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

Jasper R. L. Depotter1,2 | Xiaoqian Shi-Kunne1 | Hélène Missonnier3 | Tingli Liu4 | Luigi Faino1 | Grardy C. M. van den Berg1 | Thomas A. Wood2 | Baolong Zhang4 | Alban Jacques3 | Michael F. Seidl1 | Bart P. H. J. Thomma1

1Laboratory of Phytopathology, Wageningen University & Research, Wageningen, The Netherlands
2Department of Crops and Agronomy, National Institute of Agricultural Botany, Cambridge, UK
3Département des Sciences Agronomiques et Agroalimentaires, Equipe Agrophysiologie et Agromolécules, Institut National Polytechnique de Toulouse – Ecole d’Ingénieurs de Purpan, Université de Toulouse, Toulouse, France
4Provincial Key Laboratory of Agrobiology, Jiangsu Academy of Agricultural Sciences, Nanjing, China

Correspondence
Bart P. H. J. Thomma, Laboratory of Phytopathology, Wageningen University & Research, PB Wageningen, The Netherlands.
Email: bart.thomma@wur.nl

Michael F. Seidl, Theoretical Biology & Bioinformatics, Utrecht University, CH Utrecht, The Netherlands.
Email: m.f.seidl@uu.nl

Present address
Jasper R. L. Depoter, Botanical Institute and Cluster of Excellence on Plant Sciences (CEPLAS), University of Cologne, Cologne, Germany
Luigi Faino, Department of Environmental Biology, University La Sapienza, Rome, Italy
Michael F. Seidl, Department of Biology, Theoretical Biology & Bioinformatics, Utrecht University, CH Utrecht, The Netherlands

Funding information
Nederlandse Organisatie voor Wetenschappelijk Onderzoek; Marie Curie Industrial Training Network; European Commission, Grant/Award Number: 607178; French Ministry of Higher Education and Research

Abstract
Plant pathogens continuously evolve to evade host immune responses. During host colonization, many fungal pathogens secrete effectors to perturb such responses, but these in turn may become recognized by host immune receptors. To facilitate the evolution of effector repertoires, 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* displays regions with extensive presence/absence polymorphisms, so-called lineage-specific regions, that are enriched in in planta-induced putative effector genes. As expected, comparative genomics reveals differential degrees of sequence divergence between lineage-specific regions and the core genome. Unanticipated, lineage-specific regions display markedly higher...
sequence conservation in coding as well as noncoding regions than the core genome. We provide evidence that disqualifies horizontal transfer 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 lower nucleotide substitution rates in the plastic, lineage-specific regions of V. dahliae.

**KEYWORDS**
comparative genomics, effector, genome evolution, mutagenesis, two-speed genome, Verticillium wilt

1 | INTRODUCTION

Microbes colonize near all habitats on earth, even those that are characterized by extreme conditions with respect to salinity, temperature, pH, radiation and pressure (Pikuta, Hoover, & Tang, 2007). How microbes can adapt to their environment, especially when it concerns heterogeneous environments, is a fundamental question in evolutionary biology. It remains a major challenge to determine what the molecular consequences are of adaptive mutations in response to dynamic environments, and how these translate into conditionally adaptive phenotypes. During their life cycle, many microbes colonize other organisms that act as their hosts. These symbiotic interactions between microbes and their hosts can range from commensalistic to either beneficial or parasitic. Plant-pathogen interactions are often exquisite models for the study of the molecular processes of (microbial) adaptation to fluctuating (host) environments, due to the short generation time of the microbe, the typically well-characterized interaction in the wild and in the laboratory, and more and more frequently the genetic tractability of the pathogen and the host.

To establish their parasitic relationships, pathogenic microbes evolve repertoires of secreted proteins, so-called effectors, that mediate host colonization often by deregulating host immunity (Cook, Mesarich, & Thomma, 2015; Dodds & Rathjen, 2010). As plants have evolved immune receptors that recognize various molecular patterns that betray microbial invasion as so-called invasion patterns, receptors evolved that can detect effectors or their activities (Cook et al., 2015; Dodds & Rathjen, 2010). Consequently, pathogens and their hosts are typically engaged in co-evolutionary arms races in which plant pathogen effector repertoires are subject to selective forces that often result in rapid diversification. Interestingly, effector genes are often not randomly organized in genomes of filamentous plant pathogens (Dong, Raffaele, & Kamoun, 2015). For instance, effector genes of the oomycete 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 filamentous pathogens have a bipartite genome architecture with housekeeping genes residing in a conserved core genome and effector genes in dynamic and repeat-rich compartments; a “two-speed” genome (Croll & McDonald, 2012; Raffaele & Kamoun, 2012). It is hypothesized that this compartmentalization facilitates the rapid evolution of effector repertoires to mediate continued symbioses between pathogen and the plant host. 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 (Faino et al., 2016; de Jonge et al., 2013). Furthermore, such regions can also display increased substitution rates (Cuomo et al., 2007; van de Wouw et al., 2010), including increased levels of nonsynonymous substitutions (Raffaele et al., 2010; Sperschneider et al., 2015; Stukenbrock et al., 2010).

Verticillium is a genus of soil-borne Ascomycete fungi containing notorious plant pathogens of numerous crops (Inderbitzin & Subbarao, 2014) that infect their hosts via the roots and then colonize xylem vessels, resulting in vascular occlusion and wilt disease (Fradin & Thomma, 2006). Currently, 10 Verticillium species are described, which are divided in two phylogenetic clusters, i.e., clade Flavexudans and clade Flavnonexudans (Inderbitzin et al., 2011). Verticillium species 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, Gurung, Hu, Inderbitzin, & Subbarao, 2014). Nevertheless, Verticillium dahliae still appears to have the machinery for sexual recombination, as mating types and meiosis-specific genes occur (Short et al., 2014). However, a severely skewed mating type ratio in the global V. dahliae population (99% vs. 1%; Short et al., 2014) combined with the low mobility of V. dahliae due to its soil-borne nature makes sexual recombination only a marginal phenomenon at most. In V. dahliae, the most notorious plant pathogen within the genus that infects hundreds of plant species (Inderbitzin & Subbarao, 2014), mechanisms different from meiotic recombination were shown to contribute to the genomic diversity, including large-scale genomic rearrangements, horizontal gene transfer, and transposable element (TE) activity (Faino et al., 2016; de Jonge et al., 2012, 2013; Seidl & Thomma, 2014; Shi-Kunne, van Kooten, Depotter, Thomma, & Seidl, 2019). These mechanisms often converge on lineage-specific (LS) regions that are enriched in TEs and in planta-induced effector genes (Faino et al., 2016; de Jonge et al., 2013; Klosterman et al., 2011). 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, Banerjee, Murugan, Gangapur, & Ravikesavan, 2013). To study the evolution of the LS regions, in the present study 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.

2 | MATERIALS AND METHODS

2.1 | Genome sequencing and assembly

Genomes of Verticillium albo-atrum PD747, Verticillium alfalfae PD683, V. dahliae JR2 and VdLs17, Verticillium isaaci PD618, Verticillium klebahnii PD401, Verticillium nubilum PD621, Verticillium tricorpus PD593 and MUCL9792, Verticillium zaregamsianum PD739 were previously assembled (Faino et al., 2015; Klosterman et al., 2011; Seidl et al., 2015; Shi-Kunne et al., 2018), except for V. isaaci JR2 strain and other Verticillium species were obtained from previous studies (Faino et al., 2016; Shi-Kunne et al., 2018), except for V. isaaci strain PD660 that was annotated with the Maker2 pipeline in this study according to Shi-Kunne et al. (2018) and Holt and Yandell (2011).

2.2 | Comparative genome analysis

Repetitive elements were identified using repeatmodeler (version 1.0.8) based on known repetitive elements and on de novo repeat identification, and genomes were subsequently masked using repeatmasker (version 4.0.6; sensitive mode; Smit, Hubley, & Green, 2015). To prevent assigning high sequence identities to repetitive elements, sequence alignments were performed to a repeat-masked reference genome using nucmer (option --maxmatch), which is part of the mummer package (version 3.1; Kurtz et al., 2004).

Linear plots showing alignments within and closely adjacent JR2 LS regions were plotted with the x package genoPlotR (version 0.8.7; Figure 2; Guy, Kultima, & Andersson, 2011). Alignments >7.5 kb in

FIGURE 1 Locations of lineage-specific (LS) regions in the genome of Verticillium 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-group species. (b) LS regions of V. dahliae strain JR2 were determined by individual comparisons to V. dahliae strains VdLs17 (red), CQ2 (blue) and 85S (green). Sequences of minimum 7.5 kb without alignment to at least one of the other isolates are depicted in colour at their respective position on the V. dahliae strain JR2 genome [Colour figure can be viewed at wileyonlinelibrary.com]
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 corresponding sequences to (a) *Verticillium 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. zaregamsianum* PD739, *V. albo-atrum* PD747, *V. dahliae* JR2, *V. alfalfae* PD683, *V. nonalfalfae* TAB2 and *V. nubilum* PD621 [Colour figure can be viewed at wileyonlinelibrary.com]

length were depicted along the reference genome with the `r` package Rcircos (version 1.2.0; Figure 3; Zhang, Meltzer, & Davis, 2013). Lineage-specific sequences were defined by alignment of different strains to a reference using nucmer (version 3.1, option `--maxmatch`; Kurtz et al., 2004) and regions were determined using bedtools version 2.25.0 (options sort, merge and genomcov; Quinlan & Hall, 2010).

Lineage-specific regions of *V. dahliae* and *V. tricorpus* were delimited based on interspecific nucleotide alignments with nucmer, which is part of the `nucmer` package (version 3.1, option `--maxmatch`; Kurtz et al., 2004). Genome regions were high sequence conservations started or ended in combination with absence/presence polymorphisms were determined as start or end of the LS region, respectively (Table S6). The pairwise identity of the genome-wide and LS regions between *V. dahliae*/*V. tricorpus* and other haploid *Verticillium* species was calculated using nucmer (option `--maxmatch`) by dividing the respective *V. dahliae*/*V. tricorpus* query sequences into nonoverlapping windows of 500 bp and align them to the repeat-masked genomes of *Verticillium* species (Table 1). Only sequences with a 1-to-1 alignments and with a minimum alignment length of 500 were considered (delta-filter -1 -1500). In this fashion, genome-wide sequence identity between *Verticillium* species was also calculated to determine the sequence divergence across the *Verticillium* phylogenetic tree (Figure 4). Instead of using *V. tricorpus* and *V. dahliae* as reference genomes, *V. klebahnii* and *V. alfalfae* were used as reference for clade Flavexudans and clade Flavnonexudans respectively, to calculate the genome-wide nucleotide identity differences with the remaining species within the same clade. Nucleotide identity differences with remaining species were calculated in increasing order of phylogenetic distance with the reference species. Identity differences were evenly assigned to the phylogenetic branches in increasing order of phylogenetic distance.

A 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 four species with increased sequence conservation were combined (Table S6). Repeat masked regions were removed from the pan-LS-genome using `bedtools` version 2.25.0 (Quinlan & Hall, 2010). Additionally, duplicated regions (≥90% identity, ≥100 bp) in the pan-LS-genome were determined using nucmer (version 3.1, option `--maxmatch`; Kurtz et al., 2004) and subsequently removed with using `bedtools` version 2.25.0 (options merge, subtract and getfasta; Quinlan & Hall, 2010). The fraction of pan-LS-genome that is present in every individual *Verticillium* strain was determined using nucmer (version 3.1, option `--maxmatch`; 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.
To compare the rate of synonymous and nonsynonymous substitutions between the core and LS regions, Ka and Ks of orthologs of JR2 and TAB2 were determined using the Nei and Gojobori method (Nei & Gojobori, 1986) in paml (version 4.8; Yang, 2007). Significance of positive selection was tested using a Z test (Stukenbrock & Dutheil, 2012). Z-values > 1.65 were considered significant with p < .05. Secreted proteins were predicted by SignalP4 (Petersen, Brunak, Von Heijne, & Nielsen, 2011). To compare coding regions of genes and intergenic regions, sequence identities were retrieved by blast (version 2.2.31+) searches between strains V. dahliae JR2 and V. nonalfalfae TAB2 (Altschul, Gish, Miller, Myers, & Lipman, 1990). The coding regions of genes were aligned to each other and the best hits (sequence identity) 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 version 2.25.0 (options makewindows and getfasta) and similarly queried to the genome of V. nonalfalfae strain TAB2 (Quinlan & Hall, 2010). Hits with a maximal bit-score and minimal alignment of 500 bp to a window were selected.

2.3 Tree building and ortholog analysis

The phylogenetic trees of the Verticillium genus were 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 JR2, CQ2, VdLs17 and 85S with V. alfalfa as outgroup was constructed using realphy (default settings; Bertels, Silander, Pachkov, Rainey, & Van Nimwegen, 2014).

Orthologous groups of Verticillium species were determined by using orthofinder (version 2.2.7, option -S diamond; Emms & Kelly, 2015), including Sodiomyces alkalinus as outgroup (Grum-Grzhimaylo et al., 2018). Orthologous groups that are shared among at least three species were selected for phylogenetic tree construction. Individual ortholog groups were aligned using mafft (LINSi; version 7.04b; Katoh, Misawa, Kuma, & Miyata, 2002; Katoh & Standley, 2013). A maximum likelihood phylogeny was inferred using raxml (version 8.2.4) with the GAMMA model of rate heterogeneity and the General Time Reversible (GTR) model of nucleotide substitutions (Stamatakis, 2014). The robustness of the inferred phylogeny was assessed by 100 rapid bootstrap approximations.

3 RESULTS

3.1 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...
**TABLE 1**  Sequence identities between *V. dahliae* and *V. tricorpus* with other haploid *Verticillium* species (excluding repetitive regions)

| Species/strain       | Core-genome (%) | LS regions (%) | Number of windows aligned to LS regions<sup>a</sup> | p-Value identity<sup>b</sup> | p-Value variance<sup>c</sup> |
|----------------------|-----------------|----------------|---------------------------------------------------|-----------------------------|------------------------------|
|                      | Md              | VAR            | Md                   | VAR                   |                             |
| *Verticillium dahliae* |                 |                |                      |                      |                             |
| *V. albo-atrum*      | 88.0            | 14.1           | 97.6                 | 14.2                 | 153                         | <2.2e−16                    | .0028                      |
| *V. alfalfae*        | 94.2            | 15.2           | 100.0                | 12.5                 | 454                         | <2.2e−16                    | <2.2e−16                   |
| *V. nonalfalfae*     | 94.4            | 15.5           | 100.0                | 7.3                  | 1,018                       | <2.2e−16                    | 2.65e−05                   |
| *V. nubilum*         | 88.4            | 11.5           | 95.2                 | 30.8                 | 98                          | <2.2e−16                    |                             |
| *V. tricorpus*       | 88.2            | 12.5           | 97.6                 | 13.8                 | 153                         | <2.2e−16                    |                             |
| *V. isaaci*          | 94.4            | 15.5           | 98.6                 | 22.5                 | 170                         | <2.2e−16                    | 2.65e−05                   |
| *V. klebahnii*       | 88.4            | 12.5           | 97.9                 | 23.5                 | 82                          | <2.2e−16                    |                             |
| *V. zaregamsianum*   | 88.8            | 13.5           | 92.5                 | 17.0                 | 31                          | 2.12e−06                    | .51                         |
| *Verticillium tricorpus* |              |                |                      |                      |                             |
| *V. albo-atrum*      | 88.9            | 16.1           | 100                  | 11.3                 | 609                         | <2.2e−16                    | <2.2e−16                   |
| *V. alfalfae*        | 88.2            | 16.0           | 90.9                 | NA                   | 1                           | 0.38                        | NA                         |
| *V. dahliae*         | 88.5            | 14.5           | 97.8                 | 11.1                 | 161                         | <2.2e−16                    | .00021                     |
| *V. nonalfalfae*     | 88.2            | 15.3           | 99.4                 | 30.4                 | 26                          | 3.0e−12                     | .018                       |
| *V. nubilum*         | 88.2            | 14.4           | 95.4                 | 11.9                 | 74                          | <2.2e−16                    | .39                        |
| *V. isaaci*          | 94.6            | 16.2           | 98.7                 | 23.4                 | 146                         | <2.2e−16                    | .39                        |
| *V. klebahnii*       | 94.8            | 16.1           | 98.4                 | 15.3                 | 189                         | <2.2e−16                    | .036                       |
| *V. zaregamsianum*   | 94.8            | 15.7           | 92.4                 | 12.3                 | 47                          | 0.0018                      | .40                        |

Note: The strains used in this table: *V. albo-atrum*, PD747; *V. alfalfae*, PD683; *V. dahliae*, JR2; *V. nonalfalfae*, TAB2; *V. nubilum*, PD621; *V. tricorpus*, PD593; *V. isaaci*, PD660; *V. klebahnii*, PD401; *V. zaregamsianum*, PD739.

Abbreviations: Md, median sequence identity; VAR, sequence variation.

<sup>a</sup>Windows of 500 bp.

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

<sup>c</sup>The p-value was calculated with the Fligner–Killeen test of homogeneity of variances.
et al., 2015, 2016; de Jonge et al., 2013). By definition, LS 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, 2016; de Jonge et al., 2013). Since *V. dahliae* strains JR2 and VdLs17 only recently diverged (de Jonge et al., 2013), we sequenced two additional *V. dahliae* strains that diverged earlier from JR2 (Figure 1a), namely strains CQ2 and 85S that were isolated from cotton in China and sunflower in France, respectively, 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 of the JR2 sequences without alignment (88%, 82%, and 91% for VdLs17, CQ2 and 85S, respectively) localize in the four previously identified LS regions that collectively contain 476 predicted protein-coding genes and 85S, respectively) localize in the four previously identified LS regions.

### 3.2 | LS regions share increased sequence identity to other *Verticillium* species

Next, we extended our analysis to other *Verticillium* species. While most of the *V. dahliae* strain JR2 genome aligns with *V. nonalfafae* strain TAB2 with an average sequence identity of ~92%, particular regions display increased sequence identity, even up to 100% (Figure S1). Intriguingly, the regions with increased sequence identity co-localize with LS regions (Faino et al., 2015, 2016). To perform interspecific comparisons with all other haploid *Verticillium* species, genomic sequences of *V. dahliae* were aligned in windows of 500 bp to the other species, excluding repetitive regions, displaying median identities ranging from 88% to 94%. These identities 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 92.5% median sequence identity for *V. zaregamanium*, one of the phylogenetically most distantly related species, to 100% for the most closely related species *V. alfafae* and *V. nonalfafae* (Figure 2a, Table 1).

To assess whether high interspecific sequence identity concerns only LS regions, we aligned *Verticillium* sequences to the complete *V. dahliae* JR2 genome. For several species we used multiple strains at this stage (Table S2). Nearly all (99%–100%) of the *V. alfafae* and *V. nonalfafae* 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* species 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 LS regions of *V. dahliae*.

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

To investigate whether other *Verticillium* species 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 probably 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 MUC19792 (Seidl et al., 2015). In total, 98% of the PD593 genome could be aligned to MUC19792. However, 48% of the sequences that are specific for *V. tricorpus* strain PD593 reside...
high interspecific sequence identity of particular genomic regions, such as observed here for LS regions, could occur through two different mechanisms: (a) horizontal transfer of sequences between species, in this case horizontal transfer of LS sequences between *Verticillium* species, or (b) differences in nucleotide substitution rates between genomic regions, in this case between core and LS genome regions. To explain the high sequence identity through the occurrence of interspecific horizontal DNA transfers, a minimum of five such transfers must have occurred that involve *V. dahliae* to explain the differences in median sequence identities of LS and core genome regions with the other *Verticillium* species (Figure 4, Table 1). Similarly, at least five such transfers are required to explain the composition of the *V. tricorpus* genome relative to its sister species (Figure 4, Table 1). Consequently, dozens of horizontal DNA transfers among sister species are required to explain the composition of the LS regions throughout the species of the whole *Verticillium* genus.

If LS regions were transferred between species through a limited number of events, a depletion of interspecific sequence identity variation in LS regions may be observed. Thus, we calculated the variance in interspecific sequence identity for *V. dahliae* LS regions and genome-wide (Table 1). Sequence identities in LS regions varied less than in the core genome in alignments to *V. albo-atrum* and to *V. nubilum* and *V. isaacii* displayed more variation in LS genome regions than within the core genome regions (Table 1). No significant differences between LS and core genome regions were found in alignments with *V. tricorpus*, *V. klebahnii* and *V. zaregamsianum*. A similar pattern is observed for *V. tricorpus* as LS regions display lower sequence identity variation than the core genome in alignments to *V. albo-atrum* and *V. dahliae*, whereas the opposite is true for alignments to *V. nonalfalfa* and *V. klebahnii* (Table 1). Thus, although it needs to be noted that there is a significant difference in size between core and LS genomic regions, there seems to be no genus-wide trend towards lower interspecific sequence identity variation in LS regions compared to the core genome, which may be interpreted as an argument against horizontal transfer of these regions.

### 3.4 LS regions are unlikely to originate from horizontal DNA transfers among *Verticillium* species

In only a single genomic region of 41 kb on scaffold 1 (Figure 3b). Like for *V. dahliae* strain JR2, sequences of other *Verticillium* species aligned with high identity to *V. tricorpus* PD593: *V. isaacii*, *V. klebahnii* and *V. zaregamsianum* display a median genome identity of ~95%, while other haploid *Verticillium* species display ~88%-89% median genome identity (Table 1). 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 (Figures 2b and 3b). It is likely this concerns an LS region that could not be identified based on the two *V. tricorpus* strains used in our analysis, and thus will also be referred to as LS region. For *Verticillium* strains with total alignments of at least 100 kb of high-identity sequences, the fraction of high-identity sequences that aligned to LS genome regions ranged from 49% for *V. nubilum* (PD621) up to 84% for *V. albo-atrum* (PD747; Table S4). As expected, the sequence identity to six of the eight other haploid *Verticillium* species was significantly higher in LS genome regions compared to the core genome (Table 1). No increase in sequence identity was found in alignments with *V. albo-atrum* strain PD683 and *V. zaregamsianum* strain PD739 as only few sequences of LS genome regions could be aligned (Table 1).

**FIGURE 5** The evolution of lineage-specific (LS) region genes. The left tree shows the phylogenetic relationships between the haploid *Verticillium* species and their division into clade Flavexudans (FE) and clade Flavnonexudans (FNE). The middle and right phylogenetic trees are of a particular LS region gene from *V. dahliae* and from *V. tricorpus*, respectively. Genes indicated in red have particularly high sequence identity. The robustness of the phylogeny was assessed using 100 bootstrap replicates. For all *V. dahliae* and *V. tricorpus* LS region genes with at least one homolog in clade Flavexudans and one homolog in clade Flavnonexudans, phylogenetic trees were constructed including all their *Verticillium* homologs. The number of trees in congruence and incongruent with the *Verticillium* clade dichotomy are indicated in the green and red box, respectively. In these assessments, complete trees were analysed in case the tree did not contain paralogs for any of the *Verticillium* species. In case of paralogs that indicate gene duplication events, only one subclade of the phylogenetic tree was used for assessment. The number in the white box is the number of trees that was inconclusive [Colour figure can be viewed at wileyonlinelibrary.com]
In case of horizontal transfers, phylogenetic trees of LS genes and their orthologs in other *Verticillium* species are likely to display incongruences with the general *Verticillium* phylogeny. Genes in LS regions often have only few orthologs due to the abundant presence/absence polymorphisms that characterize these regions but, nonetheless, 50 genes could be identified that have at least one ortholog in clade Flavexudans and one in clade Flavnonexudans. To identify (in)congruences with the *Verticillium* phylogenetic tree, we assessed phylogenetic trees for these genes. Due to the generally low resolution that can be attributed to the limited amount of orthologs and their generally high degree of sequence conservation, we focussed our analysis on the dichotomy of homologs from Flavexudans and Flavnonexudans species. For example, Chr5g02240 has homologs in seven other *Verticillium* species, with five highly conserved orthologs in species that generally differ from 5.2% to 11.2% in genome-wide nucleotide identity with *V. dahliae*, whereas sequence identities between the highly conserved Chr5g02240 orthologs differs only from 0% to 0.8% (2,460 total nucleotides; Figure 5). Moreover, in accordance with the overall *Verticillium* phylogeny, the *V. nonalfafae* homolog clusters with that of *V. dahliae*. In total, 80% of the trees displayed a phylogeny where the Flavexudans and Flavnonexudans species clustered, whereas only 5% of the trees did not yield this typical dichotomy and 15% of the trees did not contain sufficient homologous sequences to be conclusive (Figure 5). Similarly, the majority (56%) of phylogenetic trees for *V. tricorpus* LS genes obey to the Flavexudans and Flavnonexudans dichotomy (Figure 5). Like for *V. dahliae* gene Chr5g02240, the phylogenetic tree for homologs of *V. tricorpus* gene Chr006g10380 shows the expected segregation.
of Flavnonexudans homologs from the Flavexudans homologs. Thus, the majority of LS genes can be demonstrated to diverge in accordance with the *Verticillium* clade dichotomy.

Taken together, neither the comparison of interspecific sequence identity variation within and outside LS regions, nor the analysis of phylogenetic trees of LS genes, seems to support a scenario where horizontal DNA transfer events was fundamental to the evolution of LS regions. Moreover, as dozens of such events would be required among the various *Verticillium* species, many of which over a short evolutionary time scale, and considering the low-mobile, soil-borne nature of the various species, we deem it very unlikely that interspecific horizontal DNA transfers shaped LS regions.

### 3.5 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, and that lower numbers of DNA substitutions seem to occur in the LS regions. To determine the distribution of conserved LS sequences across the *Verticillium* genus, we constructed a pan-LS-genome by combining regions with high interspecific sequence identity of the four *Verticillium* species *V. dahliae* strain JR2, *V. alfalfa* strain PD683, *V. tricorum* strain PD593, and *V. klebahnii* strain PD401 that were selected because of their high assembly contiguity and spread over the *Verticillium* genus (Inderbitzin et al., 2011; 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 (Figure 6). The proportion of the pan-LS-genome differed markedly between *Verticillium* strains and ranged from 12% for *V. nubilum* strain PD621 up to 58% for *V. dahliae* strain JR2 (Figure 6, Table S5). Notably, by using a limited number of isolates in the consensus reconstruction, retentions are probably 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. zaregaminsanum* and *V. isaacii* strains, despite its phylogenetically more distant relation to *V. klebahnii* and *V. tricorum* (Figure 6, Table S5). Moreover, LS contents do not only differ considerably between species but also within 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 (Figures 1a and 4; Faino et al., 2015). Thus, sequences with high interspecific identity are associated with dynamic genomic regions of *Verticillium* species throughout the genus as their contents vary greatly between and within species.
3.6 Increased sequence conservation is probably not driven by negative selection

To study the impact of the increased sequence conservation on gene evolution in more detail, substitution rates of LS region genes were compared with those of core region genes. We determined the rates of nonsynonymous (Ka) and synonymous (Ks) substitutions for LS versus core genes. In total, 48% (68 out of 142) of the LS genes could not be used for Ka and Ks 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 Ka is not different (two-sided Wilcoxon rank-sum test, \( p < .05 \)) between genes in LS regions (median = 0.015, \( n = 74 \)) and the core genome (median = 0.015, \( n = 8,583 \)), the Ks of genes in LS regions genes (median = 0.12, \( n = 74 \)) is significantly lower than of core genes (median = 0.16, \( n = 8,583 \); Figure 7a). Consequently, LS genes (median = 0.38, \( n = 60 \)) display significantly higher Ka/Ks values than core genes (median = 0.09, \( n = 8,289 \)), calculated for genes that have both synonymous and nonsynonymous 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, \( p < .05 \)). 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 (\( p < .05 \)). Thus, a similar fraction of nucleotide alterations originates from nonsynonymous substitutions in LS and core genome regions, whereas synonymous substitutions occur less frequent in LS regions when compared with the core genome. The resulting increase in Ka/Ks ratio’s for genes in LS regions suggests that negative selection is probably not responsible for the slower divergence of gene sequences. Rather, LS regions seem to encounter a lower rate of synonymous nucleotide substitutions that lead to lower numbers of neutral/synonymous substitutions in genes in LS regions when compared with core genes. In congruence with this hypothesis, sequence comparisons of coding and intergenic regions between V. dahliae and V. nonalfalfae revealed that increased sequence conservation similarly occurs in coding regions as well as intergenic sequences (Figure 7b), suggesting that the increased sequence conservation is driven by a mechanism that affects whole LS regions, rather than by selection.

4 DISCUSSION

Genomes of many filamentous plant pathogens are thought to obey to a two-speed evolution model (Croll & McDonald, 2012; Dong et al., 2015; Möller & Stukenbrock, 2017). Verticillium dahliae is similarly thought to evolve under a two-speed regime, as LS regions that are of significance for host interactions display increased structural variation and TE activity (Faino et al., 2015, 2016; de Jonge et al., 2013). Additionally, LS regions are extremely plastic with abundant presence/absence polymorphisms (Figure 1; Faino et al., 2016; de Jonge et al., 2013). Intriguingly, although LS regions are enriched in segmental duplications (Faino et al., 2016), which can be an incentive for evolutionary diversification (Magadum et al., 2013), LS sequences display a remarkably high degree of sequence identity to other Verticillium species (Figures 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 & 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., 2006). However, horizontal transfer is not likely to explain our observations, as the increased sequence identity is observed genus-wide and concerns every species within the genus (Table 1). If the increased sequence conservation would be caused by horizontal transfers, V. dahliae and V. tricorpus each should have been involved in at least five such transfers during their evolution to explain the difference in median sequence identity between LS and core genome regions (Figure 4, Table 1). Thus, by only considering these two species at least nine interspecific DNA transfers must have occurred during the evolution of the Verticillium genus. Phylogenetic trees of LS genes suggest that the majority of genes evolved according the Verticillium clade Flavononexudans and Flaveuxudans dichotomy, indicating that most genes do not originate from horizontal transfer between these two phylogenetic clades. Furthermore, individual cases of LS genes for which (near-)identical orthologs can be found in up to six of the nine haploid Verticillium species (Figure 5) further argue against horizontal transfer, especially when considering the low-mobile, soil-borne nature of the various species.

Considering that horizontal transfer is extremely unlikely, we argue that V. dahliae LS regions are subject to processes that mediate increased sequence conservation. This increased sequence conservation of LS regions would not be a consequence of negative selection on coding regions, as intergenic regions display similarly increased conservation levels (Figure 7b). Moreover, LS genes display a similar fraction of nonsynonymous substitutions (Ka) as core genes, but carry significantly less synonymous substitutions (Ks; Figure 7a). Consequently, high interspecific sequence identity of LS genome regions is probably due to generally lower synonymous nucleotide substitution rates in these regions when compared with the core genome.

Repressed 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). As sex is
probably rare to nonexistent in *Verticillium*, selective sweeps are an unlikely explanation for increased sequence conservation of LS region as, in case of *Verticillium* species with low to nonexistent meiotic recombination, entire genomes would sweep to fixation and not specific genome regions (Shapiro, Leducq, & Mallet, 2016). Moreover, sequence conservation is observed between species and fixation of the same genome region across species boundaries is unlikely, as different *Verticillium* species conceivably encounter distinct selection pressures due to differences in their biological features, such as pathogenicity and host ranges (Inderbitzin & Subbarao, 2014). 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. alfalfaef* 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. alfalfaef* and *V. nonalfalfaef* 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 (Faino et al., 2016; de Jonge et al., 2012).

Mechanisms that can explain the observed increased sequence conservation in repeat-rich LS regions remain unknown. Mutations mostly originate from DNA polymerase errors and there is no immediate reason why these error rates would diverge in LS regions. Possibly, the depletion of nucleotide substitutions can be associated with a differential chromatin 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 & 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, Cook, & Thomma, 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*.

5 | 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 nucleotide substitutions (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.

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).

AUTHOR CONTRIBUTIONS

X.S.K., H.M., T.L., L.F., G.C.M.B., B.Z., and A.J. contributed to the data generation of this study. J.R.L.D. performed the comparative genic analysis and wrote the manuscript with the help of X.S.K., M.F.S., and B.P.H.J.T., J.R.L.D., X.S.K., M.F.S., and B.P.H.J.T. contributed to the study design. T.W., M.F.S., and B.P.H.J.T. supervised this study.

DATA AVAILABILITY STATEMENT

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.

ORCID

Bart P. H. J. Thomma https://orcid.org/0000-0003-4125-4181
