Albugo candida race diversity, ploidy and host-associated microbes revealed using DNA sequence capture on diseased plants in the field

Jouet, Agathe ; Saunders, Diane G. O. ; McMullan, Mark ; Ward , Ben ; Furzer, Oliver; Jupe, Florian; Cevik, Volkan ; Hein, Ingo; Thilliez, Gaëtan J. A.; Holub, Eric ; Oosterhout, Cock van ; Jones, Jonathan D. G.

Published in:
New Phytologist

DOI:
10.1111/nph.15417

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Jouet, A., Saunders, D. G. O., McMullan, M., Ward , B., Furzer, O., Jupe, F., ... 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 Phytologist, 221(3), 1529-1543. https://doi.org/10.1111/nph.15417

Albugo candida race diversity, ploidy and host-associated microbes revealed using DNA sequence capture on diseased plants in the field

Agathe Jouet[1,2], Diane G. O. Saunders[3], Mark McMullan[4], Ben Ward[2,4], Oliver Furzer[1,5], Florian Jupe[1,6], Volkan Cevik[1,7], Ingo Hein[8,9], Gaetan J. A. Thilliez[8,10], Eric Holub[11], Cock van Oosterhout[2] and Jonathan D. G. Jones[1]

[1]The Sainsbury Laboratory, Norwich Research Park, Norwich, NR4 7UH, UK; [2]School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, UK; [3]John Innes Centre, Norwich Research Park, Norwich, NR4 7UH, UK; [4]The Earlham Institute, Norwich Research Park, Norwich, NR4 7UZ, UK; [5]University of North Carolina, Chapel Hill, NC 27599-2200, USA; [6]Plant Molecular and Cellular Biology Laboratory, Salk Institute, La Jolla CA 92037, USA; [7]The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, BA2 7AY, UK; [8]The James Hutton Institute, CMS, Dundee, DD2 5DA, UK; [9]Division of Plant Sciences at the James Hutton Institute, the University of Dundee, Dundee, DD2 5DA, UK; [10]Quadram Institute Bioscience, Norwich Research Park, Colney Lane, NR4 7UH, Norwich, UK; [11]School of Life Sciences, Warwick Crop Centre, University of Warwick, Warwick, CV35 9EF, UK

Authors for correspondence:
Jonathan Jones
Tel: +44 (0)1603 450400
Email: jonathan.jones@tsl.ac.uk

Cock van Oosterhout
Tel: +44 (0)1603 59 2921
Email: c.van-oosterhout@uea.ac.uk

Received: *17 May 2018*
Accepted: *20 July 2018*

New Phytologist (2018)
doi: 10.1111/nph.15417

Key words: *Albugo candida*, DNA sequence capture, microbiome, oomycete, PenSeq, plant–pathogen coevolution, suppression of plant defence.

Summary

- Physiological races of the oomycete *Albugo candida* are biotrophic pathogens of diverse plant species, primarily the Brassicaceae, and cause infections that suppress host immunity to other pathogens. However, *A. candida* race diversity and the consequences of host immunosuppression are poorly understood in the field.
- We report a method that enables sequencing of DNA of plant pathogens and plant-associated microbes directly from field samples (Pathogen Enrichment Sequencing: PenSeq). We apply this method to explore race diversity in *A. candida* and to detect *A. candida*-associated microbes in the field (91 *A. candida*-infected plants).
- We show with unprecedented resolution that each host plant species supports colonization by one of 17 distinct phylogenetic lineages, each with an unique repertoire of effector candidate alleles. These data reveal the crucial role of sexual and asexual reproduction, polyploidy and host domestication in *A. candida* specialization on distinct plant species. Our bait design also enabled phylogenetic assignment of DNA sequences from bacteria and fungi from plants in the field.
- This paper shows that targeted sequencing has a great potential for the study of pathogen populations while they are colonizing their hosts. This method could be applied to other microbes, especially to those that cannot be cultured.

Introduction

Localized suppression of plant defence may predispose the host tissue to secondary infection by other strains of the same or of different pathogen species. In 1951, Yarwood (1951) reported localized suppression of nonhost resistance to viruses in leaves that were pre-infected by a compatible basidiomycete rust, and other examples of nonhost resistance suppression to viral, fungal and oomycete pathogens have since demonstrated that this is a common feature of infection by rusts and powdery mildew fungi (Moseman & Greely, 1964; Gill, 1965; Yarwood, 1977; Heath, 1980; Lyngkjær & Carver, 2000; Olesen *et al.*, 2003).

Suppression of nonhost resistance also has been demonstrated for the oomycete *Albugo* (Cooper *et al.*, 2008; Belhaj *et al.*, 2017; Prince *et al.*, 2017) and several Brassicaceae species can lose resistance to fungal and oomycete pathogens after pre-inoculation with *Albugo* spp. For example, *Arabidopsis thaliana* can be colonized by *Phytophthora infestans* (late blight of potato and tomato) and *Bremia lactuca* (lettuce downy mildew) after pre-inoculation with either *Albugo laibachii* (Belhaj *et al.*, 2017) or *A. candida* (Prince *et al.*, 2017), and *Brassica juncea* by *Hyaloperonospora arabidopsidis* (*Arabidopsis* downy mildew) after pre-inoculation with *A. candida* (Cooper *et al.*, 2008).

Suppression of nonhost resistance by *Albugo* species is likely to be beneficial for some microorganisms (Agler *et al.*, 2016; Ruhe *et al.*, 2016), but it could also give an evolutionary advantage to the pathogen. For example, *A. candida* comprises at least 10 physiological races that have specialized on different *Brassicaceae* species (Hiura, 1930; Pound & Williams, 1963; Hill *et al.*, 1988; Kaur *et al.*, 2008; Meena *et al.*, 2014) including destructive pathogens of oilseed mustard (*B. juncea,* Race 7), *Brassica oleracea* (Race 9), *B. rapa* (Race 2), radish (*Raphanus sativus,* Race 1), and wild species (*Armoracia rusticana*, Race 3; *Capsella bursa-pastoris*, Race 4; *Sisymbrium officinale,* Race 5 and *Rorippa islandica,* Race 6). McMullan *et al.* (2015) used this race variation to demonstrate that defense suppression by a virulent race of *A. candida* enables subsequent co-colonization by an avirulent race. Suppression of plant defence could therefore provide a means for cohabitation for different physiological races on the same host, and potentially the emergence of new hybrid races as a consequence of inter-race mating. This idea is reinforced by evidence for hybridization between pathogen races or species (Brasier, 2001; Olson & Stenlid, 2002; Ioos *et al.*, 2006; Stukenbrock *et al.*, 2012). More specifically, sexual mating has been demonstrated under laboratory conditions between *A. candida* Races 2 and 7 (Adhikari *et al.*, 2003); and comparative genomics has provided evidence for historical recombination amongst Races 2, 7 and 9 (McMullan *et al.*, 2015). Because *A. candida* can impact both wild and cultivated hosts, populations of susceptible, genetically uniform crop varieties growing adjacent to related wild host species may create suitable conditions for co-infection by different races, and the generation and spread of recombinant variants with novel properties.

Development of methods for the rapid identification of (new) pathogen races is therefore imperative for sustainable disease management. With the advent of current DNA sequencing technologies, researchers can now generate high-resolution genotypic data from any organism. However, the quality of the data depends on the quality and purity of the DNA sample, which means that the pathogen under study needs to be in excess compared to the host plant or cultivated axenically, which is not possible for the biotroph *A. candida*. Recently, the field pathogenomics of wheat yellow rust has been investigated using cDNA sequencing from infected leaves in the field, resulting in *c.* 37% of reads being mapped to the pathogen reference genome (Hubbard *et al.*, 2015). In addition, complexity reduction methods have been developed for sequencing of specific subfractions of plant genomes (e.g. RenSeq (Jupe *et al.*, 2013), MutChromSeq (Sánchez-Martín *et al.*, 2016) and exome capture (e.g. in barley, Mascher *et al.*, 2013, and in wheat, (Henry *et al.*, 2014). In this paper, we report a complexity reduction method (Pathogen Enrichment Sequencing or 'PenSeq') that enables capture and sequencing of plant-associated microbial sequences directly from field samples. We apply this to explore race diversity in *A. candida* and the impact of *A. candida* on the host leaf microbiome.

Materials and Methods

Design of Pathogen Enrichment Sequencing for the investigation of microbial DNA sequences from field sampled leaves

We designed 120mer biotinylated RNA baits based on both mitochondrial genes and nuclear 'housekeeping' genes of 49 microbial species, comprising two rhizaria, five oomycetes, 12 fungi and 30 bacteria with the aim of identifying these species from environmental samples (Supporting Information Table S1). We expect, however, to be able to identify many more species, because baits only require *c.* 80% nucleotide identity to hybridize with and capture DNA before sequencing (Jupe *et al.*, 2013). In addition, we designed baits to capture putative RxLR (Rehmany *et al.*, 2005) and CHxC class effectors (Kemen *et al.*, 2011; Links *et al.*, 2011), and other secreted proteins, from oomycetes. In *Albugo candida*, baits also were designed both to 32 neutrally evolving loci and to a contig of *c.* 400 kb with the aim of investigating the genetic diversity of wild pathogen populations. Neutrally evolving loci were selected using three independent neutrality tests performed on whole-genome data from seven laboratory isolates: Tajima's *D* (Tajima, 1989), Fu's *Fs* (Fu, 1997) and d_N/d_S (Kimura, 1977; Table S2) whereas the *c.* 400 kb contig was the longest contiguous sequence that could be assembled from *A. candida* at the time of the bait library design, using reads from isolate *Ac*Nc2 (Race 4, 'contig 1' (in McMullan *et al.*, 2015), 246 genes including 11 with a predicted secretion signal, no putative effectors). *Ac*Nc2 was one of three races analysed in a previous study that reported on the genetic diversity of five *A. candida* isolates. We also targeted the nuclear internal transcribed spacer (ITS) sequence in plants to facilitate host species identification. In total, 18 348 120mer baits were synthesized that anneal specifically to targeted sequences (> 2 Mb) without overlap (Fig. S1; Table S1 for list of targeted loci). These baits were used to capture DNA from a total of 115 samples, including 91 *A. candida* isolates (Table S3). DNA samples were barcoded and sequenced on three Hiseq lanes with 150 bp paired-end reads.

Preparation of samples for reconstruction experiments

We set up reconstruction experiments to verify that DNA from targeted loci was enriched and sequenced in control samples, and assess the relationship between read depth and DNA abundance. To do this, two sets of controls were prepared (Table S4). In samples #18 to 22, DNA from *Erysiphe cruciferarum*, *Hyaloperonospora arabidopsidis*, *Phytophthora infestans*, *Albugo laibachii*, *Pseudomonas syringae* and *A. candida* was combined in various ways (see also Methods S1). These DNA samples were isolated either from infected leaves, or, for *P. syringae*, grown on medium before DNA extraction. We also included a noninfected *A. thaliana* sample as a negative control (#114). In samples #56 to 59, varying relative molar amounts of DNA from *P. syringae* and *A. candida* were mixed so that the amount of DNA from *P. syringae* increased, and from *A. candida* decreased. Seven additional samples collected on various host species were used to investigate whether the ITS is sufficient for host species identification (#1, 97, 14, 17, 20, 38 and 40). Reads were mapped to all targets and average read depth as well as the breadth of coverage (the percentage of targeted base pairs with at least 10× read depth) was computed for each control organism (see 'Read alignment and consensus sequences calling').

Read alignment and consensus sequences calling

Reads were aligned to targets using the BWA-MEM algorithm of the Burrow–Wheelers Aligner Bwa v.0.7.4 (Li & Durbin, 2009). Read duplicates which may have arisen from amplification of adapter-ligated DNA and post-capture enrichment were discarded using SAMtools v.0.1.19 (rmdup command). Bam files were further processed using ngsutils-0.5.7 to remove reads with > 25% clipped bases and with the highest edit distance to the reference and lowest alignment score in case of multiple alignments (NM and AS tags). Read depth, the number of reads per base, as well as the breadth of coverage, the percentage of bases covered by at least 10 reads, was estimated using the depth command in SAMtools. These statistics were averaged across loci or organisms depending on the analysis performed. Consensus sequences of *A. candida* loci were generated before nucleotide divergence and phylogenetic analyses. To do this, variants (> 10×) were called from the bam file produced by Bwa using the mpileup command of SAMtools and stored in BCF format. Conversion from BCF to VCF and then Fastq format was performed using the 'bcftools view' and 'vcfutils.pl vcf2fq' commands in SAMtools. A short shell script was written in Linux to convert Fastq files into Fasta files. We also independently fed aforementioned bam and vcf files to Shapeit2 v.2.20 (O'Connell *et al.*, 2014) to obtain phased vcf files that were used to generate neutral and putative effector haplotypes using BCFtools v.1.3.1.

Sequences were finally aligned using the multiple alignment program Mafft v7.127 (Katoh & Standley, 2013). The reference *A. candida* isolate used in this study was Nc2 (Race 4; McMullan *et al.*, 2015), except for some effectors that are absent from that particular isolate. There was a total of 6493, 24 249, 202 390 and 398 508 base positions in the final datasets of conserved loci, neutrally evolving loci, putative effectors and the *c.* 400 kb contig, respectively.

Phylogenetic, nucleotide diversity and recombination analyses

Phylogenies were inferred using RAxML v.7.7.3 (Stamatakis, 2014) with the Generalized Time Reversible (GTR) model of nucleotide substitution, gamma distributed rate variation among sites (Gtrgamma) and 100 bootstrap replicates. Pairwise nucleotide divergence was evaluated using Mega v.6.06 (Tamura *et al.*, 2013). It was computed as the number of base differences per site from averaging over all sequence pairs between or within lineages. Ambiguous positions were ignored in each sequence pair. The split network was inferred with SplitsTree v.4.14.2 (Huson & Bryant, 2006) using the UncorrectedP and the NeighborNet methods and 500 bootstrap replicates. A more in-depth genetic diversity and selection analysis was also performed using DNAsp v.5.10.01 (Librado & Rozas, 2009) on neutral and putative effector haplotypes to compute six genetic diversity estimates (number of segregating sites, of haplotypes, of pairwise differences, heterozygosity, pi and theta) and three population genetic statistics (Tajima, 1989; Fu & Li, 1993; Fu, 1997). Finally, recombination analysis was performed using the software HybridCheck (Ward & van Oosterhout, 2016). Using a sliding window approach, this software scans for sudden changes in nucleotide divergence between sequences and identifies potential recombination events when nucleotide identity is significantly increased. Recombination blocks were then dated using the same software and assuming a strict molecular clock with a mutation rate of 10^{-9} mutations/site/generation.

Evaluation of *A. candida* heterozygosity and ploidy level

Throughout this paper, heterozygosity is expressed as the proportion of heterozygous sites within individuals. Before estimating heterozygosity in *A. candida* isolates, sites where read depth was < 50× and/or with a Phred-scaled quality (QUAL) score < 100 were first removed from vcf files prepared above using vcflib (https://github.com/vcflib/vcflib#vcflib) and VCFtools v.0.1.10 (Danecek *et al.*, 2011) and isolates where > 10% sites were removed were discarded. Heterozygosity was then estimated for the remaining isolates as the proportion of heterozygous sites at the *c.* 400 kb contig. The mean and standard deviation of the mean percentage of heterozygous sites per *A. candida* lineage are reported. In addition, the ploidy level of *A. candida* was evaluated using the *c.* 400 kb contig as in Yoshida *et al.* (2013). The distribution of the read proportion of each single nucleotid polymorphism (SNP) at heterozygous sites was graphed for each isolate using R v.3.1.2. This analysis was repeated using whole-genome data from laboratory isolates *Ac*Nc2, *Ac*Em2, *Ac*Ex1, *Ac*2v, *Ac*7v, *Ac*BoT and *Ac*BoL. In this case, reads were aligned to *Ac*Nc2 (McMullan *et al.*, 2015) and to *Ac*2v (Links *et al.*, 2011) contigs. In all ploidy analyses, heterozygous sites where read depth was higher than 1.5 the average read depth were removed. This is to avoid biases due to mismapping of reads from paralogous sequences.

Microbiome analysis

Albugo candida can infect its host both symptomatically and asymptomatically. Therefore, to compare the microbiome of *A. candida*-infected and noninfected samples, we used plant samples from which *A. candida* locus Ev1786 could not be amplified ('healthy' in Table S3, see Methods S1) and symptomatically infected plants. In total, we used 82 infected samples with high read depth at the 400 kb contig and 12 noninfected plants (#62–68, 108–112).

Reads generated for the above samples were fed into Kraken (v.0.10.5). First, the proportion of reads classified as Fungi or as Bacteria was averaged and compared between both groups of samples. In addition, the number of operational taxonomic units (OTUs) detected by Kraken was compared between the infected and noninfected plants using a Mann–Whitney *U*-test. Common and rare microbial OTUs on the *A. candida*-infected plants were identified using a binomial test. In order to correct for an inflated type I error rate due to multiple testing, the critical value was divided by the total number of OTUs identified by Kraken, resulting in a Bonferroni corrected $\alpha' = 1.92 \times 10^{-5}$.

Results

Sequence capture enables detection of defined pathogen sequences from infected leaves in reconstruction experiments

Although some targets were not sequenced in some samples due to presence/absence polymorphisms between the strains used in the bait design and those in the samples, almost all targets were captured and sequenced from control organisms when present (min breadth of coverage = 94.14%, max = 100%; Fig. 1a). Read depth was variable between samples and organisms but in all cases sufficient for downstream analyses (mean read depth (± SD) = 1003 × (± 696), min = 111, Table S4). Additionally, targets were not sequenced in sample #114 where control organisms were absent (mean breadth of coverage (± SD) = 5.87% (± 10.7), mean read depth (± SD) = 7× (± 10)). In all samples, however, a small proportion of reads were wrongly assigned to control organisms likely due to the presence of closely related species (Table S4). Read misalignment was also observed at the ITS sequences of the hosts which were all fully covered by reads due to high homology between species (Fig. 1b). The highest read depth at the ITS was used to identify the host species.

After verification that targeted loci were captured and sequenced to reasonable depth and coverage, the relationship between input DNA and output reads was examined using samples #56–59. Because read depth is highly correlated with the total number of reads generated per sample (Pearson $R^2 = 0.880$, $P < 0.001$), we report on the percentage of reads mapping to each of the target organisms (Fig. 2). We also normalize this against the number of targeted base pairs in *A. candida* and *P. syringae*, respectively, to account for differences between them. Linear regression analyses revealed a positive association between the proportion of reads mapping to control organisms and the amount of DNA that was used during library preparation for both *A. candida* and *P. syringae* ($R^2 = 0.88$ and 0.78). Due to the small sample size, *P*-values were not significant for individual sets of samples ($P = 0.06$ and 0.12, respectively). However, combining both datasets renders the correlation significant ($R^2 = 0.83$, $P < 0.01$). Therefore, although additional samples are needed to generate more robust statistics, Pathogen Enrichment Sequencing (PenSeq) may provide an indication of the relative abundance of pathogen DNA across samples, as well as within samples, providing that the number of targeted base pairs per pathogen is taken into account. Next, we investigated whether PenSeq could be used to investigate both the species diversity and genetic diversity of organisms within field samples.

PenSeq from *A. candida*-infected leaves reveals correspondence of physiological race structure with phylogenetic lineage based on four sets of loci

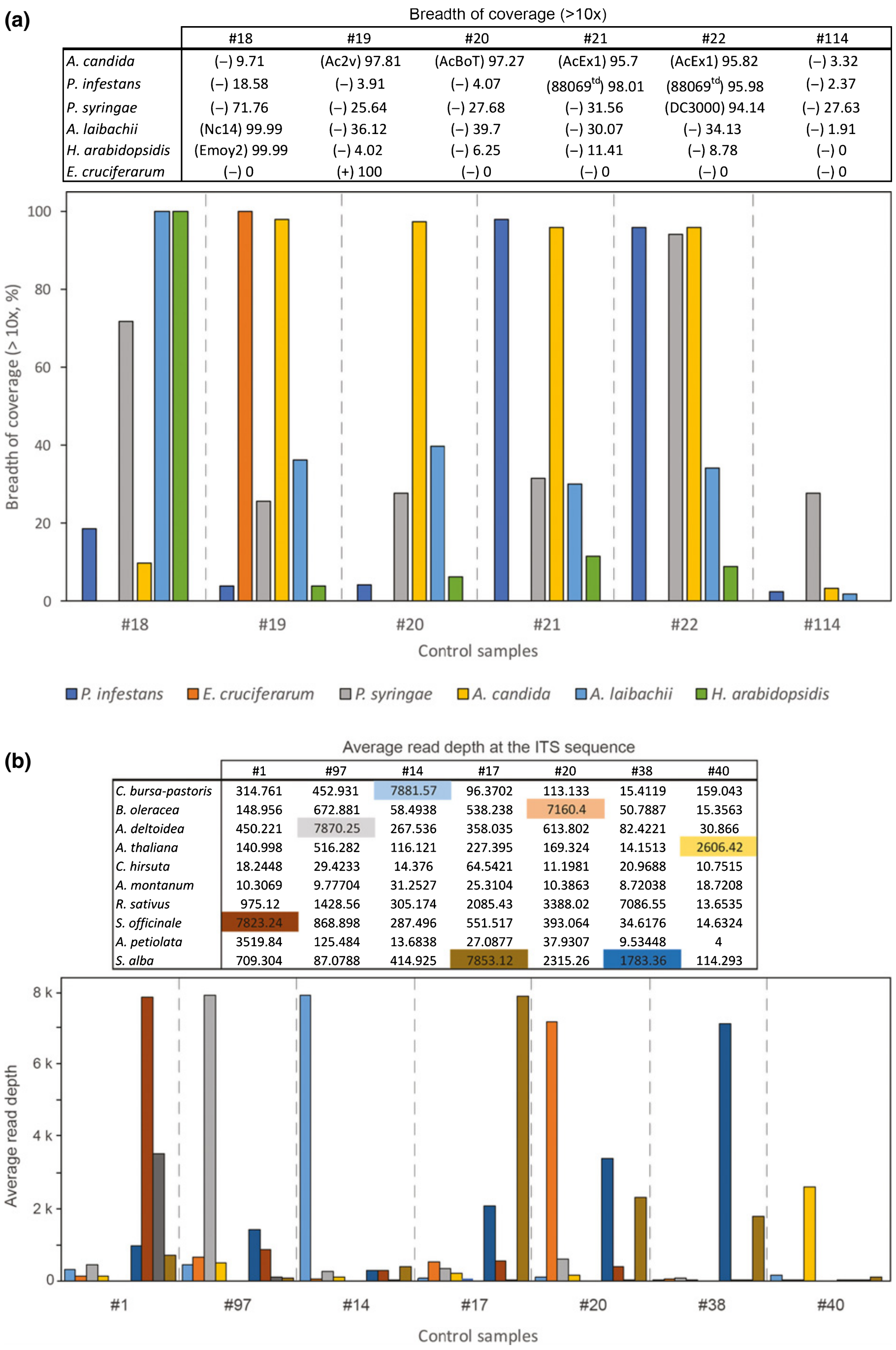
PenSeq data were used to investigate the genetic diversity of *A. candida* natural populations. To do this, 85 isolates were analysed that had high read depth at the *c.* 400 kb contig (mean read depth (± SD) 479 × (± 529); min = 18, max = 2336; Fig. 3; Table S3). Initially, 32 neutrally evolving loci were used to build a phylogeny and measure genetic diversity (Fig. 4a; Table S2). Two additional phylogenetic trees were built using the *c.* 400 kb contig and seven conserved loci (mitochondrial and nuclear 'housekeeping' genes), respectively (Figs 4b, S2). Discrimination of *A. candida* isolates was not possible due to low genetic variation within conserved loci (Fig. S2; Table S5). However, phylogenies built with the neutral loci and the 400 kb contig identified 17 *A. candida* lineages (I–XVII in Fig. 4 and throughout the paper, Table S5 for a detailed description; phylogenetic divergence was inferred if observed in both trees). These lineages consist of isolates collected on the same host or closely related hosts, irrespective of country of origin (mean within lineages p-distance = 0–0.058%). Conversely, isolates collected on different hosts formed distinct lineages (mean between lineages p-distance = 0.36–1.40%), with the exception of isolates collected on radish (*R. sativus*). In accordance with this observation, phylogenetic clustering mostly agreed with physiological race nomenclature with groups X and V corresponding to Races 5 and 9 in both trees, respectively. However, what would be considered as Races 1 (lineages VI and VII), 2 (XIII) and 4 (I) consist of genetically more diverged isolates. In particular, isolates collected on radish (*R. sativus*) are so diverged that isolate #5 clusters distinctly from other radish isolates in both phylogenetic trees.

A phylogenetic tree was also constructed using presence/absence polymorphisms of putative CHxC and related effectors of *A. candida* (Fig. S3). Isolates collected on the same or closely related hosts share similar effector candidate repertoires. Thus, as with the two previous trees, phylogenetic clustering based on these sequences mainly agrees with physiological race designation, including Races 1 (lineages VI and VII) and 4 (I). However, Races 2 and 11 were not discriminated, as in the tree built using the neutral loci (lineages XIII and XIV). Genetic diversity at putative effectors was higher compared to other loci, emphasizing the important role that these proteins may have in host adaptation (Table S5).

The relatedness of some lineages to others is similar in the phylogenies in Fig. 4 with for example, the *B. juncea* virulent races, including the reference genome Race 2 (isolate *Ac*2v), clustering distantly from most others in both datasets. However, this is not always the case and in Fig. 4(a) (neutral loci), isolates infecting *Brassica oleracea* (lineage V, Race 9) appear closely related to *Capsella-*, *Camelina-* and *Arabidopsis*-infecting races (I, II and III), whereas they appear more distantly related in the tree created using the 400 kb contig (Fig. 4b). This observation may be explained either by incomplete lineage sorting (the random sorting of ancestral alleles into the descendant host-specific races) or hybridization by secondary contact of isolates from two host-specific races.

PenSeq enables detection of historical recombination in a 400 kb contig

In order to discriminate between incomplete lineage sorting and hybridization by secondary contact, the software HybridCheck (Ward & van Oosterhout, 2016) was used on the *c.* 400 kb



(a)

Breadth of coverage (>10x)

	#18	#19	#20	#21	#22	#114
A. candida	(–) 9.71	(Ac2v) 97.81	(AcBoT) 97.27	(AcEx1) 95.7	(AcEx1) 95.82	(–) 3.32
P. infestans	(–) 18.58	(–) 3.91	(–) 4.07	(88069^{td}) 98.01	(88069^{td}) 95.98	(–) 2.37
P. syringae	(–) 71.76	(–) 25.64	(–) 27.68	(–) 31.56	(DC3000) 94.14	(–) 27.63
A. laibachii	(Nc14) 99.99	(–) 36.12	(–) 39.7	(–) 30.07	(–) 34.13	(–) 1.91
H. arabidopsidis	(Emoy2) 99.99	(–) 4.02	(–) 6.25	(–) 11.41	(–) 8.78	(–) 0
E. cruciferarum	(–) 0	(+) 100	(–) 0	(–) 0	(–) 0	(–) 0

(b)

Average read depth at the ITS sequence

	#1	#97	#14	#17	#20	#38	#40
C. bursa-pastoris	314.761	452.931	7881.57	96.3702	113.133	15.4119	159.043
B. oleracea	148.956	672.881	58.4938	538.238	7160.4	50.7887	15.3563
A. deltoidea	450.221	7870.25	267.536	358.035	613.802	82.4221	30.866
A. thaliana	140.998	516.282	116.121	227.395	169.324	14.1513	2606.42
C. hirsuta	18.2448	29.4233	14.376	64.5421	11.1981	20.9688	10.7515
A. montanum	10.3069	9.77704	31.2527	25.3104	10.3863	8.72038	18.7208
R. sativus	975.12	1428.56	305.174	2085.43	3388.02	7086.55	13.6535
S. officinale	7823.24	868.898	287.496	551.517	393.064	34.6176	14.6324
A. petiolata	3519.84	125.484	13.6838	27.0877	37.9307	9.53448	4
S. alba	709.304	87.0788	414.925	7853.12	2315.26	1783.36	114.293

Fig. 1 (a) Reconstruction experiment for the detection of control organisms, using samples #18–22 and #114. (Upper panel) Description of the organisms present in control samples (#18–22) as well as the negative control (#114). The name of the strain used in the experiment is provided when the species is present in the sample and a minus sign indicates that the species is absent. In both cases, the breath of coverage at targeted loci is provided (proportion of targeted bases covered by at least 10 reads). (Lower panel) Histogram of the breadth of coverage at targeted loci for each control organism and sample. (b) Reconstruction experiment for the identification of the host species, using samples #1, 97, 14, 17, 20, 38 and 40. (Upper panel) The average read depth is provided for all internal transcribed spacer (ITS) sequences targeted in this study. The host corresponding to each control sample is highlighted in a colour matching the histogram below. (Lower panel) Histogram of the average read depth at the ITS sequence of Brassicaceae in each sample (see Supporting Information Tables S1 and S4 for a detailed description of the samples used in reconstruction experiments).

contig to identify and date regions of high nucleotide identity between *A. candida* races (i.e. putative recombinant regions). If hybridization has occurred recently, after the split between the races, the number of mutations accumulated in the recombinant region will be significantly lower compared to the nucleotide divergence elsewhere in the genome. In total, 159 913 putative recombinant regions of an average of 9569 bp (± SD = 13425) were detected by HybridCheck. These were dated from 5798 (5–95% CI = 0–6104) to 474 852 (5–95% CI = 383 186–579 694) generations ago, assuming a base mutation rate of 10^{-9}

	DNA concentration (mol l^{-1})			
	#56	#57	#58	#59
Albugo candida 2v	1.46E-17	7.30E-18	3.65E-18	1.83E-18
Pseudomonas syringae DC3000	1.42E-17	6.63E-17	9.25E-17	1.05E-16

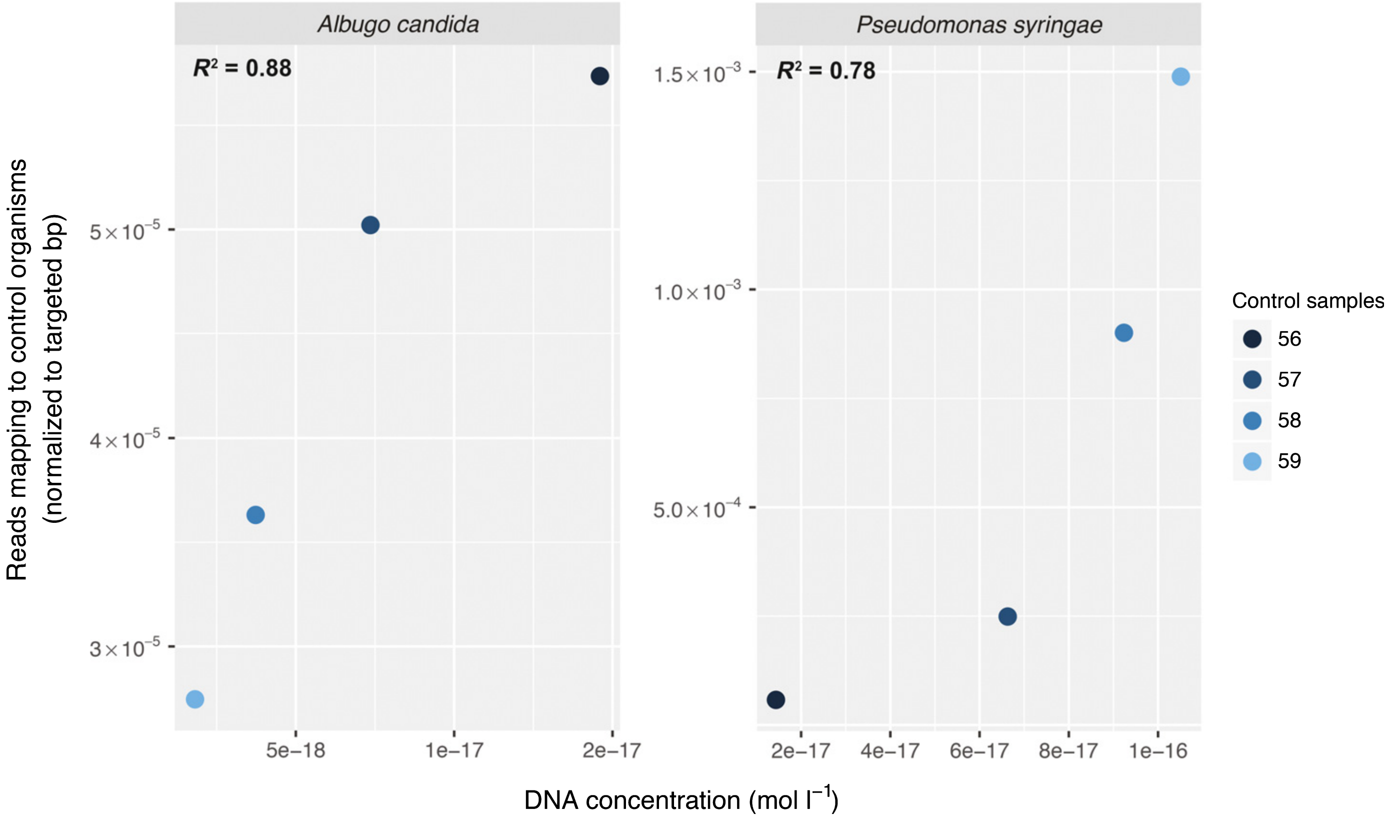


Fig. 2 Correlation between pathogen DNA concentration and pathogen read abundance for two control organisms: (left) *Albugo candida* and (right) *Pseudomonas syringae*. (Upper panel) Molar concentration of DNA (mol l^{-1}) from both *A. candida* and *P. syringae* in samples #56–59. (Lower panels) Scatterplot of the proportion of reads mapping to targeted loci of control organisms (*y*-axis) vs molar concentration (*x*-axis). The proportion of reads mapping to targeted loci was first normalized against the number of targeted base pairs in the control organisms to account for differences between them. A positive correlation was found in both cases and the *r*-square values are provided. Linear regression analyses were conducted in R v3.3.1 (see Supporting Information Tables S3 and S4 for a detailed description of the samples used in reconstruction experiments).

mutations/site/generation (Fig. 5). Although the detection of these ancient regions may be explained by incomplete lineage sorting, recombinant regions that were dated to coalesce recently are best explained by recent hybridization followed by recombination, given that only few or even no mutations distinguish isolates from distinct lineages (Fig. S4).

Finally, as bifurcating phylogenetic trees can only poorly describe the evolutionary history of taxa when recombination-like processes are frequent, a neighbour-net network was built using SplitsTree v.4.14.2 (Huson & Bryant, 2006); Fig. 6). Although based on the *c.* 400 kb contig alone, the network represents a combination of both of the trees shown above confirming some of the divisions suggested in Fig. 4(a,b). Consistent with the recombination analysis performed with HybridCheck, the network identifies events of reticulate evolution (i.e. the partial merging of ancestor lineages), early (long branches, right) or late (short branches, left) during the divergence of *A. candida* lineages.

Genotypes in the 400 kb contig reveal between-race variation in rates of clonal and sexual reproduction

Although there is evidence for sexual reproduction between *A. candida* lineages, verifying the prevalence of sexual reproduction within them is rendered more difficult by low within-race genetic diversity in *A. candida* (p-distance = 0–0.58%). In McMullan *et al.* (2015), the percentage of heterozygous sites shared between isolates of the same race was used to estimate clonality in *A. candida*. The rationale behind this is that the percentage of heterozygous sites shared between two genetically identical isolates of the same lineage only reduces gradually as each isolate accumulates novel mutations (which are mostly in a heterozygous state). By contrast, shared heterozygosity of isolates that reproduce sexually will be reduced by 50% after one generation, and by 75% after two generations of sexual reproduction, assuming diploidy and Mendelian segregation of the alleles.

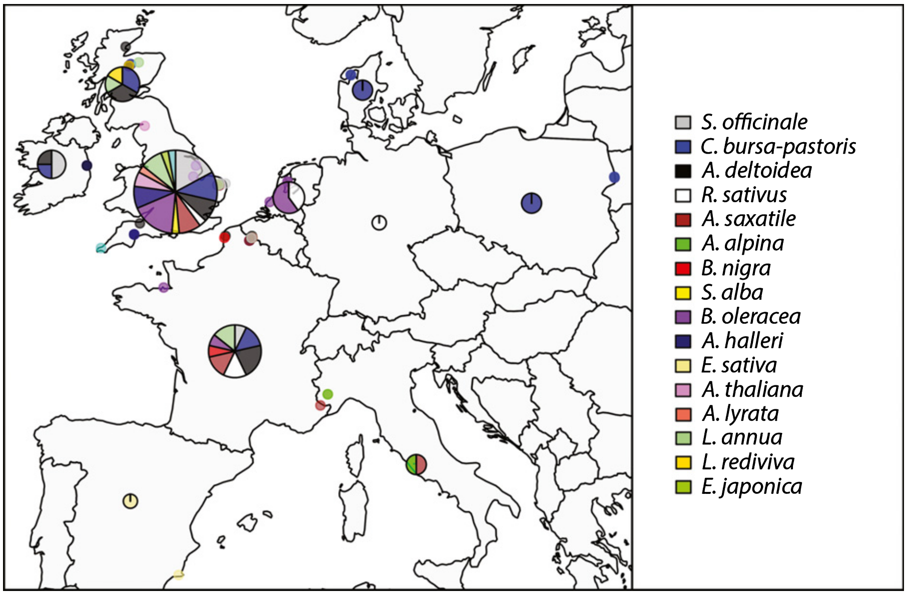


Fig. 3 Distribution of the 91 *Albugo candida* isolates collected during this work. Colours represent the hosts from which the isolates were collected. Dots indicate the positions of isolates as given by their GPS coordinates. When the exact location is not known, the isolate is shown at the centre of the country from which it was collected. Pie charts summarize data from each country and their sizes are scaled to the number of samples collected. Samples from *Brassica juncea* (India), *Brassica carinata, Camelina sativa and Brassica rapa* (Canada) were provided by collaborators and are not represented here (see Supporting Information Table S3 for a detailed description and Table S5 for a description of the dataset per host).

Here, we estimated the level of shared heterozygosity between individuals of the same lineage, as defined above. Throughout the text, the level of heterozygosity is expressed as the proportion of heterozygous sites within individuals at the *c.* 400 kb contig. Although isolates in some lineages had high levels of shared heterozygosity (84.4–97.8%; Fig. 7), others shared few of their heterozygous sites (24.8–55.4%; Fig. 7) which suggests that the importance of sexual reproduction varies between *A. candida* lineages, and we postulate that a 44.6% and 75.2% loss in shared heterozygosity is consistent with one and two generations of sexual reproduction, respectively. The *A. candida* races were previously thought to be evolving mainly clonally based on the sequencing of five isolates (McMullan *et al.*, 2015). Although this could be confirmed for isolates collected on *B. oleracea* (BoT, BoL; Race 9), those from *A. thaliana*, *C. bursa-pastoris* and *B. juncea* (Nc2, Em2; Race 4 and 2v; Race 2) appear to be able to reproduce sexually.

Allele frequency analysis suggests widespread polyploidy in some *A. candida* races

The level of observed heterozygosity at the *c.* 400 kb contig also varied between lineages and although heterozygosity was low in most lineages (e.g. 0.05% for *C. sativa*-specific isolates; Fig. 8), other lineages had intermediate to high levels of heterozygosity (isolates collected on *B. juncea* (0.14% (± 0.09)) and *B. oleracea* (0.65% (± 0.02)) – *R. sativus* (1.15% (± 0.03)), respectively). To better understand this, we investigated the distribution of SNPs at heterozygous sites along the *c.* 400 kb contig using the method published in Yoshida *et al.* (2013). The rationale behind this method is that, for diploid organisms such as *A. candida*, each bi-allelic SNP should account for *c.* 50% of the reads. Shifted proportions of reads per SNP could either be explained by polyploidy or mixed infections. However, this is unlikely to be a consequence of mixed infections if the same pattern is observed in all isolates within a race, collected at different times and locations.

It was not possible to determine ploidy in 11 samples due to the lack of heterozygous sites in the *c.* 400 kb contig (Fig. S5), suggesting mechanisms such as loss-of-heterozygosity (Lamour *et al.*, 2012) or high levels of selfing in these isolates. However, heterozygosity was sufficient for analysis in 72 isolates and a diploid pattern was observed for isolates with low levels of heterozygosity (Fig. 8, yellow). Isolates with intermediate and high levels of heterozygosity showed a tetraploid and a triploid pattern, respectively (*c.* 33 and 67% of reads for each SNP in triploids (red, lineages V and VII; Races 9 and 1) and both 50–50 and 25–75% in putative tetraploids (blue, lineage XIII; Race 2)). This analysis, repeated using whole-genome data of several laboratory isolates (single-spored), could not confirm tetraploidy of isolates collected on *B. juncea* (*Ac*2v, Race 2; Fig. S6 using two reference genomes *Ac*2v (Links *et al.*, 2011) and *Ac*Nc2 (McMullan *et al.*, 2015). This may be because of duplicated regions in the 400 kb contig of *Ac2v* that are not well assembled in the reference genomes. If that is the case, homologous regions would map to the same location, altering the per-SNP read proportion at heterozygous sites in such a way that it resembles that of a tetraploid (i.e. a double diploid). By contrast, triploidy of isolates collected on *B. oleracea* (*Ac*BoT/*Ac*BoL; lineage V, Race 9) was confirmed, based on both single-spore propagated lab strains, and field isolates.

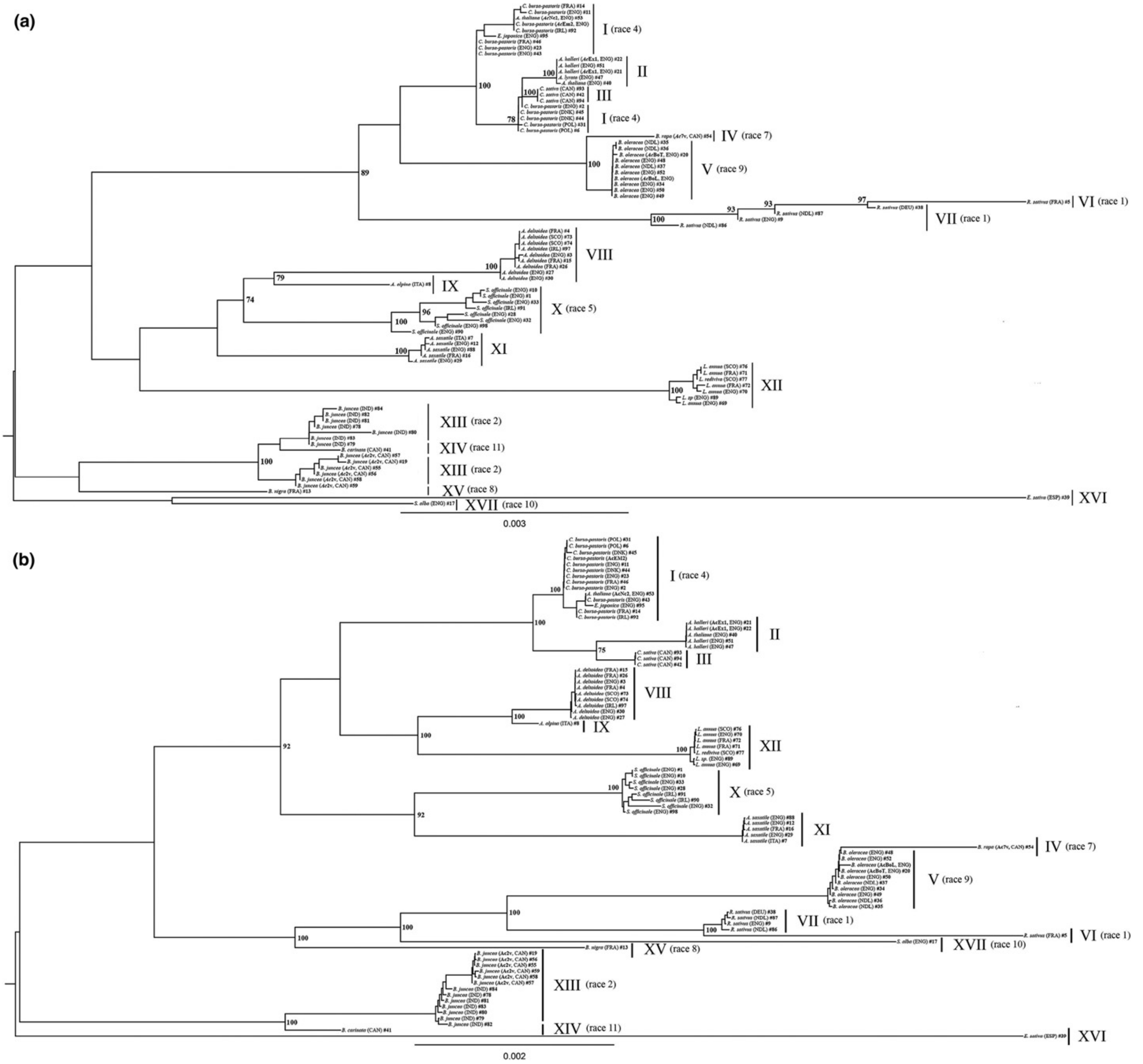


Fig. 4 Maximum-likelihood tree based on (a) the 32 concatenated neutrally evolving loci (24 249 bp) and (b) the *c.* 400 kb contig of *Albugo candida* (398 508 bp) including 85 *A. candida* isolates collected from 20 host species and showing the 17 phylogenetic clusters identified in this study (I–XVII; see Supporting Information Table S5 for a detailed description). The tree was built with RAxML v.7.7.3 using gamma distributed rate variation among sites and the GTR model of DNA evolution. Bootstrap > 70 are shown (100 replicates). The tree was viewed in FIGTREE v.1.3.1 and rooted at midpoint. The scale represents the number of substitutions per site.

Population genetic analysis of effector and neutral genes across host–pathogen associations

We assessed the impact of host lifestyle on *A. candida* evolution in more detail. For this, we evaluated the genetic diversity and selection pressure on both the neutral and putative effector loci of *A. candida* growing on wild hosts, garden escapes and crops. Fig. 9(a) and Table S6 reveal that effectors are significantly more polymorphic than neutral loci, a pattern that is consistent across all summary statistics used to describe genetic variation (Fig. S7). Significant genetic variation was also observed between plant types, and crop-infecting *A. candida* races are on average significantly more polymorphic than those infecting garden and wild plants (Figs 9a, S7; Table S6). The higher level of polymorphism of these crop-infecting races is consistent with their triploidy and (likely) clonal mode of reproduction, which in turn will result in 'frozen heterozygosity' (Schwarz, 2017).

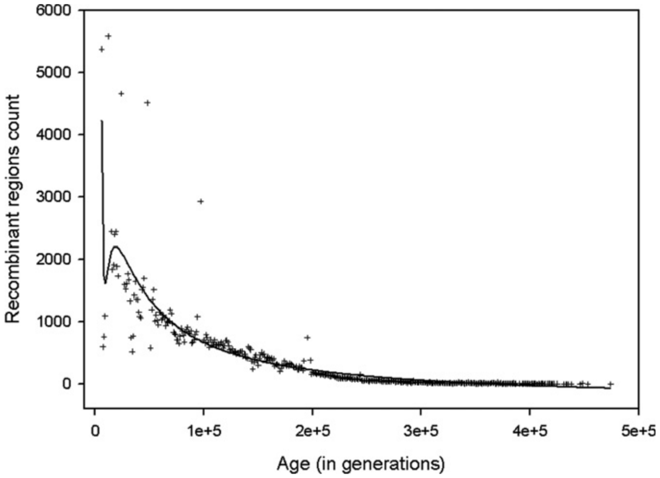


Fig. 5 Estimated age of the recombinant regions detected in the *c.* 400 kb contig of *Albugo candida*. Recombinant regions were identified using HybridCheck. The count of recombinant regions is estimated by sets of 1000 generations. The solid line is a polynomial inverse third order regression line ($R^2 = 0.72$).

Next we analysed the frequency spectrum of segregating sites of effector and neutral genes of *A. candida* using three population genetic statistics (Tajima's *D*, Fu & Li's *D* and Fu's *Fs*; Tajima, 1989; Fu & Li, 1993; Fu, 1997). These statistics can be used to identify selection as well as changes in the population's demography (Tajima, 1989; Slatkin & Hudson, 1991). By including both neutral and effector genes, we designed our experiment with the aim to delineate the demographic signal (expected both in the neutral and the effector loci) from the signal of diversifying selection (effectors only). Surprisingly, neither Tajima's *D* nor Fu & Li's *D* differed significantly between neutral and effector genes (Table S7), and Fu's *Fs* revealed only a marginally significant difference ($P = 0.036$, Fig. 9b; Table S7), which suggests that demographic processes outweigh signatures of diversifying selection. This interpretation is further supported by the consistency across the host plants in how both classes of genes deviate from neutrality (Fig. 9b); in 24 of 27 pairwise comparisons between these genes the deviation from neutrality is in the same direction for neutral as it is for effector genes. Furthermore, all three statistics show a highly significant difference between host types, with *A. candida* of crop plants showing the highest positive values, indicating an excess of intermediate frequency polymorphisms (Fig. 9b; Table S7). Such an excess is consistent with balancing selection (Slatkin & Hudson, 1991), but given that it is observed also at neutral genes, we rule out this explanation. Rather, we propose that this observation is caused by a population bottleneck, which may be the result of resistance-breaking by a small number of genotypes to a genetically relatively uniform crop.

Detection of colonization by other microbes in *Albugo*-infected field samples

We used baits designed to capture sequences from 12 genes from 42 bacterial and fungal species to investigate the presence of microorganisms on both *A. candida*-infected and noninfected plants (Table S3). We classified reads from each sample to taxonomic units (OTUs) using the software KRAKEN (Wood & Salzberg, 2014). We could detect and phylogenetically assign microbial reads from both *A. candida*-infected and noninfected plants (Fig. 10). Interestingly, we found on average more reads from bacteria on plants infected by *A. candida* compared to noninfected plants (infected: median (1st and 3rd quartile) = 1.695% (0.885–3.030); noninfected: 0.670% (0.153–2.185)), Mann–Whitney *U*-test; $W = 5.74$, df = 1, $P < 0.017$). A similar result was observed with fungi-derived reads (infected: 1.973% (1.205–2.433); noninfected: 0.090% (0.045–0.165), Mann–Whitney *U*-test; $W = 78.0$, df = 1, $P < 0.00001$). In addition, the mean (± SEM) number of OTUs detected is larger on *A. candida*-infected plants (infected: 473 (± 14); noninfected: 358 (± 51), Mann–Whitney *U*-test; $W = 167.0$, $P < 0.00001$), and some of these OTUs appear to be associated with the presence of *A. candida* (e.g. Rickettsiales and Chytridiomycetes, Table S8). However, these preliminary results warrant a more focused, quantitative analysis to assess the extent with which *A. candida* infection elevates the colonization by other microbes in the field.

Discussion

PenSeq enables cost-effective investigation of DNA sequence variation in pathogen genotypes from field samples

We report here a sequence capture-based protocol (PenSeq) that enables cost-effective assessment of pathogen genetic diversity in field samples. We designed baits based on specific sequences from *c.* 50 microbial species, and although we may have introduced biases associated to baits, we anticipated we would recover targeted gene sequences from essentially all likely plant-associated bacteria because only 80% nucleotide identity is required for baits to hybridize with DNA (Jupe *et al.*, 2013). We were able to recover all targets, at high read depth, from control pathogens when present in reconstruction experiments. In addition, we showed that the more abundant the microbe in the sample, the higher the proportion of reads observed. PenSeq thus reveals not only presence–absence differences, but also relative abundance of microbes from field samples. With this in mind, we compared the phyllosphere of *A. candida*-infected and noninfected plants and found an increase in the abundance and diversity of bacteria and fungi, which supports the hypothesis that *A. candida* suppresses the host's immune response, impacting microbial communities. The most striking examples are the Rickettsiales (Proteobacteria) and Chytridiomycetes that were detected in all 82 *A. candida*-infected plants but on none of the 12 noninfected plants. A caveat in this association analysis is that many variables, other than infection by *A. candida,* may play a role in the presence or absence of microbes. We therefore advise conducting microbiome-focused studies to confirm and quantitatively assess the impact of *A. candida* infection on the various associated microbes.

During the same experiment, we sequenced a total of *c.* 660 kb of the *A. candida* genome from 91 isolates (187 loci including putative effectors). We used these data to identify distinct

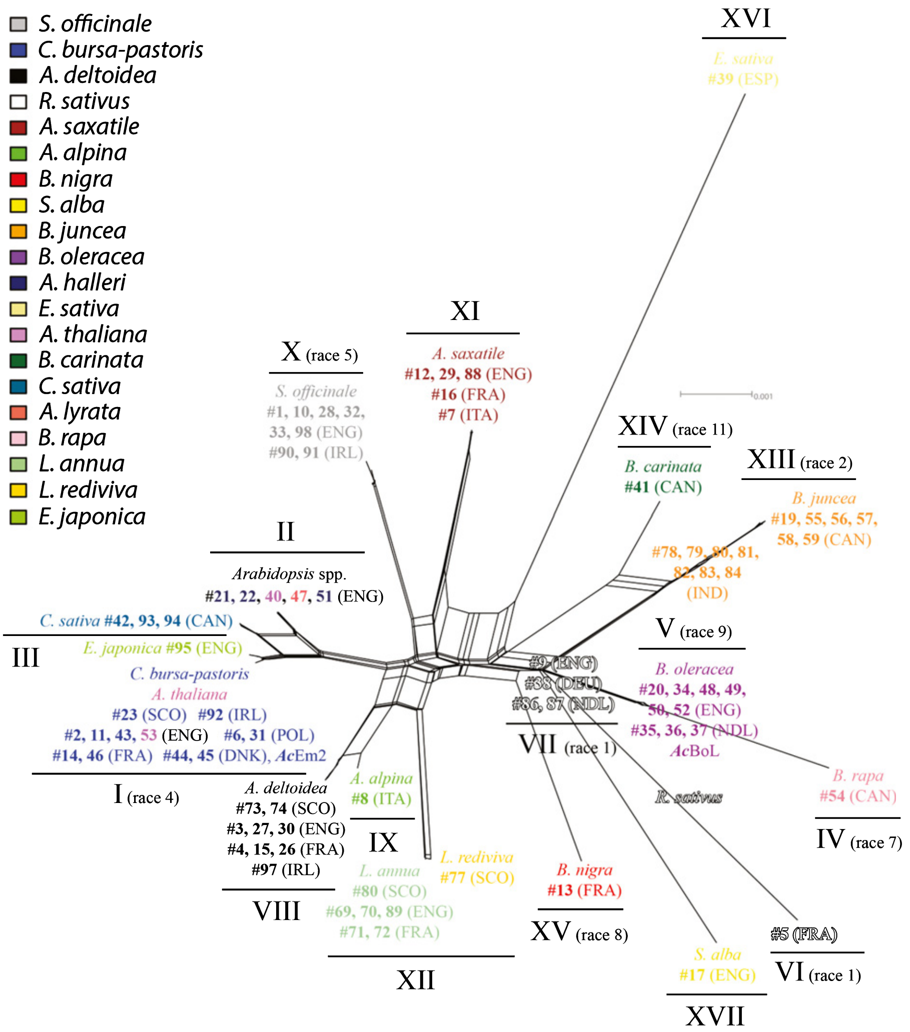


Fig. 6 Neighbour-Net network built using SplitsTree v.4.11.3 and based on the *c.* 400 kb contig (398 508 bp) of 85 *Albugo candida* isolates (including *Ac*BoL and *Ac*Em2). The scale represents distances estimated using the uncorrected p-distance. Isolates are colour-coded according to the hosts they were collected on and phylogenetic lineages identified in Fig. 4 are shown. A detailed description of the *A. candida* isolates is available in Supporting Information Table S1.

lineages that had diverged from each other by 0.29–1.15%. This could not be achieved using a limited number of loci (Saharan *et al.*, 2014) or would have been very expensive and laborious using whole-genome sequencing of purified isolates. These lineages appeared to be mostly specialized on one host species and to be consistent with what has been defined as a pathotype or race in *A. candida* (Biga, 1955; Pound & Williams, 1963; Verma, 2012). Conceivably, some *A. candida* lineages can infect closely related host species that were not sampled in this study, as observed for lineage I (Race 4) infecting *C. bursa-pastoris*, *A. thaliana* and *Eutrema japonica* and lineage XII infecting *Lunaria annua* and *Lunaria rediviva*.

With PenSeq, we were able to sequence *c.* 115 samples on three HiSeq lanes, by multiplexing up to 47 samples in a lane. We could investigate not only the genetic diversity in *A. candida*, but also the presence and abundance of other microbes from samples collected in the field. This method enables inspection of many genetic marker loci and is a good alternative to expensive whole-genome sequencing of microbiota and to PCR-based methods such as internal transcriber spacing (ITS) and 16S sequencing. Potential applications would be to study the coevolution of host resistance genes and pathogen effectors in natural populations or of microorganisms within specific environments. Sequence capture has been used to sequence viral genomes from clinical samples (VirCapSeq; Briese *et al.*, 2015) and in a companion paper, PenSeq was used to investigate putative effector sequences in *P. infestans* and *Phytophthora capsici* (Thilliez *et al.*, 2018). Although field pathogenomics can also be very helpful for assigning races to lineages (Hubbard *et al.*, 2015), sequence capture enables the sequences under investigation to be predefined.

A rich history of recombination breakpoints revealed in comparisons of 17 different races

Hybridization between *A. candida* lineages may have enabled adaptation to the numerous hosts on which the pathogen has been reported (Adhikari *et al.*, 2003; McMullan *et al.*, 2015).

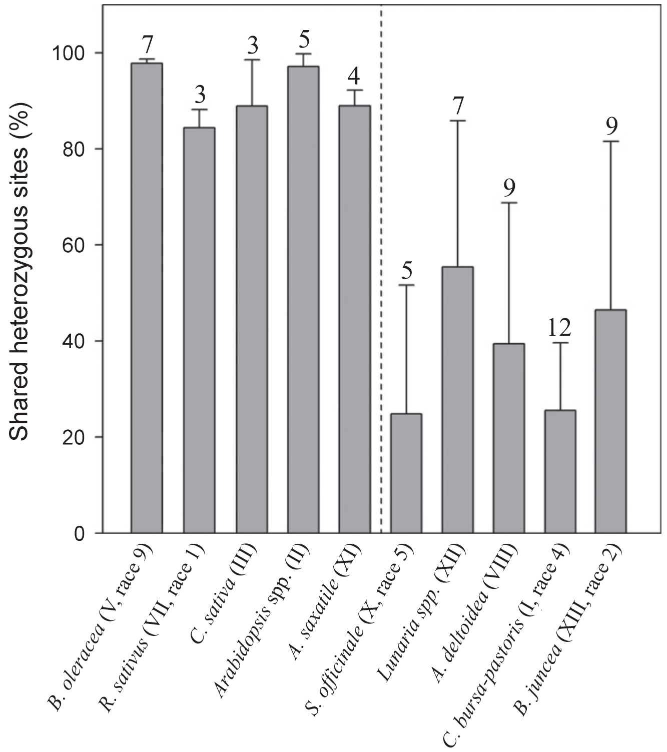


Fig. 7 Mean percentage of shared heterozygous sites (*y*-axis) per *Albugo candida* phylogenetic cluster (*x*-axis). The number of isolates used in each group is provided at the top of each bar (races with less than two sampled isolates were discarded from the analysis). Left of the dashed line are races sharing most of their heterozygous sites (clonal, > 84%) and right are races that share few of their heterozygous sites (sexual, < 55%). Standard deviation is also shown. It is larger for races that reproduce mainly sexually suggesting stochasticity in segregation.

Genetic exchange between lineages may occur if they share a common host (Pound & Williams, 1963; Saharan *et al.*, 2014) or if the immune system of the host is compromised by a compatible isolate (Cooper *et al.*, 2008; Belhaj *et al.*, 2017; Prince *et al.*, 2017). By sampling more isolates from different host species, we set out to better understand the role of hybridization in *A. candida* evolution.

In the present study, we identified many putative recombinant regions between lineages within the *c.* 400 kb contig. Some of these regions are quite polymorphic and are probably due to incomplete lineage sorting or represent trans-species polymorphism. However, other regions are (nearly) identical between races over long stretches of sequence, suggesting that they arose by recent hybridization. In addition, although the existence of tetraploid races of *A. candida* is unproven, the high levels of heterozygosity in triploid lineages (*B. oleracea* and *R. sativus*; lineages V and VII, Races 9 and 1) suggests that they originated from intercrosses between diploids and tetraploids rather than from the absence of reduction division in either male or female germ cells. This idea is reinforced by the mid-branch placement of triploids in the network shown above (lineages V and VII (Races 9 and 1) mid-branch of lineages IV and VI (Races 7 and 1)). However, no mixed infections were detected which suggests that this scenario is rare. Conceivably, genetic exchanges between lineages that are adapted to diverged hosts (each with their unique sets of resistance genes) would most likely be maladaptive, in comparison with combinations of alleles which have been selected over extended evolutionary timescales (Brasier, 2001).

Nonetheless, evidence for the emergence of novel pathogen races or species through hybridization continues to accumulate, particularly for crop species (Stukenbrock & McDonald, 2008; Gladieux *et al.*, 2011; Leroy *et al.*, 2016; Stukenbrock, 2016). For example, *Phytophthora alni* on *Alnus* spp. arose via sexual reproduction between *Phytophthora cambivora* and a species related to *Phytophthora fragariae*, both nonpathogenic to *Alnus* spp. (Brasier *et al.*, 2004). Likewise, *Blumeria graminis* f. sp. *triticale*, a powdery mildew found on triticale and wheat, is a hybrid between *B. graminis* f. sp. *tritici* pathogenic to wheat and *B. graminis* f. sp. *secalis* to rye (Menardo *et al.*, 2015).

Polyploidy is a significant contributor to race diversification and evolution

Polyploidy in *A. candida* may have important impacts on both the occurrence of sexual reproduction within and between lineages and the adaptive potential of the lineages. First, polyploid isolates may have reduced fertility or be strictly asexual (especially triploids; Comai, 2005). In triploid isolates from *B. oleracea* (V, race 9) and *R. sativus* (VII, race 1), heterozygous sites are mostly shared between isolates, suggesting strict clonal reproduction. The hypothesis of asexuality in these lineages is reinforced by the lack of observable oospores from isolates collected on *B. oleracea* and propagated in the laboratory (*Ac*BoT and *Ac*BoL; V. Cevik, pers. comm.; McMullan *et al.*, 2015). Second, polyploid isolates may not be able to hybridize with other lineages (Pannell *et al.*, 2004; Köhler *et al.*, 2010). This may, in the long term, lead to speciation of the polyploid lineages. Finally, polyploidy may allow for relaxed selection pressure and therefore increased mutation rates at the duplicated genome(s) which may lead to the acquisition of novel gene functions (Comai, 2005; Madlung, 2012). It may also increase vigour (increased biomass, growth rate) in comparison to diploid races. Interestingly, polyploid races identified during this work had previously been classified as *A. candida macrospora* due to large sporangia, compared to those observed in *A. candida microspora* (e.g. isolates collected on *C. bursa-pastoris* or *Sisymbrium* spp.; Biga, 1955; Pound & Williams, 1963). It would be interesting to test for polyploidy in other lineages classified as *macrospora* (other *Brassica* spp. or *Erucastrum* spp.) and for an increased sporangia/zoospore resistance or survival in the polyploids.

In *A. candida,* polyploidy appears to have been selected for only in cultivated host species. This is interesting because polyploidy may allow for (and indeed impose a need for) rapid clonal expansion of an adapted race on the relatively uniform genotypes of cultivated host populations. We indeed observe this with crop-infecting (mostly polyploid) *A. candida* races showing a relative excess of intermediate frequency variants, both at neutral as well as effector genes. Conceivably, this observation is caused by a genetic bottleneck resulting from resistance-breaking by a small number of genotypes during the widespread cultivation of

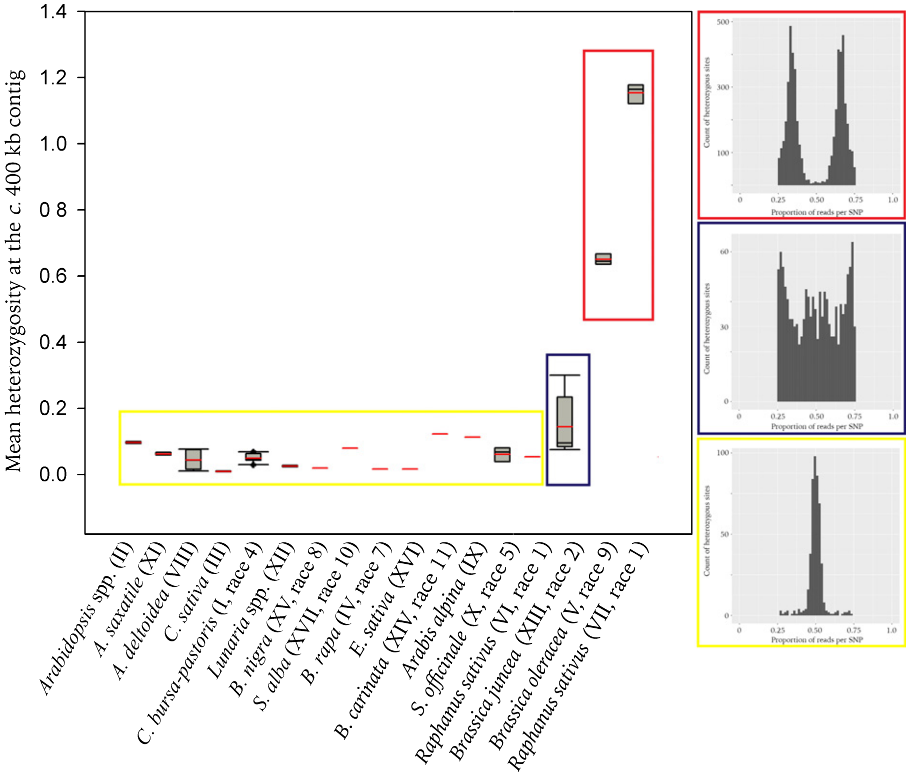


Fig. 8 Mean heterozygosity of *Albugo candida* phylogenetic lineages at the *c*. 400 kb contig (398 508 bp) along with representative ploidy graphs for races with low (yellow), intermediate (blue) and high levels of heterozygosity (red) based on that same contig. Heterozygosity is expressed as the percentage of observed heterozygous sites. Median (black solid line) and mean (red solid line) heterozygosity are provided. In the ploidy graphs, the *x*-axis represents the proportion of reads per single nucleotide polymorphism (SNP) at heterozygous positions and the *y*-axis is the count of heterozygous sites.

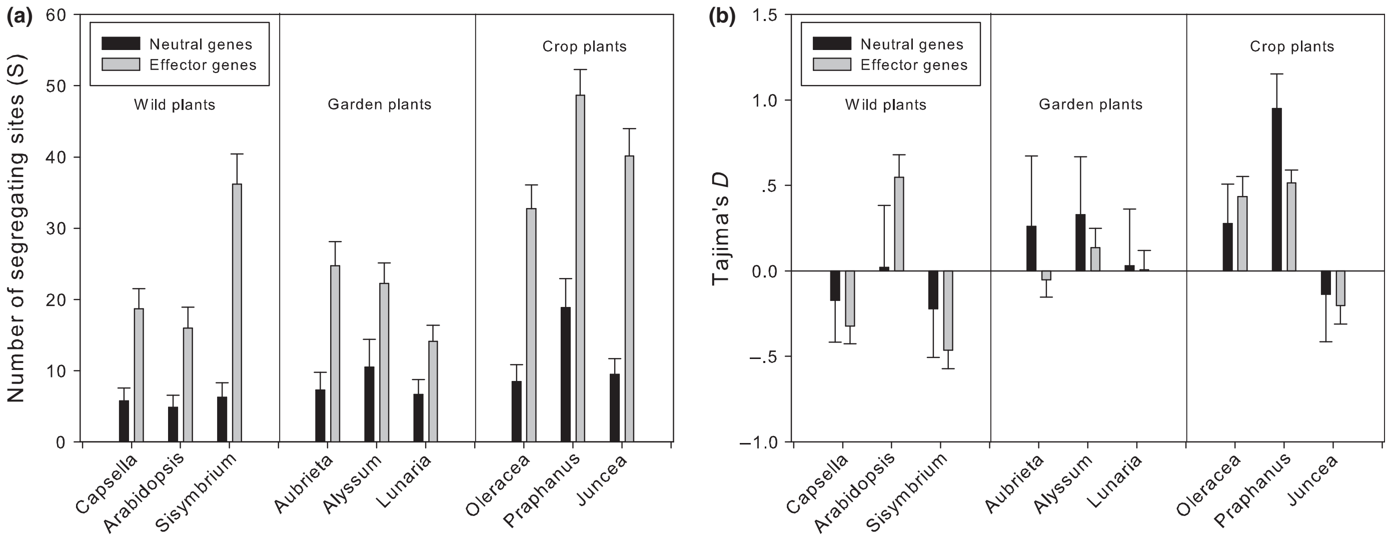


Fig. 9 Genetic diversity and selection at neutral and putative effector loci of *Albugo candida* races infecting wild plants, garden escapes and crops. Shown are the mean and SD of (a) The number of segregating sites *S* and (b) Tajima's *D*, both computed using DNAsp v5.10.01. Other summary statistics for genetic diversity (the number of haplotypes (*H*), heterozygosity (H_t), the average number of nucleotide differences, and the mutation-drift parameter theta (Θ)) and selection (Fu & Li, 1993; Fu, 1997) are shown in Supporting Information Figs S7 and S8, respectively.

Brassica crops in Northern Europe over the last 2000 yr (Maggioni *et al*., 2000). In addition, asexuality associated with polyploidy may be beneficial as sexual reproduction would break up combinations of alleles that are adapted to the host genotypes. It also can result in a phenomenon known as 'frozen heterozygosity' (Schwarz, 2017), in which the polymorphism contained within genetic loci becomes fixed in the asexual lineage, which enables it to resist the eroding effects of genetic drift. By contrast, clonal reproduction may be less advantageous for the pathogen in a coevolutionary arms race with host species that are genetically

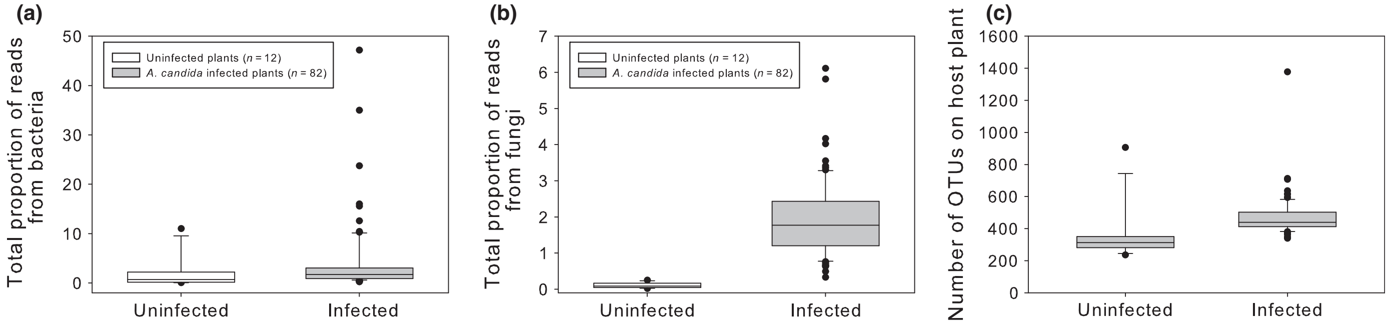


Fig. 10 Microbial abundance and diversity in both *Albugo candida*-infected and noninfected plants. Kraken v.0.10.5 was used for read classification. (a) Proportion of reads classified as Bacteria. (b) Proportion of reads classified as Fungi. (c) Number of operational taxonomic units (OTUs) identified by Kraken. The proportion of microbial reads and the number of microbial OTUs are significantly larger on *A. candida*-infected plants compared to noninfected plants. The distributions are shown in boxplots, in which the dots represent the outliers, the bars the lower and upper limits, and the box the first and third quartile value around the median.

diverse, such as, for example, pathogens infecting wild populations of *C. bursa-pastoris* or *S. officinale*.

In summary, we have established PenSeq, a cost-effective sequence capture-based method to investigate genetic diversity on pathogens while they are colonizing their hosts. We used this method to substantially enhance our understanding of *A. candida* race diversity, and to reveal associated leaf microbiota in infected leaves. We believe that this method has broad potential applications ranging from investigating obligate biotrophic pathogens such as rusts, downy mildews and powdery mildews, to diagnostics of symptomless pathogens and even to investigating mammalian diseases, such as genotyping parasites in blood samples.

Acknowledgements

This work was supported by Earth & Life Alliance Systems (to AJ and CvO); the Gordon and Betty Moore foundation (to AJ and JDGJ); the BBSRC (BB/L011646/1 to EH; BB/M003809/1, BB/L011646/1 to JDGJ and BB/N02317X/1 to CvO); EMBO (ALTF 842-2015 to JDGJ) and by the Norwich Biosciences Institutes Computing infrastructure for Science (CiS) group. Reads generated using PenSeq were made publicly available at the European Nucleotide Archive (ENA) under the study accession number: PRJEB21033. We thank Tove H Jorgensen (Aarhus University, Århus, Denmark), Ulrike Miersch and Annemarie Lokerse (Rijk Zwaan, De Lier, the Netherlands), Deepak Pental and Abha Agnihotri (University of Delhi and Amity University, India, respectively), Hossein Borhan (Agriculture and Agri-Food Canada, Saskatoon), Yan Ma and Chih-Hang Wu (The Sainsbury Laboratory, Norwich, UK), Mark Wilson (through the White Rust Survey website: http://agathejouet.wixsite.com/white-rust-survey) and Quentin Dupriez for collecting *A. candida* samples. Thanks to Marco Thines and Young-Joon Choi for their help with *A. candida* samples collection (Goethe University, Frankfurt, Germany).

Author contributions

AJ conducted fieldwork; AJ, GJAT, IH and JDGJ designed the bait library; AJ, OF and FJ performed sequence capture; VC prepared *A. candida* lab samples; AJ, MM, OF, DGOS, BW and CvO conducted bioinformatics analyses; AJ, EH, CvO and JDGJ wrote the manuscript; and AJ, CvO and JDGJ designed the research.

ORCID

Oliver Furzer http://orcid.org/0000-0002-3536-9970
Ingo Hein http://orcid.org/0000-0002-0128-2084
Eric Holub http://orcid.org/0000-0003-3341-3808
Jonathan D. G. Jones http://orcid.org/0000-0002-4953-261X
Agathe Jouet http://orcid.org/0000-0003-4998-9596
Florian Jupe http://orcid.org/0000-0001-5741-4931
Mark McMullan http://orcid.org/0000-0002-0711-5666
Cock van Oosterhout http://orcid.org/0000-0002-5653-738X
Ben Ward http://orcid.org/0000-0001-6337-5238

References

Adhikari TB, Liu JQ, Mathur S, Wu CX, Rimmer SR. 2003. Genetic and molecular analyses in crosses of race 2 and race 7 of *Albugo candida*. *Phytopathology* **93**: 959–965.

Agler MT, Ruhe J, Kroll S, Morhenn C, Kim S-T, Weigel D, Kemen E. 2016. Microbial hub taxa link host and abiotic factors to plant microbiome variation. *PLoS Biology* **14**: e1002352.

Belhaj K, Cano LM, Prince DC, Kemen A, Yoshida K, Dagdas YF, Etherington GJ, Schoonbeek HJ, van Esse HP, Jones JDG *et al.* 2017. *Arabidopsis* late blight: infection of a nonhost plant by *Albugo laibachii* enables full colonization by *Phytophthora infestans*. *Cellular Microbiology* **19**: e12628.

Biga ML. 1955. Riesaminazione delle specie del genere *Albugo* in base all morfologia dei conidi. *Sydowia* **9**: 339–358.

Brasier CM. 2001. Rapid evolution of introduced plant pathogens via interspecific hybridization. *BioScience* **51**: 123–133.

Brasier CM, Kirk SA, Delcan J, Cooke DE, Jung T, Man in't Veld WA. 2004. *Phytophthora alni* sp. *nov.* and its variants: designation of emerging heteroploid hybrid pathogens spreading on *Alnus* trees. *Mycological Research* **108**: 1172–1184.

Briese T, Kapoor A, Mishra N, Jain K, Kumar A, Jabado OJ, Lipkin W. 2015. Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis. *mBio* **6**: 1–11.

Comai L. 2005. The advantages and disadvantages of being polyploid. *Nature Reviews Genetics* **6**: 836–846.

Cooper AJ, Latunde-Dada AO, Woods-Tor A, Lynn J, Lucas JA, Crute IR, Holub EB. 2008. Basic compatibility of *Albugo candida* in *Arabidopsis thaliana* and *Brassica juncea* causes broad-spectrum suppression of innate immunity. *Molecular Plant–Microbe Interactions* **21**: 745–756.

Danecek P, Auton A, Abecasis GR, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST *et al.* 2011. The variant call format and VCF tools. *Bioinformatics* **27**: 2156–2158.

Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**: 915–925.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* **133**: 693–709.

Gill CC. 1965. Increased multiplication of viruses in rusted bean and sunflower tissue. *Phytopathology* **55**: 141–147.

Gladieux P, Vercken E, Fontaine MC, Hood ME, Jonot O, Couloux A, Giraud T. 2011. Maintenance of fungal pathogen species that are specialized to different hosts: allopatric divergence and introgression through secondary contact. *Molecular Biology and Evolution* **28**: 459–471.

Heath MC. 1980. Effects of infection by compatible species or injection of tissue extracts on the susceptibility of nonhost plants to rust fungi. *Phytopathology* **70**: 356–360.

Henry IM, Nagalakshmi U, Lieberman MC, Ngo KJ, Krasileva KV, Vasquez-Gross H, Akhunova A, Akhunov E, Dubcovsky J, Tai TH *et al.* 2014. Efficient genome-wide detection and cataloging of EMS-Induced mutations using exome capture and next-generation sequencing. *Plant Cell* **26**: 1382–1397.

Hill CB, Crute IR, Sherriff C, Williams PH. 1988. Specificity of *Albugo candida* and *Peronospora parasitica* pathotypes toward rapid-cycling Crucifers. *Cruciferae Newsletter* **13**: 112–113.

Hiura M. 1930. Biologic forms of *Albugo candida* (Pers.) Kuntze on cruciferous plants. *Journal of Japanese Botany* **5**: 1–20.

Hubbard A, Lewis CM, Yoshida K, Ramirez-Gonzalez RH, de Vallavieille-Pope C, Thomas J, Kamoun S, Bayles R, Uauy C, Saunders DGO. 2015. Field pathogenomics reveals the emergence of a diverse wheat yellow rust population. *Genome Biology* **16**: 23.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23**: 254–267.

Ioos R, Andrieux A, Marçais B, Frey P. 2006. Genetic characterization of the natural hybrid species *Phytophthora alni* as inferred from nuclear and mitochondrial DNA analyses. *Fungal Genetics and Biology* **43**: 511–529.

Jupe F, Witek K, Verweij W, Śliwka J, Pritchard L, Etherington GJ, Maclean D, Cock PJ, Leggett RM, Bryan GJ, *et al.* 2013. Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. *Plant Journal* **76**: 530–544.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772–780.

Kaur P, Sivasithamparam K, Barbetti MJ. 2008. Host range and phylogenetic relationships of *Albugo candida* from cruciferous hosts in Western Australia, with special reference to *Brassica juncea. Plant Disease* **95**: 712–718.

Kemen E, Gardiner A, Schultz-Larsen T, Kemen A, Balmuth AL, Robert-Seilaniantz A, Bailey K, Holub EB, Studholme DJ, Maclean D *et al.* 2011. Gene gain and loss during evolution of obligate parasitism in the white rust pathogen of *Arabidopsis thaliana. PLoS Biology* **9**: e1001094.

Kimura M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**: 275–276.

Köhler C, Mittelsten Scheid O, Erilova A. 2010. The impact of the triploid block on the origin and evolution of polyploid plants. *Trends in Genetics* **26**: 142–148.

Lamour KH, Mudge J, Gobena D, Hurtado-Gonzales OP, Schmutz J, Kuo A, Miller NA, Rice BJ, Raffaele S, Cano LM *et al.* 2012. Genome sequencing and mapping reveal loss of heterozygosity as a mechanism for rapid adaptation in the vegetable pathogen *Phytophthora capsici. Molecular Plant–Microbe Interactions* **25**: 1350–1360.

Leroy T, Caffier V, Celton JM, Anger N, Durel CE, Lemaire C, Le Cam B. 2016. When virulence originates from nonagricultural hosts: evolutionary and epidemiological consequences of introgressions following secondary contacts in *Venturia inaequalis. New Phytologist* **210**: 1443–1452.

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

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics (Oxford, England)* **25**: 1451–1452.

Links MG, Holub EB, Jiang RHY, Sharpe AG, Hegedus D, Beynon E, Sillito D, Clarke WE, Uzuhashi S, Borhan MH. 2011. *De novo* sequence assembly of *Albugo candida* reveals a small genome relative to other biotrophic oomycetes. *BMC Genomics* **12**: 503.

Lyngkjær MF, Carver TL. 2000. Conditioning of cellular defence responses to powdery mildew in cereal leaves by prior attack. *Molecular Plant Pathology* **1**: 41–49.

Madlung A. 2012. Polyploidy and its effect on evolutionary success: old questions revisited with new tools. *Heredity* **110**: 99–104.

Maggioni L, von Bothmer R, Poulsen G, Branca F. 2000. Origin and domestication of cole crops (*Brassica oleracea* L.): linguistic and literary considerations. *Economic Botany* **64**: 109–123.

Mascher M, Richmond TA, Gerhardt DJ, Himmelbach A, Clissold L, Sampath D, Ayling S, Steuernagel B, Pfeifer M, D'Ascenzo M *et al.* 2013. Barley whole exome capture: a tool for genomic research in the genus *Hordeum* and beyond. *Plant Journal* **76**: 494–505.

McMullan M, Gardiner A, Bailey K, Kemen E, Ward BJ, Cevik V, Robert-Seilaniantz A, Schultz-Larsen T, Balmuth A, Holub EB *et al.* 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**: e0455.

Meena PD, Verma PR, Saharan GS, Hossein Borhan M. 2014. Historical perspectives of white rust caused by *Albugo candida* in oilseed *Brassica. Journal of Oilseed Brassica* **5**: 1–41.

Menardo F, Praz C, Wyder S, Bourras SA, McNally KE, Parlange F, Riba A, Roffler S, Schaefer L, Shimizu KK *et al.* 2015. Hybridization of powdery mildew strains gives raise to pathogens on novel agricultural crop species. *Nature Genetics* **48**: 1–24.

Moseman JG, Greely LW. 1964. Predisposition of wheat by *Erysiphe graminis* f. sp. *tritici* to infection with *Erysiphe graminis* f. sp. *hordei. Phytopathology* **54**: 618.

O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, Traglia M, Huang J, Huffman JE, Rudan I *et al.* 2014. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genetics* **10**: e1004234.

Olesen KL, Carver TLW, Lyngkjær MF. 2003. Fungal suppression of resistance against inappropriate *Blumeria graminis* formae speciales in barley, oat and wheat. *Physiological and Molecular Plant Pathology* **62**: 37–50.

Olson Å, Stenlid J. 2002. Pathogenic fungal species hybrids infecting plants. *Microbes and Infection* **4**: 1353–1359.

Pannell JR, Obbard DJ, Buggs RJA. 2004. Polyploidy and the sexual system: what can we learn from *Mercurialis annua*? *Biological Journal of the Linnean Society* **82**: 547–560.

Pound GS, Williams PH. 1963. Biological races of *Albugo candida. Phytopathology* **53**: 1146–1149.

Prince DC, Rallapalli G, Xu D, Schoonbeek H, Çevik V, Asai S, Kemen E, Cruz-Mireles N, Kemen A, Belhaj K *et al.* 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 Biology* **15**: 20.

Rehmany A, Gordon A, Rose L, Allen R, Armstrong M, Whisson S, Kamoun S, Tyler B, Birch P, Beynon J. 2005. Differential recognition of highly divergent downy mildew avirulence gene alleles by RPP1 resistance genes from two *Arabidopsis* lines. *Plant Cell* **17**: 1839–1850.

Ruhe J, Agler M, Placzek A, Kramer K, Finkemeier I, Kemen E. 2016. Obligate biotroph pathogens of the genus *Albugo* are better adapted to active host defense compared to niche competitors. *Frontiers in Plant Science* **7**: 820.

Saharan GS, Verma PR, Meena PD, Kumar A. 2014. *White rust of crucifers: biology, ecology and management.* New Delhi, India: Springer India.

Sánchez-Martín J, Steuernagel B, Ghosh S, Herren G, Hurni S, Adamski N, Vrána J, Kubaláková M, Krattinger SG, Wicker T *et al.* 2016. Rapid gene isolation in barley and wheat by mutant chromosome sequencing. *Genome Biology* **17**: 221.

Schwarz EM. 2017. Evolution: a parthenogenetic nematode shows how animals become sexless. *Current Biology* **27**: R1064–R1066.

Slatkin M, Hudson RR. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**: 555–562.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Stukenbrock EH. 2016. The role of hybridization in the evolution and emergence of new fungal plant pathogens. *Phytopathology* **106**: 104–112.

Stukenbrock EH, Christiansen FB, Hansen TT, Dutheil JY, Schierup MH. 2012. Fusion of two divergent fungal individuals led to the recent emergence of a unique widespread pathogen species. *Proceedings of the National Academy of Sciences, USA* **109**: 10 954–10 959.

Stukenbrock EH, McDonald BA. 2008. The origins of plant pathogens in agro-ecosystems. *Annual Review of Phytopathology* **46**: 75–100.

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

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution* **30**: 2725–2729.

Thilliez GJA, Armstrong MR, Lim TY, Baker K, Jouet A, Ward B, van Oosterhout C, Jones JDG, Huitema E, Birch PRJ *et al.* 2018. Pathogen enrichment sequencing (PenSeq) enables population genomic studies in oomycetes. *New Phytologist*. Accepted. doi: 10.1111/nph.15441.

Verma PR. 2012. White rust of crucifers: an overview of research progress. *Journal of Oilseed Brassica* **3**: 78–87.

Ward BJ, van Oosterhout C. 2016. HybridCheck: software for the rapid detection, visualization and dating of recombinant regions in genome sequence data. *Molecular Ecology Resources* **16**: 534–539.

Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* **15**: R46.

Yarwood CE. 1951. Associations of rust and virus infections. *Science* **114**: 127–128.

Yarwood CE. 1977. *Pseudoperonospora cubensis* in rust-infected bean. *Phytopathology* **67**: 1021–1022.

Yoshida K, Schuenemann VJ, Cano LM, Pais M, Mishra B, Sharma R, Lanz C, Martin FN, Kamoun S, Krause J *et al.* 2013. The rise and fall of the *Phytophthora infestans* lineage that triggered the Irish potato famine. *eLife* **2**: e0073.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article:

Fig. S1 Pie chart representing the proportion of baits targeting each species (or groups of species) in Penseq.

Fig. S2 Maximum-likelihood tree based on the using conserved loci of *Albugo candida* (6943 bp) including 83 isolates.

Fig. S3 Presence/absence polymorphisms of putative CHxC effectors in 83 *Albugo candida* isolates.

Fig. S4 Genetic distance between representatives of each *Albugo candida* host-specific group (p-distance).

Fig. S5 Distribution of the proportion of reads per SNP at heterozygous sites in *Albugo candida* isolates which ploidy level could not be determined.

Fig. S6 Ploidy graphs of *Albugo candida* lab isolates using whole-genome data.

Fig. S7 Genetic variation at neutral genes and effector genes of *Albugo candida* expressed in the mean (± SE) of various population genetic summary statistics.

Fig. S8 Analyses of the frequency spectrum of segregating sites of effector and neutral genes of *Albugo candida* using two population genetic statistics; Fu & Li (1993) and Fu (1997) (mean and SD).

Table S1 Details about species and loci to which oligos were designed.

Table S2 Neutrality test statistics of *Albugo candida* neutrally evolving loci used in this study, based on seven laboratory isolates (Nc2, Em2, Ex1, BoT, BoL, 2v and 7v).

Table S3 Detailed list of the samples sequenced using PenSeq.

Table S4 Breadth of coverage and read depth at loci of control organisms used in the reconstruction experiment.

Table S5 Within- and between-lineage p-distance (in %) at the four sets of loci sequenced in *Albugo candida*.

Table S6 General linear model (GLM) with summary statistics for genetic diversity of *Albugo candida* genes as response variable, and plant species nested within plant type (i.e. wild, garden or crop) and gene type (effector or neutral gene) as fixed factors.

Table S7 General linear model (GLM) with summary statistics for selection of *Albugo candida* genes as response variable, and plant species nested within plant type (i.e. wild, garden or crop) and gene type (effector or neutral gene) as fixed factors.

Table S8 List of overrepresented OTUs in *Albugo candida*-infected plants and identified using Kraken.

Methods S1 Detailed description of the methodology used in this paper and including sample collection, control preparation, DNA extraction, library preparation, DNA capture and enrichment.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.