Molecular markers reveal cryptic sex in the human pathogen *Coccidioides immitis*

(population genetics/clonality/recombination/single-strand conformation polymorphism/sequencing with arbitrary primer pairs)

AUSTIN BURT*†, DEIDRE A. CARTER‡§, GINA L. KOENIG‡, THOMAS J. WHITE‡, AND JOHN W. TAYLOR*

*Department of Plant Biology, University of California, Berkeley, CA 94720; and ‡Roche Molecular Systems, 1145 Atlantic Avenue, Alameda, CA 94501

ABSTRACT   *Coccidioides immitis*, cause of a recent epidemic of "Valley fever" in California, is typical of many eukaryotic microbes in that mating and meiosis have yet to be reported, but it is not clear whether sex is truly absent or just cryptic. To find out, we have undertaken a population genetic study using PCR amplification, screening for single-strand conformation polymorphisms, and direct DNA sequencing to find molecular markers with nucleotide-level resolution. Both population genetic and phylogenetic analyses indicate that *C. immitis* is almost completely recombining. To our knowledge, this study is the first to find molecular evidence for recombination in a fungus for which no sexual stage has yet been described. These results motivate a directed search for mating and meiosis and illustrate the utility of single-strand conformation polymorphism and sequencing with arbitrary primer pairs in molecular population genetics.

Unlike most plants and animals, the vast majority of eukaryotic microorganisms can reproduce asexually, and this together with their small size can make it very difficult to determine the relative importance of sexual reproduction in nature by direct observation alone. Rather, molecular markers must be used to test for the clonal population structure expected if sex is absent and the recombinant genotypes expected if sex is present (1–3). It is important in such studies that marker identity reflect common descent and that identities due to convergences, parallelisms, and reversals be minimized; in this respect, the most informative markers are DNA sequences (4, 5). However, most studies of human pathogens have used allozymes or random amplified polymorphic DNAs (RAPDs) (1–3, 6). Neither approach is completely satisfactory, as allozyme patterns can be misleading because of natural selection (7), and RAPD patterns can be difficult both to repeat and to interpret in terms of Mendelian loci (8–10). Recently, we have developed an approach for finding DNA sequence-level polymorphisms useful as molecular markers in population genetics (called SWAPP, for sequencing with arbitrary primer pairs; ref. 11). In this paper we use this method to test for evidence of sex in the human pathogen *Coccidioides immitis* (Ascomycetes: Onygenales).

C. immitis is a dimorphic pathogen endemic to semi-arid regions of the New World which can live either as a hyphal saprobe in the soil or as a unicellular systemic pathogen of mammals. Human infections are thought to be acquired by inhalation of spores produced by the saprophytic stage; direct human-to-human transmission does not occur (12). Recently there has been an epidemic of Valley fever (coccidioidomycosis) in California, with case reports ≈10 times more frequent than normal (13, 14). Like some 20% of fungi (15), no sexual stage has ever been reported in *C. immitis* (16). Unfortunately, the inability to cross strains has hindered basic and applied research on this species.

MATERIALS AND METHODS

For our study we have analyzed 30 clinical isolates from 25 patients at a single hospital in Tucson, Arizona. Three patients contributed multiple samples, collected up to 3 weeks, 16 months, and 8 years apart, respectively (Table 1).¶ All isolates were collected in 1979–1990, before the epidemic.

Our strategy for finding molecular markers begins with low-stringency PCR amplification from genomic DNA using arbitrary primers (≈20-mers) in various pairwise combinations (11). Genomic DNA was isolated following heat treatment to kill the fungus, as described (22). PCRs were as described (11) except that the primer concentrations were 0.2 μM; they were subjected to three cycles of 94/35/72°C for 5/5/5 min, 30 cycles of 94/50/72°C for 1/1/1 min, and 7 min at 72°C. Amplicons were sought which (*a*) disappeared if either primer was used alone (indicating that it had a different primer at each end), (*b*) were consistently amplified from a subsample of 6 different isolates, and (*c*) showed polymorphism among the 6 isolates by single-strand conformation polymorphism (SSCP; ref. 23) analysis. For SSCP, 1 μl of the PCR mixture was reamplified in 25 μl of the same cocktail except that the dNTPs were diluted 300-fold and 0.15 μl of [α-[^{35}S]thio]dATP (12.5 mCi/ml; 1 mCi = 37 MBq) was added. The reamplification went for 8 cycles of 94/50/72°C for 1/1/1 min and then 7 min at 72°C. The reactions were run on an MDE gel (AT Biochem, Malvern, PA) according to the manufacturer's instructions except that 3 μl of labeled product was mixed with 3 μl of stop solution. Loci that were polymorphic among the 6 tester isolates were then amplified from all 30 isolates and scored by SSCP (Fig. 1).

Three representatives of each allele were then sequenced to determine the nucleotide basis for each SSCP; for alleles found in only one or two isolates, all were sequenced. Amplicons were gel-purified, reamplified, and sequenced with the Applied Biosystems 373 automated sequencer according to the manufacturer's instructions, using labeled terminators and the same primers as were used in the initial amplification.

RESULTS AND DISCUSSION

We tried 21 primers in 127 arbitrary pairs and found 17 polymorphic loci, of which 12 were phylogenetically informative and were characterized further. These were supplemented by two informative loci found by screening known regions of the genome, the internal transcribed spacer region of the rRNA gene (*af*) (24), and a fragment of the chitin synthase gene (*cs*) (25) (Table 1). Of the 14 informative loci, 8 had two SSCP alleles (*a1, a2, e2, z, aa, b1, af, cs*) and 6 had more than

Table 1. Multilocus genotypes for 25 clinical isolates of *C. immitis*

Locus	Isolate abcdefghijklmnopqrstuvwxy	Primers*	Size*, bp	Polymorphism(s)†	Enzyme(s)‡
a1	0000111000010100011100101	NS1, ITS5	282	144 C/T	*Bsr* I, *Mae* I
a2	0000001000010000000000000	NS1, ITS5	144	40 T/C	
e2	0000000000100010000000000	NS23, ITS5	184	108 TT/C—	
k	0010100001000111010010001	NS2, NS22	≈590	189 G/A	
q2	0000001000000000010000000	CNS19, ITS4	165	20 C/T	*Alu* I
z	0111101110111101111100011	NS24, CNS19	≈260	13 A/T	*Hinf*I, *Ple* I
aa	0111111101111101111100000	NS24, CNS25	207	11 A/G	*Hae* III, *Nae* I
am	0000100000010100000000000	NS23, ML5.1	315	66 G/A; 97 C/T	*Ava* II, *Mbo* I
bg	0101000000000001110000000	ITS2, ML3.5	161	34 CTC/—	*Dde* I
bk	0000000000000000100100000	ITS4, ML5.1	369	146 C/A	*Mbo* I, *Cla* I
bl	0010000010000000000111000	NS26, MS1	≈830	233 T/C; 733 A/—	*Dde* I, *Msc* I
bq	0011000000000100100001000	ML5.1, MS1	≈430	170 A_7/$A_{8,9}$; 235 A/G 260 T/C; 293 A/G	*Hinf*I, *Nsi* I, *Fok* I
af	0000100010000101110101001	ITS4, ITS1	630	46 T/A	
cs	0000000011000000010101011	CS2, CS4A	260	53 G/A	*Mbo* II

The patient contributing isolate a contributed two more isolates, 4 and 8 years later; the patient contributing isolate j contributed two more isolates, 2 and 16 months later; and the patient contributing isolate s contributed one more isolate, 3 weeks later. In no case was the genotype of the subsequent isolates different from the originals. Primer sequences are in refs. 17–19 except for NS23, 5′-GACTCAACACGGGAAACTC-3′; CNS19, 5′-GTTTCTCAGGCTCCCTCTCCGG-3′; CNS25, 5′-ATGTATTAGCTCTAGAATTACCAC-3′; ML3.5, 5′-GAGAATGCTGACAT-GAGTAAC-3′; ML5.1, 5′-ACCGTAACTTCGGGATAAGG-3′; CS2, 5′-AACTGAAAGGCAACCAGGTCC-3′; and CS4A, 5′-CATAACCT-TGATCTCTCCAC-3′. Since the primers used to amplify anonymous loci were originally designed to amplify rRNA genes (which we did not want), only primers designed for the same strand or for nuclear and mitochondrial genes were used together. Sequences for the 0 allele are in GenBank (accession nos. X94130–X94143). BLASTX searches (20, 21) with the 12 randomly amplified loci found only one suggestive homology (score > 70), between locus *am* and myosin genes of various organisms (e.g., 26% identical and 53% similar over 71 amino acids to the myosin heavy chain of chicken gizzard smooth muscle). If this open reading frame is functional, the G/A polymorphism translates into Arg/Lys and the C/T polymorphism is silent. The chitin synthase (*cs*) polymorphism is also silent.
*Includes the primers.
†Position in GenBank sequence; nucleotide shown for 0 allele/1 allele.
‡Selected restriction enzymes predicted to differentially cut the two alleles.

two (*k, q2, am, bg, bk, bq*). All 6 loci with more than two SSCP alleles were polymorphic at more than one nucleotide position—in no case were there more than two nucleotides at a site. Of these 6, 4 had only one informative position, and it is this position that is represented in the table and was included in the analyses. The remaining two loci (*k, bg*) each had two informative sites, and only the most informative (closest to 50:50 in frequency) was used, for ease of interpretation. For three loci, nucleotides at two or more variable sites were completely correlated in the isolates sequenced, and so any or all could have been responsible for the SSCP (*am, bl, bq*). There was no evidence of recombination between information positions within any locus. The overall sequence divergence among isolates was low (<1%).
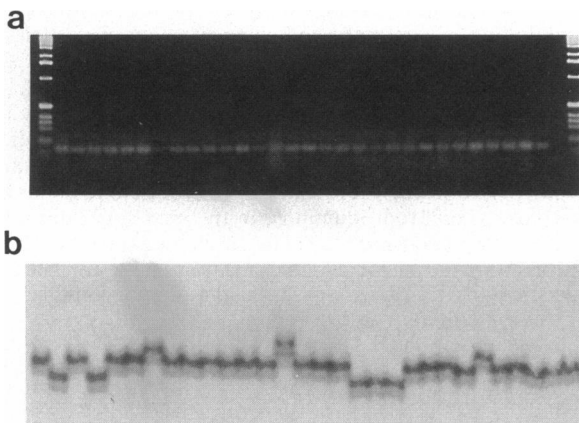

FIG. 1. Amplicons produced consistently from all 30 isolates and which appear monomorphic on an agarose gel (*a*) may show variation on an SSCP gel (*b*). Pictured is locus *bg* (Table 1). We never observed more than one allele per isolate, suggesting that, like most ascomycetes, this fungus is haploid.

While isolates from the same patient had identical multilocus genotypes, all isolates from different patients were distinct. This in itself suggests a recombining population, but it is important to examine the structure of the variation, not just the total amount. For example, if reproduction is clonal, then for most pairs of biallelic loci no more than three of the four possible genotypes should be observed. In fact, for all possible pairs of the six most informative loci (allele frequencies between 25% and 75%), all four genotypes are observed. If reproduction is clonal then there should be strong correlations between alleles at different loci (i.e., linkage disequilibrium) (1–3, 26). In fact, there are no such correlations detectable in our dataset (Fig. 2*a*). Phylogenetic analyses can also be used, for if reproduction is clonal then the ancestry of any set of isolates can be represented as a tree, and one should be able to make a good fit of a tree to the data set, whereas if reproduction is sexual, then the tree will be poorly resolved and rife with homoplasy. In fact, the fit is very poor. The strict consensus of 62 most parsimonious trees is almost completely unresolved, with only one internal branch (Fig. 3). Furthermore, these trees are much longer than the minimum possible (38 vs. 14 steps using only the informative sites; 52 vs. 28 steps using all sites), requiring a high level of homoplasy (reversals, parallelisms, and convergences), which is incompatible with the low level of sequence divergence observed and clonal reproduction. However, this is exactly what one would expect of a recombining population. Interestingly, trees fitted to randomized data sets (28) are slightly longer still (Fig. 2*b*). Thus some deviation from complete panmixia is apparent, perhaps because we have sampled from heterogeneous populations.

To further test the idea of recombination, partial sequence information was obtained for a subsample of six isolates (a, b, c, r, t, v) and seven loci (*a1, k, aa, af, bl, bq, cs*) chosen so that all were phylogenetically informative for the six isolates. Loci were 157–311 bp in length (total, 1504 bp) and within each
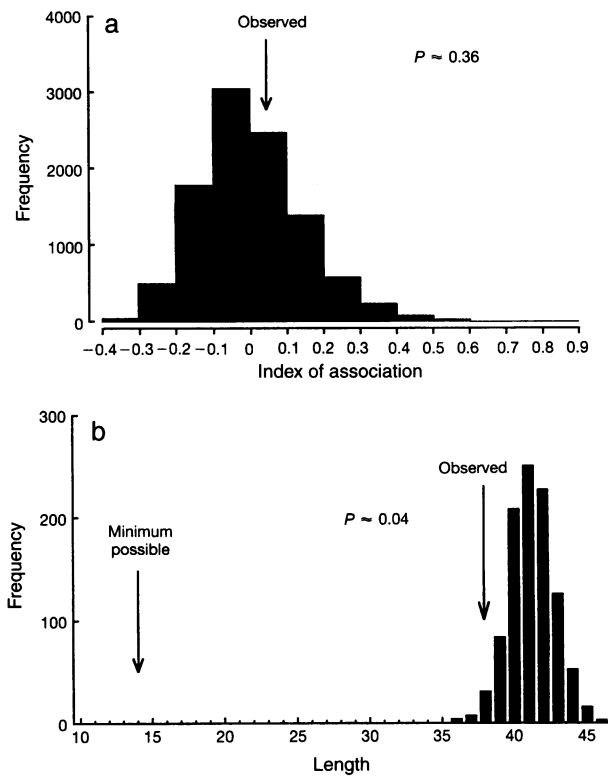
FIG. 2. Randomization tests comparing the observed dataset (arrows) with datasets in which alleles have been randomly shuffled across isolates for each locus separately (histograms). (*a*) The observed index of association (I_A) across loci (3, 26) falls well within the distribution for randomized datasets, indicating no significant correlation of alleles across loci ($P \approx 0.36$). (*b*) The length (L) of the shortest tree fitted by parsimony analysis to the observed dataset is much greater than the number of characters but is at the low end of the distribution of tree lengths fitted to randomized data sets (HENNIG86 heuristic searches with commands mh; bb*; $P \approx 0.041$). Histogram in *a* is based on 10,000 randomizations, and that in *b* on 1000. The correlation between I_A and L across randomizations is statistically significant, but weak ($r = -0.20$, $n = 1000$, $P < 0.001$). A computer program for *a* and for generating an input file for *b* is available on request from A.B.

there was one informative site (or three that were perfectly correlated), and zero to two singletons (i.e., autapomorphies); all other sites were invariant. The overall nucleotide diversity was 0.39% (29). Likelihoods for the data were calculated under two models—a null hypothesis that all loci have the same evolutionary topology (as expected under clonality) and an alternative in which each locus is allowed a different topology (as expected with recombination)—and the latter gives a much better fit to the data (Table 2). To calculate a *P* value, note that 1492 of the 1504 sites are invariant and that at least 4 of the 12 polymorphic sites must have multiple substitutions (i.e., are homoplastic) under model 1. If the number of changes is Poisson-distributed across sites, then the mean number of changes per site can be estimated as $\mu = -\ln(1492/1504) = 0.008$; the probability of there being two or more changes at a
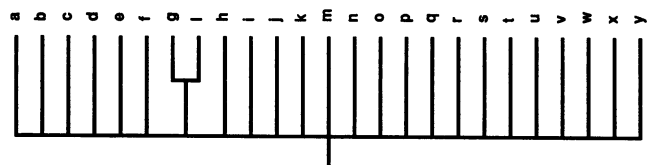

FIG. 3. Strict consensus of 62 most parsimonious trees, calculated by using PAUP (27).

Table 2. Likelihood analysis of partial sequence data for six isolates and seven loci

Model	Maximum ln likelihood	Difference	SD*
1. Same topology	-2178.7	34.0	12.28
2. Different topologies	-2144.7		

For six isolates there are 105 possible unrooted trees, and the ln likelihood of each tree was calculated separately for each locus by using the program DNAML (30), the observed nucleotide frequencies, and a transition/transversion ratio of 2.0. Under model 1, these values are summed across loci, and then the tree with the maximum value is determined; under model 2, the maximum value is determined separately for each locus, and then these are summed across loci. The data are under the model allowing different topologies than under the model allowing only one topology, $e^{34} = 6.0 \times 10^{14}$ times more likely a substantial increase in goodness-of-fit (31).

*Standard deviation of the difference between models, calculated across sites (32).

site is $(1 - e^{-\mu} - \mu e^{-\mu}) = 0.00003$, and the probability of there being two or more changes given that there is at least one change is $(1 - e^{-\mu} - \mu e^{-\mu})/(1 - e^{-\mu}) = 0.004$. The probability of observing 4 or more such sites out of the 12 polymorphisms is $P \approx 0.00000012$. Thus, there is more homoplasy than expected under clonality.

It is not clear what process could account for these results other than recombination. It might be argued that there are two types of nucleotide sites in the genome of *C. immitis*: sites that change so slowly that they are monomorphic in our dataset (>99% of sites) and sites that change so rapidly that they are saturated for changes in our dataset, and so are uninformative on patterns of descent (<1% of sites). However, there is no known molecular basis for such a dichotomy, and several lines of evidence further suggest that it is unlikely: (*i*) The polymorphic sites have monomorphic sites on both sides, and so the putative instability would have to be very precisely localized. (*ii*) For none of the polymorphic sites did we find more than two different nucleotides, so the sites would have to be alternating between just two nucleotides, not all four. (*iii*) Isolates collected from the same patient up to 8 years apart did not differ at any of the sites. (*iv*) There was no evidence of recombination within any of the loci with two or more informative sites. (*v*) Most of the loci have been scored for two other populations, and allele frequencies differ significantly between populations, including some fixed differences (unpublished data); it is very difficult to see how these sites could be changing so rapidly as to be uninformative within a population but still be informative between populations. Recombination seems the only plausible explanation.

To see what these results imply about the frequency of recombination, consider the following simple model of clonal diversity under the opposing forces of recombination and random drift (see ref. 33, pp. 319–327, for a similar model of genic diversity under mutation and random drift). Let *f* be the probability in a particular generation that two isolates taken at random are clonemates (i.e., asexual descendants of the same outcrossed ascospore, where matings between clone mates are counted as asexual reproduction). In the next generation, the probability that two random isolates are clonemates will be *f′* = [(probability that they have the same parent) × (probability that neither isolate is sexually derived)] + [(probability that they have different parents) × (probability that those two parents are clonemates) × (probability that neither is sexually derived)], or

$$f' = \frac{1}{N}(1 - s)^2 + \left(1 - \frac{1}{N}\right)f(1 - s)^2,$$

where *N* is the population size and *s* is the proportion of individuals that are sexually derived (by outcrossing). At equilibrium,

$$f' = f = \frac{(1-s)^2}{1 + 2s(N-1) - s^2(N-1)},$$

or, approximately,

$$f \approx \frac{1}{1 + 2sN} \quad \text{and} \quad sN \approx \frac{1-f}{2f}.$$

(For $N > 1000$ and $0 \leq s \leq 1$, the maximum possible difference between exact and approximate expressions for f is 0.0008, which is much smaller than the measurement error.) In our sample, no two isolates from different patients were the same, even when only shared polymorphisms were considered, and the most likely interpretation is that each isolate is a different clone, and $f \approx 0$. If we say $f < 0.05$, then $sN > 9.5$. That is, regardless of population size, only a few individuals need be sexually derived each generation in order to generate the observed diversity. While the model is simplistic in not, for example, considering loss-of-sex mutations, the conclusion remains that frequencies of sex too low to be directly observed might nonetheless give a recombining population structure.

CONCLUSIONS

We conclude that biparental sex is a regular part of the *C. immitis* life cycle, at least in our study population. This study is, to our knowledge, the first to find molecular evidence for recombination in a fungus for which no sexual stage has yet been described. These results motivate a directed search for mating and meiosis, as the ability to cross isolates would greatly facilitate genetic studies of this human pathogen. The closest known relative of *C. immitis, Uncinocarpus reesii* (25, 34), is a soil saprobe which mates by hyphal fusion and produces ascospores in loosely woven fruiting bodies (35), and so the saprophytic stage would be the natural place to start looking for sex in *C. immitis*. Alternatively, nonmeiotic, "parasexual" recombination could also account for the data; if so, then this would be the first demonstration of its importance in a natural population of any fungus. Our results also provide a baseline with which to compare isolates from the recent *C. immitis* epidemic. Ten of the 14 polymorphic loci are predicted to be differentially cut by restriction enzymes (Table 1), which should facilitate scoring of additional isolates and populations.

More generally, we conclude that phylogenetic analyses are ideally suited for addressing the issue of sex vs. clonality and, as such, are a very useful complement to the more commonly used population genetic analyses based on linkage disequilibrium and Hardy–Weinberg equilibrium (diploids only). In addition, the molecular methods used to find and characterize genetic markers with nucleotide-level resolution (11) should be widely applicable in studies of molecular ecology, epidemiology, and conservation genetics (7).

1. Tibayrenc, M., Kjellberg, F. & Ayala, F. J. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 2414–2418.
2. Tibayrenc, M., Kjellberg, F., Arnaud, J., Oury, B., Brenière, S. F., Dardé, M.-L. & Ayala, F. J. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 5129–5133.
3. Maynard Smith, J., Smith, N. H., O'Rourke, M. & Spratt, B. G. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 4384–4388.
4. DuBose, R. F., Dykhuizen, D. E. & Hartl, D. L. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 7036–7040.
5. Dykhuizen, D. E. & Green, L. (1991) *J. Bacteriol.* **173**, 7257–7268.
6. Tibayrenc, M., Neubauer, K., Barnabe, C., Guerrini, F., Skarecky, D. & Ayala, F. J. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 1335–1339.
7. Avise, J. C. (1994) *Molecular Markers: Natural History and Evolution* (Chapman & Hall, New York).
8. Riedy, M. F., Hamilton, W. J. I. & Aquadro, C. F. (1992) *Nucleic Acids Res.* **20**, 918.
9. Ellsworth, D. L., Rittenhouse, K. D. & Honeycutt, R. L. (1993) *BioTechniques* **14**, 214–216.
10. Clark, A. G. & Lanigan, C. M. (1993) *Mol. Biol. Evol.* **10**, 1096–1111.
11. Burt, A., Carter, D. A., White, T. J. & Taylor, J. W. (1994) *Mol. Ecol.* **3**, 523–525.
12. Pappagianis, D. (1988) *Curr. Top. Med. Mycol.* **2**, 199–238.
13. Pappagianis, D., Sun, R. K., Werner, S. B., Rutherford, G. W. I., Elsea, R. W., Miller, G. B. J., Bootwala, N. & Hopkins, R. S. (1993) *Morbid. Mortal. Week. Rep.* **42**, 21–24.
14. Pappagianis, D. (1994) *Clin. Inf. Dis.* **19**, Suppl. 1, S14–S18.
15. Reynolds, D. R. (1993) in *The Fungal Holomorph: Mitotic, Meiotic and Pleomorphic Speciation in Fungal Systematics*, eds. Reynolds, D. R. & Taylor, J. W. (CAB International, Wallingford, CT), pp. 15–25.
16. Rippon, J. W. (1988) *Medical Mycology* (Saunders, Philadelphia).
17. White, T. J., Bruns, T., Lee, S. & Taylor, J. (1990) in *PCR Protocols: A Guide to Methods and Applications*, eds. Innis, M. A., Gelfand, D. H., Sninsky, J. J. & White, T. J. (Academic, San Diego), pp. 315–322.
18. Gargas, A. & Taylor, J. W. (1992) *Mycologia* **84**, 589–592.
19. Swann, E. C. & Taylor, J. W. (1993) *Mycologia* **85**, 923–936.
20. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.
21. Gish, W. & States, D. J. (1993) *Nat. Genet.* **3**, 266–272.
22. Burt, A., Carter, D. A., Carter, G. L., White, T. J. & Taylor, J. W. (1995) *Fungal Genet. Newsl.* **42**, 23.
23. Orita, M., Iwahana, H., Kanazawa, H., Hayashi, K. & Sekiya, T. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 2766–2770.
24. Berbee, M. L., Yoshimura, A., Sugiyama, J. & Taylor, J. W. (1995) *Mycologia* **87**, 210–222.
25. Pan, S., Sigler, L. & Cole, G. T. (1994) *Microbiology* **140**, 1481–1494.
26. Brown, A. H. D., Feldman, M. W. & Nevo, E. (1980) *Genetics* **96**, 523–536.
27. Swofford, D. L. (1993) PAUP: Phylogenetic Analysis Using Parsimony (Ill. Nat. Hist. Surv., Champaign), Version 3.1.
28. Archie, J. W. (1989) *Syst. Zool.* **38**, 239–252.
29. Nei, M. (1987) *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York).
30. Felsenstein, J. (1990) PHYLIP Manual (University Herbarium, Univ. of California, Berkeley), Version 3.3.
31. Edwards, A. W. F. (1992) *Likelihood* (Johns Hopkins Univ. Press, Baltimore).
32. Kishino, H. & Hasegawa, M. (1989) *J. Mol. Evol.* **29**, 170–179.
33. Crow, J. F. & Kimura, M. (1970) *An Introduction to Population Genetics Theory* (Harper & Row, New York).
34. Bowman, B. H. & Taylor, J. W. (1993) in *The Fungal Holomorph: Mitotic, Meiotic and Pleomorphic Speciation in Fungal Systematics*, eds., Reynolds, D. R. & Taylor, J. W. (CAB International, Wallingford, CT), pp. 169–178.
35. Sigler, L. & Carmichael, J. W. (1976) *Mycotaxon* **4**, 349–488.