TITLE: Local population structure and patterns of Western Hemisphere dispersal for *Coccidioides spp*., the fungal cause of Valley Fever

Running Title: Phylogeographic patterns of *Coccidioides* fungi

David M. Engelthaler[a,1,2], Chandler C. Roe[a,2], Anastasia P. Litvintseva[b], Elizabeth M. Driebe[a], James M. Schupp[a], Lalitha Gade[b], Victor Waddell[c], Kenneth Komatsu[c], Eduardo Arathoon[d], Heidi Logemann[e], George R. Thompson III[f], Tom Chiller[b], Bridget Barker[a], Paul Keim[a,g]

Affiliations:

[a] TGen North, Translational Genomics Research Institute, 3051 W. Shamrell Blvd., Flagstaff, AZ, 86005

[b] Mycotic Diseases Branch, National Center for Emerging and Zoonotic Infectious Diseases, Centers For Disease Control and Prevention, 1600 Clifton Rd Atlanta, GA, 30333

[c] Division of Public Health Services, Arizona Department of Health Services, 150 N. 18th Avenue, Suite 100, Phoenix, Arizona, 85007

[d] Asociación de Salud Integral, 2da avenida 11-53 Zona 1, Guatemala City 01001, Guatemala

[e] Universidad de San Carlos, Ciudad Universitaria, 11 Av, Guatemala 01012, Guatemala.

[f] Division of Infectious Diseases, Department of Medicine, University of California Davis, 1 Shields Avenue, Tupper Hall, Rm. 3146, Davis, CA 95616

[g] Microbial Genetics and Genomics Center, Northern Arizona University, Flagstaff, AZ, 86011-4073

[1]To whom correspondence should be addressed: dengelthaler@tgen.org

[2]These authors contributed equally to this manuscript

Keywords: *Coccidioides*, molecular epidemiology, phylogenetics, mycology

Abstract

Coccidioidomycosis (or Valley Fever) is a fungal disease with high morbidity and mortality that affects tens of thousands of people each year. This infection is caused by two sibling species, *Coccidioides immitis* and *C. posadasii*, which are endemic to specific arid locales throughout the Western Hemisphere, particularly the desert southwest of the United States. Recent epidemiological and population genetic data suggest that the geographic range of coccidioidomycosis is expanding as new endemic clusters have been identified in the state of Washington, well outside of the established endemic range. The genetic mechanisms and epidemiological consequences of this expansion are unknown and require better understanding of the population structure and evolutionary history of these pathogens. Here we perform whole genome SNP analyses of 64 new and 18 previously published genomes. The results provide evidence of substantial population structure in *C. posadasii* and demonstrate presence of distinct geographic clades in central and southern Arizona as well as dispersed populations in Texas, Mexico, South America and Central America. Although a smaller number of *C. immitis* strains were included in the analyses, some evidence of phylogeographic structure was also detected in this species, which has been historically limited to California and Baja Mexico. SNP-based analyses indicated that *C. posadasii* is the more ancient of the two species and southern Arizona contains the most diverse subpopulations. We propose a southern Arizona-northern Mexico origin for *C. posadasii* and describe a pathway for dispersal and distribution out of this region.

Data Access for Reviewers: All WGS data files have been deposited in the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/bioproject, project # PRJNA274372)

Introduction

C. immitis and *C. posadasii* are the etiological agents of coccidioidomycosis, or Valley Fever, a primarily pulmonary disease that causes tremendous morbidity in the Southwestern US and other focal regions in the Americas. *C. posadasii* was first identified as non-California *C. immitis* found in AZ, NM, TX, Mexico and sporadic locales in South America, and named as a distinct species in 2002 (Fisher et al. 2002a). *C. immitis* is primarily found in California's Central Valley and the southern California-Baja Mexico region and has been recently described as endemic in southeastern Washington state (Marsden-Haug et a. 2014). The two species are nearly indistinguishable by clinical and microbiological phenotype, although specific differences have been reported (Cox and Magee 2004; Barker et al. 2007). *Coccidioides* is a dimorphic fungus, with a saprobic phase consisting of mycelia in the soil and a parasitic phase from inhaled arthroconidia, resulting in mammalian infection (Pappagianis 1967; Brown et al. 2013). Death and burial of the mammalian host can lead to re-infection of the soil (Barker et al. 2012). Both species of *Coccidioides* follow these stages and appear to lead to the same clinical outcomes (Sharpton et al. 2009).

Previous genomic analyses of targeted loci (e.g., microsatellites) led to the understanding that the California populations and the non-California populations are genetically distinct species (Fisher et al. 2001; Fisher et al. 2002a). Whole genome analysis has supported this concept, with hundreds of thousands of single nucleotide polymorphisms (SNPs) separating the two species (Engelthaler and Balajee 2010). In a comparative analysis of one genome of each species, Sharpton et al (2009) identified 1.1-1.5 Mb (~4-5%) of dissimilar genetic content between the two genomes. A later comparative genomics study,

based on the analysis of ten genomes of each species, determined that the separation of the two species has not remained fully complete, and documented introgression from likely hybridization events following species separation (Neafsey et al. 2010). As rapid DNA sequencing analysis becomes more accessible to the research and public health laboratories, approaches like whole genome SNP typing (WGST) have shown to be useful for identifying clonal fungal outbreaks (Engelthaler et al. 2011; Etienne et al. 2012; Litvintseva et al. 2014; Litvintseva et al. 2015a). However, genomic epidemiology is also needed to help establish location of exposure of a non-clonal outbreak, and for in-depth analyses of local population structure to better understand pathogen emergence, dispersal and expansion (Litvintseva et al. 2015b).

It is currently thought that wind, water, mammal hosts and anthropogenic causes are important mechanisms for local and geographic-scale dispersal of arthroconidia (Pappagianis and Einstein 1978; Fisher et al. 2007). Fisher et al (2001) using microsatellite data, described a phylogeographic pattern of *C. posadasii* that was primarily driven by human migration, strongly suggesting a movement of *Coccidioides* into South America that was contemporary with the first human movements onto the continent. While this groundbreaking study was critical to understanding large-scale phylogeographic features of *Coccidioides* epidemiology, no evidence was identified for local population structure. An additional microsatellite analysis of Arizona-only isolates was also unable to identify local structure within that state (Jewell et al. 2008).

Given the limited number of genomes previously available, an accurate understanding of population structure and phylogeography has been problematic. Here we provide comparative whole genome SNP and recombination analysis of 82 *Coccidioides* genomes with some insight into the population structure and phylogeographic differences between and within the two species, as well as a hypothetical model for the global dispersal and distribution of *C. posadasii*.

Results

Sixty-eight newly sequenced genomes were assembled and compared to 18 of the existing 28 published genomes (Neafsey et al. 2010; Engelthaler et al. 2011; Litvintseva et al. 2015a), resulting in a total of 86 genomes (26 *C. immitis* and 60 *C. posadasii*) available for analysis (Table S1). Depth of sequence coverage for the new genomes ranged from 23X to 228X (avg. 67X), resulting in an assembled genome size range of 27.1Mb to 28.5 Mb. As sequence data produced by different sequencing platforms provided differing levels of coverage and sequencing errors, phylogenetic analyses were performed multiple times including and excluding previous genome data. The publically available Sanger sequencing data (http:// www.broad.mit.edu/annotation/genome/coccidioides group/-MultiHome.html) only included final assemblies, making it difficult to distinguish true SNPs from sequence error (which is likely the cause of the inordinately long SNP branches) (Supplemental Figures S1-S3); as such, these genomes were not included in all population analyses. Six of the previously published Illumina sequences (Neafsey et al. 2010) had low depth of coverage (3.4-8.5X) and the three previously published SOLiD sequences (Engelthaler et al. 2011) used short reads (35-50bp) resulting in larger error rates (data

not shown) and these sets were removed from the analyses altogether. Phylogenies that include previously sequenced strains are provided in Supplemental Figures 1-3. The final "all species tree", included 69 genomes (18 *C. immitis* and 51 *C. posadasii*) and identified 405,244 shared SNP loci, that is, loci in genetic content shared by all strains, with 296,632 of those being parsimony informative (Figure 1). The majority of the parsimony informative SNPs (187,442; 63%) separate the two species. Additionally, an analysis of molecular variance (AMOVA) showed significantly high levels of genomic separation ($\Phi PT = 0.91$, p,0.001) providing the clearest evidence to date for their taxonomical separation. Homoplasy was present, as the consistency index (CI) was 0.427, which is likely due to recombination within each species as well as possible limited introgression between the two species (Neafsey et al. 2010), however the high retention index (RI) of 0.998 indicates that most of the non-homoplasious SNPs are synapomorphic rather than autapomorphic (or lineage specific). *C. posadasii* demonstrated a larger average SNP distance (i.e., the number of base substitutions per site from averaging overall pairwise distance between isolates within each species) than *C. immitis* (0.062 versus 0.024 respectively) and individual lineages within *C. posadasii* contained an average of over twice as many autapomorphic SNPs as *C. immitis* (Table 1).

Phylogenetic histories of each species were interrogated separately. The *C. posadasii* analysis was primarily based on 51 genomes: 38 isolates from the US, six isolates from Mexico, five from Central America (Guatemala) and three from South America (Brazil, Paraguay and Argentina). Isolate locales are based on available metadata and may represent locale of sample isolation, location of laboratory that stored isolate, or home

locale of patient. A total of 253,291 SNPs from shared loci (134,752 parsimony informative) were identified among the *C. posadasii* isolates. CI and RI values decreased when looking at the single species (0.214 and 0.364), suggesting a significant impact of both autapomorphy (i.e., isolate specific SNPs) and recombination within *C. posadasii*. Despite the presence of recombination, three major clades were clearly separated and well supported by the bootstrap analysis: one included all isolates from Arizona, one included only Guatemala isolates and the third contained nearly all other non-Arizona isolates (Figure 2).

Several distinct smaller clades were identified within the Arizona group. Isolates from Tucson formed multiple distinct clades (i.e., "Tucson Clades 1 and 2"), along with an additonal parental clade that contained multiple Tucson subpopulations (i.e., "Tucson Clade 3") and a group that contained nearly all the strains obtained from Phoenix (i.e., the "Phoenix Clade"). A closely related subpopulation contained both Phoenix and Tucson isolates and a single isolate from Nevada (i.e., the "Phx-Tuc Clade"). Only one Phoenix isolate was found within the Tucson clades, suggesting a genomically distinct Phoenix subpopulation, which likely originated from a Tucson population. A more distant clade was composed of all isolates from Texas, Mexico and South America, except for the "Sonora_2" strain, from northern Mexico, which was identified as most basal in the "all-species tree" and was therefore used to root the *C. posadasii* tree. An additional publically available Sanger-sequenced *C. posadasii* isolate from Argentina grouped closely with the isolate from Paraguay (Supplemental Figure 2) in this clade. Four of the five Central American isolates from Guatemala formed their own distinct clade.

We identified 64, 096 shared SNP loci (31,372 parsimony informative) among 18 assembled *C. immitis* genomes (Figure 3). The CI and RI values (0.345 and 0.384, respectively) also describe homoplasy in the population, likely caused by recombination within and between populations. The maximum parsimony tree separated the analyzed genomes into two distinct clades, one with membership predominantly from San Joaquin Valley (Clade 1) and one with mixed membership of San Joaquin Valley, San Diego, Washington and Mexico locales (Clade 2). A single isolate labeled as originating from Argentina (B0727) also grouped in this clade, although the patient was treated in Buenos Aries (not thought to be an endemic zone) and possibly represents an exposure outside of Argentina (although autochthonous South American exposures to *C. immitis* have been recently reported (Canteros et al. 2015)). The Washington isolates were previously established to be clonal (Mardsen-Haug et al. 2014) and while they appear to belong to Clade 2, they are clearly endemic to southeast Washington (Litvintseva et al. 2015a). The inclusion of the publically available Sanger sequenced San Diego strains added an additional San Joaquin Valley/San Diego sub-clade (Supplemental Figures 1 and 3), although the strength of these groups is limited due to fewer *C immitis* genomes in this study.

These results, as well as reports by others, indicate evidence of recombination in both species of *Coccidioides*. The analysis of molecular variance (AMOVA), while providing strong evidence of the two species as distinct non-breeding populations, did not establish significant genomic separation within species (Table 1), providing further evidence of local recombination. In order to account for recombination in the population analysis, (Taylor et

al. 1999; Fisher et al. 2002b) we applied *fineStructure*, a model-based Bayesian approach (Lawson et al. 2012) for incorporating recombination signal into population structure, to the genus-level phylogenetic analysis. The *fineStructure* analysis appropriately separated the 69 *Coccidioides* genomes into the two species and assigned within-species "populations" or groups, related by shared genomic regions, which were similar to those found in the maximum parsimony trees (Figure 4). Genomic isolation (shown as yellow on the heat map) was demonstrable for the two species, but such a lack of shared genetic history was rarely seen among members within each species.

Additionally, we produced neighbor-net "splits tree" networks for each species (Figures 5 and S2 for *C. posadasii* and Figures S3 for *C. immitis*). The neighbor-net phylogenies are displayed as networks of radiating events with sub-populations joined by splits. This analysis also identified relationships similar to those described by the respective maximum parsimony trees. All the distinct subpopulations previously identified in the *C. posadasii* tree were also distinguished in the network, including separation of the central Arizona (i.e., Phoenix) and southern Arizona (i.e. Tucson) clades, and a pronounced separation between Arizona strains and Texas-Mexico-South American (Tex-Mex-SA) strains (Figure 5). The basal "Sonora_2" strain from the maximum parsimony tree appears as the most proximal genome in this clade. In both the maximum parsimony tree and the neighbor-network, the two Guatemalan isolates demonstrated strong similarity to each other (although not completely clonal) and were distinct from the Arizona and Tex-Mex-SA populations. The maximum parsimony tree suggests a relatedness of this Guatemalan clade to one Arizona sample ("Tucson_24") that is not seen in the neighbor network.

The *fineStructure* analysis of the *C. posadasii* genomes again revealed a similar population structure as the phylogenetic analyses, with identified groups, or populations, based on identifiable shared haplotype regions (Figure 6). The primary central Arizona, multiple southern Arizona, and Tex-Mex-SA groups are pronounced. The "Phx-Tuc Clade" primarily grouped with the Phoenix Clade members, suggesting that the two groups are likely a single more diverse Phoenix Clade. For several of these sub-species groups, members shared many genome regions with members from other groups, indicating the presence of historical admixture or incomplete or recent separation between groups within the species. Additionally, evidence was provided for an extensive genome sharing history of both the "Tucson_24" and "Sonora_2" strains with the remainder of the *C. posadasii* isolates, suggesting these genomes are representative of older, more basal lineages. This analysis also displayed the relative isolation of the Guatemalan Clade, as well as limited genome sharing of the individual strains from Brazil, Paraguay and Argentina, with the latter two displaying the most between-strain genomic sharing outside of the Guatemalan group.

Discussion

Here we present results of the whole genome SNP analysis of geographically diverse strains of *Coccidioides sp*. that has further confirmed presence of two cryptic species, *C. immitis* and *C. posadasii*, first established by Fisher *et al*. (2002a) and further described in subsequent studies (Neafsey et al. 2010; Barker et al. 2012). We also provide evidence of further genetic subdivision within the major clades of *C. posadasii* and *C. immitis* and demonstrate the presence of genetically distinct subpopulations associated with specific geographic

locales. A phylogeographic map, with a focus on *C. posadasii*, helps to clearly visualize the connection between genomic and geographic space (Figure 7); e.g., the extensive genomic diversity of the Arizona subpopulation contained within a limited geographic space contrasts with the limited genomic diversity of the geographically vast Texas-Mexico-South America subpopulation. These data and analyses further our ability to conduct genomic epidemiology, make more informed hypotheses of likely ancestry and pathways of dispersal and better understand the changing distribution of *Coccidioides*.

With the high resolution genotyping view provided by WGST we are now able to see multiple distinct populations of *C. posadasii* within Arizona, including a clear separation between the two major urban areas of Tucson and Phoenix, located in southern and central Arizona respectively. Such fine scale resolution of subpopulations allows for a better understanding of originating source or location of non-endemic cases and recent emergences in new geographical regions. For example, epidemiological data suggest that the "Colorado Springs_1" case was not likely a local exposure, as there is no known endemic coccidioidomycosis occurring in that region of Colorado; however, whole genome SNP analyses place this strain within a Tucson subclade within the Arizona population of *C. posadasii*, likely representing the originating environmental source for that infection; however no other isolates from Colorado were available in our study. In the same respect, the *C. immitis* cases from Lower Mexico (Guerrero and Michoacan) and Argentina (Buenos Aires), locales that are not known to support endemic populations of *C. immitis* (Castañón-Olivares et al. 2007; , Sifuentes-Osornio et al. 2012), are possibly epidemiologically linked to their phylogeographic placement within one or more central California (e.g. San Joaquin

Valley) clades. Such phylogeographic associations should still be confirmed with local environmental sampling. For example, while the Washington isolates appear to fall within a larger San Joaquin Valley clade, it is now understood, through careful examination of clinical and environmental isoaltes, that this highly clonal population emanates from a new endemic focus in southeastern Washington state (Marsden-Haug et al. 2014; Litvintseva et al. 2014)), with likely original dispersal out of California's Central Valley. Of note, the Washington isolate and the Argentina isolate have essentially identical "recombining" profiles as a central California isolate ("SJV_10") (Figure 4).

The genome resolution provided by whole genome SNP analysis also provides for an understanding of the possible ancestry of this fungus. *C. posadasii* displays significantly greater mean SNP distance than *C. immitis* (Table 1). This affords evidence of an older and more diverse population, with *C. posadasii* likely being ancestral to *C. immitis*. Neafsey et al. (2010) established that *C. posadasii* has a two-fold larger effective population size than *C. immitis*, suggesting that this was due to the larger geographic range, allowing for the development of more subpopulations. This, however, may also be explained by *C. posadasii's* being an older population, with more time for mutation and divergence.

Previous studies demonstrated evidence of introgression between *C. posadasii* and *C. immitis*, which may have originated from a possible hybridization event in some Southern California and Baja Mexico isolates of *C. immitis* (Neafsey et al. 2010). In this study, we observed limited cross species homoplasy (Figure 1), which is consistent with limited hybridization and indicates that the two populations are largely reproductively isolated

outside of the Southern California region where suspected hybridizations events are thought to have occurred (Neafsey et al. 2010).

When analyzing subpopulation diversity within *C. posadasii*, the Arizona clade contained significantly more lineage-specific diversity than the clade containing the Texas, Mexico and South American isolates (Table 1). As Arizona *C. posadasii* strains are more genomically diverse than the other sampled *C. posadasii* populations, in that the individual members have acquired more mutations, this suggests that the Arizona group is a more ancestral population than non-Arizonan *C. posadasii*. An alternate hypothesis would suggest an increased mutation rate in Arizona *C. posadasii*, over non-Arizona strains, although there is no additional evidence to suggest a mechanism for such an evolutionary disparity. A third possible mechanism for this lineage diversity is ongoing recombination within groups; however, we only see moderate levels of evidence of recombination events with the *fineStructure* analysis (Figure 6). This limited recombination may represent historical mating events. Genomic mating types are known for *Coccidioides* however no laboratory-controlled genetic recombination has been accomplished to date (Lewis et al. 2015) and the role of sexual reproduction remains unknown. Multiple subpopulations occurring in Southern Arizona (i.e., Tucson), and one derived clade that contains nearly all the Central Arizona (i.e., Phoenix) strains, indicates that the Southern Arizona area is likely the source of the Central Arizona population, although all Arizona populations have similar SNP distances, suggesting they are likely of similar age.

Conversely, the Tex-Mex-SA clade appears to be a highly diverged subpopulation from the primary Arizona population. "Sonora_2" (Mexico) and Texas isolate "B10813" appear to be the most basal members in this clade in the neighbor-net tree (Figure 5), although greatly diverged from each other. The same Texas isolate ("B10813") and the "Michoacan_1" (Mexico) isolate are the two most basal to the Tex-Mex-SA clade in the maximum parsimony tree (where "Sonora_2" was used at the root) (Figure 2). Additionally, these two strains appeared to have the most "shared" genome space within the *fineStructure* analysis of this clade, suggesting they represent common ancestral lineages. The South American isolates overall appear to be the most derived in this clade. The Argentina ("RMSCC_3700") and Paraguay ("GT_1078") isolates are closely related and are closer to other Mexico strains than the Brazil strain ("B5773"), suggesting more than one independent founding events into South America.

The Guatemala clade appears to be a distinct subpopulation, and not a member of the Tex-Mex-SA population. While tropical Guatemala would typically be considered to be outside the arid endemic regions, multiple accounts of endemic transmission have been recorded in its arid Motagua River valley (Mayorga and Espinoza 1970, Sifuentes-Osornio et al. 2012). These data lend evidence to a local distinct clonal population of *Coccidioides* in Central America, with more recent divergence between individuals than seen in other locales. One isolate (#730334) from Guatemala did group in the Tex-Mex-SA clade, representing a likely exposure outside of Guatemala. An epidemiologic follow up determined that this case contracted coccidioidomycosis while travelling in Texas.

Evidence exists for an additional Central American population in Honduras (Mayorga 1967, Brown et al. 2013), although no isolates were available for analysis in this study.

The distinct populations identified in the differing geographic regions likely reflect single or limited founder population events. Such events would conflict with the hypothesis of deposition of spores by wind as a likely mechanism for large-scale geographic dispersal (Pappagianis and Einstein 1978). A more likely hypothesis is that primary dispersal over large regions occurs through movement by mammalian hosts, similar to the previously proposed mechanism for emergence of *C. posadasii* in South America via the Central American land bridge between the continents (Fisher et al. 2001). The *Coccidioides* genomic adaptations to mammalian host (e.g., expansion of protease and keratinase gene families) (Sharpton et al 2009) and the hypothesis that the patchy distribution of *Coccidioides* in soil is due to the fungal association with rodent carcasses (i.e., dead hosts) and burrows (Whiston and Taylor 2014; Lewis et al. 2015) comports with a theory that distribution is related to distinct movement events of infected animals, whether over great distances (North America to South America) or shorter distances (e.g., Southern Arizona to Central Arizona).

A genomic interpretation of geographic dispersal from the data presented here would then tend to reiterate Fisher's animal dispersal model (Fisher et al. 2001), with notable additions, such as the ability to see a clear delineation between southern Arizona (Tucson) and central Arizona (Phoenix) populations. The limited presence of Tucson isolates in the Phoenix clade, and vice versa, may represent instances of wind dispersal and/or exposure

(i.e., infection in Tucson patient by wind-borne Phoenix-originating *Coccidioides* spores), although it is more plausible that patients infected in Phoenix are occasionally diagnosed in Tucson, and vice versa, as there is a high degree of travel between the two population centers. Cases of patients living in one endemic area but having exposure in another endemic area have been well documented (Burt et al. 1996; Taylor et al. 1999). The seemingly derived nature of the Phoenix, Tex-Mex-SA, and Guatemalan clades suggests independent dispersal events, and given the more diverse and likely older populations from southern Arizona, the most parsimonious explanation would be original dispersal from this locale.

A proposed *C. posadasii* dispersal model (Figure 8) would then suggest that: A) the central Arizona population originated from one of southern Arizona sub-populations; B) Texas and Mexico populations also came from this region, likely concurrently from the same original subpopulation; C) Mexico (and possibly Texas) subsequently fed the South American populations (more than once); and D) the Guatemalan population independently, and more recently, emerged from the Southern Arizona region. The evidence of historical genome sharing between the "Tucson_24" and "Sonora_2" strains and the rest of the species population, and the basal nature of the "Sonora_2" strain in the whole genus phylogeny, would suggest that these strains are from older lineages and may represent a historical southern Arizona-northern Mexico origin for the species.

We recognize the limitations of this study include sampling bias, especially with regards to analyzing fewer *C. immitis* and South American *C. posadasii* genomes, and hence a

possibility for both over and under-representation of isolates from some geographic regions. Further conclusions regarding the *C. immitis* population structure should only be drawn from the inclusion of additional genomes from more locales. The vast majority of the isolates studied here originated from clinical samples. It is possible that environmental soil isolates may provide differing genotypes than those from infecting strains, although this was not seen in previous analyses (Barker et al. 2012), nor was this seen with the inclusion of previously sequenced genomes from limited soil isolates (Supplemental Figure S2). Additionally, carefully collected environmental isolates would provide a more reliable representation of locally present strains (Lewis et al. 2015). Lastly, it is important to note that early sequence data sets may have less accuracy and coverage affecting their utility for whole genome SNP typing. Current sequencing provides excellent depth and quality that should allow their use for studies well into the future.

The findings of this study provide a strong argument to continue to conduct large-scale population level sequencing of *Coccidioides*, particularly for both clinical and soil isolates from under represented areas, especially Sonora/northern Mexico populations, to further understand the geographic extent of a possible southern Arizona-northern Mexico founding population. Even more importantly, our data demonstrate that local population structure does occur, even in recombining organisms, and that WGS approaches can be readily used for fungal molecular epidemiology of not only suspect clonal outbreaks, but also for linking cases to likely exposure sites and better understanding the patterns of emergence and dispersal.

Methods

Genome Sequencing

The genomes of 16 *C. immitis* and 48 *C. posadasii* isolates (Table S1) were sequenced using Illumina HiSeq and MiSeq sequencing platforms, as previously described (Engelthaler et al. 2014; Litvintseva et al. 2014). High molecular weight DNA was extracted using the ZR Fungal/Bacterial DNA Mini Prep kit (Zymo Research, Irvine, CA, USA, catalog #D6005). DNA samples were prepared for paired-end sequencing using the Kapa Biosystems library preparation kit (Woburn, MA, USA, catalog #KK8201) protocol with an 8bp index modification. Briefly, 2 µg of dsDNA sheared to an average size of 600 bp was use for the Kapa Illumina paired-end library preparation, as described by the manufacturer. Modified oligonucleotides (Integrated DNA Technologies, Coralville, IA, USA) with 8bp indexing capability (Kozarewa and Turner 2011), were substituted at the appropriate step. Prior to sequencing, the libraries were quantified with qPCR on the 7900HT (Life Technologies Corporation, Carlsbad, CA, USA) using the Kapa Library Quantification Kit (Woburn, MA, USA, catalog #KK4835). Libraries were sequenced to a read length of 100bp or 150bp on the Illumina HiSeq or to 250bp on the Illumina MiSeq.

Genome Assembly

The "San Diego_1" (*C. immitis*) and "B10813_Tx" (*C. posadsasii*) sequenced genomes were both *de novo* assembled using the SPAdes assembler v2.5.0 (Bankevich et al 2012). The "San Diego_1" assembly was used as the reference for the "All Species" and *C. immitis* SNP matrices and the "B10813_Tx" assembly was used as the *C. posadasii* SNP matrix (see SNP Variant Detection below)

SNP Variant Detection

Illumina read data were aligned against the respective reference assemblies using Novoalign 3.00.03 (www.novocraft.com) and then SNP variants were identified using the GATK Unified Genotyper v2.4 (DePristo et al. 2011). SNP calls were then filtered using an in-house tool (http://tgennorth.github.io/NASP/) to remove positions with less than 10x coverage, with less than 90% variant allele calls, or identified by Nucmer as being within duplicated regions in the reference. SNP matrices were produced for the "all species", *C. posadasii*, and *C. immitis* analyses.

Phylogenetic Analysis

To understand relationships between isolates, we conducted whole genome SNP analysis on a total of 82 genomes (57 *C. posadasii* and 25 *C. immitis*), which included newly sequenced and previously published genomes (Table 1), similar to previously described WGST analyses (Engelthaler et al. 2014; Litvintseva et al. 2014; Gillece et al. 2011). Maximum parsimony SNP trees, based on each of the SNP matrices, were constructed, using PAUP* v.4.0b10 (Swofford 2003) and visualized in Figtree v.1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/). The final all species *Coccidioides* tree was mid-point rooted and used "San Diego_1" as a reference. The final *C. immitis* tree was rooted using "Coahuila_1", based on its basal position in the all species tree, and "San Diego_1" was used as the alignment reference. The *C. posadasii* tree was rooted using "Sonora_2", again based on its basal position in the all species tree, and "B10813_Tx" was used as the alignment reference. Due to likely recombination present in the SNP data, a

neighbor joining splits tree network or "neighbor-net" tree (Bryant and Moulton 2004) was drawn to visualize the relationship between isolates each of the species trees. The neighbor-net was drawn using SplitsTree4 (Huson and Bryant 2006) with the uncorrected P distance transformation, as previously described (Engelthaler et al. 2014). *fineStructure* population analysis (Lawson et al. 2012) was implemented on the all species SNP matrix in order to infer recombination and population structure based on the presence or absence of shared genomic haplotype regions, as described elsewhere (Engelthaler et al. 2014). In brief, the SNP matrix was reduced to a pairwise similarity matrix using Chromopainter, within the *fineStructure* program, employing the linkage model, where the underlying model assumes individuals within populations will share more regions of their genome with each other and have a similar amount of admixture with individuals from different populations. Identified populations are merged one at a time, at each step using the most likely population merge, to relate populations to each other through a tree.

In order to determine the between population and within population variance we applied an "analysis of molecular variance" (AMOVA) using the GenAlEx program (Peakall and Smouse 2012). The AMOVA produces an Fst score and a ΦPT score, which is an analogue of Wright's Fst. ΦPT = 0 is considered indicative of no genetic difference among populations and ΦPT = 1 indicates 100% genetic variance. MEGA (Kumar et al. 1994) was used to calculate the average SNP distance within each identified population, (i.e., the number of base substitutions per site from averaging over all genome pairs within each population).

Data Access: All WGS data files have been deposited in the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/bioproject, project # PRJNA274372)

Acknowledgements: This research was funded in part by the National Institutes of Health (Grant#: R21AI076773), the Centers for Disease Control and Prevention (Contract#: 200201461029), and the Arizona Biomedical Research Commission (Grant#: 20080816). We thank Blanca Samayoa, Dalia Lau, Ligia Figueroa, Danicela Mercado and Brenda Guzman for assistance with sample collection and shipment.

The findings and conclusions in this article are those of the authors and do not necessarily represent the views of the CDC.

The authors declare no conflicts of interest.

Author contributions: D.M.E., E.M.D., J.M.S., B.B., and P.K. designed research; D.M.E., C.C.R., E.M.D., A.P.L., L. G., E.A., H.L., V.W., K.K., G.R.T., T. C., B.B. and P.K. performed research; D.M.E., C.C.R., A.P.L., B.B., and P.K. analyzed data; D.M.E., C.C.R., A.P.L., and P.K. wrote the paper

Figure Legends

Figure 1. Phylogenetic analysis of *C. immitis* and *C. posadasii* isolates from all known endemic regions. Maximum parsimony phylogenetic analysis was performed on WGS SNP data from 69 *Coccidioides* genomes: A) 18 *C. immitis* and B) 51 *C. posadasii*. The analysis identified 405,244 shared SNPs, with 296,632 being parsimony informative, with a consistency index (CI) of 0.412 and a Retention Index (RI) of 0.889. The tree shown is mid-point rooted. Branch lengths represent numbers of SNPs between taxa, with the unit bar in the figure.

Figure 2. SNP phylogeny from whole genomes from *C. posadasii* isolates. Maximum parsimony phylogenetic analysis was performed on 51 *C. posadasii* isolate genomes. The tree includes 253,291 SNPs (134,752 parsimony informative) (CI = 0.214 and RI = 0.364). Major phylogeographically-clustered subpopulations are shaded and labeled by region. Numbers on top of branches are SNP length with the unit bar at bottom of figure. Bootstrap values less than 90 are shown in red below branches. All other branches have bootstrap values of 90 or greater.

Figure 3. SNP phylogeny from whole genomes from *C. immitis* isolates. Maximum parsimony phylogenetic analysis was performed on 18 *C. immitis* isolate genomes. The tree includes 64,096 SNPs (31,372 parsimony informative) (CI = 0.345 and RI = 0.384). Numbers on top of branches are SNP length with the unit bar at bottom of figure.

Figure 4. finestructure analysis of *Coccidioides* species complex. *fineStructure* analysis was performed using the SNP matrix developed for Figure 1. The SNPs from 66 *Coccidioides* spp. genomes were reduced to a pairwise similarity matrix, which was used to identify population structure based on shared haplotype regions of genome. The x-axis analysis represents the strain as a "recipient" and the y-axis represents the strain as a "donor" of genomic regions. The scale bar represents the number of shared genome regions with blue being the greatest amount of sharing and yellow being the least.

Figure 5. Phylogenetic network of *C. posadasii*. A neighbor-net representation of the relationships among the 51 *C. posadasii* isolates in Figure 3 based on SNP data, using the uncorrected P distance transformation. Each band of parallel edges indicates a split. Splits of major phylogeographically clustered subpopulations are shaded according to the key.

Figure 6. finestructure analysis of *C. posadasii.* *fineStructure* analysis was performed using the *C. posadsasii*-only SNP matrix developed for Figure 2, with the inclusion of the Jujury, Argentina strain. The SNPs from the 51 *Coccidioides* genomes were reduced to a pairwise similarity matrix, which was used to identify population structure, as in Figure 4. The subpopulations identified in Figures 2 and 5 are shaded accordingly on the y-axis.

Figure 7. Phylogeography of *Coccidioides*, with an emphasis on *C. posadasii.* The *C. posadasii* subpopulation clades, defined by the maximum parsimony, analyses are shaded as in Figure 2 and placed according to the geographic origin of the isolates. The *C. immitis*

population is also shown but not separated by local Phylogeography, with the exception of the Washington subpopulation. The generalized endemic region of both species in North America is marked by hatching.

Figure 8. Model for *C. posadasii* dispersal out of Arizona-Mexico border region. A proposed dispersal model for *C. posadasii* from a hypothetical founder population in Southern Arizona-Northern Mexico: A) the Phoenix/Maricopa population originated from one of Tucson sub-populations; B) Texas and Mexico populations also originated from this region, likely concurrently from the same Southern Arizona (and possibly Northern Mexico) subpopulation, and both subsequently fed the South American populations; and C) the Guatemalan population independently emerged from the Southern Arizona region.

Figure 1. Phylogenetic analysis of *C. immitis* and *C. posadasii*

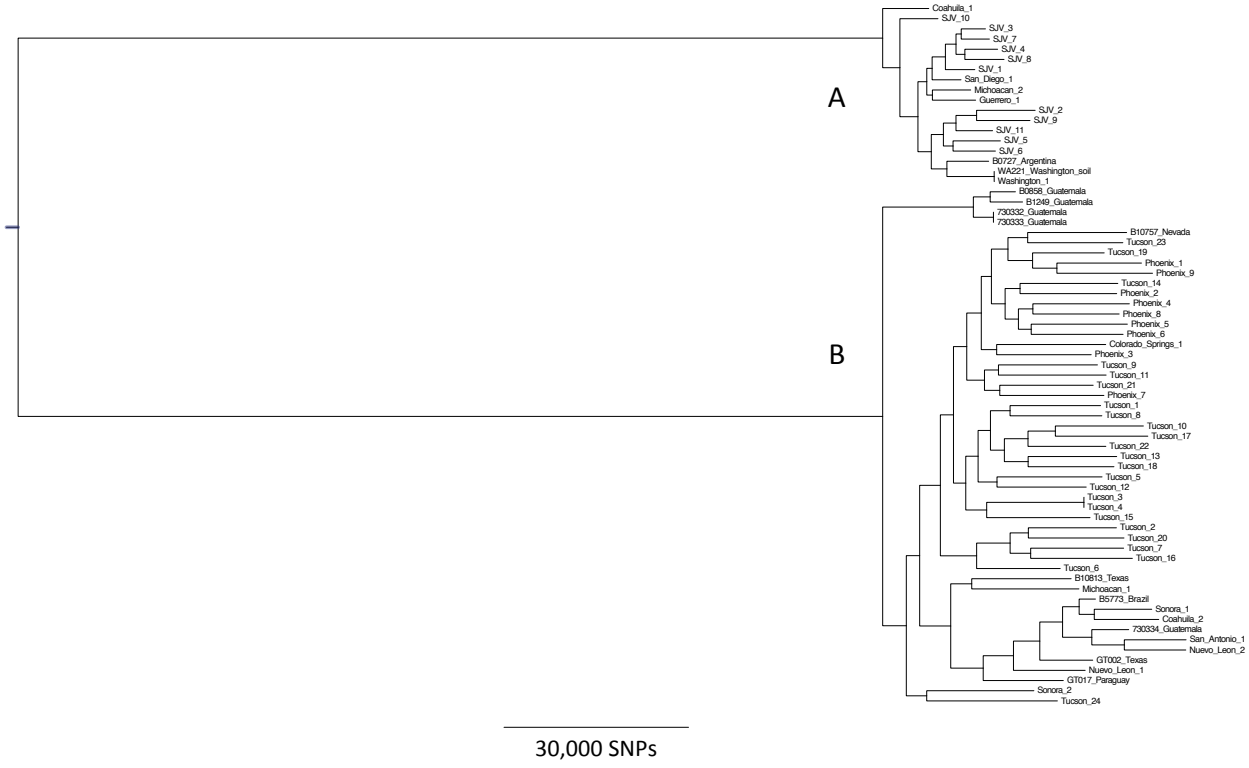

Figure 2. SNP phylogeny from whole genomes from *C. posadasii* isolates.

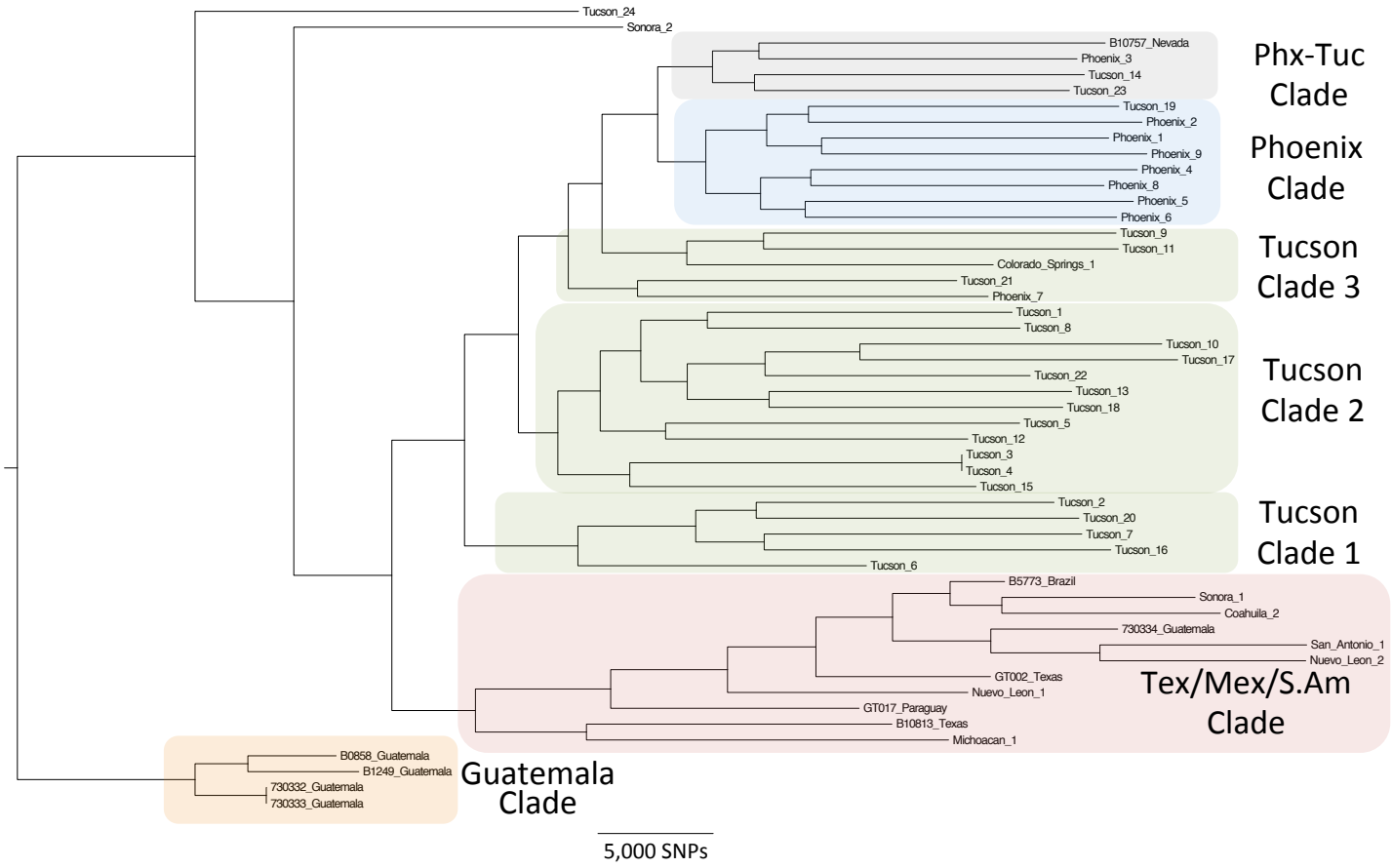

Figure 3. SNP phylogeny from whole genomes from *C. immitis* isolates.

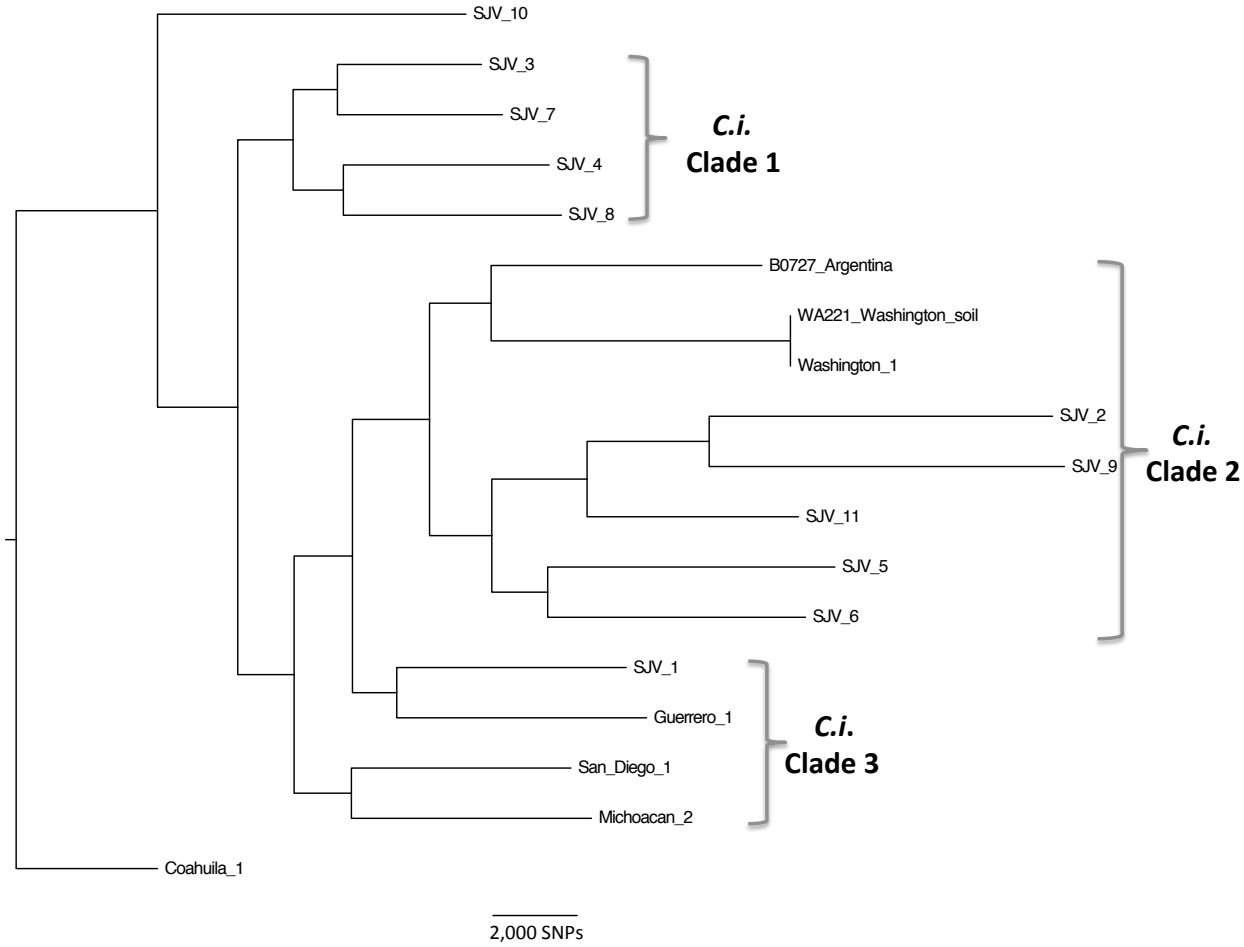

Figure 4. finestructure analysis of *C. immitis* and *C. posadasii*.

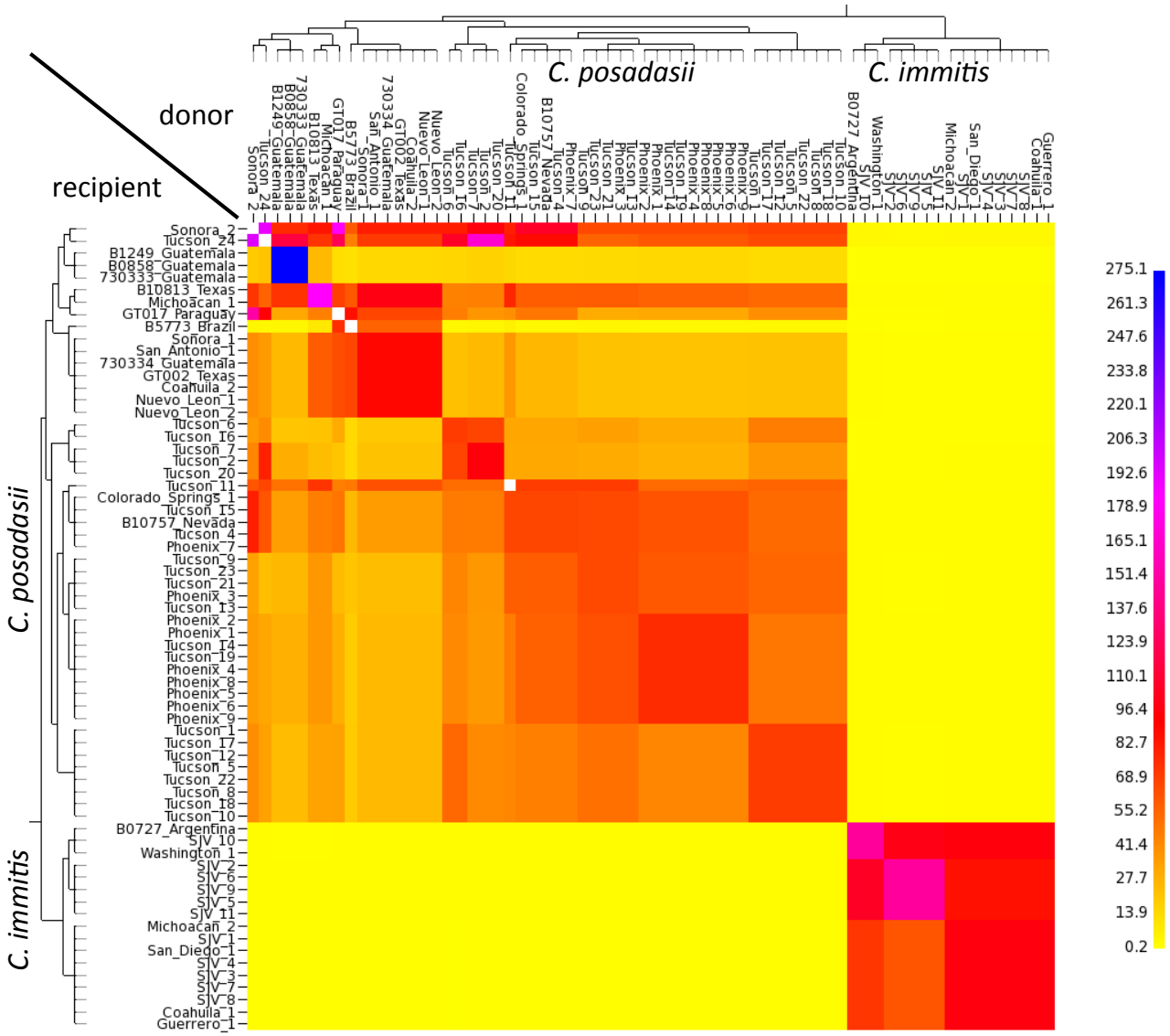

Figure 5. Phylogenetic network of *C. posadasii*.

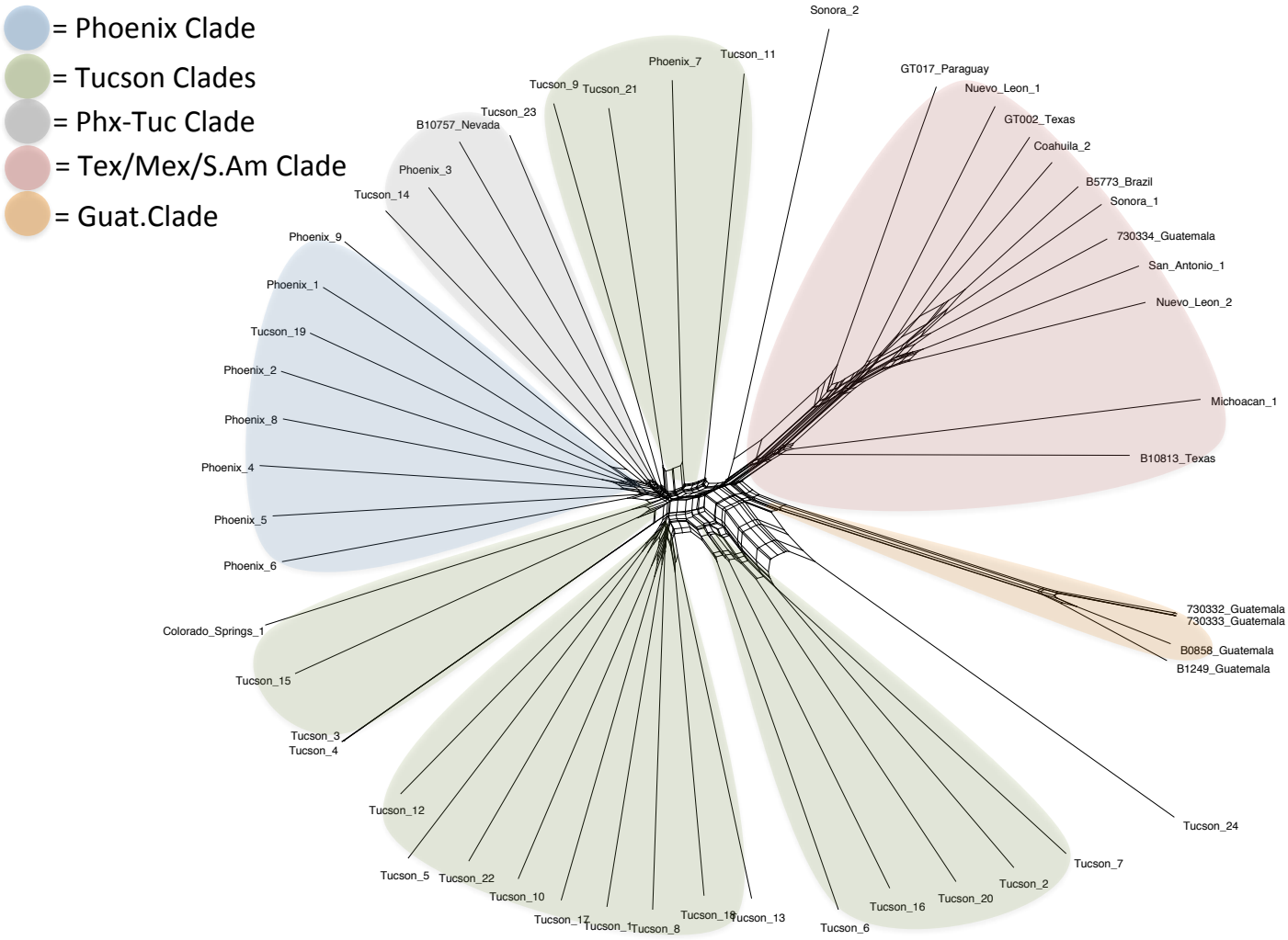

Figure 6. finestructure analysis of *C. posadasii*.

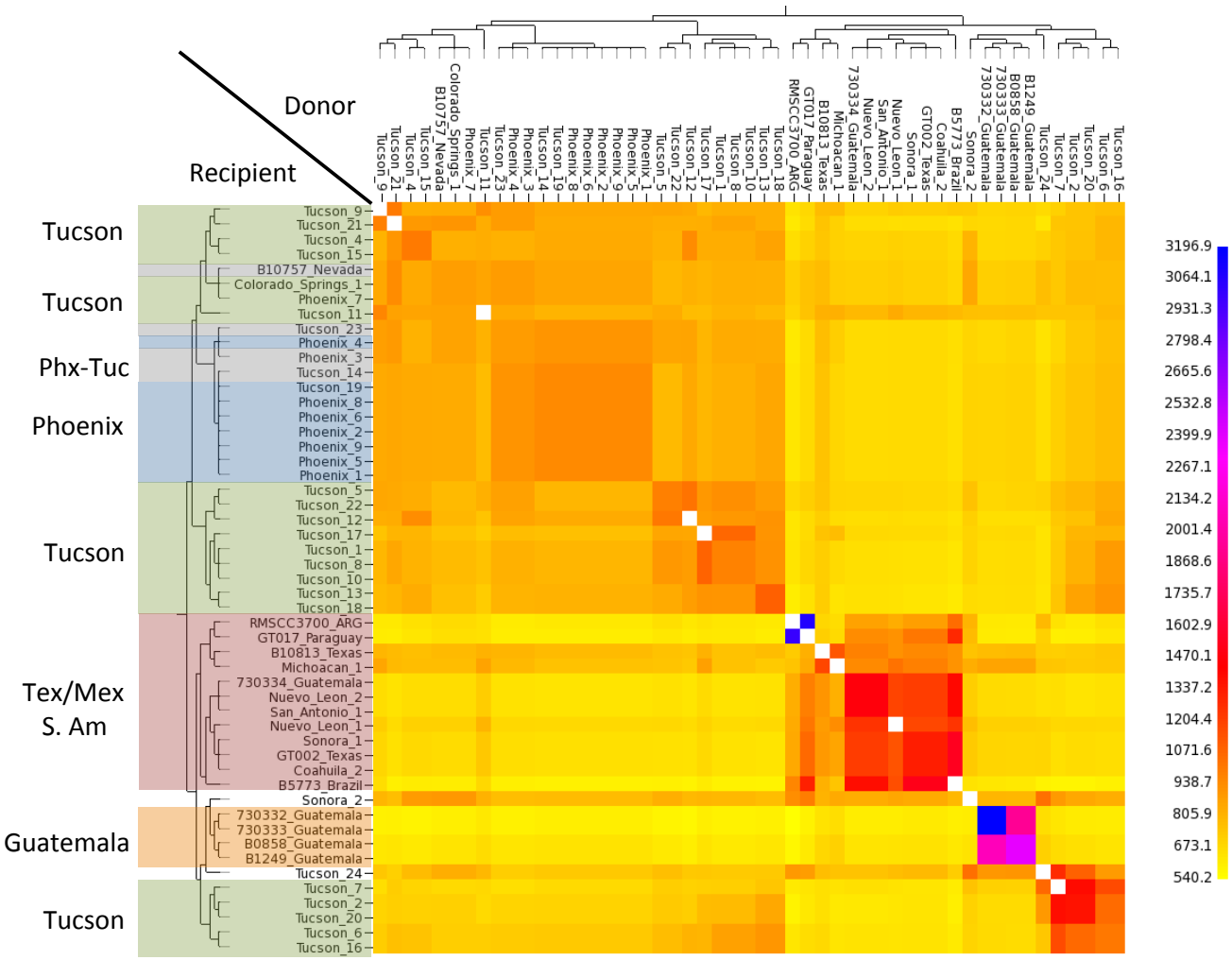

Figure 7. Phylogeography of *C. posadasii*

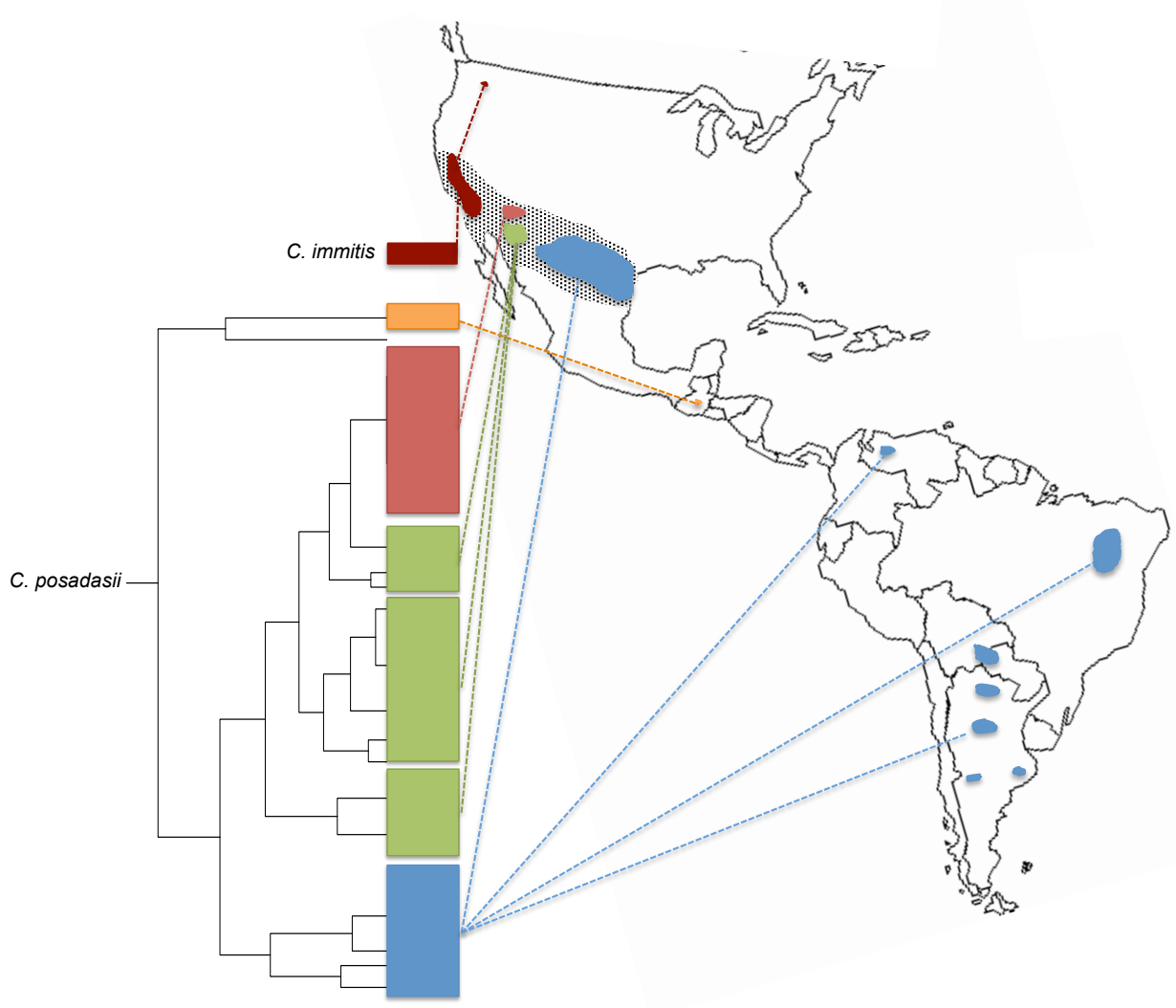

Figure 8. Proposed model of dispersal of *C. posadasii*

Table 1. Summary statistics of population analyses

Coccidioides Population	Avg. SNP Distance (SEM)	Avg. No. of Linage Specific SNPs	Population Comparison	Fst	ΦPT
C. immitis – all	0.024 (0.004)	12,149	*C.i.* vs *C.p.**	0.905	0.905
C. posadasii – all	0.062 (0.009)	4,581	Tucson vs Phoenix	0.043	0.039
C.p. – Tucson	0.159 (0.080)	13,031	Tucson vs Tex-Mex-SA*	0.206	0.210
C.p. – Phoenix	0.142 (0.063)	12,967	Tucson vs Guatemala*	0.287	0.291
C.p. – Tex-Mex-SA	0.119 (0.051)	9,140			
C.p. - Guatemala	0.037 (0.014)	4,083			

C.i. = C. immitis, *C.p. = C. posadasii*, SEM = standard error mean. * Population comparisons with significantly different SNP distances at $p<0.001$ via ANOVA.

References

1. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**: 455–477.
2. Barker BM, Jewell KA, Kroken S, Orbach MJ. 2007. The population biology of *Coccidioides*: epidemiologic implications for disease outbreaks. *Ann N Y Acad Sci* **1111**: 147-63.
3. Barker BM, Tabor JA, Shubitz LF, Perrill R, Orbach MJ. 2012. Detection and phylogenetic analysis of *Coccidioides posadasii* in Arizona soil samples. *Fungal Ecol* **5**: 163-176.
4. Brown J, Benedict K, Park BJ, Thompson GR 3rd. 2013. Coccidioidomycosis: epidemiology. *Clin Epidemiol* **5**: 185-97.
5. Bryant D, Moulton V. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol* **21**: 255-65.
6. Burt AD, Carter A, Koenig GL, White TJ, Taylor JW. 1996. Molecular markers reveal cryptic sex in the human pathogen *Coccidioides* immitis. *Proc Natl Acad Sci USA* **93**: 770–773.
7. Canteros CE, Vélez H A, Toranzo AI, Suárez-Alvarez R, Tobón O Á, Del Pilar Jimenez A M, Restrepo M Á. 2015. Molecular identification of *Coccidioides immitis* in formalin-fixed, paraffin-embedded (FFPE) tissues from a Colombian patient. *Med Mycol* **53**: 520-7.
8. Castañón-Olivares LR, Güereña-Elizalde D, González-Martínez MR, Licea-Navarro AF, González-González GM, Aroch-Calderón A. 2007. Molecular identification of *Coccidioides* isolates from Mexican patients. *Ann N Y Acad Sci* **1111**:326-35.

9. Cox RA, Magee DM. 2004. Coccidioidomycosis: host response and vaccine development. *Clin Microbiol Rev* **17**: 804-39.
10. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**: 491-498.
11. Engelthaler DM, Balajee SA. 2010. Forensics and Epidemiology of Fungal Pathogens**.** In *Microbial Forensics*, (ed. Budowle, et al.), pp. 297-313 Elsevier Press, San Diego, CA.
12. Engelthaler DM, Chiller T, Schupp JA, Colvin J, Beckstrom-Sternberg SM, Driebe EM, Moses T, Tembe W, Sinari S, Beckstrom-Sternberg JS, et al. 2011. Next-generation sequencing of *Coccidioides immitis* isolated during cluster investigation. *Emerg Infect Dis* **17**: 227-232.
13. Engelthaler DM, Hicks ND, Gillece JD, Roe CC, Schupp JM, Driebe EM, Gilgado F, Carriconde F, Trilles L, Firacative C, et al. 2014. *Cryptococcus gattii* in North American Pacific Northwest: whole population genome analysis provides insights into species evolution and dispersal. *mBio* **5**: e01464-14.
14. Etienne KA, Gillece J, Hilsabeck R, Schupp JM, Colman R, Lockhart SR, Gade L, Thompson EH, Sutton DA, Neblett-Fanfair R, et al. 2012. Whole genome sequence typing to investigate the *Apophysomyces* outbreak following a tornado in Joplin, Missouri, 2011. *PLoS One* **7**: e49989.
15. Fisher FS, Bultman MW, Johnson SM, Pappagianis D, Zaborsky E. 2007. *Coccidioides* niches and habitat parameters in the southwestern United States: a matter of scale. *Ann N Y Acad Sci* **1111**:47-72.

16. Fisher MC, Koenig GL, White TJ, San-Blas G, Negroni R, Alvarez IG, Wanke B, Taylor JW. 2001. Biogeographic range expansion into South America by *Coccidioides immitis* mirrors New World patterns of human migration. *Proc Natl Acad Sci USA* **98**: 4558–4562.
17. Fisher MC, Koenig GL, White TJ, Taylor JW. 2002a. Molecular and phenotypic description of *Coccidioides posadasii* sp nov., previously recognized as the non-California population of *Coccidioides immitis*. *Mycologia* **94**: 73–84.
18. Fisher MC, Rannala B, Chaturvedi V, Taylor JW. 2002b. Disease surveillance in recombining pathogens: Multilocus genotypes identify sources of human *Coccidioides* infections. *Proc Natl Acad Sci USA* **99**: 9067-9071.
19. Gillece JD, Schupp JM, Balajee SA, Harris J, Pearson T, Yan Y, Keim P, DeBess E, Marsden-Haug N, Wohrle R, et al. 2011. Whole genome sequence analysis of *Cryptococcus gattii* from the Pacific Northwest reveals unexpected diversity. *PLoS One* **6**: e28550.
20. Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* **23**: 254-67.
21. Jewell K, Cheshire R, Cage GD. 2008. Genetic diversity among clinical *Coccidioides spp*. isolates in Arizona. *Med Mycol* **46**: 449-55.
22. Kozarewa I, Turner DJ. 2011. 96-plex molecular barcoding for the Illumina Genome Analyzer. *Methods Mol Biol* **733**: 279-98.
23. Kumar S, Tamura K, Nei M. 1994. MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. *Comput Appl Biosci* **10**:189–191.
24. Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. Inference of population structure using dense haplotype data. *PLoS Genet* **8**: e1002453.

25. Lewis ER, Bowers JR, Barker BM. 2015. Dust devil: the life and times of the fungus that causes valley Fever. *PLoS Pathog* **11**: e1004762.

26. Litvintseva AP, Hurst S, Gade L, Frace MA, Hilsabeck R, Schupp JM, Gillece JD, Roe C, Smith D, Keim P, et al. 2014. Whole genome analysis of *Exserohilum rostratum* from the outbreak of fungal meningitis and other infections. *J Clin Microbiol* **52**: 3216-22.

27. Litvintseva AP, Marsden-Haug N, Hurst S, Hill H, Gade L, Driebe EM, Ralston C, Roe C, Barker BB, Goldoft M, et al. 2015a. Valley Fever: Finding new places for an old disease: *Coccidioides immitis* found in Washington State soil associated with recent human infection *Clin Infect Dis* **60** :e1-3.

28. Litvintseva AP, Brandt ME, Mody RK, Lockhart SR. 2015b. Investigating fungal outbreaks in the 21st Century. *PLoS Pathog* **11**: e1004804.

29. Marsden-Haug N, Hill H, Litvintseva AP, Engelthaler DM, Driebe EM, Roe CC, Ralston R, Hurst S, Goldoft M, Gade L, et al. 2014. *Coccidioides immitis* identified in soil outside of its known range — Washington, 2013. *MMWR* **63**: 450.

30. Mayorga, RP. 1967. Coccidioidomycosis in Central America. In Coccidioidomycosis. L. Ajello (ed) The University of Arizona Press. Tucson, Arizona. pp.287-291.

31. Mayorga, RP and Espinoza H. 1970. Coccidioidomycosis in Mexico and Central America. *Mycopathologia et Mycologia applicata* **40**: 13-23.

32. Neafsey DE, Barker BM, Sharpton TJ, Stajich JE, Park DJ, Whiston E, Hung CY, McMahan C, White J, Sykes S, et al. 2010. Population genomic sequencing of *Coccidioides* fungi reveals recent hybridization and transposon control. *Genome Res* **20**: 938-46.

33. Pappagianis D. 1967. Epidemiological aspects of respiratory mycotic infections. *Bacteriol Rev* **31**: 25–34.

34. Pappagianis D, Einstein H. 1978. Tempest from Tehachapi takes toll or *Coccidioides* conveyed aloft and afar. *West J Med* **129**: 527–530.
35. Peakall R, Smouse PE. 2012. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinformatics* **28**: 2537-2539.
36. Sharpton TJ, Stajich JE, Rounsley SD, Gardner MJ, Wortman JR, Jordar VS, Maiti R, Kodira CD, Neafsey DE, Zeng Q, et al. 2009. Comparative genomic analyses of the human fungal pathogens *Coccidioides* and their relatives. *Genome Res* **19**: 1722-31.
37. Sifuentes-Osornio J, Corzo-León DE, Ponce-de-León LA. 2012. Epidemiology of invasive fungal infections in Latin America. *Curr Fungal Infect Rep* **6**: 23-34.
38. Swofford DL. 2003. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods) Version 4. Sinauer Associates, Sunderland, Massachusetts.
39. Taylor JW, Geiser DM, Burt A, Koufopanou V. 1999. The evolutionary biology and population genetics underlying fungal strain typing. *Clin Microbiol Rev* **12**: 126-46.
40. Whiston E, Taylor JW. 2014. Genomics in *Coccidioides*: insights into evolution, ecology, and pathogenesis. *Med Mycol* **52**: 149-55.