Using genome resequencing to investigate racial structure, genetic diversity, sexual reproduction and hybridisation in the pine pathogen Dothistroma septosporum

Using genome resequencing to investigate racial structure, genetic diversity, sexual reproduction and hybridisation in the pine pathogen *Dothistroma septosporum*

Running heading: Population genomics of *Dothistroma septosporum*

Richard A. Ennos[1], Elisabet Ingrid Sjökvist[1,2], Marta J. Piotrowska[2,3], Carolyn Riddell[1,4], Peter N. Hoebe[2]

[1] Institute of Evolutionary Biology, University of Edinburgh, Ashworth Laboratories, Charlotte Auerbach Road., Edinburgh EH9 3FL

[2] Scotland's Rural College, West Mains Road, Edinburgh EH9 3JG

[3] Benchmak Holdings, Bush House, Edinburgh Technopole, Edinburgh EH26 0BB

[4] Forest Research, Northern Research Station, Roslin, UK EH25 9SY

Address for correspondence:

Institute of Evolutionary Biology, University of Edinburgh, Ashworth Laboratories, Charlotte Auerbach Road., Edinburgh EH9 3FL

Tel. 0131 650 5490

Fax. 0131 650 6564

Email: rennos@ed.ac.uk

Abstract

Whole genome resequencing of 25 isolates of the ascomycete pine pathogen *Dothistroma septosporum* revealed the presence of three genetically distinct races in Scotland, confirming previous results based on microsatellite markers. The genetically uniform indigenous race from Scotland (NPR), is closely related to an introduced and genetically diverse southern race (SR) from England and France. Race LPR, confined to lodgepole pine, is genetically divergent from these, but has close affinity with Canadian (CAN) isolates. Genome comparisons between individuals within CAN and SR races revealed short and scattered regions of near sequence identity, indicating significant sexual reproduction. Sequence identity patterns within LPR suggest that it derives from introduction of a single sexual fruiting body from North America. Finally a first generation interracial hybrid between indigenous NPR and introduced SR was identified, signalling an increase in the evolutionary potential of *D. septosporum* in Scotland, and an enhanced risk to native Scots pine.

Index descriptors: *Dothistroma septosporum*, genome resequencing, pine pathogen, sexual reproduction, exotic introduction, fungal hybridisation

Introduction

Population genetic analysis using a limited number of traditional markers such as microsatellites has provided us with fundamental insights into the origins, diversity, breeding systems and modes of adaptation of fungal and oomycete pathogens (Milgroom, 2015). However recent advances in next generation sequencing (NGS) now mean that it is feasible to obtain genomic level data from multiple individuals of the same taxon and conduct population genomic analyses (Grunwald et al., 2016; Moller & Stukenbrock, 2017; Bradshaw et al., 2019a, 2019b). Resequencing of whole genomes is a particularly powerful NGS methodology that yields huge numbers of reliable, high density, mapped markers. Analyses of these data provide robust, genome-wide estimates of key population genetic parameters, such as genetic diversity and differentiation, which are not compromised by the problems of ascertainment bias and hypervariability suffered by traditional microsatellite markers (Liti et al., 2009; Talas & McDonald; Rafiei et al., 2018). However the acquisition of resequencing data also offers the opportunity to conduct novel analyses that allow patterns of sequence identity and divergence between individuals to be mapped across the entire genome (Stukenbrock et al., 2012; Menardo et al., 2016).

The ability to compare the genomes of individuals, and identify chromosomal regions of near identical sequence, has already been extremely valuable for the recognition of hybrids and taxa of hybrid origin in haploid ascomycete fungi (Stukenbrock et al., 2012; Stukenbrock, 2016; Menardo et al, 2016). In an individual derived by hybridisation, half of the genome is expected to show sequence identity to one parent, and the remaining half to the second parent. The length of the chromosome blocks identical in sequence to either parent will initially be large but will decline with the number of rounds of sexual reproduction through which a hybrid population has

passed. Backcrosses to one of the parental taxa will be recognisable by an increase in the proportion of the chromosome blocks showing high sequence identity with that taxon (Stukenbrock et al., 2012; Menardo et al., 2016).

Understanding patterns of sequence identity between individuals is also potentially valuable for assessing the breeding system of haploid ascomycete fungi that practice a mixture of sexual and asexual reproduction (Taylor et al., 2015; Nieuwenhuis & James, 2016). Asexual reproduction generates clonal individuals within a population whose genomes are completely identical in sequence. If sexual reproduction occurs between unrelated clones, their offspring will, on average, show sequence identity across half of their genomes. The size of chromosomal blocks that are identical in sequence in these offspring pairs will be determined by the length of the chromosomes concerned and their recombination frequencies. Subsequent generations of sexual reproduction will further reduce the extent and size of shared chromosomal blocks of identical sequence between sexually produced individuals within the population. However if asexual reproduction occurs alongside sexual reproduction, new pairs of individuals possessing chromosomes that are completely identical in sequence will continually be created within the population. An equilibrium is expected between asexual reproduction creating chromosomal blocks of identical sequence, and sexual reproduction breaking these down. Pairs of individuals taken from populations with a high frequency of asexual reproduction will show high overall sequence identity, and this will be found in longer chromosomal blocks, than pairs of individuals taken from populations with a low frequency of asexual reproduction. Thus analysis of patterns of sequence identity between pairs of individuals within a haploid fungal population provides a measure of the relative frequency of sexual/asexual reproduction that it practices.

An ideal subject for exploiting the novel population genomic analyses made possible by resequencing is the economically important ascomycete fungus *Dothistroma septosporum*, whose 31.2MB haploid genome has been assembled and annotated (de Wit et al., 2012; Bradshaw et al., 2019a, 2019b). *D. septosporum* is currently causing devastating epidemics of needle blight in a wide variety of pine species throughout the world (Drenkhan et al., 2016). In Scotland there is particular concern about the effects of *D. septosporum* on native Scots pine, *Pinus sylvestris* L. within the Caledonian pinewoods, one of the few semi-natural forest ecosystems remaining in the UK (Steven & Carlisle, 1959, Mason et al., 2004). *D. septosporum* was first reported in these pinewoods in 2010 (Brown et al., 2012), and in order to develop appropriate management to mitigate damage from the disease, detailed information is required on the genetic composition of the *D. septosporum* population, its likely origins, its genetic diversity and mode of reproduction, and its capacity for future evolution.

Population genetic analysis using a limited number of microsatellite markers has established that three distinct races of *D. septosporum* co-occur in the Caledonian pinewoods (Mullett et al., 2017; Piotrowska et al., 2018). Race NPR is apparently indigenous, found predominantly on native Scots pine, shows low levels of microsatellite variability, and reproduces asexually. The race has not been known to cause any serious damage on native *P. sylvestris* in the past. The second race, SR, is highly genetically variable at microsatellite loci, possesses both mating types at equal frequency and shows genetic evidence of low but significant sexual reproduction (Ennos & Hu, 2019). It is believed that race SR has been introduced into England from continental Europe, where it has led to an epidemic of needle blight on exotic Corsican pine *P. nigra* subsp. *laricio* plantations (Brown & Webber,

2008). It has subsequently spread north to Scotland where it is now present on *P. sylvestris* within Caledonian pinewoods (Piotrowska et al., 2018). The final race LPR is confined to plantations of exotic lodgepole pine *P. contorta* in Scotland where it has caused very serious damage (Brown & Webber, 2008). The very low levels of microsatellite diversity, presence of only one mating type, and genetic affinity of LPR with isolates from Canada, are consistent with the race being a recent, inadvertent introduction from North America that has passed through a severe bottleneck and is reproducing asexually in Scotland (Piotrowska et al., 2018).

In this study we have resequenced isolates of *D. septosporum* previously classified, using microsatellite markers, into the three races described above, along with two isolates from Canada. We first determine whether classification into three races is supported by data from resequencing analysis. We analyse the genetic diversity within and the genetic differentiation between races using data from whole genome resequencing, and compare this with previous estimates from microsatellite analysis. Finally we analyse the patterns of sequence identity and divergence across the genome in pairs of individuals within and between the races and populations of *D. septosporum*. We use these data to compare the reproductive systems of the different races and populations, determine their likely origins, and look for evidence of hybridisation between the races. The implications of our results for managing the impact of *D. septosporum* in Scotland are then considered.

Methods

D. septosporum culture collection

In order to estimate genome-wide sequence diversity in *Dothistroma septosporum* we resequenced a subset of *D. septosporum* isolates, previously classified into different races based on microsatellite data (Table 1; Piotrowska et al., 2018). In total we used 25 isolates of *D. septosporum,* collected from either Scots or lodgepole pine. Most of these isolates (n =19) originated from our sampling in Scottish forests in 2014, while six were supplied by Forest Research (Alice Holt). The isolates comprised the following groups:

(i) Isolates classified by microsatellites as race NPR (n= 12), from four different locations in Scotland, Glen Einig (n= 3), Dundreggan (n= 3), Glen Garry (n= 3) and Inshriach Forest (n= 3), sampled from both Scots and lodgepole pines.

(ii) Isolates classified as race LPR (n= 4), collected exclusively from lodgepole pine at two locations, Glen Garry (n= 2) and Inshriach Forest (n= 2).

(iii) Isolates of uncertain race based on microsatellite genotypes but classified as NPR or SR (n= 3), collected at three sites, Glen Einig (n= 1), Dundreggan (n= 1) and Inshriach Forest (n= 1), representing samples from both Scots and lodgepole pine.

(iv) Isolates classified as race SR collected from tree nurseries in Southern Scotland (n=4) during annual DNB surveys in 2011 and 2014. These samples were provided by Forest Research (Alice Holt) and all information regarding samples remains confidential.

(v) Isolates originating from British Columbia, Canada (n =2), sampled from lodgepole pine in 2011 and provided by Forest Research (Alice Holt).

For isolates from Scottish pinewood locations we collected pine needles showing signs of *D. septosporum* infection and proceeded with single spore isolation as described by Mullet et al. (2015), obtaining a single isolate from any one tree. Detailed description of isolation procedure can be found in Piotrowska et al. (2017, 2018). For DNA extraction, isolates were cultured on porous membranes on solid Dothistroma Sporulating Medium (Bradshaw et al., 2000) at 12 hr light/12 hr dark at 20°C (Gallenkamp, INF 780C, Weiss Technik Konigswinter, Germany) and harvested once they covered 1/3 of the plate.

DNA extraction

Prior to DNA extraction, isolate material was freeze-dried overnight and ground with 3 mm stainless steel balls in the Qiagen tissue lyser LT system (Qiagen, Hilden, Germany) for 2 mins at maximum speed. Samples were then processed using the Illustra Nucleon PhytoPure Genomic DNA Extraction Kit (GE Healthcare Life Sciences, Little Chalfont, UK). To obtain a purer extraction, the lysate supernatant was transferred to a fresh tube and re-spun before taking the clear upper layer through the clean-up steps. The DNA pellet was re-suspended in 40-50 µl 1xTBE.

To remove RNA, 30 µl of each DNA sample was treated with 1.5 µl of RNase cocktail (Invitrogen, UK) incubated at 65˚C for 10 minutes and placed on ice before performing a phenol:chloroform:isoamyl alcohol clean-up. 200 µl of 1x TBE and 250 µl (>2x vol) phenol:chloroforom:isoamyl alcohol (25:24:1 ratio) was added to the sample, mixed by shaking and centrifuged on a 1-14 Sigma bench centrifuge at

16,163 x g for 10 minutes. The upper layer was collected to a fresh tube and DNA precipitated with an equal volume of ice cold isopropanol. DNA was precipitated for 1 hour at room temperature and centrifuged at 16,163 x g for 10 minutes. The pellet was washed with 500 µl of 70% ethanol, air dried and re-suspended in 40 µl 1xTBE.

Next generation sequencing

Prior to submitting samples for sequencing, DNA quality was assessed using the Nanodrop spectrophotometer ND1000 (Thermo Fisher Scientific, Wilmington, USA) to check the 260/280 ratio was >1.7 and free of contaminants. DNA concentration was calculated using flurometric quantification on the Qubit system (Thermo Fisher Scientific, Wilmington, USA) with the High Sensitivity DNA assay.

Samples were submitted to Edinburgh Genomics (Ashworth Laboratories, University of Edinburgh, EH9 3FL). A TruSeq DNA Nano gel-free library (550bp insert) with unique index was prepared for each sample before pooling all 25 libraries and sequencing on a HiSeq high output v4 125 paired end flow cell (Illumina, USA), to generate 200M x 2 reads to a read depth varying between 92x and 107x for individual samples. Standard base calling and quality control was performed by Edinburgh Genomics before further bioinformatic analysis. Sequence data are deposited in the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra/) under accession number ERS1201615.

SNP variant detection and mapping

Quality control of the 150 base paired-end reads and trimming of adapters was performed using Fastqc v0.11.4 (Andrews, 2010) and Skewer (Jiang et al., 2014). Reads were mapped to the *D. septosporum* reference genome (New Zealand strain

NZE10; http://genome.jgi.doe.gov/Dotse1/Dotse1.home.html) (de Wit et al., 2012) using bwa (Li and Durbin, 2009) and sorted and tagged for regions that were not haploid with samtools 1.3 (Li et al., 2009). Variants were called using FreeBayes (Garrison and Marth, 2012). Hard filters were applied to the variant call set with 'low quality' being defined as any variant with at least one of the following properties; QD < 2.0; FS > 60; MQ < 40; MQRankSum < -12.5; ReadPosRankSum < -8.0. 'Low quality' variants were filtered out with VCFtools (Danecek et al., 2011). The analysis resulted in a vcf file containing high quality mapped SNP variants across the genome for the 25 isolates of *D. septosporum* studied.

Preliminary clustering of genotypes

The matrix of nucleotide differences among isolates stored in the vcf file was used to conduct a Principal Component Analysis in R studio (v 1.0.136) using the adegenet package (v 2.0.1, Jombart, 2008; Jombart & Ahmed, 2011). The data were also employed to conduct a neighbour joining analysis in adegenet to further explore the relationships among the races. On the basis of these results the original classification of isolates from Scotland into three races (NPR, SR and LPR) using microsatellite markers, was reassessed before proceeding with more detailed genomic comparisons (Piotrowska et al., 2018; Table 1).

Population Genomic Analysis

i. Sequence diversity within and among races

Following assignment of isolates to races, the number of sequence differences between each pair of individuals within the total sample was determined in 50Kb windows across the 14 largest scaffolds of the *D. septosporum* genome using the procedures '*view'* and '*gtcheck'* in the program BCFtools

(http://github.com/samtools/bcftools). These data were then used to calculate π, nucleotide diversity within races and d_A, nucleotide diversity between races for each of the 14 large scaffolds of *D. septosporum*, as well as for the genome as a whole (Nei & Li, 1979). In addition, the proportion of non-reference alleles shared by pairs of races was calculated using the '*view'* procedure in BCFtools.

ii. Chromosomal patterns of sequence diversity within populations and races

Inspection of the data on nucleotide differences between pairs of individuals across the genome indicated that individuals often shared long stretches of chromosome where sequence differences were extremely low (<5bp differences within each 50Kb window). These regions of very low sequence divergence were interspersed with regions where sequence divergence was much higher. Figure S1 illustrates the distribution of number of sequence differences found in each 50Kb window for two genome comparisons within race SR, where distinct modal values close to zero and ninety base pair differences are present. To document and compare the chromosomal pattern of sequence divergence within different populations and races as well as between races SR and NPR, each 50Kb window within a genomic comparison of two isolates was assigned to one of two classes; either <5bp or >5bp difference in sequence between isolates within the 50Kb window. These data were used to create chromosomal maps showing patterns of sequence divergence across the genome for pairs of individuals within the CAN sample, within races SR, LPR and NPR, and between races NPR and race SR.

Results

SNP diversity

Initial variant calling detected 337,343 variant sites among the 25 isolates. Following hard filtering for quality with VCFtools (Danecek et al., 2011) a total of 182,442 SNPs were retained upon which all further analysis was based. The vast majority of these sites were biallelic, with only 502 loci showing more than two alleles. The ratio of transitions to transversions in this set of SNPs was 2.11.

Classification of isolates

The results of a PCA analysis based on a matrix of overall SNP differences among isolates is shown in Figure 1. The first PCA axis clearly separated the isolates into two groups. One, with low PCA1 scores, comprised the two CAN isolates together with the four isolates designated as LPR on the basis of previous microsatellite analysis. CAN and LPR isolates were separated from each other on PCA2.

The second main group, with high PCA1 scores, contained isolates designated as SR and NPR in the previous analysis. In this latter group the second PCA axis divided the isolates into a further two clusters that largely corresponded to the previously recognised SR and NPR races. However four of these 19 isolates did not fall into the grouping expected on the basis of microsatellite genotype. Two isolates originally designated as NPR (DS15, DS97) clustered with the SR group; one isolate originally classified as SR grouped with the NPR cluster (DS251); and a fourth isolate (DS63) was located midway between the SR and NPR clusters along PCA2, as expected for an interracial hybrid between NPR and SR. Neighbour joining analysis of the same data (Fig. 2) also suggested the same discrepancies between the original classification of the isolates into races and their genetic affinities.

On the grounds that the present resequencing analysis utilises far more genetic information than the microsatellite assignments, the isolates were reclassified to conform to the groupings suggested by the PCA and neighbour joining plots. Thus all further genomic analyses were conducted assuming 11 NPR isolates (DS13, DS17, DS42, DS54, DS55, DS88, DS172, DS186, DS188, DS198, DS251), 7 SR isolates (DS15, DS49, DS97, DS112, DS114, DS147, DS153), 4 LPR isolates (DS23, DS61, DS85, DS199), a putative interracial hybrid between NPR and SR (DS63), and two CAN isolates (DS643, DS644).

Sequence diversity and divergence

Table 2 (diagonal) shows the nucleotide diversity π found within each of the three races of *D. septosporum* in Scotland and in the Canadian sample. In race NPR nucleotide diversity ($\pi=4.5 \times 10^{-6}$) is extremely low, two orders of magnitude less than for any of the other races or populations. This difference is highly consistent over all 14 major chromosomes. In contrast, the SR race shows the greatest nucleotide diversity ($\pi=1.4 \times 10^{-3}$), about twice as high as that found for the LPR and CAN samples. When data from individual chromosomes are used to provide replicate measures of π within samples, ANOVA indicates that SR shows a significantly greater value of π than either LPR or CAN ($P<0.001$). It also indicates that π differs significantly among chromosomes ($P<0.004$). This is partly accounted for by a negative relationship with chromosome length ($P<0.003$); the value of π drops by roughly 0.1×10^{-3} for every 1Mb increase in chromosome length ($R^2 = 0.36$, 0.19 and 0.23 for SR, LPR and CAN respectively).

Nucleotide divergence d_A between pairs of populations and races is shown above the diagonal in Table 2, while the proportion of non-reference alleles shared by these

pairs of populations and races is given below the diagonal. Races NPR and SR are diverged least in sequence (d_A=0.57x10^{-3}) and share the highest proportion of non-reference alleles (0.53). Relatively low nucleotide divergence (d_A=1.33x10^{-3}) was also shown between LPR and SR and between LPR and CAN. However a closer genetic relationship between LPR and CAN than between LPR and SR was suggested by the higher proportion of non-reference alleles that they share (0.45 vs. 0.31). This pattern of genetic relatedness was supported by the PCA and neighbour joining analyses (Figs. 1 & 2) which both placed LPR closer to CAN than to SR.

Chromosomal patterns of sequence identity and divergence

i. Canadian isolates

Fig. 3A shows the patterns of sequence identity and divergence across the 14 major chromosomes of *D. septoporum* when the genomes of the two Canadian isolates are compared. For the vast majority of the genome (96.99%) there was large divergence in sequence between the two isolates (mean 30.62 SNPs per 50Kb). However the remaining 3.11% of the genome comprised 15 regions across the 14 principal chromosomes where sequences were near identical between the two isolates (<5 SNPs per 50Kb). The maximum and mean length of genome blocks showing near identity of sequence were 150Kb and 63.4Kb respectively.

ii. SR isolates

Fig 3B shows patterns of sequence divergence and identity between three pairs of individuals from the SR race. The figure illustrates the lowest, an intermediate and the highest identity in sequence found among the 21 possible isolate pairs. Patterns of sequence identity and divergence across the genome are quite different for each isolate pair comparison. Across the whole genome mean sequence divergence was

high when averaged over all pairs of isolates (71.31 SNPs per 50Kb). For the complete set of isolate pairs, the percentage of the genome showing near sequence identity followed a normal distribution (Fig. S2) and ranged from 3.28% to 19.21% with a mean of 10.58%, significantly higher than the value of 3.01% found for the CAN isolate pair ($P<0.001$, t test). The distribution of lengths of sequence of near identity showed a negative exponential distribution (Fig. S3) with a maximum value of 700Kb and mean of 107.3Kb, significantly higher than the mean of 63.4Kb recorded for the CAN isolate pair ($P<0.05$, Mann-Whitney U test).

iii. LPR isolates

The LPR isolates fell into two clear groups with respect to patterns of sequence identity. Isolates DS23, DS61 and DS199 were nearly identical in sequence across their whole genome, with sequence differences greater than 5 SNPs per 50Kb in only one region, of length 100Kb, on chromosome 12 (Fig. 3C(i)). In contrast isolate DS85 showed near sequence identity with the other LPR isolates over only 61.6% of its genome. Substantial sequence divergence (mean 49.67 SNPs per 50Kb) was present over the remaining 38.4% of the genome (Fig. 3C(ii)). Regions of divergent sequence were arranged in small numbers of very large blocks along chromosomes. In the case of chromosome 6 the complete chromosome of DS85 showed near sequence identity with the other three LPR isolates. In the diverged chromosomal regions, the sequences were more similar to those present in CAN isolates than to those in races present in Scotland (total mean SNP differences 16384 CAN, 31409 NPR, 31769 SR).

iv. NPR isolates

As already noted, sequence diversity in race NPR was extremely low, and there were no chromosomal regions where sequence divergence between isolates exceeded 5 SNPs per 50 Kb window (Fig 3D).

v. NPR vs. SR

On average the proportion of the genome showing near sequence identity between race NPR and the SR isolates was 14.2%. There was substantial variation in this figure among SR isolates (range 8.37% to 24.96%). However the mean value did not differ significantly from the value of 10.58% found among SR isolates (P=0.064). For illustration Fig. 3E shows the chromosomal pattern of identical sequences for three NPR/SR pairs. These three pairs shared respectively the least, an intermediate, and the greatest proportion of near identical sequence. The mean length of near identical sequence found in comparisons between NPR and SR isolates is 142 Kb, significantly greater than the mean value of 107.3 Kb found in comparisons between SR isolates (Mann-Whitney U test, $P<0.001$)

v. NPR vs. putative interracial hybrid DS63

When genome comparisons were made between the putative interracial hybrid, isolate DS63, and the NPR isolates, blocks of near sequence identity were found to cover 57.31% of the genome. This was significantly greater than the mean of 14.2% found between NPR and the SR isolates (t Test, $P<0.001$). For any one chromosome, the contrasting near identical and divergent sequences were each arranged in a small number of very large blocks measuring up to 2.6Mb in length (Fig. 3F). Those regions of the genome showing divergence from the NPR sequence were more similar to those present in race SR than in race LPR (Mean SNP differences 17945 and 31476 respectively).

Discussion

Whole genome resequencing results largely confirmed division of Scottish *D. septosporum* isolates into three genetically distinct races, a classification originally based on microsatellite markers (Piotrowska et al, 2018). Discrepancies in classification only arose with respect to the most closely related races, NPR and SR. Resequencing data suggested reassignment of three isolates between NPR and SR, and that a fourth isolate was a putative interracial hybrid between these races. Our original classifications were based on data from 11 polymorphic microsatellite loci, whereas our resequencing results made use of SNPs at 182,442 sites. Given the difference in information content between the data sets we resolved discrepancies in racial classification in favour of the resequencing analysis. We conclude that where genetic divergence is large, as between races LPR and SR, microsatellite loci are sufficient for racial classification, but for distinguishing very closely related groups resequencing data may be essential. A similar level of concordance between classifications based on different markers was found in *Verticillium dahliae* where clusters distinguished by microsatellites corresponded in all but one case with clonal lineages defined by NGS SNP markers (Rafiei et al., 2018).

In our previous population genetic analysis of the *D. septosporum* races, microsatellite data suggested that race NPR was intermediate in genetic diversity (H_t = 0.182) between race LPR (H_t = 0.041) and race SR (H_t = 0.550). However resequencing revealed that genomic sequence diversity of race NPR ($\pi = 4.5 \times 10^{-6}$) is two orders of magnitude lower than in races LPR ($\pi = 0.502 \times 10^{-3}$) and SR ($\pi = 1.424 \times 10^{-3}$). An explanation for this lack of concordance may lie in the contrasting mutation rates of microsatellite and SNP markers, and the influence this has on their population genetic behaviour.

Microsatellite markers have been developed for population genetic studies because they have high diversity that is associated with a high rate of mutation, as high as 10^{-1}/generation in some cases (Ellegren, 2000). In contrast the typical mutation rate at SNP loci is from 10^{-8} to 10^{-10} per generation (Kasuga et al., 2002; Zhu et al., 2014) with a recent estimate of 2.23 x 10^{-7}/nucleotide/year in *D. septosporum* (Bradshaw et al., 2019a). Mutation rate determines the rapidity with which mutation-drift equilibrium is restored following a population bottleneck. If the NPR population has passed through a recent bottleneck, the time since the bottleneck may have been too short to allow mutation to restore equilibrium levels of diversity at SNP loci, but adequate for restoration of relatively high levels of genetic diversity as a consequence of high mutation rate at microsatellite loci. A massive historical reduction in population size and fragmentation of native pine populations in Scotland is well documented and suggests that the endemic NPR population is likely to have suffered a severe and recent bottleneck (Steven & Carlisle, 1959).

Analysis of nucleotide divergence d_A clearly demonstrated that races SR and NPR are the most closely related. In contrast the four individuals originally classified as race LPR and isolated from lodgepole pine hosts are genetically divergent from all other Scottish isolates and grouped with the two isolates from Canada in the PCA plot and neighbour joining tree. Moreover they showed the least sequence divergence and shared the highest proportion of non-reference alleles with Canadian isolates. These patterns of relatedness among the races and isolates from Canada are in accord with our previous analyses based on microsatellite data (Piotrowska et al., 2018) and are compatible with our working hypotheses; that endemic race NPR was originally derived from race SR and has passed through a severe bottleneck

since Scottish populations of *P. sylvestris* separated from their continental European counterparts after glaciation; race LPR is a recent introduction from North America.

In *D. septosporum* each new generation can potentially be founded by both asexual and sexual propagules. Sexual fruiting bodies of *D. septosporum* have been observed in both North America and continental Europe (Funk & Parker, 1966: Butin, 1985). There is also good evidence, both from mating type frequencies and population genetic analysis using traditional markers, that the CAN and SR isolates are from populations that practice a significant amount of sexual reproduction (Dale et al., 2011; Tomsovsky et al., 2013; Mullett et al., 2015; Boron et al., 2016; Ennos & Hu, 2019). In order to compare the proportion of sexual reproduction occurring in the Canadian and SR populations we can use our genomic data to contrast the patterns of sequence identity among pairs of individuals within each population. In both populations individuals share relatively short chromosome blocks of identical genomic sequence scattered over many chromosomes. However the two individuals from Canada show both a significantly lower proportion of overall genome sequence identity, and significantly smaller lengths of near identical sequence than pairs of individuals within the SR population. This strongly suggests that both populations are practicing sexual reproduction, but that this is more prevalent in the Canadian population than in the SR population within Scotland.

When sequence identity is compared between pairs of individuals within the LPR race a completely different picture is seen from that found in the Canadian and SR populations. Three of the four isolates are almost identical in sequence over the whole genome, indicating that they are the products of asexual reproduction. Two of the three identical isolates originated from different pine trees within the same population, while the third was from a population located more than 50 km away.

Natural dispersal of asexual spores over long distances is a possible explanation for these results (Mullett et al., 2016). However given the presence of the LPR race in forest nurseries, transport of infected planting stock to distant sites could also be responsible (Piotrowska et al., 2018).

Sequence comparison of the three asexually produced LPR isolates with the remaining LPR isolate shows near identity of chromosome blocks for 61.6% of the genome. Regions of non-identical sequence are in very large blocks, and they are most similar in sequence to the same genomic regions in the Canadian isolates. These observations strongly suggest that the two distinct LPR genotypes that we have sampled represent full siblings, the product of a single sexual fruiting body. Close sequence relatedness of non-identical regions with those in the Canadian isolates indicates that the fruiting body was not the product of sexual reproduction in Scotland, but of sexual reproduction in Canada or elsewhere in North America. Thus the LPR population in Scotland could ultimately have been derived from introduction of a single sexual fruiting body on needle tissue, which then released its sexual ascospores. This hypothesis is supported by the observation that sexual reproduction is known to be common in North America (Dale et al., 2011). Moreover among the total sample of 37 LPR isolates scored for microsatellite markers in our previous study (Piotrowska et al., 2018), none of the loci showed more than two alleles, as expected if they are derived from a single fruit body. This suggests strong parallels with the invasion history of *Hymenoschyphus fraxineus*, the cause of ash dieback in Europe, which can be traced back to the introduction of only two genotypes from Asia (McMullen et al., 2018).

Our consideration of genetic diversity within races has already indicated that sequence diversity within the NPR population is extremely low, although significant

variation is found at hypervariable microsatellite loci. Previous work has shown that race NPR possesses both mating types, but mating type 1 is at a very low frequency (3.5%). It is therefore likely that reproduction in NPR is exclusively asexual. However with the introduction of closely related race SR, which contains both mating types at equal frequency, the opportunity has arisen for hybridisation to occur between these asexual and sexual races.

Evidence for such interracial hybridisation comes from our genome sequence comparison of the putative interracial hybrid isolate, DS63, and the NPR isolates. The analysis shows that the putative interracial hybrid is near identical in sequence to NPR isolates for 57.3 % of its genome. Taking into account that isolates from NPR and SR share 14.4% sequence identity on average, the expected percentage of the genome identical in sequence between NPR and an interracial hybrid with SR is (50 + (14.4/2)) = 57.2%, very close to the figure actually observed. Blocks of near sequence identity and non-identity are large and are found along all chromosomes, consistent with DS63 being a first generation interracial hybrid. Furthermore regions of non-identity are closest in sequence to individuals from the SR population. Taken together these data provide very strong evidence that DS63 represents a first generation interracial hybrid between the endemic NPR population present on native Scots pine, and the introduced SR race. It is notable that the interracial hybrid was found in the Glen Garry site where both NPR and SR isolates are present.

Results from population genetic and genomic analysis have been instrumental in demonstrating the crucial role played by hybridisation in the evolution of fungal pathogens of plants (Stukenbrock, 2016). In some cases hybridisation events took place hundreds of generations ago (Stukenbrock et al., 2012). However in many cases the anthropogenic transport of pathogenic fungi around the globe has

provided the opportunity for very recent hybridisation that would not otherwise have been possible, and that has led to the emergence of destructive new hybrid taxa (Brasier et al., 2004).

In the current example of *D. septosporum*, an endemic, predominantly asexual population of low genetic diversity causing little damage on *P. sylvestris* (NPR) has apparently hybridised with an introduced and partially sexual population that has caused severe epidemics on *P. nigra* (SR). This has been made possible by the widespread planting throughout Britain of the highly susceptible exotic *P. nigra* which shares *D. septosporum* and other pathogens with the closely related native species, *P. sylvestris* (Piotrowska et al., 2017). Hybridisation between NPR and SR has a number of potential consequences.

Hybridisation provides the genetically depauperate endemic population of NPR with access to the much more genetically diverse gene pool of SR. Hybridisation is also likely to increase the frequency of the mating type 1 allele in the gene pool of NPR, facilitating sexual reproduction. The evolutionary potential of the endemic NPR population will thus be enhanced due to an increase in genetic diversity and the opportunities for genetic recombination provided by sexual reproduction (McDonald & Linde, 2002). The net effect is that the risk posed to native *P. sylvestris* by *D. septosporum* is likely to be significantly increased as a result of hybridisation. This is in addition to the risk already posed by the presence of introduced races SR and LPR in the Caledonian pinewoods in the absence of hybridisation.

In terms of immediate management responses to the current *D. septosporum* situation in the Caledonian pinewoods, a number of measures are appropriate and have been undertaken by Forestry Commission Scotland (Forestry Commission

Scotland, 2017, 2018). The first is a cessation of pine planting in the Caledonian pinewoods from potentially infected forest nurseries. The second is removal of exotic pines, particularly *P. contorta* from the vicinity of Caledonian pinewoods. Although we have not detected hybridisation involving the LPR race present on *P. contorta*, the continuing presence of this species in the vicinity of Caledonian pinewoods raises the chances of this occurring. Given the highly destructive nature of LPR on *P. contorta*, the prospects of it transferring to *P. sylvestris* following hybridisation with NPR gives serious cause for concern.

From the viewpoint of long term management of native forest systems, our finding of increased threats arising from hybridisation of introduced pathogen races with endemic races serves to reiterate the biosecurity risks posed by widespread planting of exotic tree species that are related to native species (Ennos et al, 2019). Such practices should not be undertaken unless they have been subject to a comprehensive risk analysis. Our findings also emphasise the need to manage our native forests in such a way that they are able to evolve and adapt to novel biodiversity threats, such as those posed by new hybrid pathogen taxa (Cavers and Cottrell, 2015). Crucial to this is control of grazing and other pressures currently preventing natural regeneration so that evolution of more resistant populations can occur, fully exploiting the presence of abundant genetic diversity within native tree populations.

Data Access

Sequence reads for this project are available at NCBI Sequence Read Archive ERS1201615 '*Dothistroma septosporum* population study'.

Acknowledgements

All resequencing of isolates was carried out by Edinburgh Genomics, The University of Edinburgh. Edinburgh Genomics is partly supported through core grants from NERC (R8/H10/56), MRC (MR/K001744/1) and BBSRC (BB/J004243/1). This project was funded jointly by a grant from BBSRC, Defra, ESRC, the Forestry Commission, NERC and the Scottish Government, under the Tree Health and Plant Biosecurity Initiative. We would like to thank Dr. Katherine Tubby, Richard Baden and Dr. Martin Mullett from Forest Research in Alice Holt for kindly supplying the nursery and Canadian isolates of *D. septosporum*.

Author Contributions

MJP, CR, PNH and RAE were involved in the design of the research. CR and MJP collected and cultured isolates and prepared DNA for resequencing. ES assembled genomes and performed all quality filtering of data. ES and RAE analysed the data. RAE wrote the manuscript with significant input from ES, CR, MJP and PNH.

References

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Boron P, Lenart-Boron A, Mullett M (2016). The distribution of *Dothistroma septosporum* and its mating types in Poland. Forest Pathology, 46, 489–496.

Bradshaw RE, Ganley RJ, Jones WT, Dyer PS (2000). High levels of dothistromin toxin produced by the forest pathogen *Dothistroma pini*. Mycological Research, 104, 325–332.

Bradshaw RE, Ormond S, Dupont P-Y, Chettri P, Ozturk IK, McDougal RL, Bulman LS, Cox, MP (2019a) Reduced virulence of an introduced forest pathogen over 50 years. Microorganisms 7, 420.

Bradshaw RE, Sim AD, Chettri P, Dupont P-Y, Guo Y, Hunziker L, McDougal RL, Van der Nest A, Fourie A, Wheeler D, Cox MP, Barnes I (2019b). Global population genomics of the forest pathogen *Dothistroma septosporum* reveal chromosome duplications in high dothistromin-producing strains. Molecular Plant Pathology, 20, 784-799.

Brasier CM, Kirk SA, Declan J, Cooke DEL, Jung T, Man In't Velda WA (2004). *Phytophthora alni* sp. nov. and its variants: Designation of emerging heteroploid hybrid pathogens spreading on *Alnus* trees. Mycological Research, 108, 1172–1184.

Brown A, Webber J (2008). Red band needle blight of conifers in Britain. Forestry Commission Research Note 002.

Brown AV, Stone C, Clayden H (2012). Dothistroma needle blight: GB strategy. Edinburgh, UK: Forestry Commission.

Butin H (1985). Development of the teleomorph and anamorph of *Scirrhia pini* Funk & Parker on needles of *Pinus nigra* Arnold. Sydowia, 38, 20-27.

Cavers S, Cottrell, JE (2015). The basis of resilience in forest tree species and its use in adaptive forest management in Britain. Forestry, 88, 13– 26.

Dale AL, Lewis KJ, Murray BW (2011). Sexual reproduction and gene flow in the pine pathogen *Dothistroma septosporum* in British Columbia. Phytopathology, 101, 68-76.

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

Drenkhan R, Tomesova-Haataja V, Fraser S. et al. (2016). Global geographic distribution and host range of *Dothistroma* species: a comprehensive review. Forest Pathology, 46, 408-442.

Ellegren H (2000). Microsatellite mutations in the germline: implications for evolutionary inference. Trends in Genetics, 16, 551–558.

Ennos RA, Cottrell J, Hall J, O'Brien D (2019). Is the introduction of novel exotic forest tree species a rational response to rapid environmental change? - A British perspective. Forest Ecology and Management, 432, 718-728.

Ennos RA, Hu X-S (2019). Estimating the number of sexual events per generation in a facultatively sexual haploid population. Heredity, 122, 729–741.

Forestry Commission Scotland (2017). Planting in Caledonian pinewoods: reducing risks from Dothistroma Needle Blight (2017). Policy guidance note.

Forestry Commission Scotland (2018). Dothistroma Action Plan for Scotland (2018). Policy guidance note.

Funk A, Parker AK (1966). *Scirrhia pini* n. sp., the perfect state of *Dothistroma pini* Hulbary. Canadian Journal of Botany, 44, 1171-1176.

Garrison E, Marth G (2012). Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907

Grunwald NJ, McDonald BA, Milgroom MG (2016). Population Genomics of Fungal and Oomycete Pathogens. Annual Review of Phytopathology, 54, 23–46.

Jiang, H, Lei, R, Ding, S-W, Zhu, S (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics, 15, 182.

Jombart T (2008). Adegenet: A R package for the multivariate analysis of genetic markers. Bioinformatics, 24, 1403-1405.

Jombart T, Ahmed I (2011). Adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. Bioinformatics, 27, 3070-3071.

Kasuga T, White TJ, Taylor JW (2002). Estimation of nucleotide substitution rates in Eurotiomycete fungi. Molecular Biology and Evolution, 19, 2318–2324.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25, 1754-60.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009). The Sequence alignment/map (SAM) format and SAMtools, Bioinformatics, 25, 2078-9.

Liti G, Carter DM, Moses AM et al. (2009). Population genomics of domestic and wild yeasts. Nature, 458, 337-341.

Mason WL, Hampson A, Edwards C (2004). Managing the pinewoods of Scotland. Edinburgh, UK: Forestry Commission.

McDonald BA, Linde C (2002). Pathogen population genetics, evolutionary potential and durable resistance. Annual Review of Phytopathology, 40, 349–79.

McMullen et al. (2018). The ash dieback invasion of Europe was founded by two genetically divergent individuals. Nature Ecology and Evolution, 2, 1000–1008.

Menardo F, Praz CR, Wyder, S et al. (2016). Hybridization of powdery mildew strains gives rise to pathogens on novel agricultural crop species. Nature Genetics, 48, 201-205.

Milgroom MG (2015). Population Biology of Plant Pathogens: Genetics, Ecology, and Evolution. St. Paul, MN: APS Press.

Möller M, Stukenbrock EH (2017). Evolution and genome architecture in fungal plant pathogens. Nature Reviews Microbiology, 15, 756-771.

Mullett MS, Brown AV, Barnes I (2015). Population structure and reproductive mode of *Dothistroma septosporum* in the Brittany peninsula of France. European Journal of Plant Pathology, 143, 261–275.

Mullett MS, Tubby KV, Webber JF, Brown AV (2016) A reconsideration of natural dispersal distances of the pine pathogen *Dothistroma septosporum*. Plant Pathology, 65, 1462–1472.

Mullett MS, Brown AV, Fraser S, Baden R, Tubby KV (2017). Insights into the pathways of spread and potential origins of *Dothistroma septosporum* in Britain. Fungal Ecology, 26, 85-98.

Nei M, Li WH (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences, 76, 5269–5273.

Nieuwenhuis BPS, James TY (2016). The frequency of sex in fungi. Philosophical Transactions of the Royal Society B, 371, 20150540.

Piotrowska MJ, Ennos RA, Riddell C, Hoebe PN (2017). Fungicide sensitivity of *Dothistroma septosporum* isolates in the UK. Forest Pathology, 47, e12314.

Piotrowska MJ, Riddell C, Hoebe PN, Ennos RA (2018). Planting exotic relatives has increased the threat posed by *Dothistroma septosporum* to the Caledonian pine populations of Scotland. Evolutionary Applications, 11, 350-363.

Rafiei V, Banihashemi Z, Jimenez-Diaz RM, Navas-Cortes JA, Land BB , Jimenez-Gasco MM, Turgeone BG, Milgroom MG (2018). Comparison of genotyping by sequencing and microsatellite markers for unravelling population structure in the clonal fungus *Verticillium dahlia*. Plant Pathology, 67, 76–86.

Steven HM, Carlisle A (1959). The native pinewoods of Scotland. Edinburgh, UK: Oliver and Boyd.

Stukenbrock EH (2016). The role of hybridization in the evolution and emergence of new fungal plant pathogens. Phytopathology, 106, 104-112.

Stukenbrock EH, Christiansen FB, Hansen TT, Dutheila 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, 109, 10954–10959.

Talas F, McDonald BA (2015). Genome-wide analysis of *Fusarium graminearum* field populations reveals hotspots of recombination. BMC Genomics, 16, 996.

Taylor JW, Hann-Sodena C, Brancoa S, Sylvaina I, Ellison CE (2015). Clonal reproduction in fungi. Proceedings of the National Academy of Sciences, 112, 8901–8908.

Tomsovsky M, Tomesova V, Palovcikova D, Kostovcik M, Rohrer M, Hanacek P, Jankovsky L (2013). The gene flow and mode of reproduction of *Dothistroma septosporum* in the Czech Republic. Plant Pathology, 62, 59–68.

de Wit PJ, van der Burgt A, Ökmen B et al. (2012) The genomes of the fungal plant pathogens *Cladosporium fulvum* and *Dothistroma septosporum* reveal adaptation to different hosts and lifestyles but also signatures of common ancestry. PLoS Genetics, 8, e1003088.

Zhu YO, Siegal ML, Hall DW, Petrov DA (2014). Precise estimates of mutation rate and spectrum in yeast. Proceedings of the National Academy of Sciences, 111, E2310-E2318.

Table 1. Details of *D. septosporum* isolates used in the study.

Site	Lat/Long	Year of sampling	Isolate ID	Host species	Race (Microsatellites)	Race (Re-sequencing)
Isolates from Pinewoods in Scotland						
Glen Einig	57.957 −4.740	2014	DS13	Scots pine	NPR	NPR
			DS17	Scots pine	NPR	NPR
			DS15	Scots pine	*NPR**	*SR*
			DS42	lodgepole pine	NPR	NPR
Dundreggan	57.196 −4.736	2014	DS186	Scots pine	NPR	NPR
			DS188	Scots pine	NPR	NPR
			DS198	lodgepole pine	NPR	NPR
			DS251	lodgepole pine	*SR**	*NPR*
Glen Garry	57.051 −4.979	2014	DS61	lodgepole pine	LPR	LPR
			DS63	lodgepole pine	*NPR*	*NPR x SR interracial hybrid*
			DS97	Scots pine	*NPR*	*SR*
			DS172	Scots pine	NPR	NPR
			DS199	lodgepole pine	LPR	LPR
Inshriach Forest	57.097 −3.931	2014	DS23	lodgepole pine	LPR	LPR
			DS49	lodgepole pine	SR*	SR
			DS54	Scots pine	NPR	NPR
			DS55	Scots pine	NPR	NPR
			DS85	lodgepole pine	LPR	LPR
			DS88	lodgepole pine	NPR	NPR
Isolates from Nurseries in Southern Scotland						
		2011	DS112	*Pinus* sp.	SR	SR
		2011	DS114	*Pinus* sp.	SR	SR
		2014	DS147	*Pinus* sp.	SR	SR
		2014	DS153	*Pinus* sp.	SR	SR
Isolates from British Columbia, Canada						
Nass Valley, Brown Bear	55.730 -128.438	2011	DS643	lodgepole pine	CAN	CAN
Kispiox, Buckley Canyon	55.358 -127.687	2011	DS644	lodgepole pine	CAN	CAN

Footnotes:

Classification of the isolates into races (NPR, LPR, SR and CAN) based on microsatellite markers (Piotrowska et al., 2018) and SNP data from the present study are shown, with discrepancies in italics. Asterix indicate samples of less certain race based on microsatellite data.

Table 2. Population genomic statistics for samples of three races of *Dothistroma septosporum* from Scotland (NPR, SR, LPR) and two isolates from Canada (CAN).

	NPR	SR	LPR	CAN
NPR	**0.000005** (0.0000003)	0.00057	0.00224	0.00247
SR	*0.53*	**0.001424** (0.000081)	0.00133	0.00180
LPR	*0.28*	*0.31*	**0.000502** (0.000114)	0.00133
CAN	*0.27*	*0.29*	*0.45*	**0.000611** (0.000032)

Footnotes:

Values for nucleotide diversity within races (π) are given along the diagonal together with the standard error based on variation among chromosomes. Nucleotide diversity between races (dA) is shown above the diagonal. Below the diagonal is the proportion of non-reference alleles shared between each pair of races.

Figure 1

Principal Component Analysis based on whole genome SNP data for 25 isolates of *D. septosporum* sampled from Scotland and Canada. Classification of isolates based on SNP data is indicated: ⊜ Race NPR, ◍ Race SR, ⊘ Race LPR, ⊕ Putative NPR/SR interracial hybrid, ■ Canadian isolates.

Figure 2

Neighbour joining tree based on whole genome SNP data for 25 isolates of *D. septosporum* sampled from Scotland and Canada. Classification of isolates based on SNP data is shown: ⊜ Race NPR, ◍ Race SR, ⊘ Race LPR, ⊕ Putative NPR/SR hybrid, ● Canadian isolates

Figure 3

Patterns of sequence near identity (< 5 SNP/50Kb, open bars) and divergence (> 5 SNP/50Kb, filled bars) across the 14 major chromosomes of *D. septosporum* in genome comparisons between the following isolates. **A** two Canadian isolates; **B** i), ii) and iii) three pairs of SR isolates; **C** i) LPR isolates DS23, DS61, DS199; **C** ii) LPR isolate DS85 paired with LPR isolates DS23, DS61 and DS199; **D** all eleven NPR isolates; **E** i), ii) and iii) three SR isolates paired with NPR; **F** putative interracial hybrid DS63 and NPR.

Figure 1

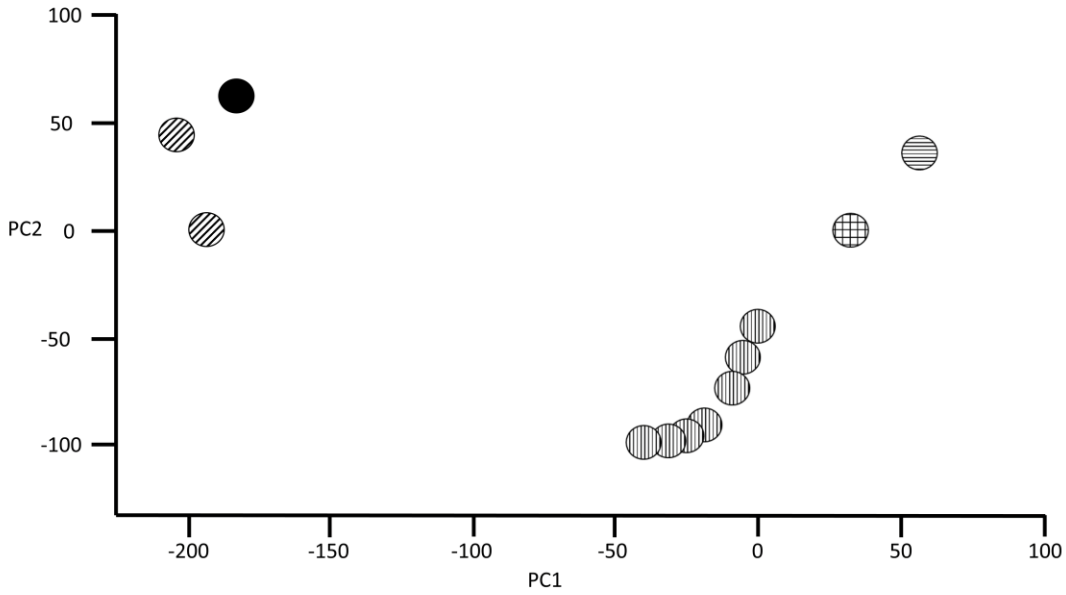


Figure 2

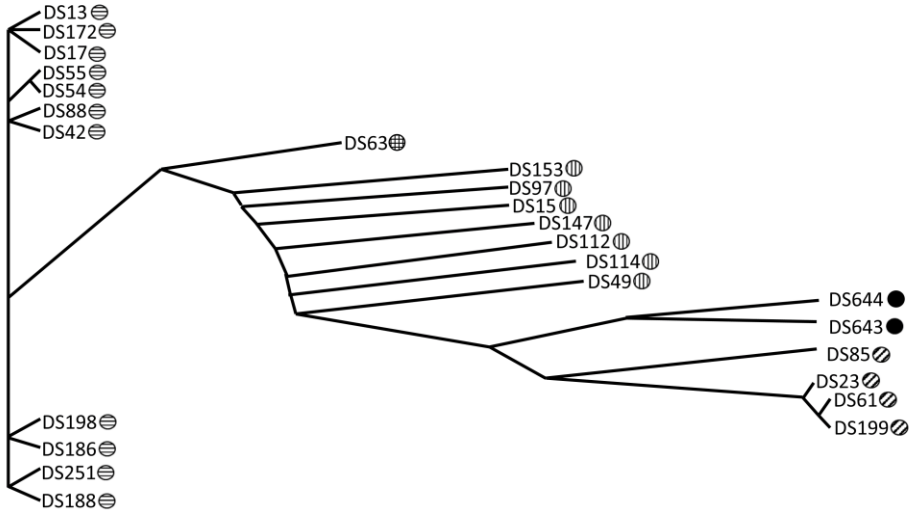

DS13
DS172
DS17
DS55
DS54
DS88
DS42
DS63
DS153
DS97
DS15
DS147
DS112
DS114
DS49
DS644
DS643
DS85
DS23
DS61
DS199
DS198
DS186
DS251
DS188

Figure 3A

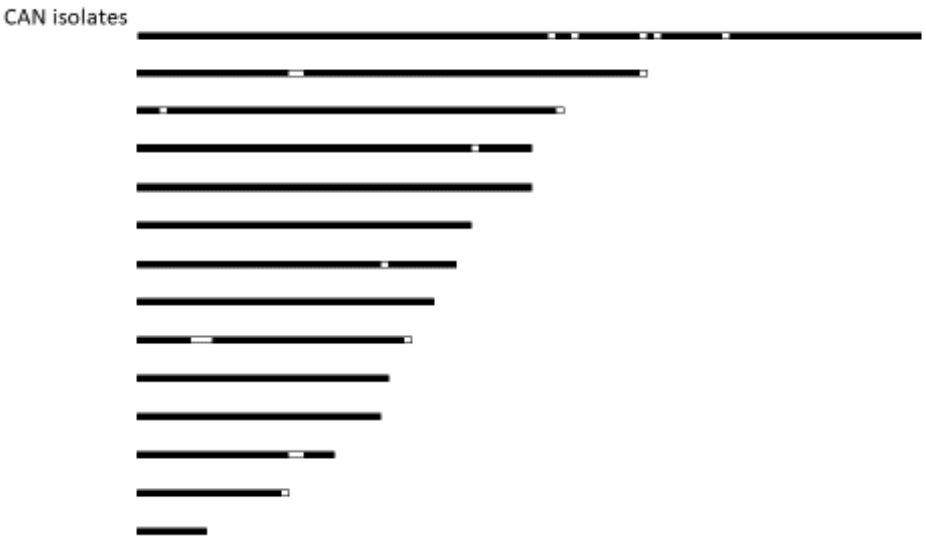
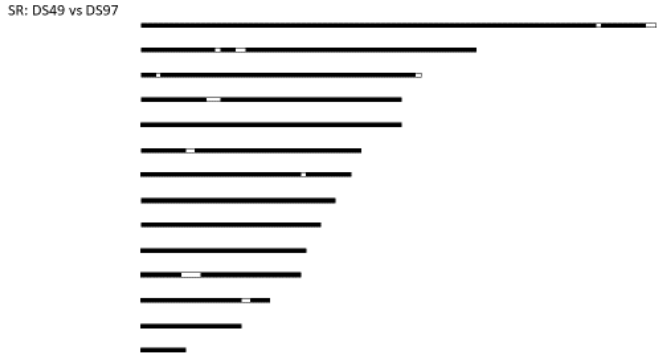


Figure 3B

i).

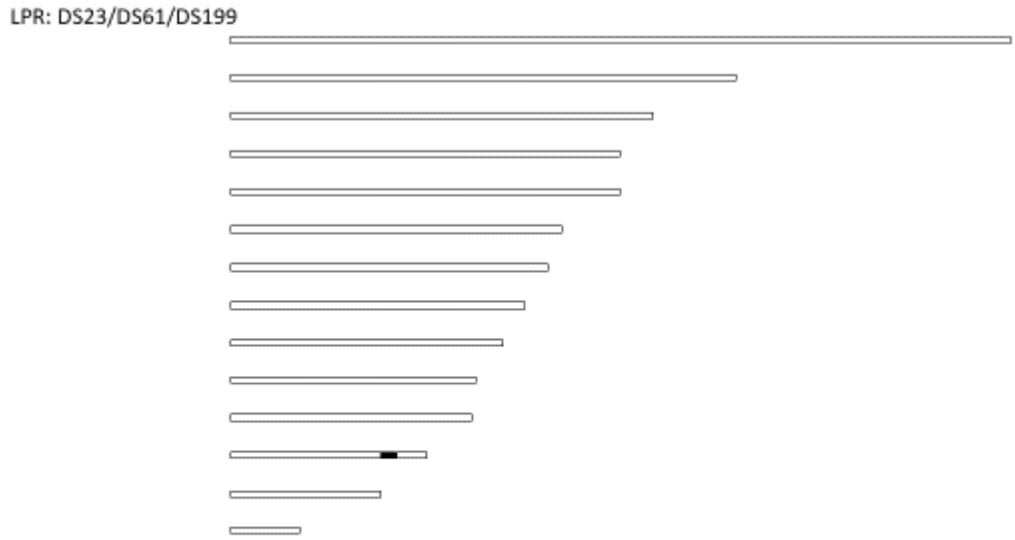


ii).

iii).

Figure 3C

i).



ii).

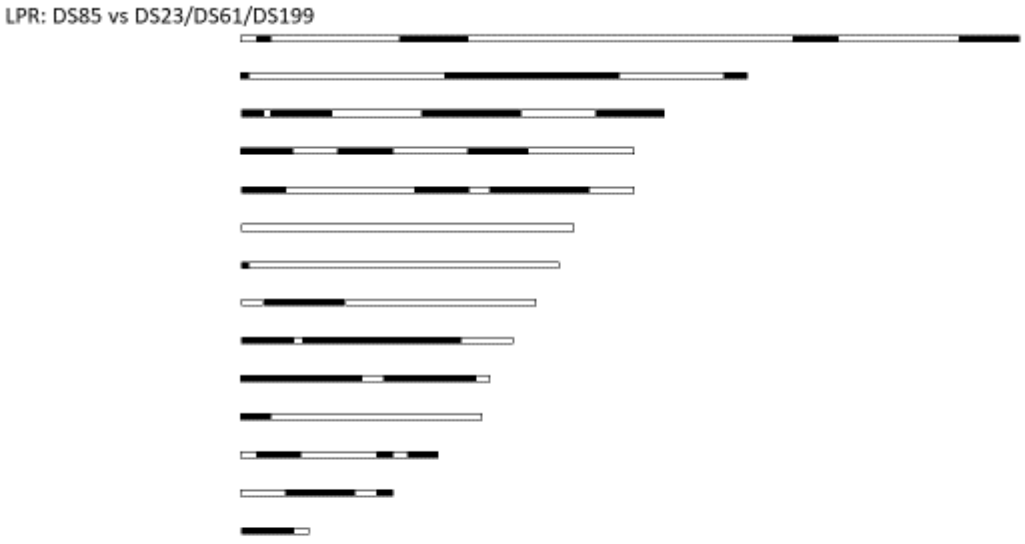


Figure 3D

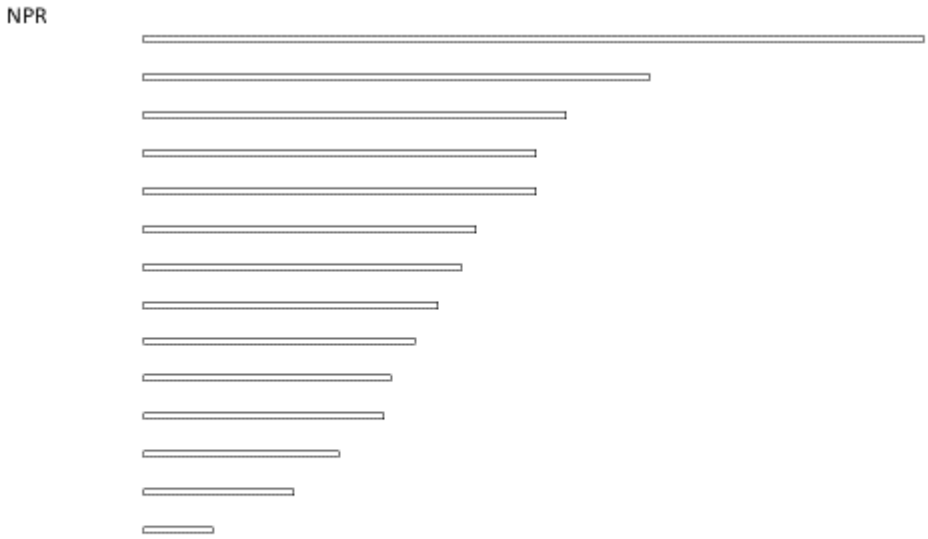


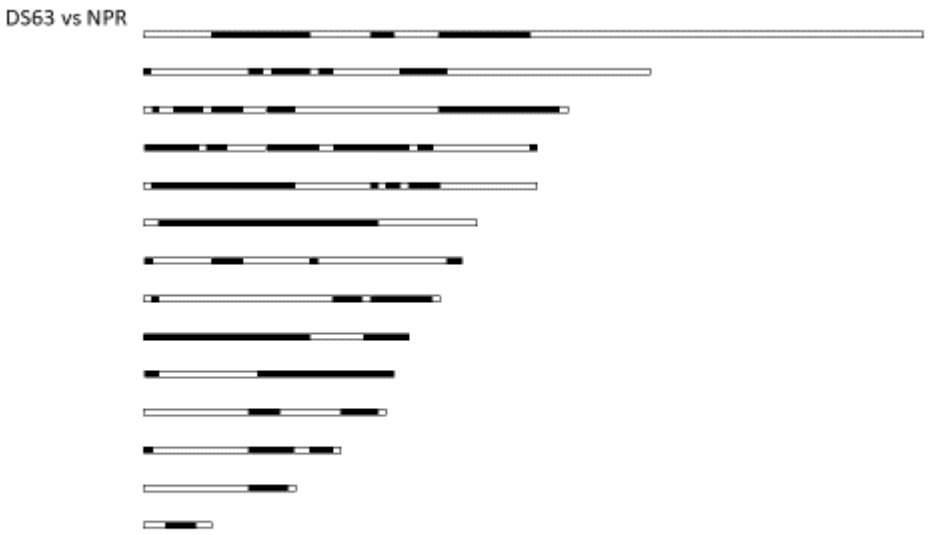
Figure 3E

i).

ii).

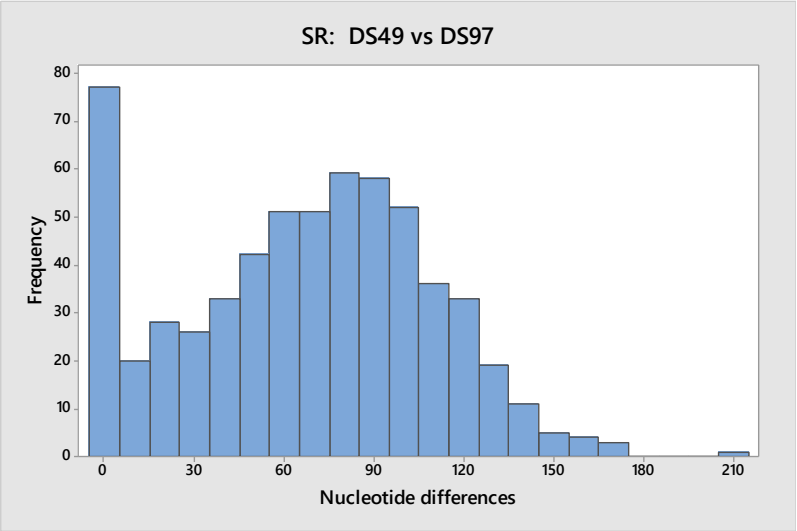
iii).

Figure 3F



Supplementary material

Figure S1. Distribution of number of nucleotide differences found in 50Kb windows across the genome when two pairs of SR isolates of *D. septosporum* are compared. A clear bimodal distribution with one mode at <5 nucleotide differences is shown.



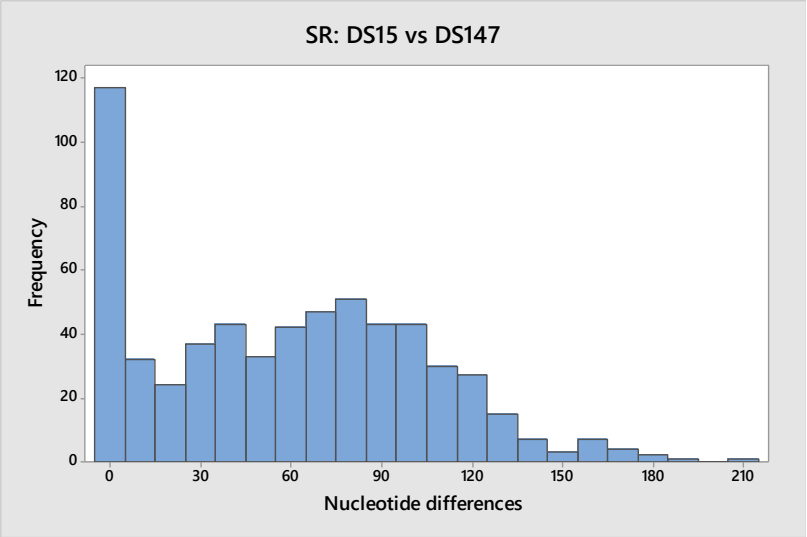


Figure S2. Distribution of the proportion of the genome that is near identical in sequence (<5 SNP/50Kb) between pairs of individuals within the SR population.

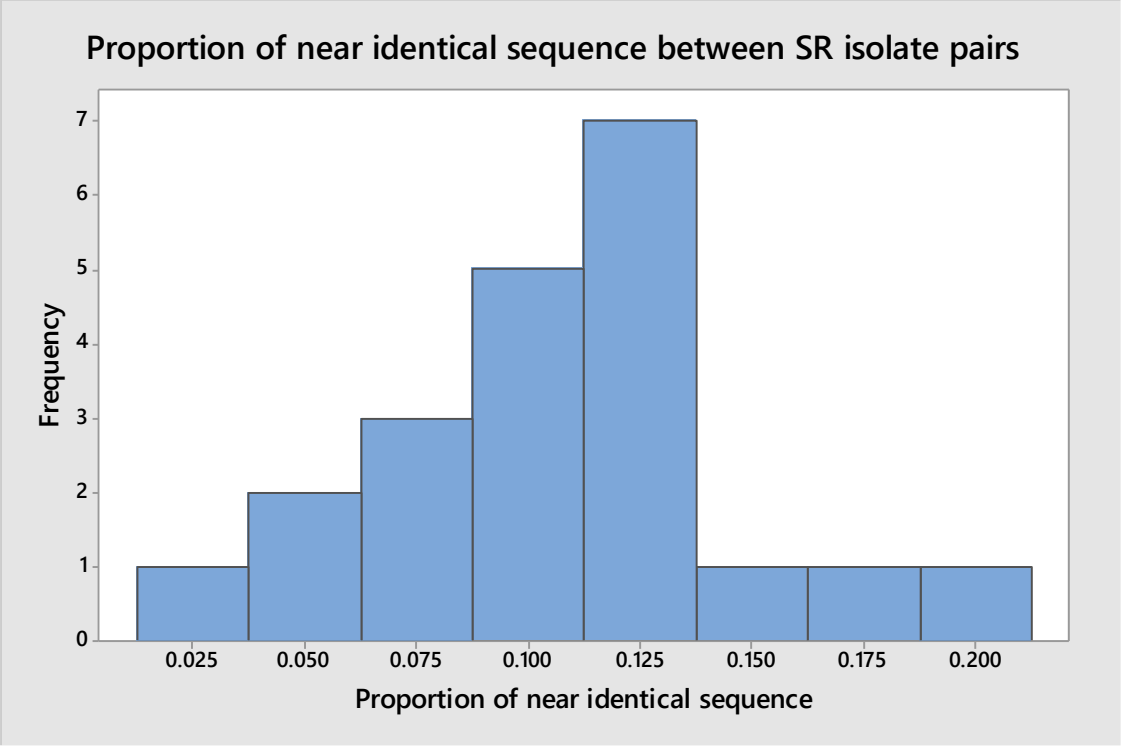


Figure S3. Distribution of lengths of chromosome blocks (50Kb units) of near identical sequence (<5 SNP/50Kb) between pairs of isolates from the SR population.