Inferences on pathogenic fungus population structures from microsatellite data: new insights from spatial genetics approaches

A. RIEUX,*† F. HALKETT,‡ M.-F. ZAPATER,* F. ROUSSET,¶ V. RAVIGNE* and J. CARLIER*
*CIRAD, UMR BGPI, Campus international de Bailleul, TA A-54K, F-34398 Montpellier Cedex 5, France,
†INRA, Nancy-Université, UMR 1136 Interactions Arbre-Microorganismes, F-54280 Champenoux, France,
‡CIRAD, UPR Syst. Banan. Ananas, TA B-26/PS4, Blvd. de la Lironde, 34398 Montpellier Cedex 5, France,
¶Université de Montpellier II, CNRS, Institut des Sciences de l’Evolution, Montpellier Cedex, France

Abstract

Landscape genetics, which combines population genetics, landscape ecology and spatial statistics, has emerged recently as a new discipline that can be used to assess how landscape features or environmental variables can influence gene flow and spatial genetic variation. We applied this approach to the invasive plant pathogenic fungus Mycosphaerella fijiensis, which causes black leaf streak disease of banana. Around 800 isolates were sampled within a 50 × 50 km area located in a fragmented banana production zone in Cameroon that includes several potential physical barriers to gene flow. Two clustering algorithms and a new \( F_{ST} \)-based procedure were applied to define the number of genetic entities and their spatial domain without a priori assumptions. Two populations were clearly delineated, and the genetic discontinuity appeared sharp but asymmetric. Interestingly, no landscape features matched this genetic discontinuity, and no isolation by distance (IBD) was found within populations. Our results suggest that the genetic structure observed in this production area reflects the recent history of \( M. \) fijiensis expansion in Cameroon rather than resulting from contemporary gene flow. Finally, we discuss the influence of the suspected high effective population size for such an organism on (i) the absence of an IBD signal, (ii) the characterization of contemporary gene-flow events through assignment methods of analysis and (iii) the evolution of the genetic discontinuity detected in this study.

Keywords: banana, clustering analysis, emergent disease, gene flow, landscape genetics, Mycosphaerella fijiensis

Received 12 July 2010; revision received 24 January 2011; accepted 1 February 2011

Introduction

Understanding the forces influencing genetic diversity and quantifying the spatial scale of their influence is a central task in evolutionary biology (Dionne et al. 2008). Among these forces, dispersal is of particular interest for at least two reasons. First, dispersal is directly related to the colonization processes of species in new environments. In the special case of pathogens, elucidation of the movement of individuals is crucial as such movement results in disease spread (Milgroom & Peever 2003). Second, dispersal processes determine the level of gene exchange between implanted populations, which condition their evolutionary potential (Slatkin 1987; Lenormand 2002).

Aerial fungal pathogens disperse by blowing microscopic spores into the air at high density, which renders them almost impossible to track accurately. Classically, spore dispersal is estimated by following the spatio-temporal progression of disease (Jeger 1989). However,
for practical reasons, most of such epidemiological studies have been restricted to the parcel scale. Because agricultural landscapes are often characterized by a high degree of heterogeneity and fragmentation, it has been suggested recently that crop protection investigations and strategies should shift from the field scale to the scale of the overall agricultural landscape (Plantegenest et al. 2007). Several new tools, including geographical information systems, global positioning systems (GPS), remote sensing and spatial statistics, have been developed to allow analysis and integration of the spatial components of epidemiology (Kitron 1998). Altogether, landscape epidemiology allows the essential characteristics of a landscape to be mapped and subsequently tested for any correlation between those elements and the spatial patterns of disease dynamics. Such an approach has thus far been applied mainly to insect-vector diseases (Fabre et al. 2005).

Landscape epidemiology may be extended by integrating tools from landscape genetics (Archie et al. 2009). This latter discipline combines population genetics, landscape ecology and spatial statistics to assess how landscape features and environmental variables influence gene flow and spatial genetic variation (Guillot et al. 2005b). A growing number of studies have been conducted in this emerging field (Coulon et al. 2006; Fontaine et al. 2007; Parisod & Bonvin 2008; Sahlsten et al. 2008), but until now, landscape genetic approaches have seldom been applied to pathogen populations. The new statistical tools used in these approaches are not based on populations defined a priori (Storfer et al. 2007) and thus appear well suited to organisms that are distributed uniformly over a given space. Landscape genetics could enhance research into landscape epidemiology because it enables the reconstruction of evolutionary relationships between parasites over a wide range of spatial scales, ranging from within individual hosts to between geographical locations (Archie et al. 2009). In addition to understanding the environmental factors influencing pathogen spread, this fundamentally new approach should improve our ability to track parasite movements, identify parasite origins and delineate population boundaries. In the case of pathogens, this latter point is of crucial importance because it should help identify management units.

Landscape genetics is based on the use of spatial genetics approaches that rely mostly on clustering and isolation-by-distance (IBD) models (Guillot et al. 2009b). Clustering approaches consider individual multilocus genotypes rather than allele frequencies to ascertain population membership or the parental origin of individuals (Rannala & Mountain 1997; Pritchard et al. 2000; Manel et al. 2005) and use clustering algorithms to sort individuals according to their genetic profile. Hence, populations comprising individuals that regularly exchange genes can be delineated without a priori knowledge of their boundaries (Pritchard et al. 2000; Guillot et al. 2005a; Francois et al. 2006). A strength, but also a potential drawback, of this kind of approach is that the measure of genetic structure used to reflect gene flow is averaged over time. It is thus of prime importance to account for historical influences when interpreting contemporary genetic structures (Balkenhol et al. 2009). Furthermore, because dispersal is generally restricted in space for many species, continuous spatial patterns of genetic differentiation with geographical distance or IBD patterns (Rousset 2000) can induce clustering bias. Thus, the existence of IBD patterns should be tested in parallel with clustering analysis (Guillot & Santos 2009).

*Mycosphaerella fijiensis*, which causes black leaf streak disease (BLSD)—the most destructive leaf disease of banana (Stover & Simmonds 1987)—is a good example of a recent fungal pandemic (Mourichon & Fullerton 1990). *M. fijiensis* is a haploid and heterothallic ascomycete fungus that spreads via three modes: movement of infected plant material, and dispersal of spores produced from either sexual (ascospores) or asexual (conidia) reproduction (Gauhl 1994). Whereas conidia are dispersed mainly over short distances around the plant, wind-transported ascospores represent the predominant dispersal mode of *M. fijiensis* (Marin et al. 2003) and viable spores can be carried over longer distances (Parmell et al. 1998). Populations (groups of individuals living in close vicinity) of *M. fijiensis* are panmictic (Rivas et al. 2004), and their demography is relatively stable throughout the year in untreated plantation areas. Because of these latter characteristics, *M. fijiensis* is considered as a model species for fungal pathogen population genetics studies. Previous studies assessing *M. fijiensis* population structure using molecular markers have been realized at the global and continental scale (Carlier et al. 1996; Rivas et al. 2004). A more recent study examining the structure of *M. fijiensis* populations in two recently colonized areas (Halkett et al. 2010) showed that current population genetic structure may carry traces of the historical colonization event. In particular, a break in gene frequencies was observed along a 300-km-long transect in Cameroon. This break occurs in a hilly region. However, the coarse sampling grain in the previous sampling (performed in one dimension, every 30 km) precludes any firm conclusion that a landscape feature (e.g., a mountain) acted as a barrier to *M. fijiensis* gene flow. Alternatively, such differences in gene frequency could have arisen simply owing to demographic processes during colonization (Rees et al. 2009; Halkett et al. 2010). *Mycosphaerella fijiensis* is distributed almost continuously over space, and the discreteness of
the sampling design may have hindered the detection of a more gradual variation in allele frequencies. Thus, the use of analytical methods assuming continuous populations — together with a finer grained two-dimensional sampling scheme — should help better delineate population boundaries and assessment of \textit{M. fijiensis} spatial population genetic structure in this area.

The main objectives of this study were to describe how \textit{M. fijiensis} populations are structured genetically over space at the agro-system scale and to assess the influence of landscape features on spatial genetic structure. For this purpose, the genetic structure of a \textit{M. fijiensis} population in an area of about $50 \times 50$ km in Cameroon was analysed using spatial genetics approaches. This area encompasses the genetic discontinuity previously detected (Halkett \textit{et al.} 2010) and harbours several landscape features that could act as barriers to gene flow (e.g., mountains and forests, Fig. 1). Here, we address the following questions: (i) Is the previously detected discontinuity in gene frequencies the result of a smooth gradient owing to IBD or of a more abrupt variation? (ii) In the latter case, can this genetic discontinuity be delineated spatially? (iii) Does its location match any specific landscape feature(s) of the agro-ecosystem? In the following, particular attention has been paid to the detection, quantification and location of discontinuities in gene frequency. Genetic analyses were performed using both clustering and a novel \textit{FST}-based approach. We compared clustering algorithms that either take spatial information into account or not. The \textit{FST}-based approach used in this study considers continuous variation in gene frequency across the population and was adapted especially for haploid organisms.

Materials and methods

Study area and sampling strategy

In 2005, \textit{M. fijiensis} isolates were collected from untreated plantain leaves infected with BLSD in a $50 \times 50$ km agricultural area located in southwest Cameroon. Plantains are cooking bananas [subgroup plantain, triploid (AAB) interspecific hybrid between \textit{Musa acuminata} (AA) and \textit{Musa balbisiana} (BB)]. The sampled area encompassed two sites (previously sampled in 2002: sites 5 and 6, Fig. 1) that displayed sharp genetic discontinuity (Halkett \textit{et al.} 2010). Plantains are grown extensively by small farmers in the sampled area in scattered small plots without the use of any fungicide for BLSD control. Plantain (a very important food crop in this part of Cameroon) is distributed mostly around anthropic activities, i.e., residential areas, villages, roads. The crop is practically nonexistent in natural forests and at high altitudes (>1500 m). The spatial domain sampled encompasses a high mountain (Mont Koupé, 2064 m), many forests and a broad river (Mungo River). Thus, banana plantain repartition in this area is fragmented by several natural factors that could act as potential natural barriers to \textit{M. fijiensis} gene flow. The samples were collected from easily accessible areas (roadside verges). At each sampling site, infected leaf fragments were collected from a few closely spaced banana plants, and GPS coordinates were recorded with a precision of 10 m (WGS84 data format). Between 2 and 8 \textit{M. fijiensis} individuals were isolated per sampling site. Two specific sampling designs were used according to the type of analysis (see Fig. 1 for location and distribution of sampling sites).

![Fig. 1 Sampling distribution over the study area: Left: study site (50 × 50 km) in a region in the South-West of Cameroon. Single black dots correspond to sampled sites, white circles correspond to populations no 5 and no 6 sampled during the previous study (Halkett \textit{et al.}, 2010). A total of 566 individuals (lesions) were sampled in 105 different sites. Right: Schematic representation of the 33-km-long transect showing the three classes of nested sites.](image)
Two-dimensional sampling for genetic clustering analysis. We employed a 2D sampling strategy on 105 sites spread throughout the 50 × 50 km area studied, without any a priori knowledge of population units and limits. Each site was sampled every 3 km along all roads in the spatial domain considered. A total of 567 isolates were isolated from this ‘2D sampling’.

One-dimensional sampling for IBD analysis. In the middle of the sampling area, 90 sites were sampled along a continuous 33-km-long transect according to three distance classes: 1 km, 250 and 50 m (embedded transect). A total of 321 isolates were isolated from this ‘1D sampling’. Sites more than 1 km distant (one-third of the one-dimensional sampling) were also included in the 2D sampling analysis.

Isolation procedure and molecular analysis

Mycelium cultures initiated by single ascospores isolated from necrotic lesions bearing perithecia were identified as belonging to the species *M. fijiensis* and stored as described in Carlier et al. (2002). Mycelium from each isolate was grown on solid medium (300 mL V8, 3g CaCO₃, 20 g agar per litre, pH 6) for 2 weeks at 25 °C, dried for 2 days at 55 °C and ground. Genomic DNA was extracted using a modification of the method described by Garnier Gangrat & Dillmann (1992) implemented in Amersham. Genomic DNA was extracted using a modification of the method described by Rogers & Bendich (1988) as detailed in Halkett et al. (2010). Genotyping was carried out using two polymerase chain reaction (PCR) multiplex sets comprising 17 microsatellite loci (Neu et al. 1999). PCR conditions are detailed in Halkett et al. (2010). Finally, amplified fragments were detected and separated on a 96 capillary MegaBACE 1000 (Amersham) and analysed for length variation using GENETIC PROFILER v.1.5 Software (Amersham).

Data analysis

Basic population genetic analyses on the global sample. Population gene diversity was estimated according to the expected heterozygosity [\(H_E\), unbiased estimate calculated according to (Nei 1978)] and the number of alleles (\(N_A\)). These two indices were calculated using FSTAT 2.9.3 (Goudet 1995). Linkage disequilibrium between all pairs of loci was tested using Fisher’s exact tests (Garnier-Gere & Dillmann 1992) implemented in GENEPOP (Rousset 2008) run on the web (http://wbiom ed.curtin.edu.au/genepop/) [Markov chain parameters: dememorization number = 2000, number of batches = 250, number of iterations per batch = 2000]. We used the false discovery rate (FDR) procedure implemented in the R package QVALUE (Storey 2002) to control for multiple testing. The resulting adjusted P-values are called Q-values. We measured multilocus linkage disequilibrium \(r_{Q}\) based on the index of association (Brown et al. 1980). This statistic is corrected for the number of loci. Departure from random association of alleles across loci was assessed by bootstrapping alleles among individuals 1000 times independently for each locus. The genotypic diversity, defined as the probability that two individuals taken at random have different multilocus genotypes, was measured. Multilocus linkage disequilibrium and genotypic diversity were calculated using MULTILOCUS (Agapow & Burt 2001).

IBD analysis. Isolation by distance was assessed using an ad hoc haploid version of the ‘IBD between individuals’ suboption available in the forthcoming version 4.1 of GENEPOP (Rousset 2008). For continuous populations, the original version considers \(Q_r\) as the probability of identity of genes between individuals at (geographical) distance \(r\) and \(Q_o\) as the probability of identity of two genes within an individual to calculate the parameter \(a_r\) so that \(a_r \equiv \frac{Q_o - Q_r}{1 - Q_r}\). Here, this original model was modified for use with haploid organisms collected over a continuous spatial domain, where \(Q_o\) should be replaced by the probability of identity of genes in individuals that would compete for the same position (Robledo-Arnuncio & Rousset 2010). This cannot be estimated from real data but can be approximated by the degree of identity among the closest individuals. An analogue of the \(a_r\) parameter is then calculated to estimate the quantities \(I_f = \frac{Q_0 - Q_r}{1 - Q_r}\) where \(Q_0\) is the probability of identity between two genes sampled on two sites separated by a given distance \(r\) and \(Q_0\) is the probability of identity of two genes within sites. Because the individuals in this case are haploids, at least two individuals per sampling site are necessary to estimate \(Q_0\). The significance of regression slopes between genetic distances \(I_f\) and the logarithm of geographical distances was evaluated in GENEPOP from Mantel tests performed across 100 000 permutations. Because the linear relationship of these variables is expected to hold best at distances greater than the mean parent—offspring dispersal distance (\(\sigma\)) and to progressively deviate from linearity at distances larger than 10–15\(\sigma\) (Rousset 2000; Leblois et al. 2003), and because we do not have a real measure of \(\sigma\) for our pathogenic fungus, we considered a 33-km-long east—west nested transect according to three distances classes: 1 km (\(L = \text{large scale represented by 33 sites}\)), 250 m (\(M = \text{medium scale represented by 41 sites}\)) and 50 m (\(S = \text{small scale represented by 32 sites}\); see Fig. 1). IBD was tested at
Assessing genetic structure using clustering methods. We applied two clustering algorithms to the 2D data set to infer population structure and assign individuals to genetic entities. First, a nonspatial algorithm, routinely used for such analysis (\textsc{structure} 2.2; Pritchard \emph{et al.} 2000), and, second, a clustering algorithm that can take spatial information into account (\textsc{geneland} 3.1.4; Guillot \emph{et al.} 2005a). The base units of both of these methods are individual multilocus genotypes. Subdivision of the data into clusters in these methods is carried out by minimizing Hardy–Weinberg disequilibrium and gametic phase disequilibrium between loci within groups.

The \textsc{structure} estimation method relies on a two-step procedure. First, the number of clusters (\(K\)) that supposedly best explains the data set is inferred from the posterior probability distribution of the data given the number of clusters, \(P(X|K)\). We tested all values from \(K = 1\) to 10 with 600 000 iterations of the Markov chain Monte Carlo (MCMC) algorithm following a burn-in of 60 000 iterations. We used the admixture model considering the correlated allele frequencies model but tested some runs by varying the underlying model (i.e., no admixture, and uncorrelated allele frequencies model). The number of clusters, \(K\), is calculated according to the method presented in Evanno \emph{et al.} (2005). Moreover, for each \(K\), we calculated the mean posterior probability over its runs and used this value to estimate the posterior probability of each \(K\) using the formula given by Pritchard & Wen (2003). Second, we ran 1 000 000 iterations with a burn-in of 100 000 iterations for the best fixed value of \(K\). Regardless of the processing step used, five independent runs were performed to check for the consistency of Markov chain runs.

In the \textsc{geneland} procedure, all parameters (including \(K\)) are co-estimated simultaneously by the MCMC algorithm. We used both the independent and the correlated model of allele frequencies and long MCMC runs to ensure convergence of the chain. The difference between the independent (uncorrelated) and the correlated model of allele frequencies comes from the prior distributions placed on allele frequencies. The uncorrelated model turns out to give the same a priori probability to any allele frequencies (independent across loci and populations), whereas the correlated model considers that allele frequencies tend to be similar across populations (for instance, alleles that are rare in one population might also be rare in other populations). Guillot recommended that data analysis begins with the uncorrelated frequency model, before checking how these initial results are modified by the use of the correlated frequency model (Guillot \emph{et al.} 2009a, software manual). We followed these recommendations by performing five independent runs with 1 000 000 iterations (thinning = 100) allowing \(K\) to vary from 1 to 10 with both models. The maximum rate of the Poisson process was fixed at 500, the maximum number of nuclei in the Poisson-Voronoi tessellation at 300 and the potential error for spatial coordinates at 2 m.

The putative number of clusters was inferred from the distribution of the model value of \(K\) along the run with the highest likelihood. The run with the highest PPD (posterior probability of the data) was post-processed (with a burn-in of 5 \(\times\) 10\(^4\) iterations) in order to obtain posterior probabilities of population membership for each individual and each pixel of the spatial domain. We also used \textsc{geneland} without taking spatial information into account.

Spatial genetic structure assessed through an \(F_{ST}\)-based approach. We used the modified estimator of between-site genetic differentiation \(F_{ST} = Q/Q_0\) presented above to assess the spatial re-partition of gene frequency differentiation between each site of the 2D sampling and a reference population (instead of considering all pairs of between-site comparisons as previously). We gathered all samples as the reference population but considered only those sites where at least three individuals have been genotyped to properly assess local genetic diversity. We then mapped the variation in the local genetic differentiation estimates of \(F_{ST}\) using the interpolation MBA package implemented in the R statistical software with the aim of comparing the maps obtained with clustering analyses.

Estimates of basic statistics and genetic differentiation between clusters. The expected heterozygosity (\(H_e\)), the number of alleles (\(N_a\)), the linkage disequilibrium between all pairs of loci and the multilocus linkage disequilibrium (\(r_D\)) were calculated within each cluster as presented in the section on ‘Basic total population genetic analysis’. We used two-sided group comparisons with 1000 permutations implemented in \textsc{fstat}2.9.3 to test for significant differences in multilocus gene diversity between the two clusters. The level of genetic differentiation at microsatellite loci among clusters was estimated as \(F_{ST}\) (Weir & Cockerham 1984) using \textsc{genepop}. We used the exact likelihood ratio (G) test implemented in this software to test for significant values of \(F_{ST}\) between inferred clusters.

© 2011 Blackwell Publishing Ltd
Results

Whole data set genetic characterization

All 17 microsatellites were polymorphic but displayed large differences in allelic variability. The mean number of alleles ranged between 2 and 22 (5.71 on average), and gene diversity for each locus ranged from 0.02 to 0.83, with an average value of 0.40 for all loci (Table 1). Only two pairs of loci (Fb05&Fh12b and F26b&Fe05b) showed significant linkage disequilibrium after FDR control (Q-value <0.05). Genotypic diversity was very high (0.9991), and we observed low and nonsignificant values of multilocus estimates of linkage disequilibrium (rD = −0.037). These results are consistent with the high rate of sexual reproduction of *M. fijiensis.*

Clustering results

**STRUCTURE results.** The estimated logarithm of likelihood for data analysed with STRUCTURE was highest for K = 1, and the posterior probability of having one population was equal to 1 whatever the underlying model chosen (results for admixture and the correlated allele frequencies model are presented in Table 2). For K > 1, none of the 567 individuals could be assigned to a given cluster as each individual had a similar probability of belonging to each cluster.

**GENELAND results.** Taking into account the spatial locations of samples in the analysis (spatial mode), we obtained slightly different results according to the underlying model of allele frequencies considered. Considering the uncorrelated model of allele frequencies (UFM), GENELAND clearly inferred the presence of two clusters (Table 2). In these runs, 543 isolates were assigned to the first cluster and 24 isolates were assigned to the second cluster (assignment threshold or posterior probability of memberships = 0.99). These two clusters appeared to be significantly genetically differentiated (FST = 0.157, P < 0.001) (Table 3) and had distinct spatial domains (Fig. 2a), with the second cluster being restricted to the extreme northwestern part of the sampling area. Considering the correlated model of allele frequencies (CFM), GENELAND inferred the presence of four clusters (Table 2). In these runs, 24 isolates (those assigned to the second cluster using the uncorrelated model of allele frequencies) were still assigned to a distinct cluster (Fig. 3a, same location as in Fig. 2a) but with a lower assignment threshold (0.90). Applying the same assignment threshold, most of the remaining individuals (450 isolates) were gathered in another cluster, located at the east side of the sampling area (Fig. 3d). These two clusters are largely differentiated (FST = 0.204, P < 0.001). The last few individuals can be confidently assigned to the two remaining clusters.

Table 1 Summary of basic information and overall population genetic analyses of *Mycosphaerella fijiensis*

<table>
<thead>
<tr>
<th>Locus</th>
<th>N</th>
<th>HE</th>
<th>NA</th>
</tr>
</thead>
<tbody>
<tr>
<td>F96</td>
<td>556</td>
<td>0.02</td>
<td>3</td>
</tr>
<tr>
<td>F19</td>
<td>567</td>
<td>0.03</td>
<td>5</td>
</tr>
<tr>
<td>F62</td>
<td>566</td>
<td>0.10</td>
<td>4</td>
</tr>
<tr>
<td>F09</td>
<td>548</td>
<td>0.10</td>
<td>3</td>
</tr>
<tr>
<td>N194</td>
<td>562</td>
<td>0.28</td>
<td>3</td>
</tr>
<tr>
<td>Fb05</td>
<td>534</td>
<td>0.28</td>
<td>6</td>
</tr>
<tr>
<td>F28</td>
<td>558</td>
<td>0.36</td>
<td>2</td>
</tr>
<tr>
<td>F26b</td>
<td>558</td>
<td>0.37</td>
<td>6</td>
</tr>
<tr>
<td>F04</td>
<td>558</td>
<td>0.41</td>
<td>4</td>
</tr>
<tr>
<td>N203</td>
<td>566</td>
<td>0.46</td>
<td>3</td>
</tr>
<tr>
<td>F40</td>
<td>562</td>
<td>0.49</td>
<td>3</td>
</tr>
<tr>
<td>Fe09</td>
<td>554</td>
<td>0.51</td>
<td>4</td>
</tr>
<tr>
<td>Fe05b</td>
<td>565</td>
<td>0.56</td>
<td>3</td>
</tr>
<tr>
<td>N244</td>
<td>555</td>
<td>0.64</td>
<td>7</td>
</tr>
<tr>
<td>N137</td>
<td>566</td>
<td>0.72</td>
<td>9</td>
</tr>
<tr>
<td>F01</td>
<td>520</td>
<td>0.72</td>
<td>10</td>
</tr>
<tr>
<td>Fh12b</td>
<td>563</td>
<td>0.83</td>
<td>22</td>
</tr>
<tr>
<td>All loci</td>
<td>562</td>
<td>0.40</td>
<td>5.71</td>
</tr>
</tbody>
</table>

N, sample size; HE, unbiased estimate of gene diversity (Nei 1978); NA, allele number, are computed over the entire sampled area among the different loci.

© 2011 Blackwell Publishing Ltd
which appeared to be somewhat intermediate between the two contrasted genetic entities (Fig. 3b, c). These two additional clusters may represent additional subtle substructuring as revealed by lower differentiation levels (all pairwise $F_{ST} < 0.10$). Ignoring the spatial locations of samples (nonspatial mode) and considering the uncorrelated model of allele frequencies, GENELAND inferred the presence of two populations (Table 2) and the map of the posterior probabilities of population membership (after interpolation using the MBA procedure) was similar to that obtained with the spatial mode (results not shown).

### Spatial genetic structure assessed through an $F_{ST}$-based approach

**$F/(1–F)$ statistic interpolation.** In parallel to clustering analysis, the interpolation of local genetic differentiation into the reference population allows the spatial variation in gene frequencies to be represented (Fig. 2b). Three sites located in the uppermost left part of the map stood out, as they displayed the highest value of genetic differentiation compared to the rest of the sampling domain. Conspicuously, the localization of this discontinuity matched that detected using GENELAND.

**$F/(1–F)$ and $1–Q_0$ statistics distribution around the genetic discontinuity.** To better characterize the genetic discontinuity near this area, we further analysed the distribution of the $F/(1–F)$ statistic and within-site gene diversity $1–Q_0$ of the sites around the genetic discontinuity (Fig. 4). In this graph, each point represents a site, and the four sites at higher latitude (at the right of the graph) correspond to sites with individuals assigned to the second cluster by GENELAND using the UFM (in the northwest of the map). We can see that values of $1–Q_0$ are lower for sites located near the discontinuity.

### Table 3  Summary of basic information and genetic analyses realized on the two clusters of *Mycosphaerella fijiensis* inferred by the GENELAND analysis when considering a spatial and uncorrelated allele frequencies model

<table>
<thead>
<tr>
<th>Locus</th>
<th>Cluster 1 ($N = 543$)</th>
<th>Cluster 2 ($N = 24$)</th>
<th>$F_{ST}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_E$</td>
<td>$N_A$</td>
<td>$H_E$</td>
</tr>
<tr>
<td>F96</td>
<td>0.04</td>
<td>3</td>
<td>0.00</td>
</tr>
<tr>
<td>F19</td>
<td>0.06</td>
<td>5</td>
<td>0.00</td>
</tr>
<tr>
<td>F01</td>
<td>0.68</td>
<td>10</td>
<td>0.78</td>
</tr>
<tr>
<td>Fh12b</td>
<td>0.79</td>
<td>21</td>
<td>0.87</td>
</tr>
<tr>
<td>Fb05</td>
<td>0.30</td>
<td>5</td>
<td>0.25</td>
</tr>
<tr>
<td>F62</td>
<td>0.19</td>
<td>4</td>
<td>0.00</td>
</tr>
<tr>
<td>F09</td>
<td>0.05</td>
<td>3</td>
<td>0.16</td>
</tr>
<tr>
<td>N203</td>
<td>0.52</td>
<td>3</td>
<td>0.34</td>
</tr>
<tr>
<td>F04</td>
<td>0.51</td>
<td>4</td>
<td>0.23</td>
</tr>
<tr>
<td>N194</td>
<td>0.45</td>
<td>3</td>
<td>0.00</td>
</tr>
<tr>
<td>N137</td>
<td>0.55</td>
<td>9</td>
<td>0.78</td>
</tr>
<tr>
<td>N244</td>
<td>0.44</td>
<td>7</td>
<td>0.72</td>
</tr>
<tr>
<td>F28</td>
<td>0.49</td>
<td>2</td>
<td>0.08</td>
</tr>
<tr>
<td>Fe05b</td>
<td>0.61</td>
<td>3</td>
<td>0.30</td>
</tr>
<tr>
<td>F40</td>
<td>0.45</td>
<td>3</td>
<td>0.29</td>
</tr>
<tr>
<td>Fe09</td>
<td>0.38</td>
<td>4</td>
<td>0.42</td>
</tr>
<tr>
<td>F26b</td>
<td>0.07</td>
<td>6</td>
<td>0.54</td>
</tr>
<tr>
<td>All loci</td>
<td>0.39</td>
<td>5.59</td>
<td>0.34</td>
</tr>
</tbody>
</table>

$N$, sample size; $H_E$, unbiased estimate of gene diversity (Nei 1978); $N_A$, allele numbers, are computed over the two clusters with related $F_{ST}$ values (Fisher’s exact test: *$P < 0.05$; **$P < 0.01$, ***$P < 0.001$) among the different loci.

---

**Fig. 2** Spatial genetic structure of *Mycosphaerella fijiensis* populations over the study area: (a) Map of posterior probability of belonging to cluster 1 obtained with GENELAND considering a spatial uncorrelated allele frequencies model (left); (b) Map of the genetic differentiation ($F/(1–F)$) parameter interpolated over the spatial domain (R) (right). Values are represented by a white-to-red graded scale; black dots represent sampled sites.

© 2011 Blackwell Publishing Ltd
beyond the discontinuity. This result shows that allelic frequencies are homogeneous in the genetic entity located in the north. However, the high values of $1-Q_0$ detected in some sites around the discontinuity reveals a high genetic variance between individuals in these sites. This high variance that is not attributable to smaller sample size at these sites could be explained by the presence of mixed genotypes coming from two genetic origins. Such a structure could explain the detection in this area of the two supplementary and intermediate clusters when using GENELAND with a correlated model of allele frequencies.

Fig. 3 Maps of posterior probability of belonging to clusters 1 (a), 2 (b), 3 (c) and 4 (d) and posterior mode of population membership (e). Results obtained with GENELAND considering a spatial and correlated allele frequencies model. Values are represented by a white-to-red graded scale; black dots represent sampled sites.

Fig. 4 Relationship between the genetic diversity measured along the site around the genetic discontinuity and latitude. The genetic differentiation inter-site parameter $F/(1-F)$ is represented by the solid line, whereas the intra-site genetic differentiation parameter $(1-Q_0)$ is represented by the dashed line.
Isolation by distance. The test for IBD was significant when considering all samples, but the IBD signal disappeared when samples from cluster 1 and 2 were considered independently (Table 4). The significant IBD detected over the whole sampling domain would thus seem to stem from the second genetic cluster being located in the northwest corner of the area under study. Thus, many pairs of points that are distant at a genetic level are also distant at a geographical level.

When focusing on the 33-km-long transect situated inside the first cluster, the Mantel test of IBD was not significant whatever distance class was taken into account (Table 4).

Genetic characteristics of the inferred clusters

Given the agreement for $K = 2$ considering the spatial or nonspatial mode in GENELAND (uncorrelated model of allele frequencies) and the good concordance with the genetic differentiation analysis, we will present here the genetic characteristics of only the two main genetic entities (i.e., not considering the additional substructuring detected by GENELAND using the correlated model of allele frequency).

Multilocus values of genetic diversity (heterozygosity $H_E$) did not reveal any significant difference between the two clusters ($P = 0.341$). Large and significant genetic differentiation between the two clusters was observed ($F_{ST} = 0.157$, $P = 0.00002$). Again, $F_{ST}$ values and significance varied greatly among microsatellite loci (Table 3). No pairs of loci appeared to be linked within cluster 1, but one pair (Fb05&Fh12b) showed significant linkage disequilibrium after FDR control ($Q_{value} < 0.05$) in cluster 2, maybe because of the small size of this population ($N = 24$). We observed low, nonsignificant values of multilocus estimate of linkage disequilibrium in both clusters ($r_D = 0.041$ and 0.001 for clusters 1 and 2, respectively).

Discussion

We used two complementary spatial genetics approaches to study the dispersal and population genetic structure of the fungal pathogen Mycosphaerella fijiensis within a production area (50 $\times$ 50 km) in Cameroon. To our knowledge, this is the first study in which this methodology has been applied to a fungal pathogen. Our results revealed the presence of two highly differentiated populations in the study area. This genetic structure is in accordance with the results obtained in a previous study, in which Halkett et al. (2010) suspected a discontinuity in gene frequencies to occur along a 300-km-long transect in Cameroon. Because sampling sites were separated by approximately 30 km, it was not possible to precisely localize this discontinuity or to determine its strength. In the present study, we confirmed the presence of this genetic discontinuity and furthermore assessed the boundary of the genetic entities. It is interesting to note that the genetic discontinuity observed in 2001 can still be reported 4 years later and that the level of genetic differentiation ($F_{ST} = 0.145$) detected between the two continuously distributed populations was similar to that previously measured ($F_{ST} = 0.150$) between samples from two sites 30 km apart. Here, because of our original sampling design, we can also argue that the genetic discontinuity appeared to be sharp rather than gradual, as we observed a break in allele frequency variation that clearly delineated two genetic entities. These two populations seem to be genetically homogeneous as no IBD signal and no mutation—migration—drift disequilibrium were detected within each
genetic entity. Hence, the fact that the IBD signal disappeared when considering both populations independently instead of analysing all samples appears to be an instance of the presence of spatially structured HWE populations (further exemplified in Guillot et al. 2009b; Fig. 1). However, the site located south of the genetic discontinuity presented higher levels of within-site diversity (1−Q₀) and could account for more subtle substructuring as evidenced by GENELAND results when considering the correlated allelic frequency model. These results might indicate that the area south of the genetic discontinuity has admixed genotypes from both populations.

We analysed the spatial genetic structure of our data set using different clustering algorithms, which did not always converge to the same clustering pattern. The spatially explicit method implemented in GENELAND detected several clusters, whereas the nonspatial software STRUCTURE failed to detect any genetic structure. Discrepancy among clustering methods has been reported frequently (Chen et al. 2007; Rowe & Beebee 2007; Frantz et al. 2009), but we found only one report in which no genetic structure (K = 1) was opposed to several clusters (K > 1) (Coulon et al. 2006). Such inconsistency can arise from differences in the underlying models, in the statistical estimators or in approximations in the algorithm used to compute this estimator, and it is generally difficult to disentangle the relative effects of these three sources of disagreement (Guillot et al. 2009b). It is important to keep in mind that the two programs used in this study are based on MCMC and are hence sensitive to convergence failures. This means that the outputs of these programs might, in certain cases, not be the exact solution of the mathematical equations but rather an approximation, the quality of which remains unknown. Here, it seems that STRUCTURE performs less well in detecting population structure than GENELAND.

Because we did not obtain fully congruent results from the different clustering methods (and in order to assess whether the population delineation is an artefact of the clustering method), we also used an interpolation procedure based on the F/(1−F) classic genetic differentiation statistic. This method, which was developed for continuous population sampling, allowed us to confirm the main result obtained with GENELAND, i.e., an abrupt switch in gene frequency in the northeast part of the sampled domain, which clearly delineates two populations. Contrary to the clustering algorithms, this F/(1−F) interpolation method is not prone to MCMC convergence issues. This result suggests that applying such an approach to different clustering analysis offers a convenient means of gathering solid information on spatial genetic population structures.

On the other hand, we found large similarities in GENELAND clustering results among the underlying genetic models. Interestingly, we did not find an effect of the spatial model, which attests that the genetic structure observed is not a by-product (e.g., edge effect) of considering spatial information in the clustering algorithm (same number of clusters and posterior probability membership values). Second, even if the number of clusters detected increases from two to four, when considering the correlated frequency model, a main discontinuity in gene frequency is still detected at the same place. Inconsistency between clustering concerning CFM and UFM in biological data has already been reported (Rosenberg et al. 2002) and illustrated by simulated data sets (Falush et al. 2003). The increase in the number of clusters detected (K = 2 under CFM vs. K = 4 under UFM) results from the CFM algorithm being more able to ascertain subtle structures (Guillot 2008) that potentially go undetected when using UFM. Accordingly, the genetic structure observed using CFM is simply a subpartition of K = 2 (detected by UFM) characterized by lower differentiation levels. This substructure could be attributable to the few sites with putative admixed genotypes (high values of 1−Q₀).

Gene flow in natural populations can be influenced strongly by landscape features. For instance, two recent landscape genetics studies on animals (aphids and minks) (Lavandero et al. 2009; Zalewski et al. 2009) concluded that the presence of major rivers and mountains is barriers to gene flow. However, in the case of an airborne pathogen such as M. fijiensis, it is not easy to determine a priori which elements could represent an effective barrier to gene flow as no study to date has focused on this point (Archie et al. 2009). In the spatial domain sampled here, we considered three types of natural obstacle to M. fijiensis dispersal: a large river, forests and a high mountain. However, we found no correlation between genetic structure and any landscape feature or ecological variable. First, no obvious barrier matches the genetic discontinuity. Second, none of the hypothetical barriers occurring in the sampled domain (mountain, river, deep forests, etc.) appeared to have any effect on the population genetic structure of M. fijiensis (all obstacles to dispersal were included in the larger population). Such patterns, in which the landscape is not correlated with the genetic structure, have already been reported in other landscape genetics studies (Gautié et al. 2008; Sahlsten et al. 2008). For instance, Sahlsten et al. (2008) explained the absence of geographical barriers between two genetically differentiated populations of hazel grouse in the Scandinavian Peninsula by the post-glacial reinvasion history of this species. We propose a similar argument. Indeed, we hypothesize that the present-day genetic structure
observed during this study could be better explained by the recent history of *M. fijiensis* expansion than by current gene flow. The earliest report of *M. fijiensis* in Africa was in 1973 in Zambia (Jones 2000). It arrived in Cameroon in 1980 from neighbouring Gabon and was restricted for a while to the southeastern area near the coast (Kribi region; Fouré & Lescot 1988) before being reported in 1983 close to the study area. North of this area, we could find no information on the spread of BLSD in Cameroon. The disease was then described for the first time in Nigeria in 1986. The pattern observed in this study, with the presence of two differentiated populations without any obvious barrier to gene flow, might reflect this history of northward expansion according to two conceivable scenarios. The first combines the effect of independent founder effects caused by the movement of a few individuals by long-distance spore dispersal (LDD or movement of infected material) as already suggested from analysis at the continental scale (Rivas et al. 2004), followed by natural gradual diffusion (GD). The genetic structure observed would then be attributable to secondary contact among two expanding patches. This original pattern has been recently empirically observed (termed ‘stratified dispersal combination’ by Parisod & Bonvin (2008) and supported theoretically by simulations (Nichols & Hewitt 1994; Ibrahim et al. 1996). However, we cannot reject a second likely scenario to explain the observed genetic structure. Indeed, recent studies have shown that low-frequency alleles can sometimes ‘surf the wave’ of the advance of a population range expansion, reaching high frequencies and spreading over large areas (Edmonds et al. 2004; Bialozyt et al. 2006; Excoffier & Ray 2008). This pattern, called ‘gene surfing’, has already been observed empirically in cultured bacterial populations (Hallatschek & Nelson 2008), although not yet in natural populations. In our study, the high variance of *F*$_{ST}$ values detected among loci is congruent with the two hypotheses outlined above (stratified dispersal combination and gene surfing). Indeed, the drastic effects of genetic drift on genetic diversity during founder effects should randomly affect the different loci observed. Spatial genetic diversity patterns could theoretically help to better elucidate the origin of the observed genetic differentiation. Indeed, in both historical scenarios, we expect a decline in genetic diversity with increasing distance from the source population as a common pattern reflecting the recurrent founder effect influencing expanding populations (Flewitt 2000). Different genetic diversity patterns could result from these two scenarios, but these may not differ enough for statistical discrimination. Indeed, if the genetic differentiation we detected was caused by a secondary contact among two expanding patches, then two sources of expansion would be present and we would expect a decrease in genetic diversity from the bottom to the top of the core of both populations. In the other case (gene surfing), we hypothesize the existence of only one source of expansion (in the south of the study area according to historical data), and in this case, a steady decrease in genetic diversity along the range of expansion (from south to north) would be predicted. However, in the present study, the number of sites in the northeastern population was too small to interpret the spatial distribution of the within-site diversity (*1−Q*$_{0}$) in this population and thus differentiate between the two scenarios. At a larger scale, the distribution of genotypes according to such patterns (stratified dispersal combination or gene surfing) could form a mosaic of relatively homogeneous patches (patchy genetic structure). Additional genetic structure analysis studies at a larger scale to test for such patterns would be of interest and could allow the origin of the observed genetic differentiation to be elucidated.

The extent of geographical variation in gene frequency is thought to result from a balance between forces tending to produce local genetic differentiation and forces tending to produce genetic homogeneity (Slatkin 1987). In our study, as discussed above, historical gene flow combined with genetic drift owing to small finite population size at the time of *M. fijiensis* introduction to Cameroon might be responsible for the current genetic structure observed (stratified dispersal combination or gene surfing). However, we assume high effective sizes in current *M. fijiensis* populations. Lesion-counting experiments on plantain leaves in Costa Rica revealed more than 15 000 lesions/leaf (Gauhl 1994). Even if not all lesions participate in the next generation, those that do might represent a large effective population size. Such a characteristic would have several consequences for both the evolution of *M. fijiensis* populations and the results we obtained. First, the absence of correlation between landscape features (a mountain, a deep forest and a river) and the genetic structure observed in this study should be interpreted taking into account the high effective sizes. Indeed, a recent simulation study showed that, for species with high effective size, recent barriers to effective dispersal can be difficult to detect through analysis of genetic variation (Gauffre et al. 2008). These landscape features could then act as effective barriers to contemporary gene flow although the high densities do not allow their detection by clustering methods. Second, the very high densities assumed might have prevented the detection of a significant IBD pattern on our specific transect. Indeed, according to the equation of the slope of the IBD regression line given by (Rousset 2000),  \( b = \frac{1}{2N_{eff}} \), we can easily predict that high effective size (*N*) and/or
high dispersal abilities ($\sigma^2$) might be responsible for a weak IBD signal even with low dispersal abilities ($\sigma^2$). As a consequence, no inference on dispersal distance could be gained from the embedded transect using the IBD approach.

In summary, we have uncovered a very interesting situation, i.e., one in which two populations characterized by high effective size encounter each other. Our results could provide vital information related to the dispersal abilities of individuals. Indeed, if there is no differential selection between the two populations, each group will spread smoothly into the other’s range and the genetic differentiation produced at the moment of contact will dissolve, leaving only a smooth, gradually falling cline (Endler 1977). If the population size is high enough to avoid the effects of genetic drift, then the rate of decay is thought to be affected only by quasi-deterministic spatial diffusion processes. Further analysis of the genetic discontinuity detected in the present study should be carried out after just a few more generations in order to test for the presence of such a falling cline pattern in an attempt to better characterize M. fijiensis dispersal processes.

In conclusion, our study has shown that individual-based population genetics analyses realized at a small scale can provide information on processes that have driven the recent colonization of an invasive plant pathogen. Indeed, using clustering analysis and a new $F_{ST}$-based procedure, we found evidence of a population structure in the M. fijiensis population in the southwestern plantain production area of Cameroon but found no evidence that geographical features affect current gene flow and dispersal patterns. We hypothesize that historical demographic events are responsible for the genetic structure detected and that high effective population sizes might have masked the contribution of contemporary processes to the genetic structure of these populations. This confusion between historical and contemporary processes has been poorly discussed in landscape genetic studies to date because the majority of such studies have focused on species characterized by moderate population sizes. However, plant pathogen populations are often characterized by recent expansion events and high densities. In this study, we showed that this latter demographic parameter should be taken into account more explicitly in future landscape genetics research.

Acknowledgements

We thank CARBAP (Oscar Nguidjo, Robert Dongmo and Josué Essoh Essando) for technical and logistical support during sampling in Cameroon; Luc Pignonot for isolation of fungal strains; Arnaud Estoup, Renaud Vitalis, Thomas Lenormand, Etienne Klein, Denis Bourguet, Benoit Barres and Eric Bazin for helpful discussions; Dany Lo Seen for technical help in mapping georeferenced individuals; and Helen Rothnie for her attentive reading and careful English language checking. Data used in this work were produced partly through the molecular genetic analysis technical facilities of the IFR119 “Montpellier Environnement Biodiversité”. Adrien Rieux was supported by a CIFRE doctoral fellowship from Bayer CropScience and the Agence Nationale de la Recherche et de la Technologie (ANRT-123/2008); and the Agence Nationale de la Recherche (ANR 07-BDIV-003/Emerfundis). This work was also supported by a Europaid contract ATF/UE- No 146-762/786/798/801.

References


© 2011 Blackwell Publishing Ltd


Gauth F (1994) *Epidemiology and Ecology of Black Sigatoka*, PhD Thesis. INIBAP, Montpellier, p. 120.


This study was conducted by A.R. and F.H. during their PhD thesis and post-doctoral position in the team BECo of the research unit BGPI, respectively. The main objective of A.R.’s PhD thesis is to characterize *Mycosphaerella fijiensis* dispersal processes using different population genetics and epidemiological approaches. F.H. now works on the ecology and the population genetics of forest fungal pathogens, especially the poplar rust *Melampsora larici-populina*. M.-F.Z., L.L.B. and J.C. have a common research programme dealing mainly with the dynamics and evolution of populations of the fungus *Mycosphaerella fijiensis* pathogenic on banana. V.R. adapts population genetic methods and develops models to study the evolution of plant pathogenic fungi. One of F.R.’s research interests concerns estimation methods of dispersal rates under isolation by distance and statistical analysis of spatial genetic patterns.