

**TABLE S5. Overview of the SSCP that are encoded by the genomes of the analyzed *Verticillium* spp. as well as related fungi, and their functional annotation, if available.**

[http://www.scientificsocieties.org/MPMIXtras/2014/MPMI-06-14-0173-R\\_TableS5.xlsx](http://www.scientificsocieties.org/MPMIXtras/2014/MPMI-06-14-0173-R_TableS5.xlsx)



# Chapter **3**

## **Genus-wide comparative genomics reveals similar features between pathogenic and non-pathogenic *Verticillium* species**

Xiaoqian Shi-Kunne, Zhuyun Bian, Luigi Faino, Melvin D. Bolton,  
Michael F. Seidl\* and Bart P.H.J. Thomma\*

\*These authors contributed equally

## Abstract

The fungal genus *Verticillium* contains ten species, some of which are notorious causal agents of vascular wilt diseases, while others are mainly known as saprophytes that only cause disease as opportunistic pathogens. Whereas the genome of the most notorious plant pathogen, *Verticillium dahliae*, has been quite well characterized, genome characteristics of the other species have received relatively little attention thus far. In this study, we sequenced the genomes of all haploid *Verticillium* spp. with the aim to identify genomic features that can be associated with pathogenicity. We demonstrate that all these species display similar genomic features, including the occurrence of extensive genomic rearrangements and the presence of extensive effector catalogs. Overall, this study reveals that no particular genomic feature can be associated to pathogenicity in the genus *Verticillium*.

## Introduction

Many plants and microbes live in close interaction and establish symbiotic relationships, ranging from mutualistic through endophytic to pathogenic (Rovenich et al., 2014). The antagonistic nature of pathogenic interactions between plants and pathogens is thought to result in an everlasting arms race for continued symbiosis (Jones and Dangl, 2006; Cook et al., 2015). In such arms race, plant pathogens utilize effector repertoires for immunity subversion in order to achieve successful colonization while plant hosts, in turn, attempt to reinstall immunity through recognition of effectors and other invasion-associated molecular patterns (Cook et al., 2015).

It is generally believed that microbial pathogenicity, including plant pathogenicity, evolved multiple times independently during evolution (Shang et al., 2016). Although microbial plant pathogens are found in many phylogenetically distantly related species, they tend to exhibit similar genomic features. Eukaryotic pathogen genomes often harbor higher numbers of genes than genomes of closely related non-pathogenic species (Galagan et al., 2003; Armbrust et al., 2004; Dean et al., 2005; Tyler et al., 2006; Cuomo et al., 2007; Martinez et al., 2008). Particularly, pathogenic species often encode larger secretomes than their saprophytic relatives (Lo Presti et al., 2015). Interestingly, effector genes are frequently found in regions that are enriched for repetitive elements and it has been hypothesized that repetitive elements are involved in chromosomal rearrangements, segmental duplications and genome size increase (Raffaele and Kamoun, 2012; Seidl and Thomma, 2017). As a result, effector genes that are located at repeat rich regions may be rapidly lost, but also gained (de Jonge et al., 2013; Seidl and Thomma, 2014; Faino et al., 2016; Dong et al., 2015; Raffaele and Kamoun, 2012). Typically, the genomes of pathogens are larger and contain higher amounts of repeats than the genomes of their non-pathogenic relatives (Raffaele and Kamoun, 2012).

*Verticillium* is a small genus that consists of ten species of soil-borne fungi (Inderbitzin et al., 2011a). Whereas *V. longisporum* is an allodiploid hybrid species, all other *Verticillium* species are haploids (Inderbitzin et al., 2011b; Depotter et al., 2017). *Verticillium dahliae* is the most notorious plant pathogen of the genus since it can cause vascular wilt disease on hundreds of plant species, resulting in large economic losses every year (Fradin and Thomma, 2006). Furthermore, also *V. longisporum*, *V. albo-atrum*, *V. alfalfae* and *V. nonalfalfae* are plant pathogens, albeit with narrower host ranges than *V. dahliae* (Inderbitzin et al., 2011a; Inderbitzin et al., 2011b; Depotter et al., 2016). Although the remaining species *V. tricorpus*, *V. zaregamsianum*, *V. nubilum*, *V. isaacii* and *V. klebahnii* have only been incidentally reported as opportunistic plant pathogens, they are able to enter plant roots and establish plant-associated relationships (Ebihara et al., 2003; Inderbitzin et al., 2011a; Gurung et al., 2015).

The genomes of few *Verticillium* spp. have been studied in detail over the last years. Various strains of *V. dahliae* have been sequenced (Klosterman et al., 2011; de Jonge et al., 2013; Seidl et al., 2015), and for two strains a gapless genome assembly is available (Faino et al., 2015; Faino et al., 2016). Comparative genomics has revealed the occurrence of extensive large-scale genomic rearrangements, likely mediated by erroneous double-stranded break repair, that gave rise to lineage-specific genomic regions (de Jonge et al., 2013; Seidl et al., 2015; Faino et al., 2016). These lineage-specific regions are enriched for *in planta*-expressed effector genes that encode secreted proteins that are involved in mediating successful host colonization (de Jonge et al., 2013). This finding suggests that genomic plasticity provides an important basis for *V. dahliae* adaptation (de Jonge et al., 2013; Seidl and Thomma, 2014; Faino et al., 2016). In addition to *V. dahliae*, genomes of a single strain of both *V. alfalfae* and *V. tricorpus* have been sequenced as well (Klosterman et al., 2011; Seidl et al., 2015). Unexpectedly, it was observed that, despite being a saprophyte, *V. tricorpus* encodes a similar effector repertoire as *V. dahliae* and similarly shows co-localization (clustering) of genes that encode secreted proteins (Seidl et al., 2015). Here, we report genome assemblies for all haploid *Verticillium* species. We investigated various genomic features among all *Verticillium* species, focusing on traits that are commonly associated with pathogenicity. By assessing these genomic features, we show that all haploid *Verticillium* species evolved similar genomic traits regardless of their lifestyle.

## Results & Discussion

### High-quality *de novo* *Verticillium* genome assemblies

Previously, we sequenced and assembled the genomes of various strains of *V. dahliae* as well as the genome of a strain of *V. tricorpus* (Klosterman et al., 2011; de Jonge et al., 2013; Faino et al., 2015; Seidl et al., 2015). Furthermore, the genomes of two *V. nonalfalfae* strains are publicly available from the short-reads archive (Accession numbers: SRX1020629 and SRX1020613). To try and infer evolutionary relationships among *Verticillium* spp., we decided to additionally sequence the genomes of strains of the remaining haploid species, namely *V. albo-atrum* and *V. alfalfae*, *V. isaacii*, *V. klebahnii*, *V. tricorpus* and *V. zaregamsianum*. We obtained sequences of two strains for *V. dahliae*, *V. nonalfalfae*, *V. isaacii*, *V. klebahnii*, *V. tricorpus* and *V. zaregamsianum*, and one strain for the remaining species.

All genome sequences were determined by using the Illumina HiSeq2000 platform. In this manner, ~17 million paired-end reads (150 bp read length of a 500 bp insert size library) and mate-paired reads (150 bp read length of a 5 kb insert size library) were produced per strain. Subsequently, reads were *de novo* assembled into 34–37 Mb draft genomes



(Table 1), which is comparable to the assemblies of *V. dahliae*, *V. tricorpus* and *V. alfalfae* (Klosterman et al., 2011; de Jonge et al., 2013; Faino et al., 2015; Seidl et al., 2015). All assemblies resulted in less than 100 scaffolds ( $\geq 1$  kb), except for *V. isaacii* strain PD618, *V. nonalfalfae* strain TAB2, *V. nonalfalfae* strain Recica91 and *V. nubilum* strain PD621 that were assembled into 122, 167 795 and 189 scaffolds, respectively. From all genome assemblies, that of *V. tricorpus* strain PD593 had the best quality as the genome was assembled into the lowest number of contigs (71) and scaffolds (9) with the highest N50 (4.4 Mb) and an overall relatively low amount of unknown nucleotides (Ns; 14.34 per 100 kb). We did not use the publicly available genome assembly of *V. alfalfae* strain VaMs.102 since it is considerably more fragmented (4,098 contigs) than any of the assemblies generated in this study (Klosterman et al., 2011).

**Table 1. Assembly statistics of the *Verticillium* genomes used in this study.**

Species	Strain name	Genome size (Mb)	# Ns/100 kb	N50 (Mb)	# Contigs ( $\geq 0$ bp)	# Scaffolds ( $\geq 1000$ bp) (%)	Repeats (%)	CEGMA (%)	BUSCO (%)
<i>V. albo-atrum</i>	PD670	37.4	56.67	3.7	107	72	3.20	94	99
	PD747	36.5	16.26	3.9	34	19	3.10	94	99
<i>V. alfalfae</i>	PD683	32.7	19.36	4.5	40	14	4.55	95	99
<i>V. dahliae</i>	VdLs17	36.0	0	5.9	8	8	12.05	95	99
	JR2	36.2	0	4.2	8	8	12.30 <sup>1</sup>	96	99
<i>V. nonalfalfae</i>	Recica91	33.0	35.4	0.9	1026	795	3.68	94	98
	TAB2	34.3	897.53	1.8	793	167	5.90	96	98
<i>V. nubilum</i>	PD621	37.9	9.13	4.7	246	189	10.92	97	99
<i>V. tricorpus</i>	MUCL9792	36.0	229.64	4.7	255	53	2.70	96	99
	PD593	35.0	14.52	4.4	71	9	3.14	96	99
<i>V. isaacii</i>	PD618	35.8	62.48	3.1	239	122	3.19	96	99
	PD660	36.0	37.53	2.5	114	43	3.42	96	99
<i>V. klebahnii</i>	PD659	36.2	59.47	3.6	120	60	3.04	96	99
	PD401	36.0	35.3	3.2	79	37	2.79	93	99
<i>V. zaregamsianum</i>	PD736	37.1	62.7	2.0	125	62	3.04	96	99
	PD739	37.1	55.38	3.5	75	32	3.31	94	99

<sup>1</sup> Repeat contents of *V. dahliae* strains JR2 and VdLs17 were estimated by Faino et al. (2016).

To assess the completeness of the assembled gene space for each of the assemblies, the orthologs of 248 core eukaryotic gene families were queried using the CEGMA pipeline (Parra et al., 2007). All assembled *Verticillium* genomes contain 93%-96% of these eukaryotic core genes (Table 1). Additionally, we also used the Benchmarking Universal Single-Copy Orthologs (BUSCO) software to assess assembly completeness with 1,438 fungal genes as queries (Simão et al., 2015). BUSCO resulted in over 98% of completeness for each of the assembled genomes. Considering that we found CEGMA and BUSCO scores of 96% and 99%, respectively, when assessing the previously generated complete and gapless

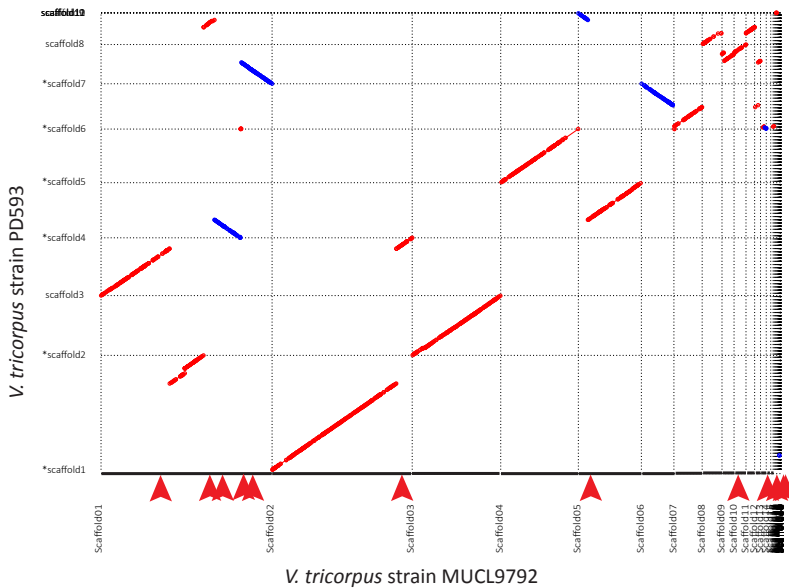
genome assembly of *V. dahliae* strain JR2 (Faino et al., 2016), it can be concluded that a (nearly) complete gene space was assembled for all nine *Verticillium* species.

### Identification of genomic rearrangements

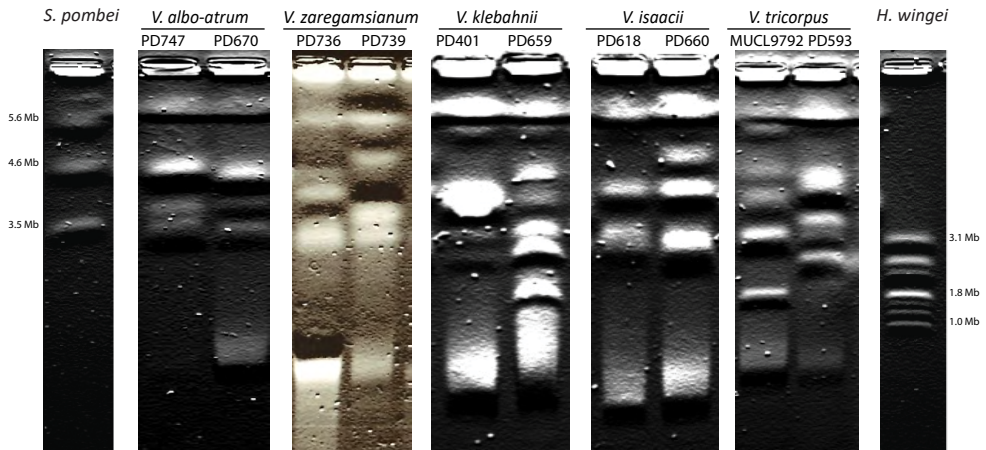
To investigate whether large-scale chromosomal rearrangements are confined only to *V. dahliae*, we performed whole-genome alignments between the two strains of each *Verticillium* spp. Firstly, we performed a pairwise alignment of the two genome assemblies from *V. tricorpus* since these assemblies are of superior quality. Despite a high degree of overall sequence identity (99.6%), we observed repeated interruption of synteny blocks in the alignment, indicative of chromosomal rearrangements (Figure 1). In total, 15 inter-chromosomal rearrangements could be identified in the pairwise alignment of the two assemblies.

To investigate occurrence of rearrangements in the remaining *Verticillium* spp., we subsequently extended the assessment of the occurrence of chromosomal rearrangements to the other species of which we have two strains with high quality assemblies. As the corresponding assemblies are considerably more fragmented than those of *V. tricorpus*, we only assessed rearrangements between contigs that are larger than 0.5 Mb. In total, 16, 11, 14 and 15 inter-chromosomal rearrangements were observed in *V. albo-atrum*, *V. zaregamsianum*, *V. isaacii* and *V. klebahnii*, respectively (Figure S2, Table S1). Thus, similar to *V. dahliae* and *V. tricorpus*, chromosomal rearrangements are also found in the other *Verticillium* species.

Unlike the previously generated genome assemblies of the two *V. dahliae* strains, for which optical mapping was used (Klosterman et al., 2011; de Jonge et al., 2013), the current assemblies are solely based on short reads. Although all assemblies are of high quality and chromosomal break points were not accompanied by gaps, the occurrence of rearrangements due to mis-assembly cannot be excluded completely. In order to support our observations, we performed karyotyping by pulsed-field gel electrophoresis (Figure 2). Extensive karyotype variations were observed between the two stains of each species, confirming the occurrence of extensive chromosomal rearrangements. Collectively, our findings show that extensive chromosomal rearrangements occur in all *Verticillium* species and are not confined to the most aggressive plant pathogen, *V. dahliae*.



**FIGURE 1. Whole-genome alignment between two *Verticillium tricorpus* strains reveals frequent chromosomal rearrangements.** The whole-genome dot-plot displays structural polymorphisms between scaffolds of two (near) complete genome assemblies of *V. tricorpus* strains MUCL9792 and PD593. Forward alignments are shown in red and reverse alignments in blue. Identified synteny breakpoints in *V. tricorpus* strain MUCL9792 are indicated by triangular arrows. For clarity, scaffolds indicated by an asterisk of *V. tricorpus* strain PD593 have been inverted.



**FIGURE 2. Pulsed-field gel electrophoresis of two strains of various *Verticillium* species.** Pulsed-field gel electrophoresis was performed on selected strains to confirm our findings of chromosome. Chromosomal DNA of *Schizosaccharomyces pombe* and *Hanseula wingei* were loaded as size markers.

## Repeat content

To assess whether the repeat content differs among genomes of *Verticillium* spp., we identified repeats in all assemblies using RepeatMasker, which uses previously identified repeat databases and performs *de novo* repeat identification (Jurka et al., 2005; Smit and Hubley, 2010). We found that repeats comprise approximately 3% of most genome assemblies (Table 1). This repeat content is similar to that of *V. tricorpus* strain MUCL9792 (2.7%) (Seidl et al., 2015), but slightly less than the initially reported repeat content of *V. dahliae* that was assessed at 4.33% and 4.05% for strains JR2 and VdLs17, respectively (Seidl et al., 2015). Moreover, this estimation of the repeat content of the *V. dahliae* genomes was compromised by the relatively high level of fragmentation of the assemblies (1,520 and 3,000 contigs for the JR2 and VdLs17 genome assemblies, respectively), which may account for the slightly higher estimations for those genomes (Klosterman et al., 2011; de Jonge et al., 2013).

In general, repetitive regions are more difficult to be assembled using short reads. Consequently, it is generally appreciated that the final size of a genome assembly tends to be an underestimation of the actual genome size. Interestingly, a recent re-assessment of the repeat content based on novel gapless genome assemblies revealed a repeat content of around 12% for both *V. dahliae* strains JR2 and VdLs17 (Faino et al., 2015). This shows that the initial estimation of the repeat content based on short-read assemblies at ~4% for both strains is a significant underestimation. To determine the genome size of each strain independent from its genome assembly, we analyzed the distribution of k-mer frequencies in the raw sequencing reads. An estimated genome size that is lower than the assembly size would suggest underestimation of the repeat content. Based on the relationship between base number, k-mer number and sequencing depth, genome sizes can be estimated (Li and Waterman, 2003). In order to validate this method, we first estimated the genomic size of *V. dahliae* strain JR2. Indeed, we obtained the same genome size (36.2 Mb) as that of the gapless genome assembly (Faino et al., 2015). Subsequently, we determined the sizes of all genomes and compared these to the generated assemblies. The difference with the assemblies was less than two Mb for each strain, except for strains of *V. nonalfalfae*, TAB2 and Recica91, for which the difference was 2.3 and 3.4 Mb, respectively (Table S2). This suggests that the sizes of the genome assemblies are relatively close to the estimated genome sizes.

Interestingly, *V. nubilum* has an increased repeat content of 10% compared to the other species that have repeat contents of 3%-4%. The repeat content of *V. nubilum* is comparable to the repeat content of *V. dahliae* (12%) that was based on novel gapless genome assemblies. Notably, whereas *V. dahliae* is a successful plant pathogen, *V. nubilum* is not and thus an increased repeat content cannot be associated with a particular life style.

**TABLE 2. Summary of protein numbers, secreted proteins, LysM and NLP proteins for each of the *Verticillium* genome assemblies.**

Species	Strain name	# proteins	# secreted proteins (% of total protein)	# LysM effectors	# of NLPs
<i>V. dahliae</i>	JR2	10,719	858 (7.5)	3	7
	VdLs17	10,885	895 (8.2)	4	8
<i>V. alfalfae</i>	PD683	10,852	966 (8,9)	3	7
<i>V. nonalfalfae</i>	TAB2	11,029	979 (8,9)	3	7
	Recica91	11,032	978 (8,9)	3	7
<i>V. nubilum</i>	PD621	10,55	962 (9,1)	5	6
<i>V. tricorpus</i>	MUCL9792	11,509	937 (8,1)	5	6
	PD593	10,636	1092 (10,3)	5	6
<i>V. isaacii</i>	PD618	10,852	1035 (9,5)	7	5
	PD660	11,116	1029 (9,3)	7	5
<i>V. klebahnii</i>	PD659	10,998	1057 (9,6)	7	5
	PD401	10,947	1055 (9,6)	7	5
<i>V. albo-atrum</i>	PD670	10,991	1120 (10,2)	7	8
	PD747	11,202	1092 (9,8)	7	8
<i>V. zaregamsianum</i>	PD736	11,244	1082 (9,6)	6	6
	PD739	11,274	1090 (9,6)	7	6

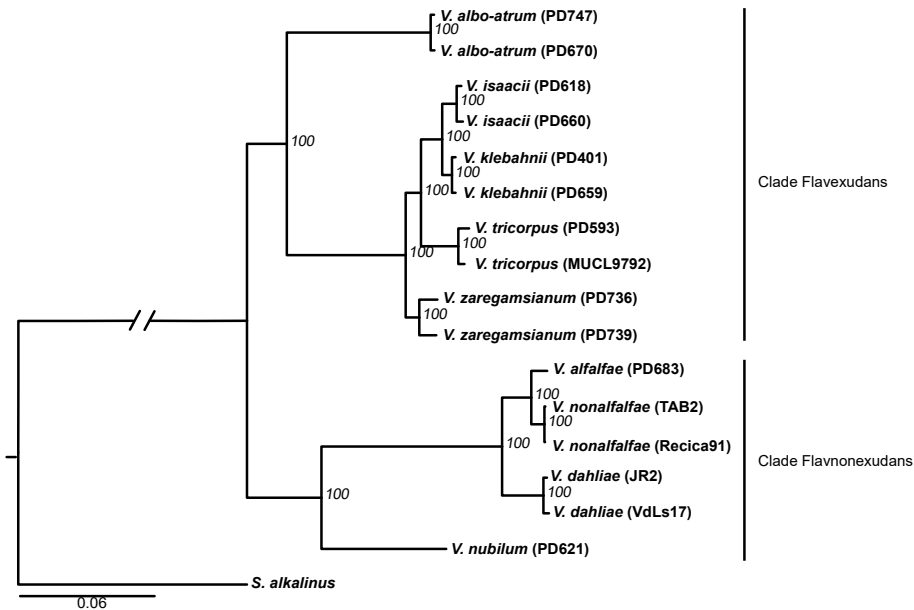
### Comparison of predicted proteome characteristics

In order to compare the number of protein-coding genes among the *Verticillium* spp., we inferred a reference gene annotation for each of the genomes by integrating *de novo* and homology-based data using the Maker2 pipeline (Holt and Yandell, 2011). In addition to proteomes of *V. dahliae* and *V. tricorpus*, which were manually annotated previously, we used 35 predicted fungal proteomes that represent a broad phylogenetic distribution to guide gene structure annotation (Klosterman et al., 2011; Faino et al., 2015; Seidl et al., 2015). This gene prediction approach yielded around 11,000 protein-coding genes for each of the genomes, with the highest number of 11,430 for *V. dahliae* strain JR2 and lowest number of 10,636 genes predicted for *V. tricorpus* strain PD593. The second *V. dahliae* strain that was analyzed (VdLs17) only contains 10,885 genes while the second *V. tricorpus* strain that was analyzed (MUCL9792) contains 11,509 genes (Table 2). Thus, while the gene count can differ for the two strains of the same species, the overall gene numbers appear to be similar for each of the *Verticillium* spp.

### Reconstruction of phylogeny

Previously, the phylogenetic relationship among *Verticillium* spp. was constructed based on the combined alignment of protein sequences of actin (ACT), elongation factor 1-alpha (EF), glyceraldehyde-3-phosphate dehydrogenase (GPD) and tryptophan synthase (TS),

showing that the species fall into two major clades: clade Flavexudans with species producing yellow-pigmented hyphae including *V. albo-atrum*, *V. isaacii*, *V. klebahnii*, *V. tricorpus* and *V. zaregamsianum*, and clade Flavnonexudans with species devoid of yellow-pigmented hyphae, including *V. alfalfae*, *V. dahliae*, *V. nonalfalfae*, and *V. longisporum* (Inderbitzin et al., 2011a). Based on whole-genome information we now constructed a robust phylogenetic relationship using concatenated protein sequences of 4,653 single-copy orthologs of each of the *Verticillium* spp. and *Sodiomyces alkalinus* as an out-group for a maximum-likelihood analysis. The single-copy ortholog-based tree forms two major clades (Figure 3), which is consistent with the previously constructed tree (Inderbitzin et al., 2011a). Moreover, the clade Flavonoexudans contains the majority of the pathogenic species, although pathogenic species are not clustered within a single clade.



**FIGURE 3. Reconstruction of the evolutionary trajectories of *Verticillium* species.** Maximum-likelihood phylogeny analysis of *Verticillium* species rooted with *Sodiomyces alkalinus*. The phylogenetic tree is based on a concatenated maker (4,653 single-copy orthologs), and the robustness of the tree was assessed using 1,000 bootstrap replicates.

### Comparison of predicted carbohydrate active enzymes

It has previously been noted that *V. dahliae* has a relatively large complement of carbohydrate-active enzyme (CAZyme) genes when compared with other fungi (Klosterman et al., 2011). Moreover, the previously sequenced *V. tricorpus* genome carried an even larger complement of CAZyme genes, owing to an expansion of the glycoside hydrolase (GH) and carbohydrate-binding module (CBM) families (Seidl et al., 2015).

To determine the CAZyme complement of each of the *Verticillium* spp., the predicted proteomes were queried for signature domains that belong to each of the five catalytic classes; glycoside hydrolase (GH), polysaccharide lyases (PL), carbohydrate esterases (CE), glycosyltransferases (GT) and carbohydrate-binding modules (CBM). This prediction resulted in 499 to 700 CAZymes for each strain (Table 3), with the highest number for the two *V. zaregamsianum* strains (690 and 700 for PD736 and PD739, respectively) and the lowest number for the two *V. dahliae* strains (499 and 532 for JR2 and VdLs17, respectively). Overall, we observed that the species of clade Flavexudans contain more CAZymes than the Flavnonexudans species. Thus, the expansion or contraction of CAZyme families cannot be associated with pathogenicity in *Verticillium*.

**TABLE 3. Summary of CAZyme numbers for each of the *Verticillium* genome assemblies.**

Species	Strain name	CBMs <sup>1</sup>	CEs <sup>2</sup>	GHs <sup>3</sup>	GTs <sup>4</sup>	PLs <sup>5</sup>	Total
<i>V. dahliae</i>	JR2	55	106	251	92	31	499
	VdLs17	59	117	261	98	36	532
<i>V. nonalfalae</i>	TAB2	69	115	285	101	37	561
	Recica91	69	116	286	101	37	563
<i>V. alfalae</i>	PD683	64	114	281	98	37	551
<i>V. nubilum</i>	PD621	63	123	298	95	36	573
<i>V. tricorpus</i>	MUCL9792	70	135	348	103	38	646
	PD593	75	130	359	103	39	656
<i>V. isaacii</i>	PD618	76	129	360	103	38	655
	PD660	74	130	360	103	38	654
<i>V. klebahnii</i>	PD659	79	130	362	103	40	663
	PD401	77	132	361	105	40	664
<i>V. albo-atrum</i>	PD747	74	148	357	100	42	668
	PD670	74	142	355	102	40	661
<i>V. zaregamsianum</i>	PD736	88	139	381	99	42	690
	PD739	88	140	387	103	42	700

<sup>1</sup> CBMs = carbohydrate binding modules

<sup>2</sup> CEs = carbohydrate esterases

<sup>3</sup> GHs = hydrolases

<sup>4</sup> GTs = glycosyltransferases

<sup>5</sup> PLs = polysaccharide lyases

## Comparison of predicted secondary metabolite gene clusters

Many filamentous fungi produce secondary metabolites (SM), which are not essential for primary growth but are important for colonizing a specific niche (Keller et al., 2005). In plant pathogenic fungi, SMs often play an important role in plant infection and virulence (Keller et al., 2005; Bolton and Thomma, 2008; Bolton et al., 2014). Fungal SMs can be classified into four groups based on key enzymes and precursors involved in their biosynthesis. Polyketides

and non-ribosomal peptides represent the most widespread groups in fungi. The first step of their biosynthesis relies on key enzymes: polyketide synthases (PKSs), non-ribosomal peptide synthases (NRPSs) and hybrid PKS-NRPSs (Keller et al., 2005). The two remaining groups are terpenes and alkaloids whose biosynthesis relies on terpene cyclases (TCs) and dimethylallyl tryptophan synthases (DMATs), respectively (Keller et al., 2005). However, SM biosynthesis is complex and typically involves several tailoring steps that are catalyzed by discrete decorating enzymes (Martín et al., 2005). In fungi, genes encoding the enzymes involved in the production of a particular secondary metabolite are usually organized in clusters, meaning that they are located at the same locus in the genome and that they are co-regulated (Keller and Hohn, 1997). We predicted secondary metabolite gene clusters for each of the genomes using antiSMASH (Medema et al., 2011), which resulted in 22 to 33 clusters for each of the genomes (Table 4). The lowest number of secondary metabolite clusters was found in both *V. alfalfae* strains and in *V. nonalfalfae* strain TAB2 (22 clusters for both strains). Subsequently, all clusters were sub-categorized into NRPs, PKS type 1, PKS type 3, terpene, hybrid indole-PKS, hybrid NRP-PKS, and uncharacterized clusters. The lower numbers of SM cluster in both *V. alfalfae* strains and in *V. nonalfalfae* strain TAB2 are due to lower numbers of uncharacterized SM clusters and T1PKS, as well as lower numbers of NRP-T1PKS and indole-T1PKS clusters than in the other strains.

**TABLE 4. Prediction of secondary metabolism gene clusters for each of the *Verticillium* genome assemblies.**

Species	Strain name	PKS <sup>1</sup>		NRPS <sup>2</sup>	NRP-PKS <sup>3</sup>	Indole-PKS <sup>4</sup>	Terpene	Other	Total clusters
		PKS type1	PKS type3						
<i>V. albo-atrum</i>	PD670	6	1	4	5	1	4	8	29
	PD747	6	1	4	5	1	4	8	29
<i>V. tricorpus</i>	MUCL9792	8	1	4	5	1	3	6	28
	PD593	8	1	4	5	1	3	7	29
<i>V. isaacii</i>	PD618	8	1	4	5	1	3	7	29
	PD660	8	1	4	5	1	3	7	29
<i>V. klebahnii</i>	PD659	11	1	4	5	1	4	7	33
	PD401	11	1	4	5	1	4	7	33
<i>V. zaregamsianum</i>	PD736	10	1	4	6	1	3	6	31
	PD739	11	1	4	4	1	4	6	31
<i>V. nubilum</i>	PD621	8	1	6	1	1	3	8	28
<i>V. alfalfae</i>	PD683	6	1	6	0	0	4	5	22
<i>V. nonalfalfae</i>	TAB2	6	1	6	0	0	4	5	22
	Recica91	7	1	4	2	0	4	9	27
<i>V. dahliae</i>	JR2	9	1	3	1	0	4	8	26
	VdLs17	9	1	3	1	0	4	9	27

<sup>1</sup> PKS = polyketide synthases

<sup>2</sup> NRPS = non-ribosomal peptide synthases

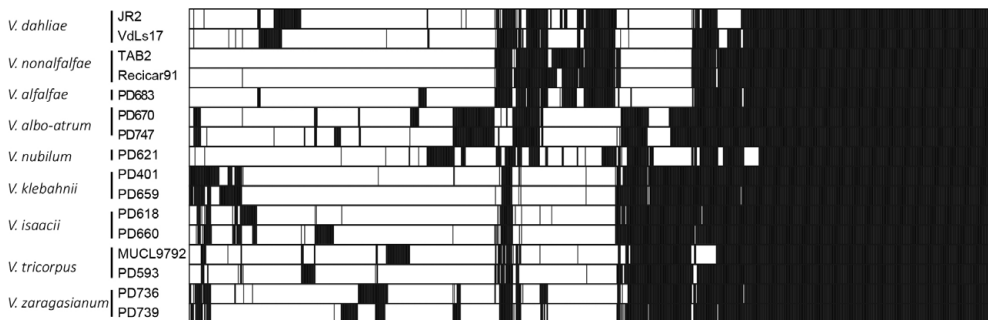
<sup>3</sup> NRP-PKS = hybrid polyketide synthases-non-ribosomal peptide synthases.

<sup>4</sup> Indole-PKS = hybrid indole-polyketide synthases



## Prediction and comparison of effector repertoires

We previously showed that *V. tricorpus* encodes an effector repertoire that largely resembles that of its pathogenic relatives *V. dahliae* and *V. alfalfae* (Seidl et al., 2015). To compare secretome sizes and compositions among each of the *Verticillium* spp., we identified N-terminal signal peptides in the predicted protein catalogues. The prediction resulted in 858 -1120 proteins potentially belonging to the secretome for each of the strains (Table 2). Intriguingly, species of the clade Flavnonexudans, which contains most of the pathogenic species, tend to have lower numbers of secreted proteins than species of the clade Flavexudans. There are in total 819 secreted protein families that are shared by at least two *Verticillium* species, suggesting that a high proportion of secreted proteins are shared between species. Among the shared secreted protein families, 336 are shared by all *Verticillium* strains (Figure 4). Only a small proportion of secreted proteins are species-specific in each species, with *V. alfalfae* (11) and *V. albo-atrum* (62) containing the lowest and highest portion of species-specific secreted proteins, respectively. This suggests that pathogenic species do not necessarily contain higher numbers of species-specific secreted proteins.



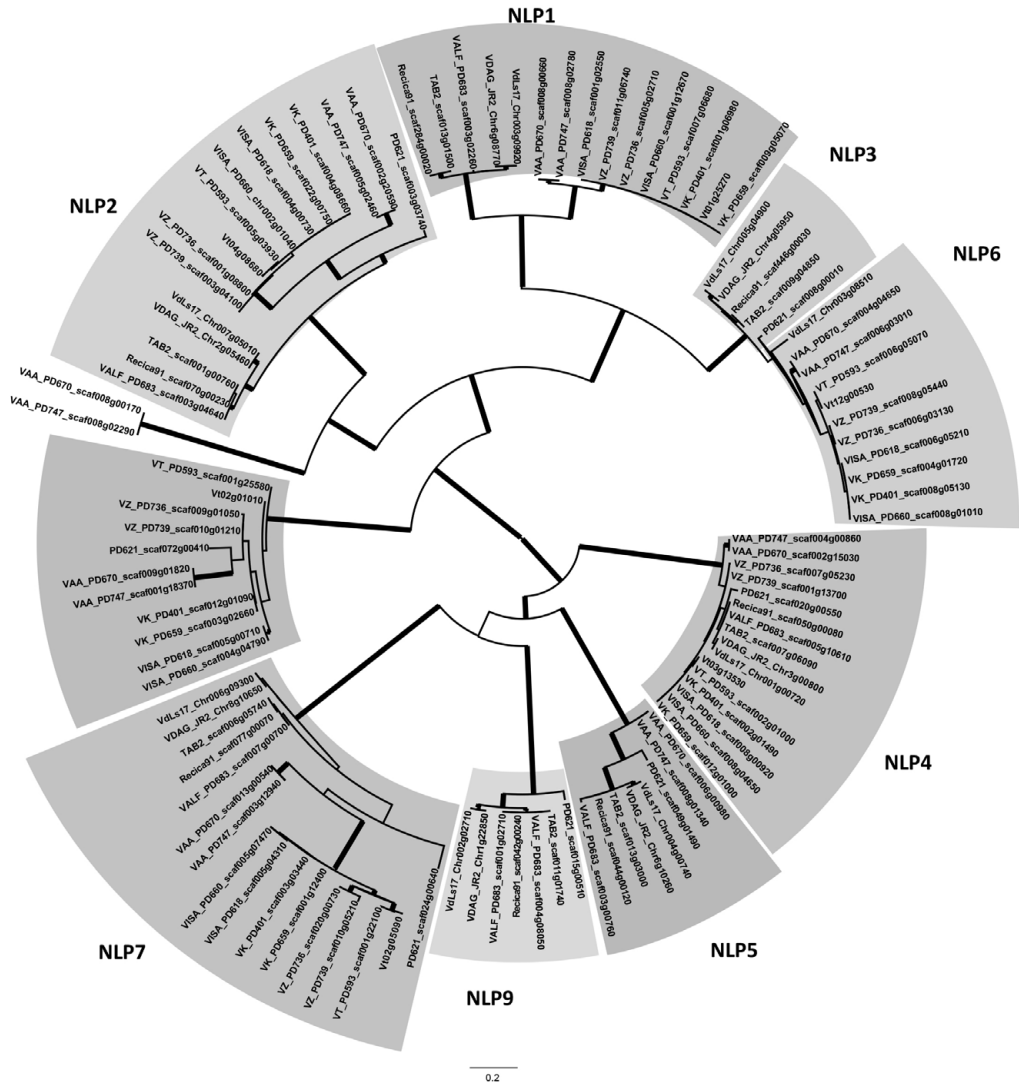
**FIGURE 4. Conservation pattern of all secreted protein-encoding proteins among the *Verticillium* spp.** Presence or absence of secreted proteins among *Verticillium* species is indicated with black and white, respectively.

Typically, pathogen effectors are species- or even strain-specific. However, some effectors are more broadly conserved, such as LysM effectors that are found in fungal species with diverse life styles (Bolton et al., 2008; de Jonge and Thomma, 2009; de Jonge et al., 2010; Kombrink and Thomma, 2013). In pathogenic fungi, including *V. dahliae*, LysM effectors have been implicated in pathogenicity (Bolton et al., 2008; de Jonge and Thomma, 2009; de Jonge et al., 2010; Kombrink and Thomma, 2013). Initially, seven LysM effector candidates were found in the genome of *V. dahliae* strain VdLs17 (Klosterman et al., 2011), of which only four were subsequently qualified as *bona fide* LysM effector candidates (Kombrink et al., 2017). Comparative genomics of several *V. dahliae* strains revealed that only one

of these four LysM effectors (Vd2LysM) is strain-specific and contributes to virulence during infection (de Jonge et al., 2013; Kombrink et al., 2017). To determine the presence of putative LysM effectors in each of the *Verticillium* strains, the predicted secretomes were queried for the presence of LysM domains (de Jonge and Thomma, 2009; Kombrink, 2014). We observed that the species in the clade Flavexudans encode more putative LysMs effectors when compared to the species from the clade Flavnonexudans (Table 2).

Another class of widely conserved effectors comprises the necrosis and ethylene-inducing-like proteins (NLPs), which are not only found in fungi but also in bacteria and oomycetes (Gijzen and Nürnberger, 2006). Many of these NLPs are particularly known for their phytotoxicity (Gijzen and Nürnberger, 2006; de Jonge et al., 2011). Compared to many fungal species that contain one or two NLPs, several *Verticillium* spp. (*V. dahliae*, *V. alfalfa* and *V. tricorpus*) were shown to have expanded NLP families containing six to eight NLPs (Klosterman et al., 2011; Seidl et al., 2015). Two *V. dahliae* NLPs, NLP1 and NLP2, are required for virulence on tomato and *Arabidopsis* (Santhanam et al., 2013). We predicted the presence of NLPs in each of the *Verticillium* genomes by searching for the conserved NLP domains. Orthologs of only four NLPs (NLP1, NLP2, NLP4, and NLP7) can be found in all genomes (Figure 5). A previous study showed that NLP8 from *V. tricorpus* strain MUCL9792 has no orthologs in *V. dahliae* and *V. alfalfae* (Seidl et al., 2015). Here, we found that *V. nubilum* and all species from the clade Flavexudans carry NLP8 orthologs, suggesting that orthologs of NLP8 were likely lost in *V. dahliae*, *V. nonalfalfae* and *V. alfalfae*.

In summary, we illustrated that the *Verticillium* species contain similar number of secreted proteins. Members of some of the identified effector families are also observed in non-pathogenic species and, thus, do not represent genuine pathogenicity factors but, arguably, the functions of most secreted protein families remain unknown. Nevertheless, the *Verticillium* species that are considered as non-pathogenic occasionally occur as opportunistic pathogens (Isaac, 1953; Usami et al., 2011; Gurung et al., 2015). For instance, it has been demonstrated that *V. tricorpus* and *V. nubilum* are able to induce wilt symptoms when plants have first been challenged with additional stresses, such as high soil nitrogen levels or waterlogging (Isaac, 1953).



**FIGURE 5. Phylogenetic relationship of necrosis and ethylene inducing-like proteins (NLPs).** Orthologous groups of previously described NLP classes in *Verticillium* are indicated (Santhanam et al., 2013). Thicker branches indicate bootstrap values higher than 60.

## Conclusions

We compared genomic features that have commonly been associated with microbial plant pathogenicity among all haploid species of the *Verticillium* genus that contains pathogenic as well as non-pathogenic species. We demonstrate that all species contain similar genomic features, suggesting that no particular features can be linked to *Verticillium* pathogenicity. However, it needs to be noted that the *Verticillium* species that are considered as non-pathogenic are occasionally identified as opportunistic plant pathogens (Isaac, 1953; Usami et al., 2011; Gurung et al., 2015). For instance, it has been demonstrated that *V. tricorpus* and *V. nubilum* are able to induce wilt symptoms when plants have first been challenged with additional stresses, such as high soil nitrogen levels or waterlogging (Isaac, 1953). On the one hand, our study suggests that the ability to cause disease is caused by subtle genomic traits that do not easily become apparent from whole-genome comparisons. On the other hand, perhaps all species have the ability to cause disease, yet the phenotype of the interaction is determined more by environmental than by intrinsic genomic factors.

## Materials & Methods

### DNA extraction

DNA of each strain was isolated from mycelium fragments that were filtered from 3-day-old cultures grown in potato dextrose broth (PDB) at 28°C. Next, the mycelium was suspended in water in a 50 ml tube, briefly vortexed and centrifuged at maximum speed (13,000 rpm) for 10 minutes. After removal of the supernatant, 0.5 mL 0.1 M Tris HCl (pH 8) was added to the pellet followed by mixing and transfer to a 2 mL screw cap tube. Three metal beads (3 mm) were added to the tube and the tube was shaken in a bead beater at maximum speed for 45 seconds. Next, 1 mL of phenol:chloroform:isoamylalcohol (24:24:1) was added. After mixing and centrifugation at 13,000 rpm for 7 minutes, the water phase was transferred to a new tube. Subsequently, 800 µl isopropanol was added and mixed by gentle inversion, then the tube was centrifuged. After removal of the isopropanol, the DNA pellet was air dried and subsequently dissolved in 50 µL of deionized water.

### Genome sequencing and assembly

Of each *Verticillium* strain, two libraries (500 bp and 5 kb insert size) were prepared and sequenced using the Illumina High-throughput sequencing platform (KeyGene N.V., Wageningen, the Netherlands). Genome assemblies were performed using A5 pipeline (Tritt et al., 2012), and the remaining sequence gaps were subsequently filled using SOAP GapCloser (Luo et al., 2012). After obtaining the final genome assemblies, QUASt (Gurevich et al., 2013) was used to calculate genome statistics.

Genome sizes were calculated by formula:  $G$  (genome size) =  $Nk$  (total number of bases) /  $Ck$  (highest peak of 17-mer) (Li and Waterman, 2003).

### Gene prediction and annotation

Protein-coding genes were annotated with the Maker2 pipeline (Holt and Yandell, 2011). Thirty-five predicted fungal proteomes that represent a broad phylogenetic distribution were used to guide gene structure annotation (Seidl et al., 2015). Additionally, the previously annotated proteomes of *V. dahliae* and *V. tricornutus* were used (Seidl et al., 2015; Faino et al., 2016).

Secreted proteins were predicted by using a combination of TMHMM (Krogh et al., 2001), SignalP4 (Petersen et al., 2011) and TargetP (Emanuelsson et al., 2000). Proteins were considered secreted when positively predicted by SignalP4 (containing signal peptide, default setting) and TargetP (predicted localization “secreted”; default setting). We further refined this set by excluding proteins that contain transmembrane domains as predicted by TMHMM, except for proteins that contain only a single transmembrane domain that starts within the first 35 amino acids or overlaps for more than 10 amino acids with the predicted signal peptide. NLPs and LysMs were identified based on NLP (PF05630) and LysM domains (PF01476) using Interproscan (Jones et al., 2014) and hmmscan (HMMER 3.0, E-value < 0.001) (Eddy, 2009), respectively.

### Ortholog analysis and tree building

Ortholog groups were determined using OrthoMCL (Li et al., 2003). To this end, similarity between proteins was established by an all-versus-all analysis using BLASTp (E-value cutoff  $1e-5$ ). The species phylogenetic tree was generated using 4,653 single-copy orthologs that are conserved among all of the genomes. Individual families were aligned using mafft (LINSi; v7.04b) (Katoh et al., 2002) and subsequently concatenated. Maximum likelihood phylogeny was inferred using RAxML (v7.6.3) (Stamatakis, 2014) with the GAMMA model of rare heterogeneity and the WAG model of amino acid substitutions. The robustness of the inferred phylogeny was assessed by 1,000 rapid bootstrap approximations.

### Repeat and rearrangement identification

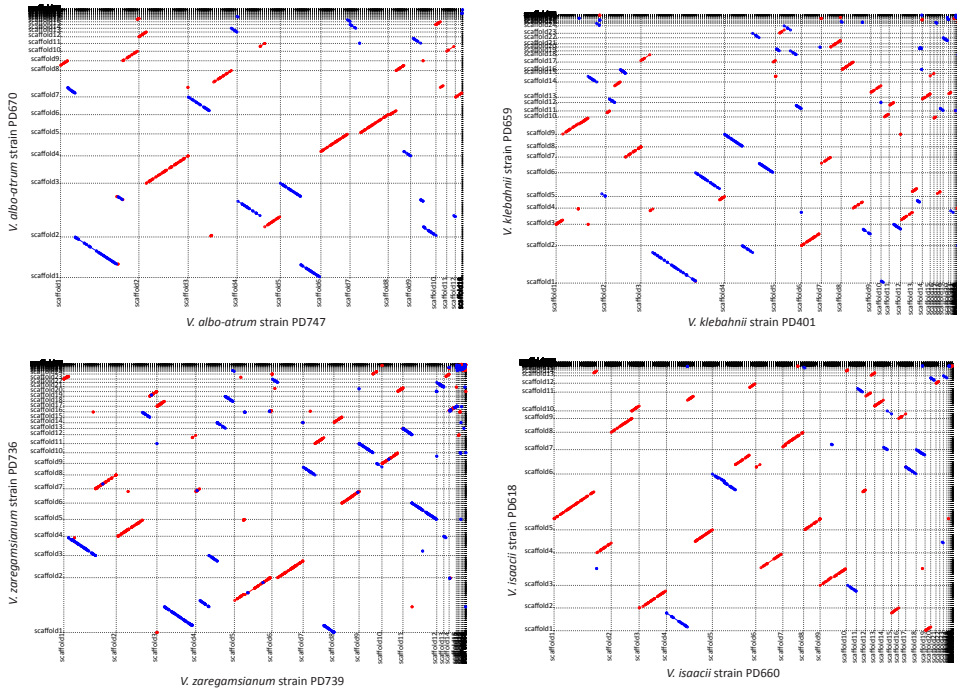
Repeats were identified and characterized using RepeatModeler (version 1.0.7 and default setting) (Smit and Hubley). De-novo-identified repeats were combined with the repeat library from RepBase (Jurka et al., 2005) to annotate families of repetitive elements using RepeatMasker (Smit et al., 1996).

Whole-genome alignments and dot plots were generated with MUMMer 3.0 (Kurtz et al., 2004) using NUCmer with default setting (setting: `-maxmatch`). A custom Python script was used to count rearrangements in each pairwise genome alignment.

### **Karyotyping**

Karyotyping was performed as previously described (de Jonge et al., 2013). Protoplasts of *Verticillium* strains were prepared following the mycelium-based fungal biomass preparation method (Mehrabi et al., 2012). Chromosome size makers from *Hansenula wingei* and *Schizosaccharomyces pombe* were loaded as reference markers.

## Supplementary material



**FIGURE S1. Whole-genome alignments between two strains of *V. albo-atrum*, *V. isaacii*, *V. klebahnii* and *V. zaregamsianum*.** Forward alignments are shown in red and reverse alignments in blue.

**TABLE S1. Number of rearrangements between strains per species.**

Species	# rearrangements
<i>V. dahliae</i>	12
<i>V. albo-atrum</i>	16
<i>V. zaregamsianum</i>	11
<i>V. tricornis</i>	15
<i>V. isaacii</i>	15
<i>V. klebahnii</i>	14

TABLE S2. Genome size estimations for each of the *Verticillium* genomes.

Species	Strain name	K-mer coverage peak	Total k-mer size	Estimated genome size (Mb)	Assembled genome size (Mb)	Genome size difference (Mb)
<i>V. klebahnii</i>	PD659	23	815,015,558	35.4	36.2	0.8
	PD401	31	1,124,860,922	36.3	36.0	0.3
<i>V. tricorpus</i>	PD593	24	831,900,618	34.7	35.0	0.3
	MUCL9792	24	884,648,660	36.9	36.0	0.9
<i>V. isaacii</i>	PD618	29	1,032,254,137	36.9	35.8	1.1
	PD660	26	971,674,348	37.4	36.0	1.4
<i>V. zaregamsianum</i>	PD736	24	899,280,993	37.5	37.1	0.4
	PD739	22	851,634,611	38.7	37.1	1.6
<i>V. nubilum</i>	PD621	44	1,666,117,828	37.9	37.9	0
<i>V. alfalfae</i>	PD683	33	1,124,662,219	34.1	32.7	1.4
<i>V. albo-atrum</i>	PD747	22	826,699,968	37.6	37.4	0.2
	PD670	24	905,070,427	37.7	36.5	1.2
<i>V. nonalfalfae</i>	Recica91	21	770,308,300	36.7	33.0	3.7
	TAB2	23	838,284,578	36.5	34.3	2.2
<i>V. dahliae</i>	JR2	13	471,892,460	36.3	36.2	0.1



# Chapter 4

## **Evolution within the fungal genus *Verticillium* is characterized by chromosomal rearrangement and gene loss**

Xiaoqian Shi-Kunne, Luigi Faino, Grardy C.M. van den Berg,  
Bart P.H.J. Thomma\* and Michael F. Seidl\*

\*These authors contributed equally

This chapter has been published as:

Shi-Kunne X, Faino L, van den Berg GCM, Thomma BPHJ\*, Seidl MF\* (2018) Evolution within the fungal genus *Verticillium* is characterized by chromosomal rearrangement and gene loss. *Environ Microbiol* 20: 1362–1373 (\*equal contribution)

## Abstract

The fungal genus *Verticillium* contains ten species, some of which are notorious plant pathogens causing vascular wilt diseases in host plants, while others are known as saprophytes and opportunistic plant pathogens. Whereas the genome of *V. dahliae*, the most notorious plant pathogen of the genus, has been well characterized, evolution and speciation of other members of the genus received little attention thus far. Here, we sequenced the genomes of the nine haploid *Verticillium* spp. to study evolutionary trajectories of their divergence from a last common ancestor. Frequent occurrence of chromosomal rearrangement and gene family loss was identified. In addition to ~11,000 genes that are shared at least between two species, only 200-600 species-specific genes occur. Intriguingly, these species-specific genes show different features than the shared genes.

## Introduction

Species continuously evolve by genetic variation that enables adaptation to changing and novel environments. In many eukaryotes, such genomic variation is established during sexual reproduction where genetic material of two parents is combined and novel genetic combinations are formed during meiotic recombination (Bell, 1982). Thus, sexual recombination is considered an important driver to establish genetic diversity (Barton and Charlesworth, 1998). However, many species, including fungi, are thought to reproduce strictly asexually (McDonald and Linde, 2002; Heitman et al., 2007; Flot et al., 2013), and have long been considered limited in their capacity to establish genetic variation. Importantly, even though asexual organisms lack meiotic recombination, adaptive evolution occurs and is established by various mechanisms ranging from single-nucleotide polymorphisms to large-scale structural variations that affect chromosomal shape, organization and gene content (Seidl and Thomma, 2014). Over longer evolutionary time-scales, all these processes establish genetic divergence that may ultimately lead to the emergence of novel species. Typically, these processes can be especially well studied in fungi that have relatively small genomes, which greatly facilitates the establishment of high-quality genome assemblies (Thomma et al., 2016).

Previously, gene loss was often neglected as an evolutionary driver, mostly because it was associated with the loss of redundant gene duplicates without apparent functional consequences (Olson, 1999). However, more and more genomic data suggest that gene loss acts as a manifest source of genetic change that may underlie phenotypic diversity (Albalat and Cañestro, 2016). Moreover, reduction or complete loss of gene families has been associated with ecological shifts of fungi (Casadevall, 2008). Human *Malassezia* pathogens that are phylogenetically closest related to plant pathogens such as *Ustilago maydis*, lack fatty acid synthase genes but instead produce secreted lipases to obtain fatty acids from human skin (Xu and Wang, 2007). Gene losses, for instance concerning cell wall-degrading enzymes or secondary metabolite production, have previously also been associated with obligate biotrophic and symbiotic lifestyles of plant-associated fungi (Martin et al., 2008; Spanu et al., 2010; Duplessis et al., 2011).

The fungal genus *Verticillium* consists of ten soil-borne asexual species with different life-styles and host ranges (Inderbitzin et al., 2011a; Klosterman et al., 2011). Among these, *Verticillium dahliae* is a notorious plant pathogen that causes vascular wilt disease on hundreds of plant species, resulting in large economic losses every year (Fradin and Thomma, 2006; Klimes et al., 2015). Furthermore, also *V. longisporum*, *V. albo-atrum*, *V. alfalfae* and *V. nonalfalfae* are plant pathogens, albeit with narrower host ranges (Inderbitzin et al., 2011a). The remaining species *V. tricorpus*, *V. zaregamsianum*, *V. nubilum*, *V. isaacii* and *V. klebahnii* are mostly considered saprophytes that thrive on dead organic material and that occasionally cause opportunistic infections (Ebihara et al., 2003; Inderbitzin et

al., 2011a; Gurung et al., 2015). Of the ten *Verticillium* species, nine are haploids, while *V. longisporum* is a hybrid that arose from inter-specific hybridisation (Inderbitzin et al., 2011b; Depotter et al., 2016).

In addition to the genomes of a single strain of *V. alfalfae* and *V. tricorpus* (Klosterman et al., 2011; Seidl et al., 2015), various strains of *V. dahliae* have been sequenced (Klosterman et al., 2011; de Jonge et al., 2012). Moreover, for two *V. dahliae* strains a gapless genome assembly has been generated (Faino et al., 2015). Comparative genomics between *V. dahliae* strains revealed the occurrence of extensive large-scale genomic rearrangements, likely mediated by erroneous double-stranded break repair, that gave rise to lineage-specific (LS) genomic regions that are enriched for *in planta*-expressed effector genes encoding secreted proteins that mediate host colonization (de Jonge et al., 2013; Faino et al., 2016). These results raise the question whether pathogenic *Verticillium* spp. evolved by prompting extensive chromosomal rearrangements, thus enabling rapid development of effector gene catalogs that are required to be competitive in arms races with host plants and their immune systems. However, whether these chromosomal rearrangements only occur within pathogenic *Verticillium* spp. remained unknown.

Despite the recent advances in *Verticillium* genomics, the evolutionary history of this genus remains unknown so far. Here, we report high-quality genome assemblies of all haploid *Verticillium* species. We reconstructed ancestral *Verticillium* genomes, and reveal processes of genomic diversification that occurred during *Verticillium* evolution.

## Results

### High-quality *de novo* genome assemblies of the haploid *Verticillium* species

To infer evolutionary relationships among *Verticillium* spp., we performed comparative genomics with minimum one genome of each of the nine haploid *Verticillium* species. Previously, we sequenced several *V. dahliae* strains as well as a single *V. tricorpus* strain (Klosterman et al., 2011; de Jonge et al., 2013; Faino et al., 2015; Seidl et al., 2015). Additionally, the genome of *V. alfalfae* has previously been sequenced (Klosterman et al., 2011), as well as the genomes of several *V. nonalfalfae* strains (Bioproject PRJNA283258) (Jelen et al., 2016). Here, we sequenced the genomes of a strain of *V. albo-atrum*, *V. isaacii*, *V. klebahnii*, *V. nubilum* and *V. zaregamsianum* using the Illumina HiSeq2000 platform. In total, ~18 million paired-end reads (150 bp read length of a 500 bp insert size library) and ~16 million mate-paired reads (150 bp read length of a 5 kb insert size library) were produced per strain. Subsequently, reads were *de novo* assembled into 34–37 Mb draft genomes (Table 1), which is comparable to the assemblies of *V. dahliae*, *V. tricorpus* and *V. alfalfae* (Klosterman et al., 2011; de Jonge et al., 2013; Faino et al., 2015; Seidl et al., 2015). All assemblies resulted in less than 100 scaffolds ( $\geq 1$  kb), except for *V. isaacii* strain PD618,

*V. nonalfalfae* strain TAB2 and *V. nubilum* strain PD621 that were assembled into 122, 167 and 189 scaffolds, respectively (Table 1). Notably, the previously obtained assemblies of *V. tricorpus* strain MUCL9792 and *V. alfalfae* strain VaMs.102 were of significantly lower quality than the assemblies of the newly sequenced *Verticillium* species in this study (Klosterman et al., 2011; de Jonge et al., 2013; Faino et al., 2015; Seidl et al., 2015). To obtain better assemblies we decided to sequence an additional strain of each of these two species. Sequencing of *V. tricorpus* strain PD593 and *V. alfalfae* strain PD683 yielded significantly better assembly qualities, with lower numbers of scaffolds (9 for *V. tricorpus* and 14 for *V. alfalfae*) when compared with the previous assemblies (Table 1). Moreover, the assembly of *V. tricorpus* strain PD593 contained seven scaffolds with telomeric repeats at both ends, suggesting the assembly of seven complete chromosomes (Figure S1).

To assess completeness of gene space, the assemblies were queried for the presence of orthologs of 248 core eukaryotic gene families using the CEGMA pipeline (Parra et al., 2007). All assemblies contained 93%-97% of these core genes (Table 1). Additionally, we also used the Benchmarking Universal Single-Copy Orthologs (BUSCO) software to assess assembly completeness with 1,438 fungal genes as queries (Simão et al., 2015). BUSCO indicated >98% completeness for each assembly. Considering that we found CEGMA and BUSCO scores of 95.97% and 98.75%, respectively, when assessing the gapless genome assembly of *V. dahliae* strain JR2 (Faino et al., 2015), we concluded that a (near) complete gene space was assembled for all *Verticillium* species.

**TABLE 1. Assembly statistics for the various *Verticillium* genomes.**

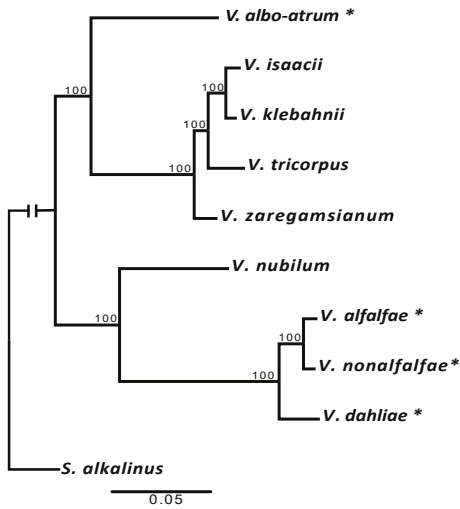
Species	Strain name	Genome size (Mb)	# of Ns per 100_kb	N50 (Mb)	# of contigs ( $\geq 0$ bp)	# of scaffolds ( $\geq 0$ bp)	# of scaffolds ( $\geq 1000$ bp)	CEGMA (%)	BUSCO (%)	# of proteins
<i>V. dahliae</i>	JR2	36.10	0	4.2	8	8	8	95.97	98.75	10,719
<i>V. alfalfae</i>	PD683	32.70	19.36	4.5	40	14	14	94.76	98.82	10,852
<i>V. nonalfalfae</i>	TAB2	34.30	897.53	1.8	793	349	167	95.97	98.47	11,029
<i>V. nubilum</i>	PD621	37.90	9.13	4.7	246	189	189	96.77	99.10	10,550
<i>V. albo-atrum</i>	PD747	36.50	16.26	3.9	34	20	19	94.35	98.96	11,202
<i>V. tricorpus</i>	PD593	35.00	14.52	4.4	71	11	9	95.97	98.82	10,636
<i>V. isaacii</i>	PD618	35.80	62.48	3.1	239	188	122	95.56	98.96	10,798
<i>V. klebahnii</i>	PD401	36.00	35.30	3.2	79	44	37	93.01	99.10	10,998
<i>V. zaregamsianum</i>	PD739	37.10	55.38	3.5	75	46	32	94.35	98.96	11,274

Next, we inferred reference gene annotations by integrating *de novo* and homology-based data using the Maker2 pipeline, making use of 35 predicted fungal proteomes that represent a broad phylogenetic distribution to further guide gene structure annotation

(Klosterman et al., 2011; Faino et al., 2015; Seidl et al., 2015). This approach yielded around 11,000 protein-coding genes for each of the genomes, with the highest number of 11,274 genes for *V. zaregamsianum* and the lowest number of 10,636 genes for *V. tricorpus* (Table 1), which is similar to the number of genes identified in previous assemblies of *Verticillium* spp. (Faino et al., 2015; Seidl et al., 2015). However, automatic annotation of *V. dahliae* strain JR2 yielded fewer genes (10,719) than the previously generated annotation (11,430) that, next to RNA-seq data, also involved manual annotation (Faino et al., 2015). In addition to the difference in the number of predicted genes, we also observed differences in genetic features between both annotation methods, such as the overall GC% and the gene, intergenic and intron lengths (Figure S2). However, for the coding sequences no significant differences in GC% or length were observed (Figure S2). Thus, besides the number of predicted protein-coding genes, manual annotation mainly increased gene lengths by leveraging UTRs. As all following analyses are based on protein-coding genes, and under-estimation of gene numbers is likely similar for each of the genomes, we used Maker2 annotations for all genome assemblies.

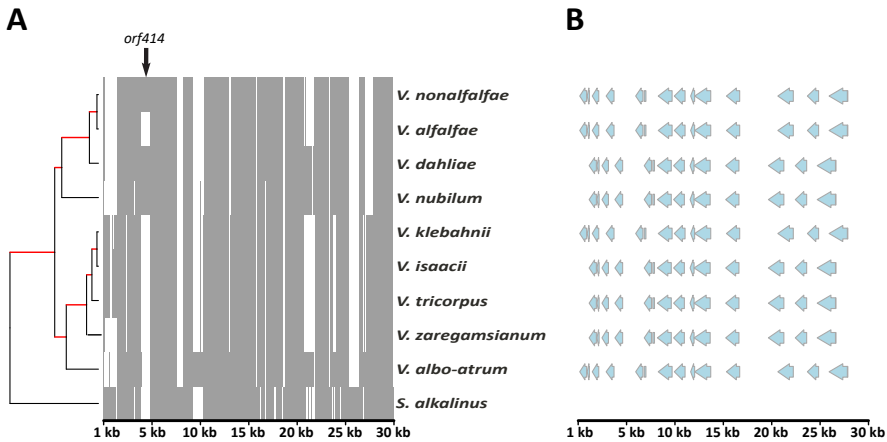
### Phylogenetic relationships within the *Verticillium* genus

To better understand the evolutionary events during the evolution of the *Verticillium* genus, determining a robust phylogenetic relationship between *Verticillium* species is crucial. Previously, a phylogeny was constructed that was inferred from a Bayesian analysis of concatenated alignments of four protein-coding marker genes; *actin* (*ACT*), *elongation factor 1-alpha* (*EF*), *glyceraldehyde-3-phosphate dehydrogenase* (*GPD*) and *tryptophan synthase* (*TS*) (Inderbitzin et al., 2011a). To construct the phylogenetic relationships between the nine *Verticillium* species based on whole-genome data, we used concatenated protein sequences of 5,228 single-copy orthologs and the out-group species *Sodiomyces alkalinus* to construct a maximum-likelihood phylogeny. The resulting phylogeny reveals two major clades (Figure 1), which is consistent with the previous analysis (Inderbitzin et al., 2011a). The major clades are the clade Flavexudans (clade A) containing *V. albo-atrum*, *V. isaacii*, *V. klebahnii*, *V. tricorpus* and *V. zaregamsianum*, which are species producing yellow-pigmented hyphae, and the clade Flavnonexudans (clade B) containing *V. alfalfae*, *V. dahliae*, *V. nubilum* and *V. nonalfalfae*, which are species devoid of yellow-pigment.



**FIGURE 1. Phylogenetic tree of *Verticillium* species.** Maximum-likelihood phylogeny analysis of *Verticillium* species rooted by *Sodiomyces alkalinus*. The phylogenetic tree is based on 5,228 concatenated single-copy orthologs, and the robustness of the tree was assessed using 100 bootstrap replicates. Pathogenic *Verticillium* spp. are marked by asterisks.

Mitochondrial genomes have several unique characteristics, such as a conserved gene content and organization, small size, lack of extensive recombination, maternal inheritance and high mutation rates (Taanman, 1999). This makes them ideal for studying evolutionary relationships among species that diverged during a relatively short period of time. We assembled the complete mitochondrial genomes of each of the *Verticillium* species from paired-end reads using GRABb (Brankovics et al., 2016), which extracts reads derived from the mitochondrial genomes from the total pool of paired-end sequencing reads using 182 published fungal mitochondrial genome sequences as bait. We assembled all mitochondrial reads per strain into a single circular sequence containing 25 to 28 Kb with a GC content of 26-27% (Table 2), in agreement with previous assemblies for *V. dahliae* and *V. nonalfalfae* (Klosterman et al., 2011; Jelen et al., 2016). We subsequently annotated each of the mitochondrial genomes, identifying 15 protein-coding genes. For each mitochondrial genome, all genes were encoded on a single strand and in the same direction (Jelen et al., 2016) (Figure 2B). The whole-mitochondrial-genome alignments were used to construct a maximum likelihood phylogeny (Figure 2A), revealing the same topology as the phylogenetic tree that is based on the nuclear genome (Figure 1). Thus, the mitochondrial genome-based phylogenetic tree further supports the robustness of the *Verticillium* whole-genome phylogeny based on the nuclear genome assemblies, and further corroborates previous phylogenies derived by a limited set of marker genes (Inderbitzin et al., 2011a).



**FIGURE 2. Mitochondrial genome alignments.** (A) Whole mitochondrial genome alignment of all *Verticillium* species and *S. alkalinus*. Grey and white colors represent presence and absence of genomic regions, respectively. The whole mitochondrial genome alignments were used for constructing a maximum likelihood phylogenetic tree. The robustness of the topology was assessed using 100 bootstrap replicate (branches with maximum bootstrap values are in red). (B) Graphic presentation of positions of mitochondrial protein-coding genes and their orders.

It was reported recently that the mitochondrial genomes of both *V. nonalfalfae* and *V. dahliae* carry a *Verticillium*-specific region with a length of 789 bp, named *orf414* (Jelen et al., 2016). When we aligned the mitochondrial genomes of each of the *Verticillium* species and the close relative *S. alkalinus*, we found that *orf414* is absent from *S. alkalinus* (Figure 2A). However, *orf414* is not conserved in the mitochondrial genomes of all *Verticillium* species, as it is only found in *V. nonalfalfae*, *V. dahliae* and *V. nubilum*. Thus, *orf414* cannot be used as a *Verticillium*-specific diagnostic marker (Jelen et al., 2016).

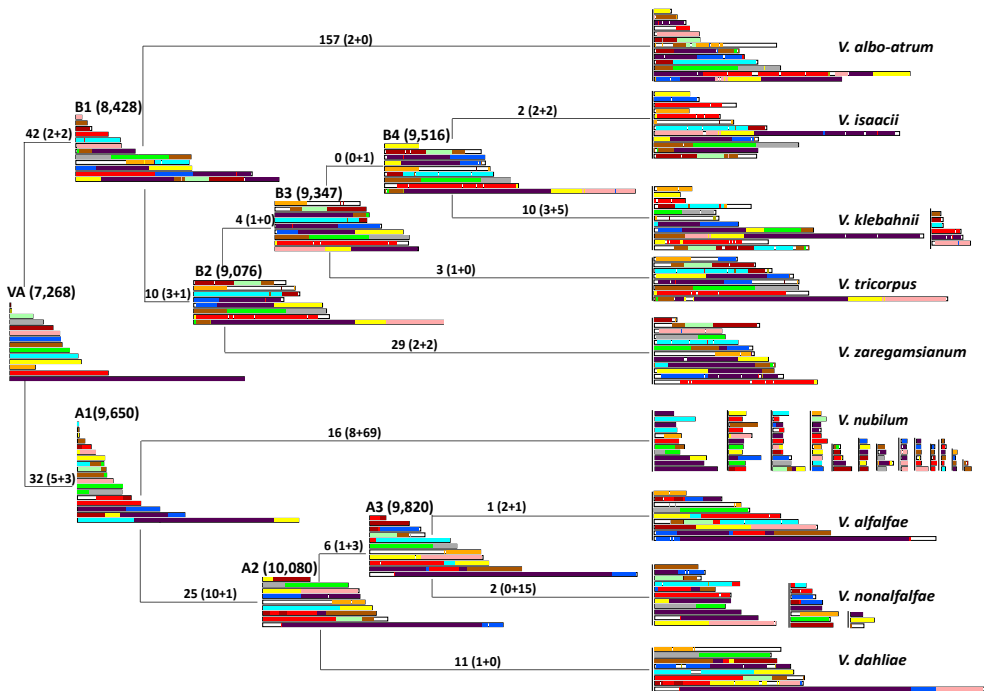
### Reconstruction of ancestor genomes

The degree of synteny between extant species reveals essential information about the evolution of species from their last common ancestor (Hane et al., 2011; Lv et al., 2011). Based on the length of the sequence stretch showing a relatively high degree of sequence similarity and on the degree of co-linearity, synteny can be defined as macro-, micro- or mesosynteny (Hane et al., 2011). Macrosynteny occurs when large (>20 kb) blocks of genes are shared between species in the same order and orientation, while microsynteny involves smaller (<20 kb) segments with relatively few genes. Like microsynteny, mesosynteny involves relatively small (<20 kb) segments, but whereas microsynteny implies that genes are conserved with the same order and orientation, mesosynteny indicates that genes within the segment are conserved with absence of co-linearity. In order to determine the type of synteny between the extant *Verticillium* species, we performed pairwise



alignments of the genomes of the two most divergent *Verticillium* species, *V. dahliae* and *V. albo-atrum* (Hane et al., 2011). These pairwise alignments showed clear regions of macro- and microsynteny, but no mesosynteny (Figure S3). Subsequently, we performed similar analyses for the other *Verticillium* species, confirming the occurrence of macro- and microsynteny, and the absence of mesosynteny between *Verticillium* spp. (Figure S4).

Considering that extensive macrosynteny occurs between the extant *Verticillium* spp. we aimed to reconstruct the evolution of the *Verticillium* genus from its last common ancestor. Inferring the genome organization and gene content of ancestral species has the potential to provide detailed information about the recent evolution of descendant species. However, reconstruction of ancestral genome architectures, followed by integrating into evolutionary frameworks, has only been achieved for a limited number of species (Nakatani et al., 2007; Gordon et al., 2009; Vakirlis et al., 2016). This is largely due to either the unavailability of sequences from multiple closely related species, or due to the high fragmentation of the genome assemblies used. In order to minimize the influence of fragmented draft genome assemblies, we only considered the largest scaffolds that comprise 95% of the total set of protein-coding genes for each of the genomes (Table S1), and constructed ancestral genome organizations that preceded each speciation event using SynChro (Drillon et al., 2014) and AnChro (Vakirlis et al., 2016). SynChro identifies conserved synteny blocks between pairwise comparisons of extant genomes, after which AnChro infers the ancestral gene order by comparing these synteny blocks. In order to validate the accuracy of AnChro, we first reconstructed the genome of the last common ancestor of *V. dahliae* and *V. alfalfae*. We did this separately for two complete and gapless genome assemblies of the extant *V. dahliae* strains JR2 and VdLs17 that, despite extensive genomic rearrangements and the presence of LS sequences (de Jonge et al., 2012; de Jonge et al., 2013; Faino et al., 2016), each contain eight chromosomes (Faino et al., 2015). Irrespective whether the genome of strain JR2 or VdLs17 was used, the resulting ancestor has nine scaffolds that are similarly organized. To compare the two ancestors, synteny blocks were constructed and aligned using SynChro, revealing an overall identical genome structure with only a few genes that lack homolog (Figure S5). Owing to the different lineage-specific regions in the JR2 and VdLs17 genomes, the number of genes in the ancestor varies slightly, with 9,165 and 9,250 protein-coding genes based on the genome of JR2 or VdLs17, respectively. Thus, we concluded that the AnChro software is suitable for reconstruction of ancestral genomes in the genus *Verticillium*.

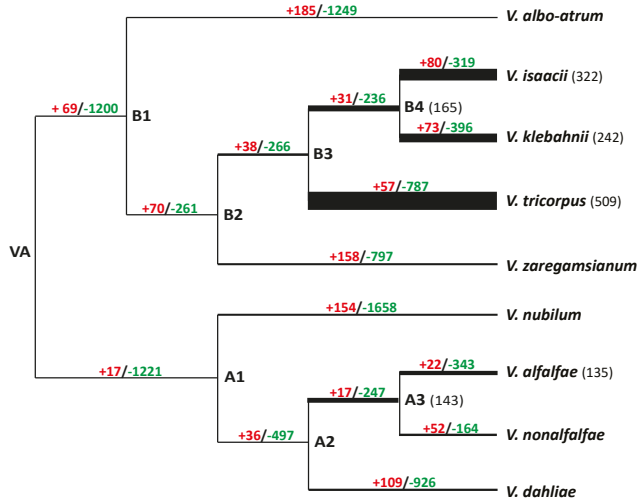


**FIGURE 3. Chromosomal history of *Verticillium* genomes.** Genome structure changes from the most common ancestor (VA) to the nine extant *Verticillium* species. The number of genes of each ancestor is indicated in brackets above the chromosomes. The number of chromosome rearrangements, i.e. the sum of translocations and inversions is indicated above each branch. The number of fusions and fissions between two genomes, respectively, are indicated in brackets.

Ancestral genomes were reconstructed for all the nodes in the phylogenetic tree, resulting in less than 20 scaffolds at each individual node (Figure 3). Using SynChro to determine synteny blocks between the last common ancestor and the ancestor-derived genomes that served as input for ReChro (Vakirlis et al., 2016), the number of rearrangements that occurred in each branch of the tree was determined. In total, 498 rearrangements, including chromosomal fusions and fissions, occurred during the evolution from the last common ancestor to the nine extant haploid *Verticillium* species (Figure 3). Yet, considerable variation occurred between species ranging from 69 rearrangements for *V. tricorpus* to up to 205 for *V. albo-atrum* (Figure 3).

To assess whether genomic rearrangements occurred in a clock-like fashion during the evolution of *Verticillium* species, we related the number of rearrangements per branch with the evolutionary time approximated by the branch lengths inferred from the phylogenetic tree (Figure 1). The branch lengths were represented by either the numbers of substitutions per site (Figure 1), or the relative divergence time that was estimated based on artificial dating of the last common ancestor of *Verticillium* and *S. alkalinus* to 100

units of time (Figure S6). The number of rearrangements per branch showed significant correlation with both representations of branch length ( $R^2= 0.3679$ ,  $P= 0.007541$  for substitution per site and  $R^2= 0.7623$ ,  $P= 6.174e-06$  for relative divergence time) (Figure S7).



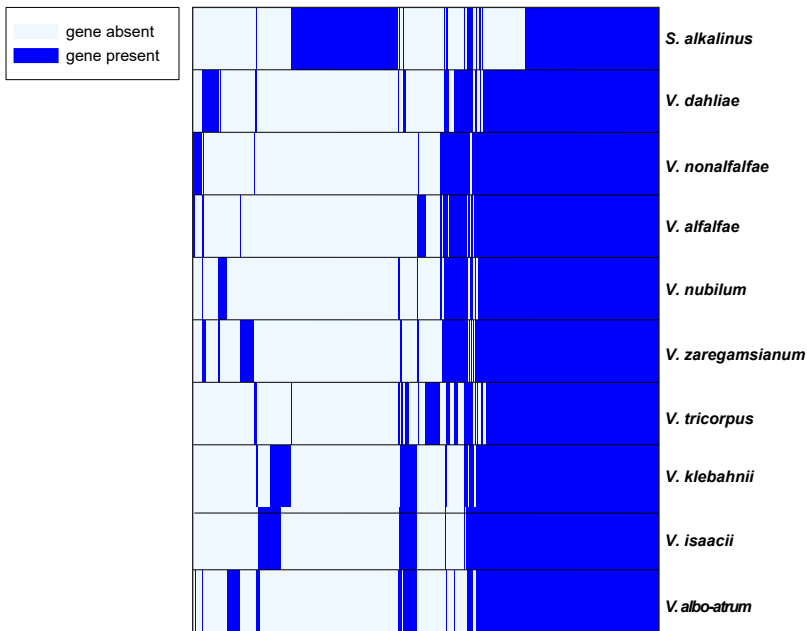
**FIGURE 4. Evolution of *Verticillium* gene repertoire.** The number of expanded (in red) and contracted (in green) gene families was estimated on each branch of the tree under a birth–death evolutionary model. The thickened branches represent the abundance of gene families that evolved rapidly ( $P<0.05$ ) and the exact number of gene families is indicated after each node name.

### Determination of gene family expansions and contractions

To monitor gene family changes during evolution, we estimated gene family size expansion (gains) and contraction (losses) on each branch using CAFE (Han et al., 2013). In this analysis, we considered all gene families that are present in at least two *Verticillium* species. CAFE models the evolution of gene family size across a species phylogeny under a birth–death model of gene gain and loss and simultaneously reconstructs ancestral gene family sizes for all internal nodes, allowing the detection of expanded or contracted families within lineages. The analysis revealed that the last common *Verticillium* ancestor contained 11,902 families with 12,631 genes. Intriguingly, *Verticillium* species generally underwent more extensive gene losses than gains (Figure 4). This pattern of more extensive gene losses than gains was further confirmed by inferring gene family gains and losses by reconciliation of 10,071 gene trees with established *Verticillium* species phylogeny (Figure 1, Figure S8). In order to obtain functional descriptions of gained and lost gene families, we searched for the corresponding Cluster of Orthologous Group (COG) functional categories in the eggNOG database (Huerta-Cepas et al., 2015). In both the A and B clade, the most prevalent COG functional category among gained and lost gene families is “carbohydrate metabolism and transport” (Figure S9). In order to obtain more detailed functional predictions of the gene families

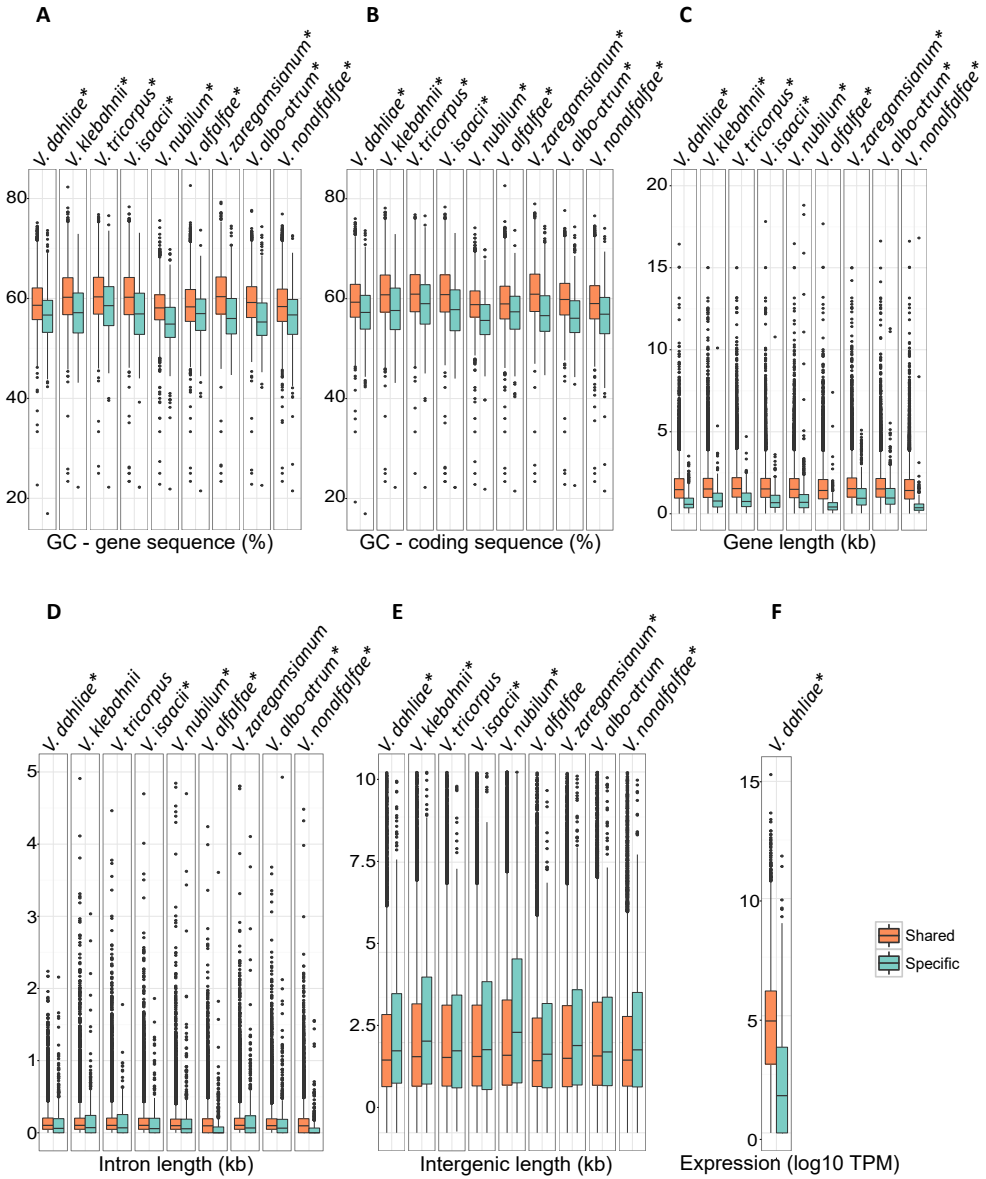
that underwent gains and losses, we first annotated Pfam domains for each gene family (Supplementary Data 1; Supplementary Data 3). Subsequently, using a Hypergeometric test with a false discovery rate (FDR)-corrected P-value of  $<0.05$ , we identified nine and 14 Pfam domains that are enriched in the gained and lost gene families, respectively (Table S2). Among the enriched Pfam domains of clade B, two belong to carbohydrate enzyme gene families, GH3 and GH35. The remainder of the Pfam domains relates to transporter activities, metabolic processes or unknown functions. The enriched Pfam domains of the lost gene families show more diverse functions, including transcription factors and cytochrome P450 activity. Among all expanded and contracted genes families, we found 1,081 gene families that are significantly more variable in size during evolution ( $P < 0.05$ ). Subsequently, we estimated in which branches these gene families evolved more rapidly by using the Viterbi algorithm of CAFE. This algorithm calculates a P-value for each of these 1,081 gene families on each of the branches of the phylogenetic tree. A P-value that is below the threshold ( $P < 0.05$ ) indicates that the respective gene family evolved more rapidly on that particular branch. This analysis revealed that the branches between *V. tricorpus*, *V. isaacii*, *V. klebahnii* and their last common ancestor (B4) evolved most rapidly (involving 509, 322, 242 and 165 gene families, respectively), followed by the branches of *V. alfalfae* and *V. nonalfalfae* with their last common ancestor (A3) (143 and 135 for *V. alfalfae* and A3, respectively) (Figure 4). Next, the overrepresentation of Pfam domains in the most rapidly evolved gene families was assessed using a Hypergeometric test with a FDR-corrected P-value of  $<0.05$ . Under these conditions, only a single Pfam domain PF01636 was found to be enriched ( $P = 1 \times 10^{-6}$ ), concerning a phosphotransferase enzyme family (APH) that is predicted to consist of antibiotic resistance proteins (Sarwar and Akhtar, 1990; Trower and Clark, 1990; Nurizzo et al., 2003). This Pfam domain was also found to be enriched in the lost gene families in *V. isaacii* (Table S2).

Although we observed considerable gene losses and gains along the different branches of the tree, we also observed a large number of shared genes among all extant species. In total, 7,538 genes are common among all *Verticillium* spp., of which 1,689 are absent from *S. alkalinus* (Figure 5). Interestingly, only as little as 123 genes are clade A-specific and 288 genes clade B-specific. These genes underwent losses in the reciprocal clade in a single event only at the last common ancestor of the respective clades, rather than multiple independent losses. For each genome, less than 5% of the total protein-coding genes are species-specific (Figure 5, Supplementary Data 2). These species-specific genes were neither enriched for any Pfam domain nor for genes that encode secreted proteins. To infer the evolutionary trajectory of these species-specific genes, we searched for homologs of these genes in the proteomes of 383 fungal species. Up to 50% of these species-specific genes had homologs in non-*Verticillium* species, suggesting that the species-specific occurrence in *Verticillium* is the result from losses from other *Verticillium* species during evolution (Table S3).



**FIGURE 5. *Verticillium* gene family conservation.** Presence or absence of gene families among *Verticillium* species are indicated by dark and light blue respectively. The gene families (columns) were ordered by hierarchical clustering.

To assess the characteristics of the species-specific genes versus shared genes, we compared features such as GC content, gene lengths, inter-genic lengths and intron lengths. Interestingly, when compared with core genes, species-specific genes significantly differ in GC content of gene sequences as well as of coding sequences, and in gene length (Figure 6A-C). Some species also show significant differences in intron lengths and intergenic lengths between species-specific and core genes (Figure 6D-E). Intriguingly, most of the species-specific genes of *V. dahliae* strain JR2 are not, or only lowly, expressed *in vitro* (Figure 6F).



**FIGURE 6. Gene feature comparisons between specific and shared genes.** The Wilcoxon rank sum test ( $P < 0.01$ , indicated by asterisk) was used to detect significant differences between species-specific and shared genes of each species. (A) GC content based on gene sequences. (B) GC content based on coding sequences. (C) Intron length. (D) Intergenic length. (E) *V. dahliae* *in vitro* expression.

## Discussion

In this study, we investigated genomic changes that occurred during evolution of the *Verticillium* genus from the last common ancestor to the currently recognized extant species. During evolution, two or more independent populations within a single species may diverge and slowly start to separate, which may ultimately lead to reproductive isolation and thus to speciation. Initially the separated species will have chromosomes that share the gene content (synteny) as well as the structure and order (co-linearity). Over time, the degree of synteny and co-linearity will degrade through various processes, including chromosomal rearrangements, segmental duplications, gene losses and gene gains until, ultimately, orthologous genes in one species occur randomly in the genome of the other.

We previously noted an unexpectedly high number of chromosomal rearrangements between strains of *V. dahliae* (de Jonge et al., 2013; Faino et al., 2015; Faino et al., 2016). We have speculated that the extent of rearrangements may be associated with the fact that, despite being asexual, *V. dahliae* is a successful broad host-range pathogen, reasoning that it would permit for the rapid adaptations that are required to be compatible in the arms race with host immune systems (Seidl and Thomma, 2014; Faino et al., 2016). From this hypothesis it would follow that the other species, being much less ubiquitous and successful pathogens, would not experience such drastic genomic rearrangements. However, our genomic reconstructions revealed that large-scale genomic rearrangements frequently occurred during evolution of the *Verticillium* genus. Moreover, our data seems to suggest that rearrangements occurred even more frequent in clade B that mostly harbors non-pathogenic species. This observation makes it unlikely that the occurrence of the rearrangements themselves is a major contributor to the pathogenicity of *V. dahliae*. Previously, we have shown that the highly variable LS regions of *V. dahliae* evolved by genomic rearrangements (de Jonge et al., 2013; Faino et al., 2016). We also showed that LS regions that harbor *in planta*-expressed effector genes arose by segmental duplications, likely generating the genetic material that subsequently has the freedom to diverge into novel functions (de Jonge et al., 2013; Faino et al., 2016). Thus, LS regions play important roles in adaption of *V. dahliae*. However, whether such LS regions also occur in other *Verticillium* spp. still needs to be investigated.

Besides extensive genomic rearrangements, conspicuous gene losses occurred during *Verticillium* evolution. This suggests that gene losses contribute to genetic divergence across the various *Verticillium* lineages. The impact of gene loss on evolution and pathogenicity is perhaps most evident for obligate biotrophic plant pathogens whose growth and reproduction entirely depends on living plant cells, such as the powdery mildews (Spanu et al., 2010). Their genomes show massive genome-size expansion, owing to retrotransposon proliferation, and extensive gene losses concerning enzymes

of primary and secondary metabolism that are responsible for loss of autotrophy and dependence on host plants in an exclusively biotrophic life-style (Spanu et al., 2010). Recently, the evolutionary history of lineages of grass powdery mildew fungi that have a rather peculiar taxonomy with only one described species (*Blumeria graminis*) that is separated into various *formae speciales* (*ff. spp.*) was reconstructed. It was observed that different processes shaped the diversification of *B. graminis*, including co-evolution with the host species for some of the *ff. spp.*, host jumps, host range expansions, lateral gene flow and fast radiation (Menardo et al., 2017). Arguably, as highly adapted and obligate biotrophic pathogens, powdery mildews underwent many host species-specific adaptations, which are perhaps not or less required for saprophytes and facultative and broad host range pathogens such as *Verticillium* species. Among the lost gene families in *Verticillium* species, we found that enriched Pfam domains relate to various functions, suggesting that the lost gene families of *Verticillium* spp. do not revolved around a particular function.

Functions of species-specific genes are often found to be associated to species-specific adaptations to a certain environment (Domazet-Loso and Tautz, 2003). For example, morphological and innate immune differences among *Hydra* spp. are controlled by species-specific genes (Khalturin et al., 2008; Khalturin et al., 2009). In plant pathogens, many effectors characterized so far are species-specific and facilitate virulence on a particular host plant (de Jonge et al., 2011) However, these species-specific effectors are frequently mutated, or even purged, in order to overcome host recognition (Cook et al., 2015). Moreover, it has been shown that compared to shared genes, species-specific genes are frequently shorter (Lipman et al., 2002), evolving quicker and less expressed (Domazet-Loso and Tautz, 2003; Plissonneau et al., 2016). Consistent with this observation, our results show that species-specific genes differ in their characteristics when compared with core genes. It has often been claimed that these distinct characteristics are hallmarks of gene structure degeneration and that these species-specific genes are more likely to go extinct (Palmieri et al., 2014; Plissonneau et al., 2016).

## Materials & Methods

### Genome sequencing and assembly

DNA was isolated from mycelium of 3-day-old cultures grown in potato dextrose broth (PDB) at 28°C as described previously (Seidl et al., 2015). Of each strain, two libraries (500 bp and 5 Kb insert size) were sequenced using the Illumina High-throughput sequencing platform (KeyGene N.V., Wageningen, The Netherlands). Genome assemblies were performed using the A5 pipeline (Tritt et al., 2012), and sequence gaps were filled using SOAP GapCloser (Luo et al., 2012). Next, QCAST (Gurevich et al., 2013) was used to



calculate genome statistics. Illumina sequence reads and assemblies were deposited in NCBI (Bioproject PRJNA392396).

### Gene prediction and annotation

Protein-coding genes were *de novo* annotated with the Maker2 pipeline (Holt and Yandell, 2011) using 35 predicted fungal proteomes and the previously annotated proteomes of *V. dahliae* and *V. tricorpus* to guide gene structure annotation (Faino et al., 2015; Seidl et al., 2015). Secretome prediction was described previously (Seidl et al., 2015).

### Ortholog analysis and tree building

Orthologous groups were determined using OrthoMCL (Li et al., 2003). The species phylogenetic tree was generated using 5,228 single-copy orthologs that are conserved among all of the genomes. Individual families were aligned using mafft (LINSi; v7.04b) (Katoh et al., 2002) and subsequently concatenated. Maximum likelihood phylogeny was inferred using RAxML (v8.2.4) with the GAMMA model of rate heterogeneity and the Whelan and Goldman (WAG) model of amino acid substitutions (Stamatakis, 2014). The robustness of the inferred phylogeny was assessed by 100 rapid bootstrap approximations.

### Mitochondrial genome assembly and comparison

Sequencing reads of mitochondrial genomes of each *Verticillium* species were extracted from raw paired-end reads using GRABb using 182 already published fungal mitochondrial genome sequences as bait (Brankovics et al., 2016). Subsequently, GRABb assembled all the mitochondrial reads into single circular sequences. To determine the mitochondrial protein-coding genes of each species, we queried the already published *Verticillium dahliae* mitochondrial protein sequences using BLAST (tblastn) against each mitochondrial genome. Whole mitochondrial genome alignment was performed with mafft (default setting) (Katoh et al., 2002), and the likelihood phylogenetic tree was built using RAxML (v7.6.3) with the GAMMA model of rate heterogeneity and the GTR model of nucleotide substitutions (Stamatakis, 2014). The robustness of the inferred phylogeny was assessed by 100 rapid bootstrap approximations.

### Ancestor genome reconstruction and comparison

Ancestral genomes were constructed by using CHRONicle package that comprises SynChro (Drillon et al., 2014), ReChro and Anchro (Vakirlis et al., 2016). Conserved synteny blocks were identified between pairwise combinations of genomes with SynChro (Drillon et al., 2014). The synteny block stringency delta, which determines the maximum number of intervening Reciprocal Best Hits (RBH) allowed between anchors within a synteny block

was set to three. Subsequently, ReChro was used for estimating genome rearrangements from ancestral to extant genomes.

### Gene family gains and losses

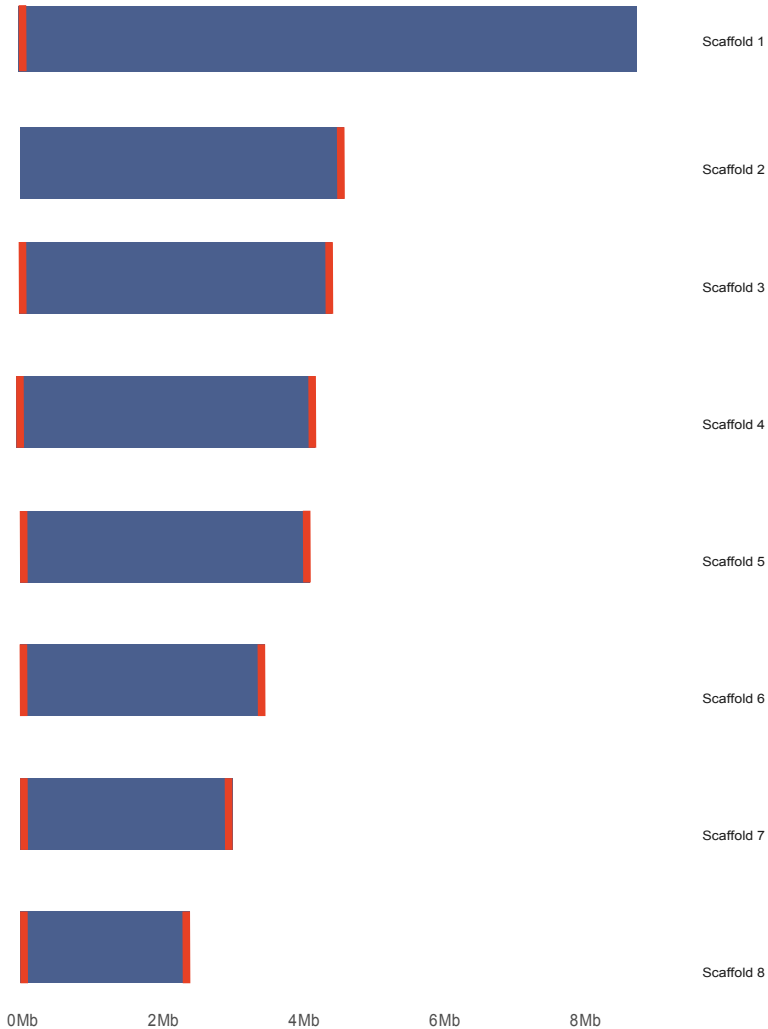
Gene family gain/loss analysis by inferring ancestral gene number counts was carried out using CAFE (Han et al., 2013). Additionally, we inferred gene family gains and losses by reconciliation of gene trees. The sequences of the gene families (>3 members) were aligned using mafft with default settings (Katoh et al., 2002). We constructed phylogenetic trees for each gene family using RAxML (v8.2.4) with the GAMMA model of rate heterogeneity and the Whelan and Goldman (WAG) model of amino acid substitutions (Stamatakis, 2014). Subsequently, we reconciled the gene family trees with the *Verticillium* species tree using NOTUNG (Chen et al., 2000). The trees were reconciled using a cost of 1.5 for a duplication event and 1 for a loss event. The trees were rooted so that the number of duplication and loss events is minimized. Cluster of Orthologous Groups (COG) functional categories were predicted using eggNOG (Huerta-Cepas et al., 2015). Pfam function domains were predicted using InterProScan (Jones et al., 2014). For analyses of Pfam domain associated with gene families (OrthoMCL clusters), Pfam domains were assigned to specific families (clusters) only when present in at least half of the homologs within each gene family. Pfam enrichment of gene families of interests was carried out using hypergeometric tests, and significance values were corrected using the Benjamini-Hochberg false discovery method (Benjamini and Hochberg, 1995).

### Acknowledgements

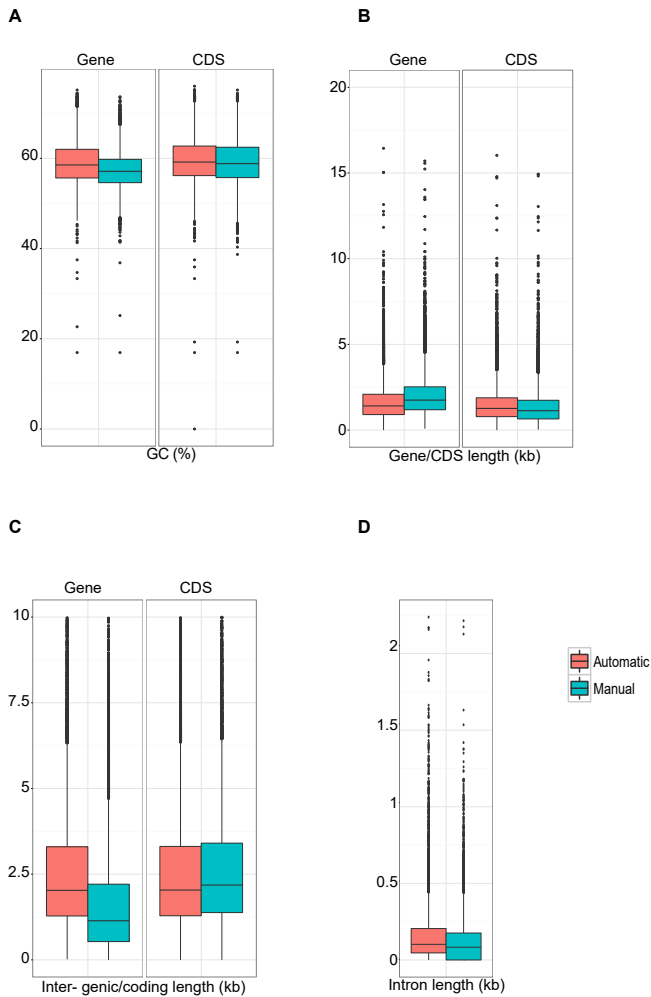
We thank Drs Inderbitzin and Subbarao for sharing fungal isolates and Sander Rodenburg for help with scripting.

## Supplementary material

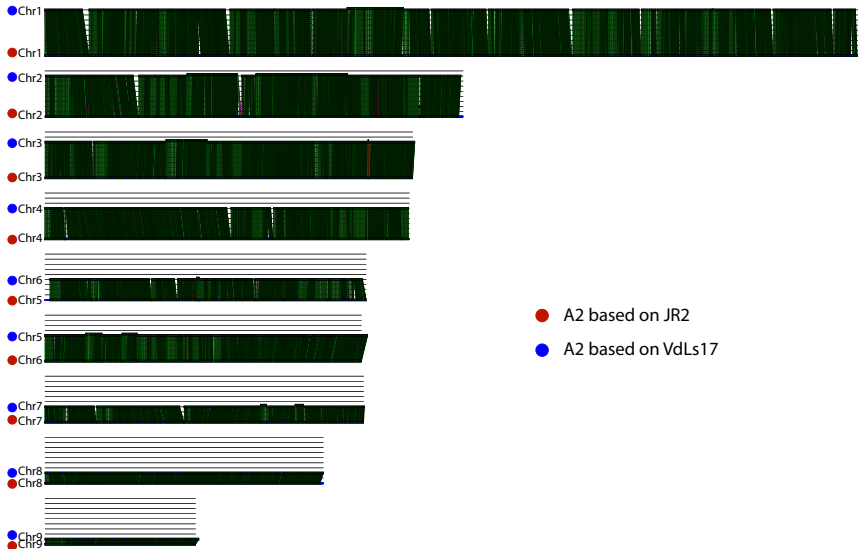
Due to the amount of data, some of the supplementary files, Supplementary Data S1, Data S2 and Data S3 are only accessible online at *Environmental Microbiology* (<https://onlinelibrary.wiley.com/doi/full/10.1111/1462-2920.14037>).



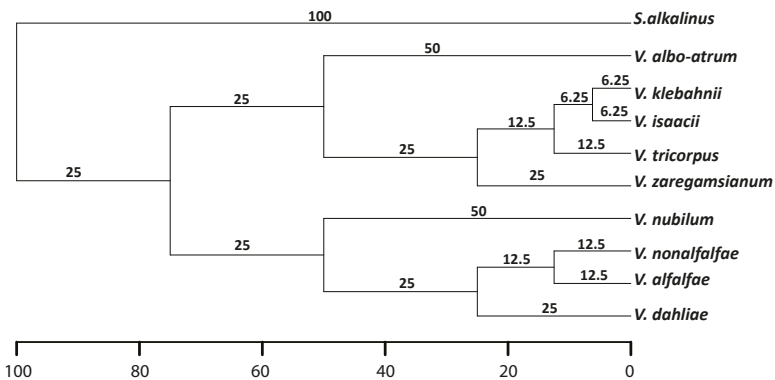
**FIGURE S1. Overview of the draft genome assembly of *V. tricorpus*, strain PD593.** Schematic representation of the eight largest scaffolds in the genome assembly of *V. tricorpus* strain PD593. Characteristic fungal telomeric repeats are displayed on the ends of the scaffolds (indicated by red color).



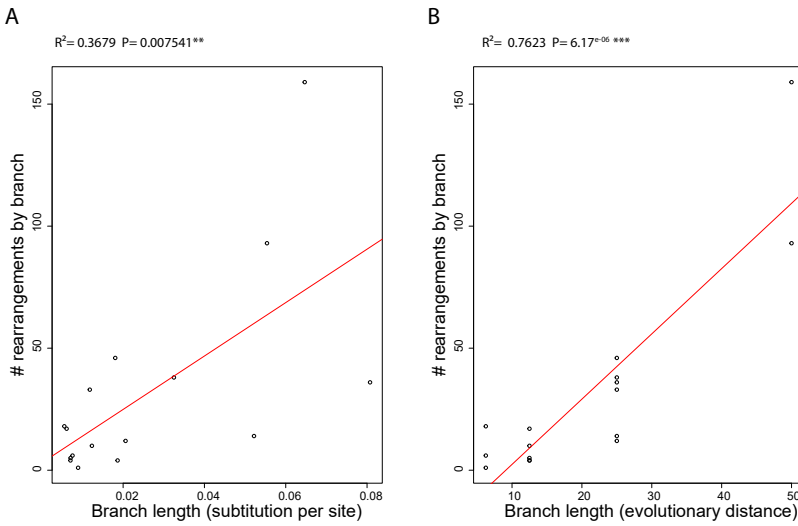
**FIGURE S2. Gene feature comparisons between automatically and manually annotated genes of *V. dahliae* strain JR2.** (A) GC content. (B) Gene/coding sequence length. (C) Intergenic length. D Intron length.



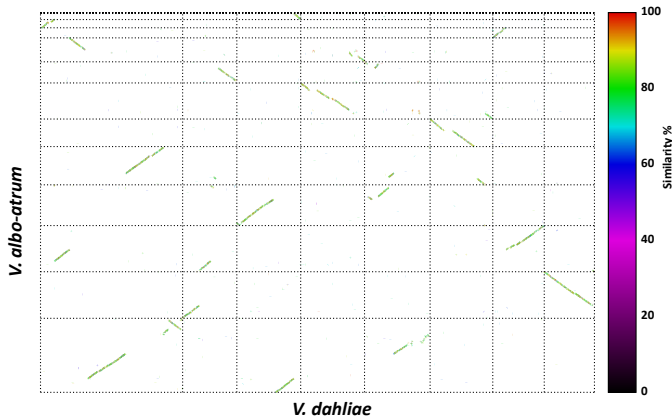
**FIGURE S3. Chromosome alignments between two ancestor genomes.** Two ancestor genomes were reconstructed based on the genome of two *V. dahliae* strain JR2 and VdLs17 respectively (indicated by red and blue dots respectively). Green and red plain lines highlight homology relationships with same and inverted directions, respectively.



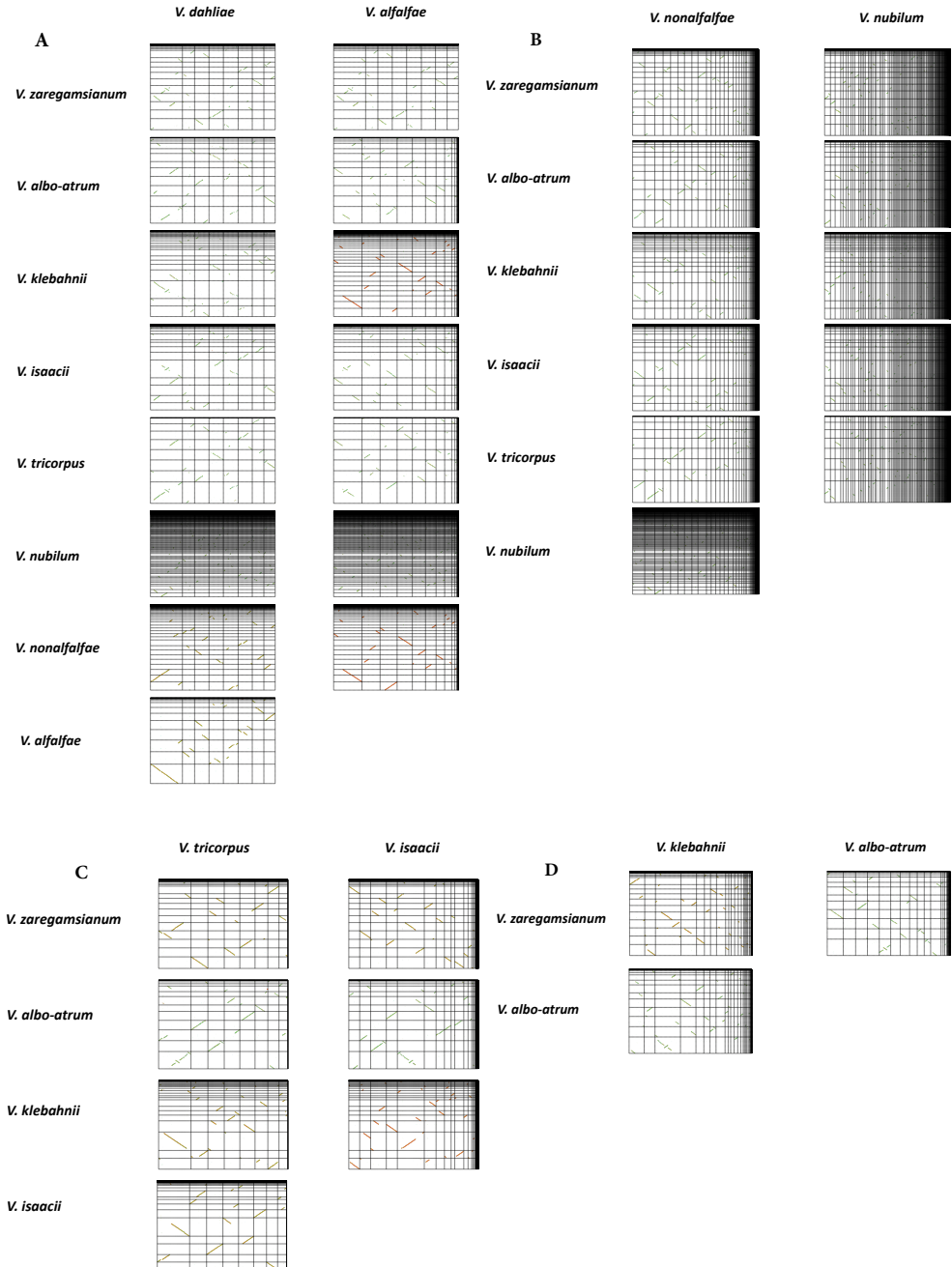
**FIGURE S4. Ultrametric phylogeny of *Verticillium* species and *S. alkalinus*.** The ultrametric tree derived by a maximum likelihood analysis of concatenated single-copy orthologs (Figure 1). Averaged divergence times per branch are reported relative to an arbitrary age of the last common ancestor of *Verticillium* spp. and *S. alkalinus* (set at 100).



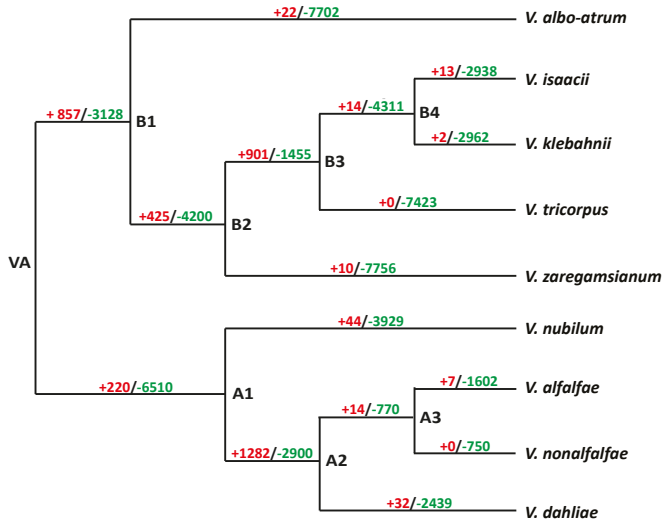
**FIGURE S5. Correlations between the branch lengths and number of rearrangements.** (A) Branch length is represented by number of substitutions per site (Figure 1), (B) Branch length is represented by evolutionary distance (Figure 3).



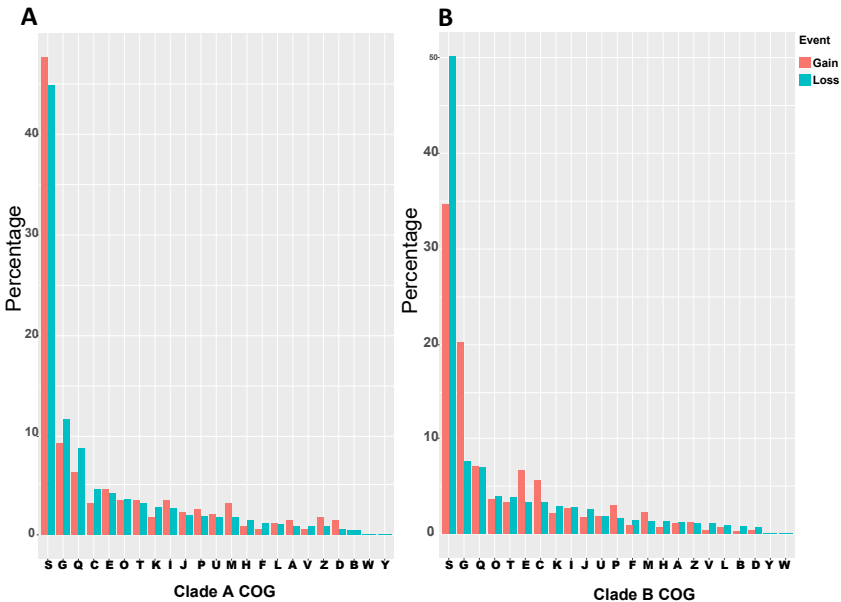
**FIGURE S6. Dot-plot comparison between *V. dahliae* and *V. albo-atrum*.** The six-frame translations of both genomes were compared using Promer (Mummer 3.0). Homologous regions are plotted as dots and are color coded for percentage similarity. Macro- or micro-synteny is indicated by long or short diagonal lines with a slope that is positive or negative depending on whether the genes align in the same or inverted order.



**FIGURE S7. A-D Pairwise Dot-plot comparison between nine haploid *Verticillium* species.** Macro- or micro-synteny is indicated by long or short diagonal lines with a slope that is positive or negative depending on whether the genes align in the same or inverted order.



**FIGURE S8. Evolution of *Verticillium* gene repertoire.** Numbers of expanded (in red) and contracted (in green) gene families were inferred by a tree reconciliation approach



- [D] Cell cycle control, cell division, chromosome partitioning
- [M] Cell wall/membrane/envelope biogenesis
- [N] Cell motility
- [O] Post-translational modification, protein turnover, and chaperones
- [T] Signal transduction mechanisms
- [U] Intracellular trafficking, secretion, and vesicular transport



[V] Defense mechanisms  
 [W] Extracellular structures  
 [Y] Nuclear structure  
 [Z] Cytoskeleton  
 [A] RNA processing and modification  
 [B] Chromatin structure and dynamics  
 [J] Translation, ribosomal structure and biogenesis  
 [K] Transcription  
 [X] Replication, recombination and repair  
**[C]** Energy production and conversion  
 [E] Amino acid transport and metabolism  
 [F] Nucleotide transport and metabolism  
 [G] Carbohydrate transport and metabolism  
 [H] Coenzyme transport and metabolism  
 [I] Lipid transport and metabolism  
 [P] Inorganic ion transport and metabolism  
 [Q] Secondary metabolites biosynthesis, transport, and catabolism  
 [R] General function prediction only  
 [S] Function unknown

**FIGURE S9. Percentages for each COG category of gained or lost gene families.** (A) Clade A, (B) Clade B.

**TABLE S1. Minimum number of scaffolds that comprises at least 95% of total proteomes of each *Verticillium* species.**

Species	Strain name	# scaffolds	% total proteomes
<i>V. dahliae</i>	JR2	8	100.0
<i>V. alfalfae</i>	PD683	9	97.4
<i>V. nonalfalfae</i>	TAB2	23	95.6
<i>V. nubilum</i>	PD621	102	95.4
<i>V. albo-atrum</i>	PD747	10	96.1
<i>V. tricorpus</i>	PD593	8	99.9
<i>V. isaacii</i>	PD618	12	95.9
<i>V. klebahnii</i>	PD401	17	95.8
<i>V. zaregamsianum</i>	PD739	12	95.3

**TABLE S2. Descriptions of the enriched Pfam domains.**

<b>Pfam ID</b>	<b>Pfam name</b>	<b>Description</b>	<b>Gain/Loss</b>
PF13802	Gal_mutarotas_2	Galactose mutarotase-like	Gain
PF00171	Aldedh	Aldehyde dehydrogenase family	Gain
PF01301	Glyco_hydro_35	Glycoside hydrolase family 35	Gain
PF10435	BetaGal_dom2	Beta-galactosidase, domain 2	Gain
PF13363	BetaGal_dom3	Beta-galactosidase, domain 3	Gain
PF13364	BetaGal_dom4_5	Beta-galactosidase jelly roll domain	Gain
PF01915	Glyco_hydro_3_C	Glycosyl hydrolase family 3 C-terminal domain	Gain
PF14310	Fn3-like	Fibronectin type III-like domain	Gain
PF00083	Sugar_tr	Sugar (and other) transporter	Gain
PF00083	Sugar_tr	Sugar (and other) transporter	Gain
PF00083	Sugar_tr	Sugar (and other) transporter	Gain
PF06441	EHN	Epoxide hydrolase N terminus	Loss
PF01636	APH	Phosphotransferase enzyme family	Loss
PF05368	NmrA	NmrA-like family	Loss
PF04479	RTA1	RTA1 like protein	Loss
PF06985	HET	Heterokaryon incompatibility protein	Loss
PF08240	ADH_N	Alcohol dehydrogenase GroES-like domain	Loss
PF00106	adh_short	short chain dehydrogenase	Loss
PF00107	ADH_zinc_N	Zinc-binding dehydrogenase	Loss
PF04082	Fungal_trans	Fungal specific transcription factor domain	Loss
PF00067	p450	Cytochrome P450	Loss
PF07690	MFS_1	Major Facilitator Superfamily	Loss
PF00083	Sugar_tr	Sugar (and other) transporter	Loss
PF00172	Zn_clus	Fungal Zn(2)-Cys(6) binuclear cluster domain	Loss
PF12796	Ank_2	Ankyrin repeats (3 copies)	Loss

GO term	P value	Branch
carbohydrate metabolic process, GO:0005975	4,00E-02	<i>V. albo-atrum</i>
metabolic process, GO:0008152; oxidoreductase activity, GO:0016491; oxidation-reduction process, GO:0055114	4,00E-02	<i>V. albo-atrum</i>
hydrolase activity; hydrolyzing O-glycosyl compounds, GO:0004553; carbohydrate metabolic process, GO:0005975	3,00E-02	B2
metabolic process, GO:0008152; oxidation-reduction process, GO:0055114; oxidoreductase activity, GO:0016491	3,00E-02	B2
metabolic process, GO:0008152; oxidation-reduction process, GO:0055114; oxidoreductase activity, GO:0016491	3,00E-02	B2
metabolic process, GO:0008152; oxidation-reduction process, GO:0055114; oxidoreductase activity, GO:0016491	3,00E-02	B2
hydrolase activity; hydrolyzing O-glycosyl compounds, GO:0004553; carbohydrate metabolic process, GO:0005975	3,00E-02	B2
NA	3,00E-02	B2
transmembrane transporter activity, GO:0022857; transmembrane transport, GO:0055085; integral component of membrane, GO:0016021	4,00E-02	<i>V. albo-atrum</i>
transmembrane transporter activity, GO:0022857; transmembrane transport, GO:0055085; integral component of membrane, GO:0016021	3,00E-02	B2
transmembrane transporter activity, GO:0022857; transmembrane transport, GO:0055085; integral component of membrane, GO:0016021	7,00E-03	<i>V. zaregamsianum</i>
NA	4,00E-02	<i>V. tricorpus</i>
NA	9,00E-05	<i>V. isaacii</i>
NA	4,00E-06	A1
response to stress, GO:0006950; integral component of membrane, GO:0016021	2,00E-03	A1
NA	9,00E-05	A1
oxidation-reduction process, GO:0055114	8,00E-03	A1
NA	2,00E-05	A1
oxidation-reduction process, GO:0055114	3,00E-02	A1
DNA binding, GO:0003677; zinc ion binding, GO:0008270; transcription; DNA- templated, GO:0006351; nucleus, GO:0005634	1,00E-06	A1
iron ion binding, GO:0005506; oxidoreductase activity; acting on paired donors; with incorporation or reduction of molecular oxygen, GO:0016705; heme binding, GO:0020037; oxidation-reduction process, GO:0055114	2,00E-02	A1
transmembrane transport, GO:0055085; integral component of membrane, GO:0016021	1,00E-06	A1
transmembrane transporter activity, GO:0022857; transmembrane transport, GO:0055085; integral component of membrane, GO:0016021	2,00E-02	A1
RNA polymerase II transcription factor activity; sequence-specific DNA binding, GO:0000981; zinc ion binding, GO:0008270; regulation of transcription; DNA-templated, GO:0006355; nucleus, GO:0005634	2,00E-03	A1
NA	5,00E-03	B1

## Chapter 4

**TABLE S3. Numbers of species-specific genes and numbers of these genes that have homologs in other fungal species.**

<b>Species</b>	<b>Strain name</b>	<b># species-specific genes</b>	<b># species-specific genes have homologs in other fungi</b>
<i>V. dahliae</i>	JR2	564	139
<i>V. alfalfae</i>	PD683	483	50
<i>V. nonalfalfae</i>	TAB2	406	47
<i>V. nubilum</i>	PD621	628	359
<i>V. albo-atrum</i>	PD747	538	362
<i>V. tricorpus</i>	PD593	271	116
<i>V. isaacii</i>	PD618	276	138
<i>V. klebahnii</i>	PD401	278	126
<i>V. zaregamsianum</i>	PD739	585	383

# Chapter 5

## **Dynamic virulence-related regions of the plant pathogenic fungus *Verticillium dahliae* display enhanced sequence conservation**

Jasper R.L. Depotter\*, Xiaoqian Shi-Kunne\*, H el ene Missonnier†, Tingli Liu†, Luigi Faino, Grardy C.M. van den Berg, Thomas A. Wood, Baolong Zhang‡, Alban Jacques‡, Michael F. Seidl#, Bart P.H.J. Thomma#

\* These authors contributed equally to this work.

† These authors contributed equally to this work.

‡ These authors contributed equally to this work.

# These authors contributed equally to this work.

## Abstract

Plant pathogens continuously evolve to evade host immune responses. During host colonization, pathogens secrete effectors to perturb such responses, but these may become recognized by host immune receptors in turn. To facilitate effector repertoire modifications, such as the elimination of recognized effectors, effector genes often reside in genomic regions that display increased plasticity, a phenomenon that is captured in the two-speed genome hypothesis. The genome of the vascular wilt fungus *Verticillium dahliae* carries plastic, lineage-specific, regions that are enrichment in *in planta*-induced effector genes. As expected, comparative genomics reveals differential degrees of sequence divergence between lineage-specific regions and the core genome. However, unanticipated, lineage-specific regions display markedly higher sequence conservation if orthologous sequences are found in other *Verticillium* species. Intriguingly, also non-coding regions within the lineage-specific regions show enhanced sequence conservation. We provide evidence that disqualifies horizontal transfer to explain the observed sequence conservation and conclude that that sequence divergence occurs at a slower pace in lineage-specific regions of the *V. dahliae* genome. We hypothesize that differences in chromatin organisation may explain lower rates of SNP frequencies that occur in the plastic LS regions of *V. dahliae*.

**Key words:** comparative genomics; effector; genome evolution; mutagenesis; two-speed genome; *Verticillium* wilt

## Introduction

Numerous microbes engage in symbiotic relationships with plants, comprising beneficial, commensalistic and parasitic relationships where each partner evolves towards its optimal fitness (Jones and Dangl, 2006). In their interactions with host plants, pathogens evolve repertoires of effector proteins, many of which deregulate host immunity, to enable host colonization (Dodds and Rathjen, 2010). Plants, in turn, evolve immune receptors that recognize various molecular patterns that betray microbial invasion; so-called invasion patterns that can also include effectors (Cook et al., 2015). Consequently, pathogen effector repertoires are typically subject to selective forces that often result in rapid diversification.

Effector genes are often not randomly organized in genomes of filamentous plant pathogens (Dong et al., 2015). For instance, effector genes of the potato late blight pathogen *Phytophthora infestans* reside in repeat-rich regions that display increased structural polymorphisms and enhanced levels of positive selection (Haas et al., 2009; Raffaele et al., 2010). Consequently, it has been proposed that many pathogens have a bipartite genome architecture with household genes residing in a conserved core genome and effector genes in dynamic and repeat-rich compartments; a “two-speed” genome (Croll and McDonald, 2012; Raffaele and Kamoun, 2012). Often, repeat-rich genome regions display signs of such accelerated evolution with structural variations such as presence/absence polymorphisms (Raffaele et al., 2010) or chromosomal rearrangements (de Jonge et al., 2013; Faino et al., 2016). Furthermore, such regions can also display increased substitution rates (Cuomo et al., 2007; van de Wouw et al., 2010), including increased levels of non-synonymous substitutions (Raffaele et al., 2010; Stukenbrock et al., 2010; Sperschneider et al., 2015).

*Verticillium* is a genus of soil-borne Ascomycete fungi containing notorious plant pathogens of numerous crops (Inderbitzin and Subbarao, 2014) that infect their hosts via the roots and then colonize xylem vessels, resulting in vascular occlusion and wilt disease (Fradin and Thomma, 2006). Currently, ten *Verticillium* species are described, which are divided in two phylogenetic clusters, i.e. clade Flavexudans and clade Flavnonexudans (Inderbitzin et al., 2011a). *Verticillium* spp. are thought to have a predominant, if not exclusive, asexual reproduction as a sexual cycle has never been described for any of the species (Short et al., 2014). Nevertheless, mechanisms different from meiotic recombination contribute to the genomic diversity of *Verticillium dahliae*, the most notorious plant pathogen with the genus that infects hundreds of plant species (Inderbitzin and Subbarao, 2014), including large-scale genomic rearrangements, horizontal gene transfer and transposable element (TE) activity (de Jonge et al., 2012; de Jonge et al., 2013; Seidl and Thomma, 2014; Faino et al., 2016). These mechanisms converge on lineage-specific (LS) regions that are enriched in TEs and in *in planta*-induced effector genes (Klosterman et al., 2011; de Jonge et al.,

2013; Thomma et al., 2016).

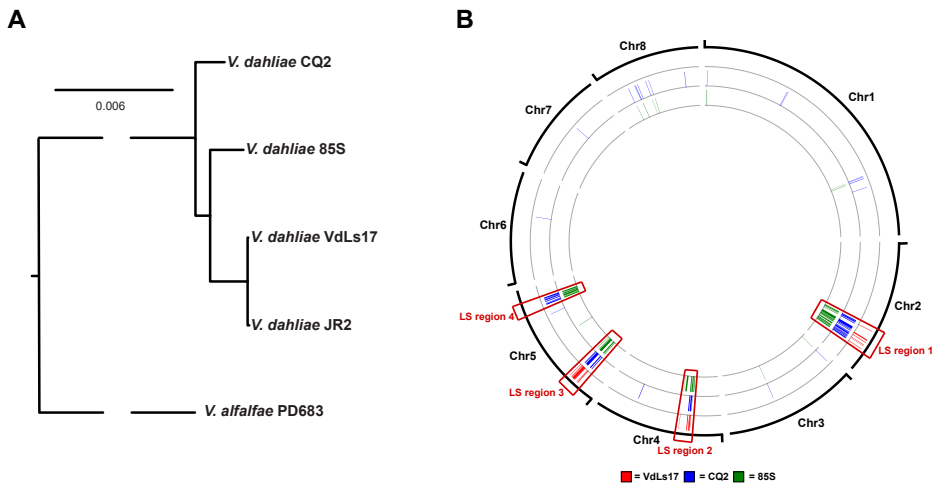
We previously reported that LS regions of *V. dahliae* are largely derived from segmental duplications (Faino et al., 2016) that are known as important sources for functional diversification (Magadum et al., 2013). Here we exploited comparative genomics across the *Verticillium* genus to identify differential rates of sequence diversification to further characterize the two-speed genome of *V. dahliae*.

## Results

### LS sequences reside in four regions of the genome of *V. dahliae* strain JR2

Previously, four LS regions were characterized for *V. dahliae* strain JR2; one on chromosome 2 and 4, and two on chromosome 5 (Faino et al., 2015; Faino et al., 2016). Lineage-specific regions in *V. dahliae* isolate JR2 are enriched for sequences that lack synteny to various other *V. dahliae* strains, including the completely sequenced genome of strain VdLs17 (Faino et al., 2015; Faino et al., 2016). Since *V. dahliae* strains JR2 and VdLs17 only recently diverged (de Jonge et al., 2013), we sequenced two *V. dahliae* strains that diverged earlier from JR2 (Figure 1A), namely strains CQ2 and 85S isolated from cotton in China and sunflower in France, resulting in assemblies of 35.8 and 35.9 Mb in 17 and 40 contigs, respectively (Table S1). Subsequent alignment revealed that the JR2 genome was not covered by 2.0%, 7.1% and 6.6% of sequences from VdLs17, CQ2 and 85S respectively, while 1.4% of the JR2 genome sequence could not be identified in any of the three other *V. dahliae* strains. The vast majority (88%, 82% and 91% for VdLs17, CQ2 and 85S, respectively) of the JR2 sequences without alignment localize in the previously identified four LS regions that collectively contain 476 genes (Figure 1B). Thus, despite the addition of more diverged *V. dahliae* strains, intraspecific presence/absence polymorphisms converge on the four previously identified genomic regions that are thus significantly more dynamic than other parts of the genome.





**FIGURE 1. Locations of lineage-specific (LS) regions in the genome of *V. dahliae* strain JR2.** (A) The phylogenetic relationship between *V. dahliae* strains JR2, VdLs17, CQ2 and 85S is shown. The phylogenetic relationship of all *V. dahliae* strains was inferred by single-copy orthologs. *V. alfalfae* was used as out-groups species. (B) *V. dahliae* strain JR2 LS regions were determined by individual comparisons to *V. dahliae* strains VdLs17 (red), CQ2 (blue) and 85S (green). Sequences of minimum 7.5 kb without an alignment to at least one of the other isolates are depicted in color at their respective position on the *V. dahliae* strain JR2 genome.

### LS regions share increased sequence identity to other *Verticillium* spp.

Next, we extended our analysis to other *Verticillium* spp. as well. While most of the *V. dahliae* strain JR2 genome aligns with *V. nonalfalfae* strain TAB2 with an average sequence identity of ~92%, particular regions display increased sequence identity, even up to 100% (Figure S1). Intriguingly, these regions co-localize with LS regions (Faino et al., 2015; Faino et al., 2016), suggesting that these regions in *V. dahliae* JR2 are derived from a recent horizontal transfer. Thus, we performed interspecific comparisons with all other haploid *Verticillium* spp. To this end, genomic sequences of *V. dahliae* were aligned in windows of 500 bp to the other *Verticillium* spp., excluding repetitive regions, displaying median identities ranging from 88 to 95% that correspond to the phylogenetic distances to *V. dahliae* (Table 1). Sequence identities were similarly calculated in windows for the LS regions. Intriguingly, the LS regions displayed significantly increased sequence identities when compared with the core genome, ranging from 92.3% median sequence identity for *V. zaregamsianum*, one of the phylogenetically most distantly related species, to 100% for the most closely related species *V. alfalfae* and *V. nonalfalfae* (Figure 2A, Table 1).

**TABLE 1. Sequence identities between *V. dahliae* strain JR2 and other haploid *Verticillium* species (excluding repetitive regions).**

Species/strain	Genome-wide <sup>†</sup> (%)	LS regions <sup>†</sup> (%)	# windows aligned to LS regions <sup>‡</sup>	<i>p</i> -value <sup>§</sup>
<i>V. albo-atrum</i> /PD747	88.8 (4.5)	97.6 (3.7)	146	2.2e-16
<i>V. alfalfae</i> /PD683	94.6 (3.9)	100.0 (3.5)	465	2.2e-16
<i>V. nonalfalfae</i> /TAB2	94.8 (4.1)	100.0 (2.9)	1037	2.2e-16
<i>V. nubilum</i> /PD621	88.4 (3.7)	95.6 (4.3)	144	2.2e-16
<i>V. tricorpus</i> /PD593	89.0 (4.3)	97.2 (4.1)	162	2.2e-16
<i>V. isaacii</i> /PD660	88.8 (4.5)	98.6 (4.4)	189	2.2e-16
<i>V. klebahnii</i> /PD401	88.8 (4.0)	97.8 (5.3)	87	2.2e-16
<i>V. zaregamsianum</i> /PD739	88.8 (3.6)	92.3 (4.2)	38	1.56e-5

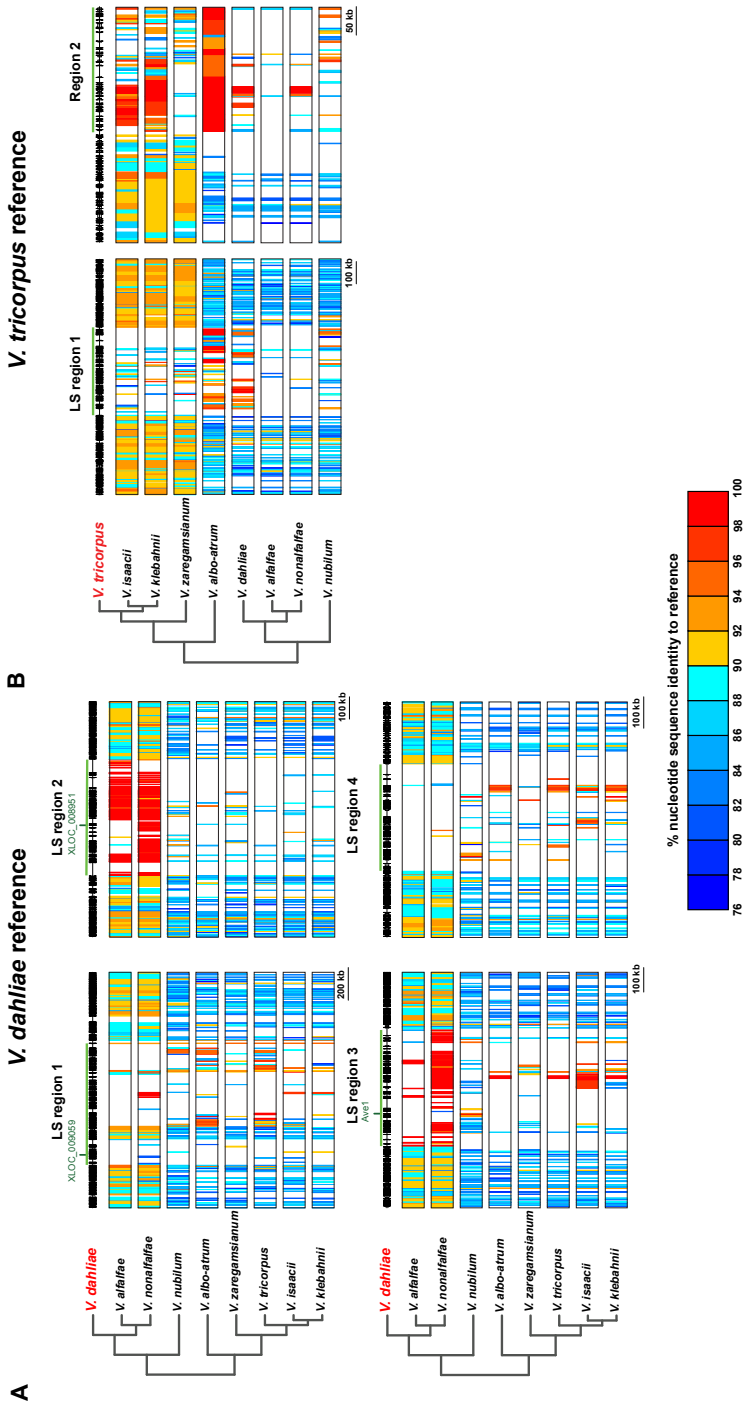
<sup>†</sup>The percentage is the median sequence identity and the number between the brackets is the standard deviation.

<sup>‡</sup>Windows of 500 bp

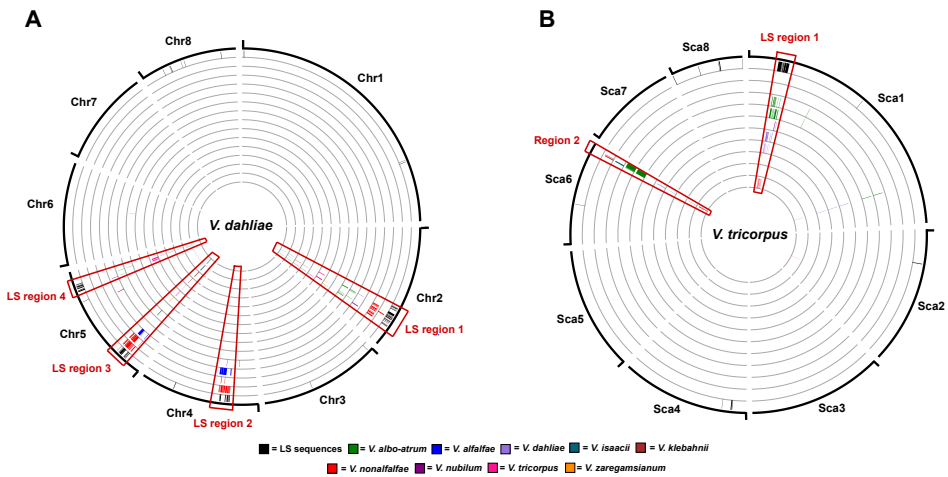
<sup>§</sup>The *p*-value was calculated with a two-sided Wilcoxon rank-sum test

To assess whether high interspecific sequence identity concerns only LS regions, we aligned *Verticillium* sequences of high identity to the complete *V. dahliae* JR2 genome. For several species we used multiple strains at this stage. Nearly all (99–100%) of the *V. alfalfae* and *V. nonalfalfae* sequences that display >96% identity to *V. dahliae* strain JR2 sequences localized in LS regions (Figure 3A). Similarly, sequences of at least 100 kb with >90% identity of other *Verticillium* spp. mapped to *V. dahliae* strain JR2 LS regions, ranging from 70% of the sequences in *V. nubilum* PD621 to 95% in *V. albo-atrum* PD670 and *V. tricorpus* PD593 (Table S3). Thus, high interspecific sequence identity is specifically associated with *V. dahliae* LS regions.

Finally, we constructed phylogenetic trees for particular LS region genes, obviously with the limitation that only few LS genes are present in multiple species. Nevertheless, this analysis revealed that Flavexudans and Flavonoxudans orthologs clustered separately from each other (Figure S2), confirming that the phylogenies of LS genes correspond to the overall species phylogeny. Overall, the genus-wide occurrence of sequence coverage combined with a differential degree of sequence correspondence that reflects the phylogenetic distance to *V. dahliae* strongly suggests that the content of the LS regions does not originate from horizontal transfer events. This hypothesis would require multiple sequential horizontal transfer events in the exact order of decreasing phylogenetic distance, which we deem extremely unlikely.



**FIGURE 2. Interspecific alignments and sequence identity within and immediately adjacent to regions with high interspecific sequence identity.** The green line indicates regions with high sequence identity. Coloured blocks are orthologous sequences to (A) *V. dahliae* and (B) *V. tricorpus* with the colour indicating the sequence identity. The black, vertical stripes represent gene positions of the reference strains. Locations of characterized *V. dahliae* effector genes are indicated: Ave1, XLOC\_008951 and XLOC\_009059 (de Jonge *et al.* 2012, 2013). Strains used in this figure are *V. tricorpus* PD593, *V. isaacii* PD660, *V. klebahnii* PD401, *V. zaregamstanum* PD739, *V. albo-atrum* PD747, *V. dahliae* JR2, *V. alfalfae* PD683, *V. nonalfalfae* TAB2 and *V. nubilum* PD621.

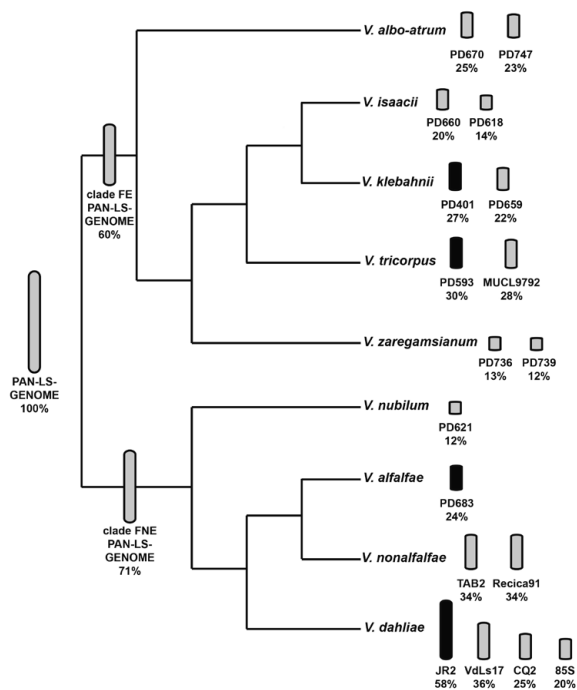


**FIGURE 3. Regions of particular high interspecific sequence identity.** All *Verticillium* strains mentioned in Table S2 were used. (A) Black bars correspond to lineage-specific (LS) sequences of *V. dahliae* strain JR2 (for details, see Figure 1B). Sequences ( $\geq 7.5$  kb) with high sequence identity in any of the other *Verticillium* spp. ( $\geq 96\%$  for *V. alfalfae* and *V. nonalfalfae*,  $\geq 90\%$  for all other *Verticillium* spp.) are plotted at the corresponding position on the genome of *V. dahliae* strain JR2. (B) The eight biggest scaffolds of *V. tricorpus* strain PD593 are depicted as these comprise over 99.5% of the genome. Black bars correspond to LS sequences ( $\geq 7.5$  kb) in the *V. tricorpus* strain PD593 genome without alignments to *V. tricorpus* strain MUCL9792. Sequences ( $\geq 7.5$  kb) with relatively high sequence identity in any of the other *Verticillium* spp. ( $\geq 96\%$  *V. isaacii*, *V. klebahnii* and *V. zaregamsianum*,  $\geq 90\%$  for all other *Verticillium* spp.) are plotted at the corresponding position on the genome of *V. tricorpus* strain PD593. Non-depicted *Verticillium* strains did not have sequences ( $\geq 7.5$  kb) with previously mentioned identity to *V. tricorpus* strain PD593.

### High interspecific sequence identity of LS regions is not unique to *V. dahliae*

To investigate whether other *Verticillium* spp. similarly carry LS regions that display high interspecific sequence identity, we performed alignments using *V. tricorpus* strain PD593 as a reference because of its high degree of completeness with seven of the nine scaffolds likely representing complete chromosomes (Table S2) (Shi-Kunne et al., 2018). Furthermore, this species belongs to the Flavexudans clade, in contrast to *V. dahliae* that belongs to Flavnonexudans. LS sequences of *V. tricorpus* strain PD593 were determined by comparison to *V. tricorpus* strain MUCL9792 (Seidl et al., 2015). In total, 98% of the PD593 genome could be aligned to MUCL9792. However, 48% of the sequences that are specific for *V. tricorpus* strain PD593 reside in only a single genomic region of 41 kb on scaffold 1 (Figure 3B). Like for *V. dahliae* strain JR2, sequences of other *Verticillium* spp. aligned with high identity to *V. tricorpus* PD593: *V. isaacii*, *V. klebahnii* and *V. zaregamsianum* display a median genome identity of  $\sim 95\%$ , while other haploid *Verticillium* spp. display  $\sim 88$ - $89\%$  median genome identity (Table S4). Notably, regions that display significantly higher sequence identity localized at the LS region on scaffold 1, but also to an additional region of 23 kb on scaffold 6 (Figure 2B, Figure 3B). Likely, this concerns an LS region that

could not be identified based on the two *V. tricorpus* strains in our analysis. For *Verticillium* strains with total alignments of at least 100 kb of high-identity sequences, the fraction of high-identity sequences that aligned to the scaffold 1 and 6 genome loci ranged from 49% for *V. nubilum* (PD621) up to 84% for *V. albo-atrum* (PD747) (Table S5). As expected, the sequence identity to six of the eight other haploid *Verticillium* spp. was significantly higher in these two genome loci compared to the genome-wide median (Table S4). No increase in sequence identity was found in alignments with *V. alfalfae* strain PD683 and *V. zaregamsianum* strain PD739 as only few regions with high sequence identity could be aligned (Table S5). Strains PD683 and PD739 only aligned two and 37 windows of 500 bp, respectively, to the scaffold 1 and 6 loci (Figure 2B, Table S4). Like for *V. dahliae*, the phylogenies of orthologs to *V. tricorpus* LS region genes separate clade Flavexudans and Flavnonexudans genes (Figure S3). In conclusion, LS regions with high interspecific sequence identity that diverge with phylogenetic distance do not only occur in *V. dahliae* but also in *V. tricorpus*.



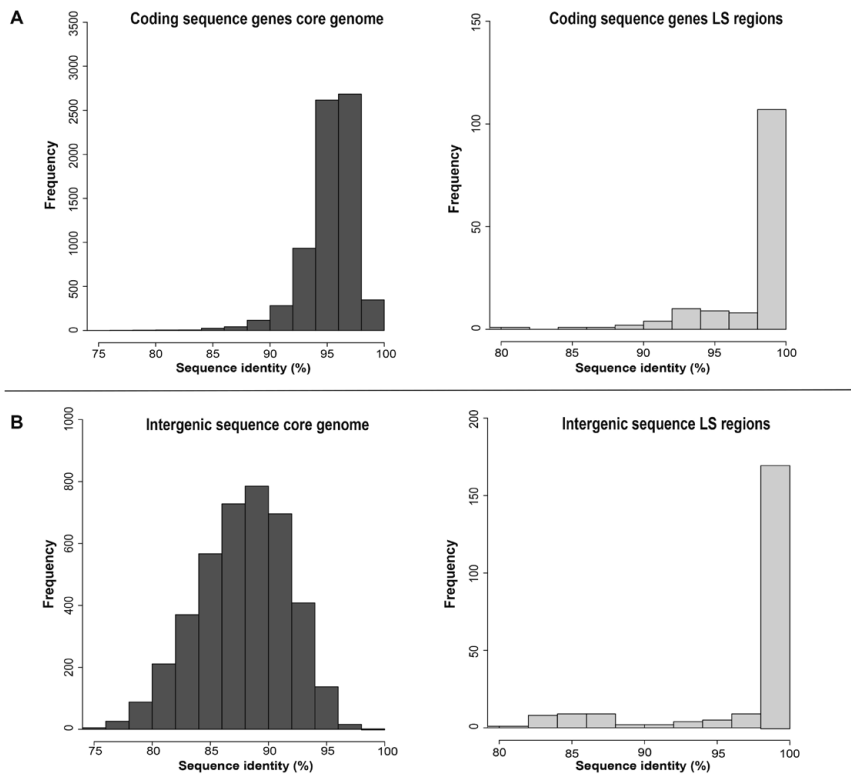
**FIGURE 4. Diversity of pan-LS-genome contents across the *Verticillium* genus.** A pan-LS-genome was constructed based on sequences from *V. dahliae* JR2, *V. alfalfae* PD683, *V. tricorpus* PD593 and *V. klebahnii* PD401 (black bars). The bar size next to the species names in the *Verticillium* phylogenetic tree is representative for the amount of the pan-LS-genome that is present in the individual isolates. All isolates of the clade Flavexudans (FE) in this study were used to calculate the percentage of the pan-LS-genome that is present in clade Flavexudans. Similarly, the portion of the Flavnonexudans (FNE) in the pan-LS-genome was calculated with all isolates of the clade Flavnonexudans used in this study.

### Pan-LS-genome distribution across the *Verticillium* genus

Considering that horizontal transfer is unlikely, the high sequence identity between *Verticillium* LS sequences indicates that their origin is ancestral and predates speciation. Hence, we constructed a pan-LS-genome to determine the distribution of conserved sequences across the *Verticillium* genus. To compose a pan-LS-genome, we combined regions with high interspecific sequence identity of four *Verticillium* spp., namely *V. dahliae* strain JR2, *V. alfalfae* strain PD683, *V. tricorpus* strain PD593 and *V. klebahnii* strain PD401 because of their high assembly contiguity and spread throughout the *Verticillium* genus (Inderbitzin et al., 2011a; Shi-Kunne et al., 2018). After removal of repetitive and duplicated sequences, we obtained a pan-LS-genome of ~2 Mb, of which 60% occurs in clade Flavexudans and 72% in clade Flavnonexudans (clade pan-LS-genomes) (Figure 4). Next, the distribution of the pan-LS-genome and the clade pan-LS-genomes was evaluated for each *Verticillium* strain individually (Figure 4, Table S6). The proportion of the LS-pan-genome differed markedly between *Verticillium* strains and ranged from 12% for *V. nubilum* strain PD621 up to 58% for *V. dahliae* strain JR2 (Figure 4). Notably, by using a limited number of isolates in the consensus reconstruction, retentions are likely biased towards strains that are phylogenetically closer related to the species that were used to compose the pan-genome. However, *V. albo-atrum* strains contained considerably more of the pan-LS-genome compared to *V. zaregamsianum* and *V. isaacii* strains, despite its phylogenetically more distant relation to *V. klebahnii* and *V. tricorpus* (Figure 4). Moreover, LS contents do not only differ considerably between species but also within species as we also observed large differences between strains of the same species. For example, the genome of *V. dahliae* strain VdLs17 contains less than two thirds of the content present in the JR2 genome despite the recent divergence of the two strains (Figure 1A, 4) (Faino et al., 2015). Thus, sequences with high interspecific identity are genus-wide associated with dynamic genomic regions of *Verticillium* spp. as their contents vary greatly between and within species.

### Increased sequence conservation is likely not driven by negative selection

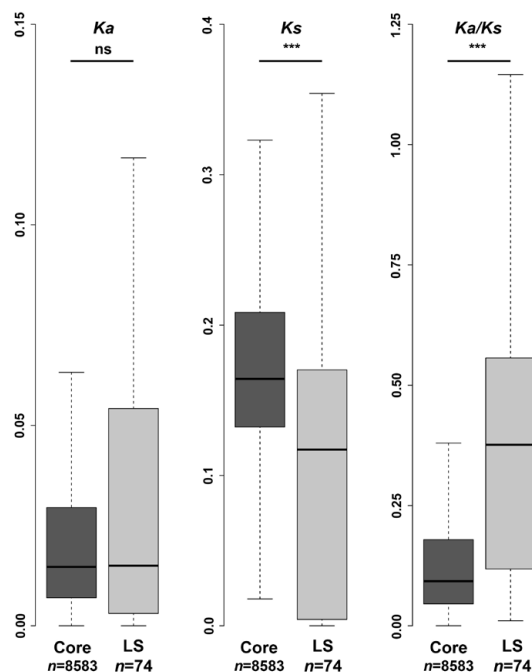
If negative selection is responsible for observed high interspecific sequence identities, depletion of polymorphisms should only concern protein-coding sequences in LS regions. However, sequence comparisons of coding and intergenic regions between *V. dahliae* and *V. nonalfalfae* revealed that increased sequence conservation is also observed in intergenic regions (Figure 5), indicating that high sequence identities are likely not driven by negative selection acting on protein-coding genes.



**FIGURE 5. Sequence identity of *V. dahliae* strain JR2 core and lineage-specific (LS) regions with *V. nonalfalfae* strain TAB2 for (A) coding and (B) intergenic sequences.** (A) Coding sequence of *V. dahliae* strain JR2 genes were aligned to coding sequences of *V. nonalfalfae* strain TAB2 genes and the sequence identity was determined. (B) For the intergenic regions, windows of 5 kb were constructed for *V. dahliae* strain JR2 core and LS regions. The sequence identity distribution is significantly different between core and LS regions and this for both the coding sequence of genes and intergenic regions (two-sided Wilcoxon rank-sum test,  $P < 0.0001$ ).

To see how selection impacts the evolution of LS region genes, we determined the rates of non-synonymous ( $K_a$ ) and synonymous ( $K_s$ ) substitutions for LS versus core genes. In total, 48% (68 out of 142) of the LS genes could not be used for  $K_a$  and  $K_s$  determination, as we did not observe any substitutions when compared to their corresponding *V. nonalfalfae* orthologs. In contrast, almost all core genes (8,583 out of 8,584) display nucleotide substitutions when compared with their *V. nonalfalfae* orthologs. Whereas the  $K_a$  was not different (two-sided Wilcoxon rank-sum test,  $P < 0.05$ ) between LS (median=0.015,  $n=74$ ) and core (median=0.015,  $n=8583$ ) genes, the  $K_s$  of LS genes (median=0.12,  $n=74$ ) was significantly lower than of core genes (median=0.16,  $n=8583$ ) (Figure 7). Consequently, LS genes (median=0.38,  $n=60$ ) have significantly higher  $K_a/K_s$  values than core genes (median=0.09,  $n=8289$ ), calculated for genes that have both synonymous and non-

synonymous substitutions compared with their *V. nonalfalfae* orthologs. In total, 15 of the 74 tested LS genes displayed  $Ka/Ks > 1$ , which is a higher proportion than the 100 of the 8,583 core genes with  $Ka/Ks > 1$  (Fisher's exact test,  $P < 0.05$ ). Two LS and two core genes with  $Ka/Ks > 1$  were predicted to contain an N-terminal signal peptide, which is a typical characteristic of effector proteins. However, due to the limited sequence divergence in the LS regions, positive selection on the genes with  $Ka/Ks > 1$  was not significant based on a Z-test, whereas in the core genome 21 genes were found to be under positive selection ( $P < 0.05$ ). In conclusion, despite high interspecific sequences identities, LS genes display more diversifying selection than the core genes as  $Ka/Ks$  ratios are significantly higher.



**FIGURE 6. Comparison of substitutions of *V. dahliae* strain JR2 and *V. nonalfalfae* strain TAB2 orthologs between core and lineage-specific (LS) regions.** The distribution of non-synonymous substitution rates ( $Ka$ ), synonymous substitution rates ( $Ks$ ) and  $Ka/Ks$  ratios are depicted for *V. dahliae* genes aligned to *V. nonalfalfae* orthologs. Outliers are not depicted. Significance of the different distributions was calculated with the two-sided Wilcoxon rank-sum test (ns = not significant, \*\*\* =  $P < 0.001$ ).



## Discussion

Genomes of many filamentous plant pathogens are thought to obey to a two-speed evolution model (Croll and McDonald, 2012; Dong et al., 2015; Möller and Stukenbrock, 2017). *V. dahliae* is similarly thought to evolve under a two-speed regime, as LS regions that are of significance for host interactions display signs of accelerated evolution with increased structural variation and TE activity (de Jonge et al., 2013; Faino et al., 2015; Faino et al., 2016). Additionally, LS regions are extremely plastic with abundant presence/absence polymorphisms (Figure 1) (de Jonge et al., 2013; Faino et al., 2016). Intriguingly, although LS regions are enriched in segmental duplications (Faino et al., 2016), LS sequences display high sequence identity to other *Verticillium* spp. (Figure 2-3, Table 1). Principally, sequences with increased identities between distinct taxa can originate from horizontal transfer, a phenomenon that has been implicated in the pathogenicity of various filamentous plant pathogens (Soanes and Richards, 2014). For instance, *Pyrenophora tritici-repentis*, the causal agent of wheat tan spot, acquired a gene from the fungal wheat pathogen *Phaeosphaeria nodorum* enabling the production of the host-specific toxin ToxA that mediates pathogenicity on wheat (Friesen et al., 2006a). However, horizontal transfer is not likely to explain our observations, as the increased sequence identity is observed genus-wide and concerns every species of the genus, and the degree of nucleotide conservation precisely corresponds to the phylogenetic distance between the species (Figure S2-S3, Table 1 and S4). As horizontal transfer is extremely unlikely, we argue that *V. dahliae* LS regions are subject to increased sequence conservation. However, this increased sequence conservation of LS regions is not a consequence of negative selection on coding regions, as intergenic regions display similarly increased conservation levels (Figure 5). Moreover, genes residing in LS regions show a tendency to display higher  $Ka/Ks$  ratios when compared with core genes, indicative of diversifying selection (Figure 6) (Sperschneider et al., 2015). Intriguingly, LS genes display similar rates of non-synonymous substitutions ( $Ks$ ) as core genes, but carry significantly less synonymous substitutions ( $Ks$ ). Consequently, high interspecific sequence identity is likely caused by fewer SNPs in LS regions.

Lower levels of synonymous substitutions were previously found for repeat-rich dispensable chromosomes of the fungal wheat pathogen *Zymoseptoria tritici* (Stukenbrock et al., 2010). However, this observation was not attributed to lower substitution rates, but rather the consequence of a lower effective population size of these dispensable chromosomes (Stukenbrock et al., 2010). Thus, the increased sequence conservation as observed here is unprecedented and perhaps counter-intuitive. Previously, increased substitution rates have been associated with two-speed genome evolution (Cuomo et al., 2007; Dong et al., 2015). For example, repeat-induced point (RIP) mutagenesis increases sequence divergence of particular effector genes of the oilseed rape pathogen

*Leptosphaeria maculans* that are localized near TEs (van de Wouw et al., 2010). However, accelerated evolution through increased SNP frequencies is not consistently observed for two-speed genomes, as no significant difference in SNP frequencies between core and repeat-rich genomic regions was found in *P. infestans* (Raffaele et al., 2010). Nevertheless, accelerated evolution of LS regions can also be established through other means, such as presence/absence polymorphisms. The well-characterized *V. dahliae* LS effector *Ave1* is highly conserved, as an identical copy occurs in *V. alfalfae* strain VaMs102 that displays a genome-wide average sequence identity of 92% (de Jonge et al., 2012). Moreover, no *Ave1* allelic variation is hitherto found in the *V. dahliae* population as well as in *V. alfalfae* and *V. nonalfalfae* populations (de Jonge et al., 2012; Song et al., 2017). Since *Ave1* is recognized by the tomato immune receptor Ve1 (Fradin et al., 2009), evasion of recognition occurs through various *Ave1* deletion events from the population (de Jonge et al., 2012; Faino et al., 2016).

Mechanisms that can explain the observed increased sequence conservation in repeat-rich LS regions remain unknown. Substitutions mostly originate from DNA polymerase errors and there is no immediate reason why error rates would diverge in LS regions. Possibly, the depletion of SNPs can be associated with a differential epigenetic organisation of LS regions. Intriguingly, a study into chromatin structure in the human genome noted that regions of open chromatin displayed lower mutation rates which was hypothesized to be a result of these regions being more accessible to repair mechanisms (Prendergast et al., 2007). However, repeat-rich regions such as the LS regions in *V. dahliae* are thought to be associated with densely organised chromatin, referred to as heterochromatin (Galazka and Freitag, 2014). In *Z. tritici*, repeat-rich conditionally dispensable chromosomes are enriched for histone modifications associated with heterochromatin, in contrast to core chromosomes that are largely euchromatic and transcriptionally active (Schotanus et al., 2015). Generally, heterochromatin is associated with suppression of genomic structural alterations such as recombination. Nevertheless, heterochromatic regions of *Z. tritici* are enriched for structural variations as they are enriched for duplications and deletions (Seidl et al., 2016). Thus, further research is needed to investigate whether differences in chromatin organisation can explain lower rates of SNP frequencies that occur in the plastic LS regions of *V. dahliae*.

## Conclusion

The two-speed genome is an intuitive evolutionary model for filamentous pathogens, as genes important for pathogenicity benefit from frequent alternations to mediate continued symbiosis with the host. However, filamentous pathogens comprise a heterogeneous group of organisms with diverse lifestyles (Dean et al., 2012; Kamoun et al., 2015). Consequently, it is not surprising that accelerated evolution is driven by

different mechanisms between species. In *V. dahliae*, acceleration evolution is merely achieved through presence/absence polymorphisms, as nucleotide sequences are highly conserved in LS regions. Perhaps, deletion of recognized effectors leads to a more rapid immunity evasion than sequence alterations through SNPs (Daverdin et al., 2012). Thus, the quick fashion of host immunity evasion through the deletion of effector genes can be evolutionary advantageous over allelic diversification, especially for soil-borne pathogens with a small effective population size that have little means of mobility.

## Materials & Methods

### Genome sequencing and assembly

Genomes of *V. albo-atrum* PD747, *V. alfalfae* PD683, *V. dahliae* JR2 and VdLs17, *V. isaacii* PD618, *V. klebahnii* PD401, *V. nubilum* PD621, *V. tricorpus* PD593 and MUCL9792, *V. zaregamsianum* PD739 were previously assembled (Klosterman et al., 2011; Faino et al., 2015; Seidl et al., 2015; Shi-Kunne et al., 2018) and sequence reads of *V. nonalfalfae* isolates TAB2 and Rec are publicly available (Bioproject PRJNA283258) (Jelen et al., 2016). *Verticillium* strains CQ2, 85S, PD670, PD660, PD659 and PD736 were newly sequenced. To this end, we isolated genomic DNA from potato dextrose broth cultures as previously described (Seidl et al., 2015). *V. dahliae* strains CQ2 and 85S were sequenced on the PacBio RSII platform (Pacific Biosciences of California, CA, USA) (Faino et al., 2015). Briefly, DNA was mechanically sheared and size selected using the BluePippin preparation system (Sage Science, Beverly, MA, USA) to produce ~20 kb size libraries. The sheared DNA and final library were characterized for size distribution using an Agilent Bioanalyzer 2100 (Agilent Technology, Inc., Santa Clara, CA, USA). The PacBio libraries were sequenced on four SMRT cells per *V. dahliae* isolate on a PacBio RS II instrument using the P6-C4 polymerase-Chemistry combination and a >4 h movie time and stage start. Filtered sub-reads for CQ2 and 85S, were assembled using the HGAP v3 protocol (Table S1) (Chin et al., 2013).

For PD670, PD660, PD659 and PD736, two libraries (500 bp and 5 kb insert size) were prepared and sequenced using an Illumina High-throughput sequencing platform. In total, ~18 million paired-end reads (150 bp read length; 500 bp insert size library) and ~16 million mate-paired read (150 bp read length; 5 kb insert size library) were produced per strain. We assembled the genomes using the A5 pipeline (Tritt et al., 2012), and we subsequently filled the remaining sequence gaps using SOAPdenovo2 (Luo et al., 2012). After obtaining final assemblies, we used QUAST (Gurevich et al., 2013) to calculate genome statistics. Gene annotation for *V. dahliae* strain JR2 and other *Verticillium* spp. were obtained from previous studies (Faino et al., 2016; Shi-Kunne et al., 2018). *V. isaacii* strain PD660 was annotated with the Maker2 pipeline according to (Holt and Yandell, 2011).

## Comparative genome analysis

Alignments to a repeat-masked genome as a reference, in order to prevent assigning high sequence identities to repetitive elements, were performed with nucmer that is part of the mummer package (v3.1) (Kurtz et al., 2004). Repetitive elements were identified using RepeatModeler (v1.0.8) based on known repetitive elements and on *de novo* repeat identification, and genomes were subsequently masked using RepeatMasker (v4.0.6; sensitive mode) (Smit et al., 2015).

Linear plots showing alignments within and closely adjacent JR2 LS regions were plotted with the R package genoPlotR (Guy et al., 2011). The *Verticillium* phylogenetic tree adjacent to the genoPlotR plots was previously generated using 5,228 single-copy orthologs that are conserved among all of the genomes (Shi-Kunne et al., 2018). The phylogenetic tree of *V. dahliae* strains was constructed using REALPHY (Bertels et al., 2014).

Alignments > 7.5 kb in length were depicted along the reference genome with the R package Rcircos (Figure 3) (Zhang et al., 2013). LS sequences were defined by alignment of different strains to a reference using nucmer (v3.1) (Kurtz et al., 2004) and regions were determined using BEDTools v2.25.0 (Quinlan and Hall, 2010).

Lineage-specific regions of *V. dahliae* and *V. tricorpus* were delimited based on the abundance of LS sequences and increased sequence conservation (Table S7). The pairwise identity of the genome-wide and LS regions between *V. dahliae*/*V. tricorpus* and other haploid *Verticillium* spp. was calculated using nucmer (mum) by dividing the respective query sequences into non-overlapping windows of 500 bp (Table 1). Sequence identities of the coding regions of genes and intergenic regions were retrieved by BLAST (v2.2.31+) searches between strains *V. dahliae* JR2 and *V. nonalfalfae* TAB2 (Altschul et al., 1990). Hits with a minimal coverage of 80% with each other were selected. Intergenic regions of *V. dahliae* strain JR2 were fractioned in 5 kb windows with BEDTools v2.25.0 and similarly blasted to the genome of *V. nonalfalfae* strain TAB2 (Quinlan and Hall, 2010). Hits with a maximal bit-score and minimal alignment of 500 bp to a window were selected. To compare the rate of synonymous and non-synonymous substitutions between the core and LS regions,  $K_a$  and  $K_s$  were of orthologs of JR2 and TAB2 were determined using the Nei and Gojobori method (Nei and Gojobori, 1986) in PAML (v4.8) (Yang, 2007). Significance of positive selection was tested using a Z-test (Stukenbrock and Dutheil, 2012). Z-values >1.65 were considered significant with  $P < 0.05$ . Secreted proteins were predicted by SignalP4 (Petersen et al., 2011).

The pan-LS-genome was constructed based on following *Verticillium* isolates: JR2 (*V. dahliae*), PD683 (*V. alfalfae*), PD593 (*V. tricorpus*) and PD401 (*V. klebahnii*). Genome regions of these for species with increased sequence conservation were combined (Table S7). Repeat masked regions were removed from the pan-LS-genome using BEDTools v2.25.0 (Quinlan and Hall, 2010). Additionally, regions in duplicate ( $\geq 90\%$  identity,  $\geq 100$ bp)

in the pan-LS-genome were determined using nucmer (v3.1) (Kurtz et al., 2004) and subsequently removed with using BEDTools v2.25.0 (Quinlan and Hall, 2010). As result, a pan-LS-genome was constructed without regions in duplicate. The fraction of pan-LS-genome that is present in every individual *Verticillium* strain was determined using nucmer (v3.1) (Kurtz et al., 2004). The clade pan-LS-genomes were constructed by combining all the pan-LS-genome regions that are present in the *Verticillium* clade isolates, which was then also removed from duplicate regions.

### Ortholog analysis and tree building

Orthologous groups were determined by using OrthoMCL (Li et al., 2003). *Sodiomyces alkalinus* was used as outgroup (Grum-Grzhimaylo et al., 2018). Orthologs groups that are shared among at least three species were selected for phylogenetic tree construction. Individual ortholog groups were aligned using mafft (LINSi; v7.04b) (Kato et al., 2002; Kato and Standley, 2013) and subsequently concatenated. Maximum likelihood phylogeny was inferred using RAXML (v8.2.4) with the GAMMA model of rate heterogeneity and the Whelan and Goldman (WAG) model of amino acid substitutions (Stamatakis, 2014). The robustness of the inferred phylogeny was assessed by 100 rapid bootstrap approximations.

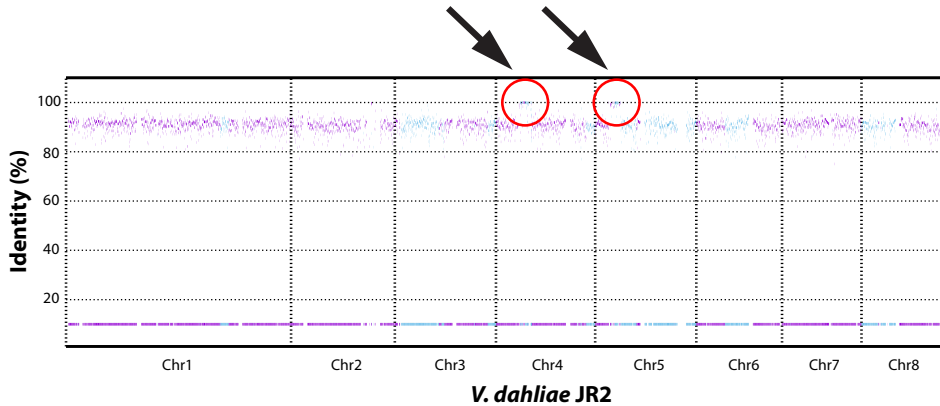
### Acknowledgements

This work was supported by the Marie Curie Actions program of the European Commission (FP7-PEOPLE-2013-ITN, grant agreement number 607178). H.M. was supported by French Ministry of Higher Education and Research (CIFRE 2013/1431). Work in the laboratories of B.P.H.J.T. and M.F.S is supported by the Research Council Earth and Life Sciences (ALW) of the Netherlands Organization of Scientific Research (NWO).

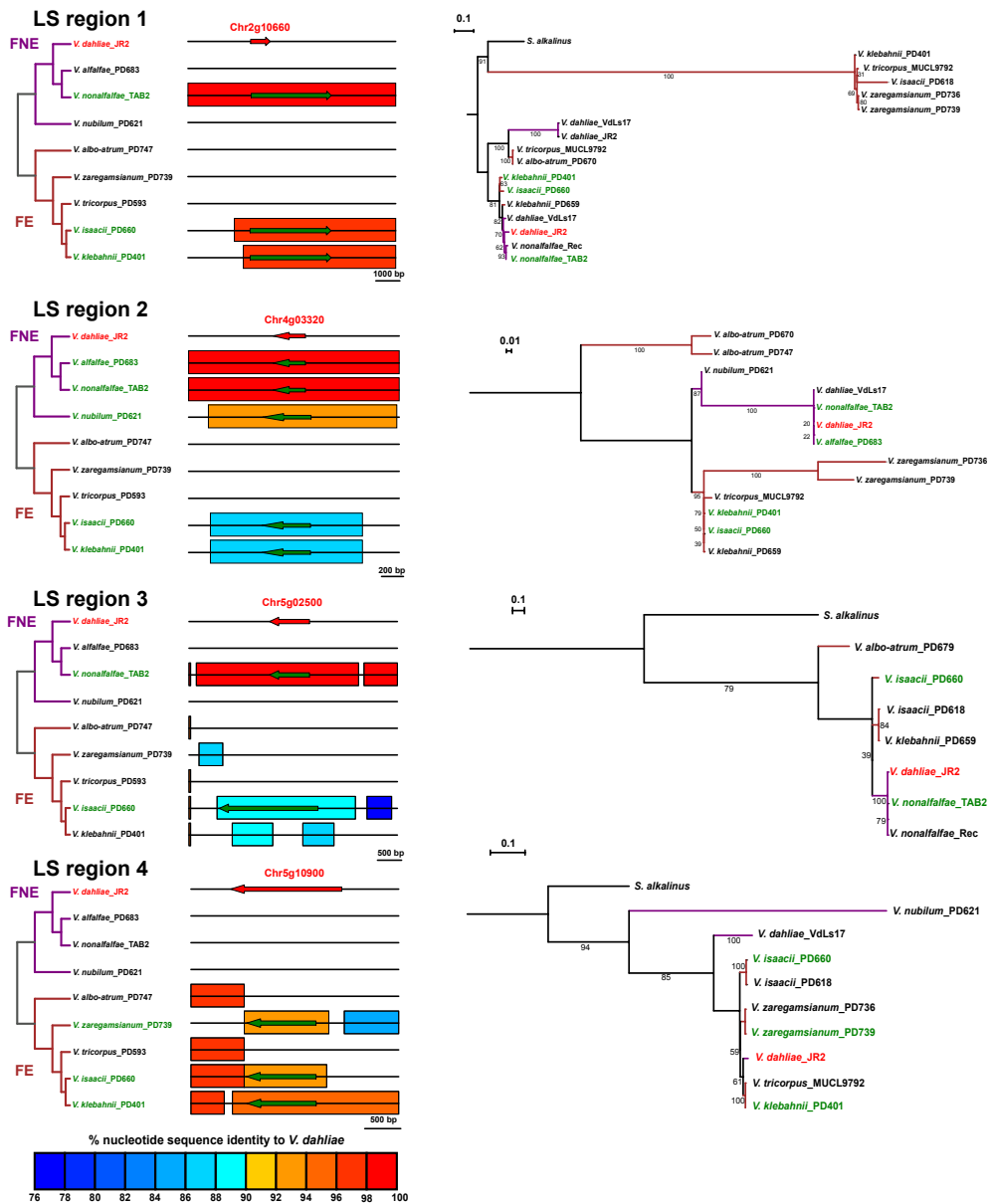
### Data accessibility

The Whole Genome Shotgun projects have been deposited at DDBJ/ENA/GenBank as accessions PRLI00000000 and PRLJ00000000 for *V. dahliae* strains CQ2 and 85S, respectively.

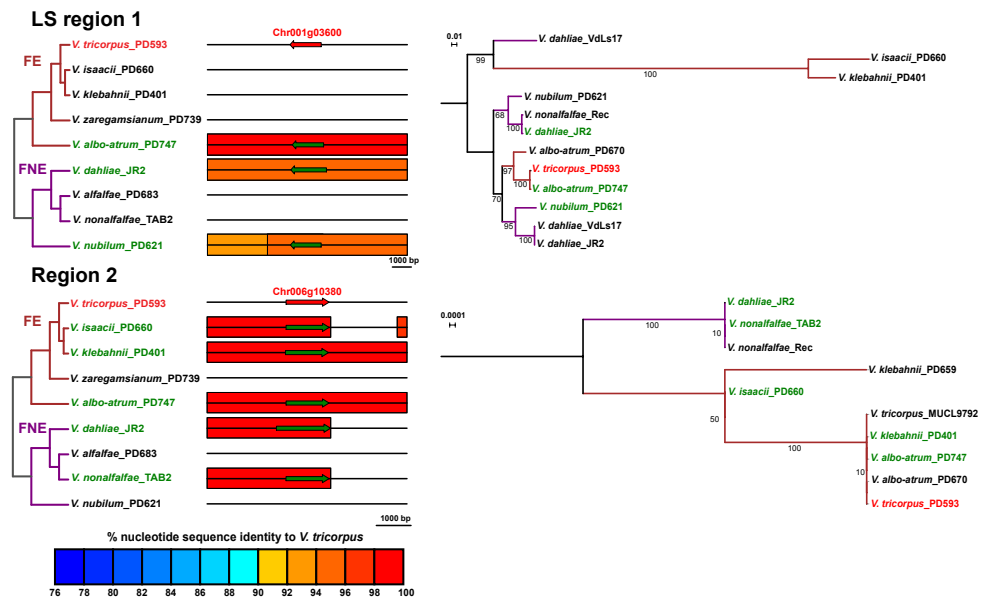
## Supplementary material



**FIGURE S1. Whole-genome coverage plot displaying identity of aligned sequences of *V. nonalfalfae* TAB2 to *V. dahliae* JR2.** Aligned genomic regions are displayed and purple and blue colours indicate forward-forward alignments and forward-reverse alignments (inversions), respectively. Genome regions with black arrows indicate 99-100% sequence identity.



**FIGURE S2. Phylogenetic relationship of *V. dahliae* LS region genes with *Verticillium* homologs.** One gene was analysed for every LS region in *V. dahliae* JR2. On the left panel are nucleotide alignments depicted of sequences orthologous to the analysed gene. The colour bar on the black line represents the nucleotide sequence identity to the orthologous genome region of *V. dahliae* JR2. On the right panel is the phylogenetic relationship with *Verticillium* homologs depicted for the analysed gene. Here homology is based on protein sequence, which is less stringent than the nucleotide alignments on the left panel. The robustness of the phylogeny was assessed using 100 bootstrap replicates. The corresponding genes between the left and right panel are depicted in red and green for *V. dahliae* and other *Verticillium* spp., respectively.



**FIGURE S3. Phylogenetic relationship of *V. tricorpus* genes with high interspecific sequence identity.** One gene was analysed for the LS region in *V. tricorpus* PD593 and one for the region on scaffold 6 with high interspecific sequence identity. On the left panel are nucleotide alignments depicted of sequences orthologous to the analysed gene. The colour bar on the black line represents the nucleotide sequence identity to the orthologous genome region of *V. tricorpus* PD593. On the right panel is the phylogenetic relationship with *Verticillium* homologs depicted for the analysed gene. Here homology is based on protein sequence, which is less stringent than the nucleotide alignments on the left panel. The robustness of the phylogeny was assessed using 100 bootstrap replicates. The corresponding genes between the left and right panel are depicted in red and green for *V. tricorpus* and other *Verticillium* spp., respectively.

**TABLE S1. *Verticillium dahliae* strain CQ2 and 85S genome assemblies.**

		CQ2	85S
SMRT cells		4	4
Filtered subreads		430,378	500,428
Coverage (n-fold)		110x	130x
HGAP3 assembly	Size (bp)	35,818,019	35,931,336
	Contigs	17	40
	Longest contig	7,864,170	6,633,769
	N <sub>50</sub> (bp)	3,754,186	3,176,090
	No. of Ns/100 kb	0	0



**TABLE S2. Assembly statistics of the *Verticillium* genomes used in this study.**

Species	Strain name	Genome size (Mb)	#Ns/100 kb	N50 (Mb)	# Contigs (≥0 bp)	# Scaffolds (≥1000 bp)
<i>V. albo-atrum</i>	PD670	37.4	56.67	3.7	107	72
	PD747	36.5	16.26	3.9	34	19
<i>V. alfalfae</i>	PD683	32.7	19.36	4.5	40	14
<i>V. dahliae</i>	VdLs17	36.0	0	5.9	8	8
	JR2	36.2	0	4.2	8	8
	CQ2	35.8	0	3.8	17	17
	85S	35.9	0	3.2	40	40
<i>V. nonalfalfae</i>	Rec	33.0	35.4	0.9	1026	795
	TAB2	34.3	897.53	1.8	793	167
<i>V. nubilum</i>	PD621	37.9	9.13	4.7	246	189
<i>V. tricorpus</i>	MUCL9792	36.0	229,64	4,7	255	53
	PD593	35.0	14.52	4.4	71	9
<i>V. isaacii</i>	PD618	35.8	62.48	3.1	239	122
	PD660	36.0	37.53	2.5	114	43
<i>V. klebahnii</i>	PD659	36.2	59.47	3.6	120	60
	PD401	36.0	35.30	3.2	79	37
<i>V. zaregamsianum</i>	PD736	37.1	62.7	2.0	125	62
	PD739	37.1	55.38	3.5	75	32

**TABLE S3. Sequences of other haploid *Verticillium* spp. with high identity to *Verticillium dahliae* JR2.**

Species	Strain name	Amount of high identity sequences (kb)	Fraction that align to LS regions (%)
<i>V. albo-atrum</i> <sup>b</sup>	PD670	209	94.6
	PD747	160	91.9
<i>V. alfalfae</i> <sup>a</sup>	PD683	473	99.5
<i>V. nonalfalfae</i> <sup>a</sup>	Rec	662	100.0
	TAB2	705	99.9
<i>V. nubilum</i> <sup>b</sup>	PD621	147	69.6
<i>V. tricorpus</i> <sup>b</sup>	MUCL9792	92	76.1
	PD593	304	95.3
<i>V. isaacii</i> <sup>b</sup>	PD618	153	89.5
	PD660	202	89.9
<i>V. klebahnii</i> <sup>b</sup>	PD659	193	86.6
	PD401	142	81.9
<i>V. zaregamsianum</i> <sup>b</sup>	PD736	54	52.4
	PD739	68	86.8

<sup>a</sup> Sequences with an identity of >95% with *V. dahliae* JR2

<sup>b</sup> Sequences with an identity of >90% with *V. dahliae* JR2

**TABLE S4. Sequence identities between *V. tricolor* (PD593) and other haploid *Verticillium* species (excluding repetitive regions).**

Species/strain	Genome-wide <sup>a</sup> (%)	Conserved regions <sup>a</sup> (%)	# windows aligned to conserved regions <sup>b</sup>	<i>p</i> -value <sup>b</sup>
<i>V. albo-atrum</i> /PD747	89.6 (5.2)	100 (3.1)	606	2.2e-16
<i>V. alfalfae</i> /PD683	88.2 (4.0)	86.4 (1.9)	2	0.35
<i>V. dahliae</i> /JR2	88.8 (4.5)	97.6 (3.4)	172	2.2e-16
<i>V. nonalfalfae</i> /TAB2	88.4 (4.1)	99.4 (4.8)	25	7.9e-11
<i>V. nubilum</i> /PD621	88.3 (3.9)	95.1 (4.2)	64	2.2e-16
<i>V. isaacii</i> /PD660	94.6 (4.0)	98.8 (4.0)	133	2.2e-16
<i>V. klebahnii</i> /PD401	94.8 (4.0)	98.8 (3.5)	182	2.2e-16
<i>V. zaregamsianum</i> /PD739	94.8 (4.0)	94.0 (3.5)	37	0.14

<sup>a</sup>The percentage is the median sequence identity and the number between the brackets is the standard deviation.

<sup>b</sup>Windows are 500bp in size.

<sup>c</sup>The *p*-value was calculated with a two-sided Wilcoxon rank sum test

**TABLE S5. Sequences of other haploid *Verticillium* spp. with high identity to *Verticillium tricolor* PD593.**

Species	Strain name	Amount of high identity sequences (kb)	Fraction that align to the two regions with high sequence identity (%)
<i>V. albo-atrum</i> <sup>b</sup>	PD670	510	83.4
	PD747	527	83.8
<i>V. alfalfae</i> <sup>b</sup>	PD683	46	1.3
<i>V. dahliae</i> <sup>b</sup>	VdLs17	153	80.6
	JR2	285	81.1
	CQ2	36	3.1
	85S	67	2.1
	Rec	322	70.0
<i>V. nonalfalfae</i> <sup>b</sup>	TAB2	60	30.0
	PD621	133	49.4
<i>V. nubilum</i> <sup>b</sup>	PD618	63	32.6
	PD660	100	60.4
<i>V. isaacii</i> <sup>a</sup>	PD659	105	62.5
	PD401	49	18.8
<i>V. zaregamsianum</i> <sup>a</sup>	PD736	44	9.8
	PD739	40	7.3

<sup>a</sup> Sequences with an identity of >95% with *V. tricolor* PD593

<sup>b</sup> Sequences with an identity of >90% with *V. tricolor* PD593

# Chapter 6

## **The genome of the fungal pathogen *Verticillium dahliae* reveals extensive bacterial to fungal gene transfer**

Xiaoqian Shi-Kunne, Mathijs van Kooten<sup>\*</sup>, Jasper R.L. Depotter<sup>\*</sup>,  
Bart P.H.J. Thomma<sup>#</sup> and Michael F. Seidl<sup>#</sup>

<sup>\*</sup>These authors contributed equally to this work.

<sup>#</sup>These authors contributed equally to this work

A modified version of this chapter has been accepted as:

Shi-Kunne X, van Kooten M<sup>\*</sup>, Depotter JRL<sup>\*</sup>, Thomma BPHJ<sup>#</sup>, Seidl MF<sup>#</sup> (2019) The genome of the fungal pathogen *Verticillium dahliae* reveals extensive bacterial to fungal gene transfer. *Genome Biol Evol* pii: evz040 (\*equal contribution, #equal contribution)

## Abstract

Horizontal gene transfer (HGT) involves the transmission of genetic material between distinct evolutionary lineages and can be an important source of biological innovation. Reports of inter-kingdom HGT to eukaryotic microbial pathogens have accumulated over recent years. *Verticillium dahliae* is a notorious plant pathogen that causes vascular wilt disease on hundreds of plant species, resulting in high economic losses every year. Previously, the effector gene *Ave1* and a glucosyltransferase-encoding gene were identified as virulence factor-encoding genes that were proposed to be horizontally acquired from a plant and a bacterial donor, respectively. However, to what extent HGT contributed to the overall genome composition of *V. dahliae* remained elusive. Here, we systematically searched for evidence of inter-kingdom HGT events in the genome of *V. dahliae* and provide evidence for extensive horizontal gene acquisition from bacterial origin.

## Introduction

Genetic information is generally vertically transferred from parents to their offspring. However, genetic information can also be transmitted laterally between reproductively isolated species, often referred to as horizontal gene transfer (HGT). It has been well established that HGT plays a significant role in the adaptive evolution of prokaryotic species (Eisen, 2000; Koonin et al., 2001; Bapteste et al., 2009). Three well-characterized mechanisms contribute to DNA uptake by prokaryotes, namely transformation, conjugation and transduction. With transformation, a DNA fragment from a dead, degraded bacterium or another donor enters a competent recipient bacterium (Johnston et al., 2014), while conjugation is the active DNA transfer between prokaryotic cells by direct cell-to-cell contact or by a bridge-like connection between two cells (Norman et al., 2009). Finally, transduction involves the transfer of a DNA fragment from one prokaryotic cell to another by a virus or viral vector. HGT in prokaryotes takes place at all taxonomic levels: from individuals of the same population up to inter-kingdom transfers.

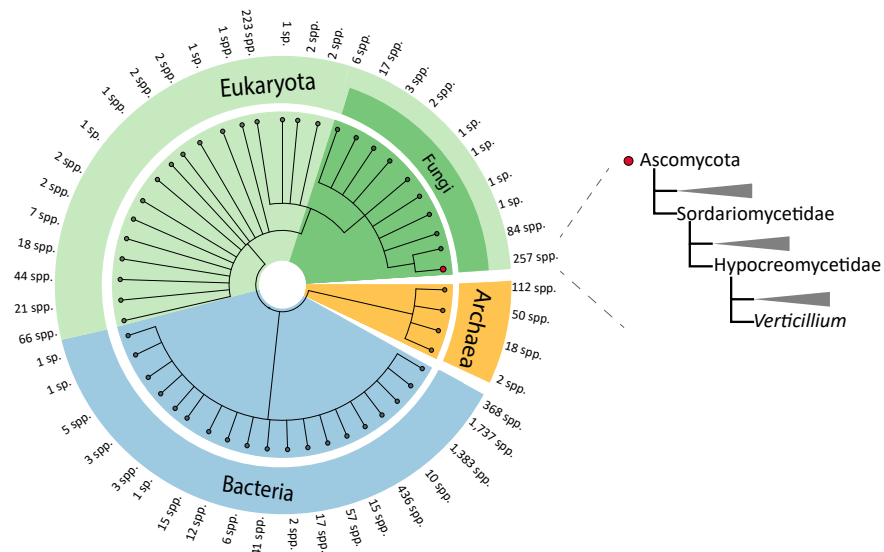
HGT also contributes to the evolution of eukaryotes, although it occurs much less frequent than in prokaryotes (Kurland et al., 2003; Bock, 2010). Horizontal gene transfer has played an important role in the evolution of pathogenic traits in plant pathogens (Soanes and Richards, 2014). For instance, the wheat pathogenic fungi *Pyrenophora tritici-repentis* and *Phaeosphaeria nodorum* both contain the near-identical effector gene *ToxA* that encodes a host-specific toxin that acts as pathogenicity factor (Friesen et al., 2006b). Compelling evidence revealed that *ToxA* was acquired by *P. tritici-repentis* from *P. nodorum* via HGT, giving rise to pathogenicity of the former species on wheat plants (Friesen et al., 2006b). Interestingly, *ToxA* was similarly found in the genome of yet another wheat pathogen, *Bipolaris sorokiniana* (McDonald et al., 2018).

Although knowledge on mechanisms of HGT in eukaryotes remains limited, HGT in filamentous plant pathogens is thought to occur more often between closely related species since donor and recipient species share similar genomic architectures (Mehrabi et al., 2011). For the filamentous plant pathogenic fungus *Fusarium oxysporum*, an entire chromosome can be transferred from one strain to another through co-incubation under laboratory conditions *in vitro* (Ma et al., 2010; van Dam et al., 2017).

Intriguingly, inter-kingdom HGT also contributed to the genome evolution of filamentous plant pathogens despite the evolutionary distant relation with the donor. For example, phylogenetic analyses of plant pathogenic oomycete species revealed 34 gene families that have undergone HGT between fungi and oomycetes (Richards et al., 2011). The repertoire of HGT candidates includes genes encoding proteins with the capacity to break down plant cell walls and effector proteins for interacting with host plants (Richards et al., 2011). Furthermore, genomic and phylogenetic analyses of the fungus

*Fusarium pseudograminearum*, the major cause of crown and root rot of barley and wheat in Australia, has revealed that a novel virulence gene was horizontally acquired from a bacterial species (Gardiner et al., 2012).

The fungal genus *Verticillium* contains nine haploid species plus the allodiploid *V. longisporum* (Inderbitzin et al. 2011a; Depotter et al. 2017). These ten species are phylogenetically subdivided into two clades; Flavexudans and Flavnonexudans (Inderbitzin et al. 2011a; Shi-Kunne et al. 2018). The Flavnonexudans clade comprises *V. nubilum*, *V. alfalfae*, *V. nonalfalfae*, *V. dahliae*, and *V. longisporum*, while the Flavexudans clade comprises *V. albo-atrum*, *V. isaacii*, *V. tricorpus*, *V. klebahnii* and *V. zaregamsianum* (Inderbitzin et al. 2011a). Among these *Verticillium* spp., *V. dahliae* is the most notorious plant pathogen that is able to cause disease in hundreds of plant species (Fradin and Thomma, 2006; Inderbitzin and Subbarao, 2014). Furthermore, *V. albo-atrum*, *V. alfalfae*, *V. nonalfalfae* and *V. longisporum* are pathogenic, albeit with narrower host ranges (Inderbitzin and Subbarao, 2014). Although the remaining species *V. tricorpus*, *V. zaregamsianum*, *V. nubilum*, *V. isaacii* and *V. klebahnii* have incidentally been reported as plant pathogens too, they are mostly considered saprophytes that thrive on dead organic material and their incidental infections should be seen as opportunistic (Inderbitzin et al. 2011a; Gurung et al. 2015; Ebihara et al. 2003). *Verticillium* spp. are considered to be strictly asexual, yet various mechanisms contributing to the genomic diversity of *V. dahliae* have been reported (Seidl and Thomma, 2014; Faino et al., 2016; Seidl and Thomma, 2017), including HGT (Klosterman et al., 2011; de Jonge et al., 2012). Previously, *Ave1* and a glucosyltransferase encoding gene were found to be acquired from a plant and a bacterial donor, respectively, both of which were found to contribute to *V. dahliae* virulence during plant infection (Klosterman et al., 2011; de Jonge et al., 2012). These two inter-kingdom HGT events inspired us to study the extent and potential impact of inter-kingdom HGT to *Verticillium dahliae*.



**FIGURE 1. Phylogenetic tree of the species in the selected local data base.** Different colors represent different taxonomical groups. The number of species (spp.) on each branch is indicated. The phylogenetic position of *Verticillium* spp. is indicated on the right.

## Results

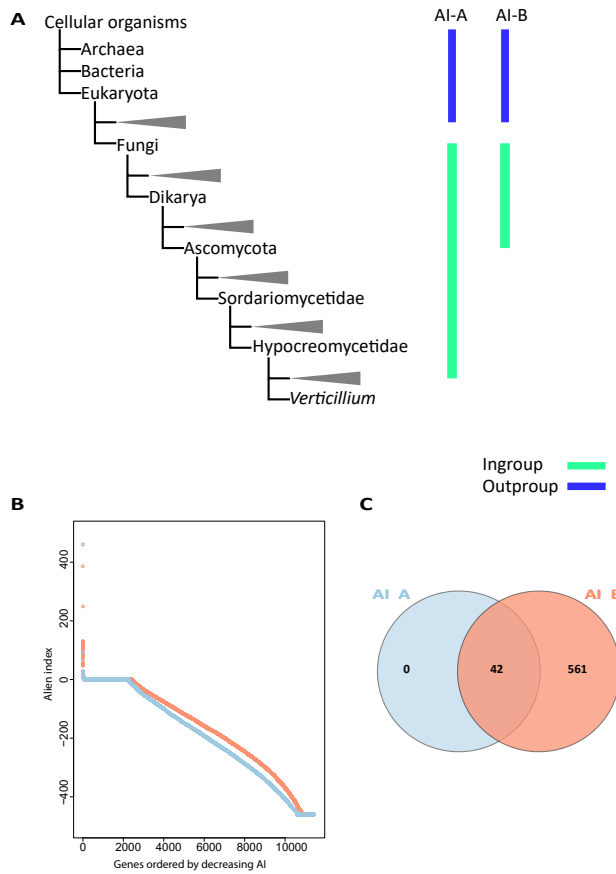
### Inter-kingdom HGT to *V. dahliae* and other Ascomycete fungi

To systematically search for genes in the *V. dahliae* genome that are derived from inter-kingdom HGT, we downloaded a recent UniProtKB proteome database that contains proteomes of 5,074 species, including 4,123 bacteria, 182 archaea and 769 eukaryotes (Figure 1). The database contains proteomes of 373 fungal species, including *V. dahliae* (strain VdLs17), *V. alfalfae* (strain VaMs102) and *V. longisporum* (strain VL1) (Klosterman et al., 2011). In our analyses, we focused on the complete telomere-to-telomere genome assembly of *V. dahliae* strain JR2 (Faino et al., 2015). We first queried each of the 11,430 protein sequences of *V. dahliae* strain JR2 using BLAST against the aforementioned UniProtKB protein database. Subsequently, we utilized the Alien Index (AI) method (Gladyshev et al., 2008; Alexander et al., 2016) that generates AI scores for each *V. dahliae* gene based on the comparison of best BLAST hits with in-group and out-group species; in this case non-*Verticillium* fungal species and non-fungal species, respectively (Figure 2A). In this manner, we queried for HGT candidates that were acquired only by *Verticillium* spp., or that were acquired by ancestral fungal species but retained only by *Verticillium* spp. during evolution. When the most significant BLAST hit of a gene within the out-group is more significant than the most significant hit within the in-group, the AI score is positive (i.e.  $AI > 1$ ) and this gene is considered a potential HGT candidate. This

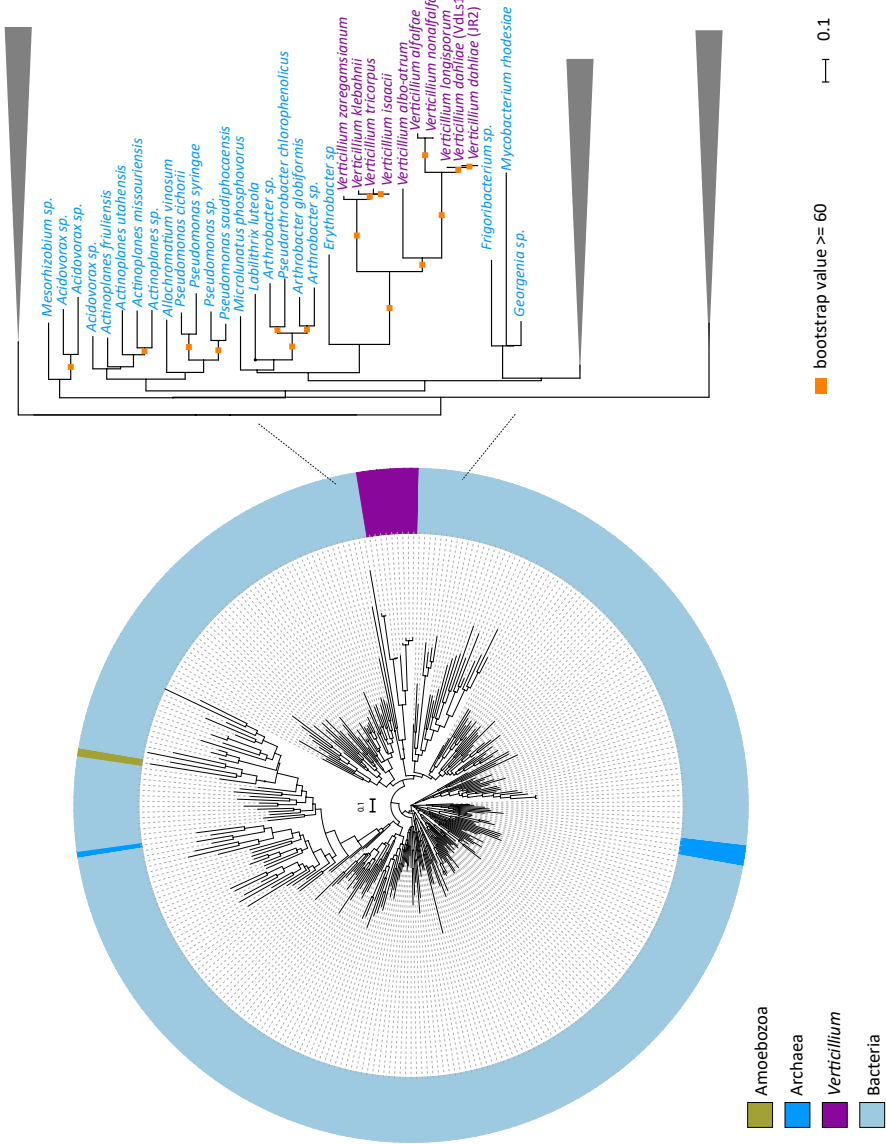
AI analysis yielded 42 HGT candidates with AI positive scores ( $AI > 1$ ) for *V. dahliae* JR2 (Figure 2B), which were further analyzed. To this end, we aligned protein sequences of all the homologs for each candidate (from in-, out-groups) and constructed phylogenetic trees that were automatically evaluated to remove phylogenetic trees with candidate genes that are directly adjacent to non-*Verticillium* fungal species (in-group species) rather than non-fungal species (out-group species). However, candidate genes that are closely related to the homologs of out-group species in a phylogenetic tree can also result from duplication followed by multiple gene losses in the in-group species. Typically, in such a phylogenetic tree, paralogs of the candidate gene cluster with genes of non-*Verticillium* fungal species (in-group species) and form separate branches from the branch of the candidate gene. Thus, we removed the candidates with phylogenetic trees that can be explained by gene losses in multiple species. In the end, we obtained two HGT candidates (HGT-1 and HGT-2) that were acquired only by *Verticillium* spp., or that were acquired by ancestral fungal species but kept only by *Verticillium* spp. during evolution. To check whether these two *V. dahliae* candidates are also present in other *Verticillium* spp., we queried them using BLAST against the proteomes of previously published *Verticillium* spp. (Shi-Kunne et al. 2018; Depotter et al. 2017). We found that candidate VDAG\_JR2\_Chr4g10560 (HGT-1) is present in all *Verticillium* species and that VDAG\_JR2\_Chr8g03340 (HGT-2) is only present in *V. alfalfae*, *V. nonalfalfae* and *V. longisporum* (Figure 3, Figure 4).

In order to search for HGT candidates that are not only present in *Verticillium* spp., but also may be present in other ascomycete species, we again used the AI method, albeit with adjusted AI group settings. In this case, we set non-fungal species and non-ascomycete fungal species as out- and in- groups, respectively. AI scanning resulted in 603 genes with AI positive scores ( $AI > 1$ ), which were selected for further assessment through phylogenetic tree analysis (Table 1). After assessing phylogenetic trees, we identified 43 additional HGT candidates with non-fungal origins that are also present in other fungal ascomycete species. Among these candidates, two appeared as paralogs in the phylogenetic tree. The clusters of these two genes are close to each other in the tree, which indicates that the two paralogs are likely the result of a duplication and not of independent acquisitions. Therefore, we only considered 42 HGT candidates for further analyses. Of these, one is of plant origin and 41 are of bacterial origin (Table 1). The candidate of plant origin is the previously identified HGT gene, *Ave1* (Figure 5) (de Jonge et al., 2012). The previously reported glucosyltransferase gene (Klosterman et al., 2011) likewise was identified as HGT event in our study (Figure 6). Among these 42 candidates, 32 are present in all of the *Verticillium* species. Eight of the remaining ten are only found in the *Verticillium* species of the Flavnonexudans clade (Figure 7).

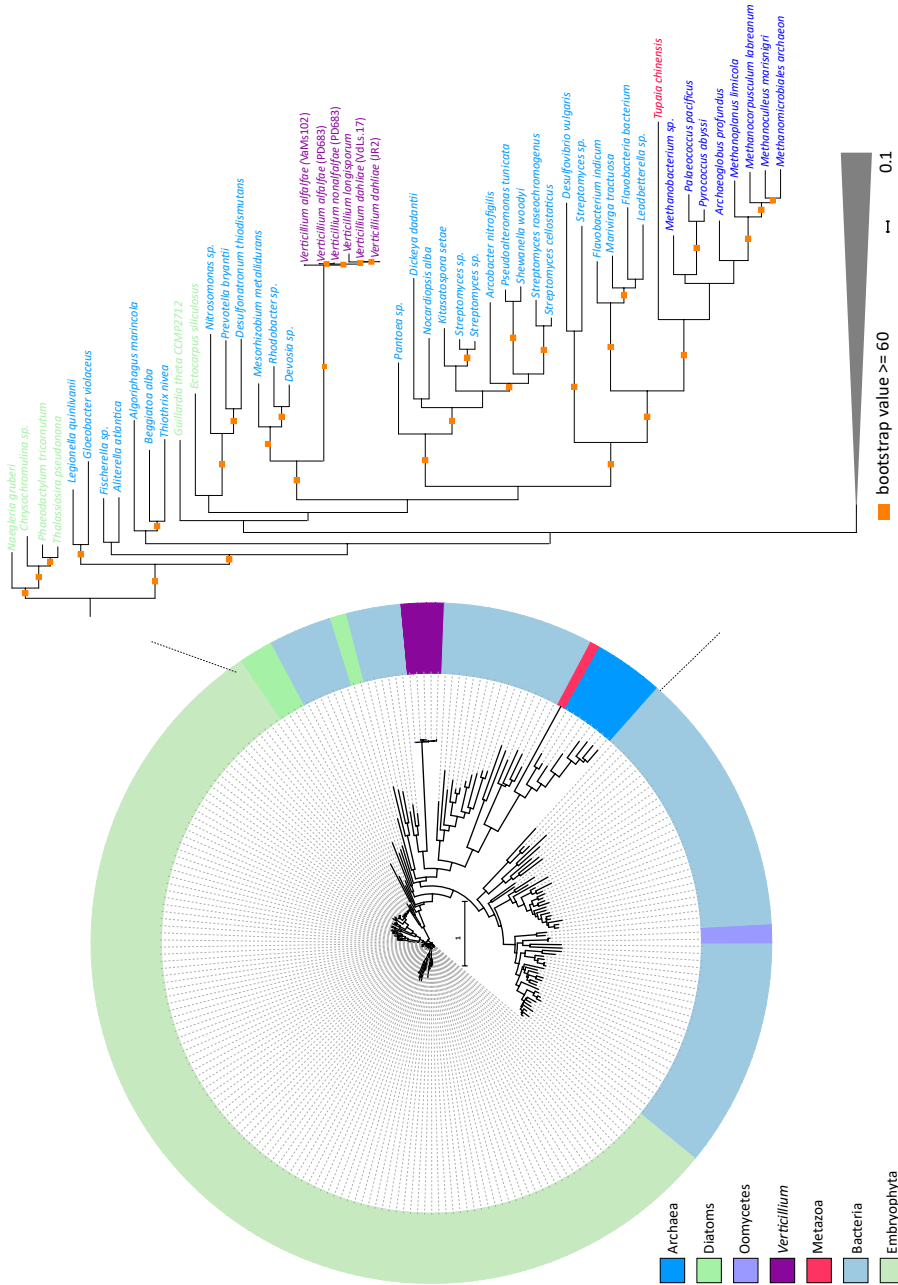




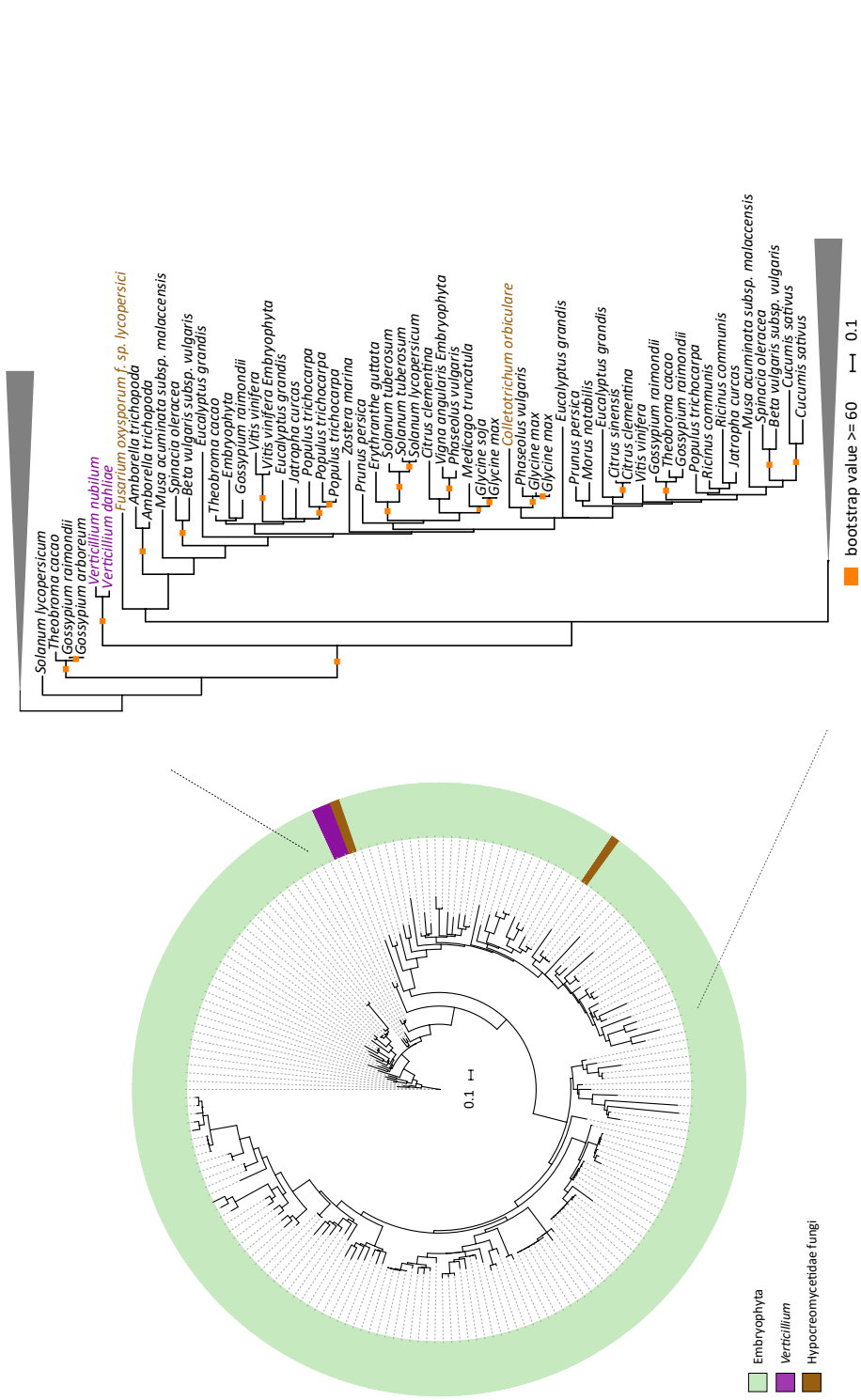
**FIGURE 2. Alien Index based HGT detection method for detecting HGT candidates.** (A) Simplified phylogenies of in- and out-group. In AI-A setting, in-group and out-group species are non-*Verticillium* fungal species and non-fungal species, respectively. In AI-B setting, in-group and out-group species are non-ascomycete fungal species and non-fungal species, respectively. (B) Distribution of AI scores. AI scores from two settings were calculated for every gene in the genome of *V. dahliae* strain JR2 and they were ordered by decreasing AI score. (C) Numbers of genes with AI positive scores from two different settings.



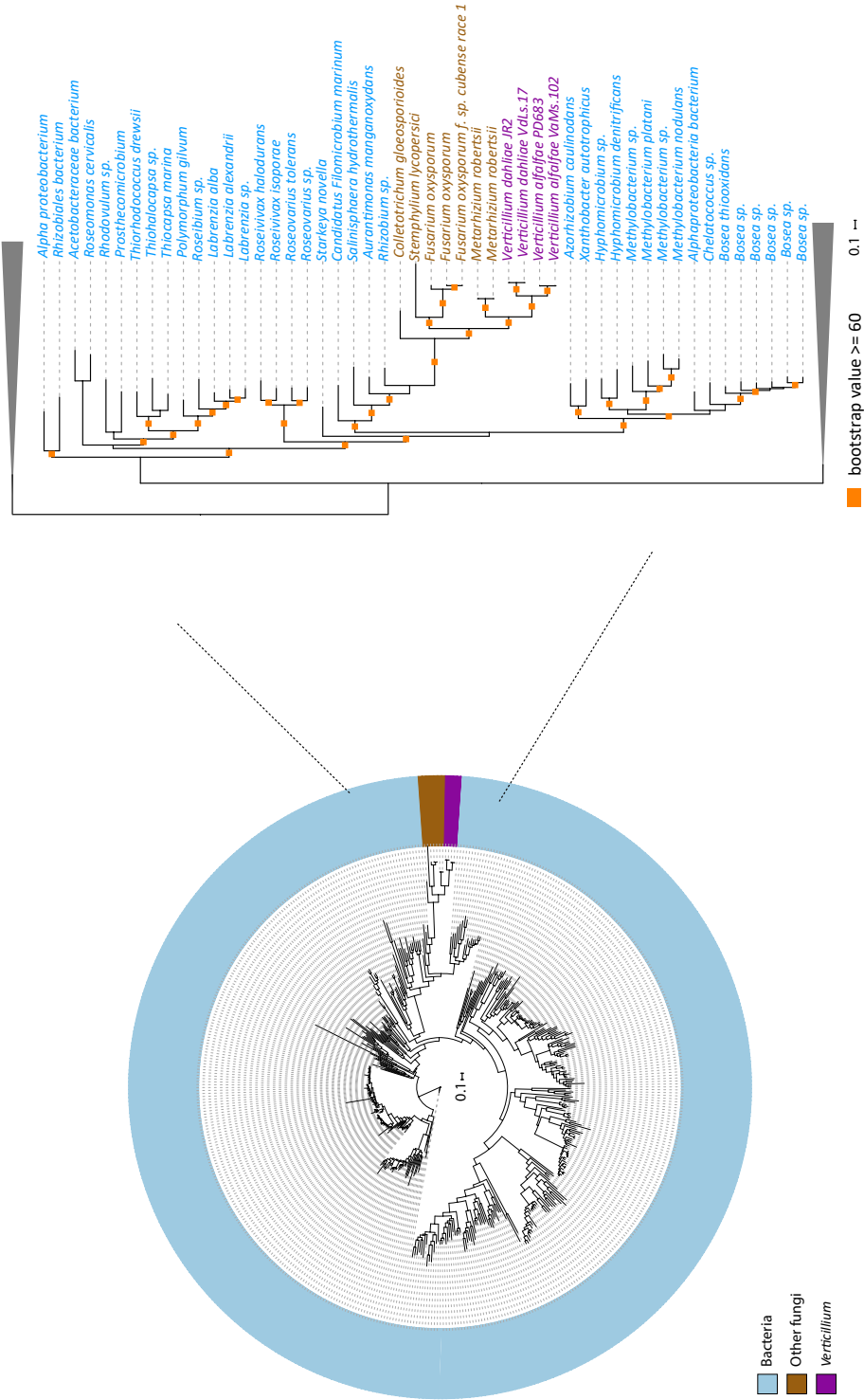
**FIGURE 3. Evolutionary relationship of HGT-1 homologs.** Protein sequences of HGT-1 orthologs were aligned and the resulting alignment was used to infer a maximum-likelihood phylogeny. The phylogeny suggests that *HGT-1* is transferred from a bacterial species. Different colors depict different groups or species. A more detailed part of the tree that contains *Verticillium* species is shown on the right. Orange squares indicate branches with bootstrap values  $\geq 60$ .



**FIGURE 4. Evolutionary relationship of HGT-2 homologs.** Protein sequences of HGT-2 orthologs were aligned (using a MAFFT) and the resulting alignment was used to infer a maximum-likelihood phylogeny (using RAxML). Different colors depict different groups or species. The phylogeny suggests that HGT-2 is transferred from a bacterial species. A more detailed part of the tree that contains *Verticillium* species is shown on the right. Orange squares indicate branches with bootstrap values  $\geq 60$ .



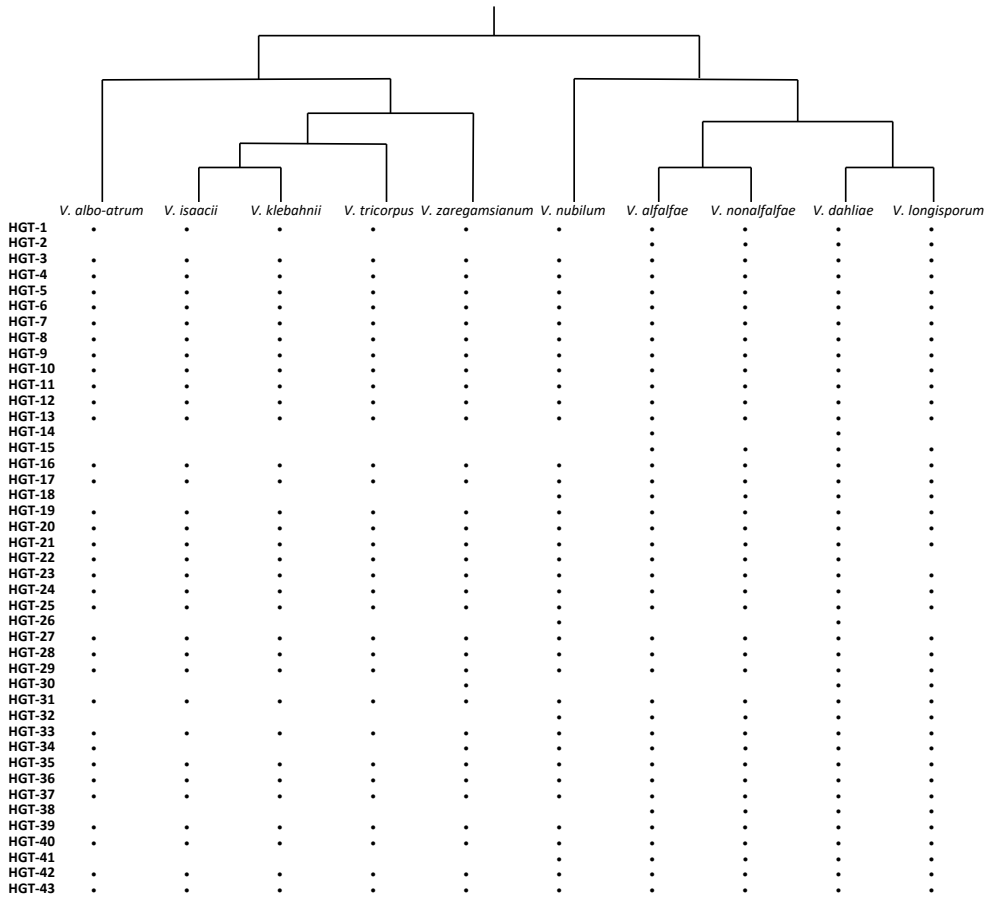
**FIGURE 5. Evolutionary relationship of Ave1 homologs.** Protein sequences of Ave1 orthologs were aligned and the resulting alignment (using a MAFFT) was used to infer a maximum-likelihood phylogeny (using RAxML). Different colors depict different groups or species. A more detailed part of the tree that contains *Verticillium* species is shown on the right. Orange squares indicate branches with bootstrap values  $\geq 60$  or above.



**FIGURE 6. Evolutionary relationship of *V. dahliae* glucosyltransferase homologs.** Orthologs of glucosyltransferase were aligned (using a MAFFT) and the resulting alignment was used to infer a maximum-likelihood phylogeny (using RAXML). Different colors depict different groups or species. A more detailed part of the tree that contains *Verticillium* species is shown on the right. Orange squares indicate branches with bootstrap values  $\geq 60$ .

## Proportionately high number of *V. dahliae* HGT candidates encode secreted proteins

Filamentous plant pathogens secrete large numbers of proteins, including plant cell wall-degrading enzymes and effector proteins, to interact with host plants (Cook et al., 2015). We used a combination of SignalP, TMHMM and TargetP (Shi-Kunne et al., 2018) to identify HGT-derived genes that encode such secreted proteins. This analysis showed that 12 of the 44 candidates encode secreted proteins. Compared to the total number of the secreted proteins in the genome of *V. dahliae* strain JR2 (858 out of 11,430), HGT candidates are significantly enriched for genes that are predicted to encode secreted proteins (Fisher exact test,  $p < 0.0001$ ). To know more about the putative functions of these proteins, we searched for conserved domains in each of these proteins using Interproscan (Jones et al., 2014). Among the 12 secreted proteins, five are glycoside hydrolases (GHs) and two are carbohydrate esterases (CEs), which are all carbohydrate-active enzymes (CAZymes). CAZymes are responsible for the synthesis and breakdown of glycoconjugates, oligo- and polysaccharides. CAZymes of plant associated fungi usually comprise a large number of plant cell-wall degrading enzymes (Zhao et al., 2013). The remaining five are Ave1, a chondroitin AC/alginase lyase, a phosphoinositide phospholipase, an amidohydrolase and a protein without functional annotation. Subsequently, we also searched for the functional annotation of the 30 non-secreted proteins. Besides the previously identified glucosyltransferase and seven proteins without functional annotation, we obtained diverse predicted functions including transcription regulation, DNA repair, hydrolase, and metabolic activities. Overall, we observed that the full set of HGT candidate proteins (secreted and non-secreted) is also enriched for GHs (10 out of 44) (Fisher exact test,  $p < 0.0001$ ) when considering to the total amount of the GH genes in the genome of *V. dahliae* strain JR2 (265 out of 11,430).



**FIGURE 7. Presence and absence of all HGT candidates in all *Verticillium* species.** We searched for the presence of the HGT candidates in one strain per species that were studied previously (Depotter et al., 2017; Shi-Kunne et al., 2018) using BLASTp (with default settings). Black dots indicate the presence of an HGT candidate in the corresponding *Verticillium* species.

TABLE 1. Information on HGT candidates.

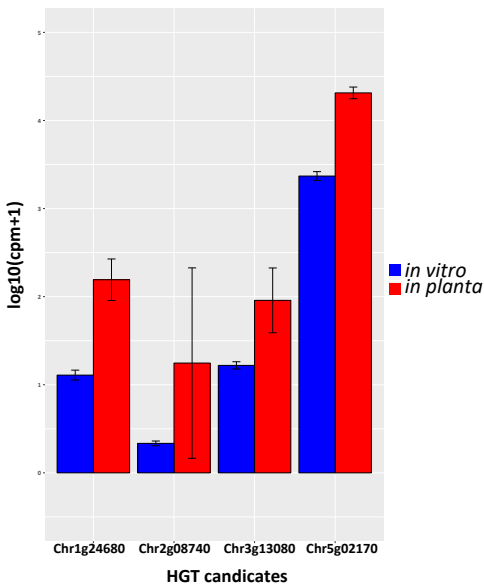
HGT ID	Gene ID	Putative donor	Putative gene product <sup>3</sup>	Functional category	Secreted	In planta expression
HGT-1	Chr4g10560	bacterial	Unknown	Unknown	No	No
HGT-2	Chr8g03340	bacterial	Pyridoxal-dependent decarboxylase	Pyridoxal phosphate binding	No	No
HGT-3	Chr1g00380	bacterial	Glycosyl hydrolase	Carbohydrate metabolism	Yes	No
HGT-4	Chr1g00710	bacterial	Unknown	Unknown	Yes	No
HGT-5	Chr1g17450	bacterial	Glycosyl hydrolase (GH27)	Carbohydrate metabolism	Yes	No
HGT-6	Chr1g24620	bacterial	Luciferase-like monooxygenase	Oxidoreductase activity	No	No
HGT-7	Chr1g24680	bacterial	Glutathione-dependent formaldehyde-activating enzyme	Detoxification of formaldehyde	No	Yes
HGT-8	Chr1g25610	bacterial	Glutathione S-transferase	Detoxification of reactive Electrophilic compounds	No	No
HGT-9	Chr1g28650	bacterial	Glycosyl hydrolases	Carbohydrate metabolism	No	No
HGT-10	Chr2g00370	bacterial	Unknown	Unknown	No	No
HGT-11	Chr2g02230	bacterial	Carbohydrate esterase (CE1)	Carbohydrate metabolism	No	No
HGT-12	Chr2g02370	bacterial	3-demethylubiquinone-9 3-methyltransferase	Unknown	No	Yes
HGT-13	Chr2g02510	bacterial	Methyltransferase	Unknown	No	No
HGT-14 <sup>1</sup>	Chr2g04700	bacterial	Glycosyl transferase family group 2	Carbohydrate metabolism	No	No
HGT-15	Chr2g07220	bacterial	Pyridine nucleotide-disulphide oxidoreductase	Oxidoreductase activity	No	No
HGT-16	Chr2g08740	bacterial	Enolase	Carbohydrate metabolism	No	Yes
HGT-17	Chr3g05430	bacterial	Chondroitin AC/alginatase lyase	glycosaminoglycans degradation	Yes	Yes
HGT-18	Chr3g06830	bacterial	Acetyltransferase (GNAT) family	transcriptional regulation	No	No
HGT-19	Chr3g12010	bacterial	Glycosyl hydrolase (GH74)	Hydrolase activity	Yes	Yes
HGT-20	Chr3g13080	bacterial	Phosphoinositide phospholipase	Hydrolase activity	Yes	Yes
HGT-21	Chr4g04370	bacterial	Membrane protein	Unknown	No	No
HGT-22	Chr4g11370	bacterial	Glycosyl hydrolase (GH74)	Carbohydrate metabolism	No	No
HGT-23	Chr4g11740	bacterial	Amidohydrolase	Hydrolase activity	Yes	No
HGT-24	Chr4g11940	bacterial	Carbohydrate esterases (CE1)	Carbohydrate metabolism	Yes	No
HGT-25	Chr5g01710	bacterial	L-asparaginase	Hydrolase activity	No	No



HGT-26 <sup>2</sup>	Chr5g02170	plant	Plant natriuretic peptide		Virulence factor	Yes	Yes
HGT-27	Chr5g05890	bacterial	Glycosyl hydrolases (GH43/29)		Carbohydrate metabolism	Yes	No
HGT-28	Chr5g09290	bacterial	Protein of unknown function		Unknown	No	Yes
HGT-29	Chr6g02080	bacterial	Bacteriocin-protection, Ydel or OmpD-Associated protein		Oxidase activity	No	No
HGT-30	Chr6g03000	bacterial	Pyridoxal-phosphate dependent enzyme		Metabolic activity	No	No
HGT-31	Chr6g05320	bacterial	OsmC-like protein		Oxidative stress regulation	No	No
HGT-32	Chr6g05650	bacterial	Luciferase-like monooxygenase		Oxidoreductase activity	No	Yes
HGT-33	Chr6g10680	bacterial	Acetyltransferase (GNAT) family		N-acetyltransferase activity	No	Yes
HGT-34	Chr7g01340	bacterial	M6 family metalloprotease		Peptidase activity	No	No
HGT-35	Chr7g01630	bacterial	Glycosyl hydrolases (GH136)		Carbohydrate metabolism	No	No
HGT-36	Chr7g03210	bacterial	Aminotransferases		Catalytic activity	No	No
HGT-37	Chr7g03340	bacterial	Alpha/beta hydrolase		Hydrolytic activity	No	Yes
HGT-38	Chr7g03590	bacterial	Insecticide toxin TcdB		Unknown	No	No
HGT-39	Chr8g00620	bacterial	Glycerate kinase		Glycerate kinase activity	No	No
HGT-40	Chr8g06130	bacterial	Unknown		Unknown	No	No
HGT-41	Chr8g08880	bacterial	Nucleotidyltransferase		DNA repair	No	No
HGT-42	Chr8g11000	bacterial	Glycosyl hydrolase (GH67)		Carbohydrate metabolism	Yes	Yes
HGT-43	Chr8g11270	bacterial	Glycosyl hydrolases (GH43/26), (CBM42)		Carbohydrate metabolism	Yes	No
HGT-44	Chr8g01080	bacterial	Peptidase m20		Peptidase activity	No	No

<sup>1</sup> previously identified candidate (Klosterman et al., 2011)<sup>2</sup> previously identified candidate (de Jonge et al., 2012)<sup>3</sup> functional annotation was performed by searching for conserved domains in each of these proteins using Interproscan (Jones et al. 2014)

To investigate the putative role of the HGT-derived genes in plant pathogen interactions, we compared expression patterns *in planta* and *in vitro*. It was previously shown that *Ave1* is expressed *in planta* during interaction of *V. dahliae* with *Nicotiana benthamiana* plants (de Jonge et al., 2012). We assessed transcription (RNA-seq) data of *V. dahliae* during colonization of *A. thaliana* and found that 12 HGT candidates showed *in planta* expression, one of which is *Ave1*. Four of these 12 were found to be induced when compared with the expression in *in vitro*-cultured mycelium (Figure 8). This suggests that besides *Ave1*, three additional HGT candidates might play a role during host colonization. Moreover, one of the three induced candidates encodes a secreted protein.

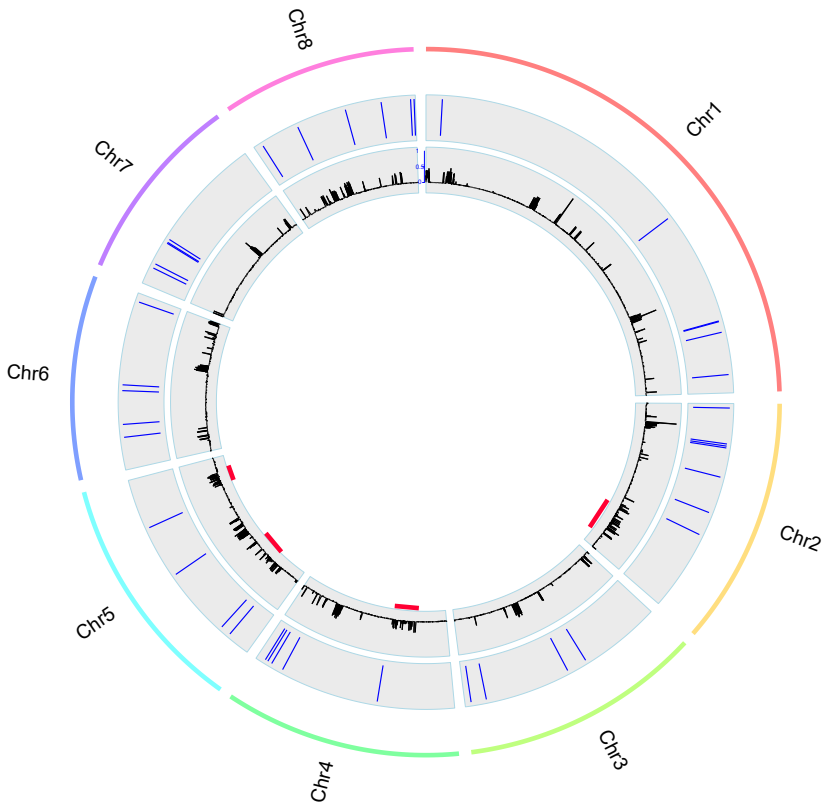


**FIGURE 8. Pair-wise comparison of HGT candidates with differential expression *in vitro* and *in planta*.** Gene expressions are depicted for *V. dahliae* strain JR2 cultured in liquid medium and upon *A. thaliana* colonization, respectively. Bars represent the mean gene expression with standard deviations. The significance of difference in gene expression was calculated using t-tests relative to a threshold (TREAT) of log<sub>2</sub>-fold-change  $\geq 1$  (McCarthy and Smyth, 2009).

### *V. dahliae* HGT candidates mostly localize at repetitive genomic regions

In principle, foreign DNA can be incorporated anywhere in a recipient genome as long as it does not disrupt an essential genomic element (Husnik and Mccutcheon, 2017). Several studies suggest that horizontally transferred sequences are often integrated into genomic regions that are enriched in transposable elements (TE) (Gladyshev et al., 2008; McNulty et al., 2010; Acuña et al., 2012; Pauchet and Heckel, 2013; Husnik and Mccutcheon, 2017). Previously, we observed that *Ave1* is located in a lineage-specific (LS) region of *V. dahliae* that is enriched in repeats (de Jonge et al., 2013; Faino et al., 2016). In order to assess whether the other HGT candidates are lineage-specific and associated with repeats as well, we plotted genomic repeat densities and the genomic location of all HGT candidates. Although only *Ave1* is located at a genuine LS region (Depotter et

al. 2018), most of the HGT candidates (34 out of 44) reside in proximity (within 1 kb) to a repetitive sequence (Figure 9). Subsequently, we compared repeat counts within 1 kb genomic windows around each of the 44 HGT candidates to the repeat counts in 1 kb genomic windows of 44 randomly picked genes (1000 permutations), which confirmed that the HGT candidates are significantly more often adjacent to repeats than expected by chance ( $P = 0.0331$ ).



**FIGURE 9. Genomic location of the HGT candidates.** The outer lane represents the chromosomes of *V. dahliae* strain JR2. The middle lane shows the relative position of the HGT candidates in each chromosome. The inner lane shows the repeat density. The red lines indicate the locations of the lineage specific (LS) regions (Depotter et al., 2018).

## Discussion

In this study, we systematically searched for evidence of inter-kingdom HGT events in the genome of *V. dahliae* using an Alien Index (AI) based method. Subsequently, we verified HGT candidates using a phylogenetic approach. In general, there are two classical ways to identify HGT candidates: intrinsic and extrinsic methods. Intrinsic methods focus on

patterns in the primary structure of genes and genome sequences and aim to find genes or genomic regions with composition patterns that differ significantly from the rest of the genome. For example, deviation in GC content and codon usage can be considered a sign for horizontal acquisition (Lawrence and Ochman, 1997). In contrast, extrinsic methods focus on comparing similarity metrics between closely related and distant taxa. For example, when a gene from a species of interest (recipient) shows higher similarity (lower E-value, higher bit-score in BLAST) to sequences from distantly related (donor) species than to genes of close relatives, this gene may be horizontally acquired. Another frequently applied extrinsic approach is to assess the phylogenetic distribution of genes across a panel of diverse organisms. These phylogeny-based methods are generally thought to be more accurate (Poptsova and Gogarten, 2007; Poptsova, 2009). In principle, an HGT event will create a discrepancy between the gene and species trees. However, phylogeny-based methods are much more computationally intensive, because they require ortholog predictions, sequence alignments and phylogeny reconstructions (Dupont and Cox, 2017). Thus, we chose to utilize an AI based method to pre-select candidates that can be further analyzed through phylogenetic analyses.

Although the AI method provided a rapid identification of HGT candidates, some limitations of the method should be kept in mind. The AI method relies on the quality of the reference database, the accuracy of the taxonomic assignment and the diversity of taxa covered by the reference database. The two HGT candidates (*HGT-1* and *HGT-2*) that appear to be *Verticillium*-specific could also be explained by the absence of the ortholog-containing species. In other words, the more closely related species are included in the database, the less *Verticillium*-specific HGT candidates might be identified. Furthermore, AI based methods rely on pre-defining in- and out- groups and can only identify HGT candidates that are absent in the in-group species. In our case, the method is unable to identify candidates that were acquired before the formation of ascomycetes, and candidates that were also acquired by fungal non-ascomycete species independently at more recent events from the same donors.

Nevertheless, the 42 HGT candidates that were also found in other ascomycete species were most likely acquired during ancient transfer events. The extent of gene transfer from prokaryotes to fungi is highly different across diverse ascomycete fungal lineages. A whole-genome study of HGT in *Aspergillus fumigatus* revealed that 40% of 189 transferred regions (containing 214 genes) were of bacterial origin (Mallet et al., 2010). In contrast, a study with a similar objective found only 11 genes with a convincing signature of transfer from bacterial origin in the genomes of three *Colletotrichum* species (Jaramillo et al., 2015). Three of the 11 *Colletotrichum* HGT candidates were found in *V. dahliae* (strain VdLs17) and *V. alfalfae* (strain VaMs102) (Klosterman et al., 2011). In our study, we also found these three candidates (*HGT-14*, *HGT-15* and *HGT-44*) in *V. dahliae* strain JR2.

We found that HGT candidates are enriched for genes encoding secreted proteins, many of which belong to a family of carbohydrate-active enzymes (CAZymes) or glycoside hydrolases (GHs). GHs play a fundamental role in the decomposition of plant biomass (Murphy et al., 2011), which suggests that bacterial genes might contribute to the plant-associated life styles of *Verticillium* species as well as of many other ascomycetes. Although these *V. dahliae* GH genes are not induced during infection of *A. thaliana* when compared with *in vitro* expression data, *V. dahliae* is able to cause disease in hundreds of plant species (Fradin and Thomma, 2006; Inderbitzin and Subbarao, 2014), and it might be possible that the GH genes are induced during interaction with other host plants. Alternatively, these proteins might play roles during life stages outside the host. Overall, when assessing transcription (RNA-seq) data of *V. dahliae*-infected *A. thaliana*, we identified four HGT candidates (including *Ave1*) that were induced when compared with expression in *in vitro*-cultured mycelium. One of the induced candidates encodes a secreted protein that is predicted to be a phosphoinositide phospholipase, a key metabolic enzyme that is needed by all living organisms to hydrolyze phospholipids into fatty acids and lipophilic substances (Ghannoum, 2000). Phospholipases are also signaling molecules that elicit stress tolerance and host immune responses in fungi (Ghannoum, 2000; Soragni et al., 2001; Köhler et al., 2006). Phospholipases have been studied in some plant pathogenic fungi, such as *Fusarium graminearum* (Zhu et al., 2016) and *Magnaporthe oryzae* (Ramanujam and Naqvi, 2010; Zhang et al., 2011; Yin et al., 2016). The phospholipase (FgPLC1) regulates development, stress responses and pathogenicity of *F. graminearum* (Zhu et al., 2016). Collectively, this may suggest that the *V. dahliae* phosphoinositide phospholipase might play an important role during interaction with host plants.

## Conclusions

In this study, we applied a conservative approach combining an Alien Index (AI) method and phylogenetic analysis to analyze inter-kingdom HGT to *Verticillium dahliae*. Besides the previously identified effector gene *Ave1* and a glucosyltransferase-encoding gene, we revealed 42 additional HGT candidates, all of which are of bacterial origin, indicating a high number of inter-kingdom gene acquisitions.

## Materials & Methods

### Construction of local protein database and protein BLAST

The protein database used for BLAST analyses was generated using the reference proteomes from UniProtKB excluding viruses (downloaded at 07-07-2016). Entries from UniProt were renamed to include their UniProt ID and NCBI taxonomy ID. Sequences shorter than ten amino acids were removed from the database (E-value = 0.001, max target seqs = 1,000). We searched for the presence of the HGT candidates in one strain per species that were studied previously (Depotter et al., 2017; Shi-Kunne et al., 2018). Protein sequences of each HGT candidate were queried against the proteome of each strain using BLASTp (with default settings).

### Identification of HGT candidates

Alien Index (AI) scores were calculated using a custom-made python script which applied the following formula:  $AI = (\ln(bbhG + 1 \times 10^{-200}) - \ln(bbhO + 1 \times 10^{-200}))$ , where bbhG and bbhO are the E-values ( $<10^{-3}$ ) of the best BLAST hit from the in-group and out-group, respectively. Best BLAST hits were determined by the highest bit score. Genes with AI score  $>1$  were classified as positive AI genes. For each AI positive gene, homologs were selected based on the BLAST criteria that at least 70% of query length is covered by at least 60% of hit sequence using a custom-made Python script. These homologs were then aligned using MAFFT with default setting (Katoh and Standley, 2013). Alignments were subsequently curated using Gblocks (Castresana, 2000) with non-stringent parameters. Phylogenetic trees of the curated alignments were constructed using FastTree (Price et al., 2010) with gamma likelihood and wag amino acid substitution matrix. Phylogenetic trees were automatically evaluated using a custom made Python script that removes phylogenetic trees with candidate genes that are directly adjacent to non-*Verticillium* fungal species (in-group species) rather than non-fungal species (out-group species). Subsequently, new phylogenetic trees (for manual inspection) were reconstructed using RAxML (Stamatakis, 2014) with automatic substitution model determination for branching values.

### Gene expression analysis

To obtain RNA-seq data for *V. dahliae* grown in culture medium, isolate JR2 was grown for three days in potato dextrose broth (PDB) with three biological replicates. To obtain RNA-seq data from *V. dahliae* grown *in planta*, roots of three-week-old *A. thaliana* (Col-0) plants were dipped in a suspension of  $10^6$  conidiospores per mL of water for 5 minutes. After root inoculation, plants were grown in individual pots in a greenhouse under a cycle of 16 h of light and 8 h of darkness, with temperatures maintained between 20 and 22°C

during the day and a minimum of 15°C overnight. Three pooled samples (10 plants per sample) of complete flowering stems were used for total RNA extraction. Total RNA was extracted based on TRIzol RNA extraction (Simms et al., 1993). cDNA synthesis, library preparation (TruSeq RNA-Seq short insert library), and Illumina sequencing (single-end 50 bp) was performed at the Beijing Genome Institute (BGI, Hong Kong, China). In total, ~2 Gb and ~1.5 Gb of filtered reads were obtained for the *V. dahliae* samples grown in culture medium and *in planta*, respectively. RNAseq data were submitted to the SRA database under the accession number: SRP149060.

The RNA sequencing reads were mapped to their previously assembled genomes using the Rsubread package in R (Liao et al., 2013). The comparative transcriptomic analysis was performed with the package edgeR in R (v3.4.3) (Robinson et al., 2010; McCarthy et al., 2012). Genes are considered differently expressed when P-value < 0.05 with a log<sub>2</sub>-fold-change ≥ 1. P-values were corrected for multiple comparisons according to Benjamini and Hochberg (Benjamini and Hochberg, 1995).

## Acknowledgements

Work in the laboratories of B.P.H.J.T. and M.F.S is supported by the Research Council Earth and Life Sciences (ALW) of the Netherlands Organization of Scientific Research (NWO).





# Chapter 7

## ***In silico* prediction and characterisation of secondary metabolite clusters in the plant pathogenic fungus *Verticillium dahliae***

Xiaoqian Shi-Kunne\*, Roger de Pedro Jové\*, Jasper R.L. Depotter†, Malaika K. Ebert†, Michael F. Seidl# and Bart P.H.J. Thomma#

\*These authors contributed equally to this work.

†These authors contributed equally to this work.

#These authors contributed equally to this work.

A modified version of this chapter has been submitted

## Abstract

Fungi are renowned producers of natural compounds, also known as secondary metabolites (SMs) that display a wide array of biological activities. Typically, the genes that are involved in the biosynthesis of SMs are located in close proximity to each other in so-called secondary metabolite clusters (SMCs). Many plant-pathogenic fungi secrete SMs during infection in order to promote disease establishment, for instance as cytotoxic compounds. *Verticillium dahliae* is a notorious plant pathogen that can infect over 200 host plants worldwide. However, the SM repertoire of this vascular pathogen remains mostly uncharted. To unravel the SM potential of *V. dahliae*, we performed *in silico* predictions and in-depth analyses of its SM clusters (SMC). We identified 25 potential SMCs in the *V. dahliae* genome, including loci that can be implicated in DHN-melanin, ferricrocin, triacetyl fusarinine and fujikurin production.

## Introduction

Filamentous fungi are known for their ability to produce a vast array of distinct chemical compounds that are also known as secondary metabolites (SMs) (Keller et al., 2005). In contrast to primary metabolites, SMs are often considered as non-essential for fungal growth, development or reproduction. However, SMs can be crucial for long-term survival in competitive fungal niches (Fox and Howlett, 2008; Pons, 2015; Derntl et al., 2017). SMs produced by plant pathogenic fungi are of particular interest as they may contribute to virulence, leading to crop losses and threatening food security (Pons 2015; Pusztahelyi et al. 2015). For example, T-toxin from *Cochliobolus heterostrophus*, a maize pathogen that caused the worst epidemic in U.S. agricultural history, has been reported to be a crucial pathogenicity factor (Inderbitzin et al. 2010).

Fungal SMs are classified into four main groups based on core enzymes and precursors involved in their biosynthesis: polyketides, non-ribosomal peptides, terpenes and indole alkaloids (Keller et al. 2005). Production of the chemical scaffold of each class requires core enzymes named polyketides synthases (PKSs), non-ribosomal peptide synthetases (NRPSs), terpene cyclases and dimethylallyl tryptophan synthases, respectively. Additionally, hybrid enzymes such as PKS-NRPSs have been identified as builders of structurally complex molecules with combined properties (Boettger and Hertweck, 2013). PKSs and NRPSs are the most abundant and are extensively studied in fungi (Cox, 2007). PKSs can be further divided into three different types (I, II and III), of which type I PKSs and type III PKSs are found in fungi. Type I PKSs are predominant in fungi whereas type III PKSs are found only rarely (Cox, 2007; Gallo et al., 2013; Hashimoto et al., 2014). Genes involved in the synthesis of SMs are frequently located in close proximity to each other, forming so-called secondary metabolite clusters (SMCs) (Keller and Hohn, 1997; Brakhage and Schroeckh, 2011; Wiemann and Keller, 2014). Most of these SMCs contain one biosynthetic core gene that is flanked by transporter proteins, transcription factors, and genes encoding tailoring enzymes that modify the SM structure (Keller and Hohn 1997; Keller et al. 2005).

The genomics era has provided new tools to study fungal SMs and their biosynthesis at the whole genome scale (Wiemann and Keller, 2014; Medema and Fischbach, 2015). The distinctive traits of gene clusters (e.g. gene distance) and the conserved signatures of core genes (e.g. conserved domains) can be exploited to identify putative loci involved in SM production. Moreover, phylogenetic and comparative genomics analyses are very informative as the number of fungal genomes and characterized SM pathways increases. These two approaches are very helpful to identify gene clusters that are involved in the production of SMs that have been characterized in other fungal species and allow subsequent predictions of identical or related compounds that a particular fungal species might produce (Medema et al. 2013; Cairns and Meyer 2017).

The fungal genus *Verticillium* contains nine haploid species plus the allodiploid *V. longisporum* (Inderbitzin et al., 2011a; Depotter et al., 2017). These ten species are phylogenetically subdivided into two clades; Flavexudans and Flavnonexudans (Inderbitzin et al., 2011a; Shi-Kunne et al., 2018). The Flavnonexudans clade comprises *V. nubilum*, *V. alfalfae*, *V. nonalfalfae*, *V. dahliae*, and *V. longisporum*, while the Flavexudans clade comprises *V. albo-atrum*, *V. isaacii*, *V. tricorpus*, *V. klebahnii* and *V. zaregamsianum* (Inderbitzin et al., 2011a). Among these *Verticillium* spp., *V. dahliae* is the most notorious plant pathogen that is able to cause disease in hundreds of plant species (Fradin and Thomma, 2006; Inderbitzin and Subbarao, 2014). Furthermore, *V. albo-atrum*, *V. alfalfae*, *V. nonalfalfae* and *V. longisporum* are pathogenic, albeit with narrower host ranges (Inderbitzin and Subbarao, 2014). Although the remaining species *V. tricorpus*, *V. zaregamsianum*, *V. nubilum*, *V. isaacii* and *V. klebahnii* have incidentally been reported as plant pathogens, they are mostly considered as saprophytes that thrive on dead organic material and their incidental infections should be seen as opportunistic (Ebihara et al., 2003; Inderbitzin et al., 2011a; Gurung et al., 2015). Previously, studies of three genes that are involved in SM biosynthesis in *V. dahliae* suggested that SMs may play a role in *V. dahliae* virulence. The deletion mutants of the putative secondary metabolism regulators *VdSge1* (Santhanam and Thomma, 2012) and *VdMcm1* (Xiong et al., 2016) displayed reduced virulence when compared with the wild type *V. dahliae* strain. Likewise, a reduction in virulence was observed for deletion mutants of the cytochrome P450 monooxygenase *VdCYP1* (Zhang et al., 2016), a common tailoring enzyme in SMC production. In this study, we conducted an *in silico* analysis to unravel the potential secondary metabolism of *V. dahliae* by making use of the gapless genome assembly of strain JR2 (Faino et al., 2015)

## Results & Discussion

### The *V. dahliae* strain JR2 genome contains 25 putative secondary metabolite clusters

To assess the potential secondary metabolism of *V. dahliae* strain JR2, we mined its genome sequence to predict SMCs using antiSMASH (Weber et al., 2015). A total of 25 putative SMCs were predicted, containing a total of 364 genes within their boundaries (Figure 1, Table 1). The putative SMCs were classified as nine type I PKSs, one type III PKS, one PKS-NRPS, three NRPSs and four terpenes. Seven clusters were classified as “other”, a generic class of SMCs containing core enzymes with unusual domain architecture, also known as non-canonical. We found that each of these SMCs contains one core gene. Disrupted SMCs frequently occur in the genomes of filamentous fungi (Collemare et al., 2014). However, none of the clusters identified in *V. dahliae* strain JR2 is evidently disrupted, as all identified clusters have no insertions of transposable elements and include genes encoding tailoring

enzymes such as methyltransferases, cytochrome P450 or dehydrogenases (Cacho et al. 2015). Several clusters also comprise transporter and transcription factor encoding genes that might be involved in SM secretion and local gene cluster regulation respectively (Collemare et al., 2014) (Table 1). Collectively, these results suggest that all the analyzed SMCs of *V. dahliae* strain JR2 are potentially functional.

**TABLE 1. Predicted secondary metabolite gene clusters (SMCs) in *V. dahliae* strain JR2.**

SM class	Key biosynthetic gene	Gene name	Sub-telomeric <sup>5</sup>	Transporter	Transcription factor	Cluster location	No. of genes
T1 PKS <sup>1</sup>	Chr1g1100	VdPks1	NO	YES	YES	1:3545717:3591771	16
	Chr1g1588	VdPks2	NO	NO	YES	1:5097044:5143749	15
	Chr2g0045	VdPks3	YES	YES	YES	2:92065:145583	18
	Chr2g0695	VdPks4	NO	YES	YES	2:2119540:2167273	14
	Chr2g0719	VdPks5	NO	YES	NO	2:2182651:2230558	20
	Chr3g0093	VdPks6	NO	YES	NO	3:316303:358028	14
	Chr4g1125	VdPks7	NO	NO	NO	4:3709416:3756014	17
	Chr5g1015	VdPks8	NO	YES	YES	5:3384693:3432789	11
	Chr8g1031	VdPks9	YES	YES	YES	8:3058903:3104627	21
PKS-NRPS <sup>2</sup>	Chr1g2388	VdHyb1	NO	YES	YES	1:7618782:7670067	20
NRPS <sup>3</sup>	Chr3g1324	VdNrps1	YES	YES	YES	3:4066099:4121170	19
	Chr6g0600	VdNrps2	NO	YES	YES	6:1768180:1819114	18
	Chr7g1025	VdNrps3	YES	NO	YES	7:3179393:3227591	13
T3PKS <sup>4</sup>	Chr4g0955	VdT3pks	NO	YES	YES	4:3205819:3247084	14
Terpene	Chr1g1245	VdTerp1	NO	NO	NO	1:4010610:4031875	7
	Chr1g1723	VdTerp2	NO	NO	NO	1:5472865:5493695	8
	Chr2g0913	VdTerp3	NO	NO	NO	2:2797712:2819868	10
	Chr7g0295	VdTerp4	NO	NO	YES	7:844777:866572	5
Other	Chr2g0388	Other1	NO	YES	YES	2:1232486:1276127	15
	Chr4g0468	Other2	NO	YES	YES	4:1587321:1632662	16
	Chr5g1148	Other3	YES	YES	YES	5:3902983:3948100	14
	Chr5g1176	Other4	YES	YES	YES	5:3993259:4036240	15
	Chr6g0066	Other5	YES	YES	YES	6:150534:194469	14
	Chr6g0109	Other6	YES	YES	NO	6:289905:335580	18
	Chr8g0117	Other7	YES	YES	NO	8:289821:333602	13

<sup>1</sup> T1PKS = type 1 polyketide synthase

<sup>2</sup> NRP-PKS = hybrid polyketide synthase-non-ribosomal peptide synthase.

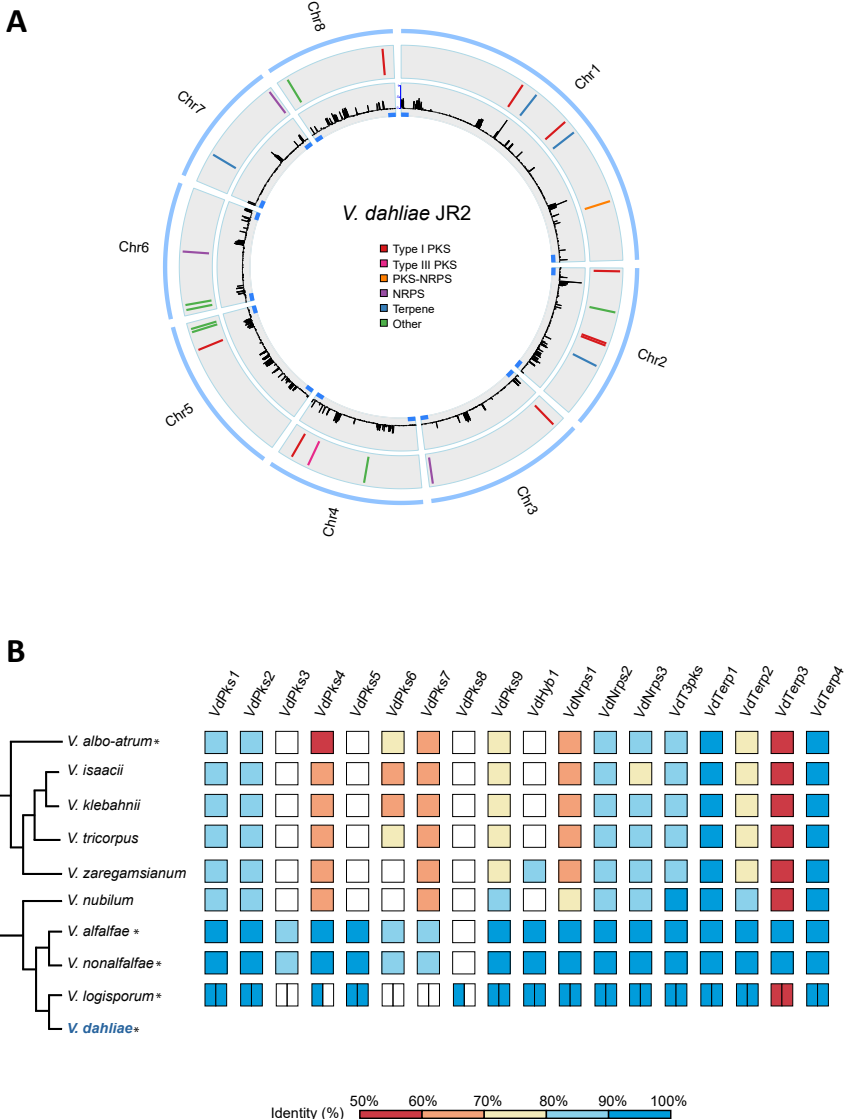
<sup>3</sup> NRPS = non-ribosomal peptide synthase

<sup>4</sup> T3PKS = type 13polyketide synthase

<sup>5</sup> Sub-telomeric clusters were defined as any cluster predicted to reside within 300 kb of a chromosome end

In several species, secondary metabolite genes are enriched at chromosomal ends adjacent to the telomeres (Cairns and Meyer 2017; Farman 2007; McDonagh et al. 2008). Thus, we assessed whether the SMCs of *V. dahliae* are located in sub-telomeric regions,

here defined as within 300 kb from the chromosomal end. We found that 36% of the predicted clusters are in sub-telomeric regions (Figure 1A, Table 1). As these sub-telomeric regions harbour 1,606 genes in total, which is 13.8% of total gene repertoire (Table 1) they are significantly enriched ( $\chi^2$ -test,  $p < 0.0001$ ) in secondary metabolism genes (Figure 1A).



**FIGURE 1. *In silico* predicted *V. dahliae* SMCs.** (A) Genomic location of *V. dahliae* SMCs. The outer blue lane represents the chromosomes. The middle grey lane shows the relative position of the predicted SMCs on each chromosome. The inner grey lane shows the repeat density in the JR2 genome. The blue rectangles indicate the regions that are defined as sub-telomeric (300 kb from each chromosomal end). (B) Conservation of *Verticillium dahliae* core SMC genes throughout the *Verticillium* genus. The colour gradient represents the % identity range of the high scoring. Species described as plant pathogens are indicated with an asterisk.

To check whether the SMCs identified in *V. dahliae* strain JR2 are also present in other *V. dahliae* strains, we assessed the presence/absence of the core enzymes in 22 *V. dahliae* strains (Klosterman et al., 2011; Faino et al., 2015; Kombrink et al., 2017; Depotter et al., 2018). Among these 22 strains is the gapless genome assembly of strain VdLs17 (Klosterman et al., 2011; Faino et al., 2015) and the nearly complete genome assemblies of strains CQ2 and 855 (Depotter et al., 2018). The remaining genome assemblies are considerably fragmented, with over 500 contigs for each of the assemblies. Nevertheless, we found that each of the core enzymes is present in all 22 strains, except for VdPks8 which was next to JR2 only found in VdLs17, CQ2 and 855. However, the absence of VdPks8 from the other strains may be due to the fragmented genome assemblies. Subsequently, we assessed the genome assemblies of strains VdLs17, CQ2 and 855 for the presence of complete clusters, revealing that all SMCs identified in *V. dahliae* strain JR2 are also found in the genomes of these three strains, therefore suggesting that SMCs are highly conserved in *V. dahliae* strains.

To examine whether the SMCs identified in *V. dahliae* strain JR2 are also present in other *Verticillium* spp., we queried core enzymes of each cluster using BLAST against the proteomes of previously published *Verticillium* spp. (Depotter et al., 2017; Shi-Kunne et al., 2018). In total, 12 SMC core enzymes are present in all *Verticillium* spp., nine of which showed considerable sequence conservation (>80% sequence identity) (Figure 1B). The other 15 core enzymes are not present in all species, but display a mosaic presence/absence pattern with presence in at least two other species. Of these, VdPks7 is conserved in all species except for *V. longisporum* (Figure 1B). VdPks3 is only conserved in the closely related species *V. alfalfae* and *V. nonalfalfae* and VdPks5 is conserved in all pathogenic species except *V. albo-atrum*. Interestingly, the VdPks8 core enzyme was only found in a single copy in the hybrid *V. longisporum* genome, presumably derived from its *V. dahliae* progenitor. Thus, based on the widespread presence of these core genes within the *Verticillium* genus, we predict that most of the SMCs are conserved throughout this genus.

### Phylogenomic analysis of *V. dahliae* secondary metabolite core enzymes

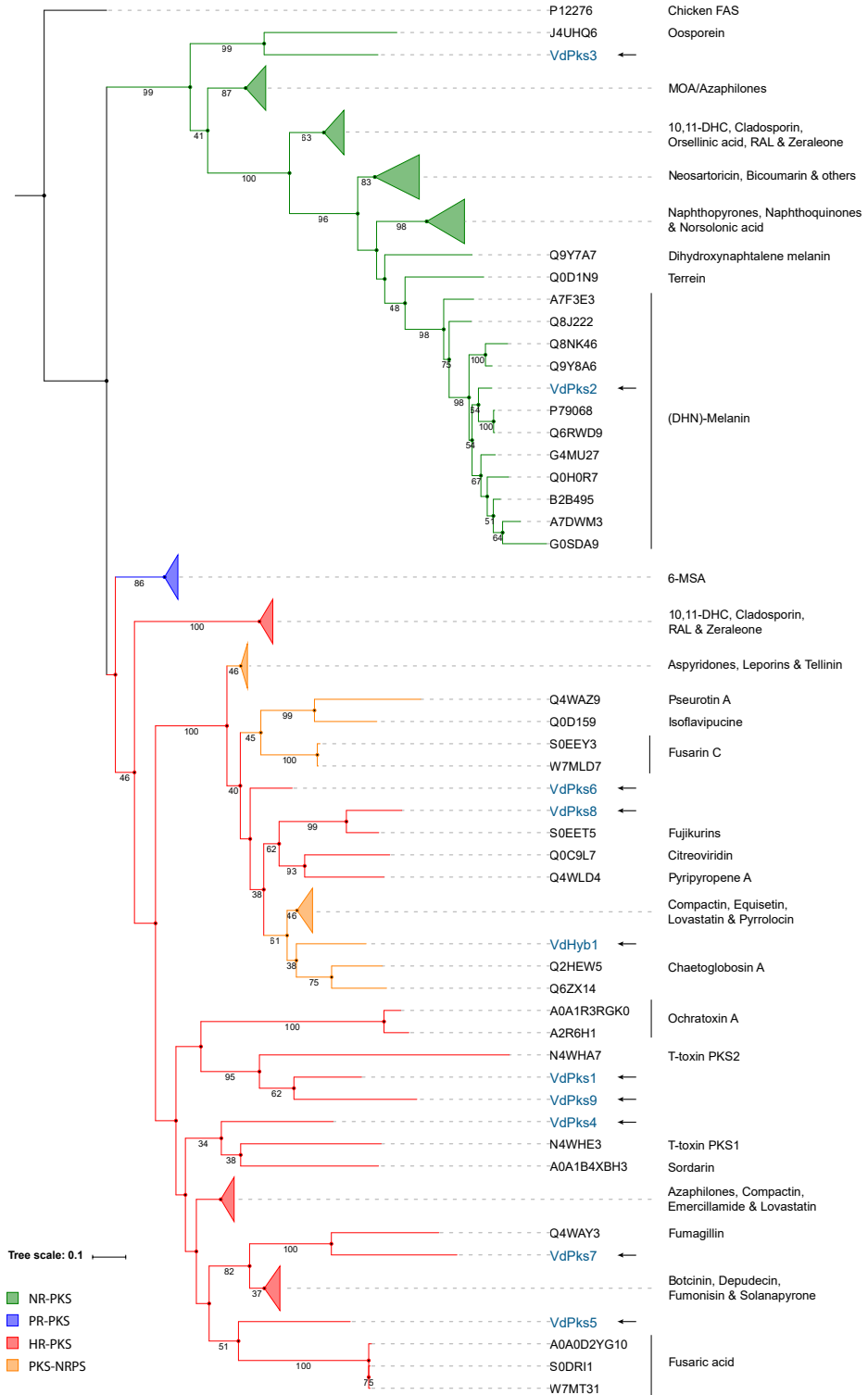
In ascomycetes, type I PKSs, NRPSs and PKS-NRPSs are known to produce most of the SMs that are involved in virulence (Pusztahelyi et al. 2015). Thus, we focused on identifying putative functions of type I PKSs, NRPSs and PKS-NRPSs in *V. dahliae* with a phylogenomics approach. To this end, we aligned KS domains of *V. dahliae* PKSs to KS domains of functionally described PKSs, and subsequently constructed a phylogenetic tree that comprises three major clades that correspond to the NR-PKSs, PR-PKSs and HR-PKSs, respectively (Figure 2). The NR-PKS clade contains two predicted *V. dahliae* PKSs, VdPks2 and VdPks3. VdPks2 clusters with PKSs that have been implicated in dyhydroxynaphthalene (DHN)-melanin formation (Tsuji et al., 2002; Yu et al., 2015). VdPks3

clustered with the Orsellinic acid synthase OpS1, which is involved in production of the toxic metabolite oosporein by the entomopathogenic fungus *Beauveria bassiana* (Feng et al., 2015). The HR-PKS clade showed that four out of the seven predicted *V. dahliae* HR-PKSs (VdPks1, VdPks7, VdPks8 and VdPks9) grouped with previously described enzymes. VdPks7 and VdPks8 clustered with the fumagillin synthase from *Aspergillus fumigatus* and fujikurin synthase from *Fusarium fujikuroi*, respectively (Lin et al., 2013; von Bargen et al., 2015; Niehaus et al., 2016). VdPks1 and VdPks9 clustered with the T-toxin synthase PKS2 from *Cochliobolus heterostrophus* (Inderbitzin et al. 2010). VdPks4 grouped in a clade that contains sdnO and PKS1 synthase, which are involved in the production of Sordarin by *Sordaria araneosa* (Kudo et al., 2016), an antifungal agent that inhibits protein synthesis in fungi by stabilizing the ribosome/EF2 complex (Justice et al., 1998), and in T-toxin production in *C. heterostrophus* (Inderbitzin et al. 2010), respectively. VdPks5 is in a clade that only contains FUB1 fusaric acid synthase orthologs of three *Fusarium* spp. (Brown et al., 2015). The remaining *V. dahliae* PKS core enzyme, VdPks6, is not directly grouping adjacent to any previously described enzyme (Figure 2). Thus, eight of the nine *V. dahliae* PKS core enzymes (VdPks1, VdPks2, VdPks3, VdPks4, VdPks5, VdPks7, VdPks8, VdPks9 and VdHyb1) group with previously characterised enzymes, thereby allowing us to infer their putative function.

**FIGURE 2. Phylogenetic tree of type I PKS and PKS-NRPS enzymes.** KS domains of PKS and PKS-NRPS enzymes were aligned to construct the maximum likelihood tree with 100 bootstrap replicates. The chicken fatty acid synthase (Chicken FAS) sequence was used as outgroup. Only bootstrap values above 30 are shown below the branches. *V. dahliae* KS domains are highlighted in blue and indicated with an arrow. Protein codes correspond to Uniprot IDs.

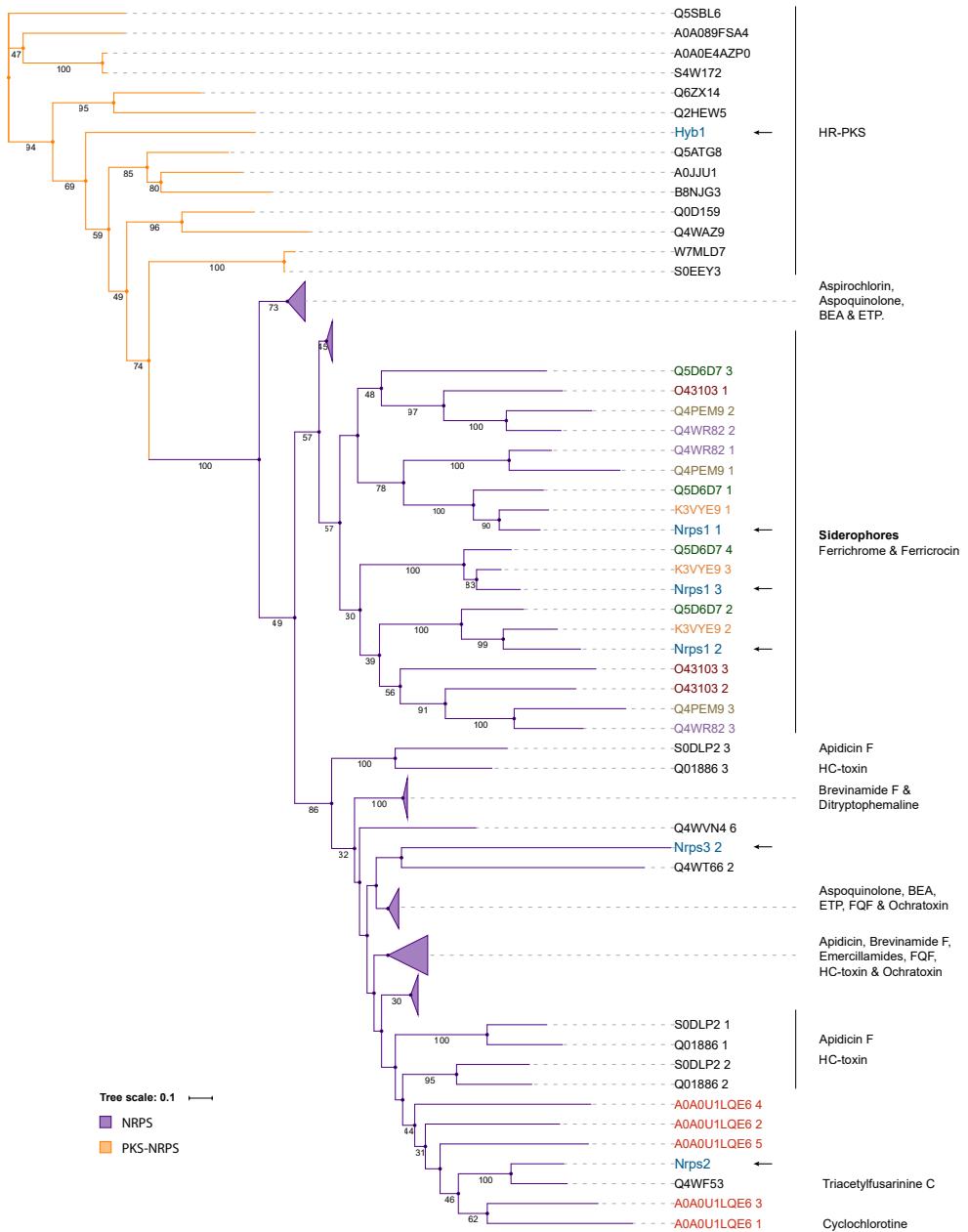


## Secondary metabolite prediction and characterisation



Like for PKSs, we similarly performed phylogenomic analysis to get more insight into the putative products of NRPSs. The conserved A-domain sequences of *V. dahliae* NRPS core enzymes were aligned with previously described enzymes of other fungal species to construct a phylogenetic tree. A distinct clade clearly separated NRPSs from the PKS-NRPSs in the phylogenetic tree (Figure 3). VdNprs1 grouped in a clade with NPS2, which is involved in the synthesis of the intracellular siderophore ferricrocin of *Fusarium pseudograminearum* and *Cochliobolus heterostrophus* (Tobiasen et al., 2007; Sieber et al., 2014; Oide et al., 2015). VdNprs2 clusters with NRPS4, which is responsible for the synthesis of the extracellular siderophore triacetylfusarinine C (TAFC) by *A. fumigatus* (Schrettl et al., 2007). The clade that contains VdNprs3 has low bootstrap values and long branches, indicating considerable divergence of this enzyme (Figure 3). Thus, only two of the *V. dahliae* NRPS core enzymes (VdNprs1 and VdNprs2) group with previously characterised enzymes.

The fusion of PKS and NRPS domains results in PKS-NRPS enzymes that stand out due to their structural complexity (Boettger and Hertweck, 2013). In the genome of *V. dahliae* (strain JR2), only one PKS-NRPS (VdHyb1) was detected. Since it contained a KS domain characteristic for HR-PKSs and an A-domain as commonly observed in NRPSs, we included VdHyb1 in both phylogenetic trees. In the PKS phylogeny (Figure 2) VdHyb1 was found in the same clade (38%, bootstrap value) as the PKS-NRPSs chaetoglobosin A synthase from *Chaetomium globosum* and the avirulence protein ACE1 from *Magnaporthe grisea* (Collemare et al., 2008; Ishiuchi et al., 2013). In contrast, VdHyb1 did not cluster with any previously characterized PKS-NRPS in the A-domain phylogenetic tree (Figure 3).



**FIGURE 3. Phylogenetic tree of NRPS and PKS-NRPS enzymes.** A-domains of NRPS and PKS-NRPS enzymes were aligned to construct the maximum likelihood tree with 100 bootstrap replicates. Only bootstrap values over 30 are shown below the branches. *V. dahliae* A-domains are highlighted in blue and indicated with an arrow

## Comparative analysis of gene clusters

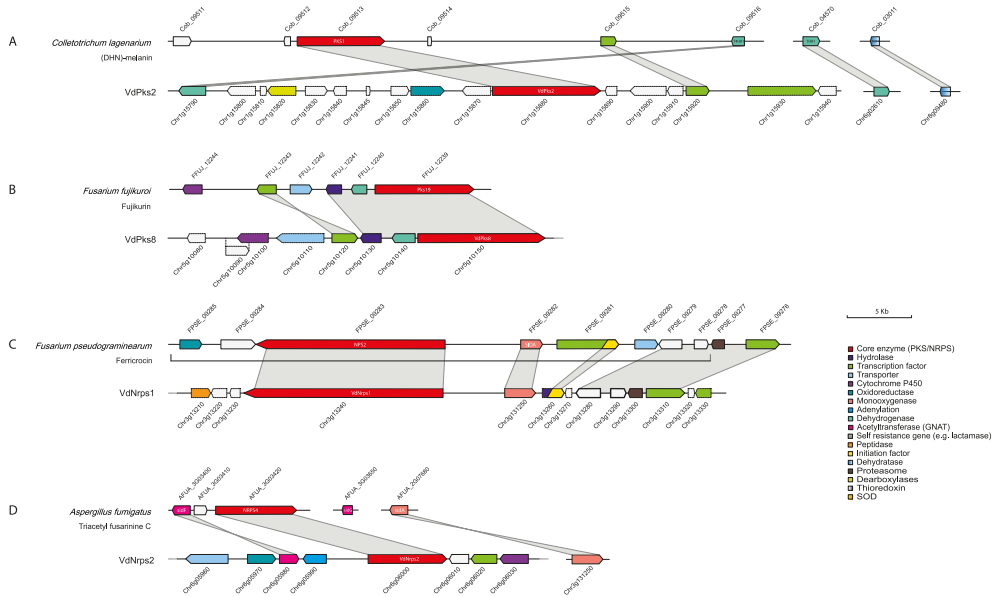
Based on the phylogenetic analyses, 11 core enzymes (VdPks1, VdPks2, VdPks3, VdPks4, VdPks5, VdPks7, VdPks8, VdPks9, VdNrps1, VdNrps2 and VdHyb1) were identified that cluster with previously characterised enzymes from other fungal species (Figure 2, Figure 3). Subsequently, we queried for the other genes besides the core genes from these previously characterised clusters in other fungal species to find homologs in the corresponding *V. dahliae* clusters. However, only the VdPks2, VdPks8, VdNrps1 and VdNrps2 clusters of *V. dahliae* share more homologs (more than 50% of the whole cluster) in addition to the core genes with other fungal species. The remaining clusters contain less than 50% of genes that share homologs with other fungal species. In other fungi, conserved gene clusters of VdPks2, VdPks8, VdNrps1 and VdNrps2 are responsible for the biosynthesis of DHN-melanin, fujikurin, ferricrocin and TAFC, respectively (Tsuiji et al., 2002; Tobiasen et al., 2007; Sieber et al., 2014; von Bargen et al., 2015; Niehaus et al., 2016).

The gene cluster for DHN-melanin biosynthesis in the fungal pathogen *C. lagenarium* comprises six genes, including three functionally characterized genes encoding polyketide synthase *CIPKS1*, reductase *T4HR1* and transcription factor *cmr1*. Moreover, two genes (scytalone dehydratase *SCD1* and *THR1* reductase) residing at another chromosome were identified to be involved in the biosynthesis of DHN-melanin as well (Tsuiji et al., 2002) (Figure 4A). We observed amino acid identities of 70%-90% between *CIPKS1*, *T4HR1*, *SCD1* and *THR1* of *C. lagenarium* and their orthologs in *V. dahliae*. The transcription factor *CMR1* only shares 55% identity with its counterpart in *C. lagenarium*.

*VdPks8* is an ortholog of the fujikurin synthase gene (*FfuPks19*) of *F. fujikuroi*. The fujikurin cluster contains six genes (von Bargen et al., 2015; Niehaus et al., 2016), four of which have homologs in the *VdPks8* locus in *V. dahliae*. The homolog of the MFS transporter *FFUJ\_12242* of *F. fujikuroi* was found on a different chromosome in *V. dahliae*, and the cytochrome P450 gene in *F. fujikuroi* has no *V. dahliae* homolog (Figure 4B). Interestingly, the *VdPks8* locus contains two other genes that are annotated as cytochrome P450 and MFS transporter, but these genes were not detected as homologs of the genes in the *FfuPks19* cluster (Figure 4B).

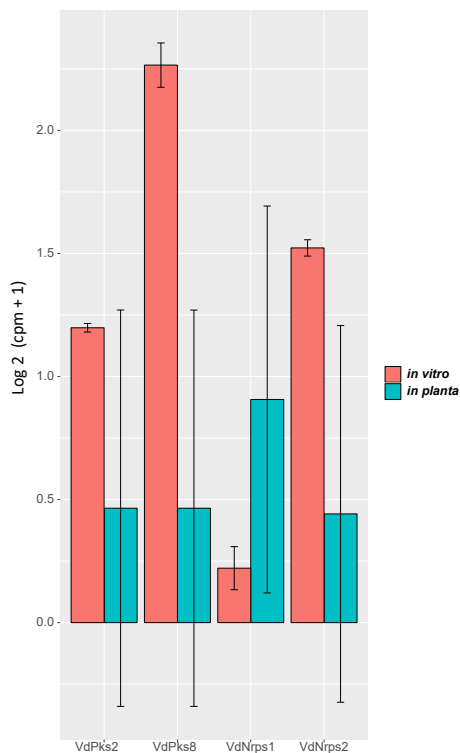
The biosynthesis of ferricrocin requires two genes that are located at the same locus in the genome of *F. pseudograminearum*, which encode an L-ornithine N5-oxygenase (*SIDA*) and an NRPS (*FpNRPS2*) (Tobiasen et al., 2007; Sieber et al., 2014). Similarly, in *V. dahliae* homologs of L-ornithine N5-oxygenase (*SIDA*) and NRPS (*FpNRPS2*) genes are located next to each other. Moreover, homologs of a proteasome subunit, a transcription factor and two uncharacterised genes in the same cluster of *F. pseudograminearum* were found in *V. dahliae*. In addition, genes encoding an MFS transporter and an oxireductase in *F. pseudograminearum* have no homologs in *V. dahliae* (Figure 4C).

VdNrps2 is an ortholog of the extracellular siderophore TFAC synthase gene *NRPS4* in *A. fumigatus*, which belongs to a cluster of two genes (*NRPS4* and *sidF*) (Schrettl et al., 2007). Another two extracellular siderophore TFAC synthase genes (*sidG* and *sidA*) are located at another chromosome (Schrettl et al., 2007). Except for *sidG*, all described genes have homologs in *V. dahliae* (Figure 4D).



**FIGURE 4. Synteny of conserved SMCs in *V. dahliae*.** *V. dahliae* putative clusters were compared to previously described clusters. Ensembl gene IDs are shown above or below the genes. (A) DHN-melanin, (B) Fujikirins, (C) Ferricrocin, (D) Triacetyl fusarinine C.

To investigate the putative role of the SM clusters in plant pathogen interactions, we compared expression patterns of each core gene *in planta* and *in vitro*. We assessed transcription (RNA-seq) data of *V. dahliae* during colonization of *Arabidopsis thaliana* and found that all four core genes of the four conserved clusters showed *in planta* expression. Whereas *VdNrps1* was found to be induced, *VdPks2*, *VdPks8*, and *VdNrps2* were repressed when compared with the expression in *in vitro*-cultured mycelium (Figure 5). Nevertheless, the expression *in planta* suggests that these SM cluster may play a role during host colonization.



**FIGURE 5. Pair-wise comparison of core SMC genes with differential expression *in vitro* and *in planta*.** Gene expressions are depicted for *V. dahliae* strain JR2 cultured in liquid medium and upon *A. thaliana* colonization, respectively. Bars represent the mean gene expression with standard deviations. The significance of difference in gene expression was calculated using t-tests relative to a threshold (TREAT) of log<sub>2</sub>-fold-change  $\geq 1$  (McCarthy and Smyth, 2009).

## Conclusions

In this study, we have used an *in silico* approach to identify 25 putative SMCs in the genome of *V. dahliae* strain JR2, all of which appear complete and thus potentially functional. Our predictions state that two putative siderophores, ferricrocin and TAFC, DHN-melanin and fujikurin compounds may belong to the active SM repertoire of *V. dahliae*.

## Materials & Methods

### Secondary metabolite cluster prediction, annotation and conservation

Putative SMCs were identified with antiSMASH fungal version 4.0.2 (Weber et al., 2015). The predicted borders from antiSMASH were directly used to retrieve all protein sequences contained within the clusters. The Bedtools intersect command (Quinlan and Hall, 2010) was used to obtain the file containing the gene locations, followed by gffread from the Cufflinks package (Trapnell et al., 2010) to retrieve the protein sequences. Sub-telomeric regions were defined as 300 kb of the chromosomal ends, as similarly used for other filamentous fungi (McDonagh et al., 2008; Cairns and Meyer, 2017). Genes within

the genomic range were counted using BioMart from Ensembl (Kersey et al., 2016). A  $\chi^2$  test was performed to determine the significance of enrichment.

The conservation of predicted SMCs among *Verticillium* spp. was assessed based on core enzyme conservation, using BLAST+ tool protein blast (blastp) (Camacho et al., 2009) on predicted protein databases (e-value  $< 1 \times 10^{-5}$ , query coverage  $> 60\%$  and identity  $> 50\%$  (Sbaraini et al., 2017).

### Phylogenetic analysis

The previously described type I PKSs, NRPSs and PKS-NRPSs enzymes used in this study for phylogenetic analysis were derived from the curated database of UniProt, SwissProt and literatures (Gallo et al. 2013; Yu et al. 2015). The amino acid alignment was built using MAFFT version 7.205 (Kato and Standley, 2013). We used the G-INS-i strategy, global alignment (--globalpair) and 1000 cycles of iterative refinement (--maxiterate 1000). Aligned sequences were visualised with Aliview version 1.20 (Larsson, 2014) and manually curated by removing non-aligned sequences. Preceding the phylogenetic analysis, the alignments were trimmed to remove poorly aligned regions using TrimAl version 1.4 (Capella-Gutiérrez et al. 2009). First, all positions in the alignment with gaps in 90% or more of the sequence were removed (-gt 0.1), followed by the automated1 parameter (-automated1). RaxML version 8.1.1 (Stamatakis, 2014) was used to construct Maximum-likelihood phylogenetic tree (-f a). The automated protein model selection (-m PROTGAMMAAUTO) was used applying 100 rapid bootstrapping (-#100). The number of seeds for parsimony inferences and rapid bootstrap analysis was set to 12345 (-p 12345 -x 12345, respectively). The output tree (RaxML\_bipartitionBranchLabels) was visualised using iTOL webtool version 3.0 (Letunic and Bork, 2016).

### Comparative cluster analysis

Protein sequences of described clusters were blasted (BLASTp, E-value cutoff  $1e-5$ , query coverage  $> 60\%$  and identity  $> 25\%$ ) against the *V. dahliae* strain JR2 protein database. We considered a cluster to be conserved in *V. dahliae* when at least 50% of the queried proteins from previously described clusters were found in *V. dahliae*.

### Gene expression analysis

To obtain RNA-seq data for *V. dahliae* grown in culture medium, strain JR2 was grown for three days in potato dextrose broth (PDB) in three biological replicates. To obtain RNA-seq data from *V. dahliae* grown *in planta*, three-week-old *A. thaliana* (Col-0) plants were inoculated with strain JR2. After root inoculation, plants were grown in individual pots in a greenhouse under a cycle of 16 h of light and 8 h of darkness, with temperatures

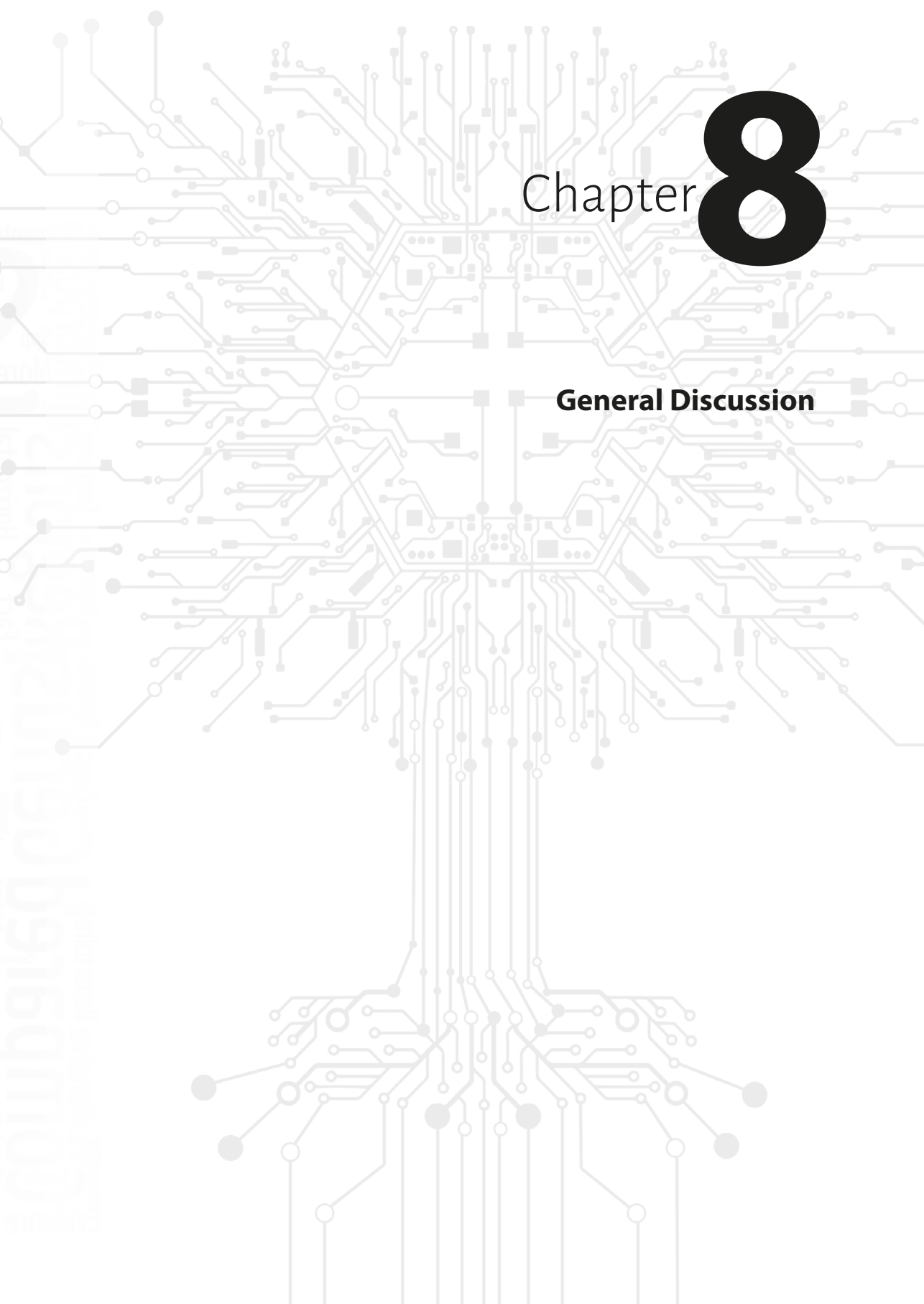
maintained between 20 and 22°C during the day and a minimum of 15°C overnight. Three pooled samples (10 plants per sample) of complete flowering stems were used for total RNA extraction. Total RNA was extracted based on TRIzol RNA extraction (Simms et al., 1993). cDNA synthesis, library preparation (TruSeq RNA-Seq short insert library), and Illumina sequencing (single-end 50 bp) was performed at the Beijing Genome Institute (BGI, Hong Kong, China). In total, ~2 Gb and ~1.5 Gb of filtered reads were obtained for the *V. dahliae* samples grown in culture medium and *in planta*, respectively. RNAseq data were submitted to the SRA database under the accession number: SRP149060.

The RNA sequencing reads were mapped to their previously assembled genomes using the Rsubread package in R (Liao et al. 2013). The comparative transcriptomic analysis was performed with the package edgeR in R (v3.4.3) (Robinson et al. 2010; McCarthy et al. 2012). Genes are considered differentially expressed when P-value < 0.05 with a log<sub>2</sub>-fold-change ≥ 1. P-values were corrected for multiple comparisons according to Benjamini and Hochberg (Benjamini and Hochberg, 1995).

### Acknowledgements

Work in the laboratories of B.P.H.J.T. and M.F.S. is supported by the Research Council Earth and Life Sciences (ALW) of the Netherlands Organization of Scientific Research (NWO).





# Chapter 8

## General Discussion

## Abstract

Fungal pathogens are the most important causal agents of disease in plants. In order to successfully establish their parasitic relationship with host plants, these pathogens rely on the secretion of effector molecules that facilitate the establishment of infections, for instance through the perturbation of host immune responses. Thus, a full understanding of the mechanisms underlying fungal infections of plants requires the identification and characterization of complete effector catalogs and an understanding of the evolutionary trajectory of pathogen adaptation. The development of next-generation sequencing technologies has greatly advanced our understanding of plant pathogens over recent years as these have facilitated the generation of (near-)complete genome assemblies. As advanced genome analyses have revealed a tight correlation between repetitive genomic regions and effector genes, the discovery of novel effectors has been boosted by advanced *in silico* identification strategies from genome sequences. Thus, genomics studies enable us to detect features that are associated with the adaptive evolution of fungal plant pathogens. Intriguingly, however, genomic features such as compartmentalized genome structures that were identified in diverse plant pathogenic fungi are also found in non-pathogenic species. Furthermore, genus-wide comparative genomics analyses revealed that there is no obvious specific genomic feature that can be associated with pathogenicity in the *Verticillium* genus. The ability to cause disease within the *Verticillium* genus is most likely caused by subtle genomic traits that do not easily become apparent from whole-genome comparisons.

## Introduction

Like for other domains of life, research into the biology of fungal species has greatly benefited from the advent of whole-genome sequencing. Initially, genomes were sequenced with Sanger sequencing, requiring a considerable investment of time, labor and costs (Goffeau et al., 1996). Initial microbial genome projects were mainly focused on human pathogens, besides model species such as the yeast *Saccharomyces cerevisiae* (Goffeau et al., 1996) and the filamentous fungus *Neurospora crassa* (Galagan et al., 2003). However, since next-generation sequencing (NGS) technologies have increased the speed and scalability of genomic sequencing at a significantly reduced cost, this pattern has changed. Nowadays, genome sequencing is within reach for most scientists within the field of biology. A large proportion (almost 50%) of sequenced fungi consists of pathogens, of which 50% are plant pathogens (Aylward et al., 2017). Thus, also research into the biology of fungal plant pathogens has greatly benefited from the advent of whole-genome sequencing (Thomma et al., 2016; Aylward et al., 2017).

NGS comprises second- and third-generation sequencing technologies. Second generation sequencing (SGS) technologies are also known as short-read technologies that can produce high amounts (in the range of gigabases) of sequence reads of up to 500 bp in only a few days. Because of their limited read size, these technologies are typically not able to span repetitive genomic regions as these regions are typically longer than the SGS read length (Kchouk et al., 2017). Consequently, genome projects that are based on short-read sequencing strategies typically yield highly fragmented genome assemblies (Koboldt et al., 2013). However, as many early genome sequencing projects aimed at identifying the coding regions of a genome of interest to advance research aimed at characterizing the physiology or molecular biology of particular cellular processes, fragmented genome assemblies were considered sufficient by many researchers as long as they comprised a full catalog of the genic spaces.

In contrast to SGS, third generation sequencing (TGS) or long-read sequencing technology can produce reads with average lengths of (over) 10 kb and individual reads of 100 kb or more (Tyson et al., 2018). Consequently, TGS is able to span repeats which greatly aids in the assembly of these regions. Based on TGS, (near-)complete genome assemblies have been obtained that have widely revealed the significance of non-coding regions and repetitive elements for the life style, adaptability and evolution of species, including plant pathogenic fungi (King et al., 2015; Faino et al., 2016; Thomma et al., 2016; Dallery et al., 2017; van Kan et al., 2017; Plissonneau et al., 2018).

In this chapter, I outline how our understanding of the biology of fungal plant pathogens has been boosted by advances in sequencing technologies. I summarize key examples from fungal plant pathogen genomics studies and discuss evolutionary process in

pathogenic fungi, highlighting the importance of determining high-quality genome assemblies. In addition, I also discuss examples showing that genome architectures associated with rapid adaptation and pathogenic success in plant-associated fungi may similarly be associated with adaptation in non-pathogenic species.

## Effector identifications

### Effector identification through functional screens

The antagonistic nature of interactions between plants and fungal plant pathogens is known to result in everlasting arms races (Jones and Dangl 2006; Cook and Thomma 2015). In such arms races, plant pathogens secrete effectors, which are molecules that manipulate host cell biology, repress host immune responses or shield the pathogen to support fungal growth and colonization, while plants try to halt pathogen invasion through the activation of immune responses upon the detection of pathogen-derived or pathogen-induced molecules (Rovenich et al., 2014; Cook et al., 2015; Rodriguez-Moreno et al., 2018). In this process, effectors can become pathogen invasion-association molecular patterns that get recognized by plant immune receptors (Cook and Thomma 2015). Thus, effectors can play a dual role in host-pathogen interactions and can impact the outcome of an infection both positively and negatively, depending on the host genotype. Consequently, research on plant-pathogen interactions has focused on discovery and functional analysis of novel effector molecules. Historically, many effectors were identified through labor-intensive and time-consuming methods such as cDNA library screening (Luderer et al., 2002) and random insertional mutagenesis (Seong et al., 2005; Betts et al., 2007; Blaise et al., 2007; Jeon et al., 2007; Michielse et al., 2009). For example, the effector *Avr2* of *Cladosporium fulvum* was discovered by using a functional screen based on the HR-inducing activity of the AVR2 protein in tomato plants (Luderer et al., 2002). Random insertional mutagenesis through *Agrobacterium tumefaciens*-mediated transformation, followed by screening for reduced virulence on the respective host plant, has been used to identify novel effectors and pathogenicity factors in several fungal pathogens including *Fusarium oxysporum* (Michielse et al., 2009), *Colletotrichum graminicola* (Muench et al., 2011), *Botrytis cinerea* (Giesbert et al., 2012) and *Verticillium dahliae* (Santhanam et al., 2017).

### Genomics-based effector identification

Genome sequences enable *in silico* predictions of candidate effectors utilizing their common characteristic features such as small size, secretion signals, and high cysteine content. However, many well-characterized effectors do not possess all of these properties. This means that predictions using these features as strict criteria often do not reveal full effector catalogs (Gibriel and Seidl 2016). Enhanced automated identification methods do

not solely rely on strict cut-off values, but rather rank proteins according to their likelihood of being effectors by scoring multiple effector features and can therefore further aid in effector identification (Saunders et al., 2012). Recently, a first machine learning-based tool, EffectorP, was introduced to predict effector candidates from secretomes (Sperschneider et al., 2016; Sperschneider et al., 2018). However, EffectorP relies on the availability or prediction of secretomes and fails to predict effectors that do not encode typical signals for extracellular secretion. EffectorP uses sequence-derived features of experimentally verified fungal effectors as a positive training set and secreted non-effectors as a negative set to predict fungal effectors. However, the negative training set consists not only of secreted non-effectors but likely also comprises uncharacterized effectors, and thus the current version of EffectorP (Sperschneider et al., 2018) is only predicting effectors with similar properties as the already experimentally verified ones. With the availability of an increasing number of experimentally verified effectors, the tool will have to be frequently re-trained to achieve more accurate predictions (Sperschneider et al., 2018).

In comparative genomics, genome sequences are compared through alignment in order to study the relationship and to reveal the similarities and differences between genomes (Raffaele and Kamoun, 2012; Plissonneau et al., 2017). For example, a comparative genomics analysis of race 1 and race 2 strains of *Verticillium dahliae* revealed the effector gene *Ave1*, the product of which is recognized by tomato plants that carry the cell surface immune receptor gene *Ve1* (de Jonge et al., 2012). Similarly, the effector *AvrFom2* of the melon wilt fungal pathogen *Fusarium oxysporum f.sp. melonis* that is recognized by melon immune receptor Form-2 was identified by comparative genomics between pathogenic and non-pathogenic strains (Schmidt et al., 2016).

### Effectors of non-pathogenic species

Like pathogens, particular non-pathogenic fungal species that include endophytes and mutualists develop intimate associations with host plants in a similar fashion as plant pathogens. During initiation of such symbioses, endophytes and mutualists are recipients of host immune responses. Conceivably, similar to pathogens, they also use effectors to suppress host immunity (de Jonge and Thomma, 2009; Rovenich et al., 2014). For example, the endophytic fungus *Rhizophagus irregularis* expresses many effector-like small secreted proteins during interacting with host plants (Tisserant et al., 2013). Among the secreted proteins, the SP7 effector was shown to attenuate ethylene-mediated immune responses of the host plant. Similarly, the genome of the mutualistic fungus *Laccaria bicolor* encodes hundreds of secreted proteins, including Mycorrhiza-induced Small Secreted Proteins (MiSSPs). Of these, MiSSP7 was shown to perturb jasmonic acid-mediated immune signaling (Plett et al., 2011; Plett et al., 2014). Arguably, most non-pathogens are saprophytes, which generally reside within the soil where they thrive on decaying

organic matter in the presence of a rich microbiota. Threats are posed by mycoparasites and microbial competitors that produce antagonistic molecules that include hydrolytic enzymes, antibiotics, toxins and volatiles (Compant et al., 2005). Furthermore, many mycoparasites secrete hydrolytic enzymes, including protease, chitinase and glucanases to target fungal cell walls. Effectors may be employed by saprophytic species to act in self-defense and competition (Rovenich et al., 2014; Snelders et al., 2018). Presumably, chitin-binding effectors protect hyphal cell walls against mycoparasite-derived chitinases, which may explain abundant LysM effector catalogs of saprophytic fungi (Kubicek et al., 2011; Kombrink and Thomma, 2013). It was demonstrated that the localization of the bacterial species *Bacillus subtilis* to fungal hyphae is largely inhibited by a LysM effector, suggesting that this effector may play a role in the interaction of fungi with bacteria (Kombrink and Thomma, 2013). Thus, it is not surprising that genome analyses further revealed that non-pathogenic species encode similarly sized effector catalogs as pathogens (Druzhinina and Kubicek 2012; Suh et al. 2012). This has also been observed in the *Verticillium* genus, which comprises soil-borne asexual species that differ significantly in their lifestyles, ranging from saprophytic to pathogenic (Fradin and Thomma, 2006; Inderbitzin et al., 2011a; Klosterman et al., 2011; Inderbitzin and Subbarao, 2014). More specifically, the saprophyte *V. tricorpus* has an effector reservoir size that resembles that of *V. dahliae* (**Chapter 2**). A genus-wide comparative analysis between pathogenic and non-pathogenic *Verticillium* spp. further established that both groups comprise similarly sized effector catalogs (**Chapter 3**).

## Genomic variation of fungal pathogens

### Genome size variation

Genome sequencing of fungal plant pathogens has revealed variability in genome size, ranging from approximately 20 to 200 Mb (Aylward et al., 2017), often correlating with the amount of repetitive elements (Mohanta and Bae, 2015; Raffaele and Kamoun 2012). Repetitive elements, or repeats, are DNA sequences that are present multiple times within a genome, ranging from few to many (hundreds) copies (Biscotti et al., 2015). Sequencing with short-read technologies, such as Illumina, commonly limits the contiguity of genome assemblies, since short reads are typically not able to span repetitive regions (Kchouk et al., 2017). Moreover, *de novo* sequence assembly software collapses identical repeats, resulting in reduced genome size and complexity (Alkan et al., 2010). Long-read technologies, such as single-molecule real-time (SMRT) sequencing, drastically improves the assembly of reads into contiguous sequences that ideally comprise complete chromosomes (Ashton et al., 2014; Huddleston et al., 2014; Laszlo et al., 2014). Finished genomes using long-read technology revealed that the repeat content

was initially often underestimated in genome assemblies that were generated based on short-read technologies only. For instance, the first genome assembly of *V. dahliae* that was produced using Sanger sequencing suggested a repeat content of 4% (Klosterman et al., 2011), which similarly was observed for other *V. dahliae* strains based on Illumina sequencing (de Jonge et al., 2013). However, both methods turned out to significantly under-estimate the repeat content when compared to the final determination of 12% for finished genome assemblies based on SMRT sequencing (Faino et al., 2015). Similarly, the finished genome assembly of *C. higginsianum* has a repeat content of 7%, whereas it was initially estimated to carry only 1.2% of repeats (Dallery et al., 2017).

### Repeat-rich genome compartments

Facilitated by (near-)complete genome assemblies, comparative genomics analyses have revealed that fungal plant pathogens often exhibit compartmentalized genomes consisting of gene-rich core regions and accessory genome sections that are gene-sparse and repeat-rich (Raffaele and Kamoun, 2012; Dong et al., 2015). Many virulence-associated genes, including effector genes, reside in these accessory compartments in various pathogens (Raffaele and Kamoun, 2012; Dong et al., 2015). Genome compartmentalization is considered to improve the evolutionary efficiency by facilitating the rapid evolution of virulence-associated genes without affecting household genes that typically localize in the core genome. Such bipartite genome structure is often referred to as a two-speed-genome (Croll and McDonald, 2012; Raffaele and Kamoun, 2012; Dong et al., 2015).

An obvious appearance of genome compartmentalization is displayed by species that carry accessory chromosomes that are also known as conditionally dispensable chromosomes (CDCs) (Ma et al., 2010; Goodwin et al., 2011; Croll and McDonald, 2012). For instance, in the fungal tomato vascular wilt pathogen *Fusarium oxysporum f.sp lycopersici* (Fol), CDCs comprise about one quarter of the total genome, while they contain 74% of the repeats and TEs (Ma et al., 2010). These repeat-rich CDCs contain important effector genes enabling infection on the host plant tomato (Ma et al., 2010). Moreover, horizontal transfer of CDCs from a pathogenic to a non-pathogenic strain of *F. oxysporum* enabled the latter to become pathogenic on tomato (Ma et al., 2010). CDCs similarly occur in other *formae speciales* of *F. oxysporum* that infect other host plants (Alexander et al., 2016), and also in other fungal pathogens, such as various species of the *Alternaria* genus (Thomma, 2003).

Accessory genome sections can also be dispersed throughout the core genome. For instance, for the maize smut pathogens *Ustilago maydis* and *Sporisorium reilianum*, effector genes reside in small, repeat-rich genomic clusters devoid of housekeeping genes (Schirawski et al., 2010). Similarly, in the genome of the vascular wilt pathogen *V. dahliae*, *in planta*-induced effector genes are located at lineage-specific (LS) regions that

are embedded within the eight chromosomes (de Jonge et al., 2013; Faino et al., 2016). In other pathogens, such as the rice blast fungus *Magnaporthe oryzae*, effector genes reside in repeat-rich sub-telomeric regions (Orbach et al., 2000; Khang et al., 2008).

### Impact of TE activities

The correct assembly of repeats is essential for the correct identification of transposable element (TE) catalogs in a genome assembly. TEs are mobile genetic elements that can move through the genome in a copy-paste (retro-transposons) or a cut-paste (DNA transposon) fashion, and can disrupt gene activity through insertion in promoters or open reading frames (Ayarpadikannan and Kim, 2014). They are identified based on their repetitiveness in the genome (Price et al., 2005), the identification of specific open reading frames in a DNA sequence, or based on known terminal repeats (Xu and Wang, 2007; Ellinghaus et al., 2008). Barbara McClintock's discovery of these "jumping genes" earned her the Noble Prize in 1983 (McClintock, 1950).

Although TEs are often thought to have mainly negative effects on their hosts, their association with rapidly evolving accessory genome regions suggests that they play a pivotal role in the accelerated evolution of these regions (Raffaele and Kamoun, 2012; Dong et al., 2015; Seidl and Thomma, 2017). In some cases, it has been demonstrated that TE activity can influence the outcome of host-pathogen interactions. TEs can disrupt gene activity through insertion in promoter or open reading frame sequences. Such insertions can facilitate immune evasion as observed for the tomato pathogen *C. fulvum* (Westerink et al., 2004; Stergiopoulos and de Wit, 2009). For example, interruption of the promoter of the *Avr4E* effector by the insertion of a transposon compromised expression of this effector gene and led to evasion of *C. fulvum* recognition by tomato plants harboring the resistance gene *Hcr9-4E* (Westerink et al., 2004). Similarly, the barley smut fungus *Ustilago hordei* gained virulence on barley plants harboring the resistance gene *Ruh1* by insertion of a TE in the promoter of the *UhAvr1* gene (Ali et al., 2014). Furthermore, TE activity can shape genomic regions by causing double-strand DNA breaks that may lead to non-faithful repair, especially between genomic regions with high sequence similarity (Seidl and Thomma, 2014; Faino et al., 2016). In *V. dahliae*, LS regions originate from structural rearrangements that resulted from unfaithful DNA repair (de Jonge et al., 2013; Faino et al., 2016; Depotter et al., 2018). Another mechanism of how TEs can lead to accelerated evolution can be found in the Brassicaceae blackleg pathogen *Leptosphaeria maculans* (Rouxel et al., 2011; Daverdin et al., 2012). TE-rich regions in the genome of this fungus are subject to repeat induced point mutation (RIP), which is a fungal defense mechanism against TEs (Rouxel et al., 2011). Effectors in these TE-rich regions can be affected by "leakage" of the RIP process from TEs into neighboring sequences, leading to the occurrence of SNPs, and thus accelerated effector diversification (Rouxel et al. 2011;



Daverdin et al. 2012). However, accelerated diversification of effectors through RIP is currently only demonstrated for *L. maculans*, as RIP only operates in sexually propagating Ascomycetes (Selker, 2002). There are currently no examples of RIP-independent increases of nucleotide diversity in accessory regions, and thus this does not seem to be a common mechanism to accelerate genome evolution in fungal plant pathogens. Even though *Verticillium* species are ascomycete fungi, they are asexual, and there is no evidence for the occurrence of RIP in these species.

Surprisingly, we recently observed that fast evolving accessory regions (LS regions) of *V. dahliae* contain lower levels of nucleotide diversity when compared with the core genome (**Chapter 5**). However, the underlying mechanisms that are responsible for this phenomenon are not known. Increased sequence conservation in effectors seems evolutionary disadvantageous as these effectors are ideal targets for plant immune recognition (Jones and Dangl, 2006). For instance, *Ave1* displays high sequence conservation, yet is recognized by the Ve1 immune receptor of tomato plants. Today no allelic variation of *Ave1* in the *V. dahliae* population has been described (de Jonge et al., 2012). However, *V. dahliae* establishes evasion of host recognition through presence/absence polymorphisms, as all race 2 strains that have evaded Ve1 recognition are characterized by loss of the *Ave1* gene, which occurred in several independent events (de Jonge et al., 2012; de Jonge et al., 2013; Faino et al., 2015).

### Genomic plasticity in non-pathogenic *Verticillium* species

Arguably, any organism can benefit from rapid evolution of particular genes to facilitate environmental adaptation. Concomitantly, structured genomes that may mediate the accelerated evolution of some genes, but not others, are also observed in non-pathogenic fungal species. In the endophytic fungus *Piriformospora indica*, small secreted protein genes that are induced during symbiotic growth in host plants are located in gene-poor and repeat-rich genomic regions (Zuccaro et al., 2011; Gill et al., 2016). Similarly, the genome of ubiquitous saprophyte and mammalian opportunistic pathogen *Aspergillus fumigatus* encompasses TE-rich LS regions that contain gene clusters involved in mycotoxin biosynthesis (Fedorova et al., 2008). Accelerated evolution of mycotoxin gene clusters may give *A. fumigatus* a competitive advantage over other microbes (Fedorova et al., 2008). Thus, structured genomes are not confined to pathogenic species. A structured genome is also observed for the saprophytic species *V. tricorpus*. In this thesis, we showed that *V. dahliae* and *V. tricorpus* share a genomic structure with genes encoding secreted proteins that are clustered in repeat-rich genomic compartments (**Chapter 2**).

To identify genomic features that are associated with pathogenic *Verticillium* spp., we investigated the occurrence of genomic features that have been described for *V. dahliae* in other *Verticillium* species (**Chapter 3-5**). Considering that chromosomal rearrangements

are important drivers of genomic variation in *V. dahliae*, we speculated that the extent of rearrangements may be associated with the fact that, despite being asexual, *V. dahliae* is a successful broad host-range pathogen (Seidl and Thomma, 2014; Faino et al., 2016). From this hypothesis it would follow that other species that are less successful pathogens would not undergo such drastic genomic rearrangements. However, we identified extensive chromosomal rearrangements in all *Verticillium* species (**Chapter 3**). Moreover, reconstruction of ancestral *Verticillium* genomes revealed that genomic rearrangements occurred throughout evolution, suggesting they have contributed to speciation in the *Verticillium* genus (**Chapter 4**). This observation makes it unlikely that the occurrence of the rearrangements themselves is a major determinant for the pathogenicity of *V. dahliae*. It has been demonstrated that the LS regions of *V. dahliae* evolved by genomic rearrangements and play important roles in adaptation (de Jonge et al., 2013; Faino et al., 2016). In this thesis, we furthermore showed that the accelerated evolution of LS regions is achieved through presence/absence polymorphisms, as nucleotide sequences are highly conserved (**Chapter 5**). We also showed that the increased sequence conservation and the clustering of presence/absence polymorphisms in rapidly evolving sections of the *V. dahliae* genome similarly occurred in *V. tricorpus* (**Chapter 5**). Collectively, we found no obvious genomic differences that can be linked to pathogenicity in the *Verticillium* genus. Therefore, the question what determines the ability to cause disease in the *Verticillium* genus still remains unresolved.

Given the fact that genes encoding secreted proteins are highly conserved between *Verticillium* spp., and well-characterized effector families that have been implicated in fungal pathogenicity, such as LysM effectors (de Jonge et al., 2010; Marshall et al., 2011; Mentlak et al., 2012; Kombrink et al., 2017) and NLP effectors (Santhanam et al., 2012) are found in non-pathogenic *Verticillium* spp. as well (**Chapter 2-3**), the ability of pathogenic *Verticillium* spp. to cause disease is unlikely determined by shared effectors alone. Still, it is possible that homologs of effectors are found both in pathogenic and non-pathogenic species, but the encoded proteins have different functions that lead to different outcomes of host interaction. For example, a homolog of the effector *Ave1* of *V. dahliae* (*VdAve1*) was also found in the saprophyte *V. nubilum* (*VnAve1*), and an amino acid sequence alignment revealed that they share a high degree of sequence similarity (92%). However, introducing *VnAve1* into a *VdAve1* deletion strain of *V. dahliae* did not restore virulence on tomato plants (Boshoven, 2017). Moreover, expression of shared effectors can be different between pathogenic and non-pathogenic species, which might be influenced by a number of genetic differences. Gene expression is often controlled by complex regulatory networks (Davidson and Levin, 2005). Therefore, small differences in regulatory sequences, transcription factors or epigenetic modifications might lead to large changes in gene expression. In addition to the shared effectors, different effectors that are only present in pathogenic species might fuel the ability to cause disease. We found

that the only 5% of the total effector repertoire is species-specific (**Chapter 3**). Further experiments that introduce pathogen-specific effectors into non-pathogenic species might be interesting. In conclusion, the ability to cause disease within the *Verticillium* genus might be caused by subtle genomic traits that do not easily become apparent from whole-genome comparisons.

## Conclusions & Remarks

Developments in sequencing technologies have significantly advanced our understanding of plant pathogens and their evolution. Genomic studies of plant pathogens based on high-quality genome assemblies revealed genome structures that promote rapid evolution of pathogenicity-related traits. However, such structured genomes can also be found in non-pathogenic species. Furthermore, genus-wide comparative genomics analyses indicated that all haploid *Verticillium* spp. display similar genomic features, including the occurrence of extensive genomic rearrangements and the presence of LS regions (**Chapter 3-5**). In conclusion, there is no obvious specific genomic feature that can be associated with pathogenicity in the *Verticillium* genus (**Chapter 2-5**). The ability to cause disease within the *Verticillium* genus is most likely caused by subtle genomic traits that do not easily become apparent from whole-genome comparisons. Thus, different approaches should be undertaken to reveal the basis of pathogenicity, such as the comparison of *V. dahliae* strains that are either pathogenic or non-pathogenic on a particular plant species.





**Summary**  
**References**  
**Acknowledgements**  
**Curriculum vitae**  
**List of publications**  
**Education statement**



## Summary

Fungi are organisms that form a distinct kingdom within the eukaryotes. Although most fungi are saprophytes that decay organic matter, the fungal kingdom also includes species that are able to cause considerable yield losses in crop production systems worldwide. In fact, fungi are the most significant type of plant pathogens. **Chapter 1** is a general introduction about fungal plant pathogens. Recent advances in genome sequencing and assembly strategies have facilitated the availability of high-quality genome assemblies of many fungal species. Consequently, comparative genomics studies have provided novel insight in the evolution of pathogen genomes.

The fungal genus *Verticillium* contains ten species, some of which are notorious causal agents of vascular wilt diseases, while others are mainly known as saprophytes that have only been reported to cause disease occasionally as opportunistic pathogens. Whereas the genome of the most notorious plant pathogen in the genus, *Verticillium dahliae*, has been quite well characterized, genome characteristics of the other species have received little attention. With comparative genomics, we aimed to identify genomic features in pathogenic *Verticillium* species that confer the ability to cause vascular wilt disease. In **Chapter 2**, we sequenced and assembled one of the saprophytic *Verticillium* species, *V. tricorpus*, based on a hybrid approach that combines second and third generation sequencing. The resulting near-gapless genome assembly was subsequently used for a comparative genomics analysis with *V. dahliae*. Unexpectedly, both species encode similar effector repertoires and share a genomic structure with genes encoding secreted proteins clustered in genomic islands. In conclusion, we highlight the technical advances of a hybrid sequencing and assembly approach and reveal that the saprophyte *V. tricorpus* shares many hallmark features with the pathogen *V. dahliae*. Subsequently, in **Chapter 3**, we extended our comparative genomics study to the whole genus-level to try and identify genomic features that can be associated with pathogenicity. We sequenced the genomes of all haploid *Verticillium* spp. and demonstrated that all species display similar genomic features, including the occurrence of extensive genomic rearrangements and the presence of extensive effector catalogs. Overall, this chapter failed to identify particular genomic features that can be associated with pathogenicity in the genus *Verticillium*.

Previously, comparative genomics of *V. dahliae* revealed that lineage-specific genomic regions are hotspots for presence/absence polymorphisms, chromosomal rearrangements active transposable elements and *in planta*-induced effector genes. This finding suggests that *V. dahliae* evolved according to a so-called two-speed genome model as is similarly observed for other filamentous plant pathogens. In **Chapter 4**, we demonstrate the occurrence of differential sequence divergence between core and lineage-specific genomic regions of *V. dahliae*. Surprisingly, we observed that lineage-specific regions display markedly increased sequence conservation, suggesting that host adaptation

is merely achieved through presence/absence polymorphisms. Increased sequence conservation of genomic regions that are important for pathogenicity is an unprecedented finding in filamentous pathogens and sheds new light on genomic dynamics in host-pathogen co-evolution. We provide evidence that disqualifies horizontal transfer between *Verticillium* species to explain the observed sequence conservation and conclude that sequence divergence occurs at a slower pace in lineage-specific regions of the *V. dahliae* genome. We hypothesize that differences in chromatin organisation may explain reduced SNP frequencies that occur in the plastic LS regions of *V. dahliae*. Furthermore, we also show that the clustering of absence/presence polymorphisms in rapid-evolving genome sections of the *V. dahliae* genome was similarly found for *V. tricorpus*, suggesting that this increased sequence conservation is not confined to pathogenic *Verticillium* species.

One of the previously characterized *V. dahliae* lineage-specific effector genes, *Ave1*, previously shown to be horizontally acquired from a plant donor. In **Chapter 5**, we systematically searched for evidence of inter-kingdom HGT events in the genome of *V. dahliae* using an alien index score that provides a score for genes that are phylogenetically more related to a defined outgroup when compared with an in-group. This analysis resulted in 42 likely HGT events, with 41 being of bacterial origin and one (*Ave1*) being of plant origin. Although most HGT candidates were not found to localize in lineage specific regions, they were significantly more likely to be found within 1 kb distance of repetitive elements than other genes. Overall, we show a high number of inter-kingdom gene acquisitions in *V. dahliae*.

In **Chapter 6**, we reveal the secondary metabolite (SM) gene cluster catalog of *V. dahliae*. We first performed *in silico* predictions using distinctive traits of gene clusters and the conserved signatures of core genes, resulting in 25 potential SM gene clusters. Subsequently, we used phylogenetic- and comparative genomics analyses, revealing that two putative siderophores, ferricrocin and TAFC, DHN-melanin and fujikurin compounds may belong to the SM repertoire of *V. dahliae*.

In **Chapter 7**, the major results in this thesis are discussed and placed in a broader perspective. I show that genomics studies enable us to detect features underlying various mechanisms involved in the adaptive evolution of fungal plant pathogens. Intriguingly, however, most of the features that were identified in diverse plant pathogenic fungi can also be found in non-pathogenic species. This finding suggests that the basis of pathogenicity is rather subtle and that pathogens and non-pathogens should not be seen as two distinct classes of microorganisms but rather form a continuum.



## References

- Acuña R, Padilla BE, Flórez-ramos CP, Rubio JD, Herrera JC, Benavides P** (2012) Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. *Proc Natl Acad Sci U S A* **109**: 2–7
- Agrios GN** (2005) chapter five - How pathogens attack plants. *Plant Pathol.* (Fifth Ed. Academic Press, San Diego, pp 175–205
- Albalat R, Cañestro C** (2016) Evolution by gene loss. *Nat Rev Genet* **17**: 379
- Alexander WG, Wisecaver JH, Rokas A, Hittinger CT** (2016) Horizontally acquired genes in early-diverging pathogenic fungi enable the use of host nucleosides and nucleotides. *Proc Natl Acad Sci U S A* **113**: 4116–4121
- Ali S, Laurie JD, Linning R, Cervantes-Chávez JA, Gaudet D, Bakkeren G** (2014) An immunity-triggering effector from the barley smut fungus *Ustilago hordei* resides in an ustilaginaceae-specific cluster bearing signs of transposable element-assisted evolution. *PLoS Pathog* **10**: e1004223
- Alkan C, Sajjadian S, Eichler EE** (2010) Limitations of next-generation genome sequence assembly. *Nat Methods* **8**: 61
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ** (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403–410
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ** (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402
- Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M** (2004) The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**: 79–86
- Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, Wain J, O’Grady J** (2014) MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat Biotechnol* **33**: 296
- Ayarpadikannan S, Kim H-S** (2014) The impact of transposable elements in genome evolution and genetic instability and their implications in various diseases. *Genomics Inform* **12**: 98
- Aylward J, Steenkamp ET, Dreyer LL, Roets F, Wingfield BD, Wingfield MJ** (2017) A plant pathology perspective of fungal genome sequencing. *IMA Fungus* **8**: 1–45
- Ba ANN, Pogoutse A, Provart N, Moses AM** (2009) NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. *BMC Bioinformatics* **10**: 202
- Baptiste E, O’Malley MA, Beiko RG, Ereshefsky M, Gogarten JP, Franklin-Hall L, Lapointe F-J, Dupré J, Dagan T, Boucher Y, et al** (2009) Prokaryotic evolution and the tree of life are two different things. *Biol Direct* **4**: 34
- von Barga KW, Niehaus E-M, Krug I, Bergander K, Würthwein E-U, Tudzynski B, Humpf H-U** (2015) Isolation and structure elucidation of fujikurins A–D: products of the PKS19 gene cluster in *Fusarium fujikuroi*. *J Nat Prod* **78**: 1809–1815
- Barton NH, Charlesworth B** (1998) Why sex and recombination? *Science* **281**: 1986–1990
- Bell G** (1982) *The Masterpiece of Nature: The Evolution and Genetics of Sexuality*. CUP Archive
- Benjamini Y, Hochberg Y** (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 289–300
- Bernáldez V, Córdoba JJ, Rodríguez M, Cordero M, Polo L, Rodríguez A** (2013) Effect of *Penicillium nalgiovense* as protective culture in processing of dry-fermented sausage “salchichón.” *Food Control* **32**: 69–76
- Bertels F, Silander OK, Pachkov M, Rainey PB, van Nimwegen E** (2014) Automated reconstruction of whole-genome phylogenies from short-sequence reads. *Mol Biol Evol* **31**: 1077–1088
- Betts MF, Tucker SL, Galadima N, Meng Y, Patel G, Li L, Donofrio N, Floyd A, Nolin S, Brown D** (2007) Development of a high throughput transformation system for insertional mutagenesis in *Magnaporthe oryzae*. *Fungal Genet Biol* **44**: 1035–1049
- Biscotti MA, Olmo E, Heslop-Harrison JS (Pat)** (2015) Repetitive DNA in eukaryotic genomes. *Chromosom Res* **23**: 415–420

- Blaise F, Rémy E, Meyer M, Zhou L, Narcy J-P, Roux J, Balesdent M-H, Rouxel T** (2007) A critical assessment of *Agrobacterium tumefaciens*-mediated transformation as a tool for pathogenicity gene discovery in the phytopathogenic fungus *Leptosphaeria maculans*. *Fungal Genet Biol* **44**: 123–138
- Bock R** (2010) The give-and-take of DNA: horizontal gene transfer in plants. *Trends Plant Sci* **15**: 11–22
- Boettger D, Hertweck C** (2013) Molecular diversity sculpted by fungal PKS–NRPS hybrids. *ChemBioChem* **14**: 28–42
- Bolton MD, de Jonge R, Inderbitzin P, Liu Z, Birla K, Van de Peer Y, Subbarao K V, Thomma BPHJ, Secor GA** (2014) The heterothallic sugarbeet pathogen *Cercospora beticola* contains exon fragments of both MAT genes that are homogenized by concerted evolution. *Fungal Genet Biol* **62**: 43–54
- Bolton MD, Thomma BPHJ** (2008) The complexity of nitrogen metabolism and nitrogen-regulated gene expression in plant pathogenic fungi. *Physiol Mol Plant Pathol* **72**: 104–110
- Boshoven J** (2017) Virulence contribution and recognition of homologs of the *Verticillium dahliae* effector Ave1.
- Brakhage AA, Schroeckh V** (2011) Fungal secondary metabolites—strategies to activate silent gene clusters. *Fungal Genet Biol* **48**: 15–22
- Brankovics B, Zhang H, van Diepeningen AD, van der Lee TAJ, Waalwijk C, de Hoog GS** (2016) GRAbB: selective assembly of genomic regions, a new niche for genomic research. *PLoS Comput Biol* **12**: e1004753
- Braslavsky I, Hebert B, Kartalov E, Quake SR** (2003) Sequence information can be obtained from single DNA molecules. *Proc Natl Acad Sci U S A* **100**: 3960–3964
- Brown DW, Lee S-H, Kim L-H, Ryu J-G, Lee S, Seo Y, Kim YH, Busman M, Yun S-H, Proctor RH** (2015) Identification of a 12-gene fusaric acid biosynthetic gene cluster in *Fusarium* species through comparative and functional genomics. *Mol Plant-Microbe Interact* **28**: 319–332
- Brown NA, Antoniw J, Hammond-Kosack KE** (2012) The predicted secretome of the plant pathogenic fungus *Fusarium graminearum*: a refined comparative analysis. *PLOS One* **7**: e33731
- Cacho RA, Tang Y, Chooi Y-H** (2015) Next-generation sequencing approach for connecting secondary metabolites to biosynthetic gene clusters in fungi. *Front Microbiol* **5**: 774
- Cairns T, Meyer V** (2017) In silico prediction and characterization of secondary metabolite biosynthetic gene clusters in the wheat pathogen *Zymoseptoria tritici*. *BMC Genomics* **18**: 631
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL** (2009) BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421
- Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T** (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973
- Casadevall A** (2008) Evolution of intracellular pathogens. *Annu Rev Microbiol* **62**: 19–33
- Castresana J** (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**: 540–552
- Chen K, Durand D, Farach-Colton M** (2000) NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J Comput Biol* **7**: 429–447
- Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE** (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**: 563
- Choi J, Kim S-H** (2017) A genome tree of life for the fungi kingdom. *Proc Natl Acad Sci U S A* **114**: 9391–9396
- Choi J, Park J, Kim D, Jung K, Kang S, Lee Y-H** (2010) Fungal secretome database: integrated platform for annotation of fungal secretomes. *BMC Genomics* **11**: 105
- Collemare J, Griffiths S, Iida Y, Jashni MK, Battaglia E, Cox RJ, de Wit PJGM** (2014) Secondary metabolism and biotrophic lifestyle in the tomato pathogen *Cladosporium fulvum*. *PLOS One* **9**: e85877
- Collemare J, Pianfetti M, Houle A, Morin D, Camborde L, Gagey M, Barbisan C, Fudal I, Lebrun M, Böhnert HU** (2008) *Magnaporthe grisea* avirulence gene ACE1 belongs to an infection-specific gene cluster involved in secondary metabolism. *New Phytol* **179**: 196–208
- Compant S, Duffy B, Nowak J, Clément C, Barka EA** (2005) Use of plant growth-promoting bacteria for biocontrol of plant diseases: principles, mechanisms of action, and future prospects. *Appl Environ Microbiol* **71**: 4951–4959

- Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M** (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**: 3674–3676
- Cook DE, Mesarich CH, Thomma BPHJ** (2015) Understanding plant immunity as a surveillance system to detect invasion. *Annu Rev Phytopathol* **53**: 541–563
- Cox RJ** (2007) Polyketides, proteins and genes in fungi: programmed nano-machines begin to reveal their secrets. *Org Biomolecular Chem* **5**: 2010–2026
- Croll D, McDonald BA** (2012) The accessory genome as a cradle for adaptive evolution in pathogens. *PLOS Pathog* **8**: e1002608
- Cuomo CA, Güldener U, Xu J-R, Trail F, Turgeon BG, Di Pietro A, Walton JD, Ma L-J, Baker SE, Rep M** (2007) The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science* **317**: 1400–1402
- Dallery J-F, Lapalu N, Zampounis A, Pigné S, Luyten I, Amselem J, Wittenberg AHJ, Zhou S, De Queiroz M V, Robin GP** (2017) Gapless genome assembly of *Colletotrichum higginsianum* reveals chromosome structure and association of transposable elements with secondary metabolite gene clusters. *BMC Genomics* **18**: 667
- van Dam P, Fokkens L, Ayukawa Y, Gragt M, Horst A, Brankovics B, Houterman PM, Arie T, Rep M** (2017) A mobile pathogenicity chromosome in *Fusarium oxysporum* for infection of multiple cucurbit species. *Sci Rep* **7**: 9042
- Daverdin G, Rouxel T, Gout L, Aubertot J-N, Fudal I, Meyer M, Parlange F, Carpezat J, Balesdent M-H** (2012) Genome structure and reproductive behaviour influence the evolutionary potential of a fungal phytopathogen. *PLOS Pathog* **8**: e1003020
- Davidson E, Levin M** (2005) Gene regulatory networks. *Proc Natl Acad Sci U S A* **102**: 4935 LP-4935
- Dean R, Van Kan JAL, Pretorius ZA, Hammond-Kosack KE, Di Pietro A, Spanu PD, Rudd JJ, Dickman M, Kahmann R, Ellis J** (2012) The Top 10 fungal pathogens in molecular plant pathology. *Mol Plant Pathol* **13**: 414–430
- Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, Orbach MJ, Thon M, Kulkarni R, Xu J-R, Pan H** (2005) The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature* **434**: 980
- Depotter JRL, Seidl MF, van den Berg GCM, Thomma BPHJ, Wood TA** (2017) A distinct and genetically diverse lineage of the hybrid fungal pathogen *Verticillium longisporum* population causes stem striping in British oilseed rape. *Environ Microbiol* **19**: 3997–4009
- Depotter JRL, Seidl MF, Wood TA, Thomma BP** (2016) Interspecific hybridization impacts host range and pathogenicity of filamentous microbes. *Curr Opin Microbiol* **32**: 7–13
- Depotter JRL, Shi-Kunne X, Missonnier H, Liu T, Faino L, van den Berg GCM, Wood TA, Zhang B, Jacques A, Seidl MF, et al** (2018) Dynamic virulence-related regions of the fungal plant pathogen *Verticillium dahliae* display remarkably enhanced sequence conservation. *bioRxiv* 277558
- Derntl C, Kluger B, Bueschl C, Schuhmacher R, Mach RL, Mach-Aigner AR** (2017) Transcription factor Xpp1 is a switch between primary and secondary fungal metabolism. *Proc Natl Acad Sci U S A* **114**: E560–E569
- Dix NJ, Webster J** (1995) Fungi of Extreme Environments - Fungal Ecology. *In* NJ Dix, J Webster, eds, Springer Netherlands, Dordrecht, pp 322–340
- Dodds PN, Rathjen JP** (2010) Plant immunity: towards an integrated view of plant–pathogen interactions. *Nat Rev Genet* **11**: 539
- Domazet-Lošo T, Tautz D** (2003) An evolutionary analysis of orphan genes in *Drosophila*. *Genome Res* **13**: 2213–2219
- Dong S, Raffaele S, Kamoun S** (2015) The two-speed genomes of filamentous pathogens: waltz with plants. *Curr Opin Genet Dev* **35**: 57–65
- Drillon G, Carbone A, Fischer G** (2014) SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. *PLOS One* **9**: e92621
- Druzhinina IS, Shelest E, Kubicek CP** (2012) Novel traits of *Trichoderma* predicted through the analysis of its secretome. *FEMS Microbiol Lett* **337**: 1–9

- Duplessis S, Cuomo CA, Lin Y-C, Aerts A, Tisserant E, Veneault-Fourrey C, Joly DL, Hacquard S, Amselem J, Cantarel BL** (2011) Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc Natl Acad Sci* **108**: 9166–9171
- Dupont P-Y, Cox MP** (2017) Genomic data quality impacts automated detection of lateral gene transfer in fungi. *G3* **7**: g3.116.038448
- Ebihara Y, Uematsu S, Nagao H, Moriwaki J, Kimishima E** (2003) First report of *Verticillium tricorpus* isolated from potato tubers in Japan. *Mycoscience* **44**: 481–488
- Eddy SR** (2009) A new generation of homology search tools based on probabilistic inference. *Genome Informatics 2009 Genome Informatics Ser. Vol. 23*. World Scientific, pp 205–211
- Eisen JA** (2000) Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Curr Opin Genet Dev* **10**: 606–611
- Ellinghaus D, Kurtz S, Willhoeft U** (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**: 18
- Emanuelsson O, Nielsen H, Brunak S, Von Heijne G** (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Evol* **300**: 1005–1016
- von Esenbeck CGN** (1817) *Das system der pilze und schwämme*. Stahelfchen
- Faino L, Seidl MF, Datema E, van den Berg GCM, Janssen A, Wittenberg AHJ, Thomma BPHJ** (2015) Single-molecule real-time sequencing combined with optical mapping yields completely finished fungal genome. *MBio* **6**: e00936-15
- Faino L, Seidl MF, Shi-Kunne X, Pauper M, Van Den Berg GCM, Wittenberg AHJ, Thomma BPHJ** (2016) Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen. *Genome Res* **26**: 1091–1100
- Faino L, Thomma BPHJ** (2014) Get your high-quality low-cost genome sequence. *Trends Plant Sci* **19**: 288–291
- Felsenstein J** (2002) PHYLIP (Phylogeny Inference Package) Version 3.6 a3.
- Feng P, Shang Y, Cen K, Wang C** (2015) Fungal biosynthesis of the bibenzoquinone oosporein to evade insect immunity. *Proc Natl Acad Sci U S A* **112**: 11365–11370
- Flot J-F, Hespels B, Li X, Noel B, Arkhipova I, Danchin EGJ, Hejnol A, Hennissat B, Koszul R, Aury J-M** (2013) Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. *Nature* **500**: 453
- Fox EM, Howlett BJ** (2008) Secondary metabolism: regulation and role in fungal biology. *Curr Opin Microbiol* **11**: 481–487
- Fradin EF, Thomma BPHJ** (2006) Physiology and molecular aspects of *Verticillium* wilt diseases caused by *V. dahliae* and *V. albo-atrum*. *Mol Plant Pathol* **7**: 71–86
- Fradin EF, Zhang Z, Ayala JCJ, Castroverde CDM, Nazar RN, Robb J, Liu C-M, Thomma BPHJ** (2009) Genetic dissection of *Verticillium* wilt resistance mediated by tomato Ve1. *Plant Physiol* **150**: 320–332
- Friesen TL, Stukenbrock EH, Liu Z, Meinhardt S, Ling H, Faris JD, Rasmussen JB, Solomon PS, McDonald BA, Oliver RP** (2006a) Emergence of a new disease as a result of interspecific virulence gene transfer. *Nat Genet* **38**: 953
- Friesen TL, Stukenbrock EH, Liu Z, Meinhardt S, Ling H, Faris JD, Rasmussen JB, Solomon PS, McDonald BA, Oliver RP** (2006b) Emergence of a new disease as a result of interspecific virulence gene transfer. *Nat Genet* **38**: 953–956
- Fulnečková J, Ševčíková T, Fajkus J, Lukešova A, Lukeš M, Vlček Č, Lang BF, Kim E, Eliaš M, Sýkorova E** (2013) A broad phylogenetic survey unveils the diversity and evolution of telomeres in eukaryotes. *Genome Biol Evol* **5**: 468–483
- Gaderer R, Bonazza K, Seidl-Seiboth V** (2014) Cerato-platanins: a fungal protein family with intriguing properties and application potential. *Appl Microbiol Biotechnol* **98**: 4795–4803
- Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma L-J, Smirnov S, Purcell S** (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* **422**: 859
- Galazka JM, Freitag M** (2014) Variability of chromosome structure in pathogenic fungi—of ‘ends and odds.’ *Curr Opin Microbiol* **20**: 19–26

- Gallo A, Ferrara M, Perrone G** (2013) Phylogenetic study of polyketide synthases and nonribosomal peptide synthetases involved in the biosynthesis of mycotoxins. *Toxins* **5**:717-42
- Gardiner DM, McDonald MC, Covarelli L, Solomon PS, Rusu AG, Marshall M, Kazan K, Chakraborty S, McDonald BA, Manners JM** (2012) Comparative pathogenomics reveals horizontally acquired novel virulence genes in fungi infecting cereal hosts. *PLOS Pathog* **8**: e1002952
- Ghannoum MA** (2000) Potential role of phospholipases in virulence and fungal pathogenesis. *Clin Microbiology Rev* **13**: 122–143
- Gibriel HAY, Thomma BPHJ, Seidl MF** (2016) The age of effectors: genome-based discovery and applications. *Phytopathology* **106**: 1206–1212
- Giesbert S, Schumacher J, Kupas V, Espino J, Segmüller N, Haeuser-Hahn I, Schreier PH, Tudzynski P** (2012) Identification of pathogenesis-associated genes by T-DNA-mediated insertional mutagenesis in *Botrytis cinerea*: a type 2A phosphoprotein phosphatase and an SPT3 transcription factor have significant impact on virulence. *Mol plant-microbe Interact* **25**: 481–495
- Gijzen M, Nürnberger T** (2006) Nep1-like proteins from plant pathogens: recruitment and diversification of the NPP1 domain across taxa. *Phytochemistry* **67**: 1800–1807
- Gill SS, Gill R, Trivedi DK, Anjum NA, Sharma KK, Ansari MW, Ansari AA, Johri AK, Prasad R, Pereira E, et al** (2016) *Piriformospora indica*: potential and significance in plant stress tolerance. *Front Microbiol* **7**: 332
- Gladyshev EA, Meselson M, Arkipova IR** (2008) Massive horizontal gene transfer in bdelloid rotifers. *Science* **320**: 1210–1214
- Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S** (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* **108**: 1513–1518
- Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M** (1996) Life with 6000 genes. *Science* **274**: 546–567
- Goodwin SB, Ben M'Barek S, Dhillon B, Wittenberg AHJ, Crane CF, Hane JK, Foster AJ, Van der Lee TAJ, Grimwood J, Aerts A, et al** (2011) Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. *PLOS Genet* **7**: e1002070
- Gordon JL, Byrne KP, Wolfe KH** (2009) Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. *PLOS Genet* **5**: e1000485
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q** (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**: 644
- Grum-Grzhimaylo AA, Falkoski DL, van den Heuvel J, Valero-Jiménez CA, Min B, Choi I, Lipzen A, Daum CG, Aanen DK, Tsang A** (2018) The obligate alkalophilic soda-lake fungus *Sodiomyces alkalinus* has shifted to a protein diet. *Mol. Ecol* **27**:4808-4819
- Gunde-Cimerman N, Sonjak S, Zalar P, Frisvad JC, Diderichsen B, Plemenitaš A** (2003) Extremophilic fungi in arctic ice: a relationship between adaptation to low temperature and water activity. *Phys Chem Earth, Parts A/B/C* **28**: 1273–1278
- Gurevich A, Saveliev V, Vyahhi N, Tesler G** (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**: 1072–1075
- Gurung S, Short DPG, Hu X, Sandoya G V, Hayes RJ, Koike ST, Subbarao K V** (2015) Host range of *Verticillium isaacii* and *Verticillium klebahnii* from artichoke, spinach, and lettuce. *Plant Dis* **99**: 933–938
- Guy L, Roat Kultima J, Andersson SGE** (2011) genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**: 2334–2335
- Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith Jr RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD** (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**: 5654–5666

## References

- Haas BJ, Kamoun S, Zody MC, Jiang RHY, Handsaker RE, Cano LM, Grabherr M, Kodira CD, Raffaele S, Torto-Alalibo T** (2009) Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* **461**: 393
- Han M V, Thomas GWC, Lugo-Martinez J, Hahn MW** (2013) Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol Biol Evol* **30**: 1987–1997
- Hane JK, Rouxel T, Howlett BJ, Kema GHJ, Goodwin SB, Oliver RP** (2011) A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. *Genome Biol* **12**: R45
- Hashimoto M, Nonaka T, Fujii I** (2014) Fungal type III polyketide synthases. *Nat Prod Rep* **31**: 1306–1317
- Heitman J, Kronstad JW, Taylor JW, Casselton LA** (2007) Sex in fungi: molecular determination and evolutionary implications. ASM Press
- Holt C, Yandell M** (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**: 491
- Horton P, Park K-J, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K** (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res* **35**: W585–W587
- Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, Hon L, Sudmant PH, Graves TA, Alkan C, Dennis MY, et al** (2014) Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res* **24**:688-96
- Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M** (2015) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* **44**: D286–D293
- Husnik F, Mccutcheon JP** (2017) Functional horizontal gene transfer. *Nat Rev Microbiol* **6**:67-79.
- Inderbitzin P, Asvarak T, Turgeon BG** (2010) Six new genes required for production of T-toxin, a polyketide determinant of high virulence of *Cochliobolus heterostrophus* to Maize. *Mol Plant-Microbe Interact* **23**: 458–472
- Inderbitzin P, Bostock RM, Davis MR, Usami T, Platt HW, Subbarao K V** (2011a) Phylogenetics and taxonomy of the fungal vascular wilt pathogen *Verticillium*, with the descriptions of five new species. **6**: e28341
- Inderbitzin P, Davis RM, Bostock RM, Subbarao K V** (2011b) The ascomycete *Verticillium longisporum* is a hybrid and a plant pathogen with an expanded host range. *PLOS One* **6**: e18260
- Inderbitzin P, Subbarao K V** (2014) *Verticillium* systematics and evolution: how confusion impedes *Verticillium* wilt management and how to resolve it. *Phytopathology* **104**: 564–574
- Ip CLC, Loose M, Tyson JR, de Cesare M, Brown BL, Jain M, Leggett RM, Eccles DA, Zalunin V, Urban JM** (2015) MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Research* **4**:1075
- Isaac I** (1967) Speciation in *Verticillium*. *Annu Rev Phytopathol* **5**: 201–222
- Isaac I** (1956) Some soil factors affecting *Verticillium* wilt of *Antirrhinum*. *Ann Appl Biol* **44**: 105–112
- Isaac I** (1953) A further comparative study of pathogenic isolates of *Verticillium*: *V. nubilum* Pethybr. and *V. tricorpus* sp. nov. *Trans Br Mycol Soc* **36**: 180-IN2
- Ishiyuchi K, Nakazawa T, Yagishita F, Mino T, Noguchi H, Hotta K, Watanabe K** (2013) Combinatorial generation of complexity by redox enzymes in the chaetoglobosin A biosynthesis. *J Am Chem Soc* **135**: 7371–7377
- Jaramillo VDA, Sukno SA, Thon MR, Pdf TP, Genomics BMC, Article I, Url A, Central P, Central B** (2015) Identification of horizontally transferred genes in the genus *Colletotrichum* reveals a steady tempo of bacterial to fungal gene transfer. *BMC Genomics* **16**: 2
- Jelen V, De Jonge R, Van de Peer Y, Javornik B, Jakše J** (2016) Complete mitochondrial genome of the *Verticillium*-wilt causing plant pathogen *Verticillium nonalfalfae*. *PLOS One* **11**: e0148525
- Jeon J, Park S-Y, Chi M-H, Choi J, Park J, Rho H-S, Kim S, Goh J, Yoo S, Choi J** (2007) Genome-wide functional analysis of pathogenicity genes in the rice blast fungus. *Nat Genet* **39**: 561
- Johnston C, Martin B, Fichant G, Polard P, Claverys J-P** (2014) Bacterial transformation: distribution, shared mechanisms and divergent control. *Nat Rev Microbiol* **12**: 181–196
- Jones JDG, Dangl JL** (2006) The plant immune system. *Nature* **444**: 323



- Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G** (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**: 1236–1240
- de Jonge R, Bolton MD, Kombrink A, Van Den Berg GCM, Yadeta KA, Thomma BPHJ** (2013) Extensive chromosomal reshuffling drives evolution of virulence in an asexual pathogen. *Genome Res* **23**: 1271–1282
- de Jonge R, Bolton MD, Thomma BPHJ** (2011) How filamentous pathogens co-opt plants: the ins and outs of fungal effectors. *Curr Opin Plant Biol* **14**: 400–406
- de Jonge R, Peter van Esse H, Kombrink A, Shinya T, Desaki Y, Bours R, van der Krol S, Shibuya N, Joosten MHAJ, Thomma BPHJ** (2010) Conserved fungal LysM effector Ecp6 prevents chitin-triggered immunity in plants. *Science* **329**: 953 LP-955
- de Jonge R, Peter van Esse H, Maruthachalam K, Bolton MD, Santhanam P, Saber MK, Zhang Z, Usami T, Lievens B, Subbarao K V., et al** (2012) Tomato immune receptor Ve1 recognizes effector of multiple fungal pathogens uncovered by genome and RNA sequencing. *Proc Natl Acad Sci U S A* **109**: 5110–5115
- de Jonge R, Thomma BPHJ** (2009) Fungal LysM effectors: extinguishers of host immunity? *Trends Microbiol* **17**: 151–157
- Jurka J, Kapitonov V V, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J** (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**: 462–467
- Justice MC, Hsu M-J, Tse B, Ku T, Balkovec J, Schmatz D, Nielsen J** (1998) Elongation factor 2 as a novel target for selective inhibition of fungal protein synthesis. *J Biol Chem* **273**: 3148–3151
- Kamoun S, Furzer O, Jones JDG, Judelson HS, Ali GS, Dalio RJD, Roy SG, Schena L, Zambounis A, Panabières F** (2015) The Top 10 oomycete pathogens in molecular plant pathology. *Mol Plant Pathol* **16**: 413–434
- van Kan JAL, Stassen JHM, Mosbach A, Van Der Lee TAJ, Faino L, Farmer AD, Papisotiriou DG, Zhou S, Seidl MF, Cottam E** (2017) A gapless genome sequence of the fungus *Botrytis cinerea*. *Mol Plant Pathol* **18**: 75–89
- Karapapa VK, Bainbridge BW, Heale JB** (1997) Morphological and molecular characterization of *Verticillium longisporum* comb. nov., pathogenic to oilseed rape. *Mycol Res* **101**: 1281–1294
- Katoh K, Misawa K, Kuma K, Miyata T** (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**: 3059–3066
- Katoh K, Standley DM** (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780
- Kchouk M, Gibrat J-F, Elloumi M** (2017) Generations of sequencing technologies: From first to next generation. *Biol. Med* **9**: 395
- Keller NP, Hohn TM** (1997) Metabolic pathway gene clusters in filamentous fungi. *Fungal Genet Biol* **21**: 17–29
- Keller NP, Turner G, Bennett JW** (2005) Fungal secondary metabolism—from biochemistry to genomics. *Nat Rev Microbiol* **3**: 937
- Kersey PJ, Allen JE, Armean I, Boddu S, Bolt BJ, Carvalho-Silva D, Christensen M, Davis P, Falin LJ, Grabmueller C, et al** (2016) Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Res* **44**: D574–D580
- Khalturin K, Anton-Erxleben F, Sassmann S, Wittlieb J, Hemmrich G, Bosch TCG** (2008) A novel gene family controls species-specific morphological traits in Hydra. *PLOS Biol* **6**: e278
- Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG** (2009) More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet* **25**: 404–413
- Khang CH, Park S-Y, Lee Y-H, Valent B, Kang S** (2008) Genome organization and evolution of the AVR-Pita avirulence gene family in the magnaporthe grisea species complex. *Mol Plant-Microbe Interact* **21**: 658–670
- Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL** (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**: R36
- King R, Urban M, Hammond-Kosack MCU, Hassani-Pak K, Hammond-Kosack KE** (2015) The completed genome sequence of the pathogenic ascomycete fungus *Fusarium graminearum*. *BMC Genomics* **16**: 544

## References

- Klimes A, Amyotte SG, Grant S, Kang S, Dobinson KF** (2008) Microsclerotia development in *Verticillium dahliae*: Regulation and differential expression of the hydrophobin gene VDH1. *Fungal Genet Biol* **45**: 1525–1532
- Klimes A, Dobinson KF, Thomma BPHJ, Klosterman SJ** (2015) Genomics spurs rapid advances in our understanding of the biology of vascular wilt pathogens in the genus *Verticillium*. *Annu Rev Phytopathol* **53**: 181–198
- Klosterman SJ, Atallah ZK, Vallad GE, Subbarao K V** (2009) Diversity, pathogenicity, and management of *Verticillium* species. *Annu Rev Phytopathol* **47**: 39–62
- Klosterman SJ, Subbarao K V., Kang S, Veronese P, Gold SE, Thomma BPHJ, Chen Z, Henrissat B, Lee YH, Park J, et al** (2011) Comparative genomics yields insights into niche adaptation of plant vascular wilt pathogens. *PLOS Pathog* **7**: e1002137.
- Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER** (2013) The next-generation sequencing revolution and its impact on genomics. *Cell* **155**: 27–38
- Köhler GA, Brenot A, Haas-Stapleton E, Agabian N, Deva R, Nigam S** (2006) Phospholipase A 2 and phospholipase B activities in fungi. *Biochim Biophys Acta (BBA)-Molecular Cell Biol Lipids* **1761**: 1391–1399
- Kombrink A** (2014) Functional analysis of LysM effectors secreted by fungal plant pathogens. Wageningen University
- Kombrink A, Rovenich H, Shi-kunne X, Rojas-padilla E, van den Berg GCM, Domazakis E, de Jonge R, Valkenburg D-J, Sánchez-Vallet A, Seidl MF, et al** (2017) *Verticillium dahliae* LysM effectors differentially contribute to virulence on plant hosts. *Mol Plant Pathol* **8**: 596–608
- Kombrink A, Thomma BPHJ** (2013) LysM effectors: secreted proteins supporting fungal life. *PLOS Pathog* **9**: e1003769
- Koonin E V, Makarova KS, Aravind L** (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol* **55**: 709–742
- Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED** (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol* **30**: 693
- 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**: 567–580
- Kubicek CP, Herrera-Estrella A, Seidl-Seiboth V, Martinez DA, Druzhinina IS, Thon M, Zeilinger S, Casas-Flores S, Horwitz BA, Mukherjee PK** (2011) Comparative genome sequence analysis underscores mycoparasitism as the ancestral life style of *Trichoderma*. *Genome Biol* **12**: 1
- Kudo F, Matsuura Y, Hayashi T, Fukushima M, Eguchi T** (2016) Genome mining of the sordarin biosynthetic gene cluster from *Sordaria araneosa* Cain ATCC 36386: characterization of cycloaraneosene synthase and GDP-6-deoxyaltrose transferase. *J Antibiot* **69**: 541
- Kulski JK** (2016) Next-generation sequencing—an overview of the history, tools, and “omic” applications. *Next Gener. Seq. Appl. Challenges*
- Kurland CG, Canback B, Berg OG** (2003) Horizontal gene transfer: a critical view. *Proc Natl Acad Sci U S A* **100**: 9658–9662
- Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL** (2004) Versatile and open software for comparing large genomes. *Genome Biol* **5**: R12
- Laszlo AH, Derrington IM, Ross BC, Brinkerhoff H, Adey A, Nova IC, Craig JM, Langford KW, Samson JM, Daza R, et al** (2014) Decoding long nanopore sequencing reads of natural DNA. *Nat Biotechnol* **32**: 829
- Lawrence JG, Ochman H** (1997) Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* **44**: 383–397
- 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–W245
- Li L, Stoekert CJ, Roos DS** (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**: 2178–2189
- Liao Y, Smyth GK, Shi W** (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res* **41**: e108–e108



- Ligon BL** (2004) Penicillin: its discovery and early development. *Semin. Pediatr. Infect. Dis.* Elsevier, pp 52–57
- Lin H-C, Chooi Y-H, Dhingra S, Xu W, Calvo AM, Tang Y** (2013) The fumagillin biosynthetic gene cluster in *Aspergillus fumigatus* encodes a cryptic terpene cyclase involved in the formation of  $\beta$ -trans-Bergamotene. *J Am Chem Soc* **135**: 4616–4619
- Lipman DJ, Souvorov A, Koonin E V, Panchenko AR, Tatusova TA** (2002) The relationship of protein conservation and sequence length. *BMC Evol Biol* **2**: 20
- Little CR, Carris LM, Stiles CM** (2012) Introduction to Fungi. *Plant Heal Instr*
- Luderer R, Takken FLW, Wit PJGM de, Joosten MHAJ** (2002) *Cladosporium fulvum* overcomes Cf-2-mediated resistance by producing truncated AVR2 elicitor proteins. *Mol Microbiol* **45**: 875–884
- Lukashin A V, Borodovsky M** (1998) GeneMark. hmm: new solutions for gene finding. *Nucleic Acids Res* **26**: 1107–1115
- Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y** (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**: 18
- Lv J, Havlak P, Putnam NH** (2011) Constraints on genes shape long-term conservation of macro-synteny in metazoan genomes. *BMC Bioinformatics*. *BioMed Central*, p S11
- Ma L-J, Van Der Does HC, Borkovich KA, Coleman JJ, Daboussi M-J, Di Pietro A, Dufresne M, Freitag M, Grabherr M, Henrissat B** (2010) Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature* **464**: 367
- Magadam S, Banerjee U, Murugan P, Gangapur D, Ravikesavan R** (2013) Gene duplication as a major force in evolution. *J Genet* **92**: 155–161
- Mallet L V, Becq J, Deschavanne P** (2010) Whole genome evaluation of horizontal transfers in the pathogenic fungus *Aspergillus fumigatus*. *BMC Genomics* **11**: 171
- Marshall R, Kombrink A, Motteram J, Loza-Reyes E, Lucas J, Hammond-Kosack KE, Thomma BPHJ, Rudd JJ** (2011) Analysis of two in planta expressed LysM effector homologs from the fungus *Mycosphaerella graminicola* reveals novel functional properties and varying contributions to virulence on wheat. *Plant Physiol* **156**: 756 LP-769
- Martin F, Aerts A, Ahrén D, Brun A, Danchin EGJ, Duchaussoy F, Gibon J, Kohler A, Lindquist E, Pereda V** (2008) The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature* **452**: 88
- Martín JF, Casqueiro J, Liras P** (2005) Secretion systems for secondary metabolites: how producer cells send out messages of intercellular communication. *Curr Opin Microbiol* **8**: 282–293
- Martinez D, Berka RM, Henrissat B, Saloheimo M, Arvas M, Baker SE, Chapman J, Chertkov O, Coutinho PM, Cullen D** (2008) Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). *Nat Biotechnol* **26**: 553
- McCarthy DJ, Chen Y, Smyth GK** (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res Res* **40**: 4288–4297
- McCarthy DJ, Smyth GK** (2009) Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics* **25**: 765–771
- McClintock B** (1950) The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci* **36**: 344 LP-355
- McDonagh A, Fedorova ND, Crabtree J, Yu Y, Kim S, Chen D, Loss O, Cairns T, Goldman G, Armstrong-James D, et al** (2008) Sub-telomere directed gene expression during initiation of invasive aspergillosis. *PLOS Pathog* **4**: e1000154
- McDonald BA, Linde C** (2002) Pathogen population genetics, evolutionary potential, and durable resistance. *Annu Rev Phytopathol* **40**: 349–379
- McDonald MC, Ahren D, Simpfendorfer S, Milgate A, Solomon PS** (2018) The discovery of the virulence gene ToxA in the wheat and barley pathogen *Bipolaris sorokiniana*. *Mol Plant Pathol* **19**: 432–439
- McGovern PE, Glusker DL, Exner LJ, Voigt MM** (1996) Neolithic resinated wine. *Nature* **381**: 480
- McGovern PE, Zhang J, Tang J, Zhang Z, Hall GR, Moreau RA, Nuñez A, Butrym ED, Richards MP, Wang C** (2004) Fermented beverages of pre-and proto-historic China. *Proc Natl Acad Sci U S A* **101**: 17593–17598
- McNulty SN, Foster JM, Mitreva M, Hotopp JCD, Fischer K, Wu B, Davis PJ, Kumar S, Brattig NW, Slatko BE, et al** (2010) Endosymbiont DNA in Endobacteria-Free Filarial Nematodes Indicates Ancient Horizontal Genetic Transfer. *PLOS One* **5**: e11029

- Medema MH, Fischbach MA** (2015) Computational approaches to natural product discovery. *Nat Chem Biol* **11**: 639
- Medema MH, Takano E, Breitling R** (2013) Detecting sequence homology at the gene cluster level with MultiGeneBlast. *Mol Biol Evol* **30**: 1218–1223
- Mehrabi R, Bahkali AH, Abd-Elsalam KA, Moslem M, Ben M'Barek S, Gohari AM, Jashni MK, Stergiopoulos I, Kema GHJ, de Wit PJGM** (2011) Horizontal gene and chromosome transfer in plant pathogenic fungi affecting host range. *FEMS Microbiol Rev* **35**: 542–554
- Mehrabi R, Taga M, Aghaee M, De Wit PJGM, Kema GHJ** (2012) Karyotyping methods for fungi. *Plant Fungal Pathog.* Springer, pp 591–602
- Menardo F, Wicker T, Keller B** (2017) Reconstructing the evolutionary history of powdery mildew lineages (*Blumeria graminis*) at different evolutionary time scales with NGS data. *Genome Biol Evol* **9**: 446–456
- Mentlak TA, Kombrink A, Shinya T, Ryder LS, Otomo I, Saitoh H, Terauchi R, Nishizawa Y, Shibuya N, Thomma BPHJ, et al** (2012) Effector-mediated suppression of chitin-triggered immunity by *Magnaporthe oryzae* is necessary for rice blast disease. *Plant Cell* **24**: 322 LP-335
- Metzker ML** (2010) Sequencing technologies—the next generation. *Nat Rev Genet* **11**: 31
- Michiels CB, van Wijk R, Reijnen L, Cornelissen BJC, Rep M** (2009) Insight into the molecular requirements for pathogenicity of *Fusarium oxysporum f. sp. lycopersici* through large-scale insertional mutagenesis. *Genome Biol* **10**: R4
- Mohanta TK, Bae H** (2015) The diversity of fungal genome. *Biol Proced Online* **17**: 8
- Möller M, Stukenbrock EH** (2017) Evolution and genome architecture in fungal plant pathogens. *Nat Rev Microbiol* **15**: 756
- Muench S, Ludwig N, Floss DS, Sugui JA, Koszucka AM, Voll LM, Sonnewald UWE, Deising HB** (2011) Identification of virulence genes in the corn pathogen *Colletotrichum graminicola* by *Agrobacterium tumefaciens*-mediated transformation. *Mol Plant Pathol* **12**: 43–55
- Murphy C, Powlowski J, Wu M, Butler G, Tsang A** (2011) Curation of characterized glycoside hydrolases of fungal origin. Database 2011: bar020
- Nakatani Y, Takeda H, Kohara Y, Morishita S** (2007) Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res* **17**: 1254–65
- Nei M, Gojobori T** (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**: 418–426
- Niehaus E-M, Münsterkötter M, Proctor RH, Brown DW, Sharon A, Idan Y, Oren-Young L, Sieber CM, Novák O, Pěnčík A, et al** (2016) Comparative “Omics” of the *Fusarium fujikuroi* species complex highlights differences in genetic potential and metabolite synthesis. *Genome Biol Evol* **8**: 3574–3599
- Norman A, Hansen LH, Sorensen SJ** (2009) Conjugative plasmids: vessels of the communal gene pool. *Philos Trans R Soc B Biol Sci* **364**: 2275–2289
- Nurizzo D, Shewry SC, Perlin MH, Brown SA, Dholakia JN, Fuchs RL, Deva T, Baker EN, Smith CA** (2003) The crystal structure of aminoglycoside-3-O-phospho-transferase-IIa, an enzyme responsible for antibiotic resistance. **2836**: 491–506
- Oide S, Berthiller F, Wiesenberger G, Adam G, Turgeon BG** (2015) Individual and combined roles of malonichrome, ferricrocin, and TAFC siderophores in *Fusarium graminearum* pathogenic and sexual development. *Front Microbiol* **5**: 759
- Olson M V** (1999) When less is more: gene loss as an engine of evolutionary change. *Am J Hum Genet* **64**: 18–23
- Orbach MJ, Farrall L, Sweigard JA, Chumley FG, Valent B** (2000) A telomeric avirulence gene determines efficacy for the rice blast resistance gene *Pi-ta*. *Plant Cell* **12**: 2019 LP-2032
- Palmieri N, Kosiol C, Schlötterer C** (2014) The life cycle of *Drosophila* orphan genes. *Elife* **3**: e01311
- Parra G, Bradnam K, Korfi I** (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**: 1061–1067
- Pauchet Y, Heckel DG** (2013) The genome of the mustard leaf beetle encodes two active xylanases originally acquired from bacteria through horizontal gene transfer. *Proc R Soc B* **280**: 20131021
- Petersen TN, Brunak S, von Heijne G, Nielsen H** (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* **8**: 785

- Pignatelli M, Serras F, Moya A, Guigó R, Corominas M** (2009) CROC: finding chromosomal clusters in eukaryotic genomes. *Bioinformatics* **25**: 1552–1553
- Plett JM, Daguette Y, Wittulsky S, Vayssières A, Deveau A, Melton SJ, Kohler A, Morrell-Falvey JL, Brun A, Veneault-Fourrey C** (2014) Effector MiSSP7 of the mutualistic fungus *Laccaria bicolor* stabilizes the Populus JAZ6 protein and represses jasmonic acid (JA) responsive genes. *Proc Natl Acad Sci U S A* **111**: 8299–304
- Plett JM, Kempainen M, Kale SD, Kohler A, Legué V, Brun A, Tyler BM, Pardo AG, Martin F** (2011) A secreted effector protein of *Laccaria bicolor* is required for symbiosis development. *Curr Biol* **21**: 1197–1203
- Plissonneau C, Benevenuto J, Mohd-Assaad N, Fouché S, Hartmann FE, Croll D** (2017) Using population and comparative genomics to understand the genetic basis of effector-driven fungal pathogen evolution. *Front Plant Sci* **8**: 119
- Plissonneau C, Hartmann FE, Croll D** (2018) Pangenome analyses of the wheat pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome. *BMC Biol* **16**: 5
- Plissonneau C, Stürchler A, Croll D** (2016) The evolution of orphan regions in genomes of a fungal pathogen of wheat. *MBio* **7**: e01231-16
- Ponts N** (2015) Mycotoxins are a component of *Fusarium graminearum* stress-response system. *Front Microbiol* **6**: 1234
- Poptsova M** (2009) Testing phylogenetic methods to identify horizontal gene transfer. *Horiz. Gene Transf.* Springer, pp 227–240
- Poptsova MS, Gogarten JP** (2007) The power of phylogenetic approaches to detect horizontally transferred genes. *BMC Evol Biol* **7**: 45
- Powell M, Gundersen B, Miles C, Coats K, Inglis DA** (2013) First Report of Verticillium Wilt on Lettuce (*Lactuca sativa*) in Washington Caused by *Verticillium tricorpus*. *Plant Dis* **97**: 996
- Prendergast JGD, Campbell H, Gilbert N, Dunlop MG, Bickmore WA, Semple CAM** (2007) Chromatin structure and evolution in the human genome. *BMC Evol Biol* **7**: 72
- Lo Presti L, Lanver D, Schweizer G, Tanaka S, Liang L, Tollot M, Zuccaro A, Reissmann S, Kahmann R** (2015) Fungal effectors and plant susceptibility. *Annu Rev Plant Biol* **66**: 513–545
- Price AL, Jones NC, Pevzner PA** (2005) De novo identification of repeat families in large genomes. *Bioinformatics* **21**: i351–i358
- Price MN, Dehal PS, Arkin AP** (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLOS One* **5**: e9490
- Pusztahelyi T, Holb I, Pócsi I** (2015) Secondary metabolites in fungus-plant interactions. *Front Plant Sci* **6**: 573
- Qin Q-M, Vallad GE, Subbarao K V** (2008) Characterization of *Verticillium dahliae* and *V. tricorpus* Isolates from Lettuce and Artichoke. *Plant Dis* **92**: 69–77
- Quinlan AR, Hall IM** (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842
- Raffaele S, Farrer RA, Cano LM, Studholme DJ, MacLean D, Thines M, Jiang RHY, Zody MC, Kunjeti SG, Donofrio NM, et al** (2010) Genome evolution following host jumps in the Irish potato famine pathogen lineage. *Science* **330**: 1540 LP-1543
- Raffaele S, Kamoun S** (2012) Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat Rev Microbiol* **10**: 417
- Ramanujam R, Naqvi NI** (2010) PdeH, a high-affinity cAMP phosphodiesterase, is a key regulator of asexual and pathogenic differentiation in *Magnaporthe oryzae*. *PLOS Pathog* **6**: e1000897
- Rasoul Z, Walter G, Schroers H-J** (2004) The type species of *Verticillium* is not congeneric with the plant-pathogenic species placed in *Verticillium* and it is not the anamorph of ‘Nectria’ inventa. *Mycol Res* **108**: 576–582
- Rehner SA, Samuels GJ** (1995) Molecular systematics of the Hypocreales: a teleomorph gene phylogeny and the status of their anamorphs. *Can J Bot* **73**: 816–823
- Richards TA, Soanes DM, Jones MDM, Vasieva O, Leonard G, Paszkiewicz K, Foster PG, Hall N, Talbot NJ** (2011) Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. *Proc Natl Acad Sci U S A* **108**: 15258–15263

## References

- Robinson MD, McCarthy DJ, Smyth GK** (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140
- Rodriguez-Moreno L, Ebert MK, Bolton MD, Thomma BPHJ** (2018) Tools of the crook-infection strategies of fungal plant pathogens. *Plant J* **93**: 664–674
- Rouxel T, Grandaubert J, Hane JK, Hoede C, van de Wouw AP, Couloux A, Dominguez V, Anthouard V, Bally P, Bourras S, et al** (2011) Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. *Nat Commun* **2**: 202
- Rovenich H, Boshoven JC, Thomma BPHJ** (2014) Filamentous pathogen effector functions: of pathogens, hosts and microbiomes. *Curr Opin Plant Biol* **20**: 96–103
- Salque M, Bogucki PI, Pyzel J, Sobkowiak-Tabaka I, Grygiel R, Szmyt M, Evershed RP** (2013) Earliest evidence for cheese making in the sixth millennium BC in northern Europe. *Nature* **493**: 522
- Samuel D** (1996) Investigation of ancient Egyptian baking and brewing methods by correlative microscopy. *Science* **273**: 488–490
- Sánchez-Vallet A, Saleem-Batcha R, Kombrink A, Hansen G, Valkenburg D-J, Thomma BPHJ, Mesters JR** (2013) Fungal effector Ecp6 outcompetes host immune receptor for chitin binding through intrachain LysM dimerization. *Elife* **2**: e00790
- Santhanam P, Boshoven JC, Salas O, Bowler K, Islam MT, Saber MK, van den Berg GCM, Bar-Peled M, Thomma BPHJ** (2017) Rhamnose synthase activity is required for pathogenicity of the vascular wilt fungus *Verticillium dahliae*. *Mol Plant Pathol* **18**: 347–362
- Santhanam P, van Esse HP, Albert I, Faino L, Nürnberger T, Thomma BPHJ** (2013) Evidence for functional diversification within a fungal NEP1-like protein family. *Mol Plant-Microbe Interact* **26**: 278–286
- Santhanam P, van Esse HP, Albert I, Faino L, Nürnberger T, Thomma BPHJ** (2012) Evidence for functional diversification within a fungal NEP1-like protein family. *Mol Plant-Microbe Interact* **26**: 278–286
- Santhanam P, Thomma BPHJ** (2012) *Verticillium dahliae* Sge1 differentially regulates expression of candidate effector genes. *Mol Plant-Microbe Interact* **26**: 249–256
- Saravanamuthu R** (2010) Industrial exploitation of microorganisms. IK International Pvt Ltd
- Sarwar M, Akhtar M** (1990) Cloning of aminoglycoside phosphotransferase (APH) gene from antibiotic-producing strain of *Bacillus circulans* into a high-expression vector, pKK223-3. Purification, properties and location of the enzyme. *Biochem J* **268**: 671–677
- Saunders DGO, Win J, Cano LM, Szabo LJ, Kamoun S, Raffaele S** (2012) Using hierarchical clustering of secreted protein families to classify and rank candidate effectors of rust fungi. *PLOS One* **7**: e29847
- Savary S, Ficke A, Aubertot J-N, Hollier C** (2012) Crop losses due to diseases and their implications for global food production losses and food security. *Food Sec* **4**: 519
- Sbaraini N, Andreis FC, Thompson CE, Guedes RLM, Junges Â, Campos T, Staats CC, Vainstein MH, Ribeiro de Vasconcelos AT, Schrank A** (2017) Genome-wide analysis of secondary metabolite gene clusters in *Ophiostoma ulmi* and *Ophiostoma novo-ulmi* reveals a fujikurin-like gene cluster with a putative role in infection. *Front Microbiol* **8**: 1063
- Schirawski J, Mannhaupt G, Münch K, Brefort T, Schipper K, Doehlemann G, Di Stasio M, Rössel N, Mendoza-Mendoza A, Pester D, et al** (2010) Pathogenicity determinants in smut fungi revealed by genome comparison. *Science* **330**: 1546 LP-1548
- Schmidt SM, Lukasiewicz J, Farrer R, van Dam P, Bertoldo C, Rep M** (2016) Comparative genomics of *Fusarium oxysporum* f. sp. *melonis* reveals the secreted protein recognized by the Fom-2 resistance gene in melon. *New Phytol* **209**: 307–318
- Schotanus K, Soyer JL, Connolly LR, Grandaubert J, Happel P, Smith KM, Freitag M, Stukenbrock EH** (2015) Histone modifications rather than the novel regional centromeres of *Zyoseptoria tritici* distinguish core and accessory chromosomes. *Epigenetics Chromatin* **8**: 41
- Schrettl M, Bignell E, Kragl C, Sabiha Y, Loss O, Eisendle M, Wallner A, Arst Jr. HN, Haynes K, Haas H** (2007) Distinct roles for intra- and extracellular siderophores during *Aspergillus fumigatus* infection. *PLOS Pathog* **3**: e128
- Schulz MH, Zerbino DR, Vingron M, Birney E** (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**: 1086–1092

- Seidl MF, Cook DE, Thomma BPHJ** (2016) Chromatin biology impacts adaptive evolution of filamentous plant pathogens. *PLOS Pathog* **12**: e1005920
- Seidl MF, Faino L, Shi-Kunne X, van den Berg GCM, Bolton MD, Thomma BPHJ** (2015) The genome of the saprophytic fungus *Verticillium tricorpus* reveals a complex effector repertoire resembling that of its pathogenic relatives. *Mol Plant-Microbe Interact* **28**: 362–373
- Seidl MF, Thomma BPHJ** (2014) Sex or no sex: Evolutionary adaptation occurs regardless. *BioEssays* **36**: 335–345
- Seidl MF, Thomma BPHJ** (2017) Transposable elements direct the coevolution between plants and microbes. *Trends Genet* **33**: 842
- Seong K, Hou Z, Tracy M, Kistler HC, Xu J-R** (2005) Random insertional mutagenesis identifies genes associated with virulence in the wheat scab fungus *Fusarium graminearum*. *Phytopathology* **95**: 744–750
- Shang Y, Xiao G, Zheng P, Cen K, Zhan S, Wang C** (2016) Divergent and convergent evolution of fungal pathogenicity. *Genome Biol Evol* **8**: 1374–1387
- Shi-Kunne X, Faino L, van den Berg GCM, Thomma BPHJ, Seidl MF** (2018) Evolution within the fungal genus *Verticillium* is characterized by chromosomal rearrangement and gene loss. *Environ Microbiol* **20**: 1362–1373
- Short DPG, Gurung S, Hu X, Inderbitzin P, Subbarao K V** (2014) Maintenance of sex-related genes and the co-occurrence of both mating types in *Verticillium dahliae*. *PLOS One* **9**: e112145
- Sieber CMK, Lee W, Wong P, Münsterkötter M, Mewes H-W, Schmeitzl C, Varga E, Berthiller F, Adam G, Güldener U** (2014) The *Fusarium graminearum* genome reveals more secondary metabolite gene clusters and hints of horizontal gene transfer. *PLOS One* **9**: e110311
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V, Zdobnov EM** (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210–3212
- Simms D, Cizdziel PE, Chomczynski P** (1993) TRIzol: A new reagent for optimal single-step isolation of RNA. *Focus* **15**: 532–535
- Simpson AGB, Roger AJ** (2004) The real ‘kingdoms’ of eukaryotes. *Curr Biol* **14**: R693–R696
- Smit AFA, Hubley R** (2010) Open-1.0. Repeat Masker Website. <http://www.repeatmasker.org/>
- Smit AFA, Hubley R** (2008) RepeatModeler Open-1.0. <http://www.repeatmasker.org/RepeatModeler/>
- Smit AFA, Hubley R, Green P** (1996) RepeatMasker Open-3.0. <http://www.repeatmasker.org/>
- Smit AFA, Hubley R, Green P** (2015) RepeatMasker Open-4.0. <http://www.repeatmasker.org/>
- Snelders NC, Kettles GJ, Rudd JJ, Thomma BPHJ** (2018) Plant pathogen effector proteins as manipulators of host microbiomes? *Mol Plant Pathol* **19**: 257–259
- Soanes D, Richards TA** (2014) Horizontal gene transfer in eukaryotic plant pathogens. *Annu Rev Phytopathol* **52**: 583–614
- Song Y, Zhang Z, Boshoven J, Rovenich H, Seidl M, Jakse J, Maruthachalam K, Liu C-M, Subbarao K, Javornik B, et al** (2017) Tomato immune receptor Ve1 recognizes surface-exposed co-localized N- and C-termini of *Verticillium dahliae* effector Ave1. *bioRxiv* 103473
- Sonjak S, Uršič V, Frisvad JC, Gunde-Cimerman N** (2007) *Penicillium svalbardense*, a new species from Arctic glacial ice. *Antonie Van Leeuwenhoek* **92**: 43–51
- Soragni E, Bolchi A, Balestrini R, Gambaretto C, Percudani R, Bonfante P, Ottonello S** (2001) A nutrient-regulated, dual localization phospholipase A2 in the symbiotic fungus *Tuber borchii*. *EMBO J* **20**: 5079–5090
- Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM, Stüber K, van Themaat EVL, Brown JKM, Butcher SA, Gurr SJ** (2010) Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science* **330**: 1543–1546
- Sperschneider J, Dodds PN, Singh KB, Taylor JM** (2018) ApoplastP: prediction of effectors and plant proteins in the apoplast using machine learning. *New Phytol* **217**: 1764–1778
- Sperschneider J, Gardiner DM, Dodds PN, Tini F, Covarelli L, Singh KB, Manners JM, Taylor JM** (2016) EffectorP: predicting fungal effector proteins from secretomes using machine learning. *New Phytol* **210**: 743–761
- Sperschneider J, Gardiner DM, Thatcher LF, Lyons R, Singh KB, Manners JM, Taylor JM** (2015) Genome-wide analysis in three *Fusarium* pathogens identifies rapidly evolving chromosomes and genes associated with pathogenicity. *Genome Biol Evol* **7**: 1613–1627

## References

- Stamatakis A** (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688–2690
- Stamatakis A** (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313
- Stamatakis A, Hoover P, Rougemont J** (2008) A Rapid Bootstrap Algorithm for the RAxML Web Servers. *Syst Biol* **57**: 758–771
- Stanke M, Steinkamp R, Waack S, Morgenstern B** (2004) AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res* **32**: W309–W312
- Stergiopoulos I, de Wit PJGM** (2009) Fungal effector proteins. *Annu Rev Phytopathol* **47**: 233–263
- Stukenbrock EH, Duteil JY** (2012) Comparing fungal genomes: Insight into functional and evolutionary processes. *Plant Fungal Pathog*. Springer, pp 531–548
- Stukenbrock EH, Jørgensen FG, Zala M, Hansen TT, McDonald BA, Schierup MH** (2010) Whole-genome and chromosome evolution associated with host adaptation and speciation of the wheat pathogen *Mycosphaerella graminicola*. *PLoS Genet* **6**: e1001189
- Suh M-J, Fedorova ND, Cagas SE, Hastings S, Fleischmann RD, Peterson SN, Perlin DS, Nierman WC, Pieper R, Momany M** (2012) Development stage-specific proteomic profiling uncovers small, lineage specific proteins most abundant in the *Aspergillus fumigatus* conidial proteome. *Proteome Sci* **10**: 30
- Taanman J-W** (1999) The mitochondrial genome: structure, transcription, translation and replication. *Biochim Biophys Acta (BBA)-Bioenergetics* **1410**: 103–123
- Talbot NJ, Ebbole DJ, Hamer JE** (1993) Identification and characterization of MPG1, a gene involved in pathogenicity from the rice blast fungus *Magnaporthe grisea*. *Plant Cell* **5**: 1575 LP-1590
- Thomma BPHJ** (2003) *Alternaria* spp.: from general saprophyte to specific parasite. *Mol Plant Pathol* **4**: 225–236
- Thomma BPHJ, Nürnberger T, Joosten MHAJ** (2011) Of PAMPs and effectors: the blurred PTI-ETI dichotomy. *Plant Cell* **23**:4-15
- Thomma BPHJ, Seidl MF, Shi-Kunne X, Cook DE, Bolton MD, van Kan JAL, Faino L** (2016) Mind the gap; seven reasons to close fragmented genome assemblies. *Fungal Genet Biol* **90**: 24–30
- Tisserant E, Malbreil M, Kuo A, Kohler A, Symeonidi A, Balestrini R, Charron P, Duensing N, dit Frey NF, Gianinazzi-Pearson V** (2013) Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. *Proc Natl Acad Sci U S A* **110**: 20117–20122
- Tobiasen C, Aahman J, Ravnholt KS, Bjerrum MJ, Grell MN, Giese H** (2007) Nonribosomal peptide synthetase (NPS) genes in *Fusarium graminearum*, *F. culmorum* and *F. pseudograminearum* and identification of NPS2 as the producer of ferricrocin. *Curr Genet* **51**: 43–58
- Trapnell C, Pachter L, Salzberg SL** (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105–1111
- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L** (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511
- Tritt A, Eisen JA, Facciotti MT, Darling AE** (2012) An integrated pipeline for de novo assembly of microbial genomes. *PLOS One* **7**: e42304
- Trower MK, Clark KG** (1990) PCR cloning of a streptomycin phosphotransferase (aphE) gene from *Streptomyces griseus* ATCC 12475. *Nucleic Acids Res* **18**: 4615
- Tsuji G, Kenmochi Y, Takano Y, Sweigard J, Farrall L, Furusawa I, Horino O, Kubo Y** (2002) Novel fungal transcriptional activators, Cmr1p of *Colletotrichum lagenarium* and Pig1p of *Magnaporthe grisea*, contain Cys2His2 zinc finger and Zn(II)2Cys6 binuclear cluster DNA-binding motifs and regulate transcription of melanin biosynthesis. *Mol Microbiol* **38**: 940–954
- Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RHY, Aerts A, Arredondo FD, Baxter L, Bensasson D, Beynon JL** (2006) Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* **313**: 1261–1266
- Tyson JR, O’Neil NJ, Jain M, Olsen HE, Hieter P, Snutch TP** (2018) MinION-based long-read sequencing and assembly extends the *Caenorhabditis elegans* reference genome. *Genome Res* **28**: 266–274



- Usami T, Kanto T, Inderbitzin P, Itoh M, Kisaki G, Ebihara Y, Suda W, Amemiya Y, Subbarao K V** (2011) *Verticillium tricorpus* causing lettuce wilt in Japan differs genetically from California lettuce isolates. *J Gen plant Pathol* **77**: 17–23
- Vakirlis N, Sarilar V, Drillon G, Fleiss A, Agier N, Meyniel J-P, Blanpain L, Carbone A, Devillers H, Dubois K** (2016) Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus. *Genome Res* **26**:918–32
- Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Müller R, Wohlleben W, et al** (2015) antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res* **43**: W237–W243
- Webster J, Weber R** (2007) Introduction to fungi. Cambridge University Press
- Westerink N, Brandwagt BF, De Wit PJGM, Joosten MHAJ** (2004) *Cladosporium fulvum* circumvents the second functional resistance gene homologue at the *Cf-4* locus (*Hcr9-4E*) by secretion of a stable *avr4E* isoform. *Mol Microbiol* **54**: 533–545
- Wicker T, Oberhaensli S, Parlange F, Buchmann JP, Shatalina M, Roffler S, Ben-David R, Doležel J, Šimková H, Schulze-Lefert P** (2013) The wheat powdery mildew genome shows the unique evolution of an obligate biotroph. *Nat Genet* **45**: 1092
- Wiemann P, Keller NP** (2014) Strategies for mining fungal natural products. *J Ind Microbiol Biotechnol* **41**: 301–313
- Wolpert TJ, Dunkle LD, Ciuffetti LM** (2002) Host-selective toxins and avirulence determinants: what's in a name? *Annu Rev Phytopathol* **40**: 251–285
- van de Wouw AP, Cozijnsen AJ, Hane JK, Brunner PC, McDonald BA, Oliver RP, Howlett BJ** (2010) Evolution of linked avirulence effectors in *Leptosphaeria maculans* is affected by genomic environment and exposure to resistance genes in host plants. *PLOS Pathog* **6**: e1001180
- Xiong D, Wang Y, Tian L, Tian C** (2016) MADS-Box transcription factor VdMcm1 regulates conidiation, microsclerotia formation, pathogenicity, and secondary metabolism of *Verticillium dahliae*. *Front Microbiol* **7**: 1192
- Xu Z, Wang H** (2007) LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* **35**: W265–W268
- Yang Z** (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–1591
- Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y** (2012) dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* **40**: W445–W451
- Yin Z, Tang W, Wang J, Liu X, Yang L, Gao C, Zhang J, Zhang H, Zheng X, Wang P** (2016) Phosphodiesterase MoPdeH targets MoM ck1 of the conserved mitogen-activated protein (MAP) kinase signalling pathway to regulate cell wall integrity in rice blast fungus *Magnaporthe oryzae*. *Mol Plant Pathol* **17**: 654–668
- Yu X, Huo L, Liu H, Chen L, Wang Y, Zhu X** (2015) Melanin is required for the formation of the multi-cellular conidia in the endophytic fungus *Pestalotiopsis microspora*. *Microbiol Res* **179**: 1–11
- Zalar P, Gostinčar C, De Hoog GS, Uršič V, Sudhaham M, Gunde-Cimerman N** (2008) Redefinition of *Aureobasidium pullulans* and its varieties. *Stud Mycol* **61**: 21–38
- Zare R, Gams W, Starink-Willemse M, Summerbell RC** (2007) *Gibellulopsis*, a suitable genus for *Verticillium nigrescens*, and *Musicillium*, a new genus for *V. theobromae*. *Nov Hedwigia* **85**: 463–489
- Zdobnov EM, Apweiler R** (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**: 847–848
- Zhang D-D, Wang X-Y, Chen J-Y, Kong Z-Q, Gui Y-J, Li N-Y, Bao Y-M, Dai X-F** (2016) Identification and characterization of a pathogenicity-related gene VdCYP1 from *Verticillium dahliae*. *Sci Rep* **6**: 27979
- Zhang H, Liu K, Zhang X, Tang W, Wang J, Guo M, Zhao Q, Zheng X, Wang P, Zhang Z** (2011) Two phosphodiesterase genes, PDEL and PDEH, regulate development and pathogenicity by modulating intracellular cyclic AMP levels in *Magnaporthe oryzae*. *PLOS One* **6**: e17241
- Zhang H, Meltzer P, Davis S** (2013) RCircos: an R package for Circos 2D track plots. *BMC Bioinformatics* **14**: 244
- Zhao Z, Liu H, Wang C, Xu J-R** (2013) Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. *BMC Genomics* **14**: 274

## References

- Zhu Q, Sun L, Lian J, Gao X, Zhao L, Ding M, Li J, Liang Y** (2016) The phospholipase C (FgPLC1) is involved in regulation of development, pathogenicity, and stress responses in *Fusarium graminearum*. *Fungal Genet Biol* **97**: 1–9
- Zuccaro A, Lahrmann U, Guldener U, Langen G, Pfiffi S, Biedenkopf D, Wong P, Samans B, Grimm C, Basiewicz M** (2011) Endophytic life strategies decoded by genome and transcriptome analyses of the mutualistic root symbiont *Piriformospora indica*. *PLoS Pathog* **7**: e1002290
- Zupančič J, Babič MN, Zalar P, Gunde-Cimerman N** (2016) The black yeast *Exophiala dermatitidis* and other selected opportunistic human fungal pathogens spread from dishwashers to kitchens. *PLOS One* **11**: e0148166



## Acknowledgements

Almost to my surprise, this thesis here finds its final form and shape. It is great to sit down for a moment and contemplate the time that led to this moment. I happily appreciate this opportunity to direct some words to those who helped me out, and worked with me during these years. You are all so connected to the time I worked on my thesis, which would not be completed without you.

First of all, I would like to thank my supervisor and promoter. Bart, thank you for giving me the opportunity to join your group. I enjoyed and benefited a lot from this multidisciplinary environment. Bart, I appreciate the tremendous freedom you have given me, although sometimes I felt I was “drowning” in it. You told me to follow my heart. You were convinced that I could learn “how to swim” by myself eventually. At the same time, you were patient with the learning process. It took a bit longer, but we all made it. Bart, I am very impressed by your dedication to work. I learnt from you that every job should start with a passion. I am sure you will continue your passion during your next journey.

Luigi, thanks for “luring” me into the “computer gang”. You taught me a lot, you made me from a Linux system “noob” into a scripting “nerd”. You are full of energy and crazy ideas. You are kind and helpful as a friend, bossy as a supervisor. I am lucky to have you as my mentor and friend. Michael, I am glad you are my co-promotor. You are a true walking library. I don’t know how you manage to remember all the relevant old and new papers. You also enjoy informing others with freshly published papers that are relevant to their topics. This is just one of the examples of your helpfulness. Your contribution to this PhD thesis is highly valued. I was spoiled in a way that I could go to you almost anytime to either ask you for help when I was stuck or share new results when new analyses just worked out. You are very critical in science, or maybe you are a perfectionist. We sometimes ran into discussions about which orientation or colour of certain figures could fit better.

A lot of thanks also go to other members of the “computer gang”. Sander, thanks for spotting and fixing all bugs in my scripts. I sometimes felt that you enjoyed reading scripts more than papers. Jasper, thanks for listening to and commenting on any crazy or stupid ideas I had. Hesham, thanks for your quick maths and free music that leaked out from your headphones, which very often triggered Jasper to sing along.

There were others as well that were a great support during my PhD. Among these, I would like to thank the rest of the *Verticillium* members. Grady, thanks for providing me with high quality DNA for sequencing in less than half of the time I would have spent. Also, I want to thank the former postdocs, Luigi, Michael, David, Mireille Andrea and Luis who looked at the progresses of PhDs critically. Also, all the PhDs during my time, Anja, Yin, Hanna, Jordi, Eduardo, Hui, Jinling, Malaika, Martin and Nick who contributed to the stimulating and helpful atmosphere of the group. A lot of thanks to my students, Bart, Yachi, Mathijs and Roger, who helped me during my PhD.

## Acknowledgements

I would like to thank all people from Phytopathology. Thank you all for this interactive, stimulating working environment. At the same time, I enjoyed a lot of social activities with fellow PhDs, through sushi evenings, bowling nights and Zumba Fridays etc.

Apart from work, I am so lucky to have great friends, many of you directly or indirectly helped me through my PhD. Su and Rob, I am happy for your new chapter in life, which unfortunately made our weekly dinner into almost monthly. Hanna and Ale, I really disliked the fact that you moved away. I enjoyed talking and sharing great food with you guys. I have to say I enjoyed watching Hanna struggling with chopsticks as well. Jordi and Lliana, you were one of the first to get babies among our close friends. I am so amazed how you guys balance work and family life so well. Pingping and Cheng, thanks for supporting me mentally during my PhD by listening to all my happy and unhappy stories. Benoit, or BenBen, thanks for your positive attitude and a lot of bioinformatics advice, most importantly thanks for giving me the perfect kitten who accompanied me with purring sounds during my thesis writing. Kun, thanks for adding dramas to my life. Jasper, Laura, Sander and Kim, thanks for all the board game nights in combination with tasting many kinds of beers, wines and whiskies. Also, the rest of my friends who I am not specifically mentioning here, I want to address my sincere gratitude for making my life so colourful.

Special thanks to my parents.我明白你们能够尊重并且支持我常年在海外求学并且定居海外是一件非常不易的事。谢谢爸爸妈妈多年无条件的付出。谢谢你们的爱。因为你们的支持。我的人生有了新的高度。 Although I live far away from my parents, I was lucky to have a family close by. Olaf and Andrea, thanks for all your love and support over all the years. I could easily talk to you, which changed my perception that the communication between two generations and two nationalities can be so free and joyful. Also, thanks to the rest of the family members, I appreciate a lot that you all made great efforts to speak English when I was around. I am glad our family has grown during the last years, still more dogs than babies though. I enjoyed our family gathering events a lot, until the noise levels exploded me from inside. And finally, though he can't possibly know how much of a help he has been, I'd like to thank Sir Inti MacPurrface (yes, I thanked my cat), who puts the "companion" in "companion animal".

Of course, I want to thank my husband, Tim, without whom this thesis would not have been finished. Not because you overused the meme, "are you doctor yet?", nor because you made it look like doing a PhD is such an easy task. I could share all my thoughts with you, including ideas of my new experiments, results of my analyses, layout of my thesis, etc. You even proof read all my chapters. Because of these passive training sessions, I am pretty sure you can follow my defence discussion very well. Mimi, Thanks for all your understanding and love. In Life, it's not where you go, it's who you travel with. I enjoy all the journeys and adventures we have taken together.

## Curriculum vitae

Xiaoqian Shi was born on the 5<sup>th</sup> of August, 1988 in Ningxia, China. In 2006, Xiaoqian started a “2+2” bachelor program on Horticulture and Marketing, which allowed her to study two years in China Agriculture University in Beijing, China, and two years in Van Hall Larenstein in Wageningen, the Netherlands. After her Bachelor study, Xiaoqian continued her study with a double master in Plant science (specialized in Plant breeding) and Plant biotechnology (specialized in functional genomics). Xiaoqian did her first major MSc thesis on “near-reverse breeding” in the group of Prof. Hans de Jong, supervised by Dr. Erik Wijnker. During this thesis, Xiaoqian focused on using natural (non-GMO) methods to simplify meiosis of hybrid *Arabidopsis* such that homozygous parental lines can be generated from a vigorous hybrid individual. Afterwards, Xiaoqian performed her minor thesis in the group of prof. Gary Loake of the University of Edinburgh, UK, working on “localization of GAPDH under pathogen challenges in *Arabidopsis*”. After returning from Scotland, Xiaoqian started her second major thesis in the group of prof. Bart Thomma, supervised by Dr. Luigi Faino, working on “genome assembly and comparative genomics of *Verticillium* species”. She focused on comparing different assembly software in order to obtain the best genome assembly from Illumina short reads. This thesis study incited Xiaoqian’s interest in pathogen genomics. Coincidentally, prof. Thomma opened two PhD positions, one of which was about genomics in *Verticillium* species. Xiaoqian obtained this position and the result of this project is presented in this thesis.

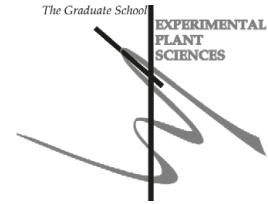
## List of publications

- Shi-Kunne X\***, de Jove R\*, Depotter JRL<sup>†</sup>, Ebert MK<sup>†</sup>, Seidl MF<sup>#</sup>, Thomma BPHJ<sup>#</sup> (2018) In silico prediction and characterisation of secondary metabolite clusters in the plant pathogenic fungus *Verticillium dahliae* bioRxiv 481648
- Depotter JRL\*, **Shi-Kunne X\***, Missonnier H<sup>†</sup>, Liu T<sup>†</sup>, Faino L, van den Berg GCM, Wood TA, Zhang B<sup>†</sup>, Jacques A<sup>†</sup>, Seidl MF<sup>#</sup>, Thomma BPHJ<sup>#</sup> (2018) Dynamic virulence-related regions of the fungal plant pathogen *Verticillium dahliae* display remarkably enhanced sequence conservation. bioRxiv 277558
- Shi-Kunne X**, van Kooten M\*, Depotter JRL\*, Thomma BPHJ<sup>#</sup>, Seidl MF<sup>#</sup> (2019) The genome of the fungal pathogen *Verticillium dahliae* reveals extensive bacterial to fungal gene transfer. Genome Biol Evol pii: evz040
- Shi-Kunne X**, Faino L, van den Berg GCM, Thomma BPHJ<sup>#</sup>, Seidl MF<sup>#</sup> (2018) Evolution within the fungal genus *Verticillium* is characterized by chromosome rearrangement and gene loss. Environ. Microbiol 20:1362-1373
- Kombrink A, Rovenich H\*, **Shi-Kunne X\***, Rojas-padilla E\*, van den Berg GCM, Domazakis E, de Jonge R, Valkenburg D-J, Sánchez-Vallet A, Seidl MF<sup>#</sup>, Thomma BPHJ<sup>#</sup> (2017) *Verticillium dahliae* LysM effectors differentially contribute to virulence on plant hosts. Mol Plant Pathol 18: 596–608
- Faino L\*, Seidl MF<sup>#</sup>, **Shi-Kunne X**, Pauper M, van den Berg GC, Wittenberg AH, Thomma BPHJ (2016): Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen. Genome Res 26:1091-1100
- Thomma BPHJ, Seidl MF, **Shi-Kunne X**, Cook DE, Bolton MD, van Kan JAL, Faino L (2016) Mind the gap; seven reasons to close fragmented genome assemblies. Fungal Genet Biol 90: 24–30
- Shi-Kunne X**, Seidl M, Faino L, Thomma BPHJ (2015) Draft genome sequence of a strain of cosmopolitan fungus *Trichoderma atroviride*. Genome Announc 3: 3–4
- Seidl MF\*, Faino L\*, **Shi-Kunne X**, van den Berg GCM, Bolton MD, Thomma BPHJ (2015) The genome of the saprophytic fungus *Verticillium tricorpus* reveals a complex effector repertoire resembling that of its pathogenic relatives. Mol Plant-Microbe Interact 28: 362–373

\*, †, # equal contribution

## Education Statement of the Graduate School Experimental Plant Sciences

Issued to: Xiaoqian Shi  
Date: 03 May 2019  
Group: Laboratory of Phytopathology  
University: Wageningen University & Research



<b>1) Start-Up Phase</b>	<i>date</i>	<i>cp</i>
▶ <b>First presentation of your project</b>		
Genomic rearrangements in <i>Verticillium</i> genus (presentation at phytopathology)	3 Oct 2014	1,5
▶ <b>Writing or rewriting a project proposal</b>		
PhD project proposal	Aug 2013	1,5
▶ <b>Writing a review or book chapter</b>		
▶ <b>MSc courses</b>		
<i>Subtotal Start-Up Phase</i>		3,0

<b>2) Scientific Exposure</b>	<i>date</i>	<i>cp</i>
▶ <b>EPS PhD student days</b>		
EPS PhD day "Get2Gether", Soest, NL	29-30 Jan 2015	0,6
EPS PhD day "Get2Gether", Soest, NL	28-29 Jan 2016	0,6
▶ <b>EPS theme symposia</b>		
EPS Theme 4: Genome biology	13 Dec 2013	0,3
EPS Theme 2: Interactions between Plants and Biotic Agents	25 Feb 2014	0,3
EPS Theme 4: Genome biology	3 Dec 2014	0,3
EPS Theme 2: Interactions between Plants and Biotic Agents	22 Jan 2016	0,3
EPS Theme 4: Genome biology	15 Dec 2016	0,3
▶ <b>National meetings (e.g. Lunteren days) and other National Platforms</b>		
'NBIC 2014' - Netherlands Bioinformatics Conference, Lunteren	08-09 Apr 2014	0,6
ALW meeting 'Experimental Plant Sciences', Lunteren	14-15 Apr 2014	0,6
ALW meeting 'Molecular genetics', Lunteren	09-10 Oct 2014	0,6
ALW meeting 'Experimental Plant Sciences', Lunteren	13-14 Apr 2015	0,6
Dutch bioinformatics & systems biology conference, Lunteren	19-20 Apr 2016	0,6
Molecular genetics meeting, Wageningen, NL	21 Oct 2016	0,3
▶ <b>Seminars (series), workshops and symposia</b>		
<i>Symposia</i>		
Symposium: Pacbio seminar	26 Mar 2014	0,3

Farewell symposium Pierre de Wit, Wageningen, NL	5 Jun 2014	0,3
Crop Pathology and Plant-Microbe Interactions Symposium	8 May 2015	0,3
WUR symposium "From Big Data to Biological Solutions"	18 Jun 2015	0,3
Symposium 'WURomics Technology- Driven Innovation for Plant Breeding'	15 Dec 2016	0,3
MINI-SYMPOSIUM APPLIED PHYTOPATHOLOGY from the lab to the field, Wageningen, NL	1 Mar 2017	0,2
Dies Natalis Symposium, Wageningen, NL	9 Mar 2017	0,3
<i>Workshops</i>		
COST SUSTAIN workshop "Evolutionary genomics of plant pathogens"	26-28 Aug 2015	0,9
Workshop BU Biointeractions & Plant Health - WU Phytopathology, Wageningen, NL	10 Feb 2015	0,2
Focus meeting New sequencing technology, Utrecht, NL	8 Mar 2016	0,2
Sino Dutch vegetable breeding sector innovation workshop	22 Sep 2016	0,2
<i>Seminars</i>		
Soilborn pathogens and their natural biocontrol agents in cereal-based production system (David Weller)	25 Sep 2013	0,1
Genome and effector evolution in the Irish potato famine pathogen lineage (Sophien Kamoun)	28 May 2014	0,1
Back to the roots (Jos Raaijmakers)	7 Jan 2014	0,1
Public lecture 'The Bonobo and the Atheist' (Frans de Waal)	12 Jun 2014	0,1
Dissecting the interactions between Phytophthora sojae and soy bean: making sense of signalling and effectors (Yunchao Wang)	16 Jul 2014	0,1
Evolution of plant-herbivore interactions: insights from genomics (Noah Whiteman)	17 Jul 2014	0,1
WEES seminar "Adaptation and Epistasis in Laboratory Budding Yeast" (Michael Desai)	16 Oct 2014	0,1
Chromatin structure controls centromeres and secondary metabolism in filamentous fungi (Michael Freitag)	21 Oct 2014	0,1
EPS Flying Seminar "Inferring species trees given coalescence and reticulation" (Michael D. Pirie)	18 Mar 2015	0,1
EPS Plant Sciences Seminar 'Tomato metabolomics in 2015, the difference a genome makes' (Alisdair Fernie)	11 Mar 2015	0,1
Seminar 'Structure and evolution of centromeres: lessons learned from plants' (Jiming Jiang)	1 Apr 2015	0,1
EPS Flying Seminar 'Long-distance endosome trafficking drives fungal effector production during plant infection' (Gero Steinberg)	5 Jun 2015	0,1
Flying Seminar "Polyploidy in wild relatives of soybean and other legumes: systematics, comparative and functional genomics, and nodulation" (Jeff Doyle)	12 May 2015	0,1
EPS Flying Seminar "And yet they oscillate: functional analysis of circadian long non-coding RNAs" (Rossana Henriques)	16 Nov 2015	0,1
EPS Flying Seminar 'Plant intracellular immunity: evolutionary and molecular underpinnings' (Jane Parker)	21 Jan 2016	0,1
PSI seminar "the importance of the plant soil interaction" (Olga Kostenko)	3 Feb 2016	0,1

Towards understanding rice brown spot, a disease induced by physiological stress	6 Feb 2015	0,1
Microbial Population Biology seminar (Fons Debets and Bart Thomma)	21 Apr 2017	0,1
The evolutionary significance succeeds in plant xylem vessels (Caitilyn Allen)	29 Apr 2016	0,1
Effectors are molecular probes to understand pathogenesis (Wenbo Ma)	20 Jun 2016	0,1
Public lecture "Rewriting our genes?" (Jenifer Doudna and Edze Westra)	30 Sep 2016	0,1
WEES seminar "Evolving immunity: Genomic basis of the evolution and variation in parasitoid resistance"	19 Jan 2017	0,1
Plant Microbiome Network seminar (Je-Seung Jeon and Michael Seidl)	18 Apr 2017	0,1
Evolution of symbiotic gene networks in land plants (Pierre-Marc Delaux)	8 Apr 2016	0,1
WPMN seminar (Gerlinde de Deyn, Martinus Schneiderberg)	11 Oct 2016	0,1
WPMN seminar (Irene de Bruijn, Sven Warris)	21 Feb 2017	0,1
WEES seminar "Evolving immunity: Genomic basis of the evolution and variation in parasitoid resistance" (Bregje Wertheim)	19 Jan 2017	0,1
"Coffee rust <i>Hemileia vastatrix</i> : its history, economic significance, fungus-host interactions and control measures" (Frits Rijkenberg)	10 May 2017	0,1
"Genome wide association as a tool for identifying fungal effectors important in virulence" (Timothy Friesen)	18 Apr 2018	0,1
► <b>Seminar plus</b>		
Michael Freitag, Oregon State University, USA	21 Oct 2014	0,1
Caitilyn Allen, University of Wisconsin-Maison, USA	29 Apr 2014	0,1
Wenbo Ma, UC Riverside, USA	20 Jun 2016	0,1
► <b>International symposia and congresses</b>		
28th Fungal Genetics Conference Pacific Grove, CA, USA	17-22 Mar 2015	1,5
► <b>Presentations</b>		
<i>Posters</i>		
The genome of <i>Verticillium tricorpus</i> facilitates comparative genomics to reveal the evolution of virulence, PhD spring school Host-Microbe Interactomics	2 Jun 2014	1,0
Chromosomal rearrangements are not confined to pathogenic <i>Verticillium</i> species, 28th Fungal Genetics Conference Pacific Grove, CA, USA	20 Mar 2015	1,0
<i>Presentations</i>		
A hybrid-assembly yields a near gapless draft genome, EPS Theme 4 (Genome biology), Wageningen, NL	13 Dec 2013	1,0
The occurrence of chromosomal rearrangements in the fungal genus <i>Verticillium</i> , EPS Theme 4 (Genome biology), Wageningen, NL	3 Dec 2014	1,0
The occurrence of chromosomal rearrangements in the fungal genus <i>Verticillium</i> , COST SUSTAIN workshop, Kiel, Germany	28 Aug 2015	1,0
► <b>IAB interview</b>		
► <b>Excursions</b>		
Company visit In2care and Genetwister	19 Sep 2014	0,2

Subtotal Scientific Exposure

19,4

<b>3) In-Depth Studies</b>	date	cp
<b>► EPS courses or other PhD courses</b>		
PhD course: R for Statistics analysis	22-24 Oct 2014	0,6
PhD spring school Host-Microbe Interactomics	02-04 Jun 2014	0,9
Data analyses and visualizations in R (for biologists)	12-13 Dec 2016	0,6
<b>► Journal club</b>		
Phytopathology journal club	2013-2017	3,0
<b>► Individual research training</b>		
<i>Subtotal In-Depth Studies</i>		5,1

<b>4) Personal Development</b>	date	cp
<b>► Skill training courses</b>		
WGS Course Carousel	17 Apr 2015	0,3
Reviewing a Scientific Paper	23 Mar 2017	0,1
Scientific writing	Oct-Dec 2015	1,8
Scientific artwork with photoshop and illustrator	01-02 Mar 2016	0,6
Pitch training for 99th Dies Natalis celebration of Wageningen UR	23 Feb 2017	0,2
Lunch lecture: Working in Industry after your PhD From the perspective of Human Resources	18 Sep 2017	0,1
Keygene career event, Recruit your Employer	22 Jun 2017	0,3
Wageningen University Career Day 2017	7 Feb 2017	0,3
Open access policy and how to publish open access (Marco van Veller)	8 Jun 2018	0,1
<b>► Organisation of PhD students day, course or conference</b>		
<b>► Membership of Board, Committee or PhD council</b>		
<i>Subtotal Personal Development</i>		3,8

<b>TOTAL NUMBER OF CREDIT POINTS*</b>	<b>31,3</b>
---------------------------------------	-------------

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS with a minimum total of 30 ECTS credits.

\* A credit represents a normative study load of 28 hours of study.



This work was supported by a VICI grant from the Research Council for Earth and Life sciences (ALW) of the Netherlands Organization for Scientific Research (NWO). The work was carried out in the Laboratory of Phytopathology, Wageningen University & Research, the Netherlands. Financial support from Wageningen University & Research for printing this thesis is gratefully acknowledged.

