The adaptive evolution of the plant pathogen *Albugo candida*

Agathe Jouet

Thesis submitted to the University of East Anglia for the degree of Doctor of Philosophy

ENV/TSL

May 2016

TABLE OF CONTENTS

ABSTRACT

Albugo candida is a plant pathogen that has been reported on many host species. While multiple host-specific races have long been recognized in *A. candida*, the genetic variation of these races has never been explored in nature and little is known about how the pathogen has adapted to its many hosts.

Recently, evidence of genetic exchanges between races suggested that hybridization played an important role in the evolution of *A. candida* races. The authors also demonstrated that host-specific races of *A. candida* can co-occur, provided the immune system of the host is compromised by a compatible race. This immunosuppression by *A. candida* had previously been shown to allow growth of other pathogens.

To study both these phenomena (the evolution of and the host immunosuppression imposed by *A. candida*), a capture array was designed to sequence 187 loci (~660,000 bp) from *A. candida* and loci from 47 other plant pathogens. In **Chapter 3**, I explain the rationale and methodology behind this approach. I show that it is cost-effective and that it may be used to identify microorganisms directly from a leaf and make inference about pathogen abundance within samples. In **Chapter 4**, genetic diversity of *A. candida* is analysed at a 400 kb contig and 32 diversity-tracking genes. Races are identified based on genetic divergence and recombination is investigated within and between races. In **Chapter 5**, I investigate genetic diversity at heterozygous sites to study the ploidy level and the reproductive mode of *A. candida* races as well as to detect mixed *A. candida* infections and loss-of-heterozygosity events.

In this thesis, I demonstrate that *A. candida* races adapt to their hosts using complex mechanisms and that some may, in the long term, speciate. I also provide a novel method which may be used to interrogate microbial diversity directly from the field.

ACKNOWLEDGEMENTS

I would first like to thank my supervisor Prof. Jonathan Jones who kindly welcomed me in his lab and who made me a better, more rigorous and passionate scientist.

I also thank my supervisor Prof. Cock van Oosterhout with whom I have liked sharing my enthusiasm every time I thought had a breakthrough in my PhD and who has helped me keep my confidence throughout. I feel lucky to have met him.

I thank my supervisor Dr. Tove Jorgensen for her kindness and her words of reassurance when I needed them. I also thank her for sharing her tips on how to sample *Albugo candida*.

Great thanks also go to Dr. Diane Saunders who suggested some of the analyses presented in this thesis. She gave her time to discuss experiments but also always took care of my well-being overall.

I could not write the acknowledgements without also thanking Dr. Mark McMullan to whom I must have asked hundreds if not thousands of questions and who always provided me with great answers, those that make you think even more. He also turned out to be a great friend.

I also thank Dr. Oliver Furzer and Dr. Christian Schudoma for all the help they provided with the bioinformatics analyses. In addition, I thank Dr. Volkan Cevik, Dr. Kamil Witek and Dr. Florian Jupe for very useful discussions and for always helping me even when I kept coming back with more questions. They, but also all the members in the lab, make Norwich a great place to work and to live in. Special thanks go to Baptiste Castel and Jenna Loiseau for they bring a bit of home to work.

Of course, I want to thank all the people who have looked for *Albugo candida*; those who provided samples (Prof. Marco Thines, Dr. Young-Joon Choi, Chi-Hang Wu, Quentin Dupriez, Prof. Jonathan Jones, Dr. Panagiotis Sarris, Yan Ma) but also those who provided me with snail-slimed, bird-pooed or white painted leaves.

Finally, I want to thank my family and friends who encouraged me in moments of despair. They helped me keep in mind that there is a life outside of research and that it is beautiful. Particular thanks go to Quentin Dupriez for giving his unconditional support during these turbulent times.

This work is funded by the Earth and Life Systems Alliance.

CHAPTER 1: GENERAL INTRODUCTION

Microbial organisms account for most of the biodiversity on earth and researchers estimate that most microbial species still await discovery (Hawksworth & Rossman 1997; Achtman & Wagner 2008). They appear to be mostly harmless to the eukaryotic organisms they interact with but some can cause devastating effects on their hosts or the environment. This is the case for pathogenic microbes that affect important crops or wild plant species (Anderson *et al.* 2004).

Plant pathogens cover organisms from all branches of the tree of life, including nematodes, viruses, bacteria, fungi, oomycetes and rhizaria (Scholthof *et al.* 2011; Dean *et al.* 2012; Hwang *et al.* 2012; Jones *et al.* 2013; Kamoun *et al.* 2015). They are usually characterized by a high evolutionary potential due to their large effective population size, short generation times and the ability for long-distance dispersal which is further reinforced by human activities such as globalization and the industrialization of food production (Fisher *et al.* 2012). Also, their ability to exchange genetic information across species boundaries (i.e. genetic introgression and horizontal gene transfer) expedites their rate of evolution compared to many eukaryotes (Stukenbrock & McDonald 2008; McMullan *et al.* 2015). In the past few decades, scientists have documented several cases of emerging infectious diseases of plants (EIDs, Lederberg *et al.* (1992); Daszak *et al.* (2000)). Of these, the oomycete *Phytophthora infestans*, a pathogen responsible for the Irish potato famine in the 19th century, might well be the most famous and well-studied example, and it has become infamous by its recurrent evolution of novel aggressive lineages (Cooke *et al.* 2012; Peters *et al.* 2014). Research on many other destructive plant pathogens is progressing rapidly, largely due to the availability of whole genome sequence data. Such data has become readily available for a plethora of plant pathogens, including fungi such as *Magnaporthe oryzae* (rice blast; Dean *et al.* (2005)) and *Botrytis cinerea* (grey mould; Amselem *et al.* (2011)), bacteria (*Pseudomonas syringae* (Baltrus *et al.* 2011), *Ralstonia solanacearum* (Salanoubat *et al.* 2002)) and viruses (e.g. Tobacco mosaic virus (Goelet *et al.* 1982) and Tomato spotted wilt virus (Tsompana *et al.* 2005). Yet, the mechanisms by which these and other plant pathogens are evolving resistance breaking are still largely unknown although precisely this knowledge is imperative for sustainable disease management.

In this thesis, I aim to investigate the processes by which oomycete *Albugo candida* is able to infect many diverged Brassicaceae species, including some of our most important crops.

Additionally, a new method is developed to cost-effectively interrogate microbial diversity in the field. By doing so, I aim to highlight important evolutionary and ecological processes in the adaptation of plant pathogens, and to propose a method that could facilitate future work on the evolution of microbes and microbial communities in the wild.

To bring background for this research, I will provide detailed information on (i) the mechanisms by which plant pathogens keep up with or adapt to new hosts and environments and (ii) on the current understanding of *A. candida* evolution.

1.1 EVOLUTION OF PLANT PATHOGENS

1.1.1 GENE-FOR-GENE INTERACTIONS

Plants have developed several ways to resist pathogens. The most basal, called pattern-triggered immunity (PTI), includes recognition of pathogen-associated molecular patterns (PAMPs, e.g. flagellin from bacteria and chitin from fungi) by plant cell receptors known as pattern-recognition receptors or PRRs (Jones & Dangl 2006). Upon PAMP perception, a cascade of signalling events is triggered that leads to defence mechanisms such as callose deposition (Luna *et al.* 2011) or the accumulation of reactive-oxygen species (ROS, Bailey-Serres & Mittler (2006)). The second, called effector-triggered immunity (ETI), involves direct or indirect recognition of pathogen effector proteins by disease resistance (R) proteins which in turn usually lead to a hyper-sensitive response (HR), or cell death, of the infected cells. Effector proteins can be secreted into the plant cells via the type-III secretion system of bacteria or the hyphae or haustoria in fungi and oomycetes (Jiang & Tyler 2012). Their functions are diverse but include interference with PTI, plant development, plant metabolism or stomatal closure (Hogenhout *et al.* 2009). There are many resistance-effector proteins interactions that have been well defined but which will not be detailed in this thesis (but see Chisholm *et al.* (2006); Jones & Dangl (2006) or Bent & Mackey (2007) for reviews).

To be successful, plant pathogens must evade recognition by evolving novel but functional PAMP or effector alleles. If advantageous, these may become fixed in the pathogen population until recognized by an R-gene allele, as in the arms race model (Stahl & Bishop 2000). In this model, a gain of recognition mutation in an R-gene allele will cause the corresponding effector allele to be selected against and to be replaced by a novel effector that can evade recognition. Novel beneficial alleles may also increase in frequency while

polymorphisms are retained by balancing selection as potential reservoir of genetic diversity, as depicted in the trench warfare model (Stahl *et al.* 1999; Kamoun 2001). In this model, effector alleles might coexist as long as they are not recognized in some host populations. In the literature, examples can be found supporting one or the other model. For instance, low allelic and haplotype diversity at the *AvrLm1* locus in *Leptosphaeria maculans* and strong linkage disequilibrium between the *AvrLm1* locus and the effector gene *AvrLm2* suggest that selective sweeps may have occurred in this region due to strong selection on advantageous *AvrLm2* alleles (Gout *et al.* 2007), supporting the arms race model. Conversely, high levels of polymorphisms in *AvrP4* and *AvrP123* of fungus *Melampsora lini* and variation in allele frequencies over time suggest strong diversifying selection and maintenance of rare alleles by balancing selection (Barrett *et al.* 2009; Thrall *et al.* 2012). Similarly, Allen *et al.* (2004) demonstrated that the *ATR13* effector gene from *Hyaloperonospora parasitica* and the R-gene *RPP13* from *Arabidopsis thaliana* were both highly polymorphic and under diversifying selection (with high levels of non-synonymous over synonymous mutations ($d_N/d_S>1$)), suggesting that the fate of these two genes are tightly linked and supporting the trench warfare model. However, while these models focus on gene interactions between pathogens and their hosts, other processes which I will describe below may have profound impacts on the patterns of variation observed in pathogen populations.

1.1.2 HOST DEMOGRAPHY AND PATHOGEN EVOLUTION

Besides selection, genetic diversity in pathogen population may be affected by host demography. Successful invasion of pathogens depends heavily on the availability of susceptible hosts (Cunniffe & Gilligan 2010). In fragmented or seasonal wild host populations, pathogens may experience high rates of extinction events as well as (re-) colonization from neighbouring pathogen populations through dispersal (Pannell & Charlesworth 2000). Depending on the number of migrants involved in (re-)colonization, genetic diversity may be lost locally due to reductions in effective population size (Asch & Collie 2008). Additionally, although migrants may by chance be sufficiently-adapted for colonization, they are less likely to be successful in cases of highly connected host populations with high levels of genetic diversity and disease resistance. This was shown for example in the *Plantago lanceolata* - *Podosphaera plantaginis* system for which spatiotemporal dynamics were recorded for 12 years in the Åland archipelago, southwest of Finland (Jousimo *et al.* 2013). However, *Albugo candida* produces resting oospores that may be able to survive intercrop seasons so that

populations may not necessarily go extinct (Lakra & Saharan 1989b; Saharan *et al.* 2014). Conversely, higher levels of gene flow between pathogen compared to host populations may lead to pathogens being better adapted to their host populations (Penczykowski *et al.* 2016). Similarly, pathogen colonization and expansion may be facilitated in agricultural ecosystems by the high density and uniformity of host genotypes (Stukenbrock & McDonald 2008). This has probably led to an increase in the number of reports of pathogen emergence in crops such as wheat and citrus (e.g. Ug99 races of stem rust fungus *Puccinia graminis f. sp. tritici* (Singh *et al.* 2011) and citrus canker bacterium *Xanthomonas axonopodis pv. citri* (Graham *et al.* 2004), respectively; see Anderson *et al.* (2004) for review).

1.1.3 THE EVOLUTIONARY POTENTIAL OF SEXUAL AND ASEXUAL PATHOGENS

Many pathogens exploit their host during a phase of intra-population expansion accomplished by rapid clonal (or asexual) reproduction. This may be required when suitable sexual partners are rare, particularly when only a limited number of pathogen genotypes end up colonising a particular host. Asexual reproduction is also faster than sexual reproduction and if a pathogen genotype is well adapted to its host, sexual reproduction may not be essential in the short term. However, although there are some advantages to being strictly asexual, for example avoiding the costs of sex (energetic cost, demographic constraints of finding a suitable partner, separation of favourable combinations of alleles…)), truly asexual organisms are often considered an evolutionary dead-end (Comai 2005; de Jonge *et al.* 2013; Seidl & Thomma 2014). This is mainly due to the lack of meiotic recombination and subsequent decreased ability to quickly adapt to changing environments (in diploids) as well as the accumulation of deleterious mutations which may reduce pathogen fitness, a phenomenon known as the Muller's ratchet (Felsenstein 1974).

There are some textbook examples of organisms reproducing strictly clonally (bdelloid rotifers (Welch & Meselson 2000), arbuscular mycorrhizal fungi (Kuhn *et al.* 2001), certain plants or nematodes (Judson & Normark 1996)). However, increasingly, evidence suggests that organisms that were initially thought to be strictly asexual may in fact also reproduce sexually on rare occasions (Schurko *et al.* 2009; Halary *et al.* 2011; Rabeling *et al.* 2011; Schwander *et al.* 2011). In pathogens, reproduction often involves a mixed system with the alternation of both asexual and more or less regular sexual cycles (for example, *Alternaria brassicicola* (Bock *et al.* 2005), *Aphanomyces euteiches* (Grünwald & Hoheisel 2006), *Phytophthora infestans* (Danies *et al.* 2014) or *Albugo candida* (Saharan & Verma 1992)). This allows

pathogens to have the best of both worlds (as described by Ellison *et al.* (2011)); novel combinations of alleles may be generated during sexual reproduction for quick adaptation, and once well-adapted genotypes have been formed, they can propagate rapidly via clonal reproduction. An example of this would be *P. infestans* lineage US-11 which was shown to have originated from sexual reproduction between lineages US-6 and US-7 during the 1993 Columbia Basin epidemic and to have clonally expanded in certain agroecosystems, particularly with tomatoes (Gavino *et al.* 2000).

However, investigating the importance of sexual and asexual reproduction of pathogens in the wild is often a complicated matter, especially in non-model organisms where cryptic sex may occur (Schurko *et al.* 2009). While one could use organismal signs of sex to infer whether a species can reproduce sexually (e.g. presence of males and females, mating behaviour), this often does not provide sufficient evidence for sex and may be misleading. For example, although the rotifer *Brachionus calyciflorus* produces stimuli which are known to induce sexual reproduction in other rotifer species, it was found to be an obligate parthenogen and lack responsiveness to those stimuli (Stelzer 2008).

As a result, to infer sexual reproduction, it is desirable to combine these organismal signs of sex with other methods, for example using molecular data. In sexual organisms, where meiotic recombination occurs during the formation of gametes (diploids) or during the formation of spores (haploids), genetic diversity is reassorted from one generation to the next. This results in the free exchange of alleles among individuals and consequently high genotypic diversity at the population level (Heitman 2010). This also results in phylogenetic incongruence between loci (Schurko *et al.* 2009; Jouet *et al.* 2015). In contrast, asexual organisms reproduce through mitotic duplication of their genetic material and therefore, strong linkage among loci is expected resulting in low genotypic diversity at the population level as well as congruent phylogenetic inferences (Balloux *et al.* 2003). Additionally, within locus heterozygosity is predicted to be higher in ancient asexual organisms due to the independent evolution of alleles, a process known as the Meselson effect (Welch & Meselson 2000). In sexual organisms however, within-locus heterozygosity (in diploids) and nucleotide diversity is expected to be much reduced due to the homogenizing effect of meiotic recombination, which shortens the coalescence time between allelic copies. Furthermore, in selfing organisms (where syngamy occurs between gametes produced by one individual), this is reinforced due to the non-random segregation of gametes (diploids) or of nuclei (haploids).

Evaluating molecular evidence for sexual and asexual reproduction in pathogens is crucial for the development of suitable management strategies. In addition, increased and

adequate sampling of pathogens combined with careful molecular investigation of their natural genetic diversity will probably shed light on the importance of other means of evolution such as hybridization, polyploidization, host jumps or extensive genomic rearrangements. These mechanisms are discussed in the next section.

1.1.4 RAPID ADAPTATION THROUGH HYBRIDIZATION, HOST JUMPS, POLYPLOIDIZATION AND GENOMIC REARRANGEMENTS

While sexual reproduction can quickly generate genetic diversity to facilitate rapid adaptation to changing environments, other mechanisms are being put forward as potentially equally important evolutionary forces. These mechanisms may not only generate diversity within species but also, in contrast with sexual reproduction, be the drivers of speciation.

Among these, hybridization has been suggested as being an important evolutionary process in plants (Soltis & Soltis 2009), oomycetes (Brasier *et al.* 1999) and fungi (Oberhofer & Leuchtmann 2012; Menardo *et al.* 2015). In *Phytophthora* species for instance, hybridization is thought to have led to the formation of species with novel host specificities. This is the case of *Phytophthora alni*, a hybrid species pathogenic to *Alnus* spp. which was generated via sexual reproduction between *Phytophthora cambivora* and a species related to *Phytophthora fragariae*, both non-pathogenic to *Alnus* spp. (Brasier *et al.* 2004). Similarly, *B. graminis f. sp. triticale*, a powdery mildew which grows on triticale and wheat, was found to be a hybrid between *B. graminis f. sp. tritici* pathogenic to wheat and *B. graminis f. sp. secalis* to rye (Menardo *et al.* 2015).

Another example of processes that may help generate genetic diversity and therefore rapid evolution is polyploidization. (Soltis *et al.* 2004; Madlung 2012; Wendel *et al.* 2016). Polyploidization may derive from whole genome duplication (autopolyploids) or hybridization (allopolyploids; Otto & Whitton (2000)) and like hybridization, it has been shown to be important in plants. However, polyploidization has been poorly documented in other organisms; yet, many fungi appear to be polyploid (*Phyllactinia*, *Stephensia*, *Xylaria*, *Botrytis* and *Zygosaccharomyces* genera, see Albertin & Marullo (2012) for a review) and certain races of the oomycete *Phytophthora infestans* were shown to have recently undergone polyploidization (Yoshida *et al.* 2013). Although polyploidization is sometimes thought to be detrimental due to abnormalities that may arise during the pairing of chromosomes in mitosis and meiosis and to a decreased likelihood of finding suitable mating partners (in the case of sexual organisms), it may also play an important role in evolution (Madlung 2012). Indeed, the

increased number of alleles at each locus may allow for reduced selection pressure and therefore increased responsiveness to environmental changes (Hegarty & Hiscock 2007). Polyploidy may also reduce the effect of Muller's ratchet in the case of recessive deleterious mutations (Soltis & Soltis 2000) as well as stabilize hybrid vigour (Chen 2010).

Possibly linked to the processes described above, many pathogens have been shown to have experienced host shifts (when a pathogen evolves to colonize a closely-related host) or host jumps (when evolution is to a distant host). However, host jumps and host shifts may also occur when a pathogen is introduced to a new environment where potential so-called naïve hosts do not possess the molecular machinery for resistance (Stukenbrock & McDonald 2008). For example, Couch *et al.* (2005) suggested that rice-infecting *Magnaporthe oryzae* lineages arose from a host shift from a *Setaria* millet-infecting species. Similarly, *Phytophthora* species are thought to have evolved through host jumps to adapt to many diverged species including *Solanum* spp. (*P. infestans*), *Ipomoea longipedunculata* (*Phytophthora ipomoeae*) and *Mirabilis jalapa* (*Phytophthora mirabilis*, Raffaele *et al.* (2010)).

Lastly, extensive genomic rearrangements may also enable quick adaptation and potentially speciation (Seidl & Thomma 2014). These may arise from the processes described above and may occur in both sexual (through for example allopolyploidization or hybridization) and asexual organisms (autopolyploidization via somatic doubling or mitotic recombination). For example, the *AVR-Pita* effector gene which is located in the telomeric region of chromosome 3 in *Magnaporthe grisea* was found to be frequently lost in spontaneous mutants. If this also occurs in nature, isolates where *AVR-Pita* is absent would avoid recognition by the corresponding *Pi-ta* resistance gene in rice (Orbach *et al.* 2000). In the asexual pathogen *Verticillium dahliae*, extensive chromosomal rearrangements among strains were revealed that are associated with lineage-specific regions enriched with effector genes (de Jonge *et al.* 2013).

In this thesis, I am interested in the mechanisms involved in the evolution of the plant pathogen *Albugo candida*. In the next sections, I will provide details on *A. candida* biology and describe publications that report on its evolution.

1.2 BIOLOGY AND EVOLUTION OF *ALBUGO CANDIDA*

1.2.1 WHITE BLISTER RUST DISEASE

Albugo candida is a eukaryotic plant pathogen that belongs to the kingdom Stramenopila, phylum Oomycota (Beakes *et al.* 2012) and as such, it is genetically closely-related to important plant pathogens such as *Phytophthora infestans* (potato late blight), *Phytophthora ramorum* (sudden oak death, sudden larch death and ramorum blight; various hosts) or *Plasmopara viticola* (grape downy mildew). Although perhaps not as economically important as the above oomycetes, *A. candida*, causal agent of the white blister rust, is considered the most important disease of vegetable and oil-yielding *Brassica* crops worldwide (Saharan *et al.* 2014). These destructive impacts on agricultural crops, for example oilseed rape *Brassica napus*, colza *Brassica juncea* and mustard greens *Brassica rapa*, have partly fuelled research on *A. candida*, especially in Canada, Australia and India (Bernier 1972; Petrie & Vanterpool 1974; Barbetti 1981; Rimmer *et al.* 2000; Kaur *et al.* 2008; Kolte *et al.* 1981; Lakra & Saharan 1989; Sandhu *et al.* 2015). However, other characteristics make it an interesting organism to study the evolution of plant pathogenic oomycetes as explained below.

1.2.2 A BIOTROPH WITH A BROAD HOST RANGE

A. candida has a biotrophic lifestyle and therefore needs to establish an intimate relationship with its host on which it depends for development, nutrition and reproduction (O'Connell & Panstruga 2006). To do this, cœnocytic hyphae of *A. candida* develop around palisade mesophyll cells and intracellular haustoria form that are necessary for nutrient uptake and effector secretion. Interestingly, although obligate biotrophy requires the formation of a compatible interaction between the oomycete and its host, *A. candida* has been reported to be able to infect more than 200 species of Brassicaceae as well as species from the Cleomaceae and Capparaceae families (Biga 1955; Saharan & Verma 1992; Choi *et al.* 2009; Meena *et al.* 2014). These include crops (e.g. *Brassica juncea*, *Brassica oleracea*, *Raphanus sativus*, *Capparis spinosa*), ornementals (*Aubrieta deltoidea*, *Alyssum saxatile*, *Lunaria annua*, *Cleome hassleriana*) as well as wild species (*A. thaliana*, *Capsella bursa-pastoris*, *Sisymbrium officinale*, *Cleome anomala*). Although not a unique feature in plant pathogenic oomycetes (see e.g. *Pseudoperonospora cubensis* (Runge *et al.* 2012) or *Phytophthora capsici* (Lamour *et al.* 2012b)), this appears to be in contrast with other obligate biotrophs which are specialized on one or a restricted number of hosts such as *Hyaloperonospora arabidopsidis* (Baxter *et al.* 2010) and *Albugo laibachii* (Thines *et al.* 2009) on *Arabidopsis thaliana* or *Plasmopara viticola* on *Vitis* spp. (Schröder *et al.* 2011)).

1.2.3 NEXT-GENERATION ERA: INSIGHTS INTO *A. CANDIDA* RACES EVOLUTION

As with many groups of plant pathogens, defining the different taxonomic units in *A. candida* has proved difficult and in the 20th century, many classifications were proposed that rationalized the broad host range of *A. candida* (Eberhardt 1904a; b; c; d; Hiura 1930; Napper 1933; Togashi & Shibasaki 1934; Pound & Williams 1963). These were based on sporangial size and host specificity and distinguished two biological forms of *A. candida*, microspora (12.5-15 µm sporangia) and macrospora (15-17.5 µm sporangia), within which several host-specialized races called pathotypes were identified via cross-inoculation experiments (Pound & Williams 1963). Lacking additional discriminative criteria however, *A. candida* species integrity was preserved. In particular, although the life cycle of *A. candida* has been described in details (asexual reproduction via vegetative zoosporangia and sexual reproduction via the formation of oospores), little is known about the modes of reproduction in place in natural populations. Therefore, it is not possible to use this information to delineate species from diverged races within a species.

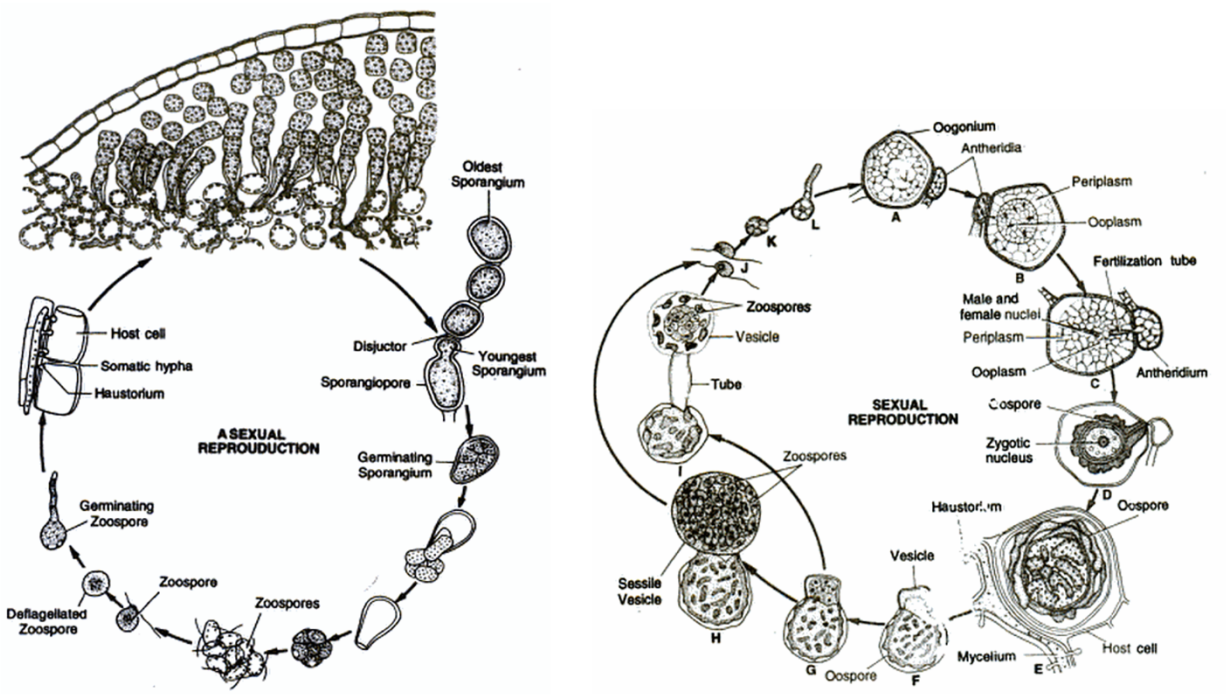


Figure 1.1 Life cycle of *Albugo candida* with (left) asexual reproduction and right (sexual reproduction). During asexual reproduction, intercellular sporogeneous hyphae develop (sporangiophores) and produce multinucleate sporangia in a basipetal succession. Primary sporangia attach to the host epidermal layer and undergo cytological changes, possibly delivering cell wall degrading enzymes and releasing secondary sporangia (Heller *et al.* 2009).

Upon suitable environmental conditions, sporangia discharge 4-12 biflagellate zoospores that will germinate and enter the host via stomata. During sexual reproduction, oogonia and antheridia develop (female and male reproductive organ, respectively). Through gametangial contact, a fertilization tube is formed that carries the male nucleus into the central uninucleate part of oogonia (ooplasm). The oospore is formed with a thick outer layer which is thought to be critical for intercrop season survival. As during asexual reproduction, the oospore will discharge biflagellate zoospores that will germinate and enter the host in the form of hyphae (figures from Saxena 2010).

Advances in DNA sequencing technologies finally made it possible to combine physiology and morphology with genetics. Using mitochondrial sequences ITS (Internal transcribed spacer) and *cox2* (Cytochrome c oxidase subunit 2) from numerous isolates, several apparently host-specific *Albugo* species could be described (*Albugo laibachii* (Thines *et al.* 2009), *A. koreana* (Choi *et al.* 2007), *A. hohenheimia* (Ploch *et al.* 2010)). However, genetic diversity at these loci was not sufficient to discriminate between most *A. candida* pathotypes (Choi *et al.* 2006, 2007; Kaur *et al.* 2008; Petkowski *et al.* 2010; Ploch *et al.* 2010). It was only in 2015, four years after the whole genome of *A. candida* was published (Links *et al.* 2011) that McMullan *et al.* (2015) provided first insights into *A. candida* races evolution. By comparing the whole genome sequences of five isolates representing three pathotypes (*Ac*Nc2/*Ac*Em2, *Ac*BoT/*Ac*BoL and *Ac*2v), the authors could show that isolates within pathotypes are probably mostly clonal. Most interestingly, they revealed that not only *A. candida* pathotypes are genetically diverged (~1%) but also that there is evidence for historical recombination between pathotypes. This is exciting because this highlights recombination as a driving force for broad host adaptation in *A. candida*. Indeed, recombination between *A. candida* pathotypes could generate novel sets of effector alleles which, in rare cases, could enable adaptation to a new host. Once this happens, the recombinant isolate would be expected to propagate rapidly through asexual reproduction (McMullan *et al.* 2015).

Finally, this is surprising because although the host ranges of *A. candida* pathotypes were long shown to be somehow overlapping (Pound & Williams 1963), it was not known that pathotypes could meet in the wild to reproduce. Using sequential inoculations, McMullan *et al.* (2015) could show that infection by a virulent race could enable subsequent colonization by another otherwise avirulent race, increasing the chance of co-occurrence and potentially enabling reproduction between pathotypes. Although not in such a context, suppression of the host immune system had already been demonstrated for *A. candida* by Cooper *et al.* (2008) and

for other plant pathogens by Yarwood (1951, *Uromyces phaseoli*, *Puccinia helianthi* and *Puccinia antirrhini*), Gill (1965, *Uromyces phaseoli* and *Puccinia helianthi*) and more recently by Olesen *et al.* (2003, *Blumeria graminis*) and by Belhaj *et al.* (2015, *Albugo laibachii*).

1.2.4 FUTURE RESEARCH ON *ALBUGO CANDIDA*

While the study by McMullan *et al.* (2015) provided first insights into how *A. candida* adapted to many diverged hosts, it also opened up new questions that need to be answered. For example, while isolates *Ac*Nc2 and *Ac*Em2 are genetically similar and considered as belonging to the same pathotype, they were originally found on different host species (*A. thaliana* and *C. bursa-pastoris*, respectively). This challenges the notion of host-specific races in *A. candida* and it would be interesting to know whether other races are capable of infecting multiple hosts in the wild. The authors also found that heterozygosity was variable between races. Indeed, *Ac*Em2, *Ac*Nc2 and *Ac*2v have low levels of heterozygosity (<0.047% heterozygous sites in a ~400 kb contig called 'contig 1') whereas *Ac*BoT and *Ac*BoL are highly heterozygous (~0.65%). This suggests that a mechanism exists to either preserve heterozygosity in the latter isolates or reduce heterozygosity in the former. Moreover, while the study by McMullan *et al.* (2015) suggests that isolates within a race are mostly clonal, *A. candida* has long been hypothesized to survive intercrop periods in the form of sexual oospores (Lakra & Saharan 1989b; Saharan *et al.* 2014). It is possible that the authors failed to detect signature of sexual reproduction within races and collecting a wide range of *A. candida* isolates would help investigate this apparent contradiction. Finally, while the authors found evidence for sexual reproduction between races and proposed host immunosuppression as a mechanism for *A. candida* races co-occurrence, it has yet to be shown that *A. candida* races can co-occur in the wild. It would also be interesting to investigate the incidence of other microorganisms in relation to *A. candida* infection.

1.3 AIMS OF THIS THESIS

In this thesis, I used a sequence capture method to investigate the genetic diversity of *A. candida* in the wild as well as to detect other microorganisms that might be in association with *A. candida*. I first explain the rationale behind this method called PathSeq and test its efficiency to detect microorganisms from whole plant leaves (Chapter 3). Later, I focus on *A.*

candida natural genetic diversity to gather information on the number of genetically diverged races that can be found in the wild, the nucleotide diversity as well as the incidence of recombination within and between races (Chapter 4). I then use per-individual heterozygous sites in *A. candida* sequenced loci to explore other mechanisms that might be at play in *A. candida* races evolution. In particular, I use these sites to investigate the ploidy level of *A. candida* races, the incidence of sexual reproduction within races and of mixed *A. candida* infection in the wild. I also investigate potential loss-of-heterozygosity events (Chapter 5).

CHAPTER 2: MATERIAL AND METHODS

2.1 SAMPLE PREPARATION FOR SEQUENCE CAPTURE

2.1.1 COLLECTION OF FIELD *ALBUGO CANDIDA* ISOLATES

White rust samples were collected in the wild from 2013 through to 2015 from several Brassicaceae species: *Capsella bursa-pastoris*, *Aubrieta deltoidea*, *Alyssum saxatile*, *Brassica nigra* (by Quentin Dupriez), *Sinapis alba* (by Chih-Hang Wu, The Sainsbury Laboratory, Norwich, England), *Sisymbrium officinale*, *Brassica oleracea* (by Prof. Cock van Oosterhout, University of East Anglia, Norwich, England), *Raphanus sativus*, *Lunaria annua*, *Lunaria rediviva*, *Arabis alpina*, *Alyssum montanum*. In all cases, symptomatic tissues (leaves, stalks, flowers and/or fruits) were placed in Falcon® tubes and stored at -80°C.

Additional samples were provided from *Brassica juncea* by Prof. Deepak Pental (University of Delhi, India) and Prof. Abha Agnihotri (Amity University, Noida, India), *Arabidopsis lyrata*, *Arabidopsis halleri*, *B. oleracea* by Dr. Sebastian Fairhead and Prof. Eric Holub, (Warwick University, England), *Arabidopsis thaliana* by Dr. Volkan Cevik (The Sainsbury Laboratory, Norwich, England), *Brassica carinata*, *Camelina sativa* by Dr. Hossein Borhan and Dr. Colin Kindrachuk (Agriculture and Agri-Food Canada, Saskatoon, Canada), *Eutrema japonica* by Yan Ma (The Sainsbury Laboratory, Norwich, England), *Raphanus sativus*, *Eruca sativa* and *Brassica oleracea* by Ulrike Miersch and Dr. Annemarie Lokerse (Rijk Zwaan, De Lier, The Netherlands). These samples were either collected very recently or propagated in the laboratory for several years. They were sent as fresh material or as DNA extracted from whole infected leaves and stored at -80°C. For a detailed list of A. candida isolates used in this study, see Table S4.1.

2.1.2 LIST OF *ALBUGO CANDIDA* LABORATORY ISOLATES

The following *Albugo candida* isolates were obtained from the laboratory: *Ac*Nc2 (originally collected *on Arabidopsis thaliana* in Norwich, UK (2007)), *Ac*Em2 (from *Capsella bursa-pastoris*, Kent, UK (1993)), *Ac*2v (from *Brassica juncea,* see Links *et al.* (2011)), *Ac*BoT and *Ac*BoL (from *Brassica oleracea*, Lincolnshire, UK (2009)), published in McMullan *et al.*

(2015) as well as *Ac*Ex1 (collected from *Arabidopsis halleri* in Exeter by Eric Holub, unpublished), *Ac*7v (from *Brassica rapa*).

2.1.3 PREPARATION OF CONTROLS

Controls in this study aim at: (i) evaluating whether sequence capture can be used to identify microorganisms directly from the field, (ii) investigating a potential quantitative correlation between the abundance of and the number of reads generated per organisms and (iii) detecting potential biases that may arise from sequence capture compared to whole-genome sequencing.

Briefly, controls were prepared as follows: *Albugo candida* and *Albugo laibachii* isolates that are routinely used in the laboratory were obtained by collecting whole symptomatic leaves which were placed at -80°C until DNA extraction. *Ac*2v was collected from *Brassica juncea*, *Ac*7v from *Brassica rapa*, *Ac*BoT from *Brassica oleracea*, *Ac*Ex1, *Ac*Nc2 and *Al*Nc14 from *Arabidopsis thaliana*.

Asymptomatic wild leaves were collected at the same time and location as wild *Albugo candida*-infected leaves (samples #62-68 and #108-110, see Table S3.1). The absence of *Albugo candida* in asymptomatic leaves was later confirmed by PCR (see section 2.1.5). Both infected and healthy samples were processed using the sequence capture method (PathSeq) and were included so as to compare the microbiome of healthy leaves versus *Albugo candida*-infected leaves.

*Ac*2v zoospores were collected by gently tapping heavily infected leaves. After DNA extraction, these were mixed with varying amounts of DNA from *Pseudomonas syringae* DC3000 and sterilized leaves from *Aubrieta deltoidea* (see Chapter 3, Figure 3.4). *Pseudomonas syringae* DC3000 (Buell *et al.* 2003) was grown on King's B medium with Rifampicin and Kanamycin and resuspended in sterile water; sterilization was achieved with two five-minute bleach baths followed by a ten-minute sterile water bath. Zoospores from *Ac*7v and *Pustula tragopogonis* (a close relative of *Albugo candida*, from the Albuginaceae family) were collected the same way as those from *Ac*2v.

*Ac*Ex1-infected leaves were drop inoculated with 100 µl of 30,000 spores per ml *Phytophthora infestans* 88069 td (Chaparro-Garcia *et al.* 2011). DNA was either extracted and processed directly or mixed with previously prepared DNA from *P. syringae*. Similarly, *Albugo laibachii* Nc14 infected leaves (Kemen *et al.* 2011) were spray-inoculated with

Hyaloperonospora arabidopsidis Emoy2 (Holub 2006) and DNA was extracted 5 days post-inoculation (dpi).

Finally, *Hyaloperonospora brassicae* and *Phytophthora infestans* samples were provided by Dr. Laura Baxter (University of Warwick, UK) and Jeroen Stellingwerf (Oak Park Crop Research Centre, Carlow, Ireland) who will analyse the data as part of their respective research.

2.1.4 DNA EXTRACTION

DNA from whole infected tissues was extracted using phenol/chloroform. Tissues were first placed in an Eppendorf tube before they were transferred in liquid nitrogen. A TissueLyser was then used with 3 mm tungsten carbide beads from Qiagen to grind the samples. Ground material was resuspended in 500 µl of Shorty buffer (20% 1M Tris HCl pH 9, 20% 2M LiCl, 5% 0.5M EDTA, 10% SDS 10%, 45% dH2O) and one volume of phenol:chloroform:isoamyl alcohol (25:24:1) was added. The tube was vortexed briefly and spun at 13,200 rpm for five minutes. The upper aqueous phase containing DNA was pipetted into a new Eppendorf and one volume of 100% ice cold isopropanol was added before it was spun again at 13,200 rpm for 10 minutes to precipitate DNA. The pellet was washed twice using 70% ethanol, heated at 70°C for 2-5 minutes to completely remove the ethanol and resuspended in sterile water. Resuspended DNA was then heated at 65°C for 20 minutes to inactivate DNases before an RNase treatment was performed using 2 µl of 10 $mg.ml^{-1}$ RNAse A at 37°C for an hour.

2.1.5 CONFIRMATION OF THE PRESENCE OF PATHOGENS IN CONTROLS

The presence of *Albugo candida* in wild samples and of other pathogens used in control samples was confirmed by amplifying species-specific loci. In *Albugo* spp., a gene coding for a putative cAMP-binding protein, *Ev1786*, was amplified using the *Taq* DNA polymerase from New England Biolabs (NEB). A 1,129 bp product was obtained using primers provided by Dr. Volkan Cevik (forward: 5'-GCCGTCGACGTGATATCTTTGC-3', reverse: 5'-GCGATC ACATCGGCTTGTCGTGG-3') with an annealing temperature of 58°C, one minute extension time and 35 cycles in a thermal cycler. The restriction enzyme HindIII-HF® from NEB was then used to discriminate between *Albugo candida* and other *Albugo* species. This amplicon carries two restriction sites in *A. candida* so that the PCR product is digested into three DNA fragments of 654, 253 and 222 bp while there are none in *A. laibachii* (Thines *et al.* 2009), *A.*

lepidii (Choi *et al.* 2007) and *A. hohenheimia* (Ploch *et al.* 2010). Other *Albugo* species could not be tested. To confirm the presence of *Hyaloperonospora arabidopsidis*, primers provided by Dr. Lennart Wirthmueller (forward: 5'-ATTGCGCCTTTTGCTCT AAACTG-3', reverse: 5'-ACTGAAGCAGTGCAAGGG C-3') were used with an annealing temperature of 56°C to amplify the putative effector gene *RxLL445*. Similarly, primers provided by Dr. Florian Jupe (forward: 5'-GGCTTAAUAAGATTCAGACAAGCTTAAT-3', reverse: 5'-GGTTTAAUT TATCCGGAGGGGTTTAGC-3') and those published in Ferrante & Scortichini (2010, forward: 5'-AAGGCGARATCGAAATCGCCAAGCG-3', reverse: 5'-GGAACWKGCGCA GGAGTCGGCACG-3') were used to amplify the effector gene *AvrSmira1* from *Phytophthora infestans* (Rietman *et al.* 2012) and the RNA polymerase sigma factor (*rpoD*) gene from *Pseudomonas syringae*, respectively.

2.1.6 LIBRARY PREPARATION

Genomic DNA extracted from whole symptomatic leaves was sheared into ~500 bp fragments using the S220 Ultrasonicator from Covaris®. To ensure that all fragments were blunt-ended with 5' phosphate and 3' hydroxyl groups, end repair was performed using the NEBNext® Ultra™ DNA Library Prep Kit for Illumina®. 3' dA-tails were then added to each fragment and Illumina adapters were ligated: /5Phos/GATCGGAAGAGCACACGTC TGAACTCCAGTC/ideoxyU/ACACTCTTTCCCTACACGACGCTCTTCCGATC*T, from Integrated DNA Technologies (IDT)). Finally, adapter-ligated DNA was amplified with a Universal PCR primer for Illumina (required for the enrichment of DNA fragments, with flow cell bound oligo P5: AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACA CGACGCTCTTCCGATC*T, ordered from IDT) and index primers (required for sample multiplexing, with flow cell bound oligo P7). These latter primers were designed with 8 bp indices (Table 2.1) using Python scripts provided on https://bioinf.eva.mpg.de /multiplex/ (Meyer & Kircher 2010) and ordered from IDT. DNA was purified after both adaptor ligation and PCR amplification using a 0.8:1 Agencourt AMPure XP beads (Beckman Coulter®) to DNA ratio. These purification steps not only remove contaminants which may be present in the samples but also small DNA fragments that would cluster preferentially during the sequencing process (<300 bp).

ID	Index sequence	Barcode (8nt)
1	CAAGCAGAAGACGGCATACGAGATcgttggttGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	AACCAACG
2	CAAGCAGAAGACGGCATACGAGATttctggttGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	AACCAGAA
3	CAAGCAGAAGACGGCATACGAGATtggcggttGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCT	AACCGCCA
4	CAAGCAGAAGACGGCATACGAGATtagtcgttGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	AACGACTA
5	CAAGCAGAAGACGGCATACGAGATtggttcttGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	AAGAACCA
6	CAAGCAGAAGACGGCATACGAGATagagttctGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	AGAACTCT
7	CAAGCAGAAGACGGCATACGAGATtccattggGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	CCAATGGA
8	CAAGCAGAAGACGGCATACGAGATccagctggGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCT	CCAGCTGG
9	CAAGCAGAAGACGGCATACGAGATgcagacggGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCT	CCGTCTGC
10	CAAGCAGAAGACGGCATACGAGATaccggaggGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCT	CCTCCGGT
11	CAAGCAGAAGACGGCATACGAGATccgtcaggGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCT	CCTGACGG
12	CAAGCAGAAGACGGCATACGAGATatgaatagGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCT	CTATTCAT
13	CAAGCAGAAGACGGCATACGAGATtataactcGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	GAGTTATA
14	CAAGCAGAAGACGGCATACGAGATtacgtagcGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCT	GCTACGTA
15	CAAGCAGAAGACGGCATACGAGATcatactccGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	GGAGTATG
16	CAAGCAGAAGACGGCATACGAGATataccgccGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCT	GGCGGTAT
17	CAAGCAGAAGACGGCATACGAGATcaactaccGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCT	GGTAGTTG
18	CAAGCAGAAGACGGCATACGAGATgctgaaccGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCT	GGTTCAGC
19	CAAGCAGAAGACGGCATACGAGATatataagaGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCT	TCTTATAT
20	CAAGCAGAAGACGGCATACGAGATcgtcgccaGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCT	TGGCGACG
21	CAAGCAGAAGACGGCATACGAGATgtcaaccaGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCT	TGGTTGAC
22	CAAGCAGAAGACGGCATACGAGATgaccggaaGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCT	TTCCGGTC
23	CAAGCAGAAGACGGCATACGAGATgttcagaaGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCT	TTCTGAAC
24	CAAGCAGAAGACGGCATACGAGATggtctcaaGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCT	TTGAGACC
25	CAAGCAGAAGACGGCATACGAGATacttcgttGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	AACGAAGT
26	CAAGCAGAAGACGGCATACGAGATatggcgttGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	AACGCCAT
27	CAAGCAGAAGACGGCATACGAGATgaatccttGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	AAGGATTC
28	CAAGCAGAAGACGGCATACGAGATtagcgagtGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCT	ACTCGCTA
29	CAAGCAGAAGACGGCATACGAGATccggttctGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	AGAACCGG
30	CAAGCAGAAGACGGCATACGAGATgttcaactGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	AGTTGAAC
31	CAAGCAGAAGACGGCATACGAGATagaggttgGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCT	CAACCTCT
32	CAAGCAGAAGACGGCATACGAGATtaagatggGTGACTGGAGTTCAGACGTGTGCTCT TCCGATCT	CCATCTTA

33	CAAGCAGAAGACGGCATACGAGATcgcctcggGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	CCGAGGCG
34	CAAGCAGAAGACGGCATACGAGATgagataggGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	CCTATCTC
35	CAAGCAGAAGACGGCATACGAGATtctcctagGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	CTAGGAGA
36	CAAGCAGAAGACGGCATACGAGATataacgagGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	CTCGTTAT
37	CAAGCAGAAGACGGCATACGAGATaacggttcGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	GAACCGTT
38	CAAGCAGAAGACGGCATACGAGATaccaatgcGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	GCATTGGT
39	CAAGCAGAAGACGGCATACGAGATagaaccgcGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	GCGGTTCT
40	CAAGCAGAAGACGGCATACGAGATcggctgccGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	GGCAGCCG
41	CAAGCAGAAGACGGCATACGAGATatctgaccGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	GGTCAGAT
42	CAAGCAGAAGACGGCATACGAGATgagaatacGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	GTATTCTC
43	CAAGCAGAAGACGGCATACGAGATtattgaacGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	GTTCAATA
44	CAAGCAGAAGACGGCATACGAGATcgcaggcaGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	TGCCTGCG
45	CAAGCAGAAGACGGCATACGAGATctagtcaaGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	TTGACTAG
46	CAAGCAGAAGACGGCATACGAGATtggatcaaGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	TTGATCCA
47	CAAGCAGAAGACGGCATACGAGATtctgccaaGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	TTGGCAGA
48	CAAGCAGAAGACGGCATACGAGATggagccaaGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	TTGGCTCC

Table 2.1 Index primers used in this study. These primers were designed with 8 bp indices using the "create_index_sequences.py" Python script from https://bioinf.eva.mpg.de /multiplex/.

2.1.7 DNA QUANTIFICATION AND SAMPLE POOLING

MYcroarray® recommends pooling samples just before sequencing. However, because it would not be cost-effective, equimolar pooling of previously prepared libraries was performed prior to sequence capture (see Furzer (2014), Shearer *et al.* (2012) or Rohland & Reich (2012)). To do this, DNA from all libraries was quantified using the fluorescent dye PicoGreen (Quant-iT™ PicoGreen® from Invitrogen™). Briefly, the dye was mixed with both samples of known concentration ("standards") and aliquots of the libraries. After excitation at ~485 nm, the fluorescent probes that are bound to double-stranded DNA emit light at ~530 nm which can be measured using the fluorometer VarioSkan™ Flash from ThermoFisher. Both the standards and the libraries were read as triplicates and readings were averaged. A standard curve was then built to infer the libraries concentration. Finally, 6.10^{-13} mole of each library (which is roughly the number of mole in 200 ng DNA with an average fragment size of 500 bp) was mixed into five pools of ~24 samples. In total, 115 samples were prepared for sequence capture (see Table S4.1 for a list of all samples).

2.2 CAPTURE, ENRICHMENT AND SEQUENCING

2.2.1 BAIT DESIGN

120 bp biotinylated RNA baits were designed to capture genomic regions from *Albugo candida* but also from other pathogens (*A. laibachii*, *Hyaloperonospora arabidopsidis*, *Phytophthora infestans*, various plant pathogenic fungi and bacteria, see Chapter 3 Table 3.1 for a detailed list of the pathogens from which baits were designed). Baits were designed to capture putative effector genes in oomycetes, but also other loci including mitochondrial and nuclear housekeeping genes in all targeted organisms (see Chapter 3, Table 3.2 for a detailed list of the genes that were targeted in the various organisms). To do this, a fasta file was created with genomic sequences from all targeted loci and a Perl script provided by Dr. Graham Etherington (TGAC, Norwich, UK) was used to break down the loci into 120 bp sequences (baits) with no overlapping (or tiling). In total, 18,348 120 bp baits covering ~2 Mb of sequences were sent to MYcroarray® for manufacture (http://mycroarray.com/).

2.2.2 SEQUENCE CAPTURE AND ENRICHMENT OF CAPTURED DNA

Sequence capture was performed on the five pooled libraries prepared above. Each pool was mixed with ~30.10^{12} baits (~$1.5.10^{09}$ of each bait) and placed at 65°C in a thermocycler for ~36 hours. The principle behind this is that biotinylated RNA baits will hybridize with DNA sequences that are at least 80% identical (Jupe *et al.* 2013). After sequence capture, RNA baits (either free or hybridized to complementary DNA molecules) were recovered using streptavidin-coated magnetic beads (Dynabeads® MyOne™ Streptavidin C1 from Invitrogen™). Recovered targets were then amplified using the KAPA HiFi DNA polymerase (Kapa Biosystems) with both P5 and P7 primers for 14 PCR cycles to generate enough DNA material for sequencing. DNA was quantified again using PicoGreen (as described above) before enriched libraries were mixed into 3 final pools of 22, 46 and 47 samples prior to size selection and sequencing.

2.2.3 SIZE SELECTION OF ENRICHED DNA AND FINAL QUALITY ASSESSMENT

Small DNA fragments preferentially bind to the flow cell during sequencing (Head *et al.* 2014). Therefore, uniformity of library fragment size is desirable to avoid over-

representation of small fragments in the data. To achieve this, enriched DNA was size-selected using the time-based mode of the Electrophoretic Lateral Fractionator SageELF^{TM} (Sage Science, 2% agarose cassette) and ~500 bp fractions were recovered. These fractions were then purified using a 1:1 Agencourt AMPure XP beads (Beckman Coulter®) to DNA ratio before quality assessment of the final pooled libraries using the 2100 Bioanalyzer from Agilent.

2.2.4 ILLUMINA SEQUENCING

Final pooled libraries were sequenced at TGAC (Norwich, UK) using the rapid mode of the Illumina HiSeq 2500 with 150 bp paired-end reads. Samples were demultiplexed by the bioinformatics pipeline group at TGAC. Also provided by TGAC were general information on sequencing quality including the number of reads per sample and mean Q30 to base which is the average number of bases per read with a Phred score larger than 30 (one-in-a-million chance that the base has been called incorrectly).

2.3 BIOINFORMATICS

2.3.1 READ ALIGNMENT

Depending on the analysis, reads were aligned to all targets from which baits were designed or to particular sequences (e.g. 'contig 1' or diversity-tracking genes of *Albugo candida* race *Ac*Nc2, see Chapter 4). In all cases, the Burrow Wheelers Aligner BWA version 0.7.4 (Li & Durbin 2009) was used with the BWA-MEM algorithm to create an alignment (BAM) file. Read duplicates which may have arisen from amplification of adapter-ligated DNA and post-capture enrichment were discarded using SAMtools version 0.1.19 (rmdup command).

2.3.2 EVALUATION OF THE MAPPING/SEQUENCING QUALITY

Read depth, the number of reads per base, as well as read coverage, the percentage of bases covered by reads, was evaluated using the depth command in SAMtools version 0.1.19. These statistics were averaged across loci or organisms depending on the analysis performed. To avoid potential biases due to contamination or misalignment, bases covered by less than 10

reads were regarded as unsequenced. SAMtools version 0.1.19 was also used to compute the percentage of reads that were found to map on targets (flagstat command).

2.3.3 GENERATION OF CONSENSUS SEQUENCES

Consensus sequences of *Albugo candida* 'contig 1' and diversity-tracking genes (see Chapter 4) were generated prior to nucleotide divergence and phylogenetic analyses. To do this, reads from all samples were aligned to reference sequences of race *Ac*Nc2 using BWA version 0.7.4 (as described above). Variants were then called using SAMtools version 0.1.19 (mpileup command) 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 version 0.1.19. A short shell script was written in Linux to convert FASTQ files into FASTA files and diversity-tracking genes from each sample were concatenated. Sequences were finally aligned using the multiple alignment program MAFFT version 7.127 (Katoh & Standley 2013). There were a total of 398,508 and 21,439 base positions in the final datasets of 'contig 1' and concatenated diversity-tracking genes, respectively.

2.3.4 PHYLOGENETIC ANALYSES

Two maximum-likelihood phylogenetic trees were inferred from aligned sequences of *Albugo candida* 'contig 1' and concatenated diversity-tracking genes. The phylogenetic software RaxML version 7.7.3 (Stamatakis 2014) was used with the Generalised Time Reversible (GTR) model of nucleotide substitution, gamma distributed rate variation among sites (GTRGAMMA) and 100 bootstrap replicates. Additionally, a split network was inferred from *Albugo candida* 'contig 1' using SplitsTree version 4.14.2 (Huson & Bryant 2006). The UncorrectedP and the NeighborNet methods were used to infer nucleotide diversity between sequences and to compute the split network, respectively. 500 bootstrap replicates were performed. Ambiguous positions were either partially (RaxML) or totally ignored (SplitsTree).

2.3.5 VARITALE PIPELINE

All evolutionary forces (selection, drift, gene flow, recombination and mutation) have an impact on the pattern of variation observed in genetic sequences (Hartl & Clark 1997). In particular, selection and gene flow may bias the accurate reconstruction of the evolutionary history of species or populations (Jouet *et al.* 2015). Therefore, neutrally evolving loci whose

fate is expected to be determined through random genetic drift only are typically used in phylogenetics. In Chapter 3, the phylogenetic relationships of *Albugo candida* isolates were inferred based on a ~400 kb contig ('contig 1'). Additionally, 32 (concatenated) so-called diversity-tracking loci were used to control for the above evolutionary processes. To select these, several neutrality tests (Fu's Fs, Tajima's D and d_N/d_S) were performed using the varitale pipeline (Ishaque 2012): a suite of Perl scripts integrated with PAML 4 (Yang 2007), PHASE (Stephens & Scheet 2005) and DNAsp (Librado & Rozas 2009). These tests were based on all gene models identified in the seven laboratory isolates from which whole genome data was available (*Ac*Nc2, *Ac*Em2, *Ac*BoT, *Ac*2v, *Ac*7v, *Ac*Ex1 and *Ac*BoL). Genes were selected when Fu's Fs and Tajima's D statistics were between -0.3 and 0.3 and d_N/d_S between 0.8 and 1.2 (see Table S2.1 for parameter estimates).

2.3.6 NUCLEOTIDE DIVERGENCE ANALYSIS

Pairwise nucleotide divergence was evaluated for *Albugo candida* 'contig 1' and concatenated diversity-tracking genes using MEGA version 6.06 (Tamura *et al.* 2013). Nucleotide divergence was computed as the number of base differences per site from averaging over all sequence pairs between and within races which were defined using the phylogenetic analyses described above. Ambiguous positions were removed from each sequence pair.

2.3.7 RECOMBINATION ANALYSIS

Recombination analysis was performed on *Albugo candida* 'contig 1' using the software HybridCheck (Ward & van Oosterhout 2015). 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. In *Albugo candida*, recombination blocks were identified both within and between races. Using the same software, these blocks were dated based on the number of mutations at the block between isolates and a strict molecular clock with a mutation rate of 10.10^{-09} mutations/site/generation (see Ward & van Oosterhout (2015) for detailed methodology). The mean age (5-95% confidence interval) of the recombination blocks (in generations) is reported in Chapter 4.

2.3.8 DETECTION OF HETEROZYGOUS SITES AND HETEROZYGOSITY ANALYSIS

Heterozygosity of *Albugo candida* was investigated at 'contig 1'. Throughout this thesis, heterozygosity is calculated as the proportion of heterozygous sites within loci and individuals. To do this, reads from all samples were aligned to reference 'contig 1' from *Ac*Nc2. Variants were called using the mpileup command in SAMtools version 0.1.19 and stored in BCF format before conversion in VCF (using the bcftools view command in SAMtools version 0.1.19). Sites where read depth was lower than 50x and/or with a Phred-scaled quality score lower than 100 (~1 chance in 1.10^{10} that the base was called incorrectly) were then removed using vcflib (https://github.com/vcflib/vcflib#vcflib) and VCFtools version 0.1.10 (Danecek *et al.* 2011). Isolates where more than 11% sites were removed were discarded from all analyses on heterozygous sites (throughout Chapter 5). Heterozygosity was estimated for each isolate as the proportion of heterozygous sites at 'contig 1' (unfiltered bases only). The mean and standard deviation of the mean percentage of heterozygous sites were estimated for the entire 'contig 1' and are reported in Chapter 5 for each *Albugo candida* race.

2.3.9 EVALUATION OF PLOIDY IN *ALBUGO CANDIDA*

Ploidy level of *Albugo candida* was evaluated for all samples using 'contig 1'. To do this, reads from all samples were aligned to reference 'contig 1' from *Ac*Nc2; variants were called and stored in pileup format. Using Perl scripts provided by Dr. Diane Saunders, the read proportion of each SNP at heterozygous sites was then calculated (scripts also used in Yoshida *et al.* (2013) to evaluate the ploidy level of *Phytophthora infestans* isolates). The rationale behind this is that, for diploid organisms, each bi-allelic SNP should account for ~50% of the reads. For triploids, the percentages are ~33 and 67% of each SNP and for tetraploids, it can be either 50-50% (when both bases occur twice) and 25-75% (when one base is present in one copy, and the other in three copies). The distribution of the read proportion of each SNP at heterozygous sites was then graphed for each isolate using R version 3.1.2 (ploidy graphs). 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 contigs (35,029,411 bp).

2.3.10 EXPERIMENTAL SIMULATION OF MIXED INFECTIONS

To simulate mixed infections, reads from *Albugo candida* wild isolates were merged according to six combinations: two diploids (#72 and 97), two triploids (#37 and 87), two

tetraploids (#78 and 84), one diploid and one triploid (#97 and 87), one diploid and one tetraploid (#97 and 84) and finally, one triploid and one tetraploid (#87 and 84). Analysis of ploidy at 'contig 1' was then repeated (described above) and ploidy graphs were built. Heterozygosity was then calculated for each simulated mixed infection as the percentage of heterozygous sites at 'contig 1'.

2.3.11 EVALUATION OF SHARED HETEROZYGOUS SITES WITHIN *ALBUGO CANDIDA* RACES

A script was written in Minitab version 12.1 (Minitab Inc.) to compute the proportion of heterozygous sites that are shared between all pairs of isolates within a race at 'contig 1'. The mean and standard deviation of the mean percentage of shared heterozygous sites over the entire 'contig 1' was then graphed for each *Albugo candida* race (Chapter 5, Figure 5.7).

2.3.12 ISOLATION BY DISTANCE ANALYSIS

A potential correlation was investigated between the proportion of heterozygous sites that are shared between isolates at 'contig 1' (see above) and the distance, in kilometres, from which they were collected. To do this, the R package "fields" ("rdist.earth" function) was first used to calculate the distance between any two isolates, based on their GPS coordinates (in degrees). A mantel test was then performed to compare both the distance and percentage of heterozygous sites matrices, using the "ade4" package (function "mantel.rtest") in R version 3.1.2 (R Core Team 2014) and 500 replicates. This analysis was carried out independently for isolates collected on *Sisymbrium officinale*, *Aubrieta deltoidea*, *Capsella bursa-pastoris* and *Lunaria annua*.

2.3.13 STATISTICAL ANALYSES AND GRAPHS

Most statistical analyses were performed in Minitab version 12.1 (Minitab Inc., Pearson correlation analyses, paired and 2-sample t-tests as well as computation of basic statistics). Regression as well as post-hoc power analyses were performed in Sigmaplot version 10.0. All graphs were produced in Sigmaplot version 10.0 unless otherwise stated.

All analyses were performed using the High Performance Computing Resources of the Norwich BioScience Institute Partnership Computing infrastructure for Science group.

CHAPTER 3: RATIONALE FOR AND DESIGN OF PATHSEQ: A CAPTURE-BASED METHOD TO INTERROGATE MICROBIAL DIVERSITY

3.1 INTRODUCTION

Since the discovery of microorganisms in the 17th century (Gest 2004), microbiologists have been steadily indexing them (Woese 1987). Unlike animals and plants however, microorganisms usually have simple morphologies and sexual reproduction is lacking or poorly described (Amann *et al.* 1995). These fundamental differences hampered their phylogenetic classification and other, more subtle traits such as physiological traits were first used (Cohan 2002). In the 1970s, microbiologists adopted DNA:DNA hybridization methods to delineate between microbial species and in 1973, Johnson (1973) defined that if two strains share no more than 70% of their genomes, they could be considered distinct species. Although still a golden standard in the field of microbiology, this method unfortunately cannot not be applied to as-yet unculturable microorganisms (Gevers *et al.* 2005).

With the advent of the PCR and sequencing technologies, several marker genes have become routinely used in microbial classification (Pontes *et al.* 2007). While the ribosomal Internal Transcribed Spacer (ITS) has been proposed as universal barcode marker for fungi (Chase & Fay 2009; Schoch *et al.* 2012), the 16S ribosomal DNA (rDNA) sequence has been adopted by many as a standard genetic marker for bacteria. In the latter case, two strains with ≤97% identity are often considered distinct species (Drancourt *et al.* 2004; Gevers *et al.* 2005; Naser *et al.* 2005; Turnbaugh *et al.* 2009). Although these markers have many advantages and are still being used today, they are now often complemented by MLSA (MultiLocus Sequence Analysis; Gevers *et al.* 2005; Almeida *et al.* 2010; Doroghazi & Buckley 2010; Wicker *et al.* 2012; Bouvet *et al.* 2014). This more recent method consists of using concatenated housekeeping genes to infer relationships between individuals (Hanage *et al.* 2006; Macheras *et al.* 2011; Wicker *et al.* 2012). By increasing the number of genetic markers, biases associated with PCR as well as with potential recombination events are reduced while sequence information is increased (Hanage *et al.* 2006).

MLSA can now easily be performed for a high number of individuals by enriching for and sequencing specific genomic regions (targets) using carefully designed probes (baits; Turner *et al.* 2009). Such genomic partitioning techniques include capture-by-circularization

and capture-by-hybridization methods (Hardenbol *et al.* 2003; Gnirke *et al.* 2009) and allow SNP genotyping, targeted sequencing or the selective sequencing of microbial versus host/eukaryotic DNA. Using both barcode markers and carefully selected genes, these methods may also be used to identify and study the genetic diversity of microbial species from environmental samples without the need to grow them in the lab (Cohan 2002; Hongoh *et al.* 2003; Baker *et al.* 2003; Gevers *et al.* 2005).

Albugo candida is a plant pathogen that cannot yet be cultured in the laboratory and although a reference genome is available (Links *et al.* 2011), little is known about the species diversity in nature. It is responsible for the white rust disease of many Brassicaceae (>300 species according to the fungal database of the Systematic Mycology and Microbiology Laboratory at the USDA-ARS, Farr *et al.* 2004) but it is organized in physiological races that are specialized on a single or closely-related plant species (Hiura 1930; Pound & Williams 1963; Petrie 1988). Depending on the classification method, up to 20 pathotypes of *A. candida* have been identified (Saharan & Verma 1992).

A. candida can reproduce both sexually and asexually (Saharan *et al.* 2014) although the relative importance of these two modes of reproduction is still largely unknown. Of the few studies investigating sexual reproduction in *A. candida*, only one reported outcrossing between distinct (physiological) races (Adhikari *et al.* 2003), while another provided evidence for genetic exchange between races over many generations (McMullan *et al.* 2015). Not surprisingly, this would be expected to keep genetic variability in wild populations of *A. candida*. It may also lead to new virulent races by generating novel sets of effector alleles allowing evasion of immunity and colonization of new hosts (Thines 2014). Yet, if *A. candida* races are highly restricted to particular hosts, how could they meet in the wild to reproduce sexually?

It has long been recognized that infection by a virulent pathogen can enhance susceptibility of the host to secondary infections (Yarwood 1951; Kalra *et al.* 1989; Meyer & Pataky 2010). In 1992, Saharan & Verma (1992) highlighted many studies that reported frequent associations between *A. candida* and *Hyaloperonospora parasitica*. Later in 1998, Cooper *et al.* (1998) showed that two Brassicaceae could lose resistance to several pathogens including *H. parasitica*, *H. arabidopsidis* and *Bremia lactucae* after pre-inoculation with *A. candida*. Although the suppression of the immune system by *A. candida* can be beneficial for other microorganisms, it also could give an evolutionary advantage to the pathogen. Recently, McMullan *et al.* (2015) demonstrated that infection by a virulent race of *A. candida* enables subsequent co-colonization by an otherwise avirulent race. This co-occurrence of genetically

diverged *A. candida* races is a likely starting point for outcrossing events that may help maintain high genetic variation within *A. candida* wild populations as well as create novel sets of alleles allowing host jumps.

During my PhD, I developed a solution-based capture-by-hybridization method (PathSeq) with the aim of detecting microorganisms from the field as well as establishing *A. candida* intraspecific evolutionary relationships. This method is based on the RenSeq method published by Jupe *et al.* (2013) where the authors used biotinylated RNA baits to capture, enrich and sequence resistance genes (R-genes) from *Solanum tuberosum* clone DM. In PathSeq, the RNA baits target DNA sequences from a large variety of plant pathogens including *A. candida*, *H. arabidopsidis*, *Phytophthora infestans*, *Fusarium oxysporum*, *Ralstonia solanacearum* and *Pseudomonas syringae*. This novel method combines both the advantages of the DNA barcode markers and MLSA and allows the simultaneous study of genetic variation in multiple microbial species. Another advantage of the method is that it allows up to 20% mismatches between the RNA baits and target sequences (Jupe *et al.* 2013). Therefore, compared to PCR-based methods it does not require as much prior knowledge of the microbes that are present in the environment although there may be ascertainment bias causing this method to potentially miss rare, highly diverged organisms.

In this chapter, I introduce the rationale behind and design of the PathSeq method. I also investigate whether PathSeq can be used to (i) detect pathogens directly from the field, (ii) analyse the genetic diversity of multiple organisms in a high-throughput manner and (iii) to estimate the abundance of microorganisms in an environmental sample.

3.2 RESULTS

3.2.1 DESIGN OF THE PATHSEQ BAIT LIBRARY

In PathSeq, baits were designed so as to hybridize with sequences from 48 microbial species, including 2 rhizaria, 5 oomycetes, 12 fungi and 29 bacteria as well as from several Brassicaceae (Table 3.1). Although most of the species included in the capture are pathogenic to plants, several bacteria were added for their potential role in the induction of resistance and growth promotion in plants (Chen *et al.* 2007; Lugtenberg & Kamilova 2009; Han *et al.* 2011; Choi *et al.* 2015; Manzanera *et al.* 2015). Importantly, although baits were designed based on

the sequences of ~50 species, many more are potential targets as baits only need ~80% nucleotide identity to hybridize with DNA (Jupe *et al.* 2013).

Rhizaria	Oomycetes	Fungi
- *Bigelowiella natans* - *Plasmodiophora brassicae**	- *Albugo candida** - *Albugo laibachii** - *Hyaloperonospora arabidopsidis** - *Phytophthora infestans** - *Phytophthora capsici**	- *Fusarium oxysporum** - *Leptosphaeria maculans** - *Erysiphe cruciferarum** - *Erysiphe pisi** - *Marssonina brunnea** - *Pyrenopeziza brassicae** - *Verticillium dahliae** - *Mycosphaerella pini** - *Mycosphaerella brassicicola** - *Botrytis cinerea** - *Sclerotinia sclerotiorum**

Plants	Bacteria	
- *Aubrieta deltoidea* - *Sinapis arvensis* - *Capsella bursa-pastoris* - *Arabidopsis thaliana* - *Alliaria petiolata* - *Raphanus sativus* - *Brassica oleracea* - *Cardamine hirsuta* - *Alyssum montanum*	- *Candidatus Liberibacter spp.** - *Acidovorax spp.** - *Actinoplanes missouriensis* - *Agrobacterium tumefaciens** - *Arthrobacter spp.* ∓ - *Bacillus amyloquefaciens*∓ - *Clavibacter michiganensis subsp. michiganensis** - *Curtobacterium flaccumfaciens** - *Duganella spp.* - *Pectobacterium carotovora** - *Flavobacterium johnsoniae* - *Leifsonia xyli** - *Massilia niastensis* - *Methylobacterium extorquens*	- *Pseudomonas spp.** ∓ - *Ralstonia solanacearum** - *Rathayibacter spp.** - *Rhodococcus spp.** - *Sphingomonas phylloshaerae* - *Spiroplasma citri** - *Streptomyces spp.** - *Variovorax paradoxus*∓ - *Xanthomonas campestris pv. campestris** - *Xylella fastidiosa** - *Acetobacter aceti* - *Actinomyces naeslundii* - *Clostridium botulinum* - *Micrococcus luteus* - *Pantoea spp.**

Table 3.1 Species from which PathSeq baits were designed. Microbial species that are known to be pathogenic to plants are marked with a star (*) while the ∓ symbol is added to bacteria that are known to promote plant growth and resistance to pathogens. The sequences that were used for each type of organisms are provided in Table 3.2.

For each type of organisms, baits were designed based on the universal barcode marker that has been proposed in the literature and several other housekeeping genes (Table 3.2). Sequences were downloaded from the Genbank database when available (Benson *et al.* 2009) and combined in a fasta file. Further oomycete sequences were added including putative effectors from all target species. Sequences from *Hpa* and *Phytophtora spp.* were provided by Dr. Liliana Cano (NC state University, USA) and Dr. Ingo Hein (James Hutton Institute, Scotland), respectively. In *Albugo spp.*, effectors were predicted based on the presence of a N-terminal Cys-His-x-Cys (CHxC) motif followed by a conserved Glycine (Kemen *et al.* 2011) and a signal peptide. An additional ~400 kb contig from *A. candida* was included to allow

recombination analyses as well as 32 "diversity-tracking" genes. These 32 genes were selected to study natural variation in *A. candida* wild populations because they seem to accumulate non-synonymous and synonymous mutations at approximately the same rate, based on the sequences of 6 isolates (d_N/d_S ~1, Tajima's D and Fu's Fs ~0 (see Table S2.1) using the varitale pipeline (Ishaque 2012)).

Rhizaria	Oomycetes	Fungi	Bacteria
- ITS1, 5.8S rDNA, ITS2 - COX1‡[a] - COX2 - COX3 - cytochrome b - beta-tubulin - Hsp70 - actin	- ITS1, 5.8S rDNA, ITS2‡[b] - COX1‡[b] - COX2 - COX3 - cytochrome b - beta-tubulin - Hsp70 - actin - putative effectors	- ITS1, 5.8S rDNA, ITS2‡[c] - COX1 - COX2 - COX3 - cytochrome b - beta-tubulin - Hsp70 - actin	- 16S rDNA‡[d] - rpoD - rpoB - gyrase b
Plants			
- ITS1, 5.8S rDNA, ITS2‡[e]	*Albugo candida*: - ~400kb contig - diversity tracking genes		

Table 3.2 Marker genes used for the design of baits. Marker genes that have been proposed as universal barcode are marked with the symbol ‡. a: Saunders & Mcdevit 2012; b: Robideau *et al.* 2011; c: Schoch *et al.* 2012; d: Links *et al.* 2012; e: Kress *et al.* 2005. Sequences from *Hpa*, *P. infestans*, *A. laibachii* and *A. candida* are mainly from strain Emoy2, T30-4, Nc14 and Nc2, respectively. Diversity tracking genes are genes where non-synonymous and synonymous substitutions seem to be accumulating at approximately the same rate (d_N/d_S ~1, Fu's Fs and Tajima's D ~0). *Abbreviations*: COX = Cytochrome c oxidase; ITS = Internal transcribed spacer; Hsp70 = 70 kDa Heat shock proteins; rpoD = RNA polymerase sigma factor rpoD; rpoB = RNA polymerase beta subunit.

All sequences were visually inspected in MEGA6 (Tamura *et al.* 2013) and reversed and complemented when necessary, using the reverse complement tool at http://arep.med.harvard.edu/labgc/adnan/projects/Utilities/revcomp.html. Targets were finally broken down into k-mers (120 bp) using a custom Perl script (Etherington, unpublished) and sent to MYcroarray® for manufacture (http://mycroarray.com/). In total, 18,348 120 bp baits were synthesized that cover all targets (>2Mb) without overlap (Figure 3.1). These were received as part of a Mybaits® customized target enrichment kit containing 6.10^{12} baits per microliter.

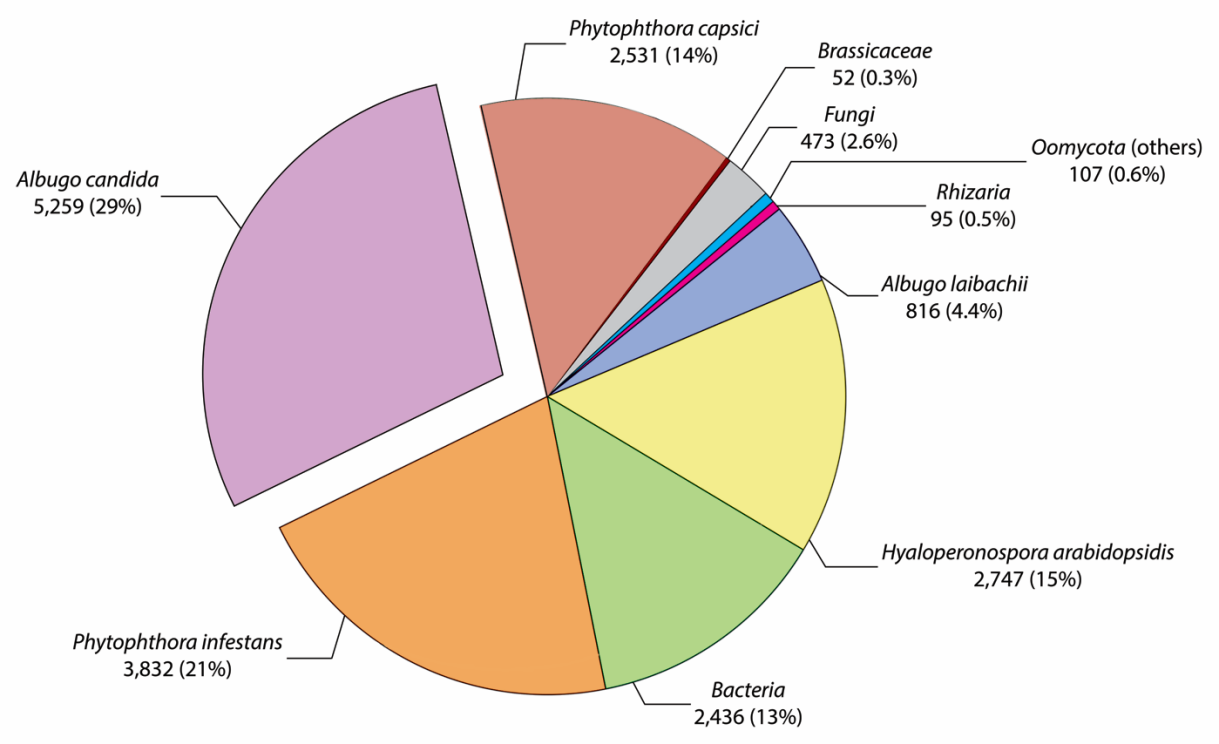


Figure 3.1 Percentage and number of 120 bp baits that are dedicated to the various microorganisms targeted in PathSeq.

3.2.2 TESTING PATHSEQ

The PathSeq method was first tested on a set of 22 samples including *A. candida* isolates collected from various hosts and locations (Table 3.3, PathSeq 1 in Table S3.1). Importantly for this test run, several Brassicaceae that had been infected by known strains of pathogens were included as both positive and negative controls (Table 3.4). DNA was first extracted from whole infected leaves using a phenol-chloroform extraction protocol and sheared to fragments of ~500 bp by ultrasonication. Adaptors were then ligated to each sample followed by sample-specific barcodes, required for sample identification during the multiplex sequencing run. Although MYcroarray® does not recommend pooling samples before capture, 6.10^{-13} moles of DNA per sample were mixed at this stage to reduce costs. This has already been done using 3-4 samples (Furzer 2014) but in this implementation I pooled 22 samples. Next, the capture of targets using the PathSeq bait library was carried out as advised by MYcroarray®, at 65°C for ~36 hours in a thermocycler. Hybridized material was recovered using streptavidin beads and amplified by PCR for 14 cycles. The quality of the enriched pooled library was finally visualised using the Agilent 2100 Bioanalyzer and sequenced on one HiSeq™ lane to generate 150 bp paired-end reads.

To assess the efficiency of the capture, reads were mapped to all targets using BWA version 0.7.4 and several statistics were estimated using SAMtools version 0.1.19 (Li *et al.* 2009). The proportion of reads on target as well as the average read depth at the 400 kb contig from *A. candida* is reported in Table 3.3. Although this contig is not representative of the whole bait library, there should be limited read misalignment due to homology with other species. It should have also been consistently captured and enriched in all but one sample, providing a mean of comparison between samples.

#	Samples	# reads	Mean Q30 to base	% reads on target	Average depth at *Ac* 400 kb contig
1	*Ac* on *S. officinale* (ENG)	13,883,380	150	48.67	1077
2	*Ac* on *C. bursa-pastoris* (ENG)	16,472,932	150	57.87	1541
3	*Ac* on *A. deltoidea* (ENG)	10,608,596	150	52.53	837
4	*Ac* on *A. deltoidea* (FRA)	16,119,514	150	55.40	1194
5	*Ac* on *R. sativus* (FRA)	15,327,772	150	56.29	1305
6	*Ac* on *C. bursa-pastoris* (POL)	20,025,168	150	55.72	1632
7	*Ac* on *A. saxatile* (ITA)	4,177,486	150	28.17	141
8	*Ac* on *A. deltoidea* (ITA)	4,828,750	150	29.14	213
9	*Ac* on *R. sativus* (ENG)	7,800,086	150	35.08	143
10	*Ac* on *S. officinale* (ENG)	12,303,036	150	53.55	699
11	*Ac* on *C. bursa-pastoris* (ENG)	18,264,106	150	56.63	1212
12	*Ac* on *A. saxatile* (ENG)	7,150,974	150	44.69	403
13	*Ac* on *B. nigra* (FRA)	8,919,184	150	41.40	266
14	*Ac* on *C. bursa-pastoris* (FRA)	8,780,834	150	54.58	530
15	*Ac* on *A. deltoidea* (FRA)	10,594,584	150	49.18	535
16	*Ac* on *A. saxatile* (FRA)	4,730,746	150	36.97	144
17	*Ac* on *S. alba* (ENG)	4,656,906	150	36.75	157
18	*Hpa* Emoy2 + *Al*Nc14 on *A. thaliana* Col-0	16,006,684	150	45.72	56
19	*Ac*2v + *Erysiphe* sp. on *B. juncea*	17,847,200	150	58.83	1625
20	*Ac*BoT on *B. oleracea*	16,284,674	150	53.16	1388
21	*Ac*Ex1 + *P. infestans* on *A. thaliana* Col-0	18,350,820	150	56.20	1262
22	*Ac*Ex1 + *P. infestans* + *P. syringae* DC3000 on *A. thaliana* Col-0	24,860,952	150	51.11	2049

Table 3.3 Read and mapping statistics for samples included in PathSeq 1. The first 17 samples are *A. candida* (*Ac*)-infected plants from the wild. The sampling location is provided as the first three letters of the country they were collected in. The last five samples are the controls. Mean Q30 to base is the number of bases, in a read, with a Phred quality score ≥ 30 (a base with less than 0.1% chance of being called incorrectly).

#	Pathogens			Hosts
18	*H. peronospora* Emoy2	*A. laibachii* Nc14	-	*A. thaliana* Col-0
19	*A. candida* 2v	*Erysiphe* sp.	-	*B. juncea*
20	*A. candida* BoT	-	-	*B. oleracea*
21	*A. candida* Ex1	*P. infestans* 88069^{td}	-	*A. thaliana* Col-0
22	*A. candida* Ex1	*P. infestans* 88069^{td}	*P. syringae* DC3000	*A. thaliana* Col-0

Table 3.4 Controls included in the test run of PathSeq. The numbers provided in the first column correspond to the samples in Tables 3.3 and S3.1.

A large variation in the number of reads generated per sample can be observed, probably due to pipetting error, inaccurate quantification of DNA before pooling, differentials in the adaptor ligation efficiency or in the size of the sample libraries (Quail *et al.* 2008). Moreover, the proportion of unmapped reads is quite high, from ~40-70%. To investigate the origin of these off-targets, reads from sample #2 (Table 3.3) were mapped to the genomes of *A. candida* and *Capsella rubella*, a close relative of the host plant, *C. bursa-pastoris* (extracted from Genbank, WGS project ANNY01). 74.94% of the reads corresponded to *A. candida* while 21.17% could align to *Capsella rubella*, suggesting that most off-targets (~42% of the reads for this sample) originate from these two species. Nevertheless, the average read depth at the 400 kb contig is more than sufficient for variant detection, with a minimum of 141x and the proportion of reads on target is comparable or higher to that found in Jupe *et al.* (2013, ~30% compared an average of 48% (±SD = 9.5%) in this first PathSeq run). To validate the method further, the ability of PathSeq to detect microorganisms in a sample was put to the test. Read coverage, the percentage of targeted base pairs that are covered by ≥10 reads, was estimated in the control samples for the following organisms: *A. candida*, *A. laibachii*, *P. infestans*, *Erysiphe* sp. and *P. syringae* (Figure 3.2).

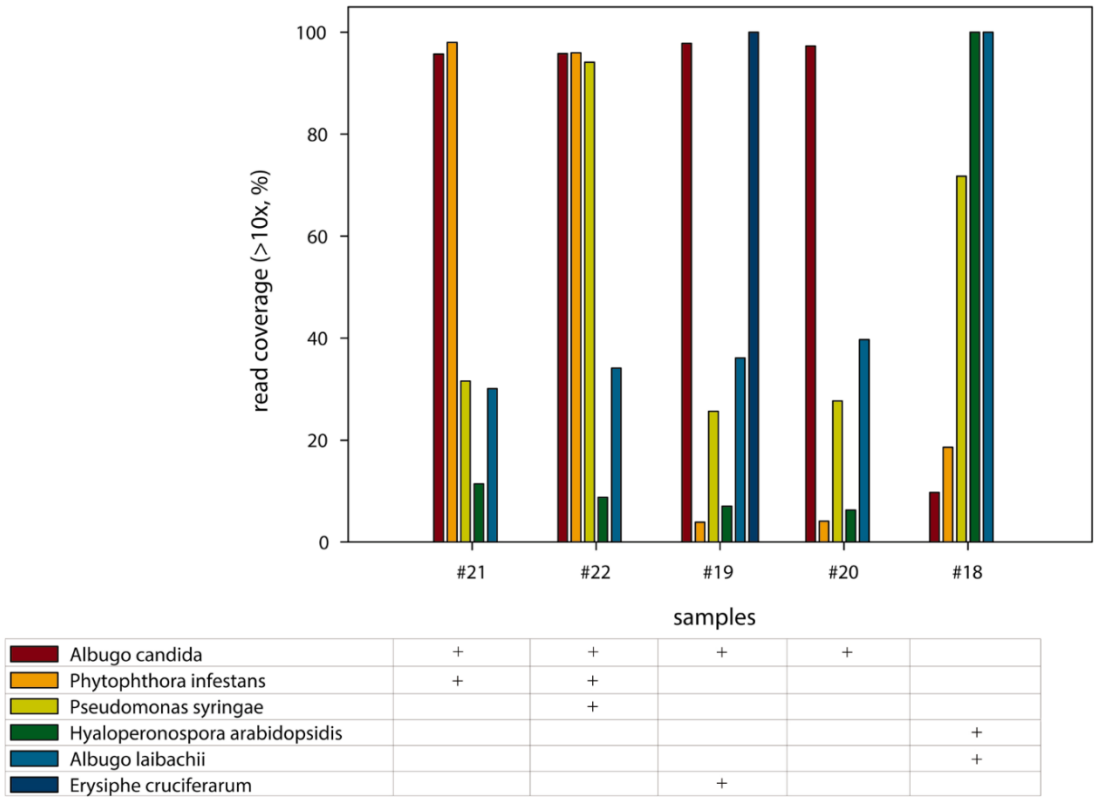


Figure 3.2 Proportion of targets that are covered by ≥10 reads in control samples. The presence of control organisms is indicated by plus signs. The numbers on the x-axis represent the samples in Tables 3.4 and S3.1.

Although some targets were not sequenced due to presence/absence polymorphism between the strains used in the bait design and those in the sample, close to 100% of the targets were captured and sequenced from control organisms when present. Not surprisingly however, a small proportion of reads were wrongly assigned to control organisms due to the likely presence of closely-related species. This sequence homology can lead to misalignment errors as illustrated by the high read depth at the 400 kb contig of *A. candida* in sample #18 (Table 3.4).

Another extreme example of read misalignment is the internal transcribed spacer sequences from the reference plants, all covered by reads due to high homology between species. In that case, it is the highest read depth that highlights the correct plant host (Figure 3.3). Mapping reads to reference sequences is therefore not a good strategy for the identification of organisms in an environmental sample and possible alternatives are discussed below.

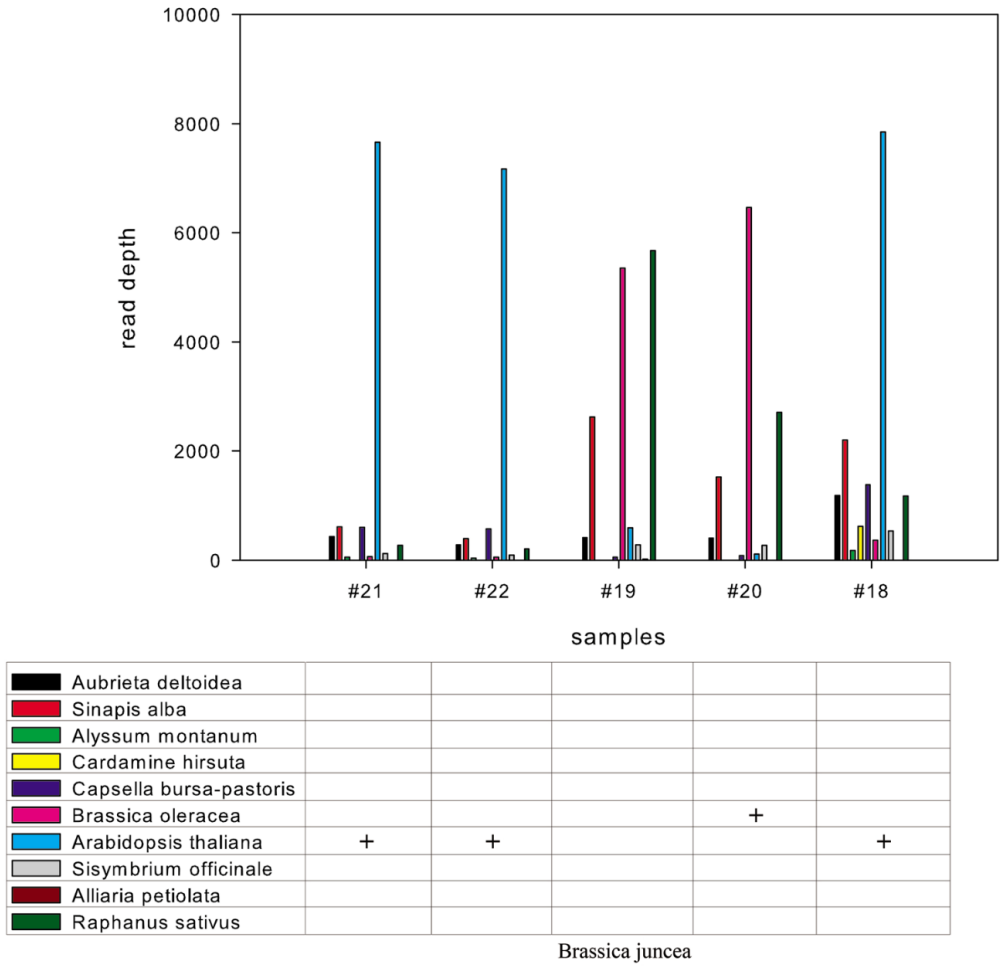


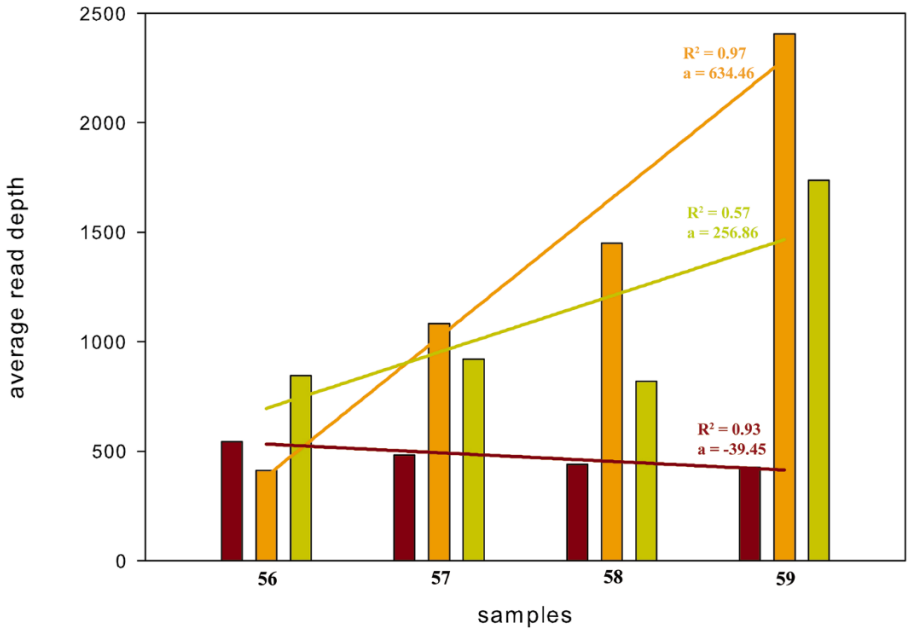
Figure 3.3 Read depth at the internal transcribed spacer sequence of the reference plants in control samples. The plant species included in the controls are indicated by plus signs. The numbers on the x-axis represent the samples in Tables 3.4 and S3.1.

3.2.3 RELATIVE ABUNDANCE OF PATHOGENS USING PATHSEQ

After a successful PathSeq run, 93 additional samples were included in two HiSeq™ lanes so that 46 and then 47 samples were pooled before sequencing. In total, 91 *A. candida* infected leaves and 24 controls were sequenced (Table S3.1). As with the 22 initial samples, mean Q30 to base was high (mean (+/-SD) = 142.1 (+/-15.3)). The proportion of reads on target was variable between samples but close to or greater than 30% (mean (±SD) = 41.79% (±15.5)). Finally, although *A. candida* failed to be sequenced from a few infected leaves samples (24, 25, 75, 85, 96, 99, 100 and 101), the 400 kb contig was captured and enriched always entirely and read depth was high enough to counterbalance sequencing errors and allow accurate and robust variant detection. These statistics are provided in Table S3.1.

As a final test, I investigated whether PathSeq could provide information about the abundance of pathogens using a series of controls. DNA from sterilized *Aubrieta deltoidea*, *P. syringae* DC3000 and *A. candida* 2v zoospores was prepared as described in Chapter 2 (section 2.1.3), and quantified after ligation of sample-specific barcodes. Varying amounts of DNA were then mixed into a set of 4 controls before additional quantification and pooling with other samples. Correlation between DNA quantity and read depth, the number of times a base pair was sequenced, was finally investigated (Figure 3.4).

Because read depth is correlated with the total number of reads generated per sample (Pearson's $r = 0.880$, $P < 0.001$), it was first normalized against the control with the highest number of reads. After regression analyses in SigmaPlot version 10.0, the slope coefficients were found to be both in agreement with increasing amounts of *P. syringae* and decreasing amounts of *A. candida* from left to right ($a = +634.46$ ($P < 0.05$) and -39.45 ($P < 0.05$), respectively), suggesting that read depth is correlated with the abundance of a pathogen. This is further supported by high regression coefficients ($R^2 = 0.98$ and 0.93). However, analysing data from *A. deltoidea* proved inconclusive with a low R^2 (0.57) and a slope coefficient with $P > 0.05$. This is probably due to the small sample size used in this analysis (post-hoc power analysis in SigmaPlot: 0.2291). Altogether, although additional samples are needed to generate more robust statistics, PathSeq may be used to estimate the relative abundance of a pathogen across samples.



Albugo candida 2v	$1.46.10^{-17}$	$7.3.10^{-18}$	$3.65.10^{-18}$	$1.83.10^{-18}$
Pseudomonas syringae DC3000	$1.42.10^{-17}$	$6.63.10^{-17}$	$9.25.10^{-17}$	$1.05.10^{-16}$
Aubrieta deltoidea	$8.08.10^{-20}$	$8.08.10^{-20}$	$8.08.10^{-20}$	$8.08.10^{-20}$

Figure 3.4 Average read depth at *A. candida*, *P. syringae* and *A. deltoidea* targets in a series of 4 controls. The quantity of DNA (in moles) is provided in the table below the graph, for each organism. Regression analyses were performed in SigmaPlot version 10.0 and both the regression (R^2) and slope coefficients (a) are reported close to the corresponding regression lines. The numbers on the x-axis represent the sample numbers provided in Table S3.1.

3.3 DISCUSSION

PathSeq is an in-solution capture-by-hybridization method which aims to detect microorganisms directly from the field. Based on a multilocus sequence analysis approach (MLSA), it further allows the study of molecular evolution under natural conditions in a high-throughput manner.

In this work, I was interested in both the mechanisms underlying *A. candida* evolution and the potential interactions it may have with other plant-associated microbes. With these aims in mind, baits were designed to hybridize with several *A. candida* loci including putative effectors, diversity-tracking genes and a ~400 kb contig as well as with housekeeping genes from a wide variety of pathogens. *A. candida* infected leaves were collected throughout Europe and prepared for sequencing. Up to 47 samples were processed in a single sequencing lane and

in total, 83 *A. candida* isolates could partially be sequenced and data from many microorganisms colonizing leaves in the field were generated.

To assess the efficiency of the method, control samples were prepared with known microbial species. Using a simple mapping approach, PathSeq was shown to be able to capture close to 100% of the targets from control organisms when present. Due to sequence homology however, a small proportion of reads was wrongly assigned to other organisms. This is due to reads mapping to a reference alignment when sequence identity is high, e.g. at conserved loci. For this reason, more suitable approaches may be used to identify microorganisms such as sequence similarity searches against large databases (BLAST, Benson *et al.* 2009) or taxonomic assignment of DNA sequences (e.g. MEGAN, Taxator-tk or Kraken (Huson *et al.* 2007; Dröge *et al.* 2014; Wood & Salzberg 2014)). By screening through large sequence databases, these methods allow the identification and subsequent treatment of conserved ambiguous regions. Using Kraken for example, sequences are assigned to the lowest common ancestor taxon (LCA) of all genomes containing them. This software can also process millions of reads at high speed.

Using the 400 kb contig from *A. candida* as representative of the whole target set, PathSeq was furthermore shown to achieve high read depth even when 47 samples were multiplexed. This enables to correct for sequencing errors, allowing robust variant detection required for analysis of genetic diversity. The method was also successful in enriching for particular pathogen loci although some undesired regions were sequenced. On average, ~41.79% of the reads were on target which is comparable to what was found in Jupe *et al.* (2013) and as expected, off-target reads were obtained from the most abundant organisms, *A. candida* and its host. These off-target reads, also called "near target" sequences, could be reduced by the use of smaller fragments in the DNA libraries (Mertes *et al.* 2011). In addition to allowing pathogen detection, PathSeq could therefore also be used to study the molecular evolution of gene families (e.g. effector genes in oomycetes or fungi) or microbial species (e.g. emerging or important plant diseases).

Not only are we interested in detecting pathogen species and studying their genetic diversity, we also aim at a better understanding of the interactions in play within microbiomes. Of special interest is the evolutionary advantage one pathogenic species may have in inducing broad host susceptibility. We therefore tested whether the abundance of a pathogen is correlated with that of reads generated using PathSeq. The hypothesis is that *A. candida* races can suppress the immune system of their hosts to coexist and exchange genetic material. While some microorganisms could exploit this newly gained susceptibility (Saharan & Verma 1992;

Cooper *et al.* 2008; Belhaj *et al.* 2015), others may be protective and promote plant health (Mendes *et al.* 2011; Berendsen *et al.* 2012). As a result, their occurrence or abundance should vary between *A. candida* infected and healthy leaves. By quantifying microbial DNA in a series of four controls, I demonstrated that PathSeq could be used to estimate the relative abundance of a pathogen between samples. With this bait library however, it is not yet possible to compare microorganisms within a sample. To do so, future bait designs should restrict targets to the same set of housekeeping, single-copy genes for all species. Finally, additional samples will be required to confirm this read depth-abundance relationship and, in any case, one will always need to cautiously interpret its findings due to variability introduced by pipetting errors or PCR biases.

In the next two chapters, I use data generated using PathSeq to investigate the genetic diversity of wild *A. candida* populations. In particular, my analyses will be based on the 400 kb contig and 32 diversity-tracking genes of *A. candida*.

CHAPTER 4: *ALBUGO CANDIDA* GENETIC DIVERSITY AND POPULATION BIOLOGY

4.1 INTRODUCTION

A. candida was first described as *Aecidium candidum* by Christiaan Hendrick Persoon in the 13th edition of Linnaeus' Systema Naturae (Gmelin 1792). Since then, its taxonomic status has undergone numerous changes. In 1801, the species was moved to the genus *Uredo* as *U. candida* Pers., "white spores", according to the symptoms it produces upon reproduction (Persoon 1801). In 1806, Henri-François-Anne de Roussel relocated *U. candida* as well as ~50 obligate biotrophs of Dicotyledonae to the genus *Albugo*. It was first placed in the Protomyceae fungal family by Samuel Frederick Gray (1821) before being transferred to the Peronosporaceae under the name of *Cystopus*, although the name *Albugo* persisted (de Bary 1863). Later, Shröter (1893) provided *Albugo* spp. with their own family, the Albuginaceae, due to their unique basipetal mode of sporangiogenesis (see also Heller *et al.* 2009). Still placed in the Peronosporales at the time, it is only in 2005 that Thines & Spring raised *Albugo* to ordinal level and introduced the Albuginales based on morphological data. In addition to sporangiogenesis, the authors highlighted features that are unique to *Albugo* spp. such as the multinucleate formation of oospores as well as the germination of oospore without formation of a germ tube. This was further supported by phylogenetic work which placed *Albugo* basal to the Pythiales (Petersen & Rosendahl 2000; Cooke *et al.* 2002; Riethmüller *et al.* 2002) or even more distant to the Peronosporales, basal to the Rhipidiales (Hudspeth *et al.* 2003).

Although *Albugo* spp. are highly similar in morphology, scientists investigated whether they could be separated into distinct groups. Among the features that were used, the presence or absence of an equatorial wall thickening of sporangia divided *Albugo* spp. into two sections: *Aequales* and *Annulatae* (the latter including *Albugo bliti* and *Albugo tragopogonis*), respectively (de Bary, 1863). More recently, Thines & Spring (2005) introduced an alternative to dividing the monogeneric *Albuginaceae* by combining the above feature with novel morphological data and proposed three genera: *Albugo*, *Pustula* and *Wilsoniana*. Sporangia in *Pustula* spp. have the largest equatorial wall thickening and are subglobose or cylindrical with a reticulate to striate surface ornamentation. The terminal sporangium of *Wilsoniana* spp. is smaller or larger than the others; it is yellowish and thick-walled. Other sporangia have a wall of uniform thickness or with a slight equatorial thickening; they are pyriform or cylindrical and

their surface ornamentation is never reticulate. Finally, the wall of sporangia in *Albugo* spp. is of uniform thickness; sporangia are mainly subglobose with a verrucose surface ornamentation. These groupings were further validated by sequence analysis of the ITS marker (Thines & Spring 2005).

Not only is this classification supported by both molecular and morphological data, it also seems to agree with the parasites host specificity. While *Wilsoniana* spp. can infect the Caryophyllidae, *Pustula* spp. are restricted to the Asteridae, and *Albugo* spp. to the Brassicales (Thines & Spring 2005; Thines *et al.* 2009; Thines 2014). This notion of host specialization also seems to apply to species within genera (*Wilsoniana portulacae* on *Portulaca* spp., *Pustula spinulosus* on *Cirsium* spp., *Albugo resedae* on *Reseda* spp.) although at least one species appears to be able to infect many diverged hosts.

Albugo candida has been reported on 241 species of 63 Brassicaceae genera (Biga 1955) and it was once considered as the only Brassicaceae-infecting white rust (Choi & Priest, 1995). In the last decade however, several host-specialized species were discovered. In 2006, Choi *et al.* formally described isolates from *Lepidium* spp. as *Albugo lepidii* using the ITS and *cox2* markers. In the years to follow, other species were described such as *Albugo koreana* on *C. bursa-pastoris* from Korea (Choi *et al.* 2007), *Albugo voglmayrii* on *Draba nemorosa* (Choi *et al.* 2008), *A. laibachii* on *A. thaliana* (Thines *et al.* 2009) and *Albugo rorippae* on *Rorippa* spp. (Choi *et al.* 2011). The oospore wall ornamentation was also used, when possible, as a discriminative criterion (Choi & Priest 1995; Choi *et al.* 2007; Ploch *et al.* 2010). Yet, the molecular and morphological uniformity of many isolates collected from diverged Brassicaceae suggested that *A. candida* is a generalist pathogen capable of infecting at least 20 genera (Saharan *et al.* 2014).

As type species of the genus *Albugo* and a generalist pathogen infecting some of our most important crops, *A. candida* isolates were further studied and divided into several host-specific races (Hiura 1930; Pound & Williams 1963; Hill *et al.* 1988; Kaur *et al.* 2008; Meena *et al.* 2014). However, although host-specificity clearly suggested the existence of several pathotypes, the mechanisms by which they evolved remained unclear. In 2003, Adhikari *et al.* were the first to show hybridization between host-specific isolates *Ac*2 and *Ac*7 (mainly restricted to *B. juncea* and *B. rapa*, respectively) using RAPD markers. High genetic diversity between *Ac*2v and *Ac*7v was also demonstrated. Later in 2011, the draft genome of *A. candida Ac*2 (*Ac*2v) was published by Links *et al.* who reported a small genome of ~45Mb and fewer pathogenicity-related genes compared to other oomycetes but a last major advance was provided when five *A. candida* isolates were sequenced and compared (McMullan *et al.* 2015).

Confirming work done by Adhikari *et al.*, the authors reported on three genetically diverged host-specific races (~1% polymorphic sites) that seem to recombine. These results represent an exciting opportunity to further explore the mechanisms by which *A. candida* pathotypes evolve.

By collecting *A. candida* isolates in the years 2013-15, I set out to undertake the first study on *A. candida* genetic diversity in the wild. Using PathSeq to generate informative data from many isolates, I will explore the genetically diverged races that can be found in the wild as well as the nucleotide diversity and the amount of sexual reproduction there is within and between these races.

4.2 RESULTS

4.2.1 COLLECTION OF *ALBUGO CANDIDA* ISOLATES

I collected white rust samples in Europe between 2013 and 2015 from several Brassicaceae species. To confirm the presence of *A. candida*, a gene coding for a putative cAMP-binding protein that is specific to *Albugo* spp. was amplified from all samples. This gene is also polymorphic at a restriction site between *A. laibachii* and *A. candida* and was used to discriminate between the two species when isolates were collected on *A. thaliana*. However, all 21 isolates collected on this host were confirmed as *A. laibachii*. Similarly, isolates collected on *Cardamine hirsuta* and *Lepidium* sp. were confirmed as a separate species, probably *A. hohenheimia* (Ploch *et al.* 2010) and *A. lepidii* (Choi *et al.* 2007). The rest of the isolates were considered as *A. candida* when similar to *A. candida* at the restriction site and if the host had previously been found susceptible to *A. candida* using the ITS or *cox2* markers (Choi *et al.* 2007, 2011).

Isolates were selected to optimise the representation of genetic diversity of different *A. candida* races across Europe (Table S4.1). They were selected from various hosts and when possible, themselves collected in different countries (*C. bursa-pastoris* in France, England, Poland, Ireland, Denmark, and Scotland or *S. officinale* in France, England and Ireland). Other hosts could only be found infected in one location (*S. alba* in England or *B. nigra* in France). Furthermore, several infected populations were sampled at different time points to investigate genetic variation through time (e.g. samples #1 & 98 or #12 & 29, Table 4.1). Finally, lab isolates were included (Nc2, BoT, 2v (McMullan *et al.* 2015), 7v (Borhan *et al.* 2008) and Ex1 (unpublished, collected by E. Holub)) as well as additional isolates provided by collaborators

from India, England, Canada and The Netherlands (§, ǂ, *, ʊ in Tables 4.1 & S4.1). DNA from whole infected leaves was extracted; DNA libraries were prepared and processed through the PathSeq method as described in Chapters 2 and 3. In total, 91 *A. candida* isolates were collected and partially sequenced from 21 host species, representing 13 genera of Brassicaceae (Figure 4.1).

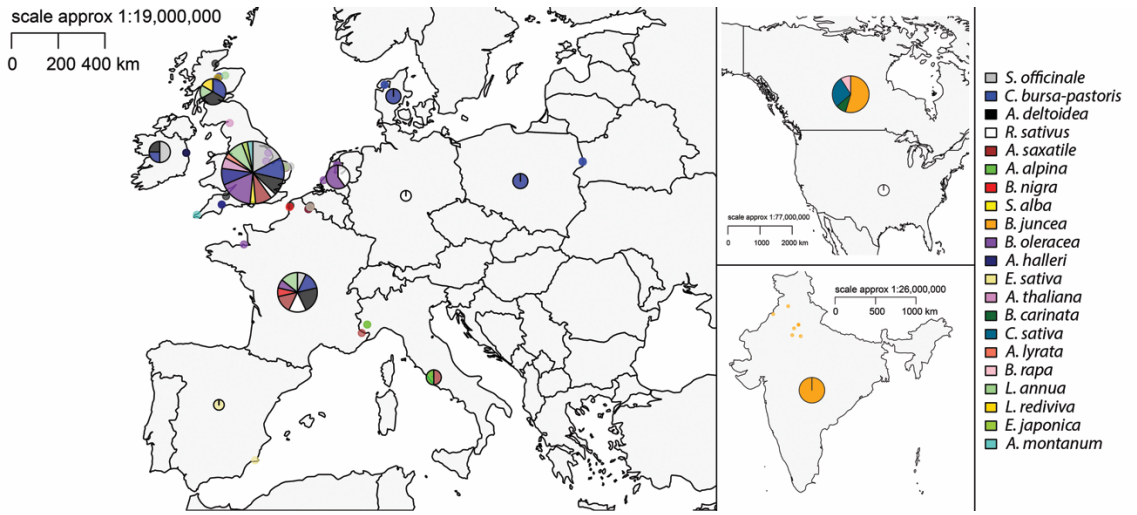


Figure 4.1 Distribution of *A. candida* isolates used in this study. Colours represent the hosts from which the isolates were collected. Dots indicate the positions of isolates as given by their GPS coordinates. When the exact location is not known, the isolate is shown at the centre of the country from which it was collected. Pie charts summarize data from each country and their sizes are scaled to the number of samples collected. A more thorough description of the samples is provided in Table S4.1.

4.2.2 THE *ALBUGO CANDIDA* COMPLEX GENETIC DIVERSITY

To evaluate genetic diversity within the *A. candida* complex, a contig of 398,508 bp was reconstructed from the selected isolates. This contig is the longest continuous sequence that could be assembled from *A. candida*, using reads from isolate Nc2 ('contig 1' in McMullan *et al.* 2015). It represents a little less than 1% of *A. candida*'s estimated genome and in subsequent analyses, we use this contig assuming it is an unbiased representation of the *A. candida* genome. Reads were aligned to the reference Nc2 contig and consensus sequences were obtained using vcfutils (a module of SAMtools, Li 2011). Sequences from samples #24, 25, 75, 85, 96, 99, 100 and 101 were discarded as read depth was too low (<10x) to allow robust variant detection (see Chapter 3, Table S3.1) and those from isolates *Ac*BoL and *Ac*Em2 were added (McMullan *et al.* 2015) so that 85 isolates were analysed in total. To assess the

phylogenetic relationship between these isolates, a maximum-likelihood tree was built in RaxML v7.7.3 (Pfeiffer & Stamatakis 2010, Figure 4.2).

Sequences showed high sequence similarity when isolates were collected on the same host or closely-related hosts, no matter the country of origin. Conversely, isolates collected on different hosts looked quite diverged. In Figure 4.2, at least 15 genetically diverged groups could be identified with high confidence (bootstrap values >70) that are host-specific to: *Capsella bursa-pastoris* and closely-related species (including *A. candida* isolates Nc2 and Em2), *Arabidopsis* spp. (including *Ac*Ex1), *Camelina sativa*, *Aubrieta deltoidea*, *Arabis alpina*, *Lunaria* spp., *Sisymbrium officinale*, *Alyssum saxatile*, *Brassica* spp. (including *Ac*BoT, *Ac*BoL and *Ac*7v), *Raphanus sativus*, *Sinapis alba*, *Brassica nigra*, *Brassica juncea* (including *Ac*2v), *Brassica carinata* and *Eruca sativa*. However, several groups could further be divided. For example, isolates collected on *C. bursa-pastoris* and related species may be separated in two groups (one with *A. candida* isolate Nc2 and the other with *Ac*Em2). Similarly, one isolate collected on *R. sativus* (#5) seems to have diverged from other isolates collected on this host.

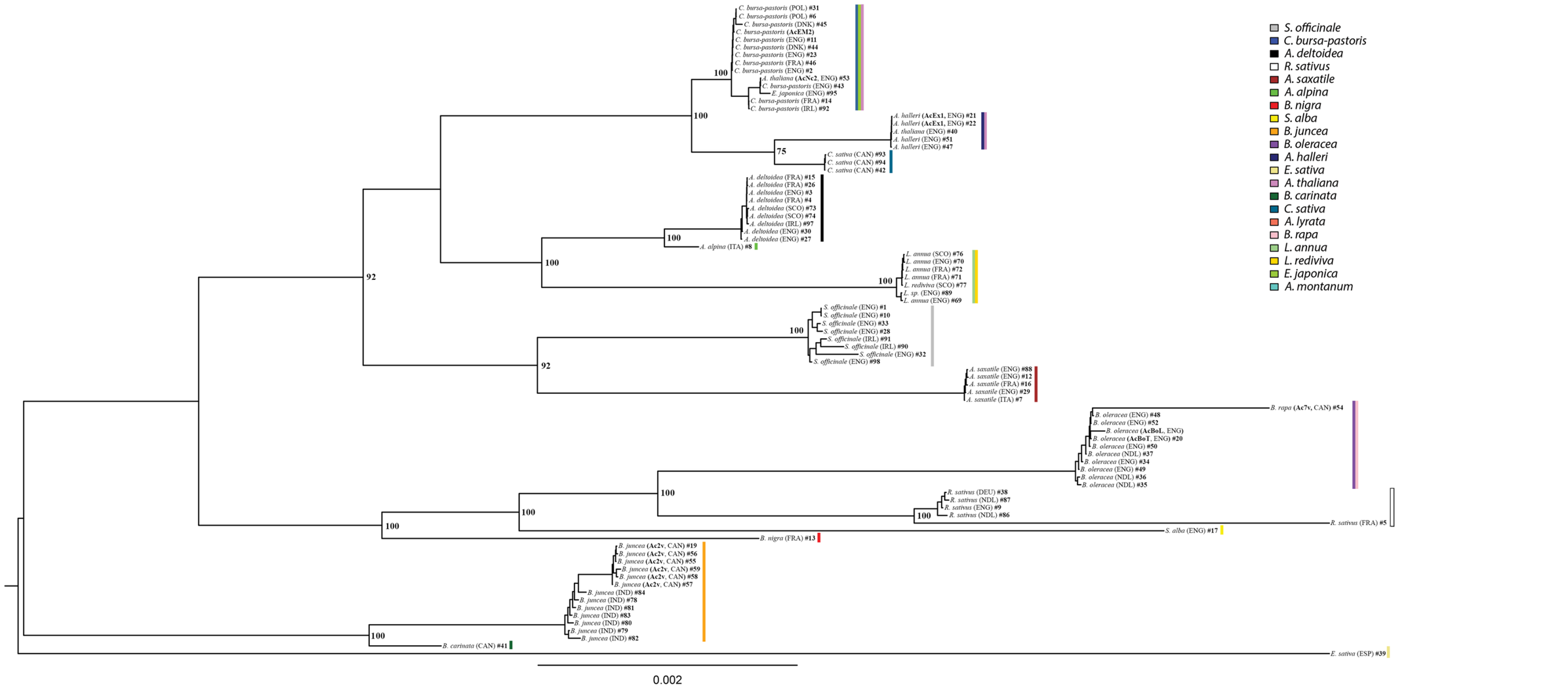


Figure 4.2 Maximum-likelihood tree based on 'contig 1' (398,508 bp), including 85 *A. candida* isolates collected from 21 hosts (one colour per host species). The tree was built with RaxML v7.7.3 using gamma distributed rate variation among sites and the GTR model of DNA evolution. Bootstrap > 70 are shown (100 replicates). The tree was viewed in Figtree v1.3.1 and rooted at midpoint. The scale represents the number of substitutions per site. Vertical bars indicate the hosts from which the clustered isolates were collected from.

Because loci may be subject to various evolutionary forces such as recombination, positive or negative selection, different gene trees may have conflicting genealogies (Maddison 1997; Degnan & Rosenberg 2009). To confirm and/or resolve the phylogeny defined above, consensus sequences from 32 diversity-tracking genes (discussed in Chapter 3) were concatenated for the 85 *A. candida* samples and a new tree was built using the same settings (Figure 4.3). In this new tree, 16 genetically diverged groups could be identified, most of which are in accordance with the previous tree: *Capsella bursa-pastoris* and related species ("I", including *A. candida* isolates Nc2 and Em2), *Arabidopsis* spp. (including *Ac*Ex1), *Camelina sativa*, *Capsella bursa-pastoris* and related species ("II"), *Brassica rapa* (including *Ac*7v), *Brassica oleracea* (including *Ac*BoT and *Ac*BoL), *Raphanus sativus*, *Aubrieta deltoidea*, *Arabis alpina*, *Sisymbrium officinale*, *Alyssum saxatile Lunaria* spp., *Brassica juncea* and *B. carinata* (including *Ac*2v), *Brassica nigra*, *Eruca sativa* and *Sinapis alba*.

Yet, certain inconsistencies should be considered. For example, isolates collected on *B. oleracea* (e.g. *Ac*BoT, *Ac*BoL) were clearly separated from *Ac*7v (mainly restricted to *B. rapa*). This is important because it has long been recognized that these isolates represent distinct pathotypes with little overlapping in their host range (discussed in Meekes *et al.* (2004)). Similarly, while *B. juncea* isolates clustered in the first tree, two groups may be identified in Figure 4.3 with isolates propagated in the laboratory for several years (#19, 55-59) on the one side and isolates collected in India this year on the other. Interestingly, the isolate collected on *B. carinata* appears closely-related to the latter group. Finally, isolates collected on *C. bursa-pastoris* and related species were divided in two groups, one of which was closely-related to isolates collected on *Arabidopsis* spp. (including *Ac*Ex1) and *C. sativa*.

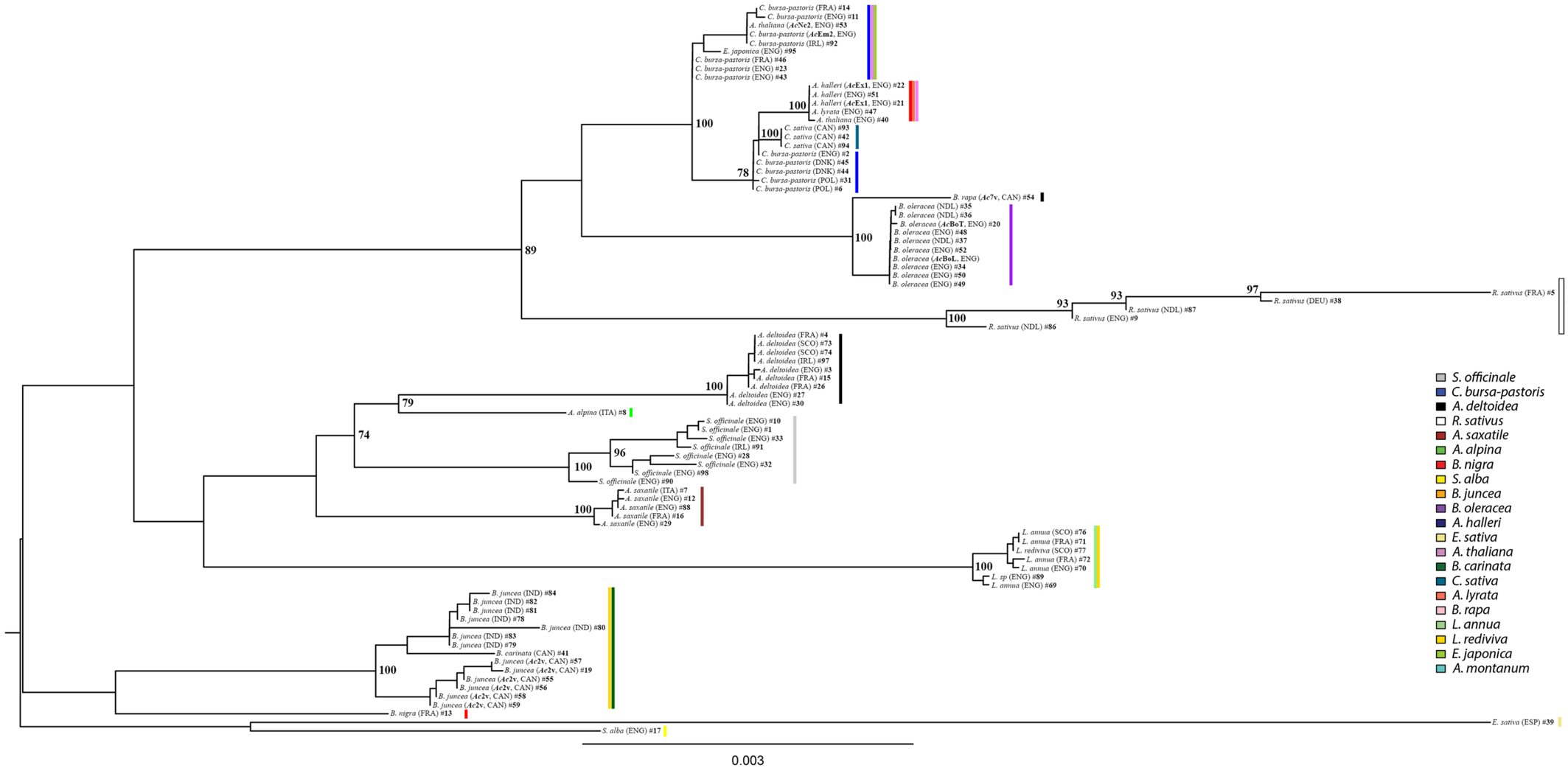


Figure 4.3 Maximum-likelihood tree based on 32 diversity-tracking loci (21,439 bp), including 85 *A. candida* isolates collected from 21 hosts (one colour per host species). The tree was built with RaxML v7.7.3 using gamma distributed rate variation among sites and the GTR model of DNA evolution. Bootstrap > 70 are shown (100 replicates). The tree was viewed in Figtree v1.3.1 and rooted at midpoint. The scale represents the number of substitutions per site. Vertical bars indicate the hosts from which the clustered isolates were collected from.

Although both trees identified similar clusters, it appears that diversity-tracking loci would allow the discrimination of *A. candida* isolates at a finer scale. This is probably due to a low selection constraint compared to 'contig 1' which facilitates the accumulation of polymorphic sites. To confirm this, the number of polymorphic sites as well as the average p-distance within and between groups defined using 'contig 1' was estimated for both trees in MEGA6 (Tamura *et al.* 2013, Table S4.2). As expected, the number of polymorphic sites was higher in the diversity-tracking loci compared to 'contig 1' (3.99% vs. 2.99%). Similarly, p-distance between groups was higher in the second tree ((mean (±SD) = 0.00757 (±0.00034) vs. 0.00655 (±0.00273)), with 162 SNPs on average out of 21,439 bp vs. ~2,610 SNPs out of 398,508 bp (paired t test: $T = 2.37$, $df = 196$, $P < 0.01$). Although not significant, p-distance was also higher within groups (mean (±SD) = 0.0003 (±0.00031) vs. 0.00017 (±0.00035)), with 6 SNPs on average out of 21,439 bp vs. ~68 SNPs out of 398,508 bp (paired t test: $T = 2.37$, $df = 196$, $P < 0.01$).

4.2.3 GENETIC EXCHANGES BETWEEN *ALBUGO CANDIDA* ISOLATES

Another major discrepancy between the two trees is the relationship between the genetically diverged host-specific races. For example, while isolates collected from *A. deltoidea* were closely-related to those collected on *C. bursa-pastoris* in the tree built using 'contig 1' (Figure 4.2), they were more closely-related to isolates collected on *S. officinale* in the tree built using diversity-tracking loci (Figure 4.3). This observation may either be explained by incomplete lineage sorting which is the random sorting of ancestral alleles into the descendant host-specific races or hybridization by secondary contact. To discriminate between the two, the R package HybridCheck (Ward & van Oosterhout 2015) was used on 'contig 1' to identify and date regions of high nucleotide identity between *A. candida* races (recombinant regions). While many mutations would have had time to accumulate in the case of incomplete lineage sorting, few mutations should be observable after recent hybridization events. In total, 159,913 inter-group pairwise recombinant regions of an average of 9,569 bp (±SD = ±13,425) were detected by HybridCheck that covered every single base pair in 'contig 1' (Figure 4.4A). These were dated from 5,798 (5-95% CI = 0-6,104) to 474,852 (5-95% CI = 383,186-579,694) generations ago (Figure 4.4B). Although the detection of some recombinant regions may be explained by incomplete lineage sorting, others appear to have occurred recently, given that only few or even no mutations are separating the races. Such putative recombinant blocks are most likely due to recent hybridization events between two races.

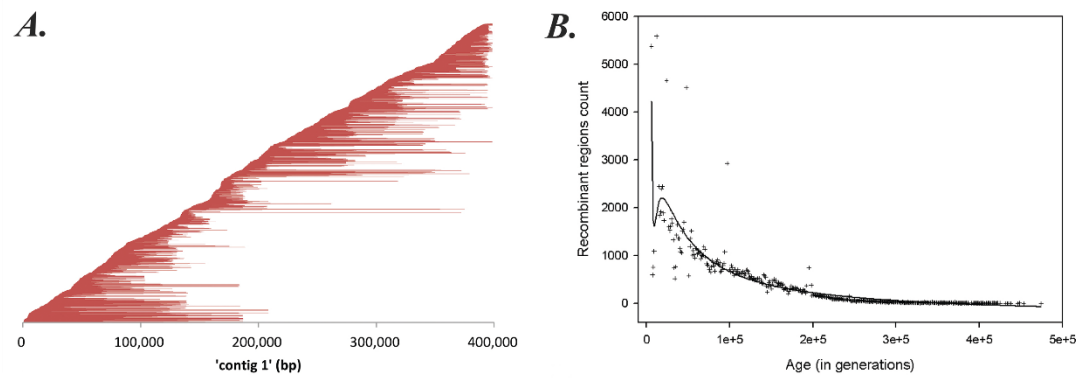


Figure 4.4 A. Inter-race recombinant regions detected by HybridCheck along 'contig 1'. These regions may be shared by more than two *A. candida* races and are shown in order, from base pair 1 to 398,508. **B. Estimated age of the recombinant regions** (the 5-95% CI is not provided here). The count of recombinant regions is estimated by sets of 1,000 generations. The solid line is a polynomial inverse third order regression line ($R^2 = 0.72$).

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.4.11.3 (Huson & Bryant 2006; Figure 4.5). Although based on 'contig 1' alone, the network appeared to be a combination of both the trees shown above confirming some of the divisions suggested in Figure 4.3. In total, 18-20 groups were defined by SplitsTree that are highly congruent with the host species isolates were collected on: *Alyssum saxatile*, *Sisymbrium officinale*, *Arabidopsis* spp. (including *Ac*Ex1), *Camelina sativa*, *Eutrema japonica*, *Capsella bursa-pastoris* and related species (including *Ac*Nc2 and *Ac*Em2), *Aubrieta deltoidea*, *Arabis alpina*, *Lunaria annua*, *Lunaria rediviva*, *Brassica nigra*, *Sinapis alba*, *Raphanus sativus* (probably two groups), *Brassica rapa* (including *Ac*7v), *Brassica oleracea* (including *Ac*BoL and *Ac*BoT), *Brassica juncea* (probably two groups: one with *Ac*2v propagated in the lab for several years, the other with wild Indian isolates), *Brassica carinata* and *Eruca sativa* which is the most diverged race as in the two trees above.

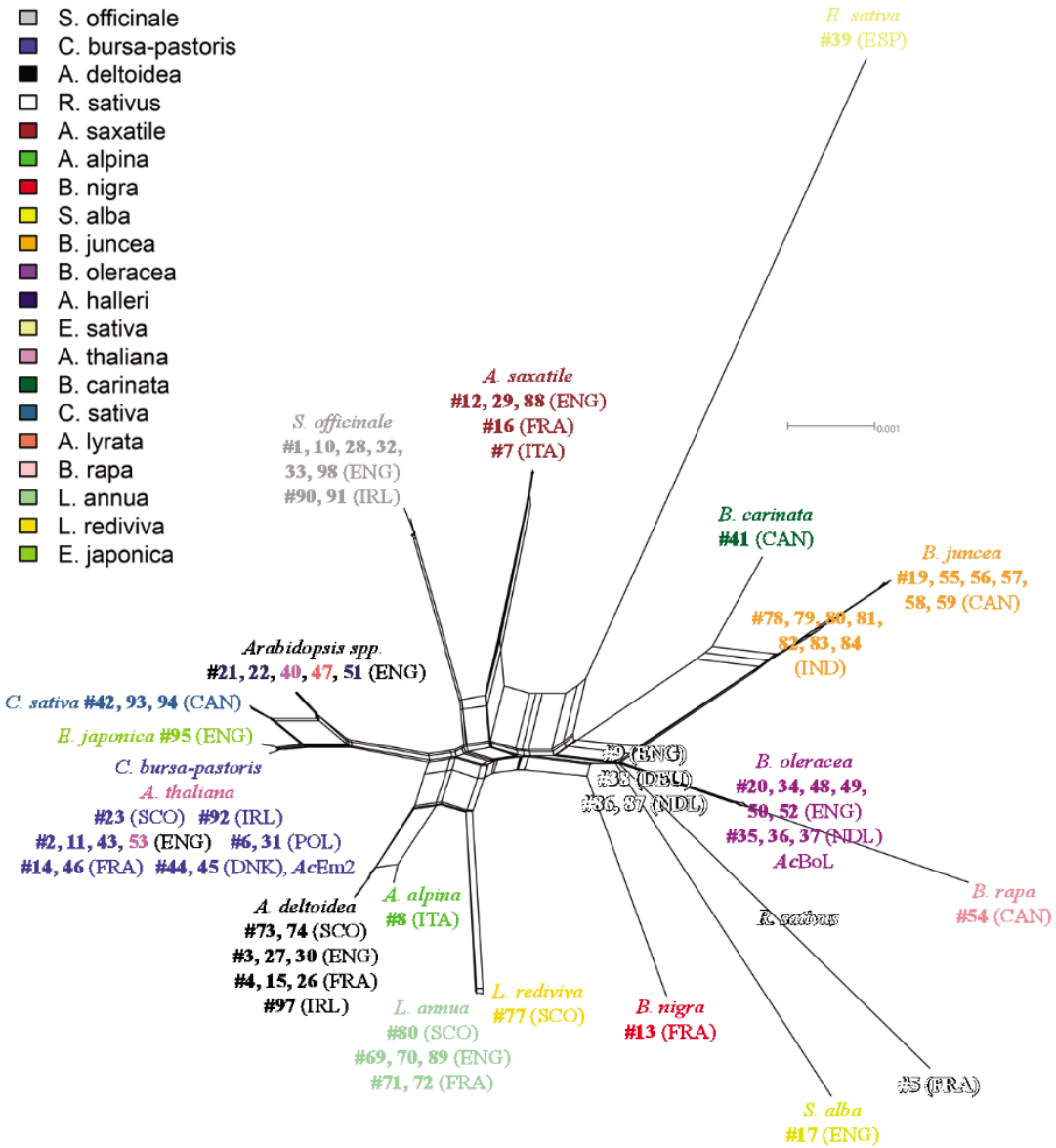


Figure 4.5 Neighbor-Net network built using SplitsTree v.4.11.3 and based on the 'contig 1' sequence (398,508 bp) of 85 *A. candida* isolates. Most bootstrap values were > 70 but are not shown for the purpose of clarity (500 replicates). Ambiguous sites are ignored. The scale represents distances estimated using the uncorrected p-distance. The isolates numbers are indicated with "#" signs as in Tables S3.1 & S4.1 and isolates are colour-coded according to the hosts they were collected on.

These groups had low levels of nucleotide diversity (Table S4.2), suggesting that races are mostly clonal. However, some within-race recombinant regions (462) could be detected of an average length of ~89,636 bp. These regions could only be detected for isolates collected on *B. juncea*, *B. oleracea*, *S. officinale* and *R. sativus* and were dated to be quite recent (mean (5-95% CI) = 6,156 generations ago (582-7,492)), suggesting either occasional sex within these races or the accumulation of mutations in some isolates in otherwise genetically low diverged races.

4.3 DISCUSSION

Until recently, analyses of *A. candida* genetic diversity were mainly restricted to the ITS and *cox2* markers. Although these markers were useful to distinguish between the various *Albugo* species, they revealed little to no genetic diversity within the *A. candida* complex (Thines & Spring 2005; Choi *et al.* 2006, 2007, 2011). Yet, it has long been recognized that *A. candida* is organized into multiple host-specific pathotypes or "races" (Hiura 1930; Pound & Williams 1963; Hill *et al.* 1988; Meena *et al.* 2014). However, it is only in 2015 that several isolates were fully sequenced, revealing three genetically diverged host-specific races that showed signature of historical exchanges of genetic material (McMullan *et al.*).

Building on the 2015 study, I set out to collect *A. candida* isolates at various locations, different time points and from multiple hosts with the aim of further exploring the mechanisms by which pathotypes evolve. In total, 91 isolates were selected and sequenced using the PathSeq method, representing 21 plant species from 13 genera of Brassicaceae (although only 83 isolates could be sequenced, with the addition of two isolates from the lab, *Ac*BoL and *Ac*Em2). In this chapter, I was particularly interested in the number of genetically diverged races that can be identified in the wild at a spatial scale across Europe and at a temporal scale, in samples collected in three years (2013-2015). In addition, I aimed to quantify the within-race genetic diversity as well as the incidence of sexual reproduction within and between these races.

To investigate the genetic structure of the selected isolates, phylogenetic analyses were performed using both 'contig 1' and a set of 32 concatenated diversity-tracking loci. Because there can be incompatibilities between gene trees, this approach is more likely to allow an accurate reconstruction of the "species tree" (assuming this exists), and it enables the identification of processes such as hybridization and incomplete lineage sorting (Degnan & Rosenberg 2009; Jouet *et al.* 2015). The diversity-tracking loci were significantly more polymorphic which made them more suitable to track recent evolutionary events. Nevertheless, both trees were remarkably congruent (despite the high incidence of hybridization in *A. candida*) and at least 16 genetically diverged groups could be identified that are host-specific to: *Capsella bursa-pastoris* and closely-related species (including *A. candida* isolates Nc2 and Em2), *Arabidopsis* spp. (including *Ac*Ex1), *Camelina sativa*, *Aubrieta deltoidea*, *Arabis alpina*, *Lunaria* spp., *Sisymbrium officinale*, *Alyssum saxatile*, *Brassica oleracea* (including *Ac*BoT, *Ac*BoL), *Brassica rapa* (including *Ac*7v), *Raphanus sativus*, *Sinapis alba*, *Brassica*

nigra, *Brassica juncea* (including *Ac*2v), *Brassica carinata* and *Eruca sativa*. I argue that these groups represent distinct and diverged *A. candida* races as the average inter-group nucleotide diversity was estimated between ~0.65 to 0.76% polymorphic sites which is comparable to the ~1% found in McMullan *et al.* (2015). Within each race, the isolates were diverged by less than 0.03% (0.017% at 'contig 1' and 0.03% at the diversity-tracking genes). It is also likely that isolates collected on *R. sativus* and *B. juncea* could further be divided as nucleotide diversity within these groups was estimated up to ~0.1% (Table S4.2). In any case, although there appears to be a strong genetic support for host races, host specificity should be tested to allow formal descriptions of pathotypes.

Although genetically diverged, these races did not seem to be evolving independently from one another as many regions of high nucleotide identity were detected between races (recombinant regions). Some of these regions are probably due to incomplete lineage sorting or trans-species polymorphism, i.e. the conservation of similar orthologous alleles over evolutionary time, since the divergence of the two races (Klein *et al.* 1998). However, other regions in which two races have near-identical nucleotides are most likely due to recent hybridization events. An example of such an event is the shared 26,647 bp region that has only two polymorphic sites (0.0075%) between isolates collected on *S. officinale* and *A. alpina* (#28 and #8) even though isolates were found to be ~0.37% diverged at 'contig 1'. In that context, hybridization may allow *A. candida* to rapidly generate polymorphisms so as to keep up with the host evolution and/or colonize new hosts (Stukenbrock & McDonald 2008; Fisher *et al.* 2012). Indeed, *A. candida* harbors considerable genetic variation across its constituent genepools of ecologically distinct host races. The exchange of genetic material between these races could be an important driver of adaptive evolution in *A. candida*. Theoretically, hybridisation is thought to be maladaptive and unlikely to introduce useful genetic variation (Castric *et al.* 2008), because the genetic background into which the foreign genes are introgressed is probably already well adapted (Rieseberg *et al.* 1995). In addition, Dobzhansky-Muller incompatibilities due to negative epistatic interactions are predicted to evolve over time, resulting in the loss of hybrid fitness and infertility (Orr & Turelli 2001)**.** On the other hand, genetic introgression can provide a source of novel alleles that have already been tried and tested by natural selection (Seehausen 2004). The inflow of this variation after hybridisation allows recombination to instantly generate an almost infinite number of novel genotypes. Like in other plant pathogens such as *Ophiostoma ulmi* (Brasier 2001), *Heterobasidion annosum* (Gonthier *et al.* 2007) or *Verticillium longisporum* (Inderbitzin *et al.* 2011), the rate of adaptive

evolution of *A. candida* may be expedited, allowing this pathogen to colonise multiple distinct host plants across three different families.

Although I provide evidence for sexual reproduction between *A. candida* races, hypothesizing about the prevalence of sex within races proves more difficult, particularly because *A. candida* within-race genetic diversity is low. On the one hand, the detection of recombinant regions would support the hypothesis that gene flow is frequent between populations, at least in some races. However, as in the case of inter-group recombination, detection of such regions may also be due to common ancestry where mutations have not yet had time to accumulate. On the other hand, the observation that isolates collected several years apart are identical at 'contig 1' seems to support the idea that asexual or clonal reproduction is dominant. Indeed, isolate Nc2 collected in England in 2007 (#53) is indistinguishable from isolates #11, 14, 23, 44 & 45 collected between 2013 and 2015 in England, France, Scotland and Denmark. Perhaps, both hypotheses need to be considered equally and the prevalence of sexual over asexual reproduction is dependent on the race of *A. candida*. For example, while isolates collected on *A. saxatile* are highly similar at 'contig 1' (little sex), those collected on *S. officinale* are more diverged even though they were sampled in fewer countries (more sex, see Table S4.2). This is further supported by the fact that recombinant regions could be detected between isolates collected on *S. officinale* but not on *A. saxatile*. It is also possible that some races outcross while others self-fertilize, depending on the abundance of the host species in Europe.

In McMullan *et al.* (2015), the percentage of heterozygous sites shared between isolates of a same race was used to infer clonality. The rationale behind this is that high levels of shared heterozygosity would be removed within few or even just one generation of sexual reproduction. Unfortunately, one important limitation of the phylogenetic, recombination and nucleotide diversity analyses performed above is that heterozygous sites (UIPAC codes) are not treated to their fullest potential, which can result in polymorphisms being missed or ignored. Importantly, throughout this thesis, heterozygosity is calculated as the proportion of heterozygous sites within loci and individuals. This will be investigated in the next chapter.

CHAPTER 5: HETEROZYGOSITY IN *ALBUGO CANDIDA* PATHOTYPES

5.1 INTRODUCTION

As described in the general introduction, sexual reproduction has many advantages including the generation of novel allelic combinations or the elimination of deleterious mutations. However, sexual reproduction is also costly. For example, outcrossing individuals need to allocate time and energy for meiosis as well as find a suitable mating partner to eventually pass on only 50% of their genes to the progenies. In addition, sexual reproduction may break favourable combinations of alleles that have been shaped by selection over long evolutionary time (Heitman 2006; Otto 2009; Billiard *et al.* 2012; Seidl & Thomma 2014).

To have "the best of both worlds" (Ellison *et al.* 2011), some organisms evolved a mixed system where cycles of sexual and asexual reproduction intersperse (Heitman 2006, 2010), including many microbial pathogens such as *Alternaria brassicicola* (Bock *et al.* 2005), *Aphanomyces euteiches* (Grünwald & Hoheisel 2006) and *Phytophthora infestans* (Danies *et al.* 2014). Under these conditions, pathogens may benefit from the advantages of sex but also from those of clonal reproduction, including the rapid demographic expansion of well-adapted genotypes (Fisher *et al.* 2012). This has been shown for example in *Phytophthora infestans* where sexual reproduction between lineages US-6 and US-7 have led to the emergence of a novel lineage US-11 that have clonally expanded since (Gavino *et al.* 2000).

In *A. candida*, a mixed reproductive system has also been described (Saharan & Verma 1992). In this system, the hyphae of *A. candida* may differentiate in the host intercellular space into sporangiophores to produce asexual zoospores, or into sexual organs (the female oogonium and the male antheridum) to form sexual oospores. However, the relative importance of both modes of reproduction in nature is still largely unknown, probably due to the lack of molecular data from *A. candida* wild populations. It is also not known whether *A. candida* can self-fertilize (homothallism, the oomycete can initiate the production of gametes in the absence of a compatible mating partner) or if it can only outcross or both (homothallism or heterothallism, gametes production cannot be initiated in the absence of a compatible mating partner, Billiard *et al.* 2012).

In Chapter 4, I provided evidence for genetic exchanges between *A. candida* pathotypes, confirming the findings of McMullan *et al.* (2015). Unfortunately, the performed analyses did not allow me to conclude on the importance of sexual reproduction within pathotypes. This

was mainly due to heterozygous sites being fully or partially ignored and low nucleotide diversity within races. Therefore, in this chapter, information from heterozygous sites is incorporated to evaluate the incidence of gene flow within *A. candida* races. This will be particularly interesting because although McMullan *et al.* (2015) concluded on a clonal propagation of *A. candida* races, sexual oospores are often being observed in the hypertrophied parts of the hosts, at least for some races (Liu & Rimmer 1993; Choi *et al.* 2008; Meena & Sharma 2012). Oospores are also thought to be crucial for survival in-between host seasons (Saharan *et al.* 2014). Furthermore, it can be expected that races that are adapted to cultivated hosts with low genotypic diversity (*B. oleracea*, *B. juncea* or *R. sativus*) have lower rates of sexual reproduction compared to races adapted to genetically diverse wild host populations (*C. bursa-pastoris*, *S. officinale*, *A. alpina*).

In this chapter, I also use heterozygous sites to investigate whether I can detect mixed infections with two host-specific *A. candida* races. This would provide further confirmation that races can co-occur in the wild to potentially hybridize, as was hypothesized in McMullan *et al.* (2015). Additionally, I use a method published by Yoshida *et al.* (2013) to verify the ploidy level of *A. candida* races. Indeed, while *A. candida* is often presented as a diploid organism, only Sansome & Sansome (1974) investigated the cytology of *A. candida* and reported hexa- or octoploidy (races collected on *C. bursa-pastoris* and *L. annua*). Polyploidy may provide an evolutionary advantage to some *A. candida* races (see Chapter 1, section 1.1.4) and it has already been observed in other oomycetes such as certain lineages of *P. infestans* (Yoshida *et al.* 2013). Finally, I investigate potential loss-of-heterozygosity events in *A. candida*. This has recently been identified in *P. capsici* (Lamour *et al.* 2012a) and may represent an important evolutionary process for the rapid evolution of both sexual and asexual oomycetes.

5.2 RESULTS

5.2.1 HETEROZYGOSITY IN *ALBUGO CANDIDA* PATHOTYPES

The proportion of heterozygous sites within *A. candida* races was evaluated at 'contig 1' using vcftools v. 0.1.10 (Danecek *et al.* 2011). Bases with low Phred scores were discarded from the analysis (<100) and samples with more than 11% missing bases were not taken into account. Mean heterozygosity expressed as the percentage of observed heterozygous sites was

then estimated for each race defined in Chapter 4 (Figure 5.1). According to previous phylogenetic analyses, isolates from *B. juncea* and *R. sativus* were also divided into two groups (see Figure 4.5). In total, heterozygous sites from 71 *A. candida* samples were analysed (See Table S4.1).

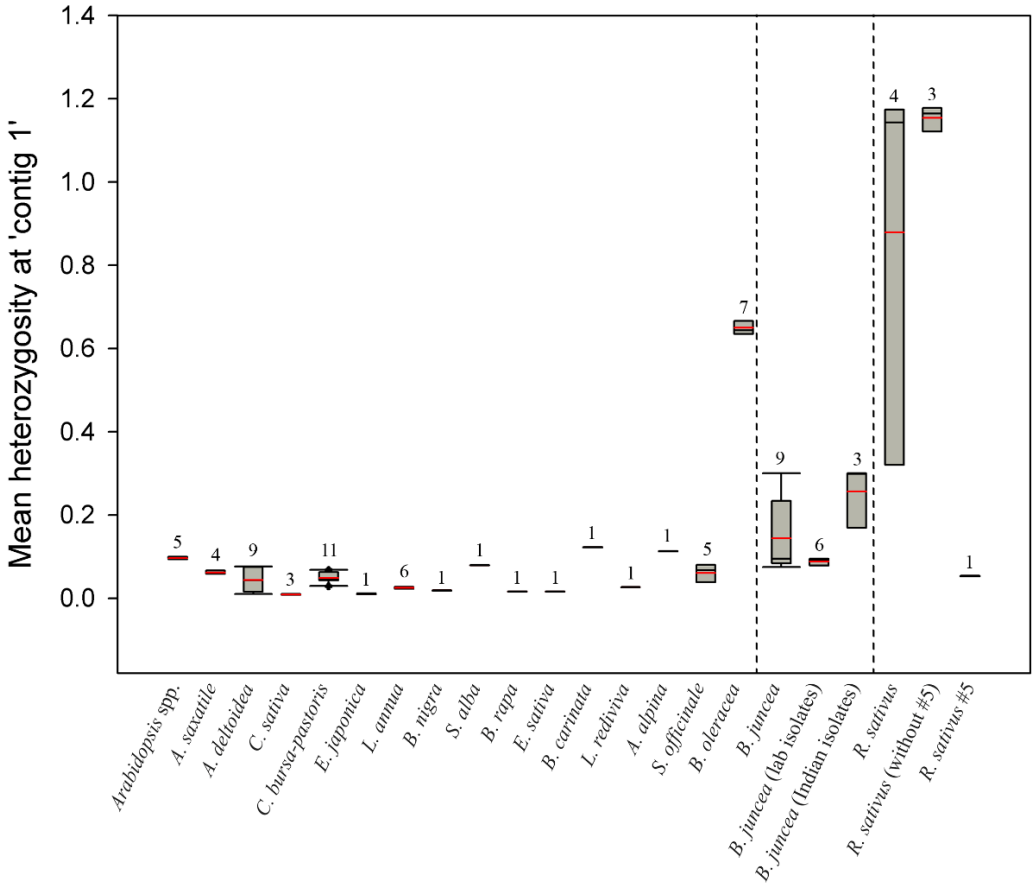


Figure 5.1 Mean heterozygosity of *A. candida* pathotypes at 'contig 1' (398,508 bp). Heterozygosity is expressed as the percentage of observed heterozygous sites. Median (black solid line) and mean (red solid line) heterozygosity are provided for each of the 18-20 *A. candida* races defined in Chapter 4 by SplitsTree (see Figure 4.5). The number of samples per race used in this analysis is provided on top. *"C. bursa-pastoris*" includes isolates collected from *C. bursa-pastoris* and one isolate collected on *A. thaliana* (#53). Isolates collected on *B. juncea* and *R. sativus* (right side of the dotted lines) are shown as one group (first bar) or as two groups as suggested by previous phylogenetic analyses.

Heterozygosity was variable within races, especially for isolates collected on *A. deltoidea* and *S. officinale* (coefficient of variation CV = 66.52 and 47.63%, respectively; see Figure 5.1 and Table S5.1). Although it also appeared to be variable in isolates collected on *B. juncea* and *R. sativus*, variation in heterozygosity dramatically reduced when both groups were separated into two, further supporting the existence of two *A. candida* races or populations on

these hosts (CV from 63.93 for *B. juncea* to 9.96 (*B. juncea* lab isolates) and 29.34% (*B. juncea* India) and from 62.66% for *R. sativus* to 2.54% (*R. sativus* without #5)). Similarly, there were large discrepancies between races and while most isolates had a low number of heterozygous sites (mean heterozygosity (±SD) = 0.009 (±0.00118) for isolates collected on *C. sativa* to 0.113 (*n.a.*) on *A. alpina*), others appeared much more heterozygous (mean heterozygosity (±SD) = 0.65 (± 0.016) for isolates collected on *B. oleracea* and 1.15 (±0.029) on *R. sativus* (without #5)).

In the next sections of this chapter, I test four non-mutually exclusive hypotheses for the observed variation in heterozygosity within and between *A. candida* races: (i) highly heterozygous *A. candida* races may be polyploid; (ii) multiple isolates may co-occur in several samples, falsely inflating the number of heterozygous sites; (iii) the incidence of asexual and sexual reproduction may vary between *A. candida* races and/or populations and (iv) loss-of-heterozygosity may occur in *A. candida*. This mechanism has already been proposed in McMullan *et al.* (2015) where they found that *Ac*BoT and *Ac*BoL (two isolates collected on *B. oleracea*) were highly heterozygous compared to *Ac*2v (collected on *B. juncea*), *Ac*Em2 (*C. bursa-pastoris*) and *Ac*Nc2 (*A. thaliana*).

5.2.2 PLOIDY IN *A. CANDIDA* RACES

To evaluate ploidy in *A. candida*, the proportion of reads per SNP at heterozygous sites was investigated using the same method as in Yoshida *et al.* (2013). The rationale behind this method is that, for diploid organisms, each bi-allelic SNP should account for ~50% of the reads. Conversely, ~33 and 67% of reads for each SNP may be observed in triploid organisms and both 50-50 and 25-75% of reads in tetraploids. This analysis was performed for each isolate with less than 11% missing sites (71 isolates in total, see Table S4.1) with the help of Diane Saunders (TGAC, Norwich, UK) who provided scripts used in Yoshida *et al.* (2013) as well as guidance. Distributions of the per-SNP read proportions (or ploidy graphs) were built using R v. 3.1.2 (R Core Team 2014) and are provided in Figures 5.2 and S5.1.

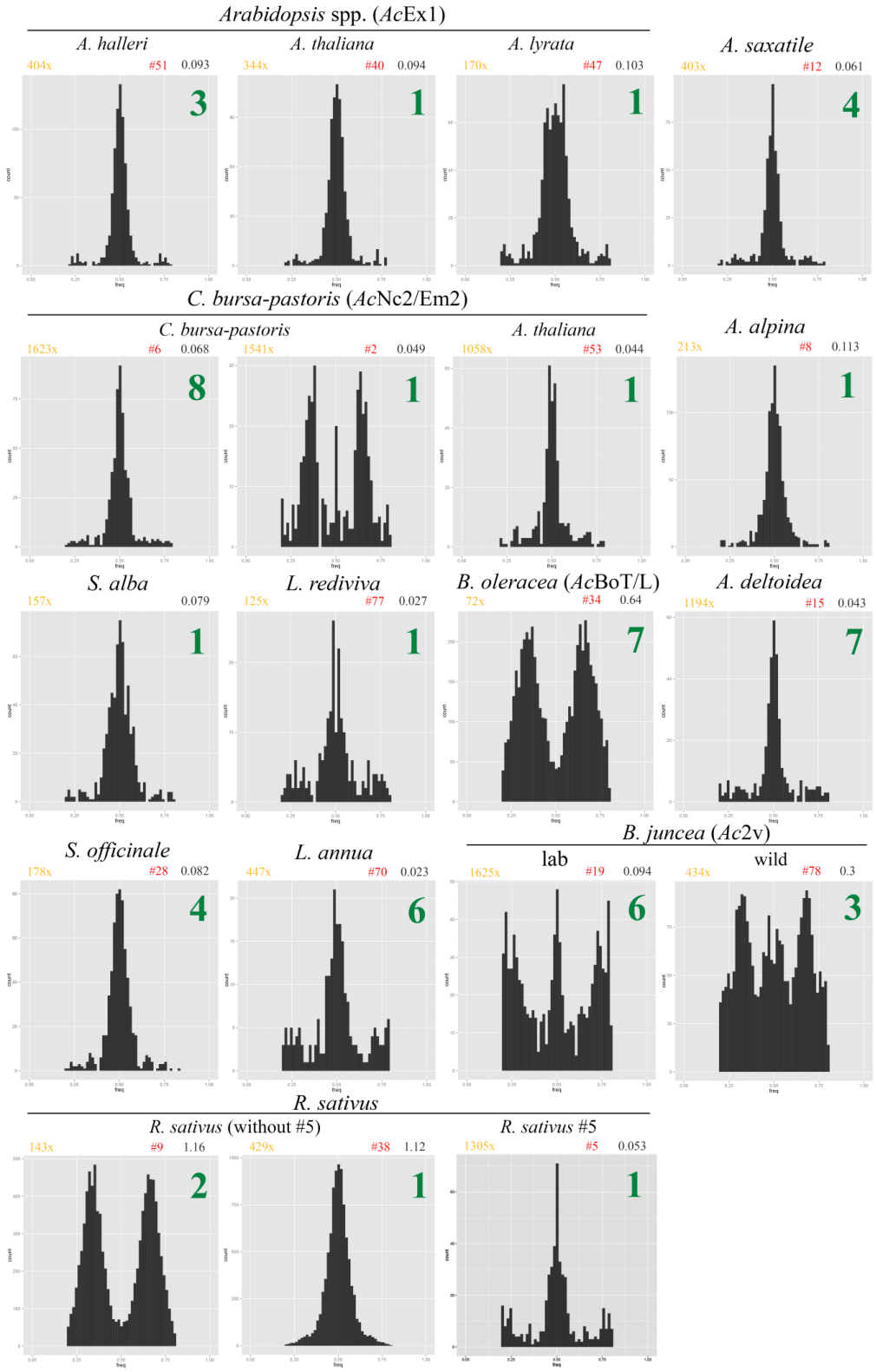


Figure 5.2 Ploidy graphs of *A. candida* wild isolates based on 'contig 1' (398,508 bp). The x-axis represents the proportion of reads per SNP at heterozygous positions and the y-axis is the count of heterozygous sites. The hosts from which isolates were collected are provided at the top of each graph. Average depth at 'contig 1' is provided in yellow, sample number in red (see Tables S3.1 and S4.1) and per-individual heterozygosity in black. In green is the number of isolates with similar distributions.

While most distributions clearly indicated the presence of diploid organisms (isolates collected on *Arabidopsis* spp., *A. saxatile*, *S. alba*, *L. annua*, *L. rediviva*, *A. deltoidea, S. officinale, A. alpina, R. sativus* #5), the observation of triploid- and tetraploid-looking distributions suggested that several *A. candida* isolates have undergone polyploidization. Interestingly, polyploidy appeared to either be restricted to a few isolates within a race (*C. bursa-pastoris* isolate #2) or to be common, if not general, in other races (all isolates collected on *B. oleracea* and *B. juncea* and most on *R. sativus* (without #5)), suggesting that although selected for only occasionally, polyploidization in *A. candida* may be frequent.

To confirm these results, the same workflow was repeated on whole-genome data. To do this, reads from previously sequenced lab isolates *Ac*Nc2 (collected on *A. thaliana*), *Ac*Em2 (*C. bursa-pastoris*), *Ac*Ex1 (*A. halleri*), *Ac*BoT and *Ac*BoL (*B. oleracea*), *Ac*7v (*B. rapa*) and *Ac*2v (*B. juncea*) were mapped to race *Ac*Nc2 (35,029,411 bp) and new distributions were built (Figure 5.3). Probably due to low-quality sequencing data, ploidy level in isolate *Ac*Nc2 could not be confirmed. However, diploidy in isolates collected on *C. bursa-pastoris* (*Ac*Em2) and *A. halleri* (*Ac*Ex1) and more importantly, polyploidy on *B. oleracea* (*Ac*BoT and *Ac*BoL, triploidy) and *B. juncea* (*Ac*2v, tetraploidy) were validated by this new analysis, leading to the first report of polyploidy in *A. candida*.

Unfortunately, while whole genome data seemed to also support tetraploidy in isolate *Ac*7v (*B. rapa*), analysis using 'contig 1' proved inconclusive (Figure S5.1). Ploidy level could also not be determined for several wild isolates using 'contig 1' (from *E. sativa*, *B. nigra*, *B. rapa*, *C. sativa*, *E. japonica*, *B. carinata* as well as *A. candida* ex *A. deltoidea* #73 & 74 and *S. officinale* #32, see Figure S5.1) but interestingly, these isolates had low heterozygosity compared to the others (two sample t-test: $T = -4.84$, $p > 0.0001$, $df = 60$).

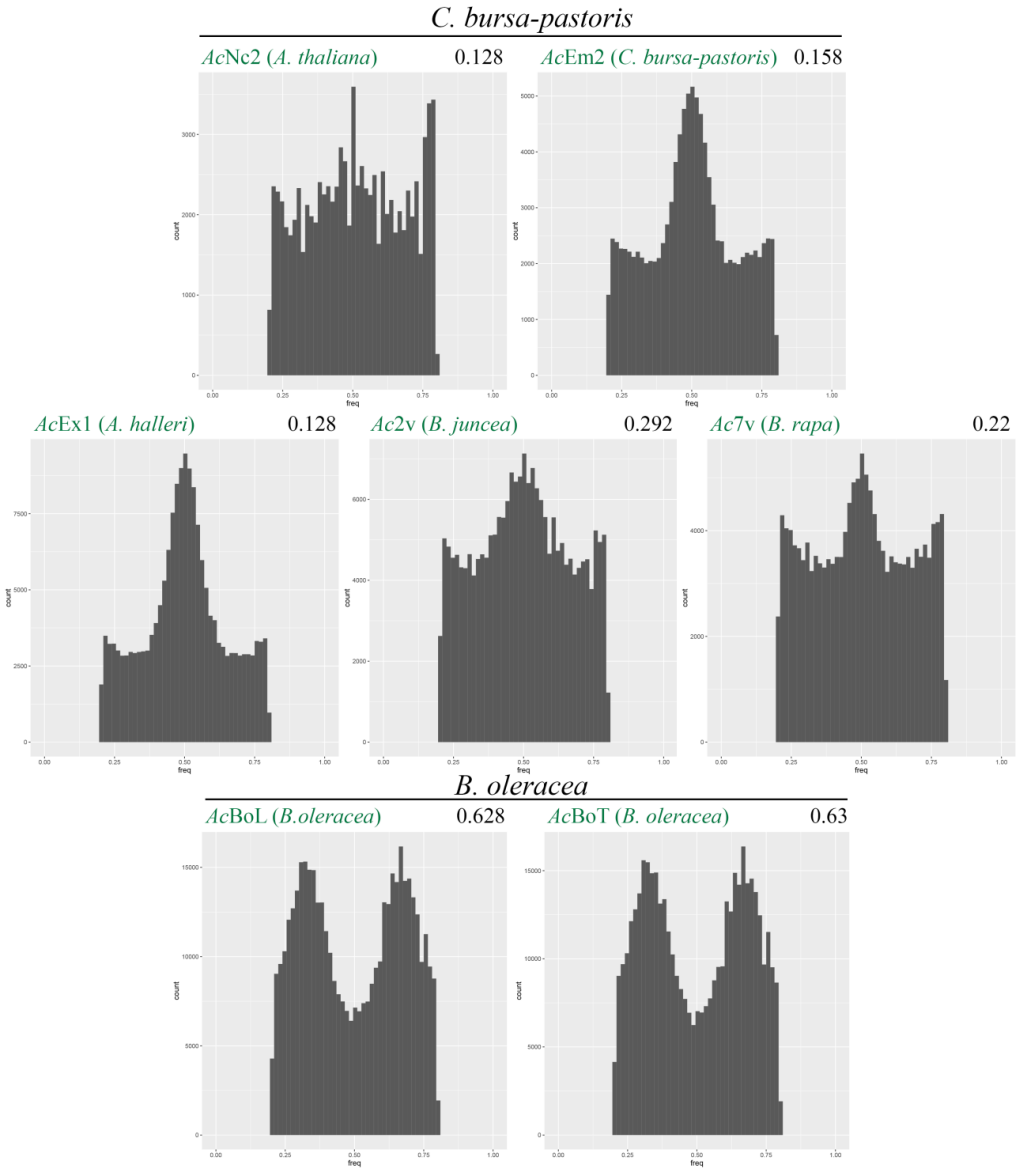


Figure 5.3 Ploidy graphs of *A. candida* lab isolates using whole genome data (35,029,411 bp). These represent five genetically diverged races. The x-axis is the proportion of reads per SNP and the y-axis is the count of heterozygous sites. The name of the isolate is provided in green as well as the host from which it was originally collected. Heterozygosity is provided in black.

To investigate this further, heterozygous sites in isolates which ploidy could not be determined were plotted along 'contig 1' and their distribution was examined (Figure 5.4). Surprisingly, 89.36% of the heterozygous sites in isolate #41 (*B. carinata*) were located between base positions 184,000-191,050 (representing just 1.77% of all base pairs in the analysis). Similarly, in the other isolates 76.46% (±SD = 16.94) of the heterozygous sites were located at the very beginning of 'contig 1', between base positions 1 to 900 (0.226%). In both cases, high heterozygosity at these loci was correlated with an increase of the read depth,

suggesting that these regions are duplicated in *A. candida* (mean read depth = 816 (1-900 bp) vs. 454 (901-398,508 bp), paired t-test: T = 4.10, p = 0.001; on *B. carinata*: mean read depth (±SD) = 107 (±75.5) vs. 75 (±23.1), two sample t-test: T = 35.17, p < 0.0001, df = 7073). Therefore, I argue that due to homology with 'contig 1', untargeted loci were captured by the baits and then sequenced. By forcedly mapping these off-target reads to 'contig 1', per-SNP read proportion at heterozygous sites was probably altered in such a way that it resembled that of a tetraploid (i.e. a double diploid). The signal of ploidy was further obscured by an overall low heterozygosity at 'contig 1', diminishing the number of data points and hence robustness of the analysis.

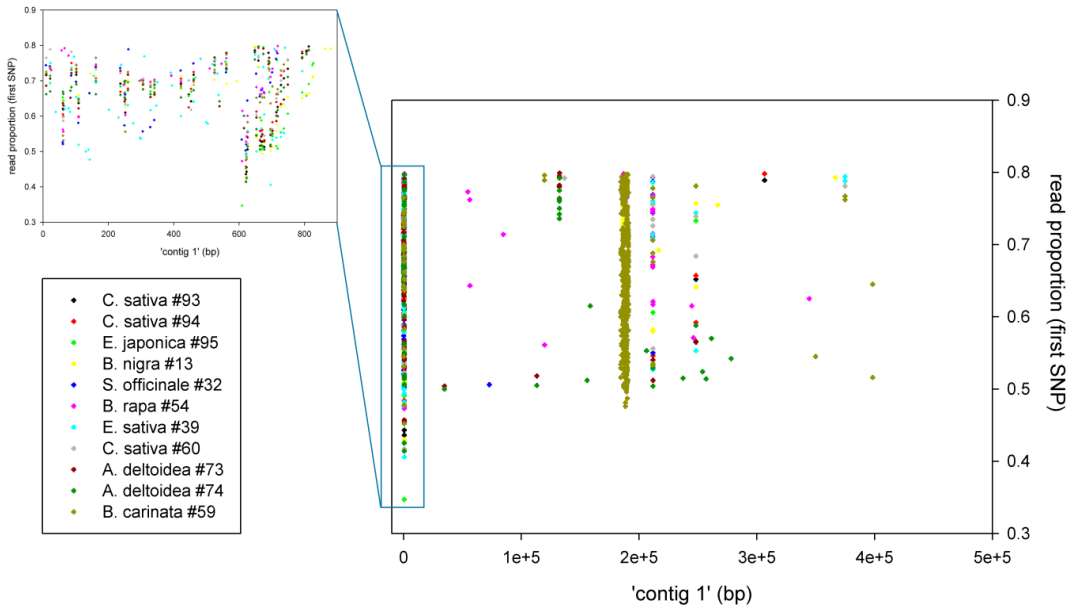


Figure 5.4 Heterozygous sites in *A. candida* isolates for which the ploidy level could not be determined ('contig 1'). Base positions in 'contig 1' are on the x-axis and read proportions of the most abundant SNP per site are on the y-axis. In diploids, each SNP should represent about 50% of the reads (0.5 on the y-axis). Although tetraploids may also have 50% reads of each SNP (0.5), proportions of ~25-75% are also observed (0.75 on the y-axis for the most abundant SNP). A close-up of heterozygous sites in the first 900 bp of 'contig 1' is provided on the left. Data for ten isolates is shown; the hosts from which they were collected as well as their names are given in the legend (bottom right, see Tables S3.1 and S4.1 for the samples' names).

Despite the technical challenges in establishing ploidy level for some isolates, variation in ploidy seemed to play a crucial role in shaping the genetic diversity of *A. candida* isolates. As a matter of fact, although it is possible that heterozygosity was underestimated at 'contig 1', compared to that of the whole genome (Table 5.1; paired t test: T = -2.13, df = 4, P = 0.05), heterozygosity was lowest in diploids (mean heterozygosity (±SD) = 0.071 (±0.021)), highest

in triploids (0.766 (±0.077)) while it was intermediate in tetraploids (0.135 (±0.029)). Yet, the existence of multiple ploidy levels in *A. candida* still cannot explain variation in heterozygosity within races. With this in mind, the remainder of the hypotheses enunciated above will be tested (mixed infections, differences in the reproductive mode between races and loss-of-heterozygosity).

A. candida isolates	Heterozygosity at 'contig 1'	Whole genome heterozygosity
AcNc2 (#40)	**0.044**	**0.128**
AcEm2	-	0.158
AcEx1 (#21, 22)	**0.0955***	**0.128**
AcBoT (#20)	**0.66**	**0.63**
AcBoL	-	0.628
Ac2v (lab, #19, 55-59)	**0.0883***	**0.292**
Wild "Ac2v" (#78, 83, 84)	0.256*	-
Ac7v (#38)	**0.0166**	**0.22**

Table S5.2 Comparison of the heterozygosity observed in 'contig 1' vs. the whole genome. Heterozygosity in 'contig 1' was estimated based on isolates that have been cultivated in the laboratory for several years and those that are highly similar but which were collected in the wild. Whole genome heterozygosity was estimated for isolates which whole genome was available. A paired t-test was performed to test for a difference between heterozygosity at 'contig 1' and based on the whole genome when both data were available (in bold; H0: no difference, H1: heterozygosity at 'contig 1' is underestimated; paired t test: $T = -2.13$, $df = 4$, $P = 0.05$). A star indicates where heterozygosity is based on multiple isolates.

5.2.3 CO-OCCURRENCE OF *A. CANDIDA* ISOLATES IN THE WILD

In Chapter 4, I argue that host-specific races of *A. candida* have reproduced sexually in the past and that it is likely that they still reproduce occasionally. However, although scientists have shown that two *A. candida* races can co-occur on the same plant in the laboratory (McMullan *et al.* 2015), they have yet to demonstrate that this can happen in the field. With this in mind, I compared the ploidy graphs of 71 wild *A. candida* samples (see Figure 5.2 and Table S4.1) with those of simulated mixed infections. To do this, reads from diverged wild samples were merged according to six combinations: two diploids (#72 and 97), two triploids (#37 and 87), two tetraploids (#78 and 84), one diploid and one triploid (#97 and 87), one diploid and one tetraploid (#97 and 84) and finally, one triploid and one tetraploid (#87 and 84). Ploidy graphs were then built using the same method as described above (Figure 5.5) and in Yoshida *et al.* (2013). Here, I hypothesize that the co-occurrence of two *A. candida* isolates

has an impact on both the number of observed heterozygous sites and the proportion of reads per SNP at those sites.

Surprisingly, it is not always possible to visually separate mixed infections from single isolates although mixed infections with two polyploids does tend to increase the number of peaks (see Figure 5.5). Importantly however, heterozygosity was largely increased when two isolates co-occurred, no matter the ploidy level (see Figure 5.5 (in red) for estimation of heterozygosity at 'contig 1' in both single and mixed isolates; paired t-test: $T = 3.66$, $p = 0.007$, average percentage increase (±SD) = 846.5% (±1,371.4), minimum = 11.6%, maximum = 3,515.7%). However, no such variation could be observed between isolates within a race in the present study (average percentage increase (±SD) = 26.9% (±26.8), minimum = 1.28%, maximum = 92%). Therefore, it appears that no co-occurrences of *A. candida* isolates were found in the wild.

However, it is possible that in natural mixed infections, one race is rarer (or less abundant) in the host tissue compared to the other. If that was indeed the case, the read depth of the base(s) of the rare race might simply have been insufficient to be scored as a polymorphism (i.e a heterozygous site) at sites where that race differed from the more abundant race. In summary, the absence of proof (i.e. not detecting a significant variation in the level of heterozygosity between isolates of a given race in the field) is not proof of absence (i.e. that mixed infections do not commonly occur in the field). Yet, it is also possible that co-occurrence of two races in the field is rare as hybridization is also probably a rare process in the evolution of *A. candida* races.

So what about sexual reproduction within races? In the next section, I attempt to evaluate the incidence of sexual reproduction in the various *A. candida* races. Provided a difference in the reproductive modes in place, this may account for the variation in heterozygosity between and within races.

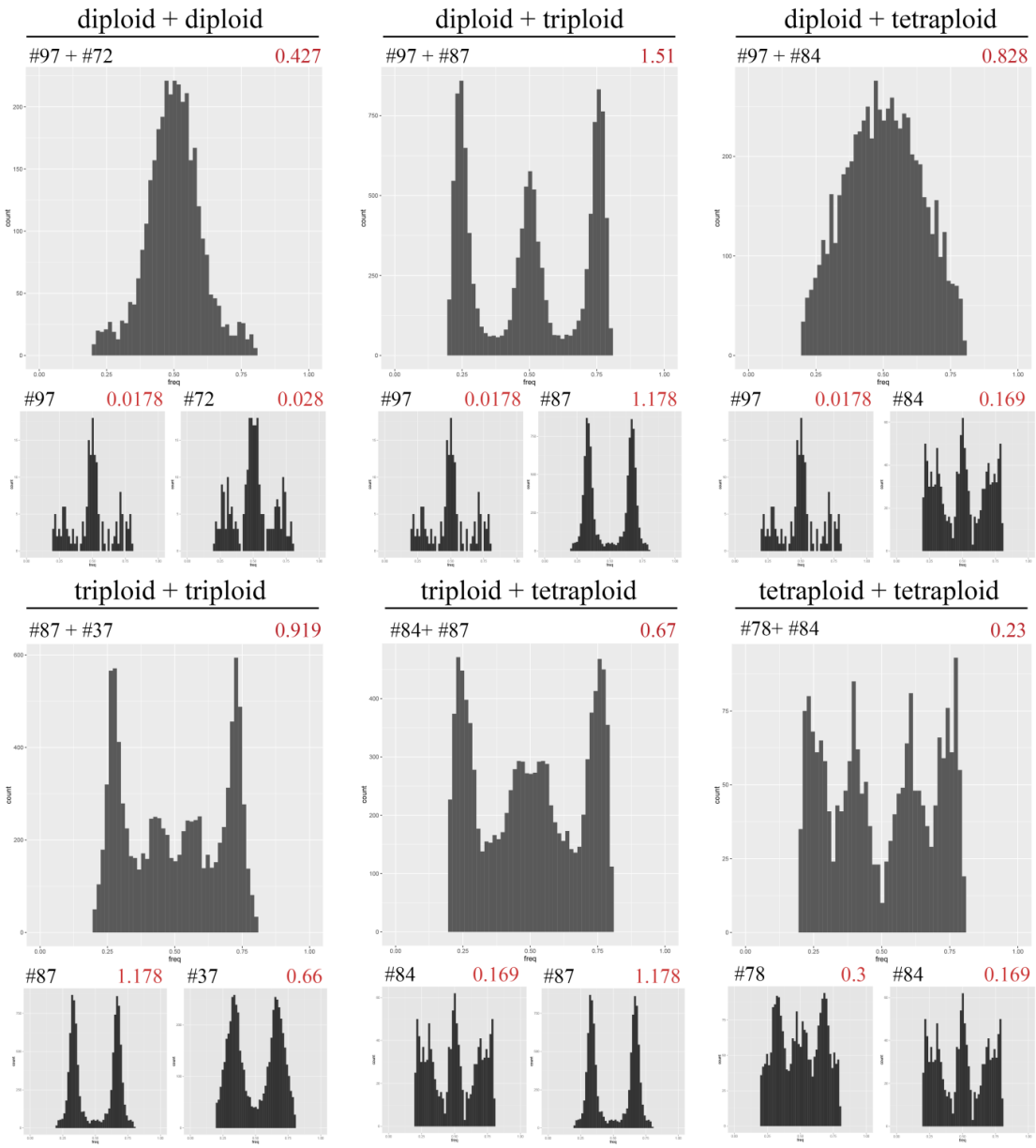


Figure 5.5 Ploidy graphs of simulated mixed infections. For each set of graphs, mixed infection is shown at the top and isolates that were mixed are provided separately at the bottom. The names of the isolates or combination of isolates are provided in black and heterozygosity in red.

5.2.4 GENE FLOW WITHIN *A. CANDIDA* RACES

Although the life cycle of *A. candida* has been described extensively (Saharan & Verma 1992; Meena *et al.* 2014; Saharan *et al.* 2014), the incidence of sexual versus asexual reproduction is still largely unknown. This is probably because genetic markers in *A. candida* have long remained limited, revealing little of the genetic diversity between and within races. It was only in 2015 that the genomes of several *A. candida* isolates were sequenced and

compared (McMullan *et al.* 2015). By doing so, the authors showed that not only isolates within a race had low nucleotide divergence but they also shared most of their heterozygous sites, suggesting that *A. candida* races propagate mainly clonally.

In Chapter 3, I could confirm low nucleotide divergence within *A. candida* races using data from 85 isolates representing 18-20 races. While this seems to support the above hypothesis, ignoring heterozygous sites may bias our evaluation of the incidence of sexual reproduction in *A. candida*, whether be it selfing or outcrossing. Therefore, in this section, I continue on the idea developed in McMullan *et al.* (2015) and I estimate the proportion of shared heterozygous sites at 'contig 1' in each race where a minimum of two isolates were sampled (Figure 5.7). Combining this new data with previous results (nucleotide diversity within races, levels of heterozygosity and ploidy, see Table 5.2), the reproductive modes and mating systems that may be in place in each race are then evaluated.

Here, I hypothesize that isolates that are selfing or predominantly clonal should share most of their heterozygous sites while those that are outcrossing should be more diverged. Furthermore, while diploids and tetraploids may reproduce sexually and/or asexually, triploids are more likely to be asexual, and although it may not be frequent, races where mixed ploidy was observed probably can reproduce sexually. In total, 62 *A. candida* isolates (representing 10-11 races) were included in this analysis.

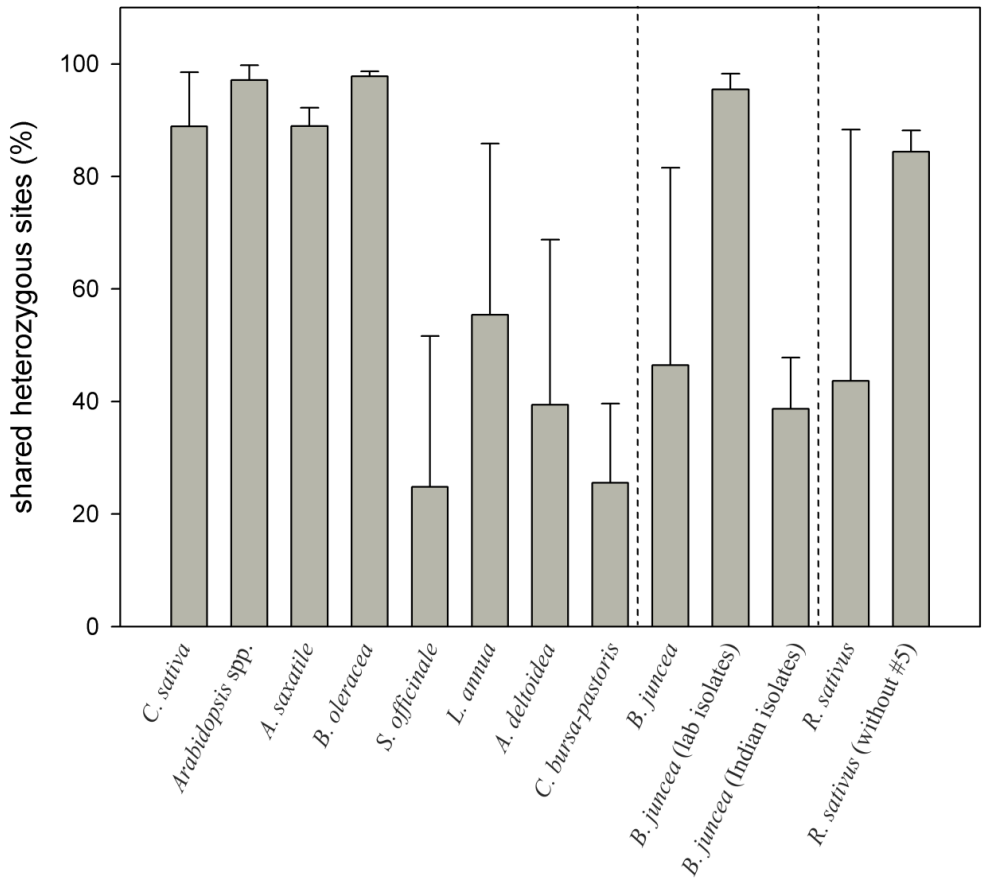


Figure 5.7 Percentage of shared heterozygous sites within *A. candida* races at 'contig 1' (398,508 bp). Using Minitab v. 12.1, the proportion of shared heterozygous sites was computed for each pair of isolates within a race and averaged. Mean and standard deviation are depicted in the figure. Data is provided for *B. juncea* and *R. sativus* whether they are considered as a unique or separated group(s).

While isolates shared most of their heterozygous sites in some races (>80%; isolates collected on *C. sativa*, *Arabidopsis* spp., *A. saxatile*, *B. oleracea, R. sativus* (without #5), *B. juncea* lab isolates), they were much more variable in others (mean percentage < 60%, large standard deviation; isolates collected on *S. officinale*, *L. annua*, *A. deltoidea*, *C. bursa-pastoris*), supporting the hypothesis that the prevalence of sexual and asexual reproduction vary between *A. candida* races. As expected, this analysis also supported the idea that isolates collected on *B. juncea* and *R. sativus* are to be divided into two races or populations (SD = 35% vs. 2.7 and 9.1% in *B. juncea* and 44.6 vs. 3.8% in *R. sativus*). Interestingly, *B. juncea* isolates collected in India in 2015 appeared quite diverged (mean percentage shared heterozygous sites (±SD) = 38.7% (±9.1)) whereas those propagated in the laboratory for several years were highly similar (95.5% (±2.76) shared sites). To summarize all data obtained so far from heterozygous sites, *A. candida* races could be partitioned into three groups (colours in Table 5.2)

	Host	P-distance	Ploidy	Heterozygosity	Shared heterozygosity	MH	MSH	Main reproductive mode?
7	*B. oleracea*	3.30E-05	3x	↑↑ (0.65 ± 0.0164)	Y (97.8 ± 0.84)	0.9 ± 0.35	91.1 ± 9.48	Strict asexual
3	*R. sativus* (without #5)	2.89E-04	2x, 3x	↑↑↑ (1.15 ± 0.029)	Y (84.4 ± 3.8)			
5	*S. officinale*	3.56E-04	2x	↓ (0.0611 ± 0.029)	N (24.8 ± 26.8)	0.087 ± 0.094	36.76 ± 12.53	Outcrossing + asexual
6	*L. annua*	1.40E-04	2x	↓ (0.025 ± 0.002)	N (55.4 ± 30.4)			
9	*A. deltoidea*	3.29E-05	2x	↓ (0.0436 ± 0.029)	N (39.4 ± 29.3)			
11	*C. bursa-pastoris*	1.46E-04	2x, 3x	↓ (0.0489 ± 0.013)	N (25.5 ± 14)			
3	*B. juncea* (India)	1.51E-04	4x	↑ (0.256 ± 0.075)	N (38.7 ± 9.13)			
3	*C. sativa*	2.09E-06	-	↓ (0.0089 ± 0.0012)	Y (88.8 ± 9.64)	0.064 ± 0.039	92.55 ± 4.32	Selfing/LOH + asexual
5	*Arabidopsis spp.*	1.14E-05	2x	↓ (0.096 ± 0.0039)	Y (97 ± 2.6)			
4	*A. saxatile*	7.05E-05	2x	↓ (0.062 ± 0.0045)	Y (88.9 ± 3.3)			
6	*B. juncea* (lab isolates)	1.36E-04	4x	↓ (0.088 ± 0.0088)	Y (95.5 ± 2.76)			
1	*R. sativus* (#5)	n.a.	2x	↓ (0.0536)	n.a.	-	-	-
1	*B. nigra*	n.a.	-	↓ (0.0194)	n.a.	-	-	-
1	*S. alba*	n.a.	2x	↓ (0.0798)	n.a.	-	-	-
1	*E. sativa*	n.a.	-	↓ (0.0165)	n.a.	-	-	-
1	*L. rediviva*	n.a.	-	↓ (0.0266)	n.a.	-	-	-
1	*B. rapa*	n.a.	4x?	↓ (0.0166)	n.a.	-	-	-
1	*B. carinata*	n.a.	-	↓ (0.1225)	n.a.	-	-	-
1	*A. alpine*	n.a.	2x	↓ (0.1132)	n.a.	-	-	-
1	*E. japonica*	n.a.	-	↓ (0.0105)	n.a.	-	-	-

Table 5.2 Summary of the parameters defined in *A. candida.* The number of isolates that could be analysed in each race is provided in the first column. P-distance was obtained from averaging p-distances estimated using 'contig 1' and diversity-tracking loci in Chapter 4 (see Table S4.2). 2x isolates are diploid, 3x triploid and 4x tetraploid. The mean and standard deviation of the mean heterozygosity as well as of the mean percentage of shared heterozygous sites are provided in the 4th and 5th columns. Two categories were also provided for the latter parameter: Y when isolates shared >88% of their heterozygous sites and N when they shared <60%. Shared heterozygosity could not be estimated for races where only one isolate was sampled, therefore it was not possible to hypothesize on the main reproductive mode in these races. Similarly, ploidy could not be defined for several isolates or races (see Chapter 5 section 2). Three groups could be defined based on these parameters (blue, red and green) and are discussed in the text. Mean heterozygosity (MH) and mean percentage of shared heterozygous sites (MSH) were estimated for these three groups.

The first group (green in Table 5.2) includes all-triploid isolates collected on *B. oleracea* and both diploid and triploid isolates from *R. sativus* (without #5). In these two races, heterozygous sites were mostly shared (97.8 and 84.4%, respectively) and heterozygosity was high, suggesting that gene diversity is being maintained by strict clonal reproduction.

The second group of races were found on wild or cultivated hosts (*S. officinale*, *L. annua*, *C. bursa-pastoris*, cultivated *B. juncea* and *A. deltoidea*, red in Table 5.2). Heterozygosity was low due to regular sexual reproduction between related individuals, which results in inbreeding. Low levels of heterozygosity are typically associated with drift, which in this case may be due to founder events during host colonization. Due to sexual reproduction, however, isolates within these races did not share much of their heterozygous sites (~25-55%). In this group, bouts of sexual reproduction between closely related individuals appeared to be interspersed by clonal reproduction. This interpretation is illustrated, for example, by analysis of isolates in *S. officinale* which shared ~98.9% of their heterozygous sites when collected at the same time and location (#1 and 10, 17/05/2013), but only shared ~26% of their heterozygous sites if collected a little more than a year apart. Interestingly, these two year-groups retain similar levels of heterozygosity (#28, 28/01/2015, mean $heterozygosity_{\#1+10}$ = 0.067, $heterozygosity_{\#28}$ = 0.084), which suggests that the transition between both modes of reproduction (sexual and clonal) might be a regular occurrence resulting in a dynamic equilibrium of changes in gene diversity.

Furthermore, in this group, the percentage of shared heterozygous sites was quite variable (SD = 9-30%, see Table 5.2). This may be explained by an increased likelihood of outcrossing events between geographically close populations. Using the ade4 package in R v. 3.1.2, a mantel test was therefore performed to investigate a potential correlation between the proportion of shared heterozygous sites between two isolates and the distance, in kilometres, that separate them (Figure 5.8). Surprisingly, while geographically close isolates were sometimes also genetically similar (e.g. 91% heterozygous sites shared between *L. annua* #71 and #72 collected 6 km apart), others were more diverged (e.g. 27% heterozygous sites shared between *L. annua* #69 and #70, also collected 6 km apart). Similarly, while distant isolates shared little of their genetic diversity (e.g. 28% shared heterozygous sites between *L. annua* #76 and #89 collected 728 km apart), others were genetically closely related (e.g. 81% shared heterozygous sites between *L. annua* #72 and #76 collected 1,040 km apart). This absence of a correlation between genetic and geographic distances (p-values ~ 0.13-1) may be explained by long distance dispersal of the pathogen. *A. candida* appears to have a huge potential for gene

flow, which in combination with its ability to reproduce clonally, make it a formidable pathogen that can rapidly colonise and expand its population size across a large geographic range.

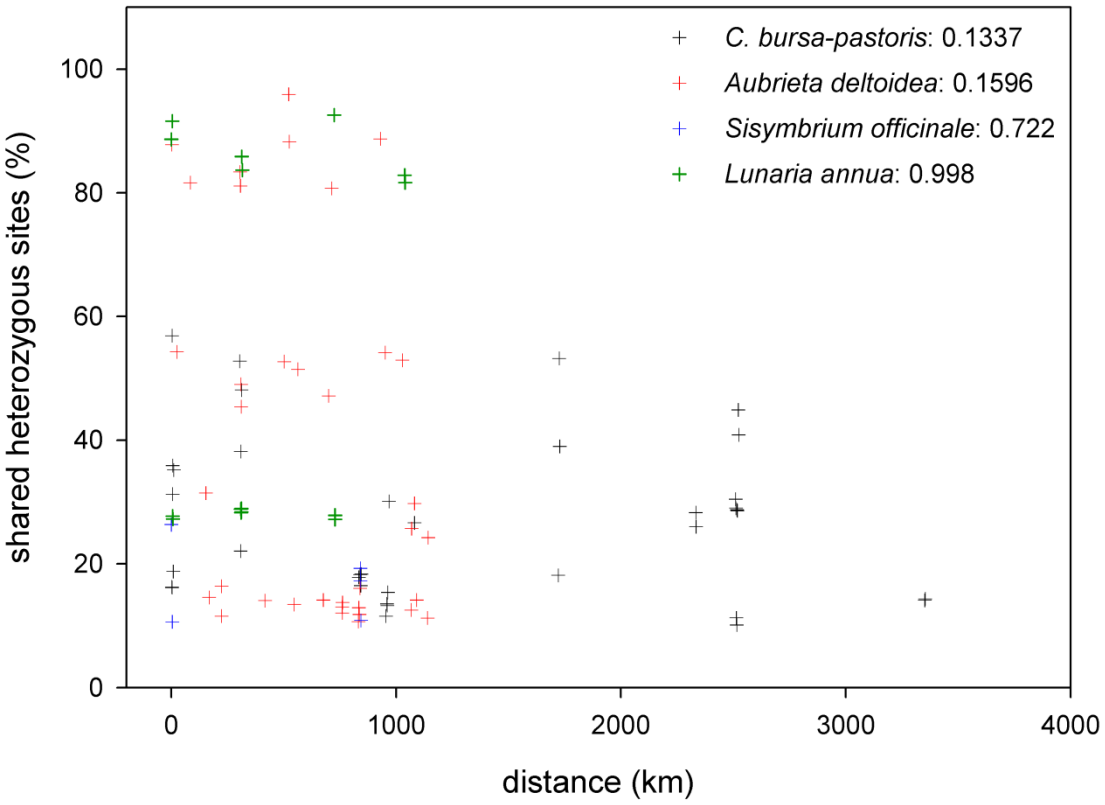


Figure 5.8 Correlation between the percentage of shared heterozygous sites between two isolates within a race and the distance from which they were collected. This analysis was performed for each race where sexual the percentage of heterozygous sites between isolates was <60%. Isolates collected on cultivated *B. juncea* could not be included in the analysis because only two isolates had known GPS coordinates. A mantel test was performed in R v. 3.1.2 using the ade4 package and p-values are provided at the top right of the graph. No statistically significant correlation was found between the two parameters.

Finally, the third group of *A. candida* races mostly consists of isolates that have been propagated in the laboratory for some time (*C. sativa*, *Arabidopsis* spp. and laboratory *B. juncea*, blue in Table 5.2). Although heterozygosity was as low as in the second group (two sample t-test: $T = -0.49$, $p > 0.5$, $df = 5$), isolates within these races shared most of their heterozygous sites (>88%). This is consistent with clonal reproduction after inbreeding and possible loss-of-heterozygosity events. This latter hypothesis is explored in the last section of this chapter.

5.2.5 LOSS-OF-HETEROZYGOSITY

Loss-of-heterozygosity (LOH) is due to mitotic recombination or gene conversion events and is responsible for homozygosis of genomic regions of variable lengths (Bennett *et al.* 2014). While it is well studied in cancer genomes (Ha *et al.* 2012; Pedersen & De 2013; Chen *et al.* 2014), this mechanism has only recently been reported in oomycetes (in *Phytophthora ramorum*: Vercauteren *et al.* 2011 and *P. capsici*: Lamour *et al.* 2012) and may have important implications for our understanding of pathogens adaptation.

To detect LOH, researchers need to compare heterozygosity between parental isolates and their progenies and identify regions of homozygosis in the progenies that could not have arisen from sexual or clonal reproduction alone. While it is not possible to achieve this using the samples that were sequenced with PathSeq, I would like to highlight several observations that may shed light on potential LOH events in *A. candida*. These are based on the distribution of heterozygous sites along 'contig 1'.

While the average homozygous tract length was ~130 bp (±SD = 31) in clonal races, suggesting that these races are more heterozygous overall (Table 5.2 green group, Figure 5.9A), it was much higher in races where selfing or LOH has been hypothesized (Table 5.2 blue group, mean homozygous tract length (±SD) = 2,163 (±2,376), Figure 5.9B). Similarly, the longest homozygous sequence tract observed in 'contig 1' was shorter in the clonal races (6,864 bp (±1,895) vs. 63,801 bp (±63,162)). This excludes the potentiality of LOH in the first group and a contrasting and puzzling pattern is revealed. If both groups of races are clonal, then, why is there such a difference in heterozygosity and in the way heterozygous sites are distributed?

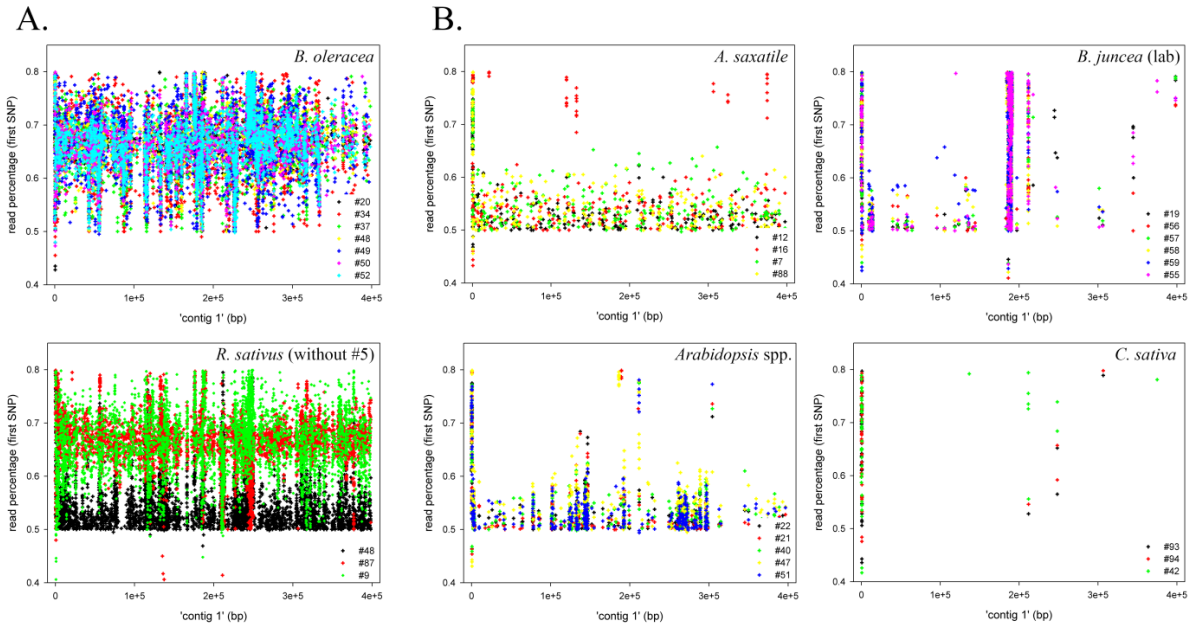


Figure 5.9 Distributions of heterozygous sites along 'contig 1' in *A. candida* races. Each dot represents a heterozygous site for which the read percentage of the most abundant SNP is given on the y-axis. **A.** Strictly clonal races as defined in Table 5.2 (green) show high levels of heterozygosity, resulting in a high density of dots. **B.** Races for which selfing or LOH have been hypothesized (Table 5.2, blue). In these races, large areas without polymorphisms (no dots) are interspersed by more polymorphic regions that appeared to be largely shared between isolates. The hypothesis most consistent with this observation is that LOH has erased nucleotide variation across large sequence tracks and/or inbreeding has reduced heterozygosity across the genome and that such identical genotypes occur in multiple isolates because they were derived from clonal reproduction.
In this graph, the most abundant SNPs at heterozygous sites in triploid isolates usually represent ~67% of the reads (0.67 on the y-axis) while those from diploid isolates represent ~50% (0.5 on the y-axis).

The first possible reason that comes to mind is the difference in ploidy level which could account for a higher number of heterozygous sites and therefore shorter homozygous tracts in the clonal races. However, the observation of similar heterozygous sites distributions in the diploid (#48) and triploid isolates from *R. sativus* (#9 and 87) appears to falsify this hypothesis. Another suggestion would be that LOH has occurred in isolates in Figure 5.9B, possibly due to stresses induced by their propagation in the laboratory (Forche *et al.* 2011; Bennett *et al.* 2014). However, again, all isolates collected from *B. oleracea* except #34 and from *R. sativus* except #9 have also been propagated in the laboratory and homozygous tracts were similarly distributed in the wild and laboratory isolates (one-way ANOVA: $F_{(6,4)} = 0.15$, $p = 0.989$ in *B. oleracea* isolates and $F_{(2,1)} = 1.31$, $p = 0.271$ in *R. sativus* isolates). It is therefore likely that isolates in Figure 5.9B (blue in Table 5.2) do not reproduce strictly clonally like

isolates in Figure 5.9A (green in Table 5.2) but rather that sexual reproduction between closely related individuals (in this case self-fertilization) has reduced the level of heterozygosity over 'contig 1'. If that is the case, selfing in these isolates may be triggered by particular environmental conditions, for example, geographic isolation due to laboratory propagation.

Finally, I would like to draw the attention on isolates #32 from *S. officinale* and #73 and 74 from *A. deltoidea*. In contrast with other isolates collected on these hosts (Figure 5.10, left), these appear to have lost almost all of their heterozygous sites at 'contig 1' (Figure 5.10, right). This observation is also consistent with LOH and/or more prolonged inbreeding that has erased gene diversity across 'contig 1'. Interestingly, however, for *A. deltoidea* there are two isolates which both share the few remaining heterozygous sites. This observation suggests that both isolates are derived from clonal reproduction.

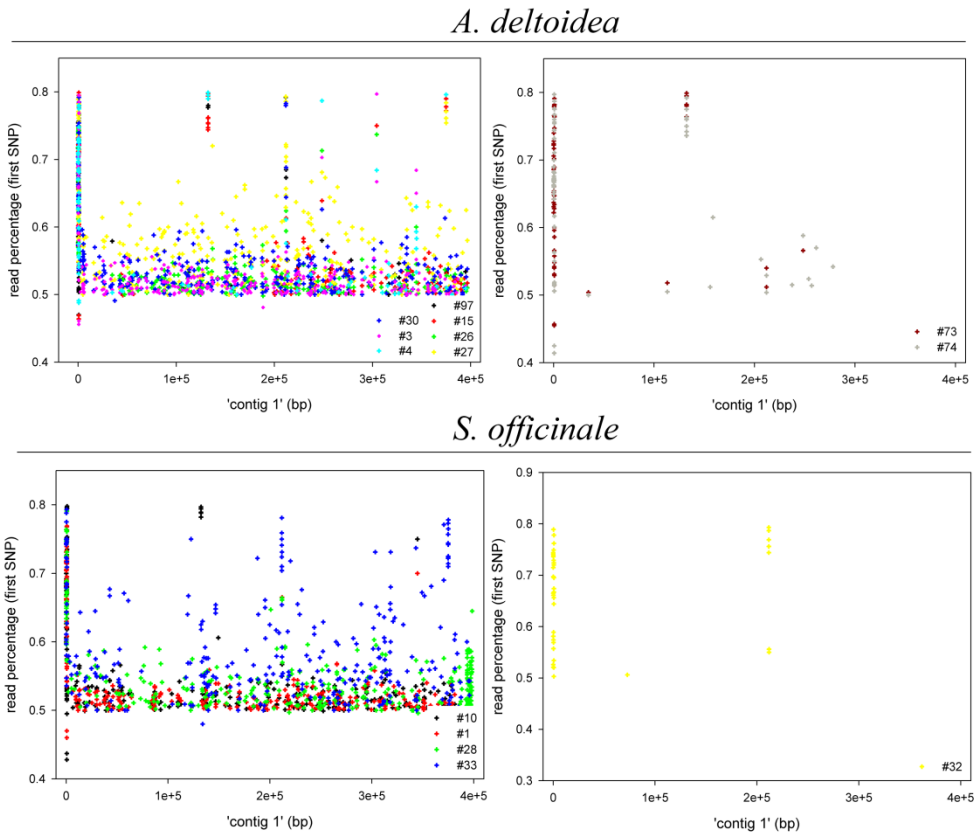


Figure 5.10 Extreme variations in the distribution of heterozygous sites within *A. candida* races collected on *A. deltoidea* and *S. officinale*. Each dot represents a heterozygous site for which the read percentage of the most abundant SNP is given on the y-axis. While typical isolates had moderate levels of heterozygosity (left), several isolates were found where heterozygosity was severely reduced, potentially indicating the occurrence of large LOH events.

5.3 DISCUSSION

In Chapter 4, I investigated genetic variation between *A. candida* isolates and could establish the presence of 18-20 mainly host specialised races out of the 83 isolates which were successfully sequenced using PathSeq. Further supporting the hypothesis from McMullan *et al.* (2015) that *A. candida* mainly reproduce asexually, genetic variation within these races appeared to be low (from 2.09E-06 to 3.56E-04 mutations per sites in pairwise comparisons of 'contig 1', see Table 5.2). Yet, thick-walled egg cells known as oospores are often observed in hypertrophied parts of the host (Lakra & Saharan 1989b; Nath *et al.* 2000) and are thought to be crucial for survival between flowering seasons or during unfavourable environmental conditions (Meena & Sharma 2012; Saharan *et al.* 2014).

In this chapter, I therefore decided to investigate the incidence of sexual reproduction in *A. candida*. To do this, I incorporated information from heterozygous sites at 'contig 1' which were ignored in previous analyses. Here, I hypothesize that while mutations should be shared among individuals in a clonal population, they are reassorted each sexual cycle and may rapidly increase or decrease genetic diversity between isolates in a sexual population (Signorovitch *et al.* 2005; Schurko *et al.* 2009). Consequently, heterozygosity should be higher in clonal, intermediate in outcrossing and low in inbreeding populations (Schurko *et al.* 2009).

Previously unexplored, heterozygous sites revealed extra nucleotide diversity between and within races and while most isolates had low levels of heterozygosity, others were highly heterozygous (from 0.009% observed heterozygous sites on *C. sativa* to 1.15% on *R. sativus* (without #5)). Interestingly, isolates with high levels of heterozygosity were found to be triploid (*R. sativus*, *B. oleracea*) and those with low and intermediate levels of heterozygosity were identified as diploids and tetraploids (*B. juncea*, *B. rapa*, possibly *B. carinata*), respectively. Polyploidization in *A. candida* may be due to either whole-genome duplication, syngamy of (reduced and) unreduced gametes of isolates of the same race (autopolyploidy) or of different races (allopolyploidy) (Albertin & Marullo 2012). Reconstructing haplotypes using PacBio sequencing for example or investigating shared heterozygosity between races could allow the identification of the mechanisms of polyploidization in *A. candida*. In the case of allopolyploidy, it may also help identify putative parental races of the polyploids and may provide indirect evidence for hybridization between *A. candida* races.

Polyploidy in *A. candida* may have important impacts on both the occurrence of sexual reproduction within and between races and the adaptive potential of the races. First, polyploid

races may have reduced fertility or may even be strictly asexual (especially triploids; Comai 2005). This is what is observed in triploid races from *B. oleracea* and *R. sativus* (without #5) where heterozygous sites were mostly shared between isolates collected in various times and locations, suggesting strict clonal reproduction (mean % shared heterozygous sites (± SD) = 91.1% (± 9.48)). The hypothesis of asexuality in these races 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, personal communication).

Second, polyploid races may not be able to hybridize with other races due to pre-zygotic reproductive barriers (such as loss of sexual reproduction) or to produce viable or fertile hybrids due to the imbalance in the ratio of both parental genomes in the hybrid (Pannell *et al.* 2004; Köhler *et al.* 2010). This may, in the long term, lead to speciation of the polyploid races, particularly of the triploid races which appear to not be able to reproduce sexually. In this chapter, I could not obtain direct evidence for hybridization (co-occurrence of isolates from different races). This may reflect the fact that hybridization in *A. candida* may be a very rare process, possibly because genetic exchange between races that are adapted to different hosts is likely to be maladaptive.

Finally, polyploidy may allow for relaxed selection pressure and therefore increased mutation rate 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, stature) in comparison to diploid races and interestingly, all polyploid races identified during this work had previously been classified as *A. candida macrospora* (isolates collected on *Brassica* and *Raphanus* spp.) due to sporangia which are larger than 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 races classified as *macrospora* (other *Brassica* spp. or *Erucastrum* spp.) and for an increased sporangia/zoospore resistance or survival in the polyploids.

Although one polyploid isolate was found on *C. bursa-pastoris* (#2), polyploidy appeared to have been selected for only in cultivated host species. This is interesting because polyploidy may allow for quick adaptation to the uniform genotypes of cultivated host populations. In addition, asexuality (or reduced fertility) associated with polyploidy may be beneficial as sexual reproduction would break up combinations of alleles that are adapted to the host genotypes. In contrast, asexual reproduction may not be advantageous when the host species are genetically diverse, for example in wild populations of *C. bursa-pastoris* or *S. officinale*. In this thesis, all but triploid races appeared to be able to reproduce sexually.

However, while isolates collected in the wild appeared to be outcrossing ((mean % shared heterozygous sites (± SD) = 36.76% (± 12.53)), isolates that have been propagated in the laboratory (*C. sativa*, *B. juncea*, *Arabidopsis* spp.) or that were collected on an ornamental plant (*A. saxatile*) appeared to be selfing (mean % heterozygous sites at 'contig 1' (SD) = 0.064 (0.039)) and mean % shared heterozygous sites (± SD) = 92.55% (± 4.32)). This apparent transition to selfing may be due to the lack of a mating partner when isolates are cultivated in isolation. This is possibly what has happened in the race adapted to *B. juncea* which appeared to be outcrossing or selfing depending on whether isolates were collected from the field ((mean % shared heterozygous sites (± SD) = 38.7% (± 9.13)) or in the laboratory ((mean % shared heterozygous sites (± SD) = 95.5% (± 2.76)), respectively. Outcrossing isolates appeared to be capable of long-distance dispersal, at least through Eastern Europe.

Finally, analysis of heterozygous sites also revealed that some isolates seemed to have lost most of their heterozygous sites at 'contig 1' (isolates #73 & #74 from *A. deltoidea* and #32 from *S. officinale*). This may be due to either selfing in some *A. candida* races or loss-of-heterozygosity events. However, the former hypothesis is unlikely in *S. officinale* because isolate #32 was collected only a few hundred meters away from #33 which appeared to be outcrossing. In *A. deltoidea*, isolates with low heterozygosity were both collected in Northern Scotland. The host plant, which is cultivated as an ornamental and is very frequent in England and France, appeared to be less frequent in Scotland which may have induced selfing in these isolates. However, it is difficult to discriminate between these two mechanisms.

CHAPTER 6: GENERAL DISCUSSION

With the advent of next-generation sequencing technologies, scientists can now easily explore the natural diversity of plant pathogens to investigate the mechanisms by which they adapt to their environments and their hosts. In addition to investigating particular gene interactions between pathogens (avirulence or effector genes) and their hosts (resistance genes), we can now try to assess the significance of other, more complex, evolutionary and ecological processes that may have been disregarded in the past. For example, while polyploidization has been put forward as an important mechanism for species diversification in plants (Madlung 2012; del Pozo & Ramirez-Parra 2015), few studies report on its potentially important role in the evolution of other organisms such as pathogenic fungi or oomycetes (Albertin & Marullo 2012; Yoshida *et al.* 2013). Similarly, while hybridization is known to be frequent in plants and to a lesser extent in animals (James *et al.* 2007; Schwenk *et al.* 2008), its significance is only recently recognized in fungal and oomycete pathogens (Schardl & Craven 2003; Stukenbrock *et al.* 2012), probably due to the "fuzzy" nature of species boundaries in microbes and to the lack of molecular data from natural populations.

Although many studies have been published since 2000 on the evolution of plant pathogenic oomycetes, these were mainly focusing on a few economically relevant or model pathogens such as those infecting important crops (e.g. potato late blight *P. infestans* (Gavino *et al.* 2000; Yoshida *et al.* 2013; Peters *et al.* 2014), grapevine downy mildew *Plasmopara viticola* (Schröder *et al.* 2011; Rouxel *et al.* 2013) or the sudden oak death pathogen *Phytophthora ramorum* (Goss *et al.* 2009; Prospero *et al.* 2009)). In particular, research has often been directed towards species from the Peronosporales, and by doing so, it failed to take advantage of the insights that can be gained by investigating the evolution of other (more ancient) orders (e.g. Albuginales, Eurychasmales and Haliphthorales; Beakes *et al.* (2012)).

The Albuginales is an early-diverging order in the Peronosporomycetes clade (the latter including the Peronosporales) and consists exclusively of obligate biotrophic pathogens of angiosperms (Beakes *et al.* 2014). Included in this order, *Albugo candida* is the most-well studied species, probably due to its apparently broad host range (Choi *et al.* 2009). However, while the mechanisms for resistance against *A. candida* have been partly characterized in various Brassicaceae and while many host-specific races have been identified (see Saharan *et al.* 2014 for a recent review), little is known about *A. candida* natural diversity and the mechanisms by which it adapted to its many hosts. Only one report addresses this question

using five isolates representing three host-specific races (McMullan *et al.* 2015). The authors provided evidence for hybridization between races as well as clonal propagation of well-adapted isolates. They could also show that immunosuppression of the host caused by *A. candida* could enable the co-occurrence of diverged races on the same plant which could potentially facilitate hybridization. However, many questions still remained. For example McMullan *et al.* (2015) found that *Ac*Nc2 and *Ac*Em2, collected on different host species (*A. thaliana* and *C. bursa-pastoris*), were genetically highly similar and part of the same race, but they were unable to address the question of whether there are other races capable of infecting multiple diverged hosts. Furthermore, it remained unknown whether there are hosts in nature that can be colonized by multiple races. In addition, while *A. candida* is known to produce sexual oospores, for example, to survive intercrop seasons, McMullan *et al.* (2015) suggested that isolates within a race propagate mainly clonally. The limited number of isolates (five in total, with a maximum of two isolates per race) was not enough to properly analyse the rate of sexual reproduction within races. Other questions remained unanswered, for example, whether sex within races is a very rare process or whether oospores are principally produced when races hybridize This latter hypothesis is unlikely as several papers report on frequent oospore formation in *A. candida* (Lakra & Saharan 1989b; Liu & Rimmer 1993; Meena & Sharma 2012) and as oospores are regularly observed in the races that are maintained in the lab (*Ac*2v and *Ac*7v; V. Cevik, personal communication). Moreover, high variation in the level of heterozygosity was observed between races (*Ac*BoT/*Ac*BoL vs *Ac*Nc2/*Ac*Em2 and *Ac*2v). While McMullan *et al.* (2015) suggested loss-of-heterozygosity in the latter races to explain this variation, alternative mechanisms should be explored such as the effect of outcrossing, selfing and clonal reproduction in *A. candida* races or potential polyploidization events. Indeed, although *A. candida* is usually described as a diploid organism, the ploidy level of the various races has not been investigated through flow cytometry or cytology work (except in Sansome & Sansome (1974) who inferred hexa- or octoploidy). Finally, if hybridization occurs in nature, can we find two races on the same plant? Can we find hybrids?

In this work, I investigated the genetic diversity and evolution of natural populations of *Albugo candida*. To do this, I developed a novel method based on sequence capture to selectively sequence loci from *A. candida* but also from diverse microorganisms. By doing so, I hope to highlight important processes in *A. candida* but also, more generally, in plant pathogen adaptation and to propose a method that could facilitate future work on the evolution of microbes and microbial communities. In this discussion, I will first report on the efficiency,

advantages and flaws of the sequence capture method (PathSeq). I then discuss my findings regarding within-race diversity and the importance of hybridization in *A. candida*.

6.1 DESIGN OF A METHOD TO EXPLORE MICROBIAL DIVERSITY

In-solution sequence capture methods are based on the hybridization between RNA biotinylated baits and target DNA that are at least ~80% identical in sequence (Jupe *et al.* 2013). These methods enable selective sequencing of loci from an infected leaf from the field. This is useful for large genomes and for mixtures of biotroph and host DNA or when the aim of the research is focused on particular genes such as resistance genes in plants (RenSeq; Jupe *et al.* (2013)) or pathogen effectors. It may also be used to sequence loci from multiple organisms for example, by combining baits targeting resistance and effector genes from both a pathogen and its host. In PathSeq, I designed baits to capture, amplify and sequence diagnostic loci from 48 microbial species including bacteria, fungi and oomycetes. By doing so, I hoped to cost-effectively generate data from multiple *A. candida* isolates collected in the wild as well as to interrogate microbial diversity in *A. candida*-infected field sample leaves. MLST loci were selected for microbial species or genus identification and other loci, including effector genes, were included to investigate the natural genetic diversity of several important plant pathogenic oomycetes (e.g. *A. candida*, *H. arabidopsidis*, *P. infestans*).

Although MYcroarray, manufacturer of the baits, suggests pooling samples after capture, I could enrich and sequence up to 47 multiplexed samples in one HiSeq lane whilst still generating high quality reads (mean Q to base > 130). The number of reads per sample was quite variable (~4,000-20,000) but in almost all cases, it seemed enough to get high read depth at selected loci (18-2,336 x based on *A. candida* 'contig 1', mean read depth at 'contig 1' (±SD) = 478.8 x (±528.4)) which is essential for downstream analyses (e.g. variant detection). However, although comparable with previous publications, the percentage of off-target reads was quite high (38-77%, Nadeau *et al.* (2012); Jupe *et al.* (2013)). Not surprisingly, these mainly originated from the most abundant organisms (the host and *A. candida* when present) and are probably due to the presence of large fragments in the libraries (larger than the actual targets) and the use of long external barcode and adapter sequences which may have hybridized with each other, reducing the overall capture efficiency (Mamanova *et al.* 2010; Rohland & Reich 2012).

PathSeq enabled efficient identification of both the host and the pathogens that were included in the control samples. Indeed, using a mapping-based approach I could show that (close to) 100% of the targets base pairs were captured and sequenced from control organisms when present. However, due to the likely presence of species that are closely-related to control organisms, a small proportion of reads wrongly mapped to those that were expected to be absent. Therefore, while a mapping-based approach appears sufficient to quickly test for the presence or absence of targeted organisms, it does not seem to be a good strategy to interrogate the whole (mostly unknown) microbial diversity from a sample. To do this, metagenomics methods would be preferred such as sequence classification tools (MEGAN: Huson *et al.* 2007; Kraken: Wood & Salzberg 2014) that rely on the comparison between reads generated from a sample and databases of known sequences.

In this thesis, I do not report on any of these methods as priority was given to the development of the method first and the analysis of *A. candida* diversity. However, my aim, in the near future, is to make use of the huge amount of data generated using PathSeq and identify, at least to the genus level, the microbial diversity there is in all samples. In particular, I am interested in the influence *A. candida* may have on the host microbial community. Indeed, as plant pathogens can dramatically influence host physiology, it is expected that they will also influence the whole microbial community associated with the host (Kemen 2014; Agler *et al.* 2016). This has already been shown in wild *A. thaliana* populations for example, where the oomycete *Albugo laibachii* and the yeast fungus *Dioszegia* were found to have a strong impact on the phyllosphere microbial diversity (Agler *et al.* 2016). In PathSeq, I have included samples of *A. candida* infected leaves as well as of healthy leaves that were collected at the same time and location. Comparing microbial diversity between these samples will help, I believe, to highlight microorganisms (bacteria, fungi, oomycetes) that either benefit from *A. candida* colonization of the host or that increase in abundance during infection to protect the host.

6.2 *ALBUGO CANDIDA* NATURAL VARIATION

For the first time, the extent of genetic diversity in natural populations of *A. candida* has been documented. Using PathSeq, I could sequence ~660,000 bp (187 loci) from 83 *A. candida* isolates collected on 21 Brassicaceae host species. In this thesis, I analysed nucleotide diversity at both 'contig 1' (~400 kb) and 32 diversity-tracking genes (~20 kb). Depending on the genomic region used in the analysis, 15-19 highly host specific races were identified that

are ~0.7% diverged (similar to what was found in McMullan *et al.* (2015)). These races appeared to be mostly specialized on one host species (e.g. *A. candida* on *A. deltoidea*, *A. saxatile* or on *S. officinale*) although more sampling is required for confirmation. Indeed, *A. candida* races may be able to infect closely-related host species that have not been sampled in this study. If that is the case, adaptation to multiple hosts would probably increase the chances of being recognized but it should also facilitate the search for a suitable host therefore increasing colonization success. This is what we see for example for race *Ac*Nc2/*Ac*Em2 which was found on *C. bursa-pastoris*, *A. thaliana* and *E. japonica* and for the race infecting both *L. annua* and *L. rediviva* (although this needs to be confirmed using host-specificity assays). Similarly, while most hosts appeared to be colonized by one race only, others may be susceptible to several races. This would induce competition between races but could also facilitate hybridization events. In this study, both races *Ac*Ex1 and *Ac*Nc2/*Ac*Em2 were found on *A. thaliana*. However, they were most probably collected from different *A. thaliana* ecotypes as only one race (*Ac*Ex1, unpublished) can overcome resistance conferred by the *WRR4* locus of some ecotypes (White Rust Resistance; Borhan *et al.* (2008)). Two races also appeared capable of colonizing *R. sativus*.

Remarkably, these host-specific *A. candida* races not only were genetically diverged but they also appeared to differ in ploidy level. This is very interesting because polyploidy in *A. candida* may have important impacts on the occurrence of sexual reproduction within and between races and therefore, on race specialization and speciation. Indeed, although not well-studied in oomycetes, polyploidy may reduce fertility (Burton & Husband 2001; Stöck *et al.* 2002; Comai 2005; Otto 2007) and act as a reproductive barrier between populations and races (Madlung 2012). In fact, triploidy in *A. candida* appeared to be associated with a strict clonal reproductive system (isolates collected on *B. oleracea* and *R. sativus*) which may, in the long-term, lead to speciation. This idea is further reinforced by the absence of observable oospores in race *Ac*BoT compared to other races (V. Cevik, personal communication). In contrast, wild tetraploid *Ac*2v-like isolates appeared to be able to reproduce sexually and may still hybridize with other races (although perhaps with lower success with diploids). Again, this is further reinforced by the frequent observation of oospores on plants infected by *Ac*2v (V. Cevik, personal communication but see also Adhikari *et al.* (2003)).

In all cases, nucleotide divergence within races was low (~0.00025%) suggesting high levels of clonal reproduction. All except the triploid races appeared to also be able to reproduce sexually. Races collected in the wild seemed to be mostly outcrossing and capable of long distance dispersal (on *C. bursa-pastoris*, *A. deltoidea*, *S. officinale*, *B. juncea*). Sampling

isolates on other continents or using spore trap experiments combined with targeted sequencing would certainly allow dispersal in *A. candida* to be studied more thoroughly. In contrast, isolates that have been propagated in the lab (*B. juncea*, *Arabidopsis* spp, *C. sativa*) or collected from the ornamental plant *A. saxatile* appeared to be mostly selfing. This suggests that *A. candida* is a homothallic oomycete and that self-fertilization may be "switched on" when there are no suitable mating partners available. A great example for this is the lower proportion of shared heterozygous sites between isolates collected on wild *B. juncea* compared to those collected in the lab.

Surprisingly, although polyploidy was occasionally observed in races adapted to wild host species (*C. bursa-pastoris*), it seemed that it was selected for only in races infecting cultivated host species (*B. oleracea*, *B. juncea*, *B. rapa*, *R. sativus*, and possibly *B. carinata*). Although this may truly be coincidental, it is possible that polyploidy represents a good evolutionary strategy in agricultural systems where selection imposed by the cultivated host is strong (Stukenbrock & McDonald 2008). Indeed, it has been hypothesized that additional gene copies generated by polyploidization may assume new (neofunctionalization) functions compared to the parental diploids, potentially increasing responsiveness to environmental changes (here, the cultivated host genotype, see Adams & Wendel (2005); Madlung (2012)). Moreover, the likely asexual reproduction associated with polyploidy may also be beneficial in agricultural systems as sex would break up combinations of alleles that are adapted to the cultivated host genotype.

Finally, extremely low levels of heterozygosity in some isolates compared to others of the same race (*A. candida* in *S. officinale* and *A. deltoidea*) either suggests that normally outcrossing races may self-fertilize (e.g. when no mating partner is available, as when propagated in the lab) or that loss-of-heterozygosity is occurring in *A. candida*. While it would be difficult to test for LOH by cultivating and crossing *A. candida* isolates/races in laboratory conditions (as was done for *Phytophthora capsici* (Lamour *et al.* 2012a)), it would be interesting to collect *A. candida* once or twice a year in the same host population over several years to test for rapid reduction of heterozygosity between generations.

In this thesis, I gained insight on *A. candida* evolution and diversity by analysing nucleotide variation at 'contig 1' and diversity tracking genes. However, PathSeq was also designed to capture and sequence MLST genes like *cox1*, *cox2*, ITS and most importantly putative effector genes of *A. candida*. In the near future, it will be important to make use of this data and investigate genetic diversity in *A. candida* putative effector genes. This may allow the identification of new putative CCG effectors that are not present in races whose genomes have

been sequenced and it could also highlight those that have an important role for host colonization or virulence.

6.3 HYBRIDIZATION BETWEEN *ALBUGO CANDIDA* RACES

One of the most important hypotheses that was proposed in McMullan *et al.* (2015) was that hybridization between *A. candida* races may have enabled adaptation to the numerous hosts from which the pathogen has been described. The authors could identify many recombination blocks that were shared between races and also demonstrate that *A. candida* races can co-occur, at least in laboratory conditions, provided the immune system of the host is compromised by a compatible race. By sampling more isolates from different host species, I set out to first confirm the results obtained by McMullan *et al.* (2015) as well as to gather direct evidence for hybridization (mixed infections or natural hybrids).

In Chapter 4, I could identify many recombinant regions between races at 'contig 1'. Some of these regions were quite polymorphic and are probably due to incomplete lineage sorting or trans-species polymorphism. However, other regions are (nearly) identical between races over long stretches of sequence, suggesting that they may be due to recent hybridization through secondary contact. Yet, no direct evidence for hybridization could be obtained during my work on *A. candida*. Indeed, in Chapter 5 I showed that the co-occurrence of isolates from two different races would dramatically increase the level of heterozygosity detected at 'contig 1', no matter the ploidy level. However, this was not observed from the samples sequenced with PathSeq suggesting that they either contained a single *A. candida* isolate or possibly (I cannot exclude this hypothesis due to the very low nucleotide divergence within races), two isolates of the same race. Although disappointing, the lack of evidence for mixed *A. candida* infections is perhaps not so surprising given that hybridization is probably very rare (McMullan *et al.* 2015). This is because frequent genetic exchanges between races that are adapted to diverged hosts (each with their unique sets of resistance genes) would most likely be maladaptive. Indeed, although in rare occasions combinations of effectors would form that are not recognized by and enhance virulence in a former or novel host, recombination would likely bring together alleles that are not as efficient as those which have been selected for over long evolutionary times.

To continue forward in confirming the role of hybridization in the evolution of *A. candida*, more samples would therefore be required. Moreover, as successful hybridization

events are likely to affect genes that are important for host colonization and virulence, the presence/absence and nucleotide diversity of putative effector genes could also potentially be used to identify hybrids. Furthermore, investigation of polyploidy may help unravel indirect evidence for hybridization between races. Indeed, although the mechanisms for the formation of polyploids in *A. candida* are unknown and potentially numerous, it is possible that they have arisen from hybridization between races. For example, tetraploids may have arisen from syngamy between unreduced gametes of two *A. candida* races. Using PacBio (Pacific Biosciences) to generate long reads from targeted or whole genome sequences of *A. candida* would help reconstitute haplotypes which could in turn enable the identification of putative parental races of polyploids.

6.4 CONCLUSIONS AND OUTLOOK

Although the existence of highly host-specific *A. candida* races has been known for a long time, the extent of genetic diversity in the *A. candida* complex was left completely unexplored up until very recently. Using targeted sequencing, I could cost-effectively generate useful data from many *A. candida* isolates and develop our understanding of the mechanisms by which *A. candida* races evolve.

Probably one of the most exciting aspect of this work is that it highlights the fact that *A. candida* races may be even more diverged than expected. Indeed, not only were they found to be ~0.7% diverged but they also seemed to differ in ploidy level and in the way they reproduce in nature. This information is not only crucial for the comprehension of (plant) pathogen evolution but it is also of fundamental importance for appropriate disease management (Burdon & Thrall 2008). For example, the fact that agricultural hosts are infected by polyploids which potential for sexual reproduction, and therefore the production of resting oospores, is reduced (especially in triploids) could be used to argue in favour of a crop rotation strategy. It also calls for a new study to investigate the mechanisms by which *A. candida* races infecting *B. oleracea* and *R. sativus* survive intercrop seasons without the production of resting oospores. Sampling and sequencing of wild Brassicaceae associated with these crops would be the step forward.

Moreover, this thesis demonstrates that there is much to be discovered about plant pathogen evolution, and that time should be taken to explore that of diverse pathogenic species even though they may not represent an immediate economic interest or danger for agricultural

yields. Although not of great threat to agriculture in Europe, *A. candida* is pathogenic to many hosts that may be profitable in the future. For example, *A. camelina* is considered an emerging source of biofuel (Kagale *et al.* 2014).

Finally, I believe this work is important because it paves the way for a new method to study microbial diversity directly from infected plant or animal hosts. This is interesting for organisms that cannot be cultured in the laboratory such as *A. candida* but also many microorganisms. In addition, this technique may allow the study of the molecular interactions in play between multiple organisms outside of the laboratory and in a cost-effective way. It has great potential, for example, for the investigation of the co-evolutionary mechanisms between a pathogen and its host and also the impact of a pathogen on the microbial community of its host.

REFERENCES

Achtman M, Wagner M (2008) Microbial diversity and the genetic nature of microbial species. *Nature reviews. Microbiology*, **6**, 431–440.

Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology*, **8**, 135–141.

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**, 959–65.

Agler MT, Ruhe J, Kroll S *et al.* (2016) Microbial Hub Taxa Link Host and Abiotic Factors to Plant Microbiome Variation. *PLOS Biology*, **14**, e1002352.

Albertin W, Marullo P (2012) Polyploidy in fungi: evolution after whole-genome duplication. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 2497–2509.

Allen RL, Bittner-Eddy PD, Grenville-Briggs LJ *et al.* (2004) Host-Parasite Coevolutionary Conflict Between Arabidopsis and Downy Mildew. *Science*, **306**, 1957–1960.

Almeida NF, Yan S, Cai R *et al.* (2010) PAMDB, a multilocus sequence typing and analysis database and website for plant-associated microbes. *Phytopathology*, **100**, 208–15.

Amann RI, Ludwig W, Schleifer K-H (1995) Phylogenetic Identification and In Situ Detection of Individual Microbial Cells without Cultivation. *Microbiological Reviews*, **59**, 143–169.

Amselem J, Cuomo CA, van Kan JAL *et al.* (2011) Genomic analysis of the necrotrophic fungal pathogens sclerotinia sclerotiorum and botrytis cinerea. *PLoS Genetics*, **7**.

Anderson PK, Cunningham AA, Patel NG *et al.* (2004) Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. *Nature*, **19**.

Anderson JP, Gleason CA, Foley RC *et al.* (2011) Plant versus pathogens: an evolutionary arms race. *Functional Plant Biology*, **37**, 499–512.

Asch RG, Collie JS (2008) Changes in a benthic megafaunal community due to disturbance from bottom fishing and the establishment of a fishery closure. *Fishery Bulletin*, **106**, 438–456.

Bailey-Serres J, Mittler R (2006) The roles of reactive oxygen species in plant cells. *Plant Physiology*, **141**, 900191.

Baker GC, Smith JJ, Cowan D a. (2003) Review and re-analysis of domain-specific 16S primers. *Journal of Microbiological Methods*, **55**, 541–555.

Balloux F, Lehmann L, De Meeus T (2003) The population genetics of clonal and partially

clonal diploids. *Genetics*, **164**, 1635–1644.

Baltrus DA, Nishimura MT, Romanchuk A *et al.* (2011) Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 pseudomonas syringae isolates. *PLoS Pathogens*, **7**.

Barbetti MJ (1981) Effects of sowing date and oospore seed contamination upon subsequent crop incidence of white rust (Albugo candida) in rapeseed. *Australasian Plant Pathology*, **10**, 44–46.

Barrett LG, Thrall PH, Dodds PN *et al.* (2009) Diversity and evolution of effector loci in natural populations of the plant pathogen melampsora lini. *Molecular Biology and Evolution*, **26**, 2499–2513.

de Bary A (1863) Recherches sur le developpement de quelques champignons parasites. *Annales des sciences naturelles, Botanique*, **20**.

Baxter L, Tripathy S, Ishaque N *et al.* (2010) Signatures of adaptation to obligate biotrophy in the Hyaloperonospora arabidopsidis genome. *Science (New York, N.Y.)*, **330**, 1549–1551.

Beakes GW, Glockling SL, Sekimoto S (2012) The evolutionary phylogeny of the oomycete " fungi ." *Protoplasma*, **249**, 3–19.

Beakes GW, Honda D, Thines M (2014) Systematics of the stramenopila: Labyrinthulomycota, Hyphochytriomycota and Oomycota. In: *Systematics and Evolution* (eds McLaughlin DJ, Spatafora JW), pp. 39–97. Springer-Verlag, Berlin Heidelberg.

Belhaj K, Cano LM, Prince DC *et al.* (2015) Arabidopsis late blight: Infection of a nonhost plant by Albugo laibachii enables full colonization by Phytophthora infestans. *bioRxiv*.

Bennett RJ, Forche A, Berman J (2014) Rapid mechanisms for generating genome diversity: whole ploidy shifts, aneuploidy, and loss of heterozygosity. *Cold Spring Harbor perspectives in medicine*, **4**.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2009) GenBank. *Nucleic Acids Research*, **37**, D26–D31.

Bent AF, Mackey D (2007) Elicitors, effectors, and R genes: the new paradigm and a lifetime supply of questions. *Annual review of phytopathology*, **45**, 399–436.

Berendsen RL, Pieterse CMJ, Bakker PAHM (2012) The rhizosphere microbiome and plant health. *Trends in Plant Science*, **17**, 478–486.

Bernier CC (1972) Diseases of rapeseed in Manitoba in 1971. *Canadian Plant Disease Survey*, **52**, 108.

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

Billiard S, Lopez-Villavicencio M, Hood ME, Giraud T (2012) Sex, outcrossing and mating types: Unsolved questions in fungi and beyond. *Journal of Evolutionary Biology*, **25**, 1020–1038.

Birky CW (1996) Heterozygosity, heteromorphy, and phylogenetic trees in asexual eukaryotes. *Genetics*, **144**, 427–437.

Bock CH, Thrall PH, Burdon JJ (2005) Genetic structure of populations of Alternaria brassicicola suggests the occurrence of sexual recombination. *Mycological research*, **109**, 227–236.

Borhan MH, Gunn N, Cooper AJ *et al.* (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*. *Molecular Plant-Microbe Interactions*, **21**, 757–768.

Bouvet P, Ferraris L, Dauphin B *et al.* (2014) 16S rRNA gene sequencing, multilocus sequence analysis, and mass spectrometry identification of the proposed new species "Clostridium neonatale." *Journal of clinical microbiology*, **52**, 4129–4136.

Brasier CM (2001) Rapid Evolution of Introduced Plant Pathogens via Interspecific Hybridization. *BioScience*, **51**, 123.

Brasier CM, Cooke DE, Duncan JM (1999) Origin of a new Phytophthora pathogen through interspecific hybridization. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 5878–5883.

Brasier CM, Kirk S a, Delcan J *et al.* (2004) Phytophthora alni sp. nov. and its variants: designation of emerging heteroploid hybrid pathogens spreading on Alnus trees. *Mycological research*, **108**, 1172–1184.

Buell CR, Joardar V, Lindeberg M *et al.* (2003) The complete genome sequence of the Arabidopsis and tomato pathogen Pseudomonas syringae pv. tomato DC3000. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 10181–10186.

Burdon JJ, Thrall PH (2008) Pathogen evolution across the agro-ecological interface: implications for disease management. *Evolutionary applications*, **1**, 57–65.

Burton TL, Husband BC (2001) Fecundity and offspring ploidy in matings among diploid, triploid and tetraploid Chamerion angustifolium (Onagraceae): Consequences for tetraploid establishment. *Heredity*, **87**, 573–582.

Castric V, Bechsgaard J, Schierup MH, Vekemans X (2008) Repeated adaptive introgression at a gene under multiallelic balancing selection. *PLoS genetics*, **4**, e1000168.

Chaparro-Garcia A, Wilkinson RC, Gimenez-Ibanez S *et al.* (2011) The receptor-like kinase serk3/bak1 is required for basal resistance against the late blight pathogen Phytophthora infestans in Nicotiana benthamiana. *PLoS ONE*, **6**.

Chase MW, Fay MF (2009) Barcoding of Plants and Fungi. *Science*, **325**, 682–683.

Chen ZJ (2010) Molecular mechanisms of polyploidy and hybrid vigor. *Trends in plant science*, **15**, 57–71.

Chen XH, Koumoutsi A, Scholz R *et al.* (2007) Comparative analysis of the complete genome sequence of the plant growth–promoting bacterium Bacillus amyloliquefaciens FZB42. *Nature Biotechnology*, **25**, 1007–1014.

Chen H-T, Wu Y-C, Chen S-T, Tsai H-C, Chien Y-C (2014) Androgen receptor CAG repeats, haplotypes, non-random X chromosome inactivation, and LOH at Xq25 in relation to breast cancer risk. *BMC Cancer*, **14**, 144.

Chisholm ST, Coaker G, Day B, Staskawicz BJ (2006) Host-microbe interactions: Shaping the evolution of the plant immune response. *Cell*, **124**, 803–814.

Choi Y-J, Hong S-B, Shin H-D (2006) Genetic diversity within the Albugo candida complex (Peronosporales, Oomycota) inferred from phylogenetic analysis of ITS rDNA and COX2 mtDNA sequences. *Molecular phylogenetics and evolution*, **40**, 400–9.

Choi SY, Kim S, Lyuck S, Kim SB, Mitchell RJ (2015) High-level production of violacein by the newly isolated Duganella violaceinigra str. NI28 and its impact on Staphylococcus aureus. *Scientific Reports*, **5**, 15598.

Choi D, Priest MJ (1995) A key to the genus Albugo. *Mycotaxon*, **53**, 261–272.

Choi Y-J, Shin H-D, Hong S-B, Thines M (2007) Morphological and molecular discrimination among Albugo candida materials infecting Capsella bursa - pastoris world-wide. *Fungal Diversity*, **27**, 11–34.

Choi Y-J, Shin H-D, Ploch S, Thines M (2008) Evidence for uncharted biodiversity in the Albugo candida complex, with the description of a new species. *Mycological Research*, **112**, 1327–1334.

Choi Y-J, Shin H-D, Ploch S, Thines M (2011) Three new phylogenetic lineages are the closest relatives of the widespread species Albugo candida. *Fungal biology*, **115**, 598–607.

Choi Y-J, Shin H-D, Thines M (2009) The host range of Albugo candida extends from Brassicaceae through Cleomaceae to Capparaceae. *Mycological Progress*, **8**, 329–335.

Cohan FM (2002) What are Bacterial Species? *Annual Review of Microbiology*, **56**, 457–487.

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

Cooke DE, Cano LM, Raffaele S *et al.* (2012) Genome analyses of an aggressive and invasive lineage of the Irish potato famine pathogen. *PLoS pathogens*, **8**, e1002940.

Cooke DE, Williams NA, Williamson B, Duncan JM (2002) An Its-Based Phylogenetic Analysis of the Relationships Between Peronospora and Phytophthora. In: *Advances in downy mildew research* (eds Spencer-Phillips PTN, Gisi U, Lebeda A), pp. 161–165. Kluwer, New York, Boston, Dodrecht, London, Moscow.

Cooper AJ, Latunde-Dada AO, Woods-Tor A *et al.* (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.

Couch BC, Fudal I, Lebrun MH *et al.* (2005) Origins of host-specific populations of the blast pathogen Magnaporthe oryzae in crop domestication with subsequent expansion of pandemic clones on rice and weeds of rice. *Genetics*, **170**, 613–630.

Cunniffe NJ, Gilligan C a (2010) Invasion, persistence and control in epidemic models for plant pathogens: the effect of host demography. *Journal of the Royal Society Interface*, **7**, 439–451.

Danecek P, Auton A, Abecasis GR *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Danies G, Myers K, Mideros MF *et al.* (2014) An ephemeral sexual population of Phytophthora infestans in the Northeastern United States and Canada. *PloS one*, **9**, e116354.

Daszak P, Cunningham AA, Hyatt AD (2000) Emerging infectious diseases of wildlife - threats to biodiversity and human health. *Science*, **287**, 443–449.

Dean R, Van Kan J a L, Pretorius Z a. *et al.* (2012) The Top 10 fungal pathogens in molecular plant pathology. *Molecular Plant Pathology*, **13**, 414–430.

Dean R, Talbot NJ, Ebbole DJ *et al.* (2005) The genome sequence of the rice blast fungus Magnaporthe grisea. *Nature*, **434**, 980–986.

Degnan JH, Rosenberg N a (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in ecology & evolution*, **24**, 332–40.

Doroghazi JR, Buckley DH (2010) Widespread homologous recombination within and between Streptomyces species. *The ISME journal*, **4**, 1136–43.

Drancourt M, Berger P, Raoult D (2004) Systematic 16S rRNA Gene Sequencing of Atypical

Clinical Isolates Identified 27 New Bacterial Species Associated with Humans. *Journal of Clinical Microbiology*, **42**, 2197–2202.

Dröge J, Gregor I, Mchardy AC (2014) Taxator-tk: Precise Taxonomic Assignment of Metagenomes by Fast Approximation of Evolutionary Neighborhoods. *Bioinformatics*.

Eberhardt A (1904a) Contribution a letude de Cystopus candidus Lev. *Zentralbl Bakteriol Parasitenkd Infektionskr Hyg Abt*, **12**, 235–249.

Eberhardt A (1904b) Contribution a letude de Cystopus candidus Lev. *Zentralbl Bakteriol Parasitenkd Infektionskr Hyg Abt*, **12**, 426–439.

Eberhardt A (1904c) Contribution a letude de Cystopus candidus Lev. *Zentralbl Bakteriol Parasitenkd Infektionskr Hyg Abt*, **12**, 614–631.

Eberhardt A (1904d) Contribution a letude de Cystopus candidus Lev. *Zentralbl Bakteriol Parasitenkd Infektionskr Hyg Abt*, **12**, 714–727.

Ellison A, Cable J, Consuegra S (2011) Best of both worlds? association between outcrossing and parasite loads in a selfing fish. *Evolution*, **65**, 3021–3026.

Felsenstein J (1974) The evolutionary advantage of recombination. *Genetics*, **78**, 737–756.

Ferrante P, Scortichini M (2010) Molecular and phenotypic features of Pseudomonas syringae pv. actinidiae isolated during recent epidemics of bacterial canker on yellow kiwifruit (Actinidia chinensis) in central Italy. *Plant Pathology*, **59**, 954–962.

Fisher MC, Henk DA, Briggs CJ *et al.* (2012) Emerging fungal threats to animal, plant and ecosystem health. *Nature*, **484**.

Forche A, Abbey D, Pisithkul T *et al.* (2011) Stress alters rates and types of loss of heterozygosity in candida albicans. *mBio*, **2**, 1–9.

Furzer O (2014) The Genetic Basis of Resistance and Susceptibility in the Albugo laibachii -- Arabidopsis thaliana pathosystem.

Gavino PD, Smart CD, Sandrock RW *et al.* (2000) Implications of sexual reproduction for Phytophthora infestans in the United States: Generation of an aggressive lineage. *Plant Disease*, **84**, 731–735.

Gest H (2004) The discovery of microorganisms by Robert Hooke and Antoni Van Leeuwenhoek, fellows of the Royal Society. *Notes and records of the Royal Society of London*, **58**, 187–201.

Gevers D, Coince A, Lawrence JG *et al.* (2005) Re-evaluating prokaryotic species. *Nature Reviews Microbiology*, **3**, 733–739.

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

Gmelin JF (1792) *Systema Naturae*. Beer, G E, Leipzig.

Gnirke A, Melnikov A, Maguire J *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, **27**, 182–189.

Goelet P, Lomonossoff GP, Butler PJ *et al.* (1982) Nucleotide sequence of tobacco mosaic virus RNA. *Proceedings of the National Academy of Sciences of the United States of America*, **79**, 5818–22.

Gonthier P, Nicolotti G, Linzer R, Guglielmo F, Garbelotto M (2007) Invasion of European pine stands by a North American forest pathogen and its hybridization with a native interfertile taxon. *Molecular Ecology*, **16**, 1389–1400.

Goss EM, Carbone I, Grünwald NJ (2009) Ancient isolation and independent evolution of the three clonal lineages of the exotic sudden oak death pathogen Phytophthora ramorum. *Molecular Ecology*, **18**, 1161–1174.

Gout L, Kuhn ML, Vincenot L *et al.* (2007) Genome structure impacts molecular evolution at the AvrLm1 avirulence locus of the plant pathogen Leptosphaeria maculans. *Environmental Microbiology*, **9**, 2978–2992.

Graham JH, Gottwald TR, Cubero J, Achor DS (2004) Xanthomonas axonopodis pv . citri : factors affecting successful erradication of citrus canker. *Molecular and Plant Pathology*, **5**, 1–15.

Gray SF (1821) *A natural arrangement of British plants, according to their relations to each other as pointed out by Jussieu, De Candolle, Brown, etc*. Baldwin, Cradock and Joy, London.

Grünwald NJ, Hoheisel G-A (2006) Hierarchical Analysis of Diversity, Selfing, and Genetic Differentiation in Populations of the Oomycete Aphanomyces euteiches. *Phytopathology*, **96**, 1134–41.

Ha G, Roth A, Lai D *et al.* (2012) Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Research*, **22**, 1995–2007.

Halary S, Malik SB, Lildhar L *et al.* (2011) Conserved meiotic machinery in Glomus spp., a putatively ancient asexual fungal lineage. *Genome Biology and Evolution*, **3**, 950–958.

Han JI, Choi HK, Lee SW *et al.* (2011) Complete genome sequence of the metabolically versatile plant growth-promoting endophyte Variovorax paradoxus S110. *Journal of Bacteriology*, **193**, 1183–1190.

Hanage WP, Fraser C, Spratt BG (2006) Sequences, sequence clusters and bacterial species.

Philosophical Transactions of the Royal Society B: Biological Sciences, **361**, 1917–1927.

Hardenbol P, Banér J, Jain M *et al.* (2003) Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nature biotechnology*, **21**, 673–8.

Hartl DL, Clark AG (1997) *Principles of population genetics*. Sinauer, Sunderland.

Hawksworth DL, Rossman AY (1997) Where Are All the Undescribed Fungi ? *Phytopathology*, **87**, 888–891.

Head SR, Kiyomi Komori H, LaMere SA *et al.* (2014) Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, **56**, 61–77.

Hegarty M, Hiscock S (2007) Polyploidy: Doubling up for Evolutionary Success. *Current Biology*, **17**, 927–929.

Heitman J (2006) Sexual Reproduction and the Evolution of Microbial Pathogens. *Current Biology*, **16**, 711–725.

Heitman J (2010) Evolution of eukaryotic microbial pathogens via covert sexual reproduction. *Cell Host and Microbe*, **8**, 86–99.

Heller A, Thines M, Money NP (2009) Evidence for the importance of enzymatic digestion of epidermal walls during subepidermal sporulation and pustule opening in white blister rusts (Albuginaceae). *Mycological Research*, **113**, 657–667.

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.

Hogenhout S a, Van der Hoorn R a L, Terauchi R, Kamoun S (2009) Emerging concepts in effector biology of plant-associated organisms. *Molecular plant-microbe interactions : MPMI*, **22**, 115–122.

Holub EB (2006) Evolution of parasitic symbioses between plants and filamentous microorganisms. *Current Opinion in Plant Biology*, **9**, 397–405.

Hongoh Y, Yuzawa H, Ohkuma M, Kudo T (2003) Evaluation of primers and PCR conditions for the analysis of 16S rRNA genes from a natural environment. *FEMS Microbiology Letters*, **221**, 299–304.

Hudspeth DSS, Stenger D, Hudspeth MES (2003) A cox2 phylogenetic hypothesis for the downy mildews and white rusts. *Fungal Diversity*, 47–57.

Huson D, Auch A, Qi J, Schuster S (2007) MEGAN analysis of metagenome data. *Gennome*

Res., **17**, 377–386.

Huson DH, Bryant D (2006) Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolutin*, **23**, 254–267.

Hwang S-FS, Strelkov SES, Feng J, Gossen BD, Howard RJ (2012) Plasmodiophora brassicae: a review of an emerging pathogen of the Canadian canola (Brassica napus) crop. *Molecular plant pathology*, **13**, 105–13.

Inderbitzin P, Davis RM, Bostock RM, Subbarao K V. (2011) The ascomycete Verticillium longisporum is a hybrid and a plant pathogen with an expanded host range. *PLoS ONE*, **6**, e18260.

Ishaque N (2012) An Investigation into the Signatures of Evolution in Pathogen Effector Genes.

James M, Mallet J, James M (2007) Hybrid speciation. *Nature*, **446**, 279–283.

Jiang RHY, Tyler BM (2012) Mechanisms and Evolution of Virulence in Oomycetes. In: *Annual Review of Phytopathology, Vol 50*, pp. 295–318.

Johnson JL (1973) Use of nucleic-acid homologies in the taxonomy of anaerobic bacteria. *Int J Syst Bacteriol*, **23**, 308–315.

Jones JDG, Dangl JL (2006) The plant immune system. *Nature Reviews*, **444**, 323–329.

Jones JT, Haegeman A, Danchin EGJ *et al.* (2013) Top 10 plant-parasitic nematodes in molecular plant pathology. *Molecular Plant Pathology*, **14**, 946–961.

de Jonge R, Bolton MD, Kombrink A *et al.* (2013) Extensive chromosomal reshuffling drives evolution of virulence in an asexual pathogen. *Genome research*, **23**, 1271–82.

Jouet A, McMullan M, Van Oosterhout C (2015) The effects of recombination, mutation and selection on the evolution of the Rp1 resistance genes in grasses. *Molecular Ecology*, **24**, 3077–3092.

Jousimo J, Tack AJM, Ovaskainen O *et al.* (2013) Ecological and evolutionary effects of fragmentation on infectious disease dynamics. *Science*, **15153**, 1289–1294.

Judson OP, Normark BB (1996) Ancient asexual scandals. *Trends in Ecology and Evolution*, **11**, 41–46.

Jupe F, Witek K, Verweij W *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. *The Plant Journal*, **76**, 530–544.

Kagale S, Koh C, Nixon J *et al.* (2014) The emerging biofuel crop Camelina sativa retains a highly undifferentiated hexaploid genome structure. *Nature communications*, **5**, 3706.

Kalra A, Grover RK, Rishi N (1989) Interaction between Phytophthora infestans and potato viruses X and Y in potato. *The Journal of Agricultural Science*, **112**, 33–37.

Kamoun S (2001) Nonhost resistance to Phytophthora: Novel prospects for a classical problem. *Current Opinion in Plant Biology*, **4**, 295–300.

Kamoun S, Furzer O, Jones JDG *et al.* (2015) The Top 10 oomycete pathogens in molecular plant pathology. *Molecular Plant Pathology*, **16**, 413–434.

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 (2014) Microbe–microbe interactions determine oomycete and fungal host colonization. *Current Opinion in Plant Biology*, **20**, 75–81.

Kemen E, Gardiner A, Schultz-Larsen T *et al.* (2011) Gene gain and loss during evolution of obligate parasitism in the white rust pathogen of Arabidopsis thaliana. *PLoS biology*, **9**, e1001094.

Klein J, Sato A, Nagl S, O'hUigin C (1998) Molecular trans-species polymorphism. *Annual Review of Ecology and Systematics*, **29**, 1–21.

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.

Kolte SJ, Sharma KD, Awasthi RP (1981) Yield losses and control of downy mildew and white rust of rapeseed and mustard. In: *Third international symposium plant pathology. December 14–18, IARI, New Delhi, India. Session 12*, pp. 70–71.

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 8369–8374.

Kuhn G, Hijri M, Sanders I (2001) Evidence for the evolution of multiple genomes in arbuscular mycorrhizal fungi. *Nature*, **414**, 745–748.

Lakra BS, Saharan GS (1989a) Correlation of leaf and staghead infection intensities of white rust with yield and yield components of mustard. *Indian Journal of Mycology and Plant Pathology*, **19**, 279–281.

Lakra BS, Saharan GS (1989b) Location and estimation of oospores of Albugo candida in infected plant parts of mustard. *Indian Phytopathology*, **42**, 467.

Lamour KH, Mudge J, Gobena D *et al.* (2012a) 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.

Lamour KH, Stam R, Jupe J, Huitema E (2012b) The oomycete broad-host-range pathogen Phytophthora capsici. *Molecular Plant Pathology*, **13**, 329–337.

Lederberg J, Shope RE, Oaks SC (Eds.) (1992) *Emerging Infections: Microbial Threats to Health in the United States*. Institute of Medicine, National Academy Press.

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

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

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics (Oxford, England)*, **25**, 1451–2.

Links MG, Dumonceaux TJ, Hemmingsen SM, Hill JE (2012) The chaperonin-60 universal target is a barcode for bacteria that enables de novo assembly of metagenomic sequence data. *PloS one*, **7**, e49755.

Links MG, Holub EB, Jiang RHY *et al.* (2011) De novo sequence assembly of Albugo candida reveals a small genome relative to other biotrophic oomycetes. *BMC Genomics*, **12**.

Liu J, Rimmer SR (1993) Production and germination of oospores of Albugo candida. *Canadian Journal of Plant Pathology*, **15**, 265–271.

Lugtenberg B, Kamilova F (2009) Plant-Growth-Promoting Rhizobacteria. *Annual Review of Microbiology*, **63**, 541–556.

Luna E, Pastor V, Robert J *et al.* (2011) Callose deposition: a multifaceted plant defense response. *Molecular plant-microbe interactions : MPMI*, **24**, 183–193.

Macheras E, Roux AL, Bastian S *et al.* (2011) Multilocus sequence analysis and rpoB sequencing of Mycobacterium abscessus (sensu lato) strains. *Journal of Clinical Microbiology*, **49**, 491–499.

Maddison WP (1997) Gene Trees in Species Trees. *Systematic Biology*, **46**, 523–536.

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

Mamanova L, Coffey AJ, Scott CE *et al.* (2010) Target-enrichment strategies for next-generation sequencing. *Nature Methods*, **7**, 111–118.

Manzanera M, Narváez-Reinaldo JJ, García-Fontana C, Vílchez JI, González-López J (2015) Genome Sequence of Arthrobacter koreensis 5J12A, a Plant Growth-Promoting and Desiccation-Tolerant Strain. *Genome Announcements*, **3**, e00648-15.

McMullan M, Gardiner A, Bailey K *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**.

Meekes ETM, Jeger MJ, Raaijmakers JM (2004) Host specialisation of the oomycete Albugo candida. In: *Advances in downy mildew research, Vol. 2*, pp. 119–139.

Meena PD, Sharma P (2012) Methodology for prodution and germination of oospores of Albugo candida infecting oilseed Brassica. *Vegetos*, **25**, 115–119.

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**.

Menardo F, Praz C, Wyder S *et al.* (2015) Hybridization of powdery mildew strains gives raise to pathogens on novel agricultural crop species. *Nature genetics*, **48**, 1–24.

Mendes R, Kruijt M, de Bruijn I *et al.* (2011) Deciphering the Rhizosphere Microbiome for Disease-Suppressive Bacteria. *Science*, **332**, 1097–1100.

Mertes F, ElSharawy A, Sauer S *et al.* (2011) Targeted enrichment of genomic DNA regions for next-generation sequencing. *Briefings in Functional Genomics*, **10**, 374–386.

Meyer M, Kircher M (2010) Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harbor Protocols*, **2010**, pdb.prot5448-prot5448.

Meyer MD, Pataky JK (2010) Increased Severity of Foliar Diseases of Sweet Corn Infected with Maize Dwarf Mosaic and Sugarcane Mosaic Viruses. *Plant Disease*, **94**, 1093–1099.

Nadeau NJ, Whibley A, Jones RT *et al.* (2012) Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **367**, 343–53.

Napper ME (1933) Observations on spore germination and specialization of parasitism in Cystopus candidus. *Journal of Pomological and Horticultural Science*, **11**, 81–100.

Naser SM, Thompson FL, Hoste B *et al.* (2005) Application of multilocus sequence analysis (MLSA) for rapid identification of Enterococcus species based on rpoA and pheS genes.

Microbiology (Reading, England), **151**, 2141–50.

Nath MD, Sharma SL, Kant U (2000) Growth of Albugo candida infected mustard callus in culture. *Mycopathologia*, **152**, 147–153.

O'Connell RJ, Panstruga R (2006) Tete a tete inside a plant cell: Establishing compatibility between plants and biotrophic fungi and oomycetes. *New Phytologist*, **171**, 699–718.

Oberhofer M, Leuchtmann A (2012) Genetic diversity in epichloid endophytes of Hordelymus europaeus suggests repeated host jumps and interspecific hybridizations. *Molecular ecology*, **21**, 2713–26.

Olesen KL, Carver TLW, Lyngkj??r 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.

Orbach MJ, Farrall L, Sweigard JA, Chumley FG, Valent B (2000) A telomeric avirulence gene determines efficacy for the rice blast resistance gene Pi-ta. *The Plant cell*, **12**, 2019–32.

Orr HA, Turelli M (2001) The Evolution of Postzygotic Isolation : Accumulating Dobzhansky-Muller Incompatibilities. *Evolution*, **55**, 1085–1094.

Otto SP (2007) The evolutionary consequences of polyploidy. *Cell*, **131**, 452–62.

Otto SP (2009) The Evolutionary Enigma of Sex. *The American naturalist*, **174**, S1–S14.

Otto SP, Whitton J (2000) Polyploid incidence and evolution. *Annual Review of Genetics*, **34**, 401–437.

Pannell JR, Charlesworth B (2000) Effects of metapopulation processes on measures of genetic diversity. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **355**, 1851–1864.

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.

Pedersen BS, De S (2013) Loss of heterozygosity preferentially occurs in early replicating regions in cancer genomes. *Nucleic Acids Research*, **41**, 7615–7624.

Penczykowski RM, Laine AL, Koskella B (2016) Understanding the ecology and evolution of host-parasite interactions across scales. *Evolutionary Applications*, **9**.

Persoon CH (1801) *Synopsis methodica fungorum. Part I, II.* Gottingen.

Peters RD, Al-Mughrabi KI, Kalischuk ML *et al.* (2014) Characterization of Phytophthora infestans population diversity in Canada reveals increased migration and genotype recombination. *Canadian Journal of Plant Pathology*, **36**, 73–82.

Petersen AB, Rosendahl Sø (2000) Phylogeny of the Peronosporomycetes (Oomycota) based

on partial sequences of the large ribosomal subunit (LSU rDNA). *Mycological Research*, **104**, 1295–1303.

Petkowski JE, Cunnington JH, Minchinton EJ, Cahill DM (2010) Molecular phylogenetic relationships between Albugo candida collections on the Brassicaceae in Australia. *Plant Pathology*, **59**, 282–288.

Petrie GA (1988) Races of Albugo candida (white rustand staghead) on cultivated Cruciferae in Saskatchewan. *Canadian Journal of Plant Pathology*, **10**, 142–150.

Petrie GA, Vanterpool TC (1974) Fungi associated with hypertrophies caused by infection of Cruciferae by A. cruciferatum. *Canadian Plant Disease Survey*, **54**, 37–42.

Pfeiffer W, Stamatakis A (2010) Hybrid MPI/Pthreads Parallelization of the RAxML Phylogenetics Code. In: *Accepted for publication at HICOMB workshop, held in conjunction with IPDPS 2010*, p. . Atlanta, Georgia.

Ploch S, Choi Y-J, Rost C *et al.* (2010) Evolution of diversity in Albugo is driven by high host specificity and multiple speciation events on closely related Brassicaceae. *Molecular phylogenetics and evolution*, **57**, 812–20.

Pontes DS, Lima-Bittencourt CI, Chartone-Souza E, Amaral Nascimento AM (2007) Molecular approaches: Advantages and artifacts in assessing bacterial diversity. *Journal of Industrial Microbiology and Biotechnology*, **34**, 463–473.

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

del Pozo JC, Ramirez-Parra E (2015) Whole genome duplications in plants: an overview from Arabidopsis. *Journal of Experimental Botany*, erv432.

Prospero S, Grünwald NJ, Winton LM, Hansen EM (2009) Migration patterns of the emerging plant pathogen Phytophthora ramorum on the West Coast of the United States of America. *Phytopathology*, **99**, 739–49.

Quail MA, Kozarewa I, Smith F *et al.* (2008) A large genome center's improvements to the Illumina sequencing system. *Nature Methods*, **5**, 1005–1010.

Rabeling C, Gonzales O, Schultz TR *et al.* (2011) Cryptic sexual populations account for genetic diversity and ecological success in a widely distributed, asexual fungus-growing ant. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 12366–12371.

R Core Team (2014) R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.*

Raffaele S, Farrer RA, Cano LM *et al.* (2010) Genome Evolution Following Host Jumps in

the Irish Potato Famine Pathogen Lineage. *Science*, **330**, 1540–1543.

Rieseberg LH, Van Fossen C, Desrochers AM (1995) Hybrid speciation accompanied by genomic reorganization in wild sunflowers. *Nature*, **375**, 313–316.

Riethmüller A, Voglmayr H, Göker M, Weiß M, Oberwinkler F (2002) Phylogenetic relationships of the downy mildews (Peronosporales) and related groups based on nuclear large subunit ribosomal DNA sequences. *Mycologia*, **94**, 834–49.

Rietman H, Bijsterbosch G, Cano LM *et al.* (2012) Qualitative and quantitative late blight resistance in the potato cultivar Sarpo Mira is determined by the perception of five distinct RXLR effectors. *Molecular Plant-Microbe Interactions*, **25**, 910–9.

Rimmer SR, Mathur S, Wu CR (2000) Virulence of isolates of Albugo candida from western Canada to Brassica species. *Canadian Journal of Plant Pathology*, **22**, 235.

Robideau GP, De Cock AW a M, Coffey MD *et al.* (2011) DNA barcoding of oomycetes with cytochrome c oxidase subunit I and internal transcribed spacer. *Molecular ecology resources*, **11**, 1002–11.

Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, **22**, 939–946.

Roussel HFA (1806) *Flore du Calvados et des terreins adjacens*.

Rouxel M, Mestre P, Comont G *et al.* (2013) Phylogenetic and experimental evidence for host-specialized cryptic species in a biotrophic oomycete. *New Phytologist*, **197**, 251–263.

Runge F, Ndambi B, Thines M (2012) Which Morphological Characteristics Are Most Influenced by the Host Matrix in Downy Mildews? A Case Study in Pseudoperonospora cubensis. *PLoS ONE*, **7**.

Saharan GS, Verma PR (1992) *White rusts: a review of economically important species*. Ottawa, Ontario: International Development Research Centre.

Saharan GS, Verma PR, Meena PD, Kumar A (2014) *White Rust of Crucifers: Biology, Ecology and Management*. Springer.

Salanoubat M, Genin S, Artiguenave F *et al.* (2002) Genome sequence of the plant pathogen Ralstonia solanacearum. *Nature*, **415**, 497–502.

Sandhu PS, Brar KS, Chauhan JS *et al.* (2015) Host–pathogen interactions of Brassica genotypes for white rust (Albugo candida) disease severity under aided epiphytotic conditions in India. *Phytoparasitica*, **43**, 197–207.

Sansome E, Sansome FW (1974) Cytology and life-history of Peronospora parasitica on Capsella bursa-pastoris and of Albugo candida on C. bursa-pastoris and on Lunaria

annua. *Transactions of the Britishi Mycological Society*, **62**, 323–332.

Saunders GW, Mcdevit DC (2012) *DNA Barcodes (Chapter 10)* (WJ Kress, DL Erickson, Eds,). Humana Press, Totowa, NJ.

Saxena NP (2010) *Objective botany*. Krishna Prakashan Media.

Schardl CL, Craven KD (2003) Interspecific hybridization in plant-associated fungi and oomycetes: A review. *Molecular Ecology*, **12**, 2861–2873.

Schoch CL, Seifert K a., Huhndorf S *et al.* (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences*, **109**, 6241–6246.

Scholthof KBG, Adkins S, Czosnek H *et al.* (2011) Top 10 plant viruses in molecular plant pathology. *Molecular Plant Pathology*, **12**, 938–954.

Schröder S, Telle S, Nick P, Thines M (2011) Cryptic diversity of Plasmopara viticola (Oomycota, Peronosporaceae) in North America. *Organisms Diversity and Evolution*, **11**, 3–7.

Schurko AM, Neiman M, Logsdon JM (2009) Signs of sex: what we know and how we know it. *Trends in Ecology and Evolution*, **24**, 208–217.

Schwander T, Henry L, Crespi BJ (2011) Molecular evidence for ancient asexuality in timema stick insects. *Current Biology*, **21**, 1129–1134.

Schwenk K, Brede N, Streit B (2008) Introduction. Extent, processes and evolutionary impact of interspecific hybridization in animals. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **363**, 2805–2811.

Seehausen O (2004) Hybridization and adaptive radiation. *Trends in Ecology & Evolution*, **19**, 198–207.

Seidl MF, Thomma BPHJ (2014) Sex or no sex: Evolutionary adaptation occurs regardless. *BioEssays*, **36**, 335–345.

Shearer AE, Hildebrand MS, Ravi H *et al.* (2012) Pre-capture multiplexing improves efficiency and cost-effectiveness of targeted genomic enrichment. *BMC Genomics*, **13**, 618.

Shröter J (1893) Peronosporinae. In: *Die Natürlichen Pflanzenfamilien*, pp. 108–119. Engler and Prantl, Leipzig.

Signorovitch AY, Dellaporta SL, Buss LW (2005) Molecular signatures for sex in the Placozoa. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 15518–22.

Singh RP, Hodson DP, Huerta-Espino J *et al.* (2011) The Emergence of Ug99 Races of the

Stem Rust Fungus is a Threat to World Wheat Production. *Annual review of phytopathology*, **49**, 465–481.

Soltis PS, Soltis DE (2000) The role of genetic and genomic attributes in the success of polyploids. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 7051–7.

Soltis PS, Soltis DE (2009) The Role of Hybridization in Plant Speciation. *Annual Review of Plant Biology*, **60**, 561–588.

Soltis DE, Soltis PS, Tate J a. (2004) Advances in the study of polyploidy since Plant speciation. *New Phytologist*, **161**, 173–191.

Stahl E a, Bishop JG (2000) Plant – pathogen arms races at the molecular level. *Current Opinion in Plant Biology*, **3**, 299–304.

Stahl EA, Dwyer G, Mauricio R, Kreitman M, Bergelson J (1999) Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. *Nature*, **400**, 667–71.

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

Stelzer CP (2008) Obligate asex in a rotifer and the role of sexual signals. *Journal of Evolutionary Biology*, **21**, 287–293.

Stephens M, Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American journal of human genetics*, **76**, 449–462.

Stöck M, Lamatsch DK, Steinlein C *et al.* (2002) A bisexually reproducing all-triploid vertebrate. *Nature genetics*, **30**, 325–328.

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 of the United States of America*, **109**, 10954–9.

Stukenbrock EH, McDonald BA (2008) The origins of plant pathogens in agro-ecosystems. *Annual review of phytopathology*, **46**, 75–100.

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.

Thines M (2014) Phylogeny and evolution of plant pathogenic oomycetes—a global overview. *European Journal of Plant Pathology*, **138**, 431–447.

Thines M, Choi Y-J, Kemen E *et al.* (2009) A new species of Albugo parasitic to Arabidopsis

thaliana reveals new evolutionary patterns in white blister rusts (Albuginaceae). *Persoonia*, **22**, 123–8.

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

Thrall PH, Laine AL, Ravensdale M *et al.* (2012) Rapid genetic change underpins antagonistic coevolution in a natural host-pathogen metapopulation. *Ecology Letters*, **15**, 425–435.

Togashi K, Shibasaki Y (1934) Biometrical and biological studies of A. candida (Pers.) O. Kuntze in connection with its specialization. *Bulletin of the Imperial College of Agriculture and Forestry*, **18**, 88.

Tsompana M, Abad J, Purugganan M, Moyer JW (2005) The molecular population genetics of the Tomato spotted wilt virus (TSWV) genome. *Molecular Ecology*, **14**, 53–66.

Turnbaugh PJ, Hamady M, Yatsunenko T *et al.* (2009) A core gut microbiome in obese and lean twins. *Nature*, **457**.

Turner EH, Ng SB, Nickerson DA, Shendure J (2009) Methods for Genomic Partitioning. *Annual Review of Genomics and Human Genetics*, **10**, 263–284.

Vercauteren A, Boutet X, D'hondt L *et al.* (2011) Aberrant genome size and instability of Phytophthora ramorum oospore progenies. *Fungal Genetics and Biology*, **48**, 537–543.

Ward BJ, van Oosterhout C (2015) HybridCheck : software for the rapid detection, visualization and dating of recombinant regions in genome sequence data. *Molecular Ecology Resources*.

Welch DM, Meselson M (2000) Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. *Science (New York, N.Y.)*, **288**, 1211–1215.

Wendel JF, Jackson SA, Meyers BC, Wing RA (2016) Evolution of plant genome architecture. *Genome Biology*, **17**, 37.

Wicker E, Lefeuvre P, de Cambiaire J-C *et al.* (2012) Contrasting recombination patterns and demographic histories of the plant pathogen Ralstonia solanacearum inferred from MLSA. *The ISME Journal*, **6**, 961–974.

Woese CR (1987) Bacterial Evolution. *Microbiological Reviews*, **51**, 221–271.

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

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, **24**, 1586–91.

Yarwood CE (1951) Associations of Rust and Virus Infections. *Science*, **114**, 127–128.

Yoshida K, Schuenemann VJ, Cano LM *et al.* (2013) The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. *eLife*, **2**.

Gene (AcNc2)	Sites (bp)	Tajima's D	Fu's Fs	d_N/d_S (codeml, mean)
Sg6400_tran2	1551	0.0931	0.235	1.2319
Sg3074_tran3	1770	-0.1104	0.127	0.7097
Sg1891_tran1	2745	0.0931	0.235	0.9832
Sg127_tran4	1755	0.198	0.05	0.8316
M3013	294	-0.279	0.166	0.7612
Gg6204	285	-0.279	0.166	0.7646
Gg4124	243	0.1203	-0.101	1.0754
Gg3224	237	0.1203	-0.101	0.7614
Gg25920	879	0.0944	0.003	0.821
Gg2479	159	0.0189	-0.156	0.7056
Gg24391	531	0.1432	-0.077	1.2642
Gg23928	282	-0.0101	0.184	0.934
Gg23675	183	-0.1839	-0.272	0.7254
Gg21542	186	-0.1782	0.223	0.8778
Gg20137	252	0.0189	-0.156	1.0519
Gg18456	291	0.1203	-0.101	0.7287
Gg15244	201	-0.1675	0.005	1.0946
Gg14391	162	0.0189	-0.156	1.1127
Gg12863	372	-0.1782	0.223	0.9621
Gg11862	198	-0.2415	-0.252	1.0856
Gg10824	207	-0.0425	0.164	1.0871
Ev7359	882	-0.0191	-0.048	0.7821
Ev6069	402	-0.2415	-0.252	0.759
Ev4751	1434	0.1367	0.259	1.1801
Ev434	3912	-0.0053	-0.173	0.7443
Ev4322	600	-0.2547	0.058	0.9071
Ev352	282	0.1203	-0.101	0.8474
Ev2368	372	-0.0835	0.044	1.0004
Ev1337	2178	0.0844	0.119	0.7404
Cg42	282	-0.2814	0.04	1.2587
Ag2511	549	-0.2799	0.033	1.011
Ag2417	636	0.0189	-0.156	1.1832

Table S2.1. The 32 diversity-tracking genes of *A. candida* that were used in this study. These were selected based on several neutrality tests (Fu's Fs, Tajima's D and d_N/d_S) performed using the varitale pipeline (Ishaque 2012): a suite of Perl scripts integrated with PAML 4 (Yang 2007), PHASE (Stephens & Scheet 2005) and DNAsp (Librado & Rozas 2009). These tests were based on all gene models identified in the seven laboratory isolates from which whole

genome data was available (*Ac*Nc2, *Ac*Em2, *Ac*BoT, *Ac*2v, *Ac*7v, *Ac*Ex1 and *Ac*BoL). Genes were selected when Fu's Fs and Tajima's D statistics were between -0.3 and 0.3 and d_N/d_S between 0.8 and 1.2 (see Table S2.1 for parameter estimates).

	#	Samples	# reads	Mean Q30 to base	% reads on target	Average depth at *Ac* 400 kb contig
PathSeq 1	1	*Ac* on *S. officinale* (ENG)	13883380	150	48.67	1077
	2	*Ac* on *C. bursa-pastoris* (ENG)	16472932	150	57.87	1541
	3	*Ac* on *A. deltoidea* (ENG)	10608586	150	52.53	837
	4	*Ac* on *A. deltoidea* (FRA)	16119514	150	55.40	1194
	5	*Ac* on *R. sativus* (FRA)	15327772	150	56.29	1305
	6	*Ac* on *C. bursa-pastoris* (POL)	20025168	150	55.72	1632
	7	*Ac* on *A. saxatile* (ITA)	4177486	150	28.17	141
	8	*Ac* on *A. deltoidea* (ITA)	4828750	150	29.14	213
	9	*Ac* on *R. sativus* (ENG)	7800086	150	35.08	143
	10	*Ac* on *S. officinale* (ENG)	12303036	150	53.55	699
	11	*Ac* on *C. bursa-pastoris* (ENG)	18264106	150	56.63	1212
	12	*Ac* on *A. saxatile* (ENG)	7150974	150	44.69	403
	13	*Ac* on *B. nigra* (FRA)	8919184	150	41.40	266
	14	*Ac* on *C. bursa-pastoris* (FRA)	8780834	150	54.58	530
	15	*Ac* on *A. deltoidea* (FRA)	10594584	150	49.18	535
	16	*Ac* on *A. saxatile* (FRA)	4730746	150	36.97	144
	17	*Ac* on *S. alba* (ENG)	4656906	150	36.75	157
	18	*Hpa* Emoy2 + *Al*Nc14 on *A. thaliana* Col-0	16006684	150	45.72	56
	19	*Ac*2v + *Erysiphe* sp. on *B. juncea*	17847200	150	58.83	1625
	20	*Ac*BoT on *B. oleracea*	16284674	150	53.16	1388
	21	*Ac*Ex1 + *P. infestans* on *A. thaliana* Col-0	24860952	150	56.20	1262
	22	*Ac*Ex1 + *P. infestans* + *P. syringae* DC3000 on *A. thaliana* Col-0	18350820	150	51.11	2049
PathSeq 2	23	*Ac* on *C. bursa-pastoris* (SCO)	6972510	148	16.33	70
	24	*Ac* on *C. bursa-pastoris* (SCO)	5128514	138	17.23	2
	25	*Ac* on *C. bursa-pastoris* (SCO)	4744896	145	14.65	4
	26	*Ac* on *A. deltoidea* (FRA)	7595634	150	44.50	628
	27	*Ac* on *A. deltoidea* (ENG)	1621138	143	41.26	78
	28	*Ac* on *S. officinale* (ENG)	3077858	145	32.29	179
	29	*Ac* on *A. saxatile* (ENG)	2723098	143	7.79	25
	30	*Ac* on *A. deltoidea* (ENG)	5023310	120	29.28	202
	31	*Ac* on *C. bursa-pastoris* (POL)	6979818	150	45.70	610
	32	*Ac* on *S. officinale* (ENG)	4602680	148	27.42	226
	33	*Ac* on *S. officinale* (ENG)	1807758	137	29.81	59
	34	*Ac* on *B. oleracea* (ENG)	3401906	150	14.08	72
	35	*Ac* on *B. oleracea* (NLD)[§]	1077532	145	22.07	36
	36	*Ac* on *B. oleracea* (NLD)[§]	1455494	145	20.82	48
	37	*Ac* on *B. oleracea* (NLD)[§]	1662042	150	43.15	146

	38	*Ac* on *R. sativus* (DEU)[§]	6734122	150	37.63	429
	39	*Ac* on *E. sativa* (ESP)[§]	1454894	150	37.46	93
	40	*Ac* on *A. thaliana* (ENG)	3864788	148	46.13	344
	41	*Ac* on *B. carinata* (CAN)[ǂ]	1176428	145	36.99	78
	42	*Ac* on *C. sativa* (CAN)[ǂ]	2984816	148	52.38	233
	43	*Ac* on *C. bursa-pastoris* (ENG)	7239720	148	23.83	137
	44	*Ac* on *C. bursa-pastoris* (DNK)	5634062	140	98.64	325
	45	*Ac* on *C. bursa-pastoris* (DNK)	13992730	145	34.21	553
	46	*Ac* on *C. bursa-pastoris* (FRA)	14611894	150	38.68	716
	47	*Ac* on *A. lyrata**	2705472	148	47.39	170
	48	*Ac* on *B. oleracea**	5154252	148	31.84	293
	49	*Ac* on *B. oleracea**	3220656	148	17.01	89
	50	*Ac* on *B. oleracea**	5749204	150	36.52	388
	51	*Ac* on *A. halleri**	6311020	150	36.47	405
	52	*Ac* on *B. oleracea**	5286890	150	39.76	424
	53	*Ac*Nc2 on *A. thaliana*	17095326	143	42.96	1058
	54	*Ac*7v zoospores	22075330	150	54.11	2221
	55	*Ac*2v zoospores	33351026	150	52.00	2336
	56	*Ac*2v zoospores + *P. syringae* DC3000 + sterilized *A. deltoidea*	5685520	148	57.90	525
	57	*Ac*2v zoospores + *P. syringae* DC3000 + sterilized *A. deltoidea*	6954204	140	52.91	556
	58	*Ac*2v zoospores + *P. syringae* DC3000 + sterilized *A. deltoidea*	2738914	130	41.61	155
	59	*Ac*2v zoospores + *P. syringae* DC3000 + sterilized *A. deltoidea*	1961550	120	38.56	83
	60	*Hpb* on *B. oleracea**	7072556	145	23.31	2
	61	*Hpb* on *B. oleracea**	1277206	140	23.13	2
	62	Sterilized healthy plant from sample #3 location	3047382	145	9.022	2
	63	Sterilized healthy plant from sample #2 location	6618204	148	14.12	2
	64	Healthy plant from sample #2 location	3987778	143	14.37	2
	65	Healthy plant from sample #3 location	3305114	145	11.54	2
	66	Healthy plant from sample #5 location	3360612	145	15.83	2
	67	Healthy plant from sample #9 location	2788846	148	13.38	18
	68	Healthy plant from sample #32 location	6369832	140	14.77	3
PathSeq 3	69	*Ac* on *L. annua* (ENG)	3744278	148	43.44	111
	70	*Ac* on *L. annua* (ENG)	5317210	143	67.92	447
	71	*Ac* on *L. annua* (FRA)	7262564	148	58.55	323
	72	*Ac* on *L. annua* (FRA)	9268608	148	59.73	336
	73	*Ac* on *A. deltoidea* (SCO)	6072504	150	53.92	275
	74	*Ac* on *A. deltoidea* (SCO)	5113550	150	59.55	305
	75	*Ac* on *A. saxatile* (FRA)	262924	135	42.97	4
	76	*Ac* on *L. annua* (SCO)	10248904	145	54.44	587
	77	*Ac* on *L. rediviva* (SCO)	1410636	148	70.65	125
	78	*Ac* on *B. juncea* (IND)[ʊ]	9744604	145	52.94	434
	79	*Ac* on *B. juncea* (IND)[ʊ]	3440228	133	43.74	43
	80	*Ac* on *B. juncea* (IND)[ʊ]	6007124	138	42.66	34

	81	*Ac* on *B. juncea* (IND)ʊ	6397780	143	42.68	72
	82	*Ac* on *B. juncea* (IND)ʊ	4062792	135	38.61	18
	83	*Ac* on *B. juncea* (IND)ʊ	6133992	140	43.29	125
	84	*Ac* on *B. juncea* (IND)ʊ	8222520	143	48.70	301
	85	*Ac* on *B. oleracea* (FRA)§	1781408	133	62.22	8
	86	*Ac* on *R. sativus* (NDL)§	2156504	134	33.77	35
	87	*Ac* on *R. sativus* (NDL)§	19169934	150	57.40	784
	88	*Ac* on *A. saxatile* (ENG)	3294138	150	49.49	132
	89	*Ac* on *Lunaria sp.* (ENG)	7707734	150	50.99	286
	90	*Ac* on *S. officinale* (IRL)	7392996	140	47.16	63
	91	*Ac* on *S. officinale* (IRL)	7755378	135	46.25	108
	92	*Ac* on *C. bursa-pastoris* (IRL)	15762514	150	71.01	1085
	93	*Ac* on *C. sativa* (CAN)ǂ	9602120	140	58.30	513
	94	*Ac* on *C. sativa* (CAN)ǂ	10510174	143	53.06	334
	95	*Ac* on *E. japonica* (ENG)	4842484	150	49.04	212
	96	*Ac* on S. officinale (FRA)	5217562	138	44.89	7
	97	*Ac* on *A. deltoidea* (IRL)	6092328	148	58.35	264
	98	*Ac* on *S. officinale* (ENG)	6575970	148	37.67	46
	99	*Ac* on *A. montanum* (ENG)	5032046	135	38.87	7
	100	*Ac* on *R. sativus* (USA)§	1726944	135	37.80	4
	101	*Ac* on *R. sativus* (FRA)§	2283578	110	45.83	4
	102	*P. infestans* on *S. tuberosum* (IRL)Ŧ	5074132	145	36.63	1
	103	*P. infestans* on *S. tuberosum* (NDL)Ŧ	4079280	145	35.20	1
	104	*P. infestans* mycelium Ŧ	2047102	143	24.13	1
	105	*P. infestans* mycelium Ŧ	12585414	138	36.18	1
	106	*P. infestans* mycelium Ŧ	7182366	140	38.21	1
	107	*P. tragopogonis* on *Helianthus* sp.	3287632	145	23.93	2
	108	Healthy plant from sample #6 location	215034	94	43.73	1
	109	Healthy plant from sample #103 location	4252926	114	49.19	1
	110	Healthy plant from sample #15 location	4745726	140	36.39	1
	111	*B. juncea*	6879692	138	42.30	1
	112	*B. rapa*	8287766	135	42.31	1
	113	*A. laibachii* Went1 on *A. thaliana*	5693058	150	73.29	14
	114	Healthy *A. thaliana*¤	4958558	135	41.76	1
	115	*A. laibachii* on *A. thaliana*¤	2790346	138	58.17	8

Table S3.1 Read and mapping statistics for all samples included in the study. Samples highlighted in orange should not contain *A. candida*, those in blue are discarded throughout the thesis due to low read depth and those that were provided by collaborators are marked with: § Ulrike Miersch and Dr. Annemarie Lokerse at Rijk Zwaan, De Lier, The Netherlands; Ŧ Dr. Jeroen Stellingwerf, Dr. Ewen Mullins' group at Oak Park Crops Research Centre, Carlow, Ireland; ǂ Dr. Hossein Borhan at Agriculture and Agri-Food Canada, Saskatoon, Canada; * Dr. Sebastian Fairhead, Prof. Eric Holub's group, Warwick University, England; ¤ Dr. Derek Lundberg, Prof. Detlef Weigel's group, Max Planck Institute for developmental biology, Tübingen, Germany; ʊ Prof. Deepak Pental (University of Delhi, India) and Prof. Abha Agnihotri (Amity University, Noida, India). Samples #17, 89, 95 and 13 were collected by Chih-Hang Wu, Prof. Jonathan Jones, Yan Ma (The Sainsbury Laboratory, Norwich,

England) and Quentin Dupriez, respectively. The three letters in parenthesis represent the abbreviations for the countries in which *A. candida* isolates were collected (abbreviations published by the United Nations, except for England (ENG) and Scotland (SCO)).

#	Host	Country	GPS coordinates (degrees)	Date of collection
1	*S. officinale*	England	52.63 / 1.28	17/05/2013
2	*C. bursa-pastoris*	England	52.61 / 1.32	17/03/2014
3	*A. deltoidea*	England	52.62 / 1.27	28/03/2014
4	*A. deltoidea*	France	50.66 / 1.67	29/03/2014
5	*R. sativus*	France	50.64 / 3.16	02/04/2014
6	*C. bursa-pastoris*	Poland	52.93 / 23.86	22/04/2014
7	*A. saxatile*	Italy	44.14 / 7.09	09/05/2014
8	*A. alpina*	Italy	44.57 / 7.51	11/05/2014
9	*R. sativus*	England	52.68 / 1.36	12/06/2014
10	*S. officinale*	England	52.63 / 1.28	17/05/2013
11	*C. bursa-pastoris*	England	52.62 / 1.22	03/10/2013
12	*A. saxatile*	England	52.61 / 1.32	17/03/2014
13	*B. nigra*	France	50.59 / 1.61	30/03/2014
14	*C. bursa-pastoris*	France	50.48 / 3.04	04/04/2014
15	*A. deltoidea*	France	50.48 / 3.04	04/04/2014
16	*A. saxatile*	France	50.48 / 3.04	04/04/2014
17	*S. alba*	England	52.62 / 1.27	19/07/2014
19	*B. juncea* (*Ac*2v)	Canada	56.47 / -23.55	-
20	*B. oleracea* (*Ac*BoT)	England	52.94 / -0.16	01/05/2009
21	*A. halleri* (*Ac*Ex1)	England	50.71 / -3.53	-
22	*A. halleri* (*Ac*Ex1)	England	50.71 / -3.53	-
23	*C. bursa-pastoris*	Scotland	52.62 / 1.27	05/06/2013
24	*C. bursa-pastoris*	Scotland	57.26 / -3.72	27/05/2013
25	*C. bursa-pastoris*	Scotland	57.19 / -3.83	27/05/2013
26	*A. deltoidea*	France	50.65 / 3.18	31/03/2014
27	*A. deltoidea*	England	51.14 / -3.2	02/05/2014
28	*S. officinale*	England	52.63 / 1.28	28/01/2015
29	*A. saxatile*	England	52.61 / 1.32	15/02/2015
30	*A. deltoidea*	England	52.62 / 1.24	26/02/2015
31	*C. bursa-pastoris*	Poland	52.89 / 23.88	22/04/2014
32	*S. officinale*	England	52.66 / 1.3	12/06/2014
33	*S. officinale*	England	52.67 / 1.67	12/06/2014
34	*B. oleracea*	England	53.36 / 0	2014
35	*B. oleracea* [§]	The Netherlands	51.96 / 4.2	24/10/2014
36	*B. oleracea* [§]	The Netherlands	52.74 / 5.22	21/10/2014
37	*B. oleracea* [§]	The Netherlands	52.74 / 5.22	19/03/2004
38	*R. sativus* [§]	Germany	51.16 / 10.45	1998
39	*E. sativa* [§]	Spain	37.62 / -0.99	10/02/2015
40	*A. thaliana*	England	54.89 / -2.93	-
41	*B. carinata* [ǂ]	Canada	56.47 / -23.55	-
42	*C. sativa* [ǂ]	Canada	56.47 / -23.55	-
43	*C. bursa-pastoris*	England	52.63 / 1.29	19/05/2013

44	*C. bursa-pastoris*	Denmark	56.85 / 8.83	13/11/2013
45	*C. bursa-pastoris*	Denmark	56.85 / 8.83	20/11/2014
46	*C. bursa-pastoris*	France	46.22 / 2.21	2013
47	*A. lyrata**	England	52.63 / 1.29	-
48	*B. oleracea**	England	52.63 / 1.29	-
49	*B. oleracea**	England	52.63 / 1.29	-
50	*B. oleracea**	England	52.63 / 1.29	-
51	*A. halleri**	England	52.63 / 1.29	-
52	*B. oleracea**	England	52.63 / 1.29	-
53	*A. thaliana* (*Ac*Nc2)	England	52.63 / 1.29	2007
54	(*Ac*7v)	Canada	56.47 / -23.55	-
55	(*Ac*2v)	Canada	56.47 / -23.55	-
56	(*Ac*2v)	Canada	56.47 / -23.55	-
57	(*Ac*2v)	Canada	56.47 / -23.55	-
58	(*Ac*2v)	Canada	56.47 / -23.55	-
59	(*Ac*2v)	Canada	56.47 / -23.55	-
69	*L. annua*	England	52.63 / 1.28	13/04/2015
70	*L. annua*	England	52.62 / 1.22	17/05/2015
71	*L. annua*	France	50.58 / 3.16	08/05/2015
72	*L. annua*	France	50.59 / 3.2	09/05/2015
73	*A. deltoidea*	Scotland	/7.93 / -4.01	05/05/2015
74	*A. deltoidea*	Scotland	57.19 / -3.82	10/05/2015
75	*A. saxatile*	France	50.59 / 3.2	09/05/2015
76	*L. annua*	Scotland	57.33 / -3.27	11/05/2015
77	*L. rediviva*	Scotland	57.2 / -3.82	11/05/2015
78	*B. juncea* ʊ	India	29.92 / 73.87	2015
79	*B. juncea* ʊ	India	30.88 / 75.85	2015
80	*B. juncea* ʊ	India	28.61 / 77.2	2015
81	*B. juncea* ʊ	India	28.18 / 76.61	2015
82	*B. juncea* ʊ	India	27.34 / 76.38	2015
83	*B. juncea* ʊ	India	28.61 / 77.2	2015
84	*B. juncea* ʊ	India	27.21 / 77.48	2015
85	*B. oleracea* §	France	48.67 / -1.85	20/10/2014
86	*R. sativus* §	The Netherlands	52.13 / 5.29	1997
87	*R. sativus* §	The Netherlands	52.13 / 5.29	29/06/2012
88	*A. saxatile*	England	52.62 / 1.24	01/06/2015
89	*L. annua*	England	52.62 / 1.27	01/06/2015
90	*S. officinale*	Ireland	53.36 / -6.24	14/06/2015
91	*S. officinale*	Ireland	53.36 / -6.23	14/06/2015
92	*C. bursa-pastoris*	Ireland	53.36 / -6.23	12/06/2015
93	*C. sativa* ǂ	Canada	56.47 / -23.55	-
94	*C. sativa* ǂ	Canada	56.47 / -23.55	-
95	*E. japonica*	England	52.63 / 1.29	-
96	*S. officinale*	France	50.58 / 3.16	08/05/2015
97	*A. deltoidea*	Ireland	53.36 / -6.23	14/06/2015
98	*S. officinale*	England	52.63 / 1.28	06/05/2014
99	*A. montanum*	England	50.19 / -5.42	12/07/2015

100	*R. sativus*§	USA	37.09 / -95.71	25/06/2012
101	*R. sativus*§	France	46.22 / 2.21	21/06/2013

Table S4.1 *A. candida* isolates used in this study. The date of collection is sometimes unknown when the samples were provided by collaborators. Similarly, GPS coordinates point to the centre of the country of collection when the exact location of the sample is unknown. Samples from collaborators are indicated by various symbols: § Ulrike Miersch and Dr. Annemarie Lokerse at Rijk Zwaan, De Lier, The Netherlands; ǂ Dr. Hossein Borhan at Agriculture and Agri-Food Canada, Saskatoon, Canada; * Dr. Sebastian Fairhead, Prof. Eric Holub's group, Warwick University, England; ʊ Prof. Deepak Pental (University of Delhi, India) and Prof. Abha Agnihotri (Amity University, Noida, India). Samples #17, 89, 95 and 13 were collected by Chih-Hang Wu, Prof. Jonathan Jones, Yan Ma (The Sainsbury Laboratory, Norwich, England) and Quentin Dupriez, respectively. The name of the pathotype is provided for lab isolates that have been sequenced and identified as such. The hosts these isolates were originally collected on are also given. If no host is provided, *A. candida* zoospores were used. Samples highlighted in blue were not analysed due to low read depth at 'contig 1'.

Within groups			Between groups		
	contig 1	DT loci		contig 1	DT loci
C. bursa-pastoris	3.55E-05	0.000257	*B.oleracea:C.bursa-pastoris*	0.005352	0.003145
B. oleracea (+*B. rapa*)	5.24E-05	0.000178	*S.officinale:C.bursa-pastoris*	0.005263	0.006994
S. officinale	0.000166	0.000546	*C.sativa:C.bursa-pastoris*	0.00106	0.000496
C. sativa	4.18E-06	0	*A.deltoidea:C.bursa-pastoris*	0.003429	0.006779
A. deltoidea	8.51E-06	5.74E-05	*A.saxatile:C.bursa-pastoris*	0.005188	0.003177
A. saxatile	1.41E-05	0.000127	*B.nigra:C.bursa-pastoris*	0.006565	0.00733
B. nigra	n/c	n/c	*B.juncea:C.bursa-pastoris*	0.007168	0.006506
B. juncea	0.001132	0.000585	*A.spp.:C.bursa-pastoris*	0.001693	0.000598
A. spp.	4.02E-06	1.88E-05	*S.alba:C.bursa-pastoris*	0.008609	0.009346
S. alba	n/c	n/c	*E.sativa:C.bursa-pastoris*	0.012334	0.015559
E. sativa	n/c	n/c	*B.carinata:C.bursa-pastoris*	0.006564	0.00735
B. carinata	n/c	n/c	*A.alpina:C.bursa-pastoris*	0.003491	0.006515
A. alpina	n/c	n/c	*R.sativus:C.bursa-pastoris*	0.004914	0.004046
R. sativus	0.000205	0.000973	*L.spp.:C.bursa-pastoris*	0.004998	0.011203
L. spp.	4.53E-05	0.000228	*S.officinale:B.oleracea*	0.00626	0.007697
Mean	**0.00017**	**0.00030**	*C.sativa:B.oleracea*	0.005731	0.003534
Standard deviation	**0.00035**	**0.00031**	*A.deltoidea:B.oleracea*	0.005343	0.007142
			A.saxatile:B.oleracea	0.005711	0.004253
			B.nigra:B.oleracea	0.004694	0.00707
			B.juncea:B.oleracea	0.004197	0.006645
			A.spp.:B.oleracea	0.00544	0.003479
			S.alba:B.oleracea	0.005577	0.009028

E.sativa:B.oleracea	0.011417	0.015987
B.carinata:B.oleracea	0.006524	0.007763
A.alpina:B.oleracea	0.005036	0.007097
R.sativus:B.oleracea	0.002197	0.004583
L.spp.:B.oleracea	0.005968	0.012192
C.sativa:S.officinale	0.005562	0.00745
A.deltoidea:S.officinale	0.004115	0.003689
A.saxatile:S.officinale	0.004908	0.004044
B.nigra:S.officinale	0.006863	0.006054
B.juncea:S.officinale	0.007476	0.006134
A.spp.:S.officinale	0.004947	0.007193
S.alba:S.officinale	0.009027	0.007898
E.sativa:S.officinale	0.012138	0.014169
B.carinata:S.officinale	0.007077	0.007023
A.alpina:S.officinale	0.003698	0.003458
R.sativus:S.officinale	0.004946	0.008295
L.spp.:S.officinale	0.005587	0.008943
A.deltoidea:C.sativa	0.003894	0.007289
A.saxatile:C.sativa	0.005762	0.003498
B.nigra:C.sativa	0.007004	0.007522
B.juncea:C.sativa	0.007598	0.006752
A.spp.:C.sativa	0.001182	0.000488
S.alba:C.sativa	0.008901	0.009512
E.sativa:C.sativa	0.012141	0.015393
B.carinata:C.sativa	0.006978	0.007648
A.alpina:C.sativa	0.00396	0.007096
R.sativus:C.sativa	0.00526	0.00454
L.spp.:C.sativa	0.005702	0.011366
A.saxatile:A.deltoidea	0.006195	0.003162
B.nigra:A.deltoidea	0.006093	0.006194
B.juncea:A.deltoidea	0.007063	0.006219
A.spp.:A.deltoidea	0.003339	0.006689
S.alba:A.deltoidea	0.008691	0.008041
E.sativa:A.deltoidea	0.012201	0.014776
B.carinata:A.deltoidea	0.006656	0.007088
A.alpina:A.deltoidea	0.000648	0.002761
R.sativus:A.deltoidea	0.004861	0.008241
L.spp.:A.deltoidea	0.003776	0.0093
B.nigra:A.saxatile	0.007309	0.005317
B.juncea:A.saxatile	0.008099	0.005206
A.spp.:A.saxatile	0.005401	0.003154
S.alba:A.saxatile	0.009116	0.0072
E.sativa:A.saxatile	0.011209	0.013577
B.carinata:A.saxatile	0.007671	0.005475

A.alpina:A.saxatile	0.006009	0.003368
R.sativus:A.saxatile	0.005421	0.004671
L.spp.:A.saxatile	0.006719	0.008999
B.juncea:B.nigra	0.0063	0.0042
A.spp.:B.nigra	0.00642	0.007067
S.alba:B.nigra	0.007711	0.00706
E.sativa:B.nigra	0.012692	0.013778
B.carinata:B.nigra	0.006512	0.004462
A.alpina:B.nigra	0.005836	0.005597
R.sativus:B.nigra	0.003756	0.007884
L.spp.:B.nigra	0.006626	0.008302
A.spp.:B.juncea	0.007353	0.006277
S.alba:B.juncea	0.007814	0.006603
E.sativa:B.juncea	0.012362	0.012965
B.carinata:B.juncea	0.002436	0.001056
A.alpina:B.juncea	0.006905	0.005889
R.sativus:B.juncea	0.004015	0.007358
L.spp.:B.juncea	0.007335	0.009337
S.alba:A.spp.	0.00846	0.009247
E.sativa:A.spp.	0.011371	0.015025
B.carinata:A.spp.	0.006739	0.007113
A.alpina:A.spp.	0.003391	0.006483
R.sativus:A.spp.	0.004917	0.004046
L.spp.:A.spp.	0.005091	0.011154
E.sativa:S.alba	0.012663	0.01267
B.carinata:S.alba	0.009042	0.007106
A.alpina:S.alba	0.008502	0.008158
R.sativus:S.alba	0.004833	0.008953
L.spp.:S.alba	0.009002	0.01149
B.carinata:E.sativa	0.011929	0.013118
A.alpina:E.sativa	0.01216	0.014797
R.sativus:E.sativa	0.00967	0.013497
L.spp.:E.sativa	0.012039	0.013298
A.alpina:B.carinata	0.006504	0.006445
R.sativus:B.carinata	0.005298	0.007412
L.spp.:B.carinata	0.00661	0.009807
R.sativus:A.alpina	0.004555	0.007083
L.spp.:A.alpina	0.003621	0.009101
L.spp.:R.sativus	0.005021	0.010977
Mean	**0.00655**	**0.00757**
Standard deviation	**0.00273**	**0.00034**

Table S4.2 Estimates of Average Evolutionary Divergence over Sequence Pairs within and between groups defined in Figure 4.2. Estimates are provided for both the trees in Figures 4.2 ('contig 1') and 4.3 (DT = Diversity-tracking loci). Groups were named according to the primary hosts isolates

were collected on. The number of base differences per site from averaging over all sequence pairs within each group are shown. The analysis involved 85 nucleotide sequences. All ambiguous positions were removed for each sequence pair. There were a total of 398,508 positions in the final dataset. Evolutionary analyses were conducted in MEGA6. The presence of n/c in the results denotes cases in which it was not possible to estimate evolutionary distances.

Host	Nb of samples analysed	Mean heterozygosity	SD	CV
Arabidopsis spp.	5	0.096161	0.003911	4.067392
A. saxatile	4	0.062016	0.004523	7.293569
A. deltoidea	9	0.043626	0.02902	66.51963
C. sativa	3	0.008956	0.001188	13.27056
C. bursa-pastoris	11	0.048952	0.012727	25.99782
E. japonica	1	0.010544	*n.a.*	*n.a.*
L. annua	6	0.025007	0.00205	8.198925
B. nigra	1	0.019382	*n.a.*	*n.a.*
S. alba	1	0.079846	*n.a.*	*n.a.*
B. rapa	1	0.016596	*n.a.*	*n.a.*
E. sativa	1	0.016458	*n.a.*	*n.a.*
B. carinata	1	0.122533	*n.a.*	*n.a.*
L. rediviva	1	0.026593	*n.a.*	*n.a.*
A. alpina	1	0.113188	*n.a.*	*n.a.*
S. officinale	5	0.061129	0.029117	47.63262
B. oleracea	7	0.650218	0.016389	2.520602
B. juncea	9	0.144273	0.09223	63.92708
B. juncea (lab isolates)	6	0.088314	0.008799	9.963884
B. juncea (India)	3	0.256192	0.075158	29.33662
R. sativus	4	0.879213	0.550935	62.66225
R sativus (without #5)	3	1.15442	0.029326	2.540296
R sativus #5	1	0.053592	*n.a.*	*n.a.*

Table S5.1 Mean heterozygosity within *A. candida* pathotypes as defined in Chapter 4. Heterozygosity is expressed as the percentage of observed heterozygous sites. *"C. bursa-pastoris*" includes isolates collected from *C. bursa-pastoris* and one isolate collected on *A. thaliana* (#53). Isolates collected on *B. juncea* and *R. sativus* are shown as one group or as two groups as suggested by previous phylogenetic analyses. SD = Standard deviation; CV = Coefficient of variation; *n.a.* = not applicable.

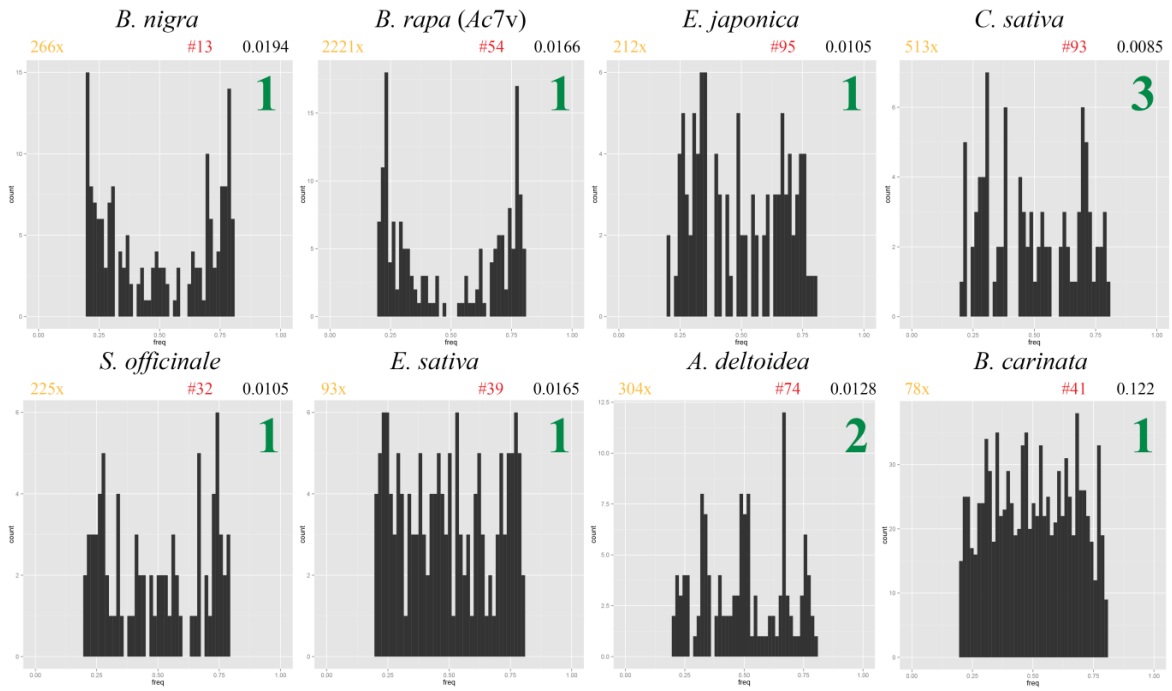


Figure S5.1 Distribution of the proportion of reads per SNP at heterozygous sites in *A. candida* isolates which ploidy level could not be determined ('contig 1'). The x-axis is the proportion of reads per SNP and the y-axis is the count of heterozygous sites. Average depth at 'contig 1' is provided in yellow, sample number in red (see Tables S3.1 and S4.1) and heterozygosity in black. In green is the number of isolates with similar distributions. The hosts from which isolates were collected are provided at the top of each graph.