Special Issue Article

Telomeres and a repeat-rich chromosome encode effector gene clusters in plant pathogenic Colletotrichum fungi

Pamela Gan,1 Ryoko Hiroyama,1 Ayako Tsushima,1,2 Sachiko Masuda,1 Arisa Shibata,1 Akiko Ueno,1 Naoyoshi Kumakura,1 Mari Narusaka,3 Trinh Xuan Hoat,4 Yoshihiro Narusaka,3 Yoshitaka Takano5 and Ken Shirasu1,2*

1RIKEN Center for Sustainable Resource Sciences, Yokohama, Kanagawa, Japan.
2Graduate School of Science, The University of Tokyo, Bunkyo, Tokyo, Japan.
3Research Institute for Biological Sciences, Okayama Prefectural Technology Center for Agriculture, Forestry, and Fisheries, Okayama, Japan.
4Plant Protection Research Institute, Ha Noi City, Vietnam.
5Graduate School of Agriculture, Kyoto University, Kyoto, Japan.

Summary

Members of the Colletotrichum gloeosporioides species complex are causal agents of anthracnose in many commercially important plants. Closely related strains have different levels of pathogenicity on hosts despite their close phylogenetic relationship. To gain insight into the genetics underlying these differences, we generated and annotated whole-genome assemblies of multiple isolates of Cf and Cs, as well as three previously unsequenced species, Ca, C. tropicale and C. viniferum with different pathogenicity on strawberry. Based on comparative genomics, we identified accessory regions with a high degree of conservation in strawberry-pathogenic Cf, Cs and Ca strains. These regions encode homologs of pathogenicity-related genes known as effectors, organized in syntenic gene clusters, with copy number variations in different strains of Cf, Cs and Ca. Analysis of highly contiguous assemblies of Cf, Cs and Ca revealed the association of related accessory effector gene clusters with telomeres and repeat-rich chromosomes and provided evidence of exchange between these two genomic compartments. In addition, expression analysis indicated that orthologues in syntenic gene clusters showed a tendency for correlated gene expression during infection. These data provide insight into mechanisms by which Colletotrichum genomes evolve, acquire and organize effectors.

Introduction

Fungi within the genus Colletotrichum can be subdivided into species complexes consisting of closely related species (Cannon et al., 2012). Among them, members of the Colletotrichum gloeosporioides species complex (CGSC) are pathogens that cause significant damage to a wide range of commercially important plants (Weir et al., 2012). For example, Colletotrichum fructicola (Cf) infects avocado, apple, pear, strawberry, lemons, cocoa, coffee and yam (Weir et al., 2012). Further, different CGSC species have geographic and host range overlaps. For instance, Cf, Colletotrichum siamense (Cs) and Colletotrichum aenigma (Ca) have been identified as causal agents of strawberry anthracnose in Japan (Gan et al., 2016).

Infection of aerial plant tissue by Colletotrichum generally occurs via asexual conidia. Upon contact with the host, a conidium germinates and forms a melanized appressorium that is involved in host penetration (Shen et al., 2001; O’Connell et al., 2004; De Silva et al., 2017). From the appressorium, a specialized fungal hypha, known as the penetration peg, emerges at the point of penetration, and develops into an infection vesicle (Moraes et al., 2013). Post-penetration, many Colletotrichum species adopt a hemibiotrophic lifestyle, initially forming bulbous, biotrophic hyphae within living host cells, followed by a necrotrophic stage, which is characterized by host cell death and the
growth of thinner, secondary hyphae (Shen et al., 2001; O’Connell et al., 2004; De Silva et al., 2017). The CGSC species Colletotrichum gloeosporioides can also adopt a quiescent, extended, biotrophic stage, where the fungus remains dormant after initial penetration of unripe fruit, until fruit ripening triggers a destructive, necrotrophic stage of infection (Alkan et al., 2015). In addition, several CGSC isolates have been documented as endophytes, living asymptptomatically on host plants (Weir et al., 2012).

The ability to maintain a hemibiotrophic lifestyle is thought to rely on effectors, which are small, secreted proteins that are hypothesized to contribute to infection by manipulation of host cell structure and function (Kamoun, 2006). In turn, host plants have evolved to recognize these proteins, resulting in host cell death and resistance (Jones and Dangl, 2006; Dodds and Rathjen, 2010). Given this scenario, effectors are often under a diversifying selection to evade recognition by hosts. An emerging paradigm is that the need to balance diversifying selection with maintenance of housekeeping genes has driven the compartmentalization of pathogen genomes into fast-evolving genomic regions, encoding effector genes, and conserved regions, encoding housekeeping genes (Dong et al., 2015; Frantzeskakis et al., 2018).

Chromosome-level assemblies of Colletotrichum lentis and Colletotrichum higginsianum, which belong to the Colletotrichum destructivum species complex, have revealed fast-evolving genomic regions in repeat-rich, pathogenicity-associated minichromosomes (Dallery et al., 2017; Pflaumann et al., 2018; Bhadauria et al., 2019). CGSC strains also harbour minichromosomes, which can be transferred between strains (He et al., 1998). Even species that lack minichromosomes, such as Colletotrichum orbiculare, which belong to the Colletotrichum destructivum species complex, have revealed fast-evolving genomic regions in repeat-rich, pathogenicity-associated minichromosomes (Dallery et al., 2017; Pflaumann et al., 2018; Bhadauria et al., 2019). CGSC strains also harbour minichromosomes, which can be transferred between strains (He et al., 1998). Even species that lack minichromosomes, such as Colletotrichum orbiculare 104-T from the C. orbiculare species complex (Taga et al., 2015), shows signatures of compartmentalization, with distinct, gene-poor, AT-rich, transposable element (TE)-dense regions, as well as gene-rich, GC-rich, TE-poor, regions (Gan et al., 2013). Similar TE-dense, AT-rich compartments in other plant pathogenic fungi are hypothesized to have undergone repeat-induced point (RIP) mutations, a fungal defence mechanism against TEs that occurs during sexual reproduction, leading to effector gene diversification (Rouxel et al., 2011). This study aims to gain insights into the evolution of CGSC genomes by comparing multiple strains from different species with shared geographical and/or host ranges. Specifically, we aimed to identify fast-evolving, non-conserved genomic regions within the species complex. By analyzing the location of identified accessory genomic regions with variable conservation patterns, we gained insights into the mechanisms by which pathogen genomes evolve, acquire and organize pathogenicity-related genes.

Results

Virulence of Colletotrichum gloeosporioides species complex strains on strawberry is not determined by species

The virulence of 14 CGSC strains was tested against wild strawberry, Fragaria vesca, and strawberry, Fragaria × ananassa var. Sachinoka (Fig. 1, Fig. S1). This revealed that pathogenicity on these hosts is not determined by species, since a subset of Cf strains (Nara gc5, S1 and S4) cause disease symptoms on both hosts, while other strains belonging to the same species (Cf413, Cf245 and Cf415) did not. Similarly, Cs strains Cg363, CAD1 and CAD5 caused lesions on both hosts, while Cs strains CAD2 and CAD4 did not. Among the other strains tested, Ca Cg56 also caused symptoms on both hosts.

Microscopy revealed that by 3 days post-inoculation (dpi), Cf Nara gc5 had penetrated F. vesca epidermal cells via melanized appressoria and formed penetration pegs, infection vesicles and bulbous, intracellular hyphae (Fig. 1B–D) resembling biotrophic hyphae in other hemibiotrophic Colletotrichum species. At 5 dpi, thinner, secondary hyphae appearing like necrotrophic hyphae of other Colletotrichum species were observed (Fig. 1E). Similar structures were observed during infection of F. ananassa var Sachinoka (Fig. S1). In contrast, even at 7 dpi, Cf413 was still restricted to conidia, germ tube and appressorium formation on the surface of F. ananassa leaves, although the fungus remained metabolically active and able to express GFP (Fig. S1).

High quality genome assemblies of C. fructicola, C. siamense and C. aenigma

Cf Nara gc5, Cs Cg363 and Ca Cg56 were selected for PacBio sequencing as they cause disease symptoms on strawberry leaves despite belonging to different species (Fig. 1, Fig. S2). In addition, Cf CF413 was also sequenced by PacBio as a representative strain that is non-pathogenic on strawberries. All 14 genome assemblies were estimated to include at least 99.3% of the gene coding space according to BUSCO analysis of Pezizomycotina conserved genes (Table 1). Identification of the telomeric repeat TTAGGG revealed that the PacBio-sequenced genomes of Cf Nara gc5, Cf413 and Cs Cg363 each possess 10, 10 and 7 contigs of greater than 100 kb enriched with telomeric repeats at both ends (>25 copies TTAGGG/terminal 10 kb), suggesting that these assemblies include 10, 10 and 7 complete telomere-to-telomere chromosomes respectively (Fig. 2A, Fig. S3).

Despite their close phylogenetic relationship, whole genome alignments revealed multiple rearrangements in Cf Nara gc5 relative to Cf413 (Fig. 2A, Fig. S3). In contrast, CF413 shares a high degree of collinearity with the
more distantly related strains, Cs Cg363 and Ca Cg56 (Fig. S3). Cf413 contigs c11 and c12 (Cf413_c11, c12) appear to represent complete minichromosomes of less than 1 Mb each with telomeric repeats enriched at both terminals. Further, a contig syntenic to Cf413_c12 is absent from Nara gc5, demonstrating its dispensable nature (Fig. 2A). Using Cs Cg363 as a reference outgroup, Nara_c01, c05, c10, c08 and c11 appear to have originated from chromosome-level translocations after divergence from Cf413 (Fig. 2A, Fig. S3). Among these rearranged sequences, Nara_c03, c05, c08, c10 and c11 also included large regions that were unique to Nara gc5 with no synteny to Cf413. Within the Nara gc5 genome, segmental duplications of greater than 10 kb were detected between the ends of contigs and at the points of large-scale structural variations, between Nara_c03 and Nara_c10 (Fig. 2B).

Transposable elements (TEs) are known to influence genomic landscapes leading to the diversification of genomes and are often compartmentalized in dispensable, minichromosomes of other fungal plant pathogens (Dong et al., 2015). Thus, we investigated the composition of repeat elements in the genomes studied. This revealed that retrotransposons, especially LTR retrotransposons, are more abundant in CGSC strains compared to C. higginsianum (Fig. 3A). However, retrotransposon content varied among CGSC strains with LINE-type repeats showing the greatest variation, ranging from 0.08% of the genome.
Cs Cg363 genome to 0.44% of the Cf Nara gc5 and Ca Cg56 genomes. Repeat element composition was highly variable even between different strains of the same species with Cf Nara gc5 having 2.3 times more Gypsy-type LTR retrotransposons as a proportion of the total genome size compared to Cf413.

© 2021 Society for Applied Microbiology and John Wiley & Sons Ltd., Environmental Microbiology

**Table 1.** Statistics of genomes assembled in this study. The completeness of the genome assemblies was assessed by estimating the conservation of BUSCO Pezizomycotina conserved genes.

<table>
<thead>
<tr>
<th>Species</th>
<th>Strain</th>
<th>Assembly size</th>
<th>Scaffold number</th>
<th>N50 (bp)</th>
<th>L50</th>
<th>% BUSCO genes</th>
<th>Complete</th>
<th>Fragmented</th>
<th>Predicted genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>C. fructicola</em></td>
<td>Nara gc5</td>
<td>59.6 Mb</td>
<td>13</td>
<td>5,461,344</td>
<td>5</td>
<td>99.5</td>
<td>0.4</td>
<td>17,388</td>
<td></td>
</tr>
<tr>
<td>Cf413</td>
<td>56.5 Mb</td>
<td>14</td>
<td>4,902,990</td>
<td>5</td>
<td></td>
<td>99.6</td>
<td>0.4</td>
<td>15,647</td>
<td></td>
</tr>
<tr>
<td>S1</td>
<td>59.0 Mb</td>
<td>600</td>
<td>1,273,051</td>
<td>16</td>
<td></td>
<td>99.4</td>
<td>0.3</td>
<td>16,137</td>
<td></td>
</tr>
<tr>
<td>Cf415</td>
<td>56.1 Mb</td>
<td>1,194</td>
<td>400,477</td>
<td>46</td>
<td></td>
<td>99.4</td>
<td>0.5</td>
<td>15,754</td>
<td></td>
</tr>
<tr>
<td>C. siamense</td>
<td>Cg363</td>
<td>62.9 Mb</td>
<td>22</td>
<td>5,441,060</td>
<td>5</td>
<td>99.5</td>
<td>0.3</td>
<td>15,190</td>
<td></td>
</tr>
<tr>
<td>CgD1</td>
<td>58.4 Mb</td>
<td>311</td>
<td>881,809</td>
<td>22</td>
<td></td>
<td>99.6</td>
<td>0.3</td>
<td>15,120</td>
<td></td>
</tr>
<tr>
<td>CgD2</td>
<td>58.1 Mb</td>
<td>214</td>
<td>709,060</td>
<td>25</td>
<td></td>
<td>99.5</td>
<td>0.5</td>
<td>15,044</td>
<td></td>
</tr>
<tr>
<td>CgD4</td>
<td>58.2 Mb</td>
<td>199</td>
<td>823,408</td>
<td>24</td>
<td></td>
<td>99.6</td>
<td>0.3</td>
<td>15,056</td>
<td></td>
</tr>
<tr>
<td>CgD5</td>
<td>57.6 Mb</td>
<td>294</td>
<td>656,815</td>
<td>27</td>
<td></td>
<td>99.6</td>
<td>0.3</td>
<td>15,097</td>
<td></td>
</tr>
<tr>
<td><em>C. aenigma</em></td>
<td>Cg56</td>
<td>59.2 Mb</td>
<td>79</td>
<td>5,181,253</td>
<td>5</td>
<td>99.7</td>
<td>0.3</td>
<td>15,211</td>
<td></td>
</tr>
<tr>
<td><em>C. viniferum</em></td>
<td>CgW01</td>
<td>68.5 Mb</td>
<td>5,864</td>
<td>146,824</td>
<td>138</td>
<td>99.3</td>
<td>0.6</td>
<td>14,535</td>
<td></td>
</tr>
<tr>
<td><em>C. tropicale</em></td>
<td>S9275</td>
<td>55.8 Mb</td>
<td>789</td>
<td>691,924</td>
<td>24</td>
<td>99.4</td>
<td>0.4</td>
<td>14,794</td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 2.** Genome rearrangements in the Colletotrichum gloeosporioides species complex.

A. Contigs are coloured according to the contig of homology in Cf413. White regions are regions without synteny in Cf413. Red asterisks indicate contig ends with ≥25 copies TTAGGG/terminal 10 kb.

B. Rearrangements detected between all nuclear genome contigs of the *C. fructicola* Nara gc5 genome assembly. Segmental duplications were detected mostly between the ends of contigs and with the repeat-rich chromosome Nara_c11. An additional segmental duplication was detected between the middle of Nara_c03 and c10 in a region corresponding to a potential chromosomal breakpoint. Ticks represent 0.5 Mb. [Color figure can be viewed at wileyonlinelibrary.com]
In the two Cf strains, LINE-type TEs were enriched at the ends of contigs and in smaller contigs (Fig. 3B, Fig. S5C) and at chromosomal breakpoints in Nara_c03, c10, c08 and c11. Notably, Nara_c11 has distinct compartmentalization with one half that is TE-rich and gene-sparse, while the other half is TE-sparse and gene-rich (Fig. 3C). The TE-rich region of Nara_c11 also has higher TpA/ApT ratios relative to the rest of the chromosome (Fig. 3C), which is a signature of repeat-induced point mutation.

Accessory regions with variable conservation are associated with TEs and subtelomeric regions in different CGSC lineages

Dispensable minichromosomes are important for pathogenicity in other Colletotrichum species (Plaumann et al., 2018; Bhaduria et al., 2019) and some CGSC strains have demonstrated considerable karyotypic diversity and the ability to exchange dispensable chromosomes (He et al., 1998). To identify dispensable regions, reads from the sequenced CGSC strains were mapped to the Cf Nara gc5 assembly. Reads from the other 13 sequenced strains mapped to 83.75% of the Nara gc5 assembly (49.91 Mb; log10(normalized depth + 1) ≥ 0.1) indicating that most of the genome is conserved (Fig. S6). In addition, 0.82 Mb (1.38%) of the genome was conserved in all Cf strains, but not in strains from other species, potentially representing Cf-specific regions. Interestingly, 3.67 Mb of the assembly was dispensable, being present in at least one Cf strain, but not all. These dispensable regions overlapped with repeat-rich regions at the ends of Nara_c08 and Nara_c06, as well as the repeat-rich region of Nara_c11 (Fig. S6, Fig. 3B).

The dispensable region of Nara_c11 (from position 910,000 to the end of Nara_c11) is TE-rich and gene-
poor (mean TE coverage of 14.32%, median gene coverage of 34.11%, median TpA/ApT ratio of 0.77, mean number of in planta up-regulated genes of 0.12 genes/10 kb) (Table S1, Fig. 3C). This was distinct from the rest of the contig (mean TE coverage of 1.88%, median gene coverage of 52.95%, median TpA/ApT ratio of 0.60, mean number of in planta up-regulated genes of 0.40 genes/10 kb), that shared homology to 413_c06. Instead, the dispensable region was more similar to Nara_c12, which encodes a potential minichromosome in terms of mean TE coverage (12.1%), median gene coverage (17.0%) and median TpA/ApT ratio (0.65, Table S1). However, the mean number of in planta up-regulated genes in Nara_c12 is lower than that of the Nara_c11 dispensable region (0.01 genes/10 kb).

Surprisingly, 0.84 Mb of Cf Nara gc5 dispensable regions was also found to have variable conservation in Cs strains, indicating the existence of sequences that are conserved in subsets of both Cf and Cs strains. We refer to these Cf and Cs variably conserved regions as CGSC accessory regions since they are dispensable in multiple CGSC lineages. In Cf Nara gc5, CGSC accessory regions are enriched with the telomeric repeat TTAGGG (Fig. 4A) and significantly overlap with LINE and Copia TEs (P < 0.05, Fig. 4A, Table S2). Similarly, in the strawberry-pathogenic Cs strain, Cg363, 0.66 Mb CGSC accessory regions were identified and found to significantly overlap with subtelomeric regions and Copia, Gypsy and TcMar-type TEs (P < 0.05, Table S2). In contrast, only 0.14% (0.08 Mb) of the Cf413 genome was variably conserved in Cf and Cs strains and no TEs were found to significantly overlap with these regions.

CGSC accessory regions harbour effector candidate clusters that are segmentally duplicated in Cf Nara gc5

A total of 205 genes were identified in Cf Nara gc5 CGSC accessory regions. OrthoFinder analysis of 14 CGSC strains and 19 additional Colletotrichum species from other species complexes and four other non-Colletotrichum ascomycetes as outgroups was performed to examine the conservation of these genes. Based on this analysis, all but 5 of the 205 genes were assigned to 80 orthogroups. Of these 80 orthogroups, 16 include at least one predicted Nara gc5 secreted protein. Further, 12 of these 16 orthogroups encode more than one Nara gc5 paralogue, indicating they are duplicated in the Nara gc5 genome (Fig. 4B). Interestingly, based on the copy number profiles of these 12 orthogroups, the 7 CGSC strains that were non-pathogenic on strawberries clustered apart from the 7 strawberry-pathogenic strains, irrespective of their species of origin (Fig. 4B). For ease of reference, these 12 orthogroups will be referred to as Gloeosporioides species Complex Accessory Secreted Paralogue (GCASP) groups 1–12.

Of the 12 GCASP orthogroups, 7 have no known function, including one consisting of C. higginsianum effector candidate EC51a homologs (Fig. 4B). On the other hand, groups with homologs of known function include orthologues of CtNudix, a previously characterized nudix hydrolase effector from C. lentis (Bhadouria et al., 2012), Git3 glucose-receptor-related proteins, enterotoxin-related proteins and proteases (Fig. 4B). Intriguingly, 10 of the 12 GCASP orthogroups include at least one EffectorP-predicted effector.

In Cf Nara gc5, members of the accessory secreted orthogroups are organized in six paralogous gene syntenic gene clusters, with one cluster each in subtelomeric regions of Nara_c06 and c08 and two pairs of tandemly duplicated clusters in Nara_c11 (Cluster I-VI in Fig. 4C, Fig. S5 and Table S4). All six clusters were located close to LINE/Tad1-type retrotransposons (Fig. 5), including sequences with homology to CgT1, a C. gloeosporioides biotype-specific TE (He et al., 1996). In addition, Copia and Gypsy-type LTR retrotransposons including reverse transcriptase-coding sequences were identified flanking Clusters I, III and V (Fig. 5).

For insight into the evolution of these accessory secreted orthogroups, the locations of these genes in Cf Nara gc5 were examined. GCASPs 1, 5 and 8 were present only in accessory syntenic gene clusters, while GCASPs 2, 7 and 9 also have paralogues in other CGSC accessory regions (Fig. 5). Interestingly, members of GCASPs 3, 4, 6 and 10 are also present in the core genome (Fig. S6 and Table S4). In contrast, members of GCASPs 11 and 12 are present in core and CGSC accessory regions but are located outside of the conserved syntenic gene clusters. Except for GCASP3, GCASPs with homologs in core regions form monophyletic clades with sequences from Cf S1, S4, Cs Cg363, CAD1, CAD5 and Ca Cg56, which are separate from clades of core paralogues, indicating a common origin of these sequences in the three species (Fig. S6). Accessory GCASP10 sequences are absent from Ca Cg56, while accessory GCASP11 and 12 sequences are absent from Cs CAD1 and CAD5. None of the GCASP paralogues in the Nara gc5 core genome are located next to another GCASP paralogue (Table S5).

To determine if the accessory clusters are conserved in other Colletotrichum spp., GCASP-encoding genomic loci were investigated in other PacBio-sequenced genomes. This revealed that, except for GCASP3, orthologues of GCASPs 1–9, are also located in syntenic gene clusters in Ca Cg56 and Cs Cg363 (Fig. 6A, Fig. S5B). In Cs Cg363, three clusters, including a pair of tandemly duplicated
clusters, were identified (Fig. S5B). Further, in Ca Cg56, an accessory cluster is present in the subtelomeric region of 56_c05. This region appears to be lineage-independent since it is absent from homologous regions in Cf Nara gc5, Cf413 and Cs Cg363 (Fig. 6A). Read mapping depths indicate that GCASP genes are also present in multiple copies in strawberry-pathogenic strains of Cf (S1 and S4), Cs (Cg363, CAD1 and CAD5) and Ca (Cg56) (Fig. 6B). This analysis also suggests that the number of clusters identified in Cs Cg363 and Ca Cg56 is underestimated, possibly due to the difficulty of assembling these regions.

The conservation of genes associated with conserved syntenic gene clusters (GCASPs 1–10) was further assessed by PCR in an additional 51 CGSC isolates from different geographical locations and hosts (Fig. 6C, Fig. S6). In total, 26 isolates were identified with at least four related sequences. This analysis revealed that these sequences are conserved in different CGSC species, namely Cf (16 out of 40), Cs (7 out of 16), Ca (1 out of 2) and Colletotrichum theobromicola (3 out of 3), with isolates in Cf, Cs and Ca showing presence/absence polymorphisms. Further, these sequences were detected in isolates from different hosts, namely, strawberry, apple,
Limonium spp. and cassava, and originated from diverse geographic locations, namely, Canada, the United States, Israel, Japan and Vietnam (Fig. 6C).

GCASP genes are upregulated in planta and GCASP paralogues associated with conserved syntenic gene clusters tend to show correlated gene expression

To assess if GCASPs have a role in infection, we examined their expression in three different species, Cf Nara gc5, Ca Cg56 and Cs Cg363. Primers for quantitative PCR were designed for selected Cf Nara gc5 sequences (Figs S7 and S8). These results indicate that GCASP2-8 are expressed in planta in Cf Nara gc5, suggesting a potential role for these genes in infection. In addition, we calculated pairwise correlations for all tested GCASP sequences within each strain (Fig. 6D, Fig. S12). This revealed that the expression of Cf Nara gc5 GCASP paralogues in conserved syntenic gene clusters tended to have a more positive correlation to other genes within the syntenic gene clusters than to paralogues located outside these regions (Fig. 6D). Interestingly, similar tendencies for correlated gene expression of cluster-associated GCASPs especially GCASPs 2, 5–7 were also shown in Cs Cg363 and Ca Cg56 (Fig. 6D, Fig. S12).

Discussion

The compartmentalization of fungal genomes into conserved and flexible regions is thought to allow the conservation of core, housekeeping genes, while allowing other genes to evolve rapidly, maintaining their pathogenicity and avoiding host recognition. Studies have shown that the genomes of Colletotrichum spp. such as C. gloeosporioides (He et al., 1998) and C. higginsianum (Plaumann et al., 2018), include core chromosomes, and repeat-rich accessory minichromosomes that can be gained or lost with little effect on vegetative growth. Despite the demonstration of minichromosome transfer between vegetative incompatible isolates more than 20 years ago (He et al., 1998), only recently has pathogenicity been linked to the presence of specific minichromosomes in certain strains (Plaumann et al., 2018; Bhadauria et al., 2019). In these recent studies, the minichromosome sequences were shown to be virulence determinants on host plants. However, as both types of chromosomes exist in the same nucleus,
TE-dense accessory chromosomes may also play an additional role, affecting ‘core’ chromosomes through the exchange of genes and promoting recombination.

Our study shows evidence for such a role of repeat-rich minichromosomes in the CGSC. The Cf Nara gc5 chromosome, Nara_c11, which is conserved in a subset of closely related Cf strains (Nara gc5, S1 and S4), has traits of both ‘core’ and ‘accessory’ chromosomes. The ‘core’ region of Nara_c11, is more like large chromosomes such as Nara_c01 and Nara_c02 in terms of TE and gene density, whereas the dispensable region is most similar to the dispensable minichromosome, Nara_c12. Based on whole genome alignments with highly contiguous genome assemblies from other CGSC strains generated in this study, we propose that this chromosome originated from the recombination of a conserved ‘core’ chromosome with homology to Cf413_c06, with a non-conserved accessory chromosome. Furthermore, our results strongly support the exchange of genes between subtelomeric regions of ‘core’ chromosomes and the repeat-rich compartment of Nara_c11, resulting in the expansion of a group of candidate effector genes that are organized in clusters. The role of subtelomeres as compartments for effector duplication and diversification has been reported in Magnaporthe oryzae, where copies of the effector gene AvrPita are present in subtelomeric regions and accessory chromosomes (Orbach et al., 2000; Chuma et al., 2011). It is thought that the presence of these genes at these loci contribute to the observed frequent loss and mutation of AvrPita, which is recognized by hosts expressing the Pi-ta R gene (Orbach et al., 2000). Intriguingly, the presence of related effector candidate gene clusters in different subtelomeric loci in Cf and Ca indicates that these have been gained and/or lost independently in different lineages, despite likely sharing a common origin. This is reminiscent of recently described subregions of Verticillium dahliae and Verticillium tricorpus effector-coding ‘lineage-specific regions’ which have absence/presence polymorphisms in different strains of the same species, but are conserved between different Verticillium species (Depotter et al., 2019). As in the case of the CGSC accessory regions, these regions show high sequence similarity between different Verticillium species and were proposed to originate prior to speciation. However, unlike the CGSC accessory regions, they were not found to be enriched in
subtelomeric regions or to be associated with multicopy effector clusters. It is noted too that in *C. higginsianum*, four out of six detected segmental duplications were identified in regions enriched with subtelomeric repeats (Dallery et al., 2017), providing evidence of a general role of these regions in driving duplications in this genus.

Effect gene clusters have been observed in other plant pathogenic fungi, such as *Ustilago maydis* (Kämper et al., 2006). Recently, the PWL2 and BAS1 effectors that are only present in subtelomer of different core chromosomes in the rice pathogen *M. oryzae* (MoO), were found to exist side-by-side only in dispensable, minichromosomes of wheat-adapted *M. oryzae* (MoT) indicating a potential role for minichromosomes as a compartment for accumulating pathogenicity-related sequences (Peng et al., 2019). Half of the GCASPs have paralogues that are also present in core genomic regions outside of accessory regions, supporting the enrichment of effector candidates from different loci in *Colletotrichum* genomes.

While there are parallels between *M. oryzae* and the CGSC genomes, the minichromosomes of *M. oryzae* do not experience significant amounts of RIP mutations (Peng et al., 2019), unlike the dispensable region of Nara_c11 and the potential minichromosome, Nara_c12. In *Leptosphaeria maculans*, RIP suppresses expression of repeat-associated genes, although neighbouring genes can be expressed (Rouxel et al., 2011). It is tempting to speculate that in the CGSC, the presence of RIP in regions flanking accessory GCASP loci suppresses the expression of genes from any single cluster, driving the need for multiple copies for increased pathogenicity. Although other genetic effects cannot be excluded, such a dosage effect may indicate why strawberry-pathogenic CGSC strains encode multiple copies of these genes. The expression analysis also indicates that GCASP paralogues in syntenic clusters tend to show similar expression dynamics compared to non-clustered paralogues. This is in line with the hypothesis that organization of these genes in clusters may be partially driven by the advantage of co-regulation, while maintaining them in a repeat-rich environment. However, this hypothesis needs to be further investigated.

The large-scale rearrangements, which may have generated Nara_c11, not only produce genetic diversity, such as chimeric gene sequences, but also affect the 3D organization of the genome, with potential effects on gene accessibility and expression (Spielmann et al., 2018). Chromosomal rearrangements have been observed in other plant pathogenic fungi such as *V. dahliae* and also *C. higginsianum* (Jonge et al., 2013; Tsushima et al., 2019). However, *V. dahliae* and *C. higginsianum* are asexual pathogens, whereas CGSC members have known sexual morphs (Weir et al., 2012). Indeed, it is noteworthy that Cs Cg363 and Ca Cg56 are highly similar to *Cf* CI413 in terms of their genome organization despite belonging to different lineages. Therefore, the rearrangements observed may result in the reproductive isolation of the *Cf* Nara gc5 lineage from other *Cf* strains. Additionally, the fact that RIP occurs during the sexual cycle, may also cause increased TE activity in Nara gc5, although this needs to be investigated. It is noted that extensive rearrangements have also been observed during vegetative growth of the sexual fungal pathogen *Zymoseptoria tritici* (Möller et al., 2018).

Taken together, our results highlight the importance of accessory sequences in subtelomeric and repeat-rich chromosomes in increasing genome plasticity in *Colletotrichum* species and add an additional dimension to the role of these sequences in affecting the genome evolution in this group of important plant pathogens.

**Experimental procedures**

**Plant infections**

Conidia from 10-day-old cultures grown on Mathur’s media or potato dextrose agar (Nissui Pharmaceutical Co. Ltd., Japan) at 24°C for 12 h under black-light blue fluorescent bulb light/12 h dark conditions were released in autoclaved distilled water, filtered through a 100 μm cell strainer (BD Biosciences, USA), pelleted, and resuspended to the final desired concentration. Detached leaves from plants grown under long day conditions (16 h light/8 h dark) at 25°C were inoculated with 5 μl of droplets of 5 × 10⁵ conidia ml⁻¹ conidial suspensions. Infected leaves were maintained at a 100% humidity under 12 h light/12 h dark conditions at 22°C. Pathogenicity was assessed from 5 days post-inoculation (dpi). For *F. vesca*, 5- to 6-week-old plants grown under long day conditions at 25°C were spray-inoculated with conidial suspensions. Plants were kept at a 100% humidity under long day conditions prior to imaging. Three independent experiments were carried out for each condition.

**Genome sequencing, assembly and annotation**

Details of all fungal strains used can be found in Table S3. Fungi were cultured in potato dextrose broth (BD Biosciences, USA) at 24°C in the dark for 2 days. Genomic DNA was isolated using CTAB and 100/G genomic tips (QIAGEN, Germany) as described in the 1000 Fungal genomes project (http://1000.fungalgenomes.org). Details on library preparation, sequencing and assembly are in Table S6. Annotations for *Cf* Nara gc5, S1 and CF413 were generated using the BRAKER1 (Hoff et al., 2016) pipeline using hints from RNAseq reads, mapped to each genome using HISAT2 with --max-intronlen set to

© 2021 Society for Applied Microbiology and John Wiley & Sons Ltd., *Environmental Microbiology*
1000. Other assemblies were annotated using the MAKER2 (Holt and Yandell, 2011) pipeline using gmes parameters trained on Cf Nara gc5 and Augustus parameters trained on each genome using BUSCO v3 with sordariomycete conserved genes. The localizations of annotated fungal proteins were predicted using DeepLoc v1 (Almagro Armenteros et al., 2017). In addition, EffectorP v1 and 2 (Sperschneider et al., 2016, 2018) was used to predict candidate effectors. PFAM domains from Pfam release 32 (Aug 2018) were assigned by performing pfam_scan (v1.6) using the cut_ga gathering threshold and -as settings. All raw and processed sequencing data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession numbers PRJNA171218 and PRJNA476648. Additional files available at: https://github.com/pamgan/colletotrichum_genome.

Analysis of repeat elements
Repeats from Cf Nara gc5, Cf CF143, Cs Cg363, Ca Cg56 and C. higginsianum IMI 349063 were predicted using RECON and RepeatScout via RepeatModeler v open-1.0.11 (http://www.repeatmasker.org), TransposonPSI (http://transposonpsi.sourceforge.net/), LTR_retriever (Ou and Jiang, 2018) and LTRPred (Benoit et al., 2019). Only sequences longer than 400 bp and with more than five hits to the reference genome (BLASTn E-value ≤1E-15) and/or with a hit to a RepBase peptide v23.12 (Bao et al., 2015) (BLASTx E-value ≤1E-5) were retained for further analysis. Sequences with ≥80% identity were combined using vsearch. Consensus sequences were classified using RepeatClassifier and the genome was masked using the custom repeat library using RepeatMasker. ‘One code to find them all’ was used to reconstruct transposable elements (Bailly-Bechet et al., 2014). Custom scripts were used to determine the location of the telomeric repeat TTAGGG and to calculate TpA/ApT dinucleotide frequencies.

Read depth analysis
Bowtie2 v2.3.4.1 (Langmead and Salzberg, 2012) was used to map Trim Galore quality-trimmed paired-end reads from 500 bp insert libraries to assemblies using the settings: ‘--end-to-end --fr --very-sensitive -X 700 -l 400’. Duplicate reads were removed using samtools v1.8 (Li et al., 2009) with default settings. Read depths were calculated using bedtools v2.29.2 (Quinlan and Hall, 2010) coverage and normalized over genome-wide median read depths. To estimate the number of GCASP-associated clusters, reads were mapped to Nara_c06, which has only a single copy of the GCASP cluster, as described and read depths were calculated using bedtools coverage. Read depths were normalized over the median read depths of Nara_c06 from position 2,000,000 to 4,000,000, since this region is present as a single copy in the genome.

Enrichment analysis
Enrichment scores were calculated as previously described (Nègre et al., 2010). Features of interest were randomly permuted within each chromosome of origin using the pybedtools randomstats function with the ‘-chrom’ setting. Enrichment scores were calculated by normalizing the actual proportions of overlaps between the two sets of features by the median of the simulated, randomized dataset (n = 10,000). Empirical P-values were obtained by determining the fraction of simulated overlaps that are greater than the observed overlap. Telomeric repeat-rich regions were defined as 100 kb windows with TTAGGG densities greater than 95% of the genome.

Identification of large-scale structural genomic rearrangements
Whole genome alignments were conducted to identify large-scale structural genomic changes between each genome assembly using nuclmer from the mummer suite of programs with the ‘maxmatch’ setting (Delcher et al., 2003). This was followed by filtering sequence lengths in the reference sequence as defined in the text. Circos (v0.69.6) and mummerplot (v3.5) were used to visualize the genomic rearrangements. Regions of synteny were identified using SynChro (Dillon et al., 2014) from the CHRONicle package with a delta of 3 as previously described (Shi-Kunne et al., 2018). Syntenic regions were plotted using EasyFig.v2.2.2 (Sullivan et al., 2011).

Orthogroup and phylogenetic analyses
Orthofinder2 (Emms and Kelly, 2019) with default settings was used to identify orthogroups and the rooted species tree of the 33 fungi tested (Table S3). For F. verticillioides, F. oxysporum, V. dahliae and N. crassa, only T0 transcript models were included. For phylogeny of all Colletotrichum strains, 199,953 SNPs were identified from nuclmer whole genome alignments by PhaME (Shakya et al., 2020) and RAXML-ng (Kozlov et al., 2019) was used to estimate the maximum likelihood phylogeny using the best model identified by modeltest-ng (Dambra et al., 2020). The same approach was taken for analysis of each GCASP orthogroup.

Expression analysis
RNA from plants was isolated using the improved 3% CTAB3 method (Yu et al., 2012), treated with RNase-free
Acknowledgements

The authors wish to thank Dr. Takeshi Suzuki and Dr. Nanako Nakata for providing strains S1, S4, Cf413, Cf415, Cf245, Cg363 and Cg56 and Dr. Anuphon Laohavisit for critical reading of the manuscript. This work was supported by the Japan Society for Promotion of Science (Grant-in-Aid for Scientific Research 17H06172 and 15H05959 to K.S., 18H02204 to Y.T., 19K15846 to P.G., JSPS DC2 fellowship to A.T.) and the Science and Technology Research Promotion Program for the Agriculture, Forestry, Fisheries and Food industry to Y.N., Y.T. and K.S.

References


Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher’s web-site:

**Fig. S1** Infection of *Fragaria × ananassa* var. Sachinoka. (A) Infection by the 14 sequenced isolates from the *Colletotrichum gloeosporioides* species complex at 7 days post-inoculation (dpi). (B) *In planta* infection by *C. fructicola* Nara gc5 and Cf413. Conidia of both strains develop appressoria by 1 dpi. At 3 dpi, intracellular hyphae are observed in Nara gc5 infections (filled arrowheads), but not in Cf413. At 6 dpi, secondary hyphae proliferate in Nara gc5 infections but not in Cf413, although conidia remain metabolically active and able to express GFP.

**Fig. S2** Relationship of strains with other known sequenced isolates. Maximum likelihood phylogeny of *Colletotrichum* fungi based on genome-wide SNPs identified by PhaMe by nucleotide alignment. 199,953 SNPs were concatenated, and the phylogeny was estimated based on the GTR + G4 model using raxml-ng. The most likely tree out of 100 random and 100 parsimony-based trees is shown with node bootstrap support values out of 100 replicates. Trees converged after 50 bootstraps. Accession numbers for all genomes used are provided in Table S3.

**Fig. S3** Synteny among the *Colletotrichum gloeosporioides* species complex. (A) Contigs are coloured according to the contig of homology in Cf413. White regions are regions without synteny in Cf413. Red asterisks indicate contig ends with ≥25 copies TTAGGG/terminal 10 kb. (B) Dot plot representing forward (red) and reverse (blue) hits with conservation between Nara gc5, Cg363 and Cg56 against Cf413. Hits were identified using nucmer with the maxmatch settings. Only hits of greater than 10 kb are shown.

**Fig. S4** Features of *Colletotrichum fructicola* Nara gc5 contigs. Features of all nuclear genome contigs in *Colletotrichum fructicola* Nara gc5. Number of reads mapping/10 kb were normalized relative to whole genome medians. *In planta* upregulated: number of genes/10 kb that are significantly up-regulated at either 1, 3, or 6 days post-inoculation (dpi) during infection of *Fragaria × ananassa* leaves or 2 dpi *F. × ananassa* root tissue compared to 3 day-old *in vitro* hyphae. Spaces between tick marks represent 1 Mb. Effector: number of genes predicted to encode effectors/10 kb.

**Fig. S5** *Colletotrichum gloeosporioides* species complex accessory regions. (A) Synteny between six GCASP-encoding gene clusters in *C. fructicola* Nara gc5. (B) Conservation of gene order in three syntenic gene clusters encoding GCASP orthologues in *C. siamense* Cg363. Only hits of 500 bp or more and with less than 0.0001 E-value are visualized. (C–D) PCR to amplify GCASP-related sequences in 51 additional *C. gloeosporioides* species complex isolates, which is shown in Fig. 6C. A 100 bp or 1 kb ladder was used to provide size estimates. See Table S3 for details of the strains used.

**Fig. S6** GCASP orthologue phylogenies. Maximum likelihood phylogenies of GCASP homologs with paralogues outside of syntenic accessory gene clusters in *Colletotrichum fructicola* Nara gc5. GCASP1, 5 and 8 are not included as these sequences are only present in these clusters. Values at nodes are percentages of support of 1000 bootstrap replicates. Purple labels: sequences used to design primers used for qPCR analysis (Table S7 and Fig. S7). The branch length of CSAL KHX65162.1 in GCASP11 is shortened 10-fold.

**Fig. S7** Expression of GCASP homologs. Quantitative PCR of selected GCASP homologs in *Colletotrichum fructicola* (Cf) Nara gc5, *C. aenigma* Cg56 and *C. siamense* Cg363 designed based on Cf Nara gc5 sequences in (A) accessory syntenic gene clusters, (B) CGSC accessory regions and (C) regions that are conserved in all 14 sequenced CGSC strains. Expression levels relative to elongation factor 1 (*CfEF*) were assessed in conidia (*n* = 3) and in *F. vesca* infected leaves at 24, 72 and 144 h post-inoculation (hpi) (*n* = 4, 5, 5 for Cf; *n* = 5, 3, 3 for Ca; *n* = 3, 4, 3 for Cs). Data from 3–5 replicates per time point are shown with a boxplot showing the distributions of gene expression. Only sequences with at least one time point showing expression in any of the three strains are shown here.

**Fig. S8** Correlation of GCASP gene expression. Correlation plots to visualize the correlation of GCASP homolog genes expression in (A) *Colletotrichum aenigma* Cg56 and (B) *C. siamense* Cg363. Mean relative expression levels in each strain were scaled within each primer set and then pairwise correlation scores were calculated.

© 2021 Society for Applied Microbiology and John Wiley & Sons Ltd., *Environmental Microbiology*
**Table S1.** Features of contigs in the *C. fructicola* Nara gc5 genome.

**Table S2.** Enrichment of different features in CGSC accessory regions. Only PacBio-sequenced *C. fructicola* and *C. siamense* strains were analyzed.

**Table S3.** Fungal strains analyzed in this study.

**Table S4.** *Colletotrichum gloeosporioides* species complex accessory secreted paralogues (GCASP) in *C. fructicola* Nara gc5.

**Table S5.** Features of genes in the *C. fructicola* Nara gc5 genome. Extracellular sequences that are not predicted to be effectors are highlighted in green, extracellular sequences that are predicted to be effectors are highlighted in blue.

**Table S6.** Sequencing and assembly settings for *C. gloeosporioides* species complex members.

**Table S7.** Sequences of primers used.