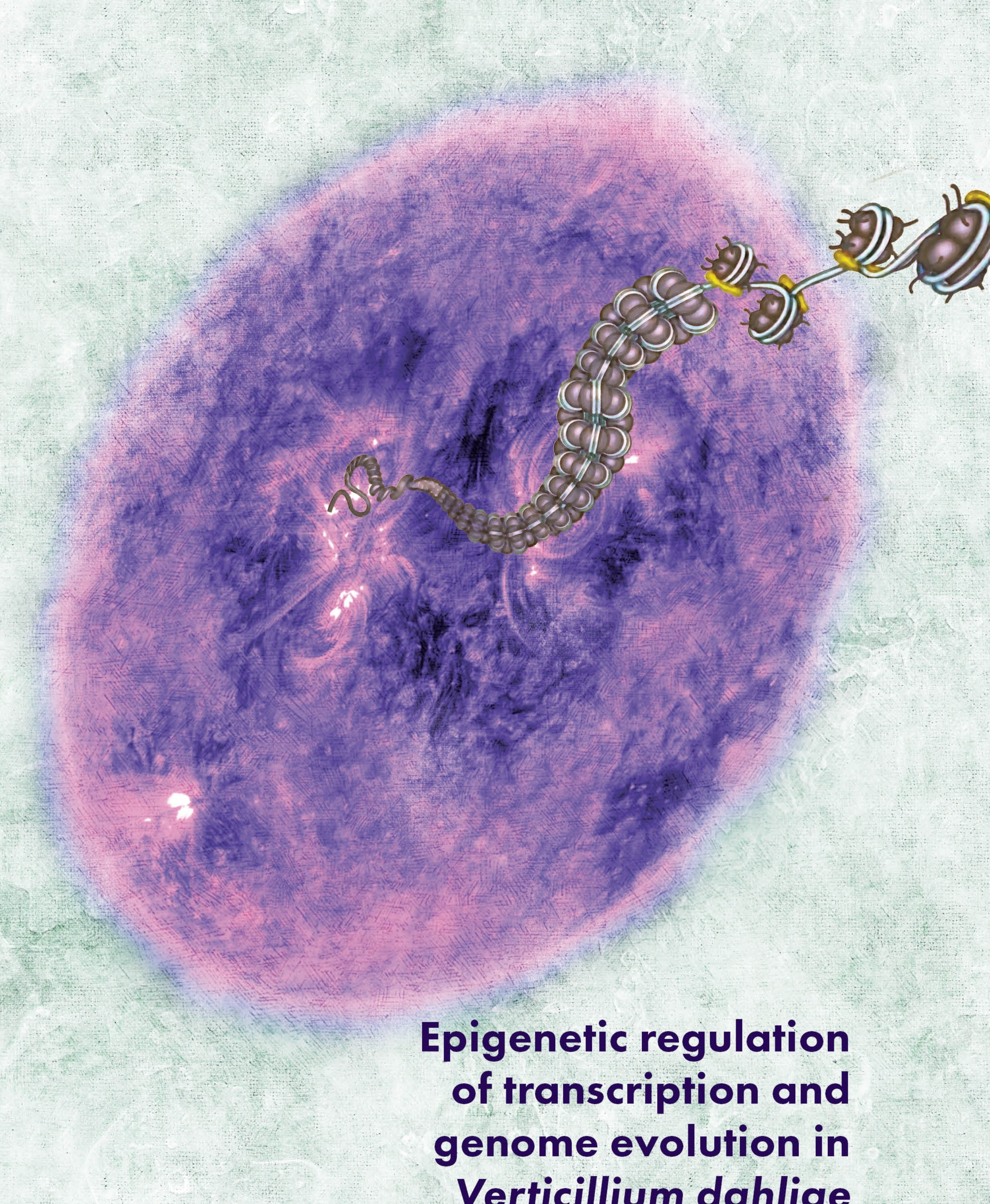


Epigenetic regulation of transcription and genome evolution in *Verticillium dahliae*

H.M. Kramer

Propositions

1. DNA methylation in *Verticillium dahliae* does not regulate transcription of transposons, but rather prevents future damage by promoting their mutation.
(this thesis)

2. *In planta* induction of *Verticillium dahliae* effector genes does not require chromatin de-condensation.
(this thesis)

3. The manipulation of microbiomes to improve plant yield in sub-optimal soils is too unpredictable to be a useful solution.

4. The highest hurdle for using bioinformatics is in the biologist's brain.

5. In the ever-globalizing world, the use of minority languages such as Dutch or German promotes xenophobia.

6. Social media platforms have to employ scientists to actively remove non-scientific "alternative facts"

Propositions belonging to the thesis, entitled

Epigenetic regulation of transcription and genome evolution
in *Verticillium dahliae*

Hisse Marten Kramer
Wageningen, 11 March 2022

Epigenetic regulation of transcription and genome evolution in *Verticillium dahliae*

Hisse Marten Kramer

Thesis committee

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

Co-promotors
Dr D.E. Cook
Assistant Professor, Department of Plant Pathology
Kansas State University, Manhattan, Kansas, USA

Dr M.F. Seidl
Assistant Professor, Theoretical Biology & Bioinformatics, Department of Biology
Utrecht University

Other members
Prof. Dr B.J. Zwaan, Wageningen University & Research
Dr A.D. van Diepeningen, Wageningen University & Research
Prof. Dr C.M.J. Pieterse, Utrecht University
Dr I. Fudal, INRAE, Thiverval-Grignon, France

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

Epigenetic regulation of transcription and genome evolution in *Verticillium dahliae*

Hisse Marten Kramer

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 11 March 2022
at 4 p.m. in the Aula.

Hisse Marten Kramer
Epigenetic regulation of transcription and genome evolution in *Verticillium dahliae*, 238 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2022)
With references, with summary in English.

ISBN: 978-94-6447-040-6
DOI: https://doi.org/10.18174/558432

Table of contents

Chapter 1

General introduction

Plant microbe interactions

During their whole life cycle, plants are exposed to numerous micro-organisms, of which many have commensalistic, some have mutualistic, and others have pathogenic interactions with plants. However, it is important to realize that plant pathogenic micro-organisms generally do not cause disease on all plants, as many pathogens have limited host ranges and only infect a single to a few hosts. Relatively few pathogens have larger host ranges that comprise dozens to, in rare cases, hundreds of host species. However, even in cases where a pathogen infects a large range of host plants, pathogenicity of a given strain or isolate of the pathogen may be limited to a single or a few host species only. For instance, whereas a very wide diversity of plant species, including most agricultural crops, is susceptible to bacterial speck disease caused by *Pseudomonas syringae*, individual strains of this pathogen typically infect only few hosts [1]. Similarly, individual strains of the pathogenic fungus *Fusarium oxysporum* infect only a single or few host species, while the fungal species collectively infects hundreds of hosts [2]. Therefore, micro-organisms that are pathogenic on a particular host can be commensalistic, endophytic, and sometimes even mutualistic on another host [3,4]. Besides depending on inherent traits of the pathogen, this host-specificity of plant pathogenic microbes is greatly dependent on the complex immune systems that have evolved in plants, as well as on environmental factors [5].

Plants prevent infection by potentially pathogenic micro-organisms through passive and active mechanisms. Firstly, the physiology of plants can deter micro-organisms from entering the plant, exemplified by the cuticular wax-layer on leaves that may be difficult to penetrate [6,7]. Additionally, during growth and development plants produce various types of antimicrobial metabolites, collectively named phytoanticipins, that are released upon damage of plant tissue to prevent microbial infection [8]. Active defense mechanisms start with the recognition of the presence of potential pathogens, or their activity, through intra- and extracellular immune receptors that bind non-self or modified-self ligands, collectively named invasion patterns (IPs) [9]. Such IPs can be structural components of pathogen cells, also known as microbe-associated molecular patterns (MAMPs), signatures of pathogen-induced plant damage, termed damage-associated molecular patterns (DAMPs), or proteins or metabolites produced by the pathogen during host invasion [10]. Invasion pattern receptors (IPRs) that get activated subsequently induce downstream responses leading to the deployment of inducible defense mechanisms aimed at the restriction of pathogen invasion. This response seems to contribute to the observation that plants are generally healthy in nature despite being surrounded by a wealth of potentially pathogenic microbes [9]. However, in particular circumstances plants still suffer from microbial disease. This can be explained by the notion that, in order to be able to infect plants, successful pathogens evolved to secrete effector molecules that support host colonization and typically aid infection by blocking IP-recognition, by interfering with defense mechanisms, by tinkering with plant physiology or by killing plant cells [10,11]. In turn, plants have evolved novel IPRs that are able to recognize such effector molecules or their activities, to re-install defense responses and again restrict pathogen dissemination, which makes that the effector that becomes recognized basically acts as IPs on the plant genotypes that evolved the appropriate IPRs [9]. This, in turn, forces pathogens to lose or mutate the recognized effector, or to evolve novel effectors to again perturb the re-installed immune response. This continuous co-evolutionary arms-race between plants and their pathogens, where single gene loss, gain or mutation can alter the outcome of the interaction from compatible to incompatible or vice versa, was first captured in the gene-for-gene hypothesis [12], subsequently refined in the zigzag model that acknowledged the immune-suppressive characteristics of pathogen effectors [10] and further tweaked in the invasion model that is built on the notion that there is no clear distinction between MAMPs and effectors regarding their immunogenic properties [9]. Rather, the invasion model proposes that any kind of ligand will act as an invasion pattern as long as it accurately betrays microbial ingress and elicits an appropriate immune response. This view is supported by recent findings that illustrate that MAMP-triggered immunity and effector-triggered immunity quickly converge and thus that a functional distinction as originally proposed is not appropriate [13–15].

Effector molecules

Based on early research on, amongst others, the extracellularly-growing apoplast-colonizing biotrophic fungus *Cladosporium fulvum*, fungal effectors were originally defined as small, cysteine-rich proteins that are secreted into the apoplast during host colonization [16]. *C. fulvum* effectors for which the bioactive function is determined are involved in preventing chitin perception by the plant and in self-defense against plant defensive enzymes [17]. However, the initial definition for effector proteins did not appear to hold true for effectors in general over time. For instance, effector proteins that are delivered into the host-cytoplasm after haustorial secretion, or through other means, are generally larger and contain fewer cysteines than effectors that remain in the apoplast [18]. Cytoplasmic effectors can inhibit defense-associated signal transduction pathways, secretion mechanisms and other cytoplasmic processes, or transit through the cytoplasm to reach particular organelles. Effector proteins that localize in the plant nucleus may directly manipulate transcription, either positively or negatively, or target transcriptional regulators to change their function [19,20]. A well-known example of such nuclear-targeted effectors is formed by the Transcription Activator-Like (TAL) effectors of *Xanthomonas* bacteria that activate the expression of plant genes that aid bacterial infection [21]. Recently, effectors in various pathogens were discovered that target chromatin-related processes to epigenetically regulate plant gene expression in their favor [22,23]. Effectors can also perform less subtle manipulations, as particularly necrotrophic pathogens, but also hemi-biotrophs and even some biotrophs, utilize cell wall-degrading enzymes, proteinaceous toxins and secondary metabolites to weaken or kill plant cells, or to induce autophagy [24]. Additionally, it has become increasingly evident that pathogens can transfer small RNAs into host cells that hijack the RNA interference machinery to suppress host immunity, and thus function as effector molecules [25,26]. Therefore, it is more appropriate that a definition of effectors encompasses any type of pathogen molecule that is secreted during infection and promotes the microbial colonization process on the host.

The wealth of effector molecules that is encoded in pathogen genomes is generally not ubiquitously released upon contact with plant cells, as often specific suites of effectors are produced and released in waves during different stages of infection [27]. For instance, while the fungal pathogen *Leptosphaeria maculans* colonizes cotyledons of oilseed rape and, after a long latent phase, spreads throughout the mature plant to cause disease symptoms, it utilizes differential suites of effectors during these different infection stages [28]. Similarly, effector proteins of the maize smut fungus *Ustilago maydis* were found to be enriched in three distinct expression modules corresponding to infection stages on the plant surface, establishment of biotrophy, and induction of tumors [29]. Conceivably, any pathogen similarly undergoes distinct infection stages, albeit not always as clearly recognizable as for *L. maculans* and *U. maydis*, that are accompanied by differential waves of expressed effector genes.

Regulation of effector gene expression

The expression of effectors in pathogen genomes requires careful regulation, as effector expression at the wrong time or place can lead to failed infection. For instance, many haustoria-forming pathogens use cell wall-degrading enzymes to specifically weaken cell walls at sites of attempted invasion (e.g., Xu & Mendgen, (1997) [30]). It is conceivable that if the pathogen produces too little of these enzymes, or not at the correct sites, successful entry of the cell may fail. In contrast, if the pathogen produces too much of the cell wall-degrading enzymes, the plant cells may be too weak to sustain the haustoria or the plant may induce defense mechanisms that arrest the fungus.

Few transcriptional regulators have been identified that are involved in effector gene expression. For instance, the transcriptional regulator Sge1 of *Fusarium oxysporum* f. sp. *lycopersici* is required for expression of particular proteinaceous effector genes during infection, and deletion of *Sge1* compromises pathogenicity on tomato [31]. In the fungal plant pathogens *F. graminearum*, *F. verticilloides* and *B. cinerea*, homologs of Sge1 control expression of genes involved in the production of secondary metabolites [32–34]. In the fungal plant pathogen *Verticillium dahliae,* an Sge1 homolog positively regulates expression of particular effector genes, while negatively regulating the expression of others [35]. However, mutants in the *Sge1* homologs of the species other than *F. oxysporum* display aberrant development, suggesting that these Sge1 homologs do not only regulate the expression of effector genes but also genes involved in developmental processes. The transcription factor Pf2 of the necrotrophic fungus *Alternaria brassicicola* is not required for spore germination and appressorium formation, but is required for pathogenicity, suggesting a potential role in effector gene expression [36]. Indeed, expression of *Pf2* during early infection stages coincides with expression of 106 fungal genes, including eight putative effector genes, that are not expressed in *Pf2* deletion mutants [36]. Pf2 homologs of the necrotrophic fungi *Parastagonospora nodorum* and *Pyrenophora tritici-repentis* were found to be similarly required for virulence by regulating the expression of proteinaceous toxins, cell wall-degrading enzymes and other effectors [37,38]. The examples of Sge1 and Pf2 suggest that different pathogens may utilize similar transcriptional regulators for effector gene expression. However, arguably these transcriptional regulators do not regulate effector gene expression on their own, as different waves of effectors likely have their own sets of transcriptional regulators.

Epigenetic regulation of gene expression

Even though transcriptional reprogramming of eukaryotes is mediated through transcriptional regulators that recruit RNA polymerase II to promoters of genes, the binding of these transcriptional regulators is influenced by epigenetic mechanisms that affect DNA characteristics without changing the genetic sequence. Thus far, relatively little research has been performed to elucidate the impact of epigenetic mechanisms on the regulation of effector gene expression in pathogenic microbes. Yet, epigenetic mechanisms are generally conserved between eukaryotes and, therefore, epigenetic studies on divergent eukaryotes provides information on how transcription may be regulated in plant pathogens as well.

Eukaryotic genomic DNA is highly structured in the nucleus through the formation of DNA-protein complexes called nucleosomes. These nucleosomes range along the chromosome, bind approximately 147 bp of DNA and are globular-shaped protein complexes consisting of two copies of histone 2a (H2A), H2B, H3 and H4, with unstructured tails sticking out from the protein complex (Fig. 1A) [39,40]. Nucleosomes can interact with each other as well as with other proteins to generate higher order structures with particular characteristics. Mainly, genomic regions that display relatively weak nucleosome interactions are known as euchromatin and are often transcriptionally active, while genome regions that display stronger nucleosome interactions are known as heterochromatin and are transcriptionally silent (Fig. 1B). The differences in chromatin compaction are largely determined by chemical modifications to amino acid residues in the unstructured histone tails. For instance, tri-methylation of lysine 4 on histone 3 (H3K4me3) is generally associated with euchromatin, whereas H3K9me3 and H3K27me3 are associated with different types of heterochromatin. H3K9me3-labeled constitutive heterochromatin remains condensed throughout the cell cycle, whereas H3K27me3-labeled facultative heterochromatin can de-condense in response to environmental stimuli.

Chromatin compactness can directly influence the transcriptional output of genomic regions by permitting or inhibiting access to the transcriptional machinery. Additionally, chromatin can affect transcription, and other nuclear processes, by dynamic presence of numerous histone modifications at specific genic regions. For instance, tri-methylation of lysine 36 on histone 3 (H3K36me3) is actively deposited at transcribed genes to recruit histone deacetylase complexes that prevent excessive transcription [41]. Another modification involved in transcriptional regulation is phosphorylation of serine 28 on histone 3 (H3S28ph). H3S28ph is specifically deposited at the promotors of stress-responsive genes, and enforces a switch from trimethylated to acetylated state of the adjacent K27 residue to allow transcription [42]. Additionally, transcription can also be affected by nucleobase modifications, most notably cytosine methylation (5-methylcytosine, 5mC). Presence of 5mC in genic promotor regions can block binding of transcriptional regulators to their binding sites [43].

Besides local DNA compaction, chromatin also determines positioning of chromosomes, or chromosome regions, throughout the nucleus. For instance, a group of H3K9me3-interacting proteins tethers the genome to the nuclear periphery into lamina-associated domains (LADs) [44]. A different set of chromatin-interacting proteins similarly forms LADs between euchromatin domains, yet these are tethered throughout the nucleus [45]. Formation of these different

LADs results in a physical separation of genome regions with heterochromatin locating at the periphery, while active euchromatin resides more centrally in the nucleus. Clustering of chromatin domains also occurs independent of the lamina into clusters that are known as topologically associating domains (TADs) (Fig. 1C,D) [46]. TADs have been described in various organisms and are often broadly delineated as alternating heterochromatic and euchromatic regions that display strong intra-TAD interactions and may interact with TADs that display a similar chromatin type [47]. As such, TADs may also be important drivers of genome evolution by inducing genome reorganization within and between chromatin domains.

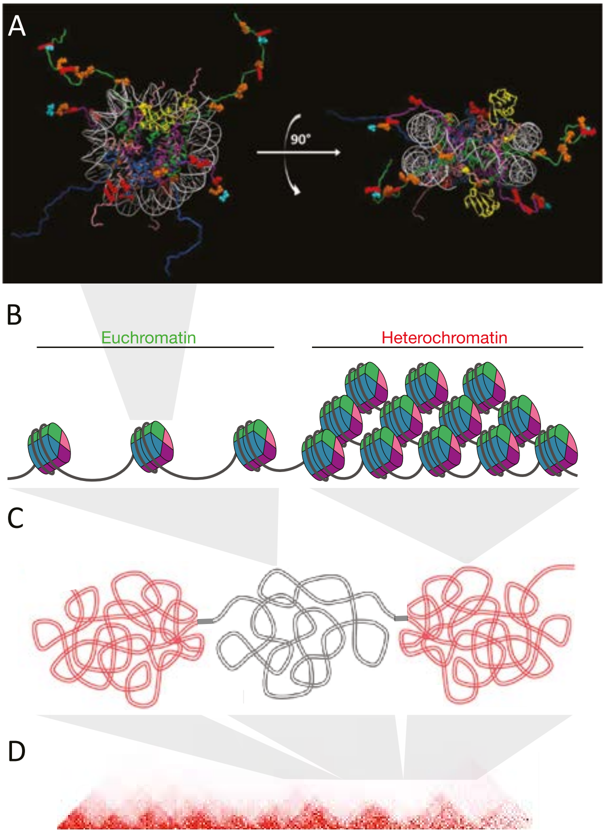


FIGURE 1 | Epigenetic organization of DNA in the nucleus. A) Top and side view of a crystal structure of DNA (grey helix) wrapped around an octamer of histone protein (two monomers of each histone 2a (H2A), H2B, H3 and H4, indicated by magenta, pink, blue and green; a nucleosome-interacting protein is indicated in yellow) that forms a nucleosome with unstructured histone tails protruding from the complex (from Luger et al., (1997) [40]). B) nucleosome-nucleosome interactions structure the DNA strand in open euchromatin and closed heterochromatin regions. C) Broad linear chromatin regions form topologically associated domains (TADs, represented as grey (euchromatin) and red (heterochromatin) clusters) that interact stronger within than between each other and separate chromosome regions in the nucleus. D) Visualization of local 3D genome interaction data on a partial human chromosome, individual TADs are visible as darker triangles.

A hypothesis on epigenetic regulation of effector gene expression

Whether epigenetics is involved in regulation of gene expression during infection was initially studied in the opportunistic human fungal pathogen *Aspergillus nidulans* and in the fungal wheat pathogen *Fusarium graminearum* [48–50]. Infection-associated genes were found to be enriched for heterochromatin-associated H3K9me3 and H3K27me3, respectively, and mutants lacking the enzymes depositing these histone modifications were found to display induced expression of infection-associated genes [49,50]. Similar results were subsequently obtained for the fungal grass endophyte *Epichloë festucae*, the fungal oilseed rape pathogen *L. maculans*, the fungal rice pathogen *Fusarium fujikuroi* and the fungal wheat pathogen *Zymoseptoria tritici* [51–54]. These findings have led to the hypothesis that genomic regions containing *in planta*-induced effector genes are in a heterochromatic state when the pathogens are not in contact with plants and, accordingly, effector genes are not expressed. Thus, in order to infect, pathogens require chromatin de-condensation at these effector gene-containing regions to mediate effector gene expression, which requires alteration of the histone modifications by depletion of H3K9me3 and H3K27me3. This hypothesis has been further reinforced by studies in *E. festucae* and *Z. tritici*, for which it was found that *in planta* induction of two secondary metabolite biosynthesis gene clusters and three effector genes, respectively, negatively correlates with the presence of H3K9me3 and H3K27me3 *in planta* [51,55].

Verticillium dahliae

V. dahliae is an ascomycete fungal pathogen that causes Verticillium wilt disease on hundreds of host plants, among which are food crops, such as potato, tomato and lettuce, as well as fiber crops, such as cotton and flax [56]. Due to yield losses in its broad range of host plants, disease caused by *V. dahliae* have been estimated to result in billions of euros of economic losses annually [57].

The disease cycle of *V. dahliae* starts in the soil, where the fungus resides in the form of microsclerotia, resting structures that germinate upon detection of nearby plant roots [58]. Hyphae from the germinating microsclerotia grow toward the root tip, sites of lateral root formation or sites of root damage, where they penetrate and cross the root epidermis to enter the xylem vessels. In these vessels, *V. dahliae* spreads through the plant by mycelial growth and by production of conidiospores, which traverse with the xylem sap and can form mycelial colonies at sites where they get trapped. During disease progression, the sap flow in the xylem becomes obstructed, leading to wilting, chlorosis and necrosis of plant tissues (Fig. 2). Upon tissue senescence, the fungus grows out of the xylem into the surrounding tissues and produces copious amounts of microsclerotia that are released in the soil upon decomposition of the plant material [58].

During infection of its hosts, *V. dahliae* produces a suite of divergent proteinaceous effectors to modulate host physiology [59]. These effectors can become recognized by plant immune receptors over the course of evolution, potentially leading to resistance. For instance, the tomato Ve1 receptor recognizes the effector Ave1 (for Avirulent on *Ve1* tomato) to cause resistance against *V. dahliae*. Over time, particular strains have overcome the resistance by losing the genomic region containing the *Ave1* gene [59,61]. These strains are assigned to race 2 while the strains that remain to be recognized are assigned to race 1. Interestingly, transcriptomic and genomic analyses revealed that *in planta* induced effector genes, including *Ave1*, are enriched in transposon-rich genomic regions that were determined to be lineage-specific (LS) within the population [59,62]. These LS regions are unique, or shared by a subset of strains only, and have evolved in the population through large-scale chromosomal rearrangements and segmental duplications, followed by reciprocal gene losses. The evolutionary forces working on LS regions have led to a high degree of presence/absence polymorphisms and increased sequence conservation within these regions between *V. dahliae* strains, and therefore different strains contain highly divergent effector repertoires that, when the genes are shared, display only relatively few SNPs. Remarkably, this increased sequence conservation is even observed for LS regions throughout the *Verticillium* genus [63]. Importantly, how *V. dahliae* effector gene expression is regulated *in planta* currently remains unknown.

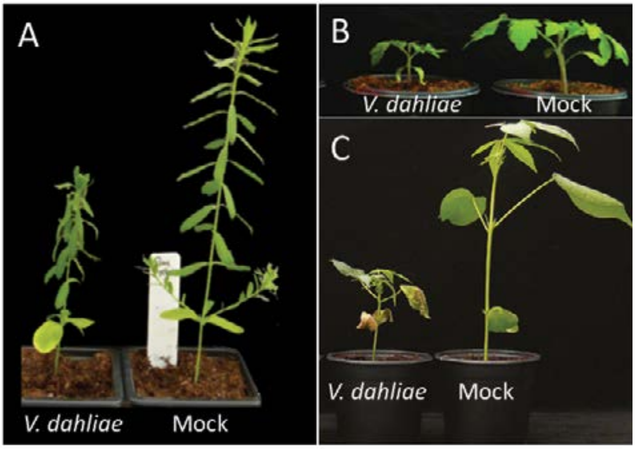


FIGURE 2 | Infection symptoms of *V. dahliae*. Symptoms of *V. dahliae* infection on (A) flax at 12 dpi, (B) tomato at 8 dpi (picture from de Jonge *et al*., (2013) [59]) and (C) cotton at 28 dpi (picture from Song *et al*., (2018) [60]), with a diseased plant on the left and a mock-infected plant on the right. Visible symptoms include growth retardation, leaf wilting and leaf yellowing.

Research objective

The observed genome structure in *V. dahliae*, where transposon-rich LS regions show a high degree of presence/absence polymorphisms between strains, while the sequence conservation is relatively high, indicates that LS regions may have a different chromatin structure than the core genome, which may also influence the expression of effector genes. According to the aforementioned hypothesis on epigenetic regulation of effector gene expression, the *V. dahliae* LS regions can be predicted to be in heterochromatic conformation before attempted host penetration, and decondense upon penetration to allow LS genes to be expressed. The main objective of my doctoral research was to investigate this hypothesis, and to study further implications of the epigenome on genome organization and gene expression.

Thesis outline

Chapter 2 presents an overview of the epigenome structure of *V. dahliae* by studying genetic features, and combining these with genome-wide DNA methylation, histone methylation and chromatin accessibility data. Furthermore, we use a machine learning approach to distinguish the evolutionary distinct LS regions from core genomic regions based on epigenome differences, and subsequently identify novel, previously unassigned, LS regions that are potentially important for fungal adaptation, including adaptation to plant hosts.

In **chapter 3**, we study the prevailing hypothesis on epigenetic effector gene expression, stating that *in planta* induced effector gene-containing LS regions are in heterochromatic conformation when the fungus is not in contact with a plant host, and decondense to a eukaryotic conformation upon infection. This is achieved by correlating differential expression of LS genes of *V. dahliae* cultured in various growth media, with differential presence of histone modifications and chromatin accessibility data.

Chapter 4 focuses on the DNA-methylation machinery of *V. dahliae*. By generating deletion mutants for predicted DNA-methyltransferases and other DNA methylation-associated proteins, we determine which proteins are important for DNA methylation and assess their impact on gene expression and transposon activity.

In **chapter 5**, we demonstrate that the genome of *V. dahliae* contains a specific transposon that only occurs in its centromeres. Using proximity ligation, we identify repeat-rich centromeres in the other species of the *Verticillium* genus as well. Intriguingly, however, these other species do not contain the transposon that populates the *V. dahliae* centromeres in their centromeres. To study centromere evolution in the *Verticillium* genus, we analyze the centromeres of all species and compare them in light of the evolution of the genus.

In **chapter 6**, we further use proximity ligation to investigate whether the local 3D-structure of chromatin can help to explain differences in transcription and genome evolution between LS regions and the core genome. Furthermore, we compare local chromatin structure of all *Verticillium* species to investigate conservation of the 3D chromatin organization within the genus.

In **chapter 7,** I discuss the implications of my findings presented in this thesis in a broader context of gene regulation, genome evolution in a plant pathogen and the impact on plant-microbe interactions.

Chapter 2

A unique chromatin profile defines adaptive genomic regions in a fungal plant pathogen

David E. Cook[1,2],
H. Martin Kramer[2],
David E. Torres[2,3],
Michael F. Seidl[2,3],
Bart P.H.J. Thomma[2,4]

[1]Department of Plant Pathology, Kansas State University, Manhattan KS, USA
[2]Laboratory of Phytopathology, Wageningen University & Research, Wageningen, The Netherlands
[3]Theoretical Biology & Bioinformatics Group, Department of Biology,
Utrecht University, Utrecht, The Netherlands
[4]University of Cologne, Institute for Plant Sciences,
Cluster of Excellence on Plant Sciences (CEPLAS), Cologne, Germany

Published in eLife 9: e62208 (2020)

Abstract

Genomes store information at scales beyond the linear nucleotide sequence, which impacts genome function at the level of an individual, but the influence on population and longer-term genome function remains unclear. Here, we addressed how physical and chemical DNA characteristics may influence genome evolution in the plant pathogenic fungus Verticillium dahliae. We identified incomplete DNA methylation of repetitive elements in the genome, associated with specific compartments of the genome defined as Lineage-Specific (LS) regions that contain genes involved in host adaptation. Further chromatin characterization shows that LS regions are associated with features such as H3 Lys-27 methylated histones (H3K27me3) and accessible DNA. Machine learning trained on chromatin data identified approximately twice as much LS DNA than previously recognized, which was validated through orthogonal analysis. Our results provide evidence that specific chromatin profiles define LS regions, and highlights how different epigenetic factors contribute to the organization of adaptive genomic regions.

Impact Statement

Assessment of DNA methylation, histone modifications and DNA accessibility revealed that physical DNA characteristics are associated with adaptive genome evolution in the broad host range plant pathogenic fungus *Verticillium dahliae*.

Introduction

Genomes are not randomly organized and comprise complex information beyond their linear nucleic acid sequence [64]. While scientific understanding of genome biology continues to grow, significant efforts in the past decade have focused on sequencing new species and additional genotypes of those species [65]. However, there is a great need to decode the complex information stored in these genomes, to understand genomic responses over various time scales, and ultimately to more fully understand how genotypes lead to phenotypes. With the growing number of high-quality, highly contiguous genome assemblies it is possible to analyze genome organization into chromosomes at high resolution [66]. Present day genome organization reflects evolutionary solutions to the challenges of information processing and adaptation; a genome must faithfully pass vast amounts of information across cell-cycles and reproduction, packaged into limited physical space, while achieving correct access to the information in response to developmental, environmental or chemical signals. In addition, there needs to be appreciable stochastic genetic variation to ensure that phenotypic variation is present for unknown future events. Organisms undergoing mainly asexual reproduction face an additional evolutionary constraint as they must generate this genetic variation in the absence of meiotic recombination [67]. Many economically important fungal plant pathogens are either asexual or undergo more frequent asexual reproduction compared to sexual reproduction [68]. Interestingly, fungal pathogens are subject to additional evolutionary pressure from their hosts, as host-pathogen interactions create dynamical systems with shifting, yet near-constant selective pressure on the two genomes [10]. These attributes make plant-fungal interactions a particularly interesting system to study aspects of genome evolution and genome organization [69,70].

Plant invading microbes use effectors to suppress, avoid or mitigate the plant immune system [9,71]. Plants in-turn use a variety of plasma-membrane bound and cytoplasmic receptors to recognize invasion, through recognition of the effector or its biochemical activity, creating a strong selective pressure on the microbe to modify the effector or its function to alleviate recognition [72,73]. The plant pathogenic fungus *Verticillium dahliae* causes vascular wilt diseases on hundreds of plant hosts. *V. dahliae* is presumed asexual and generates genomic diversity in the absence of sexual recombination through large-scale chromosome rearrangements and segmental duplications [59,62,74,75]. The regions undergoing such duplications and rearrangements are hypervariable between *V. dahliae* isolates, and consequently have been referred to as Lineage-Specific (LS) regions. These LS regions are often formally defined based on presence/absence variation (PAV) between strains of a species, and regions that are not designated LS are considered core. These LS regions are enriched for *in planta* expressed genes and harbor many effector genes contributing to host infection [59,61,76]. Similar non-random genomic arrangement of effectors have been reported across diverse plant pathogenic fungal and oomycete genomes [59,77–83]. One summary of these observations is referred to as the two-speed genome, in which repeat-rich regions harboring effectors evolve more rapidly than genes outside these regions [84].

Previous research in various plant-associated fungi has established a link between posttranslational histone modifications and transcriptional regulation of adaptive trait genes. These genes include effectors that facilitate host infection and secondary metabolite (SM) clusters that code for genes that produce chemicals important for niche fitness [85]. By removing or reducing enzymes responsible for particular repressive histone modifications, such as di- and trimethylation of Lys9 and Lys27 residues of histone H3 (H3K9me2/3 and H3K27me2/3), a disproportionally high number of effector and SM cluster genes are derepressed, although a direct role of these marks in transcriptional control was not demonstrated [49,52,53]. However, evidence from the fungus *Epichloë festucae* that forms a mutualistic interaction with its grass host *Lolium perenne* indicates that direct transcriptional regulation through histone modification dynamics is possible [51]. Although there are clear indications that the epigenome (i.e. heritable chemical modifications to DNA and histones not affecting the genetic sequence) plays a role in adaptive gene regulation, additional evidence is needed to fully understand this phenomenon.

Epigenetic modifications influence chromatin structure, defined as the DNA-RNA-protein interactions giving DNA physical structure in the nucleus [86,87]. This physical structure affects how DNA is organized in the nucleus and DNA accessibility. Methylation of H3K9 and H3K27 are hallmarks of heterochromatin; DNA that is tightly compacted in the nucleus [88–91]. H3K9 methylation is not only associated with controlling constitutive heterochromatin, but also tightly linked with DNA cytosine methylation (mC), which serves as an epigenetic mark contributing to transcriptional silencing [92]. A single DNA methyltransferase gene, termed *Dim2*, performs cytosine DNA methylation in the saprophytic fungus *Neurospora crassa* [93]. Histone methylation at H3K9 directs DNA methylation by DIM2 through another protein, termed heterochromatin protein 1 (HP1), which physically associates with both DIM2 and H3K9me3 [94,95]. Some fungi possess a unique pathway to limit the expansion of repetitive DNA such as transposable elements through repeat-induced point mutation (RIP), a mechanism that specifically mutates repetitive DNA in the genome during meiosis and induces heterochromatin formation [96,97]. The mutations occur at methylated cytosines resulting in conversion to thymine (C to T mutation) [98]. H3K27 methylation is associated with heterochromatin that is thought to be more flexible in its chromatin status and exist as bivalent chromatin that may be either transcriptionally repressed or active depending on developmental stage or environmental cues [99–102]. The deposition of H3K27me3 is controlled by a histone methyltransferase that is a member of a complex of proteins termed Polycomb Repressive Complex 2 (PRC2), with orthologs of the core machinery present across many eukaryotes [90,103].

In addition to heterochromatin playing a role in transcriptional regulation in filamentous fungi, epigenetic marks contributing to chromatin may influence genome evolution [104]. In *N. crassa*, DNA is physically arranged in the nucleus corresponding to heterochromatic and euchromatic domains, with strong inter- and intra-heterochromatin DNA-DNA interactions reported [105,106]. Recent experimental evidence using *Zymoseptoria tritici*, a fungal pathogen of wheat, suggests that H3K27me3 promotes genomic instability [54]. In the oomycete plant pathogens *Phytophthora infestans* and *Phytophthora sojae* a clear association exists between gene-sparse and transposon-rich regions of the genome and the occurrence of adenine N6-methylation (6mA) [107]. Collectively these examples point towards an unexplained connection between the epigenome, genome architecture, and adaptive evolution. To examine the hypothesis that epigenetic modifications influence the adaptive LS regions of *V. dahliae*, we performed a series of genetic, genomic, and machine learning analyses to characterize these regions in greater detail.

Results

DNA cytosine methylation occurs at transposable elements

To understand the role of DNA methylation in *V. dahliae*, whole-genome bisulfite sequencing, in which unmethylated cytosine bases are converted to uracil while methylated cytosines remain unchanged [108,109], was performed in the wild-type and a heterochromatin protein 1 deletion mutant (*Δhp1*). The overall level of DNA methylation in *V. dahliae* is low, with an average weighted methylation percentage (calculated as the number of reads supporting methylation over the number of cytosines sequenced) at CG dinucleotides of 0.4% (Table 1). The fractional CG methylation level (calculated as the number of cytosine positions with a methylated read over all cytosine positions) is higher, averaged to 9.7% over 10 kb windows. Weighted and fractional cytosine methylation (mC) levels are statistically significantly higher in the WT compared to the *Δhp1* mutant for all cytosine contexts (Table 1; Table S1A and B). This result is consistent with the requirement of HP1 for DNA methylation in *N. crassa* [94]. To understand DNA methylation in the context of the functional genome, DNA methylation was analyzed over genes, promoters, and transposable elements (TE). Despite statistically significant differences between WT and *Δhp1* for gene and promoter methylation, the bisulfite sequencing data shows virtually no DNA methylation at these two features (Fig. 1A). We attribute the difference to a marginal set of elements having a real difference between the genotypes, but the biological significance is likely negligible (Fig. 1A, Table S2). In contrast, there is a much higher degree of methylation, and a notable difference between wild-type and *Δhp1* methylation levels at TEs (Fig. 1A, bottom panel), with the average CG methylation level being five times higher in the wild-type strain.

TABLE 1 | Summary of DNA methylation in *Verticillium dahliae* wild-type (WT) and heterochromatin protein 1 deletion mutant (Δ*hp1*) as measured by whole genome bisulfite sequencing calculated over 10 kb non-overlapping windows.

Genotype	Avg. Weighted mCG	Avg. Weighted mCHG	Avg. Weighted mCHH	Avg. Fraction mCG	Avg. Fraction mCHG	Avg. Fraction mCHH
WT	0.0040	0.0037	0.0034	0.097	0.097	0.088
Δhp1	0.0030	0.0030	0.0032	0.082	0.083	0.079

Avg. Weighted**,** The average of total methylated cytosines in a given context divided by total cytosines in that context in a 10 kb windows; Avg. Fraction**,** The total cytosines positions with a read supporting methylation divided by total cytosines in a specific context in a 10 kb window; mCG, methylated cytosine residing next to a guanine; mCHG, methylated cytosine residing next to any base that is not a guanine next to a guanine; mCHH, methylated cytosine residing next to any two bases that are not a guanines.

To further analyze DNA methylation levels and confirm that the low DNA methylation levels in the wild-type strain are indeed different than those in Δ*hp1,* CG DNA methylation levels were plotted in 10 kb windows across individual chromosomes. These plots clearly show that DNA methylation is not continuously present across the *V. dahliae* genome, and DNA methylation is significantly different between wild-type and Δ*hp1* (Fig. 1B and 1C). Furthermore, regions in the genome with higher densities of TEs and lower gene numbers have higher levels of DNA methylation, consistent with the global DNA methylation summary (Fig. 1B and 1C). Interestingly, these results show that while DNA methylation is only present at TEs, not all TEs are methylated, a phenomenon that was previously described as 'non-exhaustive' DNA methylation [110]. To further understand this phenomenon, we sought to identify discriminating genomic features that could account for some TEs not being methylated. The whole-chromosome methylation data suggested a lack of DNA methylation at previously identified LS regions (Fig. 1C, grey windows). These LS regions were previously detailed for *V. dahliae,* and are characterized as regions that are highly variable between isolates of the species, are enriched for actively transcribed TEs, and contain an increased proportion of genes involved in host virulence [59,62,74]. Thus, we tested if DNA sequences at LS regions are less frequently methylated by comparing weighted mCG levels in 10 kb bins containing at least one TE for core versus LS regions. This analysis showed significantly more DNA methylation for core bins, which cannot be accounted for by a simple difference in the number of TEs in the core and LS regions analyzed (Fig. 1D and E). Higher CG methylation levels also hold true when analyzed at the level of individual TE elements (Fig. 1F; Table S2). Collectively, these analyses demonstrate that DNA methylation occurs almost exclusively at TEs and, importantly, that not all TEs are methylated. This observation can in part be explained by mCG differences for TEs in the core versus LS regions.

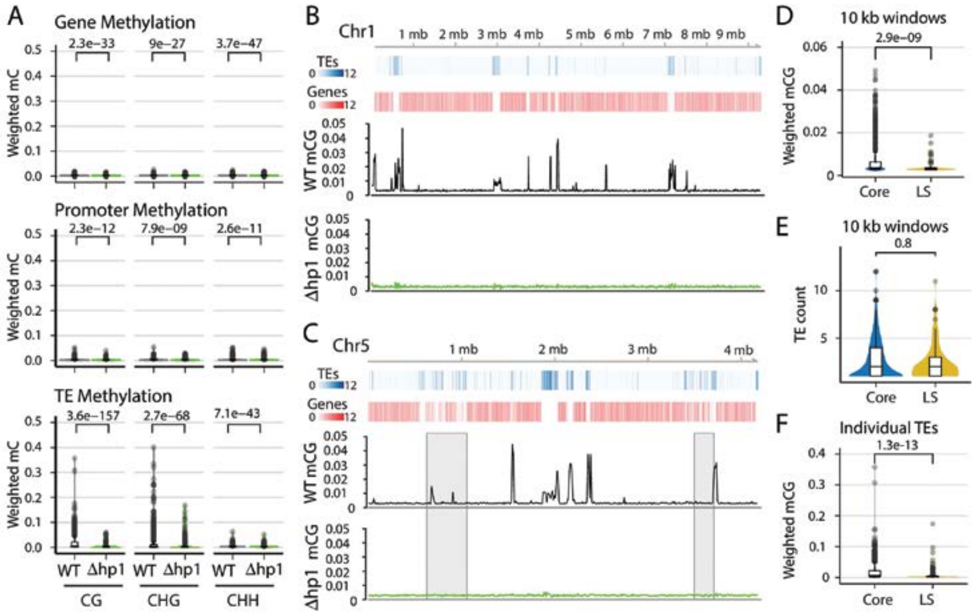


FIGURE 1 | DNA methylation is only present at transposable elements, but not at those present in LS regions. (A) Violin plot of the distribution of DNA methylation levels quantified as weighted methylation over Genes, Promoters and TEs. Cytosine methylation was analyzed in the CG, CHG and CHH sequence context. Methylation was measured in the wild-type (WT) and heterochromatin protein 1 knockout strain (Δ*hp1*). (B, C) Whole chromosome plots showing TE and Gene counts (blue and red heatmaps) and wild-type (black lines) and

Δhp1 (green line) CG methylation as measured with bisulfite sequencing. Data is computed in 10 kilobase non-overlapping windows. (C) Two previously defined LS regions [74] are highlighted by grey windows. (D) Violin plot of weighted cytosine methylation in 10 kb windows broken into core versus LS location (E) Same as D but plots are for the counts of TEs per 10 kb window. (F) Same as in D but methylation levels were computed at individual TE elements. Statistical differences for indicated comparisons were carried out using non-parametric Mann-Whitney test and Holm multiple testing correction with associated p-values shown.

Transposable element classes have distinct genomic and epigenomic profiles

To understand the functional status of the various TEs in the genome, DNA-histone modification location data were collected using chromatin immunoprecipitation followed by sequencing (ChIP-seq) against H3K9me3 and H3K27me3, which allows for the identification of DNA interacting with these modified histones. Characteristics of TE sequence, such as GC percentage, composite RIP index (CRI), and TE age, estimated as the Jukes-Cantor distance to the consensus sequence of the specific TE family, were calculated (see methods). To further classify genomic regions as eu- or heterochromatic, we performed an assay for transposase accessible chromatin and sequencing (ATAC-seq) [111]. This method uses a TN5 transposase to restrict physically accessible DNA in the nucleus and tags the DNA ends with oligonucleotides for downstream sequencing. Transcriptional activity was assayed using RNA-sequencing. To analyze all of these TE characteristics (variables) at once, dimensional reduction with principle component analysis (PCA) was employed, which facilitates data interpretation on two-dimensions to identify important variables and their relationships within large datasets. The individual TEs were grouped into four broad classes (Type I DNA elements and Type II LTR, LINEs, and Unspecified elements) and analyzed for each measured variable. The first dimension of PCA shows the largest separation of the data points and variables, and largely separates the data based on euchromatin versus heterochromatin features (Fig. 2A, PC1). This is seen by the variables ATAC-seq, %GC, RNA-sequencing, H3K9me3 ChIP, CRI and DNA methylation (mCG) being furthest separated along the x-axis (Fig. 2A). Open chromatin features such as higher ATAC-seq, %GC, and transcriptional activity are positive on the x-axis, with small angles between the vectors, indicating correlation among those variables. Conversely, features associated with heterochromatin, such as H3K9me3 association, DNA methylation and indication of RIP (CRI) are all negative on the x-axis, and the position of their vectors indicates correlation among these variables, and negative correlation to the euchromatin features (Fig. 2A). The second axis discriminates elements based on their H3K27me3 profile and sequence characteristics such as Jukes Cantor (TE age), Identity and Length (Fig. 2A). Regarding the TE classifications, there is a stronger association for the LTR and Unspecified elements with the heterochromatin features (Fig. 2A, grey and red ellipse extending along negative x-axis). Collectively, this multivariate description of TEs identifies those that are more transcribed and open as having lower association with H3K9me3, mCG, and RIP mutation. There are statistically significant differences between the TE types for each of these variables (Table S3), and the LTR elements have the highest levels of H3K9me3 and mCG, along with the highest CRI values and lowest %GC, consistent with the mechanistic link between the four variables (Fig. 2B). Interestingly, a bimodal distribution occurs for %GC and CRI within the LTR and Unspecified elements, indicating that some of the LTR elements have undergone RIP and are heterochromatic, while other elements have not been subject to this mechanism (Fig. 2B).

This delineation occurs for the Unspecified and LTR elements with a %GC sequence content less than approximately 40%, which have positive CRI values and high H3K9me3 signal (Fig. 2C). A similar distinction is seen with ATAC-seq data that show a clear break around 40% GC content, and elements below this have lower ATAC-seq signal and higher H3K9me3 signal (Fig. 2D). These trends are not observed for the LINE and DNA elements (Table S3). These results suggest that LTR and Unspecified TE elements exist in two distinct chromatin states in the genome.

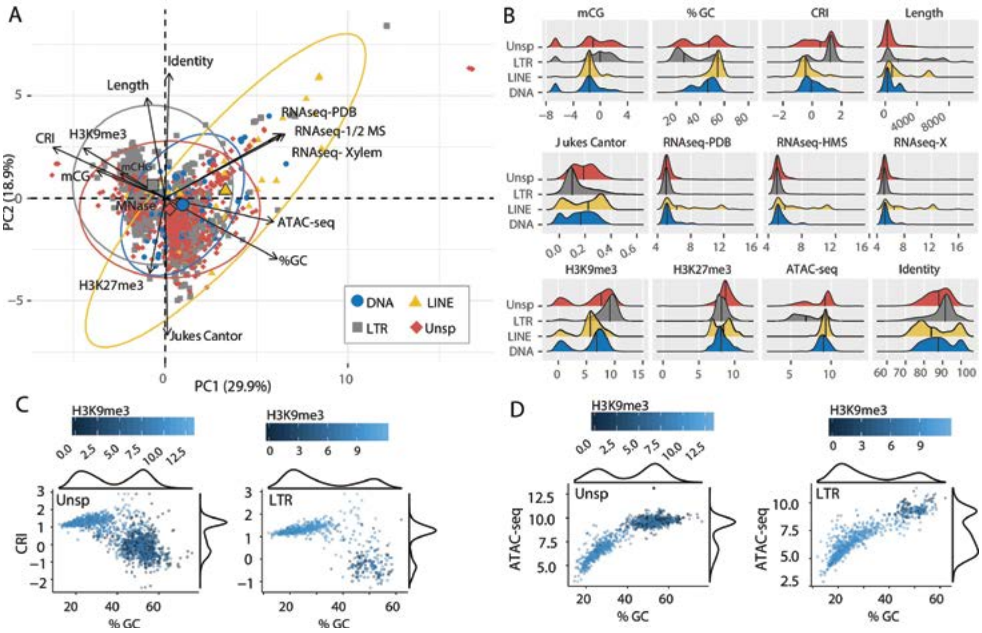


FIGURE 2 | Individual TE families have distinct epigenetic and physical compaction profiles. (A) Principle component analysis for 14 variables measured for each individual TE. Each vector represents one variable, with the length signifying the importance of the variable in the dimension. The relationship between variables can be determined by the angle connecting two vectors. For angles <90^0, the two variables are correlated, while those >90^0 are negatively correlated. Each individual element is shown and highlighted by color and symbol as indicated by the key. Colored ellipses show the confidence interval for the four families along with a single large symbol to show the mean position for the four families. mCG, weighted CG DNA methylation; mCHG, weighted CHG DNA methylation; CRI, Composite RIP index; %GC, percent GC sequence content; Identity, Nucleotide identity as percent identity to the consensus TE sequence of a family; Length, element length; Jukes Cantor, Jukes Cantor corrected distance as proxy of TE age; RNAseq, RNA-sequencing reads from (PDB), half strength MS (HMS) or tomato xylem sap (Xylem) grown fungus expressed as variance stabilizing transformed log2 values (see methods for details); H3K9me3, log2 (TPM+1) values of mapped reads from H3K9me3 ChIP-seq; H3K27me3, log2 (TPM+1) values of mapped reads from H3K27me3 ChIP-seq; ATAC-seq, log2 (TPM+1) values for mapped reads from Assay for transposase accessible chromatin. (B) Ridge plots showing the distribution of the individual TE families per variable. The median value is shown as a solid black line in each ridge. Variables same as in A except for mCG, log2(weighted cytosine DNA methylation + 0.01). (C) Scatter plot for %GC versus CRI values for individual TE elements shown as points. The two plots are for TEs characterized as Unspecified (Unsp) or LTR, labeled in the upper left corner. Each point is colored according to log2 (TPM+1) values from H3K9me3 ChIP-seq, scale shown above each plot. A density plot is shown for both variables on the opposite side from the labeled axis. (D) Same as in C, but the y-axis is now showing the log2 (TPM+1) values from ATAC-seq.

Transposable element location significantly influences the epigenetic and DNA accessibility profile

To further dissect the relationship between epigenetic modifications, chromatin status and genomic location, pair-wise comparisons were made for all TEs in core versus LS regions. All measured variables, except TE length, are significantly different for TEs in the core versus LS regions (Table S4). Further division of the TEs indicated that the LTR and Unspecified elements showed the greatest differences for core versus LS measurements (Fig. 3A), demonstrating that the major driver of core versus LS differences are driven by the LTR and Unspecified elements. The bimodal distribution for GC content, CRI, H3K9me3, and ATAC-seq can be accounted for in part by core versus LS separation (Fig. 3B, red versus grey). Collectively, the status of the LS TE elements can be characterized as devoid of DNA and H3K9 methylation, low RIP signal, generally higher than 0.5 GC content, higher levels of H3K27me3, more open with ATAC-seq signal, and higher transcription levels (Fig. 3D). The core versus LS location is not sufficient to fully explain the chromatin status, as there are many elements located in the core genome that share a similar profile with the LS elements (Fig. 3D, elements highlighted in black boxes), but as an ensemble, the core elements are statistically different than those found at LS regions.

Significantly different chromatin status between core and LS regions extends to larger DNA segments

The analysis of TEs in the genome clearly shows that a subpopulation of elements that occur in the previously defined LS regions have different epigenetic modifications and physical openness compared to those in the core genome. LS regions are significant for *V. dahliae* biology as they code many proteins which support host infection [59,74]. To capture a more global view of core versus LS regions, the genome was analyzed using 10 kb non-overlapping windows, revealing the same global patterns along the linear chromosome sequence; regions with high TE density tend to have lower GC content, higher DNA and H3K9 methylation and a lack of ATAC-seq reads. The distribution of H3K27me3 appears more complicated. This mark overlaps with that of DNA and H3K9 methylation, as nearly all regions with these two modifications also have H3K27me3, yet we observed additional regions that contain only H3K27me3 and lack DNA and H3K9 methylation (Fig. 4A). The regions that contain DNA methylation and H3K9me3 are nearly identical and for simplicity refer to these regions going forward as being marked by H3K9me3. Interestingly, regions marked by H3K27me3 that lack H3K9me3 have more open DNA than region with H3K27me3 also containing H3K9me3 (Fig. 4A, ATAC). This is apparent for the LS regions that appear to have increased H3K27me3 signal, lack H3K9me3 and are less open than the genomic background but not as closed as the regions marked by H3K9me3 (Fig. 4B, regions marked by grey boxes). PCA was again employed to combine the variables into a single analysis, with the first dimension explaining nearly 60% of the variation in the data (Fig. 4C). The first dimension largely captures the variables describing euchromatin versus heterochromatin, such that ATAC-seq and %GC are furthest separated on the x-axis from H3K9me3, DNA methylation and TE density (Fig. 4C).

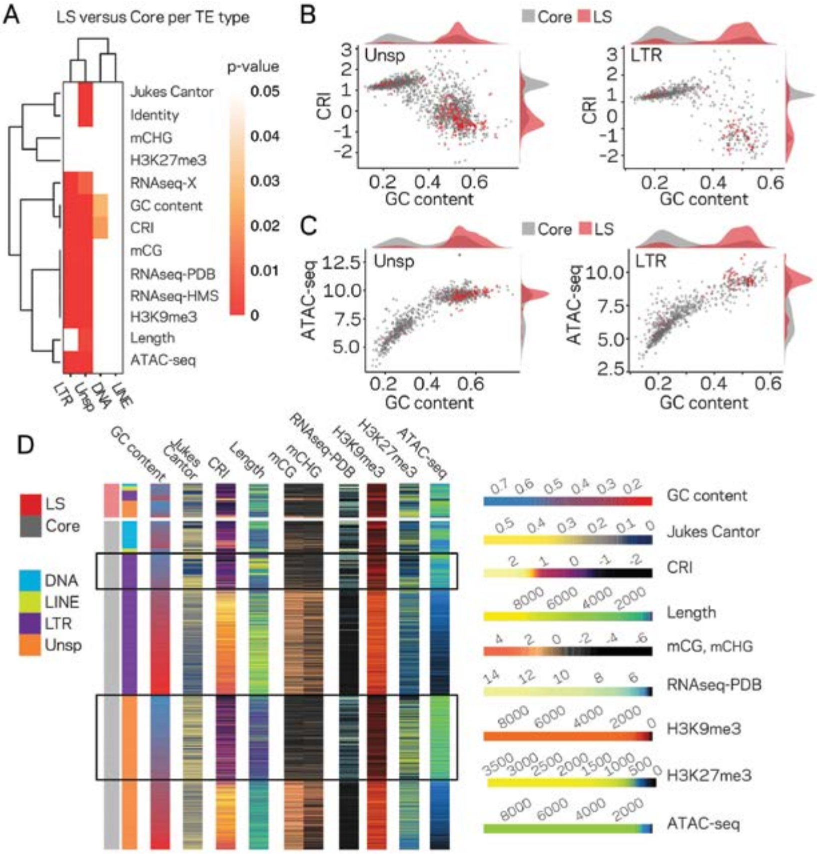


FIGURE 3 | The LTR and Unspecified elements have significantly different chromatin profiles based on core versus LS location. (A) Heatmap comparing core versus LS values within the four TE classifications for the variable listed to the right. Plot colored based on p-values from Wilcoxon rank sum test. P-values ≥ 0.05 are colored white going to red for p-value ≅ 0. (B) Scatter and density plots similar to those shown in Fig. 2c except the individual TE points are colored by core (grey) versus LS (red) location. The density plots are also constructed based on the two groupings (C) Similar to B, with the y-axis now showing the log2 (TPM+1) values from ATAC-seq (D) Multiple grouped heatmaps for ten variables collected for each TE. Each row represents a single element and the same ordering is used across all plots. The LS elements are grouped at the top, indicated by the red bar at the top left, and the core elements are grouped below, indicated by the grey bar at the left. Elements are further grouped by the four classifications indicated by the color code shown to the left. Within each element group, the elements are ordered by descending GC content. The scale for each heatmap is shown at the right. GC content, fraction of GC in sequence; Jukes Cantor, corrected distance as proxy of TE age; CRI, Composite RIP index; Length, element length; mCG and mCHG, log2(weighted cytosine DNA methylation+0.01) for CG and CHG respectively; RNAseq-PDB, variance stabilizing transformed log2 RNA-sequencing reads from PDB grown fungus; H3K9me3 and H3K27me3 and ATAC-seq, TPM values of mapped reads H3K9me3 ChIP-seq, H3K27me3 ChIP-seq, or Assay for transposase accessible chromatin respectively. Black boxes highlight LTR and Unsp elements in the core that have euchromatin profiles.

Interestingly, the DNA segments classified as core are mostly associated with this separation across the first dimension (Fig. 4C). The second and third dimensions of the PCA explained a similar amount of variation in the data; 14.4% and 10.7%, respectively. Data from the RNA-seq experiment contributed nearly all the information to the second dimension (Table S5), while the H3K27me3 ChIP-seq data contributed most of the information in the third dimension (Table S4). Interestingly, when this third dimension is considered, we observe a strong separation of the core from the LS regions (Fig. 4C, y-axis), suggesting that the LS regions of the genome are less defined by DNA openness, and DNA or H3K9 methylation but more by H3K27me3 and transcriptional activity.

Our observations can be summarized into a genome-wide model; for the core genome, regions with higher TE density have low ATAC-seq signal (closed DNA) and elevated H3K9me3 signal and thus represent the heterochromatic regions (Fig. 4D, cluster of large blue dots plotted at middle left). Core genomic regions that are gene-rich have a higher ATAC-seq and lower H3K9me3 signal and represent the euchromatic portion of the genome (Fig. 4D, cluster of small blue dots plotted in the lower-middle section). The LS regions are a hybrid of the two that contain high TE density and higher H3K27me3 signal but have higher ATAC-seq signals when compared with similar TE containing regions in the core genome (Fig. 4D, cluster of large yellow triangles plotted in the middle). This simple model of the genome accounts for many of the phenomena described here, and links the epigenome, physical genome and functional genome.

Machine learning predicts more lineage-specific genomic regions than previously considered

Given that a clear model emerges that links the epigenome and physical openness of DNA with adaptive regions of the genome, we assessed the extent to which these features can predict core or LS regions. Stimulated by our observations (Fig. 4), we used ATAC-seq, RNA-seq, H3K27me3, TE density, and H3K9me3 along with the binary classification of the 10 kb windows as core or LS for machine learning. Four supervised machine learning algorithms were used to train (i.e. learn) on 80% of the data (2890 regions), while the remaining 20% (721 regions) were used for prediction (i.e. test), using a 10-fold cross validation repeated three times. Assessing the classifier's performance using area under the receiver operating characteristic (auROC) curve suggested excellent results ranging from 0.94 to 0.95, with a value of 1 being perfect prediction (Fig. 5A). While auROC is the *de facto* standard for machine learning performance [112], it is not appropriate for assessing predictive performance of binary classification problems when the two classes are heavily skewed as it overestimates performance due to the high number of true negatives [113]. This is the case for our analysis in which the test set (721 regions) contains only 33 of the known LS regions (4.6%). To more accurately assess model performance, precision-recall curves were employed as these do not use true negatives, and are therefore less influenced by skewed binary classifications [114]. All four algorithms consistently outperformed a random classifier, with the boosted classification tree (BCT) and stochastic gradient boosting (GBM) algorithms having the same highest area under the precision-recall curve of 0.52 (Fig. 5B). However, the confusion matrix indicated that the BCT model only identified 13 of the 33 LS regions (Table 2), resulting in poor recall (Table 3). In contrast, the other three models did identify most of the known LS regions (high recall), but had lower precision caused by the high rate of false positives (Table 2 and 3). The Matthews correlation coefficient (MCC), an analogous measure to accuracy but more appropriate for unbalanced binary classification, indicated that the GBM and random forest (RF) models performed the best (Table 3).

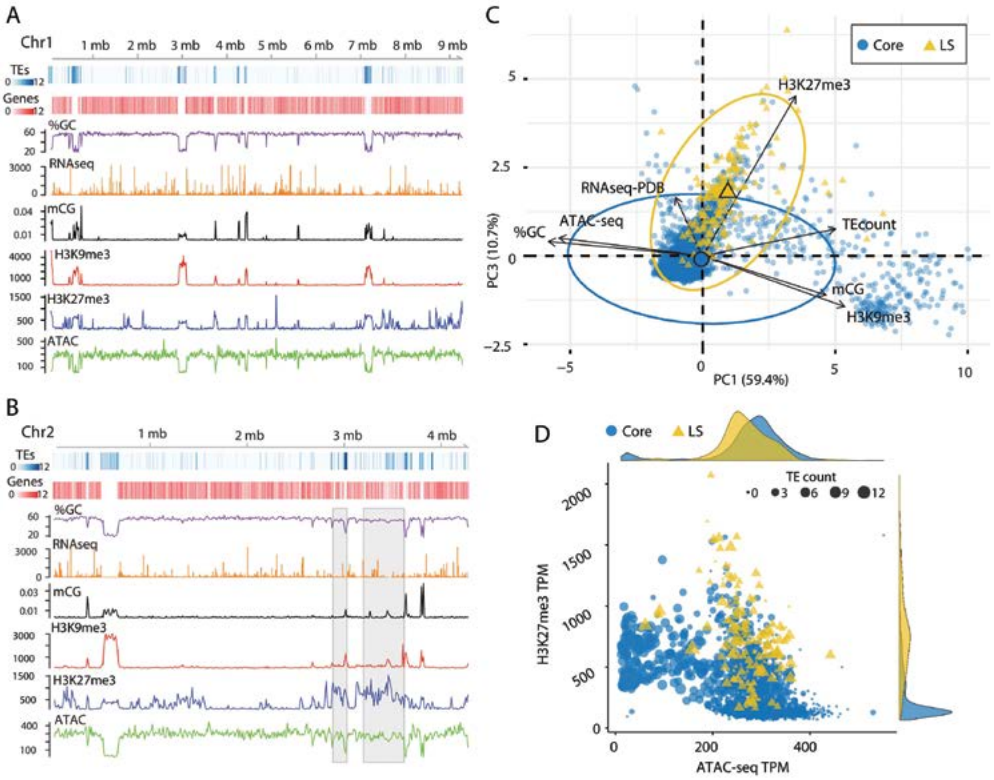


FIGURE 4 | Epigenome and physical DNA characteristics collectively define core and LS regions. (A and B) Whole chromosomes plots showing TE and gene counts over 10 kb genomic windows, blue and red heatmaps respectively. The %GC content is shown in purple, RNA-seq show in orange, CG cytosine DNA methylation shown in black, H3K9me3 and H3K27me3 ChIP-seq shown in red and blue respectively, and ATAC-seq shown in green. Values are those previously described. (B) Two LS regions are highlighted with a grey window. (C) Principle component analysis for seven variables at each 10 kb window. Dimension 1 and 3 are plotted and collective explain ~70% of the variation in the data. The individual symbols are colored by genomic location with core (blue circles) and LS (yellow triangles). Colored ellipses show the confidence interval for the core and LS elements with a single large symbol to show the mean. (D) Scatter plot of the 10 kb windows colored for core and LS location by ATAC-seq data (TPM, x-axis) and H3K27me3 (TPM, y-axis). The size of each symbol is proportional to the TE density shown in the upper right corner. The density plot of each variable is shown on the opposite axis.

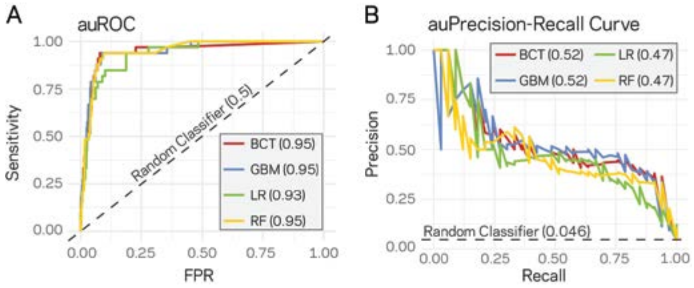


FIGURE 5 | Supervised machine learning can predict LS regions based on epigenome and physical genome characteristics. (A) Area under the Response operator curve (auROC) plotting sensitivity and false positive rate (FPR) for four machine learning algorithms, BCT- Boosted classification tree; GBM- stochastic gradient boosting; LR- logistic regression; RF- random forest. The auROC scores are shown next the algorithm key in the grey box. The black dotted line represents the performance of a random classifier. Perfect model performance would be a curve through point (0,1) in the upper left corner. (B) Area under the Precision-Recall curve for the same four models shown in A. Area under the curves are shown in the figure key in the grey box. The black dashed line shows the performance of a random classifier, calculated as the TP / (TP + FN). Perfect model performance would be a curve through point (1,1) in the upper right corner.

TABLE 2 | Confusion Matrix for LS versus core prediction in *V. dahliae*.

		Known	
	Predicted	Core	LS
LR	Core	638	7
	LS	50	26
GBM	Core	645	5
	LS	43	28
BCT	Core	672	20
	LS	16	13
RF	Core	623	2
	LS	65	31

LR, Logistic Regression; GBM, Stochastic Gradient Boosting; BCT, Boosted Classification Tree; RF, Random Forest; Core, regions of the genome defined as core; LS, regions of the genome defined as Lineage Specific.

TABLE 3 | Assessment values for the four tested machine learning algorithms used to classify genomic regions.

Models	Precision	Recall	MCC	F1	F2
LR	0.34	0.79	0.49	0.48	0.63
GBM	0.39	0.85	0.55	0.54	0.69
BCT	0.45	0.39	0.39	0.42	0.40
RF	0.32	0.94	0.52	0.48	0.68

LR, Logistic Regression; GBM, Stochastic Gradient Boosting; BCT, Boosted Classification Tree; RF, Random Forest; MCC, Matthews Correlation Coefficient.

The results indicate that the machine learning algorithms are well-suited to identify the previously known LS regions in the test data. Additionally, the algorithms identified a relatively large number of regions as LS that were previously classified core. The original classification of core and LS in *V. dahliae* was based on presence/absence variations identified from genomic information of only few strains [59,74]. Consequently, we reasoned that regions here classified as LS by ML could be genuine LS regions that were originally missed due to the then limited diversity of the *V. dahliae* strains sequenced. The two best models from the initial testing, GBM and RF, predicted a total of 96 and 81 regions as LS respectively, suggesting there could be 2 to 3 times more LS DNA than previously identified. To improve the genome-wide estimate and to further assess the robustness of ML for LS region prediction, we re-ran the GBM and RF algorithms on 15 new training-test splits, independently training and predicting on each set (see methods for details). This approach nearly saturated the genome, providing multiple predictions per window and only 124 of the 3611 regions were missed (Table S6). The average MCC performance estimate of the GBM and RF classifiers were 0.53 and 0.48 over the 15 runs, and our results indicate consistent performance across the independent predictions (Fig. 6A; Table S5, S6, and S7). The GBM classifier predicted a total of 285 of the 10 kb regions as LS, while the RF classifier predicted 388 (Table S7 and S8). The LS predictions for the two models were in agreement for 280 regions, which is 98% of the GBM predictions and 72% of those from the RF (Fig. 6B), overall showing high agreement between the two classifiers. Consensus predictions were generated from the two classifiers if a region was predicted as LS by both models, and a conservative joining step was employed in which a single predicted core region was called LS if it was flanked by LS predictions on both sides (see methods). This resulted in a total of 280 regions predicted as LS by both classifiers and an additional 41 regions due to the joining. In total, this new classification nearly doubles the total amount of LS regions compared with the original observations [59,74]. The original classification of LS regions in *V. dahliae* clustered in four larger regions [59,74]. We were interested to understand the physical genomic location of the originally identified and the newly predicted LS regions. The results of the individual classifiers reveal that the new regions are also not randomly dispersed across the genome (Table S8). The consensus prediction from the two classifiers identified the large blocks of LS regions from the original observations, along with new clusters of LS regions such as those on chromosomes 4, 6, and 8 (Fig. 6C and 6D). Importantly, the newly defined set of LS regions supports a clearer separation of the LS regions from the core regions.

Previously, LS regions were defined as genomic regions with extensive sequence variability between *V. dahliae* strains [61,62,74], and the LS regions were enriched for *in planta* induced genes known or presumed to be involved in host infection [59]. To validate that the updated LS classifications following ML still captured these characteristics, we analyzed PAV and performed genome-wide enrichment analyses for *in planta* induced genes, for genes coding secreted proteins, and for effector genes in old and new LS regions. To analyze PAV, we summarized missing DNA segments, termed absence counts, from 42 *V. dahliae* strains based on whole-genome sequencing (see methods for details). The original LS classification was defined by PAV, and thus we anticipated that the updated LS classification, if valid, should still reflect higher variability for LS regions between *V. dahliae* strains. The analysis showed that the majority (82.6%) of 100 bp windows classified as core were present in all 42 strains, and the distribution of absence counts suggested low variation (Fig. 6E, mean absence count = 2.5). In contrast, only 37.3% of LS classified regions were present in all 42 strains, and had a significantly higher mean absence count of 12.3 (Wilcoxon rank sum, $p < 2.2e-16$) (Fig. 6E). These results help validate our approach by showing the PAV disparity between LS and core elements is readily observed in this set of 42 strains. We assessed the absence counts for TEs and genes to understand what functional elements account for the differences between the LS and core. This analysis showed that the absence counts for TEs was higher than for an average 100 bp window, but there is no difference for the distribution of TE absence counts between those classified as LS versus core (Wilcoxon rank sum, p = 0.99) (Fig. 6F and G). Interestingly, this was not the case for genes, where the majority (64.4%) of core classified genes were present in all 42 strains, and only 43.1% of LS classified genes were present in all strains (Wilcoxon rank sum, $p < 2.2e-16$; core mean absence count of 0.6 and LS mean of 10.3) (Fig. 6H and 6I). These results suggest that TEs are generally variable between strains regardless of their genomic location, but the likelihood of a gene being absent varies significantly based on its location in a core or LS region. Analyzing genes based on functional categories showed that the number of *in planta* induced genes in LS regions doubled from the old to new classification, and while these genes were overrepresented in the old designations (X^2=9.94, p=.002), this overrepresentation was even more pronounced for the new LS regions (X^2=29.96, $p < .000001$). As expected, we found that *in planta* induction for LS genes was significantly higher than for core genes (Mann-Whitney test, p-value = 1.34e-50) (Fig. 6D). Even though effectors were not overrepresented in the old LS classification (X^2~0, p=1), they were present in new LS regions far greater than by chance, increasing 3.5 times compared to the old classification (X^2=11.18, p =.0008). The largest change was observed for genes coding proteins with secretion signals. Using the old classification, the LS regions were actually significantly underrepresented for genes coding secreted proteins (X^2=27.05, $p < .000001$). This result could have a number of explanations, but under the assumption that LS regions contain a significant portion of genes involved in inter-species interactions, this analysis suggests that the old classification was missing important genes. Indeed, while the new LS classification approximately doubled the number of genes designated as LS (494 to 998), the number of LS genes with secretion signals increased 5-fold (20 to 109) (X^2=3.63, p = .05) (Table S9).

Collectively, these analyses suggest that ML algorithms can be used to predict new LS regions based on epigenetic and physical DNA accessibility data. The identification of potentially new LS regions missed in the original classification provides new avenues to identify proteins important for host infection and adaptation. Our predictions were validated by showing enrichment for functional categories of genes known to be important for infection and host adaptation. Our results further show that the expanded classifications represent regions of the genome that are more variable across strains of the species and uncover the new finding that LS genes in particular experience greater PAV in *V. dahliae*. These results support that genome structure is influencing genome function and Demonstrates a ML approach for predictive biology that advances our understanding of genome biology.

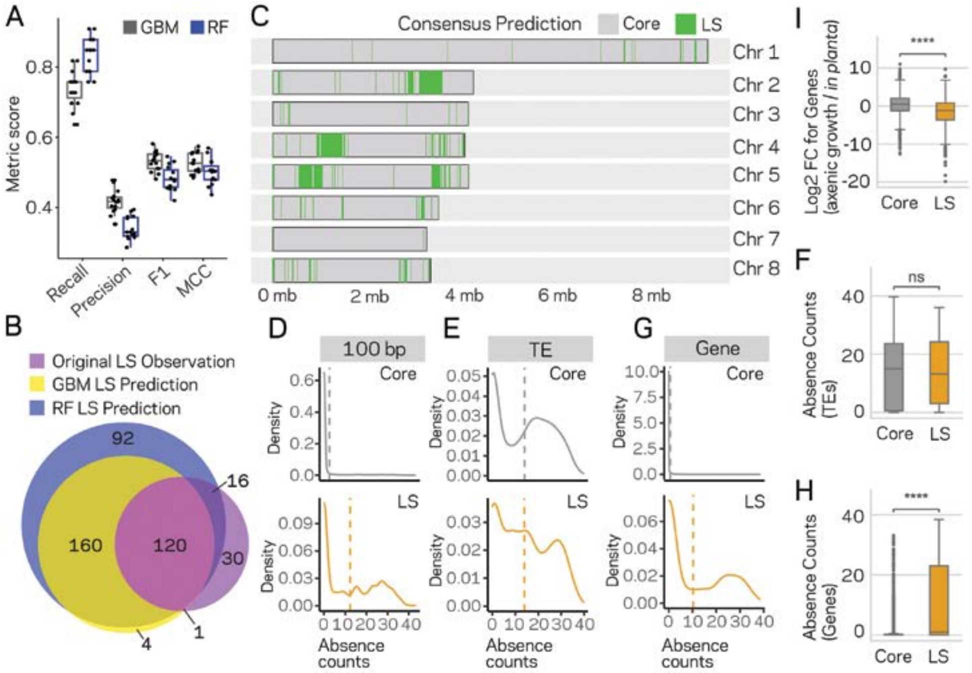


FIGURE 6 | Machine Learning predictions for genome-wide LS content. (A) Two machine learning algorithms, Stochastic Gradient Boosting (GBM) and Random Forest (RF), were used to predict LS regions from 15 independent training-test splits (80/20). Classifier performance was measured for each of the 15 trials, and summarized as a boxplot with each trial represented as a point. (B) Venn diagram showing the overlap between the results of the two classifiers and the original observations of LS regions [59,74]. Each slice of the diagram shows the number of LS regions predicted, see methods for additional details. (C) Schematic representation of the eight chromosomes (labeled on right) of *V. dahliae* strain JR2. Core (grey) and LS (green) classification for 10kb windows. The consensus predictions were those made by both the GBM and RF model (in total 280). (D) Boxplot showing a significant difference for *in planta* gene induction between core and LS genes, Mann-Whitney U test p-value = 1.34e-50. (E) Density distribution for core (grey) and LS (orange) elements based on absence counts over 100 bp windows. The mean absence counts are shown as a dashed vertical line. (F) Similar to E but the analysis was conducted for TEs. (G) Boxplot showing no significant difference between core and LS TE elements for absence counts, Mann-Whitney U test p-value = 0.92. (H) Similar to E but the analysis was conducted for genes. (I) Boxplot showing a significant difference between core and LS genes for absence counts, Mann-Whitney U test p-value = 3.82e-104. ns, non-significant; **** p-value < 1.00e-4.

Unsupervised genome clustering using chromatin data supports functional differences for core and LS classification

Using the described supervised learning approach, we were able to identify new regions of the genome as LS, and subsequently validated that these new regions fit the characteristics of LS function. To further confirm these results and define the functional genome, uniform manifold approximation and projection (UMAP) [115] was employed for dimensional reduction and clustering of TEs and genes based on transcriptional, chromatin and DNA openness data. The significance of this alternative approach is that it is unsupervised, and does not rely on, or is influenced by, prior LS and core classifications. Under the hypothesis that genome structure influences genome function, a prediction is that LS and core classification (evolutionary function) should show a non-random spatial association when layered on-top of the UMAP clustering (genome structure and function data). This approach generated three distinct UMAP clusters for TEs, which we termed Group1, 2 and 3 (Fig. 7A). When the LS and core classification were applied to the UMAP clusters, Group1 and Group2 displayed significant non-random associations of core- and LS-designated elements respectively (x-axis p-value = 1.77e-38 and y-axis p-value = 9.0e-80, Mann-Whitney rank test) (Fig. 7A). Additionally, the core and LS elements were enriched in Group1 and Group2, respectively, (X^2=348.84, p =1.78e-76) (Table S10). To understand the UMAP clustering results, each genomic variable used for UMAP was summarized across the three groups. The TEs in Group1 have the lowest GC content, transcriptional activity, and DNA openness, along with the highest CRI, H3K9me3 signal, and DNA methylation. Based on these characteristics, we conclude that Group1 TE elements from the UMAP clustering are largely heterochromatic. The TE elements in Group2 and Group3 are more similar to one another based on the per-variable analysis, although many statistically significantly differences still exist. These findings show that unsupervised genomic clustering on functional and structural data can recapitulate a large part of our previously defined core and LS regions. We interpret this data as supporting a link between genome organization on a physical level (i.e. epigenetics and DNA accessibility), genome function (i.e. transcriptional activity) and genome adaptation to the environment (i.e. LS and core regions). Interestingly, while there was no difference in PAV for TE elements classified as core and LS (Fig. 6G), we did find significant differences between all three UMAP groups. Analyzing the absence counts within the three UMAP groups revealed significant differences between LS and core elements (Fig. 7C). Specifically, the median absence count for all core TEs was 15 (Fig. 6G), but the median count is less than 1 for the UMAP Group2 and Group3 core elements, which make up 39.4% of the core TEs (Fig. 7C). It is not clear what accounts for the core TE elements in UMAP Group2 and Group3, which are less defined by heterochromatic characteristics but experience less absence variation across the analyzed *V. dahliae* strains.

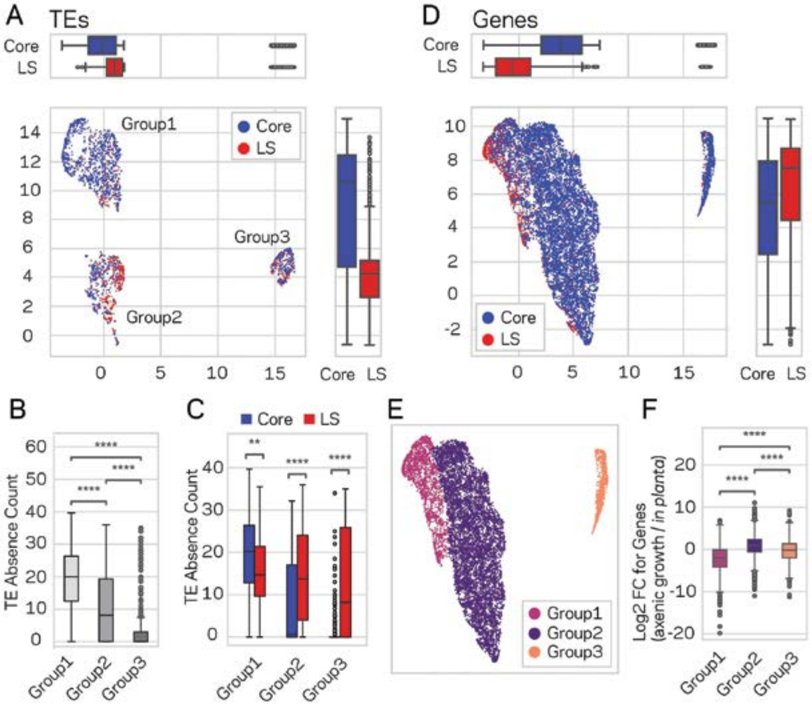


FIGURE 7 | Genome-wide UMAP clustering details that functional elements labeled core and LS have different epigenetic and DNA characteristics. (A) Uniform Manifold Approximation and Projection (UMAP) clustering of individual *V. dahliae* TEs, color coded for core (blue) and LS (red). UMAP clustering in two dimensions resulted in the identification Group1, 2, and 3 elements. Boxplots are shown opposite of the x- and y-axis to quantify the UMAP designation of the LS and core elements. Statistical difference measured using Mann-Whitney U test for UMAP labeling on the x-axis, p-value = 1.77e-38, and y-axis, p-value=9.04e-80. (B) Boxplot for TE absence counts for UMAP Group1, 2 and 3 elements. Statistical difference measured using Conover's test (C) Boxplot for TE absence counts for LS and core elements in UMAP Group1, 2 and 3. Statistical difference measured using Mann-Whitney test. (D) Similar UMAP clustering as shown in (A), but performed using genes as the clustering elements shown as individual dots. Marginal boxplots shown as in (A), x-axis p-value = 5.45e-221, and y-axis p-value = 1.84e-28. (E) UMAP gene clustering, color coded to show three groups. (F) Boxplot for *in planta* gene induction for UMAP Group1, 2 and 3 genes. Multiple comparisons performed using Conover's test. **, p-value < 0.01; ****, p-value < 0.0001. All statistical comparisons used Holm multiple-test correction of p-values.

We similarly analyzed the genes using UMAP, under the same prediction that LS and core elements would show spatial clustering within the UMAP analysis. Here, UMAP clustered the majority of the genes in the genome (89.9%) into a single cluster (Fig. 7D). Within this group, we observed significant local-clustering of LS classified genes (x-axis p-value = 5.45e-221 and y-axis p-value = 1.84e-28, Mann-Whitney rank test) (Fig. 7D). Based on this, the large group was further sub-divided, and the UMAP results were analyzed based on three groups (Fig. 7E). The UMAP Group1 contained 4 times more LS genes than expected (X^2=2119.4, p-value = 0) (Table S11), resulting in Group1 containing 75% of the LS genes. Furthermore, the Group1 genes had the lowest GC content, transcription in axenic culture, and the highest

H3K9 and H3K27 trimethylation (Fig. S11). The Group1 genes also had the highest *in planta* gene induction, which was not a variable included in the UMAP creation, indicating the UMAP clustering was associated with this characteristic (Fig. 7F). There was no significant difference for gene absence counts between the three groups (Kruskal-Wallis H statistic=4.18, p-value=.09) (Table S10A and S10B), however this was influenced by the majority of genes having a zero-absence count. When only genes with an absence count greater than 0 were analyzed (i.e. the gene was missing from at least one strain), the greatest mean absence was for Group1 genes (absence mean of 13.4, 1.9 and 3.1 for Group1, 2 and 3 respectively) (Table S10C and S10D). Thus, the UMAP Group1 genes have characteristics of heterochromatin when grown axenically, are enriched for LS classified genes, display significantly higher *in planta* gene induction, and the genes are statistically significantly more variable across the analyzed *V. dahliae* strains. These analyses identify previously unreported links between genes important for pathogenic microorganisms and specific genomic regions defined by epigenetic and DNA accessibility characteristics.

Discussion

Significant efforts to detail genomes of filamentous pathogens, to understand variation within species, and to a lesser extent to examine epigenetic modifications, have increased our understanding of genome function in this important group of organisms [75,107,116]. This is required to broaden our understanding of eukaryotic genome function and in order to combat emerging pathogens. Here we present a detailed analysis of the epigenome and physical DNA accessibility of the vascular wilt pathogen *V. dahliae* and link these analyses to previous characterizations of genomic regions contributing to host colonization and adaptation [59,62,74,75]. A clear picture emerges in which the core genome is organized into heterochromatic and euchromatic regions. The heterochromatin is characterized by a high density of TEs with low GC content, high levels of DNA and H3K9 methylation, low DNA accessibility and clear signatures of RIP mutations at repetitive sequences. The euchromatin regions are opposite in all characteristics, and this collective description is consistent with previous research in many other eukaryotic genomes [87,117,118]. Interestingly, we provide evidence that previously defined LS regions of the genome, characterized for their role in contributing to host infection, exist in a unique chromatin state. LS regions have higher TE density than the euchromatic regions, yet are devoid of DNA and H3K9 methylation. Furthermore, LS regions have higher DNA accessibility than the core heterochromatic regions and are more transcriptionally active, but they are less accessible than the 'true' euchromatic gene-rich core regions. Notably, LS regions are characterized as having a strong association with H3K27me3, similar to the discovery that SM gene clusters are enriched at H3K27me3 regions in *F. graminearum* [49]. Our results demonstrate that LS regions are by definition not heterochromatic, as they are far more accessible than the true heterochromatin, and yet they typically contain many heterochromatin features. These observations are akin to previous descriptions of facultative heterochromatin [119], but to our knowledge few studies have linked unique chromatin states to adaptive genomic regions. We believe this observation presents new hypotheses into the occurrence, formation and maintenance of adaptive regions of the genome. How chromatin interacts with evolutionary forces to shape organism fitness is an important question in the pursuit of understanding the genome.

Our results support the hypothesis that chromatin structure underlies genome function. More specifically that chromatin modifications and DNA accessibility may influence genome evolution, not just via transcriptional control but also regarding the architecture of the genome [104]. Based on the described associations, we were able to predict LS regions using machine learning. The ML algorithms, trained on H3K9 and H3K27 methylation, RNA-sequencing, TE density and DNA accessibility data, shows how these descriptions of chromatin can be used to classify DNA segments as core versus LS with high recall (i.e. sensitivity). The precision assessments of the algorithms were low, which could be due to previous miss-classification of LS and core regions. We could validate the improvement of our new classification because it doubled the amount of LS classified DNA while still maintaining an enrichment for *in planta* induced genes, presumed effectors, genes coding secreted peptides and genes with higher PAV in the species. It is a remarkable finding that, through the use of machine learning, we could significantly extend our knowledge of the *V. dahliae* genome and identify as of yet unconsidered genomic regions and genes which likely contribute to adaptive traits. In addition to the supervised binary classification, we employed unsupervised uniform manifold approximation and projection (UMAP) to cluster the genome, without any explicit information pertaining to core or LS regions. The UMAP approach reduced our multidimensional representation of the genomic data into a simple two-dimension scatterplot, where TEs and genes with a similar empirical profile are in close proximity. Analyzing the updated LS and core classifications with respect to the UMAP clustering shows again the enrichment and strong association between the physical characterization of the genome (i.e. DNA sequence, histone modifications, DNA accessibility) and its evolutionary function.

It is currently not possible to extend our machine learning predictions to additional filamentous pathogen genomes, as the necessary data are not currently publicly available. However, for many filamentous plant pathogens it is clear that genome variation on multiple scales, from SNPs to large structural variation, are not uniformly distributed in the genome [49]. Recent reports from the fungal pathogen *Z. tritici* addressed the role of genome stability and H3K27me3 during asexual reproduction [54,120]. During experimental evolution, individual strains of *Z. tritici* readily lose accessory chromosomes. The authors observed that a mutant lacking the enzyme responsible for H3K27me3 showed less accessory chromosome loss and concluded that H3K27me3 destabilizes chromosome structure [54]. However, accessory chromosome losses were clearly biased in their individual frequency and changes were not reported for core chromosomes, despite H3K27me3 being found at high levels on accessory and regions of core chromosomes [121]. Therefore, the observed genome destabilization requires additional determinants in conjunction with H3K27me3 which remain to be discovered. Results presented here suggest that DNA and histone methylation marks, along with physical DNA accessibility are important additional determinants to distinguish accessory and LS regions of the genome. However, we acknowledge that our model does not strictly differentiate all LS region in the *V. dahliae* genome, as there are LS and core regions that have very similar overall chromatin profiles, and therefore these features alone are not sufficient. One factor that could explain part of this discrepancy is that LS formation is likely not fully deterministic. Evolution is a stochastic process, and it seems unlikely that LS formation can be described in absolute terms. Rather, it is more likely to be a probabilistic process, in which specific chromatin and physical status increases the likelihood for formation and maintenance of LS regions. The results presented here offer an exciting new link between the epigenome, physical DNA accessibility and adaptive genome evolution.

Materials and methods

Fungal growth and strain construction

V. dahliae strain JR2 (CBS 143773) was used for experimental analysis [122]. The strain was maintained on potato dextrose agar (PDA) (Oxoid, Thermo Scientific, CM0139) and grown at 22^{0}C in the dark. For liquid grown cultures, conidiospores were collected from PDA plates after approximately two weeks and inoculated into flasks containing the desired media at a concentration of 2x10^4 spores per mL. Media used in this study include PDA, half-strength Murashige and Skoog plus vitamins (HMS) (Duchefa-Biochemie, Haarlem, The Netherlands) medium supplemented with 3% sucrose and xylem sap (abbreviated, X) collected from greenhouse grown tomato plants of the cultivar Moneymaker. Liquid cultures were grown for four days in the dark at 22^{0}C and 160 RPM. The cultures were strained through miracloth (22 µm) (EMD Millipore, Darmstadt, Germany), pressed to remove liquid, flash frozen in liquid nitrogen and ground to powder with a mortar and pestle. Samples were stored at -80^{0}C if required prior to nucleic acid extraction.

The Δ*hp1* strain was constructed as previously described [123]. Briefly, the genomic DNA regions flanking the 5' and 3' HP1 coding sequence were amplified (*left border*, For. Primer, 5'-GGTCTTAAUGACCTGAAGAATCGAGCAAGGA and
Rev. primer, 5'-GGCATTAAUATGAAAGCACCGGGATTTTTCT; *right border*,
For. Primer, 5'-GGACTTAAUATGCTGTTGGGAGGCAGAATAA
Rev. primer, 5'-GGGTTTAAUCCACGTAGATGGAGGGGTAGA). The PCR products were cloned in to the pRF-HU2 vector system [124] using USER enzyme following manufactured protocol (New England Biolabs, MA, USA). Correctly ligated vector was transformed into *Agrobacterium tumefaciens* strain AGL1 used for *V. dahliae* spore transformation [123]. Colonies of *V. dahliae* growing on hygromycin B selection after 5 days were moved to individual plates containing PDA and hygromycin B. Putative transformants were screened using PCR to check for deletion of the HP1 sequence (For. Primer, 5'- AATCCCGCAAGGGAAAAGAGAC and Rev. primer, 5'-CGTGTGCTTTGTCTTCTGACCA) and the integration of the hygromycin B sequence (For. Primer, 5'- TGGAATATGCCACCAGCAGTAG and Rev. primer, 5'- GGAGTCGCATAAGGGAGAGCG) at the specific locus.

Bisulfite sequencing and analysis

The wild-type *V. dahliae* strain and Δ*hp1* were grown in liquid PDA for three days, flash frozen and collected as described earlier. DNA was extracted from a single sample of wild-type strain JR2 and Δ*hp1* and sent to the Beijing Genome Institute (BGI) for bisulfite conversion, library construction and Illumina sequencing. Briefly, the DNA was sonicated to a fragment range of 100-300 bp, end-repaired and methylated sequencing adapters were ligated to 3' ends. The EZ DNA Methylation-Gold kit (Zymo Research, CA, USA) was followed according to manufacturer guidelines for bisulfite conversion of non-methylated DNA. Libraries were paired-end 100bp sequenced on an Illumina HiSeq 2000.

Whole-genome bisulfite sequencing reads were analyzed using the BSMAP pipeline (v. 2.73) and methratio script [125]. The results were partitioned into CG, CHG and CHH cytosine sites for analysis. Only cytosine positions containing greater than 4 sequencing reads were included for analysis. Methylation levels were summarized as weighted methylation percentage, calculated as the number of reads supporting methylation over the number of cytosines sequenced or as fractional methylation, calculated as the number of methylated cytosines divided by all cytosine positions [126]. For fractional methylation, a cytosine was considered methylated if it was at least 5% methylated from all the reads covering that cytosine. As such, weighted methylation captures quantitative aspects of methylation, while fractional methylation is more qualitative. Weighted and fractional methylation were calculated over intervals described in the text, including genes, promoters (defined as the 300 bp upstream of the translation start site), transposable elements and 10 kb windows. For each feature, weighted and fractional methylation were calculated from the sum of the mapped reads or the sum of the positions, respectively, over the analyzed region. Two sample comparisons were computed using base R [127] to compute the non-parametric Mann-Whitney U test (equivalent to the two-sample Wilcoxon rank-sum test) and Holm multiple testing correction was used for the associated p-values. Principle component analyses were computed in R using the packages FactoMineR (v 1.42) [128] and factoextra (v 1.0.5) [129].

Transposable element annotation

Repetitive elements were identified in the *V. dahliae* stains JR2 genome assembly [122] as well as in three other high-quality *V. dahliae* genome assemblies [75] using a combination of LTRharvest [130] and LTRdigest [131] followed by *de novo* identification of RepeatModeler [132]. Briefly, LTR sequences were identified (recent and ancient LTR insertions) and subsequently filtered, e.g. for occurrence of primer binding sites or for nested insertions (see procedure outlined by Campbell and colleagues for details [133]). Prior to the *de novo* prediction with RepeatModeler, genome-wide occurrences of the identified LTR elements are masked. Predicted LTR elements and the *de novo* predictions from RepeatModeler were subsequently combined, and the identified repeat sequences of the four *V. dahliae* strains were clustered using vsearch (80% sequence identity, search on both strands; v 2.4.4) [134]. A non-redundant *V. dahliae* repeat library that contained consensus sequences for each cluster (i.e. repeat family) was constructed by performing multiple sequence alignments using MAFFT (v7.271) [135] followed by the construction of a consensus sequence as described by Faino et al. [74]. The consensus repeat library was subsequently manually curated and annotated (Wicker classification [136]) using PASTEC (default databases and settings; search in the reverse-complement sequence enabled) [137], which is part of the REPET pipeline (v2.2) [138], and similarity to previously identified repetitive elements in *V. dahliae* [122,139]. The occurrence and location of repeats in the genome assembly of *V. dahliae* strain JR2 were determined using RepeatMasker (v 4.0.7; sensitive option). The RepeatMasker output was post-processed using 'One code to find them all' [140] which supports the identification and combination of multiple matches (for instance due to deletions or insertions) into combined, representative repeat occurrences. We only further considered matches to the repeat consensus library, and thereby excluded simple repeats and low-complexity regions. To estimate divergence time of TEs, each individual copy of a transposable element was aligned to the consensus of its family using needle, which is part of the EMBOSS package [141]. Sequence divergence between the TEs and the TE-family consensus was corrected using the Jukes-Cantor distance, with a correction term that accounts for insertions and deletions [142,143]. The composite RIP index (CRI) was calculated as previously described [97]. Briefly, CRI was determined by subtracting the RIP substrate from the RIP product index, which are defined by dinucleotide frequencies as follows: RIP product index = (TpA / ApT) and the RIP substrate index = (CpA + TpG/ ApC + GpT). Positive CRI values indicate the analyzed sequences were subjected to the RIP process. For TE analysis, elements that are less than 100 bp were removed.

RNA-sequencing and analysis

V. dahliae strain JR2 (CBS 143773) was grown in triplicate liquid media PDB, HMS and xylem sap as described. RNA extraction was carried out using TRIzol (Thermo Fisher Science, Waltham, MA, USA) following manufacturer guidelines. Following RNA re-suspension, contaminating DNA was removed using the TURBO DNA-free kit (Ambion, Thermo Fisher Science, Waltham, MA, USA) and RNA integrity was estimated by separating 2 µL of each sample on a 2% agarose gel and quantified using a Nanodrop (Thermo Fisher Science, Waltham, MA, USA) and stored at -80^{0}C. Library preparation and sequencing was carried out at BGI. Briefly, mRNA was enriched based on polyadenylation purification and random hexamers were used for cDNA synthesis. RNA-sequencing libraries were constructed following end-repair and adapter ligation protocols and PCR amplified. Purified DNA fragments were single-end 50bp sequenced on an Illumina HiSeq 2000.

Reads were mapped to the *V. dahliae* stain JR2 genome assembly [122] using STAR (v 2.6.0) with settings *(--sjdbGTFfeatureExon exon, --sjdbGTFtagExonParentTranscript Parent, --alignIntronMax 400, --outFilterMismatchNmax 5, --outFilterIntronMotifs RemoveNoncanonical*) [144]. Mapped reads were quantified using the *summarizeOverlaps* [145] and variance stabilizing transformation (*vst*) features of DESeq2 [146]. For TE analysis, the coordinates of the annotated TEs were used as features for read counting. To perform RNAseq analysis over whole genome 10 kb regions, raw mapped reads were summed over 10 kb bins using bedtools (v 2.27) [147] and converted to Transcripts Per Million (TPM) [148] and averaged over the three reps for analysis.

Chromatin immunoprecipitation and sequencing and analysis

V. dahliae strain JR2 was grown in PDB and samples collected as described. Approximately 400 mg ground material was resuspended in 4 ml ChIP Lysis buffer (50 mM HEPES-KOH pH7.5, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% NaDOC) and dounced 40 times in a 10 cm^3 glass tube with tight fitting pestle on 800 power with a RZR50 homogenizer (Heidolph, Schwabach, Germany), followed by five rounds of 20 seconds sonication on ice with 40 seconds rest between rounds with a Soniprep 150 (MSE, London, UK). Samples were redistributed to 2 ml tubes and pelleted for 2 min at max speed in tabletop centrifuge. The supernatants were combined, together with 20 µl 1M CaCl2 and 2.5µl MNase, and after 10 minutes of incubation in a 37˚C water bath with regular manual shaking, 80 µl 0.5M EGTA was added and tubes were put on ice. Samples were pre-cleared by adding 40 µl Protein A Magnetic Beads (New England Biolabs, MA, United States) and rotating at 4˚C for 60 min, after which the beads were captured, 1 ml fractions of supernatant were moved to new 2 ml tubes containing 5 µl H3K9me3 or H3K27me3 antibody (ActiveMotif ; #39765 and #39155) respectively and incubated overnight with continuous rotation at 4ºC. Subsequently, 20 µl protein-A magnetic beads were added and incubated for 3 hours at 4ºC, after which the beads were captured on a magnetic stand and subsequently washed with 1 ml wash buffer (50 mM Tris HCl pH 8, 1 mM EDTA, 1% Triton X-100, 100 mM NaCl), high-salt wash buffer (50 mM Tris HCl pH 8, 1 mM EDTA, 1% Triton X-100, 350 mM NaCl), LiCl wash buffer (10 mM Tris HCl pH8, 1 mM EDTA, 0.5% Triton X-100, 250 mM LiCl), TE buffer (10 mM Tris HCl pH 8, 1mM EDTA). Nucleosomes were eluted twice from beads by addition of 100µl pre-heated TES buffer (100 mM Tris HCl pH 8, 1% SDS, 10 mM EDTA, 50 mM NaCl) and 10 minutes incubation at 65ºC. 10mg /ml 2µl Proteinase K (10mg /ml) was added and incubated at 65ºC for 3 hours, followed by chloroform clean-up. DNA was precipitated by addition of 2 volumes 100% ethanol, 1/10th volume 3 M NaOAc pH 5.2 and 1/200th volume 120 mg/ml glycogen, and incubated overnight at -20ºC. Two ChIP replicates were performed for each antibody from independently grown samples. Sequencing libraries were prepared using the TruSeq ChIP Library Preparation Kit (Illumina) according to instructions, but without gel purification and with use of the Velocity DNA Polymerase (BioLine, Luckenwalde, Germany) for 25 cycles of amplification. Single-end 125bp sequencing was performed on the Illumina HiSeq2500 platform at KeyGene N.V. (Wageningen, the Netherlands).

Reads were mapped to the reference JR2 genome, using BWA-mem with default settings [149]. For ChIP and ATAC-seq mapping, three regions of the genome were masked due to aberrant mapping, possibly owing to sequence similarity to the mitochondrial genome (chr1:1-45000, chr2:3466000-3475000, chr3:1-4200). This is similar to what is described as blacklisted regions in other eukaryotic genomes [150]. The raw mapped reads were counted either over the TE coordinates or 10 kb intervals for the two separate analyses. The raw mapped reads were converted to TPM and the average of the two ChIP-seq replicates were used for analysis.

Assay for Transposase-Accessible Chromatin (ATAC)-sequencing and analysis

The *V. dahliae* strain JR2 was grown in PDB liquid media as described. Mycelium was collected, filtered, rinsed and flash frozen in liquid nitrogen. The ATAC-seq procedure was carried out mainly as described previously [151]. Nuclei were collected by resuspending ground mycelium in 5 mL of ice-cold Nuclei Isolation Buffer (NIB) (100 mM NaCl, 4mM $NaHSO_4$, 25mM Tris-HCl, 10mM $MgSO_4$, 0.5mM EDTA, 0.5% NP-40 including protease inhibitors added at time of extraction, 2 mM Phenylmethanesulfonyl fluoride (PMSF), 100 µM Leupeptin, 1 µg/mL Pepstatin, 10 µM E-64). The homogenate was layered onto 10-mL of an ice-cold sucrose-Ficoll gradient (bottom layer 5mL of 2.5M sucrose in 25mM Tris-HCl, 5mL 40% Ficoll 400 (GE Biosciences Corporation, NJ, USA)). Nuclei were separated into the lower phase by centrifugation at 2000g for 30 min at 4ºC. The upper layer was discarded and the lower phase (~4mL) moved to another collection tube containing 5mL of ice-cold NIB. Nuclei were pelleted at 9000g for 15 min at 4ºC and re-suspended in 3 mL of NIB. The integrity of the nuclei and their concentration in the solution were estimated by DAPI staining (DAPI Dilactate 5mg/mL, used at a 1/2000 dilution for visualization) and counted on a hemocytometer. A total of 200,000 nuclei were transferred to a 1.5mL microfuge tube, and nuclei pelleted at 13000g for 15 min at 4ºC and resuspended in the transposition reaction (20uL of 2x Nextera reaction buffer, 0.5uL of Nextera Tn5 Transposase, 19.5 uL of nuclease-free H_20) (Illumina, Nextera DNA library Preparation kit FA-121-1030) and the reaction was carried out for 5 minutes at 37ºC. Empirical testing showed this Tn5 incubation period and nuclei density resulted in optimal DNA fragmentation, and a single sample was used for further library preparation and sequencing. The reaction was halted, and fragmented DNA purified using a MinElute PCR purification kit (Qiagen, MD, USA). The eluted DNA was amplified in reaction buffer (10uL of transposased DNA, 10uL nuclease-free H_20, 2.5uL forward PCR primer (5'-AATGATACGGCGACCACCGAGATCTACACTCGTCGGCAGCGTCAGATGTG), 2.5uL reverse PCR primer (CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTCTCGTGGGCTCGGAGATGT) and 25uL NEBnext High-Fidelity 2x PCR Master Mix (New England Biolabs, MA, United States)) using thermo-cycler conditions described in [151] for a total of 9 cycles. Amplified library was purified using the MinElute PCR Purification Kit (Qiagen, MD, USA) and paired-end 100 bp sequenced on an Illumina HiSeq4000.

Reads were mapped to the reference JR2 genome with the described blacklisted regions masked, using BWA-mem with default settings [152]. The mapped reads were further processed to remove duplicates reads arising from library prep and sequencing using Picard toolkit *markDuplicates*. The mapped reads were counted either over the TE coordinates or 10 kb intervals for the two separate analyses using bedtools *multicov* (v 2.27) [147]. The reads were converted to TPM values and those numbers used for analysis.

Machine learning and assessment

The machine learning algorithms were implemented using the classification and regression training (caret) package in R [127,153]. The full set of genomic data was used to create a data frame comprising the genome in 10 kb segments as rows and the individual collected variables as columns. The regions were classified as core or LS based on the previous observations [74]. For initial model assessment and parameter tuning, the data were split into 80% for training and 20% used for testing (i.e. prediction), and the proportion of core and LS regions were kept approximately equal in the two splits. For parameter tuning, repeated cross-validation of 10-fold 3-times was used and the best model was selected based on accuracy. Four algorithms were used- logistic regression, random forest, stochastic gradient boosting, and boosted classification tree. The model for all algorithms was classification = ATAC-seq$_{TPM}$ + ChIP-H3K27me3$_{TPM}$ + ChIP-H3K9me3$_{TPM}$ + TE$_{density}$ + PDB-RNAseq$_{TPM}$. Logistic regression was run using method *glm*, family *binomial*. Random forest was run using method *rf* and tuneGrid [*mtry*= (1,2,3)]. The Stochastic Gradient Boosting was implemented with method *gbm* and tuneGrid [*interaction.depth*=(1,5,10), *n.trees*=(50,500,1000), *shrinkage*=(0.001, 0.01), *n.minobsinnode*=(1,5)]. The Boosted Classification Tree was implemented using method *ada* and tuneGrid [*iter*=(100, 1000, 3000), *maxdepth*=(1,5,20), *nu*=(0.01)]. Models were assessed using standard metrics for data retrieval, with receiver operating and precision-recall curves generated using package PRROC [154].

To saturate the genome in predictions, a total of 15 new training test splits (80:20) were generated, again maintaining the genome-wide proportion of core and LS regions in data set. The random forest and stochastic gradient boosting classifiers were used, based on their highest performance from the initial test. The settings were picked based on best performance from initial testing: random forest, method *rf* and tuneGrid [*mtry*=3]; stochastic gradient boosting, method *gbm* and tuneGrid [*interaction.depth*=(5), *n.trees*=(500), *shrinkage*=(0.01), *n.minobsinnode*=(5)]. The predictions for each of the 15 runs were assessed using the precision, recall and MCC metrics. For each genomic region, a consensus designation was assigned based on the highest occurrence of core versus LS prediction across the 15 trials. This was done independently between the two models. A region was finally classified as LS or core based on the majority classification across the 15 trails. For regions that had an equal number of core and LS predictions, the regions were designated as core to be conservative. A final high confidence LS consensus designation was determined for each genomic region if it was predicted LS by both models. Regions predicted LS by only one of the models were designated core. A conservative joining approach was used so that a single core region would be called LS if it were flanked by two LS regions. This added 41 genomic regions (410 kb) to the LS genome.

For assessment of old and new LS designations, the following three categories of coding sequence were assessed; *in planta* induced genes, putative effectors and coding sequences for secreted peptides. The *in planta* induced genes were determined by mapping RNA sequencing reads from *V. dahliae* colonizing *Arabidopsis thaliana* at 21 days post inoculation conducted in triplicate. Gene transcription levels *in planta* were compared to RNA-seq from *in vitro* cultivation in PDB using Kallisto quant with settings *--single -l* 50 *-s* 0.001 *–pseudobam* [155]. Differential gene expression between *A. thaliana* infection and PDB growth were determined using the DESeq2 package [146], and genes up regulated in *Arabidopsis* compared to media with an adjusted p-value < 0.05 were designated as *in planta* induced. Secreted peptides were predicted from the amino acid sequences of all annotated genes with SignalP (v5.0) [156]. Putative effectors were predicted by further analyzing the amino acid sequences of secreted peptides using EffectorP (v2.0) [157]. For each functional category, a 2x2 contingency table was created for the number of genes in the functional category by the LS or core location for both the old and new LS classification. Pearson's chi-squared test and Yate's continuity correction were used to determine if the observed values were significantly different than expected. Yate's error correction reduces the chi-square value and is therefore conservative and less prone to false significance. The chi-square analysis and expected values were calculated using base R *chisq.test* [127].

Analysis of Presence absence variation (PAV)

The LS and core designations were assessed for PAV across a collection of 42 *V. dahliae* strains (Table S1). PAV were identified using whole-genome alignments of DNA sequence reads from the 42 *V. dahliae* strains to the reference genome assembly of *V. dahliae* strain JR2 [122] using BWA-mem with default settings [149]. Library artifacts were marked and removed using Picard Tools with -*MarkDuplicates* followed by -*SortSam* to sort the reads. Raw read coverage was averaged per 100 bp non-overlapping window using the BEDtools -*multicov* function [147]. To estimate presence or absence of a window per strain, we transformed the raw read coverage value to a binary classifier where a region with >=10 reads indicate presence (1) and <10 reads indicate absence (0). For each window, the number of strains that were classified as absent were summed to get the 'absent count' value, which is easily interpretable as the number of strains for which the window was absent. To estimate absence counts for TEs and genes, the 100bp absence count windows were intersected with TEs and genes using BEDtools -*intersect* where > 50% of the 100 bp window had to overlap with the feature [147]. From these, a mean absence count was calculated per TE and gene and used for further analysis.

Uniform Manifold Approximation and Projection (UMAP) analysis

The UMAP algorithm for dimensional reduction [115] was implemented in Python3. For TE analysis, the following variables were included when running UMAP: Jukes cantor, fraction of GC content, CRI, variance stabilized transformed RNA-seq from ½ MS grown fungal culture, log2 transformed TPM ChIP-seq for H3K9me3 and H3K27me3, log2 transformed TPM for ATAC-seq and log2 transformed weighted CG DNA methylation with the following parameters *n_neighbor*=50, *n_components*=2, *min_dist*=0.1 and a *random_state*=42. For gene analysis, the following variables were included when running UMAP: fraction of GC content, log2 transformed RNA-seq from PDB grown fungal culture, log2 transformed TPM ChIP-seq for H3K9me3 and H3K27me3, log2 transformed TPM for ATAC-seq and log2 transformed weighted CG DNA methylation with the following parameters *n_neighbor*=100, *n_components*=2, *min_dist*=0.1 and a *random_state*=42. Additional values for *n_neighbor* were checked to balance local versus global clustering (data not shown). For genes, Group1 and Group2 were split based on visual assessment of the larger cluster, attempting to separate genes along what appeared as local clustering. The resulting two-dimensional values from UMAP *fit.transform* were used for plotting [158,159] and further statistical analysis [160–162]. Pairwise post hoc tests were computed using either Mann-Whitney or Conover's test, each using Holm's multiple-testing correction, implemented from *scikit-posthocs* as *posthoc_mannwhitney* and *posthoc_conover* [163].

Data availability

The sequencing data for this project are accessible from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under BioProject PRJNA592220.

Acknowledgements

This work was supported in part by a European Molecular Biology Organization postdoctoral fellowship (EMBO, ALTF 969-2013) and Human Frontier Science Program Postdoctoral Fellowship (HFSP, LT000627/2014-L) to D.E.C. A portion of the work was also carried out in the laboratory of D.E.C. under USDA-NIFA-PBI grant 2018-67013-28492. Work in the laboratories of M.F.S and B.P.H.J.T. is supported by the Research Council Earth and Life Sciences (ALW) of the Netherlands Organization of Scientific Research (NWO). B.P.H.J.T acknowledges support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2048/1 – Project ID: 390686111. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Supplementary data

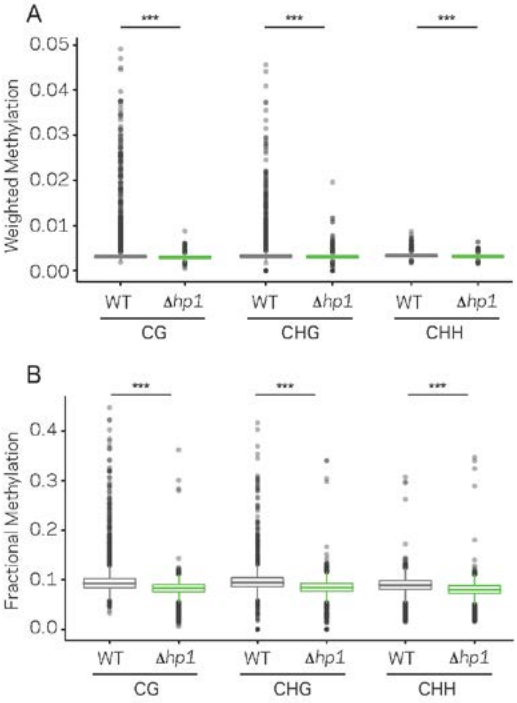


FIGURE S1-1 | Genome-wide cytosine methylation in Wild-type and Δ*hp1*. (A) Cytosine methylation was calculated using weighted methylation (see methods) in the CG, CHG and CHH sequence context in both wild-type (WT) and Δ*hp1*. Methylation levels were determined to be significantly higher in WT using the Mann-Whitney U-test. The symbol (***) indicates p < 2.2e-16. (B) Similar to (A), but the genome wide methylation level was calculated using fractional methylation. All data were summarized in 10 kb bins.

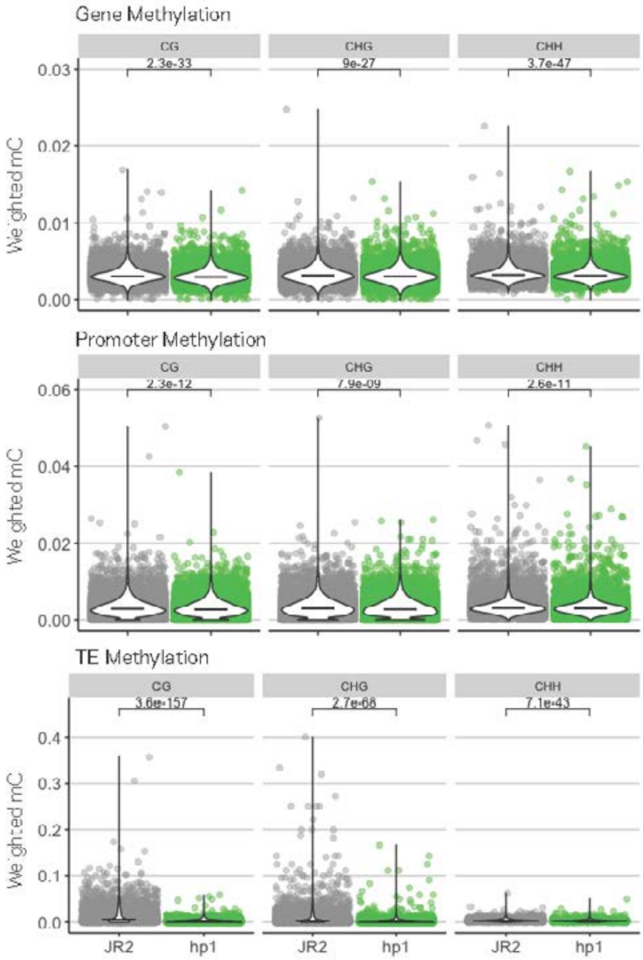


FIGURE S1-2 | Cytosine methylation for functional elements in Wild-type and Δ*hp1*. Cytosine methylation was calculated using weighted methylation (see methods) in the CG, CHG and CHH sequence context in both wild-type (JR2) and Δ*hp1*. DNA methylation was summarized over genes, promoters and TEs as labeled. The individual elements are shown as colored points, along with a violin plot showing the distribution and median as black line. Methylation levels were determined to be significantly higher in WT, Mann-Whitney U-test with Holm multiple testing correction. Associated p-values are shown

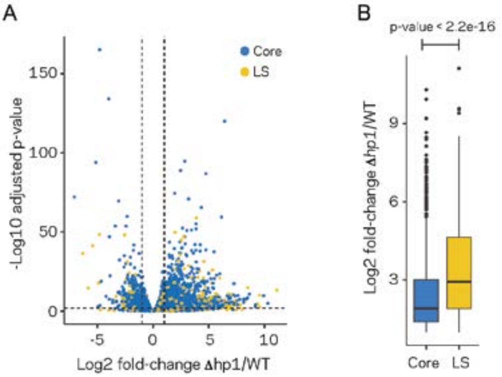


FIGURE S1-3 | Transcriptional impact of Δ*hp1*. (A) Volcano plot showing the log2 fold-change for Δ*hp1* compared to the wild-type (WT) grown in PDB culture. The adjusted p-value (-log10) is shown in the y-axis to indicate statistical significance. Individual genes are shown as colored points, with genes in the core (blue) and those in LS (yellow) regions. Genes were considered differentially expressed if they were log2 fold-change < -1 or > 1, shown as vertical dashed lines, and an adjusted p-value < 0.01, shown as horizontal dashed line. These cut-offs resulted in 1522 genes more highly expressed in Δhp1, and 587 more highly expressed in wild-type. (B) Bar plot showing the average and range of log2 fold-change values for genes (n=1522) expressed significantly higher in Δhp1 compared to wild-type from (A). The genes were grouped based on core (blue) vs LS (yellow) location. These groups were statistically different based on Mann-Whitney U-test, p-value < 2.2e-16

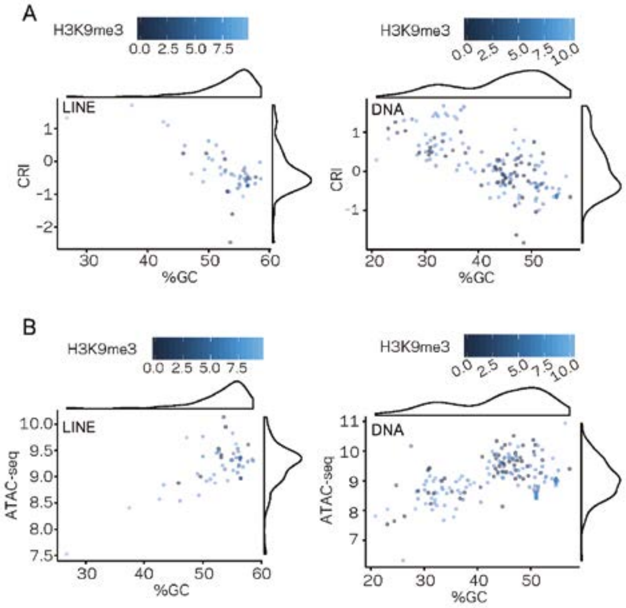


FIGURE S2-1 | Genomic distribution of DNA characteristics by TE classes. (A) scatter plot for %GC sequence content versus CRI values for individual TEs shown as points, separated by TE type, LINE and DNA, labeled in the top left of each plot. Each point is colored according to TPM values from H3K9me3 ChIP-seq, scale shown above each plot. A density plot is shown opposite to each respective labeled axis. (B) Similar to (A), but the y-axis is showing the log2 TPM values from ATAC-seq. %GC, The percent GC sequence content; CRI, Composite RIP index; ATAC-seq, Log2 of (TPM+1) values of mapped reads from Assay for Transposase Accessible Chromatin.

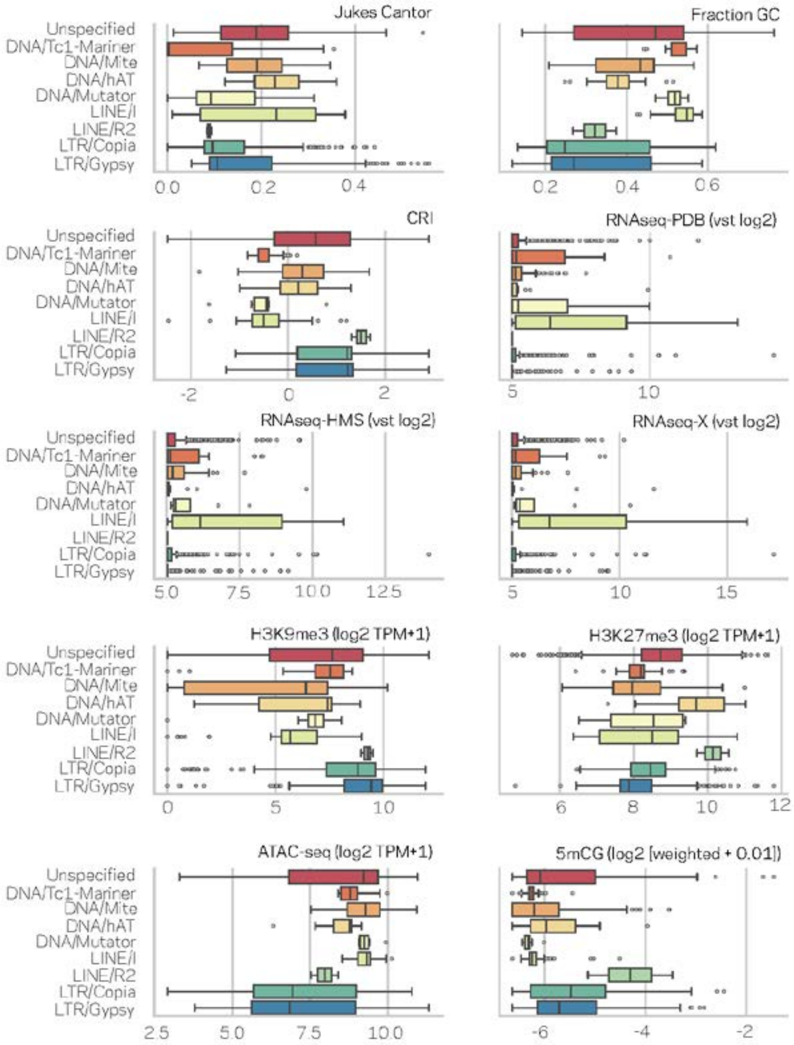


FIGURE S2-2 | Characterization of TEs in nine sub-classes across genomic variables. Each data type is shown in the upper right corner of the individual box plots. Outliers are shown as individual points. The nine subclasses of TEs are listed to the left of each figure. Test for significant differences between means of the nine subclasses per data type are shown in Table S4 (Kruskal-Wallis test) and significance of individual pair-wise differences are shown in Table S5 (Conover test and BH correction).

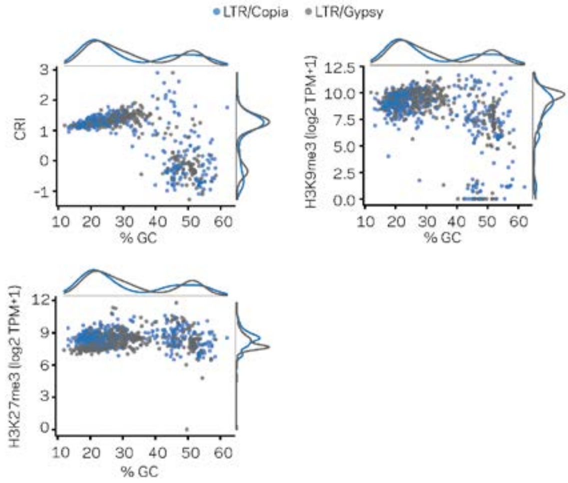


FIGURE S2-3 | The LTR subclass distinction does not account for the bimodal distribution of LTR elements in the genome. The same TEs are shown in three separate scatter plots with marginal densities. Individual Copia elements are shown as blue points, and Gypsy elements as grey points. All plots are related to those shown in Fig. 2, but only for the LTR elements. All plots show the % GC variable in the x-axis and different y-axis variables for reference. Clustered patterns of points are not simply accounted for by the two sub-classes of LTR elements.

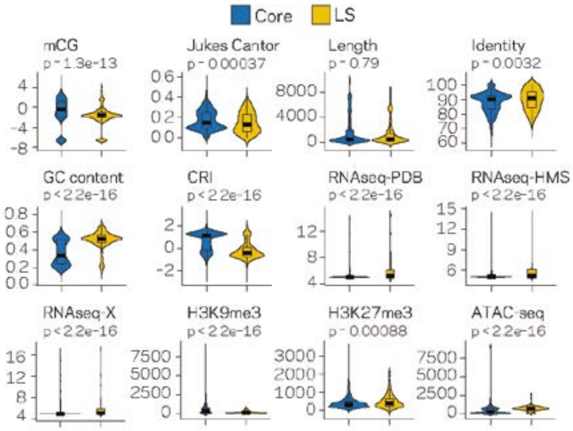


FIGURE S3-1 | Violin plots for twelve measured variables collected for the TEs located in either the core (blue) or LS (yellow) regions of the genome. Violin plots show the distribution of the values for each category, along with a box plot showing the mean (thick black line) 1st and 3rd quartiles, and whiskers extending to the furthest data point within 1.5 of the interquartile range. Differences between the core and LS values were measured using the non-parametric Mann-Whitney test and p-values adjusted using the Holm method. Adjusted p-values are shown above ach plot. mCG - Log2 weighted cytosine DNA methylation for CG; Jukes Cantor - estimate of sequence divergence from a consensus element; Length- element length in base pairs; identity - The percent identity of the elements to a family consensus; GC content - The fraction of GC sequence content; CRI - Composite RIP index; RNAseq - variance stabilizing transformed log2 RNA-sequencing reads from Potato Dextrose Broth (PDB), half-Murashige and Skoog (HMS) or Tomato Xylem (X) grown fungi; H3K9me3 and H3K27me3 – TPM values of mapped reads from ChIP sequencing using antibodies against the respective histone modifications; ATAC-seq – TPM values of mapped reads from Assay for Transposase Accessible Chromatin.

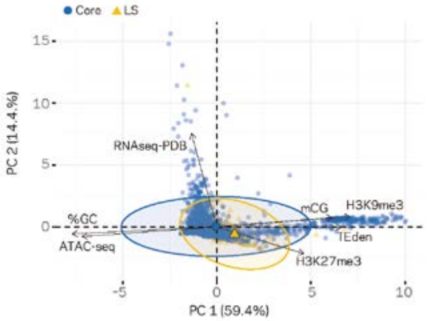


FIGURE S4-1 | Principle component analysis for seven variables genome wide at 10 kb windows. Each genomic window is shown as a point on the graph, with the windows in the core colored as blue circles and LS as yellow triangles. Colored ellipses show the confidence interval for the core and LS elements with a single large symbol to show the mean. The amount of variation for cytosine methylation for CG; %GC - The percent GC sequence content; RNAseq-PDB – TPM for RNA-sequencing reads from Potato Dextrose Broth (PDB); H3K9me3 and H3K27me3 – TPM values of mapped reads from ChIP sequencing using anti-bodies against the respective histone modifications; ATAC-seq – TPM values of mapped reads from Assay for Transposase Accessible Chromatin.

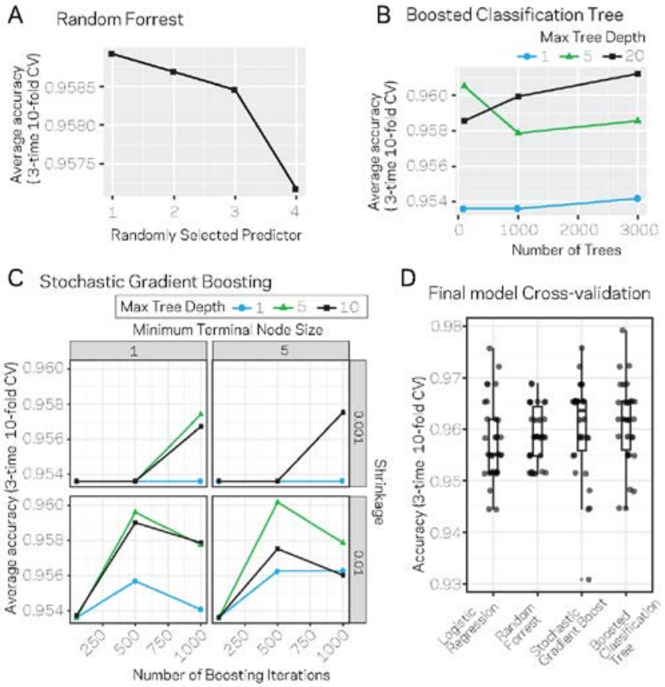


FIGURE S5-1 | Results from model parameter tuning and assessment. (A) The random forrest model was trained using 3-time 10-fold cross-validation (CV) under varying conditions for the parameter 'randomly selected predictor'. The plot shows the average accuracy across the 30 trials for each variable level as a black square. (B) Average accuracy results from 3-time 10-fold CV using the boosted classification tree algorithm. The variables 'number of trees' (x-axis) and 'max tree depth' (blue, green, black lines) were varied across the trials. Each data point represents the average accuracy across the CV. (C) Average accuracy results from 3-time 10-fold CV using the stochastic gradient boosting algorithm. The variables ‘number of boosting iterations’ (x-axis), ‘shrinkage’ (y-axis), ‘minimum terminal node size’ (columns), and max tree depth (blue, green, black lines) were varied across the trials. Each data point represents the average accuracy across the CV. (D) The individual accuracy measurements and box plot for the final models picked for each algorithm. Results are from the 30 CV runs.

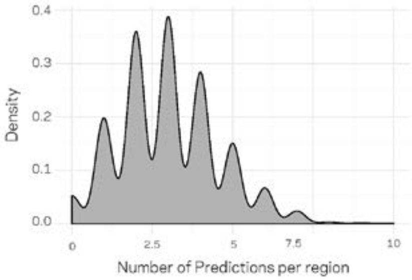


FIGURE S6-1 | Density plot for the number of distribution of predictions per genomic region. The genomic data were compiled into 3611 10 kb windows. For machine learning training and testing (related to Fig. 6), only 20% of the data could be used for prediction. To generate predictions genome wide, we randomly and independently split the data into training and testing (80:20) and generated predictions. Therefore, each region could have received more than one prediction. The above distribution profile shows that a majority of the regions received three predictions, with a large proportion of the data having received between 2 and 4 predictions. Only 124 regions received no prediction by chance. For each split, we ensured that the population distributions of ~20:1 (core:LS) was maintained in the training and testing data.

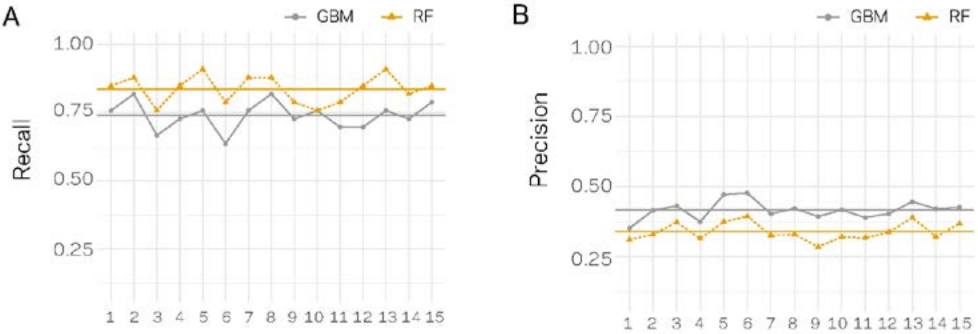


FIGURE S6-2 | Recall and Precision assessment for independent classification trials. For each trial, the data set were split 80:20, training and testing, 15 independent times. For each data split, the model was trained and tested and the performance was assessed using Recall (A) and Precision (B). the x-axes show the data split trial. Results for each trial are shown as an orange triangle connected with a dashed line for Random Forrest (RF) based classification and a grey point for Stochastic Gradient Boosting (GBM). The mean across the 15 trials is shown by a solid horizontal line of the respective color.

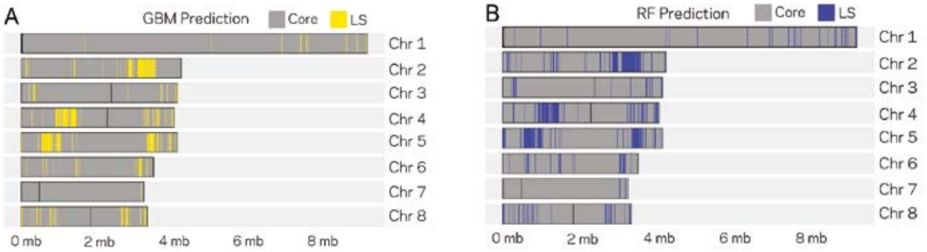


FIGURE S6-3 | Genomic location of the LS predictions from two ML models. The eight chromosomes of *V. dahliae* are labeled at the right (Chr. X) along with the physical DNA size indicated at the bottom. A) GBM model predictions for 10 kb windows as either core or LS regions are shown in grey and yellow respectively. The GBM model predicted a total of 285 LS regions. B) RF model predictions for 10 kb windows as either core and LS regions are shown in grey and blue respectively. The RF model predicted a total of 388 LS regions.

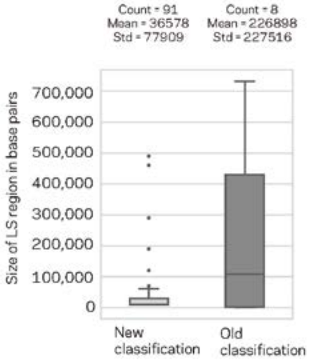


FIGURE S6-4 | Size distribution and summary description of the New and Old LS classifications. Box plots of the LS region sizes for the New classification based on model consensus and the previous LS classification. The number of regions, their mean and standard deviation (Std) are shown above the respective box plots. The means were not statistically significantly different, Mann-Whitney U-Test, p-value=0.93.

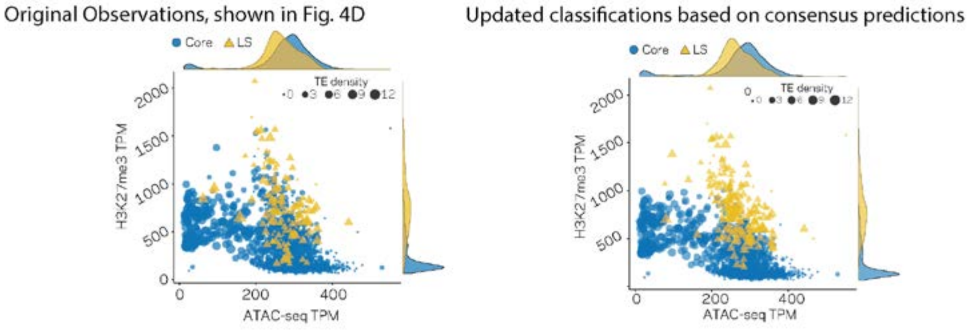


FIGURE S6-5 | Genome model of core and LS regions defined by epigenetics and chromatin status. (Top) The genome of V. dahliae was split into 10 kb windows, and labeled as core or LS based on previous observations, shown in Fig. 4D, re-shown here for comparison. (Bottom) Same 10 kb genomic windows and data, but the regions are now defined as core and LS based on the consensus machine learning predictions. The core regions are shown in blue as circles. LS regions shown as yellow triangles. Points are plotted according to TPM ATAC-seq signal (x-axis) and H3K27me3 ChIP-seq TPM (y-axis). The size of each point is proportional to the number of TEs in the 10 kb windows, shown as TE density. The marginal plots are shown opposite of the respective axis.

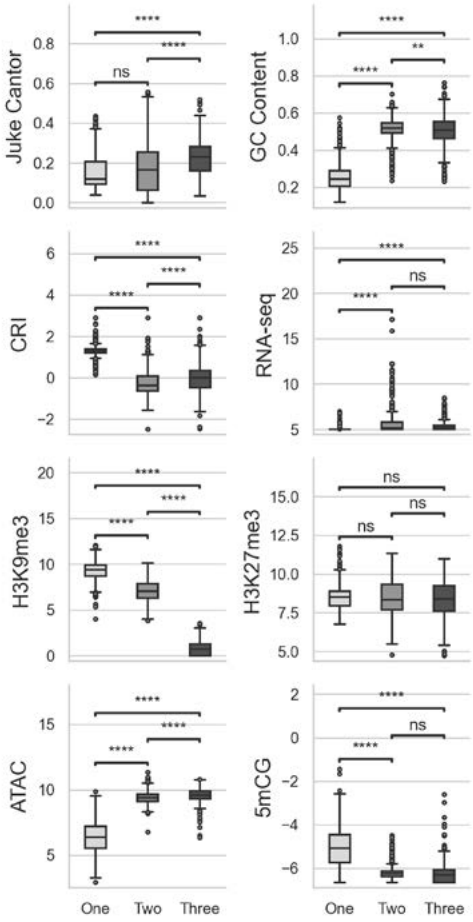


FIGURE S7-1 | Multiple comparisons of TEs in UMAP groups for genomic variables. GC content, GC sequence fraction; Jukes Cantor, corrected Jukes Cantor distance of TE comparisons; CRI, Composite RIP index; RNAseq, variance stabilizing transformed log2 RNA-sequencing reads from Xylem-media grown fungus; H3K9me3 and H3K27me3 and ATAC-seq, TPM values of mapped reads from H3K9me3 ChIP-seq, H3K27me3 ChIP-seq, or Assay for Transposase Accessible Chromatin respectively; 5mCG, log2 weighted cytosine DNA methylation+0.01 for CG. Pairwise comparisons were performed using Conover's test, with Holm multiple testing correction. **, p-value < 0.01; ****, p-value < 0.0001; ns, Non-significant p-value at $\alpha = 0.05$.

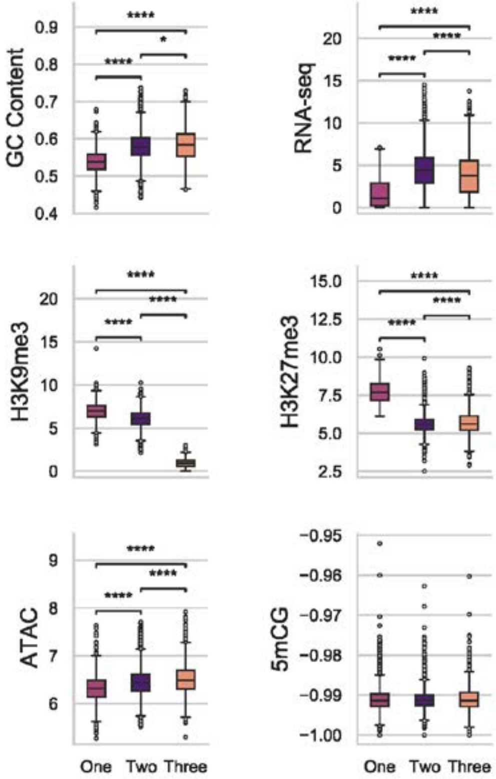


FIGURE S7-2 | Multiple comparisons of genes in UMAP groups for genomic variables. GC content, GC sequence fraction; RNAseq, variance stabilizing transformed log2 RNA-sequencing reads from PDB grown fungus; H3K9me3 and H3K27me3 and ATAC-seq, TPM values of mapped reads from H3K9me3 ChIP-seq, H3K27me3 ChIP-seq, or Assay for Transposase Accessible Chromatin respectively; 5mCG, log2 weighted cytosine DNA methylation+0.01 for CG. There were no significant differences for the comparisons of DNA methylation levels at α = 0.05. Pairwise comparisons were performed using Conover's test, with Holm multiple testing correction. *, p-value < 0.05; ****, p-value < 0.0001.

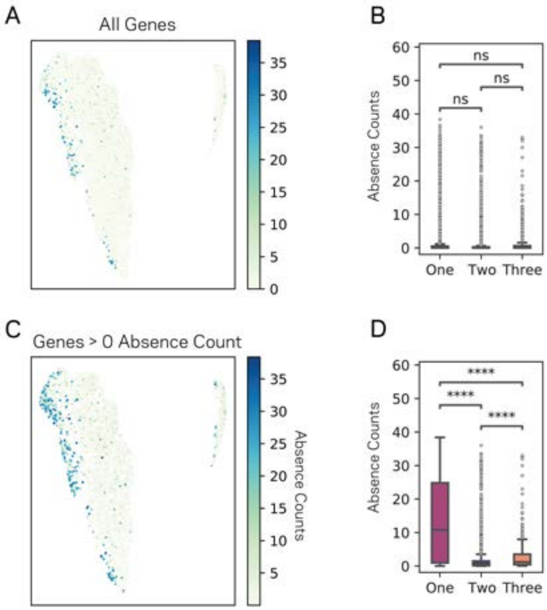


FIGURE S7-3 | UMAP groupings vary significantly for Absence across *V. dahliae* strains. (A) Scatter plot showing each 11,429 genes as a point following the UMAP results. Each gene is colored according to its absence count across 42 *V. dahliae* strains. (B) Box plot showing the distribution of gene absence counts for each of the three UMAP groups. (C) Similar plot as shown in A, but only genes that have an absence count greater than zero are plotted. (D) Similar to B, but only genes that have an absence counts greater than 0 are plotted. Pairwise comparisons were performed using Conover's test, with Holm multiple testing correction. There we 2,130, 8,140 and 1,159 genes in UMAP Group 1, 2 and 3 respectively for A and B. There were 66, 3,156 and 441 genes in UMAP Group 1, 2 and 3 respectively for C and D. ns, non-significant α = 0.05; ****, p-value < 0.0001.

TABLE 1 | Summary of Transposable elements by Family in core and LS regions.

Family	Class	Core	LS	Count (>100bp)
DNA	I	171 (0.91)	17 (0.09)	188 (0.08)
LINE	II	27 (0.50)	27 (0.50)	54 (0.03)
LTR	II	874 (0.93)	62 (0.07)	936 (0.42)
Unspecified	II	947 (0.90)	102 (0.10)	1049 (0.47)
Total		2019	208	2227

The fraction of the family total between core and LS are shown in parentheses, while the parentheses in the final column denote fraction of the total genome count.

TABLE S2 | Dunns test of pairwise differences for TE Families following Kruskal-Wallis test.

Comparison	mCG	mCHG	Juke_Gap	Length	Identity	GC%	CRI	RNAseq-PDB	RNAseq-HMS	RNAseq- X	TPM H3K9me3	TPM H3K27me3	TPM ATAC-seq
DNA - LINE	9,70E-01	8,20E-01	7,76E-02	2,81E-04	1,91E-01	5,49E-06	1,14E-02	1,57E-05	7,80E-05	2,42E-06	3,68E-01	1,63E-01	4,90E-01
DNA - LTR	2,61E-17	1,02E-01	2,16E-03	4,65E-20	2,46E-03	5,18E-23	1,85E-28	4,27E-19	1,05E-24	1,05E-18	1,76E-45	9,18E-01	1,74E-32
LINE - LTR	1,83E-06	1,94E-01	3,43E-04	2,36E-01	1,26E-03	2,05E-26	1,06E-19	1,80E-22	1,79E-24	3,69E-24	1,52E-19	1,33E-01	6,37E-14
DNA - Unsp	4,54E-06	6,87E-01	3,96E-04	3,94E-02	3,95E-01	4,11E-01	8,95E-10	2,73E-06	9,69E-10	2,69E-07	6,13E-08	3,96E-12	1,32E-02
LINE - Unsp	1,09E-02	9,76E-01	9,46E-01	1,82E-07	2,74E-01	4,59E-08	4,65E-10	1,28E-13	7,82E-15	6,99E-16	4,76E-05	3,50E-02	2,94E-02
LTR - Unsp	2,00E-12	2,30E-06	1,46E-33	1,40E-90	2,65E-12	1,78E-58	1,67E-19	1,69E-14	1,70E-14	1,66E-11	2,69E-54	4,90E-36	1,95E-61

Each cell lists the p-value for the contrast listed under comparison and the genomic variable listed as columns. The Benjamini-Hochberg (HB) multiple testing correct method was used to decrease the false discovery rate.

TABLE S3 | Summary count of TEs by sub-class.

TE Class	Class Count	TE sub-class	Sub-class Count
Unspecified	1043	Unspecified	1049
DNA	188	Mite	111
		Tc1-Mariner	54
		hAT	15
		Mutator	8
LINE	54	I	52
		R2	2
LTR	936	Gypsy	531
		Copia	405

TABLE S4 | Kruskal-Wallis test statistic for differences between TE sub-classes for genomic variables.

Genomic Variable	Kruskal-Wallis H statistic	p-value
Juke Cantor	250,51	1,34E-49
GC %	382,41	1,08E-77
CRI	259,41	1,74E-51
RNAseq-PDB	208,73	9,24E-41
RNAseq-HMS	255,54	1,15E-50
RNAseq-X	227,39	1,05E-44
H3K9me3	418,83	1,75E-85
H3K27me3	257,73	3,94E-51
ATAC	365,03	5,58E-74
5mCG	124,04	4,86E-23

TABLE S5 | P-value results from Conover test and BH multiple testing correction for genomic variables summarized over TE sub-classes.

	DNA/Mite	**DNA/Mutator**	**DNA/Tc1-Mariner**	**DNA/hAT**	**LINE/I**	**LINE/R2**	**LTR/Copia**	**LTR/Gypsy**	**Unspecified**
Jukes Cantor									
DNA/Mite		0,206199	8,13E-14	6,40E-01	1,00E+00	0,430679	2,43E-13	1,59E-05	1,00E+00
DNA/Mutator	2,06E-01		1,00E+00	3,75E-02	2,77E-01	1	1,00E+00	1,00E+00	2,48E-01
DNA/Tc1-Mariner	8,13E-14	1		3,75E-08	8,12E-10	1	8,38E-03	2,30E-07	1,74E-17
DNA/hAT	6,40E-01	0,037477	3,75E-08		6,40E-01	0,168657	1,60E-05	3,48E-03	4,01E-01
LINE/I	1,00E+00	0,277271	8,12E-10	6,40E-01		0,458269	7,64E-07	9,95E-03	1,00E+00
LINE/R2	4,31E-01	1	1,00E+00	1,69E-01	4,58E-01		1,00E+00	1,00E+00	4,58E-01
LTR/Copia	2,43E-13	1	8,38E-03	1,60E-05	7,64E-07	1		9,94E-06	9,28E-37
LTR/Gypsy	1,59E-05	1	2,30E-07	3,48E-03	9,95E-03	1	9,94E-06		2,51E-15
Unspecified	1,00E+00	0,247624	1,74E-17	4,01E-01	1,00E+00	0,458269	9,28E-37	2,51E-15	
% GC									
DNA/Mite		0,21215	3,67E-06	1,00E+00	1,48E-09	1	6,59E-08	4,54E-07	3,67E-01
DNA/Mutator	2,12E-01		1,00E+00	2,55E-01	1,00E+00	0,679528	4,80E-04	8,06E-04	4,91E-01
DNA/Tc1-Mariner	3,67E-06	1		7,83E-03	1,00E+00	0,490985	9,04E-23	6,75E-22	2,88E-05
DNA/hAT	1,00E+00	0,255386	7,83E-03		3,66E-04	1	3,67E-01	4,96E-01	1,00E+00
LINE/I	1,48E-09	1	1,00E+00	3,66E-04		0,269464	2,16E-29	1,51E-28	4,02E-09
LINE/R2	1,00E+00	0,679528	4,91E-01	1,00E+00	2,69E-01		1,00E+00	1,00E+00	1,00E+00
LTR/Copia	6,59E-08	0,00048	9,04E-23	3,67E-01	2,16E-29	1		1,00E+00	1,28E-41
LTR/Gypsy	4,54E-07	0,000806	6,75E-22	4,96E-01	1,51E-28	1	1,00E+00		1,20E-43
Unspecified	3,67E-01	0,490985	2,88E-05	1,00E+00	4,02E-09	1	1,28E-41	1,20E-43	
CRI									
	DNA/Mite	DNA/Mutator	DNA/Tc1-Mariner	DNA/hAT	LINE/I	LINE/R2	LTR/Copia	LTR/Gypsy	Unspecified
DNA/Mite		0,087138	1,79E-06	1,00E+00	1,87E-05	0,097195	1,52E-06	2,32E-10	1,77E-01
DNA/Mutator	8,71E-02		1,00E+00	4,45E-01	1,00E+00	0,007693	3,77E-04	6,70E-05	1,24E-02
DNA/Tc1-Mariner	1,79E-06	1		8,71E-02	1,00E+00	0,003025	2,06E-21	7,59E-26	1,72E-13
DNA/hAT	1,00E+00	0,444939	8,71E-02		1,53E-01	0,087138	4,26E-02	7,93E-03	7,20E-01
LINE/I	1,87E-05	1	1,00E+00	1,53E-01		0,004084	5,42E-19	3,28E-23	1,45E-11

	DNA/Mite	DNA/Mutator	DNA/Tc1-Mariner	DNA/hAT	LINE/I	LINE/R2	LTR/Copia	LTR/Gypsy	Unspecified
LINE/R2	9,72E-02	0,007693	3,02E-03	8,71E-02	4,08E-03		5,17E-01	7,20E-01	1,82E-01
LTR/Copia	1,52E-06	0,000377	2,06E-21	4,27E-02	5,42E-19	0,517235		1,93E-01	2,31E-08
LTR/Gypsy	2,32E-10	0,000067	7,59E-26	7,93E-03	3,28E-23	0,719592	1,93E-01		1,13E-18
Unspecified	1,77E-01	0,012364	1,72E-13	7,20E-01	1,45E-11	0,182472	2,31E-08	1,13E-18	
RNAseq-PDB									
DNA/Mite		1	1,00E+00	6,51E-01	3,43E-05	0,63419	1,49E-06	1,02E-14	1,80E-03
DNA/Mutator	1,00E+00		1,00E+00	4,76E-01	1,00E+00	0,392456	7,79E-02	7,15E-03	3,17E-01
DNA/Tc1-Mariner	1,00E+00	1		3,17E-01	8,89E-03	0,448094	7,51E-06	4,05E-11	1,62E-03
DNA/hAT	6,51E-01	0,475822	3,17E-01		3,58E-04	1	1,00E+00	6,47E-01	1,00E+00
LINE/I	3,43E-05	1	8,89E-03	3,58E-04		0,055853	1,35E-18	3,62E-27	7,14E-15
LINE/R2	6,34E-01	0,392456	4,48E-01	1,00E+00	5,59E-02		1,00E+00	1,00E+00	1,00E+00
LTR/Copia	1,49E-06	0,077863	7,51E-06	1,00E+00	1,35E-18	1		3,58E-04	6,35E-03
LTR/Gypsy	1,02E-14	0,00715	4,05E-11	6,47E-01	3,62E-27	1	3,58E-04		2,39E-17
Unspecified	1,80E-03	0,316775	1,62E-03	1,00E+00	7,14E-15	1	6,35E-03	2,39E-17	
RNAseq-HMS									
DNA/Mite		0,359669	1,00E+00	2,53E-02	9,14E-04	0,266296	7,19E-11	4,08E-24	5,38E-08
DNA/Mutator	3,60E-01		2,23E-01	5,43E-03	1,00E+00	0,053083	7,12E-04	7,33E-06	3,03E-03
DNA/Tc1-Mariner	1,00E+00	0,222979		1,10E-01	7,79E-04	0,370022	1,46E-04	7,19E-11	5,01E-03
DNA/hAT	2,53E-02	0,005434	1,10E-01		7,71E-06	1	1,00E+00	1,00E+00	1,00E+00
LINE/I	9,14E-04	1	7,79E-04	8,00E-06		0,027718	1,09E-19	5,09E-31	6,44E-17
LINE/R2	2,66E-01	0,053083	3,70E-01	1,00E+00	2,77E-02		1,00E+00	1,00E+00	1,00E+00
LTR/Copia	7,19E-11	0,000712	1,46E-04	1,00E+00	1,09E-19	1		6,44E-07	5,65E-02
LTR/Gypsy	4,08E-24	0,000007	7,19E-11	1,00E+00	5,09E-31	1	6,44E-07		2,88E-20
Unspecified	5,38E-08	0,003031	5,01E-03	1,00E+00	6,44E-17	1	5,65E-02	2,88E-20	
RNAseq-X									
DNA/Mite		0,144765	1,00E+00	9,96E-02	8,18E-06	0,411245	5,21E-06	5,83E-17	1,03E-04
DNA/Mutator	1,45E-01		2,34E-01	4,76E-03	1,00E+00	0,0542	1,04E-03	1,13E-05	2,84E-03

	DNA/Mite	DNA/Mutator	DNA/Tc1-Mariner	DNA/hAT	LINE/I	LINE/R2	LTR/Copia	LTR/Gypsy	Unspecified
DNA/Tc1-Mariner	1,00E+00	0,233637		9,70E-02	4,68E-04	0,381892	4,32E-04	3,39E-10	3,52E-03
DNA/hAT	9,96E-02	0,004764	9,70E-02		4,65E-06	1	1,00E+00	1,00E+00	1,00E+00
LINE/I	8,18E-06	1	4,68E-04	5,00E-06		0,024638	5,48E-19	8,46E-31	1,13E-17
LINE/R2	4,11E-01	0,0542	3,82E-01	1,00E+00	2,46E-02		1,00E+00	1,00E+00	1,00E+00
LTR/Copia	5,21E-06	0,001044	4,32E-04	1,00E+00	5,48E-19	1		4,08E-07	4,29E-01
LTR/Gypsy	5,83E-17	0,000011	3,39E-10	1,00E+00	8,46E-31	1	4,08E-07		3,27E-17
Unspecified	1,03E-04	0,002841	3,52E-03	1,00E+00	1,13E-17	1	4,29E-01	3,27E-17	
H3K9me3 (Log2 TPM +1)									
DNA/Mite		1	4,06E-02	1,00E+00	1,00E+00	0,135462	7,33E-28	7,91E-48	1,41E-09
DNA/Mutator	1,00E+00		1,00E+00	1,00E+00	1,00E+00	0,387305	2,88E-02	7,62E-04	1,00E+00
DNA/Tc1-Mariner	4,06E-02	1		1,00E+00	5,80E-02	0,602565	3,23E-06	2,22E-13	1,00E+00
DNA/hAT	1,00E+00	1	1,00E+00		1,00E+00	0,416126	3,20E-03	8,12E-06	9,16E-01
LINE/I	1,00E+00	1	5,80E-02	1,00E+00		0,123889	1,53E-16	3,18E-27	8,47E-06
LINE/R2	1,35E-01	0,387305	6,03E-01	4,16E-01	1,24E-01		1,00E+00	1,00E+00	9,05E-01
LTR/Copia	7,33E-28	0,028763	3,23E-06	3,20E-03	1,53E-16	1		1,03E-06	9,16E-21
LTR/Gypsy	7,91E-48	0,000762	2,22E-13	8,00E-06	3,18E-27	1	1,03E-06		3,57E-62
Unspecified	1,41E-09	1	1,00E+00	9,16E-01	8,47E-06	0,904847	9,16E-21	3,57E-62	
H3K27me3 (Log2 TPM +1)									
DNA/Mite		1	1,00E+00	9,04E-07	3,17E-01	0,038748	1,93E-03	1,00E+00	3,47E-12
DNA/Mutator	1,00E+00		1,00E+00	6,46E-02	1,00E+00	0,209755	1,00E+00	1,00E+00	1,00E+00
DNA/Tc1-Mariner	1,00E+00	1		9,52E-06	8,86E-01	0,049113	1,07E-01	1,00E+00	9,21E-06
DNA/hAT	9,04E-07	0,064583	1,00E-05		7,73E-04	1	4,06E-04	7,23E-08	3,63E-02
LINE/I	3,17E-01	1	8,86E-01	7,73E-04		0,122894	1,00E+00	6,46E-02	3,22E-02
LINE/R2	3,87E-02	0,209755	4,91E-02	1,00E+00	1,23E-01		1,41E-01	3,22E-02	4,09E-01
LTR/Copia	1,93E-03	1	1,07E-01	4,06E-04	1,00E+00	0,140825		1,08E-10	8,71E-08
LTR/Gypsy	1,00E+00	1	1,00E+00	7,23E-08	6,46E-02	0,032207	1,08E-10		3,90E-48
Unspecified	3,47E-12	1	9,00E-06	3,63E-02	3,22E-02	0,40895	8,71E-08	3,90E-48	

	DNA/Mite	DNA/Mutator	DNA/Tc1-Mariner	DNA/hAT	LINE/I	LINE/R2	LTR/Copia	LTR/Gypsy	Unspecified
ATAC (Log2 TPM +1)									
DNA/Mite		1	5,09E-02	9,45E-02	1,00E+00	1	9,55E-26	1,37E-32	2,98E-04
DNA/Mutator	1,00E+00		1,00E+00	8,44E-01	1,00E+00	1	1,77E-02	5,55E-03	1,00E+00
DNA/Tc1-Mariner	5,09E-02	1		1,00E+00	2,49E-01	1	2,83E-05	2,39E-07	1,00E+00
DNA/hAT	9,45E-02	0,843663	1,00E+00		2,17E-01	1	6,63E-01	2,49E-01	1,00E+00
LINE/I	1,00E+00	1	2,49E-01	2,17E-01		1	6,99E-13	4,75E-16	8,43E-02
LINE/R2	1,00E+00	1	1,00E+00	1,00E+00	1,00E+00		1,00E+00	1,00E+00	1,00E+00
LTR/Copia	9,55E-26	0,017657	2,83E-05	6,63E-01	6,99E-13	1		5,74E-01	4,56E-34
LTR/Gypsy	1,37E-32	0,005546	2,39E-07	2,49E-01	4,75E-16	1	5,74E-01		4,42E-54
Unspecified	2,98E-04	1	1,00E+00	1,00E+00	8,43E-02	1	4,56E-34	4,42E-54	
5mC (Log2 weighted +0.01)									
DNA/Mite		0,984844	4,37E-01	1,00E+00	1,00E+00	0,181967	1,28E-06	1,59E-08	3,72E-02
DNA/Mutator	9,85E-01		1,00E+00	4,10E-01	1,00E+00	0,066599	3,14E-02	1,83E-02	2,14E-01
DNA/Tc1-Mariner	4,37E-01	1		2,14E-01	1,00E+00	0,066599	3,14E-08	1,24E-09	2,82E-04
DNA/hAT	1,00E+00	0,410236	2,14E-01		6,35E-01	0,527887	1,00E+00	1,00E+00	1,00E+00
LINE/I	1,00E+00	1	1,00E+00	6,35E-01		0,120191	1,57E-05	1,16E-06	2,94E-02
LINE/R2	1,82E-01	0,066599	6,66E-02	5,28E-01	1,20E-01		7,95E-01	9,30E-01	4,10E-01
LTR/Copia	1,28E-06	0,031433	3,14E-08	1,00E+00	1,60E-05	0,795071		1,00E+00	1,53E-05
LTR/Gypsy	1,59E-08	0,01826	1,24E-09	1,00E+00	1,00E-06	0,929769	1,00E+00		1,24E-09
Unspecified	3,72E-02	0,214159	2,82E-04	1,00E+00	2,94E-02	0,410236	1,53E-05	1,24E-09	

Pairwise tests for significant differences are shown, with the TE sub-classes named by rows and columns. Each table of p-values indicate the genomic variable being tested at the top left.

TABLE S6 | Contribution of variables to the first 5 dimensions of PCA.

	Dimension1	Dimension2	Dimension3	Dimension4	Dimension5
TPM H3K9me3-ChIP	18,87	1,07	7,52	9,30	2,67
TPM H3K27me3-ChIP	8,06	7,34	74,46	0,61	7,08
TPM ATAC-seq	19,35	0,96	0,93	7,73	7,40
TPM RNA-seq PDB	0,69	89,25	9,91	0,01	0,00
mCG	14,22	0,87	4,48	79,01	0,99
TE Density	16,51	0,00	2,10	0,19	80,20
GC sequence %	22,30	0,51	0,60	3,15	1,65

TABLE S7 | Confusion Matrix results for Stochastic Gradient Boosting Machine Learning of 15 independent training-test predictions of LS and core regions.

Model1	Reference	
Prediction	Core	LS
Core	642	8
LS	46	25

Model2	Reference	
Prediction	Core	LS
Core	650	6
LS	38	27

Model3	Reference	
Prediction	Core	LS
Core	659	11
LS	29	22

Model4	Reference	
Prediction	Core	LS
Core	648	9
LS	40	24

Model5	Reference	
Prediction	Core	LS
Core	660	8
LS	28	25

Model6	Reference	
Prediction	Core	LS
Core	665	12
LS	23	21

Model9	Reference	
Prediction	Core	LS
Core	651	9
LS	37	24

Model10	Reference	
Prediction	Core	LS
Core	653	8
LS	35	25

Model11	Reference	
Prediction	Core	LS
Core	652	10
LS	36	23

Model12	Reference	
Prediction	Core	LS
Core	654	10
LS	34	23

Model13	Reference	
Prediction	Core	LS
Core	657	8
LS	31	25

Model14	Reference	
Prediction	Core	LS
Core	655	9
LS	33	24

Model7	Reference	
Prediction	Core	LS
Core	651	8
LS	37	25

Model8	Reference	
Prediction	Core	LS
Core	651	6
LS	37	27

Model15	Reference	
Prediction	Core	LS
Core	653	7
LS	35	26

TABLE S8 | Confusion Matrix results for Random Forest Machine Learning of 15 independent training-test predictions of LS and core regions.

Model1	Reference	
Prediction	Core	LS
Core	626	5
LS	62	28

Model2	Reference	
Prediction	Core	LS
Core	629	4
LS	59	29

Model3	Reference	
Prediction	Core	LS
Core	646	8
LS	42	25

Model4	Reference	
Prediction	Core	LS
Core	627	5
LS	61	28

Model5	Reference	
Prediction	Core	LS
Core	638	3
LS	50	30

Model6	Reference	
Prediction	Core	LS
Core	648	7
LS	40	26

Model9	Reference	
Prediction	Core	LS
Core	623	7
LS	65	26

Model10	Reference	
Prediction	Core	LS
Core	635	8
LS	53	25

Model11	Reference	
Prediction	Core	LS
Core	632	7
LS	56	26

Model12	Reference	
Prediction	Core	LS
Core	633	5
LS	55	28

Model13	Reference	
Prediction	Core	LS
Core	641	3
LS	47	30

Model14	Reference	
Prediction	Core	LS
Core	631	6
LS	57	27

Model7	Reference	
Prediction	Core	LS
Core	628	4
LS	60	29

Model15	Reference	
Prediction	Core	LS
Core	640	5
LS	48	28

Model8	Reference	
Prediction	Core	LS
Core	629	4
LS	59	29

TABLE S9 | Contingency tables for observed and expected LS versus core designation for *in planta* induction.

	Old LS and Core Designation		**New** LS and Core Designation	
	Not *in planta* induced	in planta induced	Not *in planta* induced	in planta induced
Core	9606 (9583)	1326 (1349)	9195 (9140)	1232 (1287)
LS	410 (433)	84 (61)	820 (874)	178 (123)

Observed values are listed in each cell, with the expected values listed in parenthesis

TABLE S10 | Contingency tables for observed and expected LS versus core designation for predicted effectors.

	Old LS and Core Designation		**New** LS and Core Designation	
	Not a predicted effector	Predicted effector	Not a predicted effector	Predicted effector
Core	10750 (10750)	182 (182)	10268 (10255)	160 (173)
LS	486 (486)	8 (8)	968 (981)	30 (17)

Observed values are listed in each cell, with the expected values listed in parenthesis

TABLE S11 | Contingency tables for observed and expected LS versus core designation for proteins with secretion signal.

	Old LS and Core Designation		**New** LS and Core Designation	
	No signalP	SignalP	No signalP	SignalP
Core	9900 (9937)	1320 (1283)	9485 (9468)	943 (960)
LS	474 (437)	20 (57)	889 (906)	109 (92)

Observed values are listed in each cell, with the expected values listed in parenthesis

Chapter 3

Local rather than global H3K27me3 dynamics associates with differential gene expression in *Verticillium dahliae*

H. Martin Kramer[1],
Michael F. Seidl[1,2],
Bart P.H.J. Thomma[1,3,#],
David E. Cook[1,4,#]

[1]Laboratory of Phytopathology, Wageningen University and Research, Droevendaalsesteeg 1, 6708 PB Wageningen, the Netherlands
[2]Theoretical Biology & Bioinformatics Group, Department of Biology, Utrecht University, Utrecht, The Netherlands
[3]University of Cologne, Institute for Plant Sciences, Cluster of Excellence on Plant Sciences (CEPLAS), 50674 Cologne, Germany
[4]Department of Plant Pathology, Kansas State University, 1712 Claflin Road, Manhattan, Kansas 66506, USA

[#]These authors contributed equally

Published in mBio: e03566 (2022)

Abstract

Differential growth conditions typically trigger global transcriptional responses in filamentous fungi. Such fungal responses to environmental cues involve epigenetic regulation, including chemical histone modifications. It has been proposed that conditionally expressed genes, such as those that encode secondary metabolites but also effectors in pathogenic species, are often associated with a specific histone modification, lysine27 methylation of H3 (H3K27me3). However, thus far no analyses on the global H3K27me3 profiles have been reported under differential growth conditions in order to assess if H3K27me3 dynamics governs differential transcriptional. Using ChIP- and RNA-sequencing data from the plant pathogenic fungus *Verticillium dahliae* grown in three *in vitro* cultivation media, we now show that a substantial number of the identified H3K27me3 domains globally display stable profiles among these growth conditions. However, we do observe local quantitative differences in H3K27me3 ChIP-seq signal that associate with a subset of differentially transcribed genes between media. Comparing the *in vitro* results to expression during plant infection suggests that *in planta*-induced genes may require chromatin remodeling to achieve expression. Overall, our results demonstrate that some loci display H3K27me3 dynamics associated with concomitant transcriptional variation, but many differentially expressed genes are associated with stable H3K27me3 domains. Thus, we conclude that while H3K27me3 is required for transcriptional repression, it does not appear that transcriptional activation requires global erasure of H3K27me3. We propose that the H3K27me3 domains that do not undergo dynamic methylation may contribute to transcription through other mechanisms or may serve additional genomic regulatory functions.

Introduction

The fungal kingdom comprises a plethora of species occupying an enormously diverse range of ecological niches [164]. As environments are typically dynamic, including the effects of daily and yearly cycles, fungi continuously need to respond and adapt to survive [165,166]. To this end, fungi have evolved various mechanisms to monitor their environment and to transcriptionally respond to environmental cues [167]. For instance, the yeast *Saccharomyces cerevisiae* senses cold stress by increased membrane rigidity, which leads to transcription of genes that, among others, encode cell damage-preventing proteins [168]. Furthermore, *S. cerevisiae* senses the quantitative availability of carbon and nitrogen sources in the environment to determine which developmental program maximizes the potential for survival [169]. In animal systems, epigenetic mechanisms (i.e. those affecting genetic output without changing the genetic sequence) are implicated in transcriptional responses to changing environments [170–172]. Such epigenetic mechanisms have similarly been proposed to contribute to environmental response in filamentous fungi. For instance, the saprotrophic fungus *Neurospora crassa* phenotypically reacts to environmental stimuli, such as changes in temperature and pH, yet *N. crassa* mutants impaired in transcription-associated epigenetic mechanisms display reduced growth in response to these stimuli [173]. Similarly, the nectar-feeding yeast *Metschnikowia reukaufii* and the ubiquitous fungus *Aureobasidium pullulans* fail to properly respond to changing carbon sources when DNA methylation or histone acetylation are inhibited [174,175]. These results suggest that epigenetic mechanisms are important for transcriptional responses to changing environments in diverse fungi, but many questions remain regarding the precise mechanisms and function of epigenetic dynamics in fungi.

Epigenetic mechanisms, such as direct modifications of DNA and histone proteins or physical changes to the chromatin architecture, can influence transcription by regulating DNA accessibility [176]. Chromatin that is accessible and potentially active is termed euchromatin, while heterochromatin is condensed and often transcriptionally silent [177]. However, heterochromatic regions are not always repressed. Heterochromatin is subcategorized into constitutive heterochromatin that remains condensed throughout the cell cycle, and facultative heterochromatin that can de-condense to allow transcription in response to developmental changes or to environmental stimuli [178–180]. In fungi, constitutive heterochromatin is often associated with repeat-rich genome regions and is typically characterized by tri-methylation of lysine 9 on histone H3 (H3K9me3), while facultative heterochromatin is characterized by tri-methylation of lysine 27 on histone H3 (H3K27me3) [97,103,181,182]. Empirically, both H3K9me3 and H3K27me3 have been implicated in transcriptional regulation in various fungi [48–54,183–186]. The majority of these studies rely on genetic perturbation of the enzymes that deposit methylation at H3K9 and H3K27, and the results consistently show that depletion of methylation at these lysine residues mainly results in transcriptional induction. However, as global depletion of a histone modification can result in pleiotropic effects, such as improper localization of other histone modifications or altered development [187,188], it is difficult to infer transcriptional control mechanisms used for natural gene regulation from these genetic perturbation experiments. Therefore, additional research is needed to directly test the hypothesis that heterochromatin-associated histone modifications directly regulate transcription either through their dynamics, or their action to form or recruit transcriptional complexes [104,189,190].

The filamentous fungus *Verticillium dahliae* is a soil-borne broad host-range plant pathogen that infects plants through the roots to invade the xylem vessels and cause vascular wilt disease [58]. Genomic and transcriptomic studies have revealed that the *V. dahliae* genome harbors lineage-specific (LS) regions that are variable between strains and enriched for genes that are *in planta* induced [59,62]. These LS regions are generally considered genomic hotspots for evolutionary adaptation to plant hosts [59,62,63,74,75,191]. Recently we explored the epigenome of *V. dahliae* and distinguished LS regions from the core genome based on particular chromatin signatures, including elevated levels of H3K27me3 accompanied with accessible DNA and active transcription [191]. Using a machine learning approach and supported by orthogonal analyses, we identified nearly twice as much LS DNA as previously considered, collectively referred to as adaptive genomic regions [191]. Given the elevated levels of H3K27me3 at adaptive genomic regions in *V. dahliae*, and previous reports that removal of H3K27me3 results in transcriptional induction [49,51,52], we now tested if H3K27me3 dynamics are required for transcriptional activation of genes under different growth conditions. Ideally, the involvement of H3K27me3 dynamics in differential gene expression of *V. dahliae* would be studied between *in vitro* and *in planta* growth, as adaptive genomic regions are enriched for *in planta* induced genes [59,191]. However, *V. dahliae* displays a low pathogen-to-plant biomass during infection [192], which impedes technical procedures to determine histone modification levels over the genome [193]. Nevertheless, as H3K27me3 is generally reported to regulate transcription in response to environmental stimuli [178,180], we hypothesize that it may be involved in transcriptional regulation in *V. dahliae* during differential growth conditions *ex planta* as well. Here, we analyze RNA-seq, ChIP-seq and ATAC-seq data of *V. dahliae* cultured in various axenic growth media to understand if transcriptional dynamics require concomitant changes H3K27me3 modification status.

Results

Chromatin features correlate with gene expression levels

To determine how general features of chromatin, such as histone modifications and DNA accessibility, impact transcriptional activity in *V. dahliae*, we mapped the occurrence of heterochromatic histone marks H3K9me3 and H3K27me3, euchromatic histone marks H3K4me2 and H3K27ac, and chromatin accessibility determined by assay for transposase accessible chromatin followed by sequencing (ATAC-seq) [151] (Fig. 1A, Fig. S1). Grouping the *V. dahliae* genes into five expression quintiles, from quintile 1 containing the highest 20% expressed genes to quintile 5 with the 20% lowest expressed genes, we are able to integrate histone modification profiles and transcriptional activity. Expressed genes (quintiles 1-4) displayed low H3K4me2 and H3K27ac coverage upstream of the transcription start site (TSS), followed by a steep increase of coverage over the start of the gene that decreases over the gene body and increases again at the transcription end site (TES) (Fig. 1B). The strength of the association corresponds to the level of transcription, also within quintiles (Fig. S2). The low H3K4me2 and H3K27ac coverage directly upstream of the TSS coincides with increased chromatin accessibility, where higher expressed quintiles have more open DNA (Fig. 1B, Fig. S2). This chromatin profile upstream of the TSS suggests occurrence of a nucleosome depleted region. There is little evidence of H3K9me3 over gene bodies and *cis*-regulatory regions (Fig. 1B, Fig. S2), which corroborates that H3K9me3 marks TEs in constitutive heterochromatic regions such as the centromeres [97,121,191,194]. During cultivation in PDB, H3K27me3 is mainly present on genes that are lowly or not expressed (quintile 4 and 5) (Fig. 1B, Fig. S2). These results, and association between chromatin features and transcriptional activity, are consistent with reports for other fungi [49,50,54].

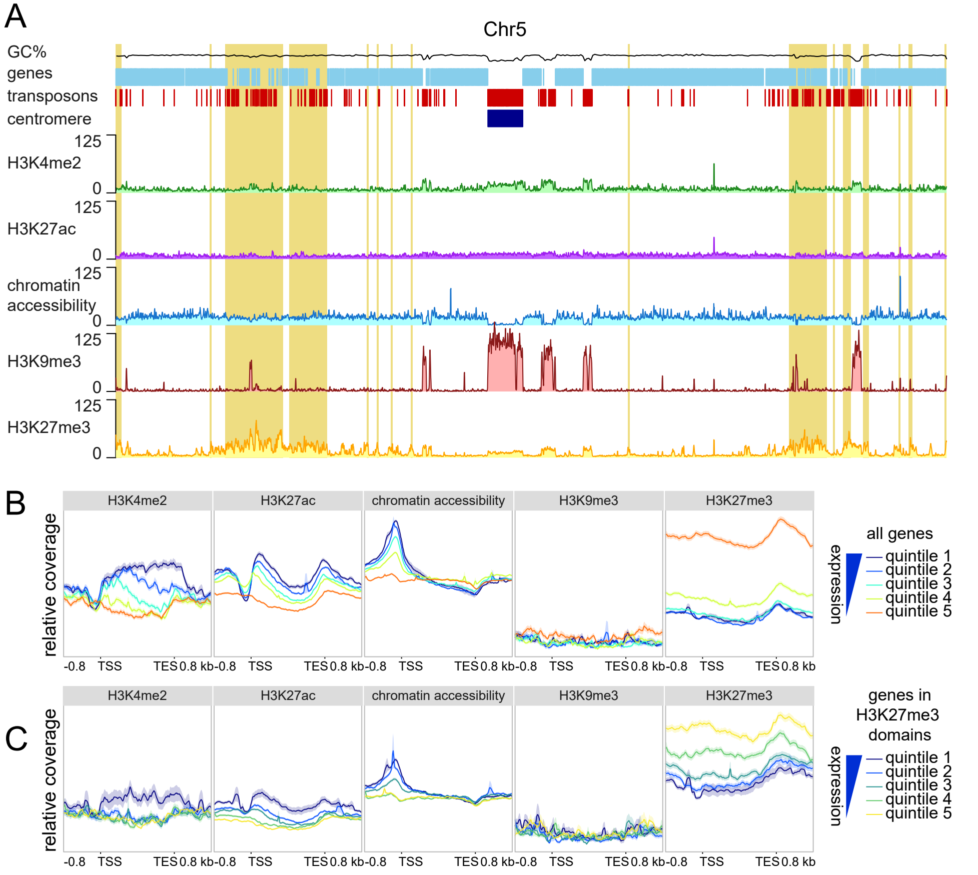


FIGURE 1 | Lowly and non-expressed genes associate with H3K27me3. A) Whole genome distribution of the euchromatin-associated histone modification H3K4me2 and H3K27ac, the constitutive heterochromatin-associated histone modification H3K9me3, the facultative heterochromatin-associated histone modification H3K27me3 and chromatin accessibility as determined by ATAC-seq, based on chromosome 5 as an example. GC percentage is indicated in black, genes are indicated in light blue, transposons are indicated in red, the centromere is indicated in dark blue and adaptive genomic regions are indicated in yellow. B,C) Relative coverage of chromatin accessibility and the histone marks H3K4me2, H3K9me3, H3K27me3 over gene bodies (between transcription start site (TSS) and transcription end site (TES)) ±800 bp of flanking sequence, grouped into quintiles based on gene expression levels upon cultivation for six days potato dextrose broth (PDB), for (B) all genes, and for (C) all genes located in an H3K27me3 domain.

To further analyze the association between H3K27me3 and gene expression, we identified 3,186 genes covered by H3K27me3 peaks (see methods) from triplicate grown *V. dahliae* in PDB. These 3,186 genes were separated into five expression quintiles as previously described. Interestingly, we found that genes in H3K27me3 domains with higher expression values (quintile 1 and 2) had higher H3K4me2 over the gene body, more accessible promoters, and lower H3K27me3 values (Fig. 1C). This suggests that genes in H3K27me3 domains are not uniformly repressed or heterochromatic, and there appears to be a quantitative, rather than qualitative, association between H3K27me3 association and gene activity. Genes with a lower association to H3K27me3 may represent loci that are not in a stable state under the tested conditions, where some cells have a more euchromatic profile and other have a more heterochromatic profile. While we cannot infer these details from the current data, it is clear that genes with lower expression in PDB are generally marked with H3K27me3, have less H3K4me3, and have less accessible DNA in the region of transcription initiation (i.e. their promoter).

Genetic perturbation of H3K27me3 induces transcription of many genes that are differentially expressed *in vitro* and *in planta*

To further characterize the influence of H3K27me3 on gene expression in *V. dahliae*, we deleted the histone methyltransferase component of the Polycomb Repressive Complex 2 (PRC2), termed *Set7* (Δ*Set7*), leading to loss of H3K27me3 (Fig. 2A, Fig. S3, S4). We do note that some background signal is present for the H3K27me3-ChIP-seq conducted in Δ*Set7*, but the signal is relatively uniform across the genome and does not correspond to the regions of H3K27me3 found in wild-type (Fig. 2A). As H3K27me3 is generally associated with facultative heterochromatin, we anticipated that the loss of H3K27me3 would mainly lead to induction of genes that were located in H3K27me3 domains in wild-type *V. dahliae*. Out of the 839 genes that are induced in Δ*Set7* ($\log_2$-fold change >2, $p<0.05$), 625 (74.5%) locate in H3K27me3 domains, which is significantly higher than expected given that only 27.9% of genes locate in H3K27me3 domains (Fisher's exact test, $p<0.00001$). In contrast, we find that 211 (27.6%) of 765 repressed genes in Δ*Set7* are in H3K27me3 domains (no association, Fisher's exact test, $p=0.94$) (Fig. 2B,C). Additionally, when comparing $\log_2$-fold changes in expression between Δ*Set7* and the wild type strain, we observed that genes located in a H3K27me3 domain in wild type are more significantly induced in Δ*Set7* than genes not locating in H3K27me3 domains (Two-sample Student's T-test, $p<0.001$) (Fig. S5). These findings support the role of H3K27me3 in transcriptional repression, and show that loss of H3K27me3 can lead to de-repression during growth *in vitro*.

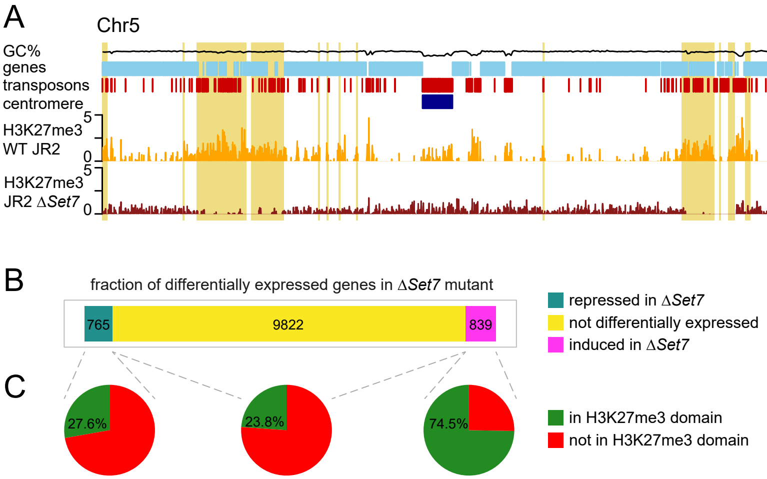


FIGURE 2 | Genetic perturbation of H3K27me3 results in induction of genes that are transcriptionally regulated in different growth conditions. A) H3K27me3 ChIP sequencing on triplicate JR2 WT (yellow) and duplicate ΔSet7 (red) samples with coverage over chromosome 5 as an example. Adaptive genomic regions are indicated in yellow. B) Fractions of induced (log2 fold change >1, p<0.05) and repressed (log2 fold change <-1, p<0.05) genes and genes that are not differentially expressed between wild-type and ΔSet7 that (C) locate in H3K27me3 domains.

H3K27me3 domains are globally stable between *in vitro* growth conditions

Given that H3K27me3 domains contribute to transcriptional repression, a key question concerns the status of H3K27me3 under growth conditions where the underlying genes are transcribed. One hypothesis is that H3K27me3 is removed or lost under growth conditions that activate gene expression, which would be noticeable as a change in H3K27me3 ChIP-seq domains between *V. dahliae* growth conditions that lead to differential expression. Here, the observed changes in H3K27me3 domains should be associated with transcriptional differences of the underlying genes. An alternative hypothesis is that H3K27me3 domain status does not change in accordance with transcriptional activity, and the repressive effects of H3K27me3 are released through alternative means. To test this hypothesis, we performed triplicates of H3K27me3 ChIP-seq on *V. dahliae* cultivated for six days in MS and CZA, in addition to the previously generated ChIP-seq data in PDB. Based on correlation between replicates (Fig. S6), we decided to continue with 3 replicates of H3K27me3 ChIP data in PDB and 2 replicates each of H3K27me3 ChIP data in CZA and MS. Control ChIP-input samples were used to normalize H3K27me3 datasets and identify H3K27me3-domains in each of the three growth media. We identified a total of 2,654 genes that were always present in H3K27me3 domains, regardless of the *in vitro* growth media (Fig. 3A). Interestingly, the 2,654 genes present in stable H3K27me3 domains display significantly stronger differential expression between all pair-wise media comparisons when compared with non-marked genes in all three growth conditions (Fig. 3B). This suggests that differential gene expression can occur without changes in global H3K27me3 coverage. We further checked whether genes that are differentially expressed between *in vitro* growth media are associated with H3K27me3 during cultivation in the growth medium with low expression levels (Fig. S7). Interestingly, we observed that for all pair-wise comparison between growth media, genes with higher log2 fold-changes in expression are more likely to locate in an H3K27me3 domain in the non-transcriptionally permissive growth medium (Fig. S7), again suggesting that H3K27me3 is involved in regulation of differential gene expression. To assess if changes in H3K27me3 domains between the media were associated with changes in transcription, we compared differences in domains and differential gene expression between media. We first analyzed differential expression for genes present in H3K27me3 domains in two media but not in an H3K27me3 domain in the third medium. For example, we identified 366 genes that were in a H3K27me3 domain in PDB and MS but not during CZA growth (Fig. 3A), and found that these genes are not differential expressed between PDB and MS (one-sample t-test, p 0.99), but they are higher expressed in CZA compared to both PDB (one-sample t-test, p 2e-4) and MS (one-sample t-test, p 5e-8) (Fig. 3C). Similarly, genes present in shared H3K27me3 domains for CZA and PDB growth (111 genes) are not differentially expressed between CZA and PDB (one-sample t-test, p 0.92) when the genes are associated with H3K27me3, they are higher expressed in MS than in PDB (one-sample t-test, p 4e-3), but not higher expressed in MS than CZA (one-sample t-test, p 0.99). For genes present in shared H3K27me3 domains between MS and CZA growth (223 genes), we did not observe statistically significant transcriptional differences between the growth conditions where the genes lacked H3K27me3 domains (Fig. 4A, C). Analyzing H3K27me3 domains unique to a medium, we found that CZA growth had the greatest number and proportion of unique genes locating in H3K27me3 domains (23.3%), followed by 8.0% unique to MS growth, and 1.7% unique to PDB growth (Fig. 3A). The genes uniquely marked in any condition did not show consistently increased expression in the condition in which the gene was not located in an H3K27me3 domain (Fig 3D). Overall, these results suggest that differential expression can be associated with differential H3K27me3 domain status, but it is not a requirement. We observed clear examples where the loss of H3K27me3 in one medium is associated with increased transcription in that medium, but this was not universally true. Many genes undergo differential gene expression between growth conditions and remain in stable H3K27me3 domains. We note that the majority of genes located in H3K27me3 domains were common to all three growth conditions, accounting for 83.3%, 75.3% and 68.2% of the identified genes in H3K27me3 domains from PDB, MS and CZA growth, respectively, indicating that the qualitative presence of H3K27me3 domains is globally stable.

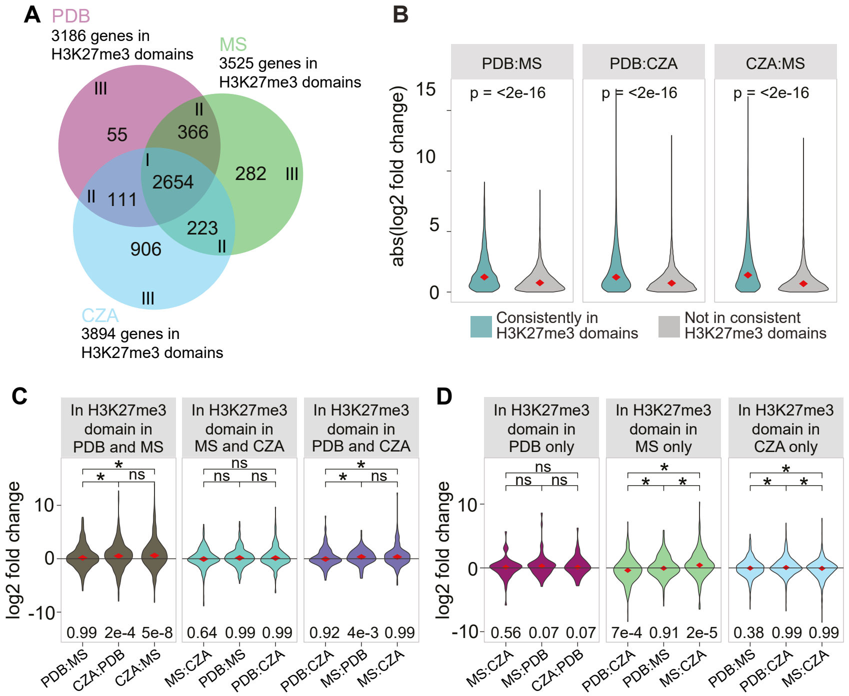


FIGURE 3 | Differential H3K27 methylation partially explains differential expression. A) The number of genes located in H3K27me3 domains for V. dahliae cultivated for 6 days in Czapec-Dox (CZA), half-strength Murashige Skoog medium (MS) and potato dextrose broth (PDB). B) Absolute log2 fold-changes in expression between pairs of growth media. The genes were grouped between those consistently localized in H3K27me3 domains (2654 genes) and those not consistently present in H3K27me3 domains. C,D) Log2 fold-change values for each pair of growth medi1 for genes in H3K27me3 domains found in C) two or D) one of three tested in vitro growth media. Differential gene expression comparisons are set such that genes higher expressed in the medium in which they do not locate in H3K27me3 domain will have positive log2 fold-change. In case differential gene expression comparisons are between growth media in which the genes locate in H3K27me3 domains in C) both or D) neither of the compared media, the orientation of positive and negative log2 fold-change is arbitrary. Median values of violin plots are indicated with diamonds and shown above the plot. Significant difference of TPM values between growth media are determined with the One-Sample Wilcoxon Signed Rank Test (*: p <= 0.05). The one-sample two-sided T-test was used to test whether sample means significantly differ from 0, p-value shown below the plot.

Local quantitative differences in H3K27me3 levels are associated with transcriptional differences

Our analysis on H3K27me3 presence/absence dynamics did not account for potential quantitative differences between growth media. Whole chromosome plots of H3K27me3 domains identified between media reflect their generally stable presence (Fig. S8), but the analysis is based on qualitative H3K27me3 domain identification. Domain calling for H3K27me3 results in broad domains, but this fails to capture higher resolution quantitative differences that may exist between media, as is seen during inspection of global chromosome plots (Fig. S8).

To understand how qualitative domain calling impacts the analysis, we examined quantitative differences between H3K27me3 ChIP-seq signal and transcriptional output between pairs of growth conditions. Genes were grouped based on differential gene expression between media, where genes that are significantly higher expressed in media A were one group and genes significantly that are higher expressed in media B another group. Subsequently, the H3K27me3 ChIP-seq signal relative to the input samples were normalized and compared between both growth media for the groups of differentially expressed genes (Fig. 4). Comparing results for PDB versus MS, we see that genes that are higher expressed in MS have a significantly lower MS H3K27me3 ChIP-signal versus the same genes from PDB ChIP (Fig. 4A). The contrasting comparison, genes that higher expressed in PDB, shows these genes do not significantly differ in ChIP signal between PDB and MS growth (Fig. 4A). Further integrating the transcriptional and ChIP fold-changes, for the 4,794 genes that have an input-corrected H3K27me3 ChIP signal above 0 in either MS and PDB, we see that genes higher expressed in PDB have a significantly lower log2 fold-change for H3K27me3 coverage, indicating these genes have lower H3K27me3-signal in PDB when compared with MS (Fig. 4B). Quantifying the number of genes in quadrant II and IV, we find that 396 (242+154) display a negative association between transcription and H3K27me3 ChIP-signal, whereas 256 (145+111) display a positive association (Fig. 4B). Thus, more genes are present in the quadrants that represent genes having lower H3K27me3 levels and higher transcription between the two growth conditions. The linear regression based on all genes is R=-0.1, also suggesting the low but significant negative association (Fig. 4B). It is clear that the association between differential expression and changes in ChIP-signal is not true for all genes, but overall we observe that the majority of genes that display changes for H3K27me3 ChIP-signal between media show the predicted transcriptional response where less H3K27me3 is associated with higher transcription (Fig. 4B).

Results for PDB versus CZA growth showed that genes higher expressed in CZA have higher H3K27me3 levels in PDB, consistent with the expected association (Fig. 4C). For genes higher expressed in PDB, however, we also observed a higher H3K27me3 level in PDB. Globally, the data indicated that genes higher expressed in CZA have more H3K27me3 ChIP signal in PDB than genes that are higher expressed in PDB, indicating a negative association between transcription and H3K27me3 presence (Fig. 4D). This is corroborated by the number of genes per quadrant, as 568 (197+371) genes locate in quadrants II and IV, whereas 469 (112+357) genes locate in quadrants I and III (Fig. 4D). The linear regression analysis indicates a slight negative association between differential expression and H3K27me3 ChIP-signal (Fig. 4D). We also observed an overall higher ChIP-signal from samples grown in CZA, but the reason for this is not clear.

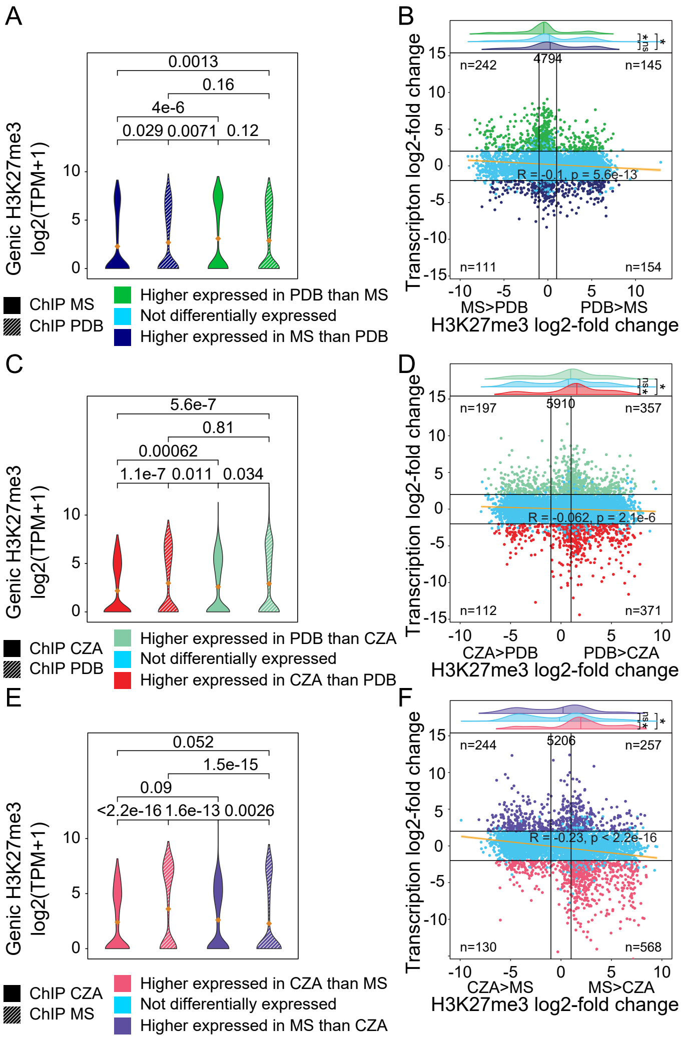


FIGURE 4 | Differentially expressed genes associate with local changes in H3K27me3 coverage. Pairwise comparisons of input-corrected H3K27me3 coverage over differentially expressed genes for V. dahliae cultivated for 6 days in A,B) half-strength Murashige Skoog medium (MS) and potato dextrose broth (PDB), C,D) Czapec-Dox (CZA) and PDB, and E,F) MS and CZA. A,C,E) H3K27me3 coverage over differentially expressed genes in each corresponding growth medium. Mean values of violin plots are indicated with orange diamonds. B,D,F) per gene comparison of log2-fold change in transcription and genic H3K27me3 coverage, for genes with input-corrected H3K27me3 coverage > 0 in either or both compared media (total number of genes indicated in top of plot). Black horizontal lines indicate transcription log$_2$-fold change of -2 and 2, black vertical lines indicate H3K27me3 log$_2$-fold change of -1 and 1. Numbers of genes in the four corner section are indicated. Significant difference of TPM values between growth media or gene sets are determined with the One-Sample Wilcoxon Signed Rank Test and indicated by their p-value or by asterisks (*: $p \le 0.05$).

Results for CZA versus MS growth showed that genes higher expressed in CZA had statistically higher levels of H3K27me3 ChIP-signal in MS medium, whereas genes higher expressed in MS have higher levels of H3K27me3 in CZA medium (Fig. 4E). The same pattern was observed in the integrated analysis, as genes higher expressed in CZA have more H3K27me3 signal in MS (Fig. 4F). There are 812 (244+568) genes in quadrant II and IV, supporting a negative association between differential transcription and H3K27me3 ChIP-signal, whereas 387 (130+257) genes are present in quadrants I and III. This is supported by the significant negative correlation (R=-0.23) (Fig. 4F). Overall, the results of the integrated analyses show that there is an association between quantitative transcriptional levels and H3K27me3 signal, where genes that are more highly expressed in a transcriptionally permissive medium have less H3K27me3 when compared with H3K27me3 levels in the repressive media. There are also many genes that were differentially expressed without an accompanying shift in H3K27me3 log2 fold-change (Fig. 4B,D,F). Collectively, these results are consistent with the observations of the qualitative H3K27me3 presence/absence comparisons, and suggest that H3K27me3 levels are generally stable, but local changes at particular genes may contribute to transcriptional dynamics.

Genes induced *in planta* are largely H3K27me3 associated across all tested growth conditions

The presented analyses compared the directionality of H3K27me3 and transcriptional changes between axenic growth media to address if H3K27me3 dynamics are associated with transcriptional dynamics. Another interesting growth condition for *V. dahliae* is host colonization, and the contrast for the genes differentially expressed *in planta* when compared with axenic culture. We have thus far been unable to perform ChIP on *V. dahliae* during host colonization due to a low pathogen-to-plant biomass, and therefore cannot compare H3K27me3 levels between these conditions. However, we can identify the genes that are significantly induced during infection when compared with *in vitro* growth media and assess their chromatin profiles in these media to assess if *in planta*-induced genes appear heterochromatic during *in vitro* growth. Genes were grouped based on differential expression between PDB and *in planta* growth and chromatin profiles showed that genes that are higher expressed *in planta* had significantly more H3K27me3 during PDB growth when compared with genes that are more highly expressed in PDB (Fig. 5A). These *in planta*-induced genes lacked H3K4me2 and the DNA was less accessible during PDB growth (Fig. 5A). The association between *in planta* induction and higher H3K27me3 levels was not only seen relative to PDB growth, but also to growth in MS and CZA (Fig. 5B,C). For each media comparison, the genes that are more highly expressed *in planta* have higher H3K27me3 levels during axenic growth. To understand what genes are driving these differences, we investigated whether *in planta*-induced genes that locate in H3K27me3 domains are overrepresented for genes that are potentially involved in infection. There were approximately 600 genes that are more highly expressed *in planta* when compared with any of the three media (Fig. 5D). We found that, depending on the medium, from 41.2% to 52.8% of the *in planta*-induced genes were located in H3K27me3 domains (Fig. 5D). We observed that the *in planta*-induced genes within H3K27me3 domains have a higher, yet not statistically significant, fraction of genes encoding secreted proteins or putative effectors when compared with all *in planta*-induced genes. For example, genes in H3K27me3 domains that were *in planta*-induced when compared with PDB growth (277 genes total), 15.5% had a secretion signal and 4.3% were predicted effectors. This is compared to all *in planta*-induced genes relative to PDB (673 genes total), where 12.3% had a secretion signal and 2.7% were predicted effectors. These results were similar between the other two media, and we conclude that *in planta*-induced genes in H3K27me3 domains have a slightly higher fraction of genes that are potentially involved in infection when compared with *in planta*-induced genes not in H3K27me3 domains. Collectively, these results indicate that genes that are higher expressed *in planta* have a heterochromatic profile (i.e. H3K27me3 association, low H3K4me2, lower accessibility) when analyzed in axenic culture. The chromatin profile of these genes during *in planta* growth will need to be directly assessed in the future.

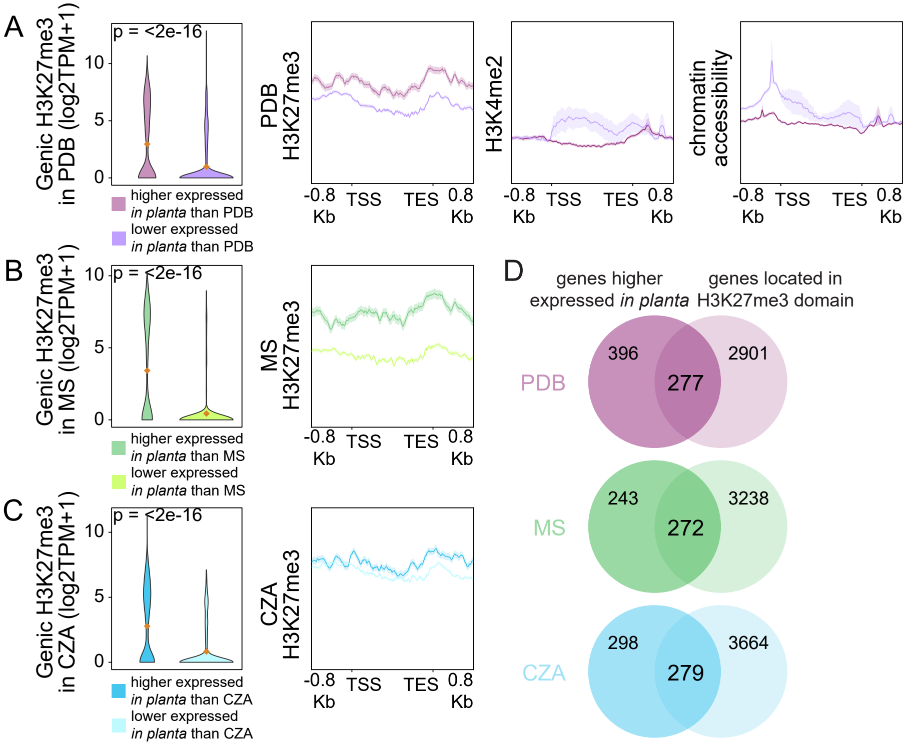


FIGURE 5 | *In planta*-induced genes are H3K27me3-associated under transcriptionally repressive conditions. Violin plots display input-corrected H3K27me3 ChIP signal over genes differentially expressed between in planta and A) 6 days of cultivation in potato dextrose broth (PDB), B) half strength Murashige Skoog medium (MS), and C) Czapec-Dox medium (CZA). Mean values are indicated by orange diamonds. Line plots display average coverage of chromatin features over the gene sets. D) overlap of genes that are higher expressed in planta than in PDB, MS or CZA with genes located in an H3K27me3 domain in the corresponding growth medium. Significant differences of H3K27me3 coverage are determined with the One-Sample Wilcoxon Signed Rank Test (*: p <= 0.05).

Discussion

Chemical histone modifications play an essential role in transcriptional regulation, but a number of mechanistic questions remain for their role in differential transcription in filamentous fungi. In this study, we show that many genes that are differentially transcribed between *in vitro* growth conditions are located in H3K27me3 domains. This is interesting because the mark contributes to transcriptional repression and it is relatively sparse in the genome, occupying approximately 33% of the genes. However, 35% to 70% of differentially expressed genes identified between growth conditions reside within H3K27me3 domains. It is not clear what mechanism might drive this association.

The global fraction of genes that locates in H3K27me3 domains in *V. dahliae* is similar to what has been reported for the ascomycete *Podospora anserina* [185], but higher than for *N crassa, Zymoseptoria tritici*, *Leptosphaeria maculans,* and *Magnaporthe oryzae* (9-16% of genes in H3K27me3 domain) [54,182,183,195]. Previous reports showed that H3K27me3 represses transcription of secondary metabolite clusters in *Fusarium spp.* and in *Epichloë festucae*, and of effectors in *Z. tritici* and *M. oryzae* [49,51,53,54,195]. However, these reports have mainly relied on genomic association with H3K27me3, genetic perturbation altering H3K27me3 deposition, and the analysis of only few genes under natural conditions. Given our results, and these previous findings, our hypothesis was that differences in transcription between *in vitro* growth conditions is in part coordinated by dynamics for H3K27 methylation. We directly addressed this phenomenon using genome-wide H3K27me3 profiling and RNA-seq under different *in vitro* growth conditions. Consistent with reports in other fungi, we see that H3K27me3 is associated with transcriptional repression, and that deleting *Set7*, encoding the histone methyltransferase that is responsible for H3K27me3, led to induction of many genes normally present in H3K27me3 domains. Importantly, our direct observations of H3K27me3 levels and transcriptional output from three *in vitro* media showed that H3K27me3 domains are generally stable globally. We see that 50% to 75% of the identified H3K27me3 domains, depending on the medium analyzed, did not change between the tested conditions. Despite many H3K27me3 domains not undergoing presence/absence dynamics, numerous genes in these domains displayed differential transcription between the tested conditions, indicating that complete loss of H3K27me3 is not strictly required for transcriptional induction.

One possibility to account for these seemingly contradictory results is that even though H3K27me3 appears stable at the level of broad-peak calling, there may be smaller regions within the broad domain that are dynamic. Assessing quantitative differences in H3K27me3 ChIP-seq levels at defined genomic locations between paired *in vitro* growth conditions indicates that this can account for some of the observed transcriptional differences. For example, some genes have a lower H3K27me3 ChIP-signal upon growth in transcriptionally permissive medium when compared with growth in a transcriptionally repressive medium. We interpret these results as evidence for local, rather than global, changes in H3K27me3 dynamics contributing to transcriptional regulation for the underlying genes. H3K27me3-associated genes that are differentially expressed while presence of the histone mark remains stable, may be transcriptionally regulated through the activity of H3K27me3 readers. For instance, the *Fusarium graminearum* histone reader BP1, which is orthologous to *N. crassa*

EPR-1, specifically binds to H3K27me3 and co-represses gene transcription [196,197]. The gene encoding BP1 is conserved within fungi, including *V. dahliae* [196]. Dynamic binding of such transcription-repressing histone reader to stable H3K27me3 domains may explain the observed transcriptional dynamics.

We note that an individual histone does not permit for a quantitative tri-methylation status, as an individual H3K27 is either tri-methylated or not. At the nucleosome level, there can be none, one, or two H3K27 tails with a tri-methyl modification. At the cell population level, there can be considerable quantitative differences because of heterogeneity for histone modification status between individual cells. Cell variability may be the source of quantitative differences observed in our experiments, arising from variation for both the percent of cells with H3K27me3 and the number of tails of a nucleosome with H3K27me3. While we cannot determine this based on our present data, our results show that some genes that are differentially expressed between growth conditions are associated with quantitative differences in H3K27me3 levels, providing evidence that chemical histone modifications dynamics can be involved in a transcriptional regulatory mechanism in *V. dahliae*. The concept of local versus global H3K27me3 changes is consistent with data from *M. oryzae*, in which direct *in planta* H3K27me3 ChIP-qPCR showed that some, but not all, analyzed regions displayed differential H3K27me3 levels consistent with increased transcription between *in planta* and *in vitro* conditions [195]. We have not been able to directly assess histone modifications for *V. dahliae* during plant infection, but we did analyze the status of H3K27 methylation during *in vitro* growth of *in planta*-induced genes. The *in planta*-induced genes have significantly higher levels of H3K27me3 and the DNA is substantially less accessible during *in vitro* growth when compared with genes that are highly expressed during *in vitro* growth. We conclude that dynamics for H3K27me3 can contribute to differential gene expression under natural conditions, but our results show this is not required. Our conclusions are limited to the tested conditions, and it is possible that analyzing H3K27me3 levels in different cell types or growth stages (e.g. spores, microsclerotia, *in planta* infection) may yield a different picture of global H3K27me3 distribution. Additionally, our experiments focused on steady-state growth on different media as these provide a clear and reproducible transcriptional difference. It is possible that H3K27me3 dynamics occur rapidly in response to environmental changes, cues, or developmental stages that we did not capture.

It appears evident that H3K27me3 contributes to additional genome functions beyond strict transcriptional repression, at least in fungi [190]. This is supported by our data showing that the majority of H3K27me3 domains were stable between the tested growth conditions, indicating these H3K27me3 domains have additional functions. Stable H3K27me3 domains may represent another form of constitutive heterochromatin, but this fact seems unlikely given that in our data many stable H3K27me3 domains harbored differentially expressed genes between growth media. Additionally, results in *N. crassa* have shown that genetic perturbation leading to the genome-wide loss of H3K9me3 causes a redistribution of H3K27me3 to previously H3K9me3 marked sites, but interestingly, this re-distribution does not result in transcriptional silencing [198]. Rather, the redistribution of H3K27me3 in *N. crassa* lacking H3K9me3 appears to contribute to genomic instability [198,199]. In *Z. tritici*, the enrichment of H3K27me3 at dispensable chromosomes and empirical evidence that H3K27me3 somehow increases genomic instability, additionally supports the hypothesis that H3K27me3 contributes to additional genomic functions beyond transcriptional regulation [54,121]. Evolutionary analysis across *Fusarium* and related species indicates that genes marked by H3K27me3 have a higher duplication rate in *Fusarium* and are less conserved with more distantly related species [200]. Additional research is needed to fully understand the mechanisms of H3K27me3 targeting, dynamics, and impact on genome stability in fungi. The results presented here show that H3K27me3 domains are largely similar between *in vitro* growth conditions, but that quantitative differences in H3K27me3 levels can be associated with concomitant transcriptional differences between the tested conditions.

Materials and methods

Fungal growth conditions

Verticillium dahliae strain JR2 (CBS 143773 [122]) was cultured on potato dextrose agar (PDA) (Oxoid, Thermo Scientific) at 22˚C in the dark. Liquid cultures were obtained by collecting conidiospores from PDA plates after approximately 2 weeks of growth followed by inoculation at a final concentration of $1x10^4$ spores per mL into liquid growth media. Media used in this study are potato dextrose broth (PDB) (Difco, Becton Dickinson, Franklin Lakes, NJ, USA), half strength Murashige & Skoog plus vitamins (MS) (Duchefa-Biochemie, Haarlem, The Netherlands) medium supplemented with 3% sucrose, and Czapec-Dox medium (CZA) (Oxoid, Thermo Scientific, Waltham, MA, USA). Liquid cultures were grown for 6 days in the dark at 22˚C at 140 RPM. Mycelium was collected by straining cultures through miracloth (22 µm) (EMD Millipore, Darmstadt, Germany) and pressing to remove liquid after which the mycelium was flash frozen in liquid nitrogen and ground to powder with a mortar and pestle. If required, samples were stored at -20˚C prior to nucleic acid extraction. All analyses were performed based on triplicate cultures that were processed individually.

RNA sequencing and analysis

RNA was isolated from ground mycelium using TRIzol (Thermo Fisher Science, Waltham, MA, USA) following manufacturer guidelines. Contaminating DNA was removed using the TURBO DNA-free kit (Ambion, Thermo Fisher Science, Waltham, MA, USA) and RNA integrity was assessed with gel-electrophoresis and quantified using a Nanodrop (Thermo Fisher Science, Waltham, MA, USA). Library were prepared and singleend 50 bp sequenced on the DNBseq platform at BGI (Hong Kong, China). Sequencing reads were mapped to the reference annotation of *V. dahliae* strain JR2 [122] using Kallisto quant [155] to calculate per gene TPM values. Differential expression between cultivation in PDB and CZA, MS or during colonization of *Arabidopsis thaliana* was determined using DESeq2 [146].

ChIP-sequencing and analysis

ChIP-seq was performed as described previously [191]. Frozen ground mycelium was added to ChIP lysis buffer, dounced 40 times and subsequently sonicated for 5 rounds of 20 seconds with 40 second rest stages on ice. Supernatants were collected after centrifugation and treated with MNase (New England Biolabs, Ipswich, MA, United States) for 10 minutes in a 37˚C waterbath. MNase activity was quenched by addition of EGTA and samples were subsequently pre-cleared by addition of 40 µl Protein A Magnetic Beads (New England Biolabs, Ipswich, MA, United States) and rotating at 4˚C for 60 min. Beads were captured and supernatant was divided over new tubes containing antibodies against either H3K4me2, H3K9me3 or H3K27me3 (ActiveMotif; #39913, #39765 and #39155) and incubated overnight with continuous rotation at 4˚C. Subsequently, the antibodies were captured, washed and nucleosomes were eluted from beads, after which DNA was treated with Proteinase-K and cleaned-up using chloroform. DNA was isolated by overnight precipitation in ethanol and DNA concentration was determined with the Qubit™ dsDNA HS Assay Kit (Thermo Fisher Science, Waltham, MA, USA). Sequencing libraries were generated using the TruSeq ChIP Library Preparation Kit (Illumina, San Diego, CA, United States) according to instructions, but without gel purification and with use of the Velocity DNA Polymerase (BioLine, Luckenwalde, Germany) for 25 cycles of amplification. Single-end 125 bp sequencing was performed on the Illumina HiSeq2500 platform at KeyGene N.V. (Wageningen, the Netherlands).

Raw reads were trimmed using TrimGalore [201] and mapped to the *V. dahliae* strain JR2 reference genome [122] using BWA-mem with default settings [152]. Three regions of the genome were masked due to aberrant mapping, (chr1:1-45000, chr2:3466000-3475000, chr3:1-4200). ATAC-seq reads of *V. dahliae* cultured in PDB [191] were treated similarly as ChIP-seq reads, but the paired end read pairs were trimmed and mapped simultaneously, and only read-pairs <100 bp were considered for further analyses, as these represent open DNA. Mapped reads were RPGC-normalized using deepTools bamCoverage [202] with binsize 1,000 and smoothlength 3,000 for plotting over the genome, and binsize 10 and smoothlength 30 for further analysis. Normalized replicate samples with high correlation were selected and mean datasets per growth media were generated with input controls for background signal correction using WiggleTools mean [203]. H3K27me3-enriched regions were determined on selected replicates with input control using epic2 with a binsize of 2,500 bp [204]. Average coverage over gene bodies per expression quintile was calculated using deepTools computeMatrix in scale-regions mode [202]. Expression quintiles were generated by sorting all genes based on their average TPM value from three replicates of *V. dahliae* cultured for 6 days in PDB. We used deepTools multiBigwigSummary [202] to determine the presence of H3K27me3 over gene bodies (region between TSS and TES) for each replicate ChIP sample, as well as for input controls. Samples and input-controls were TPM normalized, after which the normalized input-control signal was subtracted from normalized H3K27me3 TPM values and resulting negative values were set to 0. Changes in H3K27me3 levels between growth media were determined by taking the average input-normalized H3K27me3 TPM values +1 and calculating the log2-fold change for each pair-wise comparison.

Generation of *Set7* deletion mutant

The *Set7* deletion mutant (D*Set7*) was constructed as previously described [123]. Briefly, genomic DNA regions flanking the 5’ and 3’ ends of the coding sequences were amplified with PCR using primers listed in Table S1 (primers 1-4) and cloned in to the pRF-HU2 vector [124] using USER enzyme following the manufacturer's protocol (New England Biolabs, MA, USA). Sequence-verified vectors were transformed into *Agrobacterium tumefaciens* strain AGL1 and used for *V. dahliae* conidiospore transformation as described previously [123]. *V. dahliae* transformants that appeared on hygromycin B were transferred to fresh PDA supplemented with hygromycin B after five days. Putative transformants were screened using PCR to verify deletion of the target gene sequence and integration of the selection marker at the designated locus using primers listed in Table S1 (primers 5-8). Analysis and comparison of gene expression in the *Set7* deletion mutant to type *V. dahliae* was performed as described above.

Acknowledgements

This work was supported by a PhD grant of the Research Council Earth and Life Sciences (project 831.15.002) to HMK, and by Human Frontier Science Program Postdoctoral Fellowship (HFSP, LT000627/2014-L), by USDA's National Institute of Food and Agriculture (award no. 2018-67013-28492) through the Plant Biotic Interactions Program, and by the National Science Foundation (award no. 1936800) Division of Molecular and Cellular Biosciences to DEC. Work in the laboratories of M.F.S and B.P.H.J.T. is supported by the Research Council Earth and Life Sciences (ALW) of the Netherlands Organization of Scientific Research (NWO). B.P.H.J.T acknowledges funding by the Alexander von Humboldt Foundation in the framework of an von Humboldt Professorship endowed by the German Federal Ministry of Education and Research is furthermore supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2048/1 – Project ID: 390686111.

Supplementary data

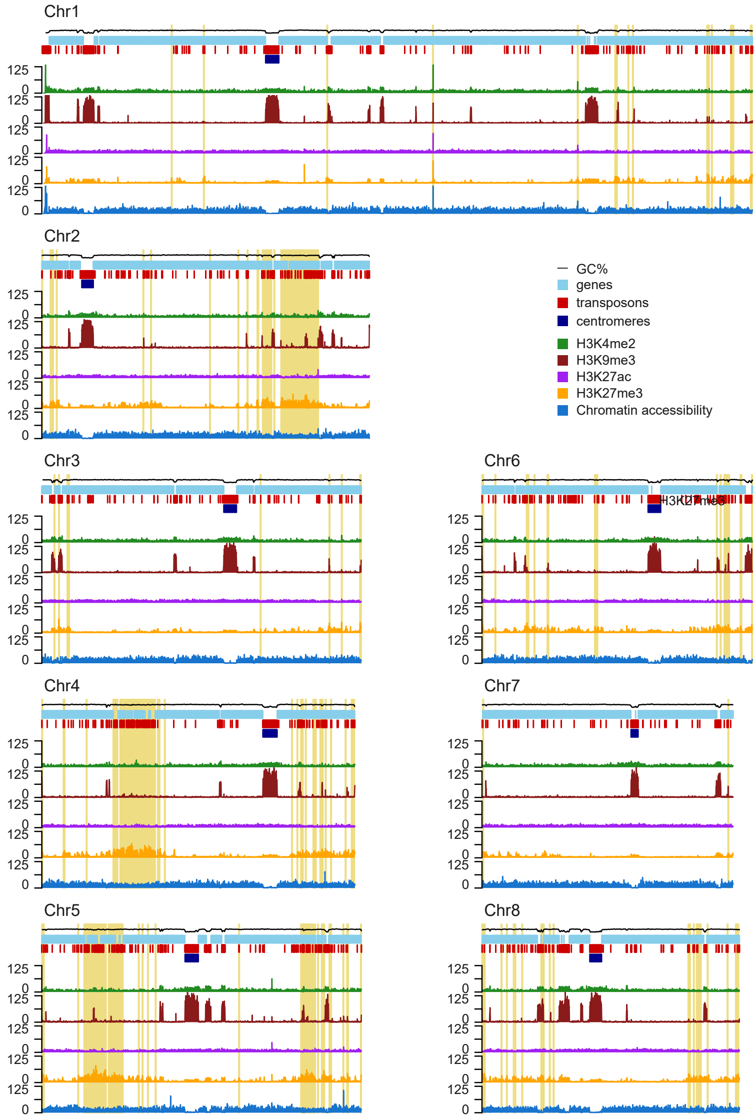


FIGURE S1 | Distribution of chromatin features over all chromosomes. Whole genome distribution of the euchromatin-associated histone modification H3K4me2 (green line), the constitutive heterochromatin-associated histone modification H3K9me3 (red line), the facultative heterochromatin-associated histone modification H3K27me3 (yellow line) and chromatin accessibility (blue line) as determined by ATAC-seq. Genes are indicated in light blue, transposons are indicated in red, centromere is indicated in dark blue and adaptive genomic regions are indicated in yellow.

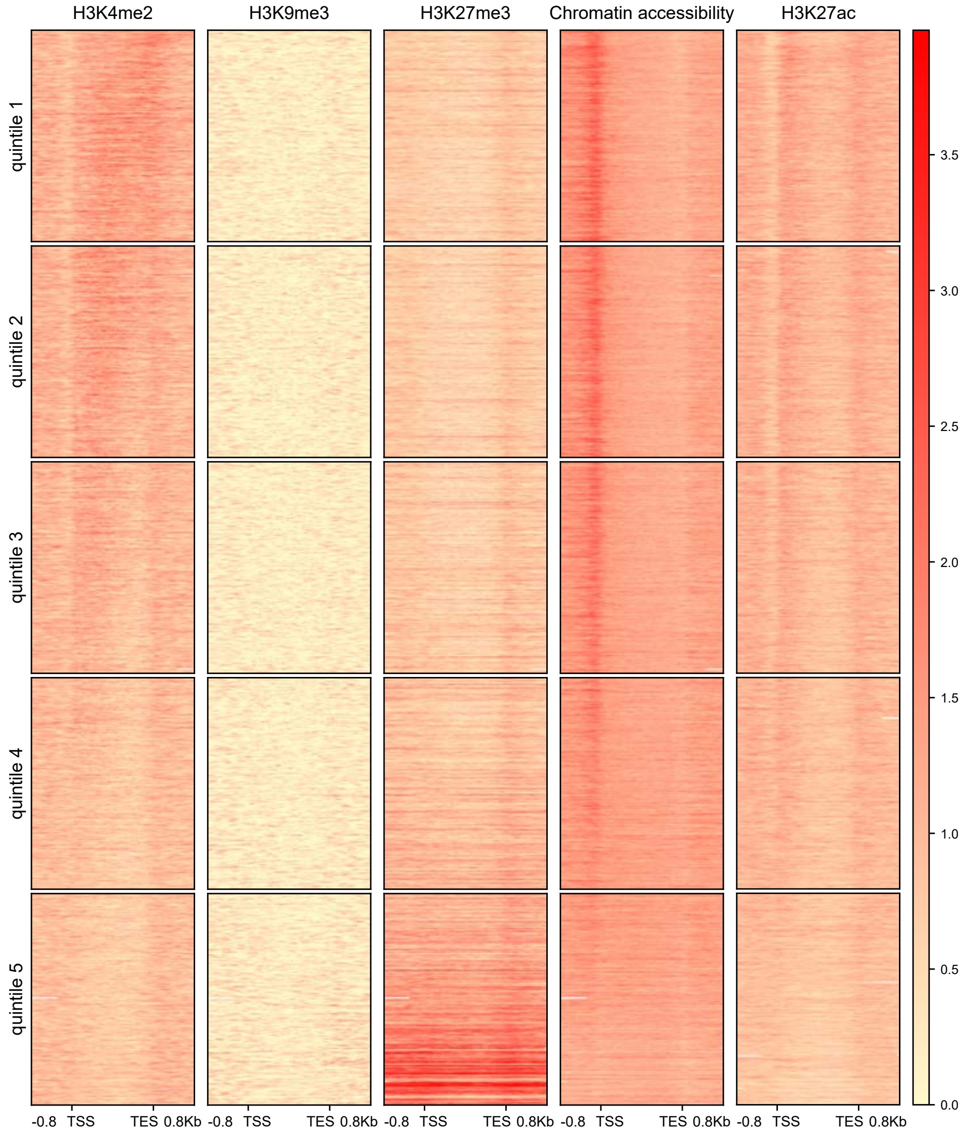


FIGURE S2 | Gene expression correlates with histone modification presence and chromatin accessibility. RPGC normalized coverage of the histone marks H3K4me2, H3K9me3, H3K27me3 and chromatin accessibility over gene bodies (between transcription start site (TSS) and transcription end site (TES)) ±800 bp of flanking sequence. Each row represents a single gene, and are sorted based on their TPM value upon cultivation for six days potato dextrose broth (PDB), with the top gene being most highly expressed. Genes are grouped in expression quintiles as in Fig. 3A.

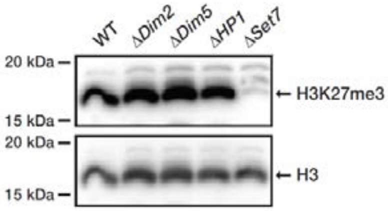


FIGURE S3 | Western blot shows loss of H3K27me3 in the *V. dahliae* Δ*Set7* deletion mutant. Histone isolations of wild-type, ΔDim2, ΔDim5, ΔHP1 and ΔSet7 were tested for presence of the H3K27me3 histone modification by Western blot. Antibody against H3 was used as loading control

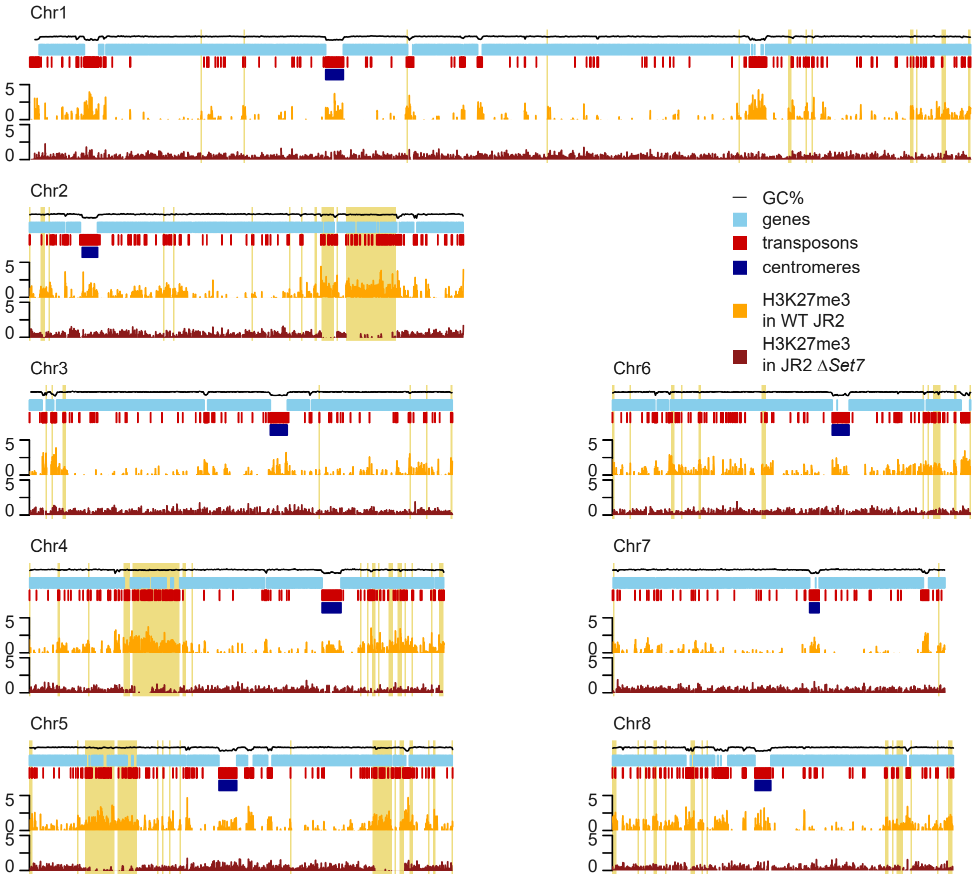


FIGURE S4 | ChIP-sequencing shows loss of H3K27me3 in the *V. dahliae* Δ*Set7* deletion mutant. H3K27me3 ChIP coverage over the genome in a triplicate of JR2 WT (yellow) and in a duplicate of JR2 Δ*Set7*, cultivated for 6 days in potato dextrose broth.

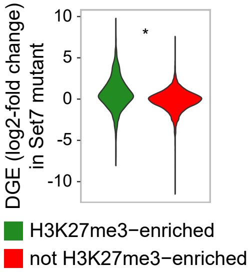


FIGURE S5 | Genes associated with H3K27me3 in wild type *V. dahliae* are stronger transcriptionally induced in the Δ*Set7* mutant than not H3K27me3 associated genes. Log2-fold change of expression between wild type and Δ*Set7* mutant for genes associated with H3K27me3 in wild type (green) and those not associated with H3K27me3 in wild type (red). Significant difference of log2-fold change between gene sets are determined with the One-Sample Wilcoxon Signed Rank Test (*: p <= 0.05).

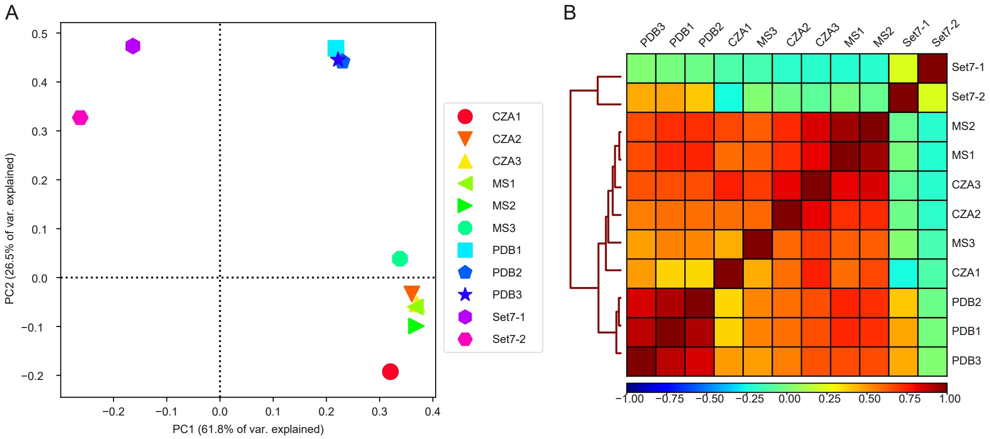


FIGURE S6 | Correlation between H3K27me3 ChIP samples. A) PCA plot and B) heatmap displaying between-sample correlation of H3K27me3 coverage for 1kb bins for triplicates of JR2 WT cultivated for 6 days in Czapec-Dox medium (CZA), half-strength Murashige-Skoog medium (MS) and potato dextrose broth (PDB), and a duplicate JR2 Δ*Set7* cultivated for 6 days in PDB.

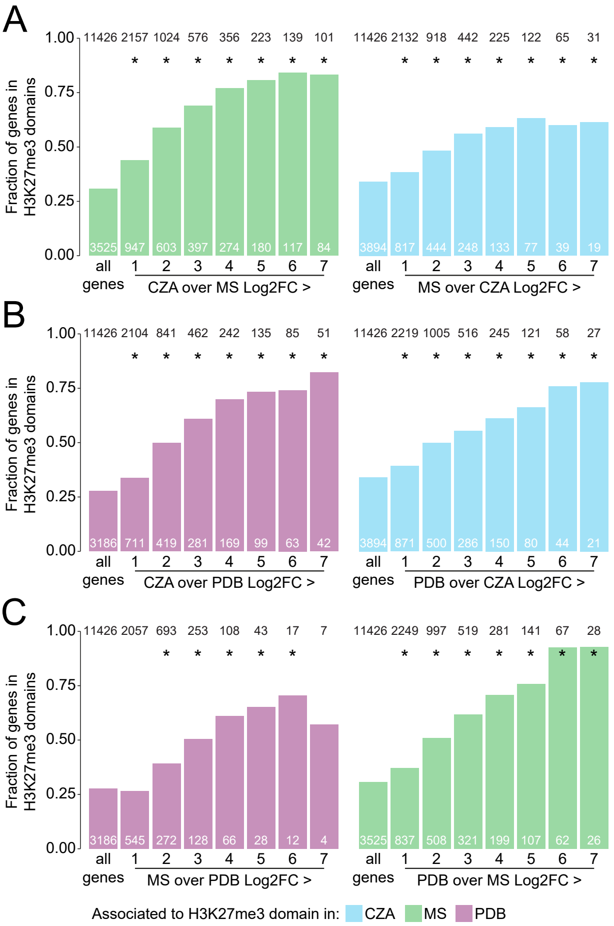


FIGURE S7 | Differentially expressed genes are enriched in H3K27me3 domains during cultivation in the non-permissive growth medium. Fractions of total and differentially expressed genes (Log2-fold changes >1, 2, 3, 4, 5, 6 and 7) that are associated with an H3K27me3 domain in the non-permissive condition Czapec-Dox medium (CZA, blue bars), half-strength Murashige-Skoog medium (MS, green bars) or Potato-Dextrose broth (PDB, red bars) between (A) cultivation for 6 days in CZA and MS, (B) cultivation for 6 days in CZA and PDB and (C) cultivation for 6 days in MS and PDB. Black numbers above the bars indicate numbers of genes per gene set. White numbers at bottom of the bars indicate the numbers of genes that locate in an H3K27me3-domain per gene set. Significance of fraction differences was calculated for each fraction of higher expressed genes in adaptive genomic regions when compared with the fraction of total genes in adaptive genomic regions, by the two-sided two-proportions Z-test (*: $p \leq 0.05$).

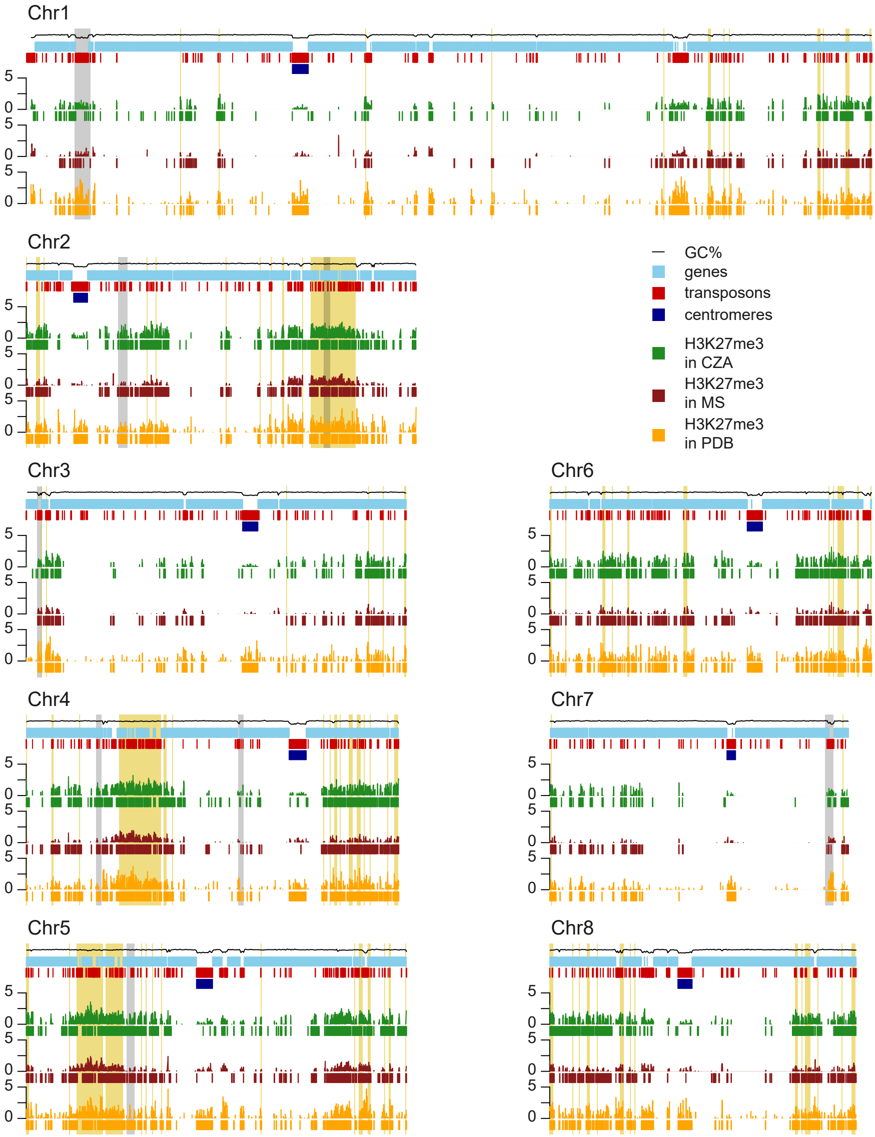


FIGURE S8 | Distribution of H3K27me3 for *V. dahliae* cultivated in in vitro growth media. Average H3K27me3 distribution for replicates of V. dahliae cultivated for 6 days in Potato dextrose broth (PDB, indicated in yellow), half strength Murashige Skoog medium (MS, indicated in red) and Czapec-Dox medium (CZA, indicated in green). Predicted H3K27me3 domains are indicated as blocks below each H3K27me3 track. Adaptive genomic regions are highlighted in yellow. Genomic regions within H3K27me3 domains that are visibly different between growth conditions are highlighted in grey.

TABLE S1 | Primers used to delete and analyze the *Set7* coding sequence in *V. dahliae*.

Name	Sequence	Purpose	No.
Set7.ko-LB_F	*GGTCTTAAU*TGAGCTTGACAGTTCAGTTGTCG	Amplify left flanking sequence	1
Set7.ko-LB_R	*GGCATTAAU*AAGTTGTGTTGTCAGCGTGCATA	Amplify left flanking sequence	2
Set7.ko-Rb_F	*GGACTTAAU*ATCAAGTCCGCCTACTTTCCAAG	Amplify right flanking sequence	3
Set7.ko-Rb_F	*GGGTTTAAU*GTGGAGAATCGTCTGGGGTTATC	Amplify right flanking sequence	4
Set7_Confirm_F	CCTCCAGCTCCTGAAGAAGAA	confirm gene replacement with selection	5
Vector_Reverse	GGAGTCGCATAAGGGAGAGCG	confirm gene replacement with selection	6
Set7.ORF.270_F	TCACAGCCGCTACATCAATCA	confirm ORF is absent in KO	7
Set7.ORF.270_R	TCGTGTTGAACCTCCTTGGAC	confirm ORF is absent in KO	8

Italic sequences at the 5' end represent those added for USER cloning

Chapter 4

Three putative DNA methyltransferases of *Verticillium dahliae* differentially contribute to DNA methylation that is dispensable for growth, development and virulence

H. Martin Kramer[1,*],
David E. Cook[1,2,*],
Grardy C.M. van den Berg[1],
Michael F. Seidl[1,3],
Bart P.H.J. Thomma[1,4]

[1]Laboratory of Phytopathology, Wageningen University and Research, Droevendaalsesteeg 1, 6708 PB Wageningen, the Netherlands
[2]Department of Plant Pathology, Kansas State University, 1712 Claflin Road, Manhattan, Kansas 66506, USA
[3]Theoretical Biology & Bioinformatics, Department of Biology, Utrecht University, Utrecht, The Netherlands
[4]University of Cologne, Institute for Plant Sciences,
Cluster of Excellence on Plant Sciences (CEPLAS), 50674 Cologne, Germany

*These authors contributed equally

Published in Epigenetics & Chromatin 14: 21 (2021)

Abstract

DNA methylation is an important epigenetic control mechanism that in many fungi is restricted to genomic regions containing transposable elements (TEs). Two DNA methyltransferases, Dim2 and Dnmt5, are known to perform methylation at cytosines in fungi. While most ascomycete fungi encode both Dim2 and Dnmt5, only few functional studies have been performed in species containing both. In this study, we report functional analysis of both *Dim2* and *Dnmt5* in the plant pathogenic fungus *Verticillium dahliae*. Our results show that Dim2, but not Dnmt5 or the putative sexual-cycle related DNA methyltransferase Rid, is responsible for the majority of DNA methylation under the tested conditions. Single or double DNA methyltransferase mutants did not show altered development, virulence, or transcription of genes or TEs. In contrast, *Hp1* and *Dim5* mutants that are impacted in chromatin-associated processes upstream of DNA methylation are severely affected in development and virulence and display transcriptional reprogramming in specific hypervariable genomic regions (so-called adaptive genomic regions) that contain genes associated with host colonization. As these adaptive genomic regions are largely devoid of DNA methylation and of Hp1- and Dim5-associated heterochromatin, the differential transcription is likely caused by pleiotropic effects rather than by differential DNA methylation. Overall, our study suggests that Dim2 is the main DNA methyltransferase in *V. dahliae* and, in conjunction with work on other fungi, is likely the main active DNMT in ascomycetes, irrespective of *Dnmt5* presence. We speculate that Dnmt5 and Rid act under specific, presently enigmatic, conditions or, alternatively, act in DNA-associated processes other than DNA methylation.

Introduction

Transcriptional control is important for regulating developmental processes and environmental responses. In eukaryotes, transcriptional control is achieved through transcription factor-mediated and epigenetic mechanisms, the latter affecting DNA accessibility and altering interactions between DNA and various proteins [205–207]. Eukaryotic DNA associates with histone-protein complexes to form nucleosomes that are the main constituents of chromatin, a highly ordered DNA-structure [40]. DNA accessibility for the transcriptional machinery is regulated in part by chemical modifications to histones that can alter chromatin structure or nucleosome positioning, and by direct DNA modifications that can alter transcription factor-binding sites [176]. One such DNA modification is mediated by DNA methyltransferases (DNMT) that covalently add a methyl group to the 5th carbon of a cytosine residue (5-methylcytosine, 5mC) [208]. Cytosine methylation can occur in symmetric CG or CHG genomic contexts, or in the asymmetric CHH genomic context, where H stands for either A, C or T. In general, 5mC occurs more commonly at symmetric sites because maintenance methylation can cause methylation of daughter strands during DNA-replication, whereas asymmetric sites require *de novo* methylation [209]. In mammals, DNA methylation is largely restricted to CG sites, while plants and fungi show methylation in each of the genomic contexts [210].

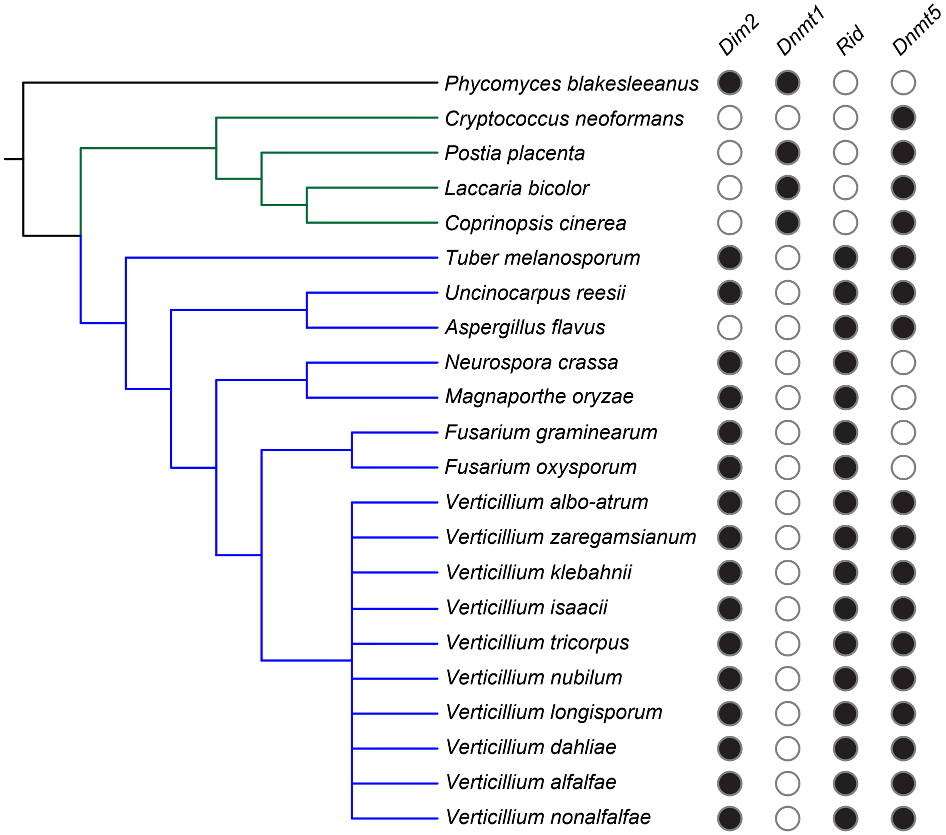
Compared to animal and plant genomes, fungi typically have smaller and less complex genomes, and they serve as important eukaryote models for various cellular processes including DNA methylation [211]. Much of the initial research on DNA methylation in fungi was performed in the saprophytic ascomycete fungus *Neurospora crassa*. In *N. crassa*, DNA methylation is restricted to transposable elements (TEs) and is dependent on a single DNMT, Deficient In Methylation-2 (Dim2), an ortholog of Human Dnmt1 that performs *de novo* as well as maintenance methylation [93]. Dim2 operates in a complex with Heterochromatin Protein-1 (Hp1) that recognizes and directs DNA methylation to genomic regions marked by tri-methylation of histone 3 lysine 9 (H3K9me3) that is deposited by the histone methyltransferase Deficient In Methylation-5 (Dim5) [92,94]. Besides Dim2, *N. crassa* encodes another DNMT domain-containing protein of the fungal-specific class Dnmt4, named Repeat-Induced Point Mutation (RIP)-Defective (Rid), which is only active during sexual reproduction [97,212]. However, Rid has not been shown to methylate DNA, but is required for the RIP mechanism that can induce C to T mutations in duplicated genomic regions, including TEs [97,212]. Similar to *N. crassa*, the ascomycete plant pathogenic rice blast fungus *Magnaporthe oryzae* encodes orthologues of Dim2 and Rid. However, in contrast to *N. crassa* Rid, *M. oryzae* Rid displays DNA methylation activity, albeit with lower activity than Dim2 [213,214]. The opportunistic human pathogenic basidiomycete *Cryptococcus neoformans* encodes neither Dim2 nor Rid, but relies on an ortholog of Human Dnmt5 for DNA methylation [215]. *C. neoformans* Dnmt5 can methylate DNA through direct binding to H3K9me3 or through association with the Hp1 homolog Swi6 [216]. Additionally, *C. neoformans* Dnmt5 performs maintenance methylation through association with the 5mC-reader Uhrf1 that recognizes hemi-methylated CG sites [216]. Recent phylogenetic analyses of DNMTs across the fungal kingdom revealed extensive diversity in the DNMT repertoires, with only few (less than 10%) species containing either both Dim2 and Rid, or only Dnmt5, whereas many contain the combination of Dim2, Rid and Dnmt5 [217]. Thus, our knowledge on DNA methylation in fungi has been primarily based on species that are not representative for the typical DNMT repertoire of most fungi.

Verticillium dahliae is a xylem-invading, soil-borne ascomycete fungus that causes Verticillium wilt disease on hundreds of plant species [57,58]. Sexual reproduction has not been reported for *V. dahliae* that is presumed to mainly reproduce asexually [59]. Recently, we demonstrated that DNA methylation in *V. dahliae* requires Hp1 and is restricted to H3K9me3-enriched TEs that localize mainly in evolutionary stable core genomic regions that are typically shared across different *V. dahliae* strains, including centromere regions [191,194]. In contrast to stable core regions, genomic regions that are important for adaptation show extensive presence-absence polymorphisms between *V. dahliae* strains, and are therefore designated as adaptive genomic regions [59,62,63,74,191]. Many genes that play critical roles in host colonization reside in adaptive genomic regions [59,61,62,74]. Adaptive genomic regions are enriched in TEs that typically lack DNA methylation, which corresponds with increased transcriptional activity when compared with TEs in the core genome [191]. Interestingly, the transcriptional activity of TEs seems instrumental for the evolution of adaptive genomic regions [74], indicating that TEs in the core genome may carry DNA methylation to supress their transcriptional activity and to prevent genomic alterations that might reduce fitness. In this study, we investigated the contribution of various putative components of the methylation machinery on the physiology and biology of *V. dahliae* by performing bisulfite sequencing (BS-seq), transcriptomic analysis (RNA-seq), and functional studies on DNA methylation-associated genes.

Results

The genome of *V. dahliae* encodes three putative DNA methyltransferases

Putative DNMTs in *V. dahliae* were identified using homology searches to known fungal DNMTs. We selected representative basidiomycete, ascomycete and phycomycete fungi that were previously shown to have DNA methylation, as well as ascomycete *Fusarium* species that are related to *Verticillium*. The predicted proteomes of the selected species were searched with a Hidden Markov Model (HMM) pfam model (PF00145) that is characteristic for Dnmt1, Dim2, Rid and Dnmt5. Whereas *N. crassa, M. oryzae* and *C. neoformans* possess either a combination of *Dim2* and *Rid*, or only *Dnmt5*, our analyses showed that several ascomycete species, including all ten species of the *Verticillium* genus, encode all three DNMTs (Fig. 1, Fig. S1). Thus, *Verticillium* spp. encode the most commonly shared DNMT complement as observed in ascomycete fungi [217].



	Dim2	Dnmt1	Rid	Dnmt5
Phycomyces blakesleeanus	●	●	○	○
Cryptococcus neoformans	○	○	○	●
Postia placenta	○	●	○	●
Laccaria bicolor	○	●	○	●
Coprinopsis cinerea	○	●	○	●
Tuber melanosporum	●	○	●	●
Uncinocarpus reesii	●	○	●	●
Aspergillus flavus	○	○	●	●
Neurospora crassa	●	○	●	○
Magnaporthe oryzae	●	○	●	○
Fusarium graminearum	●	○	●	○
Fusarium oxysporum	●	○	●	○
Verticillium albo-atrum	●	○	●	●
Verticillium zaregamsianum	●	○	●	●
Verticillium klebahnii	●	○	●	●
Verticillium isaacii	●	○	●	●
Verticillium tricorpus	●	○	●	●
Verticillium nubilum	●	○	●	●
Verticillium longisporum	●	○	●	●
Verticillium dahliae	●	○	●	●
Verticillium alfalfae	●	○	●	●
Verticillium nonalfalfae	●	○	●	●

FIGURE 1 | Presence of putative 5mC DNA methyltransferases in various fungi. Phylogenetic tree showing a phycomycete (black line), basidiomycetes (green lines) and ascomycetes (blue lines). Filled circles indicate presence of the corresponding DNA methyltransferase as identified in Fig. S1.

DNA methyltransferase mutants are not affected in growth and virulence

As *V. dahliae* encodes three potential DNA methyltransferases, we sought to determine their activity and impact on development and virulence. To this end, we constructed deletion mutants for each DNMT gene, Δ*Dim2*, Δ*Dnmt5* and Δ*Rid*, as well as the Δ*Dim2*Δ*Dnmt5* double mutant in *V. dahliae* strain JR2. We furthermore generated the H3K9 histone methyltransferase deletion mutant Δ*Dim5* that is H3K9me3 deficient (Fig. S2), and used Δ*Hp1* [191], the DNA methyltransferase-complex member that recognizes H3K9me3 [92,94].

Growth impacts were assessed for each strain under axenic growth as determined by colony size, spore production and morphology. Whereas all DNMT mutants displayed similar growth rates, spore production and colony morphology when compared with the wild-type strain, both Δ*Hp1* and Δ*Dim5* displayed decreased radial growth when compared with the wild-type and complementation strains (Fig. 2A, C, Fig. S3). However, whereas Δ*Hp1* produced statistically significant fewer spores, Δ*Dim5* produced similar amounts of spores as wild-type

V. dahliae when also considering their respective colony sizes (Fig. 2B). This is likely due to Δ*Hp1* growing relatively flat, similar to wild-type *V. dahliae*, while Δ*Dim5* colonies display a severely crinkled surface, leading to an increased surface area on the same area of cultivation medium (Fig. 2C). Both Δ*Hp1* and Δ*Dim5* displayed reduced pigmentation when compared with wild-type *V. dahliae* (Fig. 2C).

The deletion mutants were also assessed for growth under abiotic stress conditions by axenically culturing all the strains at elevated temperature, or in the presence of osmotic, oxidative and genotoxic stress agents. Under these conditions, the DNMT mutants grew similar as the wild-type strain, while both Δ*Hp1* and Δ*Dim5* displayed reduced growth (Fig. 2D, Fig. S4). Interestingly, however, Δ*Hp1* grew similar as the wild-type strain when exposed to the genotoxic compound phleomycin despite its growth retardation under all other conditions tested (Fig. 2D, Fig. S4).

The ability to infect tomato plants was also assessed for all mutants. Tomato plants inoculated with any of the DNMT mutants displayed severe stunting at a level similar to plants inoculated with the wild-type strain (Fig. 2F). Fungal biomass measurements on the infected plants confirmed that fungal colonization by the DNMT mutants was similar to that of the wild-type strain (Fig. 2E). In contrast, Δ*Hp1* and Δ*Dim5* displayed significantly reduced tomato infection, evidenced by a similar canopy area of plants inoculated with these mutants when compared with mock-inoculated plants, as well as by the finding that inoculated plants contained only low amounts of fungal biomass (Fig. 2E). Arguably, the observation of significantly reduced plant infection for both Δ*Hp1* and Δ*Dim5* should be attributed to their compromised growth characteristics (Fig. 2A).

Dim2 is the main DNA methyltransferase in *V. dahliae*

To determine the role of the putative DNMTs in cytosine methylation in *V. dahliae*, whole-genome bisulfite sequencing was conducted on the wild-type strain, along with the *DNMT* and *Hp1* mutants. We recently reported that wild-type *V. dahliae* displays relatively low levels of DNA methylation, with an average of ~0.4% methylation in CG and CHG context and essentially no DNA methylation in CHH context [191]. DNA methylation in *V. dahliae* is restricted to particular inactive TEs that locate in condensed, H3K9me3-enriched, chromatin regions in the core genome, including those localized in centromeres (Fig. 3A) [191,194]. We furthermore showed that the Δ*Hp1* mutant lost all DNA methylation, indicating that Hp1 is required for cytosine methylation and *V. dahliae* DNMTs cannot methylate DNA independently [191].

To study the extent to which the different *V. dahliae* mutants lost DNA methylation, we compared the bisulfite sequencing patterns over the genome in 10 kb windows and assessed the amount and location of hypomethylated windows when compared with the wild-type methylation pattern. Of the DNMT deletion mutants, Δ*Dim2* showed considerable loss of cytosine methylation, having 100 and 61 hypomethylated windows in the CG and CHG context, respectively (Table 1). As there is little methylation in CHH context, we combined the methylation data for CG and CHG context and also determined hypomethylation for the contexts simultaneously. This combination optimizes the number of potential methylated cytosines per window and therefore better captures differential methylation. In the combined contexts we observed 97 hypomethylated windows that locate at regions that have relatively

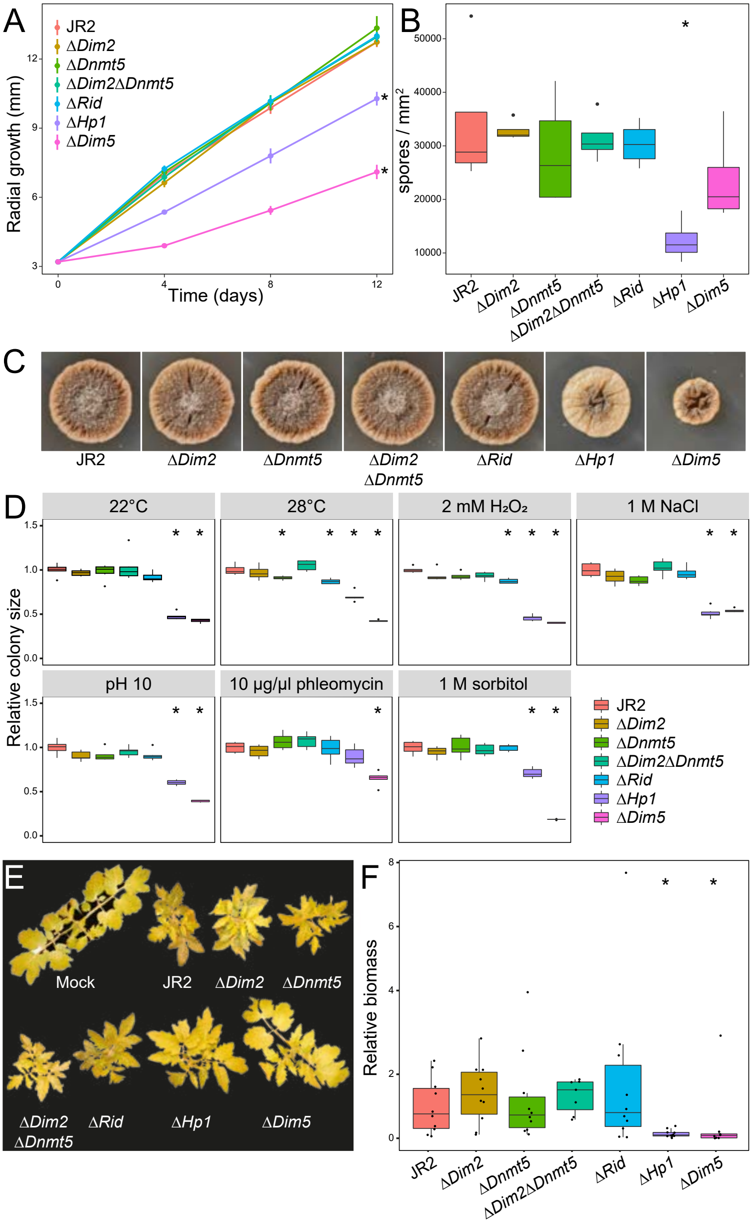


FIGURE 2 | DNMT mutants of *Verticillium dahliae* do not show altered growth under axenic conditions, stress, or host colonization. A) Radial growth of wild-type and mutants over 12 days, with B) number of spores produced per mm2 of colony and C) pictures showing representative colony morphology after 12 days of growth. D) Colony area of wild-type and mutants subjected to various stress agents, relative to average colony area of wild-type (see Fig. S4). E) Representative pictures of infected tomato plants at 21 days after inoculation, with F) biomass of wild-type and mutants, relative to wild-type infection. Statistically significant differences from wild-type (Wilcoxon Signed Rank, p < 0.01) are indicated with asterisks. For (A), statistical tests were only performed on colony diameter at 12 dpi.

high methylation percentages in wild-type *V. dahliae* (Table 1, Fig. 3, Fig. S8). Notably, additional regions showed reduced DNA methylation in Δ*Dim2*, yet these were not classified as hypomethylated because the methylation level was already low in the wild-type and therefore did not meet our criteria for calling hypomethylated regions (see methods for details). In contrast to the results for Δ*Dim2*, Δ*Dnmt5* and Δ*Rid* largely retained DNA methylation with only three and twelve windows being hypomethylated in CG context and fifteen and seventeen windows in CHG context, respectively (Table 1). When assessing CG and CHG methylation combined, Δ*Dnmt5* and Δ*Rid* have one and eight hypomethylated windows, respectively. Additionally, their genome-wide DNA methylation patterns are similar to the wild-type with no obvious loss of DNA methylation peaks (Fig. 3A). Thus, both Dnmt5 and Rid may contribute to DNA methylation in *V. dahliae*, albeit to a much lower degree as Dim2. The Δ*Dim2*Δ*Dnmt5* double mutant as well as Δ*Hp1* showed similar cytosine methylation levels over the genome as Δ*Dim2* and had hypomethylation of 93 and 99 windows in CG context and 59 and 65 windows in CHG context, respectively, and 89 and 92 hypomethylated windows when combining CG and CHG methylation data (Table 1, Fig. 3A). Loss of methylation in the Δ*Dim2,* Δ*Dim2*Δ*Dnmt5* and Δ*Hp1* mutants largely occurred in the same genomic regions, as 84 of the hypomethylated windows were shared between the mutants (Fig. 3B). Even though the chromosome plots of the Δ*Dim2*Δ*Dnmt5* double mutant as well as Δ*Hp1* are similar to those of Δ*Dim2*, the few bins with slightly elevated methylation levels in Δ*Dim2* have decreased further (Fig. 3A). This finding suggests that Dnmt5 has DNA methylation activity on particular genomic regions, albeit at a lower level. However, Δ*Dnmt5* does not display reduced methylation at the regions that remain slightly methylated in Δ*Dim2* (Fig. 3A). Thus, if Dnmt5 has DNA methylation activity, it is redundant and secondary to the DNA methylation activity of Dim2. No windows were hypomethylated for CHH in any of the mutants (Table 1, Fig. S7). The few bins with low levels of DNA methylation that locate in adaptive genomic regions behave similar as those in core regions, in that they are hypomethylated in Δ*Dim2*, Δ*Hp1* and Δ*Dim2*Δ*Dnmt5* (Fig. 3A). These results show that the methyltransferase Dim2 is responsible for the vast majority of detectable DNA methylation in *V. dahliae*.

TABLE 1 | Number of 10 kb windows that are hypomethylated in *Verticillium dahliae* DNMT mutants relative to those in the wild-type strain.

Genotype	Hypo-methylated windows (CG)	Hypo-methylated windows (CHG)	Hypo-methylated windows (CHH)	Hypo-methylated windows (CG and CHG)
Δ*Dim2*	100	61	0	97
Δ*Dnmt5*	3	15	0	1
Δ*Dim2*Δ*Dnmt5*	93	59	0	89
Δ*Rid*	12	17	0	8
Δ*Hp1*	99	65	0	92

Previous research shows that methylated TEs have accumulated more C-T mutations than non-methylated TEs [74,191]. As we observe that methylation mainly occurs in CG and CHG contexts, we investigated whether cytosine methylation is directly involved in C-T mutation. If methylated cytosines are more likely to mutate, we would expect that methylated TEs have specifically lost cytosines in CG and CHG context, while CHH sites remain intact.

Unexpectedly, however, we observe that methylated TEs only display significant depletion of CHG sites, whereas CG sites occur as frequently as would be expected based on the sequence composition of the TE (Fig. S9). In contrast, non-methylated TEs have slightly fewer CG sites than expected based on their sequence composition, yet they do not have reduced CHG sites (Fig. S9). For both methylated, as well as non-methylated TEs, we observe that CHH sites occur as frequent as would be expected based on sequence composition. Thus, although methylated TEs show increased C-T mutations [191], these mutations do not affect all methylated cytosines, as they are largely restricted to cytosines in CHG context.

We compared *V. dahliae* Dnmt5 to the homolog in *C. neoformans,* where it is the sole active DNA methyltransferase [216]. The two proteins share only 18% sequence similarity, but do share similar domain structures, except that the *V. dahliae* Dnmt5 lacks the N-terminal chromo-shadow domain found in *C. neoformans* Dnmt5 (Fig. S10). This domain is responsible for the direct binding to H3K9me3, and this histone mark is required for DNA methylation, which could explain why we observed little Dnmt5 contribution to DNA methylation in *V. dahliae*. However, *C. neoformans* Dnmt5 can also bind H3K9me3 indirectly through Hp1 [216], and it is not clear if this is also the case for *V. dahliae*. The lack of DNA methylation by Dnmt5 cannot be explained by transcriptional activity, as *Dnmt5* is expressed higher than *Dim2* during cultivation in PDB (Fig. 3B).

Loss of DNA methylation does not affect transcriptional regulation

While the *Dim2* mutant loses nearly all DNA methylation (Table 1, Fig. 3A), it displays wild-type-like growth *in vitro*, under stress conditions as well as during infection (Fig. 2), suggesting that DNA methylation is not essential under these conditions. However, given that DNA methylation is mainly restricted to TEs [191], we anticipated that loss of DNA methylation could result in activated transcription at TEs. To address this, we performed RNA sequencing on axenically grown cultures of all DNMT mutants, as well as the *Hp1* and *Dim5* mutants. Consistent with the lack of DNA methylation at coding regions, all three single DNMT mutants and the double mutant showed differential expression of only few genes (< 10) (Fig. 4A). Unanticipatedly, the four DNMT mutant strains similarly showed differential expression of only a few TEs (< 10) (Fig. 4B). In contrast, the Δ*Hp1* and Δ*Dim5* mutant strains showed considerable differential expression of genes and TEs (Fig. 4A). In total, 1,661 genes were induced and 663 repressed in Δ*Hp1*, and 1,617 genes were induced and 781 are repressed in Δ*Dim5* when compared with wild-type (Fig. 4A). Furthermore, 261 TEs were induced and 23 were repressed in Δ*Hp1*, whereas 241 TEs were induced and 47 were repressed in Δ*Dim5* when compared with wild-type (Fig. 4B). Analysis of the induced genes and TEs revealed a large overlap between the mutants, with 1,207 out of 1,617 (~75%) of the induced genes and 166 out of 241 (~69%) of the induced TEs shared between the two mutants (Fig. 4A, B). This overlap is likely related to the functional link between these heterochromatin components, as Hp1 directs DNA methylation to H3K9me3 deposited by Dim5 [94]. To study whether the genes activated in Δ*Hp1* and Δ*Dim5* represent specific biological functions, we performed GO enrichment analysis on the 1,207 induced genes. Interestingly, genes encoding secreted proteins and proteins involved in transport and metabolic processes were overrepresented among the induced genes (Fig. 4C), suggesting that the mutants impact expression of genes with roles in responses to the environment.

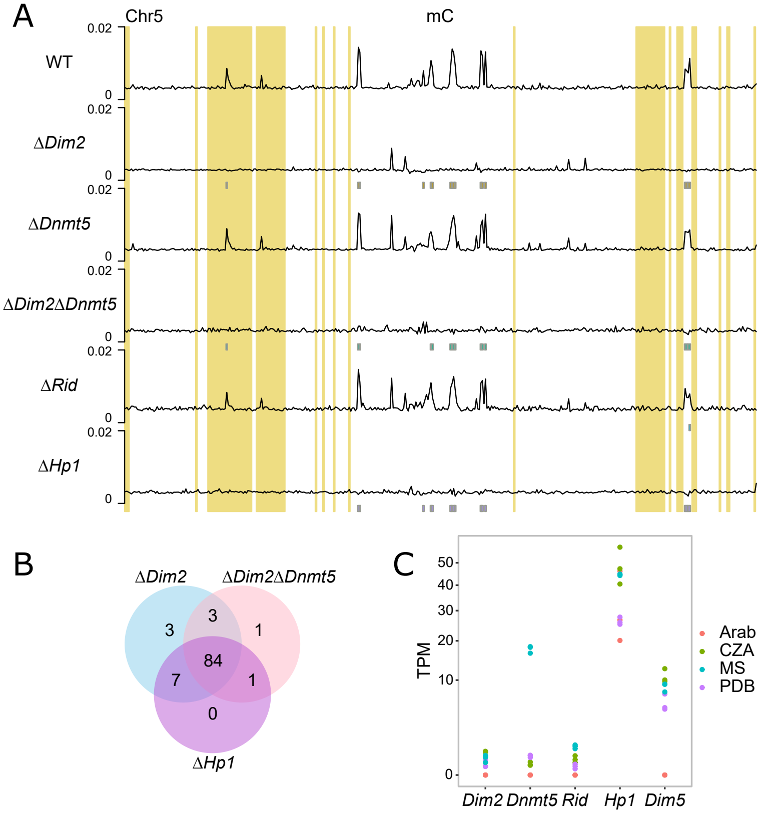


FIGURE 3 | Dim2 is the main DNA methyltransferase in *V. dahliae*. A) Whole-chromosome plot displaying the fraction of methylated cytosines for non-overlapping 10 kb windows for wild-type, and DNMT and Hp1 deletion mutants with chromosome 5 as an example. Grey boxes, displayed below the DNA methylation tracks, indicate the hypomethylated windows compared to the wild-type strain in CG and CHG context from Table 1. Previously defined adaptive genomic regions [191] are highlighted in yellow. B) Overlap of hypomethylated windows in mutant strains showing severe loss of methylation. C) Expression (TPM values) of DNA methyltransferase genes Dim2, Dnmt5 and Rid, as well as Hp1 and Dim5 of *V. dahliae* strain JR2 cultured in Czapec-Dox medium (CZA), half strength Murashige-Skoog medium (MS) and potato dextrose broth (PDB), and during Arabidopsis infection at 21 days post inoculation (Arab), in triplicates.

As Hp1 functions downstream of Dim5 during DNA methylation, we expect that the induction of genes as well as TEs in both mutants may be due to reduced recruitment of repressive complexes to previously silenced chromatin regions because either H3K9me3 or Hp1 is lacking. Based on H3K9me3 ChIP-seq, we found that approximately 2.1 Mb (~6%) of the genome is associated with H3K9me3 that occurs in 621 enriched genomic regions, of which 38 are larger than 10 kb (Fig. S12). To study whether the induced genes and TEs localize in these H3K9me3 domains, we investigated the occurrence of physical clustering of the 1,207 induced genes and the 166 induced TEs. Our hypothesis was that Dim5 deposited H3K9me3 and associated Hp1 mediate transcriptional silencing in physical proximity to H3K9me3 domains.

As such, we expected that genes and TEs induced in Δ*Hp1* and Δ*Dim5* occurred in clusters and that these clusters would be in proximity to H3K9me3-enriched genomic regions. We identified 58 clusters containing 526 of the 1,207 (~44%) induced genes and four clusters containing 37 of the 166 (~22%) induced TEs, which is more than expected by chance, as measured from 1,000 random sets of 1,207 genes (p<0.001) (Fig. S12) and 166 TEs (p = 0.024) (Fig. S13). H3K9me3 domains contain numerous TE copies (1,034 out of 2,574, ~40%) and only few genes (76 out of 11,426, ~0.6%) (Fig. 4D, Fig. S14) [191]. Next, we calculated the distance of each gene and TE to the closest H3K9me3 domain and associated this to induced and non-induced genes and TEs. When considering the smallest distance to H3K9me3 domains, we observed that induced genes are slightly closer to H3K9me3 domains than non-induced genes (Fig. 4E). In contrast, induced TEs are slightly further from H3K9me3 domains than non-induced TEs (Fig. 4E). Additionally, when considering presence of TEs in H3K9me3 domains, we observed that significantly fewer induced TEs, 48 out 165 (29.1%), than non-induced TEs, 987 out of 2,409 (41.0%), locate in H3K9me3 domains (Fisher's exact test, p = 0.0126). The relatively minor enrichment of induced genes near H3K9me3 domains and the enrichment of induced TEs away from H3K9me3 domains suggests that the transcriptional changes in Δ*Hp1* and Δ*Dim5* are not due to reduced H3K9me3 and Hp1 association. As we demonstrated that the induced genes and TEs occur clustered in the genome (Fig. S12, S13), we asked whether the clusters localize in specific genomic regions. As we found that induced genes are enriched for genes encoding secreted proteins and proteins involved in metabolic processes (Fig. 4C), we speculated that the induced genes may be involved in processes related to plant infection. Such genes are typically located in adaptive genomic regions of *V. dahliae*, which are enriched in TEs but are not associated with H3K9me3 [59,191]. Therefore, we also tested whether the induced genes and TEs locate in proximity to adaptive genomic regions. Intriguingly, genes and TEs induced in Δ*Hp1* and Δ*Dim5* are significantly closer to adaptive genomic regions than non-induced genes and TEs (Fig. 4D,F). Additionally, 180 out of 1,207 (14.9%) and 59 out of 165 (35.8%) of the induced genes and TEs locate in adaptive genomic regions, which is significantly more than the 818 out of 10,219 (8.0%) and 474 out of 2,409 (19.7%) of non-induced genes (Fisher's exact test, p <0.00001) and TEs (Fisher's exact test, p <0.00001). Consequently, both genes and TEs that are induced in Δ*Hp1* and Δ*Dim5* reside significantly closer to adaptive genomic regions than non-induced genes and TEs (Fig. 4F). Since adaptive genomic regions are not associated with H3K9me3, these findings suggest that the observed transcriptional changes are not directly related to loss of Hp1 binding at H3K9me3 domains, but rather through pleiotropic effects affecting transcription throughout the genome, and especially at adaptive genomic regions.

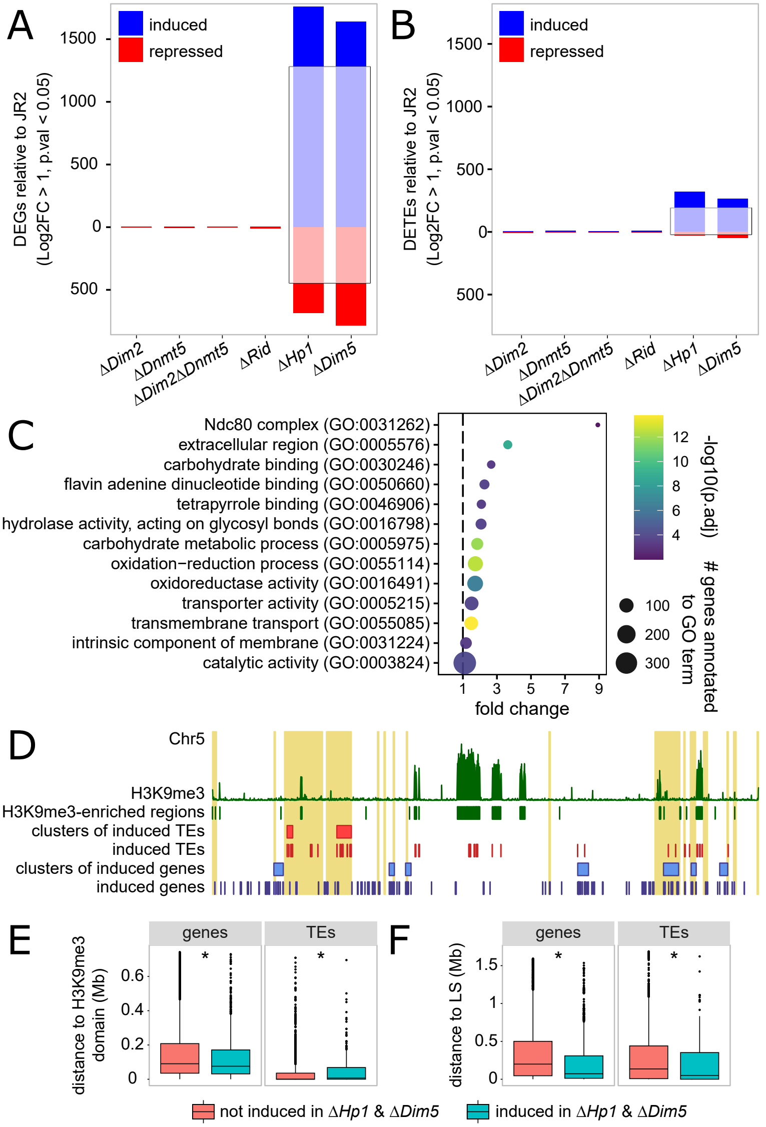


FIGURE 4 | Genes and TEs that are induced in the *Verticillium dahliae Hp1* and *Dim5* mutants do not associate with H3K9me3-marked chromatin. Differentially expressed genes (A) and TEs (B) in the mutants relative to wild-type. Induced genes and TEs are indicated in blue, repressed genes and TEs in red. The number of genes and TEs that are induced and repressed in both the *Hp1* and *Dim5* mutants are indicated by opaque coloring in black rectangles. (C) Gene Ontology (GO) terms that are enriched (fold change > 1, p.adj < 0.01) in the set of genes that are induced in both the *Hp1* and *Dim5* mutant. (D) Whole-chromosome plot displaying the location of induced genes (in blue) and TEs (in red) on chromosome 5 as an example. Clusters of induced genes and TEs are indicated as blue and red rectangles, respectively. H3K9me3-ChIP signal along the chromosome is indicated in green in the upper track. adaptive genomic regions [191] are highlighted in yellow. The minimal distance of genes and TEs to H3K9me3-enriched genomic regions (E) and to adaptive genomic regions (F). Asterisks indicate statistical differences (Wilcoxon signed rank test, p<0.01) between genes and TEs induced in both the *Hp1* and *Dim5* mutants and those that are not induced in the mutants.

Discussion

DNA methylation is essential for proper functioning of nuclear processes in many organisms [208], but various fungal species have lost or degraded their machinery for DNA methylation [217]. The most commonly found combination of DNMTs in ascomycete genomes is the presence of *Dim2*, *Rid* and *Dnmt5* [217]. As *Dim2* is the main DNA methyltransferase gene in fungal species that lack *Dnmt5*, and vice versa, it is relevant to study the importance of these DNA methyltransferase genes in fungal species that carry both. The fungal pathogen *Z. tritici* carries all three DNMTs and, similar to *V. dahliae*, loss of *Dim2* almost completely abolishes DNA methylation [218], indicating that Dnmt5 and Rid have little to no DNA methylation activity. However, a low residual DNA methylation signal remains in *Dim2* mutants of *V. dahliae* and *Z. tritici*, which may be due to a low degree of Dnmt5 activity [218]. Our results indicate that *Dnmt5* is more highly expressed during growth in nutrient-limited media, a type of environmental stress. It is possible *Dnmt5* may be more active and cause differential DNA methylation during specific growth conditions not tested here, an occurrence that has been observed for DNMTs in several plant and animal species [219,220]. Differential expression of DNMTs has previously been observed in the entomopathogenic ascomycete fungus *Cordyceps militaris* that contains orthologues of *Dim2* and *Rid* [221]. Whether *V. dahliae* Dnmt5 plays such a role requires further study.

The importance of DNA methylation in fungi that are able to perform DNA methylation remains unclear. Deletion of functional components of DNA methylation did not result in clear phenotypic alterations in *N. crassa* or the necrotrophic plant pathogenic fungus *Botrytis cinerea*, while deletion of *Dim2* in *M. oryzae* leads to aberrant colony morphology and compromised conidiospore formation [93,213,222]. In *Z. tritici*, strains collected in the centre of origin of its wheat host carry DNA methylation, while strains collected in Europe contain mutated *Dim2* copies that lack DNA methylation [218]. The *Z. tritici* strains that lack a functional copy of *Dim2* are at least as virulent as strains that perform DNA methylation [223], suggesting that the recent loss of DNA methylation in these *Z. tritici* strains does not negatively affect their infection biology. Our study reveals that DNA methylation in *V. dahliae* is not essential for growth and infection. Moreover,

we show that loss of DNA methylation does not result in altered expression of genes or TEs, an observation that could be explained by DNA methylation co-localizing with H3K9me3, which is likely sufficient for heterochromatin formation and transcriptional silencing in the absence of DNA methylation. This is further supported as the H3K9me3-deficient *V. dahliae Dim5* mutant showed significant differential expression of genes and TEs. Interestingly, some genes and TEs induced in this mutant were located in adaptive genomic regions that are not labelled with H3K9me3 in wild-type *V. dahliae*, suggesting that the removal of H3K9me3 leads to pleiotropic effects in unrelated genomic regions. Similar effects on gene and TE expression occurs in the *V. dahliae Hp1* mutant, indicating that the differential expression observed in the *Hp1* and *Dim5* mutants are related to disrupted Hp1 functioning. In *N. crassa*, and also fission yeast *Schizosaccharomyces pombe* that lacks DNA methylation, Hp1 was found to be involved in the formation of H3K9me3-associated heterochromatin [94,224,225]. As such, it is possible that the transcriptional changes in the *V. dahliae Hp1* and *Dim5* mutants are due to pleiotropic effects of chromatin de-condensation. In line with such pleiotropic effects on chromatin architecture, deletion of *Dim5* in *Z. tritici*, and *Dim5* and *Hp1* in *N. crassa* leads to re-localization of H3K27me3 to previous H3K9me3 domains [54,198].

Previously, we showed that the adaptive genomic regions in *V. dahliae* are enriched for H3K27me3 [191], suggesting that the observed transcriptional induction of adaptive genomic region-localized genes and transposons in the *V. dahliae Hp1* and *Dim5* mutants may be due to altered localization patterns of H3K27me3. Considering that experimental and natural loss of DNA methylation in various fungi does not seem to affect their proliferation, it is remarkable that the vast majority of fungal species have retained DNA methylation. One explanation for the role of DNA methylation in fungi, which accounts for the lack of reported phenotypes, is that it serves in maintaining genome integrity during evolution. In this way, DNA methylation does not functionally regulate transcription *per se*, but works in conjunction with H3K9me3 to minimize the impact of TEs in the genome. One possible mechanism is that DNA methylation may have persistent effects on TE activity through spontaneous deamination of methylated cytosines, resulting in C to T mutations [226]. The deamination process is considered an important driver of mutations in *Z. tritici* TEs as recently shown in an experimental evolution experiment in which a DNA methylation competent strain had increased in C to T mutations compared to the strain lacking DNA methylation [218]. Interestingly, in *V. dahliae* we previously observed that TEs that carry DNA methylation contain more C to T mutations than unmethylated TEs [191]. Typically, such C to T mutations are also caused by the RIP mechanism, which relies on the Rid DNA methyltransferase that is active during sexual cycles in *N. crassa* [97,212]. However, since *V. dahliae* is presumed to reproduce asexually, it may be more likely that C to T mutations in TEs are caused by spontaneous deamination. These results support that the main role for DNA methylation in fungi might be to aid in TE sequence degradation over time, not to directly supress transcriptional activity. Alternatively, it is possible DNA methylation is important for inhibiting transcriptional activity of TEs during specific developmental or cell-cycle stages which have not been reported or observed to date.

Our results show that although *V. dahliae* encodes multiple DNMTs, only *Dim2* seems to be essential for DNA methylation. As *Dim2*, *Dnmt5* and *Rid* are wide-spread among ascomycetes, it is likely that their combined presence is an ancestral state [217]. Even though only four ascomycete species have been studied with respect to the contribution of their DNMTs to DNA methylation so far, these studies suggest that species, irrespective of the presence or absence of *Dnmt5,* utilize Dim2 as the main DNMT (this study; [93,213,218]). Additional research is needed to determine if Dnmt5 and Rid play a role in DNA methylation, or possibly in other DNA-associated pathways, or if their presence is the remnant of an ancestral state that is not strongly selected against.

Materials and methods

Assessment of DNMT occurrence

To assess the presence of DNA methyltransferases in a selection of fungal species with confirmed DNA methylation performance, we downloaded predicted proteomes of *Aspergillus flavus* strain NRRL_3357 (AFL2T), *Coprinopsis cinerea* strain Okayama-7#130 (CC1G), *Cryptococcus neoformans* strain H99 (CNAG), *Fusarium graminearum* strain PH-1 (FGSG), *F. oxysporum* strain 4287 (FOXG), *Laccaria bicolor* strain S238N-H82 (lacbi2), *Magnaporthe oryzae* strain MG8 (MGG), *Neurospora crassa* strain OR74a (NCU), *Phycomyces blakesleeanus* strain NRRL 1555(-) (Phybl2), *Postia placenta* strain MAD698 (pospl1), *Tuber melanosporum* strain Mel28 (Tubme1), *Uncinocarpus reesii* strain 1704 (URET) and the *Verticillium albo-atrum* PD747, *V. alfalfae* PD683, *V. dahliae* JR2, *V. isaacii* PD618, *V. klebahni* PD401, *V. longisporum* PD589, *V. nonalfalfae* T2, *V. nubilum* 397, *V. tricorpus* PD593 and *V. zaregansianum* PD739. The predicted proteomes were scanned for the presence of a DNA methyltransferase domain with hmmsearch with --cut_ga option. Identified proteins were visually inspected. To check for presence of additional not-annotated DNA methyltransferase homologs of *Dim-2*, *Dnmt5* and *Rid* that were initially not annotated in the predicted proteomes, we manually assessed the genomes using TBLASTN and Augustus. Phylogenetic trees were constructed using IQ-tree.

Fungal growth and mutant generation

V. dahliae strain JR2 (CBS 143773 [122]) was maintained on potato dextrose agar (PDA) (Oxoid, Thermo Scientific, CM0139) and grown at 22°C in the dark. The Δ*Dim2*, Δ*Dnmt5,* Δ*Rid* and Δ*Dim5* single deletion mutants and the Δ*Dim2*-Δ*Dnmt5* double mutant were constructed as previously described [123]. Briefly, for all genes except *Dnmt5*, genomic DNA regions flanking the 5' and 3' ends of the coding sequences were amplified with PCR using primers listed in Table S1 and cloned in to the pRF-HU2 vector [124], using USER enzyme following the manufacturer's protocol (New England Biolabs, MA, USA). For *Dnmt5*, the 5' and 3' amplicons were cloned into vector pRF-NU2, a custom-made pRF-HU2 variant, containing the NAT-cassette for selection on nourseothrycin. Sequence-verified vectors were transformed into *Agrobacterium tumefaciens* strain AGL1 used for *V. dahliae* conidiospore transformation as described previously [123]. *V. dahliae* transformants that appeared on hygromycin B or nourseothrycin (for *Dnmt5*) were transferred to fresh PDA supplemented with hygromycin B or nourseothrycin after five days. Putative transformants were screened using PCR to verify deletion of the target gene sequence (Table S3) when compared with positive amplification from the wild-type strain. To further confirm integration of the selectable marker at the locus of interest, another round of PCR was conducted in which one primer was position adjacent to the deleted genomic region, and the other primer was designed to bind a portion of the inserted vector DNA (Table S3). In this manner, deletion mutants were confirmed to lack the gene of interest and contain the selectable marker at the locus of interest. Generation of the *Hp1* deletion mutant was conducted in the same way and described previously [191]. Complementation vectors were generated by amplifying the coding region of *Dim2*, *Hp1* and *Dim5* from genomic DNA using primers listed in Table S1, and ligating the amplicons into PacI-digested pFBT-005 vector using the NEBuilder HiFi DNA Assembly Cloning Kit (New England Biolabs, MA, USA). Fungal transformations were performed as described above and obtained colonies were screened by PCR to verify presence of target gene (Table S2).

Growth and inoculation assays

To check for aberrant growth phenotypes of the generated mutants, all strains were cultured as described above. To this end, conidiospores were harvested in sterile water and brought to a final concentration of 10^6 conidiospores per mL. Subsequently, 10 µL of conidiospore suspension, containing 10^4 conidiospores, was deposited in the middle of a 90 mm Petri dish containing 20 ml of PDA. Plates were stored at 22˚C in the dark and colony diameter was measured in perpendicular directions after 4, 8 and 12 days of growth. After twelve days of growth all newly formed conidiospores were harvested in 1 mL of water and counted using a hemocytometer.

Stress assays were performed by spotting 5 µL conidiospore suspension containing $5x10^3$ conidiospores on PDA without supplement, or on PDA supplemented with 1 M NaCl, 1 M Sorbitol, 2 mM H_2O_2 or 10 µg/µL phleomycin, and on PDA adjusted to pH 10. Plates were incubated at 22˚C in the dark, apart from one set of PDA plates without supplement that was incubated at 28˚C to assess heat stress responses. Pictures were taken after 6 or 10 days, depending on wild-type colony development, and colony size was determined using ImageJ software with custom settings for each stress condition.

Infection assays were performed using root dip inoculation in a conidiospore suspension of 10^6 spores per mL on 10-day-old seedlings of tomato cultivar Moneymaker. Stems of infected plants were harvested at 21 days after inoculation, cut in small pieces, frozen in liquid nitrogen and ground by reciprocal shaking in a MixerMill MM 400 (Retsch, Haan, Germany). DNA was isolated incubating the ground powder with 800 µL of CTAB lysis buffer at 65˚C for 1 hour, followed by addition of 400 µL chloroform/IAA (24:1), vigorous shaking and centrifuging for 5 minutes at ~13,000 RCF. DNA was precipitated from the aqueous layer with isopropanol and the precipitate was washed with 70% ethanol. The fungal biomass in the stem tissue was determined with real-time PCR using *V. dahliae* ITS-specific and tomato GAPDH-specific primer sets (Table S4).

Bisulfite sequencing and analysis

The *V. dahliae* wild-type strain, Δ*Dim2*, Δ*Dnmt5,* Δ*Dim2* / Δ*Dnmt5,* Δ*Rid* and Δ*Hp1* were grown in potato dextrose broth (PDB) for three days, strained through miracloth (22 µm) (EMD Millipore, Darmstadt, Germany), pressed to remove excess liquid, flash frozen in liquid nitrogen and ground to powder with a mortar and pestle. Genomic DNA was isolated as described above and sent to the Beijing Genome Institute (BGI, Hong Kong, China) for bisulfite conversion, library construction and Illumina sequencing. Briefly, the DNA was sonicated to a fragment range of 100-300 bp, end-repaired and methylated sequencing adapters were ligated to 3' ends. The EZ DNA Methylation-Gold kit (Zymo Research, CA, USA) was employed according to manufacturer's guidelines for bisulfite conversion of non-methylated DNA. Lambda DNA was used as spike-in to determine conversion efficiency, which was >99% for all samples. Libraries were paired-end 100 bp sequenced on an Illumina HiSeq 2000 machine.

Whole-genome bisulfite sequencing reads were analyzed using the BSMAP pipeline (v. 2.73) and methratio script [125]. The results were partitioned into CG, CHG and CHH cytosine sites for analysis. Only cytosine positions containing more than 4 sequencing reads were included for analysis. BSMAP datasets were further analyzed using MethylKit (v. 1.12.0) [227]. Methylation levels were summarized as the number of methylated cytosines divided by the total number of sequenced cytosines per 10 kb window. Hypomethylated windows in the mutants were determined by comparing corresponding 10 kb windows between mutants and wild-type and selecting windows with meth.diff value < -1 and a qvalue < 0.01. Genome plots displaying methylation data were generated using karyoploteR (v. 1.12.4) [228].

Analysis of CG, CHG and CHH site occurrence in methylated and non-methylated TEs

The observed occurrence of CG, CHG and CHH sites were compared to the expected occurrence of these sites for each of the 2,574 TEs based on the sequence composition (CG expected: n(C) x n(G) / (N x N-1), CHG expected: n(C) x n(C+A+T) x n(G) / (N x N-1 x N-2), CHH expected: n(C) x n(C+A+T) x n(C+A+T) / (N x N-1 x N-2)). TEs were categorized as methylated if any cytosine in the sequence showed evidence for methylation in wild-type JR2.

Protein Extraction and Western Blot

Total proteins were extracted from *V. dahliae* wild-type, D*Dim2*, D*Dim5*, and Δ*hp1* grown for 10-days in potato dextrose broth at 22˚C. Mycelium was collected by straining over a double layer of miracloth, frozen in liquid nitrogen and ground with a mortar and pestle. Approximately 0.5 gram of ground mycelium was resuspended in 12 mL lysis buffer (75 mM Tris-HCl pH 7.4, 0.5 mM EDTA, 0.3 M Sucrose, 40 mM $NaHSO_3$, 10 mM $MgSO_4$, 0.5% NP-, 2 mM Phenylmethanesulfonyl fluoride (PMSF), 100 µM Leupeptin, 1 µg/mL Pepstatin), briefly vortexed and rotated at 4˚C for 15 minutes. Samples were spun at 4˚C at 10,500 g for 20 minutes, and the pellet was resuspended in 2 ml CW buffer (10 mM Tris-HCl pH 8.0, 150 mM NaCl, 1:400 β-mercaptoethanol, 2 mM Phenylmethanesulfonyl fluoride (PMSF), 100 µM Leupeptin, 1 µg/mL Pepstatin). 0.5 mL of 0.4 M sulfuric acid was added, rotated at 4˚C for 2.5 hours and centrifuged at 4˚C at 10,500 g for 15 minutes. Supernatant was added to 25 mL aceton on ice and proteins were precipitated at -20˚C overnight. Precipitates were collected by centrifugation 4˚C at 7,500 g for 15 minutes and resuspended in 300 µL 4 M urea. To assess H3K9 methylation status, approximately 15uL of total protein was added to 2x Laemmli loading buffer (4% SDS, 20% glycerol, 0.004% bromophenol blue, 125 mM Tris HCL pH 6.8), boiled at 95˚C for 1 minute, and separated using PAGE (15% polyacrylamide gel). Proteins were transferred to PVDF membranes, blocked in 5% BSA, washed twice in TBST, and incubated with 1:3500 anti-H3K9me3 antibody (#39161, Active Motif, Carlsbad, CA, USA). Blot was subsequently stripped, washed in TBST and incubated with 1:4000 anti-H3 antibody (ab1791, Abcam, Cambridge, United Kingdom).

Dnmt5 analysis and expression of DNA methyltransferase genes

The Dnmt5 amino-acid sequences of *V. dahliae* strain JR2 (VDAG_JR2_Chr1g14260) and *C. neoformans* var *grubii* strain H99 (CNAG_07552) were retrieved from EnsemblFungi, aligned using Uniprot.org/align and their domain structure was predicted using Interproscan. To assess expression of *Dim2*, *Dnmt5*, *Rid*, *Hp1* and *Dim5* in *V. dahliae* cultured in different growth media, wild-type *V. dahliae* was cultured in Czapec-Dox medium (CZA), half-strength Murashige-Skoog medium with vitamins and supplemented with 3% sucrose (MS), and potato dextrose broth (PDB) at 22˚C at 160 RPM in the dark for six days and mycelium was collected for three replicates per growth medium and ground as described above. To obtain RNA-seq data from *V. dahliae* grown in planta, three-week-old *A. thaliana* (Col-0) plants were root dip inoculated in a conidiospore suspension of 10^6 spores per mL for 10 minutes. After root inoculation, plants were grown in individual pots in a greenhouse for 21 days, 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, respectively. Total RNA from *in vitro* cultured mycelium and was isolated using Trizol (Thermo Fisher Science, Waltham, MA, USA) following the manufacturer's guidelines. Following RNA re-suspension, contaminating DNA was removed using the TURBO DNA-free kit (Ambion, Thermo Fisher Science, Waltham, MA, USA) and RNA integrity was determined by separating 2 µL of each sample on a 2% agarose gel and quantified using a Nanodrop (Thermo Fisher Science, Waltham, MA, USA) and stored at -80°C until further use. Library preparation was carried out at BGI (BGI, Hong Kong, China) and 50 bp fragments were sequenced using the BGISEQ-500 platform. Sequenced reads were mapped to *V. dahliae* strain JR2 gene annotation using Kallisto quant (settings: --single -l 50 -s 0.001 --pseudobam) to obtain normalized TPM values [155].

Transcriptional analysis of mutants in DNA methylation-associated genes

The *V. dahliae* wild-type strain JR2 [122], Δ*Dim2*, Δ*Dnmt5,* Δ*Dim2*-Δ*Dnmt5,* Δ*Rid,* Δ*Hp1* and Δ*Dim5* were cultured in PDB at 22°C at 160 RPM in the dark for six days. Mycelium collection, RNA-isolation and sequencing are performed for three replicates per growth condition as described above.

Differential gene and TE expression between mutants and wild-type was determined by mapping sequencing reads to the *V. dahliae* strain JR2 gene and TE annotation using TEtranscripts, which uses an iterative multimapping approach [122,191,229]. Genes and TEs were considered differentially expressed when they displayed log2FoldChange of < -1 and a pvalue of < 0.05. Additionally, as transcript mapping to TEs using a multimapping approach may falsely identify transcription of highly similar sequences, we also used a mapping approach using unique reads only on TEs. The unique mapping approach resulted in slightly fewer differentially expressed TEs than the iterative multimapping approach (Table S5).

Gene ontology (GO) terms were annotated to the *V. dahliae* JR2 proteome using Blast2GO (v1.4.4) [230]. GO enrichment analysis was performed using Ontologizer (v2.1), using Parent-Child-Union calculation and Benjamini-Hochberg multiple testing correction with 1000 resampling steps [231]. Fold change enrichment was calculated by dividing the fraction of genes annotated to each GO term in the study set by the fraction of genes annotated to each GO term in the population.

Clustering of induced genes and TEs was determined using CROC (settings, for genes: -w 50000 -o 10000 -m 5, for TEs: -w 100000 -o 20000 -m 5) [232]. To analyze whether induced genes or TEs display more clustering than in random gene or TE sets, 1,000 random selections of 1,280 genes and of 191 TEs were generated. These random gene and TE sets were similarly analyzed using CROC. Overlap with H3K9me3 domains and adaptive genomic regions [191] was assessed using bedtools closest (settings -d) [147].

Availability of data and materials

BSseq data of wildtype and mutants were submitted to the SRA database under the accession numbers: PRJNA592220 and PRJNA659638. RNAseq data of *in vitro* cultivation were submitted to the SRA database under the accession numbers: PRJNA592220 and PRJNA659638. RNAseq data of *Arabidopsis* infection were submitted to the SRA database under the accession number: SRP149060. H3K9me3 ChIPseq data is submitted to the SRA database under the accession number: PRJNA592220.

Funding

This work was supported by a PhD grant of the Research Council Earth and Life Sciences (project 831.15.002) to HMK and by Human Frontier Science Program Postdoctoral Fellowship (HFSP, LT000627/2014-L) to DEC. Work in the laboratories of M.F.S and B.P.H.J.T. is supported by the Research Council Earth and Life Sciences (ALW) of the Netherlands Organization of Scientific Research (NWO) and B.P.H.J.T is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2048/1 – Project ID: 390686111.

Supplementary data

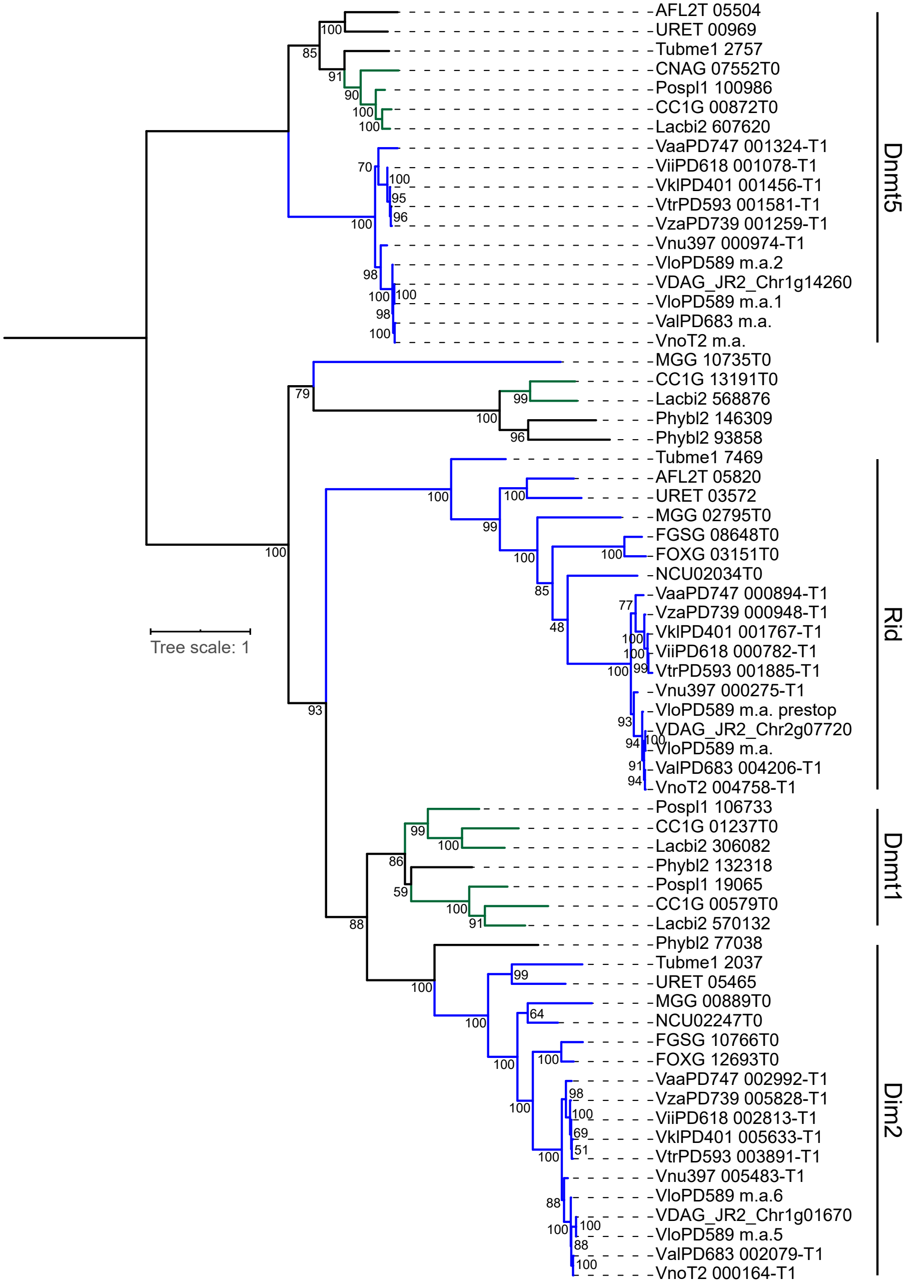


FIGURE S1 | Phylogenetic tree of DNA methyltransferases. Gene codes containing “m.a.” were manually added, as they were missed in the predicted proteomes.

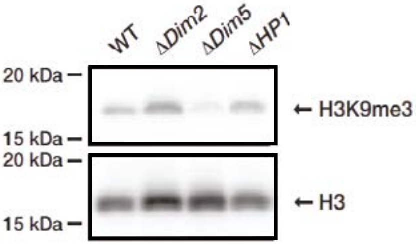


FIGURE S2 | *Verticillium dahliae* Δ*Dim5* loses H3K9me3. Western blot on nuclear protein extracts of *V. dahliae* wild-type and mutants show loss of H3K9me3 in the Δ*Dim5* deletion mutant.

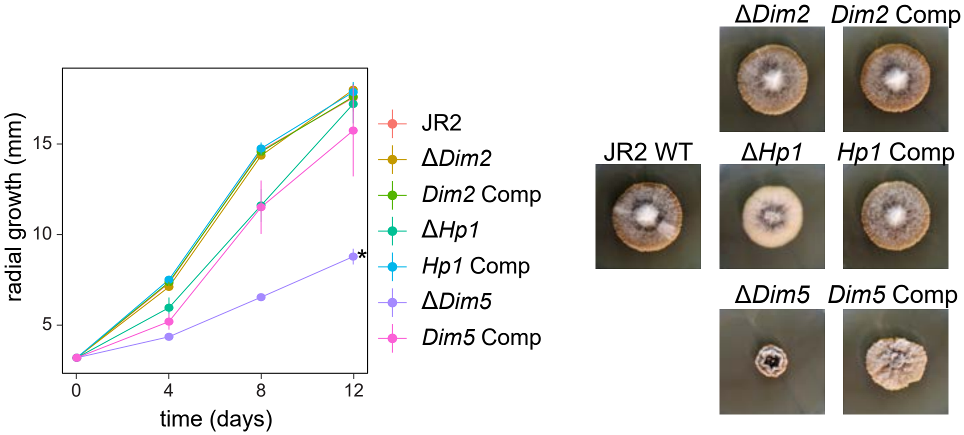


FIGURE S3 | Growth assay of complementation strains. Radial growth of wild-type, mutants and complementation strains over 12 days. Vertical lines represent the standard error of 8 measured colonies. Pictures showing representative colony morphology after 12 days of growth are shown on the right. Statistically significant differences from wild-type at 12 dpi (Wilcoxon Signed Rank, p < 0.01) are indicated with asterisks.

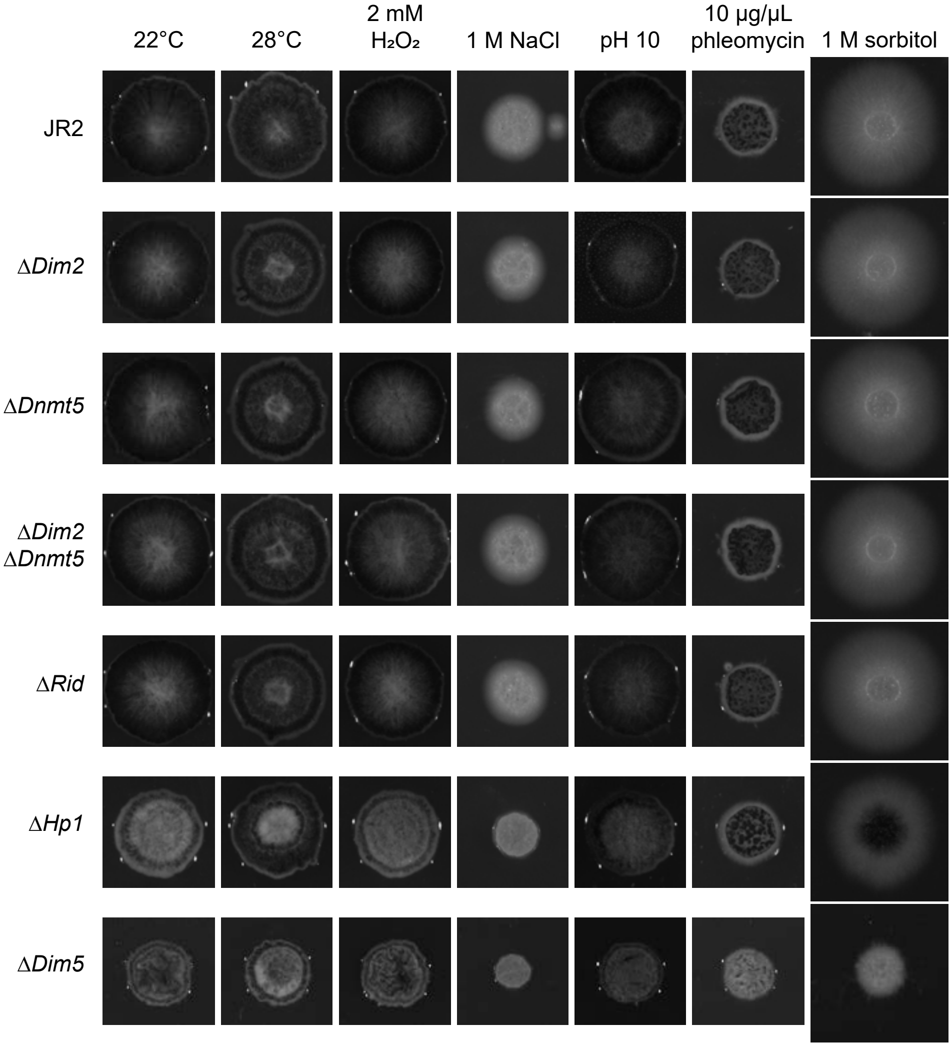


FIGURE S4 | Stress assay pictures at 10dpi. Colony photographs as taken by the ChemiDoc MP imaging system underlying data of Fig. 2D.

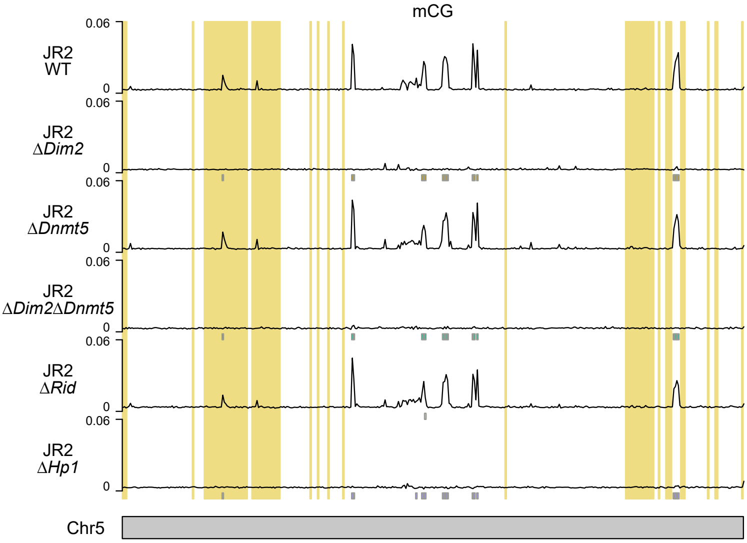


FIGURE S5 | DNA methylation in CG context. Whole-chromosome plot displaying the fraction of methylated cytosines for non-overlapping 10 kb windows in CG context for WT, and DNA methyltransferase and *Hp1* deletion mutants with chromosome 5 as an example. Grey boxes, displayed below the DNA methylation tracks, indicate the hypomethylated windows in CG context from Table 1. Previously defined LS regions [191] are highlighted in yellow.

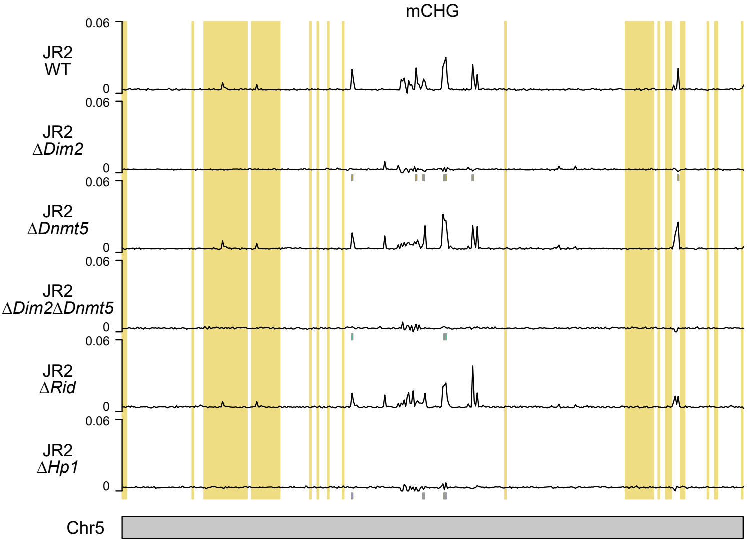


FIGURE S6 | DNA methylation in CHG context. Whole-chromosome plot displaying the fraction of methylated cytosines for non-overlapping 10 kb windows in CHG context for WT, and DNA methyltransferase and *Hp1* deletion mutants with chromosome 5 as an example. Grey boxes, displayed below the DNA methylation tracks, indicate the hypomethylated windows in CHG context from Table 1. Previously defined LS regions [191] are highlighted in yellow.

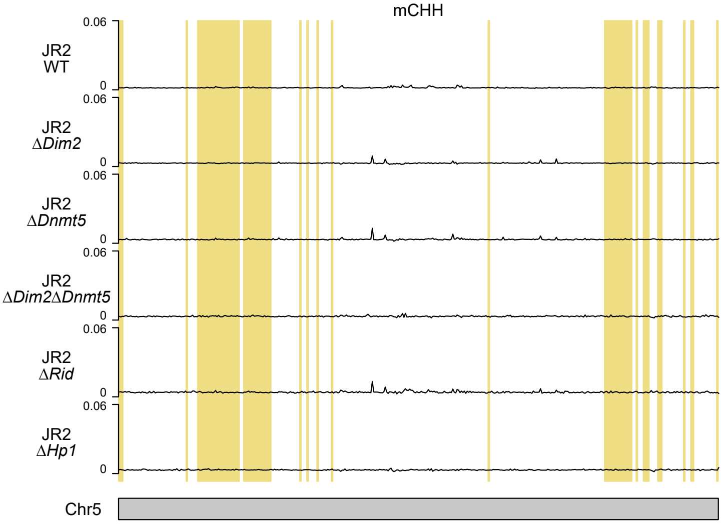


FIGURE S7 | DNA methylation in CHH context. Whole-chromosome plot displaying the fraction of methylated cytosines for non-overlapping 10 kb windows in CHH context for WT, and DNA methyltransferase and *Hp1* deletion mutants with chromosome 5 as an example. Colored boxes, displayed below the DNA methylation tracks, indicate the hypomethylated windows in CHH context from Table 1. Previously defined LS regions [191] are highlighted in yellow.

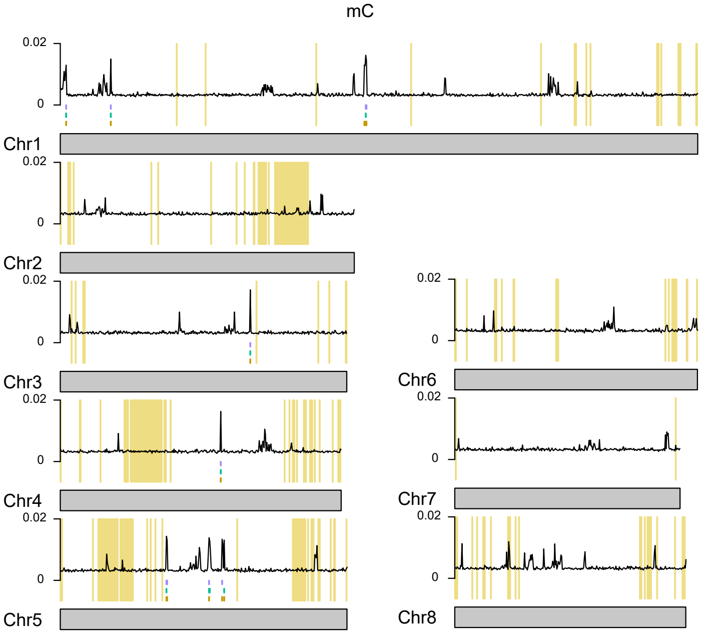


FIGURE S8 | DNA methylation over the genome. Whole-chromosome plot displaying the fraction of methylated cytosines for non-overlapping 10 kb windows for wild-type over the whole genome. Colored boxes displayed below the DNA methylation track indicate the hypomethylated windows from Table 1. Brown Δ*Dim2*, green Δ*Dnmt5*, teal Δ*Dim2*-Δ*Dnmt5*, blue Δ*Rid*, purple Δ*Hp1*. Previously defined LS regions [191] are highlighted in yellow.

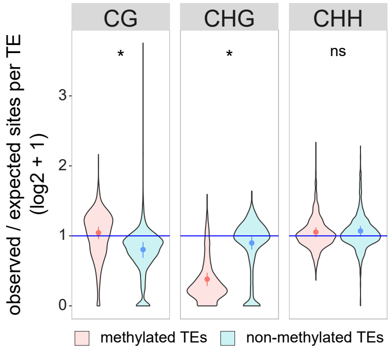


FIGURE S9 | Occurrence of CG, CHG and CHH sites in methylated and non-methylated transposable elements. The observed sites per TE were compared to the number of expected sites based on sequence composition per TE.

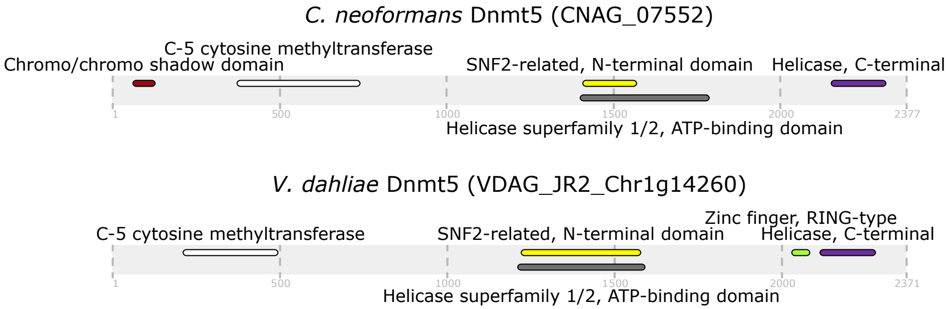

C. neoformans Dnmt5 (CNAG_07552)
C-5 cytosine methyltransferase
Chromo/chromo shadow domain
SNF2-related, N-terminal domain
Helicase, C-terminal
1
500
1000
1500
2000
2377
Helicase superfamily 1/2, ATP-binding domain
V. dahliae Dnmt5 (VDAG_JR2_Chr1g14260)
Zinc finger, RING-type
C-5 cytosine methyltransferase
SNF2-related, N-terminal domain
Helicase, C-terminal
1
500
1000
1500
2000
2371
Helicase superfamily 1/2, ATP-binding domain

FIGURE S10 | Comparison of protein domain structure of *C. neoformans* and *V. dahliae* Dnmt5.

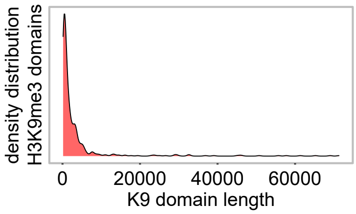


FIGURE S11 | Distribution of H3K9me3 domain lengths.

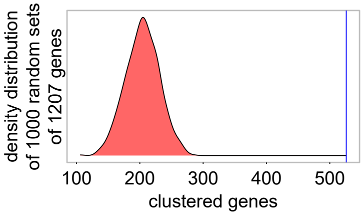


FIGURE S12 | Genes induced in *Hp1* and *Dim5* mutants cluster more often than expected based on chance. The 526 out of 1207 induced genes (blue vertical line) that cluster in the genome are more than would be expected based on 1000 random sets of 1207 genes.

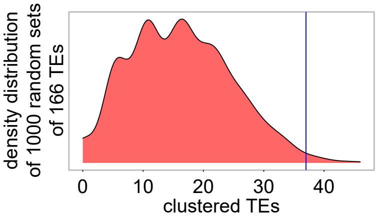


FIGURE S13 | Transposons induced in *Hp1* and *Dim5* mutants cluster more often than expected based on chance. The 37 out of 166 induced TEs (blue vertical line) that cluster in the genome are more than would be expected based on 1000 random sets of 166 TEs.

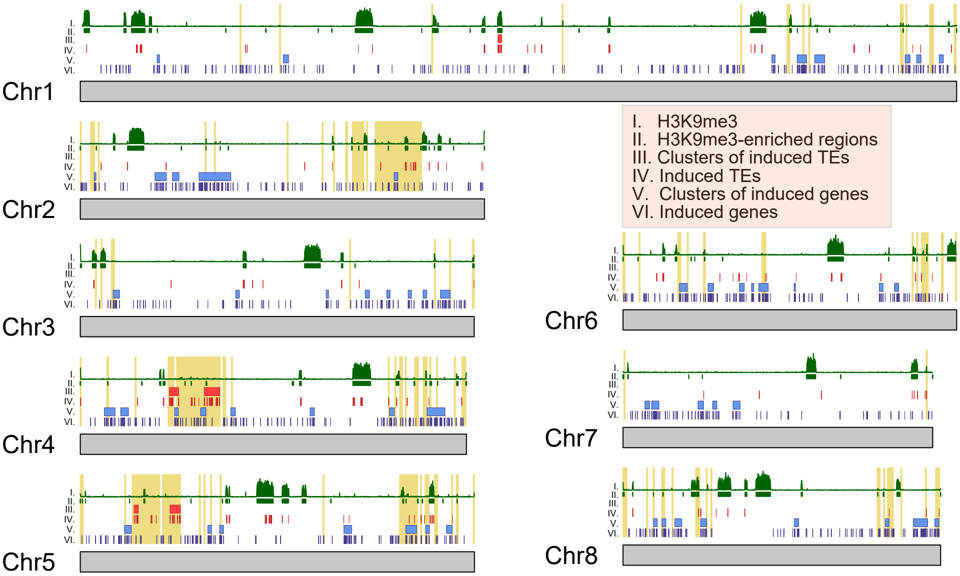


FIGURE S14 | Clusters of genes and transposons over all chromosomes. Whole-chromosome plots displaying the location of induced genes (in blue) and transposons (in red). Clusters of induced genes and transposons are indicated as blue and red rectangles, respectively. H3K9me3-ChIP signal along the chromosomes is indicated in green in the upper track. Previously defined LS regions [191] are highlighted in yellow.

TABLE S1 | Primers used to generate homologous recombination knockout and complementation vectors in *V. dahliae.*

Name	Sequence (5′-3′)
DIM2.ko-LB_F	*GGTCTTAAU*ACTACTCGGAGCTGACGGATT
DIM2.ko-LB_R	*GGCATTAAU*TTGTGACTGTCAGATTGCGGATA
DIM2.ko-RB_F	*GGACTTAAU*CGCGTTGCATAGCACTTTGATA
DIM2.ko-RB_R	*GGGTTTAAU*GAAACCAGCAAGGCAAGAGAGA
DIM5.ko-LB_F	*GGTCTTAAU*CCAAGCAGGAGTACGAGATGCT
DIM5.ko-LB_R	*GGCATTAAU*GATCTGCCTCTCATCCCAAGTG
DIM5.ko-RB_F	*GGACTTAAU*ATCAGGTTCATGGGACTTGTGG
DIM5.ko-RB_R	*GGGTTTAAU*GCGAGGTCGAGCAGAAGACTAT
Rid.KO-LB.F	*GGTCTTAAU*AGATCTGTTCTGGTTCGCTTGG
Rid.KO-LB.R	*GGCATTAAU*CAAGGCAAATAAAGCCCACAAAG
Rid.KO-RB.F	*GGACTTAAU*TTGCAGAACGAAGAGAGGTTCG
Rid.KO-RB.R	*GGGTTTAAU*ATAACGCCTTCACCAGCGTCTT
Dnmt5.LB.KO_F	*GGTCTTAAU*CGCTGCATCACGAACATCTACG
Dnmt5.LB.KO_R	*GGCATTAAU*GCCGTGCAAACTGAATGCTCTAT
Dnmt5.RB.KO_F	*GGACTTAAU*ACGAGTGCCGTGAATCCTGATAC
Dnmt5.RB.KO_R	*GGGTTTAAU*TCGTTGCCGTGACATCAAATACA
Italic sequence at 5′ end of primer corresponds to the necessary sequence for USER cloning	
Dim2_p005_f_4336bp	CGCGCCACTAGTCTCGAGTTAATATGCCGTACTTCATATATGGTC
Dim2_p005_r_4336bp	TAGAGCGGCCGCCACCGCGGTTAATCTAGGGCAGACATGTCCTTT
Hp1_p005_f_1250bp	CGCGCCACTAGTCTCGAGTTAATATGCCGCCAGGTAGGCCT
Hp1_p005_r_1250bp	TAGAGCGGCCGCCACCGCGGTTAATTTATTTGAGCTCGGCATCCGT
Dim5_p005_f_1168bp	CGCGCCACTAGTCTCGAGTTAATATGGAGGGTATCACGAGAC
Dim5_p005_r_1168bp	TAGAGCGGCCGCCACCGCGGTTAATTCACCACAGGAAGCCACG

Italic sequence at 5′ end of primer corresponds to overlap sequence with pFBT005 vector

TABLE S2 | Primers used to confirm homologous recombination knockout and complementation vectors in *V. dahliae.*

Name	Sequence (5′-3′)
Dim2_ORF.400.F	CAGACTGCTGCATCCCGTTC
Dim2_ORF.400.R	GTTGCCGACCGTCTTCCACT
Dim5_ORF.400.F	CGATCACCATCATCAACGACAT
Dim5_ORF.400.R	GTAGATTTTCTTCCGGCCCTGT
Rid.ORF.380.F	TCGCAGCTTGACTTGGAGTATC
Rid.ORF.380.R	CTCTGGCTGGGGGCTATGAAG
Vd.Dnmt5.ORF_F	CCTTGGCGTTGGTCATTGTAAGA
Vd.Dnmt5.ORF_R	GATATTTCGCACATCGGGGTTAC
Dim5_pres_434bp_f	GACACGGAGAAGCAGACTCC
Dim5_pres_434bp_r	GCGTGTGGTACCGGTAGATT
Hp1_pres_484bp_f	GGATTTGCTTCGGGTCTACA
Hp1_pres_484bp_r	AATGTGGCGCTCGTAGAACT
Dim2_pres_440bp_f	TCGTTGTCTACAGGCACTCG
Dim2_pres_440bp_r	CCACCTGTGAAAGTGGGACT

Gene specific knockout was confirmed by the attempted amplification of a PCR band using the gene specific forward and reverse primer for an indicated gene. Positive PCR products were confirmed from a wild-type strain.

TABLE S3 | Primers used to confirm gene deletion in *V. dahliae*.

Name	Sequence (5'-3')
Dim2ko_Confirm.F	CCAAGATTGCCTTCTCGTCG
Dim5ko_Confirm.F	GCGTCGAGAAGACCAACTACAT
Ridko_Confirm.F	GGCGGTTATAGGTGAGAAACACG
Dnmt5ko_Confirm.F	CAGCTTCTTCAGCAACATGCAACTA
Vector_Reverse	GGAGTCGCATAAGGGAGAGCG

Locus specific integration for a gene was confirmed by the amplification of a PCR band using the forward primer for an indicated gene and the universal reverse primer for the inserted DNA

TABLE S4 | Primers used for qPCR fungal biomass quantification of *V. dahliae* infecting tomato.

Name	Sequence (5'-3')
Vdahliae_ITS.F	AAAGTTTTAATGGTTCGCTAAGA
Vdahliae_ITS.R	CTTGGTCATTTAGAGGAAGTAA
Slyco_Gap.F	GCTCCCACAACTTAACGGCA
Slyco_Gap.R	TGCTGTCACCAACAAAGTCTGTG

TABLE S5 | Number of TEs induced and repressed for each mutant compared to wild-type using different mapping parameters.

		Iterative multimapping	Uniq-mapping
ΔDim2	induced TEs	2	1
	repressed TEs	2	0
ΔDnmt5	induced TEs	3	2
	repressed TEs	1	0
ΔDim2ΔDnmt5	induced TEs	3	1
	repressed TEs	1	0
ΔRid	induced TEs	1	1
	repressed TEs	0	0
ΔHp1	induced TEs	261	227
	repressed TEs	23	17
ΔDim5	induced TEs	241	220
	repressed TEs	47	27

Chapter 5

Repetitive elements contribute to the diversity and evolution of centromeres in the fungal genus *Verticillium*

Michael F Seidl[1,2],
H Martin Kramer[2],
David E Cook[2,3],
Gabriel Lorencini Fiorin[2],
Grardy CM van den Berg[2],
Luigi Faino[2,4],
and Bart PHJ Thomma[2,5]

[1]Theoretical Biology & Bioinformatics, Utrecht University, Utrecht, the Netherlands
[2]Laboratory of Phytopathology, Wageningen University, Wageningen, the Netherlands
[3]Plant Pathology, Kansas State University, Manhattan, United States of America
[4]Environmental Biology Department, Sapienza Università di Roma, Rome, Italy
[5]University of Cologne, Institute for Plant Sciences,
Cluster of Excellence on Plant Sciences (CEPLAS), 50674 Cologne, Germany

Published in mBio 11: e01714-20 (2020)

Abstract

Centromeres are chromosomal regions that are crucial for chromosome segregation during mitosis and meiosis, and failed centromere formation can contribute to chromosomal anomalies. Despite this conserved function, centromeres differ significantly between and even within species. Thus far, systematic studies into the organization and evolution of fungal centromeres remain scarce. In this study, we identified the centromeres in each of the ten species of the fungal genus *Verticillium* and characterized their organization and evolution. Chromatin immunoprecipitation of the centromere-specific histone CenH3 (ChIP-seq) and chromatin conformation capture (Hi-C) followed by high-throughput sequencing identified eight conserved, large (~150 kb), AT-, and repeat-rich regional centromeres that are embedded in heterochromatin in the plant pathogen *V. dahliae*. Using Hi-C, we similarly identified repeat-rich centromeres in the other *Verticillium* species. Strikingly, a single degenerated LTR retrotransposon is strongly associated with centromeric regions in some but not all *Verticillium* species. Extensive chromosomal rearrangements occurred during Verticillium evolution, of which some could be linked to centromeres, suggesting that centromeres contributed to chromosomal evolution. The size and organization of centromeres differ considerably between species, and centromere size was found to correlate with the genome-wide repeat content. Overall, our study highlights the contribution of repetitive elements to the diversity and rapid evolution of centromeres within the fungal genus *Verticillium*.

Importance

The genus *Verticillium* contains ten species of plant-associated fungi, some of which are notorious pathogens. *Verticillium* species evolved by frequent chromosomal rearrangements that contribute to genome plasticity. Centromeres are instrumental for separation of chromosomes during mitosis and meiosis, and failed centromere functionality can lead to chromosomal anomalies. Here, we used a combination of experimental techniques to identify and characterize centromeres in each of the *Verticillium* species. Intriguingly, we could strongly associate a single repetitive element to the centromeres of some of the *Verticillium* species. The presence of this element in the centromeres coincides with increased centromere sizes and genome-wide repeat expansions. Collectively, our findings signify a role of repetitive elements in the function, organization and rapid evolution of centromeres in a set of closely related fungal species.

Introduction

Centromeres are crucial for reliable chromosome segregation during mitosis and meiosis. During this process, centromeres direct the assembly of the kinetochore, a multi-protein complex that facilitates attachment of spindle microtubules to chromatids [233–235]. Failure in formation or maintenance of centromeres can lead to aneuploidy, i.e. changes in the number of chromosomes within a nucleus, and to chromosomal rearrangements [235–237]. While these processes have been often associated with disease development [238], they can also provide genetic diversity that is beneficial for adaptation to novel or changing environments [67,74]. For example, aneuploidy in the budding yeast *Saccharomyces cerevisiae* can lead to increased fitness under selective conditions, such as the presence of antifungal drugs [239,240]. Thus, centromeric instability can contribute to adaptive genome evolution [241,242].

Despite their conserved function, centromeres are among the most rapidly evolving genomic regions [243,244] that are typically defined by their unusual (AT-rich) sequence composition, low gene and high repeat density, and heterochromatic nature [243,245]. Nevertheless, centromeres differ significantly in size, composition, and organization between species [243,246]. Centromeres in *S. cerevisiae* are only ~125 nucleotides long and are bound by a single nucleosome containing the centromere-specific histone 3 variant CenH3 (also called CENP-A or Cse4) [247–250]. In contrast to these 'point centromeres', centromeres in many other fungi are more variable and larger, and have thus been referred to as 'regional centromeres' [245]. For instance, in the opportunistically pathogenic yeast *Candida albicans*, the CenH3-bound 3-5 kb long centromeric DNA regions differ significantly between chromosomes, and rapidly diverged from closely related *Candida* species [251–253]. Centromeres in the basidiomycete yeasts *Malassezia* are similar in size (3-5 kb) but contain a short AT-rich consensus sequence in multiple *Malassezia* species [241]. In *Malassezia*, chromosomal rearrangements and karyotype changes are driven by centromeric loss through chromosomal breakage or by inactivation through sequence diversification [241]. Chromosomal rearrangements at centromeres have been similarly observed in the yeast *Candida parapsilosis*, suggesting that centromeres can be fragile and contribute to karyotype evolution [241,242]. CenH3-bound centromeric regions of the basidiomycete yeast *Cryptococcus neoformans* are relatively large, ranging from 30 to 65 kb, and are rich in Long Terminal Repeat (LTR)-type retrotransposons [246]. Centromere sizes differ between *Cryptococcus* species as those lacking RNAi and DNA methylation have shorter centromeres, associated with the loss of full-length LTR retrotransposons at centromeric regions, suggesting that functional RNAi together with DNA methylation is required for centromere stability [246].

In filamentous fungi, centromeres have been most extensively studied in the saprophyte *Neurospora crassa* [245]. In this species, centromeric regions are considerably larger than in yeasts (on average ~200 kb), and are characterized by AT-rich sequences that are degenerated remnants of transposable elements and sequence repeats that lack an overall consensus sequence [245,254,255]. The increased AT-content and the degenerated nature of transposable elements in the genome of *N. crassa* are the result of a process called repeat-induced point mutation (RIP) [245,256]. RIP has been linked to the sexual cycle of ascomycetes and targets repetitive sequences by inducing C to T mutations, preferably at CpA di-nucleotides [256]. The AT-rich centromeric regions are bound by CenH3 and enriched in the heterochromatin-specific histone modification histone 3 trimethylation of lysine 9 (H3K9me3) [255]. Additionally, H3K9me3 and cytosine methylation occurs at the periphery of the centromeres [255]. Alterations in H3K9me3 localization compromise centromeric localization, suggesting that the formation and location of heterochromatin, rather than the DNA sequence itself, is essential for function and localization of centromeres in *N. crassa* [245,255]. However, heterochromatin is not a hallmark for centromeres in all filamentous fungi. Centromeres in the fungal wheat pathogen *Zymoseptoria tritici* are shorter (~10 kb) and AT-poor, and their presence does not correlate with transposable elements nor with heterochromatin-specific histone modifications such as H3K9me3 or histone 3 trimethylation of lysine 27 (H3K27me3) [121]. Thus, even though centromeric function is highly conserved, fungal centromeres differ considerably in size, sequence composition, and organization.

Knowledge on centromeres has been impaired by their repetitive nature, which hampers their assembly and subsequent analyses [66,245]. However, recent advances in long-read sequencing technologies enables to study the constitution and evolution of centromeres [241,246,257–259]. By using long-read sequencing technologies in combination with optical mapping, we previously generated gapless genome assemblies of two strains of the fungal plant pathogen *Verticillium dahliae* [122]. The genome of *V. dahliae* is characterized by lineage-specific (LS) regions [59,61,63,67,74] that are hypervariable between *V. dahliae* strains and that contain genes with crucial roles in virulence and host adaptation [59,63,67,74]. LS regions evolved by extensive chromosomal rearrangements, translocations, duplications, and deletions, that are mediated by erroneous double-strand repair pathways, often involving repetitive elements [74]. Repetitive elements within the LS regions display a distinct chromatin state when compared with other repetitive regions [191]. The *Verticillium* genus consists of ten species that are all soil-borne and presumed asexual but have different life-styles [260]. Nine of these species are haploid, while the species *Verticillium longisporum* is an allodiploid hybrid between a strain that is closely related to *V. dahliae* and an unknown *Verticillium* species [260,261]. During the evolution of the different *Verticillium* species frequent chromosomal rearrangements occurred [59,74,75], and regions with characteristics similar to LS regions have been identified in other *Verticillium* species as well [63]. Centromeres have been thought to facilitate chromosomal rearrangements and contribute to karyotype evolution [241,242,262], and thus deeper knowledge of centromeres might help to understand mechanisms that drive chromosomal rearrangements in *Verticillium* genome evolution. Facilitated by the availability of *V. dahliae* high-quality genome assemblies and of all other *Verticillium* species [63,75,122,263], we here sought to identify and study the constitution and evolution of centromeres in the *Verticillium* genus and to elucidate their impact on chromosome evolution.

Results

CenH3-binding identifies large regional centromeres in *Verticillium dahliae*

Centromeres differ significantly between fungi, but most centromeres are functionally defined by nucleosomes containing CenH3 [233]. To identify centromeres in *V. dahliae* strain JR2 by chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq), we first identified the *V. dahliae* CenH3 ortholog (Fig. S1a) and generated transformants with N-terminally FLAG-tagged CenH3 (Table S1). To this end, the coding sequence for the FLAG-tagged CenH3 was inserted in locus behind the native *CenH3* promoter (Fig. S1b-c). We subsequently used anti-FLAG antibodies to purify FLAG-tagged CenH3-containing nucleosomes from two *V. dahliae* transformants (Table S1a) and sequenced the nucleosome-associated genomic DNA. Mapping of the sequencing reads to the *V. dahliae* strain JR2 genome assembly identified a single CenH3-enriched region per chromosome (Fig. 1a; Fig. S1d-e), while mapping of the sequencing reads derived from the WT strain did not reveal any CenH3-enriched region (Fig. S1d-e). The CenH3-enriched regions, designated as *Cen1-8*, range between ~94 and ~187 kb in size (Fig. 1a; Table 1). To corroborate these centromere sizes, we assessed centromere locations based on a previously generated optical map [59,122] revealing no significant size differences (Fig. S1e). Thus, we conclude that CenH3-binding defines large regional centromeres in *V. dahliae* strain JR2.

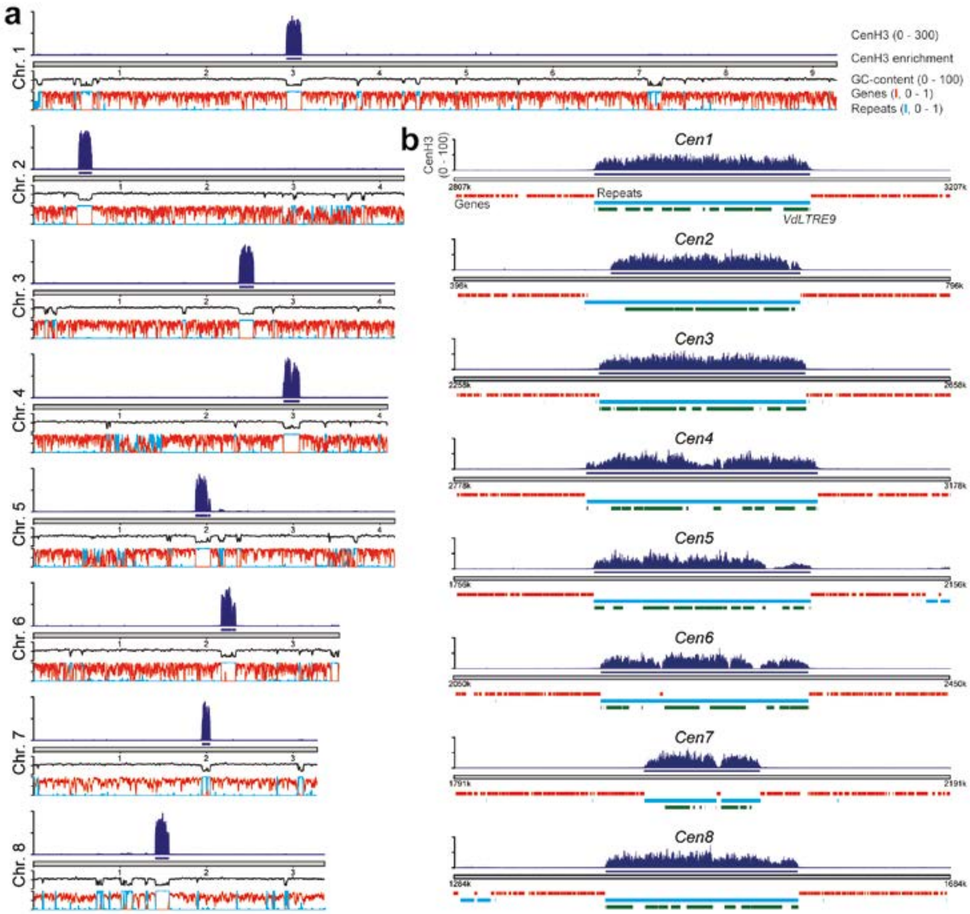


FIGURE 1 | CenH3-binding defines centromeres in *Verticillium dahliae* strain JR2. (a) Schematic overview of the chromosomes of *V. dahliae* strain JR2 showing the normalized CenH3 ChIP-seq read coverage (RPGC normalization in 1 kb bins with 3 kb smoothening), CenH3 enriched regions, GC-content, gene density (red line), and repeat density (blue line). (b) Magnification of a 400 kb region containing the centromere is shown for each of the eight chromosomes of *V. dahliae* strain JR2 (*Cen1-8*) depicting the CenH3 ChIP-seq read coverage (RPGC normalization in 10 bp bins with a 30 bp smoothening) and enrichment, as well as the presence of genes (red) and repetitive elements (blue). Regions carrying the centromere-specific long-terminal repeat element *VdLTRE9* are highlighted in green.

Centromeres in *Verticillium dahliae* are repeat-rich and embedded in heterochromatin

Centromeres are often characterized by increased AT-content, increased repeat density, and depletion of protein coding genes [243,245,257]. To characterize the centromeres in *V. dahliae* strain JR2, we queried the eight chromosomes for the presence of large AT-rich, gene-sparse, and repeat-rich regions. Seven of the eight chromosomes contain only a single large (>93 kb; average size ~150 kb) AT-rich region (~74-78% versus ~46% genome-wide), nearly completely devoid of protein-coding genes and enriched for repetitive sequences, that overlaps with the regions defined by CenH3-binding (Fig. 1a; Table 1). In contrast, chromosome 1 contains three regions with these characteristics (Fig. 1a; Table 1). However, only one of these overlaps with the centromeric regions defined by CenH3-binding (Fig. 1).

Elevated AT-levels in repeat-rich regions are caused by RIP mutations in some filamentous fungi [218,245,255,256]. Due to its presumably asexual nature [67], the occurrence of RIP in *V. dahliae* is controversial [62,74,139], although signatures of RIP have previously been reported in a subset of repeat-rich regions [191]. We assessed the occurrence of RIP signatures in centromeres using the composite RIP index (CRI) [97], which considers C to T mutations in the CpA context. Intriguingly, genomic regions located at centromeres display significantly higher CRI values than other genomic regions (e.g. genes or repetitive elements) (Fig. 2a; Fig. S2, S3a), and thus RIP signatures at repetitive elements located at centromeres likely contribute to the high AT-levels.

In most filamentous fungi and oomycetes, AT- and repeat-rich centromeres are embedded in heterochromatin that is characterized by methylated DNA and by particular histone modifications (H3K9me3 and H3K27me3) [97,243,245,246,255,258]. We recently determined chromatin states in the genome of *V. dahliae* strain JR2 and revealed that repetitive sequences outside of the LS regions display characteristics of heterochromatin [191]. To define centromeric chromatin states, we used previously generated bisulfite sequencing data to monitor DNA methylation (mC) and ChIP-seq data to determine the distribution of the heterochromatic marks H3K9me3 and H3K27me3 [191]. To also determine the distribution of euchromatin, we performed ChIP-seq with an antibody against the euchromatic mark di-methylation of lysine 4 of histone H3 (H3K4me2). We observed overall low genome-wide DNA methylation levels [191] (Fig. 2a; Fig. S2), similar to the previously reported levels for *Aspergillus flavus* [264] and lower than for *N. crassa* [265]. Nevertheless, repetitive elements and centromeres show significantly higher DNA methylation levels in all contexts when compared with genes (Fig. 2b). Methylation (in CG context) at repetitive elements at centromeres is significantly higher than at repeats located along the chromosomal arm, but not at sub-telomeric regions (Fig. 2c), and more methylation at centromeres correlates with increased CRI (Fig. 2a; Fig. S2, S3a). DNA methylation co-localizes with H3K9me3 at repeat-rich regions [191] (Fig. 2a; Fig. S2). H3K9me3 occurs predominantly at repetitive elements localized at sub-telomeres and centromeres (Fig. 2d-e; Fig. S2, S3b). In comparison, H3K4me2 and H3K27me3 are largely absent from centromeres (Fig. 2d-e; Fig. S3b). Collectively, these observations indicate that centromeres of *V. dahliae* display typical characteristics of constitutive heterochromatin.

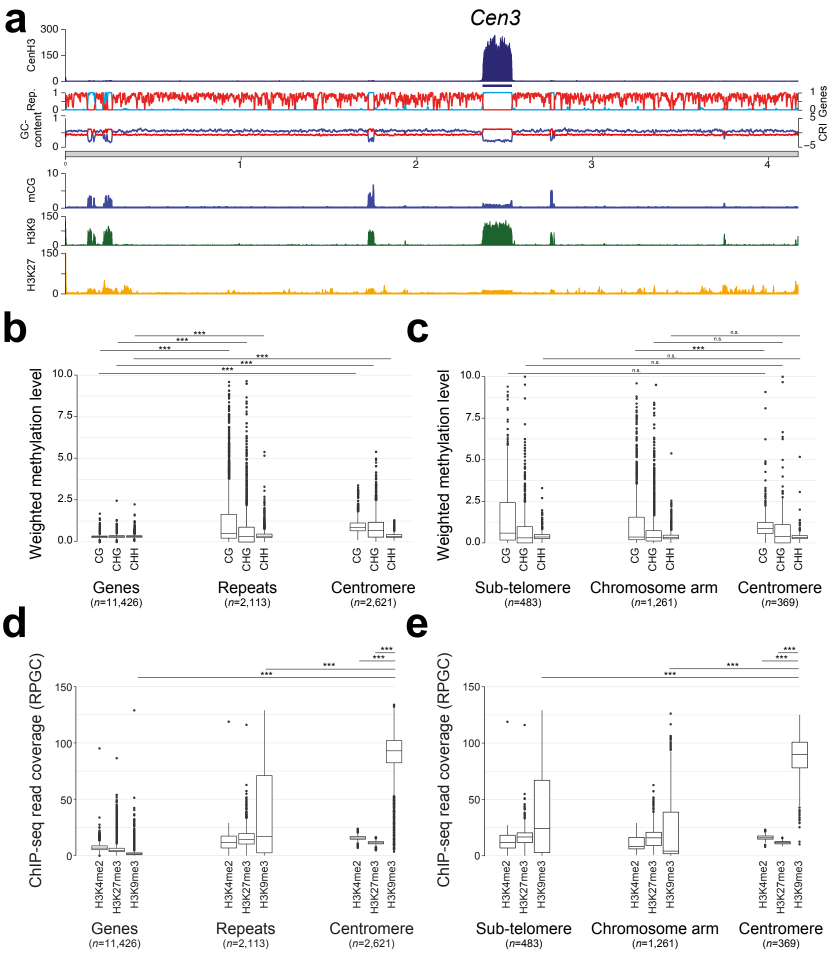


FIGURE 2 | Centromeres in *Verticillium dahliae* strain JR2 are embedded in heterochromatin. (a) Schematic overview of chromosome 3 of *V. dahliae* strain JR2, exemplifying the distribution of heterochromatin-associated chromatin modifications (mC, H3K9me3, and H3K27me3) in relation to the centromeres. The different lanes display the FLAG-CenH3 ChIP-seq read coverage (RPGC normalization in 1 kb bins with 3 kb smoothening), the FLAG-CenH3 enriched regions, the repeat- and gene-density (light blue and red, respectively), the GC-content (blue), the CRI (red) as well as the weighted cytosine methylation (all summarized in 5 kb windows with 500 bp slide), and the normalized H3K9me3 and H3K27me3 ChIP-seq read coverage (RPGC normalization in 1 kb bins with 3 kb smoothening). The schematic overview of all chromosomes is shown in Fig. S2. (b) Box plots of weighted DNA methylation levels per genomic context (CG, CHG, or CHH) are summarized over genes, repetitive elements, or 5 kb genomic windows (500 bp slide) overlapping with the centromeric regions. (c) Weighted DNA methylation levels per genomic context (CG, CHG, or CHH) are summarized over repetitive elements that have been split based on their genomic location; sub-telomeres (within the first or last 10% of the chromosome), centromeres, or the remainder of the chromosome arm. (d) ChIP-seq read coverage (RPGC normalized; see (a)) for H3K4me2, H3K27m3, and H3K9me3 is summarized over genes, repetitive elements, or 5 kb windows (500 bp slide) overlapping with the centromeric regions. (e) ChIP-seq read coverage (RPGC normalized; see (a)) for H3K4me2, H3K27m3, and H3K9me3 is summarized over repetitive elements that have been split based on their genomic location; sub-telomeres (within the first or last 10% of the chromosome), centromeres, or the remainder of the chromosomal arm. Statistical differences for the indicated comparisons were calculated using the one-sided non-parametric Mann-Whitney test; p-values < 0.001: ***.

A single repeat associates with centromeres of *Verticillium dahliae* strain JR2

Centromere identity and function is typically defined by CenH3-binding and not by specific DNA sequences, although various types of repetitive sequences, such as transposable elements, are commonly observed in centromeres of plants, animals, and fungi [243,245,266,267]. Unsurprisingly, CenH3-bound centromeres are repeat-rich in *V. dahliae* (Fig. 1). A detailed analysis of the eight centromeres revealed a near-complete (>96%) composition of repetitive elements belonging to only ten different repeat sub-families (Fig. 1b, 3a; Table 1), of which the majority shows similarity to LTR retrotransposons of the *Gypsy*- and *Copia*-like families (Fig. 3a). These elements show signs of RIP, are highly methylated, and non-transcribed (Fig. S3c-e), and thus likely inactive. Interestingly, a single LTR retrotransposon sub-family, previously designated *VdLTRE9* [74,122], covers on average ~70% of the DNA sequences at the eight centromeres, ranging from 47% in *Cen7* to 83% in *Cen2* (Fig. 3a; Table 1). We scanned the genome for the localization of the ten repeat sub-families (Fig. 3). Intriguingly, although it is one of the most abundant repeats in the genome with 215 complete or partial matches, *VdLTRE9* is associated to centromeres as 95% of the copies (204 out of 215; one-sided Fisher's exact test; multiple-testing corrected p-value 3e-106) occur at the eight centromeres (Fig. 3b-c). The remaining eleven *VdLTRE9* copies (5%) occur outside of the CenH3-rich centromeres, yet five out of eleven copies are localized within 50 kb of the centromeric regions (Fig. 3b-c). The nine other repeat sub-families have additional matches that are located outside of the centromeres (Fig. 1a; Fig. 3b-c), and only two of these repeats are significantly enriched and consistently present in all eight centromeres; 63% and 45% of the matches of these two sub-families occur at the centromeres (Fig. 3c). Repeats at centromeres are often fragmented and most copies, with the exception of the Tc1/mariner-like elements, are similarly fragmented when located outside of the centromeres (Fig. S3f), indicating extensive degeneration of repetitive elements in *V. dahliae*. Collectively, these findings suggest that only the presence of *VdLTRE9* is strongly associated with centromeres in *V. dahliae* strain JR2.

VdLTRE9 displays similarity to LTR retrotransposons. The consensus sequence of *VdLTRE9* is ~7.3 kb long (the two LTR sequences are each ~200 bp long), and the individual matches share a high degree of sequence identity (~86%). Sequence similarity based TE-classifications using PASTEC [138] indicates that the consensus sequence displays remote similarity to *Gypsy*-like retrotransposons. Only ~25% of the *VdLTRE9* matches in the genome cover the entire (>97.5%) consensus sequence, but many of these are still fragmented as they occur as discontinuous copies. Furthermore, the *VdLTRE9* consensus sequence is AT-rich (~75% AT), which may be caused by RIP (Fig. S3d), indicating that *VdLTRE9*, similar to other repeats in *V. dahliae*, has significantly degenerated.

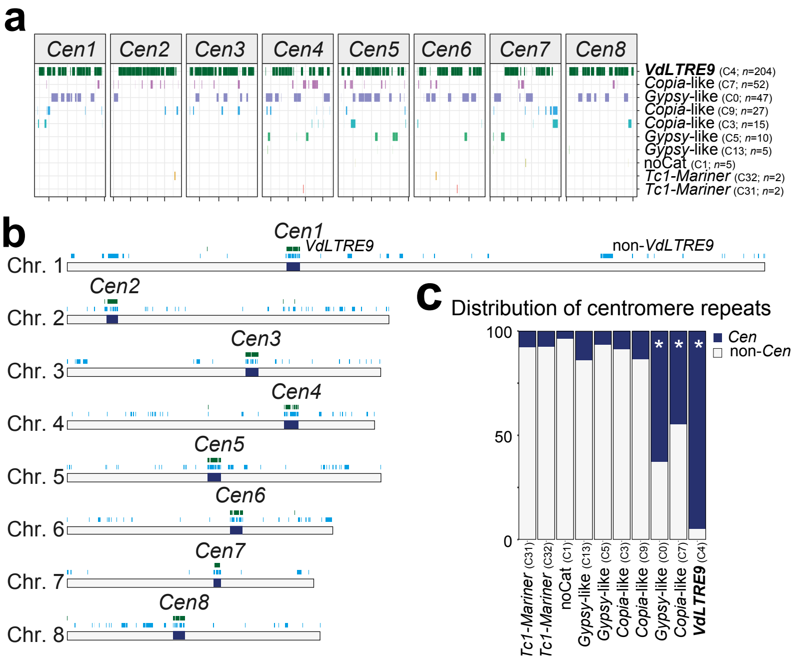


FIGURE 3 | A single repeat family associates with centromeres in *Verticillium dahliae* strain JR2. (a) The presence of different repeat sub-families is shown across the eight centromeres (*Cen1-8*), and the number of occurrences for each sub-family within the centromeres is indicated. The individual centromeres in the diagram are shown in equal scale. (b) Genome-wide distribution of the ten repeat sub-families occurring within the eight centromeres (*Cen1-8*; dark blue); the location of *VdLTRE9* is shown in green and the location of elements belonging to the other nine sub-repeat families (from panel (a)) is shown in light blue. (c) The distribution of different repeat sub-families in centromeres (*Cen*; dark blue) and across the genome (non-*Cen*; light grey). The enrichment of specific sub-families at centromeres was assessed using a one-sided Fisher's exact test. Significant enrichment (multiple-testing corrected p-value < 0.01) is denoted with an asterisk.

VdLTRE9 as hallmark of *Verticillium dahliae* centromeres

To examine if *VdLTRE9* similarly occurs at centromeres in other *V. dahliae* strains, we made use of the complete genome assembly of *V. dahliae* strain VdLs17 [59,74,122]. The evolution of *V. dahliae* is characterised by chromosomal rearrangements [59,74] (Fig. 4a; Fig. S4a-c). Nevertheless, synteny analyses between *V. dahliae* strains JR2 and VdLs17 revealed large regions of co-linearity between chromosomes and identified significant sequence and synteny conservation between the centromeres and their flanking regions (Fig. 4b-c; Fig. S4a), suggesting that centromeric sequences and their locations are conserved. We queried the genome of *V. dahliae* strain VdLs17 for the presence of *VdLTRE9* and identified a single region on each chromosome, collectively containing 186 of the 207 (90%) complete or partial matches of *VdLTRE9* in the genome (Fig. 4d) (one-sided Fisher's exact test; multiple-testing corrected p-value 3e-146). These *VdLTRE9*-rich regions are ~150 kb in size, AT-rich, gene-poor and repeat-rich, and share similarity to the previously identified CenH3-bound and *VdLTRE9*-enriched regions of *V.*

dahliae strain JR2 (Fig. 4b-c; Fig. S4d), suggesting that these regions similarly represent the centromeres of *V. dahliae* strain VdLs17.

Centromeres of *N. crassa* and some other fungi co-localize within the nucleus [105,245,268–271]. This co-localization can be experimentally determined using chromosome conformation capture (Hi-C), which can identify centromeres by their increased inter-chromosomal contacts [271]. To confirm that Hi-C can be used to identify centromeres in *V. dahliae*, we first applied Hi-C to *V. dahliae* strain JR2. As anticipated, we observed seven strong inter-chromosomal contacts for each of the eight chromosomes (Fig. 4e). Importantly, the interacting regions overlap with the CenH3-bound regions that we identified as centromeres (Table S1b), demonstrating that centromeres in *V. dahliae* strain JR2 co-localize within the nucleus and supporting that Hi-C reliably identifies centromeres [105,268]. We then applied Hi-C to *V. dahliae* strain VdLs17, and similarly identified regions with strong inter-chromosomal contacts, one for each of the chromosomes (Fig. 4f). These regions overlap with the *VdLTRE9*-enriched regions (Table S1b), suggesting that these represent functional centromeres in *V. dahliae* strain VdLs17.

The two *V. dahliae* strains JR2 and VdLs17 are closely related and differ only by ~0.05% sequence diversity [59,74]. Thus, the conservation of *VdLTRE9* at centromeres could be driven by limited divergence between the two *V. dahliae* strains rather than representing a hallmark of *V. dahliae* centromeres. Therefore, we sought to determine centromeres in an additional *V. dahliae* strain with increased sequence diversity when compared with *V. dahliae* strains JR2 or VdLs17, namely strain CQ2 that displays ~1.05 percent sequence diversity [63]. We previously obtained a long-read based genome assembly of this strain that encompasses 17 contigs [63]. We generated Hi-C data for *V. dahliae* strain CQ2 and utilized intra-chromosomal contacts to assign the contigs into eight pseudo-chromosomes, leaving ~148 kb unplaced scaffolds (Fig. 4g; Fig. S4e; Table S1c). We subsequently identified a single region with seven strong inter-chromosomal contacts for each pseudo-chromosome that is significantly enriched for *VdLTRE9* (one-sided Fisher's exact test; multiple-testing corrected p-value 3.4e-166) (Fig. 4d,g; Fig. S4e; Table S1b). Synteny analyses between *V. dahliae* strains JR2 and CQ2 revealed that the eight *VdLTRE9*-rich regions and their flanking chromosomal regions are co-linear, suggesting that centromere locations are conserved between different *V. dahliae* strains (Fig. 4; Fig. S4a-c, f). With an average size of 165 kb, the centromeres of *V. dahliae* strain CQ2 are similar in size as the 144 kb and 157 kb average sizes in *V. dahliae* strains VdLs17 and JR2, respectively (Table S1b). The sizes of the corresponding (i.e. homologous) centromeres vary between the different *V. dahliae* strains. Yet, the consistent co-occurrence of the *VdLTRE9*-rich regions with the interaction data obtained by Hi-C throughout a selection of *V. dahliae* strains demonstrates that *VdLTRE9* is a hallmark of *V. dahliae* centromeres.

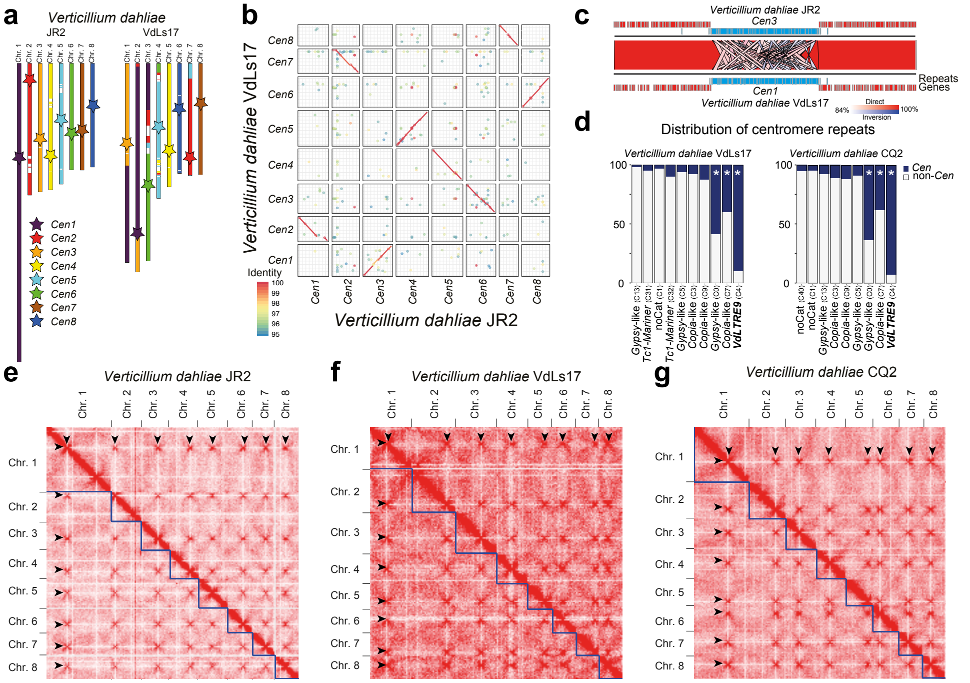


FIGURE 4 | Hi-C contact maps identify *VdLTRE9* as hallmark of centromeres in *Verticillium dahliae*. (a) Synteny analyses of the eight chromosomes of *V. dahliae* strains JR2 and VdLs17. Schematic overview of the eight chromosomes of *V. dahliae* strain JR2 (left) and the corresponding syntenic regions in *V. dahliae* strains VdLs17 (right). Approximate locations of centromeres are indicated by stars, and syntenic centromeres of *V. dahliae* strain VdLs17 are colored according to *Cen1-8* of *V. dahliae* strain JR2. (b) Sequence alignment of the centromeric regions ± 20 kb in *V. dahliae* strain JR2 and the corresponding regions in *V. dahliae* strains VdLs17 shown as dot-plot. For clarity, only alignments with >95% sequence identity are displayed. (c) Magnification of *Cen3* of *V. dahliae* strain JR2 and the syntenic *Cen1* of strain VdLs17. Synteny between regions is indicated by ribbons; entire centromeric regions *Cen1* and *Cen3* are syntenic and sequence similarity between individual *VdLTRE9* elements is visualized. The *Cen* regions ± 150 kb are shown as well as genes (red) and repeats (blue) are annotated within this region. (d) Distribution of different repeat families in centromeres (*Cen*; dark blue) and across the genome (non-*Cen*; light grey) for *V. dahliae* strains VdLs17 and CQ2. The enrichment of specific sub-families at centromeres was assessed using a one-sided Fisher's exact test. Significant enrichment (multiple-testing corrected p-value < 0.01) is denoted with an asterisk. (e-g) Hi-C contact matrix showing interaction frequencies between genomic regions in *Verticillium dahliae* strains JR2 (e), VdLs17 (f), and CQ2 (g). Regions of high inter-chromosomal interaction frequencies are indicative of centromeres and are highlighted by arrow heads. Interaction frequencies are summarized in 50 kb bins along the genome.

The evolution of *Verticillium* centromeres

In addition to *V. dahliae*, we previously generated genome assemblies of the eight haploid *Verticillium* species and the allodiploid *V. longisporum* [75,261] (Fig. 5a) that ranged from 12 to 684 scaffolds (Table S1c). These ten *Verticillium* species have been traditionally separated over two distinct clades; Flavnonexudans and Flavexudans (Fig. 5a) [260]. We generated Hi-C data to study the composition and evolution of centromeres in the different *Verticillium* species. By using intra-chromosomal interaction signals, we assigned the vast majority of the previously assembled contigs into eight pseudo-chromosomes for each of the haploid *Verticillium* species and 16 pseudo-chromosomes for the diploid *V. longisporum*, leaving between 0.5 kb and 2,022 kb unassigned (Fig. S5; Table S1c). For most genome assemblies, the pseudo-chromosomes contain one or both telomeric repeats (Table S1c), and thus we conclude that all *Verticillium* strains have eight chromosomes, and that this number doubled in *V. longisporum*. Based on the inter-chromosomal Hi-C interaction signals, we identified a single region with high inter-chromosomal contacts for each of the pseudo-chromosomes (Fig. S5; Table S1d), indicating that these are the centromeres in the different *Verticillium* species. The average centromere size in *Verticillium* is ~80 kb, yet we observed significant differences between the species (Fig. 5b; Fig. S6a-b). Centromeres within the Flavexudans clade are similarly sized and significantly smaller than the genus-wide average. By contrast, *V. dahliae* and *V. longisporum* centromeres are significantly larger.

We subsequently assessed whether *VdLTRE9* defines centromeres in the other *Verticillium* species besides *V. dahliae* as well. Interestingly, *VdLTRE9* is abundant at centromeres in the allodiploid *V. longisporum* and in *V. alfalfae*, but fewer (21) or no *VdLTRE9* copies were identified at centromeres in *V. nonalfalfae* and *V. nubilum*, respectively (Fig. 5c, e; Fig. S6c-d). Similar to *V. dahliae*, the vast majority of matches are fragmented, suggesting that *VdLTRE9* has been significantly degenerated in these species as well. Only very few partial or no matches of *VdLTRE9* consensus could be identified in the genomes of the Flavexudans species (Fig. 5c, e; Fig. S6-7). Collectively, these findings suggest that *VdLTRE9* is specific to Flavnonexudans species, yet we cannot exclude the alternative scenario in which *VdLTRE9* was present at the last common ancestor of *Verticillium* and has been lost in all Flavexudans species. Regardless of the origin, *VdLTRE9* has likely been recruited to the centromeres of Flavnonexudans species only after the divergence of *V. nubilum* (Fig. 5a; Fig. S6-7).

Since *VdLTRE9* occurs only in few *Verticillium* species, we assessed to which extent other repetitive elements contribute to centromere organization. We analyzed the repeats identified by *de novo* repeat predictions for each of the *Verticillium* species. Centromeres in all species are AT- and repeat-rich (Fig. 5d-e; S6a-b), and some repeats occur in high frequency or nearly exclusively at centromeres in species that lack *VdLTRE9*. However, in contrast to *VdLTRE9*, these repeats cover only a minority (typically less than 10%) of the centromeres. Sequence similarity-based cluster analyses of the *de novo* repeat consensus sequences revealed that divergent repeat families contribute to *Verticillium* centromere organization (Fig. S8). Thus, in contrast to *VdLTRE9* in most Flavnonexudans species, we could not identify any additional repeat family as a hallmark of centromeres in other *Verticillium* species.

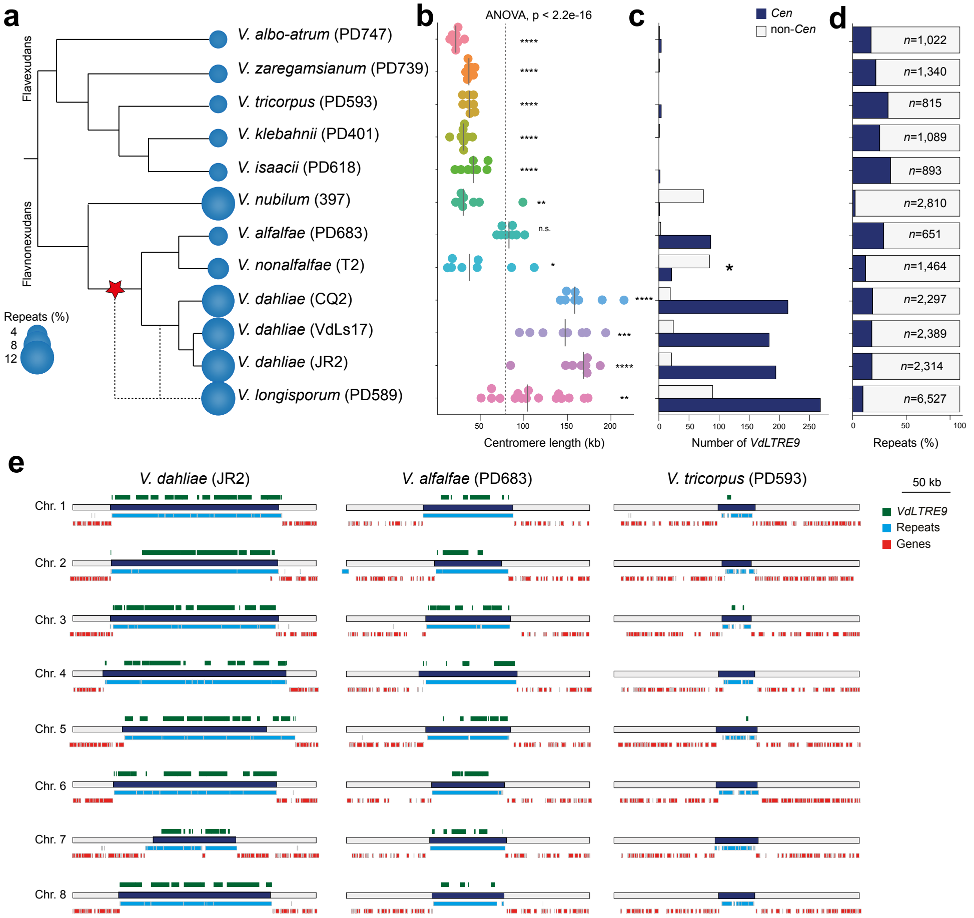


FIGURE 5 | Evolution of centromeres in the genus *Verticillium*. (a) Relationship of the ten members of the genus *Verticillium*. The predicted repeat content for each of the genomes is indicated (see Table S3 for details). The red star indicates the recruitment of *VdLTRE9* into centromeres. (b) Comparison of estimated centromere lengths (in kb) in the different *Verticillium* spp. Each dot represents a single centromere and the line represents the median size. (c) The number of (partial) *VdLTRE9* matches identified in centromeres (*Cen*; dark blue) and across the genome (non-*Cen*; light grey). The asterisk indicates the high number of *VdLTRE9* elements in unassigned contigs for *Verticillium nonalfalfae* strain T2 (see text for details). (d) Proportion of predicted repeat content localized at centromeres (*Cen*; dark blue) and across the genome (non-*Cen*; light grey). (e) Schematic overview of the eight centromeric regions (250 kb) in *Verticillium dahliae* strain JR2, and *Verticillium alfalfae* strain PD683 and *Verticillium tricorpus* stain PD593 as representatives for clade Flavnonexudans and clade Flavexudans, respectively. The centromeres are indicated by dark blue bars. The predicted genes (red) and repeats (light blue) are shown below each centromere, and location of (partial) *VdLTRE9* matches (light green) are shown above each centromere. Global statistical differences for the centromere sizes was calculated using one-way ANOVA, and differences for each species compared to the overall mean were computed using unpaired T-tests; p-values < 0.0001: ****, p-values < 0.001: ***, p-values < 0.01: **, p-values < 0.05: *.

Centromeres contribute to *Verticillium* karyotype evolution

We previously used fragmented genome assemblies to identify chromosomal rearrangements during *Verticillium* evolution [59,74,75]. We hypothesize that centromeres might have contributed to these chromosomal rearrangements. To identify genome rearrangements and to trace centromeres during *Verticillium* evolution, we used the pseudo-chromosomes of the haploid *Verticillium* species to reconstruct ancestral chromosomal configurations using AnChro (Fig 6a) [272]. We reconstructed all potential ancestors that predominantly had eight chromosomes and ~8,000 genes (Fig. S9a-b), yet the number of ancestral chromosomes and genes varied when approaching the last common ancestor (Fig. S9a-b). By balancing the number of reconstructed chromosomes and genes, we identified a single most parsimonious ancestral genome with eight chromosomes and ~8,500 genes (Fig. 6a; Fig. S9c), except for the last common ancestor within the clade Flavexudans clade that had eight major chromosomes and two additional 'chromosomes' with only six and two genes (Fig. S9d). As these two smaller 'chromosomes' likely do not represent genuine chromosomes, we conclude that all of the ancestral genomes, similar to the extant haploid *Verticillium* genomes, had eight chromosomes (Fig. 6a). Confirming our previous report [75], we observed in total 198 chromosomal rearrangements (124 inversions and 74 translocations) (Fig. 6a). The number of chromosomal rearrangements is lower than previously recorded and we did not observe any chromosomal fusion or fission events, which is likely the result of the drastically improved genome assemblies, but the rearrangement signal on each branch is sufficient to nevertheless recapitulate the known *Verticillium* species phylogeny (Fig. S9e). Importantly, we observed 17 genomic rearrangements that occurred at, or in close proximity (within ~15 genes up or downstream) to, centromeres, both in extant *Verticillium* species as well as in the ancestors (Fig. 6). For example, at the branch from the last common ancestor (VA; Fig. 6a) to the ancestor of the clade Flavexudans (B1; Fig. 6a), two centromere-associated translocations (between the ancestral chromosome 2 and 6) led to the formation of two rearranged chromosomes. In total, we observed that five out of the eight ancestral centromeres were associated with a chromosomal rearrangement at one point during evolution (Fig. 6a). Nevertheless, comparisons of protein-coding genes that flank centromeres show that these are syntenic in most extant species. Similarly, none of the recent chromosomal rearrangements observed between *V. dahliae* strains is clearly associated with centromeres (Fig. 4a-b, 6a), even though *CEN2* of *V. dahliae* strain VdLs17 is located near (20-25 genes up/downstream) a chromosomal rearrangement (Fig. 4a). Thus, while chromosomal rearrangements involving centromeres occurred during evolution, they do not account for the majority of the karyotype variation between extant *Verticillium* species.

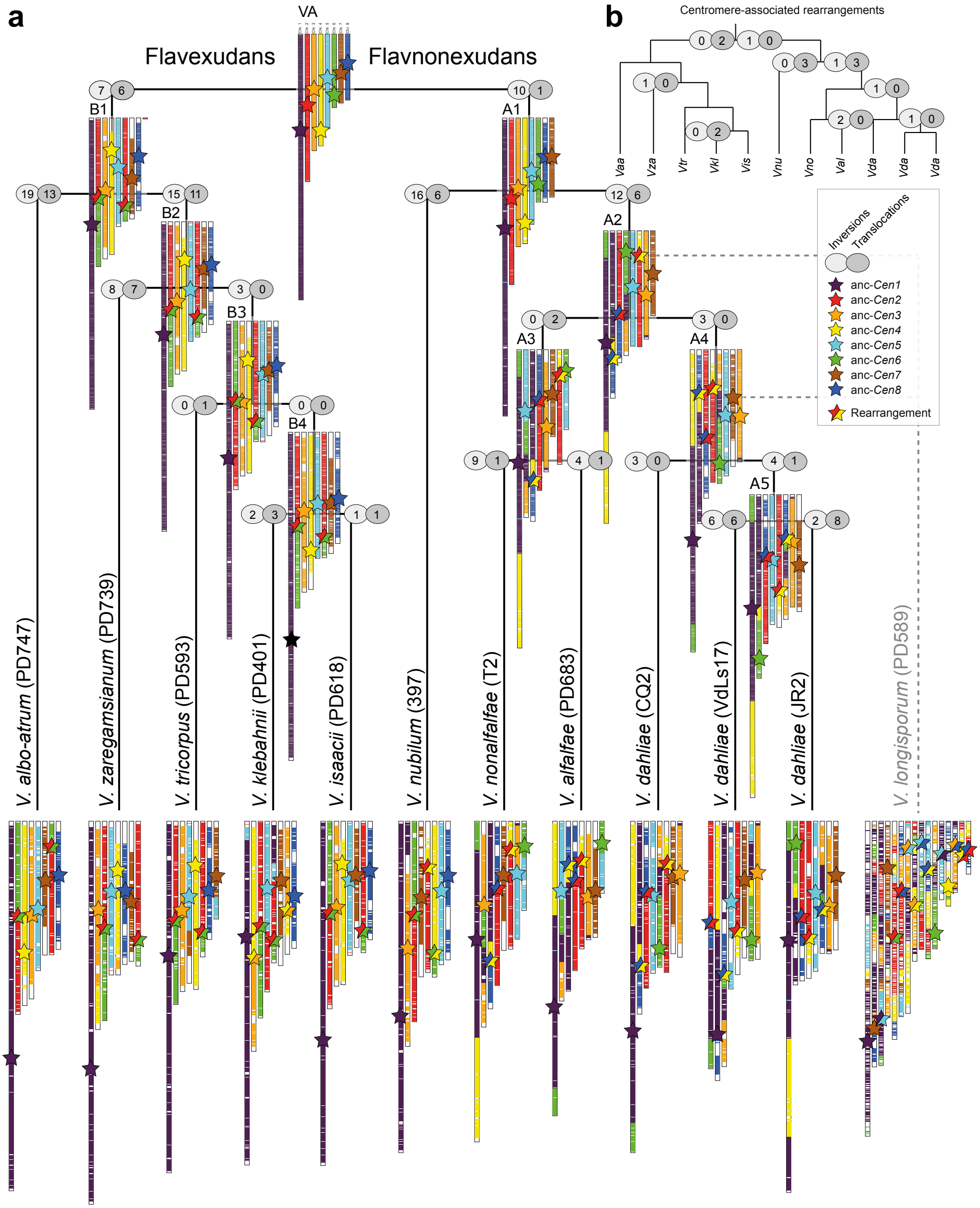


FIGURE 6 | Centromeres contribute to karyotype evolution in *Verticillium.* (a) Relationship of the ten members of the genus *Verticillium*. The allodiploidization event forming *V. longisporum* is indicated by dashed lines [260,273]. The chromosomal evolution within the haploid members of the genus was reconstructed using AnChro [272]. The chromosomal structure of the nine species is shown in relation to the last common ancestor of the genus. The approximate locations of the centromeres are indicated by stars. The number of chromosomal rearrangements (inversions and translocations) are displayed for each branch, and centromeres that co-localize in proximity to chromosomal rearrangements are highlighted by two-colored stars. (b) The number of major chromosomal rearrangements that occurred at, or in close proximity of, centromeres are shown along the branches depicting the *Verticillium* species phylogeny shown in (a).

Discussion

Centromeric regions are among the most rapidly evolving genomic regions [243–246,257], yet centromere evolution has only been systematically studied in few fungi [241,242,246,257]. Here, we took advantage of the fungal genus *Verticillium* and used a combination of genetic and genomic strategies to identify and characterize centromere organization and evolution. *Verticillium* centromeres are characterized as large regional centromeres that are repeat-rich and embedded in heterochromatin. We furthermore show that centromeres contribute to the karyotype evolution of *Verticillium*. Finally, we demonstrate that *VdLTRE9* is a hallmark of centromeres in some *Verticillium* species, while species that lack *VdLTRE9* display a divergent repeat content.

Centromeres in fungi, plants, and animals co-localize within the nucleus [105,245,268–271,274], a phenomenon that can be exploited for their identification [105,268]. Here, we used Hi-C to first establish chromosome-level genome assemblies and subsequently identify centromeres in every *Verticillium* species, and we demonstrate that centromere locations are in agreement with CenH3-binding. While we obtained chromosome-level genome assemblies for all species, Hi-C scaffolded genome assemblies could still contain partially collapsed repeats and assembly gaps, in particular for short-read assemblies [275]. With the exception of *V. nonalfalfae*, we observed only few sequencing gaps and no evidence that would point to collapsed repeats at centromeres, suggesting that the inferred centromeres are of high quality. *Verticillium* centromere sizes differ , which is likely not driven by assembly artefacts, and centromeres in most *Verticillium* species are larger than in *Z. tritici* [121], *C. neoformans, M. oryzae*, or *Fusarium graminearum* [243,246,257], yet smaller than in *N. crassa* [255]. Species of the Flavexudans clade typically encode fewer repeats than species of the clade Flavnonexudans clade [75,122,276], and *V. nubilum*, *V. longisporum*, and *V. dahliae* are particularly rich in repeats when compared with other *Verticillium* species [75,122,261,263,276]. Thus, increased centromere sizes positively correlate with overall increased repeat contents.

Using fragmented genome assemblies, we previously identified chromosomal rearrangements during *Verticillium* evolution, which contributed to the formation of hypervariable LS regions containing genes with important roles in pathogens virulence [59,74,75]. Thus, we proposed that chromosomal rearrangements in *Verticillium* contributed to genetic diversity and adaptation in the absence of sexual recombination [59,67,75]. Chromosome-level genome assemblies for an entire genus enabled unprecedented analyses of the karyotype evolution over longer evolutionary timescales. Here, we observed extensive chromosomal rearrangements and provide evidence that some rearrangements at centromeres contributed to karyotype evolution, most of which occurred early during the divergence of *Verticillium*. Chromosomal rearrangements at centromeres occur in the fungal yeasts *Candida*, *Cryptococcus,* and *Malassezia* [241,242,262], and synteny breakpoints have been identified between mammals and chicken [277], suggesting that centromeres often contribute to karyotype evolution. The emergence of chromosomal rearrangements at centromeres could be facilitated by their repeat-rich nature [241,242]. For example, centromeres in *Malassezia* are enriched with an AT-rich motif that could facilitate replication fork stalling, which leads to double strand DNA breaks [241]. Repeats localized outside of centromeres in *V. dahliae* contribute to chromosomal rearrangements [74], and thus it seems plausible that centromeric repeats similarly contribute to chromosomal rearrangements. It is tempting to speculate that the additional larger AT- and repeat-rich regions outside of the centromeres (e.g. on chromosome 1, 7, or 8 of *V. dahliae* strain JR2) might have been involved in chromosomal rearrangements. However, based on our ancestral chromosome reconstruction these regions, and even the entire chromosome (e.g. chromosome 8), are conserved and do not co-localize with any of the predicted large-scale translocations, even though smaller rearrangements might have occurred that have remained undetected. Chromosomal rearrangements often do not only lead to changes in chromosome organization but also in chromosome number [241,242]. While we observed chromosomal rearrangements, all extant and ancestral genomes contained eight chromosomes, suggesting that eight chromosomes are a stable configuration for all *Verticillium* species.

Centromere position and function are thought to be driven by the protein complement (e.g. CenH3 localization) and by heterochromatin formation rather than by specific DNA sequences [243,245,278]. In *V. dahliae*, we observed the co-occurrence of CenH3 with H3K9me3 and DNA methylation. This suggests that DNA methylation, as previously reported in *N. crassa* and in *C. neoformans* [246,255], is also a feature of centromeric DNA in *V. dahliae*. Co-localization of CenH3 with H3K9me2/3 and DNA methylation has been reported for *N. crassa* [255] and *C. neoformans* [246]. In contrast, H3K9me3 and H3K27me3 are absent from centromeres in *Z. tritici* [121]. H3K4me2 borders most centromeres in *Z. tritici* [121], and is associated with centromeres in *S. pombe* and some animals and plants [278–281]. H3K4me2 has not been observed at centromeres in most fungi, including *V. dahliae*, and in the oomycete plant pathogen *Phytophthora sojae* [258]. Changes in heterochromatin in *N. crassa* leads to altered CenH3 positioning [255], suggesting that heterochromatin is similarly required for centromere maintenance and function in *V. dahliae*. Elevated AT-levels in repeat-rich heterochromatic regions can be caused by RIP mutations [218,245,255,256]. RIP-like mutations have been previously reported in some repeats in *V. dahliae* [139,191], and we observed strong RIP signals at centromeres. Due to its presumably asexual nature [67], the occurrence of RIP in *V. dahliae* is controversial [62,74,139]. Noteworthy, mutational signatures resembling RIP have recently been observed in *Z. tritici* propagated through mitotic cell divisions, pointing to the existence of a mitotic version of a RIP-like process [218]. Thus, we conclude that RIP was an active process in *V. dahliae* at some point in evolution, or that RIP-like processes outside of the sexual cycle occur in *V. dahliae.* Furthermore, a mechanistic link between AT-rich RIP'ed DNA, H3K9me3 deposition and DNA methylation has been established in *N. crassa* [181], suggesting that these processes are also connected in *V. dahliae*.

Centromeres are often enriched for a variety of different retrotransposons and other repetitive elements [245,246,255,257,258,282–284]. We similarly observed that centromeres in all *Verticillium* species are repeat-rich. Repeats and their remnants identified at centromeres typically also occur outside of centromeres, as observed in *M. oryzae* [257] and *N. crassa* [255] for instance. Strikingly, we observed that a single degenerated LTR retrotransposon, *VdLTRE9*, is strongly associated with centromeres in some *Verticillium* species, while it is absent from LS regions in *V. dahliae*. The association of specific retrotransposons to centromeres has also been observed in the yeasts *Ogataea polymorpha* [283], *Debaryomyces hansenii* [282], and *Scheffersomyces stipitis* [284], where a retrotransposon related to *Ty5* is enriched at centromeres. Similarly, centromeres in *Cryptococcus* contain six retrotransposons (*Tcn1-6*) that occur nearly exclusively at centromeres [246]. Centromeres of *P. sojae* contain multiple types of repeats, but they are enriched for a single element called CoLT (*Copia*-Like Transposon) [258]. The strong associations of specific repeats to centromeres could directly or indirectly link these elements to centromere function. Functional centromeres as observed here are also heterochromatic and contain CenH3. AT-rich repetitive elements can direct heterochromatin formation via DNA methylation and H3K9me3 deposition in *N. crassa* [97,181], a phenomenon that can also occurs at repeats outside of centromeres [97]. Heterochromatin occurs at centromeres but also at repeat-rich regions outside of centromeres in *V. dahliae*, thus the repeat-rich nature of centromeres is likely not sufficient to direct CenH3 deposition. In *S. pombe* heterochromatin formation is directed by short interfering RNAs (siRNA) derived from flanking repetitive elements via RNAi [285,286], and RNAi and heterochromatin mediate CenH3 localization at centromeres [287,288]. RNAi is also important for centromere maintenance and evolution in *Cryptococcus*, as RNAi deficient species have smaller centromeres than RNAi proficient ones [246]. Interestingly, centromere-specific elements (*Tcn1-6*) in RNAi-proficient species are typically full-length elements while only remnants can be found in RNAi-deficient species, which could be caused by recombination between elements [246]. Furthermore, the genome size of RNAi-deficient species is smaller than of RNAi-proficient ones, and centromere size reduction is at least partially responsible for genome size differences [246]. In *Verticillium*, centromere size differences correlate with an increase in repeat content and the recruitment of *VdLTRE9*, which is highly fragmented and likely non-active. Genome size differences exist in haploid *Verticillium* (33 Mb – 36 Mb; Table S1c), yet these do not seem to correlate with centromere sizes. Even though key components of the RNAi machinery exist in all *Verticillium* species [289] (Table S1e), we know only little about their biological functions. Similarly, to *C. neoformans*, we observed no transcriptional activity of *VdLTRE9* or any other repeat at centromeres, but it is unclear if this silencing is mediated by RNAi, is a consequence of their heterochromatic nature, is due to their fragmentation, or a combination of these. Ultimately, unravelling how specific elements contribute to centromere identify necessitates future experiments. *VdLTRE9* occurs only in some *Verticillium* species and has likely been recruited to centromeres subsequent to the divergence of *V. nubilum*. Conversely, these observations raise further questions on the roles of repeats and mechanisms of centromeric identity in species without *VdLTRE9*. Repeats drive the formation of chromosomal rearrangements, which are crucial for the formation and maintenance of LS regions, and thus are important drivers of *Verticillium* genome evolution and function [74,191]. Here we highlight their contributions to centromere diversity within the fungal genus *Verticillium* and demonstrate that also centromeres contributed to chromosomal evolution. Our analyses provide the framework for future research into the diversity or convergence of mechanisms establishing centromere identity and functioning, to elucidate roles of centromeres in generating genomic diversity in fungi.

Materials and methods

Construction of *Verticillium dahliae* transformants expressing FLAG-tagged CenH3

CenH3 and H3 homologs were identified in the predicted proteomes of *V. dahliae* strain JR2 [122] and selected other fungi through a BLAST sequence similarity search (blastp v2.9.0+; default settings, e-value cutoff 1e-20) [290,291] using the *N. crassa* CenH3 (Q7RXR3) and H3 (P07041) sequences as queries. Missing homologs of CenH3 or H3 were identified using manual BLAST (tblastn v2.9.0+; default settings) [290,291] and exonerate (v2.2.0; default settings) [292] searches against the genome sequences. Protein sequences of selected CenH3 and H3 proteins were aligned using mafft (v7.271; default settings, LINSi) [135], and poorly aligned regions in the alignment were removed using trimAl (v1.2; default settings) [293]. A phylogenetic tree was inferred with maximum-likelihood methods implemented in IQ-tree (v1.6.11) [294] and robustness was assessed by 1,000 rapid bootstrap replicates.

To construct the N-terminally FLAG-tagged CenH3 strain of *V. dahliae,* a recombinant DNA fragment was constructed into the binary vector PRF-HU2 [124] or PRF-GU2 for homologous recombination. The CenH3 locus, from *V. dahliae* strain JR2, was amplified as 3 fragments with overlapping sequences (Table S1f). The 5' most fragment containing the promoter was amplified using primers A + B, the ORF with primers C+D, the Hyg promoter and ORF with primers E+F, and the 3' end of the CenH3 locus with primers G+H. The four fragments were combined by overlap PCR using primers A + H and cloned into a *Psp*OMI and *Sph*I linearized vector using Gibson Assembly. The vector construction was confirmed by Sanger sequencing. Vectors were transformed to *Verticillium* with *Agrobacterium*-mediated transformation [123]. Correct homologous recombination and replacement at the *CenH3* locus was verified by PCR amplification using primer I+J (Fig. S1b, Table S1f). Correct translation of the recombinant protein was assessed using Western analyses with anti-FLAG antibody (Fig. S1c). Briefly, proteins were extracted from 5-day old cultures grown in 100 ml Potato Dextrose Broth at 22˚C with continuous shaking at 120 rpm. Mycelium was collected by straining over a double layer of miracloth and subsequently snap-frozen in liquid nitrogen and ground with a mortar and pestle using liquid nitrogen. Approximately 0.3 g of ground mycelium was resuspended in 600 µL protein extraction buffer (50 mM HEPES pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% glycerol, 0.02% NP-40, 2 mM Phenylmethanesulfonyl fluoride (PMSF), 100 µM Leupeptin, 1 µg/mL Pepstatin), briefly vortexed, incubated on ice for 15 min and centrifuged at 4˚C at 8,000 g for 3 min. The supernatant was collected by transferring 20 µL to a new tube to serve as the input control and the remaining ~500 µL was transferred to a fresh microcentrifuge tube with 15 µL of Anti-FLAG M2 affinity gel (catalog number A2220, Sigma-Aldrich, St. Louis, Missouri, United States) and incubated while rotating at 4˚C for 1 h. Samples were centrifuged at 5,000 g, 4˚C for 3 min, after which the supernatant was discarded. Samples were washed with 500 µL of lysis buffer, and the centrifugation and washing were repeated three times. Protein was eluted from the resin by adding 15 µL of lysis buffer, 20 µL of 2x Laemmli loading buffer (4% SDS, 20% glycerol, 0.004% bromophenol blue, 125 mM Tris HCL pH 6.8) and boiled at 95˚C for 3 min. Protein samples were separated on a 12% polyacrylamide gel, and subsequently transferred to PVDF membranes, blocked in 5% BSA, washed twice in TBST, and incubated with 1:3500 anti-FLAG antibody (monoclonal anti-FLAG M2; Merck KGaA, Darmstadt, Germany).

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq)

For each *V. dahliae* genotype, one million spores were added to 100 ml Potato Dextrose Broth and incubated for 7 days at 22°C with continuous shaking at 120 rpm. Mycelium was collected by straining over a double layer of miracloth and subsequently snap-frozen in liquid nitrogen and ground with a mortar and pestle using liquid nitrogen. All ground material (0.5-1 gram per sample) was resuspended in 4 mL ChIP Lysis buffer (50 mM HEPES-KOH pH7.5, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% NaDOC) and dounced 40 times in a 10 cm^3 glass tube with tightly fitting pestle on 800 power with a RZR50 homogenizer (Heidolph, Schwabach, Germany), followed by five rounds of 20 seconds sonication on ice with 40 seconds of resting in between rounds with a Soniprep 150 (MSE, London, UK). Samples were redistributed to 2 mL tubes and pelleted for 2 min at maximum speed in a tabletop centrifuge. Supernatants were pooled per sample in a 15 mL tube together with 25 µL α-FLAG M2 magnetic beads (Sigma-Aldrich, St. Louis, Missouri, United States), incubated overnight at 4°C and continuous rotation. Beads were captured on a magnetic stand and washed with wash buffer (50 mM Tris HCl pH 8, 1 mM EDTA, 1% Triton X-100, 100 mM NaCl), high-salt wash buffer (50 mM Tris HCl pH 8, 1 mM EDTA, 1% Triton X-100, 350 mM NaCl), LiCl wash buffer (10 mM Tris HCl pH8, 1 mM EDTA, 0.5% Triton X-100, 250 mM LiCl), and TE buffer (10 mM Tris HCl pH 8, 1mM EDTA). Chromatin was eluted twice from beads by addition of 100 µL pre-heated TES buffer (100 mM Tris HCl pH 8, 1% SDS, 10 mM EDTA, 50 mM NaCl) and 10 minutes incubation at 65°C. 10 mg/mL 2 µL Proteinase K was added and incubated at 65°C for 5 hours, followed chloroform extraction. DNA was precipitated by addition of 2 volumes 100% ethanol, 1/10th volume 3 M NaOAc pH 5.2 and 1/200th volume 20mg/mL glycogen, and overnight incubation at -20°C.

Sequencing libraries were prepared using the TruSeq ChIP Library Preparation Kit (Illumina) according to the manufacturer's instructions, but without gel purification and with use of the Velocity DNA Polymerase (BioLine, Luckenwalde, Germany) for 12 cycles of amplification for the FLAG-CenH3. H3K4me2 ChIP was performed as described previously [191], using an α-H3K4me2 antibody (#39913, ActiveMotif; Carlsbad, California, United States). Single-end (125 bp) sequencing was performed on the Illumina HiSeq2500 platform at KeyGene N.V. (Wageningen, the Netherlands).

Chromatin confirmation capturing followed by high-throughput sequencing (Hi-C)

We determined the inter- and intra-chromosomal contact frequencies using Hi-C in *V. dahliae* strains CQ2, JR2, and VdLs17, as well as in *V. albo-atrum* strain PD747, *V. alfalfae* strain PD683, *V. isaacii* strain PD618, *V. klebahnii* strain PD401, *V. longisporum* strain PD589, *V. nonalfalfae* strain T2, *V. nubilum* strain 397, *V. tricorpus* strain PD593, and *V. zaregamsianum* strain PD739. For each strain, one million spores were added to 400 mL Potato Dextrose Broth and incubated for 6 days at 22˚C with continuous shaking at 120 rpm. Mycelium was collected by straining over double layer miracloth and 300 mg (fresh weight) was used as input for generating Hi-C sequencing libraries with the Proximo Hi-C kit (Microbe) (Phase Genomics, Seattle, WA, USA), according to manufacturer's instructions. Briefly, samples were firstly crosslinked for 15 minutes at room temperature. Crosslinked mycelium was treated with fungal cell lysis solution (10 mM beta-mercaptoethanol, 15 mg/mL Glucanex®, dissolved in phosphate buffered saline at pH 7.4) for 1 hour at 30˚C, followed by snap freezing in liquid nitrogen and grinding with a plastic pestle to obtain a powder. The resulting material was further lysed using the lysis buffers provided with the Hi-C kit and chromatin was collected by centrifugation. Next, chromatin was fragmented at 37˚C for 1 hour and proximity ligation was performed at room temperature for 4 hours. Reverse crosslinking was performed overnight at 65˚C. The resulting soluble DNA was purified and bound to streptavidin beads. Library preparation was then performed, followed by on-bead library amplification by PCR (72˚C for 5 min; 98˚C for 30 s; 15 cycles of: 98˚C for 10 s, 62˚C for 20 s, 72˚C for 50 s). Libraries were cleaned up and eluted from the beads. Final yields were determined by quantification using a Qubit® 2.0 fluorometer (Invitrogen). Hi-C sequencing libraries of V. *dahliae* strains CQ2, JR2. and VdLs17 were paired end (2x125 bp) sequenced on the Illumina HiSeq2500 platform at KeyGene N.V. (Wageningen, the Netherlands). Hi-C sequencing libraries of the other *Verticillium* species were paired-end (2x150 bp) sequenced on the NextSeq500 platform at USEQ (Utrecht, the Netherlands).

In vitro transcriptome profiling using RNA-seq

RNA sequencing of *V. albo-atrum* strain PD747, *V. isaacii* strain PD618, *V. klebahnii* strain PD401, *V. longisporum* strain PD589, *V. nonalfalfae* strain T2, *V. nubilum* strain 397, *V. tricorpus* strain PD593, and *V. zaregamsianum* strain PD739 as described previously [191]. Single-end (50 bp) sequencing was performed on the BGISeq500 platform at BGI (BGI Hong Kong).

Analyses of high-throughput sequencing data

High-throughput sequencing libraries (Table S1a) have been analyzed as follows: Illumina reads were quality-filtered and trimmed using trimmomatic (version 0.36) [295]. Sequencing reads were trimmed and filtered by removing Illumina TruSeq sequencing adapters (settings seed mismatches 2, palindrome clip threshold 30, and simple clip threshold 10), removal of low-quality leading or trailing bases below quality 5 and 10, respectively, and 4-base sliding window trimming and cutting when average quality per base dropped below 15. Additionally, filtered and trimmed reads < 90 nt were removed from further analyses. Filtered and trimmed reads were mapped to the corresponding genome assembly with Bowtie2 (default settings) [296], and mapping files were converted to bam-format using samtools (v 1.8) [149]. Genomic coverage was determined using deepTools (v3.4.1; bamCoverage) [297] by extending sequencing reads to 147 bp followed by RPGC normalization with a bin-size of 1,000 bp and smoothening of 3,000 bp. To assess between sample variability, we used deepTools (v3.4.1, plotPCA) [297] to generate principle component analyses. Furthermore, we employed deepTools (v3.4.1, multiBigwigSummary) [297] to summarize genomic coverages of over genes, repetitive elements, and genomic windows (5 kb windows with 500 bp slide). Genomic regions enriched for FLAG-CenH3 were identified using MACS2 (v2.1.1) (broad peak option; broad cutoff 0.0025) [298].

To determine DNA (cytosine) methylation, we utilized sequencing data of bisulfite treated genomic DNA previously generated for *V. dahliae* strain JR2 [191]. Sequencing reads were mapped to the *V. dahliae* strain JR2 genome assembly as previously described [191]. Subsequently, the number of reads supporting cytosine methylation in CG-context were extracted, and weighted CG-methylation levels were calculated over genes, repetitive elements, and genomic windows (5 kb window size with 500 bp slide) [126]; weighted CG-methylation was defined as the sum of reads supporting cytosine methylations divided by the sum of all reads occurring at all CG sites in the respective regions. Sites with less than four reads were not considered.

To improve the genome assemblies of the *Verticillium* species, we mapped Hi-C sequencing reads to genome assemblies of *V. dahliae* strain CQ2, *V. albo-atrum* strain PD747, *V. alfalfae* strain PD683, *V. isaacii* strain PD618, *V. klebahnii* strain PD401, *V. longisporum* strain PD589, *V. nonalfalfae* strain T2, *V. nubilum* strain 397, *V. tricorpus* strain PD593, and *V. zaregamsianum* strain PD739 using Juicer (v1.6) with early stage setting [299]. The contact matrices generated by juicer were used by the 3D de novo assembly (3D-DNA) pipeline [300] (v180922) with a contig size threshold of 1000bp to eliminate mis-joints in the previous assemblies and to generate improved assemblies. The genome assemblies were manually improved using Juicebox Assembly Tools (JBAT) (v1.11.08) [301] and improved genome assemblies were generated using the 3D-DNA post-review asm pipeline [300]. Centromere locations were determined using a 1 kb-resolution contact matrix in JBAT, by identifying a region per chromosome that displays strong inter-chromosomal interactions, yet weak intra-chromosomal interactions (see Fig. S12, S13).

To assess potential repeat collapses during genome assemblies at centromeric regions, we mapped previously generated short-read data *V. dahliae* strain JR2 and VdLs17, *V. albo-atrum* strain PD747, *V. alfalfae* strain PD683, *V. isaacii* strain PD618, *V. klebahnii* strain PD401, *V. longisporum* strain PD589, *Verticillium nonalfalfae* strain T2, *V. tricorpus* strain PD593, and *V. zaregamsianum* strain PD739 [61,75,261,302] to the genome assemblies using BWA (v0.7.17; mem) [149]. We first used bedtools (v2.29.2) [147] to identify few genomic regions with > 500x coverage. We then applied deepTools (v3.4.1, computeGCBias) [297] to compute GC biases of read depth across the genome, excluding the identified high coverage regions, and used deepTools (v3.4.1, correctGCBias) [297] to correct GC biases, which addresses known biases in sequencing library preparation to ensure even read coverage throughout the genome irrespective of their base composition [303]. We used deepTools (v3.4.1, bamCoverage, bins 50 bp, CPM normalization) [297] to obtain the read coverage throughout the genome, excluding regions containing sequence assembly gaps (Ns). Assuming that collapsed repeats would lead to a local increase in read depth, we used the ratio of the average read coverage at the centromeres and outside of the centromere at each chromosome to correct the inferred centromere sizes. To further validate the genome assembly of regions identified as centromeres of *V. dahliae* strain JR2, the genome assembly was compared to the previously generated optical map [122] using MapSolver (v 3.2; OpGen, Gaithersburg, MD).

The transcriptional activity for genes and repetitive elements in *V. dahliae* strain JR2 was assessed *in vitro* (in Potato Dextrose Broth) using previously generated deep transcriptome datasets [191]. To this end, single-end sequencing reads of three biological replicates were mapped to the *V. dahliae* strain JR2 genome assembly [122] using STAR (v2.4.2a; max. intron size 1 kb and outFilterMismatchNmax to 5) [304]. The resulting mapped reads were summarized per genomic feature (gene or repeat) using summarizeOverlaps [145], converted to counts per million (cpm) mapped reads, and averaged over the three biological replicates.

Sequence analyses of *Verticillium* genome assemblies, centromeres, repeat and gene content

Repetitive elements in the genomes of *V. dahliae* strains JR2, VdLs17 and CQ2 [63,122] were identified as previously described [191]. Briefly, repetitive elements were identified in each genome independently using a combination of LTRharvest [130] and LTRdigest [131] followed by identification of RepeatModeler. Identified repeats in the different *V. dahliae* strains were clustered into a non-redundant library that contained consensus sequences for each repeat family. The repeat library was, if possible, manually curated and annotated using PASTEC [137] or by sequence similarity to previously identified and characterized repeat families [122,139]. Genome-wide occurrences of repeat families were determined using RepeatMasker (v 4.0.9; sensitive option and cutoff 250), and the output was postprocessed using 'One code to find then all' [140]. We only considered matches to the repeat consensus library, and thereby excluded simple repeats and low-complexity regions.

De novo gene and repeat annotation for the Hi-C-improved *Verticillium* genome assemblies, and for *V. dahliae* strains JR2 and VdLs17 as a comparison was performed using the funannotate pipeline [305]. Briefly, repetitive elements were first *de novo* identified using RepeatModeler and masked for gene prediction using RepeatMasker. Subsequently, gene prediction parameters were estimated using *in vitro* RNA-seq data (see above for details; exception: *V. alfalfae* for which no RNA-seq data was available, *V. nonalfalfae* for which publicly available RNA-seq data was used [302], and *V. dahliae* strain JR2 for which in addition to the *in vitro* RNA-seq data generated in this study, also previously generated *in vitro* (xylem sap and half-MS; [191]) as well as long-read nanopore cDNA data [306] was used). Based on the gene prediction parameters, gene prediction was performed with funannotate using a combination of *ab initio* gene predictors, consensus predictions were obtained using Evidencemodeler (v1.1.1) [307], and gene predictions were adjusted using information from the RNA-seq data. Repeat annotation for each genome assembly was based on the *de novo* repeat family consensus sequences obtained with funannotate. Genome-wide occurrences of these repeat families as well as previously defined repeat families for *V. dahliae* (see above) were determined using RepeatMasker (v 4.0.9; sensitive option and cutoff 250), and the output was postprocessed using 'One code to find then all' [140]. *De novo* repeat families overlapping with centromeres in the different species were clustered using BLASTClust (v2.2.26; parameter '-S 60 -L 0.55 -b F -p F'), and subsequently visualized using Cytoscape (v.3.8.0) [308]. Next to RepeatMasker, genome-wide occurrences of the previously determined *VdLTRE9* [122,191] were identified by BLAST searches (blastn v2.9.0+; e-value cutoff 1e-5, no soft-masking and dust, fixed database size 10e6) [290,291], and similarity between VdLTRE9 consensus sequences and the *de novo* predicted repeat families was established using BLAST (blastn, e-value cutoff 1e-5, query coverage > 50%, no soft-masking and dust, fixed database size 10e6).

Repeat and gene density (*V. dahliae* strain JR2 and VdLs17 based on previous gene annotation [306]), GC-content, and composite RIP index were calculated along the genome sequence using sliding windows (5 kb window with 500 bp slide). The composite RIP index (CRI) was calculated according to Lewis et al. [97]. CRI was determined by subtracting the RIP substrate from the RIP product index, which are defined by dinucleotide frequencies as follows: RIP product index = TpA / ApT and the RIP substrate index = (CpA + TpG)/ (ApC + GpT). Overlaps between different genomic features (for example repetitive elements over centromeric regions) was assessed using bedtools (v2.29.2) [147]. Genome-wide data was visualized using R [127] with the packages ggplot2 [309], karyoplotR [228], or Gviz [310], as well as EasyFig [311].

Whole-genome alignments between *V. dahliae* strains JR2, VdLs17, and CQ2 were performed using NUCmer, which is part of the MUMmer package (v 3.1; --maxmatch) [312]. To remove short matches, we only considered alignments longer than 10 kb. Ancestral genome configurations were reconstructed using AnChro [272]. We first determined the synteny relationships between all possible pairs of haploid *Verticillium* genomes and two outgroup genomes (*Plectosphaerella cucumerina* and *Sodiomyces alkalinus*) using SynChro with synteny block stringency (delta parameter) ranging from 2-5 [313]. We then obtained all ancestors by calculating all possible pairs of genomes (G1 and G2) and outgroups (G3,..,G_n) and by varying the delta' (G1 and G2 comparisons) and delta'' (G1/G3..G1/G_n and G2/G3..G2/G_n comparisons) parameters for AnChro. We additionally reconstructed all ancestors starting from the extant genomes in a sequential approach with multiple successive cycles of SynChro and AnChro (delta parameters varied between 2-5). For each ancestor, we chose the optimal reconstructed by the delta parameter combination (delta' and delta'') that minimizes the number of reconstructed chromosomes and rearrangements and at the same time maximizes the number of genes, both guided by the most commonly observed number of chromosomes and genes in all rearrangements. We obtained the number of large-scale rearrangements between reconstructed ancestral genomes and the extant *Verticillium* genomes using ReChro with a delta parameter of 1 [272]. The relationship between chromosomes of the reconstructed ancestors and the extant species in relationship to the common ancestor is generated with SynChro with a delta parameter of 1 [313]. A species phylogeny that uses synteny relationships computed by SynChro (see above) as informative character between the *Verticillium* genomes and the outgroup genomes was reconstructed using PhyChro [314].

Acknowledgments

Work in the laboratories of M.F.S and B.P.H.J.T. is supported by the Research Council Earth and Life Sciences (ALW) of the Netherlands Organization of Scientific Research (NWO). Furthermore, B.P.H.J.T. would like to acknowledge the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2048/1 – Project ID: 390686111. This work was supported in part by a European Molecular Biology Organization postdoctoral fellowship (EMBO, ALTF 969-2013) and Human Frontier Science Program Postdoctoral Fellowship (HFSP, LT000627/2014-L) to D.E.C. This work was partially supported by USDA's National Institute of Food and Agriculture (award no. 2018-67013-28492) through the Plant Biotic Interactions Program, which is jointly offered by the National Science Foundation (NSF) and the National Institute of Food and Agriculture (NIFA) and by the National Science Foundation (award no. 1936800) through Models for Uncovering Rules and Unexpected Phenomena in Biological Systems (MODULUS) program of the Division of Molecular and Cellular Biosciences – Systems and Synthetic Biology to DEC. We thank Utrecht Sequencing Facility for providing sequencing service and data. Utrecht Sequencing Facility is subsidized by the University Medical Center Utrecht, Hubrecht Institute, Utrecht University, and The Netherlands X-omics Initiative (NWO project 184.034.019).

Supplementary data

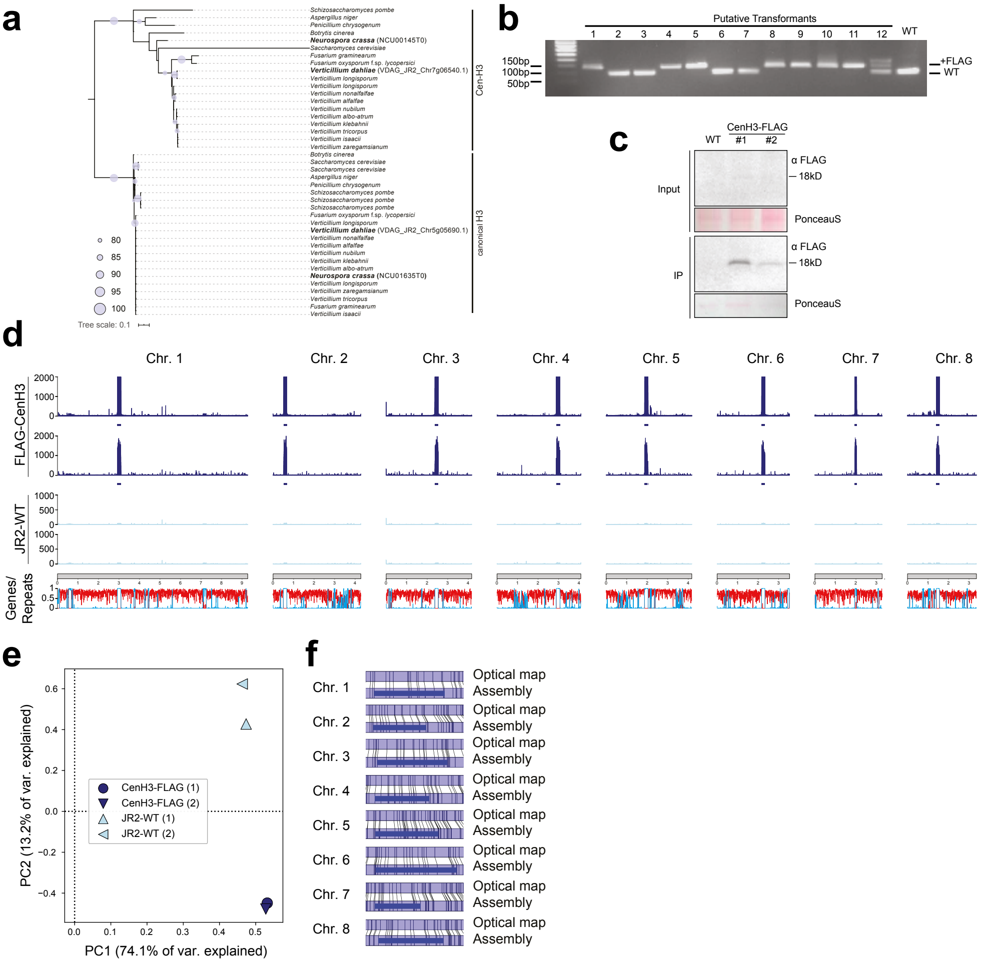


FIGURE S1 | (a) Phylogenetic analyses of the canonical H3 and the centromeric-specific CenH3 in *Verticillium dahliae* (strain JR2) and other fungal genomes. (b-c) Transformation of the coding sequence of N-terminally FLAG-tagged CenH3 directed by its native promoter at the *CenH3* locus in *Verticillium dahliae* strain JR2. (b) Correct homologous recombination and replacement at the *CenH3* locus was verified by PCR amplification was assessed using PCR and (c) Correct translation of the recombinant protein was assessed using Western Blot analyses with anti-FLAG antibody. (d) Sequencing read coverage (RPGC normalization in 1 kb bins with 3 kb smoothening) from ChIP-seq experiments using FLAG-tag antibodies on two independent transformants of *Verticillium dahliae* strain JR2 that express FLAG-tagged CenH3 and the wild-type strain are mapped to the eight chromosomes of *V. dahliae* strain JR2 [122]. Gene (red) and repeat (blue) density are shown below each chromosome. (e) Principal component analysis of the four FLAG-tag ChIP-seq samples (two wild-type and two FLAG-CenH3). (f) Comparison of the centromeric regions with the identified centromeres highlighted as blue block in the genome assembly of *Verticillium dahliae* strain JR2 with a previously generated optical map [59]. Vertical lines display corresponding (*in silico*) restriction sites and their alignment.

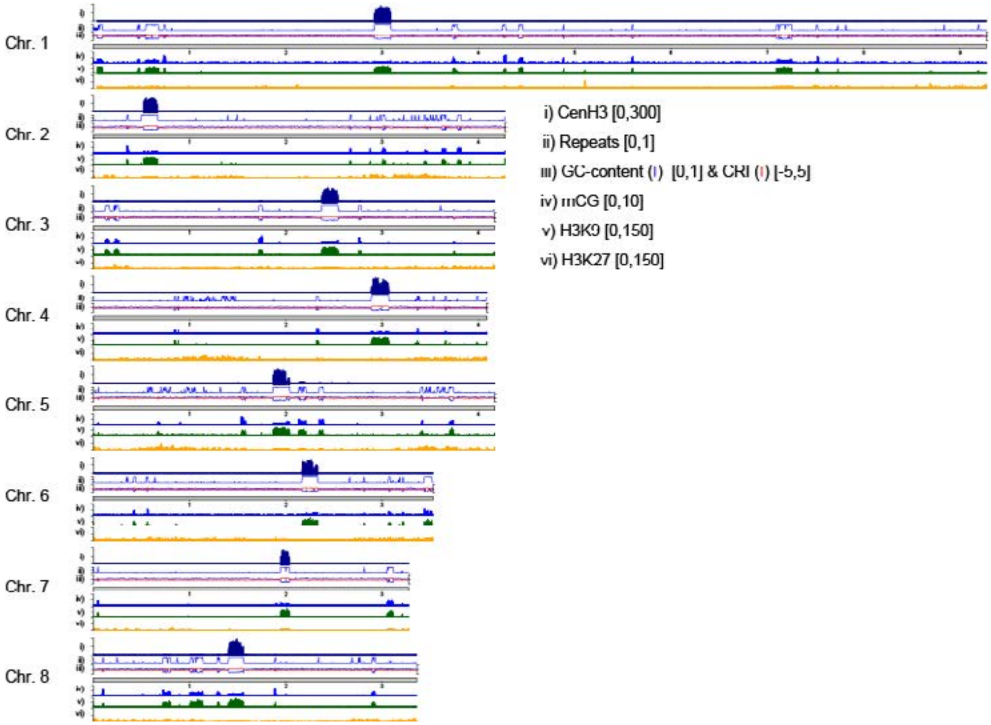


FIGURE S2 | Schematic overview of the eight chromosomes of *Verticillium dahliae* strain JR2 displaying different heterochromatin-associated chromatin modifications (mC, H3K9me3, and H3K27me3) in relation to the centromeres. The different lanes display the FLAG-CenH3 ChIP-seq read coverage (RPGC normalization in 1 kb bins with 3 kb smoothening), the repeat-density, the GC-content, the CRI as well as the weighted cytosine methylation (all summarized in 5 kb windows with 500 bp slide), and the normalized H3K9me3 and H3K27me3 ChIP-seq read coverage (RPGC normalization in 1 kb bins with 3 kb smoothening).

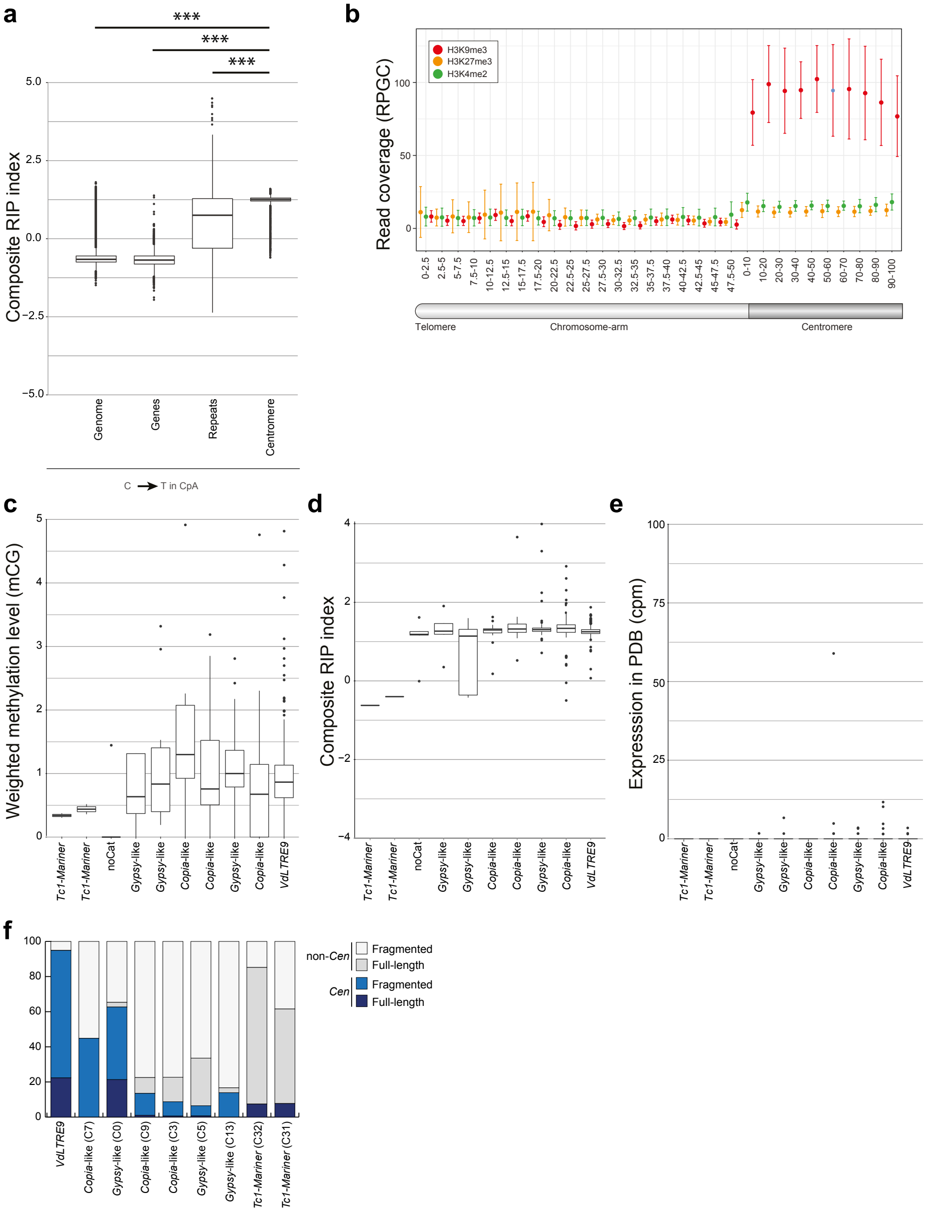


FIGURE S3 | (a) Boxplot displaying the composite RIP index (CRI) of C to T in CA recorded in genomic windows (5 kb, 500 bp slide), per gene, per annotated repeat, and per window overlapping with the CenH3-enriched centromeres. Statistical differences for the indicated comparisons were calculated using the one-sided non-parametric Mann-Whitney test; p-values < 0.001: ***. (b) Summary of H3K4me2 (green), H3K9me3 (red), and H3K27me3 (orange) normalized ChIP-seq read coverage (RPGC normalization in 1 kb bins and 3 kb smoothening) in genomic bins (2.5%) across the chromosomal arms of the eight chromosomes of *Verticillium dahliae* strain JR2 (divided into 2.5% bins) and the centromeric regions (divided into 10% bins). The dots indicate the average ChIP-seq coverage and the whiskers indicate ± 1.5 times the interquartile range. (c-e) Boxplots displaying the (c) weighted methylation levels (CG context), (d) the composite RIP index, and (e) the expression in PDB growth medium (counts per million) for repetitive elements belonging to ten repeat families identified in the eight centromeres in *Verticillium dahliae* JR2. (f) The distribution of different repeat sub-families in centromeres (*Cen*) and across the genome (non-*Cen*) and separated by full-length and fragmented elements.

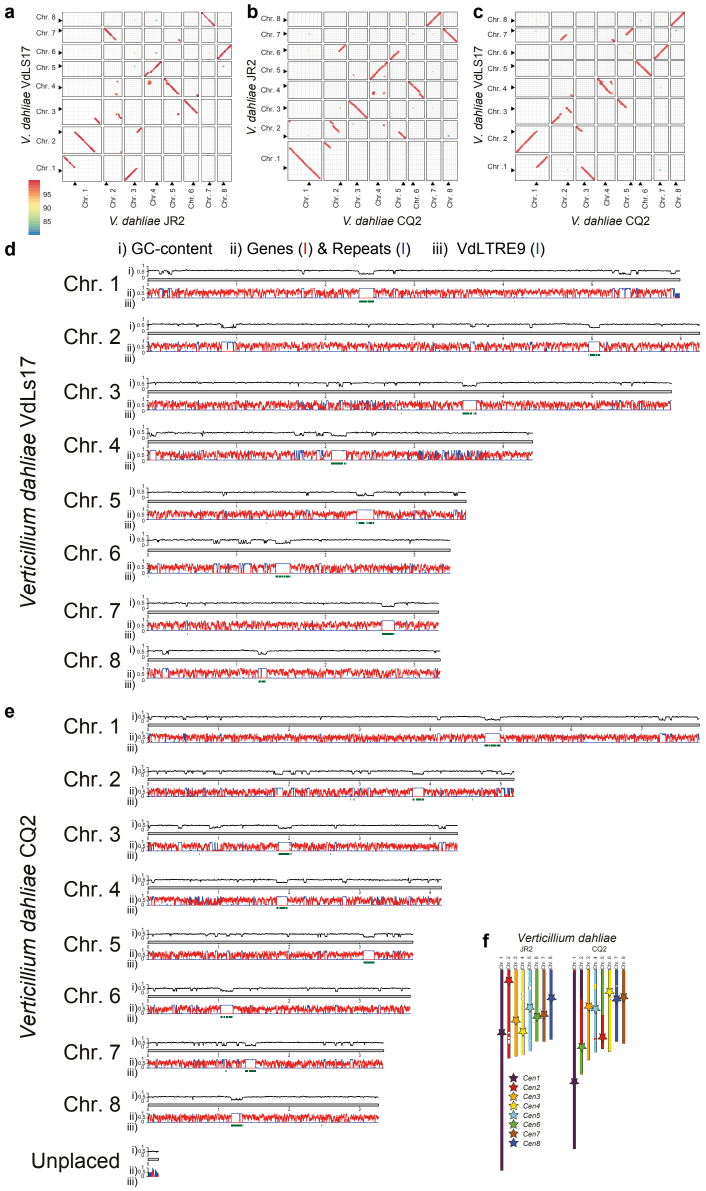


FIGURE S4 | (a-c) Whole-genome alignments between the eight chromosomes of (a) *Verticillium dahliae* strains JR2 and VdLs17 [122], (b) *V. dahliae* strains CQ2 and JR2 [63,122], and (c) *V. dahliae* strains CQ2 and VdLs17 [63,122]. (d-e) Schematic overview of the genome assemblies of *Verticillium dahliae* strains (d) VdLs17 and (e) CQ2. The individual lanes show the GC content, the gene (red) and repeat (blue) density (all summarized in 5 kb windows with 500 bp slide), and the location of the centromere associated *VdLTRE9*. (f) Synteny analyses of the eight chromosomes of *V. dahliae* strains JR2 and CQ2. Schematic overview of the eight chromosomes of *V. dahliae* strain JR2 (left) and the corresponding syntenic regions in *V. dahliae* strains CQ2 (right). Centromeres are indicated by stars, and syntenic centromeres of *V. dahliae* strain CQ2 are colored according to *Cen1-8* of *V. dahliae* strain JR2.

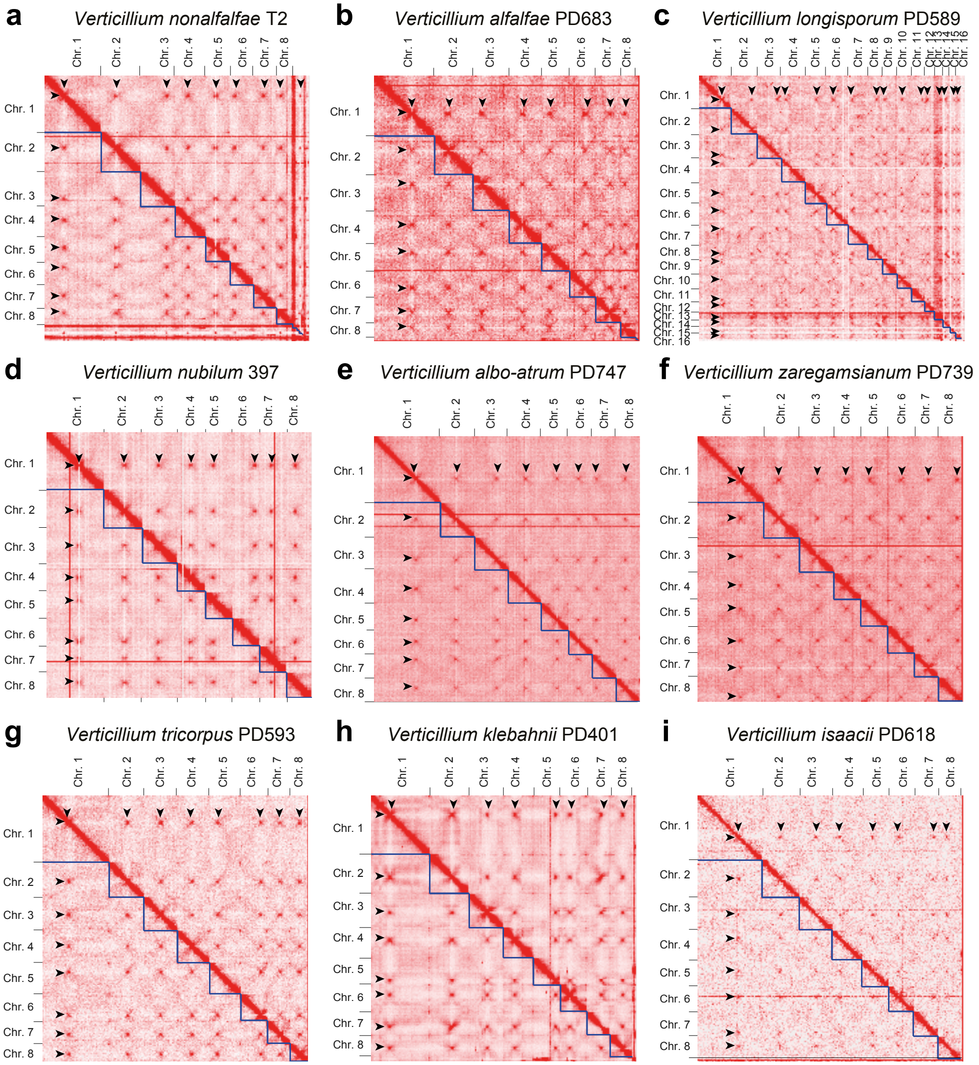


FIGURE S5 | Hi-C contact matrix showing the interaction frequencies between genomic regions in (a) *V. nonalfalfae* (T2), (b) *V. alfalfae* (PD683), (c) the allodiploid *V. longisporum* (PD589), (d) *V. nubilum* (397), (e) *V. albo-atrum* (PD747), (f) *V. zaregamsianum* (PD739), (g) *V. tricorpus* (PD593), (h) *V. klebahnii* (PD401), and (i) *V. isaacii* (PD618). Regions of high inter-chromosomal interaction frequencies are indicative of centromeres and are highlighted by arrow heads, and the blue line indicated boundaries between the pseudo-chromosomes.

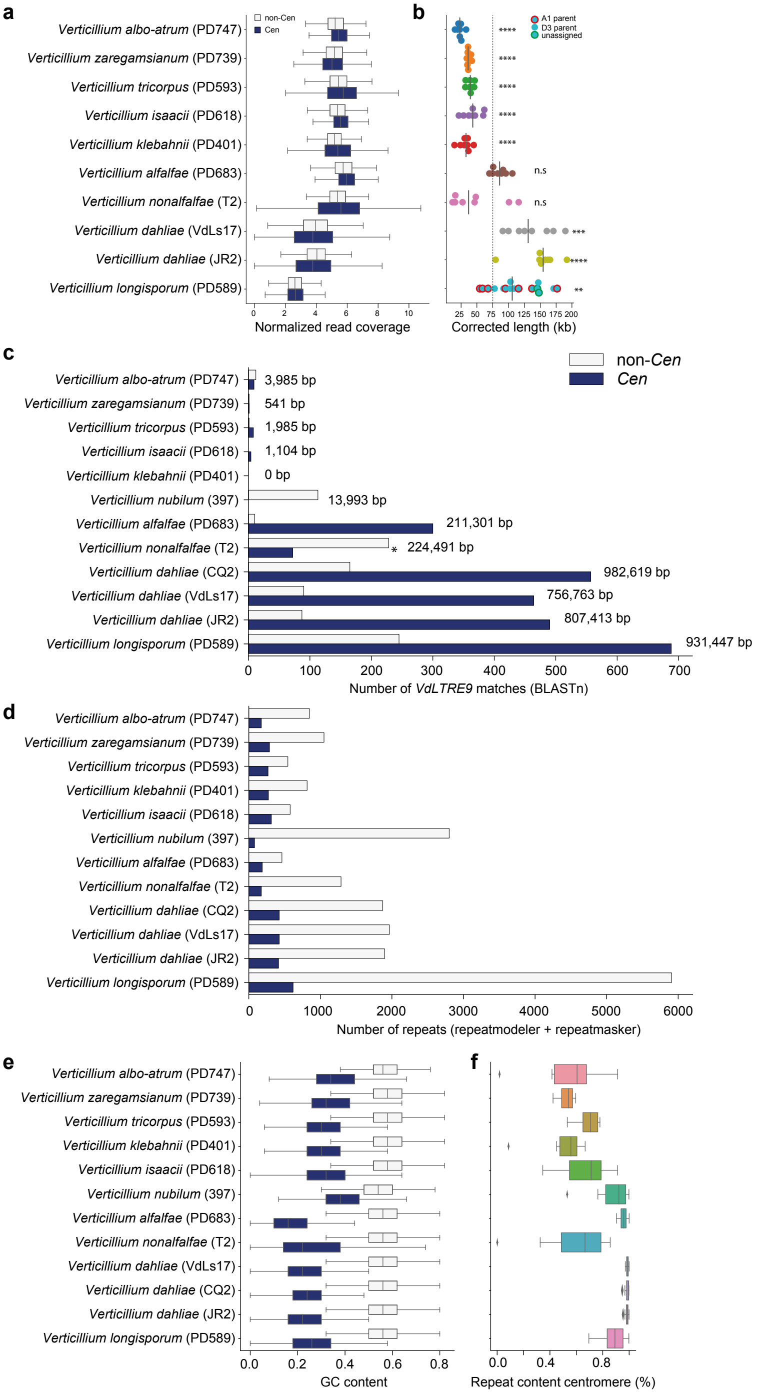


FIGURE S6 | a-b) Comparison of normalized read coverage and corrected centromere lengths for *Verticillium* species for which short-read data is available. (a) Counts per million mapped reads (CPM) normalized read coverage was calculated for GC-biased corrected short-read libraries in 50 bp genomic windows, excluding regions containing assembly gaps (Ns). Genomic windows are summarized in boxplots (outliers not shown) by genomic location, centromeric regions (*Cen*, blue) and non-centromeric regions (non-*Cen*, grey). (b) Centromeric lengths inferred by Hi-C data were 'corrected' based on the ratio of normalized read depth between centromeres and non-centromeric regions per chromosomes. Differences for each species compared to the overall mean were computed using unpaired T-tests; p-values < 0.0001: ****, p-values < 0.001: ***, p-values < 0.01: **, p-values < 0.05: *. (c) The number of BLASTn matches of the *VdLTRE9* consensus element to the genomes of the *Verticillium* species separated by their genomic location, centromeric regions (*Cen*, blue) and non-centromeric regions (non-*Cen*, grey). The overall number of base pairs (bp) covered by the BLASTn matches in each genome sequence is indicated. The asterisk denotes the high number of *VdLTRE9* matches to unassigned, non-*Cen* regions in the genome assembly of *Verticillium nonalfalfae* (T2). (d) The number of repetitive element matches identified by RepeatMasker for each *Verticillium* species based on species/strain-specific repeat libraries generated by RepeatModeler separated by their genomic location, centromeric regions (*Cen*, blue) and non-centromeric regions (non-*Cen*, grey). (e) GC-content of the *Verticillium* genomes in 50 bp windows and separated by their genomic location, centromeric regions (*Cen*, blue) and non-centromeric regions (non-*Cen*, grey). (f) The repeat content of centromeric regions in percent covered sequences in the different *Verticillium* species. Each data point summarized in the boxplot is the repeat content per centromere.

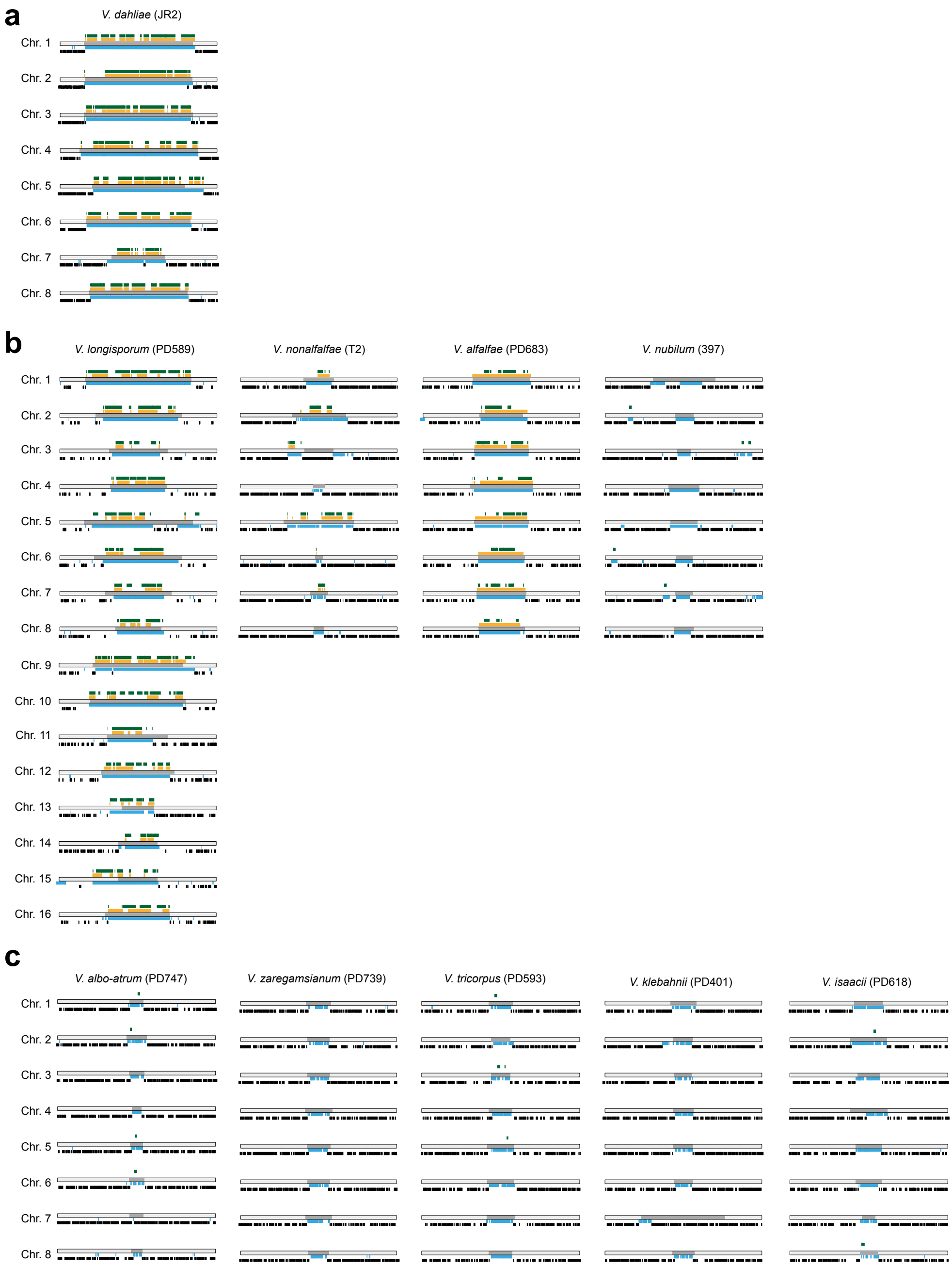


FIGURE S7 | Schematic overview of the centromeric regions (250 kb) in (a) *Verticillium dahliae* strain JR2, in (b) species belonging to clade Flavnonexudans, and in (c) species belonging clade Flavexudans. The centromeres are indicated by dark grey bars. The predicted genes (black) and repeats (blue) are shown below each centromere, and location of *VdLTRE9* (partial) matches (dark green) are shown above each centromere. Repeats that share sequence similarity (BLASTn) to the *VdLTRE9* consensus sequence are shown above each centromere (orange).

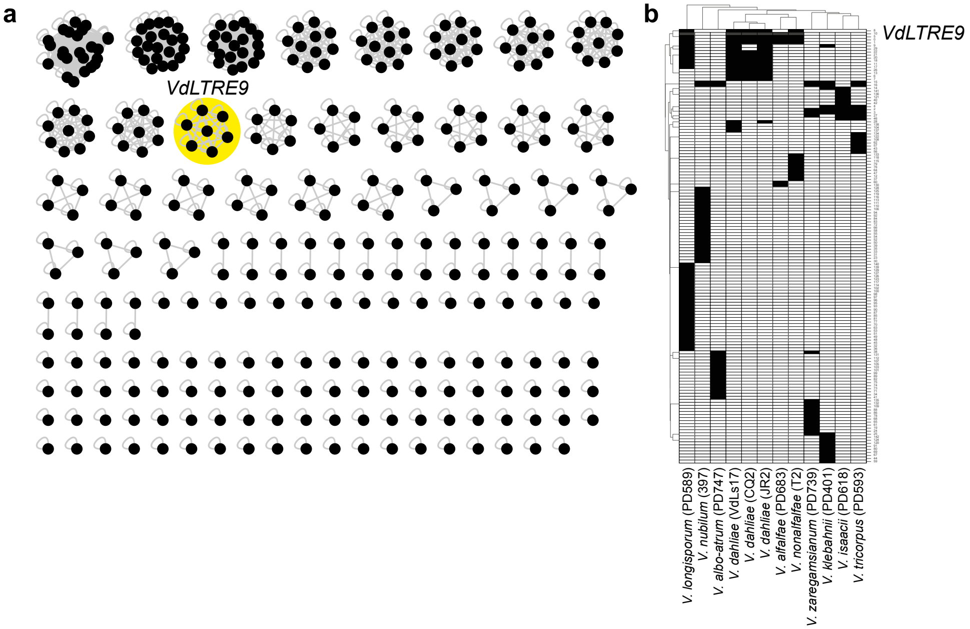


FIGURE S8 | Sequence comparisons of *de novo* repeat families identified with RepeatModeler and RepeatMasker in the genome assemblies of the different *Verticillium* species. Individual repeat family consensus sequences were clustered using BLASTClust. (a) Relationships between different repeat family consensus sequences are displayed as connected graphs. The sub-graph with the consensus sequences with similarity to *VdLTRE9* is highlighted in yellow. (b) The presence/absence matrix indicates the occurrences of different repeat families in the analyzed *Verticillium* species (black present, white absent). The cluster containing consensus sequences with similarity to *VdLTRE9* is highlighted.

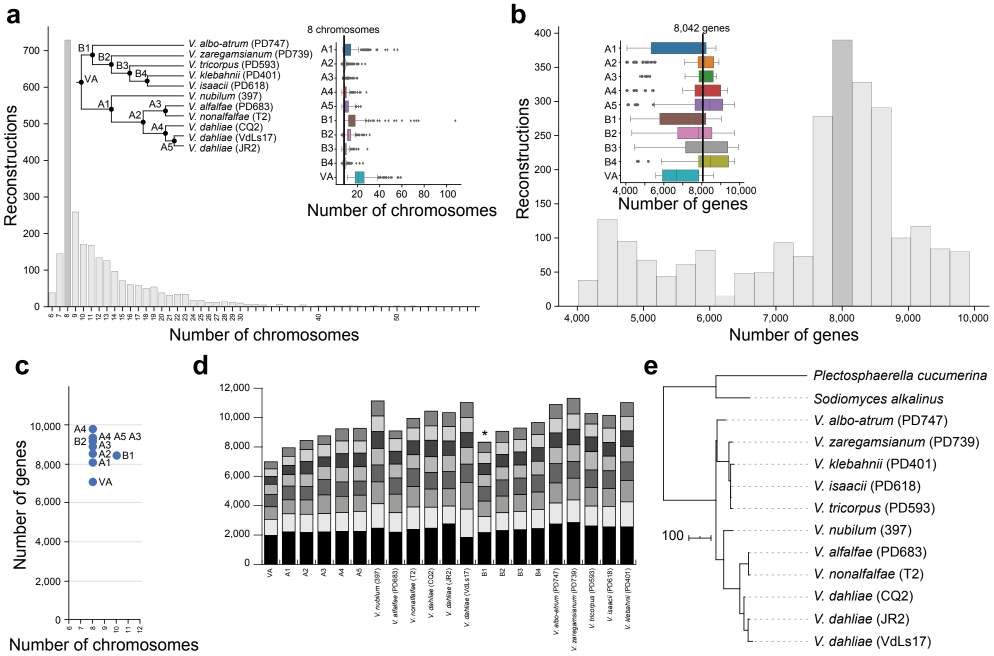


FIGURE S9 | Reconstruction of ancestral genomes within the genus *Verticillium* with AnChro [272]. The number of (a) chromosomes and (b) genes predicted by all potential ancestral reconstructions using different combinations of genomes and stringency parameters. The phylogenetic tree in (a) depicts the relationships between *Verticillium* species and the abbreviations used for the ancestors. The inlays display boxplots to summarize the number of (a) chromosomes and (b) genes per ancestral reconstruction. (c) The number of chromosomes and genes of the chosen 'optimal' reconstruction for each of the internal ancestors. (d) The number of genes per chromosome for each of the reconstructed ancestor and the extant *Verticillium* species. The star highlights the reconstruction for the B1 ancestor that had ten chromosomes, but with two chromosomes with six and two genes. (e) Reconstruction of the *Verticillium* species phylogeny based on synteny relationship using PhyChro [314].

TABLE S1A | Overview of the *Verticillium* sequencing libraries used in this study.

***Verticillium* species**	**Strain**	**Genotype**	**Growth medium**	**Antibody or procedure**	**Accession number**	**Reference**
V. dahliae	JR2	WT	PDB	α-FLAG	PRJNA641329	This study
V. dahliae	JR2	WT	PDB	α-FLAG	PRJNA641329	This study
V. dahliae	JR2	CenH3-FLAG	PDB	α-FLAG	PRJNA641329	This study
V. dahliae	JR2	CenH3-FLAG	PDB	α-FLAG	PRJNA641329	This study
V. dahliae	JR2	WT	PDB	α-H3K4me2	PRJNA641329	This study
V. dahliae	JR2	WT	PDB	α-H3K9me3	PRJNA592220	Cook *et al.* 2020 [191]
V. dahliae	JR2	WT	PDB	α-H3K27me3	PRJNA592220	Cook *et al.* 2020 [191]
V. dahliae	JR2	WT	PDB	ATAC-seq	PRJNA592220	Cook *et al.* 2020 [191]
V. dahliae	JR2	WT	PDB	ATAC-seq	PRJNA592220	Cook *et al.* 2020 [191]
V. dahliae	JR2	WT	PDB	Bisulfite-seq	PRJNA592220	Cook *et al.* 2020 [191]
V. dahliae	JR2	WT	PDB	Bisulfite-seq	PRJNA592220	Cook *et al.* 2020 [191]
V. dahliae	JR2	WT	PDB	Hi-C	PRJNA641329	This study
V. dahliae	VdLs17	WT	PDB	Hi-C	PRJNA641329	This study
V. dahliae	CQ2	WT	PDB	Hi-C	PRJNA641329	This study
V. albo-atrum	PD747	WT	PDB	Hi-C	PRJNA641329	This study
V. alfalfae	PD683	WT	PDB	Hi-C	PRJNA641329	This study
V. isaacii	PD618	WT	PDB	Hi-C	PRJNA641329	This study
V. klebahnii	PD401	WT	PDB	Hi-C	PRJNA641329	This study
V. longisporum	PD589	WT	PDB	Hi-C	PRJNA641329	This study
V. nonalfalfae	T2	WT	PDB	Hi-C	PRJNA641329	This study
V. nubilum	397	WT	PDB	Hi-C	PRJNA641329	This study
V. tricorpus	PD593	WT	PDB	Hi-C	PRJNA641329	This study
V. zaregamsianum	PD739	WT	PDB	Hi-C	PRJNA641329	This study

TABLE S1B | Position of centromeric regions inferred by Hi-C in different *Verticillium dahliae* strains.

			Hi-C centromeres			
Species	**Strain**	**Chr.**	**Position (bp)**	**Length (kb)**	**Overlap CenH3 (kb)**	**Overlap *VdLTRE9* (kb)**
Verticillium dahliae	JR2	1	2,918,001 – 3,091,001	173	171	121
Verticillium dahliae	JR2	2	500,187 – 671,187	171	150	126
Verticillium dahliae	JR2	3	2,371,628 – 2,544,628	173	167	134
Verticillium dahliae	JR2	4	2,882,422 – 3,070,422	188	186	100
Verticillium dahliae	JR2	5	1,865,995 – 2,013,995	148	146	96
Verticillium dahliae	JR2	6	2,166,518 – 2,332,518	166	166	104
Verticillium dahliae	JR2	7	1,952,284 – 2,037,284	85	85	44
Verticillium dahliae	JR2	8	1,404,514 – 1,560,514	156	154	114
Verticillium dahliae	VdLs17	1	2,372,701 – 2,544,701	172	-	129
Verticillium dahliae	VdLs17	2	4,960,001 – 5,082,001	122	-	79
Verticillium dahliae	VdLs17	3	3,542,720 – 3,692,720	150	-	97
Verticillium dahliae	VdLs17	4	2,065,712 – 2,232,712	167	-	109
Verticillium dahliae	VdLs17	5	2,348,685 – 2,542,685	194	-	102
Verticillium dahliae	VdLs17	6	1,439,821 – 1,546,821	107	-	75
Verticillium dahliae	VdLs17	7	2,630,001 – 2,775,001	145	-	118
Verticillium dahliae	VdLs17	8	1,239,898 – 1,334,898	95	-	54
Verticillium dahliae	CQ2	1	4,776,001 – 4,991,001	215	-	173
Verticillium dahliae	CQ2	2	3,749,589 – 3,898,589	149	-	96
Verticillium dahliae	CQ2	3	1,846,205 – 2,036,204	190	-	120
Verticillium dahliae	CQ2	4	1,816,985 – 1,975,985	159	-	98
Verticillium dahliae	CQ2	5	3,060,911 – 3,202,910	142	-	116
Verticillium dahliae	CQ2	6	1,035,725 – 1,198,724	163	-	112
Verticillium dahliae	CQ2	7	1,376,742 – 1,524,741	148	-	120
Verticillium dahliae	CQ2	8	1,175,722 – 1,333,722	158	-	124

TABLE S1C |Genome assemblies of ten *Verticillium* species using Hi-C.

		Genome assembly[a]			Hi-C assembly					
Species	**Strain**	**Genome size (Mb)**	**Contigs/ Scaffolds/ Chromosomes**	**Genes**	**Genome size (Mb)**	**(Pseudo-) Chromosomes**	**Telomeres[c]**	**Un-scaffolded regions (kb)**	**Genes[d]**	**Repeats (%)[f]**
V. dahliae	JR2	36,15	8	11,426	36,15	8	16 (8)	-	10,556	11,32
V. dahliae	VdLs17	35,97	8	10,885	35,97	8	14 (6)	-	11,196	11,08
V. dahliae	CQ2	35,81	17	10,672	35,77	8	16 (8)	149	10,672	11,92
V. nonalfalfae	T2	34,3	349	11,029	34,17	8	10 (2)	2,022	10,464	4,35
V. alfalfae	PD683	32,7	22	10,852	32,68	8	13 (3)	474	9,380[e]	3,21
V. longisporum[b]	PD589	72,61	27	19,197	72,63	16	16 (2)	987	18,890	12,61
V. nubilum	397	38,37	13	-	38,37	8	6 (1)	67	11,377	12,08
V. albo-atrum	PD747	36,47	21	11,202	36,46	8	14 (6)	142	11,149	2,82
V. zaregamsianum	PD739	37,14	48	11,274	37,11	8	12 (4)	132	11,566	3,01
V. tricorpus	PD593	35,13	12	10,636	35,1	8	15 (7)	0,5	10,517	2,81
V. klebahnii	PD401	36,09	45	10,998	36,08	8	10 (3)	503	11,341	2,62
V. isaacii	PD618	35,99	684	10,798	35,61	8	13 (5)	409	10,467	2,98

[a] *Verticillium dahliae* and other *Verticillium* species have been previously sequenced, assembled and/or annotated by Faino et al. 2015 [122], Shi-Kunne et al. 2018 [75], Jakše et al. 2018 [302], Depotter et al. 2019 [63], Cook et al, 2019 [306], and Depotter et al. 2021 [261]

[b] *Verticillium longisporum* PD589 is an allodiploid (see Depotter et al. 2021 [261])

[c] Telomeric repeats (TTAGGG/CCCTAA) in the first 50 nt of each assembled pseudo-chromosome were identified. The number in brackets indicate the number of chromosomes with telomeres on both ends

[d] Gene annotation was performed using the funannotate pipeline (Palmer 2020 [305])

[e] Gene annotation for *Verticillium alfalfae* was performed with funannotate but without RNA-seq data

[f] Repeat annotation (expressed as percentage genome covered by repetitive elements) is performed *de novo* using RepeatModeler and RepeatMasker. The abundance of annotated repeats is comparable to an annotation using the manual corrected *Verticillium dahliae* specific repeat library (see Material and Methods) that yielded 11.02, 11.06, and 11.61% for *V. dahliae* strain JR2, VdLs17, and CQ2, respectively.

TABLE S1D | Position of centromeric regions determined based on Hi-C interactions in different *Verticillium* species.

			Hi-C centromeres		
Species	**Strain**	**Chr.**	**Position (bp)**	**Length (kb)**	**Gaps (bp)** [a]
Verticillium alfalfae	PD683	1	4653001 - 4745001	92	2
Verticillium alfalfae	PD683	2	1776001 - 1845001	69	0
Verticillium alfalfae	PD683	3	1093743 - 1180743	87	3
Verticillium alfalfae	PD683	4	1615743 - 1716743	101	2
Verticillium alfalfae	PD683	5	987482 - 1074482	87	0
Verticillium alfalfae	PD683	6	1984629 - 2059629	75	0
Verticillium alfalfae	PD683	7	1779768 - 1858768	79	0
Verticillium alfalfae	PD683	8	521435 - 595435	74	0
Verticillium longisporum	PD589	1	6403001 - 6572001	169	0
Verticillium longisporum	PD589	2	5673333 - 5810333	137	0
Verticillium longisporum	PD589	3	5392239 - 5485239	93	0
Verticillium longisporum	PD589	4	1267050 - 1357050	90	0
Verticillium longisporum	PD589	5	3104664 - 3277664	173	0
Verticillium longisporum	PD589	6	1784226 - 1924226	140	0
Verticillium longisporum	PD589	7	788226 - 893226	105	500
Verticillium longisporum	PD589	8	2441049 - 2514049	73	0
Verticillium longisporum	PD589	9	647525 - 790525	143	0
Verticillium longisporum	PD589	10	1319942 - 1471942	152	0
Verticillium longisporum	PD589	11	3022672 - 3119672	97	0
Verticillium longisporum	PD589	12	856540 - 973540	117	0
Verticillium longisporum	PD589	13	1782323 - 1833323	51	0
Verticillium longisporum	PD589	14	706938 - 769938	63	0
Verticillium longisporum	PD589	15	1226215 - 1289215	63	0
Verticillium longisporum	PD589	16	688030 - 791030	103	0
Verticillium nonalfalfae	T2	1	2669001 - 2717001	48	7,588
Verticillium nonalfalfae	T2	2	1945484 - 2031484	86	3,213
Verticillium nonalfalfae	T2	3	3320839 - 3366839	46	200
Verticillium nonalfalfae	T2	4	1625372 - 1643372	18	3,189
Verticillium nonalfalfae	T2	5	1224696 - 1336696	112	20,432
Verticillium nonalfalfae	T2	6	629527 - 641527	12	3,027
Verticillium nonalfalfae	T2	7	1292615 - 1321615	29	8,770
Verticillium nonalfalfae	T2	8	566615 - 583615	17	1,002
Verticillium nubilum	397	1	4707001 - 4806001	99	500
Verticillium nubilum	397	2	3020108 - 3050108	30	0
Verticillium nubilum	397	3	2356532 - 2378532	20	0
Verticillium nubilum	397	4	1987044 - 2036044	49	0
Verticillium nubilum	397	5	1126247 - 1169247	43	0

			Hi-C centromeres		
Species	**Strain**	**Chr.**	**Position (bp)**	**Length (kb)**	**Gaps (bp)** [a]
Verticillium nubilum	397	6	3279564 - 3306564	27	0
Verticillium nubilum	397	7	1746411 - 1774411	28	0
Verticillium nubilum	397	8	1320388 - 1351388	31	0
Verticillium albo-atrum	PD747	1	5782001 - 5804001	22	503
Verticillium albo-atrum	PD747	2	2196232 - 2228232	32	4
Verticillium albo-atrum	PD747	3	3069289 - 3093289	24	5
Verticillium albo-atrum	PD747	4	2370289 - 2385289	15	2
Verticillium albo-atrum	PD747	5	2261832 - 2281832	20	3
Verticillium albo-atrum	PD747	6	1403541 - 1427541	24	2
Verticillium albo-atrum	PD747	7	580947 - 602947	22	500
Verticillium albo-atrum	PD747	8	1276204 - 1294204	18	501
Verticillium zaregamsianum	PD739	1	6032001 - 6072001	40	509
Verticillium zaregamsianum	PD739	2	2105027 - 2140027	35	4680
Verticillium zaregamsianum	PD739	3	2540247 - 2576247	36	5
Verticillium zaregamsianum	PD739	4	1683162 - 1727162	44	3
Verticillium zaregamsianum	PD739	5	1173874 - 1211874	38	12
Verticillium zaregamsianum	PD739	6	1900512 - 1936512	36	2426
Verticillium zaregamsianum	PD739	7	1986660 - 2028660	42	9
Verticillium zaregamsianum	PD739	8	2678514 - 2711514	33	547
Verticillium tricorpus	PD593	1	3511001 - 3549001	38	1
Verticillium tricorpus	PD593	2	2481976 - 2511976	30	1
Verticillium tricorpus	PD593	3	2248026 - 2278026	30	4
Verticillium tricorpus	PD593	4	1939836 - 1976836	37	0
Verticillium tricorpus	PD593	5	1055060 - 1097060	42	4
Verticillium tricorpus	PD593	6	2670631 - 2712631	42	2
Verticillium tricorpus	PD593	7	1756965 - 1800965	44	4
Verticillium tricorpus	PD593	8	1311474 - 1348474	37	3
Verticillium klebahnii	PD401	1	2588001 - 2629001	41	3
Verticillium klebahnii	PD401	2	3094540 - 3125540	31	500
Verticillium klebahnii	PD401	3	2523713 - 2551713	28	3
Verticillium klebahnii	PD401	4	1583592 - 1615592	32	4
Verticillium klebahnii	PD401	5	3127250 - 3158250	31	3
Verticillium klebahnii	PD401	6	1351470 - 1381470	30	2
Verticillium klebahnii	PD401	7	2070065 - 2203065	15	8
Verticillium klebahnii	PD401	8	1787795 - 1822795	35	4

			Hi-C centromeres		
Species	**Strain**	**Chr.**	**Position (bp)**	**Length (kb)**	**Gaps (bp)** [a]
Verticillium isaacii	PD618	1	5499001 - 5545001	46	2
Verticillium isaacii	PD618	2	2460985 - 2518985	58	9
Verticillium isaacii	PD618	3	2251196 - 2287196	36	0
Verticillium isaacii	PD618	4	974148 - 1033148	59	1001
Verticillium isaacii	PD618	5	1404268 - 1446268	42	0
Verticillium isaacii	PD618	6	1345140 - 1387140	42	1
Verticillium isaacii	PD618	7	2710634 - 2731634	21	2
Verticillium isaacii	PD618	8	1220645 - 1248645	28	2

TABLE S1E | Occurrence of RNAi components in the genomes of the ten different *Verticillium* species.

	Argonaute 1	**Argonaute 2**	**Dicer-like 1**	**Dicer-like 2**	**RNA-dependent RNA polymerase 1**	**RNA-dependent RNA polymerase 2**	**RNA-dependent RNA polymerase 3**
Verticillium dahliae JR2	yes	yes	yes	yes	yes	yes	yes
Verticillium dahliae CQ2	yes	yes	yes	yes	yes	yes	yes
Verticillium dahliae VdLs17	yes	yes	yes	yes	yes	yes	yes
Verticillium longisporum PD589	yes	yes	yes	yes	yes	yes	yes
Verticillium nonalfalfae T2	yes	yes	yes	yes	yes	yes	no*
Verticillium alfalfae PD683	yes	yes	yes	yes	yes	yes	no*
Verticillium nubilum 397	yes	yes	yes	yes	yes	yes	no*
Verticillium albo-atrum PD747	yes	yes	yes	yes	yes	yes	yes
Verticillium zaregamsianum PD739	yes	yes	yes	yes	yes	yes	yes
Verticillium tricorpus PD593	yes	yes	yes	yes	yes	yes	yes
Verticillium klebahnii PD401	yes	yes	yes	yes	yes	yes	yes
Verticillium isaacii PD618	yes	yes	yes	yes	yes	yes	yes

The protein sequences for the RNAi components were previously identified in *Verticillium nonalfalfae and V. dahliae* have been obtained from Jeseničnik, et al. (2019) [289]. The occurrence of the homologs in all genomes was performed using tblastn and exonerate.

*possible pseudo-gene

TABLE S1F | Primers for cloning fragments for CenH3 FLAG tag in *Verticillium dahliae* strain JR2

Name	Short Name	Sequence (5′ -> 3′)
5′_CenH3_F	A	ttaagtcctcagcgggccACAGCTTCGTTGGCTGGTCTCC
5′_CenH3_R	B	GTAGTCACCGTCGTGATCCTTGTAGTCCATGGTTCAGGACAAGTGCTGATG
CenH3_ORF_F	C	CACGACGGTGACTACAAGGATGACGACGATAAGCCACCACGCTCAGGTACAAGCA
CenH3_ORF_R	D	tcgccgacatcaccgGCCTGGGAAAAGAATCCTACGA
Hyg_F	E	CGACCTACAGTGCTTAGGCGATTAAGTTGGGTAACGC
Hyg_R	F	ATTCTTTTCCCAGGCCggtgatgtcggcgatatagg
CenH3_RB_F	G	ccaacttaatcgcctAAGCACTGTAGGTCGTACACC
CenH3_RB_R	H	ccgccgcaaggaatggtgGGCAGAAGCAAGCGAATTATGTC
Flag_check_F	I	GACCTCTCCGAAAACTCACGC
Flag_check_R	J	GTCCGACTTTCTGCTGGTACC

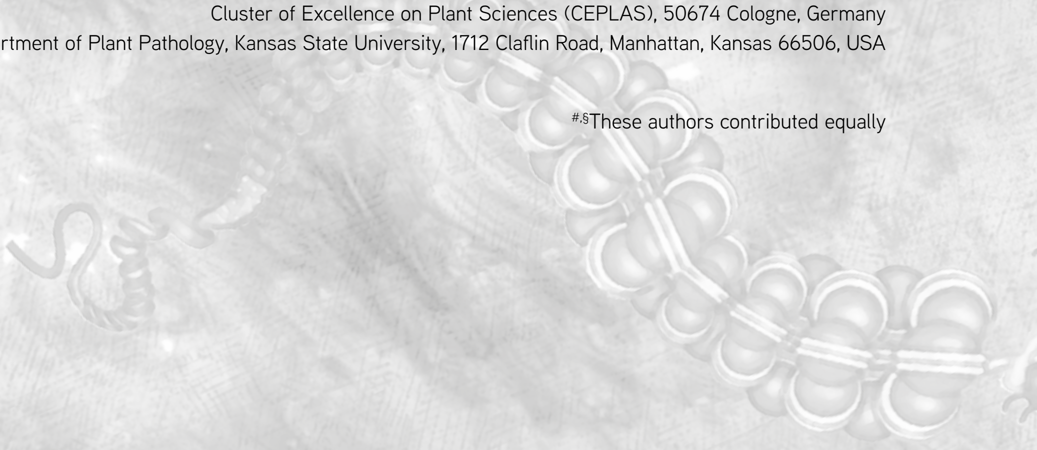
Chapter 6

Local three-dimensional chromatin organization impacts the evolution of adaptive genomic regions in *Verticillium dahliae*

H. Martin Kramer[1,#],
David E. Torres [1,2,#],
Vittorio Tracanna[3,#],
Gabriel L. Fiorin[1],
David E. Cook[1,4],
Michael F. Seidl[1,2,§],
Bart P.H.J. Thomma[1,3,§]

[1]Laboratory of Phytopathology, Wageningen University and Research, Droevendaalsesteeg 1, 6708 PB Wageningen, the Netherlands
[2]Theoretical Biology & Bioinformatics Group, Department of Biology, Utrecht University, Utrecht, The Netherlands
[3]University of Cologne, Institute for Plant Sciences, Cluster of Excellence on Plant Sciences (CEPLAS), 50674 Cologne, Germany
[4]Department of Plant Pathology, Kansas State University, 1712 Claflin Road, Manhattan, Kansas 66506, USA

[#,§]These authors contributed equally



Abstract

Three dimensional (3D) folding of DNA in the nucleus organizes chromosomes into so-called topologically associating domains (TADs). These TADs are self-interacting genomic regions that display less interaction with adjacent regions. Functionally, TADs have been implicated in transcriptional regulation as well as in genome evolution in numerous organisms, yet in fungi the functional implication of these regions remains less clear. Here, we utilize chromatin conformation capture (Hi-C) data generated for the plant pathogenic fungus *Verticillium dahliae* to investigate TAD organization and its influence on transcription. Additionally, we compare the TAD organization between two *V. dahliae* strains as well as with other *Verticillium* species to study the conservation of TADs throughout the genus. Remarkably, we find that TADs in the evolutionary dynamic adaptive genomic regions (AGRs) of *V. dahliae* are less well insulated than TADs in the core genome, indicating that TADs in AGRs are not as well established as those in the core genome. Moreover, TADs in AGRs display significantly more co-regulation of gene expression than TADs in the core genome. Furthermore, genes located in TAD boundaries, i.e. regions that delineate adjacent TADs, are generally lower expressed in AGRs *in vitro*, while stronger differentially expressed between *in vitro* conditions, than genes located in TADs in these AGRs. We find that TAD boundaries are depleted for structural variation between *Verticillium* species, and that TADs are generally conserved in the *Verticillium* genus*.* Overall, our study points towards an association between TAD organization and transcriptional regulation as well as genome evolution in *Verticillium.*

Introduction

The spatial organization of the eukaryotic nuclear genome is intimately linked to its biological functions [315,316]. Besides the organization of genetic elements on the linear DNA strands, the association of DNA and nuclear proteins, such as histones, leads to an organized nucleoprotein complex known as chromatin [317]. Chromatin states can be broadly divided in the relatively weakly compacted euchromatin, in which genetic features can be transcriptionally active, and the strongly compacted heterochromatin that is generally inaccessible to the DNA-binding components that act in the transcriptional machinery, and therefore transcriptionally repressive [90,177]. The difference in chromatin organization is mainly mediated by chemical modifications to nucleosomes, the building blocks of chromatin that consist of an octamer of four different histone proteins, wrapped by 146 bp of DNA, with unstructured tails sticking out of the complex [40]. Typically, tri-methylation of lysines 9 and 27 on the tails of histone 3 (H3K9me3 and H3K27me3) are hallmarks of heterochromatin, whereas di-methylation of lysine 4 on the tails of histone 3 (H3K4me2) is a hallmark of euchromatin [177,318]. Besides these histone modifications, many more modifications have been described, yet their role in DNA organization often remains unclear [319].

On a global scale, the genome displays a spatial three-dimensional (3D) structure that brings in close proximity genomic sites that are physically separated on the linear DNA strand, or lie on different chromosomes, and conversely, separates proximal genomic sites through specific folding barriers [315,320]. Such 3D chromosome structure has revealed multiple levels of organization across genomic scales. These levels of organization range from small-scale chromatin loops of a few kilobases that contribute to transcriptional regulation based on DNA-DNA contacts, to large-scale subdomains composed of hundreds of kilobases that arrange large chromosomal regions into active or silent chromatin regions in A and B subdomains, respectively [321–324].

Local three-dimensional chromosome interactions shape chromosome structure into discrete genomic regions commonly known as topologically associating domains (TADs). TADs are physically self-interacting genomic regions that are delineated by TAD boundaries that display less interaction to adjacent genomic regions [47,325]. Although the function of TADs is controversial and still under debate [326–329], several studies have associated the organization of TADs with transcriptional regulation by limiting the interaction of regulatory sequences with their gene targets [321,325,329]. Other studies implicate TADs in genome replication by keeping origins of replication synchronized and active within TADs [330,331]. Studies into various groups of related organisms uncovered evolutionary conservation of TAD organization [321,331–339]. In metazoans, genes and TADs co-localize by their evolutionary age, suggesting a high degree of conservation of TAD organization [340]. Moreover, conserved TADs between different species display similar transcription patterns of genes residing within TADs [321,331,335,341,342]. Furthermore, the genomic regions comprising particular TADs appear to have reshuffled integrally during metazoan evolution [47,331,335,341]. In contrast, unexpected changes in TAD organization potentially lead to variations in gene expression patterns, and therefore to important phenotypic modifications [47,343]. For example, changes in TAD organization are associated with developmental alterations and particular human diseases [344,345]. Recently transient modifications of TAD organization were also associated with quick transcriptional changes upon environmental cues, suggesting an active role of TADs in transcriptional regulation [346–348]. Thus, it seems clear that changes in 3D genome organization are potentially relevant for phenotypic variations and quick transcriptional responses in changing environments.

In the fungal model organisms *Saccharomyces cerevisiae, Schizosaccharomyces pombe*, and *Neurospora crassa*, the 3D chromosomal organization is linked to heterochromatin distribution [105,106,270,349,350]. Structurally, heterochromatic regions in TAD boundaries are associated with permissive cohesin and condensin I binding [270,316,330,349]. Cohesin and condensin I are architectural DNA-binding proteins widely known to cooperate in chromosome folding, and assist in chromosome organization during meiosis and mitosis [351,352]. Additionally, heterochromatin is typically associated with repeat-rich regions [353]. In the endophytic fungus *Epichloë festucae*, repeat-rich regions often colocalize with TAD boundaries and are therefore associated with genome folding [268]. Moreover, genes that are highly expressed *in planta* are enriched near those repeat-regions [268]. In several other fungi it has been observed that heterochromatic regions are associated with the regulation of environmentally responsive genes, such as *in planta* induced genes or genes expressed upon heat-shock [49,195,348,354,355]. Collectively, these findings suggest a structural correlation between the 3D chromosome organization, heterochromatin, and repeat-rich regions, putatively linked with transcriptional regulation in response to environmental challenges in fungal organisms.

The genomes of many plant pathogenic fungi display a two-tier organization in which particular gene-poor and transposable element (TE)-rich genomic regions are evolutionary more dynamic than the relatively stable core genome [69,84,356]. Such dynamic regions often contain environmentally responsive genes, including genes that encode *in planta* secreted proteins, and display increased frequencies of nucleotide substitutions, genomic rearrangements, presence/absence polymorphism and are typically associated with facultative heterochromatin [59,69,77–81,83,84,191,354,357]. Conversely, TE-poor regions are gene-dense, harbor primary metabolic genes, and are often associated with euchromatin [67,84,121,191,354,357]. Collectively, this two-tier organization is typically referred to as a 'two-speed genome' [69,82,84]. TE-rich genomic compartments play important roles in the coevolutionary 'arms-race' between pathogens and their hosts, as the increased frequency of genomic variation can more rapidly delete or diversify genes encoding proteins recognized by plants, or generate genes encoding proteins with novel functions in pathogenicity [356–359]. Presently, it remains unclear what role the three-dimensional genome organization plays in this genomic compartmentalization.

The asexual soil-borne fungal plant pathogen *Verticillium dahliae* is a notorious vascular wilt pathogen that can infect hundreds of plant species [58]. Comparative genomics among *V. dahliae* strains has revealed the presence of extensive large-scale genomic rearrangements associated with discrete TE-rich regions [59,74,122,360]. These rearrangements are associated with the occurrence of extensive segmental duplications that underwent substantial reciprocal gene losses leading to a high degree of presence/absence polymorphism that, collectively, contributed to the formation of the genomic compartments formerly known as lineage-specific regions [59,62,74,122]. Recent work on the chromatin landscape in *V. dahliae* revealed that these regions display unique chromatin characteristics that are shared by additional regions in the genome that were not previously recognized as lineage-specific based on comparative genomics alone [191]. Collectively, these regions are now referred to as adaptive genomic regions (AGRs) [191]. Importantly, AGRs are enriched for genes encoding *in planta*-induced effector proteins, but also for genes that are differentially expressed between *in vitro* growth media, suggesting that these regions contain conditionally responsive genes that contribute to host colonization and environmental adaptation [59,191,354,360]. While a key role of TEs in driving the formation and maintenance of AGRs in *V. dahliae* has been revealed [74,191,360], it presently remains unclear how chromosome folding correlates with the organization and evolution of the core genome and AGRs. Here, we explore the chromatin conformation of *V. dahliae* with DNA proximity ligation followed by sequencing (Hi-C) to uncover the spatial organization of the core genome and the AGRs in detail. In our analysis, we also address other members of the *Verticillium* genus. Our analysis reveals a unique chromatin conformation associated with AGRs, and further unveils ancestral conservation of chromosome organization in the core genome in the *Verticillium* genus.

Results

The *Verticillium dahliae* genome is locally organized in topologically associating domains (TADs)

We sought to determine if the differential association of particular histone marks to the core genome and to AGRs in *V. dahliae* [191] may be correlated with a differential spatial organization of DNA in the nucleus [63,338]. To investigate the chromatin organization in *V. dahliae*, we performed chromatin conformation capture sequencing (Hi-C) in two biological replicates of *V. dahliae* strain JR2 cultivated for 6 days in potato dextrose broth (PDB). As the Hi-C data between the replicates displayed a high correlation (Fig. S1A), we combined these in a single interaction matrix. As expected, we observed a negative correlation between interaction strength and genomic distance (Fig. S1B) that arises from genomic regions interacting strongly with neighboring genomic regions [361]. To further investigate the occurrence of discrete clusters of strong DNA interactions, we predicted TAD organization based on the insulation score method [362]. Building on the notion that a TAD is a self-interacting genomic region with sequences that physically interact more with each other than with sequences outside the TAD, we calculated the insulation score of each bin on the Hi-C interaction matrix (average bin size ~4 kb) by determining the interaction strength with the adjacent bins. Bins that display a low insulation score weakly interact with neighboring bins and consequently were assigned as a TAD boundary region. The bins between two TAD boundaries were therefore assigned to a single TAD. Using this approach, we identified in total 353 TADs (mean size 102,394 bp) separated by 347 TAD boundaries (mean size 4,747 bp) that are distributed along the eight chromosomes of *V. dahliae* strain JR2 (Fig. 1A,B,E, Fig. S1C). Taken together, the high reproducibility between two independent samples and the identification of clear TAD boundaries in the combined interaction matrix, suggests a confident prediction of TADs as units of chromosomal organization in *V. dahliae*.

In various organisms, it has been found that TAD boundary regions are enriched in sequence motifs for DNA-binding insulator proteins, such as the zinc-finger CTCF and the cohesin ringed complexes in vertebrates, or condensin I in yeasts [316]. To investigate the presence of similar protein-binding motifs in the boundaries of *V. dahliae* TADs, we queried the 347 boundary sequences for potential motifs *de novo* using MEME [363]. This analysis revealed two significantly enriched motifs; an 11 bp GAAG motif that is present in 68.6% of the TAD boundaries ($p = 6.5x10^{-10}$; Fig. 1C), and a 24 bp TATA motif in 18.2% of the boundaries ($p=1.2x10^{-66}$; Fig. S1D). To further analyze these two motifs, we queried these in the JASPAR 2018 and YEASTRACT 2018 databases using TomTom [363–365]. The GAAG motif showed a significant match to the Azflp transcription factor of *S. cerevisiae* ($p=8.74x10^{-5}$), a Zinc-finger protein known to regulate the expression of genes under different carbon sources [366–368]. Similarly, we found a significant match for the TATA motif ($p=2.9x10^{-8}$) to an HMG (High Mobility Group) nucleosome remodeler, know to slide and eject nucleosomes to regulate gene transcription [369–371]. Presence of the motifs coincides with a decline in chromatin accessibility, as determined by the assay for transposase-accessible chromatin (ATAC; Fig.1C; Fig. S1D), possibly suggesting DNA-protein interactions at these sites. Even though we find clear enrichment of these motifs in TAD boundaries, the putative function of their protein binding partners does not appear to be that of the DNA-binding insulator proteins CTCF, cohesin and condensin. Although we cannot rule out that these motifs in *V. dahliae* are recognized by proteins that function in TAD separation, it is more likely that they function in other boundary-associated processes.

We reasoned that the previously reported genome compartmentalization of *V. dahliae* into core genome and AGRs could be associated with a differential TAD organization between these two compartments. Therefore, we investigated the TADs and TAD boundaries within the context of both genomic compartments. Globally, of the 353 TADs and 347 TAD boundaries that we identified in the *V. dahliae* genome, 277 TADs (78.47 %) and 308 TAD boundaries (88.76 %) could be assigned to the core genome, and 76 TADs (21.53 %) and 39 TAD boundaries (11.24 %) to an AGR. Interestingly, we observed that AGRs are enriched for relative weakly insulated TAD boundaries (Fig. 1D), indicating that the insulation of TADs in AGRs is 'weaker' than in the core genome. Moreover, we observed a correlation of weak boundaries with smaller TADs in AGRs (Fig. S1F), suggesting a correlation between insulation and length of TADs.

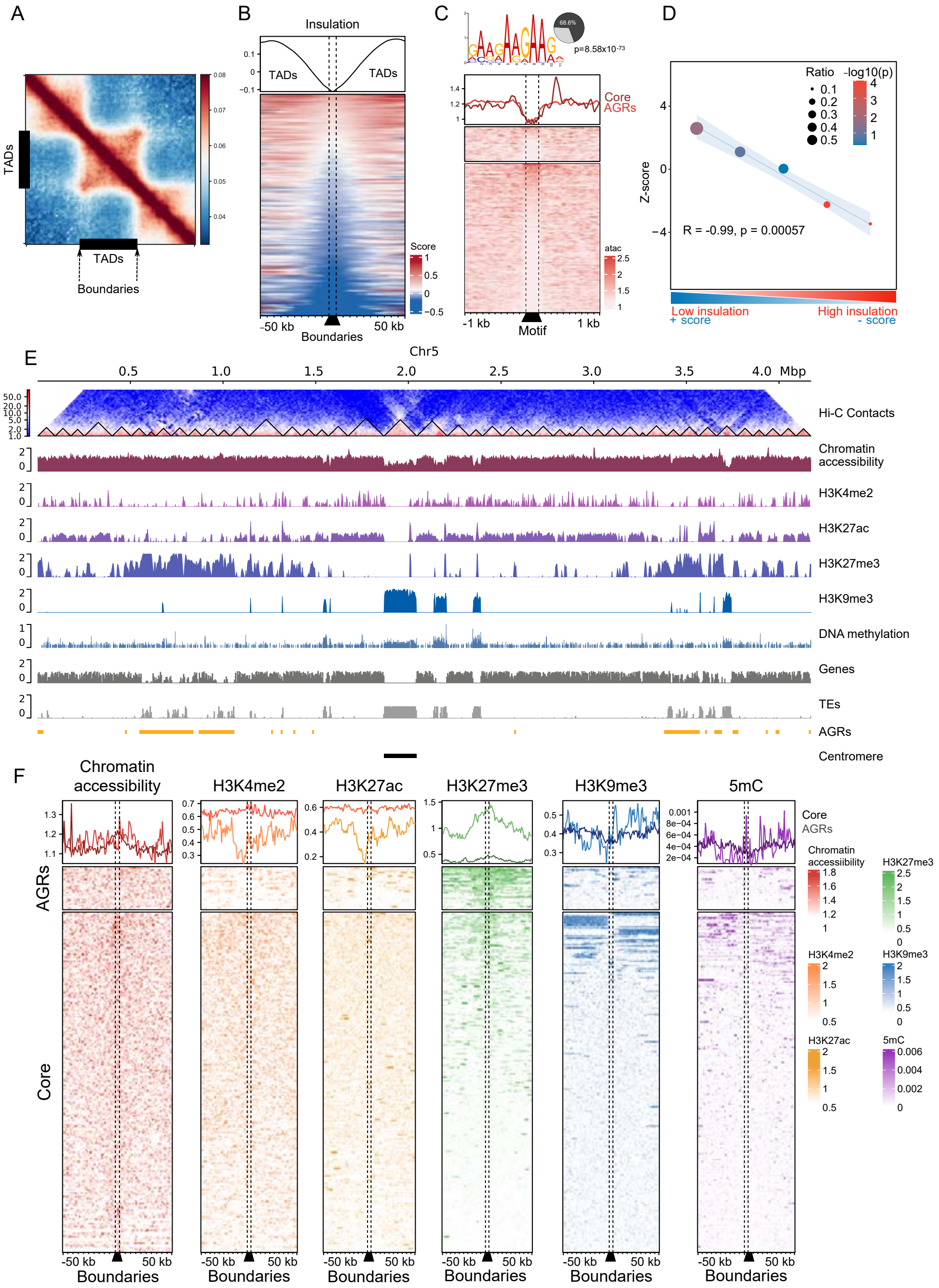


FIGURE 1 | The *Verticillium dahliae* genome is organized in topological associating domains (TADs). (A) Hi-C contact matrix showing local interaction frequency, aggregated over all TADs (black bars) with 50 kb up- and down-stream sequence. The drop in intensity at boundaries at either side of the TADs indicates stronger interaction within TADs than with neighboring genomic regions. (B) Heatmap showing insulation scores centered over boundaries with 50 kb up- and down-stream sequence as rows, ordered on insulation score with weakest insulated boundaries (i.e. having the highest insulation score) on top. The top plot displays the average insulation score as shown in the heatmap below. (C) Chromatin accessibility as determined by the assay for transposase-accessible chromatin (ATAC) over the 11 bp Zinc-finger binding motif that is located in 68.6% of all TAD boundaries. The position weight matrix displays the 11 bp motif enriched in TAD boundaries. Heatmaps at the bottom display ATAC scores for each motif occurrence with 1 kb up- and down-stream sequence in the core genome and in AGRs. The line plot displays the average ATAC score for motif occurrences in the core genome (dark red) and in AGRs (light red). (D) TADs in AGRs are less well insulated when compared with TADs in the core genome. The X-axis indicates quintiles of boundaries, separated based on insulation scores. The Y-axis indicates Z-score and the -log10(p-value) color-scale after a permutation test (10,000 iterations). The plot displays a linear regression (blue line) and confidence interval (light blue) as well as the R and p-value after linear regression. (E) TAD distribution in *V. dahliae* strain JR2, with chromosome 5 as an example. From top to bottom: Hi-C contact matrix depicting TADs as black triangles, open chromatin regions as determined with ATAC-seq, H3K4me2, H3K27ac, H3K27me3, and H3K9me3 normalized over a micrococcal nuclease digestion control, GC methylation, as well as gene and transposable element (TE) densities in 10 kb windows. Adaptive genomic regions (AGRs) [191] and the centromeric region [194] are indicated in yellow and black, respectively. (F) Chromatin characteristics are differentially associated with TAD boundaries in the core genome and in AGRs. On top, distribution of each chromatin feature in (E) centered for boundaries with 50 kb up- and down-stream sequence, for the core genome (dark color) and AGRs (light color). On the bottom the corresponding heatmaps are shown for the core genome and AGRs.

Chromosome organization is typically associated with chromatin characteristics, such as DNA methylation and histone modifications [323]. Therefore, we analyzed the distribution of a set of histone marks, and of DNA methylation, over the TADs and boundaries. In line with previous results [191,194,354,360,372], we observed that the gene-rich core genome is enriched in H3K27ac and H3K4me2, while the centromeres and TE-rich core regions are enriched in H3K9me3 and DNA methylation, and AGRs are enriched in H3K27me3 (Fig. 1E,F, Fig. S3). We observed that such broad chromatin associations are maintained similarly on TADs and boundaries, suggesting that chromatin characteristics associate with the overall separation into core genome and AGRs, rather than with TAD organization (Fig. S3). Interestingly, for the core genome, our analysis revealed that boundary regions show higher chromatin accessibility, and reduced presence of heterochromatin-associated marks (H3K9me3, DNA methylation) than TADs (Fig. 1G, Fig. S3). In contrast, we did not observe major differences in chromatin accessibility at TADs and boundaries in AGRs. However, boundary regions in AGRs are depleted in H3K9me3 and in activation marks (H3K4me2, H3K27ac), but enriched in H3K27me3 and DNA methylation when compared with TADs in AGRs (Fig. 1F, Fig. S3). These differences in chromatin state between TADs and boundaries for both the core genome and AGRs suggests that TADs and TAD boundaries may differ in functionality, not only between each other, but also between the two genomic compartments.

TAD organization impacts transcriptional regulation

Considering the differential enrichment of particular chromatin marks between TADs and boundaries in AGRs, and the enrichment of AGRs in conditionally responsive genes that contribute to host colonization and environmental adaptation [59,191,354], we hypothesized that genes located in the core genome and in AGRs differ in their transcriptional profile between TADs and boundaries. To investigate this, we first determined the occurrence of genes over TADs and boundaries for both genomic compartments. Previous studies uncovered that the core genome contains more genes and fewer TEs compared with AGRs [59,74,360]. Interestingly, within the core genome we observed a significant enrichment of genes, and a corresponding depletion of TEs, in the TAD boundaries when compared with TADs (Fig. 2A). However, within AGRs we do not observe enrichment or depletion of genes and TEs in boundaries (Fig. 2A). The genetic differences between TADs and boundaries in the core genome, and the epigenetic differences between TADs and boundaries in AGRs, suggests that TAD organization may impact transcriptional regulation.

To further study the impact of TAD organization on gene expression, we queried previously generated expression data of *V. dahliae* cultivated for 6 days in PDB [191,354], which is the same cultivation condition as used for our Hi-C data. To this end, we performed a Uniform Manifold Approximation and Projection for Dimensional Reduction (UMAP) on all *V. dahliae* genes, based on DNA methylation, CRI, H3K27ac, H3K27me3, H3K4me2, and H3K9me3. In line with previous observations, genes are mainly separated in clusters representing the core genome and AGRs (Fig. S4) [191]. Additionally, when considering only boundary genes, we observe that the UMAP cluster containing AGR genes is enriched for H3K27me3, while the cluster containing core genes is enriched for H3K27ac. Furthermore, we observe a clear separation based on transcriptional activity, with genes in boundaries of the core genome generally displaying higher transcription levels, associated with increased H3K4me2 levels, than genes in AGR boundaries (Fig. S4). As we reported above that AGRs are enriched for weak boundaries (Fig. 1D), we sought to investigate a potential relationship between gene expression and insulation strength of boundaries. Interestingly, we observed a positive correlation between boundary insulation and expression level (Fig. 2B, Fig. S4), indicating that genes in weakly insulated TAD boundaries are generally lower expressed than genes in strongly insulated boundaries. To investigate if transcriptional activation is dependent on distance to boundaries, we grouped *V. dahliae* genes based on their distance to the closest boundary and quantified their expression. We observed that genes located within TAD boundaries in the core genome and in the AGRs, are lower expressed than those located further in the TADs (Fig. 2C). However, the expression of genes within TAD boundaries in the core genome is notably higher than the expression of genes within boundaries in AGRs, whereas genes further away from boundaries in the core genome and AGRs are similarly expressed (Fig. 2C). Taken together, these findings indicate that genes localized in proximity to weak AGR boundaries are strongly silenced *in vitro*, yet strong boundaries in the core genome are equally transcriptionally active as the adjacent TADs.

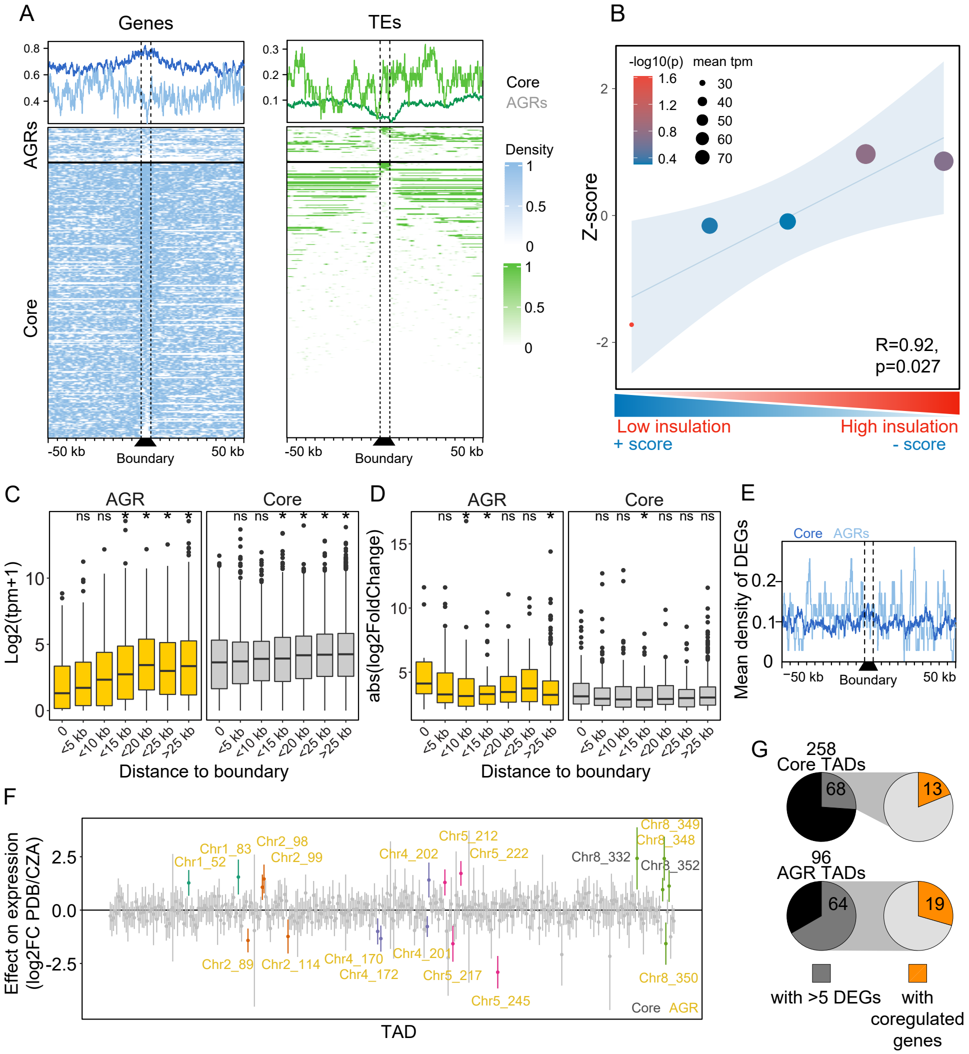


FIGURE 2 | Topological associating domain (TAD) organization affects transcription in *Verticillium dahliae*. (A) Heatmaps visualizing the density of genes and transposable elements (TEs) per 1 kb window centered over boundaries (black trapezoids) with 50 kb up- and down-stream sequence in the core genome and in adaptive genomic regions (AGRs) of *V. dahliae* strain JR2. Line plots display average presence of genes and TEs summarized for boundaries in the core genome (dark color) and in AGRs (light color). (B) Genes in lowly insulated boundaries are lower expressed than those in more highly insulated boundaries. Association between TAD boundary quintiles, separated on insulation score, and transcription of genes located in TAD boundaries. The Y-axis depicts the Z-score, color of the datapoints indicates the -log10(p-value) after a permutation test (10,000 iterations) and size of datapoints indicates mean transcription value (TPM) of represented genes. A linear regression (blue line), with 95% confidence interval (light blue), between boundary quintiles is displayed. (C) Transcription values for *V. dahliae* cultivated for 6 days in potato dextrose broth (PDB) and (D) absolute log2-fold change in expression between cultivation in PDB or in Czapec-Dox medium (CZA), for all genes grouped based on their distance to the closest boundary in the core genome (grey) or in AGRs (yellow). Statistically significant differences in average transcription level for the distance groups was compared to the group of genes located in boundaries (distance 0) and determined by the Wilcoxon Rank-Sum test (* $p < 0.05$). (E) Mean density of differentially expressed genes (log2 >2) between cultivation for 6 days in PDB or in CZA, per 1 kb window over boundaries with 50 kb up- and down-stream sequence in the core genome (dark blue) and in AGRs (light blue). (F) Linear regression effect size of each TAD on differential gene expression between cultivation for 6 days in PDB or in CZA. Mean effect size of each TAD is shown as a point, with 95% confidence interval, and TADs with a significant effect (95% confidence interval is significantly different from 0) are shown in color and labelled by corresponding chromosome and TAD number, for TADs in the core genome (black labels) and in AGRs (yellow labels). (G) Pie charts displaying the proportion of TADs in the core genome and in AGRs containing more than five differentially expressed genes (DEGs) between cultivation for 6 days in PDB or in CZA (dark grey), and for which DEGs display directionality of differential expression (orange). Grey connections between pie charts indicate that the right-side pie charts represent only the 68 TADs in the core genome and the 64 TADs in AGRs with >5 DEGs.

As TADs are thought to function as regulatory units for differential gene expression [329,373], we hypothesized that differentially expressed genes (DEGs) are enriched within TADs, and thus are depleted in boundaries. To test this hypothesis, we first examined the occurrence of DEGs between cultivation in potato dextrose broth (PDB) and in Czapec-Dox medium (CZA) *in vitro.* Previously, we have shown that DEGs are enriched in AGRs as well as in H3K27me3 domains in the core genome [59,191,354], indicating that TADs in both the core genome and in AGRs may function as regulatory units for differential gene expression. We identified a total of 1,844 DEGs (1,005 higher expressed in PDB and 839 higher expressed in CZA), yet we did not observe differences in DEG presence between TADs and boundaries in the core genome nor in AGRs (Fig. 2E). Additionally, when inspecting the absolute log2 fold-change for the expression of genes in relation to their distance to the closest boundary, we observed that only genes at a distance of 5-15 kb and >25 kb from TAD boundaries in AGRs are significantly stronger differentially expressed than genes located in AGR boundaries (Fig. 2D), suggesting that differential gene expression *in vitro* does not depend on the local physical TAD organization, and thus rather is a general feature of genes in AGRs.

To investigate whether genes localizing within the same TAD in *V. dahliae* display transcriptional co-regulation, we fitted a linear model in which differential expression between cultivation in PDB and CZA of each gene is predicted by TAD association. Co-expressed genes in TADs will result in a positive or negative score for that TAD, depending on the prevalent direction of differential gene expression, while opposite direction of differential gene expression within a TAD results in a mean effect of zero. We identified 19 TADs with confidence intervals that are not zero, and therefore have a significant effect on transcription (Fig. 2F). Of these TADs, 17 are associated with AGRs and two with the core genome (Fig. 2F), suggesting that transcriptional co-regulation of expression mainly occurs in AGRs. To corroborate these findings, we checked whether TADs in the core genome and those in AGRs contain genes that display a common transcriptional pattern of higher expression in PDB than in CZA, or vice versa. First, we selected only TADs with more than five DEGs, as we did not consider TADs with fewer DEGs to be co-regulated. In total, 68 out of 258 (26.4%) in the core genome, and 64 out of 96 (6.7%) TADs in AGRs, contain more than five DEGs (Fig. 2G). Of these TADs, 13 out of 68 (19.1%) and 19 out 64 (29.7%) in the core genome and AGRs, respectively, contain more than twice the number of genes that are higher expressed in one of the growth media than in the other one, and thus display co-regulation of differential transcription (Fig. 2G). Taken together, our results suggest that TAD organization affects transcription *in vitro*, and that although some TADs display transcriptional co-regulation of gene expression, this occurs only for a subset of TADs that predominantly locate in AGRs.

TAD boundaries are depleted of genomic variation

TADs are often considered to be conserved between closely related organisms [337,374]. To study whether TADs are conserved between strains of *V. dahliae*, we performed chromatin conformation capture sequencing (Hi-C) in two biological replicates of *V. dahliae* strain VdLs17 that is 98% syntenic to strain JR2 (Fig. 3A) [59,62,122]. Following the same methodology as for strain JR2, we combined the biological replicates (between replicate correlation >0.89; Fig. S5A) in a single interaction matrix and predicted 365 TADs (mean length=98,558 bp) and 357 boundaries (mean length=4,506 bp) (Fig. S5B). Remarkably, the TAD organization in VdLs17 displays similar patterns of insulation scores, gene-enrichment, TE-depletion, and DNA motif enrichment in boundaries as in strain JR2, suggesting that these TAD characteristics are conserved in *V. dahliae* (Fig. S5B,C,D). To investigate if also TAD localizations are conserved between the two strains, we normalized the boundary distribution over syntenic regions and tested their overlap between VdLs17 and JR2 (n=342 and n=330 TADs in syntenic portion, respectively). Interestingly, we observed a significant overlap between the boundaries of the two strains (n=225, p=9.6x10-4, one-way Fisher exact test; Fig. 3A), and an overall overlap in TAD positions (Fig. 3B; Fig. S5E). Originally, AGRs in strain JR2 were defined based on absence of synteny with genomic regions in other *V. dahliae* strains [59,62,74], whereas more recently AGRs have been defined based on their epigenetic profile [191]. As we do not presently have the required complex epigenetic data to determine AGRs in strain VdLs17, we divide the genome of VdLs17 into syntenic and non-syntenic compartments based on the genomic comparison to JR2. Non-syntenic compartments of the VdLs17 genome are enriched for weak TAD boundaries (z-score=2.3858, p=0.00001, permutation test after 10,000 iterations), which is similar to what we observed for AGRs in the JR2 strain.

As genomic rearrangements directly impact genome organization, while we observe that TAD boundaries are conserved between *V. dahliae* strains JR2 and VdLs17, we hypothesized that TAD boundaries may be depleted of such genomic variation. Previously, genomic comparisons between *V. dahliae* strains have revealed extensive genomic rearrangements and structural variations (SVs) [59,63,75,122,360]. To study the association between SVs and TAD boundaries, we used previously generated data for a set of 42 *V. dahliae* strains [191,360] to query the distribution of single nucleotide polymorphisms (SNPs) and presence/absence polymorphisms (PAVs) over TAD boundaries in *V. dahliae* strain JR2. The PAV data was generated by summarizing genomic segments of *V. dahliae* strain JR2 that were absent in the other strains and therefore concerns absence counts relative to the genome of strain JR2 [191]. Interestingly, we observed a depletion of SNPs and of PAVs in TAD boundaries in the core genome (Fig. 3C,D). However, TAD boundaries in AGRs showed more PAVs (Fig. S6A), together with a lower nucleotide diversity (Fig. S6A), indicating that TAD boundaries in the core genome are strongly depleted in genomic variation, while boundaries in AGRs are evolutionary less stable. As boundaries in the core genome and in AGRs differ with respect to genomic variation, we next questioned if presence of genomic variation in boundaries correlated with insulation. However, we observed no significant correlation between insulation on the one hand and either SNPs (R = -0.56, p = 0.16, Fig. S6B) or PAVs (R = -0.76, p = 0.068, Fig. S6B) on the other hand.

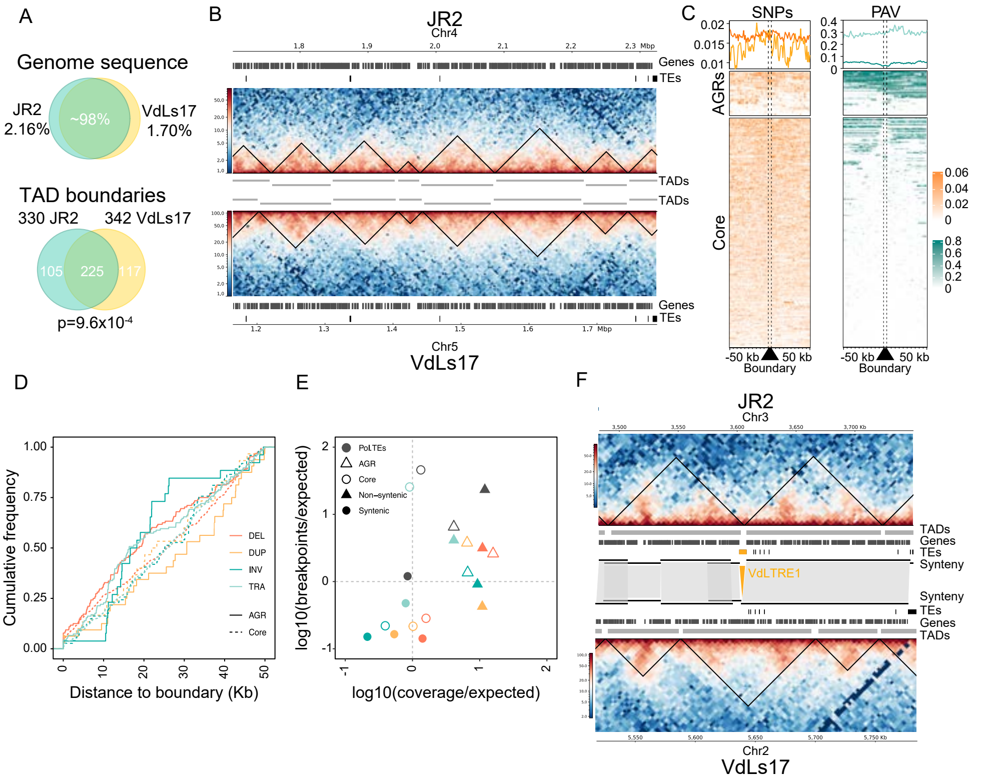


FIGURE 3 | TAD organization is conserved between two *V. dahliae* strains. (A) Top: *V. dahliae* strains JR2 and VdLs17 are highly similar as 97.84% and 98.30% of their respective genomes are syntenic [74]. Bottom: Most of the TAD boundaries overlap significantly between JR2 and VdLs17. (B) Syntenic block between JR2 chromosome and VdLs17 chromosome 5 shows conserved distribution of TADs and boundaries. Heatmaps represent contact matrixes of JR2 (top) and VdLs17 (bottom) with TADs (black triangles). TADs are also displayed as grey bars between heatmaps. Genes and transposable elements (TEs) are displayed above. (C) Boundaries are not enriched for genomic variation in a set of 42 *V. dahliae* strains. Heatmaps displaying single nucleotide polymorphisms (SNPs) [360] and presence/absence variation (PAVs) [191] for 1 kb non-overlapping windows centered around TAD boundaries of *V. dahliae* strain JR2. (D) Cumulative frequency plot of structural variant (SV) breakpoints over distance from boundaries in the core genome (dashed line) and in AGRs (solid line), overlaps with boundaries (distance = 0) are included. SVs are separated in deletions (DEL, orange), duplications (DUP, yellow), inversions (INV, green) and translocations (TRA, blue) [360]. (E) TAD boundaries in AGRs contain more SVs than expected by chance, whereas TAD boundaries in the core genome contain fewer SVs than expected by chance. Color code of SVs is similar as in (D) for SVs in boundaries in the core genome (open circles) and in AGRs (open triangles), as well as for boundaries in syntenic (solid circles) and non-syntenic (solid triangles) genomic regions and for polymorphic TEs (grey circles [360]). (F) Synteny breaks caused by the insertion of transposable elements may give rise to new TAD boundaries in *V. dahliae* strain JR2. Heatmaps represent contact matrixes of JR2 (top) and VdLs17 (bottom) with TADs (black triangles), and TADs, genes and TEs are displayed in between. Synteny between JR2 and VdLs17 indicated as grey blocks. A VdLTRE1 insertion in strain JR2 is indicated in yellow.

To continue our investigations into genomic variation in relation to TAD organization, we next studied different categories of SVs. We observed that deletions, duplications, inversions, and translocations occur more commonly in TADs than in boundaries, indicating depletion of genomic variation from TAD boundaries (Fig. 3D). To assess whether this depletion may be due to purifying selection, we calculated the expected amount of SV breakpoints and SV coverage occurring in boundaries, based on their genome wide occurrences, and compared this to the measured number of breakpoints and their coverage. SV categories displaying fewer breakpoints and lower coverage than expected in boundaries (log ratio <0) are considered to be under purifying selection, as SVs displaying negative selection are expected to display lower frequency of occurrence as well as lower total coverage [333]. As expected, the different SV categories appear to be under purifying selection in boundaries in syntenic regions (Fig. 3E). Conversely, SVs occur more commonly over TAD boundaries in non-syntenic regions and in AGRs, which is in line with previous observations that suggested that SVs in non-syntenic regions are tolerated (Fig. 3E) [360]. Thus, our results show that SVs in boundaries in the syntenic core genome undergo purifying selection, suggesting that genomic stability at TAD boundaries is important for species survival.

Prior studies have indicated that SVs in *V. dahliae* often colocalize with polymorphic TEs that display PAV between strains and are characterized as evolutionary young, scarcely methylated and highly expressed [360]. As TE activity may have been involved in the generation of SVs [74,360], we investigated whether polymorphic TEs occur more frequently in TADs than in boundaries. We identified 36 polymorphic TEs (21.8% of the total) that display PAV between *V. dahliae* strains JR2 and VdLs17, mainly concerning TEs present in *V. dahliae* strain JR2 that are absent from VdLs17. However, we observed no overrepresentation nor depletion of polymorphic TEs in boundary regions (Fig. 3E). Nevertheless, interestingly, some TE insertions in *V. dahliae* strain JR2 occur at a 'new' TAD boundary or at a site of boundary rearrangement (Fig. 3F, Fig. S7), suggesting that polymorphic TEs may drive changes in the TAD organization.

TAD organization is evolutionary conserved in the *Verticillium* genus

Genomic regions that are syntenic between species, often also display conservation of TAD organization [325,331]. As we observed that TADs and boundaries are conserved between strains of *V. dahliae* (Fig. 3A,B), we hypothesized that TAD organization is similarly conserved in other within the *Verticillium* species. Therefore, we aligned the genomes of all ten *Verticillium* species. and retained the genomic regions that aligned to *V. dahliae* strain JR2 (Fig. 4A,B). As expected, relatively distant species such as *V. albo-atrum* shared less syntenic content (78.30%) with *V. dahliae* JR2 than closely related species such as *V. alfalfae* (94.56%; Fig. 4B). The allodiploid species *V. longisporum*, which is composed by three lineages that each arose from a different hybridization event (strain VLB2=A1/D1, PDB589=A1/D3) between two *V. dahliae* strains or between *V. dahliae* and an unknown species [261], shared 50.57% and 44.70% for the strain VLB2 and PD589, respectively (Fig. 4B). To investigate conservation of TAD boundaries between the *Verticillium* species, we used the generated alignments to calculate conservation score [375]. As expected, we observed a higher conservation score in the core genome than in AGRs (Fig. 4C). Interestingly, we observed a clear peak in conservation score around TAD boundaries in the core but not in the AGRs (Fig. 4C), which is similar to our comparison between *V. dahliae* strains JR2 and VdLs17 (Fig. 3C-E), indicating that TAD organization is conserved within the *Verticillium* genus as well.

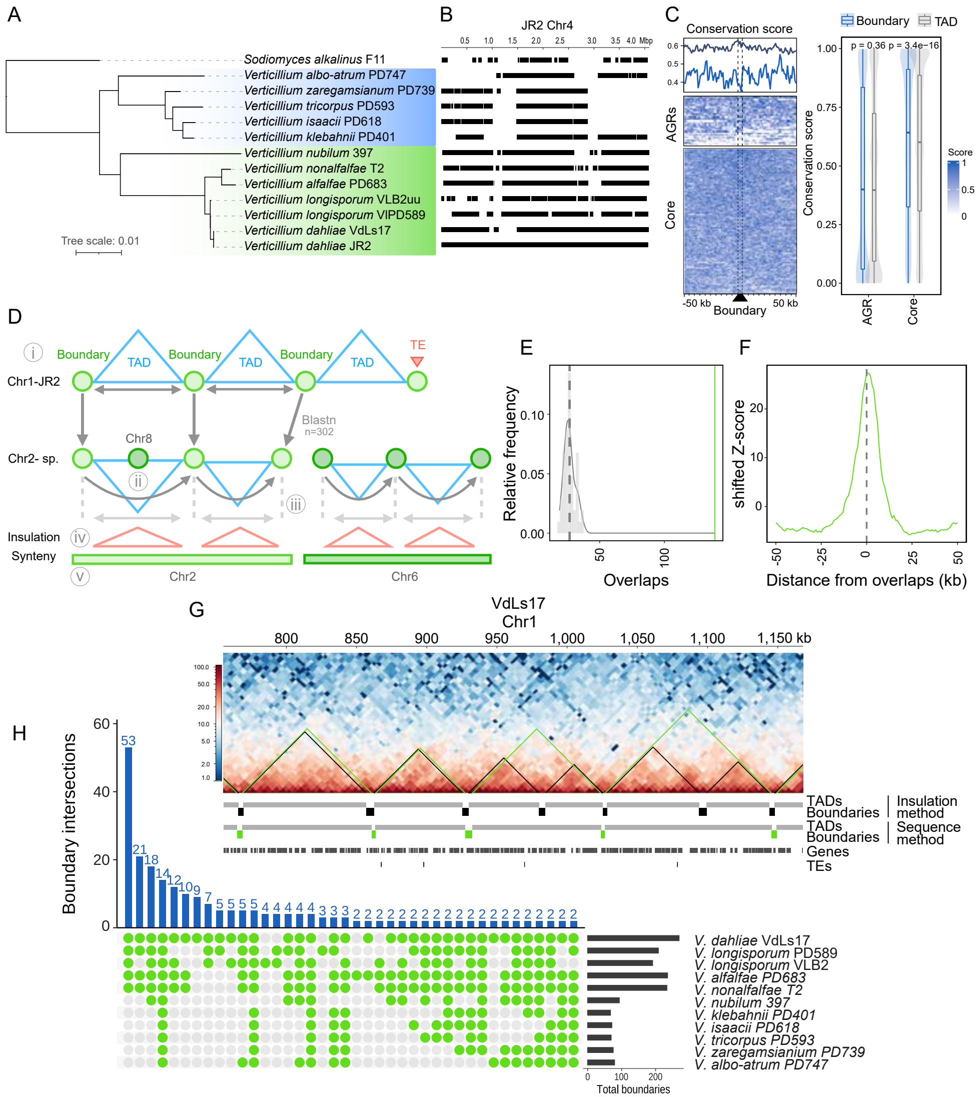


FIGURE 4 | TAD boundaries show signs of conservation in the *Verticillium* genus. (A) Phylogeny of the ten species in the *Verticillium* genus using *Sodiomyces alkalinus* as root; Flavexudans and Flavnonexudans clades are depicted in blue and green, respectively. (B) Syntenic regions (black blocks) between *Verticillium* genomes when compared with *V. dahliae* strain JR2, based on chromosome 4 of *V. dahliae* JR2 strain as example. (C) The heatmap displays conservation score centered over boundaries with 50 kb up- and down-stream sequence in the core genome and in AGRs. The line plot displays the mean conservation score of heatmap for boundaries in the core genome (dark blue) and in AGRs (light blue). Boxplots display the conservation score of each TAD (grey) and TAD boundary (blue) in the core genome and in AGRs. P-values shown represent the result of a one-way Wilcoxon rank sum test. (D) Overview of our approach to predict the TAD organization in other *Verticillium* species with a schematic representation of the *V. dahliae* JR2 reference TAD organization on top, boundaries as green circles and TADs as blue triangles. i) 302 query DNA sequences from boundaries (excluding those with TE insertions) were obtained and compared to the genomes of *Verticillium* spp. ii) Contiguous sequences belonging to same chromosome were retained and single interrupting matches from other chromosomes were removed. iii)

The distances between contiguous boundaries were calculated in reference and subject. iv) Predicted boundaries were compared to the calculated insulation score from the Hi-C data. v) Predicted boundary locations were compared to the progressive Cactus syntenic regions to crosscheck boundary distribution. (E) Distribution of 10,000 iterations of the permutation test for overlaps between the boundaries as predicted for *V. dahliae* strain VdLs17 by the sequence-based method in (D) and by the insulation method (grey distribution), green line indicates significant overlaps between predicted and 'reference' boundaries from *V. dahliae* VdLs17 (Fig. 3), Z-score=27.1264, p=9.9x10-5. (F) Z-score shifts from boundaries indicate a high enrichment of overlapping TAD boundaries for the two methods. (G) Overlap in TAD prediction for strain VdLs17 using the insulation method (black triangles) and the sequence-based method in (D) (green triangles), with partial chromosome 1 of the strain VdLs17 (755,315-1167941 bp) as example (H) TAD boundaries predicted in the ten *Verticillium* species ordered according to sequence-based method in (D), on the bottom-right the total number of boundaries predicted in each species. On the bottom, a combination matrix depicting combinations of species (in green) in which boundaries are shared. On the top, the total amount of boundaries shared f each species combination.

To further assess the conservation of TAD organization in the *Verticillium* genus, we used previously generated Hi-C data [194,261]. Using the same approach as we used for *V. dahliae* strains JR2 and VdLs17, we performed matrix normalization, correction, and insulation score calculation for each species. As TAD boundaries of *V. dahliae* strain JR2 display increased sequence conservation when compared with TADs, we hypothesize that conserved TADs can be recovered as lowly insulated regions in the Hi-C data of other *Verticillium* species. As we did not generate replicates of Hi-C data for these species, we were unable to identify their TAD organization as reliably as for *V. dahliae* strains JR2 and VdLs17 using the insulation method. Therefore, we conceived a sequence-based method to predict boundaries in the other *Verticillium* species. To this end, we compared the sequence of all TAD boundaries in *V. dahliae* strain JR2 to the genome sequences of the other *Verticillium* species. We first removed the boundaries containing TE insertions, as TEs often occur abundantly in the genome, which may lead to improper identification of TAD boundaries. Additionally, as we do not expect TAD sizes to differ fundamentally among *Verticillium* spp., we removed putative boundaries that were separated from contiguous boundaries by a distance shorter than the smallest TAD in *V. dahliae* strain JR2 (Fig. 4D). Finally, we used the Hi-C data of each species to assess whether the insulation score of the predicted boundaries is lower than for adjacent genomic regions (Fig. 4D). To verify the validity of our sequence-based method for boundary identification, we first used it on the genome of *V. dahliae* VdLs17, of which the boundaries were also annotated using the insulation method (Fig. 3). Using this method, we recovered 269 boundaries in VdLs17 (Fig. 4E-G) that display significant positional overlap with the boundaries as determined by the insulation method (z-score=27.1264, p=9.99x10-5; Fig. 4E,G, Table S1). .

As our sequence-based method for boundary identification proved reliable for *V. dahliae* strain VdLs17, we next used this method to predict boundaries in the other *Verticillium* spp. As expected, we observed that boundaries of *V. dahliae* strain JR2 are more likely to be shared with phylogenetically close species (Fig. 4H). For instance, only 80 boundaries were recovered in the more distantly related *V. albo-atrum*, whereas 254 boundaries were predicted in *V. alfalfae* and 283 boundaries in *V. longisporum* strain PD589, that are both relatively closely related to *V. dahliae* (Fig. 4H). In general, the predicted boundaries in the *Verticillium* genus depict a drop in insulation score with the adjacent genomic regions (Fig. S8), indicating that we confidently assign TAD boundaries. Collectively, our results suggest that boundary sequences display a high degree of conservation between species, especially those that are phylogenetically close, and that not only the sequence, but also the function of the boundaries is conserved.

Discussion

In eukaryotes, the DNA within the nucleus typically is subject to three-dimensional (3D) folding, resulting in the organization of chromosomes into so-called topologically associating domains (TADs). The existence of TADs has been demonstrated in numerous eukaryotes [325,341,376,377], including several fungi [105,268,270,330]. Here, we used chromatin conformation capture and high-throughput sequencing (Hi-C) to identify TADs in the genome of the fungal plant pathogen *Verticillium dahliae*, and thereby study local 3D interactions in the nucleus. We furthermore aimed to assess the impact of TADs on transcription as well as on genome evolution. We identified genomic regions with an average size of 100 kb that display stronger interactions within these regions than with neighboring genomic regions, reminiscent of TADs. The size of the regions is smaller than typical mammalian TADs (200 kb - 2.5 Mb) [325], yet similar to TADs in *Drosophila*, in the filamentous fungi *N. crassa* and *E. festucae*, and the yeast *S. cerevisiae* (100 - 150 kb) [87,105,268,330,341]. Our TAD prediction was performed for two strains of *V. dahliae*, and for each of the strains we used independent biological replicates. Importantly, among strains as well as among replicates we obtained a high degree of reproducibility, suggesting that we have confidently predicted TADs as units of chromosomal organization in *V. dahliae*.

As TADs are genomic regions that interact more strongly within the region than with neighboring regions, they separate genomes into discrete units [378,379]. In this manner, TADs function in physically separating clusters of replicating DNA [380]. Replication generally occurs with coordinated timing within genomic domains, called replication domains, and in several animal species, as well as in *S. cerevisiae*, these domains largely correlate with TAD structure [330,380,381]. Thus, TADs can be seen as the unit of replication. In addition to TADs functioning in DNA replication, for some TADs it has been shown that transcriptional co-regulation of genes within the TADs occurs [268,373,382–387]. However, even for TADs that display transcriptional co-regulation it currently remains unclear whether TAD organization facilitates such transcriptional co-regulation, or whether shared transcriptional profiles and the epigenetic status of the genes within a particular genomic region predispose a region to become organized into a TAD [329,388]. Thus, for such TADs it can presently only be concluded that a correlation between TAD organization and gene expression exists. That only few TADs display correlation with transcriptional co-regulation has also been observed in the arbuscular mycorrhizal fungus *Rhizophagus irregularis* and in *E. festucae* [268,387]. Similarly, in *V. dahliae* we find that transcriptional co-regulation occurs in only 19 out of 353 TADs. It is not known what makes that these particular TADs display transcriptional co-regulation, while most TADs do not show signs of transcription co-regulation. Transcriptional co-regulation within TADs could be determined by epigenetic characteristics of the TAD. Remarkably, we find that 17 of the 19 TADs displaying transcriptional co-regulation localize in AGRs. Previously we have shown that AGRs are epigenetically distinct from the core genome [191]. This difference between AGRs and the core genome involves a lack of DNA methylation over TEs, as well as a general enrichment of the histone modification H3K27me3, combined with accessible DNA, in AGRs [191]. The enrichment of H3K27me3 at AGRs plays a role in transcriptional regulation, as loss of H3K27me3 in a *V. dahliae* mutant strain induces the expression of particular genes [354]. Additionally, we showed that changes in H3K27me3 levels between *in vitro* growth media partially explains differential gene expression [354]. An association between H3K27me3 and transcriptional regulation has similarly been observed in various plant associated fungi [49,52–54,185,195]. Therefore, it is conceivable that the presence of H3K27me3 over TADs in *V. dahliae* is associated with transcriptional co-regulation of the genes within these TADs. Nevertheless, we observed that also within the AGRs only a fraction of TADs display transcriptional co-regulation, suggesting that the presence of H3K27me3 alone is not sufficient to regulate transcription, and consequently other factors involved in regulating transcription are required.

TAD boundaries are the regions that mediate the separation of TADs [325,389]. Such boundary regions have frequently been shown to display a relatively high degree of conservation between closely related organisms [321,333,335,337], yet their conservation has not been addressed for fungal species. Here, we show that TAD boundaries display a high degree of conservation in the *Verticillium* genus. The conservation of TAD boundaries suggests that TAD boundaries are important for evolutionary stability, or that boundaries serve additional roles, for instance in regulation of the global 3D genome organization. Interestingly, whereas we find that TAD boundaries in general are conserved in *Verticillium*, TAD boundaries located in AGRs are not particularly conserved between *V. dahliae* strains. This could mean that TAD organization in the evolutionary young AGRs has not yet been firmly established, or that a clear TAD organization is important in the core genome yet can be more relaxed in the evolutionary dynamic AGRs. However, these possibilities are not mutually exclusive. Remarkably, we found clear signs of TE insertions coinciding with putatively newly generated TAD boundaries as well as with extensively rearranged TAD boundaries. Interestingly, a recent study in cotton uncovered that cultivar-specific TAD boundaries generally harbor more TEs than conserved TAD boundaries that are shared between cultivars [390]. Moreover, it has been shown that *de novo* insertions of HERV-H TEs in humans can introduce new TAD boundaries [391]. As TE activity in *V. dahliae* is largely confined to AGRs [74,360], we speculate that such TE activity may be involved in generating new TAD boundaries. In addition to the low conservation of TAD boundaries in AGRs, we also observe that these boundaries are generally weaker insulated than those in the core genome, suggesting that TADs in AGRs are not as well established as those in the core genome. However, single-cell research in human and *Drosophila* cells show that TADs are dynamic structures that are continuously broken and formed, and that boundary positions can vary between cells, especially in a heterogeneous population containing distinct cell types [378,392–394]. Therefore, it cannot be excluded presently that the weaker insulation of TAD boundaries in AGRs than in the core genome is a consequence of TAD boundaries in AGRs being more dynamic among the cells used as input for our Hi-C experiments. In metazoans, TAD boundaries are bound by CTCF proteins, which are thought to be involved in maintaining TAD organization [395], and the amount of CTCF proteins bound to the TAD boundary correlates with the insulation score [396,397]. Therefore, the difference in TAD boundary dynamics in AGRs and in the core genome may be caused by differential binding of proteins with CTCF-like function. However, such proteins have not been identified in filamentous fungi so far, including in *V. dahliae*. Additionally, the functional implications of the here observed differential dynamics of TAD boundaries in AGRs and the core genome remain unclear. Potentially, the TAD boundary dynamics observed in AGRs confers a more flexible transcriptional response of the environmental responsive genes that predominantly locate in AGRs [59,191,354],

Besides local interactions in the context of TADs, the 3D genome is known to also display long-distance interactions, in which distant regions on the same chromosome, or regions on distinct chromosomes, interact [398]. For example, in some eukaryotes long-distance interactions occur between centromeres that, despite localizing on different chromosomes, collectively co-localize within the nucleus [245,269,274,316], a pattern that we previously also observed in *V. dahliae* [194]. Additionally, long-distance interactions among chromosomes of *N. crassa* occur between similar chromatin regions, for instance those that are marked with H3K27me3 [105,106]. Therefore, we hypothesize that the H3K27me3-marked AGRs in *V. dahliae* may similarly display long-distance interactions. Such inter-AGR interactions could potentially explain why chromosome rearrangements preferentially occur within AGRs [59,74]. Accordingly, physical clustering of AGRs in the nucleus may lead to the formation of nuclear bodies [399]. Nuclear bodies are membrane-less sub-compartments in the nucleus, in which a micro-environment exists that allows spatial segregation of nuclear activities, such as transcription and DNA-repair [399–401]. In this study, we have not yet addressed the nuclear clustering of AGRs in inter-chromosomal interactions due to the technical challenges that are associated with such analysis. As AGRs originate from large-scale duplications they are partially syntenic within the genome [74], which leads to mapping problems during Hi-C read analysis. Thus, an analysis pipeline needs to be established that rules out false-positive interactions between syntenic regions in AGRs.

AGRs were originally discovered by their high degree of PAV between *V. dahliae* strains [59,74], and subsequently further characterized and refined by their distinct epigenomic profile [191]. That plastic genome compartments are epigenetically distinct from core genomic regions is increasingly recognized for various plant pathogens [121,183,191,356,402–404]. Here, we have added yet another layer to the divergence between plastic regions and the core genome, by showing that also the local 3D organization differs between the core genome and AGRs in *V. dahliae*. The exact ramifications of this divergent local 3D organization on evolution and transcriptional regulation are still unclear. Future research into global 3D genome organization within the nucleus should show whether AGRs interact and, if so, how this differs from other global interactions. This holistic view with genetic, epigenetic, and spatial characterization of plastic genome compartments will aid further understanding of the genome function and evolution in fungi.

Materials and methods

Hi-C analysis and TAD prediction

Hi-C library preparation was performed on *V. dahliae* strains JR2 and VdLs17 as previously described [194], and paired-end (2 × 150 bp) sequenced on the NextSeq500 platform at USEQ (Utrecht, the Netherlands). Additionally, further Hi-C datasets of *V. dahliae* strains JR2 and VdLs17, *V. albo-atrum* strain PD747, *V. alfalfa* strain PD683, *V. isaacii* strain PD618, *V. klebahnii* strain PD401, *V. longisporum* strains PD589 and VLB2, *V. nonalfalfae* strain T2, *V. nubilum* strain 397, *V. tricorpus* strain PD593, and *V. zaregamsianum* strain PD739, were previously generated [194,261].

Sequenced read-pairs were quality-filtered and trimmed using Trimmomatic (v 0.36) in paired end mode with default settings [295]. Filtered and trimmed reads were mapped to the corresponding genomes [122,194] using Burrows-Wheeler aligner (BWA mem, settings: -A1 -B4 -E50 -L0) [405]. Hi-C interaction matrices were built and analyzed using HiCExplorer tools [406]. First, we used hicBuildMatrix to generate the interaction matrix based on the *in silico* *Dpn*I restriction digested corresponding genome. Matrix resolution was reduced by merging 5 adjacent bins using hicMergeMatrixBins. For *V. dahliae* strains JR2 and VdLs17, replicates were corrected separately according to the iterative correction and eigenvector decomposition (ICE) method [407] using hicCorrectMatrix, and TADs were predicted using hicFindTADs (settings: --delta 0.01). Correlation between replicates was determined by using a reproducibility score based on a stratified cross-correlation using the HiCRep package [408].

To combine replicate matrices, resolutions of raw matrices were reduced by merging 5 adjacent bins using hicMergeMatrixBins, normalized between replicates using hicNormalize (settings: --setToZeroThreshold 1), corrected separately according to the ICE method using hicCorrectMatrix, and finally combined using hicSumMatrices [406]. For the other *Verticillium* species, matrix resolution reduction and correction was performed as above, and hicFindTADs was used to generate a table with per bin insulation scores.

Characterization of epigenetic profiles

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) for H3K4me2, H3K9me3, H3K27me3, and H3K27ac, and the assay for transposase-accessible chromatin followed by sequencing (ATAC-seq) were performed for *V. dahliae* strain JR2 as described previously [191,194,354]. ChIP datasets were normalized over MNAse control samples.

The UMAP algorithm for dimensional reduction [115] was implemented in Python. We used the umap-learn implementation through the R/umap package. For the gene analysis, the following variables were used: GC content, ATAC-seq, 5mC, H3K27ac, H3K27me3, H3K9me3 and log2(PDB *in vitro* expression +1), with the following parameters random_state=42, n_neighbors=50, n_components=2, min_dist=0.01, metric=cosine. The resulting two-dimensional values from UMAP fit.transform were used for plotting and further statistical analysis using Matplolib, Numpy and Seaborn V0.8.1 [158,159,162].

Characterization of transcriptional regulation

RNA sequencing of *V. dahliae* strain JR2 cultivated for six days in potato dextrose broth (PDB) and Czapec-Dox medium (CZA) was previously performed [191,354]. Analyses of gene [122] and TE presence [194,360] over TADs and TAD boundaries were performed using the EnrichedHeatmap package in R [127,409]. To assess co-regulation of genes within TADs, we used R to fit a linear model with log2 fold-change in expression of target genes between PDB and CZA as the response variable and TAD membership as a predictor, similarly as previously described [268].

Characterization of *Verticillium dahliae* genomic variation

Structural variants (duplications, deletions, inversions and translocations), single nucleotide variants and polymorphic transposable elements were previously identified using paired-end sequencing reads of each 42 previously sequenced *V. dahliae* strains [360]. Briefly, structural variants were predicted using the 'sv-callers' workflow with few modifications that enabled parallel execution of multiple SV callers [410], an approach that is considered optimal as it exploits complementary information to predict SVs [411,412]. Single nucleotide variants were identified using the -HaplotypeCaller of the Genome Analysis Toolkit (GATK) v.4.0 [413] and transposable element PAV was analyzed using TEPID v.2.0 [414]. To investigate if SVs and polymorphic TEs co-localize with TAD boundaries, we summarized the overlap of each set of variants by their breakpoint frequency (start or ends ±1 bp of the feature) and coverage (number of bases covered) across the genome of *V. dahliae* strain JR2 [333]. Similar to Fudenberg and Pollard (2019) [333], we calculated the log10(observed/expected) of each feature representing the deviation from a uniform distribution across the genome, therefore accounting for the proportion of the genome covered by a specific genomic feature. Finally, we considered two scenarios: core genome vs AGRs, and syntenic regions between JR2 and VdLs17 versus non-syntenic regions. Syntenic regions between *V. dahliae* strains JR2 and VdLs17 were previously determined [74]. Briefly, whole-genome alignments between the eight chromosomes was performed using MUMmer 3.0 and GEvo [312,415], where only gene-coding regions were used as anchors between syntenic chromosomal regions.

To further expand our analysis of *V. dahliae* to the full genus, we used the recently available Hi-C-corrected genomes of all *Verticillium* species [194,261]. The phylogenetic tree was generated using Realphy v. 1.12 using a maximum likelihood inferred by RAxML [296,416]. We aligned the *Verticillium* genomes using ProgressiveCactus [417]. This approach allowed us to reduce the reference-bias and consider more accurate further analysis. We obtained the specific MAF alignments on JR2 and syntenic regions using the HAL package [418]. To analyze the nucleotide conservation throughout the genus, we used PhastCons, a hidden Markov model-based method that estimates the probability that each nucleotide belongs to a conserved element based on a multiple sequence alignment [375]. Briefly, for each independent JR2 chromosome we assumed a neutral evolution model and correction for indels. For further analysis, we summarized the PhastCons score over TADs and TAD boundaries in the core genome and in AGRs.

TAD boundary prediction throughout the *Verticillium* genus

The Hi-C datasets of the *Verticillium* species (excluding *V. dahliae*), were available with one biological replicate. Therefore, we decided to predict TAD boundaries based on sequence homology to boundaries in *V. dahliae* strain JR2. We first filtered the boundary sequences that do not have a TE insertion and queried them to the *Verticillium* genomes using Blastn, retaining those with >50% coverage that were contiguous in the same syntenic block. Finally, we cross-referenced those putative TAD boundary regions with the previously calculated insulation score for each independent species.

Statistical analysis and visualization

Hi-C matrix and TAD visualizations were performed using HiCExplorer and FAN-C [419]. Heatmap and enrichment visualization of insulation scores over boundaries, normalized chromatin marks, structural and nucleotide variants, as well as the PhastCons score, were performed using the R/EnrichedHeatmap v1.2 package [409]. Permutation tests were computed using R/Bioconductor regioneR v1.18.1 package [420] and performed with 10,000 iterations, using overlaps between TAD boundaries divided by the insulation score quantiles and the predefined AGRs, and circular randomization to maintain the order and distance of the regions in the chromosomes. All statistical analyses and comparison tests were performed in R v.3.6.3 [127], and visualized with ggplot2 [421].

Supplementary data

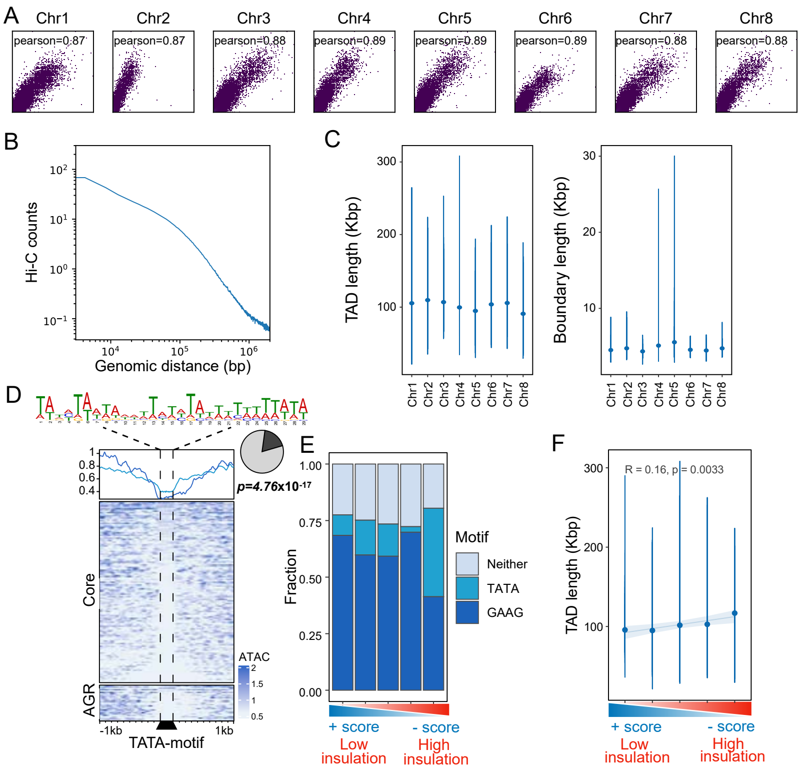


FIGURE S1 | TAD and boundary characteristics of *Verticillium dahliae* strain JR2. (A) Pearson correlation between replicates of Hi-C interactions for *V. dahliae* strain JR2 divided per chromosome. (B) Distance-dependent interaction frequency after replicates of Hi-C data for *V. dahliae* strain JR2 were merged. (C) Length of TADs and TAD boundaries predicted on the merged replicates, divided per chromosome. (D) A TATA-motif is enriched in TAD boundaries of *V. dahliae* strain JR2, correlating with a decrease in chromatin accessibility as determined by the assay for transposase accessible chromatin (ATAC). The top plot displays the average ATAC signal over TATA motifs with 1 kb up- and down-stream sequence in the core genome (light blue) and in AGRs (dark blue). Middle and bottom plots display detected TATA motifs as rows in the core genome and in AGRs, respectively. (E) Occurrence of the GAAG-motif (Fig. 1C) and the TATA-motif (Fig S1D) in boundaries separated on their insulation score. (F) Correlation between boundary insulation score and length of adjacent TADs for boundary quintiles separated based on their insulation score.

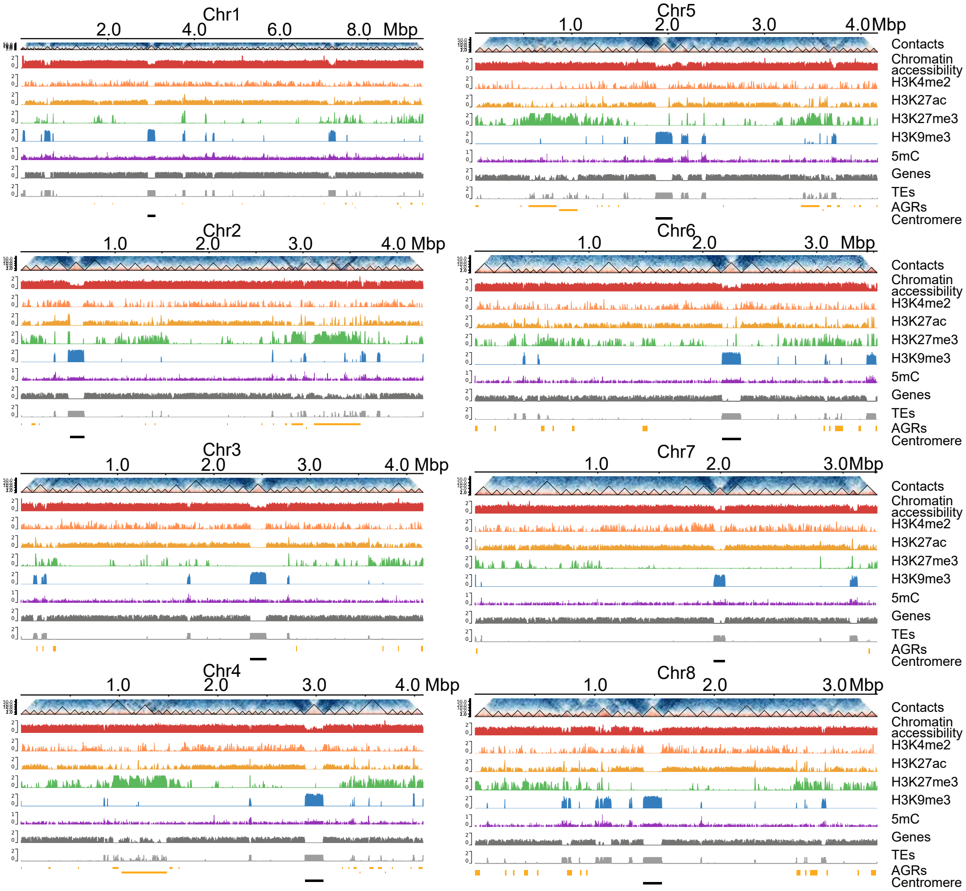


FIGURE S2 | TAD distribution in *Verticillium dahliae* strain JR2 for all chromosomes. From top to bottom: Hi-C contact matrix depicting TADs as black triangles, open chromatin regions as determined with ATAC-seq, H3K4me2, H3K27ac, H3K27me3, and H3K9me3 normalized over a micrococcal nuclease digestion control, GC methylation, gene and transposable element (TE) densities for 10 kb non-sliding windows, adaptive genomic regions (AGRs) [191] and centromeric region [194] are indicated in yellow and black, respectively. Chromosome sizes are displayed above the matrix.

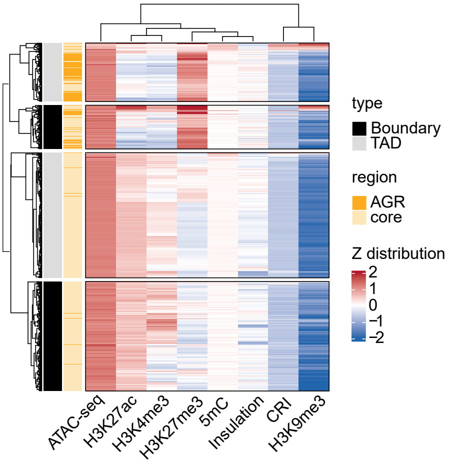


FIGURE S3 | Chromatin characteristics are globally differentially associated with TADs and boundaries. k-means clustering (*k*=2) of TADs and TAD boundaries using data presented in Fig. 1E combined with insulation scores and composite repeat-induced-point mutation (RIP) index (CRI) values averaged for 100 bp non-overlapping windows and summarized per TAD and boundary region.

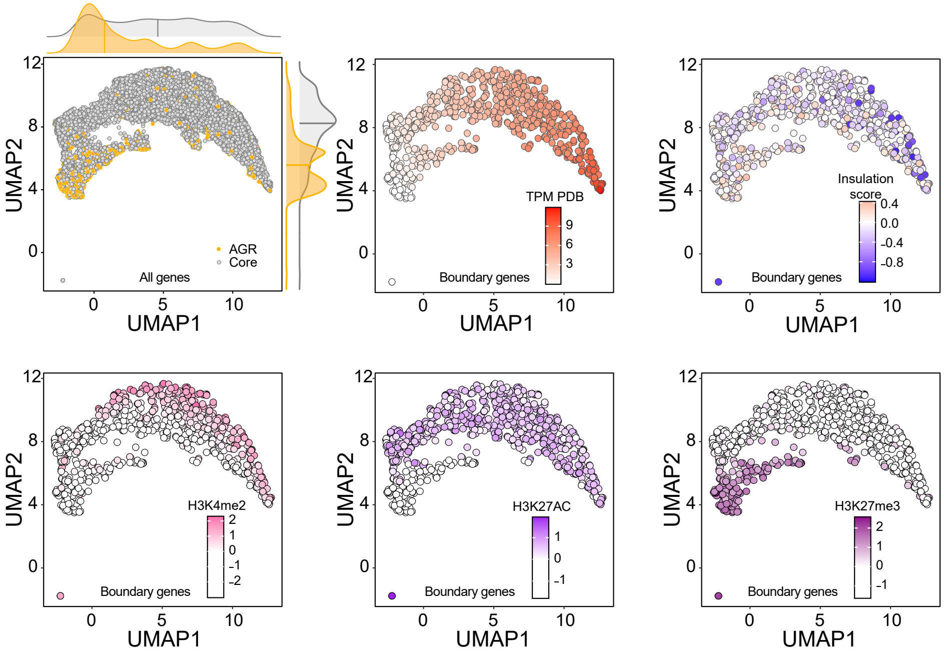


FIGURE S4 | Uniform Manifold Approximation and Projection for Dimensional Reduction (UMAP) to separate genes of *Verticillium dahliae* strain JR2 according to their epigenetic profile. The left top plot displays all genes in the core genome (grey) and in AGRs (yellow). Distributions are shown on top and right of plot. Remaining plots display only boundary located genes and are colored according to their transcription in PDB (TPM PDB), the insulation score of the boundary they locate in, their H3K4me2 ChIP coverage signal, their H3K4ac ChIP coverage signal, and their H3K27me3 ChIP coverage signal.

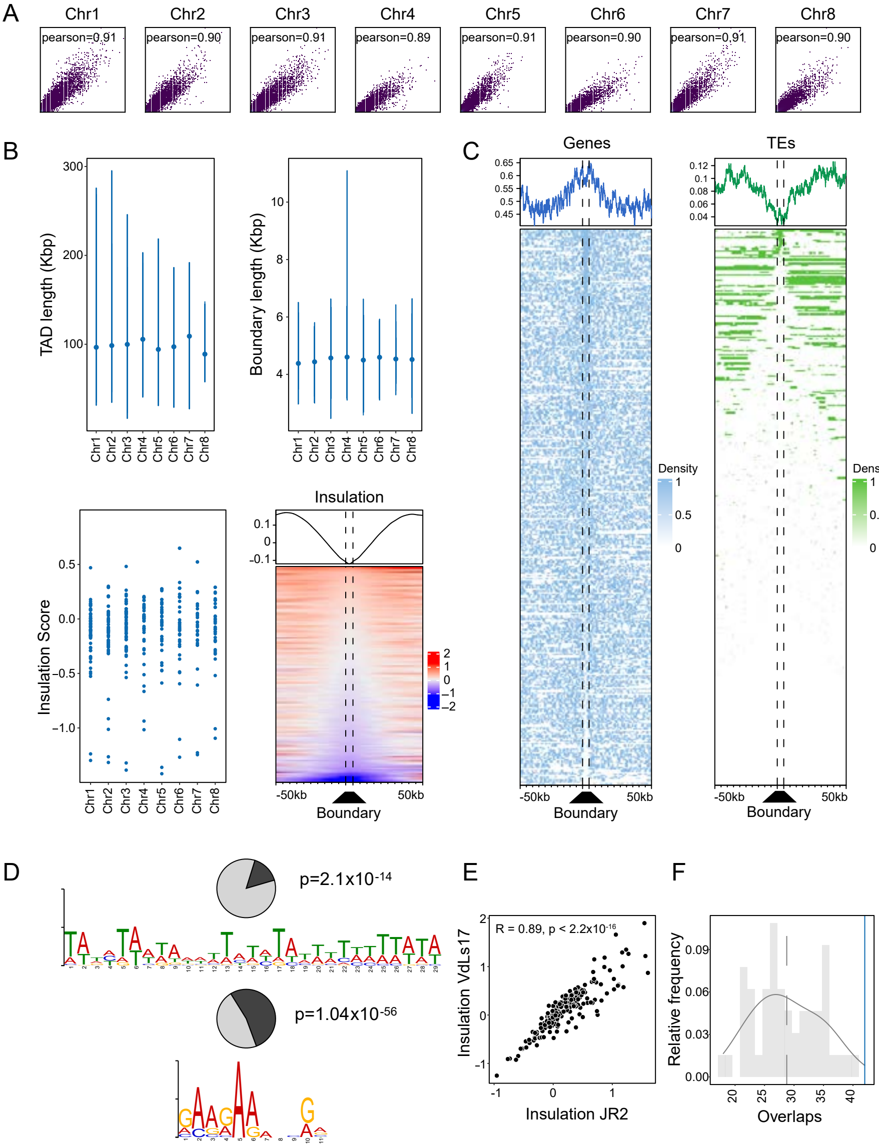


FIGURE S5 | TAD and boundary characteristics of *Verticillium dahliae* strain VdLs17. (A) Pearson correlation between replicates of Hi-C interactions for *V. dahliae* strain VdLs17, divided per chromosome. (B) Top plots: length of TADs and boundaries predicted on the merged replicates, divided per chromosome. Left bottom plot: insulation score of boundaries predicted in *V. dahliae* strain VdLs17. Right bottom plot: insulation score over boundaries with 50kb up- and down-stream sequence. (C) Enrichment of genes and depletion of transposable elements (TEs) in and near boundaries in *V. dahliae* strain VdLs17. (D) Presence of the TATA- and GAAG-motifs in boundaries of *V. dahliae* strain VdLs17. (E) Correlation of insulation score for boundaries predicted in genomic regions that are syntenic between strains JR2 and VdLs17. (F) Lowly insulated boundaries of *V. dahliae* strain VdLs17 are enriched in genomic regions that are non-syntenic between JR2 and VdLs17.

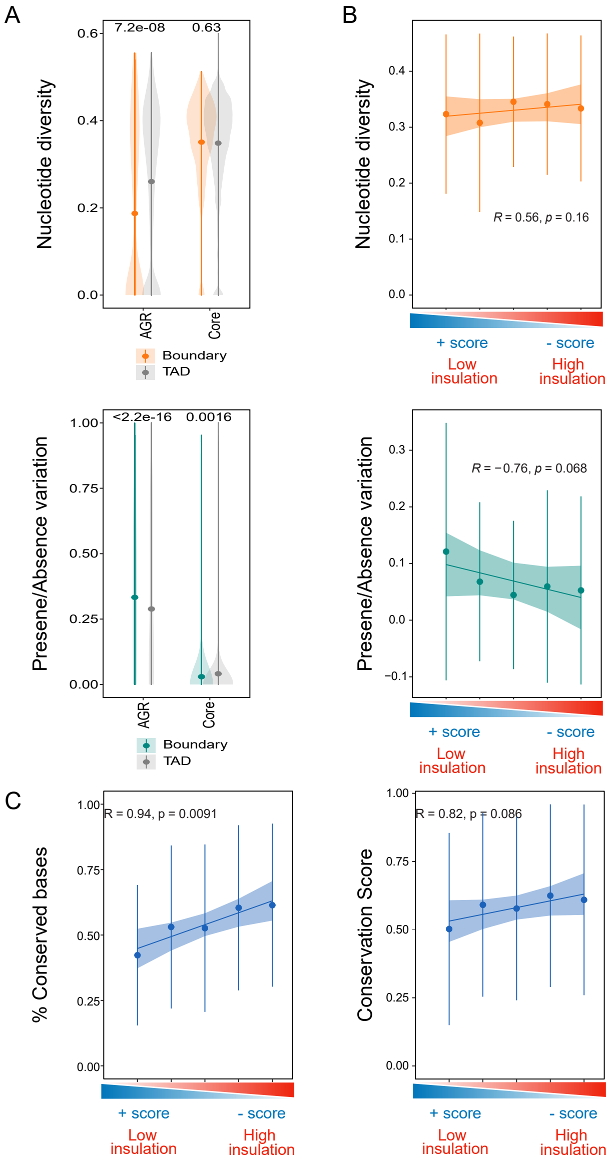


FIGURE S6 | Genetic conservation of TADs and TAD boundaries in a population of 42 *Verticillium dahliae* strains. (A) Nucleotide diversity (top) and presence/absence variation (bottom) for TADs (grey) and boundaries (yellow or green) in AGRs and in the core genome. (B) Correlation between insulation score and nucleotide diversity (top) and presence/absence variation (bottom) for boundary quintiles separated on their insulation strength. (C) Correlation between insulation score and fraction of conserved bases (left) and conservation score (right) for boundary quintiles separated on their insulation strength.

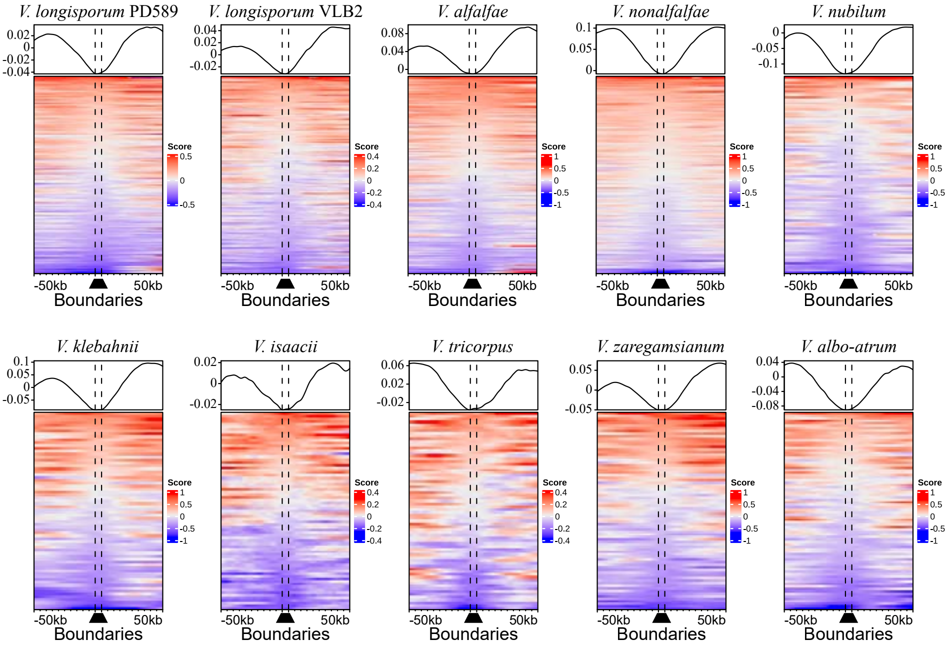


FIGURE S7 | Predicted TAD boundaries in the *Verticillium* genus display low insulation compared to adjacent genomic regions. Insulation score over TAD boundaries predicted by the sequence-based method (Fig. 4D), with 50 kb up- and down-stream sequence, for each *Verticillium* species. Line plots display average signal over boundaries and up/down-stream sequence, bottom plots display predicted boundaries in rows, ordered on insulation score.

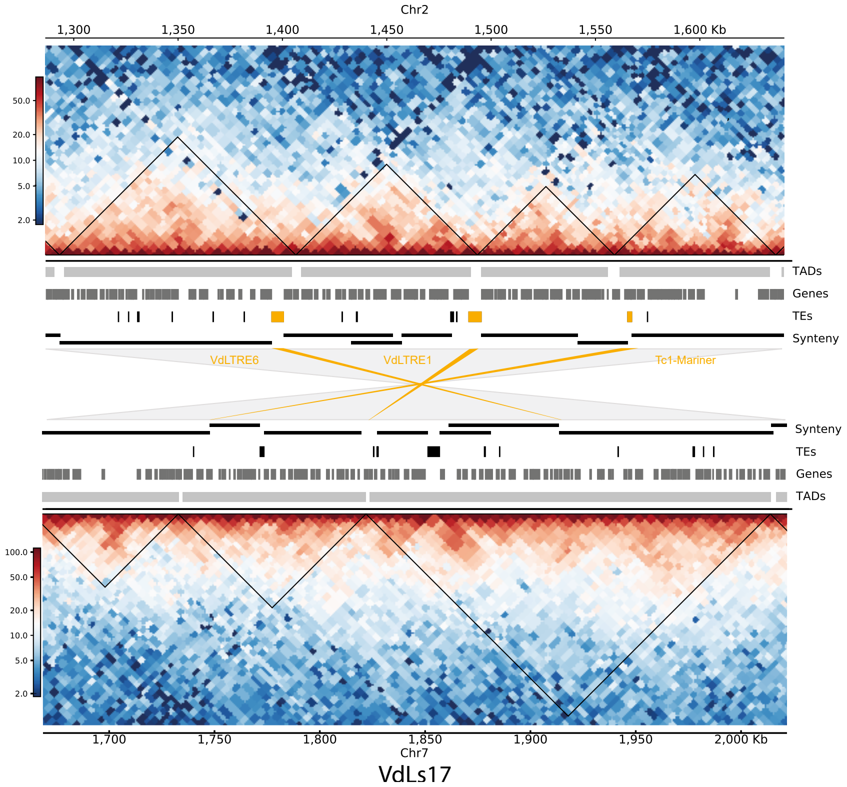


FIGURE S8 | Multiple TE insertions in *Verticillium dahliae* strain JR2 coincide with boundary rearrangements between *V. dahliae* strains. Heatmaps represent contact matrixes with TADs (black triangles) over a section of chromosome 2 of JR2 (top) and chromosome 7 of VdLs17 (bottom) that are syntenic. The region occurs inverted in VdLs17 relative to JR2. TADs, genes and TEs are displayed between heatmaps. Synteny between JR2 and VdLs17 indicated as grey blocks. TE insertions in strain JR2 are indicated in yellow.

TABLE S1 | Validation of boundary prediction using sequence-based method in *Verticillium dahliae* strain VdLs17

+		**Insulation method**		
		-	**Total**	
	+	173	96	269
Sequence method	-	130	42	
	Total	303		
True Positive Rate		0.765		
False Positive Rate		0.696		
Specificity		0.304		
Precision		0.643		
False Negative Rate		0.195		
Accuracy		0.488		

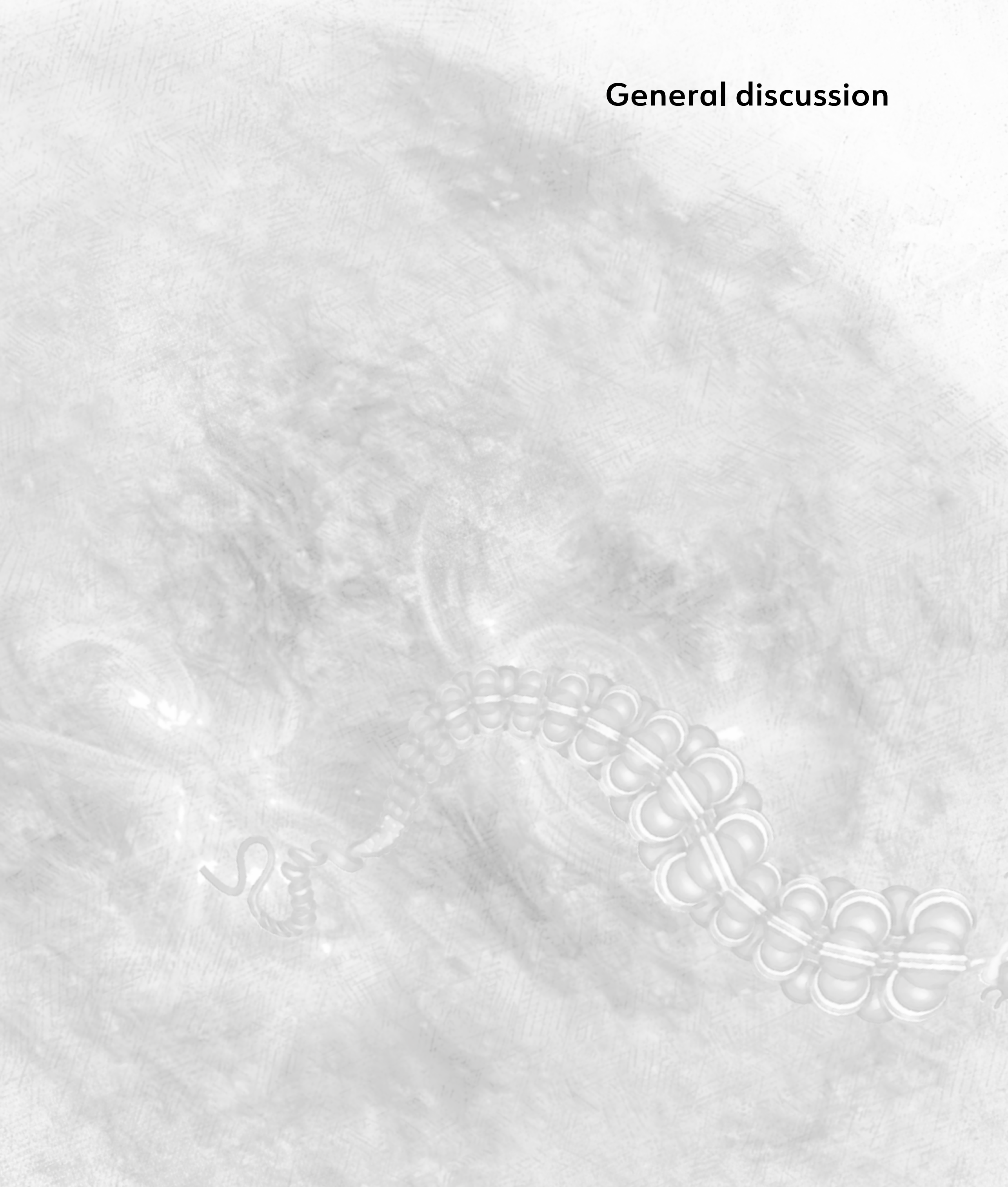
V. dahliae strain VdLs17 boundaries predicted by the sequence-based method (Fig. 4D) were compared to those predicted by the insulation method [362].

Acknowledgements

This work was supported by a PhD grant of the Research Council Earth and Life Sciences (project 831.15.002) to HMK, by the Consejo Nacional de Ciencia y Tecnología, México to DET and by Human Frontier Science Program Postdoctoral Fellowship (HFSP, LT000627/2014-L), by USDA's National Institute of Food and Agriculture (award no. 2018-67013-28492) through the Plant Biotic Interactions Program, and by the National Science Foundation (award no. 1936800) Division of Molecular and Cellular Biosciences to DEC. Work in the laboratories of M.F.S and B.P.H.J.T. is supported by the Research Council Earth and Life Sciences (ALW) of the Netherlands Organization of Scientific Research (NWO). B.P.H.J.T acknowledges funding by the Alexander von Humboldt Foundation in the framework of an Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research is furthermore supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2048/1 – Project ID: 390686111. The authors declare no conflict of interest.

Chapter 7

General discussion



Introduction

The most important transition in eukaryotic evolution has been cell compartmentalization, which allowed physical separation of cellular processes [422]. One of the organelles that arose from this compartmentalization is the nucleus; the organelle that harbors most of the DNA of the eukaryotic cell. Nuclear processes include those that are crucial for short-term cell survival, such as transcription and generation of small RNAs (sRNAs), as well as processes that are important for long-term survival of organisms and species, such as genome evolution, DNA replication and DNA repair. Regulation of all these DNA-associated processes converges at the histone code; the combination of histone modifications that regulates genome functionality [423,424].

Eukaryotes have four canonical histone proteins (histone 2A, histone 2B, histone 3 and histone 4) that form globular octameric protein complexes by incorporation of two monomers of each histone protein. These protein complexes, termed nucleosomes bind approximately 147 bp of DNA and provide the needed packaging to fit DNA into a nucleus [39,40]. Each histone protein carries a flexible N-terminal tail that sticks out from the nucleosome complex. These histone-tails are enriched for amino acid residues that can undergo chemical modification, such as methylation, acetylation and phosphorylation [318]. The combination of different histone modifications can regulate nuclear processes locally, but also more broadly. For local regulation of nuclear processes, histone modifications can serve as recognition site for enzymes that are active on DNA, whereas for broad regulation, they can affect accessibility of chromosomal regions and shape the three-dimensional (3D) genome arrangement [177,423,425,426].

Fungi represent a diverse kingdom of heterotrophic eukaryotic organisms that can be unicellular, as yeasts, or multicellular, as mycelial networks and higher order structures such as rhizomorphs, or mushrooms and other types of fruiting bodies [427,428]. Fungi thrive in many environments and are crucial for life on earth because of their important ecological role as decomposers of nearly all types of organic matter but also of certain types of non-organic matter. Through these traits, fungi have become important for human welfare not only as direct food source (mushrooms, Quorn), but also as fermenters of food (e.g. alcoholic drinks, bread and certain cheeses), as sources of useful metabolites (e.g. pigments, antibiotics and other pharmaceutical agents) and as biotechnological cell factories (e.g. for the production of citric acid that is widely used for instance in soft drinks) [429]. However, some fungi negatively affect human well-being, as they have evolved pathogenic lifestyles and cause disease in humans and animals, but also on plants, leading to reduced agricultural yields [430].

A notorious plant pathogenic fungus is *Verticillium dahliae*, which causes vascular wilt disease in hundreds of host plants [58]. *V. dahliae* is an ascomycete, haploid, fungus that reproduces predominantly asexually [62]. Genomic analyses on various strains uncovered that *V. dahliae* harbors a genome that evolved through large-scale chromosomal rearrangements, including duplications that have been followed by reciprocal gene losses [59,62,74,122,360]. Eventually, these processes resulted in a genome structure that can be characterized by core genomic regions that are shared by all strains, and adaptive genomic regions (AGRs) that show a considerable degree of plasticity among strains (Chapter 2) [61,74]. The availability of high-quality genome sequences and the relative ease of genetic analyses and manipulations in *V. dahliae* make it a well-suited organism to study the influence of various epigenetic features, such as changes to DNA, histones and chromatin, on the regulation of DNA-associated nuclear processes in plant pathogenic fungi.

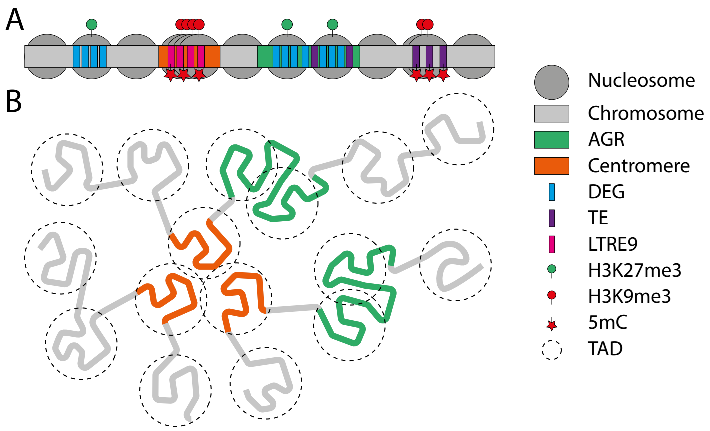


FIGURE 1 | Schematic representation of the major findings of this thesis: chromosomal regions differentially display distinct chromatin features, associated with differences in 3D-genome organization in *Verticillium dahliae.* (A) Representation of the chromatin structure on a linear *V. dahliae* chromosome. Adaptive genomic regions (AGRs, green blocks) display a distinct open (uncondensed) chromatin profile, in which the nucleosomes are marked by tri-methylation of histone 3 lysine 27 (H3K27me3, green circles) (see Chapter 2 for details). H3K27me3-marked regions, consisting of AGRs, as well as particular regions of the core genome, are enriched for differentially expressed genes in vitro and in planta (DEGs, blue blocks) (see Chapter 3 for details). Centromeres (orange blocks) in *V. dahliae*, but not in all sister species are specifically associated with the LTR retrotransposon LTRE9 (pink blocks) (see Chapter 4 for details). The chromatin profile at centromeres consists of tightly packed nucleosomes that are marked by tri-methylation of histone 3 lysine 9 (H3K9me3, red circles) and by DNA-methylation (5mC, red stars) (see Chapters 4 & 5 for details). Besides LTRE9 TEs at the centromere, additional inactive TEs in the core genome are marked by H3K9me3 and 5mC, while active TEs in the AGRs are not associated with these marks. (B) Schematic 3D-representation of the organization of three chromosomes in *V. dahliae*. Locally, genomic regions form topologically associating domains (TADs, indicated by dotted circles) that interact more strongly within the domain than with other domains. Intriguingly, TADs within AGRs are less well insulated and interact more freely with neighboring TADs (see Chapter 6 for details). Centromeres often form single TADs and display strong inter-centromeric interactions.

Epigenetic mechanisms affect transcription

To colonize their plant hosts, plant pathogens secrete effector molecules to evade host immune systems and tamper with host physiology [10]. The expression of effector molecules requires careful transcriptional regulation, as untimely expression may lead to failed infection. For instance, many haustoria-forming pathogens use cell wall-degrading enzymes to specifically weaken cell walls at sites of attempted invasion [30]. It is conceivable that if the pathogen produces too little of these enzymes, or not at the correct sites, successful entry of the cell may fail. In contrast, If the pathogen produces too much of the cell wall-degrading enzymes, the plant cells may be too weak to sustain the haustoria or the plant may induce defense mechanisms that arrest the fungus.

Regulation of transcription in eukaryotes involves binding of transcription factors to promoter regions, followed by recruitment of DNA-dependent RNA polymerases that generate mRNA molecules. The binding and recruitment of the transcriptional machinery is influenced by chromatin condensation over broad genomic regions [425,431], and more locally by particular histone modifications that recruit or inhibit binding of the transcriptional machinery [432,433]. Thus, epigenetic mechanisms are involved in the regulation of transcription. During axenic cultivation of various plant pathogenic fungi, when effector genes are generally thought to be inactive, these genes were shown to be enriched in genomic regions marked by H3K9me3 and/or H3K27me3, and mutants in which the enzymes that are responsible for deposition of these marks were deleted or silenced display induced effector gene expression [49,52–54,121,185,186,195]. As the histone modifications H3K9me3 and H3K27me3 are generally associated with inaccessible heterochromatin, these findings led to the hypothesis that genomic regions containing effector genes are heterochromatic, and therefore inaccessible to the transcriptional machinery, when the pathogen does not require effector gene expression. Consequently, in order to express the effector genes that facilitate infection, pathogens are generally hypothesized to require chromatin de-condensation at effector gene-containing regions, which requires depletion of H3K9me3 and H3K27me3 [51,52,193]. Alternatively, it is also possible that the transcriptionally repressive H3K9me3 and H3K27me3 remain stable, while activating modifications are added to regulated transcription.

The findings described in this thesis concerning epigenetic regulation of transcription are mainly based on analyses performed on *in vitro* growth conditions. Preferably, we would have investigated the regulation of effector gene expression *in planta*, the niche where the fungus naturally occurs and the context in which it is most studied. However, epigenetic studies of fungal chromatin during plant infection are typically impeded by the usually low ratio of pathogen-to-plant biomass, resulting in excessive amounts of plant-derived sequences from such mixed samples [193]. In fact, previously performed RNA sequencing experiments of *V. dahliae*-inoculated *Nicotiana benthamiana* resulted in only 0.05 till 0.9% of reads mapping to the *V. dahliae* genome [192]. During this thesis research I have attempted several procedures to retrieve sufficient *V. dahliae*-derived DNA for *in planta* chromatin immunoprecipitation assays. These attempts included sugar-gradient centrifugation to separate fungal and plant nuclei, as well as enzymatic treatments to specifically degrade plant cells and nuclei while keeping fungal cells intact. Both these methods showed merit, but I was nevertheless unable to consistently enrich for sufficient fungal chromatin.

In agreement with the notion that effectors are heterochromatically silenced *in vitro*, we find that the genomic regions harboring effector genes and other environmentally-regulated genes in *V. dahliae* are enriched for H3K27me3 *in vitro* and that loss of H3K27me3 in the *Set7* deletion mutant, lacking the gene encoding the H3K27me3 methyltransferase, leads to induction of previously marked genes (Chapters 2 & 3). We therefore conclude that transcriptional regulation of effector gene expression in *V. dahliae* likely occurs in a similar fashion as in other filamentous plant pathogens (Chapters 2 & 3). However, even though our findings suggest that H3K27me3 is involved in transcriptional regulation in *V. dahliae*, our results do not suggest that depletion of H3K27me3 is the major event leading to transcriptional activation (Chapters 2 & 3). In fact, we find that H3K27me3-marked chromatin is already in a de-condensed state during cultivation *in vitro* (Fig. 1A, Chapter 2), and that differential expression of genes located in H3K27me3-marked domains does not require concomitant changes in H3K27me3 (Chapter 3). Additionally, even though presence of H3K27me3 inhibits transcription, loss of H3K27me3 in the *Set7* deletion mutant does not lead to transcriptional induction of all previously H3K27me3-associated genes in *V. dahliae* (Chapter 3). Loss of H3K27me3 in the plant pathogenic fungi *Fusarium graminearum, Fusarium fujikuroi* and *Magnaporthe oryzae* similarly leads to induction of only a subset of previously H3K27me3-associated genes [49,53,195], showing that other factors are involved in transcriptional regulation. Thus, our results in *V. dahliae*, combined with the results in other fungi, indicate that the previously postulated hypothesis does not fully describe the role of H3K27me3 in regulation of effector gene expression. Rather, it is perhaps more likely that while H3K27me3 dynamics are not essential for transcriptional regulation, H3K27me3 presence at genomic regions increases the propensity of genes to become differentially expressed, possibly by serving as binding site for transcriptional regulators that are activated or produced upon detection of particular environmental signals.

Proteins recognizing and binding to histone modifications are generally known as histone readers and can function in various cellular processes, including transcriptional regulation [434]. For instance, in a human cell line, one such histone reader is the TAF3 subunit of the basal transcription complex TFIID that was shown to specifically bind H3K4me3, which is often associated with transcriptionally active genes [435]. In response to genotoxic stress, TAF3 is recruited to H3K4me3-marked genes to stimulate formation of the preinitiation complex and thereby regulates initiation of gene expression, yet was found to not affect constitutive gene expression [436]. In contrast, binding of H3K4me3 by the histone reader ING2, a subunit of the mSin3a histone deacetylase complex, leads to rapid repression of gene expression [437]. Alternative to histone modifications being recognized by separate readers, resulting in opposing transcriptional outputs, it is also possible that single histone readers simultaneously recognize multiple histone marks, resulting in different transcriptional outputs depending on the present histone marks. For instance, the *Arabidopsis* histone reader EBS recognizes both H3K4me3 and H3K27me3 and can switch between binding the histone marks to balance gene expression [438]. As the histone code in plant pathogens is only partially scrutinized so far, it is possible that H3K4me3 or another, yet unidentified, histone modification works in conjunction with H3K27me3 to regulate transcription. These examples show that the presence of a specific histone modification may result in opposing transcriptional outputs. Therefore, it is possible that the relatively stable presence of H3K27me3 that we observe in *V. dahliae*, may lead to different transcriptional outputs, depending on whether an activating or repressing histone reader is bound. Interestingly, recently transcription repressing H3K27me3 histone readers were identified in the filamentous fungi *F. graminearum* and *Neurospora crassa* [196,197]. As these histone readers inhibit binding of transcriptional machinery, it is conceivable that dynamic presence of the reader over stable H3K27me3 domains regulates transcription.

Besides potentially functioning as a direct regulator of transcription, H3K27me3 may also affect transcriptional regulation through shaping 3D organization of the genome within the nucleus. In various organisms, including animals, plants and fungi, genomes have been shown to display long-range intra-chromosomal and inter-chromosomal interactions [106,342,377,439]. At smaller genomic distances, 3D organization results in the formation of topologically associating domains (TADs), which are local genome regions that interact preferentially within themself, but not with adjacent genomic regions [323,373]. Although the exact functionality of TADs is still under debate [327,329,440,441], TADs have frequently been associated with transcriptional regulation, for instance by facilitating enhancer-promoter interactions [321,325,373,386]. H3K27me3 has been shown to affect both global and local 3D organization of chromosomes [105,442,443]. Thus, H3K27me3 may be involved in structuring the 3D genome such that it is conducive for expression when required, for instance by providing a local 3D-chromatin environment pre-loaded with components of the transcriptional machinery. Interestingly, in *V. dahliae* we find that TADs in the H3K27me3-associated AGRs are weaker insulated than TADs in the core genome, meaning that TADs in AGRs are less well-separated from their neighboring TADs. Therefore, TADs in the AGRs interact more freely with adjacent TADs when compared with those in the core genome (Fig. 1B, Chapter 6). Additionally, we found that genes within the H3K27me3-associated TADs in AGRs are more likely to be transcriptionally co-regulated than genes in core TADs (Chapter 6). Collectively, this suggests that H3K27me3 may be involved in generating a distinct local 3D genome organization, paired with a divergent transcriptional profile, at TADs in AGRs, when compared with TADs in the core genome. However, admittedly, we have not demonstrated a driving role of H3K27me3 in these processes. Therefore, future efforts to demonstrate a direct effect of H3K27me3 on local 3D genome organization will be important.

Besides mRNA molecules, also different types of small RNAs are encoded in the genome, including micro RNAs (miRNAs) and small interfering RNAs (siRNAs) [444,445]. These different classes of small RNAs are generated by their own specific machinery, which involves epigenetic mechanisms. For instance, siRNAs are molecules of 21-25 nucleotides that are cleaved from double-strand RNA precursor molecules [446]. Such precursor RNAs can be transcribed by RNA polymerase IV that is recruited to genomic regions with elevated DNA methylation in *Arabidopsis* [447]. Moreover, loss of DNA methylation over genomic regions from which siRNA precursors are generated enhances the production of siRNA precursors in *Arabidopsis* [448]. Similar to siRNAs, miRNAs are 21-25 nucleotides in size, but are cleaved from primary-miRNA molecules that are transcribed by RNA polymerase II in plants and animals [449,450]. Transcription of primary-miRNAs in *Arabidopsis* is promoted by activity of the chromatin remodeling factor BRM [451], indicating that generation of miRNAs is also influenced by chromatin. Additionally, the RNA-interference machinery that causes transcriptional silencing of target genes, yet also promotes sRNA amplification at these loci, is recruited to genomic regions containing H3K9me2/3 [452,453]. Collectively, these findings in a wide range of organisms indicate that epigenetic mechanisms strongly affect the transcriptional output of a genome.

Epigenetic mechanisms affect genome evolution

To survive over evolutionary timescales, organisms need to gain, lose or alter proteins performing particular functions through genome evolution. Genome evolution starts from a combination of mechanisms, including DNA replication “errors”, the activity of mutagens such as radiation, chromosomal crossovers during meiosis, transposable element (TE) activity, (partial) genome duplications, tandem duplications, chromosome gain or loss, presence-absence polymorphisms, large-scale chromosome rearrangements, *et cetera*, leading to genome variation [454]. In some cases, the generated genome variation leads to increased cell viability, potentially affecting their frequency in the population, ultimately leading to evolution. Even though genome evolution is considered a stochastic process, it was found that various plant pathogens harbor genomic regions that display accelerated evolution, while the majority of the genome, which is typically designated as the core genome, remains evolutionary rather stable [69,74,79,82,84,356–358,455]. This compartmentalization of genomes into plastic regions, that differ more frequently between strains of a species, and less plastic genomic regions is termed the two-speed genome [82,84]. The fast-evolving, plastic, compartments of a two-speed genome are typically TE-rich [84], indicating that TE activity may be one of the more important drivers for evolution of plant pathogens, including *V. dahliae.* The break-points of the large-scale chromosome rearrangements that occurred during the evolution of *V. dahliae* often locate in AGRs, and show a great degree of overlap with TE-rich duplicated regions, suggesting that the presence of these duplicated TEs may underly chromosome rearrangements [59,74]. Interestingly, the TEs located in the plastic genome are relatively young, transcriptionally active, and more frequently polymorphic when compared to TEs in the core genome, suggesting that these TEs actively contribute to shaping AGRs [74,360]. Additionally, the polymorphic TEs in AGRs can be associated with *in planta* highly expressed pathogenicity-related genes, suggesting that TEs may be involved in transcriptional regulation as well [360].

Even though TE activity is beneficial to a certain extent, TE overactivity can be detrimental to genome stability, and therefore TEs are generally epigenetically silenced [456–459]. In fungi, genomic regions that are enriched for TEs are often epigenetically silenced by H3K9me3 and cytosine methylation (5-methylcytocine, 5mC) [93,217,218,222,460]. Similarly, in *V. dahliae* we found that H3K9me3 and 5mC co-localize on TE-rich genomic regions (Fig. 1A, Chapters 2 & 4). However, even though 5mC is generally thought to be involved in transcriptional silencing, we observed that loss of 5mC did not induce TE transcription, whereas loss of H3K9me3, and the accompanying loss of 5mC, leads to the induction of numerous TEs (Chapter 4). Thus, 5mC is not strictly necessary for TE silencing. Instead 5mC may be subject to spontaneous deamination, causing C to T mutations, and thus potentially rendering affected TEs forever inactive [226]. Cytosine deamination as a driver of genome evolution is well accepted in various taxonomic groups [461–463]. For instance, simulations of DNA sequence evolution indicated that mutational pressure by cytosine deamination was vital for the evolution of isochore structures in the mammalian genomes [462]. Additionally, cytosine deamination has been proposed to constitute one of the main evolutionary forces in generating new transcription factor binding sites in the human genome [464]. However, as cytosine methylation is mainly restricted to genomic regions containing TEs in *V. dahliae*, spontaneous deamination is likely only involved in the inactivation of TEs, and not in genome evolution in a more general sense.

Previous studies in *V. dahliae* indicated that relatively young and active TEs underly the evolution of the AGRs as plastic regions (Chapter 2) [74]. Interestingly, the TEs in these AGRs have a lower fraction of C to T mutations (represented by the composite repeat-induced point mutation index (CRI)) and display lower association with H3K9me3 and 5mC (Chapter 2) [74,360]. This indicates that C to T mutations happen more frequently in TEs that are marked with 5mC, and thus that spontaneous deamination may be a true end-result of DNA methylation, but also that a particular subset of TEs is devoid of H3K9me3 and 5mC to allow for TE activity driving evolution within the plastic regions of the two-speed genome. It is unclear what makes that these TE silencing mechanisms preferentially target TEs in the core genome. It is possible that this is merely driven by the "survival of the fittest" principle, making that fungal cells with active TEs in the core genome suffer too many detrimental effects and therefore disappear from the population, while cells with active TEs in the AGRs suffer less from such detrimental effects, and thus survive more frequently. Alternatively, the presence of specific epigenetic features of AGRs may constrain the deposition of H3K9me3 and 5mC on TEs in the plastic genome, thereby only allowing for TE activity within AGRs. Furthermore, it is conceivable that the specific epigenetic profile of AGRs makes that these regions cluster in the 3D-genome, generating so-called nuclear bodies that allow concentration and spatial segregation of nuclear activities into membrane-less sub-compartments [399–401]. Various types of nuclear bodies have been described, including nucleoli, Cajal bodies, nuclear speckles, nuclear stress bodies and polycomb bodies, of which some are formed on chromatin and embed DNA while others form in the nucleosol and do not contain DNA, but rather interact with chromatin [400,465]. Chromatin-containing nuclear bodies form by a process called phase separation through the activity of self-aggregating chromatin-binding molecules or through the activity of individual chromatin bridging factors that cross-link separate chromatin sections, without self-aggregation (Fig. 2) [465]. Nuclear bodies formed with the non-aggregating bridging factors are usually less stable, as these molecules can more readily disperse into the nucleoplasm (Fig. 2A), whereas nuclear bodies formed with self-aggregating molecules are more stable (Fig. 2B), and can exist independent of chromatin [465]. Interestingly, the *Drosophila* H3K9me3-interacting protein Hp1 was shown to aggregate *in vitro* and to nucleate into foci during early heterochromatin domain formation, suggesting that aggregation of Hp1 may drive heterochromatin domain formation [466]. As such, H3K9me3-marked heterochromatin at distal genomic regions may cluster in the nucleus through presence of Hp1, for instance at centromeres [467]. Similarly, H3K27me3-marked heterochromatin is bound by the polycomb repressive complex 1 (PRC1), of which the CBX2 protein member is capable of assembly through phase separation [468]. PRC1 components are absent from fungi [199], yet other H3K27me3-readers may have an analogous function in fungi. Thus, it is possible that the H3K27me3-marked AGRs in *V. dahliae* form nuclear bodies, in which TE activity mainly causes TE insertions in the plastic regions, while leaving the core genome largely unaffected. We have started to assemble preliminary evidence, not included in this thesis, supporting the preferential interaction of AGRs located on separate chromosomes, and suggesting that AGRs indeed form sub-compartments in the nucleus.

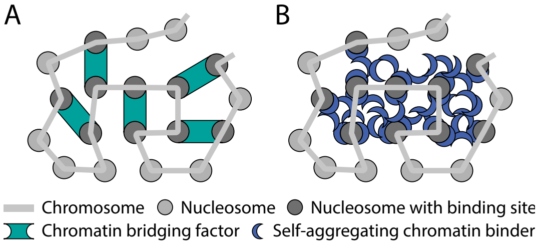

A
B
Chromosome
Nucleosome
Nucleosome with binding site
Chromatin bridging factor
Self-aggregating chromatin binder

FIGURE 2 | Models for formation of nuclear sub-compartments. Genomic regions containing nucleosomes with specific binding sites (dark grey circles) can form chromatin sub-compartments through (A) association with bridging factors (green shapes) that cross-link the chromatin, or through (B) association with self-aggregating chromatin binders (purple shapes). Sub-compartments formed by bridging factors are generally less stable, as these factors can more readily disperse into the nucleoplasm, whereas self-aggregating chromatin binders form clusters that remain stable independent of chromatin. Figure adapted from Erdel & Rippe, 2018 [465].

All organisms evolved DNA repair mechanisms that correct damage to the genome. Various DNA repair mechanisms exist, including double-strand break repair by homologous recombination or by non-homologous end joining, and nucleotide and base-excision repair pathways [469,470]. Interestingly, the histone code has been implicated in these DNA repair mechanisms. For instance, histone methylation of lysine 79 on the tail of histone 3, and of lysine 20 on histone 4, as well as phosphorylation and ubiquitination of histone variant H2AX are involved in recruitment of double-strand break machinery [470]. Additionally, budding yeast mutants lacking the N-terminal tails of histones H2A and H3 displayed increased mutation rates due to deficient base-excision repair, indicating that chromatin plays an important role in this DNA repair mechanism [471]. As these DNA-repair mechanisms do not always function faithfully they inhibit, yet do not prevent, DNA changes. In fact, research in *V. dahliae* identified sequence signatures of double strand-break repair machinery at sites of chromosomal rearrangements, indicating that the evolution of the two-speed genome in *V. dahliae* involved erroneous double strand-break repair [74]. As these mechanisms are, in part, regulated on chromatin, it is becoming more and more evident that epigenetics is important for genome evolution.

Epigenetic mechanisms affect cell division

For mitotic and meiotic replication, organisms first require DNA replication, followed by segregation of chromosome pairs, and finally cell division [472]. DNA replication starts with unwinding and separation of DNA strands, resulting in formation of replication forks in which DNA-polymerases attach to the DNA. It was shown that formation of replication forks is favored in genomic regions with hyperacetylated euchromatin [473,474], and that timing of replication is further regulated on chromatin [474–476]. Heterochromatic regions replicate later during the mitotic S-phase and this process involves histone acetylation and methylation, as well as the activity of histone readers recognizing heterochromatin-associated histones [477–479].

After DNA replication, the generated chromosome pairs need to be segregated, which happens through formation of microtubule spindles that attach to centromeric regions present on each chromosome. Fungal centromeres vary significantly in composition and size between species, ranging from point centromeres of approximately 125 bp in size to regional centromeres of a few kb up till a few hundreds of kb [243,245]. Even though fungal centromeres vary widely, their chromatin is always characterized by presence of nucleosomes carrying the histone H3 variant CenH3 [243,245]. CenH3 is essential for centromere function, as it is the chromatin component that connects chromosomes to the microtubule spindle via the proteinaceous kinetochore complex [480]. Besides CenH3, fungal centromeres are often epigenetically characterized by H3K9me3 and 5mC [233,245], which we also found to be present on *V. dahliae* centromeres (Fig. 1A, Chapter 5). Additionally, in various plants and animals, a large set of histone modifications have been associated with centromeres [481]. The conservation of epigenetic profiles at centromeres in different organisms indicates that the epigenetic landscape likely plays a crucial role in centromere function, and thus in cell division. Such role of the epigenetic landscape, and perhaps especially of H3K9me3, may entail the formation of a nuclear sub-compartment and thereby drive the physical clustering of centromeres. The crucial role of the epigenetic landscape in centromere functioning is further supported by studies into neocentromere formation, occurring upon centromere defects, showing that neocentromeres often form in H3K9me3-marked genomic regions [482]. However, neocentromeres in the human pathogenic fungus *Cryptococcus deuterogattii* are formed in genic regions that are not associated with H3K9me3 and 5mC [483,484], suggesting that these heterochromatic features are not essential. Moreover, *C. deuterogattii* chromosomes lacking their original centromere are unstable and undergo chromosomal fusions, after which the neocentromere loses its function [483,484]. These results suggest that although H3K9me3 and 5mC may not be essential for centromere function and cell viability in short-term, they are important for genome stability over longer evolutionary timescales.

Concluding remarks

In this thesis, we show that genomic regions differ in their chromatin profile, which is associated with differences in 3D-genome organization in *Verticillium dahliae* (Fig. 1). Moreover, we contest the overall starting hypothesis that genomic regions containing effector genes are heterochromatic when the pathogen does not require effector gene expression, and consequently require chromatin de-condensation to express effectors upon plant recognition. As we show that differential gene expression *in vitro* for only a subset of genes located in H3K27me3 domains involves local H3K27me3 depletion, we postulate that gene expression *in planta* may require local H3K27me3 depletion and chromatin, but that it is not strictly required. However, to irrefutably demonstrate this, future studies will require a functional, and reliable, procedure to perform *in planta* chromatin immunoprecipitation assays.

Histone proteins are not only found in eukaryotes, but are also common in archaea [485,486], indicating that the potential for epigenetic regulation using histones is evolutionary ancient. As such, it is not surprising that nuclear processes have evolved to heavily rely on epigenetic regulation. This evolution resulted in an intricate mechanism, the histone code, in which certain histone modifications may have multiple different functions depending on their genetic localization and the co-occurrence of additional modifications. Therefore, it will be difficult to predict how the function of the genome is affected by specific changes in the histone code. Advances in single cell sequencing and epigenome analyses [487,488] will deepen the understanding of epigenetics by providing increasingly more fine-grained information about the regulation and output of the epigenome. Studies into epigenetics may ultimately identify subtle differences in regulatory mechanisms between plants and their pathogens, which could be exploited to design novel strategies to combat plant diseases.

References
Summary
Acknowledgements
About the author
List of publications

References

1. Xin, X.F., Kvitko, B. and He, S.Y. (2018) *Pseudomonas syringae*: What it takes to be a pathogen. *Nat. Rev. Microbiol.*, **16**, 316–328.
2. Edel-Hermann, V. and Lecomte, C. (2019) Current status of *Fusarium oxysporum formae speciales* and races. *Phytopathology*, **109**, 512–530.
3. Wheeler, D.L., Dung, J.K.S. and Johnson, D.A. (2019) From pathogen to endophyte: an endophytic population of *Verticillium dahliae* evolved from a sympatric pathogenic population. *New Phytol.*, **222**, 497–510.
4. Malcolm, G.M., Kuldau, G.A., Gugino, B.K. and Jiménez-Gasco, M.D.M. (2013) Hidden host plant associations of soilborne fungal pathogens: An ecological perspective. *Phytopathology*, **103**, 538–544.
5. Cheng, Y.T., Zhang, L. and He, S.Y. (2019) Plant-microbe interactions facing environmental challenge. *Cell host microbe*, **26**, 183–192.
6. Martin, J.T. (1964) Role of cuticle in the defense against plant disease. *Annu. Rev. Phytopathol.*, **2**, 81–100.
7. Serrano, M., Coluccia, F., Torres, M., L'Haridon, F. and Métraux, J.P. (2014) The cuticle and plant defense to pathogens. *Front. Plant Sci.*, 10.3389/fpls.2014.00274.
8. VanEtten, H.D., Mansfield, J.W., Bailey, J.A. and Farmer, E.E. (1994) Two classes of plant antibiotics: phytoalexins versus 'phytoanticipins'. *Plant Cell*, **6**, 1191–1192.
9. Cook, D.E., Mesarich, C.H. and Thomma, B.P.H.J. (2015) Understanding plant immunity as a surveillance system to detect invasion. *Annu. Rev. Phytopathol.*, **53**, 541–563.
10. Jones, J.D.G. and Dangl, J.L. (2006) The plant immune system. *Nature*, **444**, 323–329.
11. Rovenich, H., Boshoven, J.C. and Thomma, B.P.H.J. (2014) Filamentous pathogen effector functions: Of pathogens, hosts and microbiomes. *Curr. Opin. Plant Biol.*, 10.1016/j.pbi.2014.05.001.
12. Flor, H.H. (1971) Current status of the gene-for-gene concept. *Annu. Rev. Phytopathol.*, **9**, 275–296.
13. Ngou, B.P.M., Ahn, H.-K., Ding, P. and Jones, J.D.G. (2021) Mutual potentiation of plant immunity by cell-surface and intracellular receptors. *Nature*, **592**, 110–115.
14. Yuan, M., Jiang, Z., Bi, G., Nomura, K., Liu, M., Wang, Y., Cai, B., Zhou, J.-M., He, S.Y. and Xin, X.-F. (2021) Pattern-recognition receptors are required for NLR-mediated plant immunity. *Nature*, **592**, 105–109.
15. Pruitt, R.N., Locci, F., Wanke, F., Zhang, L., Saile, S.C., Joe, A., Karelina, D., Hua, C., Fröhlich, K., Wan, W.-L., *et al.* (2021) The EDS1–PAD4–ADR1 node mediates *Arabidopsis* pattern-triggered immunity. *Nature*, **598**, 495–499.
16. Stergiopoulos, I. and de Wit, P.J.G.M. (2009) Fungal effector proteins. *Annu. Rev. Phytopathol.*, **47**, 233–263.
17. de Wit, P.J.G.M. (2016) *Cladosporium fulvum* effectors: weapons in the arms race with tomato. *Annu. Rev. Phytopathol.*, **54**, 1–23.
18. Sperschneider, J., Dodds, P.N., Singh, K.B. and Taylor, J.M. (2018) ApoplastP: prediction of effectors and plant proteins in the apoplast using machine learning. *New Phytol.*, **217**, 1764–1778.
19. Rivas, S. and Genin, S. (2011) A plethora of virulence strategies hidden behind nuclear targeting of microbial effectors. *Front. Plant Sci.*, 10.3389/fpls.2011.00104.
20. Chaudhari, P., Ahmed, B., Joly, D.L. and Germain, H. (2014) Effector biology during biotrophic invasion of plant cells. *Virulence*, **5**, 703–709.
21. Boch, J. and Bonas, U. (2010) *Xanthomonas* AvrBs3 family-type III effectors: discovery and function. *Annu. Rev. Phytopathol.*, **48**, 419–436.
22. Kong, L., Qiu, X., Kang, J., Wang, Y., Chen, H., Huang, J., Qiu, M., Zhao, Y., Kong, G., Ma, Z., *et al.* (2017) A *Phytophthora* effector manipulates host histone acetylation and reprograms defense gene expression to promote infection. *Curr. Biol.*, **27**, 981–991.
23. Vijayapalani, P., Hewezi, T., Pontvianne, F. and Baum, T.J. (2018) An effector from the cyst nematode *Heterodera schachtii* derepresses host rRNA genes by altering histone acetylation. *Plant Cell*, **30**, 2795–2812.
24. Lo Presti, L., Lanver, D., Schweizer, G., Tanaka, S., Liang, L., Tollot, M., Zuccaro, A., Reissmann, S. and Kahmann, R. (2015) Fungal effectors and plant susceptibility. *Annu. Rev. Plant Biol.*, **66**, 513–545.
25. Hua, C., Zhao, J.H. and Guo, H.S. (2018) Trans-kingdom RNA silencing in plant–fungal pathogen interactions. *Mol. Plant*, **11**, 235–244.

A

26. Weiberg, A., Wang, M., Lin, F.M., Zhao, H., Zhang, Z., Kaloshian, I., Huang, H. Da and Jin, H. (2013) Fungal small RNAs suppress plant immunity by hijacking host RNA interference pathways. *Science*, **342**, 118–123.

27. Toruño, T.Y., Stergiopoulos, I. and Coaker, G. (2016) Plant-pathogen effectors: cellular probes interfering with plant defenses in spatial and temporal manners. *Annu. Rev. Phytopathol.*, **54**, 419–441.

28. Gervais, J., Plissonneau, C., Linglin, J., Meyer, M., Labadie, K., Cruaud, C., Fudal, I., Rouxel, T. and Balesdent, M.H. (2017) Different waves of effector genes with contrasted genomic location are expressed by *Leptosphaeria maculans* during cotyledon and stem colonization of oilseed rape. *Mol. Plant Pathol.*, **18**, 1113–1126.

29. Lanver, D., Müller, A.N., Happel, P., Schweizer, G., Haas, F.B., Franitza, M., Pellegrin, C., Reissmann, S., Altmüller, J., Rensing, S.A., *et al.* (2018) The biotrophic development of *Ustilago maydis* studied by RNA-seq analysis. *Plant Cell*, **30**, 300–323.

30. Xu, H. and Mendgen, K. (1997) Targeted cell wall degradation at the penetration site of cowpea rust basidiosporelings. *MPMI*, **10**, 87–94.

31. Michielse, C.B., Van Wijk, R., Reijnen, L., Manders, E.M.M., Boas, S., Olivain, C., Alabouvette, C. and Rep, M. (2009) The nuclear protein Sge1 of *Fusarium oxysporum* is required for parasitic growth. *PLoS Pathog.*, **5**, e1000637.

32. Michielse, C.B., Becker, M., Heller, J., Moraga, J., Collado, I.G. and Tudzynski, P. (2011) The *Botrytis cinerea* Reg1 protein, a putative transcriptional regulator, is required for pathogenicity, conidiogenesis, and the production of secondary metabolites. *MPMI*, **24**, 1074–1085.

33. Jonkers, W., Dong, Y., Broz, K. and Kistler, H.C. (2012) The Wor1-like protein Fgp1 regulates pathogenicity, toxin synthesis and reproduction in the phytopathogenic fungus *Fusarium graminearum*. *PLoS Pathog.*, **8**, 1–18.

34. Brown, D.W., Busman, M. and Proctor, R.H. (2014) *Fusarium verticillioides* SGE1 is required for full virulence and regulates expression of protein effector and secondary metabolite biosynthetic genes. *MPMI*, **27**, 809–823.

35. Santhanam, P. and Thomma, B.P.H.J. (2013) *Verticillium dahliae* Sge1 differentially regulates expression of candidate effector genes. *MPMI*, **26**, 249–256.

36. Cho, Y., Ohm, R.A., Grigoriev, I. V. and Srivastava, A. (2013) Fungal-specific transcription factor AbPf2 activates pathogenicity in *Alternaria brassicicola*. *Plant J.*, **75**, 498–514.

37. Jones, D.A.B., John, E., Rybak, K., Phan, H.T.T., Singh, K.B., Lin, S.Y., Solomon, P.S., Oliver, R.P. and Tan, K.C. (2019) A specific fungal transcription factor controls effector gene expression and orchestrates the establishment of the necrotrophic pathogen lifestyle on wheat. *Sci. Rep.*, **9**, 15884.

38. Rybak, K., See, P.T., Phan, H.T.T., Syme, R.A., Moffat, C.S., Oliver, R.P. and Tan, K.C. (2017) A functionally conserved Zn2Cys6 binuclear cluster transcription factor class regulates necrotrophic effector gene expression and host-specific virulence of two major Pleosporales fungal pathogens of wheat. *Mol. Plant Pathol.*, **18**, 420–434.

39. Lowary, P.T. and Widom, J. (1997) Nucleosome packaging and nucleosome positioning of genomic DNA. *Proc. Natl. Acad. Sci.*, **94**, 1183–1188.

40. Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F. and Richmond, T.J. (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, **389**, 251–260.

41. Huang, C. and Zhu, B. (2018) Roles of H3K36-specific histone methyltransferases in transcription: antagonizing silencing and safeguarding transcription fidelity. *Biophys. Reports*, **4**, 170–177.

42. Lau, P.N.I. and Cheung, P. (2011) Histone code pathway involving H3 S28 phosphorylation and K27 acetylation activates transcription and antagonizes polycomb silencing. *Proc. Natl. Acad. Sci.*, **108**, 2801–2806.

43. Le, T.-N., Schumann, U., Smith, N.A., Tiwari, S., Au, P., Zhu, Q.-H., Taylor, J.M., Kazan, K., Llewellyn, D.J., Zhang, R., *et al.* (2014) DNA demethylases target promoter transposable elements to positively regulate stress responsive genes in *Arabidopsis*. *Genome Biol.*, **15**, 458.

44. Briand, N. and Collas, P. (2020) Lamina-associated domains: peripheral matters and internal affairs. *Genome Biol.*, **21**, 1–25.

45. Gesson, K., Rescheneder, P., Skoruppa, M.P., von Haeseler, A., Dechat, T. and Foisner, R. (2016) A-type lamins bind both hetero-and euchromatin, the latter being regulated by lamina-associated polypeptide 2 alpha. *Genome Res.*, **26**, 462–473.

46. Dekker, J. and Heard, E. (2015) Structural and functional diversity of topologically associating domains. *FEBS Lett.*, **589**, 2877–2884.
47. Szabo, Q., Bantignies, F. and Cavalli, G. (2019) Principles of genome folding into topologically associating domains. *Sci. Adv.*, **5**, eaaw1668.
48. Reyes-Dominguez, Y., Boedi, S., Sulyok, M., Wiesenberger, G., Stoppacher, N., Krska, R. and Strauss, J. (2012) Heterochromatin influences the secondary metabolite profile in the plant pathogen *Fusarium graminearum*. *Fungal Genet. Biol.*, **49**, 39–47.
49. Connolly, L.R., Smith, K.M. and Freitag, M. (2013) The *Fusarium graminearum* histone H3 K27 methyltransferase KMT6 regulates development and expression of secondary metabolite gene clusters. *PLoS Genet.*, **9**, e1003916.
50. Reyes-Dominguez, Y., Bok, J.W., Berger, H., Shwab, E.K., Basheer, A., Gallmetzer, A., Scazzocchio, C., Keller, N. and Strauss, J. (2010) Heterochromatic marks are associated with the repression of secondary metabolism clusters in *Aspergillus nidulans*. *Mol. Microbiol.*, **76**, 1376–1386.
51. Chujo, T. and Scott, B. (2014) Histone H3K9 and H3K27 methylation regulates fungal alkaloid biosynthesis in a fungal endophyte-plant symbiosis. *Mol. Microbiol.*, **92**, 413–434.
52. Soyer, J.L., El Ghalid, M., Glaser, N., Ollivier, B., Linglin, J., Grandaubert, J., Balesdent, M.H., Connolly, L.R., Freitag, M., Rouxel, T., *et al.* (2014) Epigenetic control of effector gene expression in the plant pathogenic fungus *Leptosphaeria maculans*. *PLoS Genet.*, **10**, e1004227.
53. Studt, L., Rösler, S.M., Burkhardt, I., Arndt, B., Freitag, M., Humpf, H.U., Dickschat, J.S. and Tudzynski, B. (2016) Knock-down of the methyltransferase Kmt6 relieves H3K27me3 and results in induction of cryptic and otherwise silent secondary metabolite gene clusters in *Fusarium fujikuroi*. *Environ. Microbiol.*, **18**, 4037–4054.
54. Möller, M., Schotanus, K., Soyer, J.L., Haueisen, J., Happ, K., Stralucke, M., Happel, P., Smith, K.M., Connolly, L.R., Freitag, M., *et al.* (2019) Destabilization of chromosome structure by histone H3 lysine 27 methylation. *PLoS Genet.*, **15**, e1008093.
55. Soyer, J.L., Grandaubert, J., Haueisen, J., Schotanus, K. and Holtgrewe Stukenbrock, E. (2019) *In planta* chromatin immunoprecipitation in *Zymoseptoria tritici* reveals chromatin-based regulation of putative effector gene expression. *bioRxiv*, 10.1101/544627.
56. Inderbitzin, P. and Subbarao, K. V. (2014) *Verticillium* systematics and evolution: How confusion impedes verticillium wilt management and how to resolve it. *Phytopathology*, **104**, 564–574.
57. Klosterman, S.J., Atallah, Z.K., Vallad, G.E. and Subbarao, K. V. (2009) Diversity, pathogenicity, and management of *Verticillium* species. *Annu. Rev. Phytopathol.*, **47**, 39–62.
58. Fradin, E.F. and Thomma, B.P.H.J. (2006) Physiology and molecular aspects of Verticillium wilt diseases caused by *V. dahliae* and *V. albo-atrum*. *Mol. Plant Pathol.*, **7**, 71–86.
59. de Jonge, R., Bolton, M.D., Kombrink, A., van den Berg, G.C.M., Yadeta, K.A. and Thomma, B.P.H.J. (2013) Extensive chromosomal reshuffling drives evolution of virulence in an asexual pathogen. *Genome Res.*, **23**, 1271–1282.
60. Song, Y., Liu, L., Wang, Y., Valkenburg, D., Zhang, X., Zhu, L. and Thomma, B.P.H.J. (2018) Transfer of tomato immune receptor Ve1 confers Ave1-dependent *Verticillium* resistance in tobacco and cotton. *Plant Biotechnol. J.*, **16**, 638–648.
61. de Jonge, R., Peter van Esse, H., Maruthachalam, K., Bolton, M.D., Santhanam, P., Saber, M.K., 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.*, **109**, 5110–5115.
62. Klosterman, S.J., Subbarao, K. V., Kang, S., Veronese, P., Gold, S.E., Thomma, B.P.H.J., Chen, Z., Henrissat, B., Lee, Y.H., Park, J., *et al.* (2011) Comparative genomics yields insights into niche adaptation of plant vascular wilt pathogens. *PLoS Pathog.*, **7**, e1002137.
63. Depotter, J.R.L., Shi-Kunne, X., Missonnier, H., Liu, T., Faino, L., Berg, G.C.M., Wood, T.A., Zhang, B., Jacques, A., Seidl, M.F., *et al.* (2019) Dynamic virulence-related regions of the plant pathogenic fungus *Verticillium dahliae* display enhanced sequence conservation. *Mol. Ecol.*, 10.1111/mec.15168.
64. Sexton, T. and Cavalli, G. (2015) The role of chromosome domains in shaping the functional genome. *Cell*, **160**, 1049–1059.

65. David, K.T., Wilson, A.E. and Halanych, K.M. (2019) Sequencing disparity in the genomic era. *Mol. Biol. Evol.*, **36**, 1624–1627.

66. Thomma, B.P.H.J., Seidl, M.F., Shi-Kunne, X., Cook, D.E., Bolton, M.D., van Kan, J.A.L. and Faino, L. (2016) Mind the gap; seven reasons to close fragmented genome assemblies. *Fungal Genet. Biol.*, **90**, 24–30.

67. Seidl, M.F. and Thomma, B.P.H.J. (2014) Sex or no sex: evolutionary adaptation occurs regardless. *bioEssays*, **36**, 335–345.

68. Giraud, T., Gladieux, P. and Gavrilets, S. (2010) Linking the emergence of fungal plant diseases with ecological speciation. *Trends Ecol. Evol.*, **25**, 387–395.

69. Raffaele, S. and Kamoun, S. (2012) Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat. Rev. Microbiol.*, **10**, 417–430.

70. Möller, M. and Stukenbrock, E.H. (2017) Evolution and genome architecture in fungal plant pathogens. *Nat. Rev. Microbiol.*, **15**, 756–771.

71. Oliveira-Garcia, E. and Valent, B. (2015) How eukaryotic filamentous pathogens evade plant recognition. *Curr. Opin. Microbiol.*, **26**, 92–101.

72. Couto, D. and Zipfel, C. (2016) Regulation of pattern recognition receptor signalling in plants. *Nat. Rev. Immunol.*, **16**, 537–552.

73. Liang, X. and Zhou, J.-M. (2018) Receptor-like cytoplasmic kinases: central players in plant receptor kinase–mediated signaling. *Annu. Rev. Plant Biol.*, **69**, 267–299.

74. Faino, L., Seidl, M.F., Shi-Kunne, X., Pauper, M., Van Den Berg, G.C.M., Wittenberg, A.H.J. and Thomma, B.P.H.J. (2016) Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen. *Genome Res.*, **26**, 1091–1100.

75. Shi-Kunne, X., Faino, L., van den Berg, G.C.M., Thomma, B.P.H.J. and Seidl, M.F. (2018) Evolution within the fungal genus *Verticillium* is characterized by chromosomal rearrangement and gene loss. *Environ. Microbiol.*, **20**, 1362–1373.

76. Kombrink, A., Rovenich, H., Shi-Kunne, X., Rojas-Padilla, E., van den Berg, G.C.M., Domazakis, E., De Jonge, R., Valkenburg, D., Sánchez-Vallet, A., Seidl, M.F., *et al.* (2017) *Verticillium dahliae* LysM effectors differentially contribute to virulence on plant hosts. *Mol. Plant Pathol.*, **18**, 596–608.

77. Ma, L.-J.J., van der Does, H.C., Borkovich, K.A., Coleman, J.J., Daboussi, M.-J., Di Pietro, A., Dufresne, M., Freitag, M., Grabherr, M., Henrissat, B., *et al.* (2010) Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature*, **464**, 367–373.

78. Rouxel, T., Grandaubert, J., Hane, J.K., Hoede, C., van de Wouw, A.P., 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**, doi: 10.1038/ncomms1189.

79. Dutheil, J.Y., Mannhaupt, G., Schweizer, G., MK Sieber, C., Münsterkötter, M., Güldener, U., Schirawski, J. and Kahmann, R. (2016) A tale of genome compartmentalization: the evolution of virulence clusters in smut fungi. *Genome Biol. Evol.*, **8**, 681–704.

80. Peng, Z., Oliveira-Garcia, E., Lin, G., Hu, Y., Dalby, M., Migeon, P., Tang, H., Farman, M., Cook, D. and White, F.F. (2019) Effector gene reshuffling involves dispensable mini-chromosomes in the wheat blast fungus. *PLoS Genet.*, **15**, e1008272.

81. Tsushima, A., Gan, P., Kumakura, N., Narusaka, M., Takano, Y., Narusaka, Y. and Shirasu, K. (2019) Genomic plasticity mediated by transposable elements in the plant pathogenic fungus *Colletotrichum higginsianum*. *Genome Biol. Evol.*, **11**, 1487–1500.

82. Raffaele, S., Farrer, R.A., Cano, L.M., Studholme, D.J., MacLean, D., Thines, M., Jiang, R.H.Y., Zody, M.C., Kunjeti, S.G., Donofrio, N.M., *et al.* (2010) Genome evolution following host jumps in the irish potato famine pathogen lineage. *Science*, **330**, 1540–1543.

83. Goodwin, S.B., M'Barek, S.B., Dhillon, B., Wittenberg, A.H.J., Crane, C.F., Hane, J.K., Foster, A.J., van der Lee, T.A.J., 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**, doi:10.1371/journal.pgen.1002070.

84. Dong, S., Raffaele, S. and Kamoun, S. (2015) The two-speed genomes of filamentous pathogens: Waltz with plants. *Curr. Opin. Genet. Dev.*, **35**, 57–65.

85. Macheleidt, J., Mattern, D.J., Fischer, J., Netzker, T., Weber, J., Schroeckh, V., Valiante, V. and Brakhage, A.A. (2016) Regulation and role of fungal secondary metabolites. *Annu. Rev. Genet.*, **50**, 371–392.

86. Riddle, N.C., Minoda, A., Kharchenko, P. V, Alekseyenko, A.A., Schwartz, Y.B., Tolstorukov, M.Y., Gorchakov, A.A., Jaffe, J.D., Kennedy, C. and Linder-Basso, D. (2011) Plasticity in patterns of histone modifications and chromosomal proteins in *Drosophila* heterochromatin. *Genome Res.*, **21**, 147–163.

87. Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A. and Cavalli, G. (2012) Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell*, **148**, 458–472.

88. Rea, S., Eisenhaber, F., O'Carroll, D., Strahl, B.D., Sun, Z.-W., Schmid, M., Opravil, S., Mechtler, K., Ponting, C.P. and Allis, C.D. (2000) Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature*, **406**, 593–599.

89. Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R.S. and Zhang, Y. (2002) Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science*, **298**, 1039–1043.

90. Margueron, R. and Reinberg, D. (2011) The Polycomb complex PRC2 and its mark in life. *Nature*, **469**, 343–349.

91. Janssen, A., Colmenares, S.U. and Karpen, G.H. (2018) Heterochromatin: guardian of the genome. *Annu. Rev. Cell Dev. Biol.*, **34**, 265–288.

92. Tamaru, H. and Selker, E.U. (2001) A histone H3 methyltransferase controls DNA methylation in *Neurospora crassa*. *Nature*, **414**, 277–283.

93. Kouzminova, E. and Selker, E.U. (2001) Dim-2 encodes a DNA methyltransferase responsible for all known cytosine methylation in *Neurospora*. *EMBO J.*, **20**, 4309–4323.

94. Freitag, M., Hickey, P.C., Khlafallah, T.K., Read, N.D. and Selker, E.U. (2004) HP1 is essential for DNA methylation in *Neurospora*. *Mol. Cell*, **13**, 427–434.

95. Honda, S. and Selker, E.U. (2008) Direct interaction between DNA methyltransferase DIM-2 and HP1 is required for DNA methylation in *Neurospora crassa*. *Mol. Cell. Biol.*, **28**, 6044–6055.

96. Freitag, M., Williams, R.L., Kothe, G.O. and Selker, E.U. (2002) A cytosine methyltransferase homologue is essential for repeat-induced point mutation in *Neurospora crassa*. *Proc. Natl. Acad. Sci.*, **99**, 8802–8807.

97. Lewis, Z.A., Honda, S., Khlafallah, T.K., Jeffress, J.K., Freitag, M., Mohn, F., Schübeler, D. and Selker, E.U. (2009) Relics of repeat-induced point mutation direct heterochromatin formation in *Neurospora crassa*. *Genome Res.*, **19**, 427–437.

98. Selker, E.U., Tountas, N.A., Cross, S.H., Margolin, B.S., Murphy, J.G., Bird, A.P. and Freitag, M. (2003) The methylated component of the *Neurospora crassa* genome. *Nature*, **422**, 893–897.

99. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R. and Coyne, M. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.

100. Bemer, M. and Grossniklaus, U. (2012) Dynamic regulation of Polycomb group activity during plant development. *Curr. Opin. Plant Biol.*, **15**, 523–529.

101. Gaydos, L.J., Wang, W. and Strome, S. (2014) H3K27me and PRC2 transmit a memory of repression across generations and during development. *Science*, **345**, 1515–1518.

102. Dattani, A., Kao, D., Mihaylova, Y., Abnave, P., Hughes, S., Lai, A., Sahu, S. and Aboobaker, A.A. (2018) Epigenetic analyses of planarian stem cells demonstrate conservation of bivalent histone modifications in animal stem cells. *Genome Res.*, **28**, 1543–1554.

103. Freitag, M. (2017) Histone methylation by SET domain proteins in fungi. *Annu. Rev. Microbiol.*, **71**, 413–439.

104. Seidl, M.F., Cook, D.E. and Thomma, B.P.H.J. (2016) Chromatin biology impacts adaptive evolution of filamentous plant pathogens. *PLoS Pathog.*, **12**, e1005920.

105. Galazka, J.M., Klocko, A.D., Uesaka, M., Honda, S., Selker, E.U. and Freitag, M. (2016) *Neurospora* chromosomes are organized by blocks of importin alpha-dependent heterochromatin that are largely independent of H3K9me3. *Genome Res.*, **26**, 1069–1080.

106. Klocko, A.D., Ormsby, T., Galazka, J.M., Leggett, N.A., Uesaka, M., Honda, S., Freitag, M. and Selker, E.U. (2016) Normal chromosome conformation depends on subtelomeric facultative heterochromatin in *Neurospora crassa*. *Proc. Natl. Acad. Sci.*, **113**, 15048–15053.

107. Chen, H., Shu, H., Wang, L., Zhang, F., Li, X., Ochola, S.O., Mao, F., Ma, H., Ye, W. and Gu, T. (2018) *Phytophthora* methylomes are modulated by 6mA methyltransferases and associated with adaptive genome regions. *Genome Biol.*, **19**, 1–16.

108. Clark, S.J., Harrison, J., Paul, C.L. and Frommer, M. (1994) High sensitivity mapping of methylated cytosines. *Nucleic Acids Res.*, **22**, 2990–2997.

109. Lister, R. and Ecker, J.R. (2009) Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome Res.*, **19**, 959–966.

110. Montanini, B., Chen, P.-Y., Morselli, M., Jaroszewicz, A., Lopez, D., Martin, F., Ottonello, S. and Pellegrini, M. (2014) Non-exhaustive DNA methylation-mediated transposon silencing in the black truffle genome, a complex fungal genome with massive repeat element content. *Genome Biol.*, **15**, 411.

111. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. and Greenleaf, W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.

112. Bradley, A.P. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.*, **30**, 1145–1159.

113. Davis, J. and Goadrich, M. (2006) The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*.pp. 233–240.

114. Saito, T. and Rehmsmeier, M. (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*, **10**, e0118432.

115. McInnes, L., Healy, J. and Melville, J. (2018) Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv*.

116. Mondo, S.J., Dannebaum, R.O., Kuo, R.C., Louie, K.B., Bewick, A.J., LaButti, K., Haridas, S., Kuo, A., Salamov, A., Ahrendt, S.R., *et al.* (2017) Widespread adenine N6-methylation of active genes in fungi. *Nat. Genet.*, **49**, 964–968.

117. Bannister, A.J., Zegerman, P., Partridge, J.F., Miska, E.A., Thomas, J.O., Allshire, R.C. and Kouzarides, T. (2001) Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature*, **410**, 120–124.

118. Cam, H.P., Sugiyama, T., Chen, E.S., Chen, X., FitzGerald, P.C. and Grewal, S.I.S. (2005) Comprehensive analysis of heterochromatin-and RNAi-mediated epigenetic control of the fission yeast genome. *Nat. Genet.*, **37**, 809–819.

119. Huisinga, K.L., Brower-Toland, B. and Elgin, S.C.R. (2006) The contradictory definitions of heterochromatin: transcription and silencing. *Chromosoma*, **115**, 110–122.

120. Möller, M., Habig, M., Freitag, M. and Stukenbrock, E.H. (2018) Extraordinary genome instability and widespread chromosome rearrangements during vegetative growth. *Genetics*, **210**, 517–529.

121. Schotanus, K., Soyer, J.L., Connolly, L.R., Grandaubert, J., Happel, P., Smith, K.M., Freitag, M. and Stukenbrock, E.H. (2015) Histone modifications rather than the novel regional centromeres of *Zymoseptoria tritici* distinguish core and accessory chromosomes. *Epigenetics Chromatin*, **8**, 41.

122. Faino, L., Seidl, M., Datema, E., van den Berg, G.C.M., Janssen, A., Wittenberg, A.H.J. and Thomma, B.P.H.J. (2015) Single-molecule real-time sequencing combined with optical mapping yields completely finished fungal genome. *mBio*, **6**, e00936-15.

123. Santhanam, P. (2012) Random insertional mutagenesis in fungal genomes to identify virulence factors. In *Plant fungal pathogens*. Springer, pp. 509–517.

124. Frandsen, R.J.N., Andersson, J.A., Kristensen, M.B. and Giese, H. (2008) Efficient four fragment cloning for the construction of vectors for targeted gene replacement in filamentous fungi. *BMC Mol. Biol.*, **9**, 70.

125. Xi, Y. and Li, W. (2009) BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, **10**, 232.

126. Schultz, M.D., Schmitz, R.J. and Ecker, J.R. (2012) 'Leveling' the playing field for analyses of single-base resolution DNA methylomes. *Trends Genet.*, **28**, 583–585.

127. Team, R.C. (2013) R: A language and environment for statistical computing.

128. Lê, S., Josse, J. and Husson, F. (2008) FactoMineR: an R package for multivariate analysis. *J. Stat. Softw.*, **25**, 1–18.

129. Kassambara, A. and Mundt, F. (2017) Factoextra: extract and visualize the results of multivariate data analyses. *R Packag. version*, **1**, 337–354.

130. Ellinghaus, D., Kurtz, S. and Willhoeft, U. (2008) LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics*, **9**, 1–14.

131. Steinbiss, S., Willhoeft, U., Gremme, G. and Kurtz, S. (2009) Fine-grained annotation and classification of *de novo* predicted LTR retrotransposons. *Nucleic Acids Res.*, **37**, 7002–7013.

132. Smit, A.F.A. and Hubley, R. (2015) 2015 RepeatModeler Open-1.0. *Repeat Masker Website http//www. repeatmasker. org*.

133. Campbell, M.S., Law, M., Holt, C., Stein, J.C., Moghe, G.D., Hufnagel, D.E., Lei, J., Achawanantakun, R., Jiao, D. and Lawrence, C.J. (2014) MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.*, **164**, 513–524.

134. Rognes, T., Flouri, T., Nichols, B., Quince, C. and Mahé, F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, **4**, e2584.

135. Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.

136. Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M. and Panaud, O. (2007) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.*, **8**, 973–982.

137. Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V. and Quesneville, H. (2014) PASTEC: an automatic transposable element classification tool. *PLoS One*, **9**, e91929.

138. Flutre, T., Duprat, E., Feuillet, C. and Quesneville, H. (2011) Considering transposable element diversification in *de novo* annotation approaches. *PLoS One*, **6**, e16526.

139. Amyotte, S.G., Tan, X., Pennerman, K., del Mar Jimenez-Gasco, M., Klosterman, S.J., Ma, L.J., Dobinson, K.F. and Veronese, P. (2012) Transposable elements in phytopathogenic *Verticillium* spp.: insights into genome evolution and inter- and intra-specific diversification. *BMC Genomics*, **13**, 314.

140. Bailly-Bechet, M., Haudry, A. and Lerat, E. (2014) "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. *Mob. DNA*, **5**, 1–15.

141. Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet.*, **16**, 276–277.

142. Jukes, T.H. and Cantor, C.R. (1969) Evolution of protein molecules. *Mamm. protein Metab.*, **3**, 21–132.

143. Van de Peer, Y., Neefs, J.-M. and De Wachter, R. (1990) Small ribosomal subunit RNA sequences, evolutionary relationships among different life forms, and mitochondrial origins. *J. Mol. Evol.*, **30**, 463–476.

144. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

145. Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T. and Carey, V.J. (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, e1003118.

146. Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

147. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

148. Wagner, G.P., Kin, K. and Lynch, V.J. (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.*, **131**, 281–285.

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

150. Amemiya, H.M., Kundaje, A. and Boyle, A.P. (2019) The ENCODE blacklist: identification of problematic regions of the genome. *Sci. Rep.*, **9**, 1–5.

151. Buenrostro, J.D., Wu, B., Chang, H.Y. and Greenleaf, W.J. (2015) ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.*, **2015**, doi: 10.1002/0471142727.mb2129s109.

152. Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997*.

153. Kuhn, M. (2008) Building predictive models in R using the caret package. *J. Stat. Softw.*, **28**, 1–26.

154. Grau, J., Grosse, I. and Keilwagen, J. (2015) PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics*, **31**, 2595–2597.

155. Bray, N.L., Pimentel, H., Melsted, P. and Pachter, L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.

156. Armenteros, J.J.A., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G. and Nielsen, H. (2019) SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.*, **37**, 420–423.

157. Sperschneider, J., Dodds, P.N., Gardiner, D.M., Singh, K.B. and Taylor, J.M. (2018) Improved prediction of fungal effector proteins from secretomes with EffectorP 2.0. *Mol. Plant Pathol.*, **19**, 2094–2110.

158. Hunter, J.D. (2007) Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95.

159. Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D.C., Augspurger, T., Halchenko, Y., Cole, J.B. and Warmenhoven, J. (2017) Mwaskom/Seaborn: V0. 8.1 (September 2017). *Zenodo*.

160. McKinney, W. (2010) Data structures for statistical computing in python. In *Proceedings of the 9th Python in science conference*. Austin, TX, Vol. 445, pp. 51–56.

161. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V. (2011) Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

162. van der Walt, S., Colbert, S.C. and Varoquaux, G. (2011) The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.*, **13**, 22–30.

163. Terpilowski, M.A. (2019) scikit-posthocs: pairwise multiple comparison tests in Python. *J. open source Softw.*, **4**, 1169.

164. Schmit, J.P. and Mueller, G.M. (2007) An estimate of the lower limit of global fungal diversity. *Biodivers. Conserv.*, **16**, 99–111.

165. Gibson, G. (2008) The environmental contribution to gene expression profiles. *Nat. Rev. Genet.*, **9**, 575–581.

166. Selvig, K. and Alspaugh, J.A. (2011) pH response pathways in fungi: adapting to host-derived and environmental signals. *Mycobiology*, **39**, 249–256.

167. Bahn, Y.-S., Xue, C., Idnurm, A., Rutherford, J.C., Heitman, J. and Cardenas, M.E. (2007) Sensing the environment: lessons from fungi. *Nat. Rev. Microbiol.*, **5**, 57–69.

168. Aguilera, J., Randez-Gil, F. and Prieto, J.A. (2007) Cold response in *Saccharomyces cerevisiae*: new functions for old mechanisms. *FEMS Microbiol. Rev.*, **31**, 327–341.

169. Schneper, L., Düvel, K. and Broach, J.R. (2004) Sense and sensibility: nutritional response and signal integration in yeast. *Curr. Opin. Microbiol.*, **7**, 624–630.

170. Heijmans, B.T., Tobi, E.W., Stein, A.D., Putter, H., Blauw, G.J., Susser, E.S., Slagboom, P.E. and Lumey, L.H. (2008) Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc. Natl. Acad. Sci.*, **105**, 17046–17049.

171. Zannas, A.S., Arloth, J., Carrillo-Roa, T., Iurato, S., Röh, S., Ressler, K.J., Nemeroff, C.B., Smith, A.K., Bradley, B. and Heim, C. (2015) Lifetime stress accelerates epigenetic aging in an urban, African American cohort: relevance of glucocorticoid signaling. *Genome Biol.*, **16**, 1–12.

172. Martin, E.M. and Fry, R.C. (2018) Environmental influences on the epigenome: exposure-associated DNA methylation in human populations. *Annu. Rev. Public Health*, **39**, 309–333.

173. Kronholm, I., Johannesson, H. and Ketola, T. (2016) Epigenetic control of phenotypic plasticity in the filamentous fungus *Neurospora crassa*. *G3 Genes, Genomes, Genet.*, **6**, 4009–4022.

174. Slepecky, R.A. and Starmer, W.T. (2009) Phenotypic plasticity in fungi: a review with observations on *Aureobasidium pullulans*. *Mycologia*, **101**, 823–832.

175. Herrera, C.M., Pozo, M.I. and Bazaga, P. (2012) Jack of all nectars, master of most: DNA methylation and the epigenetic basis of niche width in a flower-living yeast. *Mol. Ecol.*, **21**, 2602–2616.

176. Slotkin, R.K. and Martienssen, R. (2007) Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.*, **8**, 272–285.

177. Bannister, A.J. and Kouzarides, T. (2011) Regulation of chromatin by histone modifications. *Cell Res.*, **21**, 381–395.

178. Trojer, P. and Reinberg, D. (2007) Facultative heterochromatin: is there a distinctive molecular signature? *Mol. Cell*, **28**, 1–13.

179. Kim, J.-M., Sasaki, T., Ueda, M., Sako, K. and Seki, M. (2015) Chromatin changes in response to drought, salinity, heat, and cold stresses in plants. *Front. Plant Sci.*, **6**, doi: 10.3389/fpls.2015.00114.

180. Wang, J., Jia, S.T. and Jia, S. (2016) New insights into the regulation of heterochromatin. *Trends Genet.*, **32**, 284–294.

181. Lewis, Z.A., Adhvaryu, K.K., Honda, S., Shiver, A.L., Knip, M., Sack, R. and Selker, E.U. (2010) DNA methylation and normal chromosome behavior in *Neurospora* depend on five components of a histone methyltransferase complex, DCDC. *PLoS Genet*, **6**, e1001196.

182. Jamieson, K., Rountree, M.R., Lewis, Z.A., Stajich, J.E. and Selker, E.U. (2013) Regional control of histone H3 lysine 27 methylation in *Neurospora*. *Proc. Natl. Acad. Sci.*, **110**, 6027–32.

183. Soyer, J.L., Clairet, C., Gay, E.J., Lapalu, N., Rouxel, T., Stukenbrock, E.H. and Fudal, I. (2021) Genome-wide mapping of histone modifications during axenic growth in two species of *Leptosphaeria maculans* showing contrasting genomic organization. *Chromosom. Res.*, **29**, 219–236.

184. Wiemann, P., Sieber, C.M.K., von Bargen, K.W., Studt, L., Niehaus, E.M., Espino, J.J., Huß, K., Michielse, C.B., Albermann, S., Wagner, D., *et al.* (2013) Deciphering the cryptic genome: genome-wide analyses of the rice pathogen *Fusarium fujikuroi* reveal complex regulation of secondary metabolism and novel metabolites. *PLoS Pathog.*, **9**, e1003475.

185. Carlier, F., Li, M., Maroc, L., Debuchy, R., Souaid, C., Noordermeer, D., Grognet, P. and Malagnac, F. (2021) Loss of EZH2-like or SU (VAR) 3–9-like proteins causes simultaneous perturbations in H3K27 and H3K9 tri-methylation and associated developmental defects in the fungus *Podospora anserina*. *Epigenetics Chromatin*, **14**, 1–28.

186. Lukito, Y., Lee, K., Noorifar, N., Green, K.A., Winter, D.J., Ram, A., Hale, T.K., Chujo, T., Cox, M.P. and Johnson, L.J. (2021) Regulation of host-infection ability in the grass-symbiotic fungus *Epichloë festucae* by histone H3K9 and H3K36 methyltransferases. *Environ. Microbiol.*, **23**, 2116–2131.

187. Tian, L. and Chen, Z.J. (2001) Blocking histone deacetylation in *Arabidopsis* induces pleiotropic effects on plant gene regulation and development. *Proc. Natl. Acad. Sci.*, **98**, 200–205.

188. Fan, A., Mi, W., Liu, Z., Zeng, G., Zhang, P., Hu, Y., Fang, W. and Yin, W.-B. (2017) Deletion of a histone acetyltransferase leads to the pleiotropic activation of natural products in Metarhizium robertsii. *Org. Lett.*, **19**, 1686–1689.

189. Galazka, J.M. and Freitag, M. (2014) Variability of chromosome structure in pathogenic fungi-of 'ends and odds'. *Curr. Opin. Microbiol.*, **20**, 19–26.

190. Ridenour, J.B., Möller, M. and Freitag, M. (2020) Polycomb repression without bristles: facultative heterochromatin and genome stability in fungi. *Genes*, **11**, 638.

191. Cook, D.E., Kramer, H.M., Torres, D.E., Seidl, M.F. and Thomma, B.P.H.J. (2020) A unique chromatin profile defines adaptive genomic regions in a fungal plant pathogen. *eLife*, **9**, e62208.

192. Faino, L., de Jonge, R. and Thomma, B.P.H.J. (2012) The transcriptome of *Verticillium dahliae*-infected *Nicotiana benthamiana* determined by deep RNA sequencing. *Plant Signal. Behav.*, **7**, 1065–1069.

193. Soyer, J.L., Möller, M., Schotanus, K., Connolly, L.R., Galazka, J.M., Freitag, M. and Stukenbrock, E.H. (2015) Chromatin analyses of *Zymoseptoria tritici*: methods for chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq). *Fungal Genet. Biol.*, **79**, 63–70.

194. Seidl, M.F., Kramer, H.M., Cook, D.E., Lorencini Fiorin, G., van den Berg, G.C.M., Faino, L. and Thomma, B.P.H.J. (2020) Repetitive elements contribute to the diversity and evolution of centromeres in the fungal genus Verticillium. *mBio*, **11**, e01714-20.

195. Zhang, W., Huang, J. and Cook, D.E. (2021) Histone modification dynamics at H3K27 are associated with altered transcription of in planta induced genes in *Magnaporthe oryzae*. *PLoS Genet.*, **17**, e1009376.

196. Tang, G., Yuan, J., Wang, J., Zhang, Y.-Z., Xie, S.-S., Wang, H., Tao, Z., Liu, H., Kistler, H.C., Zhao, Y., *et al.* (2021) *Fusarium* BP1 is a reader of H3K27 methylation. *Nucleic Acids Res.*, **49**, 10448–10464.

197. Wiles, E.T., McNaught, K.J., Kaur, G., Selker, J.M.L., Ormsby, T., Aravind, L. and Selker, E.U. (2020) Evolutionarily ancient BAH–PHD protein mediates Polycomb silencing. *Proc. Natl. Acad. Sci.*, **117**, 11614–11623.

198. Basenko, E.Y., Sasaki, T., Ji, L., Prybol, C.J., Burckhardt, R.M., Schmitz, R.J. and Lewis, Z.A. (2015) Genome-wide redistribution of H3K27me3 is linked to genotoxic stress and defective growth. *Proc. Natl. Acad. Sci.*, **112**, E6339–E6348.

199. Lewis, Z.A. (2017) Polycomb group systems in fungi: new models for understanding polycomb repressive complex 2. *Trends Genet.*, **33**, 220–231.

200. Tralamazza, S.M., Abraham, L.N., Sarai Reyes-Avila, C., Corrêa, B. and Croll, D. (2021) Histone H3K27 methylation perturbs transcriptional robustness and underpins dispensability of highly conserved genes in fungi. *Mol. Biol. Evol.*

201. Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.*, **17**, 10–12.

202. Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A. and Manke, T. (2014) deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.*, **42**, W187–W191.

203. Zerbino, D.R., Johnson, N., Juettemann, T., Wilder, S.P. and Flicek, P. (2014) WiggleTools: parallel processing of large collections of genome-wide datasets for visualization and statistical analysis. *Bioinformatics*, **30**, 1008–1009.

204. Stovner, E.B. and Sætrom, P. (2019) Epic2 efficiently finds diffuse domains in ChIP-seq data. *Bioinformatics*, **35**, 4392–4393.

205. Huang, Q., Ma, C., Chen, L., Luo, D., Chen, R. and Liang, F. (2018) Mechanistic insights into the interaction between transcription factors and epigenetic modifications and the contribution to the development of obesity. *Front. Endocrinol.*, **9**, 10.3389/fendo.2018.00370.

206. Chen, Z.J. (2007) Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annu. Rev. Plant Biol.*, **58**, 377–406.

207. Burdge, G.C., Hanson, M.A., Slater-Jefferies, J.L. and Lillycrop, K.A. (2007) Epigenetic regulation of transcription: A mechanism for inducing variations in phenotype (fetal programming) by differences in nutrition during early life? *Br. J. Nutr.*, **97**, 1036–1046.

208. Bird, A. (1992) The essentials of DNA methylation. *Cell*, **70**, 5–8.

209. Law, J.A. and Jacobsen, S.E. (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.*, **11**, 204–220.

210. Schmitz, R.J., Lewis, Z.A. and Goll, M.G. (2019) DNA methylation: shared and divergent features across eukaryotes. *Trends Genet.*, **35**, 818–827.

211. Galagan, J.E., Henn, M.R., Ma, L.J., Cuomo, C.A. and Birren, B. (2005) Genomics of the fungal kingdom: Insights into eukaryotic biology. *Genome Res.*, **15**, 1620–1631.

212. Cambareri, E.B., Jensen, B.C., Schabtach, E. and Selker, E.U. (1989) Repeat-induced G-C to A-T mutations in *Neurospora*. *Science*, **244**, 1571–1575.

213. Jeon, J., Choi, J., Lee, G.-W., Park, S.-Y., Huh, A., Dean, R.A. and Lee, Y.-H. (2015) Genome-wide profiling of DNA methylation provides insights into epigenetic regulation of fungal development in a plant pathogenic fungus, *Magnaporthe oryzae*. *Sci. Rep.*, **5**, 8567.

214. Ikeda, K.I., Van Vu, B., Kadotani, N., Tanaka, M., Murata, T., Shiina, K., Chuma, I., Tosa, Y. and Nakayashiki, H. (2013) Is the fungus *Magnaporthe* losing DNA methylation? *Genetics*, **195**, 845–855.

215. Huff, J.T. and Zilberman, D. (2014) Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. *Cell*, **156**, 1286–1297.

216. Catania, S., Dumesic, P.A., Pimentel, H., Nasif, A., Stoddard, C.I., Burke, J.E., Diedrich, J.K., Cook, S., Shea, T. and Geinger, E. (2020) Evolutionary persistence of DNA methylation for millions of years after ancient loss of a *de novo* methyltransferase. *Cell*, **180**, 263–277.

217. Bewick, A.J., Hofmeister, B.T., Powers, R.A., Mondo, S.J., Grigoriev, I. V., James, T.Y., Stajich, J.E. and Schmitz, R.J. (2019) Diversity of cytosine methylation across the fungal tree of life. *Nat. Ecol. Evol.*, **3**, 479–490.

218. Möller, M., Habig, M., Lorrain, C., Feurtey, A., Haueisen, J., Fagundes, W.C., Alizadeh, A., Freitag, M. and Stukenbrock, E.H. (2021) Recent loss of the Dim2 DNA methyltransferase decreases mutation rate in repeats and changes evolutionary trajectory in a fungal pathogen. *PLOS Genet.*, **17**, e1009448.

219. Dowen, R.H., Pelizzola, M., Schmitz, R.J., Lister, R., Dowen, J.M., Nery, J.R., Dixon, J.E. and Ecker, J.R. (2012) Widespread dynamic DNA methylation in response to biotic stress. *Proc. Natl. Acad. Sci.*, **109**, e2183-91.

220. Meaney, M.J. and Szyf, M. (2005) Environmental programming of stress responses through DNA methylation: life at the interface between a dynamic environment and a fixed genome. *Dialogues Clin. Neurosci.*, **7**, 103–123.

221. Wang, Y., Wang, Z., Liu, C., Wang, S. and Huang, B. (2015) Genome-wide analysis of DNA methylation in the sexual stage of the insect pathogenic fungus *Cordyceps militaris*. *Fungal Biol.*, **119**, 1246–1254.

222. Zhang, X., Liu, X., Zhao, Y., Cheng, J., Xie, J., Fu, Y., Jiang, D. and Chen, T. (2016) Histone H3 lysine 9 methyltransferase DIM5 is required for the development and virulence of *Botrytis cinerea*. *Front. Microbiol.*, 10.3389/fmicb.2016.01289.

223. Haueisen, J., Möller, M., Eschenbrenner, C.J., Grandaubert, J., Seybold, H., Adamiak, H. and Stukenbrock, E.H. (2019) Highly flexible infection programs in a specialized wheat pathogen. *Ecol. Evol.*, **9**, 275–294.

224. Hayashi, A., Ishida, M., Kawaguchi, R., Urano, T., Murakami, Y. and Nakayama, J. (2012) Heterochromatin protein 1 homologue Swi6 acts in concert with Ers1 to regulate RNAi-directed heterochromatin assembly. *Proc. Natl. Acad. Sci.*, **109**, 6159–6164.

225. Gessaman, J.D. and Selker, E.U. (2017) Induction of H3K9me3 and DNA methylation by tethered heterochromatin factors in *Neurospora crassa*. *Proc. Natl. Acad. Sci.*, **114**, E9598–E9607.

226. Holliday, R. and Grigg, G.W. (1993) DNA methylation and mutation. *Mutat. Res.*, **285**, 61–67.

227. Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A. and Mason, C.E. (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.*, **13**, R87.

228. Gel, B. and Serra, E. (2017) karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics*, **33**, 3088–3090.

229. Jin, Y., Tam, O.H., Paniagua, E. and Hammell, M. (2015) TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics*, **31**, 3593–3599.

230. Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

231. Bauer, S., Grossmann, S., Vingron, M. and Robinson, P.N. (2008) Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*, **24**, 1650–1651.

232. Pignatelli, M., Serras, F., Moya, A., Guigó, R. and Corominas, M. (2009) CROC: finding chromosomal clusters in eukaryotic genomes. *Bioinformatics*, **25**, 1552–1553.

233. Roy, B. and Sanyal, K. (2011) Diversity in requirement of genetic and epigenetic factors for centromere function in fungi. *Eukaryot. Cell*, **10**, 1384–1395.

234. Foley, E.A. and Kapoor, T.M. (2013) Microtubule attachment and spindle assembly checkpoint signalling at the kinetochore. *Nat. Rev. Mol. Cell Biol.*, **14**, 25–37.

235. Burrack, L.S. and Berman, J. (2012) Flexibility of centromere and kinetochore structures. *Trends Genet.*, **28**, 204–212.

236. Janssen, A., van der Burg, M., Szuhai, K., Kops, G.J.P.L. and Medema, R.H. (2011) Chromosome segregation errors as a cause of DNA damage and structural chromosome aberrations. *Science*, **333**, 1895–1898.

237. Sheltzer, J.M., Blank, H.M., Pfau, S.J., Tange, Y., George, B.M., Humpton, T.J., Brito, I.L., Hiraoka, Y., Niwa, O. and Amon, A. (2011) Aneuploidy drives genomic instability in yeast. *Science*, **333**, 1026–1030.

238. Barra, V. and Fachinetti, D. (2018) The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA. *Nat. Commun.*, **9**, 1–17.

239. Rancati, G., Pavelka, N., Fleharty, B., Noll, A., Trimble, R., Walton, K., Perera, A., Staehling-Hampton, K., Seidel, C.W. and Li, R. (2008) Aneuploidy underlies rapid adaptive evolution of yeast cells deprived of a conserved cytokinesis motor. *Cell*, **135**, 879–893.

240. Pavelka, N., Rancati, G., Zhu, J., Bradford, W.D., Saraf, A., Florens, L., Sanderson, B.W., Hattem, G.L. and Li, R. (2010) Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast. *Nature*, **468**, 321–325.

241. Sankaranarayanan, S.R., Ianiri, G., Coelho, M.A., Reza, M.H., Thimmappa, B.C., Ganguly, P., Vadnala, R.N., Sun, S., Siddharthan, R. and Tellgren-Roth, C. (2020) Loss of centromere function drives karyotype evolution in closely related *Malassezia* species. *eLife*, **9**, e53944.

242. Ola, M., O'Brien, C.E., Coughlan, A.Y., Ma, Q., Donovan, P.D., Wolfe, K.H. and Butler, G. (2020) Polymorphic centromere locations in the pathogenic yeast *Candida parapsilosis*. *Genome Res.*, **30**, 684–696.

243. Yadav, V., Sreekumar, L., Guin, K. and Sanyal, K. (2018) Five pillars of centromeric chromatin in fungal pathogens. *PLoS Pathog.*, **14**, e1007150.

244. Henikoff, S., Ahmad, K. and Malik, H.S. (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. *Science*, **293**, 1098–1102.

245. Smith, K.M., Galazka, J.M., Phatale, P.A., Connolly, L.R. and Freitag, M. (2012) Centromeres of filamentous fungi. *Chromosom. Res.*, **20**, 635–656.

246. Yadav, V., Sun, S., Billmyre, R.B., Thimmappa, B.C., Shea, T., Lintner, R., Bakkeren, G., Cuomo, C.A., Heitman, J. and Sanyal, K. (2018) RNAi is a critical determinant of centromere evolution in closely related fungi. *Proc. Natl. Acad. Sci.*, **115**, 3108–3113.

247. Fitzgerald-Hayes, M., Clarke, L. and Carbon, J. (1982) Nucleotide sequence comparisons and functional analysis of yeast centromere DNAs. *Cell*, **29**, 235–244.

248. Furuyama, S. and Biggins, S. (2007) Centromere identity is specified by a single centromeric nucleosome in budding yeast. *Proc. Natl. Acad. Sci.*, **104**, 14706–14711.

249. Krassovsky, K., Henikoff, J.G. and Henikoff, S. (2012) Tripartite organization of centromeric chromatin in budding yeast. *Proc. Natl. Acad. Sci.*, **109**, 243–248.

250. Cliften, P.F., Fulton, R.S., Wilson, R.K. and Johnston, M. (2006) After the duplication: gene loss and adaptation in *Saccharomyces* genomes. *Genetics*, **172**, 863–872.

251. Baum, M., Sanyal, K., Mishra, P.K., Thaler, N. and Carbon, J. (2006) Formation of functional centromeric chromatin is specified epigenetically in *Candida albicans*. *Proc. Natl. Acad. Sci.*, **103**, 14877–14882.

252. Sanyal, K., Baum, M. and Carbon, J. (2004) Centromeric DNA sequences in the pathogenic yeast *Candida albicans* are all different and unique. *Proc. Natl. Acad. Sci.*, **101**, 11374–11379.

253. Padmanabhan, S., Thakur, J., Siddharthan, R. and Sanyal, K. (2008) Rapid evolution of Cse4p-rich centromeric DNA sequences in closely related pathogenic yeasts, *Candida albicans* and *Candida dubliniensis*. *Proc. Natl. Acad. Sci.*, **105**, 19797–19802.

254. Cambareri, E.B., Aisner, R. and Carbon, J. (1998) Structure of the chromosome VII centromere region in *Neurospora crassa*: degenerate transposons and simple repeats. *Mol. Cell. Biol.*, **18**, 5465–5477.

255. Smith, K.M., Phatale, P.A., Sullivan, C.M., Pomraning, K.R. and Freitag, M. (2011) Heterochromatin is required for normal distribution of *Neurospora crassa* CenH3. *Mol. Cell. Biol.*, **31**, 2528–2542.

256. Selker, E.U. (2002) Repeat-induced gene silencing in fungi. *Adv. Genet.*, **46**, 439–450.

257. Yadav, V., Yang, F., Reza, M.H., Liu, S., Valent, B., Sanyal, K. and Naqvi, N.I. (2019) Cellular dynamics and genomic identity of centromeres in cereal blast fungus. *mBio*, **10**, e01581-19.

258. Fang, Y., Coelho, M.A., Shu, H., Schotanus, K., Thimmappa, B.C., Yadav, V., Chen, H., Malc, E.P., Wang, J. and Mieczkowski, P.A. (2020) Long transposon-rich centromeres in an oomycete reveal divergence of centromere features in Stramenopila-Alveolata-Rhizaria lineages. *PLoS Genet.*, **16**, e1008646.

259. Navarro-Mendoza, M.I., Pérez-Arques, C., Panchal, S., Nicolás, F.E., Mondo, S.J., Ganguly, P., Pangilinan, J., Grigoriev, I. V, Heitman, J. and Sanyal, K. (2019) Early diverging fungus *Mucor circinelloides* lacks centromeric histone CENP-A and displays a mosaic of point and regional centromeres. *Curr. Biol.*, **29**, 3791–3802.

260. Inderbitzin, P., Bostock, R.M., Davis, R.M., Usami, T., Platt, H.W. and Subbarao, K. V (2011) Phylogenetics and taxonomy of the fungal vascular wilt pathogen *Verticillium*, with the descriptions of five new species. *PLoS One*, **6**, e28341.

261. Depotter, J.R.L., van Beveren, F., Rodriguez-Moreno, L., Kramer, H.M., Chavarro Carrero, E.A., van den Berg, G.C.M., Wood, T.A., Thomma, B.P.H.J. and Seidl, M.F. (2021) The interspecific fungal hybrid *Verticillium longisporum* displays subgenome-specific gene expression. *mBio*, **12**, e0149621.

262. Sun, S., Yadav, V., Billmyre, R.B., Cuomo, C.A., Nowrousian, M., Wang, L., Souciet, J.-L., Boekhout, T., Porcel, B. and Wincker, P. (2017) Fungal genome and mating system transitions facilitated by chromosomal translocations involving intercentromeric recombination. *PLoS Biol.*, **15**, e2002527.

263. Depotter, J.R.L., Seidl, M.F., van den Berg, G.C.M., Thomma, B.P.H.J. and Wood, T.A. (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.

264. Liu, S.-Y., Lin, J.-Q., Wu, H.-L., Wang, C.-C., Huang, S.-J., Luo, Y.-F., Sun, J.-H., Zhou, J.-X., Yan, S.-J. and He, J.-G. (2012) Bisulfite sequencing reveals that *Aspergillus flavus* holds a hollow in DNA methylation. *PLoS One*, **7**, e30349.

265. Seymour, M., Ji, L., Santos, A.M., Kamei, M., Sasaki, T., Basenko, E.Y., Schmitz, R.J., Zhang, X. and Lewis, Z.A. (2016) Histone H1 limits DNA methylation in *Neurospora crassa*. *G3 Genes, Genomes, Genet.*, **6**, 1879–1889.

266. Kursel, L.E. and Malik, H.S. (2016) Centromeres. *Curr. Biol.*, **26**, R487–R490.

267. Friedman, S. and Freitag, M. (2017) Evolving centromeres and kinetochores. *Adv. Genet.*, **98**, 1–41.

268. Winter, D.J., Ganley, A.R.D., Young, C.A., Liachko, I., Schardl, C.L., Dupont, P.Y., Berry, D., Ram, A., Scott, B. and Cox, M.P. (2018) Repeat elements organise 3D genome structure and mediate transcription in the filamentous fungus *Epichloë festucae*. *PLoS Genet.*, **14**, e1007467.

269. Marie-Nelly, H., Marbouty, M., Cournac, A., Flot, J.-F., Liti, G., Parodi, D.P., Syan, S., Guillén, N., Margeot, A. and Zimmer, C. (2014) High-quality genome (re) assembly using chromosomal contact data. *Nat. Commun.*, **5**, 1–10.

270. Mizuguchi, T., Fudenberg, G., Mehta, S., Belton, J.-M., Taneja, N., Folco, H.D., FitzGerald, P., Dekker, J., Mirny, L. and Barrowman, J. (2014) Cohesin-dependent globules and heterochromatin shape 3D genome architecture in *S. pombe*. *Nature*, **516**, 432–435.

271. Varoquaux, N., Liachko, I., Ay, F., Burton, J.N., Shendure, J., Dunham, M.J., Vert, J.-P. and Noble, W.S. (2015) Accurate identification of centromere locations in yeast genomes using Hi-C. *Nucleic Acids Res.*, **43**, 5331–5339.

272. Vakirlis, N., Sarilar, V., Drillon, G., Fleiss, A., Agier, N., Meyniel, J.-P., Blanpain, L., Carbone, A., Devillers, H. and Dubois, K. (2016) Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus. *Genome Res.*, **26**, 918–932.

273. Depotter, J.R.L., Deketelaere, S., Inderbitzin, P., Tiedemann, A. Von, Höfte, M., Subbarao, K. V, Wood, T.A. and Thomma, B.P.H.J. (2016) *Verticillium longisporum*, the invisible threat to oilseed rape and other brassicaceous plant hosts. *Mol. Plant Pathol.*, **17**, 1004–1016.

274. Muller, H., Gil Jr, J. and Drinnenberg, I.A. (2019) The impact of centromeres on spatial genome architecture. *Trends Genet.*, **35**, 565–578.

275. Treangen, T.J. and Salzberg, S.L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.*, **13**, 36–46.

276. Seidl, M.F., Faino, L., Shi-Kunne, X., van den Berg, G.C.M., Bolton, M.D. and Thomma, B.P.H.J. (2015) The genome of the saprophytic fungus *Verticillium tricorpus* reveals a complex effector repertoire resembling that of its pathogenic relatives. *Genomics*, **28**, 362–373.

277. International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**, 695–716.

278. Fukagawa, T. (2017) Critical histone post-translational modifications for centromere function and propagation. *Cell cycle*, **16**, 1259–1265.

279. Volpe, T.A., Kidner, C., Hall, I.M., Teng, G., Grewal, S.I.S. and Martienssen, R.A. (2002) Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science*, **297**, 1833–1837.

280. Li, X., Wang, X., He, K., Ma, Y., Su, N., He, H., Stolc, V., Tongprasit, W., Jin, W. and Jiang, J. (2008) High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression. *Plant Cell*, **20**, 259–276.

281. Blower, M.D., Sullivan, B.A. and Karpen, G.H. (2002) Conserved organization of centromeric chromatin in flies and humans. *Dev. Cell*, **2**, 319–330.

282. Lynch, D.B., Logue, M.E., Butler, G. and Wolfe, K.H. (2010) Chromosomal G+C content evolution in yeasts: systematic interspecies differences, and GC-poor troughs at centromeres. *Genome Biol. Evol.*, **2**, 572–583.

283. Hanson, S.J., Byrne, K.P. and Wolfe, K.H. (2014) Mating-type switching by chromosomal inversion in methylotrophic yeasts suggests an origin for the three-locus *Saccharomyces cerevisiae* system. *Proc. Natl. Acad. Sci.*, **111**, E4851–E4858.

284. Coughlan, A.Y. and Wolfe, K.H. (2019) The reported point centromeres of *Scheffersomyces stipitis* are retrotransposon long terminal repeats. *Yeast*, **36**, 275–283.

285. Bayne, E.H., White, S.A., Kagansky, A., Bijos, D.A., Sanchez-Pulido, L., Hoe, K.-L., Kim, D.-U., Park, H.-O., Ponting, C.P. and Rappsilber, J. (2010) Stc1: a critical link between RNAi and chromatin modification required for heterochromatin integrity. *Cell*, **140**, 666–677.

286. Bühler, M. and Moazed, D. (2007) Transcription and RNAi in heterochromatic gene silencing. *Nat. Struct. Mol. Biol.*, **14**, 1041–1048.

287. Yang, J., Sun, S., Zhang, S., Gonzalez, M., Dong, Q., Chi, Z., Chen, Y. and Li, F. (2018) Heterochromatin and RNAi regulate centromeres by protecting CENP-A from ubiquitin-mediated degradation. *PLoS Genet.*, **14**, e1007572.

288. Kagansky, A., Folco, H.D., Almeida, R., Pidoux, A.L., Boukaba, A., Simmer, F., Urano, T., Hamilton, G.L. and Allshire, R.C. (2009) Synthetic heterochromatin bypasses RNAi and centromeric repeats to establish functional centromeres. *Science*, **324**, 1716–1719.

289. Jeseničnik, T., Štajner, N., Radišek, S. and Jakše, J. (2019) RNA interference core components identified and characterised in *Verticillium nonalfalfae*, a vascular wilt pathogenic plant fungi of hops. *Sci. Rep.*, **9**, 1–12.

290. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

291. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 1–9.

292. Slater, G.S.C. and Birney, E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 1–11.

293. Capella-Gutiérrez, S., Silla-Martínez, J.M. and Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.

294. Nguyen, L.-T., Schmidt, H.A., Von Haeseler, A. and Minh, B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.

295. Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

296. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

297. Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F. and Manke, T. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.*, **44**, W160–W165.

298. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M. and Li, W. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, 1–9.

299. Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S. and Aiden, E.L. (2016) Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.*, **3**, 95–98.

300. Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S. and Aiden, A.P. (2017) *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, **356**, 92–95.

301. Dudchenko, O., Shamim, M.S., Batra, S.S., Durand, N.C., Musial, N.T., Mostofa, R., Pham, M., St Hilaire, B.G., Yao, W. and Stamenova, E. (2018) The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. *bioRxiv*.

302. Jakše, J., Jelen, V., Radišek, S., de Jonge, R., Mandelc, S., Majer, A., Curk, T., Zupan, B., Thomma, B.P.H.J. and Javornik, B. (2018) Genome sequence of a lethal strain of xylem-invading *Verticillium nonalfalfae*. *Genome Announc.*, **6**, e01458-17.

303. Benjamini, Y. and Speed, T.P. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.*, **40**, e72–e72.

304. Dobin, A. and Gingeras, T.R. (2015) Mapping RNA-seq reads with STAR. *Curr. Protoc. Bioinforma.*, **51**, 11–14.

305. Palmer, J. and Stajich, J.E. (2017) Funannotate: eukaryotic genome annotation pipeline. *http://funannotate.readthedocs.io*.

306. Cook, D.E., Valle-Inclan, J.E., Pajoro, A., Rovenich, H., Thomma, B.P.H.J. and Faino, L. (2019) Long-read annotation: automated eukaryotic genome annotation based on long-read cDNA sequencing. *Plant Physiol.*, **179**, 38–54.

307. Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R. and Wortman, J.R. (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.*, **9**, 1–22.

308. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

309. Wickham, H. (2009) ggplot2: Elegant graphics for data analysis. *Media*, **35**, 10–1007.

310. Hahne, F. and Ivanek, R. (2016) Visualizing genomic data using Gviz and bioconductor. In *Statistical genomics*. Springer, pp. 335–351.

311. Sullivan, M.J., Petty, N.K. and Beatson, S.A. (2011) Easyfig: a genome comparison visualizer. *Bioinformatics*, **27**, 1009–1010.

312. Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C. and Salzberg, S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, **5**, 1–9.

313. Drillon, G., Carbone, A. and Fischer, G. (2014) SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. *PLoS One*, **9**, e92621.

314. Drillon, G., Champeimont, R., Oteri, F., Fischer, G. and Carbone, A. (2020) Phylogenetic reconstruction based on synteny block and gene adjacencies. *Mol. Biol. Evol.*, **37**, 2747–2762.

315. Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J. and Dorschner, M.O. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

316. Hoencamp, C., Dudchenko, O., Elbatsh, A.M.O., Brahmachari, S., Raaijmakers, J.A., van Schaik, T., Cacciatore, Á.S., Contessoto, V.G., van Heesbeen, R.G.H.P. and van den Broek, B. (2021) 3D genomics across the tree of life reveals condensin II as a determinant of architecture type. *Science*, **372**, 984–989.

317. Campos, E.I. and Reinberg, D. (2009) Histones: annotating chromatin. *Annu. Rev. Genet.*, **43**, 559–599.

318. Xhemalce, B., Dawson, M.A. and Bannister, A.J. (2012) Histone modifications. In Meyers,R.A. (ed), *Epigenetic regulation and epigenomics*. Wiley VCH, Weinheim, pp. 657–703.

319. Zhao, Y. and Garcia, B.A. (2015) Comprehensive catalog of currently documented histone modifications. *Cold Spring Harb. Perspect. Biol.*, **7**, a025064.

320. Jerkovic, I. and Cavalli, G. (2021) Understanding 3D genome organization by multidisciplinary methods. *Nat. Rev. Mol. Cell Biol.*

321. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D. and Lander, E.S. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.

322. Holwerda, S. and De Laat, W. (2012) Chromatin loops, gene positioning, and gene expression. *Front. Genet.*, **3**, 217.

323. Rowley, M.J. and Corces, V.G. (2018) Organizational principles of 3D genome architecture. *Nat. Rev. Genet.*, **19**, 789–800.

324. Eagen, K.P., Aiden, E.L. and Kornberg, R.D. (2017) Polycomb-mediated chromatin loops revealed by a subkilobase-resolution chromatin interaction map. *Proc. Natl. Acad. Sci.*, **114**, 8764–8769.

325. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S. and Ren, B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.

326. Ghavi-Helm, Y., Jankowski, A., Meiers, S., Viales, R.R., Korbel, J.O. and Furlong, E.E.M. (2019) Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat. Genet.*, **51**, 1272–1282.

327. Arzate-Mejía, R.G., Cerecedo-Castillo, A.J., Guerrero, G., Furlan-Magaril, M. and Recillas-Targa, F. (2020) In situ dissection of domain boundaries affect genome topology and gene transcription in Drosophila. *Nat. Commun.*, **11**, 1–16.

328. Kaushal, A., Mohana, G., Dorier, J., Özdemir, I., Omer, A., Cousin, P., Semenova, A., Taschner, M., Dergai, O. and Marzetta, F. (2021) CTCF loss has limited effects on global genome architecture in *Drosophila* despite critical regulatory functions. *Nat. Commun.*, **12**, 1–16.

329. Cavalheiro, G.R., Pollex, T. and Furlong, E.E.M. (2021) To loop or not to loop: what is the role of TADs in enhancer function and gene regulation? *Curr. Opin. Genet. Dev.*, **67**, 119–129.

330. Eser, U., Chandler-Brown, D., Ay, F., Straight, A.F., Duan, Z., Noble, W.S. and Skotheim, J.M. (2017) Form and function of topologically associating genomic domains in budding yeast. *Proc. Natl. Acad. Sci.*, **114**, E3061–E3070.

331. Yang, Y., Zhang, Y., Ren, B., Dixon, J.R. and Ma, J. (2019) Comparing 3D genome organization in multiple species using Phylo-HMRF. *Cell Syst.*, **8**, 494–505.

332. Harmston, N., Ing-Simmons, E., Tan, G., Perry, M., Merkenschlager, M. and Lenhard, B. (2017) Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. *Nat. Commun.*, **8**, 1–13.

333. Fudenberg, G. and Pollard, K.S. (2019) Chromatin features constrain structural variation across evolutionary timescales. *Proc. Natl. Acad. Sci.*, **116**, 2175–2180.

334. Golicz, A.A., Bhalla, P.L., Edwards, D. and Singh, M.B. (2020) Rice 3D chromatin structure correlates with sequence variation and meiotic recombination rate. *Commun. Biol.*, **3**, 1–9.

335. Krefting, J., Andrade-Navarro, M.A. and Ibn-Salem, J. (2018) Evolutionary stability of topologically associating domains is associated with conserved gene regulation. *BMC Biol.*, **16**, 1–12.

336. Liu, C., Cheng, Y.-J., Wang, J.-W. and Weigel, D. (2017) Prominent topologically associated domains differentiate global chromatin packing in rice from *Arabidopsis*. *Nat. Plants*, **3**, 742–748.

337. McArthur, E. and Capra, J.A. (2021) Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *Am. J. Hum. Genet.*, **108**, 269–283.

338. Rowley, M.J., Nichols, M.H., Lyu, X., Ando-Kuri, M., Rivera, I.S.M., Hermetz, K., Wang, P., Ruan, Y. and Corces, V.G. (2017) Evolutionarily conserved principles predict 3D chromatin organization. *Mol. Cell*, **67**, 837–852.

339. Huynh, L. and Hormozdiari, F. (2019) TAD fusion score: discovery and ranking the contribution of deletions to genome structure. *Genome Biol.*, **20**, 1–13.

340. James, C., Trevisan-Herraz, M. and Rico, D. (2021) Vertebrate whole genome duplications shaped the current 3D genome architecture. *bioRxiv*.

341. Liao, Y., Zhang, X., Chakraborty, M. and Emerson, J.J. (2021) Topologically associating domains and their role in the evolution of genome structure and function in *Drosophila*. *Genome Res.*, **31**, 397–410.

342. Torosin, N.S., Anand, A., Golla, T.R., Cao, W. and Ellison, C.E. (2020) 3D genome evolution and reorganization in the *Drosophila melanogaster* species group. *PLoS Genet.*, **16**, e1009229.

343. Niu, L., Shen, W., Shi, Z., Tan, Y., He, N., Wan, J., Sun, J., Zhang, Y., Huang, Y. and Wang, W. (2021) Three-dimensional folding dynamics of the *Xenopus tropicalis* genome. *Nat. Genet.*

344. Anania, C. and Lupiáñez, D.G. (2020) Order and disorder: abnormal 3D chromatin organization in human disease. *Brief. Funct. Genomics*, **19**, 128–138.

345. Lupiáñez, D.G., Spielmann, M. and Mundlos, S. (2016) Breaking TADs: how alterations of chromatin domains result in disease. *Trends Genet.*, **32**, 225–237.

346. Li, L., Lyu, X., Hou, C., Takenaka, N., Nguyen, H.Q., Ong, C.-T., Cubeñas-Potts, C., Hu, M., Lei, E.P. and Bosco, G. (2015) Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing. *Mol. Cell*, **58**, 216–231.

347. Liang, Z., Zhang, Q., Ji, C., Hu, G., Zhang, P., Wang, Y., Yang, L. and Gu, X. (2021) Reorganization of the 3D chromatin architecture of rice genomes during heat stress. *BMC Biol.*, **19**, 1–10.

348. Kainth, A.S., Chowdhary, S., Pincus, D. and Gross, D.S. (2021) Primordial super-enhancers: heat shock-induced chromatin organization in yeast. *Trends Cell Biol.*, **58**, 216–231.

349. Schalbetter, S.A., Fudenberg, G., Baxter, J., Pollard, K.S. and Neale, M.J. (2019) Principles of meiotic chromosome assembly revealed in *S. cerevisiae*. *Nat. Commun.*, **10**, 1–12.

350. Muller, H., Scolari, V.F., Agier, N., Piazza, A., Thierry, A., Mercy, G., Descorps-Declere, S., Lazar-Stefanita, L., Espeli, O. and Llorente, B. (2018) Characterizing meiotic chromosomes' structure and pairing using a designer sequence optimized for Hi-C. *Mol. Syst. Biol.*, **14**, e8293.

351. Hirano, T. (2016) Condensin-based chromosome organization from bacteria to vertebrates. *Cell*, **164**, 847–857.

352. Kalitsis, P., Zhang, T., Marshall, K.M., Nielsen, C.F. and Hudson, D.F. (2017) Condensin, master organizer of the genome. *Chromosom. Res.*, **25**, 61–76.

353. Lippman, Z., Gendrel, A.-V., Black, M., Vaughn, M.W., Dedhia, N., McCombie, W.R., Lavine, K., Mittal, V., May, B., Kasschau, K.D., *et al.* (2004) Role of transposable elements in heterochromatin and epigenetic control. *Nature*, **430**, 471–476.

354. Kramer, H.M., Seidl, M.F., Thomma, B.P.H.J. and Cook, D.E. (2021) Local rather than global H3K27me3 dynamics associates with differential gene expression in Verticillium dahliae. *bioRxiv*.

355. Meile, L., Peter, J., Puccetti, G., Alassimone, J., McDonald, B.A. and Sánchez-Vallet, A. (2020) Chromatin dynamics contribute to the spatiotemporal expression pattern of virulence genes in a fungal plant pathogen. *mBio*, **11**, e02343-20.

356. Torres, D.E., Oggenfuss, U., Croll, D. and Seidl, M.F. (2020) Genome evolution in fungal plant pathogens: looking beyond the two-speed genome model. *Fungal Biol. Rev.*, **34**, 136–143.

357. Frantzeskakis, L., Kusch, S. and Panstruga, R. (2019) The need for speed: compartmentalized genome evolution in filamentous phytopathogens. *Mol. Plant Pathol.*, **20**, 3.

358. Seidl, M.F. and Thomma, B.P.H.J. (2017) Transposable elements direct the coevolution between plants and microbes. *Trends Genet.*, **33**, 842–851.

359. Croll, D. and McDonald, B.A. (2012) The accessory genome as a cradle for adaptive evolution in pathogens. *PLoS Pathog.*, **8**, e1002608.

360. Torres, D.E., Thomma, B.P.H.J. and Seidl, M.F. (2021) Transposable elements contribute to genome dynamics and gene expression variation in the fungal plant pathogen *Verticillium dahliae*. *Genome Biol. Evol.*, **13**, evab135.

361. Dekker, J., Marti-Renom, M.A. and Mirny, L.A. (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.*, **14**, 390–403.

362. Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K.C., Grüning, B.A., Villaveces, J., Habermann, B., Akhtar, A. and Manke, T. (2018) High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.*, **9**, 1–15.

363. Bailey, T.L., Johnson, J., Grant, C.E. and Noble, W.S. (2015) The MEME suite. *Nucleic Acids Res.*, **43**, W39–W49.

364. Fornes, O., Castro-Mondragon, J.A., Khan, A., Van der Lee, R., Zhang, X., Richmond, P.A., Modi, B.P., Correard, S., Gheorghe, M. and Baranašić, D. (2020) JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **48**, D87–D92.

365. Monteiro, P.T., Oliveira, J., Pais, P., Antunes, M., Palma, M., Cavalheiro, M., Galocha, M., Godinho, C.P., Martins, L.C. and Bourbon, N. (2020) YEASTRACT+: a portal for cross-species comparative genomics of transcription regulation in yeasts. *Nucleic Acids Res.*, **48**, D642–D649.

366. Mirzaei, H., Knijnenburg, T.A., Kim, B., Robinson, M., Picotti, P., Carter, G.W., Li, S., Dilworth, D.J., Eng, J.K. and Aitchison, J.D. (2013) Systematic measurement of transcription factor-DNA interactions by targeted mass spectrometry identifies candidate gene regulatory proteins. *Proc. Natl. Acad. Sci.*, **110**, 3645–3650.

367. Newcomb, L.L., Hall, D.D. and Heideman, W. (2002) AZF1 is a glucose-dependent positive regulator of CLN3 transcription in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.*, **22**, 1607–1614.

368. Slattery, M.G., Liko, D. and Heideman, W. (2006) The function and properties of the Azf1 transcriptional regulator change with growth conditions in *Saccharomyces cerevisiae*. *Eukaryot. Cell*, **5**, 313–320.

369. Schlichter, A., Kasten, M.M., Parnell, T.J. and Cairns, B.R. (2020) Specialization of the chromatin remodeler RSC to mobilize partially-unwrapped nucleosomes. *eLife*, **9**, e58130.

370. Hepp, M. and Gutierrez, J.L. (2014) The yeast HMG proteins Hmo1 and Nhp6 exert a differential stimulatory effect on ATP-dependent nucleosome remodeling activity. *FEBS J.*, **281**, 65–783.

371. Ragab, A. and Travers, A. (2003) HMG-D and histone H1 alter the local accessibility of nucleosomal DNA. *Nucleic Acids Res.*, **31**, 7083–7089.

372. Kramer, H.M., Cook, D.E., van den Berg, G.C.M., Seidl, M.F. and Thomma, B.P.H.J. (2021) Three putative DNA methyltransferases of *Verticillium dahliae* differentially contribute to DNA methylation that is dispensable for growth, development and virulence. *Epigenetics Chromatin*, **14**, 1–15.

373. Gonzalez-Sandoval, A. and Gasser, S.M. (2016) On TADs and LADs: spatial control over gene expression. *Trends Genet.*, **32**, 485–495.

374. Spielmann, M., Lupiáñez, D.G. and Mundlos, S. (2018) Structural variation in the 3D genome. *Nat. Rev. Genet.*, **19**, 453–467.

375. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W. and Richards, S. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

376. Wang, M., Wang, P., Lin, M., Ye, Z., Li, G., Tu, L., Shen, C., Li, J., Yang, Q. and Zhang, X. (2018) Evolutionary dynamics of 3D genome architecture following polyploidization in cotton. *Nat. Plants*, **4**, 90–97.

377. Dong, P., Tu, X., Chu, P.Y., Lü, P., Zhu, N., Grierson, D., Du, B., Li, P. and Zhong, S. (2017) 3D chromatin architecture of large plant genomes determined by local A/B compartments. *Mol. Plant*, **10**, 1497–1509.

378. Szabo, Q., Jost, D., Chang, J.-M., Cattoni, D.I., Papadopoulos, G.L., Bonev, B., Sexton, T., Gurgo, J., Jacquier, C. and Nollmann, M. (2018) TADs are 3D structural units of higher-order chromosome organization in *Drosophila*. *Sci. Adv.*, **4**, eaar8082.

379. Chang, L.-H., Ghosh, S. and Noordermeer, D. (2020) TADs and their borders: free movement or building a wall? *J. Mol. Biol.*, **432**, 643–652.

380. Pope, B.D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D.L., Wang, Y., Hansen, R.S. and Canfield, T.K. (2014) Topologically associating domains are stable units of replication-timing regulation. *Nature*, **515**, 402–405.

381. Kolesnikova, T.D., Goncharov, F.P. and Zhimulev, I.F. (2018) Similarity in replication timing between polytene and diploid cells is associated with the organization of the Drosophila genome. *PLoS One*, **13**, e0195207.

382. Le Dily, F. and Beato, M. (2015) TADs as modular and dynamic units for gene regulation by hormones. *FEBS Lett.*, **589**, 2885–2892.

383. Le Dily, F., Bau, D., Pohl, A., Vicent, G.P., Serra, F., Soronellas, D., Castellano, G., Wright, R.H.G., Ballare, C. and Filion, G. (2014) Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes Dev.*, **28**, 2151–2162.

384. Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.-A., Schmitt, A.D., Espinoza, C.A. and Ren, B. (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, **503**, 290–294.

385. Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N.L., Meisig, J. and Sedat, J. (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**, 381–385.

386. Kim, J. and Dean, A. (2021) Enhancers navigate the three-dimensional genome to direct cell fate decisions. *Curr. Opin. Struct. Biol.*, **71**, 101–109.

387. Yildirir, G., Sperschneider, J., Malar C, M., Chen, E.C.H., Iwasaki, W., Cornell, C. and Corradi, N. (2021) Long reads and Hi-C sequencing illuminate the two compartment genome of the model arbuscular mycorrhizal symbiont *Rhizophagus irregularis*. *New Phytol.*

388. Rocha, P.P., Raviram, R., Bonneau, R. and Skok, J.A. (2015) Breaking TADs: insights into hierarchical genome organization. *Epigenomics*, **7**, 523–526.

389. Van Bortle, K., Nichols, M.H., Li, L., Ong, C.-T., Takenaka, N., Qin, Z.S. and Corces, V.G. (2014) Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biol.*, **15**, 1–18.

390. Wang, M., Li, J., Wang, P., Liu, F., Liu, Z., Zhao, G., Xu, Z., Pei, L., Grover, C.E., Wendel, J.F., *et al.* (2021) Comparative genome analyses highlight transposon-mediated genome expansion and the evolutionary architecture of 3D genomic folding in cotton. *Mol. Biol. Evol.*, **38**, 3621–3636.

391. Zhang, Y., Li, T., Preissl, S., Amaral, M.L., Grinstein, J.D., Farah, E.N., Destici, E., Qiu, Y., Hu, R. and Lee, A.Y. (2019) Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat. Genet.*, **51**, 1380–1388.

392. Finn, E.H., Pegoraro, G., Brandao, H.B., Valton, A.-L., Oomen, M.E., Dekker, J., Mirny, L. and Misteli, T. (2019) Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell*, **176**, 1502–1515.

393. Luppino, J.M., Park, D.S., Nguyen, S.C., Lan, Y., Xu, Z., Yunker, R. and Joyce, E.F. (2020) Cohesin promotes stochastic domain intermingling to ensure proper regulation of boundary-proximal genes. *Nat. Genet.*, **52**, 840–848.

394. Bintu, B., Mateo, L.J., Su, J.-H., Sinnott-Armstrong, N.A., Parker, M., Kinrot, S., Yamaya, K., Boettiger, A.N. and Zhuang, X. (2018) Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*, **362**.

395. Kentepozidou, E., Aitken, S.J., Feig, C., Stefflova, K., Ibarra-Soria, X., Odom, D.T., Roller, M. and Flicek, P. (2020) Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. *Genome Biol.*, **21**, 1–19.

396. Gong, Y., Lazaris, C., Sakellaropoulos, T., Lozano, A., Kambadur, P., Ntziachristos, P., Aifantis, I. and Tsirigos, A. (2018) Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries. *Nat. Commun.*, **9**, 1–12.

397. Barutcu, A.R., Maass, P.G., Lewandowski, J.P., Weiner, C.L. and Rinn, J.L. (2018) A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus. *Nat. Commun.*, **9**, 1–11.

398. Dekker, J. and Misteli, T. (2015) Long-range chromatin interactions. *Cold Spring Harb. Perspect. Biol.*, **7**, a019356.

399. Sawyer, I.A. and Dundr, M. (2016) Nuclear bodies: Built to boost. *J. Cell Biol.*, **213**, 509–511.

400. Mao, Y.S., Zhang, B. and Spector, D.L. (2011) Biogenesis and function of nuclear bodies. *Trends Genet.*, **27**, 295–306.

401. Morimoto, M. and Boerkoel, C.F. (2013) The role of nuclear bodies in gene expression and disease. *Biology*, **2**, 976–1033.

402. Fokkens, L., Shahi, S., Connolly, L.R., Stam, R., Schmidt, S.M., Smith, K.M., Freitag, M. and Rep, M. (2018) The multi-speed genome of *Fusarium oxysporum* reveals association of histone modifications with sequence divergence and footprints of past horizontal chromosome transfer events. *bioRxiv*, 10.1101/465070.

403. Wang, Q., Jiang, C., Wang, C., Chen, C., Xu, J.-R. and Liu, H. (2017) Characterization of the two-speed subgenomes of *Fusarium graminearum* reveals the fast-speed subgenome specialized for adaption and infection. *Front. Plant Sci.*, **8**, 140.

404. Rojas-Rojas, F.U. and Vega-Arreguín, J.C. (2021) Epigenetic insight into regulatory role of chromatin covalent modifications in lifecycle and virulence of *Phytophthora*. *Environ. Microbiol. Rep.*

405. Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, **26**, 589–595.

406. Wolff, J., Rabbani, L., Gilsbach, R., Richard, G., Manke, T., Backofen, R. and Grüning, B.A. (2020) Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.*, **48**, W177–W184.

407. Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J. and Mirny, L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.

408. Yardımcı, G.G., Ozadam, H., Sauria, M.E.G., Ursu, O., Yan, K.-K., Yang, T., Chakraborty, A., Kaul, A., Lajoie, B.R. and Song, F. (2019) Measuring the reproducibility and quality of Hi-C data. *Genome Biol.*, **20**, 1–19.

409. Gu, Z., Eils, R., Schlesner, M. and Ishaque, N. (2018) EnrichedHeatmap: an R/Bioconductor package for comprehensive visualization of genomic signal associations. *BMC Genomics*, **19**, 1–7.

410. Kuzniar, A., Maassen, J., Verhoeven, S., Santuari, L., Shneider, C., Kloosterman, W.P. and de Ridder, J. (2020) sv-callers: a highly portable parallel workflow for structural variant detection in whole-genome sequence data. *PeerJ*, **8**, e8214.

411. Goerner-Potvin, P. and Bourque, G. (2018) Computational tools to unmask transposable elements. *Nat. Rev. Genet.*, **19**, 688–704.

412. Cameron, D.L., Di Stefano, L. and Papenfuss, A.T. (2019) Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat. Commun.*, **10**, 1–11.

413. Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N. and Afshar, P.T. (2018) A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.*, **36**, 983–987.

414. Stuart, T., Eichten, S., Cahn, J., Borevitz, J. and Lister, R. (2016) Population scale mapping of novel transposable element diversity reveals links to gene regulation and epigenomic variation.

415. Lyons, E., Pedersen, B., Kane, J., Alam, M., Ming, R., Tang, H., Wang, X., Bowers, J., Paterson, A. and Lisch, D. (2008) Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol.*, **148**, 1772–1781.

416. Bertels, F., Silander, O.K., Pachkov, M., Rainey, P.B. and van Nimwegen, E. (2014) Automated reconstruction of whole-genome phylogenies from short-sequence reads. *Mol. Biol. Evol.*, **31**, 1077–1088.

417. Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I.T., Novak, A.M., Deran, A., Fang, Q., Xie, D., Feng, S. and Stiller, J. (2020) Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, **587**, 246–251.

418. Hickey, G., Paten, B., Earl, D., Zerbino, D. and Haussler, D. (2013) HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics*, **29**, 1341–1342.

419. Kruse, K., Hug, C.B. and Vaquerizas, J.M. (2020) FAN-C: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data. *Genome Biol.*, **21**, 1–19.

420. Gel, B., Díez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M.A. and Malinverni, R. (2016) regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics*, **32**, 289–291.

421. Wickham, H. (2016) Programming with ggplot2. In *ggplot2*. Springer, pp. 241–253.

422. Chen, A.H. and Silver, P.A. (2012) Designing biological compartmentalization. *Trends Cell Biol.*, **22**, 662–670.

423. Dos Santos, Á. and Toseland, C.P. (2021) Regulation of nuclear mechanics and the impact on DNA damage. *Int. J. Mol. Sci.*, **22**, 3178.

424. Prakash, K. and Fournier, D. (2017) Histone code and higher-order chromatin folding: a hypothesis. *Genomics Comput. Biol.*, **3**, e41.

425. Grewal, S.I.S. and Jia, S. (2007) Heterochromatin revisited. *Nat. Rev. Genet.*, **8**, 35–46.

426. Yun, M., Wu, J., Workman, J.L. and Li, B. (2011) Readers of histone modifications. *Cell Res.*, **21**, 564–578.

427. Lücking, R., Aime, M.C., Robbertse, B., Miller, A.N., Aoki, T., Ariyawansa, H.A., Cardinali, G., Crous, P.W., Druzhinina, I.S. and Geiser, D.M. (2021) Fungal taxonomy and sequence-based nomenclature. *Nat. Microbiol.*, **6**, 540–548.

428. Richards, T.A., Leonard, G. and Wideman, J.G. (2017) What defines the “kingdom” fungi? *Microbiol. Spectr.*, **5**, 3–5.

429. Sanchez, S. and Demain, A.L. (2017) Bioactive products from fungi. In *Food bioactives*. Springer, pp. 59–87.

430. De Lucca, A.J. (2007) Harmful fungi in both agriculture and medicine. *Rev. Iberoam. Micol.*, **24**, 3.

431. Lee, D.Y., Hayes, J.J., Pruss, D. and Wolffe, A.P. (1993) A positive role for histone acetylation in transcription factor access to nucleosomal DNA. *Cell*, **72**, 73–84.

432. Chen, Y., Jørgensen, M., Kolde, R., Zhao, X., Parker, B., Valen, E., Wen, J. and Sandelin, A. (2011) Prediction of RNA Polymerase II recruitment, elongation and stalling from histone modification data. *BMC Genomics*, **12**, 1–16.

433. Angelov, D., Molla, A., Perche, P.-Y., Hans, F., Côté, J., Khochbin, S., Bouvet, P. and Dimitrov, S. (2003) The histone variant macroH2A interferes with transcription factor binding and SWI/SNF nucleosome remodeling. *Mol. Cell*, **11**, 1033–1041.

434. Musselman, C.A., Lalonde, M.-E., Côté, J. and Kutateladze, T.G. (2012) Perceiving the epigenetic landscape through histone readers. *Nat. Struct. Mol. Biol.*, **19**, 1218–1227.

435. Vermeulen, M., Mulder, K.W., Denissov, S., Pijnappel, W.W.M.P., van Schaik, F.M.A., Varier, R.A., Baltissen, M.P.A., Stunnenberg, H.G., Mann, M. and Timmers, H.T.M. (2007) Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell*, **131**, 58–69.

436. Lauberth, S.M., Nakayama, T., Wu, X., Ferris, A.L., Tang, Z., Hughes, S.H. and Roeder, R.G. (2013) H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. *Cell*, **152**, 1021–1036.

437. Shi, X., Hong, T., Walter, K.L., Ewalt, M., Michishita, E., Hung, T., Carney, D., Pena, P., Lan, F. and Kaadige, M.R. (2006) ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression. *Nature*, **442**, 96–99.

438. Yang, Z., Qian, S., Scheid, R.N., Lu, L., Chen, X., Liu, R., Du, X., Lv, X., Boersma, M.D. and Scalf, M. (2018) EBS is a bivalent histone reader that regulates floral phase transition in Arabidopsis. *Nat. Genet.*, **50**, 1247–1253.

439. Dekker, J., Rippe, K., Dekker, M., Kleckner, N., Woodcock, C.L., Dimitrov, S., Andrulis, E.D., Neiman, A.M., Zappulla, D.C., Sternglanz, R., *et al.* (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.

440. Beagan, J.A. and Phillips-Cremins, J.E. (2020) On the existence and functionality of topologically associating domains. *Nat. Genet.*, **52**, 8–16.

441. Ghavi-Helm, Y. (2020) Functional consequences of chromosomal rearrangements on gene expression: not so deleterious after all? *J. Mol. Biol.*, **432**, 665–675.

442. Donaldson-Collier, M.C., Sungalee, S., Zufferey, M., Tavernari, D., Katanayeva, N., Battistello, E., Mina, M., Douglass, K.M., Rey, T. and Raynaud, F. (2019) EZH2 oncogenic mutations drive epigenetic, transcriptional, and structural changes within chromatin domains. *Nat. Genet.*, **51**, 517–528.

443. Cheutin, T. and Cavalli, G. (2014) Polycomb silencing: from linear chromatin domains to 3D chromosome folding. *Curr. Opin. Genet. Dev.*, **25**, 30–37.

444. Kim, V.N., Han, J. and Siomi, M.C. (2009) Biogenesis of small RNAs in animals. *Nat. Rev. Mol. Cell Biol.*, **10**, 126–139.

445. Axtell, M.J. (2013) Classification and comparison of small RNAs from plants. *Annu. Rev. Plant Biol.*, **64**, 137–159.

446. Castel, S.E. and Martienssen, R.A. (2013) RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. *Nat. Rev. Genet.*, **14**, 100–112.

447. Law, J.A., Du, J., Hale, C.J., Feng, S., Krajewski, K., Palanca, A.M.S., Strahl, B.D., Patel, D.J. and Jacobsen, S.E. (2013) Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. *Nature*, **498**, 385–389.

448. Liu, X., Chen, X., Yu, X., Tao, Y., Bode, A.M., Dong, Z. and Cao, Y. (2013) Regulation of microRNAs by epigenetics and their interplay involved in cancer. *J. Exp. cinical cancer Res.*, **32**, 1–8.

449. Xie, Z., Allen, E., Fahlgren, N., Calamar, A., Givan, S.A. and Carrington, J.C. (2005) Expression of *Arabidopsis MIRNA* genes. *Plant Physiol.*, **138**, 2145–2154.

450. Nepal, C., Coolen, M., Hadzhiev, Y., Cussigh, D., Mydel, P., Steen, V.M., Carninci, P., Andersen, J.B., Bally-Cuif, L. and Müller, F. (2016) Transcriptional, post-transcriptional and chromatin-associated regulation of pri-miRNAs, pre-miRNAs and moRNAs. *Nucleic Acids Res.*, **44**, 3070–3081.

451. Wang, Z., Ma, Z., Castillo-González, C., Sun, D., Li, Y., Yu, B., Zhao, B., Li, P. and Zhang, X. (2018) SWI2/SNF2 ATPase CHR2 remodels pri-miRNAs via Serrate to impede miRNA production. *Nature*, **557**, 516–521.

452. Motamedi, M.R., Verdel, A., Colmenares, S.U., Gerber, S.A., Gygi, S.P. and Moazed, D. (2004) Two RNAi complexes, RITS and RDRC, physically interact and localize to noncoding centromeric RNAs. *Cell*, **119**, 789–802.

453. Jih, G., Iglesias, N., Currie, M.A., Bhanu, N. V, Paulo, J.A., Gygi, S.P., Garcia, B.A. and Moazed, D. (2017) Unique roles for histone H3K9me states in RNAi and heritable silencing of transcription. *Nature*, **547**, 463–467.

454. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (2002) How genomes evolve. In *Molecular biology of the cell. 4th edition*. Garland Science.

455. Haas, B.J., Kamoun, S., Zody, M.C., Jiang, R.H.Y., Handsaker, R.E., Cano, L.M., Grabherr, M., Kodira, C.D., Raffaele, S., Torto-Alalibo, T., *et al.* (2009) Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature*, **461**, 393–398.

456. Grandaubert, J., Balesdent, M.H. and Rouxel, T. (2014) Evolutionary and adaptive role of transposable elements in fungal genomes. *Adv. Bot. Res.*, **70**, 79–105.

457. Rebollo, R., Horard, B., Hubert, B. and Vieira, C. (2010) Jumping genes and epigenetics: towards new species. *Gene*, **454**, 1–7.

458. Fedoroff, N. V (2012) Transposable elements, epigenetics, and genome evolution. *Science*, **338**, 758–767.

459. Fouché, S., Oggenfuss, U., Chanclud, E. and Croll, D. (2021) A devil's bargain with transposable elements in plant pathogens. *Trends Genet.*

460. He, C., Zhang, Z., Li, B. and Tian, S. (2020) The pattern and function of DNA methylation in fungal plant pathogens. *Microorganisms*, **8**, 227.

461. Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C.R., Dopman, E.B., Dickinson, W.J., Okamoto, K., Kulkarni, S. and Hartl, D.L. (2008) A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl. Acad. Sci.*, **105**, 9272–9277.

462. Fryxell, K.J. and Zuckerkandl, E. (2000) Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol. Biol. Evol.*, **17**, 1371–1383.

463. Lu, Z., Cui, J., Wang, L., Teng, N., Zhang, S., Lam, H.-M., Zhu, Y., Xiao, S., Ke, W. and Lin, J. (2021) Genome-wide DNA mutations in *Arabidopsis* plants after multigenerational exposure to high temperatures. *Genome Biol.*, **22**, 1–27.

464. Żemojtel, T., Arndt, P.F., Behrens, S., Bourque, G. and Vingron, M. (2011) CpG deamination creates transcription factor–binding sites with high efficiency. *Genome Biol. Evol.*, **3**, 1304–1311.

465. Erdel, F. and Rippe, K. (2018) Formation of chromatin subcompartments by phase separation. *Biophys. J.*, **114**, 2262–2270.

466. Larson, A.G., Elnatan, D., Keenen, M.M., Trnka, M.J., Johnston, J.B., Burlingame, A.L., Agard, D.A., Redding, S. and Narlikar, G.J. (2017) Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. *Nature*, **547**, 236–240.

467. Singh, P.B. and Newman, A.G. (2020) On the relations of phase separation and Hi-C maps to epigenetics. *R. Soc. open Sci.*, **7**, 191976.

468. Tatavosian, R., Kent, S., Brown, K., Yao, T., Duc, H.N., Huynh, T.N., Zhen, C.Y., Ma, B., Wang, H. and Ren, X. (2019) Nuclear condensates of the Polycomb protein chromobox 2 (CBX2) assemble through phase separation. *J. Biol. Chem.*, **294**, 1451–1463.

469. Branzei, D. and Foiani, M. (2008) Regulation of DNA repair throughout the cell cycle. *Nat. Rev. Mol. Cell Biol.*, **9**, 297–308.

470. Huertas, D., Sendra, R. and Muñoz, P. (2009) Chromatin dynamics coupled to DNA repair. *Epigenetics*, **4**, 31–42.

471. Meas, R., Smerdon, M.J. and Wyrick, J.J. (2015) The amino-terminal tails of histones H2A and H3 coordinate efficient base excision repair, DNA damage signaling and postreplication repair in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **43**, 4990–5001.

472. Ohkura, H. (2015) Meiosis: an overview of key differences from mitosis. *Cold Spring Harb. Perspect. Biol.*, **7**, a015859.

473. Aggarwal, B.D. and Calvi, B.R. (2004) Chromatin regulates origin activity in *Drosophila* follicle cells. *Nature*, **430**, 372–376.

474. Sansam, C.G., Pietrzak, K., Majchrzycka, B., Kerlin, M.A., Chen, J., Rankin, S. and Sansam, C.L. (2018) A mechanism for epigenetic control of DNA replication. *Genes Dev.*, **32**, 224–229.

475. Gindin, Y., Valenzuela, M.S., Aladjem, M.I., Meltzer, P.S. and Bilke, S. (2014) A chromatin structure-based model accurately predicts DNA replication timing in human cells. *Mol. Syst. Biol.*, **10**, 722.

476. MacAlpine, D.M. and Almouzni, G. (2013) Chromatin and DNA replication. *Cold Spring Harb. Perspect. Biol.*, **5**, a010207.

477. Casas-Delucchi, C.S., van Bemmel, J.G., Haase, S., Herce, H.D., Nowak, D., Meilinger, D., Stear, J.H., Leonhardt, H. and Cardoso, M.C. (2012) Histone hypoacetylation is required to maintain late replication timing of constitutive heterochromatin. *Nucleic Acids Res.*, **40**, 159–169.

478. Schwaiger, M., Kohler, H., Oakeley, E.J., Stadler, M.B. and Schübeler, D. (2010) Heterochromatin protein 1 (HP1) modulates replication timing of the *Drosophila* genome. *Genome Res.*, **20**, 771–780.

479. Brustel, J., Kirstein, N., Izard, F., Grimaud, C., Prorok, P., Cayrou, C., Schotta, G., Abdelsamie, A.F., Déjardin, J. and Méchali, M. (2017) Histone H4K20 tri-methylation at late-firing origins ensures timely heterochromatin replication. *EMBO J.*, **36**, 2726–2741.

480. Freitag, M. (2016) The kinetochore interaction network (KIN) of ascomycetes. *Mycologia*, **108**, 485–505.

481. Achrem, M., Szućko, I. and Kalinka, A. (2020) The epigenetic regulation of centromeres and telomeres in plants and animals. *Comp. Cytogenet.*, **14**, 265.

482. Nishimura, K., Komiya, M., Hori, T., Itoh, T. and Fukagawa, T. (2019) 3D genomic architecture reveals that neocentromeres associate with heterochromatin regions. *J. Cell Biol.*, **218**, 134–149.

483. Schotanus, K. and Heitman, J. (2020) Centromere deletion in *Cryptococcus deuterogattii* leads to neocentromere formation and chromosome fusions. *eLife*, **9**, e56026.

484. Schotanus, K., Yadav, V. and Heitman, J. (2021) Epigenetic dynamics of centromeres and neocentromeres in *Cryptococcus deuterogattii*. *PLoS Genet.*, **17**, e1009743.

485. Henneman, B., Van Emmerik, C., van Ingen, H. and Dame, R.T. (2018) Structure and function of archaeal histones. *PLoS Genet.*, **14**, e1007582.

486. Stevens, K.M., Swadling, J.B., Hocher, A., Bang, C., Gribaldo, S., Schmitz, R.A. and Warnecke, T. (2020) Histone variants in archaea and the evolution of combinatorial chromatin complexity. *Proc. Natl. Acad. Sci.*, **117**, 33384–33395.

487. Harada, A., Kimura, H. and Ohkawa, Y. (2021) Recent advance in single-cell epigenomics. *Curr. Opin. Struct. Biol.*, **71**, 116–122.

488. Agbleke, A.A., Amitai, A., Buenrostro, J.D., Chakrabarti, A., Chu, L., Hansen, A.S., Koenig, K.M., Labade, A.S., Liu, S. and Nozaki, T. (2020) Advances in chromatin and chromosome research: perspectives from multiple fields. *Mol. Cell*, **76**, 881–901.

Summary

Through the association of protein complexes to DNA, the nuclear genome is broadly organized into accessible euchromatin and condensed heterochromatin. Chemical and physical alterations to these types of chromatin may impact their organization and functionality, and are therefore important regulators of nuclear processes in eukaryotes. Studies in various fungal plant pathogens have uncovered an association between chromatin organization and expression of *in planta*-induced effector genes that are important for pathogenicity. Chapter 1 of this thesis introduces interactions between plants and microbial pathogens, with a particular focus on the plant pathogenic Ascomycete fungus *Verticillium dahliae*; the subject of study of this thesis research. *V. dahliae* is a soil-borne filamentous fungus that can infect hundreds of host plants and colonizes their xylem vessels, leading to wilt diseases that can devastate crop yields. Chapter 1 outlines a prevalent hypothesis on epigenetic regulation of effector gene expression, stating that chromatin at effector gene-containing genomic regions is condensed when the fungus does not grow inside its plant hosts. Consequently, in order to express effectors *in planta*, the pathogen requires to chemically alter its chromatin, leading to chromatin de-condensation.

In a similar fashion as has been reported for the genomes of various other fungal species, the genome of *V. dahliae* can be characterized as a so-called two-speed genome, in which particular regions are more plastic, and evolve more rapidly, than the evolutionary more stable core genome. In Chapter 2, we explore epigenome features, including DNA methylation, chromatin accessibility and histone methylation, and show that the plastic genome of *V. dahliae* is associated with tri-methylation of lysine 27 on histone 3 (H3K27me3) and with accessible chromatin. Using a machine learning approach trained on chromatin, and validated through orthogonal analyses, we now identified approximately twice as much DNA in plastic regions than previously recognized. The collective plastic regions are now referred to as adaptive genomic regions (AGRs). Our results show that a specific chromatin profile defines the plastic genome, and highlight how different epigenetic factors contribute to the organization of AGRs.

H3K27me3 is generally associated with facultative heterochromatin, which represents a closed conformation of the DNA and corresponding inaccessibility to the transcriptional machinery, yet can de-condense upon recognition of external cues. In Chapter 3, we investigated the involvement of H3K27me3 in transcriptional regulation by comparing H3K27me3 coverage and transcription for *V. dahliae* cultivated in three *in vitro* cultivation media. We show that although various genes in AGRs are differentially expressed between the cultivation media, H3K27me3 domains globally display stable profiles. However, we do observe local quantitative differences in H3K27me3 coverage that associate with differentially expressed genes, although this is not a ubiquitous pattern. Overall, our results demonstrate that although some loci display H3K27me3 dynamics that can contribute to transcriptional variation, other loci do not show such dynamics. Thus, we conclude that while H3K27me3 is required for transcriptional repression, it is not a conditionally responsive global regulator of differential transcription. We propose that the H3K27me3 domains that do not undergo dynamic methylation may contribute to transcription through other mechanisms, or may serve additional genomic regulatory functions.

Methylation of cytosine nucleobases (5-methylcytosine, 5mC) is an important epigenetic control mechanism that is restricted to genomic regions containing transposable elements (TEs) in many organisms, including fungi. Two DNA methyltransferases, Dim2 and Dnmt5, are known to perform methylation at cytosines in fungi. While most ascomycete fungi encode both Dim2 and Dnmt5, only few functional studies have been performed in species that contain both genes. In Chapter 4, we use functional analyses to show that Dim2, but not Dnmt5 or the putative sexual cycle-related DNA methyltransferase Rid, is responsible for the majority of DNA methylation in *V. dahliae*. Single and double DNA methyltransferase mutants did not show altered development, virulence, or transcription of genes or TEs. In contrast, Hp1 and Dim5 mutants that are impacted in chromatin-associated processes upstream of DNA methylation are severely affected in development and virulence and display transcriptional reprogramming in AGRs. As these AGRs are largely devoid of DNA methylation and of Hp1- and Dim5-associated heterochromatin, the differential transcription is likely caused by pleiotropic effects rather than by differential DNA methylation. Overall, our results suggest that Dim2 is the main DNA methyltransferase in *V. dahliae* and, in conjunction with work on other fungi, is likely the main active DNMT in ascomycetes, irrespective of Dnmt5 presence. We speculate that Dnmt5 and Rid act under specific, presently enigmatic, conditions or, alternatively, act in DNA-associated processes other than DNA methylation.

Centromeres are chromosomal regions that are crucial for chromosome segregation during mitosis and meiosis, and failed centromere formation can contribute to chromosomal anomalies. Despite this conserved function, centromeres differ significantly between, and even within, species. Thus far, systematic studies into the organization and evolution of fungal centromeres remain scarce. In Chapter 5, we identified the centromeres in each of the ten species of the *Verticillium* genus and characterized their organization and evolution. Chromatin immunoprecipitation of the centromere-specific histone CenH3 (ChIP-seq) and chromatin conformation capture (Hi-C) followed by high-throughput sequencing identified eight conserved, large (~150 kb), AT-, and repeat-rich regional centromeres that are embedded in heterochromatin in *V. dahliae*. Using Hi-C, we similarly identified repeat-rich centromeres in the other *Verticillium* species. Strikingly, a single degenerated LTR retrotransposon is strongly associated with centromeric regions in some *Verticillium* species. Extensive chromosomal rearrangements occurred during *Verticillium* evolution, of which some could be linked to centromeres, suggesting that centromeres contributed to chromosomal evolution. The size and organization of centromeres differ considerably between species, and centromere size was found to correlate with the genome-wide repeat content. Overall, this chapter highlights the contribution of repetitive elements to the diversity and rapid evolution of centromeres within the *Verticillium* genus.

The three dimensional (3D) folding of DNA in the nucleus organizes chromosomes into so-called topologically associating domains (TADs). These TADs are self-interacting genomic regions that display less interaction with adjacent regions. Functionally, TADs have been implicated in transcriptional regulation as well as in genome evolution in numerous organisms, yet in fungi the functional implication of these regions remains less clear. In Chapter 6, we utilize Hi-C data generated for *V. dahliae* to investigate TAD organization and its influence on transcription. Additionally, we compare the TAD organization between two *V. dahliae* strains as

well as with other *Verticillium* species to study the conservation of TADs throughout the genus. Remarkably, we find that TADs in the AGRs of *V. dahliae* are less well insulated than TADs in the core genome, indicating that TADs in AGRs are not as well established as those in the core genome. Moreover, TADs in AGRs display significantly more co-regulation of gene expression than TADs in the core genome. Furthermore, genes located in TAD boundaries, i.e. regions that delineate adjacent TADs, in AGRs are generally lower expressed *in vitro*, while stronger differentially expressed between *in vitro* conditions, than genes located in TADs in AGRs. We find that TAD boundaries are depleted for structural variation between *Verticillium* species, and that TADs are generally conserved in the *Verticillium* genus. Overall, our study points towards an association between TAD organization and transcriptional regulation as well as genome evolution in *Verticillium.*

Finally, Chapter 7 revisits the prevalent hypothesis on epigenetic regulation of effector gene expression through extensive chromatin dynamics, as presented in Chapter 1. I conclude that this hypothesis is likely too simple, and therefore I bring forward alternative hypotheses to explain the potential role of H3K27me3 in transcriptional regulation of *in planta* and *in vitro* differentially expressed genes. Furthermore, the implications of the findings presented in this thesis, regarding epigenetic mechanisms and spatial genome organization, are discussed in the broader context of nuclear processes in eukaryotic organisms.

Acknowledgements

After a long time in which I wrote the words that fill all the preceding pages of this thesis, its finally time to write the last words and to express my gratitude to the people who supported me throughout my time as a PhD student. So, without further ado.

Thank you and goodbye!

Just kidding, see next page

First and foremost, I want to thank my supervisors, because without them this thesis would not have existed. **Bart**, you gave me the opportunity to start my PhD, and although I must have made it difficult for you at times, you helped me through this adventure till the end. You are not only a very good scientist, but also an accomplished psychologist, as you could prick right through my insecurities and self-imposed blockades. **David**, you are the person who put me on the path of epigenetics. I think that your excitement and enjoyment in science has been an important factor in me doing a PhD. I'm envious of the way you experience and practice science. It seems that at times you have manuscripts outlined in your head, before the first experiment started. You're not only good to have around for science, but also for the more relaxing activities. For instance, that time we ended up in an Irish pub in Munich, on St. Patrick's Day, was a great experience. **Michael**, you are a great source of knowledge and always seem to know the best solutions for problems I had with bioinformatics, analyses, and writing. It was always interesting to hear your view on presented data and the follow-up analyses you proposed.

Then, my dearest and nearest colleagues, and paranymphs. **Hui**, you are a great person and a great cook. You have made me very interested in Chinese culture, language and food. I have really enjoyed the evenings that you cooked for us, but I have to admit that your traditional sliced pig ear recipe was not my favorite food ever. I am glad that you decided to stay near the Netherlands, and perhaps even are likely to come back (with a bit of pressure from Bart, thanks Bart!). **Nick**, as you wrote in your acknowledgements: "our personalities differ as much as Friesland and Brabant are apart, for Dutch standards". However, somehow this made us quite compatible. A bit belatedly, I want to apologize for all the terrible word play jokes you had to suffer from me. I had a great time sitting next to you in the lab and the writing room that, due to your drive, you managed to escape long before I did.

For my chapters and experiments I got to work together with various people. **Grardy**, you are probably as important to the Verticillium group, if not more, than Bart is. With your experience in Phyto you have helped many PhD students to get their experiments working, and you had to perform numerous experiments to finish manuscripts. Your proficiency in doing this is proven by your amazing track record. Also, you helped me stay on the right track. When I was thinking of unnecessarily postponing an experiment or my writing, you skillfully reminded me to correct my planning. During my PhD, I had the pleasure to guide three students in the lab; Hamid, Bas and Xin. Sadly, none of your work ended up in this thesis, but I learned a lot from guiding you. **Xin**, you have been a PhD student of your own right for a while now, and I know you will be successful. Good luck with all you do. **David**, you are a passionate scientist and you developed our shared project in a way that I never could have managed on my own. And a special thanks to **Laurens**, although we never really worked together, we must have spent a lot of time discussing all kinds of random things during coffee, which made the coffee break the highlight of most days.

Besides the previously mentioned persons, there are of course a whole range of people whom I met during my stay in Phytopathology. I did not intend to write an endless list in here, but I wanted to thank all of you, so I did anyway. First of all, the past and the few remaining members of the Verticillium group; Mireille, Luigi, Luis, Dirk Jan, Jordi, Hanna, Xiaoqian, Malaika, Yin, Jinling, Jasper, Jasper, Katharina, Gabriel, Edgar and Nelia, as well as all members of the Phyto department; Klaas, Johan, Elysa, Kiki, Chara, Sander, Aranka, Shuqing, Laura, Maikel, Si, Yaohua, Weizhen, Einar, Karolina, Lorena, Michele Wen, Sergio, Jelmer, Jinbin, Ali, Esther, Ciska, Giuliana, Henriek, Petra, Anneke, Harold, Sander, Francine, Matthieu, Jan and Gert. Thank you all for making Phyto what it is and making my stay here enjoyable (p.s. sorry if I forgot someone in this list; if you feel left out, call me up and I'll personally thank you :P).

Finally, I want to thank my non-science friends and family, who likely have no clue what the heck I've been doing for the past years. Obviously, I have to send my final thanks to **Mirella** and our beautiful daughter **Julia**.

Let me finish my acknowledgements with a quote from the people to whom I probably listened most during my PhD:

"In the end, the love you take is equal to the love you make"
- John, Paul, George & Ringo -

About the author

Martin Kramer was born in the city of Hindeloopen, the Netherlands, on September 12th, 1987. In 2004, he started his Bachelor's programme in biotechnology at the Van Hall Larenstein university of applied sciences in Leeuwarden. During this study, he performed medical research-oriented internships at the VU medical center in Amsterdam and at the University Medical Center in Groningen. After some hiccups and a switch of focus, he continued his bachelor in biotechnology by doing plant-pathogen research-oriented internships at the Stichting Proefboerderijen Noordelijke Akkerbouw (SPNA) in Nieuw Beerta and at Syngenta in Enkhuizen. After graduation, Martin worked at Enza zaden in Enkhuizen, to develop detection methods for several bacterial pathogens of onions. In 2013, he started his Master's programme on plant biotechnology at Wageningen University. In the second year of the MSc programme, he joined the group of Prof. Bart Thomma to identify effector genes and to study their transcriptional regulation. Following completion of the thesis, he accepted the offer to join Prof. Thomma's group as a PhD student. Prior to the initiation of the project, Martin did an internship in the group of Prof. Eva Holtgrewe-Stukenbrock at the Max-Planck Institute for Evolutionary Biology in Plön, Germany, where he studied the epigenetic regulation of transcription in a fungal plant pathogen. Besides doing his internship, he wrote a PhD project proposal for a grant call by the Dutch Science Organization (NWO). Martin was one of the four students that year that were selected to receive a personal grant of €250.000 to perform their PhD studies. In October 2015, under the joint supervision of Bart Thomma, David Cook and Michael Seidl, Martin started his PhD study into the epigenetic regulation of transcription and genome evolution in the fungal plant pathogen *Verticillium dahliae*, using functional genetic, epigenetic and bioinformatic approaches.

List of publications

Cook, D.E., **Kramer, H.M.**, Torres, D.E., Seidl, M.F. and Thomma, B.P.H.J. (2020) A unique chromatin profile defines adaptive genomic regions in a fungal plant pathogen. eLife, 9, e62208.

Seidl, M.F., **Kramer, H.M.**, Cook, D.E., Lorencini Fiorin, G., van den Berg, G.C.M., Faino, L. and Thomma, B.P.H.J. (2020) Repetitive elements contribute to the diversity and evolution of centromeres in the fungal genus *Verticillium*. mBio, 11, e01714-20.

Kramer, H.M., Cook, D.E., van den Berg, G.C.M., Seidl, M.F. and Thomma, B.P.H.J. (2021) Three putative DNA methyltransferases of *Verticillium dahliae* differentially contribute to DNA methylation that is dispensable for growth, development and virulence. *Epigenetics Chromatin*, **14**, 1–15.

Kramer, H.M., Seidl, M.F., Thomma, B.P.H.J. and Cook, D.E. (2022) Local rather than global H3K27me3 dynamics associates with differential gene expression in *Verticillium dahliae*. mBio, e03566-21.

Depotter, J.R.L., van Beveren, F., Rodriguez-Moreno, L., **Kramer, H.M.**, Chavarro Carrero, E.A., van den Berg, G.C.M., Wood, T.A., Thomma, B.P.H.J. and Seidl, M.F. (2021) The interspecific fungal hybrid *Verticillium longisporum* displays subgenome-specific gene expression. mBio, 12, e0149621.

Kramer, H.M., Cook, D.E., Seidl, M.F. and Thomma, B.P.H.J. (2022) Epigenetic regulation of nuclear processes in fungal plant pathogens (in preparation).

Torres, D.E., **Kramer, H.M.**, Tracanna, V., Fiorin, G.L., Cook, D.E., Seidl, M.F. and Thomma, B.P.H.J. (2022) Local Three-dimensional chromatin organization impacts the evolution of adaptive genomic regions in *Verticillium dahliae* (in preparation).

The research in this thesis was financially supported by the Research Council Earth and Life Sciences (ALW) of the Netherlands Organization of Scientific Research (NWO) in the framework of the program Graduate School Horticulture and Starting Materials, project number 831.15.002.

Cover & layout design: Iliana Boshoven-Gkini | AgileColor.com
Printed by Ipskamp | proefschriften.net