Effector profiles distinguish formae speciales of Fusarium oxysporum

van Dam, P.; Fokkens, L.; Schmidt, S.M.; Linmans, J.J.J.; Kistler, H.C.; Ma, L.-J.; Rep, M.

DOI
10.1111/1462-2920.13445

Publication date
2016

Document Version
Final published version

Published in
Environmental Microbiology

License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/in-the-netherlands/you-share-we-take-care)

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Effector profiles distinguish *formae speciales* of *Fusarium oxysporum*

Peter van Dam,1 Like Fokkens,1 Sarah M. Schmidt,1 Jasper H.J. Linmans,1 H. Corby Kistler,2 Li-Jun Ma3 and Martijn Rep1*

1Molecular Plant Pathology, Swammerdam Institute for Life Sciences, University of Amsterdam, The Netherlands.
2United States Department of Agriculture, ARS Cereal Disease Laboratory, University of Minnesota, St. Paul, MN, USA.
3Department of Biochemistry and Molecular Biology, University of Massachusetts, Amherst, MA 01003, USA.

Summary

*Formae speciales* (ff.spp.) of the fungus *Fusarium oxysporum* are often polyphyletic within the species complex, making it impossible to identify them on the basis of conserved genes. However, sequences that determine host-specific pathogenicity may be expected to be similar between strains within the same *forma specialis*. Whole genome sequencing was performed on strains from five different ff.spp. (*cucumerinum*, *niveum*, *melonis*, *radicis-cucumerinum* and *lycopersici*). In each genome, genes for putative effectors were identified based on small size, secretion signal, and vicinity to a “miniature impala” transposable element. The candidate effector genes of all genomes were collected and the presence/absence patterns in each individual genome were clustered. Members of the same *forma specialis* turned out to group together, with cucurbit-infecting strains forming a supercluster separate from other ff.spp. Moreover, strains from different clonal lineages within the same *forma specialis* harbour identical effector gene sequences, supporting horizontal transfer of genetic material. These data offer new insight into the genetic basis of host specificity in the *F. oxysporum* species complex and show that (putative) effectors can be used to predict host specificity in *F. oxysporum*.

Introduction

The *Fusarium oxysporum* (Fo) species complex (FOSC) comprises an important group of filamentous fungi that includes plant-pathogenic strains. The species complex as a whole has a very wide host range, but individual pathogenic strains are restricted to one or a few host species. Accordingly, such strains are grouped into *formae specialiae* (ff.spp.) based on host specificity (Armstrong and Armstrong, 1981; Baayen et al., 2000). They cause vascular wilt (due to xylem colonization) or root, bulb or foot rot in over 120 plant species, including many economically important crops like tomato and cucurbits (Armstrong and Armstrong, 1981; Michielse and Rep, 2009).

Each *forma specialis* (f.sp.) of Fo consists of one or several clonal lineages (O'Donnell et al., 1998; Katan, 1999). Strains belonging to different *formae specialiae* may be more related than strains belonging to the same *forma specialis* (Kistler, 1997; Lievens et al., 2009), which implies that the host of a strain of *F. oxysporum* can not reliably be determined based on conserved gene sequences, such as the translation elongation factor 1-alpha gene (*EF-1α*) alone. Currently, disease assays remain the primary method of discriminating host range and races (defined by the capacity to infect different cultivars of a plant species and often based on presence/absence of or point mutations in effector genes) of a pathogenic Fo strain (Recorbet et al., 2003; Covey et al., 2014; Martyn, 2014).

Because disease assays are laborious and time consuming, a molecular screening method based on *forma-specialis*-specific DNA sequences is highly desirable. However, knowledge of the genetic basis of host-specificity is limited.

For successful infection of their host, pathogens often rely on effector proteins—small secreted proteins that facilitate the colonization process (Stergiopoulos and de Wit, 2009; Dodds and Rathjen, 2010; Giraldo and Valent, 2013). Infection of different plant hosts likely requires a different set of effectors, which renders these proteins potentially informative for discriminating *formae specialiae*. In this study we aimed to uncover which putative effector
genes are shared by strains belonging to each of four economically important (Kim et al., 1993b; Vakalounakis et al., 2005) cucurbit-infecting *formae speciales*, notably *cucumerinum* (Foc; cucumber), *melonis* (Fom; musk melon), *niveum* (Fon; watermelon), and *radicis-cucumerinum* (Forc; cucumber and other cucurbits), as well as the tomato-infecting *F. oxysporum* f.sp. *lycopersici* (Fol). We chose these ff.spp. because we wished to know whether strains that infect related plant species (*Cucurbitaceae*) possess similar effector suites.

Of all of the *formae speciales* mentioned above, Fol has been investigated most extensively. A horizontally transferable “pathogenicity” chromosome (chromosome 14 of Fol4287) harbours all but one of the 14 known Fol effector genes, named SIX (for Secreted In Xylem; Ma et al., 2010; Schmidt et al., 2013). These genes encode small, cysteine-rich, secreted proteins with no recognizable protein domain and at least some appear to be employed by Fol to manipulate the host's defense responses, thereby promoting the infection process (Rep et al., 2004; Gawehns et al., 2014; Ma et al., 2015). Together with chromosomes 1B, 2B, 3, 6, and 15, chromosome 14 forms the accessory genome of Fol4287. The accessory genome lacks synteny with other Fusarium species such as *F. verticillioides*, is relatively gene poor and harbours many repeats and transposable elements (TEs), characteristics that differentiate it from the core genome (Ma et al., 2010; 2013).

Analysis of the genomic context of SIX1 – SIX7 revealed an association of these genes with two TEs. A miniature impala (mimp) was found in all cases in the upstream region and mFot5 was found frequently downstream of the Open Reading Frame (ORF) (Schmidt et al., 2013). Both mimps and mFots are classes of Miniature Inverted-repeat TEs (MITEs). MITEs are short, nonautonomous DNA transposons, thought to be truncated derivatives of autonomous DNA transposons (Feschotte and Pritham, 2007; Lu et al., 2012). They contain 27–30 nucleotide terminal inverted repeats (TIRs). In the case of mimps, the sequence of these TIRs is conserved between different mimp-subfamilies (Bergemann et al., 2008). A total of 103 mimps are present in the genome of Fol4287 (Schmidt et al., 2013). The highest density was found on chromosome 14, the “pathogenicity chromosome,” with 54 mimps (21 mimps/Mb), compared to 45 (3 mimps/Mb) on the other accessory chromosomes and 4 (0.1 mimps/Mb) on the core chromosomes. Although mimp deletion experiments did not result in altered SIX gene expression (Schmidt et al., 2013), the consistent presence of a mimp in the promoter region of known Fo effector genes was successfully exploited to identify novel effector candidates (Schmidt et al., 2016). Sixteen candidate effector genes were identified close to a mimp, and the products of fourteen of these (SIX1 – SIX14) were found in the xylem sap of infected tomato plants using mass spectrometry (Houterman et al., 2007; Schmidt et al., 2013).

We made use of the association between mimps and effector genes in Fo genomes to predict the suite of putative effectors present in 59 *F. oxysporum* genomes (45 new assemblies), without the need to rely on genome annotation. We then compared the predicted “effectoromes” of these 59 different strains to determine whether strains that belong to the same f.sp. have similar effector repertoires and whether we could use presence/absence patterns of putative effector genes to predict the host range of a strain. We find that indeed strains cluster into *formae speciales* based on effector presence/absence profiles. Moreover, identification of ff.spp. based on effector genes is further strengthened by taking their sequences into account, since these are identical or highly similar within a f.sp., but often different between ff.spp.

**Results**

**Strain selection**

In order to make a well-founded evaluation of genome and effector variation within and between *formae speciales* of *F. oxysporum*, we selected divergent strains for our study. For the *formae speciales* Foc, Fom and Fon, a polyphyletic nature has been described (Jacobson and Gordon, 1990; Kim et al., 1992; 1993a; Lievens et al., 2007). Forc has a broader host range and is able to cause root and stem rot in various cucurbit species such as muskmelon and sponge gourd (*Luffa aegyptica*) (Vakalounakis, 1996; Vakalounakis and Fragkidakis, 1999; Vakalounakis et al., 2005). According to earlier reports it comprises two vegetative compatibility groups (VCGs) but these might constitute a single clonal lineage (Lievens et al., 2007). Fol is also polyphyletic, comprising at least four clonal lineages (Van Der Does et al., 2008; Ma et al., 2010).

Based on diversity in EF-1α sequence (indicative of clonal lineages, Supporting Information Fig. S1), geographical origin, and disease assays to confirm host-specific pathogenicity, 45 strains (9 Foc, 9 Fom, 9 Fon, 3 Forc, 14 Fol, one non-pathogenic, Table 1) were selected for whole genome Illumina paired-end sequencing. Combined with 14 previously generated Fo assemblies belonging to various ff.spp. (Table 2), these genomes formed the basis for our search for putative effector genes within the FOSC.

**Most cucurbit-infecting strains are host specific**

Previous studies have indicated that Fo strains belonging to f.sp. *cucumerinum* (Cafri et al., 2005) and *niveum* (Zhou and Everts, 2007) display mild cross-pathogenicity towards musk melon. To assess the level of (cross-) pathogenicity of the cucurbit-infecting strains in our collection, we conducted disease assays with all sequenced Foc, Forc, Fom, and Fon strains, plus one Fol strain (Fol029) on susceptible cultivars of cucumber,
musk melon, watermelon, and tomato, as well as on a
Foc-resistant cultivar of cucumber. The results are sum-
marized in Fig. 1.

Fom and Fon strains were highly specific to their
described host species, whereas several Foc strains
showed some degree of cross-pathogenicity, especially
towards musk melon. Disease symptoms on musk melon
plants caused by *formae speciales* other than *melonis*
(and *radicis-cucumerinum*) were generally not severe,
being limited to growth retardation or light wilting symp-
toms. Within Foc, strains Foc018, 021, 030 (all belonging
to VCG0180) and Foc011 (VCG0186) were pathogenic on
susceptible, but not on resistant cucumber plants; the oth-
er Foc strains also caused symptoms in the resistant

<table>
<thead>
<tr>
<th>Strain</th>
<th>Original designation</th>
<th>VCG</th>
<th>Race</th>
<th>Origin of strain</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foc001</td>
<td>Foc-1</td>
<td>0183</td>
<td>–</td>
<td>Japan</td>
<td>BL&lt;sup&gt;5&lt;/sup&gt;</td>
</tr>
<tr>
<td>Foc011</td>
<td>9903-1</td>
<td>0186</td>
<td>–</td>
<td>China</td>
<td>Liewens et al., 2007</td>
</tr>
<tr>
<td>Foc013</td>
<td>9904-1</td>
<td>0186</td>
<td>–</td>
<td>China</td>
<td>Liewens et al., 2007</td>
</tr>
<tr>
<td>Foc015</td>
<td>9906-3</td>
<td>0184</td>
<td>–</td>
<td>China</td>
<td>Liewens et al., 2007</td>
</tr>
<tr>
<td>Foc018</td>
<td>Afu-50(B)</td>
<td>0180</td>
<td>–</td>
<td>Crete, Greece</td>
<td>Liewens et al., 2007</td>
</tr>
<tr>
<td>Foc021</td>
<td>ATCC 16416</td>
<td>0180</td>
<td>–</td>
<td>Florida</td>
<td>Liewens et al., 2007</td>
</tr>
<tr>
<td>Foc030</td>
<td>FOCU-22P</td>
<td>0180</td>
<td>–</td>
<td>Israel</td>
<td>Liewens et al., 2007</td>
</tr>
<tr>
<td>Foc035</td>
<td>NETH 11179</td>
<td>0181</td>
<td>–</td>
<td>Netherlands</td>
<td>Liewens et al., 2007</td>
</tr>
<tr>
<td>Foc037</td>
<td>T1-213</td>
<td>0185</td>
<td>–</td>
<td>Japan</td>
<td>Liewens et al., 2007</td>
</tr>
<tr>
<td>Foc016</td>
<td>33</td>
<td>0260</td>
<td>–</td>
<td>Canada</td>
<td>Liewens et al., 2007</td>
</tr>
<tr>
<td>Foc024</td>
<td>Afu-11(A)</td>
<td>0260</td>
<td>–</td>
<td>Crete, Greece</td>
<td>Liewens et al., 2007</td>
</tr>
<tr>
<td>Foc031</td>
<td>AK-2</td>
<td>0261</td>
<td>–</td>
<td>Crete, Greece</td>
<td>Liewens et al., 2007</td>
</tr>
<tr>
<td>Fom002</td>
<td>CBS 418.90</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>Israel</td>
</tr>
<tr>
<td>Fom005</td>
<td>TX-471-1</td>
<td>0080</td>
<td>0</td>
<td>Texas</td>
<td>Zhou and Everts, 2007</td>
</tr>
<tr>
<td>Fom010</td>
<td>F-016-1</td>
<td>0082</td>
<td>1</td>
<td>Maryland</td>
<td>Zhou and Everts, 2007</td>
</tr>
<tr>
<td>Fom013</td>
<td>F-014-2</td>
<td>0082</td>
<td>2</td>
<td>Maryland</td>
<td>Zhou and Everts, 2007</td>
</tr>
<tr>
<td>Fom015</td>
<td>F-063-1</td>
<td>0082</td>
<td>2</td>
<td>Maryland</td>
<td>Zhou and Everts, 2007</td>
</tr>
<tr>
<td>Fom019</td>
<td>TX-X1D</td>
<td>0082</td>
<td>2</td>
<td>Texas</td>
<td>Zhou and Everts, 2007</td>
</tr>
<tr>
<td>Fom020</td>
<td>F-099-1</td>
<td>0083</td>
<td>2</td>
<td>Delaware</td>
<td>Zhou and Everts, 2007</td>
</tr>
<tr>
<td>Fom021</td>
<td>MD-ZE622</td>
<td>–</td>
<td>3</td>
<td>Maryland</td>
<td>Zhou and Everts, 2007</td>
</tr>
<tr>
<td>Fom037</td>
<td>NRRL 38539</td>
<td>–</td>
<td>–</td>
<td>Israel</td>
<td>Hadar and Katan, 1989</td>
</tr>
<tr>
<td>Fom004</td>
<td>Form 0122</td>
<td>0134</td>
<td>0</td>
<td>Spain</td>
<td>Schmidt et al., 2016</td>
</tr>
<tr>
<td>Fom005</td>
<td>Form 0123</td>
<td>0134</td>
<td>1</td>
<td>Spain</td>
<td>Schmidt et al., 2016</td>
</tr>
<tr>
<td>Fom006</td>
<td>Form 0124</td>
<td>0134</td>
<td>2</td>
<td>Spain</td>
<td>Schmidt et al., 2016</td>
</tr>
<tr>
<td>Fom009</td>
<td>–</td>
<td>0135</td>
<td>2</td>
<td>Israel</td>
<td>Schmidt et al., 2016</td>
</tr>
<tr>
<td>Fom010</td>
<td>–</td>
<td>–</td>
<td>1</td>
<td>Israel</td>
<td>Schmidt et al., 2016</td>
</tr>
<tr>
<td>Fom011</td>
<td>–</td>
<td>–</td>
<td>0</td>
<td>Israel</td>
<td>Schmidt et al., 2016</td>
</tr>
<tr>
<td>Fom012</td>
<td>ML2</td>
<td>0134</td>
<td>0</td>
<td>–</td>
<td>Schmidt et al., 2016</td>
</tr>
<tr>
<td>Fom013</td>
<td>–</td>
<td>0134</td>
<td>2</td>
<td>Spain</td>
<td>Schmidt et al., 2016</td>
</tr>
<tr>
<td>Fom016</td>
<td>Form26</td>
<td>0134</td>
<td>1</td>
<td>–</td>
<td>Schmidt et al., 2016</td>
</tr>
<tr>
<td>Fol002</td>
<td>WCS862/E241</td>
<td>0030</td>
<td>2</td>
<td>Netherlands</td>
<td>Mes et al., 1999</td>
</tr>
<tr>
<td>Fol004</td>
<td>IPO1530/B1</td>
<td>0030</td>
<td>1</td>
<td>Netherlands</td>
<td>Mes et al., 1999</td>
</tr>
<tr>
<td>Fol007</td>
<td>D2</td>
<td>0030</td>
<td>2</td>
<td>France</td>
<td>Mes et al., 1999</td>
</tr>
<tr>
<td>Fol014</td>
<td>LSU-3</td>
<td>0030</td>
<td>1</td>
<td>Louisiana</td>
<td>Mes et al., 1999</td>
</tr>
<tr>
<td>Fol016</td>
<td>BFOL-51</td>
<td>0031</td>
<td>1</td>
<td>Louisiana</td>
<td>Mes et al., 1999</td>
</tr>
<tr>
<td>Fol018</td>
<td>LSU-7</td>
<td>0030</td>
<td>2</td>
<td>Louisiana</td>
<td>Mes et al., 1999</td>
</tr>
<tr>
<td>Fol026</td>
<td>BRIP 14844 (M1943)</td>
<td>0030</td>
<td>3</td>
<td>Australia</td>
<td>Mes et al., 1999</td>
</tr>
<tr>
<td>Fol029</td>
<td>5397</td>
<td>0030</td>
<td>3</td>
<td>Florida</td>
<td>Mes et al., 1999</td>
</tr>
<tr>
<td>Fol038</td>
<td>CA92/95</td>
<td>0030</td>
<td>3</td>
<td>California</td>
<td>Liewens et al., 2009</td>
</tr>
<tr>
<td>Fol069</td>
<td>DFS0-23</td>
<td>0035</td>
<td>2</td>
<td>California</td>
<td>Cai et al., 2003</td>
</tr>
<tr>
<td>Fol072</td>
<td>DFS0-38</td>
<td>0031</td>
<td>2</td>
<td>California</td>
<td>Cai et al., 2003</td>
</tr>
<tr>
<td>Fol073</td>
<td>DFS0-40</td>
<td>0030</td>
<td>2</td>
<td>California</td>
<td>Cai et al., 2003</td>
</tr>
<tr>
<td>Fol074</td>
<td>DFS0-41</td>
<td>0030</td>
<td>3</td>
<td>California</td>
<td>Cai et al., 2003</td>
</tr>
<tr>
<td>Fol075</td>
<td>DFS0-62</td>
<td>0031</td>
<td>2&lt;sup&gt;d&lt;/sup&gt;</td>
<td>California</td>
<td>Cai et al., 2003</td>
</tr>
<tr>
<td>Fol076</td>
<td>14844 (M1943)</td>
<td>0030</td>
<td>3</td>
<td>Australia</td>
<td>Mes et al., 1999</td>
</tr>
<tr>
<td>FoMN14</td>
<td>MN-14</td>
<td>–</td>
<td>N.P.</td>
<td>California</td>
<td>Gale et al., 2003</td>
</tr>
</tbody>
</table>

<sup>a</sup> Foc: *Fo f.sp. cucumerinum*, Forc: *Fo f.sp. radicis-cucumerinum*, Fon: *Fo f.sp. niveum*, Form: *Fo f.sp. melonis*, Fol: *Fo f.sp. lycopersici*.

<sup>b</sup> Race designation and VCG was taken from the corresponding reference. In the case of Fol, inoculation on differential cultivars (differing in their I (immunity)-gene genotype: C32/KGS2201/PV2002MM (i-i2-i3; susceptible); GCR161 (i-i2); OT264 (KG324)/341F (i-I2); C295 (i-I2-i3); E779 (i2-i3-i3)) was used to confirm the race of each strain (N.P.: non-pathogenic on tomato).

<sup>c</sup> Foc001 (Foc-1) was obtained from Bart Lievens, Scientia Terrae Research Institute, Belgium.

<sup>d</sup> Fol075 (DF0-62) was reported to be non-pathogenic. Based on retesting we now designate it as a (weak) race 2 strain.
The genomes of the 45 selected strains were sequenced using paired-end Illumina libraries with different insert sizes, as described in the Experimental procedures. The genomes were sequenced to 60–200× coverage (Supporting Information Table S1), resulting in assemblies of 826 (Forc024) to 3867 (Foc018) scaffolds. The smallest de novo assembly has a cumulative size of 48.4 megabasepairs (Mb), including ambiguous bases (Foc037); the largest de novo assembly adds up to 57.7 Mb (Fom009). 97–98% of highly conserved protein-coding genes are estimated by CEGMA in all assemblies (Parra et al., 2009), which is similar to other F. oxysporum reference genome assemblies retrieved from Genbank (Supporting Information Table S1).

This highlights the completeness of the assemblies, particularly in core regions of the genomes. Repeat-density, RNA and DNA transposon abundance (Supporting Information Fig. S2) and GC content (Supporting Information Table S1) is largely similar between the assemblies.

The genome of the reference strain Fol4287 is the best-studied and assembled Fo genome to date, with a near-complete chromosome assembly through the use of Sanger sequencing combined with an optical map (Ma et al., 2010). Figure 2 shows the size and the level of fragmentation in each of the assessed assemblies. This analysis indicates that most strains contain a roughly equally sized core genome of about 41 Mb, with some exceptions that seem to have a slightly larger (e.g. Fom001, Fo f.sp. cubense [Focub] B2, Forl CL57, Fo47, FOSC-3a) or smaller (e.g. Fo5176, Fom010, Foc037) cumulative core genome size. Differences in reported core genome size between strains may be caused by addition of accessory regions to the ends of core chromosomes, as was observed in chromosomes 1 and 2 of Fol4287 (Ma et al., 2010) or by partial or entire loss of chromosomes. The size of the accessory regions varies considerably (between 4 and 19 Mb), but is usually comparable within a clonal lineage. To make an estimate of the complete genome size, we calculated the number of basepairs mapped against the assembly divided by the median coverage of contigs larger than 100 kb. This yielded estimated genome sizes that were considerably larger than the assembly (Supporting Information Table S1), indicating that indeed large-scale duplication events are a relevant factor. This also explains Fol4287’s assembly sticking out in size (Fig. 2) due to its relatively complete assembly, most notably in the aforementioned duplicated regions.

Foc, Fom, Fon, and Fol are all polyphyletic within the FOSC

To determine to what extent host specificity is polyphyletic among the selected strains, we inferred a
phylogeny based on a concatenated alignment of 1195 conserved core genes (see Experimental procedures for more detail). The resulting tree, depicted in Fig. 2, is congruent with previously published phylogenetic analyses (Baayen et al., 2000; Lievens et al., 2007). We observe three major clades, of which clade 1 corresponds to a separate phylogenetic species (Laurence et al., 2014). We find that ff.spp. are generally grouped into distinct clonal lineages and that these lineages are represented in distinct clades or subclades in the tree. One notable exception is Forc, of which all three sequenced strains group into a single clonal line. This is in accordance with previous reports in which 68 different Forc strains all belonged to a single Random Amplified Polymorphic DNA (RAPD) and Amplified Fragment Length Polymorphism group (Vakalounakis and Fragkiadakis, 1999; Vakalounakis et al., 2005; Lievens et al., 2007).

We find that Foc, Fom, and Focub are each represented in two different major clades, Foc and Focub even in two different phylogenetic species. Foc strains are distributed over five clonal lineages: one in clade 1 and four in clade 2. Focub (poorly represented in this study with only three strains) is grouped into two lineages, one that belongs to clade 1 and one strain that is placed in clade 2. Fom strains cluster into three clonal lineages, two that belong to different subclades in clade 2 and one strain (Fom001/NRRL24604) that belongs to clade 3. Fon and Fol are confined to a single major clade, but within this clade, we find that different clonal lines belong to distinct subclades. Fon is clustered into three distinct groups within clade 2. Fol harbours four

<table>
<thead>
<tr>
<th></th>
<th>cucumber</th>
<th>cucumber</th>
<th>musk melon</th>
<th>watermelon</th>
<th>tomato</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Paraiso,</td>
<td>(Melen,</td>
<td>(Charantais-T,</td>
<td>(B. Diamond,</td>
<td>(C32,</td>
</tr>
<tr>
<td>mock</td>
<td>susceptible)</td>
<td>resistant)</td>
<td>susceptible)</td>
<td>susceptible)</td>
<td>susceptible)</td>
</tr>
<tr>
<td>Foc029</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foc001</td>
<td>++ (5/5)</td>
<td>++ (5/5)</td>
<td></td>
<td></td>
<td>++ (5/5)</td>
</tr>
<tr>
<td>Foc013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foc015</td>
<td>++ (5/5)</td>
<td>+ (5/5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foc017</td>
<td>+ (5/5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foc019</td>
<td>++ (5/5)</td>
<td>+ (5/5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foc021</td>
<td>++ (5/5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foc023</td>
<td>++ (5/5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foc025</td>
<td>++ (5/5)</td>
<td>++ (5/5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foc016</td>
<td>++ (7/1)</td>
<td></td>
<td>+++ (5/5)</td>
<td>++ (3/4)</td>
<td></td>
</tr>
<tr>
<td>Foc024</td>
<td>+++ (5/5)</td>
<td>+ (5/5)</td>
<td>+++ (4/4)</td>
<td>+ (3/5)</td>
<td></td>
</tr>
<tr>
<td>Foc031</td>
<td>+++ (5/5)</td>
<td>+ (5/5)</td>
<td>+++ (5/5)</td>
<td>++ (2/4)</td>
<td></td>
</tr>
<tr>
<td>Fom001</td>
<td></td>
<td></td>
<td>+++ (0/5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fom004</td>
<td></td>
<td></td>
<td>+++ (5/5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fom005</td>
<td></td>
<td></td>
<td>+++ (5/5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fom006</td>
<td></td>
<td></td>
<td>+++ (5/5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fom009</td>
<td></td>
<td></td>
<td>+++ (5/5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fom010</td>
<td></td>
<td></td>
<td>+++ (5/5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fom011</td>
<td></td>
<td></td>
<td>+++ (5/5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fom012</td>
<td></td>
<td></td>
<td>+++ (1/5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fom013</td>
<td></td>
<td></td>
<td>+++ (5/5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fom016</td>
<td></td>
<td></td>
<td>+++ (0/5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fon002</td>
<td></td>
<td></td>
<td></td>
<td>+++ (5/5)</td>
<td></td>
</tr>
<tr>
<td>Fon005</td>
<td></td>
<td></td>
<td></td>
<td>+ (0/5)</td>
<td></td>
</tr>
<tr>
<td>Fon010</td>
<td></td>
<td></td>
<td></td>
<td>+ (4/5)</td>
<td></td>
</tr>
<tr>
<td>Fon013</td>
<td></td>
<td></td>
<td></td>
<td>+++ (5/5)</td>
<td></td>
</tr>
<tr>
<td>Fon015</td>
<td></td>
<td></td>
<td></td>
<td>+++ (4/5)</td>
<td></td>
</tr>
<tr>
<td>Fon019</td>
<td></td>
<td></td>
<td></td>
<td>+++ (4/5)</td>
<td></td>
</tr>
<tr>
<td>Fon020</td>
<td></td>
<td></td>
<td></td>
<td>+++ (5/5)</td>
<td></td>
</tr>
<tr>
<td>Fon021</td>
<td></td>
<td></td>
<td></td>
<td>++ (6/6)</td>
<td></td>
</tr>
<tr>
<td>Fon037</td>
<td></td>
<td></td>
<td></td>
<td>+++ (5/5)</td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 1.** Host-specificity of cucurbit-infecting strains used in this study. Disease index (0–4) was scored in multiple bioassays under controlled conditions in the greenhouse after two weeks (cucurbit) or three weeks (tomato). All compatible interactions were tested at least twice in individual bioassays. Between brackets is the number of surface-sterilized hypocotyls from which *F. oxysporum* outgrowth was observed on CDA after 4 days.
clonal lineages in clade 3, with the majority of the sequenced strains belonging to the same clonal lineage as the reference strain Fol4287, corresponding to VCG 30 (VCG 0030). In summary, all ff.spp. except Forc are polyphyletic, consisting of several clonal lines.

**Effector prediction pipeline yields 104 effector candidates in 59 Fo genomes**

If host preference of a strain is largely determined by its suite of effectors, we expect strains that infect the same host to have a similar effector repertoire, even if they...
belong to different phylogenetic clades. To determine the effector repertoire of different strains, we exploited the association of mimps with effectors in Fo. Each genome was scanned for the presence of mimp TIRs. Subsequently, two methods were used for ORF identification: (i) the sequence 2500 bp downstream of the mimp IR was translated in the three possible reading frames and ORFs bigger than 25 codons were extracted; (ii) AUGUSTUS 3.1 gene prediction software was run on the 5000 bp downstream of the TIR. In both cases, the threshold for distance from TIR to ATG was set to 2000bp. Method ii allowed for the prediction of putative effectors with a short first exon, like SIX10. Supporting Information Fig. S3 depicts a summary of the method.

To evaluate the influence of assembly fragmentation on the number of effectors found, we compared the number of mimps, mimp-inverted repeats and associated candidate genes identified in two different short read assemblies (based on Illumina HiSeq and Ion Proton sequencing) to the reference assembly of Fol4287 (Sanger/optical mapping). We found that the fragmented assemblies performed only slightly worse than the reference assembly (Supporting Information Tables S1 and S3) in terms of effector candidate prediction, although clearly less complete mimps were found in the Ionproton assembly (21 vs. 40).

When we applied the extended pipeline to all 59 genomes, we found in total 2242 ORFs that met the selection criteria (>25 aa; <300 aa; SignalP value >0.550; distance to closest mimp TIR <2500 bp). To reduce redundancy, we grouped these ORFs into gene families based on a self-BLAST search. This resulted in a final set of 201 mimp-associated gene families encoding small secreted proteins. This set includes all previously described SIX genes of Fol except SIX5 (Six5 returns a SignalP value of 0.444) and SIX12 (Six12 does not have a signal peptide). Subsequently, we ran BLAST2GO and InterProScan to find gene ontology (GO) terms and protein domains in our set of genes, which resulted in a functional annotation for 88 out of 201 genes. Fifty-four genes returned no significant BLAST hit and 59 returned a best BLASTX hit with a hypothetical protein (13 of these had at least one associated GO term).

For those genes that had a homolog in any of the annotated Fo genomes, we also report the corresponding gene id in this strain (Supporting Information Table S2, column L).

Due to the nature of the method, we sometimes picked up fragments of genes, originating from only the first exon or from an internal gene region. We checked for each candidate if the sequence could be aligned to the start site of the corresponding gene sequence from the Broad annotation in order to update the gene model. In 83 out of 201 candidates the reported Broad gene model could be used. For the others, we did not alter the gene model. To grom the list of candidates, we removed duplicate records (5×), very short protein products (<35aa after signal peptide removal; 23×); homologs or fragments of SIX genes (17×: all 14 SIX genes with correct gene models from Fol were manually added to the list instead, resulting in 215 records in Supporting Information Table S2) and sequences with an unlikely signal peptide based on visual inspection (67×), despite a positive (>0.550) detection of a signal peptide by SignalP. These included TEs, transcription factors and integral membrane proteins. This resulted in a final set of 104 candidate effectors (Supporting Information Table S2).

Predicted proteins in the set with recognisable domains include FOVG_19456, which contains a LysM domain, potentially protecting the fungal cell wall or preventing chitin-triggered immunity in plants (de Jonge and Thomma, 2009; Stergiopoulos and de Wit, 2009; Jiang et al., 2014), Rapid Alkalization Factor-like protein 33 (FOIG_11494) and several secreted enzymes with predicted peptidase (FOXG_17323, FOVG_19376, FOXB_07727), polygalacturonase (FOTG_18786, FOWG_18016), glycoside hydrolase (FOCG_17303), carbonic anhydrase (FOMG_18585), or peroxidase (FOVG_19731) activity that could play a role in nutrient acquisition or suppressing plant defences.

**Strains with the same host-specificity cluster together based on effector content**

Having identified the combined putative “effectorome” of the 59 Fo strains, we examined which putative effectors are shared amongst members of the same f.sp., multiple f.f.spp., all analyzed Fo genomes or with other *Fusarium* species. Presence of a candidate effector gene in a genome was defined as having at least one blastn hit with an e-value ≤1E-03 and an identity score (number of identical nucleotides divided by the query length) of at least 30%. This gave rise to a binary “effector-barcode” for each genome. Hierarchical clustering of these presence/absence patterns showed clear grouping of strains with the same host-specificity (Fig. 3), with the exception of Foc that split into two clades. We conclude that simple scoring presence/absence patterns of candidate effector genes reflects which host can be infected.

Candidate effector genes were then grouped based on their absence/presence in *formae speciales*. Genes in groups A and B are present in virtually all Fo genomes and consists mainly of secreted enzymes and hypothetical proteins, or genes without a BLASTX annotation. Group C is composed of genes that are present in most pathogenic, but not nonpathogenic strains, including some SIX genes (marked in red at the top of Fig. 3). SIX8 forms an exception in this group, it is also found as an intact ORF in FoMN14 while it is incomplete in the FOSC-3a genome assembly, situated at the end of a contig. Group D contains genes that are mostly present in cucurbit-infecting strains.
and group E represents Fol genes, including previously described Fol-specific effector genes SIX2, 3 and 5 (Lievens et al. 2009).

The putative effector repertoires for strains infecting the same host are remarkably similar to each other, suggesting a highly related (at least on the level of presence-absence) shared accessory genome. Interestingly, the cucurbit-infecting strains form a supercluster distinct from the other formae specialies, implying that a significantly overlapping set of effectors is associated with the ability to infect cucurbits. This might explain why there is some degree of cross-pathogenicity of Foc towards, for example, melon plants but not from Fom to any of the cucurbits tested. Between cucurbit-infecting formae specialies, different sets of effectors are present (particularly in groups C and D) that may be involved in specific pathogenicity towards their host plant.

While still grouping with the other cucurbit-infecting formae specialies, the three (highly similar) Forc strains lack a number of putative effectors present in the other cucurbit-infecting strains, predominantly in group D. Their effector pattern is most similar to that of Foc strains not belonging to VCG0180.

Both FolAVR1 (Houterman et al., 2008) and AVRFOOM2 (Schmidt et al., 2016) show the expected presence-absence pattern (absence of AVR1 in race 2 and 3 Fol strains and absence of AVRFOOM2 in race 2 Fom strains). Recognition of Avr2 in Fol is evaded by sequence difference rather than deletion of the gene and is present in all Fol strains in the figure.

Strikingly, the hierarchical clustering of genomes based on putative effectors also grouped strains belonging to the same clonal lineage within each polyphyletic forma specialis, irrespective of race designation (Fig. 3). This is a signature of vertical inheritance, indicating that the common ancestor of such a clonal lineage at a relatively recent point in time obtained the genetic information needed to infect a new host. Similar selection pressures on different clonal lineages likely resulted in the emergence of the same races in different lineages. This is most clear in the cases of Fol and Fom, where strains belonging to races 1, 2, 3, and races 0, 1, 2, respectively, do not cluster according to race, even though the core genome (Fig. 2), as well as the accessory genome (Fig. 3) are strongly similar within each clonal lineage.

Both Foc and Fon are clearly subdivided into two groups, whose accessory genomes are more different from each other compared to other polyphyletic forma specialies. Fon strains 005, 010, 013, 020, 021 form one clonal lineage (Fig. 2) and their putative effector profiles are also nearly identical, while the profiles of strains Fon002, 037, 015, 019 are also highly similar, but lack several effector candidates in group F. The same was found in the case of Foc, where Foc018, 021 and 030, belonging to one clonal lineage, are highly similar but quite different from the other Foc strains. This suggests relatively large differences in the accessory genomes between clonal lineages in Fon and Foc.

Two non-pathogenic strains were included in this study: Fo47, a biocontrol strain (Aimé et al., 2013) and FoMN14 that was isolated from diseased tomato plants from the same field as MN25 but turned out to be nonpathogenic towards tomato (Gale et al., 2003). These two strains group close to each other and contain relatively few candidate effector genes (47 in Fo47 and 46 in FoMN14, compared to an average of 69 in pathogenic strains). The clinical strain FOSC-3a as well as Fo f.sp. radicis-lycopersici CL57 were found in the same branch. These strains also have a relatively low number of candidate effector genes. Based on core-genome sequence (Fig. 2), as well as effector content, Focub strains B2 (race 4) and II5 (tropical race 4; TR4) are nearly identical to each other with only two differentials. This is in line with Fo f.sp. cubense TR4 being a single clonal lineage (Ordonez et al., 2015). Since these two strains do not differ that much from strain N2 in their effector pattern (race 1; six presence/absence polymorphisms), the Cavendish banana resistance breaking TR4 may have evolved from race 1 (or race 2, for which a time of this study no genome sequence was available).

To see whether clustering of strains belonging to the same forma specialis depends on effector gene sequences and not DNA sequences close to mimps per se, we scored presence/absence of 2.5 kb regions downstream of a mimp TIR. The resulting figure (Supporting Information Fig. S4) shows that although some forma specialies form a single group, much more fragmentation is found and the clustering generally follows the core genome tree (Fig. 2). For example, Fom001 is found among the Fol strains and Foc011 and 013 are close to Focub strains.

For comparison, the pipeline was also run on 29 other Fusarium (non-Fo) genome assemblies of strains for which a whole-genome assembly is available. In most cases, 0–4 mimps or mimp-TIRs could be identified and only in the species F. avenaceum, F. fujikuroi, and F. nygamai, 1–4 candidate ORFs with a secretion signal downstream of a mimp TIR were found (Supporting Information Table S4). This illustrates the specificity of this method for the FOSC and suggests that specifically within this species complex, mimps developed their contextual association with effector genes.

Sequence comparison of SIX genes shows evidence of horizontal transfer

Strong grouping of host-specificity can already be seen based on presence/absence clustering. We next
wanted to find out if a higher resolution could be achieved by comparing sequence types of (candidate) effectors. We aligned the homologs of SIX1-14 and looked at their phylogeny. Remarkably, different clonal lines within the same f.sp. possess identical or highly similar sequence variants of each of the 14 described SIX genes.

The SIX1 coding sequence (Rep et al., 2004; van der Does et al., 2008; Fig. 4A) was detected in Fol, Fom, Foucon and Fop, with strains within each f.sp.
possessing a (nearly) identical sequence for this gene. All Fon strains except Fon002 and 037 possess a SIX1 homolog that is interrupted by a Home1-TE at the same position, resulting in a pseudogene. Form009, 010 and 011 have a second copy of SIX1 that seems to have been duplicated in the ancestor of this clonal lineage, indicative of vertical inheritance. For SIX6 (Gawehns et al., 2014), five sequence variants were found: Fol, Foc, Forc-Fon-Fom, Fon015/019, and Focub (Fig. 4B). This suggests that the cucurbit-infecting formae speciales Forc (single lineage), Fom (strains from multiple lineages), and Fon (one lineage) have acquired this sequence from the same source. SIX13 (Schmidt et al., 2013) harbours more sequence variation, but again all Fon, Fol, and Focub strains have identical sequences within the respective f.s.p. (Fig. 4C). Interestingly, two different SIX13 sequence variants were found in Foc and Fon, with one of the two variants being similar to the Forc/SIX13 homolog and the other most closely related to Fom/SIX13.

SIX5, 7, 10, and 12 were only found in Fol, with no sequence variation. Earlier studies have shown that some of these genes can be used as Fol-markers (Lievens et al., 2009). SIX3 (AVR2) was also only encountered in Fol, albeit with some sequence variation in strains of Fol that evade I-2 recognition, thus becoming race 3 (V4). Of these genes can be used as Fol-markers (Lievens et al., 2009).

Most effector genes are expressed during invasive growth

To assess whether the candidates identified were indeed expressed during plant infection, we performed RNA sequencing on RNA extracted from infected plant roots 10 days after inoculation and five-day-old mycelium from an in vitro liquid KNO₃ medium culture. We assessed eight strains belonging to the ff.sp. Foc, Fom, Fon, Forc, Fol. On average, 60.6% (± 3.9% S.D.) of the candidates showed evidence for in planta transcription, and 36.7% (± 10.7% S.D.) of the candidates qualified as being clearly expressed in planta, but with no or little expression in vitro (Supporting Information Fig. S6). In Fol4287, 12 out of 13 SIX genes fall inside the latter category (Supporting Information Table S5).

Discussion

Using a bioinformatics pipeline, putative effectors were identified in 45 newly sequenced strains and 14 publicly available Fo genomes. Effector candidates were identified by searching for a small TIR sequence of miniature Impala TEs. In this way, 104 effector candidates were identified across all 59 F. oxysporum genomes. Hierarchical clustering of the presence/absence patterns of these sequences led to grouping of strains that largely coincided with host specificity. This indicates that the accessory genome and specifically the effector genes residing in these regions can be used to identify formae speciales in the FOSC. Clustering of f.s.p. was observed even without taking sequence divergence or copy number variation into consideration, underlining the robustness of the method.

We have improved an existing method to identify putative and real effector genes in F. oxysporum (Schmidt et al., 2013; 2016) by incorporating a gene prediction module into the pipeline. This at least partially solved the issue of effector genes that were not recognized due to a short first exon, so that SIX10 was now identified, too. The remaining known false negatives were SIX5 and SIX12, which were not identified due to a low SignalP score or the absence of a signal peptide, respectively. False positives in the list, like duplicate records, very short gene fragments, transcription factors, TEs, membrane proteins and hypothetical proteins with an unlikely signal peptide were identified and removed from the list. The effect of fragmentation of the genome assembly on the number of effectors identified was limited, and by combining the effector candidates from multiple short-read Illumina genomes into one large list, a sufficient resolution was achieved. The level of fragmentation in the Illumina assemblies, did, however, prevent an analysis of the degree to which putative effector genes form clusters in the same genomic regions and possible synteny of such regions between strains.

Other methods of computationally recognizing effector genes include the selection of (predicted) proteins that meet some or all of the following requirements: small size, secreted, high number of cysteines, traces of diversifying selection and in planta transcriptional induction (Sperschneider et al., 2015a;2015b). The advantage of the method presented here is that genome annotation is not necessary and that relevant contextual genome location is taken into account. However, while the association between effector genes and mimps has been demonstrated to hold true in many cases (e.g., SIX genes, AVRFOX2; Schmidt et al., 2013; 2016), it is possible that several effector candidates have been missed.
For Fom, Fon, Forc, and Fol a clear and unambiguous clustering could be observed. Foc separated into two clades. Previously, RAPD has been applied as a way of distinguishing between cucurbit-infecting formae speciales (Vakalounakis and Fragkiadakis, 1999; Wang et al., 2001; Vakalounakis et al., 2004; Lievens et al., 2007). Aside from being laborious, difficult to replicate and time-consuming, the technique focuses on both the core as well as the accessory genome, which results in different RAPD profiles for each clonal lineage. Vakalounakis and Fragkiadakis (1999) showed that by transforming RAPD patterns into a binary data matrix and calculating the genetic distance value based on presence or absence of RAPD bands, a similar dendrogram could be generated as shown in our study (Fig. 3), again with VCG 0180 of Foc (here represented by Foc018, 021, and 030) clustering in a

Fig. 4. Identical SIX gene sequence types are found in strains belonging to polyphyletic formae speciales, suggesting a combination of vertical and horizontal inheritance of these genes. A MUSCLE alignment was made of the nucleotide sequence of (a) SIX1 (881 nt), (b) SIX6 (727 nt), and (c) SIX13 (943 nt). Phylogeny was inferred using PhyML with 100 bootstrap iterations and plotted with mid-point rooting. Branches with most parsimonious bootstrap partitions below 50% were collapsed; values ≥ 50% and < 100% are indicated in red; 100% are not indicated. A coloured circle (wilting) or square (root and shoot rot) representing the strain’s forma specialis was plotted on the leaves of the dendrogram.

For Fom, Fon, Forc, and Fol a clear and unambiguous clustering could be observed. Foc separated into two clades. Previously, RAPD has been applied as a way of distinguishing between cucurbit-infecting formae speciales (Vakalounakis and Fragkiadakis, 1999; Wang et al., 2001; Vakalounakis et al., 2004; Lievens et al., 2007). Aside from being laborious, difficult to replicate and time-consuming, the technique focuses on both the core as well as the accessory genome, which results in different RAPD profiles for each clonal lineage. Vakalounakis and Fragkiadakis (1999) showed that by transforming RAPD patterns into a binary data matrix and calculating the genetic distance value based on presence or absence of RAPD bands, a similar dendrogram could be generated as shown in our study (Fig. 3), again with VCG 0180 of Foc (here represented by Foc018, 021, and 030) clustering in a
different group than the other Foc clonal lines. Both Forc (68 strains) and Fom (three related strains from Spain) cluster as single clades in the RAPD study. The fact that clustering based on all DNA sequences associated with mimps resulted in a more fragmented clustering showing more traces of the core phylogeny (Supporting Information Fig. S4) illustrates the importance and relevance of clustering based on ORFs for secreted proteins close to a mimp in *F. oxysporum*.

For about 60% of the identified effector genes, evidence for transcription was found in planta 10 days post inoculation (Supporting Information Fig. S5), indicating that many of these genes might play a role during infection. Some candidate genes might be expressed earlier in the infection process and are therefore not found to be expressed at this timepoint.

Focub strains II5, N2 and B2 and *Brassicaceae*-infecting strains PHW808 (f.sp. *conglutinans*) and Fo5176 (no f.sp. reported, but likely f.sp. *conglutinans*) also clustered together, but these *formae speciales* were not sampled extensively enough in this dataset to cover the complete variation present. In both Fon and Foc, and more subtly also in the other *formae speciales*, footprints of vertical inheritance could be seen. Strains from a single clonal lineage usually contain (near) identical putative effector presence patterns (Fig. 3) as well as SIX gene sequences (Fig. 4 and Supporting Information Fig. S5).

Foc018, 021, 030 as well as Foc011 (all lacking SIX6) were pathogenic on *C. sativus* cv. Paraiso, but not on “Foc-resistant” *C. sativus* cv. Melen plants; the other Foc strains also caused symptoms in *cv. Melen*. Since they are not in the same core genome clade (Fig. 2), the cause of this race differentiation must have developed at least twice independently, or passed on horizontally. Foc013, which is in the same phylogenetic species as Foc011 (clade 1), was aggressive on both cucumber cultivars.

Because strains belonging to the same *formae specialis* cluster together, the presence of certain putative effectors can indicate with high confidence to which *formae specialis* a strain belongs. For instance, a distinct cluster of putative effector genes is associated with cucurbit-infecting strains (Fig. 3, cluster D). Furthermore, all of the ff.spp. represented by multiple strains share sets of differentiating effector genes. In the case of Fol, it was shown in previous studies that presence of effector genes can be indicative of the host; a PCR screen with SIX1, SIX2, SIX3, and SIX5 primers resulted in a 100% success rate in identifying Fol strains in a large collection containing fifteen other Fo ff.spp. (Van Der Does et al., 2008; Lievens et al., 2009). From the clustering shown in Fig. 3 we conclude that the suite of putative effector genes present in the genomes of cucurbit-infecting strains, as well as other monophyletic or polyphyletic *formae speciales* can be applied diagnostically if the genome sequence of a novel unknown strain is available. Members of the same *forma specialis* might share parts of their (accessory) genome involved in specific pathogenic interaction with their host. An important requirement is the completeness of sampling to be able to make a well-supported choice of *forma specialis*-specific marker loci (e.g., group E for Fol, genes in groups C and D for Fom and Foc, cluster F for Fon). Such marker loci are essentially the smallest possible set of effectors that is shared by all strains of a f.sp. and absent (at least as a set) in all other strains. This would greatly improve host range identification, something that so far has proved to be a difficult objective to achieve. Applicability to other *formae speciales* will need to be investigated, for example by sequencing genomes of multiple differential strains of the same *forma specialis* in existing culture collections. A similar clustering result would be expected for other host ranges besides tomato and cucurbits. This would allow *forma specialis*-identification for crops where Fusarium wilt and root and shoot rot is a pressing problem.

**Experimental procedures**

**Plant lines and fungal strains**

The following plant cultivars were used: *Cucumis sativus* cv. Paraiso (susceptible to Foc), *C. sativus* cv. Melen (resistant to Foc), *C. melo cv. Cha-T* (susceptible to Fom), *Citrullus lanatus* cv. Black Diamond (susceptible to Fom) and *Solanum lycopersicum* C32 (susceptible to Fol).

Strains of Fo were selected based on core genome divergence, pathogenicity testing and geographical origin (Table 1). *F. oxysporum* was grown at 25°C in the dark on Czepek Dox agar (CDA, Difco) plates containing 100 mg/l penicillin and 200 mg/l streptomycin.

**Pathogenicity testing**

Pathogenicity testing was performed using the root dip method (Wellman, 1939). In short, conidia were isolated from five-day-old cultures in NO3-medium (0.17% yeast nitrogen base, 3% sucrose, 100 mM KNO3) by filtering through miracloth (Merck; pore size of 22–25 μm). Spores were centrifuged, resuspended in sterile MillQ water, counted and brought to a final concentration of 10^7 spores/ml. When the first true leaves were emerging, seedlings were uprooted, inoculated, individually potted and kept at 25°C in the greenhouse. Two weeks (cucurbits) or three weeks (tomato) after inoculation, disease was scored using a disease index ranging from 0 to 4 (Supporting Information Methods S1) based on published methods (Rep et al., 2004; Vakalounakis et al., 2004; Pavlou and Vakalounakis, 2005).

From each bioassay combination, pieces of hypocotyl were collected randomly from five plants, surface sterilized with 96% ethanol and a slice was cut of each piece to be placed on CDA plates containing 100 mg/l penicillin and 200 mg/l streptomycin. After four days of incubation at 25°C in the dark, *Fusarium* outgrowth was assessed and scored.
**DNA isolation, genome sequencing, and assembly**

Genomic DNA was isolated as described in detail in Supporting Information Methods S2. Library preparation of insert size 550 bp and Illumina HiSeq 2000 (Foc001, 011, 013, 015, 021, 035, 037, Forc016, 024, 031) and Illumina HiSeq 2500 (Fon and Foc018, 030) paired-end sequencing was performed at Keygene N.V. ( Wageningen, The Netherlands). All Fom genomes, “Fol4287-Illumina” and Fol007 were sequenced and assembled as described in Schmidt et al. (2016) at the Beijing Genome Institute (BGI, Hong Kong), using multiple insert libraries. “Fol4287-Ionproton” was sequenced at the University of Amsterdam (MAD Dutch Genomics Service and Support Provider) using an Ion Proton™ Sequencer (Thermo Fisher Scientific). Fol paired-end genome sequencing of the remaining strains was performed using Illumina HiSeq 2000 180 bp insert libraries at the Broad Institute of Harvard and MIT.

Genome assemblies were made with CLC workbench v8.0, as described in Methods Supporting Information S3. We used nucmer (with -maxmatch) from the MUMmer package (Delcher et al., 2002) to align all genome sequences to the reference genome, that of Fol4287. We designated scaffolds as “core” if its best match in terms of bp that could be aligned was a scaffold that is part of a core chromosome in Fol4287 (with a minimum overlap of either query or the reference scaffold of 30%). All scaffolds that did not meet this criterion were designated “accessory.”

In our assemblies, repeats that are longer than the length of our reads will typically be collapsed into a single sequence; the cumulative length of individual contigs in our assemblies is likely to be an underestimate of true genome size. To improve upon this estimate we determine the coverage by taking the median read depth of all contigs > 100 kb, assuming that these large contigs contain relatively few repeats and divide the number of reads that are mapped to our assembly by the coverage to obtain a putative genome size (Supporting Information Table S1).

**RNA isolation and transcriptome sequence analysis**

For transcriptome sequencing, 10-day-old melon Cha-T (Fom), cucumber Paraizo (Foc, Forc), watermelon Black Diamond (Fon), and tomato C32 (Fol) seedlings were inoculated with conidia of strain Fom001 by dipping the roots in the spore suspension for 5 min; roots of infected plants were harvested ten days after inoculation and flash-frozen in liquid nitrogen. Additionally, mycelium from five-day-old in vitro KNO₃ cultures was harvested. Total RNA was extracted as described previously (Schmidt et al., 2013). cDNA synthesis, library preparation (200 bp inserts) and Illumina sequencing was performed at BGI (Foc013, Fom001, Forc016, Forc031, Fol4287) and Keygene N.V. (Fon019, Fon020, Fon006).

The obtained reads were mapped against the set of candidate effector sequences retrieved from the genome sequence of each of the strains (only sequences starting with “ATG” were used) with CLC workbench v8.0 with both length and similarity settings set to 0.9. The number of unique mapped reads was divided by the gene length and then multiplied by 1000 to find the number of reads per kb (RPK). This value was then used to calculate relative expression (RE) compared with EF1 alpha expression. Genes with five or more reads were distinguished as showing evidence for transcription. If a RE of >2% was found, the gene was considered to be strongly expressed (Supporting Information Table S5).

**Data access**

The Whole-Genome Shotgun projects for the newly sequenced strains of Foc, Forc, Fom, and Fon have been deposited at Genbank under the BioProject PRJNA306247. Raw sequence data have been deposited into the Sequence Read Archive under the accession numbers SRP067515 (Foc, Fon, Forc, Fol007, Fol4287), SRP042982 (Fom), and SRP002087 (Fol). All publically available genome sequences that were used were obtained from Genbank and the Broad Institute of Harvard and MIT (http://www.broadinstitute.org). Detailed information on these strains and their respective accession numbers are listed in Table 2.

**Phylogenetic analysis**

We searched for homologs of 15,956 Fol4287 core genes (including introns) in the sequences of the other genomes using BLASTN with default parameters. We selected all sequences that overlap > 70% with the query sequence and more than 80% identity to the query. We then selected query genes for which we find only a single hit in each species, leaving us with 1194 genes. We used ClustalO (Siévers et al., 2014) to construct a multiple sequence alignment for each query and a custom python script to concatenate these alignments. This alignment was trimmed using trimAl -strictplus. We used RaxML (with -T2 -N 100 -m GTRGAMMAL -x 1234567 -p 123 -f a) to infer a phylogeny with 100 bootstrap replicates (Stamatakis, 2014).

For the phylogenetic analysis of the genes encoding SIX proteins, we extracted the sequences by BLASTN (with -evalue 0.001 -task = “blastn”), then manually curated them to extract complete coding sequences. We used MUSCLE (default values) to construct a multiple sequence alignment for each query, followed by PhyML to infer a phylogeny with 100 bootstrap replicates (Guindon and Gascuel, 2003).

**Putative effector identification**

Prediction of candidate effectors was performed as described (Schmidt et al., 2013) on each genome using a custom Python script. Briefly, the genome was searched for a consensus sequence of the mimp inverted repeats (IRs), “TT[T]TTGCA[N]CCACTGNN.” Subsequently, two methods were used for ORF identification: (i) the sequence 2500 bp downstream of the mimp IR was translated in the three possible ORFs and ORFs bigger than 25 codons were extracted; (ii) AUGUSTUS 3.1 gene prediction software was run on the 5000 bp downstream of the mimp IR with options “species = fusarium singlestrand = true” (Stanke et al., 2006). In both cases, the threshold for distance from TIR to ATG was set to 2000 bp. Method (ii) allowed for the
prediction of putative effectors with a short first exon, like SIX10. The identified putative ORFs were submitted to a local instance of SignalP4.1 with the option “-u 0.550” for stringent signal peptide prediction (Petersen et al., 2011). All ORFs that met the criteria were collected. After concatenating the putative effectors of the 59 genomes, they were BLASTed against themselves in order to identify and remove redundant ORFs. For each set of homologous putative effectors (criteria: e-value: 1E-03; percent identity: >60%; alignment length: >60%), the longest entry was selected. A total of 201 candidates were saved and used for the downstream hierarchical clustering analysis. For functional annotation of the found candidate genes, we performed BLASTX and InterProScan using BLAST2GO v3.1.3 with standard settings (Conesa et al., 2005).

Hierarchical clustering

Screening for presence of the putative effectors collected from multiple genomes was done by conducting a BLASTN search (e-value 0.001, -task = “blastn”) on each genome assembly. Presence of a candidate effector gene in a genome was defined as having at least one blast hit with an e-value ≤ 1E-03 and an identity score (number of identical nucleotides in the correct position in the alignment divided by the query length) of at least 30%. A binary damatrix was generated containing presence (“1”) or absence (“0”) of each candidate in each genome. This table was used as input for hierarchical clustering performed in R, using a Jaccard binary distance matrix and average linkage.

Acknowledgements

The authors are grateful to K. Broz for Fol DNA isolations and P.M. Houterman for Fol DNA isolations and restesting of Fol races. They thank B. Lievens, D.J. Vakalounakis, and K. Everts for providing their cucurbit-infecting F. oxysporum strains. Furthermore, we are grateful to ENZA seeds and Rijk Zwaan for kindly providing cucumber, melon, and watermelon seeds. This work was supported by the Horizon programme of The Netherlands Organisation for Scientific Research (project 93512007) of The Netherlands Genomics Initiative and comparison. 

References


**Supporting information**

Additional supporting information may be found in the online version of this article at the publisher’s web-site.