Molecular Ecology (2001) **10**, 947–964

A microbial population–species interface: nested cladistic and coalescent inference with multilocus data

I. CARBONE and L. M. KOHN
Department of Botany, University of Toronto, 3359 Mississauga Road North, Mississauga, Ontario, Canada L5L 1C6

Abstract

Using sequence data from seven nuclear loci in 385 isolates of the haploid, plant parasitic, ascomycete fungus, *Sclerotinia*, divergence times of populations and of species were distinguished. The evolutionary history of haplotypes on both population and species scales was reconstructed using a combination of parsimony, maximum likelihood and coalescent methods, implemented in a specific order. Analysis of site compatibility revealed recombination blocks from which alternative (marginal) networks were inferred, reducing uncertainty in the network due to recombination. Our own modifications of Templeton and co-workers' cladistic inference method and a coalescent approach detected the same phylogeographic processes. Assuming neutrality and a molecular clock, the boundary between divergent populations and species is an interval of time between coalescence (to a common ancestor) of populations and coalescence of species.

Keywords: fungi, migration, recombination, *Sclerotinia*, site compatibility, speciation

Received 2 July 2000; revision received 30 October 2000; accepted 30 October 2000

Introduction

Speciation continues the process of cessation of gene flow and divergence that begins among populations. Species concepts and the process of determining or diagnosing species do not capture this process very well (Mishler & Donoghue 1982; Mishler & Budd 1990; Avise & Wollenberg 1997; Avise & Johns 1999). For example, the biological species concept may be insufficient if gene flow ceases due to separation by time, geography or host, and the ability to interbreed is maintained (Mishler & Budd 1990; Avise & Wollenberg 1997; Dieckmann & Doebeli 1999; Kondrashov & Kondrashov 1999). Or, in interpreting the tip clades under a phylogenetic species concept (e.g. Donoghue 1985; Mishler 1985), the rank of a monophyletic lineage as an individual genotype, clone, population or species may not be explicit (Mishler & Brandon 1987; O'Donnell *et al*. 1998). Many fungi are predominantly, if not exclusively, asexually reproducing. Here it is difficult to interpret indirect evidence of genetic exchange and recombination in apparently asexual lineages. Whether such evidence reflects contemporary episodes of unobserved sexual reproduction, suggesting that the boundaries of biological species can be determined, or historical associations with a sexual ancestor is as yet unclear (Burt *et al*. 1996; Koufopanou *et al*. 1997; Geiser *et al*. 1998).

One approach that might capture more of the speciation process in determining species boundaries integrates population genetics with evolutionary systematics, ideally in studies of multiple gene loci (Templeton 1989; Avise & Wollenberg 1997). Haplotypes are the fundamental units. A haplotype is a unique combination of nucleotides in a sequence (locus). Haplotypes can be single or multilocus. Individuals sharing a haplotype have the same nucleotides at variable positions as well as at invariant positions. The integrative approach would work on two scales (Templeton 1994), haplotypes in populations and populations in species. From these perspectives, key processes could be differentiated, such as the frequency of migration or recombination, as well as dispersal, range expansion or shifts in host association, that either drive the wedge of divergence or maintain cohesion (Mishler & Budd 1990; Avise & Wollenberg 1997; Dieckmann & Doebeli 1999; Kondrashov & Kondrashov 1999). Mutational histories should converge at a population–species interface at which the more recent divergence of populations is distinct from speciation and from which the causes of divergence events can be inferred. The power of the integrative approach has not yet been evaluated empirically with a large sample of multilocus

Correspondence: I. Carbone. Fax: 01 905 828 3792; E-mail: icarbone@credit.erin.utoronto.ca

haplotypes — deterred perhaps by the logistic and computational challenges.

Here, we compare a sample of multilocus haplotypes in populations with a sample of populations in species and we identify a population–species interface. Ascomycete fungi of the genus *Sclerotinia* provide advantages for identifying this interface not afforded by many plant and animal systems. They are haploid with a significant clonal population component; even at the population scale, robust phylogenies can be inferred from nuclear loci and combined analysis of multiple loci or data sets is possible (Carbone *et al*. 1999). Recombination can be detected and localized in a multilocus or combined data set phylogeny as incompatiblity or incongruence (Carbone *et al*. 1999). That these are plant parasites of agricultural and wild plants with overlapping host specificities and wide geographical distribution makes them very appropriate for tests of association of phylogeographical and host factors with haplotypes in population divergence and speciation. Their natural histories are well understood, especially in agriculture, and they can be sampled over large, continuous geographical areas.

Our species-scale sampling began with a group of morphologically similar taxonomic species, *Sclerotinia sclerotiorum* (Lib.) de Bary, *S. minor* Jagger and *S. trifoliorum* Erikss, and a new, as yet undescribed species, *Sclerotinia* sp. 1. These species have been assumed to share recent common ancestry based on a lack of sequence polymorphism in one genomic region (Carbone & Kohn 1993; Holst-Jensen *et al*. 1998). It was not known whether the reference species sampled at the population scale, *S. sclerotiorum*, was one panmictic population, or several populations distinguished by their histories of geographical distribution, reproduction or host specialization. Using sequence data from multiple nuclear loci, we sought to determine whether divergence times of populations could be distinguished from divergence times of species.

Materials and methods

Fungus and host biology

All species included in this study are haploid, filamentous ascomycetes. Asexual reproduction is by means of sclerotia. Sclerotia are soil-borne, melanized, multicellular propagules that can undergo dormancy and remain viable for documented periods of 4–5 years (Adams & Ayers 1979) and probably longer under some conditions. After physiological conditioning, sclerotia germinate as either somatic filamentous hyphae or sexual fruitbodies (apothecia). Apothecia produce air-dispersed meiospores (ascospores). Infection can result either from hyphae developing from sclerotia, or from ascospores. Once lesions develop in the host, sclerotia are formed.

Negligible host specificity has been demonstrated in *Sclerotinia sclerotiorum*. The species has an extremely wide host range of at least 408 plant species in 75 families (Boland & Hall 1994). Apothecia of *S. sclerotiorum*, associated with a variety of crops such as soybean and other beans, sunflower, vegetable crops and canola (oilseed rape; *Brassica napus* L. or *B. rapa* L.) in temperate North America and Europe, are generally produced only during spring and early summer. This is when conditions are also conducive to flowering of spring-planted canola cultivars. Canola has been cultivated in Western Canada since the 1940s, but acreage did not expand greatly until 1970 and canola-quality rapeseed was introduced around 1980 (R. Morrall, personal communication). In the subtropical southeastern US, apothecia associated with infection of canola, cabbage and other crops are produced when conditions are conducive, potentially any time that is not too hot, too dry or too cold. Here the fungus may produce multiple sexual and asexual generations per year, depending on the host crops and how they are grown, potentially infecting one crop in the autumn and another in the spring. While susceptible crops have been grown in the southeastern US, probably predating European settlement, canola is a relative newcomer. It was introduced commercially into Georgia in the late 1980s as winter-type cultivars planted in the autumn, and further south in the early 1990s (D. Phillips, personal communication). Apothecia of *S. sclerotiorum* associated with the wild plant, *Ranunculus ficaria*, in Norway are produced under the flowering plants in May or early June, depending on seasonal conditions.

S. minor is mainly a pathogen of peanut, lettuce, sunflower (especially in Australia) and, to some extent, potato, but has a host range on dicots that overlaps with that of *S. sclerotiorum* (Melzer *et al*. 1997). *S. trifoliorum* is restricted to the Fabaceae (e.g. forage legumes), although *S. minor* and *S. sclerotiorum* have also been reported from this family. *Sclerotinia* sp. 1 is an undescribed species from wild *Taraxicum* and *Caltha* in Norway (Holst-Jensen *et al*. 1998).

The mating systems of these species differ in detail. *S. sclerotiorum* preferentially self-fertilizes and does not cross in the laboratory — although forced crosses with nutritional, antibiotic or physiological markers have not been attempted. Because selfing in fungi is not obligate, indirect evidence for genetic exchange and recombination has been interpreted as evidence that some outcrossing occurs in *S. sclerotiorum* (Kohli & Kohn 1998). The mating system of *S. minor* is not well understood, but the strong clonal component in populations is combined with some evidence of outcrossing and recombination (Patterson 1986; Carbone 2000). *S. trifoliorum* is an obligate outcrossing species with a proposed mating-type switching mechanism (Uhm & Fujii 1983a,b).

Sampling and data collection

Population-scale analysis was based on 341 *S. sclerotiorum* strains from different hosts and geographical locations (Table 1).

Table 1 Samples at population and species scales

Species	Host	Locality	Year	Size (n = 341)
Population scale				
Sclerotinia sclerotiorum	*Brassica napus*	Alberta (AB)*	1992	40
	or *B. rapa*	Saskatchewan (SK)	1991	4
	(Canola, CA)†	Ontario (ON)	1989	1
		Norway (NO)	1993	15
		Georgia (GA)	1990, 91, 92, 97	48
		Alabama (AL)	1997	15
		South Carolina (SC)	1997	6
		Florida (FL)	1997	2
		North Dakota (ND)	1997	1
	Brassica oleracea	North Carolina (NC)	1994	40
	(Cabbage, CB)	New York (NY)	1996	31
		Louisiana (LA)	1996	16
	Nicotiana tabacum (Tobacco, TB)	North Carolina (NC)	1991, 94, 95	11
	Actinidia chinensis (Kiwi, KW)	New Zealand (NZ)	1993–97	38
	Apios americana (Groundnut, GN)	Louisiana (LA)	1996	5
	Helianthus annuus (Sunflower, SF)	North Dakota (ND)	1997	1
	Ranunculus ficaria	Sandvika, Norway (SV)	1993	10
	(Buttercup, RF)		1994	41
		Vestfold, Norway (VF)	1993	13
	Geranium sp.§ (Cranesbill, GM)	Louisiana (LA)	1996	1
	Cannabis sativa (Hemp, CS)	New Zealand (NZ)	1994	1
	Raphanus sp.§ (Radish, RD)	Alabama (AL)	1997	1
Species scale				Size (n = 44)
S. sclerotiorum	*Solanum tuberosum* (potato)	Herdum, Larvik, Norway	1994	3
		Slagen, Tonsberg, Norway	1994	2
S. minor	*Lactuca sativa* (lettuce)	Palumbo farm, Holland Marsh, ON	1998	5
		Muck Crops Research Station‡	1998	5
		Visser farm, Holland Marsh, ON	1998	5
	Lactuca sativa	Miedema farm, Holland Marsh, ON	1998	5
		Cook's Bay Gardens, Keswick, ON	1998	5
	Arachis hypogaea	Virginia	1987	1
	Lactuca sativa	ON	1984	2
	Lactuca sativa	Ohio	1985	1
S. trifoliorum	*Trifolium repens*	Tasmania	1973	1
	Medicago sativa	Virginia	1984	1
	Trifolium pratense	Wisconsin	1987–93	2
	Medicago sativa	Mississippi	1987–93	2
	Vicia villosa *Pisum sativum* var. *arvense*	Louisiana	1987–93	2
Sclerotinia sp. 1	*Caltha palustris*	Norway	1992	1
	Taraxacum sp.§	Norway	1992	1

*In brackets, the two-letter abbreviation for geographical location. †In brackets, colloquial host name followed by the two-letter abbreviation for host. ‡Muck Crops Research Station, 1125 Woodchoppers Lane, Holland Marsh, ON. §Not identified to species.

The species-scale analysis was on the same sample plus 44 strains of *S. minor*, *S. trifoliorum* and an undescribed new species, designated sp. 1 (Table 1). For the population-scale analysis, strains were screened for nucleotide sequence polymorphisms in the intergenic spacer region of the nuclear ribosomal RNA (rRNA) gene repeat unit (IGS, 4000 bp), an anonymous nuclear region (44.11, 700 bp) and portions of genes encoding translation elongation factor 1 alpha (EF-1α, 350 bp), calmodulin (CAL, 500 bp), chitin synthase 1 (CHS, 300 bp), actin (ACT, 300 bp) and ras protein (RAS, 350 bp) (Carbone & Kohn 1999). For the species-scale analysis, the same loci were screened with two exceptions. First, the promoter region of the IGS (400 bp) was used instead of the entire IGS, which at the species scale presented highly divergent paralogues. Second, because 44.11 could not be amplified among closely related species, the internal transcribed spacer 1 (ITS-1, 200 bp) of the nuclear rRNA gene repeat unit was used instead. Linkage relationships among loci were determined from Southern hybridizations of electrophoretic karyotypes with probes from each locus. DNA polymorphisms were identified using single strand conformation polymorphisms (SSCPs) and DNA sequencing, as described previously (Carbone *et al*. 1999). For each locus, DNA sequences were aligned using CLUSTAL W version 1.7 (Thompson *et al*. 1994). Multiple sequence alignments were adjusted manually using SEQUENCE ALIGNMENT EDITOR, version 1.0 alpha 1 (http://evolve.zoo.ox.ac.uk/software/Se-Al/Se-Al.html) and collapsed into distinct haplotypes using CONVERT, version 1.0 Beta Build 8 (I. Carbone, Y. Barakat, J. Lee). CONVERT also generated a table of base substitutions and encoded indels for each distinct haplotype.

Haplotype network estimation

We reconstructed the evolutionary history of haplotypes on both scales using a combination of parsimony, maximum likelihood and coalescent methods, implemented in a specific order. While we were guided by the current concepts of these taxonomic species, the fundamental units in our study were haplotypes, with populations considered as groups of haplotypes and species as groups of populations. At both scales, phylogenetic trees were first inferred with unweighted parsimony using PAUP* 4.0 (Swofford 1998) and then with maximum-likelihood quartet puzzling, as implemented in TREE-PUZZLE 4.0.2 (Strimmer & von Haeseler 1996). Maximum likelihood trees were based on the Hasegawa–Kishino–Yano (HKY; Hasegawa *et al*. 1985) model of nucleotide substitutions, assuming either uniform or varying rates of substitutions among sites. A likelihood ratio test was used to test for rate heterogeneity among sites (uniform vs. gamma distributed rates) and to test for a molecular clock (clock-like DNA sequences vs. no clock).

Haplotype networks were inferred for each of the loci in the population scale sample using an approach similar to the method of statistical parsimony (Templeton *et al*. 1992). Deviations from parsimony that arose because of recombination were resolved by generating compatibility matrices for variable sites at each locus using SEQ2TR (R.C. Griffiths; http://www.maths.monash.edu.au/~mbahlo/mpg/gtree.html). If compatibility matrices revealed recombination blocks, i.e. recombining segments within which no recombination was detected, the locus was subdivided at the boundary between the segments.

Nested design and permutation analysis

Haplotype networks at the population scale were converted into a nested design following nesting rules (Templeton *et al*. 1987; Templeton & Sing 1993; Crandall 1996; Gómez-Zurita *et al*. 2000). The nested design was used to perform random, two-way, contingency permutation analysis using GEODIS 2.0 (Posada *et al*. 2000; http://bioag.byu.edu/zoology/crandall_lab/geodis.htm), to evaluate clades for significant geographical position-haplotype and host–haplotype associations. In clades with significant haplotype–geographical associations, distance analysis was performed using GEODIS and then interpreted using the biological inference key given in the appendix of Templeton *et al*. (1995). The inference key is based on two assumptions. First, older haplotypes have more mutational derivatives and are located in the interior portions of the network; more recently derived haplotypes are at the tips. Second, older haplotypes are more geographically dispersed than more recently derived haplotypes. Bearing in mind that 0-step clades are single haplotype units, the geographical distances of interior clades are contrasted with the distances of tip clades and significant differences between these distances (see Fig. 1 of Templeton *et al*. 1995). Significantly large interior distances relative to tip distances are interpreted as evidence of restricted gene flow. Range expansions and fragmentations are associated with tip distances that are significantly larger than interior clade distances. Populations were not necessarily synonymous with sampling areas. Rather, each population was defined in two steps, from the nested distance analysis as a group of haplotypes sharing a most recent common ancestor, and from the inference key as a group of haplotypes originating from a common phylogeographic event.

Migration analysis

Once we demonstrated by means of the nested analysis that our population sample of *S. sclerotiorum* was subdivided and determined the boundaries of the subpopulations, we prepared for the coalescent analysis by generating a full migration matrix. An estimate of the

amount of migration between sampling areas is needed to organize the order of coalescent events for haplotypes backwards in time; haplotypes in sampling areas sharing migrants will coalesce before those in areas unlinked by migration. We used MIGRATE, a coalescent program that estimates the effective number of migrants (haplotypes) exchanged between sampling areas each generation over the entire coalescent time for the locus, estimating the number of immigrants and emigrants exchanged among all sampling areas with similar or greater accuracy than F_{ST} methods (Beerli & Felsenstein 1999; Carbone 2000). Note that the sampling areas were used in this analysis, rather than the populations inferred from the nested networks, which would have biased the analysis by grouping the haplotypes according to common ancestry prior to determining patterns of migration among the sampling areas. MIGRATE assumes no recombination and no selection in the ancestral history of haplotypes.

Coalescent analysis

The migration matrix from MIGRATE was then specified as the starting migration matrix for coalescent analyses of a matrix of variable sites distinguishing the haplotypes and corresponding sampling areas of the haplotypes, using GENETREE, version 8.3 (R.C. Griffiths; http://www.maths.monash.edu.au/~mbahlo/mpg/gtree.html) which applies coalescent theory (Kingman 1982a,b,c) to infer the ancestral history of haplotypes. GENETREE assumes an infinitely many-sites model and no recombination. Before initiating the coalescent process, a module of the program, SEQ2TR, checks for violations of the assumptions; if there are violations, it identifies sites that are not compatible with the assumptions. As mentioned above, this was essential in determining the boundaries of recombination blocks in large genomic regions as a means of identifying sources of uncertainty in the phylogenetic network (see also Templeton *et al*. 2000). In our study this would be a crucial prerequisite to reducing uncertainty in the network, for testing clades in the network for significant geographical or host associations, and for inferring the phylogeographic history of clades in the network.

To place migration events in time as historical or contemporary and ongoing we needed to root the network. We inferred the tree with the highest root probability from the coalescent. The estimated migration matrix was specified as the backward migration matrix for ancestral inference using GENETREE. GENETREE provides a maximum-likelihood estimate of the tree with the highest root probability, an estimate of the population mean mutation rate (Ø) and the time to the most recent common ancestor (TMRCA), in coalescent time units of N_e generations for the entire tree, where N_e is the effective population size. Finally, at the population scale in *S. sclerotiorum*, estimates of divergence times (TMRCA) from GENETREE were compared with estimates under a coalescent model with recombination using RECOM58 (Griffiths & Marjoram 1996). Because RECOM58 assumes no selection and no migration (and the infinitely many-sites model) it can be used only after any population subdivision has been identified.

Using the step-by-step sequence of analyses on the population-scale sample, we were able to identify the geographical and host ranges of populations for migration and coalescent analyses on the species scale.

Results

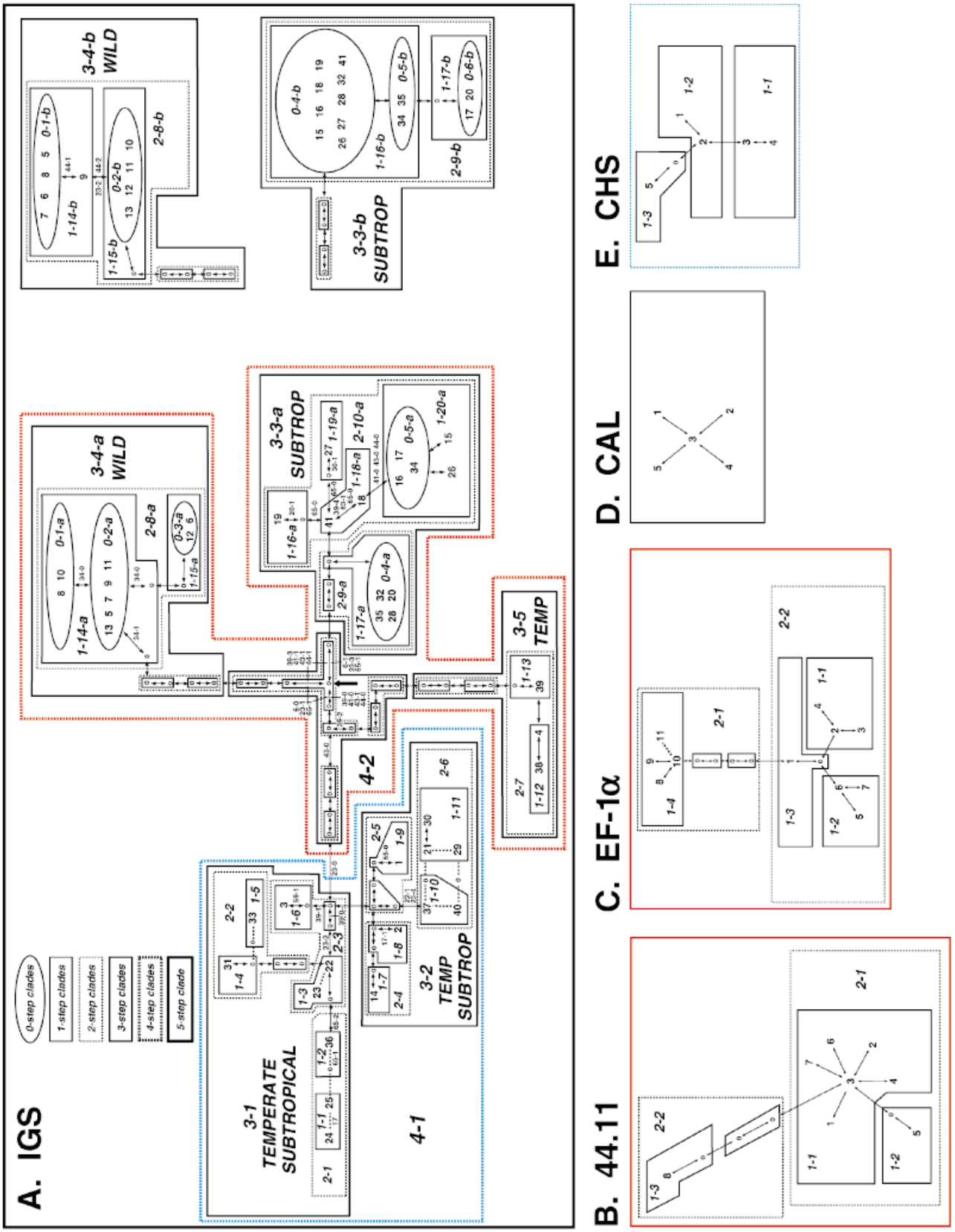
Nucleotide sequence variation on both population and species scales

At the population scale there were 55 parsimony-informative base substitutions in the IGS, eight in EF-1α, one in CHS, one in ACT, and none in CAL, RAS and 44.11. Indels were present only in the IGS, EF-1α and 44.11 regions (see supplemental Table 1, http://www.erin.utoronto.ca/~kohn/). In addition to the 65 variable sites in the IGS with base substitutions, there were 15 sites with short (< 10 bp) or long (up to 200 bp) indels, and several microsatellite and minisatellite motifs. With the indels included in the multiple alignment, all 341 strains were grouped into 47 distinct IGS haplotypes. Because the inclusion of indels introduced homoplasy, precluding even heuristic searches, indels were initially excluded from the alignment and added later after coalescent analysis to the haplotype network (Fig. 1). Indels were excluded from the IGS region for coalescent analyses because of the potential violation of the infinite sites model for microsatellites and minisatellites. At the species scale there were 18 parsimony-informative base substitutions in the IGS promoter region, 94 in EF-1α, 19 in CHS, 32 in ACT, 46 in CAL, 27 in RAS and 1 in ITS-1.

Network estimation and nested analysis on the population scale

With the indels excluded from the IGS region there were 41 distinct haplotypes. Because including the additional six haplotypes associated with indels did not affect the overall nesting topology, and because indels with more than two states at a site violate the infinitely many-sites model, indels were excluded from the analysis. The nested haplotype network for the 41 haplotypes in the IGS region is shown in Fig. 1A. Homoplasy was detected in clades 3-3 and 3-4.

In testing the assumption of no recombination under GENETREE, compatibility matrices resolved two distinct and contiguous recombination blocks in each clade (3-3 and 3-4) spanning 2000 bp. Each block was a segment


A. IGS
0-step clades
1-step clades
2-step clades
3-step clades
4-step clades
5-step clade
3-4-b WILD
3-3-b SUBTROP
3-4-a WILD
3-3-a SUBTROP
3-5 TEMP
3-1 TEMPERATE SUBTROPICAL
4-1
4-2
3-2 TEMP SUBTROP
B. 44.11
C. EF-1α
D. CAL
E. CHS

3-1 SUBTROPICAL TEMPERATE

	2	2	2	2	2	4	4	
	1	2	4	5	7	9	0	0
	6	4	6	9	1	4	5	8
	5	4	7	4	2	0	4	5
165	.	.	.	.	x	x	x	x
2244	.	.	.	.	x	x	.	.
2467	.	.	.	.	x	x	.	.
2594	.	.	.	.	x	x	.	.
2712	x	x	x	x	.	.	x	x
2940	x	x	x	x	.	.	x	x
4054	x	.	.	.	x	x	.	.
4085	x	.	.	.	x	x	.	.

3-2 SUBTROPICAL TEMPERATE

	1	1	1	1	3	
	3	2	6	7	8	7
	2	2	6	6	0	5
	0	7	6	8	2	9
320	.	.	.	x	x	.
1227	.	.	.	.	.	.
1666	.	.	.	.	.	.
1768	x	.	.	.	.	.
1802	x	.	.	.	.	.
3759	.	.	.	.	.	.

3-3 SUBTROPICAL

	3-3-b			3-3-a			
				2	2	2	3
	1	2	5	1	7	9	3
	6	1	1	2	1	4	7
	5	5	9	0	8	2	0	6
165	.	x	.	.	x	x	x	x
215	x	.	.	.	.	.	.	.
519	.	.	.	.	x	x	x	x
520	.	.	.	.	x	x	x	x
2128	x	.	x	x	.	.	.	.
2712	x	.	x	x	.	.	.	.
2940	x	.	x	x	.	.	.	.
3376	x	.	x	x	.	.	.	.

3-4 WILD

	3-4-b		3-4-a			
			1	1	2	2	2	2
	9	0	5	2	6	6	6
	9	2	6	4	3	4	5
	9	5	3	7	7	8	6
999	.	.	.	x	x	x	x
1025	.	.	x	x	x	x	x
1563	.	x	.	x	x	x	x
2247	x	x	x	.	.	.	.
2637	x	x	x	.	.	.	.
2648	x	x	x	.	.	.	.
2656	x	x	x	.	.	.	.

Fig. 2 The population-scale compatibility matrices for phylogenetically informative sites within three-step clades at the IGS locus, generated using SEQ2TR. Compatible sites are indicated with a '.' and incompatible sites with a 'x'. There was no evidence of recombination for haplotypes within clade 3-5 (temperate) or among temperate haplotypes in clade 4-1, and there was no recombination in the mutational connections joining three-step clades (see Fig. 1A). The vertical line partitions separate recombination blocks, designated as -a and -b within clades 3-3 and 3-4. The alternative nested haplotype networks for each partition in clades 3-3 and 3-4 are shown in Fig. 1A.

within which no recombination was detected (Fig. 2). From informative sites spanning both recombination blocks in each clade, > 40 equally parsimonious trees were inferred for clade 3-3 and for clade 3-4, with consistency indices < 0.750 (data not shown). From each recombination block, however, either one most parsimonious tree was inferred with a consistency index of 1.000, or at most, three equally parsimonious trees with consistency indices > 0.900 (data not shown). The alternative nested haplotype networks from each recombination block are 3-3-a and 3-3-b and 3-4-a and 3-4-b (Figs 1A, 2). The topology of the network shown in Fig. 1A, marginal for clades 3-3 and 3-4 (the 3-3-a and 3-4-a alternatives), was identical to the topology of the maximum likelihood tree inferred with the quartet puzzling algorithm (Strimmer & von Haeseler 1996). Branch lengths were consistent with a molecular clock (data not shown). A single unambiguous nested haplotype network was inferred for 44.11 (Fig. 1B), EF-1α (Fig. 1C), CAL (Fig. 1D) and CHS (Fig. 1E); ACT and RAS did not have enough variation for phylogenetic inference in the population-scale analysis.

The nested haplotype networks for each locus were used to perform 1000 random, two-way, contingency permutation analyses to evaluate clades for significant geographical position–haplotype and host–haplotype associations (Table 2). For the IGS locus, all four alternative nested haplotype networks were evaluated for significant geographical and host associations (Table 3). Biological inferences using the key in Templeton *et al*. (1995) based on the nested distance analyses of our population-scale data for those clades with significant geographical associations are summarized in Fig. 3. The predominant pattern from both alternative IGS nested haplotype networks indicated an initial fragmentation event, followed by extensive geographical and host dispersal, with a recurrent pattern of restricted gene flow resulting from isolation by distance. The fragmentation and dispersal events were also detected from the nested distance analysis of the 44.11, EF-1α and CHS loci; restricted gene flow and long-distance colonization were inferred from all nested loci (Fig. 3). Populations (3-1, 3-2, 3-3, 3-4, 3-5) were defined from the IGS nested distance analysis as haplotypes that share a common ancestry, and from the inference key as haplotypes that have arisen from a common event (fragmentation or dispersal).

Fig. 1 (Opposition) The nested unrooted haplotype networks at the population scale for the IGS (A), 44.11 (B), EF-1α (C), CAL (D) and CHS (E) loci. Haplotypes are designated by numbers. The frequency, geographical and host distribution of each haplotype is shown in Table 2. A '0' designates inferred intermediate haplotypes not present in the sample. A double-headed arrow separates haplotypes that differ by one mutational change; dashed lines identify ambiguous mutational connections giving rise to marginal networks which do not alter the nesting topology. A number next to an arrow identifies the site number and encoded state of an indel that is associated with the base-substitution (see supplemental Table 1, http://www.erin.utoronto.ca/~kohn/). The bold arrow in the middle of the network points to the position with the highest root probability, as determined from coalescent analysis (Fig. 4). Indels were superimposed on the network using this root as a reference point. Recombination within clades 3-3 and 3-4 was resolved by dividing the region into two nonrecombining segments (see Fig. 2). The alternative nested haplotype networks (marginal networks) are designated as 3-3-a or 3-3-b and 3-4-a or 3-4-b in (A). The nested haplotype network shown in (A) is marginal for clades 3-3 (3-3-a) and 3-4 (3-4-a). Haplotypes within a closed circle in 3-3 and 3-4 are identical and grouped into 0-step clades. Non-recombinant haplotypes connect the marginal networks for 3-3 (haplotypes 18 and 41) and 3-4 (haplotypes 11 and 13) to the rest of the network. Clades outlined in colour show significant geographical–haplotype associations. Using the inference key by Templeton *et al*. (1995) (Table 3, Fig. 3), a fragmentation event (red) was detected for three-step clades within 4-2 in the IGS (A) and for two-step clades in 44.11 (B) and EF-1α (C). A pattern consistent with extensive dispersal (blue) was detected for three-step clades within 4-1 in the IGS (A) and for one-step clades in CHS (E).

Table 2 Geographic and host distributions of haplotypes at the population scale in *Sclerotinia sclerotiorum*

Locus	Clade	Haplotype (frequency)	Locality (frequency)	Host (frequency)
IGS	0-1-a	8 (6)	SV (6)	RF (6)
		10 (8)	SV (2) VF (6)	RF (8)
	0-2-a	13 (6)	VF (6)	RF (6)
		5 (22)	SV (22)	RF (22)
		7 (7)	SV (7)	RF (7)
		9 (2)	SV (1) VF (1)	RF (2)
		11 (2)	SV (2)	RF (2)
	0-3-a	12 (6)	SV (6)	RF (6)
		6 (5)	SV (5)	RF (5)
	0-4-a	35 (1)	GA (1)	CA (1)
		32 (1)	SC (1)	CA (1)
		28 (16)	GA (4) AL (3) NC (9)	CA (6) CB (9) RD (1)
		20 (1)	LA (1)	GN (1)
	0-5-a	16 (3)	LA (2) NC (1)	CB (3)
		34 (1)	FL (1)	CA (1)
		17 (1)	LA (1)	GN (1)
	1-16-a	19 (11)	LA (4) GA (3) AL (2) NC (2)	CB (6) CA (5)
	1-18-a	18 (19)	LA (3) GA (6) SC (2) AL (1) NC (7)	CB (10) CA (9)
		41 (2)	NC (2)	TB (2)
	1-19-a	27 (2)	AL (1) GA (1)	CA (2)
	1-20-a	26 (4)	GA (3) AL (1)	CA (4)
		15 (16)	LA (2) GA (7) AL (4) NC (3)	CB (5) CA (11)
	1-1	24 (2)	LA (2)	GN (1) GM (1)
		25 (5)	LA (1) GA (3) AL (1)	CB (1) CA (4)
	1-2	36 (6)	FL (1) AB (1) NC (4)	CA (2) CB (1) TB (3)
	1-3	22 (7)	LA (4) NC (3)	CB (4) TB (3)
		23 (1)	LA (1)	CB (1)
	1-4	31 (1)	GA (1)	CA (1)
	1-5	33 (7)	SC (1) GA (5) AL (1)	CA (7)
	1-6	3 (32)	AB (21) NO (1) NZ (10)	CA (22) KW (10)
	1-7	14 (3)	NY (3)	CB (3)
	1-8	2 (24)	AB (5) SK (1) NC (1) NZ (17)	CA (6) TB (1) KW (17)
	1-9	1 (22)	AB (10) SK (1) ON (1) NY (9) ND (1)	CA (12) CB (9) SF (1)
	1-10	37 (15)	NC (15)	CB (15)
		40 (2)	NC (2)	CB (2)
	1-11	21 (12)	LA (1) GA (7) NC (1) SC (2) AL (1)	GN (1) CA (10) TB (1)
		29 (7)	GA (6) AL (1)	CA (7)
		30 (1)	GA (1)	CA (1)
	1-12	4 (50)	SK (2) AB (3) NO (13) NZ (12) NY (19) NC (1)	CA (18) KW (11) CB (19) TB (1) CS (1)
		38 (1)	NO (1)	CA (1)
	1-13	39 (1)	ND (1)	CA (1)
	0-1-b	7 (7)	SV (7)	RF (7)
		6 (5)	SV (5)	RF (5)
		8 (6)	SV (6)	RF (6)
		5 (22)	SV (22)	RF (22)
	0-2-b	13 (6)	VF (6)	RF (6)
		12 (6)	SV (6)	RF (6)
		11 (2)	SV (2)	RF (2)
		10 (8)	SV (2) VF (6)	RF (8)
	0-4-b	15 (16)	LA (2) GA (7) AL (4) NC (3)	CB (5) CA (11)
		16 (3)	LA (2) NC (1)	CB (3)
		41 (2)	NC (2)	TB (2)
		32 (1)	SC (1)	CA (1)
		19 (11)	LA (4) GA (3) AL (2) NC (2)	CB (6) CA (5)
		26 (4)	GA (3) AL (1)	CA (4)
		27 (2)	AL (1) GA (1)	CA (2)
		28 (16)	GA (4) AL (3) NC (9)	CA (6) CB (9) RD (1)
		18 (19)	LA (3) GA (6) SC (2) AL (1) NC (7)	CB (10) CA (9)

Table 2 *Continued*

Locus	Clade	Haplotype (frequency)	Locality (frequency)	Host (frequency)
	0-5-b	34 (1)	FL (1)	CA (1)
		35 (1)	GA (1)	CA (1)
	0-6-b	17 (1)	LA (1)	GN (1)
		20 (1)	LA (1)	GN (1)
	1-14-b	9 (2)	SV (1) VF (1)	RF (2)
44.11	1-1	1 (14)	NO (1) NC (1) ND (1) NY (11)	CA (2) CB (12)
		3 (136)	AB (25) SK (1) LA (9) SC (3) NC (27) GA (24) NY (9) NZ (17) AL (7) NO (14)	CA (74) CB (36) TB (5) KW (17) GN (3) GM (1)
		2 (19)	NZ (19)	KW (18) CS (1)
		6 (28)	ON (1) AB (14) SK (1) ND (1) NY (11)	CA (16) SF (1) CB (11)
		7 (5)	GA (3) SC (1) AL (1)	CA (5)
		4 (67)	NC (22) LA (13) GA (20) FL (2) AL (8) SC (2)	CB (28) CA (31) TB (5) GN (2) RD (1)
	1-2	5 (8)	AB (1) SK (2) GA (1) NZ (3) NC (1)	CA (4) KW (3) TB (1)
	1-3	8 (64)	SV (51) VF (13)	RF (64)
EF-1α	1-1	2 (138)	NY (21) AB (17) SK (2) NC (32) GA (17) SC (2) AL (3) FL (1) ND (1) LA (2) NZ (25) NO (15)	CA (58) CB (49) KW (24) TB (4) GN (2) CS (1)
		3 (2)	GA (2)	CA (2)
		4 (5)	AB (3) NZ (1) NC (1)	CA (3) KW (1) TB (1)
	1-2	5 (2)	GA (1) SC (1)	CA (2)
		6 (45)	NC (3) GA (21) AL (9) SC (1) LA (11)	CB (12) CA (31) GN (2)
		7 (1)	SC (1)	CA (1)
	1-3	1 (84)	ON (1) AB (20) SK (2) NC (15) GA (7) SC (1) AL (4) FL (1) LA (9) ND (1) NZ (13) NY (10)	CA (35) CB (29) SF (1) KW (13) TB (3) GN (1) GM (1) RD (1)
	1-4	8 (5)	VF (5)	RF (5)
		9 (33)	SV (31) VF (2)	RF (33)
		10 (25)	SV (19) VF (6)	RF (25)
		11 (1)	SV (1)	RF (1)
CAL	1-1	1 (100)	ON (1) AB (31) SK (3) NZ (25) ND (1) NO (12) NY (27)	CA (47) KW (24) CS (1) SF (1) CB (27)
		2 (27)	SV (27)	RF (27)
		3 (212)	VF (13) SK (1) NZ (14) GA (47) SC (5) AL (16) FL (2) ND (1) NY (4) NO (3) AB (9) NC (51) SV (24) LA (22)	RF (37) KW (14) CA (83) CB (60) GN (5) GM (1) TB (11) RD (1)
		4 (1)	SC (1)	CA (1)
		5 (1)	GA (1)	CA (1)
CHS	1-1	3 (177)	ON (1) AB (40) SK (4) LA (20) NY (20) NZ (20) NC (34) ND (1) NO (13) AL (6) GA (13) FL (2) SC (3)	CA (82) CB (61) KW (20) TB (9) GN (3) GM (1) SF (1)
		4 (1)	ND (1)	CA (1)
	1-2	1 (64)	VF (13) SV (51)	RF (64)
		2 (98)	NO (1) NC (17) GA (35) AL (10) SC (3) LA (2) NZ (19) NY (11)	CA (48) CB (26) KW (18) GN (2) CS (1) TB (2) RD (1)
	1-3	5 (1)	NO (1)	CA (1)

Coalescent analyses on the population scale

The relative ages of mutations and divergence times of populations were first examined using a coalescent model without recombination, necessitating the exclusion of recombinant haplotypes giving rise to homoplasy. Recombinant haplotypes were distinguished in reference to interior haplotypes that were stable in the network and that were therefore assumed to be ancestral and nonrecombinant (Fig. 1). A pairwise homoplasy matrix, constructed using PAUP* 4.0 identified 22 of the 41 haplotypes as recombinants (see supplemental Table 2, http://www.erin.utoronto.ca/~kohn/). The removal of these 22 recombinant haplotypes did not alter the overall distribution of haplotypes. Although the data set was pruned of recombinant haplotypes, it retained the parental haplotypes with the highest frequencies from each distinct population identified in the nesting analysis.

A migration matrix indicating the number of migrants exchanged between sampling areas was estimated using MIGRATE (Beerli & Felsenstein 1999; results shown in Table 4). All sampling areas were considered (not populations determined in the nested analysis, see Materials and Methods). To increase sample size and confidence

Table 3 Nested contingency analysis of haplotype-geography and haplotype–host associations

Locus	Clade	Geography Permutation X^2 statistic	Probability	Clade	Host Permutation X^2 statistic	Probability
IGS	1-1	3.73	0.572	1-1	7.00	0.141
	1-3	0.69	1.000	1-3	0.69	1.000
	1-11	3.54	0.839	1-12	1.72	1.000
	1-12	2.70	0.627	1-16-b	1.85	0.531
	1-14-a	3.45	0.070	1-18-a	29.00	0.003*
	1-14-b	20.49	0.052	1-20-a	8.89	0.065
	1-16-b	38.53	0.022*	2-1	5.62	0.201
	1-18-a	2.95	0.767	2-3	40.00	0.000*
	1-20-a	14.55	0.052	2-4	27.00	0.000*
	2-1	13.00	0.002*	2-6	37.00	0.000*
	2-2	0.38	1.000	2-7	1.63	1.000
	2-3	40.00	0.000*	2-9-b	78.00	0.000*
	2-4	27.00	0.001*	2-10-a	8.89	0.319
	2-6	33.20	0.000*	3-1	19.84	0.016*
	2-7	52.00	0.023*	3-2	51.95	0.000*
	2-8-a	3.39	0.103	3-3-a	5.03	0.263
	2-8-b	24.27	0.000*	4-1	15.45	0.005*
	2-9-b	10.26	0.155	4-2	225.97	0.000*
	2-10-a	15.55	0.362	Total	75.97	0.000*
	3-1	62.96	0.000*			
	3-2	123.62	0.000*			
	3-3-a	4.92	0.467			
	4-1	30.16	0.002*			
	4-2	382.03	0.000*			
	Total	113.02	0.000*			
44.11	1-1	398.35	0.000*	1-1	185.80	0.000*
	2-1	37.98	0.032*	2-1	7.77	0.227
	Total	341.00	0.000*	Total	341.00	0.000*
EF-1α	1-1	23.72	0.300	1-1	9.02	0.295
	1-2	23.36	0.017*	1-2	1.32	0.743
	1-4	24.22	0.000*	2-2	26.97	0.012*
	2-2	123.24	0.000*	Total	341.00	0.000*
	Total	341.00	0.000*			
CAL	Total	446.27	0.000*	Total	174.33	0.024*
CHS	1-1	88.50	0.027*	1-1	4.52	0.337
	1-2	162.00	0.000*	1-2	162.00	0.000*
	Total	196.20	0.002*	Total	98.64	0.005*

*Significant at $P = 0.05$.

estimates, sampling areas with only one or two sampled haplotypes were amalgamated into larger areas, for a total of nine sampling areas (Table 4). The criterion for amalgamating was that at least one isolate representing each area share a DNA fingerprint (see methods in Kohn *et al*. 1991) with one isolate from each candidate area for amalgamation. Our population-scale data set had a large number of migration parameters to be estimated (total of 81). This necessitated increasing the number of sampled genealogies (both short-sample and long-sample; see MIGRATE help manual (http://evolution.genetics.washington.edu/lamarc/migratedoc/migratedoc.html) by a factor of 20 and repeating the analysis 10 times using different random number seeds. All runs yielded similar results. A very low level of migration was detected between haplotypes sampled from *Ranunculus ficaria* in Norway and haplotypes from other sampling areas (Table 4).

To root the IGS network, the estimated migration matrix was specified as the backward migration matrix for ancestral inference with GENETREE. The probabilities for each of the 60 possible rooted IGS trees were evaluated using GENETREE (Fig. 4). The tree with the highest root probability showing the geographical distribution of mutations in the tree was determined by first performing 100 000 simulations of the coalescent with 10 different starting random number seeds. From these runs, the tree with the highest root probability was selected and the accuracy of this tree was further checked by repeating the analysis with 1 000 000 coalescent simulations using two different random number seeds. The coalescent time scale shows the TMRCA

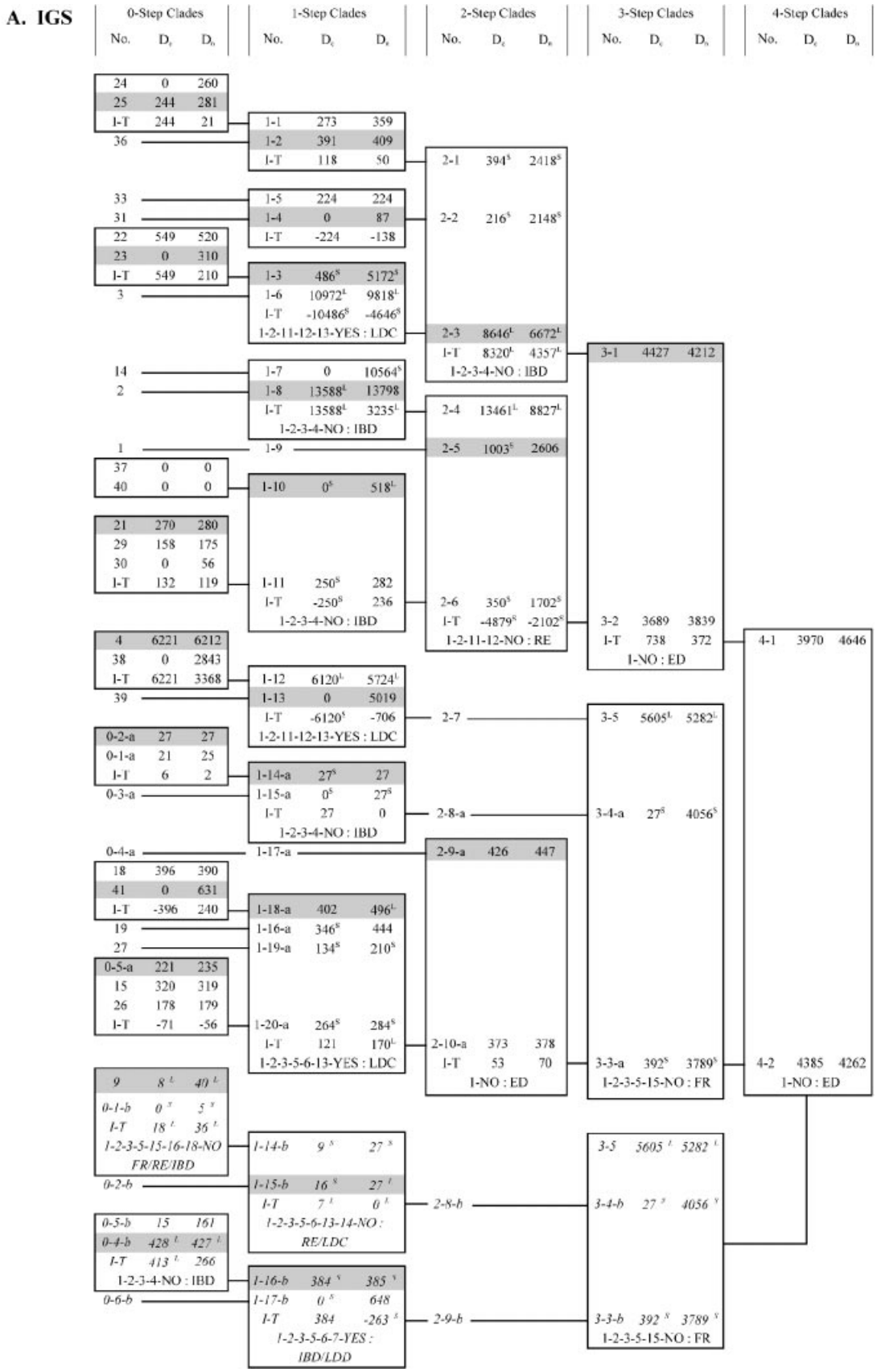


Fig. 3 *Continued*

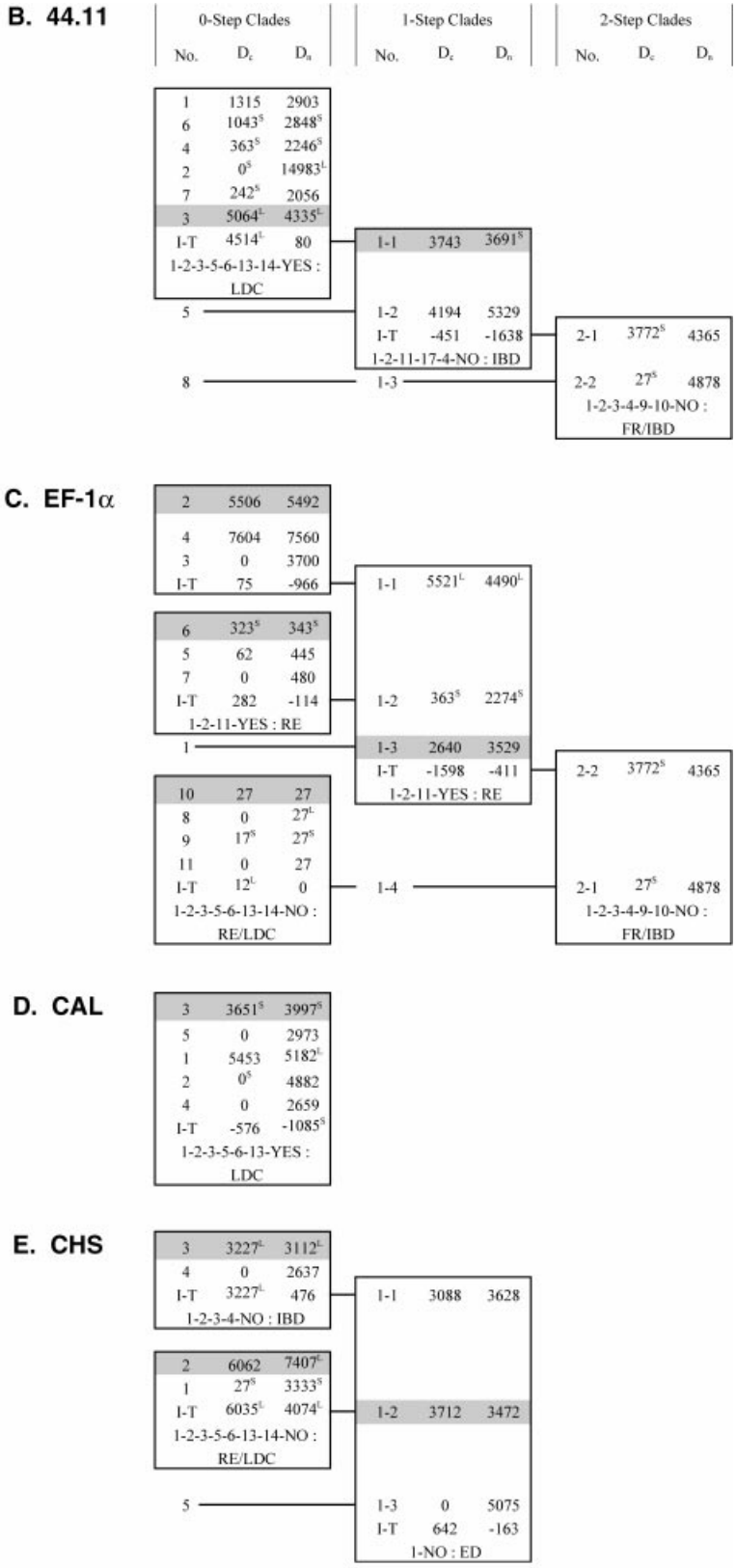


Fig. 3 Results of the nested geographical distance analysis at the population scale of *Sclerotinia sclerotiorum* for haplotypes in the IGS (A), 44.11 (B), EF-1α (C), CAL (D) and CHS (E) loci. The boxes indicate the clades found at each nesting level in the nested hierarchical structure, shown in Fig. 1. The boxes with numbers in italics in (A) indicate the alternative IGS nested clades, shown in Fig. 1A. The geographical distance analysis was performed separately for the alternative nested clades. Each clade is indicated by the clade number followed by the within (D_c) and between (D_n) nested clade distances. D_c is the average of the individual distances of each sampled isolate bearing a haplotype from the geographical centre of the nested clade. D_n is the average of the individual distances in the nested clade from the geographical centre of the nesting clade, the next highest clade level (Templeton *et al*. 1995). For clades containing both interior and tip clades, the interior clades are shaded and the average difference in distances between interior and tip clades (I–T) is given. Differences in geographical distances (D_c, D_n or I–T) that were significant at the 5% level are indicated with a superscript: either an 'S' for smaller or an 'L' for larger. The last line in each box is the biological inference for a particular set of nested clades and applies to all clades nested in the box. The inference was obtained using the results of the nested geographical distance analysis and the inference key given in Templeton *et al*. (1995). The numbers refer to the sequence of questions in the key and the answer to the final question is indicated as either YES or NO. The inference is given after the colon, where ED is extensive dispersal (inferred at higher clade levels), LDC is long-distance colonization, LDD is long-distance dispersal, IBD is restricted gene flow with isolation by distance, RE is contiguous range expansion and FR is fragmentation. When there are two or more inferences per box, the key could not distinguish among them.

Table 4 Migration estimates for haplotypes from the nine sampling areas using MIGRATE. Estimates are based on Markov chain Monte Carlo (MCMC) approximation methods

Sampling area	$2N_em$ estimates								
	1,x	2,x	3,x	4,x	5,x	6,x	7,x	8,x	9,x
1: CAN_CA	—	0.0	0.0	0.3	2.8	0.0	0.0	0.0	0.0
2: LA_CB	0.0	—	1.0	0.0	0.0	0.0	0.0	0.0	0.0
3: NC_CB	0.0	0.0	—	1.2	0.1	0.0	0.0	0.0	0.0
4: NC_TB	0.0	3.6	0.0	—	0.0	0.0	0.0	2.0	0.6
5: NO_CA	0.0	0.0	0.0	0.0	—	0.0	0.9	1.4	0.2
6: NO_RF	0.0	0.0	0.0	0.1	0.0	—	0.0	0.0	0.0
7: NY_CB	0.4	0.0	0.0	0.0	0.7	0.0	—	0.0	0.0
8: NZ_KW	0.4	0.0	0.0	0.0	0.8	0.0	0.0	—	0.0
9: SE_CA	0.0	0.4	0.0	0.3	0.0	0.0	0.0	0.0	—

—, Undefined value. The migration parameter, $2N_em$, is defined as the effective number of migrants exchanged between two sampling areas each generation, where N_e is the effective population size and m is the migration rate per generation. The following is an example of how to read the migration parameter. For sampling area 6 (NO_RF), the 4,× means that immigration from sampling area 4 (NC_TB) into sampling area 6 is $2N_em = 0.1$ (a very low level of migration).

and relative divergence time of the five distinct populations in the sample (Fig. 4). These populations coincide with the five distinct clades, 3-1, 3-2, 3-3, 3-4 and 3-5, identified in the nested analysis. The distribution and frequency of haplotypes in the nine sampling areas is shown in the grid below the tree. Haplotypes from North Carolina tobacco are found in populations 3-1, 3-2, 3-3 and 3-5, and contain mutations that may have arisen in the ancestral population (Table 4). For coalescent analysis under a model with recombination, the population mean mutation rate, Ø, and the recombination parameter, *r*, were estimated from haplotypes within each population identified in the nested analysis with the aid of RECOM58 (Griffiths & Marjoram 1996) (Table 5). The TMRCA of each population was estimated using RECOM58 and compared with estimates assuming no recombination (Table 5).

Phylogenetic and coalescent analyses at species scale

Our objective was to compare divergence times between populations and species. In the species tree for each locus, species were on long branches but populations were not fully resolved. By combining loci, we could increase resolution at the population level, but first we had to determine whether the loci were combinable. For the IGS promoter region, ITS-1, RAS and ACT loci, the parsimony and maximum likelihood tree topologies for each locus were identical, indicating an approximately equal rate of nucleotide substitution among these closely related species. Topological concordance was evaluated simultaneously, and in all pairwise combinations among these loci, using the partition homogeneity test (PHT: Farris *et al*. 1995; Huelsenbeck *et al*. 1996) implemented in PAUP* 4.0. Significance was assessed by performing 1000 random repartitions of phylogenetically informative sites. All four loci were concordant ($P > 0.05$). Three of these four loci were physically unlinked; the IGS/ITS-1, CHS/EF-1α, ACT/44.11/CAL and RAS regions hybridized to different chromosome-sized DNAs in Southern hybridizations with CHEF chromosome separations (data not shown). A combined analysis of base substitutions and encoded indels from all four loci yielded 91 variable sites and 14 distinct haplotypes for the 385 strains at the species scale (see supplemental Table 3, http://www.erin.utoronto.ca/~kohn/). With the indels excluded there were 86 variable sites and 13 distinct haplotypes yielding one most

Table 5 Maximum-likelihood estimates of population parameters at the IGS locus using GENETREE and RECOM58

	GENETREE				RECOM58				
Popn	No. haplotypes	No. individuals	Theta* (SD)	TMRCA* (SD)	No. haplotypes	No. individuals	Theta*	r*	TMRCA* (SD)
3-1	3	45	1.37 (0.66)	1.25 (5.43e-04)	8	61	2.05	0.75	3.10 (0.73)
3-2	5	73	2.06 (0.82)	0.703 (5.02e-04)	8	86	2.30	0.40	3.38 (0.89)
3-3	2	21	0.28 (0.28)	0.285 (1.14e-02)	13	78	2.90	0.40	1.66 (0.48)
3-4	6	49	1.79 (0.78)	0.244 (4.05e-04)	9	64	1.4	0.65	2.14 (0.57)
3-5	3	51	0.44 (0.33)	1.000 (2.04e-03)	3	52	1.0	0.05	1.06 (0.44)

*TMRCA is the mean time to the most recent common ancestor conditional on the data and theta and r are simulated maximum likelihood estimates.

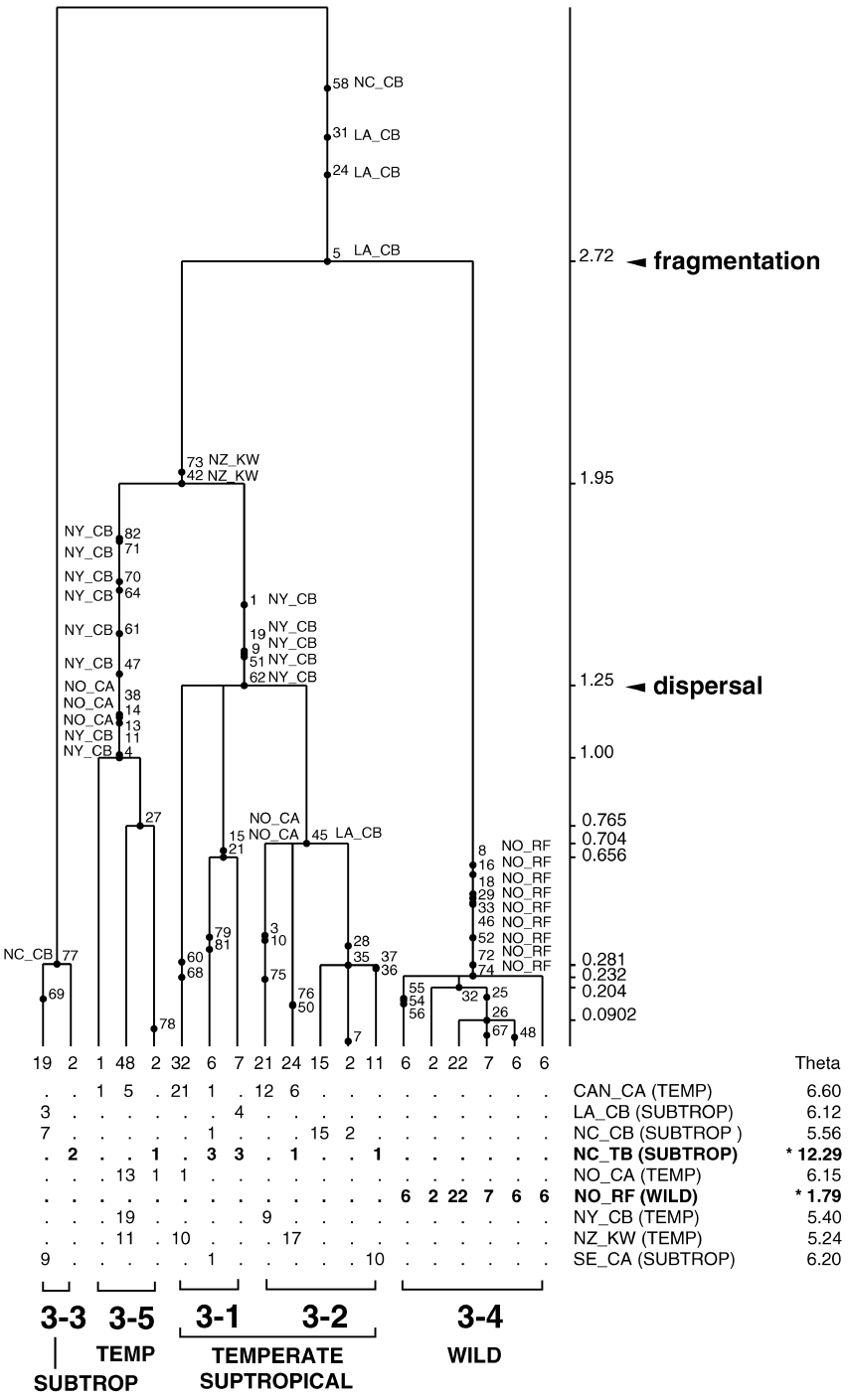


Fig. 4 The IGS genealogy with the highest root probability based on the migration matrix (Table 4) and 1 000 000 coalescent simulations with Watterson's estimate of Ø = 9.9 (Watterson 1975). The time scale on the right shows the estimated time to the most recent common ancestor (TMRCA), in coalescent time units of N_e generations, where N_e is the effective population size. Base substitutions are designated with • and identified by site number. The tree shows the ancestral distribution of mutations and coalescent events (fragmentation, dispersal) in the population history of the 19 nonrecombinant haplotypes in the sample. Each haplotype is represented at the tips of the tree by its frequency in the total sample. The grid below the tree shows the distribution and frequency of haplotypes in the nine sampling areas. The asterisks indicate sampling areas with significantly different Ø estimates. These differences suggest that the magnitude of Ø depends on the number of populations that occupy a particular sampling area. For example, haplotypes in NC_TB belong to four different populations (3-5, 3-1, 3-2 and 3-3) and have an elevated Ø, whereas haplotypes in NO_RF belong to a single population (3-4) and have a small Ø. All other sampling areas with comparable Ø estimates have haplotypes that span only two populations.

parsimonious tree with a consistency index of 0.9875 (data not shown). The topology of this tree was identical to the maximum likelihood tree. A pairwise homoplasy matrix for all haplotypes showed that *Sclerotinia trifoliorum* and *S. minor* were responsible for the small amount of homoplasy in the combined data set. A compatibility matrix identified five sites violating the infinitely many-sites model and four incompatible sites (see supplemental Table 3, http://www.erin.utoronto.ca/~kohn/). A migration matrix was estimated using MIGRATE for three broad sampling areas defined as temperate–agricultural, subtropical–agricultural and Norwegian–wild (Table 6). Similar results were also obtained by increasing the length of chains by a factor of 20 and repeating the analysis using 10 different random number seeds. These sampling areas were defined as the geographical ranges of the five populations

Table 6 Migration estimates for haplotypes at species scale

	$2N_em$ estimates		
Sampling areas	1,x	2,x	3,x
1: TEMPERATE	—	2.0	1.1
2: SUBTROPICAL	0.5	—	0.4
3: WILD	1.0	1.3	—

—, Undefined value.

determined in the population-scale analysis. Coalescent analysis under a model assuming no recombination and no selection was used to estimate the relative order of population divergence, and this was compared with the relative order of population-splitting events. The probabilities for each of the 77 possible rooted trees from the combined data set were evaluated using the migration matrix (Table 6) and GENETREE (Fig. 5). The tree with the highest root probability was determined using 1 000 000 coalescent simulations with 10 different starting random number seeds.

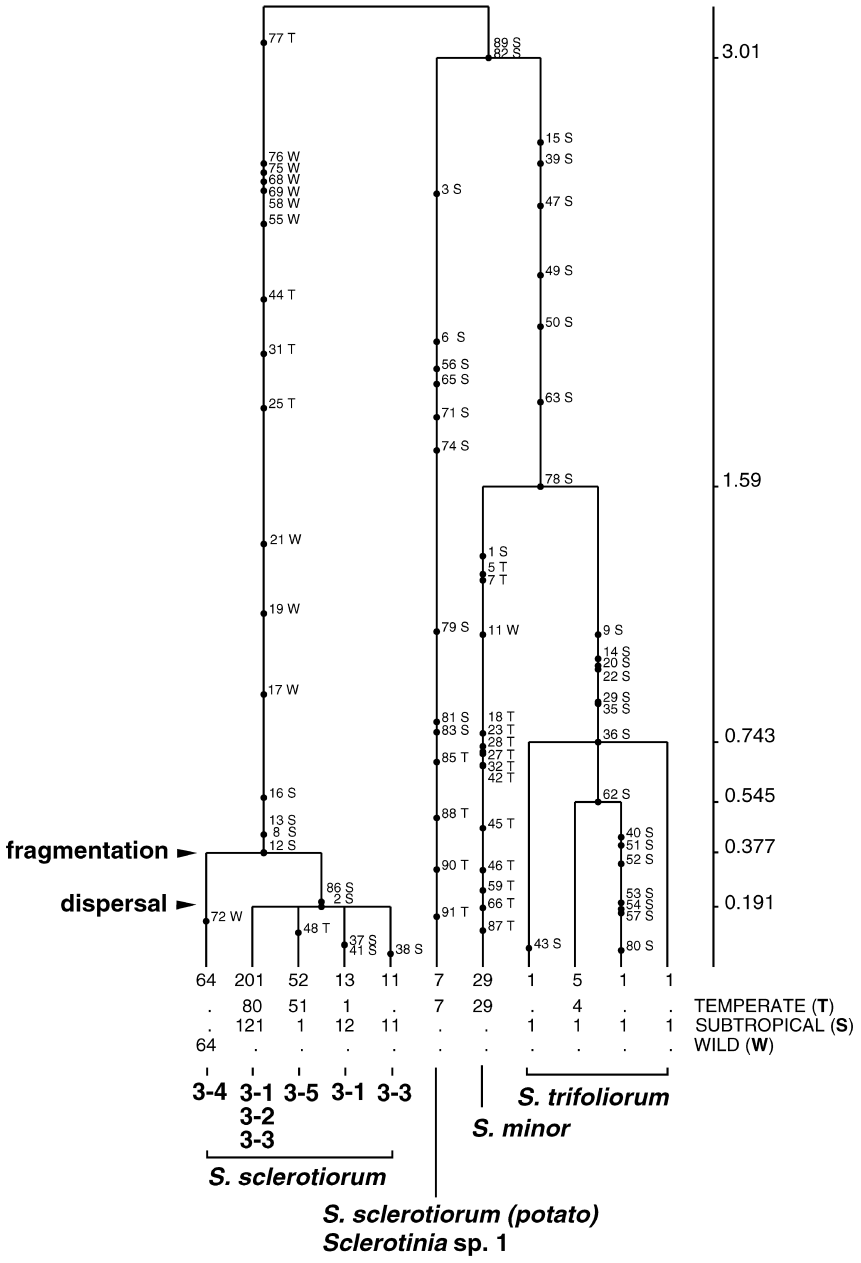


Fig. 5 The genealogy with the highest root probability from the combined species-scale analysis of the IGS promoter region, ITS-1, RAS and ACT loci. Sites that were incompatible or violated the infinitely many-sites model were excluded from the analysis. This tree was based on the migration matrix at the species scale (Table 6) and 1000 000 coalescent simulations with Watterson's estimate of Ø = 11.8 (Watterson 1975). Mutations are indicated with • and identified by site number. The tree shows the ancestral distribution of mutations and coalescent events (fragmentation, dispersal) in the population-species history of the 11 distinct haplotypes in the sample. Each haplotype is represented by its frequency in the total sample. The grid below the tree shows the distribution and frequency of haplotypes in the three sampling areas.

Discussion

Using this stepwise sequence of analyses we inferred the geographical origins of mutations in the sample. Populations coalesced before species, as would be expected. The boundary between divergent populations and species exists in this difference in coalescence time (assuming neutrality and a molecular clock); this is the population–species interface in this group of species. From the convergence of the population and species scales, we could determine the relative order of population splitting and speciation events, as well as patterns of gene flow maintaining cohesion. We recommend the compatibility matrix implemented in GENETREE as an early step in phylogenetic inference from large, noncoding loci; in detecting recombination blocks in the IGS, we found a route to reduce phylogenetic uncertainty in the nested statistical analysis of this locus. Also recommended is the full migration matrix generated with MIGRATE, which reduces the uncertainties in migration parameters and provides more information than F_{ST} methods for estimating the TMRCA of populations (Carbone 2000). Templeton and co-workers' cladistic inference method and the coalescent approach detected the same phylogeographic processes.

Both scales of analysis were needed for full resolution of both patterns of descent and phylogeographic origins of populations and species. Coalescent analyses at both scales showed populations 3-1, 3-2, 3-3, 3-4 and 3-5 as distinct evolutionary lineages, but only 3-4 (Norwegian wild buttercups) and 3-5 (temperate agriculture) were fully differentiated at the species scale. The multilocus analysis at the species scale did not completely differentiate the more recently evolved populations (3-1, 3-2) because rates of evolution among these loci are too slow to resolve events on this time scale. The two key phylogeographic events were first fragmentation, then dispersal. The population scale of coalescent analysis traces mutations back through dispersal to the population fragmentation event but not earlier (Fig. 4). The species scale traces back earlier, before the fragmentation event, but does not resolve the later dispersal events among populations (Fig. 5). For example, 3-3 is fully resolved as a clade in the population scale sampling (Fig. 4), but is only partially resolved as a clade distinct from 3-1 and 3-2 at the species scale (Fig. 5).

Our approach established a relative ranking of lineages at both scales of analysis. For example, based on both the nested and coalescent analyses at the population scale, clade 3-4 included only samples from Norwegian wild buttercups. Under the biological and phylogenetic species concepts, 3-4 would be a species. The nature of the historical divergence event, geographical fragmentation or host shift, could not be distinguished. The species-scale gene tree indicated a recent split of 3-4 from the other *Sclerotinia sclerotiorum* clades, relative to the split of 3-4 from its most recent common ancestor with *Sclerotinia* sp. 1, making it unlikely that 3-4 (or 3-3 or 3-5) is a group at species rank (Avise & Johns 1999). At this scale, there was evidence of migration between temperate climate, wild plants (*Sclerotinia* sp. 1) and agricultural crops (potato), but not between temperate wild plants and plants in warmer climates, suggesting divergence of 3-4 at the same time as the subtropical agricultural clade (3-3) in fragmentation events in North America and Northern Europe. Although the conversion of coalescent TMRCAs to real time produced huge variances and was based, of necessity, on speculative estimates of generation time and mutation rates, migration from Pleistocene refugia is an explanation of the fragmentation consistent with our data. An alternative, more recent fragmentation with human migration and associated domestication of plants, is more difficult to reconcile with the degree of sequence divergence that we observed.

A reproductively isolated lineage could be a clone, a population or a species. In some apparently asexual fungi, a multiple gene genealogy with long branches separating geographically isolated groups of haplotypes, in which recombination is detected within, but not between, groups has been interpreted as a signature of speciation (Koufopanou *et al*. 1997; Geiser *et al*. 1998), but these groups could be divergent populations. In these fungi, recombination events have been attributed either to unobserved sexual reproduction or parasexuality (Burt *et al*. 1996), but whether the signature of recombination is that of a sexual ancestor or a sign of contemporary recombination cannot be determined.

In our study, there was more recombination in some populations than in others, and it appears to have been infrequent. Recombination was evident in the subtropical agricultural (3-3 and subtropical haplotypes of 3-1 and 3-2) and temperate Norwegian wild (3-4) populations, but not in the temperate agricultural populations (3-5 and temperate haplotypes in 3-1 and 3-2; see also Carbone *et al*. 1999). This may be explained by an overall low recombination rate with a climate-dependent number of meiotic generations (ascospore-to-ascospore) per season, for example, one generation per season in temperate areas and two or more generations in subtropical locales. There would also be a longer growing season, with more mitotic generations, in a subtropical climate. Despite the large IGS recombination blocks in two clades, 3-3 on subtropical crops and 3-4 on Norwegian wild buttercups, both clades were still resolved after removing recombinant haplotypes, an indication that recombination has occurred as past or sporadic events, but not routinely and recently. If recombination were frequent there would be no blocks (Awadalla *et al*. 1999). This interpretation is further supported by the extensive clonality detected as groups of independently sampled strains sharing a DNA fingerprint in each of the five populations (see also Carbone *et al*. 1999).

Both significant and non–significant associations between haplotype and geography or host (Table 3) are informative.

Non-significant associations in two-step clades nested in 3-3-a suggest that evidence of range expansion and long-distance colonization may have been lost because the events occurred long ago. Consequently, extensive dispersal is the predominant pattern now detected in 3-3-a. Coalescent analysis (Fig. 4) indicates the progenitors of 3-3 as the root of all contemporary North American *S. sclerotiorum* populations. In comparison, populations 3-1 and 3-2, which are more recently evolved (Fig. 4), have significant associations of haplotypes with geography and host (Table 3).

The significant associations in 4-2 suggest that haplotypes in clades 3-3 (subtropical crops), 3-4 (temperate wild) and 3-5 (temperate crops) may be adapted to specific ecological conditions. These adaptations may be associated with phenotypes such as growth temperature range or sclerotial vernalization requirement, or more specific adaptations to host microclimate characteristics — all experimentally testable. Whether haplotypes in 3-4 are host specific (*Ranunculus*) or ecologically adapted in other ways cannot be distinguished. If adaptations in 4-2 are as old as the clades (Fig. 4) with which they may be associated, they would have constrained the spread of haplotypes for a long time. In contrast, populations 3-1 and 3-2 are associated with a mix of crops from both temperate and subtropical areas. These recently evolved populations (Fig. 4) have come to occupy a range of conditions. The apparent specialization of 3-3, 3-4 and 3-5 suggests that these populations may be en route to sympatric speciation. However, the relatively short time of divergence of these populations compared with divergence times among closely related species (Fig. 5) suggests that they still have a long way to go.

The long coalescent branches separating *S. sclerotiorum* from *S. minor* and *S. trifoliorum* are consistent with a long history of reproductive isolation. Avise & Wollenberg (1997) noted that reproductive barriers may result in deep biotic discontinuities distinguishable from recent evolutionary processes, such as fragmentations and bottlenecks. Although our process of inferring the TMRCA offers the opportunity, this more ancient population–species interface may still be difficult to capture in sampling of contemporary populations. What is captured in this study, based on coalescent estimates from rooted trees, is the continuum, from haplotypes in populations (3-1, 3-2, 3-3, 3-4, 3-5), to divergence in populations (4-1 to 3-1 and 3-2), to speciation (sp. 1 and potato from a common ancestor of the *S. sclerotiorum* populations). The best definition of a species may well be indistinguishable from that of a population — with the difference contingent on cessation of gene flow and relative time of divergence.

Acknowledgements

For supplying sets of isolates or assisting us in sampling, we thank T. Schumacher (University of Oslo) and A. Holst-Jensen (Norwegian National Veterinary Inst.), H.A. Magnus (Norwegian Crop Res. Inst., Aas), A. Hov, S. Hoyte (New Zealand HortResearch), D. Phillips (University of Georgia), M. Cubeta (North Carolina State University), H. Dillard (Cornell University), and G. Holcomb (Louisiana State University). R. Griffiths (Oxford University), J. Felsenstein, M. Kuhner, J. Yamato and P. Beerli (University of Washington) and D. Posada (Brigham Young University) were all instrumental in implementing their programs.

References

Adams PB, Ayers WA (1979) Ecology of *Sclerotinia* species. *Phytopathology*, **69**, 896–899.

Avise JC, Johns GC (1999) Proposal for a standardized temporal scheme of biological classification for extant species. *Proceedings of the National Academy of Sciences of the USA*, **96**, 7358–7363.

Avise JC, Wollenberg K (1997) Phylogenetics and the origin of species. *Proceedings of the National Academy of Sciences of the USA*, **94**, 7748–7755.

Awadalla P, Eyre WA, Maynard Smith J (1999) Linkage disequilibrium and recombination in hominid mitochondrial DNA. *Science*, **286**, 2524–2525.

Beerli P, Felsenstein J (1999) Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, **152**, 763–773.

Boland GJ, Hall R (1994) Index of plant hosts of *Sclerotinia sclerotiorum*. *Canadian Journal of Plant Pathology*, **16**, 93–108.

Burt A, Carter DA, Koenig GL, White TJ, Taylor JW (1996) Molecular markers reveal cryptic sex in the human pathogen *Coccidioides immitis*. *Proceedings of the National Academy of Sciences of the USA*, **93**, 770–773.

Carbone I (2000) Population history and process: nested clade and coalescent analysis of multiple gene genealogies in a parasite of agricultural and wild plants. PhD Thesis. University of Toronto, Canada.

Carbone I, Anderson JB, Kohn LM (1999) Patterns of descent in clonal lineages and their multilocus fingerprints are resolved with combined gene genealogies. *Evolution*, **53**, 11–21.

Carbone I, Kohn LM (1993) Ribosomal DNA sequence divergence within internal transcribed spacer 1 of the Sclerotiniaceae. *Mycologia*, **85**, 415–427.

Carbone I, Kohn LM (1999) A method for designing primer sets for speciation studies in filamentous ascomycetes. *Mycologia*, **91**, 553–556.

Crandall KA (1996) Multiple interspecies transmissions of human and simian T-cell leukemia/lymphoma virus type I sequences. *Molecular Biology and Evolution*, **13**, 115–131.

Dieckmann U, Doebeli M (1999) On the origin of species by sympatric speciation. *Nature (London)*, **400**, 354–357.

Donoghue MJ (1985) A critique of the biological species concept and recommendations for a phylogenetic alternative. *Bryologist*, **88**, 172–181.

Farris JS, Kallersjö M, Kluge AG, Bult C (1995) Testing significance of incongruencies. *Cladistics*, **10**, 315–319.

Geiser DM, Pitt JI, Taylor JW (1998) Cryptic speciation and recombination in the aflatoxin-producing fungus *Aspergillus flavus*. *Proceedings of the National Academy of Sciences of the USA*, **95**, 388–393.

Gómez-Zurita J, Petitpierre E, Juan C (2000) Nested cladistic analysis, phylogeography and speciation in the *Timarcha goettingensis*

complex (Coleoptera, Chrysomelidae). *Molecular Ecology*, **9**, 557–570.

Griffiths RC, Marjoram P (1996) Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, **3**, 479–502.

Hasegawa M, Kishino H, Yano TA (1985) Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22**, 160–174.

Holst-Jensen A, Vaage M, Schumacher T (1998) An approximation to the phylogeny of *Sclerotinia* and related genera. *Nordic Journal of Botany*, **18**, 705–719.

Huelsenbeck JP, Bull JJ, Cunningham CW (1996) Combining data in phylogenetic analysis. *Trends in Ecology and Evolution*, **11**, 152–158.

Kingman JFC (1982a) On the genealogy of large populations. *Journal of Applied Probability*, **19**, 27–43.

Kingman JFC (1982b) Exchangeability and the evolution of large populations. In: *Exchangeability in Probability and Statistics* (eds Koch G, Spizzichino F), pp. 97–112. North-Holland, Amsterdam.

Kingman JFC (1982c) The coalescent. *Stochastic Processes and Their Applications*, **13**, 235–248.

Kohli Y, Kohn LM (1998) Random association among alleles in clonal populations of *Sclerotinia sclerotiorum*. *Fungal Genetics and Biology*, **23**, 139–149.

Kohn LM, Stasovski E, Carbone I, Royer J, Anderson JB (1991) Mycelial incompatibility and molecular markers identify genetic variability in field populations of *Sclerotinia sclerotiorum*. *Phytopathology*, **81**, 480–485.

Kondrashov AS, Kondrashov FA (1999) Interactions among quantitative traits in the course of sympatric speciation. *Nature (London)*, **400**, 351–354.

Koufopanou V, Burt A, Taylor JW (1997) Concordance of gene genealogies reveals reproductive isolation in the pathogenic fungus *Coccidioides immitis*. *Proceedings of the National Academy of Sciences of the USA*, **94**, 5478–5482.

Melzer MS, Smith EA, Boland GJ (1997) Index of plant hosts of *Sclerotinia minor*. *Canadian Journal of Plant Pathology*, **19**, 272–280.

Mishler BD (1985) The morphological, developmental, and phylogenetic basis of species concepts in bryophytes. *Bryologist*, **88**, 207–214.

Mishler BD, Brandon RN (1987) Individuality, pluralism, and the phylogenetic species concept. *Biology and Philosophy*, **2**, 397–414.

Mishler BD, Budd AF (1990) Species and evolution in clonal organisms — introduction. *Systematic Botany*, **15**, 79–85.

Mishler BD, Donoghue MJ (1982) Species concepts: a case for pluralism. *Systematic Zoology*, **31**, 491–503.

O'Donnell K, Cigelnik E, Nirenberg HI (1998) Molecular systematics and phylogeography of the *Gibberella fujikuroi* species complex. *Mycologia*, **90**, 465–493.

Patterson CL (1986) The comparative biology, epidemiology, and control of lettuce drop caused by *Sclerotinia minor* and *S. sclerotiorum* and the genetic analysis of vegetative and sexual compatibility in *S. minor*. PhD Thesis. University of California, USA.

Posada D, Crandall KA, Templeton AR (2000) GEODIS: a program for the cladistic nested analysis of the geographical distribution of genetic haplotypes. *Molecular Ecology*, **9**, 487–488.

Strimmer K, von Haeseler A (1996) Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution*, **13**, 964–969.

Swofford DL (1998) PAUP*. *Phylogenetic Analysis Using Parsimony (*And Other Methods), Version 4.0*. Sinauer Associates, Sunderland, MA.

Templeton AR (1989) The meaning of species and speciation: a genetic perspective. In: *Speciation and its Consequences* (eds Otte D, Endler JA), pp. 3–27. Sinauer, Sunderland, MA.

Templeton AR (1994) The role of molecular genetics in speciation studies. In: *Molecular Ecology and Evolution: Approaches and Applications* (eds Schierwater B, Streit B, Wagner GP, DeSalle R), pp. 455–477. Birkhäuser Verlag, Basel, Switzerland.

Templeton AR, Boerwinkle E, Sing CF (1987) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*. *Genetics*, **117**, 343–351.

Templeton AR, Clark AG, Weiss KM, Nickerson DA, Boerwinkle E, Sing CF (2000) Recombinational and mutational hotspots within the human lipoprotein lipase gene. *American Journal of Human Genetics*, **66**, 69–83.

Templeton AR, Crandall KA, Sing CF (1992) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics*, **132**, 619–633.

Templeton AR, Routman E, Phillips CA (1995) Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. *Genetics*, **140**, 767–782.

Templeton AR, Sing CF (1993) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. *Genetics*, **134**, 659–669.

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.

Uhm JY, Fujii H (1983a) Ascospore dimorphism in *Sclerotinia trifoliorum* and cultural characters of strains from different-sized spores. *Phytopathology*, **73**, 565–569.

Uhm JY, Fujii H (1983b) Heterothallism and mating type mutation in *Sclerotinia trifoliorum*. *Phytopathology*, **73**, 569–572.

Watterson GA (1975) On the number of segregating sites in genetic models without recombination. *Theoretical Population Biology*, **7**, 256–276.

This work was part of the PhD research of Ignazio Carbone. The goal was to understand the history of population divergence and speciation in a cosmopolitan group of plant parasitic fungi, the Sclerotiniaceae by bringing a sequence of phylogenetic and coalescent approaches to bear on a large multilocus data set. This work is part of on-going research on the systematics of the Sclerotiniaceae in the Kohn laboratory.
