

# Disease surveillance in recombining pathogens: Multilocus genotypes identify sources of human *Coccidioides* infections

Matthew C. Fisher\*<sup>†</sup>, Bruce Rannala<sup>‡</sup>, Vishnu Chaturvedi<sup>§</sup>, and John W. Taylor<sup>¶</sup>

\*The Institute of Zoology, Regent's Park, London NW1 4RY, United Kingdom; <sup>†</sup>Department of Medical Genetics, University of Alberta, Edmonton, AB, Canada T6G 2H7; <sup>§</sup>Mycology Laboratory, Wadsworth Center, 120 New Scotland Avenue, Albany, NY 12201-2002; and <sup>¶</sup>Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720

Edited by John C. Avise, University of Georgia, Athens, GA, and approved May 1, 2002 (received for review March 27, 2002)

**Molecular surveillance of pathogenic microbes works by genotyping isolates with DNA fingerprinting techniques and then using these genotypes to assign individuals to populations. Clonality is assumed in many fingerprinting studies, although this assumption has been shown to be false for many organisms. To accommodate recombining organisms into surveillance programs, methods using population allele frequencies in combination with individual multilocus genotypes are necessary. Here, we develop a statistical method appropriate for haploid recombining microbes that allows individuals to be assigned to populations. We illustrate the usefulness of this technique by inferring the source populations for *Coccidioides* isolates recovered from patients treated outside the endemic area of *Coccidioides* sp., the etiological agents of human coccidioidomycosis, but with a travel history including visits to one or more endemic areas.**

Globalization has increased the rate at which organisms are being introduced into previously nonendemic areas. This traffic has spawned an increase in the number of emerging infectious diseases of animals and plants as a result of nonnative introductions (1). Central to developing policy aimed at controlling such immigration events is the implementation of surveillance systems that enable epidemiologists to identify sources of infection. An example of this type of surveillance is the nationally networked U.S. PulseNet database, which uses DNA fingerprints for detecting outbreaks of bacterial foodborne disease ([www.cdc.gov/ncidod/dbmd/pulsenet/pulsenet.htm](http://www.cdc.gov/ncidod/dbmd/pulsenet/pulsenet.htm)). Although the need for such surveillance systems exists for other infectious organisms, differences in their basic biology require alternative theoretical approaches. This fact is true especially for eukaryotic microbes where, in contrast to many bacteria, meiotic recombination is a regular and obligatory feature of most life cycles. These microbes typically differ from higher eukaryotes, however, in having a haploid, rather than a diploid, genome; recently developed methods for assigning diploid individuals to populations (2) must therefore be modified for use with recombining pathogens.

Where population genetic structure is principally clonal, single DNA sequences of sufficient variability are all that is needed to identify source populations (or strains) because of the high degree of linkage among regions of the genome. However, in recombining organisms, each locus has a separate evolutionary history. Therefore, a number of independent loci are required to estimate the relatedness between isolates (3). In this article, we develop statistical methods for determining the source populations of haploid recombining microbes from multilocus data sets, by extending the method developed by Rannala and Mountain (2). We demonstrate the utility of this test by considering the worldwide population genetic structure of the closely related fungal pathogens *Coccidioides immitis* and *Coccidioides posadasii* (4). *Coccidioides* sp. are saprobic soil fungi adapted to growth in the highly arid Lower Sonoran Life Zone. When vertebrates inhale arthroconidia, the fungus undergoes a dimorphic transi-

tion to a yeast-like phase that can cause a severe disseminated disease, coccidioidomycosis (5). The incidence of human coccidioidomycosis in endemic areas increased during the 1990s (6), and a concomitant rise in nonendemic areas seems to have occurred (7). A recent study by Chaturvedi *et al.* (7) of mycology data in New York State showed 181 hospitalizations with coccidioidomycosis. A substudy of 16 of these cases from which patient isolates were available found that all had a history of travel to *Coccidioides* endemic areas, and multilocus genotyping demonstrated that all isolates were *C. posadasii* and most likely from Arizona.

Here, we formulate a Bayesian assignment test (BAT), using Bayesian estimation of allele frequencies that assigns unknown isolates of *Coccidioides* sp. to their population of origin. The method assumes that loci are statistically independent (i.e., in linkage equilibrium) at the intrapopulation level and that all potential source populations have been sampled. The performance of the method, and the effect of violating these assumptions, is assessed. We do this by simulating populations and varying a critical set of population genetic parameters, then evaluate the strength of assignment of random samples of known origin. We subsequently field-test our method by reanalyzing the 16 clinical isolates from patients in New York State and demonstrate a great increase in sensitivity for detecting their source population. These methods provide an important advance in the epidemiological tools necessary for molecular surveillance in this, and other, haploid recombining organisms.

## Data and Analyses

**Theory and Methodology.** The statistical method used to assign individuals of uncertain origin to populations is similar to that of Rannala and Mountain (2), except that the approach is modified to accommodate haploid, rather than diploid, genotypes. Another change is that the Bayesian posterior probability that an individual derives from any one of  $I$  potential source populations is used to assign the individual to a source population rather than basing the decision on a series of likelihood ratio tests. Assuming that individual  $m$  derives from a particular population, the probability of the individual's multilocus genotype,  $Y_m = \{Y_{jm}\}$ , where  $Y_{jm}$  is the allele present at the  $j$ th locus in individual  $m$ , is used to assign the individual to a source population conditional on a preexisting sample of genotypes from all potential source populations and assuming random mating within populations.

Define  $x = \{x_{hji}\}$  to be a matrix of the allele frequencies in each of the sampled populations, where  $x_{hji}$  is the frequency of allele  $h$  at locus  $j$  in population  $i$ . The probability of allele  $h$  at locus  $j$  is

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: BAT, Bayesian assignment test;  $I_A$ , standardized index of association.

<sup>†</sup>To whom reprint requests should be addressed. E-mail: matthew.fisher@ioz.ac.uk.

$$\Pr(Y_{jm} = h | \mathbf{x}_{ji}) = x_{hji}. \quad [1]$$

The allele frequencies are, in reality, unknown and therefore we instead consider

$$\Pr(Y_{jm} | \mathbf{X}_{ji}) = \int \Pr(Y_{jm} | \mathbf{x}_{ji}) f(\mathbf{x}_{ji} | \mathbf{X}_{ji}) d\mathbf{x}_{ji} = \frac{n_{hji} + 1/k_j}{n_{ji} + 1}, \quad [2]$$

where  $X = \{X_{ijm}\}$  is a matrix of multilocus genotypes among  $n$  individuals sampled from the  $I$  populations, and  $X_{ijm}$  is the allele present at the  $j$ th locus of the  $m$ th individual sampled from the  $i$ th population. Note that  $n_{hji}$  is the total number of  $h$  alleles at locus  $j$  in the sample from population  $i$ , and  $n_{ji}$  is the total number of alleles typed for locus  $j$  from population  $i$  (both are functions of  $X_{ji}$ ). We define  $k_j$  to be the total number of alleles present at locus  $j$  (assumed to be identical in all populations and set equal to the observed number of alleles present in all populations). To generate the posterior density of  $x_{ji}$ , given  $X_{ji}$ , we have assumed a Dirichlet prior density for the allele frequencies (2). The posterior probability density that individual  $m$  arose from population  $i$  is

$$\Pr(i | \mathbf{Y}_m, \mathbf{X}) = \frac{\Pr(\mathbf{Y}_m | \mathbf{X}_i) f(i)}{\sum_{l=1}^I \Pr(\mathbf{Y}_m | \mathbf{X}_l) f(l)}, \quad [3]$$

where

$$\Pr(\mathbf{Y}_m | \mathbf{X}_i) = \prod_{j=1}^L \Pr(Y_{mj} | \mathbf{X}_{ji}). \quad [4]$$

If it is assumed that all populations are equally likely sources for individual  $m$ , *a priori*, then  $f(i) = 1/I$  for all  $i$ . The potential source populations for an unknown isolate were considered to be those populations contained in the 95% credible set. The 95% credible set is obtained by ranking potential source populations according to their posterior probabilities and sequentially placing populations into the credible set until the posterior probabilities of the populations contained in the set sum to be greater than 0.95 (see Berger ref. 8).

**Analysis with Population Simulations.** We studied the expected proportion of BATs that will correctly assign isolates by simulating haploid populations under a range of models using the software environment written by Balloux (9). For each run, 5 populations consisting of 1,000 fully recombining haploid isolates were allowed to evolve for 5,000 generations under an island model of population dispersal. Here, an isolate undergoing dispersal had an equal probability of migrating to any population, and the probability of migration per generation was varied between 0.1 and 0.0001 for each set of simulations. Loci evolved according to the stepwise (10) or infinite alleles (11) models of mutation, and the total genetic diversity was determined by setting the mutation rate to either 0.001, 0.0001, or 0 per generation. The impact of multilocus linkage disequilibrium on the population assignments was investigated by varying the per generation recombination rate between 0 and 1. For each simulation, reference samples of  $n = 30$  genotypes were randomly picked from every population, except where the effects of sample size were being investigated in which case  $n$  varied between 5 and 45. BATs were then used to determine the source population of 50 genotypes. The resulting posterior probabilities were assessed as a function of  $F_{st}$  (12), here estimated between pairs of populations with the Weir and Cockerham estimator,  $\theta$  (13), and calculated with the program FSTAT (14). The strength

**Table 1. Genetic differentiation ( $F_{st}$ ) between populations for *C. posadasii* and *C. immitis***

	<i>C. posadasii</i>				
	SA	SC	TX	AZ	MX
SA	0.000				
SC	0.2606**	0.000			
TX	0.1389**	0.1486**	0.000		
AZ	0.3984**	0.1203*	0.2785**	0.000	
MX	0.1966**	0.0820*	0.1191**	0.2021**	0.000
	<i>C. immitis</i>				
	SC	CC	MX		
SC	0.000				
CC	0.1056**	0.000			
MX	-0.0047	0.0662**	0.000		

SA = South America; SC = Southern California; TX = Texas; AZ = Arizona; MX = Mexico; CC = Central California.

\*\*Denotes significance at  $P = 0.99$ ; \* Denotes significant  $F_{st}$  at  $P = 0.95$ . Source of the data is from an online database at <http://plantbio.berkeley.edu/~taylor/ftp/genotypes.pdf> (4).

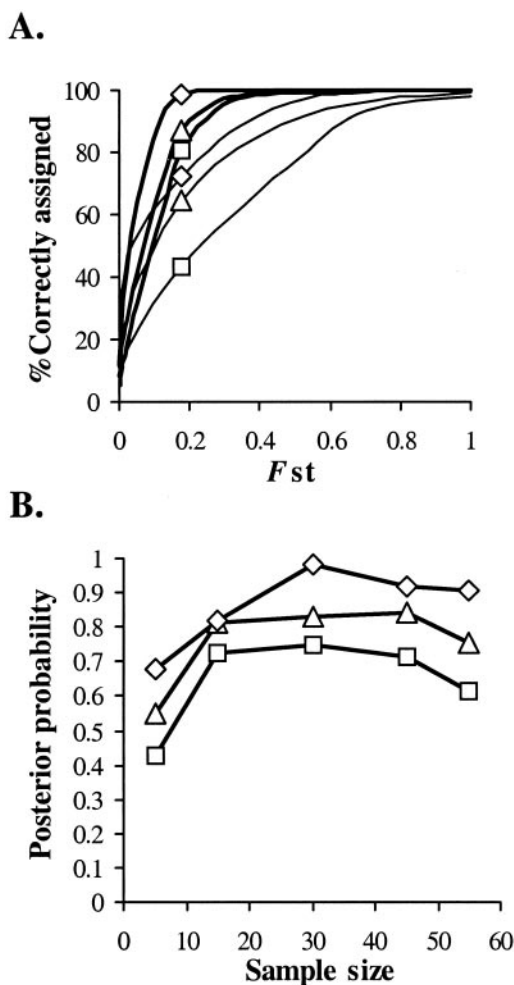
of multilocus linkage associations was assessed with the standardized index of association ( $I_A$ ; ref. 15).

**Population Allele Frequencies for *C. immitis* and *C. posadasii*.** Here, we use data from nine microsatellite loci that were isolated from *C. immitis* (loci *GAC*, *62I*, *GA37*, *GAI*, and *ACJ*) and *C. posadasii* (loci *KO3*, *KO7*, *KO1*, and *KO9*). These loci are highly informative because of their extensive polymorphism (between 8 and 26 alleles per locus), making them useful for discriminating between genetically similar isolates (16, 19). Multilocus genotypes were determined for 161 clinical and environmental isolates from 8 populations (3 from *C. immitis* and 5 from *C. posadasii*) spanning the worldwide distribution of *Coccidioides* sp., as described (4).

Within both species,  $F_{st}$  is significant between most pairs of populations (Table 1), and isolates tend to cluster in phylogenetic analyses according to where they were found, showing that pronounced phylogeographical structure exists (4). Previous analyses have shown that levels of linkage disequilibria are low within these populations (barring South American isolates), indicating that recombination has caused local panmixia (17, 18). Within South America, the likelihood that this population results from a postpleistocene radiation from North America, and has therefore undergone a population bottleneck, means that levels of linkage disequilibria are high relative those observed in North America (4). However, pairwise comparisons between unlinked loci show that recombination is (or has) occurred and the  $I_A$  for this population is low but significant ( $I_A = 0.12$ ,  $P < 0.01$ ).

**Determining the Source Population for Isolates of Unknown Origin.**

We applied BATs to determine where patients from nonendemic areas had acquired their infections. Hospitalization records from 1992–1997 were obtained from the state of New York and used to identify 161 persons who were diagnosed with coccidioidomycosis; of this set, 16 patients had positively identified *Coccidioides* sp. cultures that were available to us (7). All of these 16 patients had a clinical history that recorded travel to areas that are endemic for *Coccidioides* sp. The isolates were genotyped at all nine microsatellite loci and, based on two diagnostic loci (*62I* and *GAC2*; ref. 19), were identified as *C. immitis* or *C. posadasii*. An individual is unambiguously assigned to a population if the posterior probability that it belongs to that population is greater than 0.95, given its genotype and the sample of reference genotypes from all potential source populations. The accuracy of

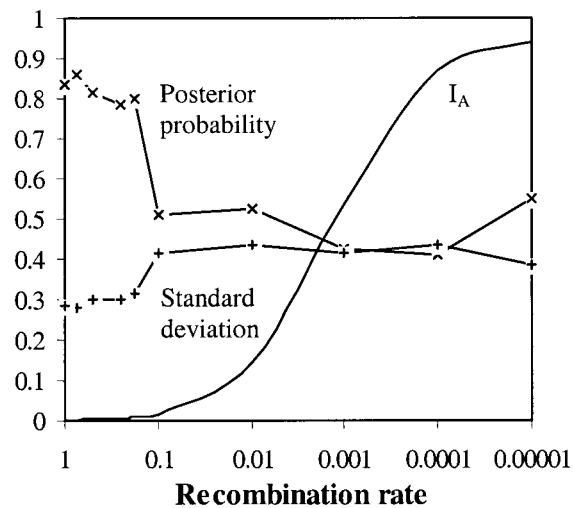


**Fig. 1.** (A) The percentage of individuals correctly assigned to their source population as a function of  $F_{st}$  for loci with high allelic diversity (an average of 10 alleles locus<sup>-1</sup>, thick lines) and low allelic diversity (an average of 4 alleles locus<sup>-1</sup>, thin lines). The numbers of loci used in the BATs are shown for 20 (diamonds), 10 (triangles), and 5 loci (squares). Curves are approximated by using logistic regression ( $P < 0.001$  for all curves). (B) Mean posterior probabilities for assignment to their original source population for 50 randomly selected isolates. Reference sample populations of size 5, 15, 30, and 45 are used, drawn from a simulated population of  $n = 5,000$ .  $F_{st} \approx 0.16$  between populations and assignments are shown for samples of 20, 10, and 5 loci (see A for the definitions of symbols).

the test for determining the correct source population was estimated by simulating novel genotypes from each population with the observed allele frequencies. BATs were then used to assign each simulated genotype, and the numbers of individuals correctly assigned to their source population were determined.

**Results**

**Assignment to Simulated Populations.** We analyzed the performance of the BATs by observing the posterior probabilities and measuring the proportion of individuals correctly assigned to their population as a function of  $F_{st}$ . Fig. 1A shows the relationship, approximated by logistic regression, between the numbers of isolates correctly assigned to the source population and  $F_{st}$  for high (an average of 10 alleles per locus) and low (an average of 4 alleles per locus) diversity populations. The relationships are also shown as a function of the numbers of loci used in the BATs. These data show that to achieve assignment of over 95% of isolates for the levels of genetic divergence seen in this study



**Fig. 2.** Mean posterior probabilities for assignment to their original source population for isolates selected from simulated populations with varying recombination rates. Reference sample populations of 30 isolates and 20 loci are used with  $F_{st}$  set to  $\approx 0.2$ . Posterior probabilities (line marked with  $\times$ ) and their associated SDs (line marked with  $+$ ) are marked, as well as a curve plotting the increase of the standardized  $I_A$  across the simulations.

(where  $F_{st}$  between populations ranges from 0 to 0.4; Fig. 1A), then locus diversity needs to be high (10 alleles per locus). At this level of diversity, only 5 loci were needed to correctly assign 95% of isolates to populations of  $F_{st}$  between 0.3 and 0.4. However, for populations that are more closely related ( $F_{st} \approx 0.2$ ), then between 10 and 20 loci were needed to correctly assign isolates. Further, these data showed that the power of the BATs critically depended on the levels of diversity found at loci. Here,  $F_{st}$  between populations needed to be in the order of 0.5–0.8 to correctly assign isolates to populations for low diversity loci (4 alleles per locus).

In a separate series of analyses, we examined the effect of increasing sample size on the accuracy of assignment for a series of populations with values of  $F_{st}$  similar to those observed in this study ( $F_{st}$  0.15–0.17; Fig. 1B). Here, posterior probabilities were seen to plateau at values above  $n = 15$  showing that little statistical power was gained by increasing the size of sampled populations. The effect of the model of mutation on posterior probabilities was also examined. The BATs were found to be relatively impervious to how loci were evolving, although assignments for loci mutating under a stepwise mutation model were found to have slightly reduced power as a result of the occurrence of back mutations that lowered the overall genetic diversity of the simulated populations (data not shown). However, these data only affect the general conclusions drawn here in that they tend to the conservative.

The effect of interlocus associations was determined by setting  $F_{st}$  to  $\approx 0.2$  and then analyzing the magnitude and variance of posterior probabilities as the  $I_A$  increased. The simulations (Fig. 2) showed a significant reduction in posterior probabilities and an associated rise in their variances, as the  $I_A$  rose above 0.02. This change corresponded to a 60% loss in the power of the analysis. Further loss in power associated with increasing linkage disequilibria was not observed and increasing the  $I_A$  until total clonality was observed demonstrated that 40% of isolates were still being correctly assigned to their source population with a probability of  $>0.95$ .

**Assignment of Clinical Isolates to Populations.** Diagnostic loci showed that the isolates of *Coccidioides* sp. from New York clinical cases consisted of both *C. posadasii* (12 isolates, Table 2

**Table 2. Results of assignment tests for nonendemic clinical cases of coccidioidomycosis found in New York State for *C. posadasii* and *C. immitis***

Isolate	Travel	Species	Potential source	Posterior probabilities
269.97*	Arizona	<i>C. posadasii</i>	SC	0.2086
			TX	<b>0.6580</b>
			AZ	0.0324
			MX	0.0998
376.95*	Arizona	<i>C. posadasii</i>	AZ	<b>0.6419</b>
			MX	0.3364
104.94	Arizona	<i>C. posadasii</i>	AZ	<b>0.9884</b>
639.97*	California/Mexico/ Southwest U.S.	<i>C. posadasii</i>	AZ	<b>0.4911</b>
			MX	0.2939
815.97	California/Arizona	<i>C. posadasii</i>	AZ	<b>0.9954</b>
515.95	California/Arizona	<i>C. posadasii</i>	AZ	<b>0.9705</b>
588.97	Arizona	<i>C. posadasii</i>	AZ	<b>0.9887</b>
229.96	Arizona	<i>C. posadasii</i>	AZ	<b>0.9500</b>
686.97	Southwest U.S.	<i>C. posadasii</i>	AZ	<b>0.9989</b>
819.96	California/Arizona	<i>C. posadasii</i>	AZ	<b>0.9508</b>
371.97	Arizona	<i>C. posadasii</i>	AZ	<b>0.9980</b>
750.95	Arizona	<i>C. posadasii</i>	AZ	<b>0.9992</b>
779.97	California	<i>C. immitis</i>	SC	<b>0.9995</b>
131.96*	Arizona/Texas	<i>C. immitis</i>	SC	0.1307
			CC	<b>0.8265</b>
366.96	Arizona	<i>C. immitis</i>	CC	<b>0.9898</b>
474.97	California	<i>C. immitis</i>	CC	<b>0.9750</b>

SA = South America; CC = Central California; SC = Southern California; TX = Texas; AZ = Arizona; MX = Mexico.

\*Signifies isolates that have a significant posterior probability of having come from one or more source populations. Isolates marked in bold are the most likely source population taken from within the 95% credible set.

*C. posadasii*) and *C. immitis* (4 isolates, Table 2 *C. immitis*). BATs assigned 12 of 16 isolates to a source population with high probability (Table 2). For four isolates, assignment was equivocal between two or more populations (isolates 269.97, 376.95, 639.97, and 131.96). We empirically determined the accuracy of the BATs for all populations and species by simulating novel genotypes and then determining the numbers that were correctly assigned to their source population (Table 3). These simulations showed that 97% of *C. posadasii* and 93% of *C. immitis* isolates were assigned to the correct source populations (76% and 66% of isolates were assigned with posterior probabilities of >0.95).

In 11 of the 16 tests, the statistically determined source population matched the travel history of the patient and often differentiated among a number of potential regions where infection might have been acquired, e.g., the patient contributing isolate 686.97 had traveled throughout the Southwest U.S. (and therefore had four potential sources of infection), but the genotype showed that it was most likely to have been acquired in Arizona. In 2 of 16 tests, the source population did not match the known travel history of the patient (isolates 269.97 and

**Table 3. Results of assigning simulated genotypes to observed populations and species by using observed allele frequencies**

	$F_{st}$ *	Alleles per locus (range)	% Simulated loci correctly assigned (with $P > 0.95$ )
<i>C. posadasii</i>	0.216 ± 0.07	9.25 (4–16)	97 (76)
<i>C. immitis</i>	0.081 ± 0.04	7.30 (4–12)	93 (66)

\*Average  $F_{st}$  between all geographic populations within species. The ±95% confidence interval is assessed by bootstrap.

366.96). In these cases it is possible that either the case history of the patient is incomplete or the patient had acquired an isolate that had undergone long-distance dispersal.

## Discussion

A number of assignment methods for diploid populations have been rigorously tested by using simulation approaches by Cornuet *et al.* (20), with the general conclusion that a Bayesian approach displayed uniformly higher probabilities of the correct assignment of an individual by comparison with the other methods. The study further concluded that knowledge of the  $F_{st}$  values among an array of populations provided a means of predicting the performance of assignment tests. This now enables genotyping experiments to be optimally designed with respect to the numbers and type of loci required for accurate assignment. The simulations described here bear out their conclusions. We show that there are considerable gains in statistical power resulting from increasing the numbers of loci and targeting highly polymorphic loci. This finding is illustrated by the observation that  $F_{st}$  would need to be on the order of 0.5–0.8 to accurately assign isolates for low-diversity loci. These levels of genetic differentiation are more representative of between-species levels of genetic differentiation (e.g., many alleles are fixed between populations). As a consequence, studies on intraspecific genetic differentiation should aim to maximize the numbers of alleles found at a locus, as well as increasing the numbers of loci themselves.

Our analyses of natural infective *Coccidioides* isolates show that genetic differentiation among spatially separated populations is sufficient to allow assignment of infections of unknown origin to a source population with high probability. For our clinical data, the expected fraction of misassigned isolates is low, as measured by the fraction of misassigned genotypes observed when simulating with the estimated population frequencies (between 3 and 7%). Moreover, simulated data sets suggest that typing a further 5–10 loci would achieve nearly perfect assignments given the levels of differentiation seen in our study populations. These data are in contrast to a previous study by Chaturvedi *et al.* (7) who, using these same 16 isolates, assigned them all to the same species (*C. posadasii*) and population (Arizona). This prior study compared the genotypes of five single nucleotide polymorphism-containing loci against their published allele frequencies (21). Problems inherent in that study were twofold: (i) the loci and alleles used had low discriminatory power between populations and species and (ii) statistical methods for gauging the significance of the associations were not available. The data presented here demonstrate the increased resolution and confidence that is possible when more, highly polymorphic, genetic markers are used in conjunction with the newly developed statistical tools.

Our clinical data show that there are two cases where clinical history and assignment are in conflict (isolates 269.97 and 366.96). That there are high levels of population subdivision for *Coccidioides* shows that effective long-distance migration between geographical areas has been low historically and suggests that effective wind dispersal of spores is rare. It is therefore unlikely that these two isolates represent cases where the patient has inhaled spores undergoing long-distance dispersal. Rather, incomplete clinical histories are a more likely explanation. One such instance has been recognized by Burt *et al.* (21), who showed that a Californian isolate with atypical genotype was the result of a patient having acquired an infection in Texas, then having moved to Central California where the infection was diagnosed.

In the case of isolate 269.97, the BATs show that four populations are the possible sources, including the region that the patient traveled through, suggesting that the Arizona population (AZ) is still a possible, albeit less likely, source of

infection. In some cases, the set of alleles found within a population may be uninformative relative to the observed genotype. In this case, the power of the test will be low, resulting in posterior probabilities that are nearly uniform (each population having posterior probability of  $1/I$ ) as is seen in the case of 269.97. For such isolates, examination of additional loci might resolve this ambiguity by increasing the power of the assignment test.

Active surveillance of infectious diseases allows rapid comparison of DNA fingerprints with centralized, and standardized, databases. Matches made in this manner promise to revolutionize the early detection and management of common-source outbreaks as well as acts of bioterrorism. Moreover, once these databases achieve sufficient coverage of natural genetic and geographical diversity, matching cases to sources of infection becomes possible. However, these techniques are relevant for organisms that reproduce mainly clonally, and hence have relatively stable genomes. Because of the effect of genetic recombination shuffling alleles, the multilocus genotypes of pathogens undergoing random genetic exchange will typically not match those in fingerprint databases, unless by chance alone. Such considerations are paramount in the medically important fungi where estimates show that 75% are morphologically sexual, and where several of the morphologically asexual species (which include *C. immitis* and *C. posadasii*) have been shown to undergo cryptic recombination in nature (3, 22, 23). Within *Coccidioides*, the rate of recombination is high enough that in the two populations that have been extensively surveyed for single nucleotide polymorphisms, only three instances of genetically identical isolates have been found (of which two could have occurred by chance alone), and tests consistently confirm the occurrence of panmixia and linkage equilibrium among markers (17, 18). For such organisms where clonal reproduction has a low impact on the population genetic structure, DNA fingerprinting studies should take into account recombination by comparing the variation in allele frequencies between populations, as described here, rather than the simple identity of multilocus genotypes (24).

There are two main caveats associated with our method. First, our simulations show that as significant linkage associations occur as a result of a decrease in recombination rates, then posterior probabilities drop by about 60%. This finding confirms

the necessity of limiting analyses to recombining populations, although loci in linkage disequilibrium can be accommodated at a significant cost to the power of the analysis. In cases where low levels of linkage occur ( $I_A = 0 - 0.1$ ), clone correcting data sets by removing identical genotypes may be applicable. Second, a wide sampling of genotypes are needed to include all potential source populations, and as it is likely that sampling will be focused on clinical isolates then some potential environmental source populations may be missed, initially at least. Unidentified population structure may also occur and further refinement of our method will focus on the development of clustering algorithms to define populations based on local genetic equilibria, as recently proposed by Dawson and Belkhir (25). However, once a scientific community has decided on a standardized set of marker loci on which to focus, multi-user submission of data then will generate data sets of useful depth and quality, as has been the case with the Multilocus Sequence Typing projects of bacteria (26). For recombining pathogens, adopting standardized markers will enable the development of high-resolution data sets that will allow the assignment of isolates to increasingly finer levels of population structure, ultimately providing a powerful tool for detecting the source of disease outbreaks.

Although this article has focused on haploid eukaryotes, the techniques described are potentially applicable to other systems. Multiclonal infections are being observed in many bacterial and viral diseases (27), and within these infections chimeric genomes are often observed, implying recombination. Even for taxa lacking meiotic capability, recombination can be frequent, as the often high levels of segregation within populations of certain bacterial species show (28). This mixing illustrates the pressing need to develop the theory and methodology that is necessary to allow the incorporation of data from recombinant populations into molecular epidemiological studies.

**Program Availability.** A computer program (BAYESASS) implementing these procedures was written by B.R. and used to perform the BATs carried out in this article. The program is available from [www.rannala.org](http://www.rannala.org).

This work was supported by National Institutes of Health Grants PO1 AI37232 (to J.W.T.) and R01 HG01988 (to B.R.) and by a Wellcome Trust Biodiversity Fellowship (to M.C.F.).

- Daszak, P., Cunningham, A. A. & Hyatt, D. (2000) *Science* **287**, 443–449.
- Rannala, B. & Mountain, J. L. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 9197–92013.
- Taylor, J. W., Geiser, D. M., Burt, A. & Koufopanou, V. (1999) *Clin. Microbiol. Rev.* **12**, 126–146.
- Fisher, M. C., Koenig, G. L., White, T. J., San-Blas, G., Negroni, R., Alvarez, I. G., Wanke, B. & Taylor, J. W. (2001) *Proc. Natl. Acad. Sci. USA* **98**, 4558–4562.
- Rippon, J. W. (1988) *Medical Mycology* (Saunders, Philadelphia).
- Kirkland, T. M. & Fierer, J. (1996) *Emerg. Infect. Dis.* **2**, 192–199.
- Chaturvedi, V., Ramani, R., Gromadzki, S., Rodeghier, B., Hwa-Ghan, C. & Morse, D. (2000) *Emerg. Infect. Dis.* **6**, 25–29.
- Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis* (Springer, New York), 2nd Ed.
- Balloux, F. (2001) *J. Hered.* **92**, 301–302.
- Ohta, T. & Kimura, M. (1973) *Genet. Res.* **22**, 201–204.
- Kimura, M. & Crow, J. F. (1964) *Genetics* **49**, 725–738.
- Wright, S. (1951) *Ann. Eugen.* **15**, 323–354.
- Weir, B. S. & Cockerham, C. C. (1984) *Evolution (Lawrence, Kans.)* **38**, 1358–1370.
- Goudet, J. (2001) FSTAT 2.9.3 (Institute of Ecology, Lausanne, Switzerland).
- Haubold, B. & Hudson, R. R. (2000) *Bioinformatics* **16**, 847–848.
- Fisher, M. C., Koenig, G. L., White, T. J. & Taylor, J. W. (1999) *Mol. Ecol.* **8**, 1082–1084.
- Burt, A., Carter, D. A., Koenig, G. L., White, T. J. & Taylor, J. W. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 770–773.
- Fisher, M. C., Koenig, G. L., White, T. J. & Taylor, J. W. (2000) *J. Clin. Microbiol.* **38**, 807–813.
- Fisher, M. C., Koenig, G. L., White, T. J. & Taylor, J. W. (2001) *Mycologia* **94**, 73–84.
- Cornuet, J.-M., Piry, S., Luikart, G., Estoup, A. & Solignac, M. (1999) *Genetics* **153**, 1989–2000.
- Burt, A., Dechairo, B. M., Koenig, G. L., Carter, D. A., White, T. J. & Taylor, J. W. (1997) *Mol. Ecol.* **6**, 781–786.
- Gräser, Y., Volovsek, M., Arrington, J., Schonian, G., Presber, W., Mitchell, T. G. & Vilgalys, R. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 12473–12477.
- Rodriguez, E., De Meeus, T., Mallie, M., Renaud, F., Symoens, F., Mondon, P., Piens, M. A., Lebeau, B., Viviani, M. A., Grillot, R., et al. (1996) *J. Clin. Microbiol.* **3**, 2559–2568.
- Soll, D. R. (2000) *Clin. Microbiol. Rev.* **13**, 332–370.
- Dawson, K. J. & Belkhir, K. (2001) *Genet. Res.* **78**, 59–77.
- Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Cagant, D. A., et al. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 3140–3145.
- Read, A. F. R. & Taylor, L. H. (2001) *Science* **292**, 1099–1101.
- Feil, A. J. & Spratt, B. G. (2001) *Annu. Rev. Microbiol.* **55**, 561–590.