MPMI Vol. 35, No. 1, 2022, pp. 39–48, https://doi.org/10.1094/MPMI-04-21-0075-R *e*-Xtra*

An Improved Assembly of the *Albugo candida* Ac2V Genome Reveals the Expansion of the "CCG" Class of Effectors

Oliver J. Furzer,[1] Volkan Cevik,[1,2] Sebastian Fairhead,[1] Kate Bailey,[1] Amey Redkar,[1,3] Christian Schudoma,[1] Dan MacLean,[1] Eric B. Holub,[4] and Jonathan D. G. Jones[1,†]

[1] The Sainsbury Laboratory, University of East Anglia, Norwich NR4 7UH, United Kingdom
[2] The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, United Kingdom
[3] Department of Genetics, University of Córdoba, Córdoba 14071, Spain
[4] University of Warwick, School of Life Sciences, Warwick Crop Centre, Wellesbourne, CV35 9EF, United Kingdom

Accepted 20 September 2021.

***Albugo candida* is an obligate oomycete pathogen that infects many plants in the Brassicaceae family. We resequenced the genome of isolate Ac2V using PacBio long reads and constructed an assembly augmented by Illumina reads. The Ac2VPB genome assembly is 10% larger and more contiguous compared with a previous version. Our annotation of the new assembly, aided by RNA-sequencing information, revealed a 175% expansion (40 to 110) in the CHxC effector class, which we redefined as "CCG" based on motif analysis. This class of effectors consist of arrays of phylogenetically related paralogs residing in gene sparse regions, and shows signatures of positive selection and presence/absence polymorphism. This work provides a resource that allows the dissection of the genomic components underlying *A. candida* adaptation and, particularly, the role of CCG effectors in virulence and avirulence on different hosts.**

Keywords: *Albugo candida*, biotrophy, effectors, genome, genomics, oomycete, effectors, oomycete–plant interactions, PacBio, population biology

Oomycetes of the order Albuginales are obligate biotrophs and cause a white rust disease on plant hosts that resembles disease caused by basidiomycete rust fungi. Leaf pustules disseminate asexual spores that germinate on new leaves to enter the host via stomata and then emerge via formation of new pustules (blisters) beneath the epidermis that rupture at maturity to redisseminate asexual spores (Holub et al. 1995). A key distinction, however, is that the asexual inoculum of oomycete rusts is a zoosporangium that releases motile spores in water, which swim and alight on stomata to initiate host penetration. Oomycete rusts also have an overwintering sexual phase, producing oospores from fertilization within leaf tissue which are released into soil as the host tissue decomposes, then reinfect the next cycle of host plants via roots.

Oomycete rusts have evolved as angiosperm pathogens and include species of the genus *Albugo* which attack *Brassica* spp. (*Albugo candida*), spinach (*A. occidentalis*), and sweet potato (*A. ipomoeae*), and genera *Pustula* and *Wilsoniana* which attack the families Compositae (e.g., sunflower) or Chenopodiaceae, respectively (Brandenberger et al. 1994; Kamoun et al. 2015; Thines and Spring 2005). Interestingly, *Arabidopsis thaliana* is a natural host of two species—namely, *Albugo laibachii*, which commonly occurs on rosette leaves (Thines et al. 2009), and *A. candida*, which can infect floral tissues (Fairhead 2016) in wild host populations. Subspecies or phylogenetic races of *A. candida* have coevolved in different wild and domesticated host species of the family Brassicaceae (Jouet et al. 2019; Pound and Williams 1963). The species was first described as a pathogen of *Capsella bursa-pastoris* (Shepherds Purse), now referred to as *A. candida* race 4, and is also commonly found on other wild relatives, including *Arabidopsis thaliana, Arabis lyrata*, and *Cardamine* spp. (Holub et al. 1995; Jouet et al. 2019). Other subspecies that have specialized on economically important vegetable and oilseed brassicas include race 2 from *Brassica juncea*, race 7 from *B. rapa*, and race 9 from *B. oleracea*.

During infection of a compatible host, *Albugo* spp. grow as a network of aseptate hyphae between mesophyll cells of the host leaf, then penetrate through cell walls to produce specialized feeding structures called haustoria by invaginations of the host plasma membrane. Haustoria provide an intimate interface for coordinated nutrient acquisition and suppression of host defense responses by delivery of effector proteins. Intracellular nucleotide-binding leucine-rich repeat (NLR) immune receptors in a resistant host enable specific detection of effectors and triggering of innate immunity.

[†]Corresponding author: J. D. G. Jones; jonathan.jones@sainsbury-laboratory.ac.uk

O. J. Furzer and V. Cevik contributed equally to this work.

Current address of O. J. Furzer: Department of Biology, University of North Carolina, Chapel Hill, NC 27599, U.S.A.

Current address of C. Schudoma: Structural and Computational Biology Unit, European Molecular Biology Laboratory (EMBL), 69117 Heidelberg, Germany.

Funding: Support was provided by the Biotechnology and Biological Sciences Research Council (grant BB/L011646/1 to V. Cevik and grant BB/M003809/1 to O. J. Furzer), European Molecular Biology Organization (long-term fellowship ALTF-842-2015 to A. Redkar), and H2020 European Research Council (grant 233376 to K. Bailey and J. D. G. Jones). The Sainsbury Laboratory is supported by the Gatsby Charitable Foundation.

*The ***e*-X**tra logo stands for "electronic extra" and indicates that supplementary data and supplementary figures are published online.

The author(s) declare no conflict of interest.

Interestingly, a single NLR protein has been reported which confers broad-spectrum white rust resistance (*WRR4A*) to *Albugo candida* races 2, 4, 7, and 9 (Borhan et al. 2008). However, natural stacking of multiple NLRs has been proposed to explain the divergence of *A. candida* subspecies as a consequence of host species-level resistance (Cevik et al. 2019).

Genome sequencing revealed that *A. candida* and *A. laibachii* have compact genomes of approximately 40 Mb that show signatures of obligate biotrophy. For example, they lack key biosynthetic enzymes, making them host dependent. They also lack necrosis- and ethylene-inducing peptides commonly present in hemibiotroph and necrotroph phytopathogen genomes. A unique class of secreted effector candidates that possess what was termed the "CHxC" amino-acid motif can, in a motif-dependent manner, translocate inside plant cells and suppress plant immunity (Kemen et al. 2011; Links et al. 2011). Analysis of variation in five additional *A. candida* genomes representing four races (races 2, 4, 7, and 9) showed that recombination followed by clonal propagation likely underpins the emergence of new strains (McMullan et al. 2015). Pathogen-enrichment sequencing on the CHxC effector repertoire and a 400-kb region of 91 field samples revealed that host plant species and *A. candida* races were assorted congruently in terms of phylogeny, suggesting that host adaptation and specialization occur in the field. That study also provided evidence that certain *A. candida* races have increased ploidy levels, with the likely outcome that these lineages can only propagate asexually (Jouet et al. 2019).

Genes encoding effector proteins are often among the most variable in the genome and can be embedded in repetitive regions that hinder genome assembly (Raffaele et al. 2010). Voglmayr and Greilhuber (1998) used Feulgen imaging to estimate the genome size of a *C. bursa-pastoris*-infecting isolate at 45.6 Mb; Links et al. (2011) produced a 34.56-Mb assembly from the *B. juncea*-infecting isolate Ac2VRR. The *k*-mer analysis of our previously generated Ac2V Illumina reads suggested that the Ac2VRR assembly lacked up to 17% of the genome. To improve our understanding of *A. candida* infection and to gain further insights into *A. candida* effector repertoires, we used long-read sequencing platforms to generate an improved genome assembly of an *A. candida* race 2 strain from Canada (Ac2V). This, combined with new RNA-sequencing (RNA-Seq) data, allowed us to generate a more complete annotation, and expand the number of candidate CHxC (now renamed CCG) effectors more than twofold. Analysis of the CCG repertoire revealed that (i) CCGs are polymorphic and show presence/absence variation among *A. candida* races, and (ii) they have expanded differentially in comparison with the related species *A. laibachii*, which may be driven by pressure to avoid recognition by host immune receptors. Consistent with this, two related articles (Redkar et al. 2021 and Castel et al. 2021) report the multiple CCG effectors are recognized by different alleles of the resistance genes *WRR4A* (Borhan et al. 2008) and *WRR4B* (Cevik et al. 2019). Our analysis of the new Ac2V genome provides an essential foundation for further investigation of *Albugo* effectors.

RESULTS

Sequencing and assembly of a PacBio based Ac2V reference genome.

Our frequency analysis of *k*-mer sets derived from Ac2V Illumina reads (McMullan et al. 2015) revealed a predicted genome size of 39.7 to 40.4 Mb, 15.1 to 17.1% larger than the Ac2VRR assembly (Links et al. 2011). We extracted high molecular weight DNA from a Canadian isolate of *A. candida* race 2 (Ac2V) (Rimmer et al. 2009) and submitted it for sequencing on the PacBio RSII platform. Before assembly, raw PacBio reads were corrected with previously obtained Illumina reads (McMullan et al. 2015). Following error correction, we obtained reads with a total length of 1,219,162,236 bp (approximately 30.5× coverage) and N_{50} of 7,093 bp. The corrected reads were then used for de novo genome assembly. We named the assembly Ac2V PacBio, hereafter "Ac2VPB".

The 38.96-Mb Ac2VPB assembly is longer and more contiguous than the SOLID-based Ac2VRR and the Illumina-based AcNc2 assemblies (Fig. 1) (Links et al. 2011; McMullan et al. 2015). It has an N_{50} of 466 kb and an average contig size of 196 kb (Table 1; Fig. 1A). Ac2VRR has a similar scaffold length distribution but has a high number of Ns (1.7 Mb), representing gaps in the sequence. Subtracting those and splitting noncontiguous scaffolds reveals the relative contiguity of the Ac2VPB assembly (Fig. 1A).

We aligned Illumina reads from Ac2V to each genome and found that the Ac2VPB genome allowed the mapping of 95.5% of total reads, compared with 78.3% for Ac2VRR (Table 1). The previous Ac2VRR genome was constructed using scaffolding and contained 1.7 million Ns, showing that large parts of the genome were unresolved, likely due to repetitiveness. To compare the repeat content of the genomes, we plotted the frequency of unique *k*-mers of length 27 bp (27mers) and compared them to the raw Illumina data. This revealed that the Ac2VRR genome is missing approximately 1.2 million 27mers which predominantly occurred repeatedly (2 to 200 occurrences), compared with the Av2VPB genome. Comparison with the *k*-mer content of the Illumina reads (average depth 250) shows that there are some highly repetitive regions (27mers occurring >300 times) unaccounted for, even in Av2VPB. We estimate the overall repeat content of the Ac2VPB assembly to be 29%, of which half is composed of retroelements (for full analysis of repeats, see Supplementary Data S1).

We further used the mapped Illumina reads to check for either misassembled regions or potential hemizygous regions. Coverage showed a normal distribution centered around 250× and no large region was enriched for either low or high depth (Fig. 1C). Likewise, gene depth showed a compact normal distribution, suggesting that most genes are represented at diploid copy number (1 copy, 2 alleles) (Supplementary Fig. S1). Of 15,445 quality control-passing single-nucleotide polymorphisms (SNPs) detected, 71% had an allele frequency >0.33 and <0.66, suggesting that Ac2V has a diploid genome (Fig. 1D).

Annotation and search for candidate effector encoding genes.

RNA was extracted from Ac2V-infected *B. juncea* 'Burgonde' plants at 2, 4, 6, and 8 days postinfection (dpi) and used for library preparations and sequencing of 100-bp paired-end reads on Illumina HiSeq2000. The RNA-Seq reads were then mapped to the Ac2VPB assembly. The overall read alignment rates for samples at 2, 4, 6, and 8 dpi were 1.8, 17.2, 40.4, and 47%, respectively. Trypan blue staining was used to visualize pathogen growth, which was correlated with the proportion of reads that mapped to Ac2VPB (Supplementary Fig. S2).

These data were also used as evidence in a holistic gene prediction process (see Materials and Methods) alongside previous cDNA and protein models from published annotations (Ac2VRR, AcNc2, and AlNc14). The result is a new annotation that contains fewer but longer predicted genes (Fig. 1E) and is more complete, as evaluated by benchmarking universal single-copy orthologs (BUSCOs) (Simão et al. 2015). Ac2VPB and AcNc2 annotations both contain 100% of stramenopile BUSCOs, whereas Ac2VRR contains 93%. Furthermore, Ac2VPB encodes 89% fungal BUSCOs (complete) compared with 60% in Ac2VRR or 82% in AcNc2 (Table 2). Despite having fewer genes, the coding space of the new genome is more than 1 Mb larger than the Ac2vRR annotation (Table 2). The AlNc14 genome was reported to lack nitrate and sulphite reductases, and the molybdopterin biosynthesis pathway (Kemen et al. 2011). These are also absent from Ac2VRR and Ac2VPB. Overall, the gene-coding "compartment" of Ac2VPB remains highly compact, with intergenic distances averaging 1.3 kb. The expanded secreted protein complement contained 13 proteins with similarity to crinkling and necrosis (CRN) class effectors (Stam et al. 2013) and at least 40 potential cell-wall-modification enzymes, including 16 candidate secreted glycosyl hydrolases (Supplementary Data S2). A search of all predicted secreted proteins (no transmembrane domain) with the RXLR HMM revealed no RXLR effector candidates (Win et al. 2007).

We searched the expanded secretome for CHxC effectors. A MEME (Bailey et al. 2009) motif was constructed using a database of CHxC effectors identified from Ac2VRR and AcNc2 and used as input for a MAST search of the Ac2VPB database of proteins with predicted secretion signal (SignalP3.0) (Bendtsen et al. 2004) and lacking additional predicted transmembrane helices (TMHMM Server, v. 2.0). From this search, we identified 110 CHxC protein candidates, 70 more than were identified in Ac2VRR and 75 more than AlNc14 (Fig. 2A), and accounting for approximately 10% of the secretome.

We generated a new motif from the newly expanded CHxC complement. Compared with the consensus motif from AlNc14, we found that the previously reported histidine residue in the CHxC motif was less conserved in Ac2VPB, but a glycine six amino acids after the second cysteine was highly conserved, making the consensus motif CxxCxxxxxG. For simplicity, we redefined the effector class as "CCGs" in *A. candida* (Fig. 2B). Within figures in this article, CHxC names from *A. laibachii* were converted to AlCCG. These names do not indicate orthology between *A. candida* and *A. laibachii* CCGs sharing the same number; CHxC numbers were carried over from Kemen et al. (2011).

Among the CCGs, only seven showed homology to functionally annotated proteins or domains. Of those seven, six were annotated as "AMP-activated protein kinase β subunit" (Supplementary Data S2). The CCGs were checked for coverage artifacts and show a coverage distribution similar to that of the overall gene complement (Supplementary Fig. S1B).

The CCGs show a significant tendency to have longer-than-average intergenic distances, as revealed by plotting the 5′ and 3′ intergenic distances (Fig. 2C) and by comparing combined intergenic distances against other genes (Student's *t* test, $P <$ 0.001). We generated an alignment using the CCG and surrounding region of 60 amino acids (the only part where all of the proteins in the family could be aligned) and produced a

Table 1. Genome statistics

Parameters	Ac2vRR[a]	Ac2vPB
Scaffolds/contigs	252	198
N_{50} size (bp)	375,021	466,138
N_{50} (scaffold/contig)	33	29
Average contig length (bp)	137,158	196,775
Assembly content (bp)	34,563,972	38,961,604
Gaps (Ns)	1,728,198	0
Ac2v reads mapping (%)	78.26	95.5
Repeat content (bp)	5,875,875*	11,287,808

[a] The asterisk (*) indicates that Ns were excluded from the repeat count in Ac2vRR.

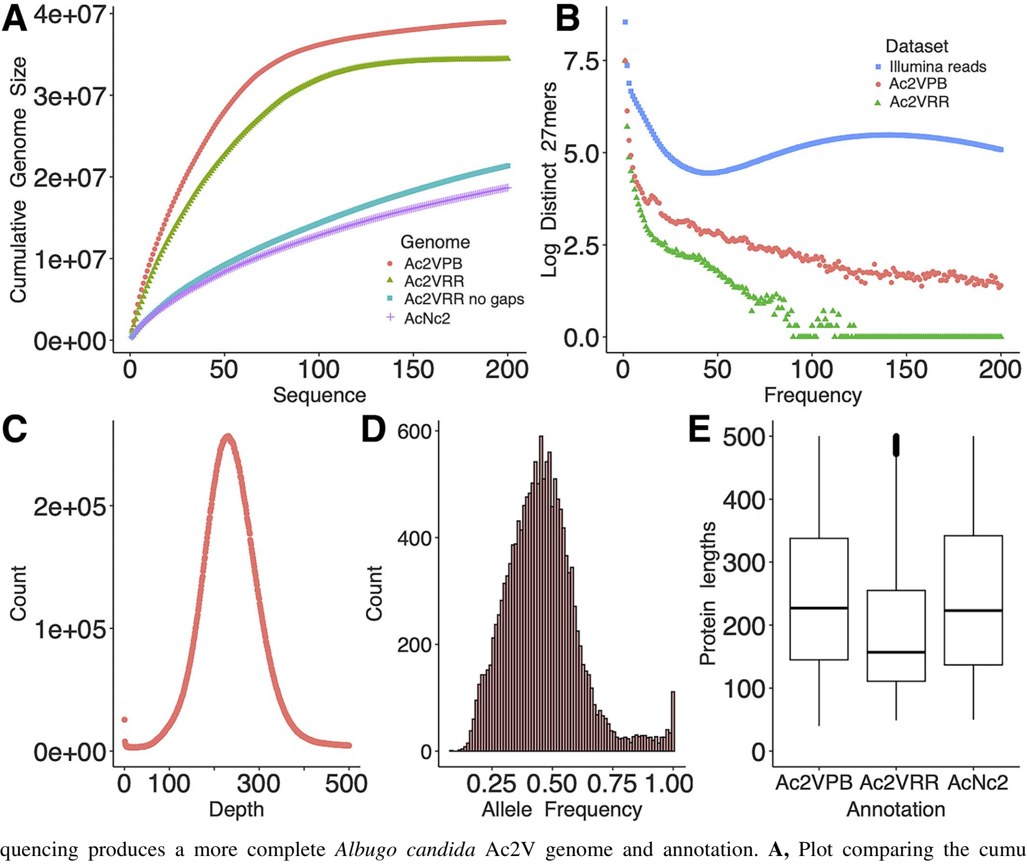

Fig. 1. PacBio sequencing produces a more complete *Albugo candida* Ac2V genome and annotation. **A,** Plot comparing the cumulative increase of genome size by size-sorted sequence (contig or scaffold) in the Ac2VPB, Ac2VRR, Ac2VRR 'no gaps', and AcNc2 assemblies. The first 200 sequences only are displayed. Ac2VRR no gaps and AcNc2 have many further smaller sequences that were thus excluded. **B,** Plot of the log number of unique *k*-mers of length 27 bp (27mers) and their frequency of occurrence in Ac2VPB and Ac2VRR assemblies compared with Ac2V Illumina reads. **C,** Plot of counts of Illumina read depth across the Ac2VPB genome. **D,** Histogram of allele frequency at all sites showing allelic variation when Illumina reads were mapped to the Ac2VPB genome. **E,** Boxplots of the distribution of protein lengths comparing the annotation of Ac2VPB, Ac2VRR, and AcNc2.

maximum-likelihood phylogeny. This revealed that the CCG effector family in Ac2V falls into seven major phylogenetic clades (Fig. 2D). As a result of the longer contigs of Ac2VPB, it became apparent that these clades largely correspond to physically colocated gene clusters, suggesting that CCGs have undergone parallel segmental duplications forming clusters at different locations across the genome (Fig. 2E; Supplementary Data S3).

Analysis of diversity in candidate effector encoding genes.

To assess diversity and selection across the genome, we used Illumina sequencing data from six additional *A. candida* isolates of race 4 (AcEx1, AcEm2, and AcNc2), race 7 (Ac7V), and race 9 (AcBoT and AcBoL) (McMullan et al. 2015; Prince et al. 2017). SNP data were used to compare gene-level nucleotide diversity and Tajima's D (a measure of balancing selection) (Tajima 1989) across the CCGs, predicting secreted proteins and remaining genes. There was no statistical difference between any of these groups (Fig. 3A and B) (analysis of variance [ANOVA] *P* values > 0.05). However, we noted the high degree of divergence between these strains which, despite being classified as the same species, are differentially adapted to diverse specific host ranges (Jouet et al. 2019) (average genome-wide identity to Ac2VPB was approximately 98%). This divergence means that, in gap-ridden cross-strain alignments of CCG regions, fewer high-quality SNPs could be assigned, resulting in artificially low diversity scores. It was possible to use predicted insertions and deletions in addition to SNP data to derive an estimate of the proportion of nonsynonymous to synonymous changes (pN/pS) or pseudogenization in all races in each gene and, in this analysis, the CCGs have a significantly higher pN/pS ratio compared with the other two categories (ANOVA, Tukey's test, $P < 0.001$) (Fig. 3C). Further taking into account zero coverage regions, CCGs showed presence/absence polymorphism across the seven races: 27 Ac2V CCGs are absent in one or more of the six additional races and, as a class, they show a stronger tendency for presence/absence polymorphism compared with other genes or genes encoding non-CCG secreted proteins (Fig. 3D), as assessed by the alignment of Illumina reads from the seven *A. candida* isolates.

The RNA-Seq reads that were mapped to the assembled Ac2VPB genome were also used to determine the expression levels of CCGs and other secreted protein-encoding genes across all colonization time points and were grouped into expression clusters (Fig. 3E). CCGs showed a full spectrum of expression patterns, from exclusively early expression to constitutive expression or expression late in the infection. These expression patterns seem to be independent of genomic cluster location, pN/pS ratio, or phylogenetic relatedness (Supplementary Data S2).

To compare CCGs in *A. candida* and *A. laibachii*, we constructed a combined phylogeny of these proteins focused around the CCG motif. We found that, at the clade level, most CCGs and CHxCs had at least one analog in the sister species; however, each had undergone a differential pattern of gene family expansion. Several clusters are greatly expanded in *A. candida* while others are more expanded in *A. laibachii*. CCGs that have evidence of recognition by WRR4A or WRR4B in Redkar et al. (2021) are highlighted and belong to clades that have specifically expanded in Ac2V (Fig. 4A). By aligning several of the clades which have expanded in either family, we confirmed that identity is retained in a restricted region around the CCG (Fig. 4B and C). We observed a frequent feature of two pairs of cysteines located at approximately 50 amino acids after the CCG motif. Additionally, we noted that complete divergence regarding both identity and length of the C-terminal region occurred in many paralogs (data not shown).

Based on the expansion of CCGs in the order Albuginales, we speculated that ancestral genes might exist in other oomycetes. We investigated CCG presence in four *Phytophthora* spp., *Hyaloperonospora arabidopsidis*, *Pythium ultimum*, and *Arabidopsis thaliana* as a plant control, using the Motif Alignment & Search Tool (MAST) (Bailey et al. 2009). We identified three potential hits (E-value < 0.1) from *Phytophthora parasitica,* two from *Phytophthora infestans*, and one weak hit (E-value = 0.2) from *H. arabidopsidis* (Supplementary Data S4). No hits were found in *Pythium ultimum* or *Arabidopsis thaliana*. Only two from *Phytophthora parasitica* contained predicted secretion signals directly prior to the CCG motif, a feature of all *Albugo* CCGs. When inserted into the overall alignment and phylogeny of CCGs from *A. candida* and *A. laibachii*, these candidates form a distant clade with, at best, 16% amino acid identity in the CCG region to the nearest CCG from Ac2vPB or AlNc14 (Supplementary Fig. S3).

DISCUSSION

We present an improved reference genome for *A. candida*. The genome is still not fully complete or assembled into full chromosomes. In the future, methods such as Bionano and Nanopore-based sequencing may enable a fully contiguous *A. candida* genome to be assembled.

The analysis of Ac2V in Jouet et al. (2019) was inconclusive about its ploidy level. Nondiploids are generally infertile. In addition to our genome-wide analysis of SNP allele frequency, it was reported that Ac2V was capable of mating with an isolate of *A. candida* race 7, with avirulence segregating 3:1 in F_2 progeny (Adhikari et al. 2003). Together, these data support Ac2V as a diploid.

The number of predicted genes in Ac2vPB is surprisingly fewer than Ac2VRR; however, the average gene length and overall coding space are larger. This suggests that the number of genes was overestimated in Ac2VRR due to fragmentation of coding sequences, possibly caused by inferior RNA-Seq results and inferior prediction algorithms (the technologies underpinning both have improved over time). A major outcome of our investigation of the improved genome is the expansion of the predicted repertoire of CCG class effectors by 175%. Other oomycete plant pathogens from the Peronosporalean lineage, which evolved plant pathogenicity independently from the Albuginales, have large families of RxLR and CRN effectors (Baxter

Table 2. Annotation statistics

Parameters[a]	Ac2vRR	Ac2vPB
Predicted genes	15,826	13,073
Total protein length	4,898,153	5,869,091
Average protein length	310	449
Predicted secreted proteins without transmembrane domains	929	1,104
Complete stramenopile BUSCOs number (percent)/fragmented (number)	84 (84%)/9	98 (98%)/2
Complete fungal BUSCOs number (percent)/fragmented (number)	173 (59.7%)/67	258 (89.0%)/19
Predicted CCG effectors	40	110

[a] BUSCOs = benchmarking universal single-copy orthologs and CCG = redefined CHxC effector class.

et al. 2010; Haas et al. 2009; Torto et al. 2003). The presence of CRN effectors in Ac2V suggests that a common ancestral CRN predates plant pathogenicity, and the expansion to nine copies may indicate some role in parasitism of plants by *A. candida*.

We performed a phylogenetic analysis of the CCGs and found a link between physical genome location and phylogenetic assortment. The clustering of effectors or virulence factors in microbes is a widespread phenomenon, allowing for the epigenetic coregulation of these clusters and facilitating mutation and duplication by unequal crossovers (Cuomo et al. 2007; Kämper et al. 2006; Raffaele and Kamoun 2012). This clustering is also seen in fungal pathogens; for example, the largest effector gene cluster encodes 24 secreted effectors in the corn smut pathogen *Ustilago maydis* (Brefort et al. 2014). In Ac2V, we found no

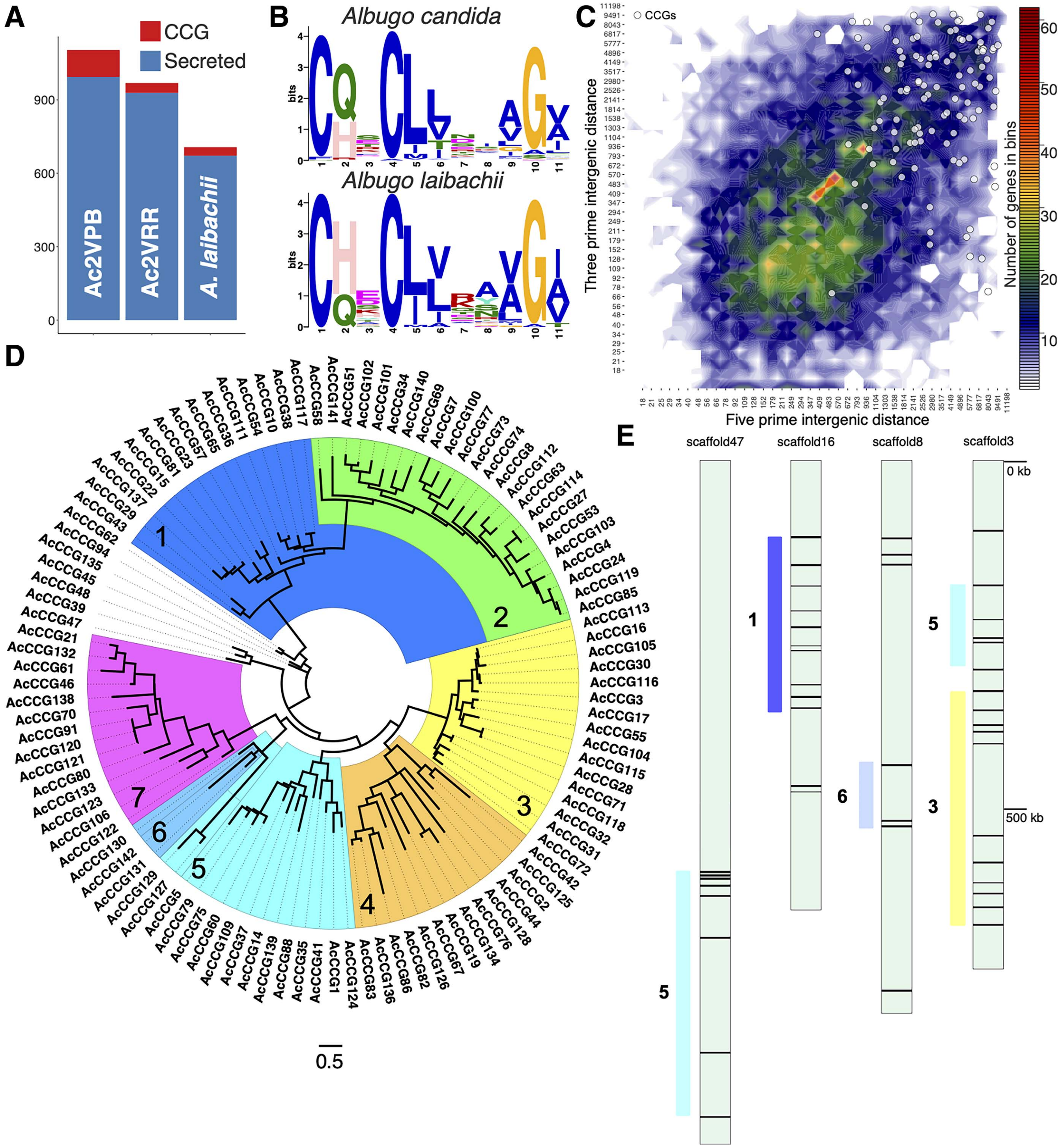

Fig. 2. CHxC effectors (CCGs) are an expanded family of secreted proteins in the *Albugo candida* Ac2V genome. **A,** Bar chart of the predicted secretome content of Ac2VPB, Ac2VRR, and AlNc14. **B,** CCG motifs in *A. candida* Ac2VPB and *A. laibachii* Nc14. **C,** Genome density volcano plot of the Ac2VPB annotation overlaid with white dots representing the CCG complement. **D,** Maximum-likelihood phylogeny of CCG proteins in Ac2VPB. This is based on an alignment of 60 amino acids surrounding the conserved CCG motif. Colors and numbers indicate manually annotated phylogenetically related clusters of CCG proteins. Scale bar represents substitutions per site. **E,** Positioning in clusters of CCGs on four selected CCG-rich scaffolds. Colored bars and numbers indicate coincidence of these clusters with those defined in D.

evidence for transcriptional coregulation of these clusters; however, we propose that the segmental duplication of these CCG clusters occurred through sequential unequal crossovers. Some clusters spanned more than one genome scaffold. Because we have not resolved the genome to chromosome level, we do not know whether these are part of the same large cluster on a single chromosome or represent interchromosomal transfer of one or more genes followed by further duplications.

In general, gene duplication is considered as a mechanism that facilitates adaptation, because increased gene copy number relaxes purifying selection on each paralog (Ohno 1970; Soskine and Tawfik 2010). The CCGs have a higher rate of pN/pS and turnover than other genes that encode secreted proteins, suggesting that adaptive evolution of the CCG sequence has been key to the coevolution of *A. candida* races with their respective host species, in line with their predicted role as effectors. Likewise, the tendency of a proportion of CCGs to be absent in various *A. candida* isolates is consistent with these effectors being recognized by host-species-specific NLR proteins. Regarding the origin of CCGs, we found a low copy number of distant CCG relatives in other oomycete genomes. More genome sequences of members of Albuginales and species that lie between them and the Peronosporales are needed to pinpoint the origin of this gene family. Our analysis suggests that, following the evolution of a useful ancestor CCG effector, specific expansions occurred as *Albugo* lineages adapted to different hosts. Although we observe the expansion of several distinct CCG clades in *A. laibachii* Nc14, in light of this work it is likely that the CCG repertoire in such Illumina assemblies is underreported.

Redkar et al. (2021) and Castel et al. (2021) discovered that certain CCG effectors are recognized by NLRs encoded by *WRR4* and related NLR-encoding genes in *Arabidopsis thaliana.* Each *Albugo laibachii* isolate can grow on approximately 88% of *Arabidopsis thaliana* accessions (Kemen et al. 2011), and resistance to *Albugo laibachii* has not been associated with the *WRR4* locus (Borhan et al. 2004). The pattern of expansion of CHxCs in *A. laibachii* is confined to clades that are distantly related to the recognized CCGs, whereas the expansions in Ac2V are in both the recognized clades and several others. We previously proposed that natural stacking of NLR-encoding genes such as *WRR4A* and *WRR4B* in *Arabidopsis thaliana* provides a nonhost-like protection against races of *Albugo candida*

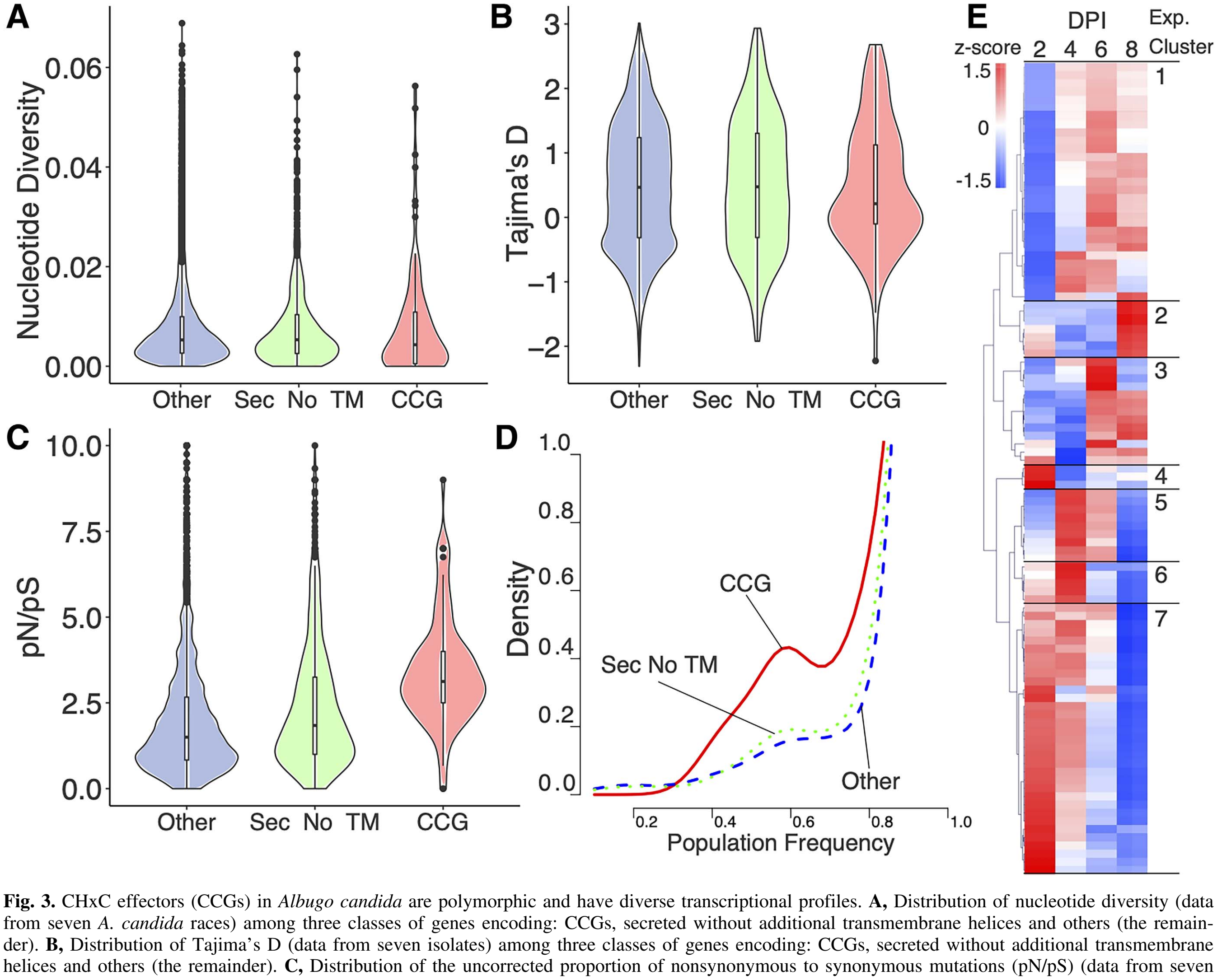

Fig. 3. CHxC effectors (CCGs) in *Albugo candida* are polymorphic and have diverse transcriptional profiles. **A,** Distribution of nucleotide diversity (data from seven *A. candida* races) among three classes of genes encoding: CCGs, secreted without additional transmembrane helices and others (the remainder). **B,** Distribution of Tajima's D (data from seven isolates) among three classes of genes encoding: CCGs, secreted without additional transmembrane helices and others (the remainder). **C,** Distribution of the uncorrected proportion of nonsynonymous to synonymous mutations (pN/pS) (data from seven races) among three classes of genes: CCGs, secreted without additional transmembrane helices and others (the remainder). **D,** Population frequency of three classes of genes among seven *A. candida* races. Most genes are present across all races (right side of graph). CCGs show an increased proportion of genes present at an intermediate frequency. The y-axis density is an arbitrary scale that is necessary to compare density plots across three unequally sized groups. **E,** Clustering of CCG encoding genes based on their expression pattern. DPI = days postinfection. *Z* score represents the number of standard deviations a given measurement is away from the normalized mean expression.

from Brassica hosts (Cevik et al. 2019). We speculate that this has exerted selective pressure on CCG effectors in *A. laibachii* to avoid patterns enriched in CQxC clades, which were free to emerge and expand in, for example, brassica-infecting lineages such as *A. candida*. Perhaps a similar NLR-based reciprocal nonhost-like barrier prevents *A. laibachii* from infecting *Arabidopsis thaliana* relatives such as *Arabis lyrata*, and adapted *Albugo* spp. on those plants have expansions in another variant of the CCG effector class. This is a further reason to obtain high-quality genomes for additional *Albugo* spp., including ones closely related to *Albugo candida* (Ploch et al. 2010). Building on this work by obtaining as many diverse samples of CCG family proteins as possible could help unlock the structural basis for CCG recognition, which could lead to the engineering of NLR proteins that confer robust resistance through the recognition of pathogen effector families as opposed to single effector proteins.

MATERIALS AND METHODS

Plant growth and laboratory techniques.

Plant and pathogen maintenance. B. juncea 'Burgonde' plants were grown on Scotts Levington F2 compost (Scotts, Ipswich, U.K.) in a controlled-environment room at 22°C with a photoperiod of 10 h/day and 14 h/night, and was used as host for *A. candida* infections. To propagate *A. candida* race Ac2V, zoosporangia were suspended in cold water and incubated on ice for 30 min. The spore suspension was then sprayed on 4-week-old *B. juncea* plants. The infected plants were kept under cycles of 10 h of light and 14 h of darkness with day and night temperatures of 21 and 14°C, respectively.

DNA extraction and genome sequencing. Zoosporangia were collected from heavily infected *B. juncea*. The spores were then ground to fine powder in a prechilled pestle and mortar with liquid nitrogen. DNA extraction was then carried out using ChargeSwitch gDNA Plant Kit (Invitrogen) following the manufacturer's instructions. *A. candida* Ac2V genomic DNA was sequenced by PacBio RSII platform (4 SMRTcells) (Pacific Biosciences) at Earlham Institute, Norwich, U.K.

RNA extraction and sequencing. Four-week-old *B. juncea* plants were sprayed with the pathogen and between 10 and 15 infected leaves were collected at 2, 4, 6, and 8 dpi, immediately flash frozen in liquid nitrogen, and stored at −80°C. RNA extraction was carried out using Direct-zol RNA Miniprep Kit (Zymo Research, Cambridge, U.K.). RNA integrity was then assessed using the Agilent 2100 Bioanalyzer with the RNA 6000 Nano Assay Kit (Agilent Technologies). Library preparation was carried out using the Illumina TruSeq RNA sample preparation kit (Illumina). Library preparations were then sequenced on an Illumina HiSeq2000 platform and 100-bp paired-end reads were generated at Earlham Institute, Norwich, U.K.

Bioinformatics and statistics.

PacBio read correction and genome assembly. Prior to the genome assembly, the Proovread (Hackl et al. 2014) error correction tool was used to correct PacBio SMRT reads (FastQ) using short Illumina reads from Ac2V (ENA: SRR1811471) (McMullan et al. 2015). Corrected reads were screened for contamination using KRAKEN2 v2-2.0.8 (databases plant, human, bacterial, fungal, and viral) (Wood et al. 2019). The reads were found to contain low levels of contamination, with 6% potential plant and 2% potential bacterial reads. The genome assembly was then conducted using the error-corrected SMRT reads with Canu V1.3 (Koren et al. 2017). The assembly was checked for contamination by breaking it down into 150mers and running KRAKEN2 v2-2.0.8 (databases plant, human, bacterial, fungal,

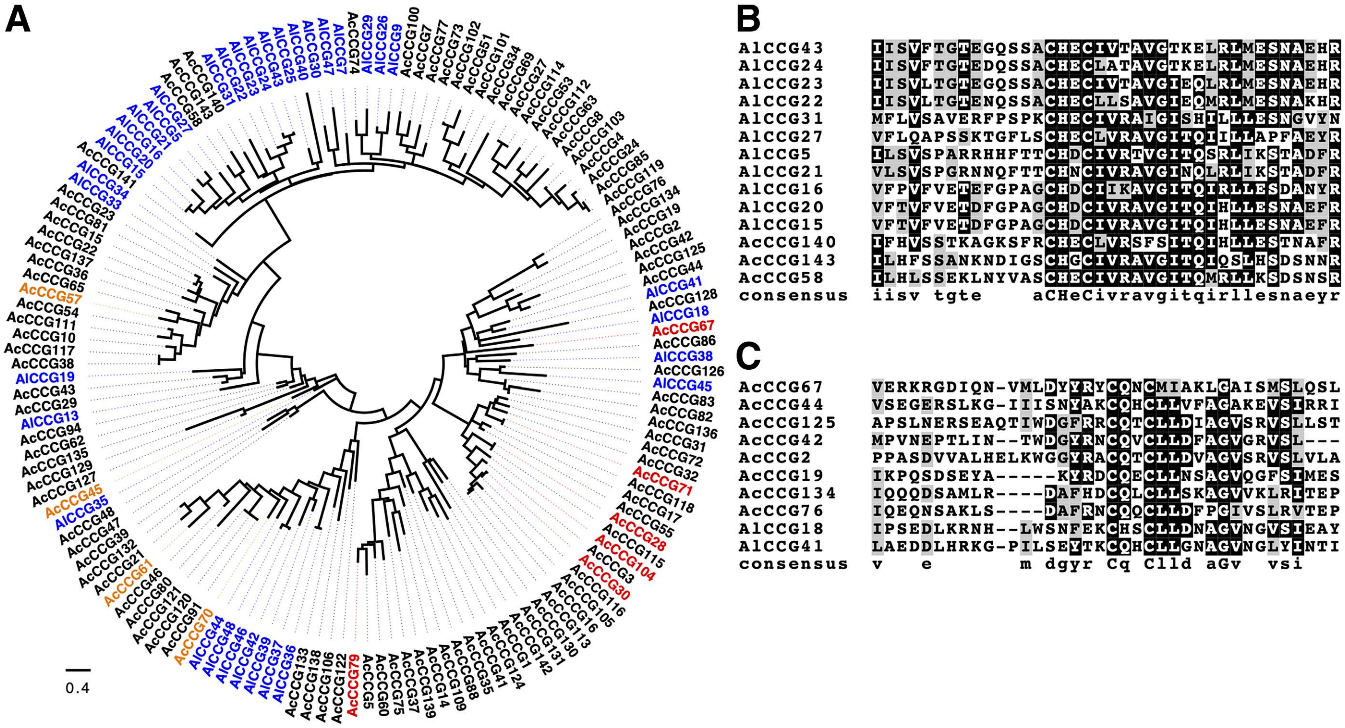

Fig. 4. Differential CHxC effector (CCG) clade expansions in *Albugo candida* and *A. laibachii.* **A,** Maximum-likelihood phylogeny incorporating CCG proteins from Ac2VPB and AlNc14. This is derived from an alignment of the 60 amino-acids surrounding (and including) the conserved CCG motif. Circles indicate the putative Ac2VPB orthologs of CCGs reported recognized by the *Arabidopsis thaliana* Col-0 R proteins WRR4A (red) and WRR4B (orange). Scale bar represents substitutions per site. **B,** Example alignment of an AlNc14 expanded clade. **C,** Example alignment of an Ac2VPB expanded clade.

and viral). It was found to contain 2.5% potential plant and <1% potential bacterial content. Most are likely false positives due to the random phylogenetic distribution of those potential plant sequences, and only 0.19% are potentially derived from Brassiceae (the family of the host plant *B. juncea*).

Repeat analysis. Repeats were modeled and detected using the RepeatMasker/RepeatModeler pipeline (RepeatMasker Open-4.0) including RECON (version 1.08) and RepeatScout (version1.06).

Annotation. The new Ac2V RNA-Seq reads from all time points were aligned to the Ac2VPB assembly using HISAT2 (Kim et al. 2015). Successfully mapped reds were extracted and assembled using three systems: velvet/oases (versions 1.2.10 and 0.2, *k*-mer sizes 33, 41, 49, and 55) (Schulz et al. 2012), soapdenovo (version 1.03, trans, *k*-mer sizes 31, 41, 51, 61, and 71) (Luo et al. 2012), and trinity (version 2.0.6, genome-guided with aligned reads) (Grabherr et al. 2011). The resulting redundant pool of transcripts was reduced to the best isoforms using EvidentialGene (tr2aacds v.2013.03.11). These transcripts were aligned to the genome with exonerate (v2.2.0) (Slater and Birney 2005), which provided evidence to predict genes using genemarkES suite v4.21 (Lomsadze et al. 2014). The assembled transcripts were also used as evidence to predict genes using webaugustus v3.2.1 (Hoff and Stanke 2013). Nc2 transcripts and proteins, cloned CCG proteins, and transcripts from the Ac2VRR genome were all used to separately train augustus and produce gene models. Independently, all mapped RNA-Seq reads were used to predict genes using braker v1.10 (Hoff et al. 2019). All of the resulting gene models and the exonerate transcript mapping data were together fed into EVidenceModeler (version 1.1.1) (Haas et al. 2008) over three iterations, which generated a maximal nonredundant genome annotation. A small number of known CCG-encoding genes were manually edited; those genes are annotated with an "m" at the end of their gene or transcript name.

Gene functional annotation. Eggnog emapper (version emapper-1.0.3-35) (Huerta-Cepas et al. 2019) with automatic taxonomic scope was used to assign functional annotations to the Ac2VPB predicted proteins.

k*-mer Analysis.* For the purpose of genome size estimation, 19-, 21-, and 23mers were generated from Ac2V Illumina reads using jellyfish v2.3.0 (Marçais and Kingsford 2011), and frequency histograms were submitted to Genomescope for analysis (Vurture et al. 2017). The 27mers were counted using kat (Mapleson et al. 2017) and plotted in RStudio using ggplot2 v3.2.1.

Read alignment and variant calls. Illumina read alignments for whole-genome data were performed using BWA v0.7.17 (Li and Durbin 2009), and Samtools v1.10 (Li et al. 2009) was used to process alignments and generate variant calls. VCFtools v0.1.15 (Danecek et al. 2011) was used for VCF formatting and quality filtering (variants with minimum quality of 20 were considered).

Presence/absence variation analysis. BUSCO analysis was performed using BUSCO v3.0.1 (Simão et al. 2015), amino-acid fasta sequences as input, and the stramenopiles_odb10 and fungi_odb9 lineage datasets. The BEDtools v2.29 (Quinlan 2014) command 'bedtools coverage' was used to compute both per-base coverage in each alignment and per-gene coverage. Because any percent cutoff would be arbitrary, the percent coverage of each gene was used to contribute to an overall population frequency score scaled from 0 to 1 used in Figure 3D. Figure 3D was generated using the sm package in RStudio v1.2.5001 (Allaire 2012).

Diversity and selection analysis. Nucleotide diversity and Tajima's D statistics were calculated using popgenome v2.7.1 (Pfeifer et al. 2014), with filtered combined variant call files and the Ac2VPB annotation as inputs.

Simple pN/pS analysis. Filtered combined variant call files were used as input for SNPEff v4.3t (Cingolani et al. 2012), which divided variants by their computed effect synonymous or nonsynonymous to produce a ratio for each gene.

ANOVA and Tukey's test. Statistical comparison of population statistics of groups of genes was performed by one-way ANOVA and post hoc Tukey's test using Online Web Statistical Calculators.

Secreted protein prediction. Ac2VPB proteins were submitted to Signalp3 (Bendtsen et al. 2004) and considered secreted. TMHMM Server (v. 2.0) was used to search these proteins post their predicted signal peptide cleavage site for additional transmembrane helices.

CCG searches. CHxC and CCGs identified by previous genome projects (Jouet et al. 2019; Links et al. 2011; McMullan et al. 2015) were used as a base to generate a CCG motif with MEME (v5.1.1) (Bailey et al. 2009). This motif was used to search the Ac2VPB-secreted no-transmembrane-domain proteins with MAST (Bailey et al. 2009). The proteins positive (E-value < 0.1, correct positioning) for the CCG motif were then fed back into MEME to produce a refined CCG motif corresponding to the Ac2V CCG signature. Certain proteins outside the *Albugo* genus, with relaxed E-value threshold < 0.2, are included in the list of hits.

Gene expression analysis. For RNA-Seq data analysis, reads (2 × 100 bp) obtained from each time point were first trimmed using Trimmomatic version 0.36 (Bolger et al. 2014) and the quality of the trimmed reads was assessed with FastQC v0.11.4 (Babraham Bioinformatics). Reads were then mapped to the assembled Ac2VPB genome using HISAT2 v2.2 (Kim et al. 2015). The counts of reads that mapped to each predicted gene were obtained using the featureCounts utility of the subread package (Liao et al. 2014). Read count data were then normalized as counts per million (CPM) with the EdgeR package (Robinson et al. 2010). Clusters and heatmaps were then made using z scores obtained from normalized [$\log_2$(CPM + 1)] data.

Genome architecture analysis. Genome architecture analysis was performed in Rstudio following the protocol of Saunders et al. (2014).

CCG phylogenies. CCG proteins were aligned using Muscle and alignments were trimmed and realigned to good quality. These alignments were used to generate maximum-likelihood phylogenies using the WAG method, with freq, three distinct γ categories, and 100 bootstraps. These analyses were performed in the MEGA X suite (Kumar et al. 2018). Phylogenies were edited using FigTree (v1. 3.1.).

Karyoplot. The Karyoplot diagram was generated using karyoplotR (Gel and Serra 2017) using the procedure described by Van de Weyer et al. (2019).

General tools for figure production. Figure panels were generated using the patchwork tool for Rstudio and all figures were edited in Inkscape 0.92. Scatter plots, box and whisker diagrams, and bar charts were generated using ggplot2 v3.2.1.

Data availability.

Data from previous studies included Illumina reads AcBoT (ENA: SRR1811472), AcEm2 (SRR1806791), AcBoL (SRR1811474), AcNc2 (SRR1811450), and Ac2V (SRR1811471). Expressed sequence tags from Ac2vRR can be downloaded from NCBI GenBank (HO914811 to HO965058, HO965059 to HO999999, and HS000001 to HS003763), plus assemblies and annotations of AcNc2, AlNc14, and Ac2vRR.

Data were submitted as follows: Ac2vPB PacBio reads (PRJEB39673), Ac2v RNA-Seq reads, Ac2VPB assembly (GCA_905220665.1), AcEx1 Illumina reads (ERR4395362), Ac7V Illumina reads (ERR5168241), and Ac2VPB annotation. The Ac2VPB annotation can be downloaded as GFF or FASTA files at GitHub.

ACKNOWLEDGMENTS

We thank R. K. Shrestha and A. Jouet for technical support and assistance with data retrieval. This research was supported in part by the NBI Computing Infrastructure for Science Group, which provides technical support and maintenance to the Sainsbury Laboratory's high-performance computing cluster and storage systems.

AUTHOR-RECOMMENDED INTERNET RESOURCES

EvidentialGene: http://arthropods.eugenes.org/EvidentialGene

FastQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

FigTree v1. 3.1: https://ci.nii.ac.jp/naid/10030433668/

GFF or FASTA files at GitHub: https://github.com/oliverjf/ac2v_genomics

Inkscape 0.92: https://inkscape.org/

Online Web Statistical Calculators: https://astatsa.com/

patchwork: The composer of plots: cran.r-project.org/web/packages/patchwork/

RepeatMasker Open-4.0: https://www.repeatmasker.org/

TMHMM Server, v. 2.0: http://www.cbs.dtu.dk/services/TMHMM/

LITERATURE CITED

Adhikari, T. B., Liu, J. Q., Mathur, S., Wu, C. X., and Rimmer, S. R. 2003. Genetic and molecular analyses in crosses of race 2 and race 7 of *Albugo candida.* Phytopathology 93:959-965.

Allaire, J. J. 2012. RStudio: Integrated development environment for R. Page 14 in: The R User Conference, useR! University of Warwick, Coventry, U.K.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. 2009. MEME SUITE: Tools for motif discovery and searching. Nucleic Acids Res. 37:W202-W208.

Baxter, L., Tripathy, S., Ishaque, N., Boot, N., Cabral, A., Kemen, E., Thines, M., Ah-Fong, A., Anderson, R., Badejoko, W., Bittner-Eddy, P., Boore, J. L., Chibucos, M. C., Coates, M., Dehal, P., Delehaunty, K., Dong, S., Downton, P., Dumas, B., Fabro, G., Fronick, C., Fuerstenberg, S. I., Fulton, L., Gaulin, E., Govers, F., Hughes, L., Humphray, S., Jiang, R. H. Y., Judelson, H., Kamoun, S., Kyung, K., Meijer, H., Minx, P., Morris, P., Nelson, J., Phuntumart, V., Qutob, D., Rehmany, A., Rougon-Cardoso, A., Ryden, P., Torto-Alalibo, T., Studholme, D., Wang, Y., Win, J., Wood, J., Clifton, S. W., Rogers, J., Van den Ackerveken, G., Jones, J. D. G., McDowell, J. M., Beynon, J., and Tyler, B. M. 2010. Signatures of adaptation to obligate biotrophy in the *Hyaloperonospora arabidopsidis* genome. Science 330:1549-1551.

Bendtsen, J. D., Nielsen, H., von Heijne, G., and Brunak, S. 2004. Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. 340: 783-795.

Bolger, A. M., Lohse, M., and Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.

Borhan, M. H., Gunn, N., Cooper, A., Gulden, S., Tör, M., Rimmer, S. R., and Holub, E. B. 2008. WRR4 encodes a TIR-NB-LRR protein that confers broad-spectrum white rust resistance in *Arabidopsis thaliana* to four physiological races of *Albugo candida.* Mol. Plant-Microbe Interact. 21:757-768.

Borhan, M. H., Holub, E. B., Beynon, J. L., Rozwadowski, K., and Rimmer, S. R. 2004. The Arabidopsis TIR-NB-LRR gene RAC1 confers resistance to *Albugo candida* (white rust) and is dependent on EDS1 but not PAD4. Mol. Plant-Microbe Interact. 17:711-719.

Brandenberger, L. P., Correll, J. C., Morelock, T. E., and McNew, R. W. 1994. Characterization of resistance of spinach to white rust (*Albugo occidentalis*) and downy mildew (*Peronospora farinosa* f. sp. *spinaciae*). Phytopathology 84:431-437.

Brefort, T., Tanaka, S., Neidig, N., Doehlemann, G., Vincon, V., and Kahmann, R. 2014. Characterization of the largest effector gene cluster of *Ustilago maydis.* PLoS Pathog. 10:e1003866.

Castel, B., Fairhead, S., Furzer, O. J., Redkar, A., Wang, S., Cevik, V., Holub, E. B., and Jones, J. D. G. 2021. Evolutionary trade-offs at the Arabidopsis WRR4A resistance locus underpin alternate *Albugo candida* race recognition specificities. Plant J. 107:1490-1502.

Cevik, V., Boutrot, F., Apel, W., Robert-Seilaniantz, A., Furzer, O. J., Redkar, A., Castel, B., Kover, P. X., Prince, D. C., Holub, E. B., and Jones, J. D. G. 2019. Transgressive segregation reveals mechanisms of *Arabidopsis* immunity to *Brassica*-infecting races of white rust (*Albugo candida*). Proc. Natl. Acad. Sci. U.S.A. 116:2767-2773.

Cingolani, P., Platts, A., Wang, L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) 6:80-92.

Cuomo, C. A., Güldener, U., Xu, J.-R., Trail, F., Turgeon, B. G., Di Pietro, A., Walton, J. D., Ma, L. J., Baker, S. E., Rep, M., Adam, G., Antoniw, J., Baldwin, T., Calvo, S., Chang, Y. L., Decaprio, D., Gale, L. R., Gnerre, S., Goswami, R. S., Hammond-Kosack, K., Harris, L. J., Hilburn, K., Kennell, J. C., Kroken, S., Magnuson, J. K., Mannhaupt, G., Mauceli, E., Mewes, H. W., Mitterbauer, R., Muehlbauer, G., Münsterkötter, M., Nelson, D., O'Donnell, K., Ouellet, T., Qi, W., Quesneville, H., Roncero, M. I., Seong, K. Y., Tetko, I. V., Urban, M., Waalwijk, C., Ward, T. J., Yao, J., Birren, B. W., and Kistler, H. C. 2007. The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. Science 317:1400-1402.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., and 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. Bioinformatics 27:2156-2158.

Fairhead, S. 2016. Translating genetics of oomycete resistance from *Arabidopsis thaliana* into Brassica production. http://wrap.warwick.ac.uk/90258/1/WRAP_Theses_Fairhead_2016.pdf

Gel, B., and Serra, E. 2017. karyoploteR: An R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics 33:3088-3090.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644-652.

Haas, B. J., Kamoun, S., Zody, M. C., Jiang, R. H. Y., Handsaker, R. E., Cano, L. M., Grabherr, M., Kodira, C. D., Raffaele, S., Torto-Alalibo, T., Bozkurt, T. O., Ah-Fong, A. M., Alvarado, L., Anderson, V. L., Armstrong, M. R., Avrova, A., Baxter, L., Beynon, J., Boevink, P. C., Bollmann, S. R., Bos, J. I., Bulone, V., Cai, G., Cakir, C., Carrington, J. C., Chawner, M., Conti, L., Costanzo, S., Ewan, R., Fahlgren, N., Fischbach, M. A., Fugelstad, J., Gilroy, E. M., Gnerre, S., Green, P. J., Grenville-Briggs, L. J., Griffith, J., Grünwald, N. J., Horn, K., Horner, N. R., Hu, C. H., Huitema, E., Jeong, D. H., Jones, A. M., Jones, J. D., Jones, R. W., Karlsson, E. K., Kunjeti, S. G., Lamour, K., Liu, Z., Ma, L., Maclean, D., Chibucos, M. C., McDonald, H., McWalters, J., Meijer, H. J., Morgan, W., Morris, P. F., Munro, C. A., O'Neill, K., Ospina-Giraldo, M., Pinzón, A., Pritchard, L., Ramsahoye, B., Ren, Q., Restrepo, S., Roy, S., Sadanandom, A., Savidor, A., Schornack, S., Schwartz, D. C., Schumann, U. D., Schwessinger, B., Seyer, L., Sharpe, T., Silvar, C., Song, J., Studholme, D. J., Sykes, S., Thines, M., van de Vondervoort, P. J., Phuntumart, V., Wawra, S., Weide, R., Win, J., Young, C., Zhou, S., Fry, W., Meyers, B. C., van West, P., Ristaino, J., Govers, F., Birch, P. R., Whisson, S. C., Judelson, H. S., and Nusbaum, C. 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans.* Nature 461:393-398.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., and Wortman, J. R. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9:R7.

Hackl, T., Hedrich, R., Schultz, J., and Förster, F. 2014. proovread: Large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics 30:3004-3011.

Hoff, K. J., Lomsadze, A., Borodovsky, M., and Stanke, M. 2019. Whole-genome annotation with BRAKER. Pages 65-95 in: Gene Prediction: Methods and Protocols. M. Kollmar, ed. Springer New York, New York, NY, U.S.A.

Hoff, K. J., and Stanke, M. 2013. WebAUGUSTUS—A web service for training AUGUSTUS and predicting genes in eukaryotes. Nucleic Acids Res. 41:W123-W128.

Holub, E. B., Brose, E., Tör, M., Clay, C., Crute, I. R., and Beynon, J. L. 1995. Phenotypic and genotypic variation in the interaction between *Arabidopsis thaliana* and *Albugo candida.* Mol. Plant-Microbe Interact. 8:916-928.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., and Bork, P. 2019. eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 47:D309-D314.

Jouet, A., Saunders, D. G. O., McMullan, M., Ward, B., Furzer, O., Jupe, F., Cevik, V., Hein, I., Thilliez, G. J. A., Holub, E., van Oosterhout, C., and Jones, J. D. G. 2019. *Albugo candida* race diversity, ploidy and host-associated microbes revealed using DNA sequence capture on diseased plants in the field. New Phytol. 221:1529-1543.

Kamoun, S., Furzer, O., Jones, J. D. G., Judelson, H. S., Ali, G. S., Dalio, R. J. D., Roy, S. G., Schena, L., Zambounis, A., Panabières, F., Cahill, D., Ruocco, M., Figueiredo, A., Chen, X. R., Hulvey, J., Stam, R., Lamour, K., Gijzen, M., Tyler, B. M., Grünwald, N. J., Mukhtar, M. S., Tomé, D. F., Tör, M., Van Den Ackerveken, G., McDowell, J., Daayf, F., Fry, W. E., Lindqvist-Kreuze, H., Meijer, H. J., Petre, B., Ristaino, J., Yoshida, K., Birch, P. R., and Govers, F. 2015. The Top 10 oomycete pathogens in molecular plant pathology. Mol. Plant Pathol. 16:413-434.

Kämper, J., Kahmann, R., Bölker, M., Ma, L.-J., Brefort, T., Saville, B. J., Banuett, F., Kronstad, J. W., Gold, S. E., Müller, O., Perlin, M. H., Wösten, H. A., de Vries, R., Ruiz-Herrera, J., Reynaga-Peña, C. G., Snetselaar, K., McCann, M., Pérez-Martín, J., Feldbrügge, M., Basse, C. W., Steinberg, G., Ibeas, J. I., Holloman, W., Guzman, P., Farman, M., Stajich, J. E., Sentandreu, R., González-Prieto, J. M., Kennell, J. C., Molina, L., Schirawski, J., Mendoza-Mendoza, A., Greilinger, D., Münch, K., Rössel, N., Scherer, M., Vranes, M., Ladendorf, O., Vincon, V., Fuchs, U., Sandrock, B., Meng, S., Ho, E. C., Cahill, M. J., Boyce, K. J., Klose, J., Klosterman, S. J., Deelstra, H. J., Ortiz-Castellanos, L., Li, W., Sanchez-Alonso, P., Schreier, P. H., Häuser-Hahn, I., Vaupel, M., Koopmann, E., Friedrich, G., Voss, H., Schlüter, T., Margolis, J., Platt, D., Swimmer, C., Gnirke, A., Chen, F., Vysotskaia, V., Mannhaupt, G., Güldener, U., Münsterkötter, M., Haase, D., Oesterheld, M., Mewes, H. W., Mauceli, E. W., DeCaprio, D., Wade, C. M., Butler, J., Young, S., Jaffe, D. B., Calvo, S., Nusbaum, C., Galagan, J., and Birren, B. W. 2006. Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis.* Nature 444:97-101.

Kemen, E., Gardiner, A., Schultz-Larsen, T., Kemen, A. C., Balmuth, A. L., Robert-Seilaniantz, A., Bailey, K., Holub, E., Studholme, D. J., Maclean, D., and Jones, J. D. 2011. Gene gain and loss during evolution of obligate parasitism in the white rust pathogen of *Arabidopsis thaliana.* PLoS Biol. 9:e1001094.

Kim, D., Langmead, B., and Salzberg, S. L. 2015. HISAT: A fast spliced aligner with low memory requirements. Nat. Methods 12:357-360.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. 2017. Canu: Scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res. 27:722-736.

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. 2018. MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. 35:1547-1549.

Li, H., and Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079.

Liao, Y., Smyth, G. K., and Shi, W. 2014. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923-930.

Links, M. G., Holub, E., Jiang, R. H. Y., Sharpe, A. G., Hegedus, D., Beynon, E., Sillito, D., Clarke, W. E., Uzuhashi, S., and Borhan, M. H. 2011. De novo sequence assembly of *Albugo candida* reveals a small genome relative to other biotrophic oomycetes. BMC Genomics 12:503.

Lomsadze, A., Burns, P. D., and Borodovsky, M. 2014. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. 42:e119.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D. W., Yiu, S. M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T. W., and Wang, J. 2012. SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. Gigascience 1:18.

Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., and Clavijo, B. J. 2017. KAT: A *k*-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics 33:574-576.

Marçais, G., and Kingsford, C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics 27:764-770.

McMullan, M., Gardiner, A., Bailey, K., Kemen, E., Ward, B. J., Cevik, V., Robert-Seilaniantz, A., Schultz-Larsen, T., Balmuth, A., Holub, E., van Oosterhout, C., and Jones, J. D. 2015. Evidence for suppression of immunity as a driver for genomic introgressions and host range expansion in races of *Albugo candida*, a generalist parasite. eLife 4: e04550.

Ohno, S. 1970. Evolution by Gene Duplication. Springer Science+Business Media, LLC, New York, NY, U.S.A.

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E., and Lercher, M. J. 2014. PopGenome: An efficient Swiss army knife for population genomic analyses in R. Mol. Biol. Evol. 31:1929-1936.

Ploch, S., Choi, Y.-J., Rost, C., Shin, H.-D., Schilling, E., and Thines, M. 2010. Evolution of diversity in *Albugo* is driven by high host specificity and multiple speciation events on closely related Brassicaceae. Mol. Phylogenet. Evol. 57:812-820.

Pound, G. S., and Williams, P. H. 1963. Biological races of *Albugo candida*. Phytopathology 53:1146-1149.

Prince, D. C., Rallapalli, G., Xu, D., Schoonbeek, H.-J., Çevik, V., Asai, S., Kemen, E., Cruz-Mireles, N., Kemen, A., Belhaj, K., Schornack, S., Kamoun, S., Holub, E. B., Halkier, B. A., and Jones, J. D. 2017. *Albugo*-imposed changes to tryptophan-derived antimicrobial metabolite biosynthesis may contribute to suppression of non-host resistance to *Phytophthora infestans* in *Arabidopsis thaliana.* BMC Biol. 15:20.

Quinlan, A. R. 2014. BEDTools: The Swiss-army tool for genome feature analysis. Curr. Protoc. Bioinf. 47:11.12.1-11.12.34.

Raffaele, S., Farrer, R. A., Cano, L. M., Studholme, D. J., MacLean, D., Thines, M., Jiang, R. H., Zody, M. C., Kunjeti, S. G., Donofrio, N. M., Meyers, B. C., Nusbaum, C., and Kamoun, S. 2010. Genome evolution following host jumps in the Irish potato famine pathogen lineage. Science 330:1540-1543.

Raffaele, S., and Kamoun, S. 2012. Genome evolution in filamentous plant pathogens: Why bigger can be better. Nat. Rev. Microbiol. 10:417-430.

Redkar, A., Cevik, V., Bailey, K., Furzer, O. J., Fairhead, S., Borhan, M. H., Holub, E. B., and Jones, J. D. G. 2021. The Arabidopsis WRR4A and WRR4B paralogous NLR proteins both confer recognition of multiple *Albugo candida* effectors. BioRxiv. https://doi.org/10.1101/2021.03.29.436918

Rimmer, S. R., Mathur, S., and Wu, C. R. 2009. Virulence of isolates of *Albugo candida* from western Canada to Brassica species. Can. J. Plant Pathol. 3:229-235.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. 2010. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139-140.

Saunders, D. G. O., Win, J., Kamoun, S., and Raffaele, S. 2014. Two-dimensional data binning for the analysis of genome architecture in filamentous plant pathogens and other eukaryotes. Methods Mol. Biol. 1127:29-51.

Schulz, M. H., Zerbino, D. R., Vingron, M., and Birney, E. 2012. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28:1086-1092.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210-3212.

Slater, G. S. C., and Birney, E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinf. 6:31.

Soskine, M., and Tawfik, D. S. 2010. Mutational effects and the evolution of new protein functions. Nat. Rev. Genet. 11:572-582.

Stam, R., Jupe, J., Howden, A. J. M., Morris, J. A., Boevink, P. C., Hedley, P. E., and Huitema, E. 2013. Identification and characterisation CRN effectors in *Phytophthora capsici* shows modularity and functional diversity. PLoS One 8:e59517.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595.

Thines, M., Choi, Y.-J., Kemen, E., Ploch, S., Holub, E. B., Shin, H.-D., and Jones, J. D. 2009. A new species of *Albugo* parasitic to *Arabidopsis thaliana* reveals new evolutionary patterns in white blister rusts (*Albuginaceae*). Persoonia 22:123-128.

Thines, M., and Spring, O. 2005. A revision of *Albugo* (Chromista, Peronosporomycetes). Mycotaxon 92:443-458.

Torto, T. A., Li, S., Styer, A., Huitema, E., Testa, A., Gow, N. A. R., van West, P., and Kamoun, S. 2003. EST mining and functional expression assays identify extracellular effector proteins from the plant pathogen *Phytophthora.* Genome Res. 13:1675-1685.

Van de Weyer, A.-L., Monteiro, F., Furzer, O. J., Nishimura, M. T., Cevik, V., Witek, K., Jones, J. D. G., Dangl, J. L., Weigel, D., and Bemm, F. 2019. A species-wide inventory of NLR genes and alleles in *Arabidopsis thaliana.* Cell 178:1260-1272.e14.

Voglmayr, H., and Greilhuber, J. 1998. Genome size determination in peronosporales (Oomycota) by Feulgen image analysis. Fungal Genet. Biol. 25:181-195.

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., and Schatz, M. C. 2017. GenomeScope: Fast reference-free genome profiling from short reads. Bioinformatics 33:2202-2204.

Win, J., Morgan, W., Bos, J., Krasileva, K. V., Cano, L. M., Chaparro-Garcia, A., Ammar, R., Staskawicz, B. J., and Kamoun, S. 2007. Adaptive evolution has targeted the C-terminal domain of the RXLR effectors of plant pathogenic oomycetes. Plant Cell 19:2349-2369.

Wood, D. E., Lu, J., and Langmead, B. 2019. Improved metagenomic analysis with Kraken 2. Genome Biol. 20:257.