The *Venturia inaequalis* effector repertoire is expressed in waves, and is dominated by expanded families with predicted structural similarity to avirulence proteins from other fungi.

Mercedes Rocafort¹, Joanna K. Bowen², Berit Hassing¹, Murray P. Cox³, Brogan McGreal², Silvia de la Rosa¹, Kim M. Plummer⁴, Rosie E. Bradshaw³ and Carl H. Mesarich¹.*

¹Laboratory of Molecular Plant Pathology/Bioprotection Aotearoa, School of Agriculture and Environment, Massey University, Private Bag 11222, Palmerston North 4442, New Zealand.

²The New Zealand Institute for Plant and Food Research Limited, Mount Albert Research Centre, Auckland 1025, New Zealand.

³Bioprotection Aotearoa, School of Natural Sciences, Massey University, Private Bag 11222, Palmerston North 4442, New Zealand.

⁴Department of Animal, Plant and Soil Sciences, La Trobe University, AgriBio, Centre for AgriBiosciences, La Trobe University, Bundoora, Victoria 3086, Australia.

*Corresponding author: Carl H. Mesarich (c.mesarich@massey.ac.nz). Postal address: School of Agriculture and Environment, Massey University, Private Bag 11222, Palmerston North 4442, New Zealand.

Abstract

**Background:** Scab, caused by the biotrophic fungus *Venturia inaequalis*, is the most economically important disease of apples worldwide. During infection, *V. inaequalis* occupies the subcuticular environment, where it secretes virulence factors, termed effectors, to promote host colonization. However, in the presence of corresponding host resistance proteins, these effectors are recognized as avirulence determinants to activate plant defences. To develop durable control strategies against scab, a better understanding of the molecular mechanisms underpinning subcuticular growth by
V. inaequalis is required. Likewise, more information is needed on the role that effectors play in activating, suppressing or circumventing resistance protein-mediated defences.

Results: We generated the first comprehensive RNA-seq transcriptome of V. inaequalis during colonization of apple. Analysis of this transcriptome revealed five in planta gene expression clusters or waves corresponding to three specific infection stages: early, mid and mid-late infection. Early infection was characterized by genes encoding plant cell wall-degrading enzymes (PCWDEs) and proteins associated with oxidative stress responses. Mid infection was characterized by genes encoding transporter proteins. Finally, mid-late infection was characterized by genes encoding PCWDEs and effector candidates (ECs), with most ECs belonging to expanded protein families. To gain insights into function, AlphaFold2 was used to predict the tertiary structures of proteinaceous ECs. Strikingly, many ECs were predicted to have structural similarity to avirulence proteins from other plant-pathogenic fungi, including members of the MAX, LARS, ToxA and FOLD structural effector families. In addition, several other ECs, including an EC family with amino acid similarity to the AvrLm6 effector from Leptosphaeria maculans, were predicted to adopt a KP6/ferredoxin-like fold. Thus, proteins with a KP6/ferredoxin-like fold may represent yet another structural family of effectors shared among plant-pathogenic fungi.

Conclusions: Our study reveals the transcriptomic profile underpinning subcuticular growth by V. inaequalis, and reinforces the idea that fungal effectors share a limited number of structural folds. Importantly, our study also provides an enriched list of V. inaequalis ECs that can be investigated for roles in virulence and avirulence, and raises the possibility that apple resistance proteins can be engineered to recognize EC family folds or host components targeted by multiple EC family members.

Keywords
Venturia inaequalis; apple scab fungus; biotrophic subcuticular pathogen; effectors and effector families; virulence and avirulence; RNA-seq transcriptome; AlphaFold2 protein tertiary structure predictions.
Background

Fungal pathogens are responsible for some of the most devastating diseases of crop plants worldwide, causing large economic losses, and threatening both food production and security [1]. Resistance to these pathogens is largely governed by the plant immune system, and is based on the recognition of invasion patterns (IPs) by plant immune receptors [2]. At the plant cell surface, many of these immune receptors are pattern recognition receptors (PRRs) of the receptor-like protein (RLP) or receptor-like kinase (RLK) classes, which recognize conserved IPs known as microbe-associated molecular patterns (MAMPs). Following the recognition of these MAMPs, plant defence responses are initiated that slow or halt growth of the fungal pathogen [3-5]. To circumvent or suppress these defence responses, plant-pathogenic fungi must secrete a collection of virulence factors, termed effectors, into their hosts [6-8]. These effectors are typically proteinaceous, but can also be secondary metabolites and small RNAs [6]. In some cases, however, effectors can be recognized as IPs by extracellular PRRs or intracellular nucleotide-binding leucine-rich repeat (NLR) immune receptors, collectively referred to as resistance (R) proteins, to activate plant defences. Such effectors are termed avirulence (Avr) effectors because their recognition often renders the fungal pathogen unable to cause disease.

Scab (or black spot), caused by Venturia inaequalis, is the most economically important disease of apple (Malus x domestica) worldwide [9, 10]. Under favourable conditions, this disease can render the fruit unmarketable and cause a yield reduction of up to 70% [11-13]. During biotrophic host colonization, V. inaequalis exclusively colonizes the subcuticular environment without penetrating the underlying plant epidermal cells [9, 14, 15]. It is here that the fungus develops specialized infection structures, known as stromata and runner hyphae [9, 14]. Stromata give rise to asexual conidia, but are also likely required for nutrient acquisition and effector secretion [9, 14, 16], while runner hyphae enable the fungus to radiate out from the initial site of host penetration, acting as a base from which
additional stromata can be differentiated [14]. At the end of the season, in autumn, *V. inaequalis* switches to saprobic growth inside fallen leaves, where it undergoes sexual reproduction [9].

Scab disease is largely controlled through fungicides, which greatly accelerate the development of fungicide resistance in *V. inaequalis* [17]. Scab-resistant apple cultivars, developed through the incorporation of *R* genes, represent a more sustainable disease control option [12]. However, as races of *V. inaequalis* have been identified that can overcome resistance in apple mediated by single *R* genes [12, 18], it is likely that multiple *R* genes will need to be stacked in these cultivars to provide durable disease resistance. For this to be successful, prior knowledge of the molecular mechanisms used by the fungus to overcome *R* gene-mediated resistance, including an understanding of how these mechanisms impact effector function and pathogen fitness, will be required. So far, however, there have been no publications describing the cloning of Avr effector genes from *V. inaequalis*. Crucially, the genomes of multiple *V. inaequalis* isolates have been sequenced [11, 19-24] and bioinformatic studies have identified a large catalogue of effector candidates (ECs) from which Avr effectors can be identified [20]. Most of these ECs belong to expanded families comprising multiple members [20]. These include ECs of the AvrLm6-like family, which share amino acid sequence similarity to AvrLm6 [16], an Avr effector from the fungal pathogen *Leptosphaeria maculans* (blackleg of canola) [25], as well as ECs of the Ave1-like family [20], which share amino acid sequence similarity to Ave1, an Avr effector from the fungal pathogen *Verticillium dahliae* (Verticillium wilt disease) [26].

To develop durable control strategies against scab disease, a better understanding of the molecular mechanisms underpinning subcuticular growth by *V. inaequalis* is also required. To date, subcuticular growth has been largely understudied, even though it is exhibited by many plant-pathogenic fungi, including other crop-infecting members of the *Venturia* genus [14, 27-29], as well as, for example, *Rhynchosporium* (scald disease of graminaceous plants) [30, 31] and *Diplocarpon* (e.g. rose black spot) [32, 33]. In recent years, host colonization by plant-pathogenic fungi has been studied by transcriptomic analysis [34-36]. However, comprehensive transcriptomic studies focusing on the subcuticular parasitic strategy are not available. Indeed, while previous expression data derived from
interactions between \textit{V. inaequalis} and susceptible apple have been published [13, 20], these data have only been based on a limited number of infection time points with no biological replicates. In this study, we provide the first comprehensive transcriptomic analysis of \textit{V. inaequalis} during colonization of susceptible apple, and identify infection-related gene expression waves associated with general fungal and effector biology. Using recent advances in \textit{de novo} protein folding algorithms, we also show that the EC repertoire of \textit{V. inaequalis} is dominated by expanded families with predicted structural similarity to Avr proteins from other plant-pathogenic fungi. Collectively, this study furthers our understanding of subcuticular growth by \textit{V. inaequalis}, and provides an enriched list of ECs from this fungus that can be investigated for potential roles in virulence and avirulence.

**Results**

**Different stages of subcuticular host infection by \textit{V. inaequalis} display distinct gene expression profiles**

To investigate changes in \textit{V. inaequalis} gene expression during host colonization, we set up an infection time course involving detached leaves of susceptible apple cultivar \textit{M. x domestica} ‘Royal Gala’. This time course involved six \textit{in planta} time points (12 and 24 hours post-inoculation [hpi], as well as 2, 3, 5 and 7 days post-inoculation [dpi]), and was compared to growth of the fungus in culture at 7 dpi.

Analysis of leaf material from the infection time course by bright-field microscopy revealed that, at 12 hpi, conidia of \textit{V. inaequalis} had germinated and formed appressoria (Fig. 1A). These appressoria were produced at the end of germ tubes or directly from conidia and were frequently surrounded by extracellular matrix (ecm) (Fig. 1A). At 24 hpi, primary hyphae had developed, indicating that subcuticular colonization was underway. Then, by 2 and 3 dpi, \textit{V. inaequalis} had developed subcuticular runner hyphae from stromata, and stromata had undergone a rapid expansion in size through non-polar division (Fig. 1A). By 5 dpi, fungal biomass had accumulated extensively in the subcuticular environment and more stromata had started to develop from runner hyphae (Fig.
Often, these stromata had formed conidiophores, from which conidia had developed (Fig. 1A). Finally, at 7 dpi, conidia of *V. inaequalis* had started to rupture through the plant cuticle (Fig. 1A) and, at this time point, the first macroscopic olive-brown lesions, indicative of heavy sporulation, were apparent. This represented the end of the biotrophic infection stage, as detached apple leaves had started to decay after this time point.

Inspection of the RNA-seq data revealed that the percentage of reads mapping to the *V. inaequalis* genome [20] increased as the infection time course progressed (Additional file 1: Table S1). Furthermore, the biological replicates clustered robustly within time points, and a clear distinction between the early and mid-late infection stages, as well as between the *in planta* and in-culture growth conditions, was observed (Fig. 1B).
**Fig. 1** Microscopic and transcriptomic profile of *Venturia inaequalis* during infection of detached leaves from susceptible apple cultivar ‘Royal Gala’. A. Microscopic evaluation of *V. inaequalis* during colonization of apple leaves. Infection structures observed by bright-field microscopy were stained with aniline blue. Infected leaves are representative of material used in the RNA-seq transcriptome sequencing experiment. c, conidium; gt, germ tube; ap, appressorium; ecm: extracellular matrix; ph: primary hyphae. B. Principal component analysis (PCA) of RNA-seq data from *V. inaequalis* during colonization of apple leaves and in culture on the surface of cellophane membranes overlaying potato dextrose agar. Four biological replicates were used for each *in planta* time point and the in culture growth condition. hpi: hours post-inoculation; dpi: days post-inoculation.
Genes of *V. inaequalis* are expressed in temporal waves during infection of apple leaves

We set out to identify which genes of *V. inaequalis* are up-regulated during host colonization, when compared to growth in culture. For this purpose, we updated the current gene catalogue for isolate MNH120 [20] to increase the total number of annotated genes, including those that encode ECs, which are notoriously difficult to predict in fungi (Additional file 2: Fig. S1). Here, our approach was to predict as many genes as possible and, as a consequence, it is expected that many spurious genes were included in the annotation. In total, 24,502 genes, excluding splice variants, were predicted and, of these, 3,563 were up-regulated at one or more *in planta* time points (*p* value of 0.01 and log₂-fold change of 1.5) (Additional file 3).

The total set of *in planta* up-regulated genes was used to identify host infection-specific gene expression clusters, henceforth referred to as temporal waves. Here, all expression data was scaled across all samples (Z-score) to visualize the gene expression deviation from the overall mean. For clustering, the parameters were set to specifically identify the minimum number of gene clusters for which a distinct gene expression profile could be observed (Fig. 2A). In total, five distinct host infection-specific gene expression waves (Fig. 2A), representing three separate infection stages (Fig. 2B), were identified (Additional file 4: Fig. S2). Waves 1 (274 genes) and 2 (550 genes) represented early infection (Figs. 2A and 2B). Genes of these two waves demonstrated their highest level of expression at 12 hpi, with expression largely plateauing (wave 1) or trending downwards (wave 2) throughout the remaining infection time points (Figs. 2A and 2B). Wave 3 (453 genes) represented mid infection. This wave contained genes that had a peak level of expression at 2 dpi (Figs. 2A and 2B). Finally, waves 4 (959 genes) and 5 (1,318 genes) represented mid-late infection (Figs. 2A and 2B). Genes of wave 4 displayed their lowest level of expression at 12 and 24 hpi, with expression strongly increasing through 2 and 3 dpi, peaking at 5 dpi (Figs. 2A and 2B). A similar profile was also observed for genes of wave 5, but with expression strongly increasing from 3 dpi and peaking at 7 dpi (Figs. 2A and 2B).
To determine which biological processes are overrepresented in the five host infection-specific gene expression waves, gene ontology (GO) (Fig. 2B, Additional file 5: Fig. S3, Additional file 6) and protein family (Pfam) enrichment analyses (Additional file 7) were performed. Early infection (waves 1 and 2) was characterized by GO terms associated with high metabolic activity and responses to oxidative stress. In contrast, mid infection (wave 3) was characterized by GO terms associated with transmembrane transport, while mid-late infection (waves 4 and 5) was mostly characterized by GO terms associated with carbohydrate metabolism (Fig. 2B).
Fig. 2 Genes of *Venturia inaequalis* up-regulated during infection of susceptible apple cultivar ‘Royal Gala’, relative to growth in culture, belong to one of five distinct expression waves. A. Expression profile of the five distinct expression waves at 12 and 24 hours post-inoculation (hpi), as well as 2, 3, 5 and 7 days post-inoculation (dpi), relative to growth in culture (7 dpi). B. Heatmap of all *V. inaequalis* genes up-regulated *in planta* when compared with growth in culture. Gene expression data are scaled rlog-normalized counts across all samples (Z-score), averaged from four biological replicates. Genes up-regulated *in planta* were clustered using hclust method Ward.D2 and Euclidian distance. Coloured block labels on the left indicate gene expression waves. Numbers in brackets indicate number of genes per wave. Black asterisk indicates the wave significantly enriched for ECs (*p* value: 3.644e-14). Grey boxes indicate enriched gene ontology (GO) terms per wave highlighted in the main text. ECs: effector candidates; PCWDEs: plant cell wall-degrading enzymes; *p*: value.

**Plant cell wall-degrading enzyme (PCWDE)- and transporter-encoding genes of *V. inaequalis* are up-regulated during colonization of the subcuticular environment**

To understand the nutritional requirements of *V. inaequalis* during colonization of the subcuticular environment, we investigated the expression profile of genes that putatively encode PCWDEs and transporters during the five host infection-specific waves (Figs. 3 and 4). Within the early infection wave 1, only four genes encoding PCWDEs, mostly of the auxiliary activity (AA) enzyme class, that are thought to aid the activity of subsequent classes of PCWDEs, were expressed. Within wave 2, an increase in the overall number of genes encoding PCWDEs, mostly of the AA and carbohydrate esterase (CE) classes (e.g. CE5 cutinases), was observed. Within wave 3 (mid infection), only six PCWDE-encoding genes were expressed, representing a mixture of AA and CE enzymes. Later, within wave 4, a striking increase in the number of genes encoding PCWDEs, mostly of the glycoside hydrolase (GH) class (e.g. GH28s), was observed. Finally, during wave 5 (mid-late infection), genes encoding GH and AA enzymes predominated.

The expression profile of genes encoding sugar and nitrogen transporters also differed during early and mid-late infection. During early infection (waves 1 and 2), only two genes encoding sugar
transporters, together with seven genes encoding nitrogen-associated transporters, were up-regulated. Most of the up-regulated genes encoding nitrogen transporters belonged to the oligopeptide (OPT) transporter family. Many OPT transporter-encoding genes were found to be co-expressed highly with aspartyl protease-encoding genes (R > 0.8). During mid infection (wave 3), only two genes encoding sugar transporters were expressed and genes encoding nitrogen transporters started increasing expression. During mid-late infection (waves 4 and 5), there was also a striking increase in the expression of genes encoding sugar and nitrogen transporters (Fig. 4).

Taken together, the differences observed between the expression of genes encoding PCWDEs and transporters across the different waves likely reflect the differences in the nutritional requirements by *V. inaequalis* during early and mid-late subcuticular host colonization.
Fig. 3 Plant cell wall-degrading enzyme (PCWDE)-encoding genes of *V. inaequalis* up-regulated during infection of susceptible apple cultivar ‘Royal Gala’, relative to growth in culture. **A.** Proportion of *in planta* up-regulated PCWDE-encoding genes in each host infection expression wave. **B.** Heatmap of PCWDE-encoding genes up-regulated during infection.
regulated in planta during early infection waves 1 and 2. C. Heatmap of PCWDE-encoding genes up-regulated during mid infection (wave 3) and mid-late infection (waves 4 and 5). Labels on the right indicate expression profiles corresponding to genes encoding carbohydrate-active enzymes (CAZymes). Block labels on the left indicate gene expression wave. Gene expression data are scaled rlog-normalized counts across all samples (Z-score), averaged from four biological replicates. hpi: hours post-inoculation; dpi: days post-inoculation. AA: auxiliary activity; GH: glycoside hydrolase; CE: carbohydrate esterase; PL: polysaccharide lyases; CBM: carbohydrate-binding module. Numbers in brackets indicate number of genes per wave.

**Fig. 4** Sugar- (A.) and nitrogen-associated transporter (B.) genes of *Venturia inaequalis* up-regulated during infection of susceptible apple cultivar ‘Royal Gala’, relative to growth in culture. Gene expression data are scaled rlog-normalized counts across all samples (Z-score), averaged from four biological replicates. hpi: hours post-inoculation; dpi: days post-inoculation. Sugar transporters in A. (PF0083), nitrogen associated transporters in B. OPT: oligopeptide transporter (PF03169); POT: proton-dependent oligopeptide transporter (PF00854); Ammonium T: ammonium transporter (PF00909); aa permease: amino acid permease (PF13620); Formate/nitrate T: formate/nitrate transporters (PF01226).
Genes encoding components of the predicted *V. inaequalis* effector repertoire are predominantly expressed in the early and mid-late waves

To determine whether the predicted *V. inaequalis* effector repertoire is expressed in temporal waves during host colonization, the expression of genes encoding putative effectors was examined, including genes responsible for the biosynthesis of secondary metabolites and genes encoding proteinaceous ECs. A total of 26 secondary metabolite gene clusters were predicted using antiSMASH v6.0.1; however, only seven were up-regulated during host-colonization. Of these, two were non-ribosomal peptide synthetase-like (NRPS-like) clusters, two were type I polyketide synthase (T1PKS) clusters, two were indole clusters, and one was a siderophore cluster (Additional file 8: Fig. S4). All of the up-regulated secondary metabolite genes, except the lowly expressed T1PKS1 core biosynthetic gene *g15330*, were expressed during the mid-late infection waves (waves 4 and 5) (Additional file 8: Fig. S4).

Ribosomally-synthesized and post-translationally modified peptide (RiPP) gene clusters cannot be predicted using antiSMASH. Therefore, RiPP biosynthetic clusters present in the *V. inaequalis* genome were manually identified by searching for genes that encode a protein with a DUF3328 (or DUF3328-like) domain, as this domain is known to be associated with RiPP biosynthesis [37]. In total, nine dikaritin RiPP clusters were identified, with each cluster containing one or more DUF3328 protein-encoding genes as well as a RiPP precursor gene (Additional file 9: Fig S5). Three of the dikaritin RiPP clusters were made up of genes that were highly expressed and up-regulated during host-colonization. Of these, all were expressed during the mid-late infection waves (4 and 5). The most highly expressed dikaritin precursor gene (*g7830*, from the dikaritin-2 cluster) was expressed during wave 4. This gene corresponds to the previously identified gene, *cellophane-induced (Cin)* 3, which was formerly considered to be a repeat-containing EC protein [14, 20].
Dikaritin cluster 2

Dikaritin cluster 4

Dikaritin cluster 9

**Fig. 5** Expression of ribosomally-synthesized and post-translationally modified peptide (RiPP) dikaritin gene clusters from *Venturia inaequalis* that are up-regulated during colonization of susceptible apple cultivar ‘Royal Gala’, relative to growth in culture. Gene expression data are DESeq2-normalized counts, averaged from four biological replicates per infection time point *in planta*, with error bars representing standard deviation (hpi, hours post-inoculation; dpi, days post-inoculation) and during growth in culture. Genes marked with bold characters encode putative DUF3382 proteins. Genes marked as ε putatively encode a protein with a DUF3382 domain or were annotated as a major-facilitator superfamily protein. Genes marked with * putatively encode a dikaritin precursor peptide. The gene marked with N encodes a protein with no characterized functional domain.

Proteinaceous ECs were investigated including genes that encode small secreted proteins of ≤400 amino acid residues or larger secreted proteins with an effector prediction by EffectorP v3.0. In total, 1,369 genes encoding putative proteinaceous ECs were identified (Additional file 10: Fig. S6).
Amino acid similarities among these ECs were investigated using BLASTp, and the ECs grouped into families using spectral clustering. Based on this analysis, 759 of the proteinaceous EC were grouped into 118 families ranging in size from 2 to 75 members. Of these, 32 families were found to be expanded with five or more members. Only 610 of the proteinaceous ECs were singletons that did not belong to any family (Additional file 11). Of the genes encoding proteinaceous ECs, 686 were both up-regulated in planta and expressed during one of the early or mid-late infection waves. More specifically, early infection (waves 1 and 2) contained transcripts from 139 EC genes, while the mid-late infection stage (waves 4 and 5) contained transcripts from 485 EC genes (Fig. 6). The early infection waves were mostly composed of singletons, while the mid-late infection waves were mostly composed of expanded families (Fig. 6).

Most of the EC genes expressed during early infection (waves 1 and 2) encoded proteins that lacked predicted functional domains. However, two genes from wave 2 encoded proteins with an Egh16-like virulence factor domain and had amino acid sequence similarity to the appressorium-specific Gas1 effector from the rice blast fungus Magnaporthe oryzae [38], hereafter named the Gas1-like family. Three of four identified expanded gene families expressed during the early infection waves also encoded proteins with predicted functional domains. These were family 8 with a stress-upregulated Nod19 domain (PF07712), hereafter named the Nod19 family (39 members), family 13 with a hydrophobic surface-binding protein A (HsbA) domain (PF12296), hereafter named the HsbA family (15 members), and family 37 with a common fold in several fungal extracellular membrane proteins (CFEM) domain (PF05730) [39], hereafter named the CFEM family (16 members).

The Nod19 family is the third most expanded EC family in V. inaequalis, and most Nod19 family members were encoded by genes that were both up-regulated in planta and expressed during early host colonization (waves 1 and 2) (Fig. 6 and Additional file 12: Fig. S7A). Alignment of all members from this family with Nod19 domain-containing proteins from Medicago truncatula (barrel medic) and the potato late blight oomycete Phytophthora infestans revealed that they all contained the conserved motif H-X-H-X-GG-X18/20-Y (Additional file 12: Fig. S7B). The HsbA and CFEM families are mostly...
expressed during early host-colonization (Fig. 6). However, each of these families had genes with contrasting expression profiles, with most members expressed during early infection (waves 1 and 2) and a few members expressed during mid-late infection (waves 4 and 5) (Additional file 13: Fig. S8).

Interestingly, the \textit{HsbA} genes expressed during early host colonization appeared to be co-expressed with cutinases (R > 0.8).

Another family with contrasting expression among members during host colonization was the \textit{Cin1} family (Additional file 13: Fig. S8) [40]. This family contains the \textit{Cin1} gene (\textit{g8385}), which encodes a cysteine-rich protein with eight repeats, and two \textit{Cin1-like} genes (\textit{Cin1-like 1}: \textit{g10529}, \textit{Cin1-like 2}: \textit{g13013}), which encode smaller proteins with only one repeat. The Cin1 family has no homology to known proteins and is specific to the \textit{Venturia} genus (Additional file 14: Fig. S9). \textit{Cin1} was expressed during mid-late infection (wave 4), and it was the most highly expressed gene during mid-late host colonization. In contrast, the \textit{Cin1-like} genes were expressed during early infection (wave 2) (Additional file 14: Fig. S9).
A. Early EC waves (n=139)
Singletons (64) + ECs from a family (75)

B. Mid-late EC waves (n=485)
Singletons (93) + ECs from a family (392)
Fig. 6 Genes encoding proteinaceous effector candidates (ECs) of *Venturia inaequalis* are expressed in waves during colonization of susceptible apple cultivar ‘Royal Gala’. Early infection (A.) waves 1 and 2 are composed of 139 EC-encoding genes, and the mid-late infection (B.) waves 4 and 5 are composed of 485 EC-encoding genes. Black annotations on the right indicate that the proteins encoded by these genes belong to an expanded EC family (≥5 members), with brackets highlighting large families and families with amino acid sequence similarity to avirulence (Avr) effector proteins from other plant-pathogenic fungi. hpi: hours post-inoculation; dpi: days post-inoculation.

The mid-late infection waves (waves 4 and 5) possessed the largest number of transcripts from genes encoding proteinaceous ECs, with 485 genes expressed from 2 to 7 dpi. Indeed, wave 4 was enriched in ECs (*p* value: 3.644e-14). Notably, most expanded families, including those with amino acid similarity to effectors or Avr effectors from other plant-pathogenic fungi, were encoded by genes that were expressed during waves 4 and 5 (Fig. 6, Additional file 15: Table S2). For example, the *AvrLm6-like* family [16] had 21 up-regulated genes *in planta*, with most expressed during waves 4 and 5 (Fig. 7A). Likewise, the *Ave1-like* family [20] had eight genes up-regulated *in planta*, with most expressed during wave 4 (Fig. 7A). One of the genes in this family, *g18550*, was the second most highly expressed EC gene during mid infection, and encoded a protein with the highest amino acid similarity to Ave1 (39.3% pair wise identity). Another example was the *Ecp10-like* family, which encoded proteins with amino acid sequence similarity to *Ecp10-1*, an Avr EC from the tomato leaf mold fungus *Cladosporium fulvum*. The *Ecp10-1* family represented the second most expanded family in *V. inaequalis* with 61 members and, of these, 45 were up-regulated during waves 4 and 5 (Fig. 7A). Similarly, a family of six proteins with amino acid sequence similarity to Ecp39 from *C. fulvum* [41], hereafter named the Ecp39-like family, were predominantly encoded by genes that were expressed during waves 4 and 5 (Fig. 7A). Finally, an *Ecp6-like* gene (singleton), which encoded a protein with amino acid similarity to the Ecp6 effector from *C. fulvum*, was very highly expressed during wave 4 (Fig. 7A).
Interestingly, most expanded EC families encoded by genes expressed during waves 4 and 5 did not have amino acid sequence similarity to other proteins. Of these, family 1 was the most expanded EC family in *V. inaequalis*. This family was encoded by 75 genes and, of these, 54 were up-regulated *in planta* (Fig. 8). Likewise, family 2 was encoded by 32 genes, and of these, 28 were up-regulated *in planta* (Fig. 8). In most cases, one or a few family members were very highly expressed, when compared with the other family members, during host colonization (Fig. 8).

Several expanded effector candidate families of *V. inaequalis* have predicted structural similarity to *Avr* effector proteins from other plant-pathogenic fungi

To gain insights into the putative function of proteinaceous ECs, we predicted their tertiary structures using AlphaFold2 [42], which has been successfully benchmarked against pathogen effectors of known tertiary structure [43, 44], and then investigated these structures for similarity to proteins of characterized tertiary structure (and in some cases, function) using the Dali server [45]. This analysis was specifically performed on the most highly expressed member from each EC family (which we referred to as the main family member) as well as each singleton, expressed during the host infection-specific waves. In total, the tertiary structure was confidently predicted for the most highly expressed member of 75% of EC families (94 families) and 64% of EC singletons (118 singletons) (Additional file 16).

We investigated the predicted tertiary structures of EC families from *V. inaequalis* that had amino acid sequence similarity to known effectors and characterized or candidate *Avr* effector proteins from other plant-pathogenic fungi (Additional file 17: Table S3) (Fig. 7). During early infection, the most highly expressed *Gas1*-like gene (Fig. 7A), as well as the similar *Gas1* effector from *M. oryzae*, were predicted to encode a protein with an immunoglobulin-like β-sandwich fold. These structures were similar to chitin-binding lytic polysaccharide monooxygenase (LPMO) proteins (Fig. 7B). During the mid-late infection waves, the most highly expressed gene of the *Ecp10*-like family (Fig. 7A) was predicted to encode a protein with a compact β-folded structure, stabilized by three disulfide
bonds, that is structurally similar to the antimicrobial PAFB protein from *Penicillium chrysogenum* [46] (Fig. 7B). *C. fulvum* Ecp10-1 on the other hand was predicted to adopt a smaller compact β-folded structure with only two disulphide bonds (Fig. 7B). The most highly expressed *AvrLm6*-like family gene (Fig. 7A), along with *AvrLm6* from *L. maculans*, were predicted to encode a protein with a ferredoxin-like fold, stabilized by two disulfide bonds, that is structurally similar to the virally encoded *Ustilago maydis* (corn smut pathogen) P6 virus killer toxin α-subunit (KP6α) [47, 48] (Fig. 7B), as well as an EC from the *Septoria tritici* blotch pathogen *Zymoseptoria tritici* [49] (Zt-PK6-1) (Fig. 8B). The protein encoded by the most highly expressed *V. inaequalis* Ave1-like gene (Fig. 7A), together with Ave1, were predicted to adopt a double-psi β-barrel fold, stabilized by two conserved disulfide bonds, with high similarity to expansins (Fig. 7B). Both the most highly expressed *Ecp39*-like gene (Fig. 7A) and *CfEcp39* gene from *C. fulvum* were predicted to encode a protein that adopts a crambin-like fold (Fig. 7B). However, no analogous protein structures were found in the RCSB PDB using the predicted tertiary structure of the Ecp39-like protein as query. Finally, the ViEcp6-like protein was predicted to adopt a similar tertiary structure to the previously reported three-domain lysin motif (LysM) structure of Ecp6 from *C. fulvum* [50] (Fig. 7B).
Fig. 7 Effector candidate (EC) families and singletons from *Venturia inaequalis* with amino acid sequence similarity to effectors and characterized or candidate avirulence (Avr) effector proteins from other plant-pathogenic fungi. A. Expression data of ECs during host colonization are DESeq2-normalized counts, averaged from four biological replicates, with error bars representing standard deviation (hpi: hours post-inoculation; dpi: days post-inoculation). Light orange box indicates that the EC family belongs to the early expression waves (wave 1 or 2) and dark orange box indicates that the EC family belongs to the mid-late expression waves (wave 4 or 5).
B. Protein tertiary structures of candidate virulence and avirulence (Avr) effector proteins predicted by AlphaFold2, except Ecp6 from Cladosporium fulvum (RCSB PDB id: 4B8V). Disulfide bonds coloured in yellow. V. dahliae is Verticillium dahliae; U. maydis is Ustilago maydis; L. maculans is Leptosphaeria maculans. Green tertiary structures represent the V. inaequalis protein; purple structures represent the EC/Avr from the other fungal pathogen; cyan, represents closest analogous structure in the RCSB PDB database. Amino acid similarity was identified by reciprocal protein searches based on BLASTp (E-value <0.05). pLD: predicted Local Distance Difference Test score (0‒100). A pLD score of 70‒100 is indicative of medium to high confidence. A Dali Z-score above 2 indicates ‘significant similarities’ between proteins.

We investigated the predicted fold of the most highly expressed member of the V. inaequalis EC families and singletons that were novel and did not share obvious amino acid similarity with other ECs and/or Avr effectors. Strikingly, many EC families were predicted to be structurally similar to one or more ECs or Avr effector proteins with solved tertiary structures from other plant-pathogenic fungi (Fig. 8, Additional file 17: Table S3, Additional file 18: Fig. S9). Of the 94 EC families up-regulated in planta, 71 could be assigned a structure with confidence and of those, 12 families were structurally analogous to ECs and/or Avr effectors (Additional file 17: Table S3). Remarkably, many of these families were among the most expanded families in V. inaequalis, while only three singletons had structural similarity to other ECs and/or Avr effectors (Table 1).

The main member of the most expanded EC family in V. inaequalis, family 1, was confidently predicted to adopt a six-stranded β-sandwich fold with structural similarity to seven Magnaporthe Avr and ToxB-like (MAX) effectors. The identified MAX effectors were the new MAX effector 6RSJ [51] and the Avr effectors AvrPiz-t [52], Avr-Pia [53, 54], Avr-Pib [55], Avr1-CO39 [53] and Avr-Pik [56] from the rice blast fungus M. oryzae, as well as the ToxB virulence effector from Pyrenophora tritici-repentis [57], the fungal pathogen responsible for tan spot of wheat (Additional file 17: Table S3). An amino acid sequence alignment constructed from all family 1 protein members from V. inaequalis, as well as the identified MAX effectors from M. oryzae, revealed that these proteins lacked significant amino
acid sequence similarity, with a maximum pairwise identity of only 13.9% observed between g13386 and Avr1-CO39 (Additional file 19). However, these sequence-diverse proteins did share the characteristic conserved disulphide bond between β1 and β5 that has previously been reported for MAX effectors [53] (Fig. 8A).

Two additional expanded families, family 7 with 21 members and family 28 with 6 members, as well as a small family with three members (family 38) and a singleton, shared a common β-sandwich fold with structural similarity to the virulence effector ToxA from P. triticirepentis [58], the Avr effectors Avr2/Six3 from the wilt fungus Fusarium oxysporum [59], and AvrL567-A from the flax rust fungus Melampsora lini [60] (Fig. 8A, Additional file 17: Table S3, Additional file 18: Fig. S9). Notably, despite sharing an overall β-sandwich fold with similar topology, these proteins were very diverse at the amino acid level, with a maximum amino acid identity of 12.7% (Additional file 19). All of the ToxA-like families had a different number of β-sheets: family 7 with 10 β-sheets, family 28 with nine β-sheets, family 38 with eight β-sheets and the singleton with nine β-sheets (Additional file 18: Fig. S9).

Another two families, family 15 with 12 members and family 47 with two members, were found to have structural similarity to AvrLm4-7 and AvrLm5-9 from L. maculans, as well as Ecp11-1 from C. fulvum [61, 62] (Fig. 8A, Additional file 17: Table S3, Additional file 18: Fig. S9), which belong to the Leptosphaeria AviRulence-Supressing (LARS) structural superfamily [62]. AvrLm4-7 has four stranded β-sheets [61]; however, the V. inaequalis LARS-like effectors had the same predicted topology but with a variable number of β-sheets. Specifically, members of family 15 had five β-sheets, while members of family 47 had three β-sheets. The identified V. inaequalis LARS-like proteins share an amino acid identity of only 7.8–16.9% with AvrLm4-7, AvrLm5-9 and Ecp11-1 (Additional file 19). Interestingly, these structures are stabilized by a different number of disulfide bonds, and they lack both the conserved disulfide bridge between the α-helix and β-strand, as well as the conserved WR(F/L/V)(R/K) motif, previously reported for the LARS effectors (Lazar et al., 2020).

Family 49, with 3 members, was predicted to adopt a two-domain fold similar to the Avr1/Six4 and Avr3/Six1 Avr effectors from F. oxysporum (Fig. 8A, Additional file 17: Table S3), which are the
found members of the Fol dual-domain (FOLD)-family effector family recently identified in F. oxysporum [43]. The V. inaequalis FOLD-like family representative shared 17.6% amino acid identity with Avr1/Six4 (Additional file 19) and, as part of this, the cysteine spacing pattern was conserved. Many expanded families had a ferredoxin-like fold similar to the KP6 protein of the U. maydis P6 virus (Additional file 20: Fig. 59) [47, 48] and the Zt-KP6-1 EC from Z. tritici [49]. This included the representative from the AvrLm6-like family (Fig. 8B). Additionally, family 2 with 32 members, family 5 with 36 members, family 23 with 5 members, family 26 with 4 members, and two singletons were predicted to adopt a KP6-like fold (Additional file 20: Fig. 59). All KP6-like proteins had two to four disulphide bridges, two α-helices and a variable number of β-sheets (Fig. 8B, Additional file 20: Fig. 59).

Many EC families were made up of proteins that were too small to be included in RCSB PDB searches using the Dali server or that did not have any significant Dali hit. We investigated the general structural classification of all proteins with a confidently predicted tertiary structure in this category and identified three EC families and one EC singleton that were predicted to adopt a knottin-like fold (Additional file 16). The most highly expressed member of family 11 (16 members) (Fig. 8C) and a singleton were predicted to adopt a knottin-like fold with two β-sheets, as well as three disulfide bonds that form an intramolecular knot (Additional file 21: Fig. S10). Additionally, family 12 with 17 members, family 24 with 12 members, and family 35 with 5 members, were predicted to adopt a knottin-like fold using the SCOPe database [63] (Additional file 21: Fig. S10). However, these proteins were not predicted to have a true intramolecular knot characteristic of knottin proteins based on our current AlphaFold2 predictions. Instead, these proteins shared a common fold with cysteine-stabilized αβ defensins, which are made up of a single α-helix and three β-strands [64]. In line with previous reports [65, 66], Avr9 from C. fulvum was predicted to adopt a knottin-like fold (Fig. 8C).

Finally, we investigated which predicted protein SCOPe folds were most frequent in the ECs expressed during the early (waves 1 and 2) and mid-late infection (waves 4 and 5) stages. During the early waves, we observed that ECs had a great diversity of predicted folds. The most frequently
predicted fold in the early infection wave with a pLDDT score of 70 or higher was the ‘STAT-like’ fold (Additional file 22: Fig. S11). The ‘STAT-like’ fold was observed for the most highly expressed member of the HsbA-like family and family 22. In contrast, during the mid-late infection waves, many ECs were predicted to share similar folds. The most frequently predicted fold observed was the ‘immunoglobulin-like β-sandwich’ fold, followed by ‘ferredoxin-like’, ‘parallel coiled-coil’ and ‘knottin-like’ folds (Additional file 22: Fig. S11). The ‘immunoglobulin-like β-sandwich’ fold is made up of the EC families with a predicted MAX-like or ToxA-like fold, and interestingly, most of the genes encoding these ECs are expressed during wave 4. Likewise, the ‘ferredoxin-like fold’, which includes the KP6-like ECs, were also all encoded by genes that are expressed during infection wave 4. In conclusion, genes expressed during the mid-late infection waves encoded more ECs with similar folds than the early infection waves, and many of these had structural similarity to Avr effectors or ECs from other plant-pathogenic fungi.
Fig. 8 Predicted tertiary structures of effector candidates (EC) from *Venturia inaequalis*. A. Representative EC family members with structural similarity to known avirulence (Avr) effector proteins from other plant-pathogenic fungi. MAX-like: representative predicted family 1 protein structure (green; g13386) aligned to an uncharacterized MAX effector from *Magnaporthe oryzae* (6R5J) (purple). ToxA-like: representative predicted family 7 (green; g4781) protein structure aligned to ToxA from *Pyrenophora tritici-repentis* (1ZLE) (purple). LARS-like: representative predicted family 15 (green; g11097) protein structure aligned to AvrLm4-7 from *Leptosphaeria maculans* (7FPR) (purple). Representative predicted family 49 (green; g3787) protein structure aligned to Avr1/Six4 from *Fusarium oxysporum* (7T6A) (purple). B. Predicted EC family members with a KP6-like fold. KP6-like: representative predicted AvrLm6-like family (green; g20030) protein structure aligned to EC Zt-KP6-1 from *Zymoseptoria tritici* (6QPK). C. Predicted EC family members with a knottin-like fold. Knottin-like representative predicted family 11 (green; g8686) protein structure. Predicted structure of Avr9 from *Cladosporium fulvum* (pink). Gene expression data of up-regulated ECs during colonization of susceptible apple cultivar ‘Royal Gala’ are DESeq2-normalized counts, averaged from four biological replicates, with error bars representing standard deviation (hpi: hours post-inoculation; dpi: days post-inoculation). Protein structures predicted by AlphaFold2 represent the most highly expressed member of each EC family from *V. inaequalis*. Disulfide bonds coloured in yellow. N: amino (N) terminus; C: carboxyl (C) terminus. pLDDT: predicted Local Distance Difference Test score (0–100). A pLDDT score of 70–100 is indicative of medium to high confidence. A Dali Z-score above 2 indicates ‘significant similarities’ between proteins. RMSD: root-mean-square deviation.

Discussion

In this study, we present the first comprehensive temporal transcriptome of *V. inaequalis* during colonization of apple, covering six biotrophic time points from early (12 hpi) to late (7 dpi) infection. In doing so, we have, to our knowledge, also provided the first comprehensive *in planta* transcriptome of a subcuticular fungal pathogen. Based on our analysis of this transcriptome, we identified five host infection-specific gene expression waves for *V. inaequalis* that represent three infection stages: early, mid and mid-late infection (Fig. 9). These host infection-specific gene expression waves were biologically distinct from each other and were enriched for different GO terms.
**Fig. 9** Summary of *Venturia inaequalis* host infection stages during biotrophic colonization of susceptible apple leaves. GO: gene ontology; RIPP: ribosomally synthesized and post-translationally modified peptide; CE: carbohydrate esterase, GH: glycoside hydrolase; hpi: hours post-inoculation; dpi: days post-inoculation.

Early infection was characterized by the expression of genes associated with metabolic processes (wave 1) and responses to oxidative stress (wave 2). More specifically, wave 1 was enriched for GO terms associated with RNA-binding and the synthesis of branched amino acids. Branched amino acids play an important role in fungal pathogenesis [67-69], with leucine metabolism shown to be crucial for host infection by *M. oryzae* and *Fusarium graminearum* [68], and methionine biosynthesis found to be associated with the regulation of multiple cellular processes [69, 70]. The early expression of genes associated with the synthesis of branched amino acids and methionine could suggest a similar crucial role for these amino acids in the pathogenesis of *V. inaequalis*.

Wave 2 was enriched for GO terms associated with cutinase activity and the mitigation of chemical and oxidative stress. Here, a large number of CE-encoding genes, specifically of the CE5 and CE1 families, were up-regulated during early host colonization, which is in line with previous reports showing that localized enzymatic hydrolysis is needed for penetration of the apple cuticle by...
Interestingly, we observed that genes encoding proteins with an HsbA domain were co-expressed with cutinase-encoding genes during early infection, suggesting that, as previously hypothesized, these putative hydrophobic surface-binding proteins may recruit cutinases to facilitate the degradation and digestion of the hydrophobic apple cuticle [20, 72, 73].

Wave 2 was additionally characterized by genes associated with oxidative stress tolerance, such as those encoding peroxidases, while genes encoding proteins with a ‘stress up-regulated Nod19’ domain were also abundant, forming part of an expanded Nod19 EC family in V. inaequalis. The function of proteins with a stress up-regulated Nod19 domain is currently unknown. However, these proteins have been suggested to be associated with responses to abiotic and biotic stress [74-77]. Based on these studies and the expression of Nod19 domain-containing genes during the early infection waves, the Nod19 proteins of V. inaequalis could represent a novel family of effectors that play a role in modulating oxidative stress during early infection of apple.

Interestingly, only around 20% of the EC genes from V. inaequalis were up-regulated during early infection and, with the exception of genes encoding the Nod19, HsbA and CFEM EC families, were mainly singletons. The CFEM family is defined by the presence of a CFEM domain with eight conserved cysteines. This domain is commonly found in proteins of pathogenic fungi [78] and has been shown to display diverse functions [79-82]. Some CFEM proteins promote cell death and chlorosis [79, 80], while others suppress plant cell death [81] or are associated with the development of appressoria [82]. The CFEM family in V. inaequalis has a diverse expression profile, indicating that genes of this family might play distinct roles during host colonization. Another EC family expressed during early host colonization was the Gas1-like family. Recently, Gas1-like proteins have been observed to form part of the widely distributed fungal effectors with chitinase activity (EWCA) family [83]. The tertiary structure of the representative V. inaequalis Gas1-like protein was predicted to be structurally similar to different chitin-binding LPMOs, for which EWCA proteins have been predicted to adopt the same fold [83]. EWCA proteins are secreted chitinases that degrade immunogenic chitin fragments to
suppress chitin-triggered immunity in plants [83]. Thus, it is tempting to speculate that members of the Gas1-like protein family from *V. inaequalis* play a similar role.

Mid infection (wave 3) was characterized by the expansion of subcuticular infection structures (stromata and runner hyphae), and was enriched for GO terms associated with transmembrane transport. Given that mid infection was characterized by the development of infection structures, we suggest that this wave contains genes that are highly expressed in stromata, mostly transporters for nitrogen metabolism and morphogenesis. However, it is important to point out that the mid infection wave was not well defined, overlapping partially with the mid-late infection waves 4 and 5. This is likely due to the asynchronous nature of *V. inaequalis* infection.

Mid-late infection (waves 4 and 5) was characterized by an extensive change in lifestyle and nutrition by *V. inaequalis*, as well as the expression of a large number of *EC* genes. More specifically, in wave 4, there was sudden increase in the number and expression of GH-encoding genes associated with the degradation of the plant cell wall, such as the pectin-degrading GH28s. Nutrients in the cuticle are likely scarce throughout host colonization; therefore, we suggest that *V. inaequalis* obtains nutrients by degradation of the pectin-rich layer between the cuticle and epidermal cells using these enzymes [84]. Another interesting PCWDE is a GH10 endoxylanase, which is encoded by a gene that is highly expressed during this mid-late infection stage. The importance of endoxylanases in virulence has been reported in some necrotrophic and hemibiotrophic fungi [85-88], and a GH10 endoxylanase from the necrotrophic fungus *Valsa mali* is important for colonization of apple [89]. Notably, the up-regulation of genes encoding PCWDEs during mid-late infection coincided with the up-regulation of genes encoding sugar and nitrogen-associated transporters. As many fungi only utilize nitrates as a source of nitrogen during nitrogen starvation, the up-regulation of genes encoding nitrate transporters at 2 dpi likely reflects the scarcity of nitrogen available in the subcuticular environment [90] and suggests the start of nitrogen starvation for *V. inaequalis*. The up-regulated genes encoding nitrogen transporters are co-expressed with six genes encoding aspartyl proteases; these proteases may degrade extracellular apple proteins to obtain oligopeptides that are then taken up by OPT
transporters to be used as a nitrogen source. Co-regulation of genes encoding OPT transporters and aspartyl proteases has been reported in *Candida albicans* and *U. maydis* [34, 91].

Finally, wave 5, which represents the end of mid-late infection, was enriched for GO terms that could be associated with the end of the biotrophic growth phase by *V. inaequalis*. In line with this, *PCWDE* genes expressed during this late infection stage mainly encoded AA enzymes. It is possible that these AA enzymes facilitate maceration of the plant cell wall at the end of the biotrophic stage, following leaf senescence, as a means of nutrition. A subset of cutinase-encoding genes was also expressed during late infection, which may assist *V. inaequalis* to rupture through the plant cuticle to enable the release of conidiospores from infection structures.

Mid-late infection was also characterized by the expression of secondary metabolite genes, especially those that encode RiPPs, a class of cyclic bioactive peptides recently discovered in fungi [92]. Three dikaritin RiPP clusters were up-regulated during host colonization, and all were highly expressed during late infection at 5–7 dpi. RiPPs have been hypothesized to play a role as effectors in promoting host colonization [6]. Notably, the host-selective toxin Victorin from the necrotrophic fungal pathogen *Cochliobolus victoriae* is a RiPP that has been reported to be essential for pathogenicity in oat cultivars resistant to the biotrophic crown rust fungus *Puccinia coronata* [93]. Additionally, two putative RiPP precursor genes from the endophytic fungus *Epichloë festucae* have been shown to be highly expressed during the mutualistic interaction with perennial ryegrass. In this case, deletion of these genes did not disrupt the mutualist interaction, suggesting that these RiPPs could have a redundant role or might be involved in antagonistic fungus–microbe interactions [94, 95]. The late-expression profile of the *V. inaequalis* RiPPs, together with the finding that many RiPPs from plants have potent antimicrobial activity, may suggest that these peptides promote host colonization through the eradication of microbial competitors in preparation for saprobic growth inside fallen leaf litter [6, 96, 97].

Remarkably, the mid-late infection wave also had the largest number of up-regulated genes encoding ECs (~70%), and many of these belonged to expanded protein families. The reason behind
the large expansion of EC families in *V. inaequalis* is still unknown; however, it has been suggested that the expansion of effectors within families might facilitate diversification of effector function or enable avoidance of recognition by cognate host immune receptors [20]. Expansion of EC families has also been reported in other lineage-specific pathogens, such as *Blumeria graminis* [98, 99]. Given that most EC families of *V. inaequalis* did not have amino acid sequence similarity to ECs from other plant-pathogenic fungi, and because their amino acid sequences did not provide any clues regarding potential functions, we used the *de novo* folding algorithm AlphaFold2 to predict their tertiary structures. One of the main limitations when using AlphaFold2 is that proteins with low numbers of homologous sequences in public databases normally result in predictions with low confidence scores. In an attempt to overcome this limitation, we generated custom multiple sequence alignments (MSAs), which greatly improved prediction scores (Additional file 2: Fig. S12). In total, we were able to confidently predict the tertiary structures of ~75% of EC families and 65% of singletons up-regulated in planta. Strikingly, many of the EC families had predicted structural similarity to Avr effectors from other plant-pathogenic fungi. This was in addition to ECs from *V. inaequalis* that had amino acid sequence similarity to Avr or candidate Avr effectors from *V. dahliae* (Ave1), *L. maculans* (AvrLm6), and *C. fulvum* (Ecp6 and Ecp10-1). Interestingly, the genes encoding these EC families were predominantly expressed during mid-late infection, suggesting that they may play a key role in the establishment and maintenance of biotrophy.

The largest EC family with structural similarity to Avr effectors was the MAX-like effector family, which has undergone drastic expansion in *V. inaequalis* (family 1, 75 members). In *M. oryzae*, Avrs in the MAX effector superfamily have been shown to be translocated into plant cells, where they are recognized by NLR R proteins [52, 55, 100-103]. Of these, Avr-PikD and Avr1-CO39 directly interact with their corresponding NLR R proteins. This is mediated through an interaction with a heavy metal-associated (HMA) domain that is integrated into the R protein itself, albeit through different interactions. Remarkably, it has recently been shown that the MAX effector Avr-Pik binds and stabilizes an HMA protein from rice to modulate host immunity [104]. Altogether, the MAX fold could
be well suited to interactions with HMA domains [105, 106]. It is therefore tempting to speculate that members of the MAX-like effector family from *V. inaequalis* are translocated into the plant cell, where they interact with HMA domain-containing proteins of apple involved in plant defence. In terms of expression, the *V. inaequalis* MAX-like genes are expressed at a later infection time point than the *M. oryzae* MAX effectors [53], which are expressed during early infection [53]. This difference in expression could be associated with the extended biotrophic phase of *V. inaequalis*, when compared to *M. oryzae*, which is an intracellular hemibiotrophic fungus.

Also identified in *V. inaequalis* with structural similarity to Avr and EC proteins from other plant-pathogenic fungi was the ToxA-like family. In total, three EC families and one singleton were predicted to adopt a ToxA-like fold in *V. inaequalis*. The ToxA-like structural superfamily includes the necrotrophic effector ToxA from *P. tritici-repentis* [58, 107], as well as the Avr effectors Avr2 from *F. oxysporum* [59, 108] and AvrL567 (AvrLm567-A and AvrLm567-D) [109, 110] from *M. lini*. Although all members of the ToxA-like structural family described above have been shown to be translocated into plant cells [108, 111-113], it remains to be determined if this is the case for members of the ToxA-like superfamily from *V. inaequalis*. Certainly, as R proteins from apple active against *V. inaequalis* are known to be either a PRR (Rvi6) or an NLR (Rvi15) [114, 115], it seems likely that a subset of Avrs from this fungus will likely be translocated into plants cells.

A third structural superfamily that was identified in *V. inaequalis* was the LARS-like superfamily [62]. We identified two families that have the general LARS fold, common to the Avr effectors AvrLm4-7 [61], AvrLm5-7 and AvrLm3 from *L. maculans*, as well as the Avr effector candidate Ecp11-1 from *C. fulvum* [62]. As in *V. inaequalis*, the LARS effector genes of *L. maculans* and *C. fulvum* are expressed during biotrophic host colonization [41, 62]. The *V. inaequalis* LARS-like effectors have the same predicted topology as the abovementioned effectors, with a variable number of β-sheets, but are quite different at the amino acid level and do not share the previously described conserved WR(F/L/V)(R/K) motif [62]. Another family in *V. inaequalis*, the Six4-like family, was predicted to adopt
the two-domain FOLD-like structure, recently identified in *F. oxysporum* [43]. The C-domains of these effectors also have structural similarity to proteins from the ToxA-like structural family [43].

Two other folds were found to be abundant among ECs of *V. inaequalis* that were encoded by genes expressed during the mid infection wave. These were the ferredoxin-like/KP6-like and knottin-like folds. Notably, the AvrLm6-like family, including the Avr effector AvrLm6 from *L. maculans*, as well as the candidate effector Zt-KP6-1 (6qpk) from *Z. tritici*, appear to adopt the KP6-like fold [49]. This fold was first described in the KP6 protein from the P6 virus from *U. maydis*, an antifungal toxic protein [47, 48]. Similarly, different ECs from *C. fulvum* and a necrosis-inducing effector from *Cercospora beticola* are predicted to adopt this fold [41, 116]. Intriguingly, through structural modelling, the ferredoxin-like fold was also found to be the most abundant fold in the *M. oryzae* secretome, and was predicted for the BAS4 effector protein from this fungus [117]. Altogether, this suggests that this ferredoxin/KP6-like fold may be a widely conserved structural family in different phytopathogens.

In terms of ECs that had a predicted knottin-like fold, family 11 of *V. inaequalis*, along with one singleton, were identified. Knottins are small, ultra-stable proteins with at least three disulfide bridges that form an intramolecular knot known to provide stability in hostile conditions such as the plant apoplast [118]. Current evidence suggests that multiple fungal effectors adopt this fold. Indeed, as previously suggested based on NMR [66] and cysteine bond connectivity [65] data, we predicted that Avr9, an Avr effector from *C. fulvum*, adopts a knottin fold. Furthermore, an EC from *Melampsora larici-populina*, MLP124266, which is a homolog of AvrPm4, an Avr effector from *M. lini*, was recently shown to adopt this fold [119]. The other knottin-like families identified in our study (families 12, 24 and 35) appear to adopt a fold resembling defensins, similar to the VdAMP3 effector of *V. dahliae*, which has antifungal activity and facilitates microsclerotia formation [97]. Whether the ECs from *V. inaequalis* with KP6/ferredoxin-like and knottin-like folds also possess antimicrobial activity remains to be determined, but could be a future focus of research.

Taken together, our study on *V. inaequalis*, along with previous studies on effector proteins from *M. oryzae*, *F. oxysporum* and other fungi [43, 53, 62, 117, 120], reinforce the idea that fungal
Effectors are often sequence-diverse, but share a limited number of structural folds. The presence of common structural folds in effectors without obvious sequence similarity could be the result of diversifying selection, where the effectors have evolved rapidly to a point where almost all amino acid sequence similarity, with the exception of residues involved in the maintenance of the overall structural fold, is lost [53]. Alternatively, the appearance of common folds could be the result of convergent evolution, indicating that similar folds have evolved independently in different fungi [53]. Common protein folds might have arisen in many fungal effector proteins as they provide a stable structural scaffold that is well suited to hostile plant environments and/or for interaction with conserved host targets. Future knowledge about the distribution of these folds among phytopathogens and the functional characterization of these effectors, will shed more light on this intriguing topic.

Conclusions

In conclusion, we have performed the first comprehensive gene expression analysis of a subcuticular pathogen during host plant colonization, which will provide valuable new insights into the molecular mechanisms underpinning subcuticular growth by this largely understudied class of fungi, including V. inaequalis. In conjunction with structural modelling, we have also provided an enriched list of ECs from which effectors and Avr effectors of V. inaequalis can be identified and functionally characterized and, in particular, have provided further evidence that many fungal effectors are highly diverse in sequence, but very conserved at the structural level. The identification of ECs from V. inaequalis with sequence and/or structural similarity to Avr effectors from other plant-pathogenic fungi is of great interest as, to date, there are no publications reporting the cloning of Avr effector genes from this fungus. The identification of these Avr effector genes is of paramount importance, as they can provide information on how V. inaequalis overcomes or suppresses R gene-mediated resistance in apple, and can enable the real-time detection of resistance-breaking strains in the orchard. Should Avr effectors of V. inaequalis belong to expanded protein families, it may be possible to engineer their cognate R...
proteins to recognize features common to the structural fold (direct recognition) or to monitor specific host components targeted by multiple members of the Avr family (indirect recognition). Indeed, such an approach could also be applied more globally to provide resistance against multiple plant-pathogenic fungi possessing effectors with shared effector folds or host virulence targets. Certainly, with the recent development of CRISPR-Cas9 technology in *V. inaequalis* [121], the functional characterization of ECs, and in particular those that form part of expanded EC families, is now possible.

**Materials and methods**

**V. inaequalis isolate**

The race (1) *V. inaequalis* isolate MNH120, also known as ICMP 13258 and Vi1 [20, 122], was used in this study.

**Growth in culture for RNA sequencing**

*V. inaequalis* was grown as a lawn culture from conidia on cellophane membranes (Waugh Rubber Bands, Wellington, New Zealand) [123] overlaying potato dextrose agar (PDA; Difco™, NJ, USA) at 20°C for 7 days under white fluorescent lights (4,300 K) with a 16 h light/8 h dark photoperiod. Four culture plates, representing four independent biological replicates, were then flooded with 1 mL sterile distilled water and the fungal biomass scraped from the membrane surface using a sterile cell spreader. Following this step, fungal suspensions were transferred to independent microcentrifuge tubes, pelleted by centrifugation at 21,000 x g for 1 min, snap frozen in liquid nitrogen, and then ground to a powder in preparation for RNA extraction.

**Plant infection assays for RNA sequencing and microscopy**

Seeds from open-pollinated *M. x domestica* cultivar ‘Royal Gala’ (Hawke’s Bay, New Zealand) were germinated at 4°C in moist vermiculite with 100 mg/ml Thiram fungicide (Kiwicare Corporation Limited; Christchurch, New Zealand) for approximately two months in the dark. Germinated seedlings
were planted in potting mix (Daltons™ premium potting mix; Daltons, Matamata, New Zealand) and grown under a 16 h light–8 h dark cycle with a Philips SON-T AGRO 400 Sodium lamp, at 20°C with ambient humidity. Inoculations were performed on freshly un-furled detached leaves from 4- to 6-week old apple seedlings, as described previously [124], with the exception that 5 µl droplets of conidial suspension (1 x 10^5 ml^{-1}) were used to cover the entire leaf surface. At 12 and 24 hpi, as well as 2, 3, 5 and 7 dpi, four infected leaves, each from an independent seedling, were sampled to give four biological replicates. A microscopic evaluation of infection was then performed on harvested tips from these leaves. Here, leaf tips were cleared and stained according to [125], and then visualised by bright-field microscopy, with images captured using a Leica DFC 295 digital camera and the Leica Application Suite X (LAS X). Immediately following tip harvesting, leaves were snap frozen in liquid nitrogen, and then ground to a powder in preparation for RNA extraction.

**RNA extraction and sequencing**

Total RNA was extracted from samples of *V. inaequalis* grown in culture, as well as infected leaves, using a Spectrum™ Plant Total RNA Kit (Sigma-Aldrich, St. Louis, MO, USA), with DNA subsequently removed using DNase I (Invitrogen™, Thermo Fisher Scientific, MA, USA). RNA concentration and purity were quantified using a Nanodrop ND-1000 Spectrophotometer (NanoDrop Technologies, Rockland, DE, USA), while RNA integrity was assessed on the Agilent 2100 Bioanalyzer (Agilent Technologies, Waldbronn, Germany) using an Agilent RNA 6000 Nano Kit in conjunction with Agilent 2100 Bioanalyzer software. Genomic DNA contamination was excluded by visualisation of RNA on a 0.8% agarose gel and absence of polymerase chain reaction (PCR) amplification products specific to the *actin* gene of apple (primers RE45 [TGACGAATGAGCAAGGAAATTACT] and RE64 [TACTCAGCTTTGGCAATCCACATC]) [14]. Following these quality control checks, total RNA from each of the samples was sequenced on a HiSeq X platform at Novogene (Beijing, China), via the Massey Genome Service facility (Palmerston North, New Zealand; project number MGS00286). Here, only those RNA samples with an RNA Integrity Number (RIN) value of ≥3.5 were sequenced.
Gene prediction

The genome sequence and associated gene annotations of *V. inaequalis* isolate MNH120 [20] were downloaded from the Joint Genome Institute MycoCosm portal (https://mycocosm.jgi.doe.gov/Venin1/Venin1.home.html). New genes were predicted to accommodate genes that may have been missed in the initial annotation (summarized in additional file 2: Fig S1). For this purpose, we used a three-step approach. In the first step, coding sequences (CDSs) from *V. inaequalis* isolate 05/172, which were predicted as part of a previous study by [19], were downloaded from the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/nuccore/QFBF00000000.1/) and mapped to the MNH120 genome using GMAP v2021-02-22 [126]. Here, CDSs from 05/172 were used instead of CDSs previously predicted for MNH120 [20], as the former was deemed to contain a higher number of CDSs corresponding to EC genes. In the second step, good quality RNA-seq reads (see RNA-seq read analysis section) from one biological replicate of each time point of *V. inaequalis* grown in planta and in culture were mapped to the MNH120 genome using HISAT2 v2.2.1 [127, 128], with unmapped reads filtered out using SAMtools v9.2.0 [129]. Then, a genome-guided *de novo* transcriptome assembly was performed with reads mapped to the *V. inaequalis* genome using Trinity v2.12.0 [130]. Likely CDS regions were identified using Transdecoder v5.5.0 (https://github.com/TransDecoder/TransDecoder) in conjunction with a minimum length open frame (ORF) of 50 amino acids. To maximize the identification of ORFs encoding proteins with characterized functional domains, translated ORFs were scanned against the Pfam database. Additionally, to capture as many ORFs encoding ECs as possible, translated ORFs were scanned against a list of putative effector proteins identified by [20], supplemented with an in-house database of putative MNH120 effector proteins (de la Rosa and Mesarich, unpublished). Finally, in the third step, the original gene prediction for *V. inaequalis* isolate MNH120 [20], together with the newly annotated CDSs from steps 1 and 2, were loaded as different tracks in Geneious v9.05 [131], and manual curation was performed to create consensus gene
predictions. The aim of this pipeline was to predict as many genes as possible. As a consequence, it is likely that many spurious genes were predicted.

Annotation of protein functions

Protein functional annotations were predicted using InterProScan v5.51-85.0 in conjunction with the Pfam, HAMAP, MOBIDB, PIRSF, PROSITE and SUPERFAMILY tools [132], while GO class predictions were carried out using Pannzer2 [133]. N-terminal signal peptides were predicted using SignalP v5.0 [134] and transmembrane (TM) domains were predicted using TMHMM v2.0 [135]. CAZymes were predicted using dbCAN2 in conjunction with the Hotpep, HMMER and DIAMOND tools [136]. Only CAZymes predicted with at least two of the three tools were retained for further analysis. Putative PCWDEs were manually identified based on their CAZy classification, KEGG description and InterPro annotation. Secondary metabolite clusters were predicted using antiSMASH v6.0.1 for fungi [137-142] and RiPP gene clusters in V. inaequalis were manually identified (Additional file 24: S1 Text), RiPP clusters were defined by the presence of a DUF3328 domain and precursor gene (i.e. a gene that encodes a precursor peptide with an N-terminal signal peptide, followed by one or more tandem sequence repeats that are often separated by putative kexin protease cleavage sites).

Prediction of effector candidates and effector candidate families

Small proteins of ≤400 amino acid residues in length with a predicted N-terminal signal peptide, but without a predicted transmembrane domain (TM) or endoplasmic reticulum (ER) retention motif (HDEL/KDEL), were annotated as ECs. The list of ECs was supplemented with proteins of >400 amino acids in length with a predicted N-terminal signal peptide, but no predicted TM domain or ER retention motif, provided that they were predicted to be an effector using effectorP v3.0 [143]. ECs were grouped into protein families using spectral clustering SCPS 0.9.8 [144]. The identified protein families were then manually curated by eye, taking into account conservation of the N-terminal signal peptide sequence, cysteine spacing, as well as conserved functional domains identified with InterProScan.
To further refine the list of ECs, proteins with an enzymatic annotation by InterProScan were discarded. In cases where only one or two ECs from a family were predicted to have an enzymatic domain, the EC was retained. To determine whether the ECs had amino acid sequence similarity to other proteins, a BLASTp analysis using an e-value threshold of 0.05 was performed against the NCBI non-redundant (nr) protein database.

**RNA-seq read analysis**

Methods associated with this section are summarized in additional file 2: Fig S1. As a starting point for the analysis of RNA-seq reads, a mappability mask was applied to the *V. inaequalis* MNH120 genome to prevent the multimapping of reads to repetitive genomic regions ([http://lh3lh3.users.sourceforge.net/download/seqbility-20091110.tar.bz2](http://lh3lh3.users.sourceforge.net/download/seqbility-20091110.tar.bz2)). Raw RNA-seq reads were then filtered, in which adapter sequences, as well as reads with >10% Ns or ≥50% low quality (Qscore: 5) bases, were removed. Quality of all reads was then checked using fastQC v0.11.9 [145].

Next, filtered RNA-seq reads from all samples were mapped to the masked MNH120 genome using HISAT2 v2.2.1 [127, 128]. Here, reads were filtered with SAMtools v9.2.0 [129] to keep only those reads that mapped to the fungal genome. Uniquely mapped reads were counted using featureCounts from SubRead package v2.0.0 [146]. Results from all steps of the RNA-seq analysis were aggregated for quality control assessment using MultiQC v1.11 [147].

**Differential gene expression and clustering**

The count matrix was imported to R and a differential gene expression analysis was performed with DESeq2 package v1.32.0 [148]. Pairwise comparisons from all samples were performed and genes with log₂ fold change of 1.5 and a p value of 0.01 during at least one *in planta* infection time point, relative to growth in culture, were considered to be significantly differentially expressed. A PCA plot was then generated using the PCA function of the DESeq2 package. Genes that were identified to be up-regulated at one or more *in planta* infection time points were selected for hierarchical clustering. For
clustering, expression counts were normalized using the rlog method from the DESeq2 package and scaled. Hierarchical clustering of the genes was performed using the hclust function, method Ward.D2 and Euclidian distance. Visualization of gene expression clusters/waves was performed using ggplots2 v3.3.5 [149], while gene expression heatmaps were generated using Complexheatmap v2.9.1 [150]. To further investigate the co-expression of genes inside a wave, their correlation was calculated using Pearson method.

GO and Pfam term enrichment analysis

GO predictions from Pannzer2 [133] with a predictive positive value (PPV) >5 were used for a GO term enrichment analysis across the five distinct gene expression waves. The GO term enrichment analysis was performed with topGO package v2.44.0 [151], using the total set of genes employed for clustering as background and a Fisher’s exact test for all GO terms: Biological Process (BP), Cellular Component (CC) and Molecular Function (MF). GO enrichment analysis results were visualized using ggplots2 v3.3.5 [149]. An enrichment test for Pfam domains was performed using a Fisher’s exact test, with all genes targeted in the clustering analysis used as background.

Structural modelling of protein tertiary structures

AlphaFold2 [152], in conjunction with the ColabFold notebook [153] and default parameters, was used to predict the protein tertiary structures of EC family members from V. inaequalis. Here, only those family members that were encoded by genes up-regulated in planta, relative to growth in culture, were used in this analysis, with only the most highly expressed member targeted for prediction (i.e. as a family representative). In each case, the mature amino acid sequence of the EC (i.e. without its predicted N-terminal peptide) was used as input. For the Avr1/Six4-like family, the published pro-domain [43] was also removed. For those ECs that had <30 proteins with amino acid sequence similarity in the NCBI database, as identified by a BLASTp analysis in conjunction with an e-value threshold of ≤0.05, a custom MSA was generated and used as input. To generate custom MSAs, mature
EC family members and similar proteins identified through the BLASTp analysis were aligned using Clustal Omega [154, 155], with alignments subsequently converted to the a3m format using ToolSeq [156, 157]. The only exception was for members of the Avr1/Six4-like family. Here, in an attempt to improve the structural prediction, the *F. oxysporum* Six4 and Six3 proteins were manually added to the input sequences to generate a custom MSA, even though these proteins were not identified in the initial BLASTp similarity search. For EC singletons, protein tertiary structures were predicted using AlphaFold2 open source code version 2.0.1 and 2.1.0 [152], with preset casp14, max_template_date: 2020-05-14, using mature protein sequences as input. Again, only those ECs that were encoded by genes up-regulated in planta, relative to growth in culture, were used in this analysis. All predicted protein tertiary structures with a pLDDT score of 70 or higher were considered confident predictions. Protein structures with an pLDDT score of 50–60 that also had an intrinsically disordered region predicted with MobiDB-lite [158] or PrDos [159] were also considered confident predictions.

Predicted EC protein tertiary structures were screened against the Research Collaboratory for Structural Bioinformatics (RCSB) PDB database to identify proteins with similar folds using the Dali server [45]. Here, all hits with a Z-score of ≥2 were considered to be similar. Protein tertiary structures were visualized and aligned using PyMol v2.5, in conjunction with the alignment plugin tool CEalign [160]. To further investigate similarities between protein tertiary structures, TM-align [161] was used to calculate root-mean-square deviation (RMSD). Finally, the general fold of confidently predicted protein tertiary structures was investigated using RUPEE [162, 163] against the SCOPe v2.08 database [63, 164]. Proteins predicted to have a knottin fold in the SCOPe database were assessed using Knotter 3D to determine whether they had a true knottin structure [118].

**Funding**

MRF and CHM were supported by the Marsden Fund Council from Government funding (project ID 17-MAU-100), managed by Royal Society Te Apārangi. JKB and BM received funding from The New...
Zealand Institute for Plant and Food Research Limited, Strategic Science Investment Fund, Project number: 12070.

Availability of data and materials

The raw RNA-seq data generated in this study, as well as the count matrix and DESeq2-normalized read counts, have been deposited in the NCBI Gene Expression Omnibus (GEO), and are accessible through GEO Series accession number (GSE198244). The V. inaequalis MNH120 gene annotations and associated proteins sequences generated in this study, as well as the output of AlphaFold2 (open source or ColabFold) with the PDB files for the predicted ECs tertiary structures are available at zenodo (10.5281/zenodo.6233645).

Acknowledgements

We acknowledge use of the New Zealand eScience Infrastructure (NeSI) high-performance computing facilities, as well as their technical support and training services. In particular, we thank Dinindu Senanayake for his consulting support on the high-throughput prediction of protein tertiary structures. New Zealand’s national computing facilities are provided by NeSI and are funded jointly by NeSI’s collaborator institutions and through the Ministry of Science & Innovation’s Research Infrastructure programme. URL: http://www.nesi.org.nz. We thank Drs Erik Rikkerink and Jay Jayaraman for critically reviewing the manuscript, and Dr Simon Williams for providing the F. oxysporum Avr1 and Avr3 protein structure files ahead of public release.

References


Barbar E: pathogen
MJ: De la Concepcion JC, Franceschetti M, Maqbool A, Saitoh H, Terauchi R, Kamoun S, Banfield rice blast fungus
Journal of Biomolecular NMR
Kamiya M: Ose T, Oikawa A, Nakamura Y, Maenaka K, Higuchi Y, Satoh Y, Fujiwara S, Demura M, Sone T, a highly diverse but structurally conserved effector family in phytopathogenic fungi
de Guillen K, Ortiz
Microbe Interactions
triggers the immunity in rice mediated by the blast resistance gene Magnaporthe oryzae


Viola odorata.

1221

1220

1219

1218

1217

1216

1215

1214

1213

1212

1211

1210

1209

1208

1207

1206

1205

1204

1203

1202

1201

1200

1199

1198

1197

1196

1195

1194

1193

1192

1191

1190

1189

1188

1187

1186

1185

1184

1183

1182

1181

1180

1179

1178

1177

1176

1175

1174

1173

1172


Dagvadorj B, Outram MA, Williams SJ, Solomon PS: The necrotrophic effector ToxA from Parastagonospora nodorum interacts with wheat NHL proteins to facilitate Tsn1-mediated necrosis. The Plant Journal 2022, n/a(n/a).


Sperschneider J, Dodds PN: **EffectorP 3.0: prediction of apoplastic and cytoplasmic effectors in fungi and oomycetes.** *Molecular Plant-Microbe Interactions* 2022, 35(2):146-156.


Alexa A, Rahnenführer J, Lengauer T: **Improved scoring of functional groups from gene expression data by decorrelating GO graph structure.** *Bioinformatics* 2006, **22**(13):1600-1607.


Shindyalov IN, Bourne PE: **Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.** *Protein Engineering* 1998, **11**(9):739-747.

