RUTH ELIZABETH FUNDYGA
The Population Genetics of Pathogenic Fungi: Empirical and Theoretical Studies
(Under the Direction of JONATHAN ARNOLD)

Pathogenic fungi are important causes of morbidity and mortality in immunocompromised patients, and agricultural pathogens cause millions of dollars in crop loss annually. This dissertation describes the empirical population genetic structure of *Candida albicans* and mathematically examines the population structure of diploid fungi in general. Thirteen populations of *C. albicans* were genetically surveyed using 5 microsatellite loci. Surprisingly, the two United States populations were most divergent, but they were also from blood stream isolates. Previous studies that did not use population genetic methods found no association between source of isolate and genotype. The third chapter describes the effect of clone correction, a common data analysis procedure, on the perception of population structure. Clone correction is the process of counting a genotype once no matter how often it appears in the sampled population and is performed to avoid bias due to sampling of related individuals. Clone correction alone can drive Wright's F_{IS} away from 0, even when the initial population is randomly mating. Further, its effects on individual loci are varied but can be large. Clone corrected data should be interpreted with caution alongside the uncorrected data. The fourth chapter is a mathematical analysis of the effects of clonal, sexual, and parasexual reproduction on allele frequencies. Even when parasexual and/or clonal reproduction represents 99% of all reproduction, allelic disequilibria will decay nearly to 0 in 10 generations.

INDEX WORDS: Population genetics, Fungi, *Candida albicans*, F statistics, Clone correction, Clonal reproduction, Parasexual reproduction

THE POPULATION GENETICS OF PATHOGENIC FUNGI:

EMPIRICAL AND THEORETICAL STUDIES

by

RUTH ELIZABETH FUNDYGA

B.S., The University of Maryland, 1991

M.A., Indiana University, 1996

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2001

THE POPULATIONS GENETICS OF PATHOGENIC FUNGI: EMPIRICAL AND THEORETICAL STUDIES

BY

RUTH ELIZABETH FUNDYGA

Approved:

Major Professor: Jonathan Arnold

Committee: John Avise
Sarah Covert
Timothy Lott
Daniel Promislow

Electronic Version Approved:

Gordhan L. Patel
Dean of the Graduate School
The University of Georgia
December 2001

DEDICATION

This dissertation is dedicated to my Grandmother, Ruth Learn. When I was small, I insisted I wanted to be a veterinarian. She knew better and always called me Professor.

ACKNOWLEDGMENTS

First I would like to thank my advisor, Jonathan Arnold, and my committee, John Avise, Sarah Covert, Daniel Promislow and Timothy Lott, for allowing me to complete a dissertation that combines population genetics with public health and for their help in fulfilling my dream. It was an honor to be with the Genetics Department as a whole and the Mycology Discussion Group, and to have the opportunity to learn from them. Mary Case, Rob Dean, James Griffith, David Hall, Hemant Kelkar, Britton Lance, Linda Lee, David Logan, Steve Oliver, Maria Sanchez, and Claudia Vargas made it a pleasure to be a part of Jonathan's lab. In particular, Mary Case and Linda Lee watched over me and cared for me when I was unable to do for myself. Paul Peteet and the University Health Center made all the difference and I wouldn't have completed school without their help.

Many friends and family members have given me unending support throughout graduate school. Alyssa Moquin, Nutan Jäger, Maria Ow, Mary Stofega, Jo-Anne Fung-a-fat, and Suzy Robinson have been dear friends. Neil’s entire family has warmly welcomed me, especially Merle and Connie Rowles, Camille and JD Elliott, and Martin Kenny. My Grandmother Betty Fundyga supported me throughout graduate school with unwavering love and cookies, and my Grandfather Pete Fundyga with love and humor. My parents have been there since day one, doing everything they could to give me a good life and later helping me to also find it on my own. Finally, I would like to thank Neil Kenny for his patience, love and understanding throughout graduate school, and for spending his life with me.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS.....v

CHAPTER

1 INTRODUCTION AND LITERATURE REVIEW.....1

2 POPULATION STRUCTURE OF *CANDIDA ALBICANS*, A MEMBER OF THE HUMAN FLORA, AS DETERMINED BY MICROSATELLITE LOCI.....14

3 WRIGHT'S F STATISTICS IN CLONAL POPULATIONS 49

4 THE EFFECT OF CLONAL, SEXUAL AND PARASEXUAL REPRODUCTION ON ALLELE FREQUENCIES.....84

5 CONCLUSIONS.....100

REFERENCES.....103

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

To understand how genes for drug resistance, pathogenicity, and virulence move in a population, we have to understand how gene and genotypic diversity are created and maintained. This includes not only the biochemical pathways responsible for meiosis and cell fusion but also the phenotypic realization of these pathways in terms of mating system and population structure. Fungi have diverse and complex life cycles with the capacity to live as haploids or diploids, the potential to generate thousands of mating incompatibility groups, and the ability to migrate hundreds of kilometers on the wind (Alexopoulos, et al.,1996; Taylor et al., 1999a). They can reproduce sexually, clonally, and parasexually in nature even if the full life cycle cannot be duplicated in the lab. They are major pathogens of immunosuppressed and critically ill patients with systematic infections associated with mortalities up to twice those of bacterial septicemia (Cornwell and Belzberg, 1995; Edmond et al., 1999; Wenzel and Edmond, 2001). Agricultural pathogens cost millions of dollars annually in crop loss (Oerke et al., 1999).

Population genetics provides a method to elucidate the actual mating system of economically and medically important fungi in their natural settings. Understanding the mating system and population structure is important for developing effective control strategies and tracking epidemiologically important characters (Milgroom, 1996; Taylor, et al., 1999b). In addition to illuminating causes of variation in population structure, methods for quantitating the relative contributions of each mechanism of reproduction in nature are needed (Milgroom, 1996; Taylor et al., 1999a; Taylor et al., 1999b). This

dissertation consists of an empirical description of *Candida albicans*' population structure, a formal assessment of clone correction (a procedure for mating system analysis), and a mathematical model that examines the effects of the relative proportions of clonal, sexual and parasexual reproduction on population structure. To provide a background for this work, the introduction will review the mating system and population structure of pathogenic fungi, although the third and forth chapters of this dissertation apply to agriculturally important fungi and diploid microorganisms as well.

Mating System

The mating system of a species is a description of the types and relative proportions of reproduction that occur in a population. Much of the theory about mating systems and population structure was developed in plants (Allard, 1975). The main types of reproduction in plants are sexual outcrossing, asexual or clonal propagation, and selfing after meiosis; species that reproduce sexually and by selfing are considered mixed maters. Fungal correlates to these are sexual and clonal reproduction. Selfing in fungi commonly leads to the production of progeny identical to the parent and is considered a form of clonal reproduction (Taylor et al., 1999a). Parasexual reproduction has been observed in the laboratory, but its frequency in nature is unknown. Parasexual reproduction involves fusion of mitotic parent cells followed by chromosome loss to restore ploidy; it occurs without meiosis but does result in progeny with new genotypes (Pontecorvo, 1956; Poulter et al., 1981; Clutterbuck, 1992; Durand et al., 1993; Gu and Ko, 1998). Mitotic recombination is a final method of generating variation and can alter

the order of allelic variants within a chromosome by crossing over, change the structure of chromosomes by translocations, and introduce mutations that alter the phenotype (Aguilera, et al., 2000; Chavez, et al., 2000; Ossman, et al., 2000; Puig, et al., 2000; Rattray, et al., 2001).

The principle difference between sexual and clonal reproduction is the relationship between a gene and the organism. In clonal reproduction, the parental genotype is duplicated in the progeny, and the entire genome is transmitted as a whole across generations. In sexual reproduction, parental genotypes combine to produce a new genotype in the offspring. Sexual reproduction occurs through the meiotic production of gametes followed by fusion of gametes from genetically different parents. The particular combination of genes in any single genome is transient and occurs only in that individual.

Both types of reproduction are thought to occur in many fungi and the unique genotypes created during sexual reproduction may later be propagated clonally. The artificial group Deuteromycota (Fungi Imperfecti) contains those fungi with no known sexual stage but its members are gradually being renamed as sexual reproduction is found (Kwon-Chung, 1976; Sigler, 1996; Rosa and Lachance, 1998; Lachance et al., 2000). While the capacity for a particular type of reproduction exists broadly within a species, the mating system is the realization of that capacity and can vary over space and time. The mating system dictates what typing methods are appropriate for epidemiological studies, allowing the movement of pathogenic and resistance genes to be followed (Avise, 1994; Milgroom, 1996; Taylor, 1999a). Diagnostic fingerprints are used to track clonal organisms because all the genes associated with a particular disease are linked. Sexual organisms must be tracked by specific, single locus markers for the pathogenic

genes of interest because these genes are transmitted more or less separately within any specific lineage (Avise 1994).

In diploid organisms such as many plant pathogens, the mating system is examined by comparing the observed genotype frequencies to those expected assuming random association of alleles (Weir, 1996). For single loci the expectations are defined as the Hardy-Weinberg proportions, and for two loci deviations from expectations can be defined in terms of linkage disequilibria. Under random, sexual mating, a population is eventually in both Hardy-Weinberg and linkage equilibria, meaning the observed genotypic proportions equal those expected with random association of alleles. A natural question is whether this will occur under other modes of reproduction.

These associations cannot be measured in organisms that are encountered and cultured in their haploid phase as with many human fungal pathogens. Concordance of multiple gene genealogies, homoplasy, phylogenetic resolution, and presence of all genotype combinations for multiple loci allow assessment of the mating system (Tibayrenc, et al., 1991; Avise, 1994; Taylor et al., 1999b). Biparental recombination is indicated with high homoplasy, poor tree resolution, and the presence of all possible genotypes although these methods do not quantify the proportion of sexual and clonal reproduction. Phylogenetic reconstructions of lineages allow the detection of cryptic species so that mating system can be assessed within biologically appropriate groups (Avise, 1994; Taylor et al., 1999a; Taylor et al., 1999b). It is used as an adjunct method in both haploid and diploid species where mating system determination is problematic, as it is in many fungi.

Population Structure

The population structure of a species describes the spatial arrangement of genetic variation and depends on the mating system (Allard, 1975; Hedrick, 1985; Weir, 1996). The most basic unit of structure is the individual; several individuals are grouped into a subpopulation; and several subpopulations together form a population. Population genetics takes advantage of this hierarchical structure of individuals and examines the hierarchical structure of their genes (Wright, 1969; Weir 1996). The classic measures of structure are Wright's F statistics in which the proportions of observed and expected heterozygosity are compared within and between subpopulations (Wright, 1969). Deviations from an F statistic of 0 occur with non-random mating, asexual reproduction, and little migration between sub-populations. Another common method is phylogeographic studies that map the locations of individuals onto a geographical map to detect genetic and geographic concordance of history (Avise et al., 1987; Avise, 1994).

The distribution of genetic variation within and between populations is influenced by the amount of gene flow between populations (Wright, 1969). Without migration, populations are subject to unique forces of drift and selection and are expected to diverge as in the initial steps of allopatric speciation (Dobzhansky, 1937). As migration increases, populations become more genetically similar and only one migrant per generation is sufficient to prevent population divergence (Wright, 1969). Gene flow measurements are estimates derived from the distribution of genetic variation within and between subpopulations and are influenced by recent migration events as well as those in

the historic past (Slatkin, 1985; Slatkin, 1987). Few gene flow estimates exist for human pathogenic fungi.

The mating system and population structure are inseparable and influence each other. Clonal reproduction tends to make individuals genetically divergent as alleles and new mutations cannot move between individuals: genetically distinct lineages become established. It can lead to subpopulations becoming genetically divergent if different genotypes are amplified in different subpopulations (Fisher, 1930). Sexual reproduction reduces the associations of genes in populations (Fisher, 1930). Subpopulations can still diverge due to drift or selection, but migration between subpopulations will tend to make them more similar (Wright, 1969). Clonal expansion of one or a few genotypes is typical of the front wave of an epidemic as the pathogen sweeps through the naïve host population. Sexual reproduction is usually established during the endemic phase when the host population has a broad exposure to the pathogen (Milgroom, 1996; Taylor, et al., 1999b).

Population Genetics of Human Fungal Pathogens

The population structure of human fungal pathogens has been of great interest in the past decade as most are opportunistic infections of immunosuppressed or severely debilitated patients. The number of immunosuppressed patients has increased with the HIV epidemic, immunosuppressive treatments for cancer and organ transplants, and improved care of debilitated patients. The number of fungal infections has increased as the size of this patient group has increased (Cornwell and Belzberg, 1995; Rees et al.,

1997). Systemic fungal infections cause significant morbidity and mortality in these patients but are difficult to treat due to drug toxicity, drug interactions, and poor absorption (Cornwell and Belzberg, 1995; Dean and Buchard, 1998; Karthaus, et al., 2000; Marques et al., 2000).

The major, severe, fungal pathogens are *Aspergillus fumigatus, Candida albicans, Coccidiodes immitis, Cryptococcus neoformans,* and *Histoplasma capsulatum* (Table 1.1). *A. fumigatus, C. immitis,* and *H. capsulatum* are Ascomycetes isolated and cultured in the haploid phase. They are acquired through inhalation of spores found in contaminated environments, such as damp buildings or pigeon droppings, and typically cause pneumonia (Fisher and Cook, 1998). *C. albicans* is a diploid Ascomycete and is a member of the human normal flora. Associated with diaper rash and thrush, it can gain access to sterile sites through catheters or incisions and disseminate through the bloodstream (Cornwell and Belzberg, 1995; Dean and Burchard, 1998). *C. neoformans* is a Basidiomycete isolated and cultured as a haploid. Exposure is through inhalation of spores in the environment and may be widespread. Goldman et al. (2001) found 56% serologic conversion by age 5 years old in a cohort of inner city children and propose that some of the upper respiratory infections common in children may actually be due to this fungus.

Studies of human pathogenic fungi have concentrated on determining the mating system and relatedness of geographically distant populations (Table 1.1). The typical mating system appears to be mixed. Recombination occurs at some level because genetic and phylogenetic indicators of sex such as homoplasy, linkage equilibrium, and the presence of all recombinant genotypes are found. Evidence for clonal reproduction

includes over-representation of multilocus genotypes and concordance of gene trees. The extent to which recombination and clonal reproduction occur is unknown.

General trends in population structure are less evident, but this may reflect different habitats. For example, Fisher et al. (2001) found South American isolates of *C. immitis* were derived from a Texas population and hypothesized that this soil fungus was carried along in human migrations. *C. albicans* is transmitted between people by direct contact and shows no clear trend in geographic population structure, although there may be structure in patient populations (Pfaller et al., 1998; Xu et al., 1999; Chapter 2). Confusion about the mating system confounds analysis of population structure. Introduction of species into a new range changes the population genetic structure by decreasing variability and increasing linkage (Taylor, et al., 1999b; Hewitt, 2000), but these are also signatures of clonality (Tibayrenc et al., 1991). The complication of discovering a close relative of *C. albicans*, *C. dubliniensis*, throws into question earlier studies of population structure. Drift can give the appearance of clonality/genetic uniformity by fixing alleles over time (Burt et al., 1997). Limited polymorphism in markers used has limited the ability to detect population structure in *C. albicans* to date. Further empirical studies of population genetic structure with better markers and clear strain identification, along with theoretical analysis of the effects of reproduction on structure, are needed to understand how genes for drug resistance and virulence move in these pathogen populations.

Current Work

This dissertation consists of an empirical description of *Candida albicans*' population structure, a mathematical examination of a data analysis procedure frequently used in the analysis of microbial population structure, and a mating system model that provides a way to examine the effect of the relative proportions of clonal, parasexual, and sexual reproduction on population structure. Chapter Two is an analysis of geographically diverse isolates obtained through a mailing campaign to *Candida* researchers and hospitals. Isolates were typed to ensure they were *C. albicans*, genotyped by microsatellites and analyzed using standard population genetic analyses. Surprisingly, the two United States populations were the most different. In Europe, the most different population was from Norway and consisted of half bloodstream isolates; both U.S. populations were entirely bloodstream isolates. Previous studies that have not used population genetic methods have found no relationship between genotype and isolate site or patient status. This study is consistent with another (Pfaller et al., 1998) and suggests that cryptic, bloodstream populations may exist and that undefined transmission routes within these patient groups may exist.

The third chapter is a mathematical examination of clone correction, a data analysis procedure in which a genotype is counted once in analysis regardless of the number of times it is sampled (Chen et al., 1994). This is done in population studies of agriculturally important plant pathogens, which are usually diploid, as well as *C. albicans* studies to avoid biasing results by the inclusion of relatives. The effects of this method on perception of structure have not been formally examined but needs to be so that

unknown effects of this method do not confound population analyses. This chapter presents an analysis of clone correction for one locus with multiple alleles, the effects of clone correction on mating system and population structure, and provides an example as applied to the data from the first chapter. Clone correction has a significant effect on the perception of the mating system and structure, giving the appearance of non-random mating even when only sexual reproduction is assumed and increasing subpopulation similarity.

The fourth chapter is a mathematical extension of the Hardy-Weinberg Law to examine the effects of different proportions of sexual, clonal and parasexual reproduction on allele frequencies. This was specifically developed with fungi in mind, although by setting parasexual reproduction to zero it applies to any diploid organism with sexual and clonal reproduction.

Table 1.1 The population structure of human pathogenic fungi.

Fungus	Population	Recombination	Clonality	Population Structure	Markers	Reference
A. fumigatus	63 isolates, 3 European hospitals		F_{IS} = 0.66, significant linkage disequilibrium	F_{ST} = 0.24, no relation between geographic and genetic distance	12 isozyme loci	Rodriguez et al., 1996
C. albicans	52 isolates Duke Univ.	less linkage disequilibrium than expected for clonal	Hardy-Weinberg deviations, overrepresented genotypes		12 single base mutations in 8 PCR fragments	Gräser et al., 1996
C. albicans	126 isolates, Duke Univ. and Brazil	Hardy-Weinberg equilibrium in about half of the loci	homozygote excess in about half of the loci	F_{ST} = 0.017, clinical isolates form a separate group from commensal isolates	16 RFLP on 9 PCR fragments	Xu et al., 1999
C. albicans	global	less linkage disequil an expected for clonal	Hardy nberg deviat., absence of segregation types	Africa less variable than Europe, USA	85 SSCP loci in 13 PCR fragments	Schonian et al., 2000

C. albicans	50 isolates		12/16 informative loci deviate from Hardy-Weinberg		21 loci, various	Lott and Effat, 2001
C. immitis	30 isolates from 25 patients Tucson, AZ	homoplasy, all recombinant types found at 6 most informative loci			12 SSCP loci	Burt et al., 1996
C. immitis	161 isolates, 8 N and S American locations			Genetic distance, local differentiation in N America, S America colonized from Texas	9 microsatellite loci	Fisher et al., 2001
C. immitis	37 isolates Bakersfield, CA epidemic	4 of 78 pairs significant linkage disequilibria, consistency indices low (0.44) in phylogenies indicating homoplasy as would be expected with recombination, index of association p = 0.28	expected 2, observed 3 identical genotypes by chance	Epidemic derived from nearly panmictic population	SNP, 2 STR, SSCP same as for Burt 1996 PNAS	Fisher et al., 2000

C. neoformans	36 clinical, 15 environmental isolates, Brazil		short, well resolved phylogeny; correlation between markers	New York isolates group with those from Brazil	CNRE-1 Southerns, RFLP and sequencing of *URA5*, phylogeny	Franzot et al., 1997
C. neoformans	34 isolates, various locations	incongruent gene genealogies indicate hybridization		no phylogeographic pattern	sequenced 4 genes	Xu et al., 2000
H. capsulatum	30 clinical isolates, Indianapolis , IN	Index of Association indicates recombination; poorly resolved tree			11 biallelic markers based on RAPD primers	Carter et al., 1996

CHAPTER 2.

POPULATION STRUCTURE OF *CANDIDA ALBICANS*, A MEMBER OF THE HUMAN FLORA, AS DETERMINED BY MICROSATELLITE LOCI[1]

[1] Fundyga, R.E., T.J. Lott and J. Arnold. 2001. To be submitted to Molecular Microbiology

ABSTRACT

This study examines the macrogeographic population structure of *Candida albicans*, a yeast commensal of humans, through a population genetic analysis of five microsatellite loci in thirteen cities. Observed heterozygosity ranged from 0.51 to 0.74, the populations were in Hardy-Weinberg equilibrium for 1.3 of the 5 loci on average, and in linkage equilibrium for 7.8 of 10 pairwise associations on average. About 5% to 15% of the genetic variation is between populations. The overall pattern is one of intermediate gene flow and differentiation, with subdivisions associated with bloodstream isolates but not HIV infection in these populations. We did not find a single widespread genotype but instead found high, macrogeographic gene flow in these clinical populations; the most common genotype was limited to Atlanta and San Francisco. Homogeneity is evident within large geographic regions, such as Europe, Asia, and the United States and indicates that overall gene flow for a member of the human flora is variable but can be extensive, with an average of 5.6 migrants per generation (Nm).

INTRODUCTION

Candida albicans is a ubiquitous, commensal yeast that occurs on the mucosal epithelia of humans and is a member of the normal flora (Odds, 1988). It is a common opportunistic infection of immunocompromised patients and is a model for pathogenic fungi (Odds 1988). Basic knowledge of population processes is needed to understand how the organism and its genes move, including genes for drug resistance, and whether distinct clinical subpopulations exist. Epidemiological studies often use DNA

fingerprinting to track outbreaks (Soll 2000) but high resolution, single locus markers are needed for population analyses (Avise 1994; Soll 2000).

We used microsatellites to describe the population structure of *C. albicans* (Ascomycota; Saccharomycetales). Haploid states have not yet been observed, but it encodes a mating type-like locus (*MTL*) whose genes complement *Saccharomyces cerevisiae* homologues (Hull 1999). When the *MTL* locus is made haploid, the otherwise diploid cells fuse and form tetraploids (Hull et al. 2000; Magee and Magee 2000). Single population studies of the mating system indicate predominantly clonal reproduction with a small amount of recombination, as some loci were in Hardy-Weinberg and linkage equilibria (Pujol et al. 1993; Gräser et al. 1996; Xu et al. 1999b). These studies likely provide minimum estimates of recombination due to Wählund effects (Tibayrenc et al. 1991): the host travels, and isolates are often only available from hospitals, which may attract members of several populations.

Most studies of *C. albicans* population structure assessed geographic diversity and relatedness of patient isolates through phylogenetic inference and have found clustering of genotypes over broad geographic regions with widespread serotypes or genotypes (Pujol et al. 1993; Clemons et al. 1997; Pujol et al. 1997; Xu et al. 1999c). Marker choice and geographical sampling scale influence perceived population structure, as markers with little variation will present populations as more homogeneous while more variable markers will enhance discrimination (Avise 1994; Taylor et al. 1999; Soll 2000).

Microsatellites were used in this study because of their variability and have been successfully used to assess plant, animal and human population structure (Devlin et al.

1990; Awadalla and Ritland 1997; Anderson et al. 1998; Becher and Griffiths 1998; Goodman et al. 1998; Oliveira et al. 1998; Piertney et al. 1998; Ross et al. 1999). Microsatellites are increasingly used in epidemiological investigations for their ability to differentiate genotypes in organisms with a high degree of clonal reproduction (Soll 2000). One concern with microsatellites is that information may be lost through marker mutation as microbes have short generation times, but the loci used in this study were stable during 300 generations (25 days) in the lab (unpublished data, Bretagne et al. 1997).

The goals of this study were to quantitate genetic differentiation, population structure, and gene flow in this microbial system. We compiled multilocus genotypes from single locus microsatellites, estimated F statistics, estimated gene flow as Nm, and tested for isolation by distance. Such analyses will assist in describing movement of clinically significant genes and identifying unique populations (Taylor et al. 1999).

MATERIALS AND METHODS

Sampling and Identification

Scientists in 13 cities (Table 2.1) kindly sent us clinical isolates, and we used one isolate per patient. The 294 isolates used were primarily obtained from oral sites, with mixed oral and mucosal isolates from Austria, Barbados, Kuwait and South Africa. All of the Atlanta and San Francisco isolates were from blood, as were 15 of those from Norway and 3 from India. HIV status of all patients was not available, but all of the

isolates from Germany, half of those from South Africa and England, and 2 from Brazil were from HIV positive patients. The Norway and United States isolates were from HIV negative patients, although the patients with bloodstream infections were immunocompromised.

We screened isolates to ensure they were all typical *C. albicans* using three phenotypic methods: green growth on CHROMagar Candida (CHROMagar Company, distributed by Hardy Diagnostics), the germ tube test (Odds 1988), and growth at 45°C to differentiate *C. albicans* from *C. dubliniensis* (Pinjon et al. 1998). Cultures were grown on yeast peptone dextrose (YPD) agar, suspended in phosphate buffered saline (PBS) with 15% glycerol and frozen at –80°C. Each culture originated from a single colony and we performed all manipulations on clonal descendants of this one colony.

DNA Extraction

DNA was extracted using a small-scale modification of a standard lyticase protocol (Mannarelli and Kurtzman 1998). Cultures for DNA extraction were newly grown from the frozen stocks and were the third culturing since we received the isolates. Five ml YPD broth was inoculated with a frozen stock and grown overnight at 37°C, 200RPM. Cells were pelleted by centrifugation, washed with 300 µl PBS (phosphate buffered saline), repelleted, resuspended in 300 µl PBS, and transferred to microfuge tubes. To weaken the tough cell wall, we added 5 µl of a 0.5 µg lyticase / µl PBS solution and incubated the cells at 37°C for about 45 to 60 min.

The resulting spheroplasts were gently pelleted and resuspended in 270µl TE pH 7.6 and 30 µl NDS (a detergent). The spheroplasts were physically lysed by adding a 200 µl volume of 300 µm glass beads and vortexing for 90 sec with a Vortex-Genie Turbo Mix attachment. To this suspension, 3µl of 10 mg/ml RNAse A was added and the samples incubated for 1 hour at 37°C. DNA was purified by a single phenol:chloroform:isoamyl alcohol extraction (25:25:1) followed by cold ethanol precipitation. DNA was resuspended in 50 µl TE pH 7.6 and stored at -20°C. A 1:10 dilution in TE pH 7.6 was used as a working stock.

Microsatellite Analysis

We assessed genotypes at five microsatellite loci: CEF3 (translation elongation factor, Bretagne et al. 1997), ERK1 (extracellular-signal-regulated kinase), KRE6 (beta-1,6-glucan synthesis), ZNF1 (zinc finger transcription factor, Field et al. 1996) and an unidentified locus previously named Locus 4 (Lott et al. 1999). From the *C. albicans* sequencing project at Stanford University, CEF3 is on chromosome 5, ERK1 and ZNF1 are on chromosome 4, and the locations of KRE6 and Locus 4 are unknown. PCR primer pairs (Table 2.2) were synthesized based on previously published sequences and the reverse primer of each pair was fluorescently labeled: ERK1, ZNF1 and Locus 4 with FAM; KRE6 with HEX; and CEF3 with TET (Applied Biosystems, Inc.). All loci were PCR amplified in 25µl reactions composed of 1X buffer with Mg+ (Roche-Boerhinger Mannheim), 0.2 µM each primer, 200 µM dNTPs, 1 unit Taq polymerase (Roche-Boerhinger Mannheim) and 1ul DNA; locus 4 reactions also contained 5% DMSO.

Cycling parameters were: 95°C denaturing for 5 min; 35 cycles of 95°C denaturing for 1 min, 55°C annealing for 1 min and 72°C extension for 1 min; and a final extension at 72°C for 5 min. The CEF3 locus required an annealing temperature of 50°C.

We sized PCR products on a capillary DNA sequencer (ABI 310) using the polyacrylamide matrix Pop 4 and the size standards TAMRA 350 and 500 according to manufacturers directions (Applied Biosystems Inc.). The use of this automated, high resolution sequencer reduced spurious allele calling: binning was reduced by its single nucleotide discriminatory ability, and splitting was reduced by incorporating the same positive marker control into each sequencer run to control for sequencer drift. For each locus, we standardized allele sizes between sequencer runs by adding to or subtracting from each allele the size differences of the positive control, typically less than one nucleotide (data not shown).

Statistical Analysis

We examined the mating system, described how genetic variation was partitioned within and between populations, estimated gene flow and described the overall population structure of these *C. albicans* isolates. The average heterozygosity at each locus, over all demes, is the fraction of genotypes with two different alleles at a single locus and was calculated as a basic measure of genetic variation (Weir 1996). We assessed the mating system by examining allelic associations within single populations. Disequilibrium measures were used to compare the observed genotypic frequencies with those expected from the allele frequencies. For each locus, Hardy-Weinberg disequilibria

were estimated using exact tests in the program GENEPOP 3.1d (Raymond and Rousset 1995). Exact tests are more appropriate than chi-squared tests for small sample sizes as the latter use the data to estimate a large sample approximation to an exact test (Weir 1996). Associations between alleles at two loci were measured as genotypic linkage disequilibria, also using GENEPOP. Recombination is indicated when observed frequencies are consistent with the expectations, producing no disequilibria.

We calculated Wright's F statistics (Wright 1969) in GENEPOP to assess how genetic variation was partitioned within and between populations. Within a single population, F_{IS} measures genetic variation due to differences between individuals and ranges from –1 to +1. F_{IS} detects deviations from random mating with negative values indicating heterozygote excess, as with clonality, and positive values indicating heterozygote deficiency, as with inbreeding. For the metapopulation, F_{ST} measures the amount of genetic variation in the metapopulation that is due to differences between subpopulations and ranges from 0 to positive infinity. A positive F_{ST} implies there are more genetic differences between subpopulations than within them and a heterogeneous population structure exists. An F_{IS} of zero indicates random mating within subpopulations and an F_{ST} of zero indicates homogeneous metapopulation structure. The average number of individuals migrating per generation between a pair of populations, Nm, was estimated using the private alleles method in GENEPOP. It measures whether gene flow is low or high: an Nm greater than 1 is sufficient to prevent populations from becoming fixed for alternate alleles and gene flow increases as Nm increases (Wright 1969). A variation of F_{ST} for microsatellites, designated $Rho_{ST,}$ and Nm were estimated using RSTCALC (Goodman 1997) before clone correction. The

hypothesis of genetic isolation with increasing geographic distance between populations was examined by a Mantel test; the test was executed in GENEPOP using Rho_{ST} values calculated in RSTCALC. All statistics were evaluated at $\alpha = 0.05$ except for the linkage disequilibria, which were evaluated at $\alpha = 0.005$ (0.05/10 comparisons) due to multiple tests on the same data.

RESULTS

Markers and Genotypes

Out of 294 isolates, we observed 20 clonal and 246 unique multilocus genotypes (Table 2.1). The number of isolates per clone ranged from 2 to 13, with 12 clones represented by 2 isolates each and 18 clones by 5 or fewer isolates. The most common genotype is composed of 13 isolates and is limited to San Francisco and Atlanta; seven genotypes differ from this most common genotype by one repeat and 16 differ by one allele. The second most common genotype is composed of 8 isolates, with one other isolate differing by one repeat and three differing by one allele. Sixteen clonal genotypes were found in more than one population and four of these genotypes were found in both eastern and western Hemispheres. Allele frequencies differed significantly by population, generally indicating that the metapopulation was heterogeneous (data not shown).

The microsatellites with simple triplet repeats, KRE6, Locus 4 and ZNF1, had between 9 and 13 alleles in the global population (Table 2.2). The locus CEF3, with a

repeat containing either 2 or 3 T's, had 27 alleles in the global population, and the locus ERK1 had 46 alleles, which was expected given its clustered (closely linked but non-contiguous) repeat motif (Table 2.2). Each locus had 3 or 4 fragment sizes that together accounted for over half of the alleles sampled at that locus. Observed heterozygosity per locus ranged from 0.51 at Locus 4 to 0.74 at CEF3 (Table 2.2), indicating that at least half of the population was heterozygous.

Allelic Associations

First we examined associations within loci to determine if genotypic frequencies were those expected under recombination. Five of the populations met Hardy-Weinberg expectations for KRE6, eight for Locus 4, two populations for ZNF1, one for CEF3 and none for ERK1 (Table 2.3). This result was expected for CEF3 and ERK1 given the large number of alleles and small sample sizes. Of the 65 tests, 48 deviated from expectations while three deviations were expected by chance.

Next we examined associations between locus pairs to determine if genotypic frequencies were those expected under recombination. All locus pairs were in linkage equilibrium in at least 10 of the 13 populations (Table 2.4), except for ERK1-ZNF1 which were in equilibrium in 9 populations and KRE6-Locus4, which were in equilibrium in 6 populations. These results were expected because of the large number of alleles at CEF3 and ERK1. Of the 130 tests, 29 deviated from expectations while one deviation was expected by chance.

Population Structure

We estimated how genetic variation was partitioned within and between populations using F statistics and also Rho_{ST}, a microsatellite specific estimator of between population variation. Averaged over all loci, F_{IS} was 0.132, with the positive sign indicating an excess of homozygotes (Table 2.5). Over all loci, F_{ST} was 0.052, meaning about 5% of the genetic variation was due to differences between populations (Table 2.5). The Rho_{ST} values were higher and indicated about 15% of genetic variation was due to differences between populations.

The Würzburg, Germany and Vienna, Austria populations were the least differentiated and statistically different only from the Barbados and United States populations (Table 2.6). Atlanta and San Francisco, United States, were the two most differentiated populations and were similar only to each other. The Barbados population was the next most divergent population overall behind the two U.S. populations, followed by Piracicaba, Brazil and Tromsø, Norway which were each different from seven other populations. Europe forms a homogeneous group except for Tromsø, Norway-Linz, Austria, (Table 2.6); Norway is the most differentiated European population. Asia and Africa form a homogeneous group except for Chandigarh, India- Johannesburg, South Africa; South Africa is the most differentiated Asian and African population.

Apart from Norway, there are only two significantly different population pairs between the Asian and African populations and those of Europe; in contrast, there is only one homogeneous pair, India-Brazil, between the Asian and African populations and those of the Western Hemisphere. In contrast to the Eastern Hemisphere, the Western

Hemisphere populations are divergent except for Atlanta-San Francisco. There are eight pairwise differences between Eurasian and African populations with five of these due to a comparison including Tromsø, Norway. There are 29 significant differences between the Eastern and Western Hemispheres, with 18 of these 29 from a United States comparison.

Migration

The overall Nm estimated by private alleles was 5.6 individuals per generation. This is an intermediate value and suggests that gene flow is occurring, but at an intermediate rate that still allows partial population divergence. Pairwise estimates of gene flow were calculated from Rho_{ST} and ranged from a low of 0.73 migrants per generation for Barbados - South Africa to unmeasurably high between six Eurasian pairs and also Atlanta - San Francisco (Table 2.6).

Given that migration occurs, we wished to determine if it followed the simple pattern of isolation by distance, in which increasing geographic distance between populations reduces the amount of gene flow between them. Isolation by distance was tested for by plotting Rho_{ST} against geographic distance, with Pearson's r = 0.54 (Fig 2.1). The Mantel test for isolation by distance was significant and accounted for 29% of the genetic variation between populations.

DISCUSSION

Genotypes

The observed heterozygosities of 0.38 to 0.74 found with these markers are similar to heterozygosities found in a previous microsatellite study of *C. albicans* (Field et al. 1996) and are less than those observed in animal microsatellites, which are typically around 0.8 (Andersen et al. 1998; Goodman et al. 1998; Piertney et al. 1998). In contrast to a PCR-RFLP study of *C. albicans* populations with heterozygosities of only 0.08 to 0.16, our study found more genotypes and each clonal genotype was composed of fewer isolates (Xu et al. 1999b). This is expected given that microsatellites are more variable than PCR-RFLPs, but it may also be due to sampling different populations.

Allelic Associations

Associations within and between loci provide different perspectives on the mating system and in this study are consistent with mixed mating. About 80% of the locus pairs were in linkage equilibrium, which is greater than expected for a strictly clonal organism (Whittam et al. 1983). Other studies have shown no disequilibria between marker pairs, even those on the same chromosome (Tibayrenc et al. 1991; Caugant and Sandven 1993; Gräser et al. 1996). Genetic drift may create linkage disequilibrium in subdivided populations with low recombination (Whittam et al. 1983), but drift is unlikely to be important here as each population experiences high gene flow with at least one other population. Tests of linkage disequilibrium compare observed genotypic frequencies

with those expected. With several alleles at each locus, it is possible some observed and expected values are not significantly different because of small sample size and low power. However, disequilibria were calculated within each subpopulation, reducing the number of alleles in the calculations (data not shown) and for KRE6-ZNF1, the loci with the fewest alleles, 11 populations were in linkage equilibrium (Table 2.4).

In contrast are the Hardy-Weinberg disequilibria in about 75% of the tests (Table 2.3) and the positive F_{IS} of 0.13 (Table 2.5) which both indicate substantial deviations from random mating. This study along with others have found about 1/4 to 1/2 of markers in Hardy-Weinberg equilibria and most locus pairs in linkage equilibria (Gräser et al. 1996; Xu et al.1999b; Lott and Effat 2001), which is consistent with predictions that Hardy-Weinberg equilibria decay faster than linkage equilibria (Weir 1996). A positive F_{IS} is usually associated with inbreeding or selfing, not clonality (Tibayrenc 1996), as it results from a heterozygote deficit. Gene conversion in mitotic recombination would reduce heterozygosity (Krafzig et al. 1993) but is also expected to lead to the rapid decay in linkage equilibria, as they are measures of statistical association. We conclude that the populations of *C. albicans* examined here are predominantly clonal with some recombination (Hull et al. 2000; Magee and Magee 2000).

Population Structure

We did not find a single widespread genotype but instead found high, macrogeographic gene flow (Table 2.6) in these clinical populations; the most common genotype was limited to Atlanta and San Francisco. Schmid et al. (1999)

phylogeographically analyzed fingerprints of patient isolates from diverse regions and found about one third of all isolates clustered together, which is consistent with widespread gene flow. Clemons et al. (1997) found European and U.S. mucosal isolates to be more similar to each other than those from Singapore. Xu et al. (1999c) found isolates from HIV populations in Brazil and the U.S. to be similar, and for all patient populations to group separately from healthy populations. The German population in this study came from HIV positive patients and was one of two least differentiated populations. Many studies have looked for host-isolate associations within patient populations, but most have found no relationship between isolate genotype and body site or patient health status (Caugant and Sandven 1993; Boerlin et al. 1996; Xu et al. 1999a; Schmid et al. 1999).

It is surprising that 27 of the 42 significant differences between population pairs would be from comparisons involving Tromsø, Atlanta, and San Francisco (Table 2.6). Tromsø is a small Norwegian city inside the Arctic Circle while Atlanta and San Francisco are large U.S. cities with substantial immigrant populations, but half of the Tromsø and all of the U.S. isolates are from blood. There is little gene flow between Tromsø and the U.S. populations (Table 2.6). Pfaller et al. (1998) found that some bloodstream isolate populations in the United States exhibit both hospital and regional specificity, with some hospitals exhibiting an endemic type different from that of the community epithelial population. This study found little gene flow between bloodstream and mucosal populations in different locations, with the U.S. populations the most differentiated overall, and Tromsø the most differentiated in the Eastern Hemisphere. In contrast, this study found high gene flow between Atlanta and San Francisco, suggesting

a homogeneous population including these large cities or the Veterans' Administration hospitals from which the isolates came.

Isolation by distance accounts for 29% of the observed genetic variation between populations, leaving 71% unexplained. This is likely an upper limit of variation accounted for as the two U.S. populations and Norway are both divergent because they contain bloodstream isolates and are also geographically distant from other populations. A more thorough sampling of both mucosal and bloodstream isolates from the same geographic locations would allow isolation by distance to be distinguished from isolation by body site. A more appropriate measure than strict geographic distance may incorporate opportunity for travel, such as flights per day between cities or the number of border crossings by car or train. Most of the populations are from large cities with airports and a more differentiated pattern is likely to emerge from an examination of small, isolated towns.

CONCLUSIONS

The overall pattern is one of intermediate gene flow and differentiation, with metapopulation subdivisions associated with bloodstream isolates but not HIV infection in these populations (Figure 2, Table 2.6). Europe forms a homogeneous group except for Tromsø, Norway-Linz, Austria and estimated gene flow between population pairs is high (Table 2.6). Asia and Africa form a homogeneous group except for Chandigarh, India- Johannesburg, South Africa. Pairwise migration rates are lower than within Europe but still high enough to prevent divergence. There is more genetic variation

between Western Hemisphere populations than within them and migration rates are low, except for Atlanta –San Francisco. Homogeneity is evident within large geographic regions, such as Europe, Asia, and the United States and indicates that overall gene flow for a commensal transferred through direct contact is high.

ACKNOWLEDGMENTS

We thank the following persons for their generous gifts of isolates: O.P de Almeida, Arunaloke Chakrabarti, Maeve Coogan, Z.U. Khan, Paul Levett, Axel Lischewski, M. Louw, Peter Müller-Uri, Julian Naglik, Elisabeth Presterl, Lars Vorland and Xu Yanying. This study would not have been possible without their assistance. For their comments, we thank John Avise, Sarah Covert, Paul Peteet, and Daniel Promislow. Brian Holloway and Nikhat Sullaiman, of the Centers for Disease Control Biotechnology Facility, provided assistance with the ABI310. REF was supported by a University of Georgia Fellowship and a Training Grant in Molecular and Cellular Mycology (T32-AI-07373) from the National Institutes of Health. JA was supported by grant MCB-9630910 from the National Science Foundation.

REFERENCES

Anderson LW, Born EW, Gjertz I, Wiig Ø, Holm LE, Bendixen C (1998) Population structure and gene flow in the eastern Atlantic walrus (*Odobenus rosmarus rosmarus*) in the eastern Atlantic Arctic based on mitochondrial DNA and microsatellite variation. *Molecular Ecology*, 7, 1323-1336.

Avise JC (1994) *Molecular markers, natural history and evolution*. Chapman and Hall, New York.

Awadalla P, Ritland K (1997) Microsatellite variation and evolution in the *Mimulus guttatus* species complex with contrasting mating systems. *Molecular Biology and Evolution*, 14, 1023-1034.

Becher SA, Griffiths R (1998) Genetic differentiation among local populations of the European hedgehog (*Erinaceus europaeus)* in mosaic habitats. *Molecular Ecology*, 7, 1599-1604.

Boerlin P, Boerlin-Petzold F, Goudet J, Durussel C, Pagani JL, Chave JP, Bille J (1996) Typing *Candida albicans* oral isolates from Human Immunodeficiency Virus-infected patients by multilocus enzyme electrophoresis and DNA fingerprinting. *Journal of Clinical Microbiology*, 34, 1235-1248.

Bretagne S, Costa JM, Besmond C, Carsique R, Calderone R (1997) Microsatellite polymorphism in the promoter sequence of the elongation factor 3 gene of *Candida albicans* as the basis for a typing system. *Journal of Clinical Microbiology*, 35, 1777-1780.

Caugant DA, Sandven P (1993) Epidemiological analysis of *Candida albicans* strains by multilocus enzyme electrophoresis. *Journal of Clinical Microbiology*, 31, 215-220.

Chen RS, Boeger JM, McDonald BA (1994) Genetic stability in a population of a plant pathogenic fungus over time. *Molecular Ecology*, 3, 209-218.

Clemons KV, Feroze F, Holmberg K, Stevens DA (1997) Comparative analysis of genetic variability among *Candida albicans* isolates from different geographic locales by three genotypic methods. *Journal of Clinical Microbiology*, 35, 1332-1336.

Devlin B, Risch N, Roeder K (1990) No excess of homozygosity at loci used for DNA fingerprinting. *Science*, 249, 1416-1420.

Field D, Eggert L, Metzgar D, Rose R, Wills C (1996) Use of polymorphic short and clustered coding-region microsatellites to distinguish strains of *Candida albicans. FEMS Immunology and Medical Microbiology*, 15, 73-79.

Goodman SJ (1997) R_{ST}Calc: a collection of computer programs for calculating estimates of genetic differentiation from microsatellite data and determining their significance. *Molecular Ecology*, 6, 881-885.

Goodman S (1998) Patterns of extensive genetic differentiation and variation among European harbor seals (*Phoca vitulina vitulina)* revealed using microsatellite DNA polymorphisms. *Molecular Biology and Evolution*, 15, 104-118.

Gräser Y, Volovsek M, Arrington J, Schönian G, Presber W, Mitchell TG, Vilgalys R (1996) Molecular markers reveal that the population structure of the human pathogen *Candida albicans* exhibits both clonality and recombination. *Proceedings of the National Academy of Sciences of the USA*, 93, 12473-12477.

Hull C, Johnson A (1999) Identification of a mating type-like locus in the asexual pathogenic yeast *Candida albicans*. *Science*, 285, 1271-1275.

Hull CM, Raisner RM, Johnson AD (2000) Evidence for mating of the “asexual” yeast *Candida albicans* in a mammalian host. *Science*, 289, 307-310.

Krafzig D, Klawonn F, Gutz H (1993) Theoretical analysis of the effects of mitotic crossover in large yeast populations. *Yeast*, 9, 1093-1098.

Lott TJ, Holloway BP, Logan DA, Fundyga R, Arnold J (1999) Towards understanding the evolution of the human commensal yeast *Candida albicans*. *Microbiology*, 145, 1137-43.

Magee BB, Magee PT (2000) Induction of mating in *Candida albicans* by construction of *MTL*a and *MTL*α strains. *Science*, 289, 310-313.

Mannarelli BM, Kurtzman CP (1998) Rapid identification of *Candida albicans* and other human pathogenic yeasts by using short oligonucleotides in a PCR. *Journal of Clinical Microbiology*, 36, 1634-41.

Odds FC (1988) *Candida and Candidosis: A review and bibliography* Baillière Tindall, Philadelphia.

Oliveira RP, Broude NE, Macedo AM, Cantor CR, Smith CL, Pena SDJ (1998) Probing the genetic population structure of *Trypanosoma cruzi* with polymorphic microsatellites. *Proceedings of the National Academy of Sciences of the USA*, 95, 3776-3780.

Pfaller MA, Lockhart SR, Pujol C, Swails-Wenger JA, Messer SA, Edmond MB, Jones RN, Wenzel RP, Soll DR (1998) Hospital specificity, region specificity, and fluconazole resistance of *Candida albicans* bloodstream isolates. *Journal of Clinical Microbiology*, 36, 1518-29.

Piertney SB, MacColl ADC, Bacon PJ, Dallas JF (1998) Local genetic structure in red grouse (*Lagopus lagopus scoticus*): evidence from microsatellite DNA markers. *Molecular Ecology*, 7, 1645-1654.

Pinjon E, Sullivan D, Salkin I, Shanley D, Coleman D (1998) Simple, inexpensive, reliable method for differentiation of *Candida dubliniensis* from *Candida albicans*. *Journal of Clinical Microbiology*, 36, 2093-2095.

Pujol C, Reynes J, Renaud F, Raymond M, Tibayrenc M, Ayala FJ, Janbon F, Mallié M, Bastide JM (1993) The yeast *Candida albicans* has a clonal mode of reproduction in a population of infected human immunodeficiency virus-positive patients. *Proceedings of the National Academy of Sciences of the USA*, 90, 9456-9459.

Pujol C, Joly S, Lockhart SR, Noel S, Tibayrenc M, Soll DR (1997) Parity among the randomly amplified polymorphic DNA method, multilocus enzyme electrophoresis, and Southern blot hybridization with the moderately repetitive DNA probe Ca3 for fingerprinting *Candida albicans*. *Journal of Clinical Microbiology* 35, 2348-2358.

Raymond M, Rousset F (1995) GENEPOP (version 1.2): Population genetics software for exact tests and ecumenicism. *Journal of Heredity*, 86, 248-249.

Ross KG, Shoemaker DD, Krieger MJ, DeHeer CJ, Keller L (1999) Assessing genetic structure with multiple classes of molecular markers: a case study involving the introduced fire ant *Solenopsis invicta*. *Molecular Biology and Evolution*, 16, 525-43.

Schmid J, Herd S, Hunter PR, Cannon RD, Yasin MS, Samad S, Carr M, Parr D, McKinney W, Schousboe M, Haris B, Ikram R, Harris M, Restrepo A, Hoyos G, Singh KP (1999) Evidence for a general-purpose genotype in *Candida albicans*, highly prevalent in multiple geographic regions, patient types and types of infection. *Microbiology*, 145: 2405-2413.

Soll DR (2000) The ins and outs of DNA fingerprinting the infectious fungi. *Clinical Microbiology Reviews*, 13, 332-370.

Taylor JW, Geiser DM, Burt A, Koufopanou V (1999) The evolutionary biology and population genetics underlying fungal strain typing. *Clinical Microbiology Reviews*, 12, 126-146.

Tibayrenc M, Kjellberg F, Arnaud J, Oury B, Brenière SF, Dardé ML, Ayala FJ (1991) Are eukaryotic microorganisms clonal or sexual? A population genetics vantage. *Proceedings of the National Academy of Sciences of the USA*, 88, 5129-5133.

Tibayrenc M (1996) Towards a unified evolutionary genetics of microorganisms. *Annual Review of Genetics*, 50, 401-429.

Weir B (1996) *Genetic Data Analysis II.* Sinauer Associates, Incorporated, Massachusetts.

Whittam TS, Ochman H, Selander RK (1983) Multilocus genetic structure in natural populations of *Escherichia coli*. *Proceedings of the National Academy of Sciences of the USA*, 80, 1751-1755.

Wright S (1969) *Evolution and the Genetics of Natural Populations*, vol. 2. The University of Chicago Press, Chicago.

Xu J, Boyd CM, Livingston E, Meyer W, Madden JF, Mitchell TG (1999a) Species and genotypic diversities and similarities of pathogenic yeasts colonizing women. J*ournal of Clinical Microbiology*, 37, 3835-3843.

Xu J, Mitchell TG, Vilgalys R (1999b) PCR-restriction fragment length polymorphism (RFLP) analyses reveal both extensive clonality and local genetic differences in *Candida albicans*. *Molecular Ecology*, 8, 59-73.

Xu J, Vilgalys R, Mitchell TG (1999c) Lack of genetic differentiation between two geographically diverse samples of *Candida albicans* isolated from patients infected with human immunodeficiency virus. *Journal of Bacteriology*, 181, 1369-137

Table 2.1 Population sample sizes N. N_{clonal} refers to the number of isolates with genotypes that occur more than once in that population. $N_{clonal\ genotypes}$ gives the number of isolates in each clone. For example, Linz had 3 isolates with genotype 150 and 2 isolates with genotype 162, for a total of 5 clonal isolates. The Subpopulation Sum is the sum of subpopulation values; the Total Population row includes clones found once in at least two subpopulations and sums to a different value. Twenty clonal genotypes were found in the total population.

Table 2.1. Population sample sizes N.

Population	**N**	**N_{clonal}**	**$N_{clonal\ genotypes}$**
AUSTRIA Linz	25	5	3-150; 2-162
AUSTRIA Vienna	17	4	2-87, 145
BARBADOS	29	10	2-87, 130, 131; 4-152
BRAZIL Piracicaba	16	2	2-152
CHINA Beijing	17	0	0
ENGLAND London	19	2	2-152
GERMANY Würzburg	13	2	2-152
INDIA Chandigarh	15	0	0
KUWAIT Jabriya	25	0	0
NORWAY Tromsø	23	2	2-211
SOUTH AFRICA Johannesburg	46	4	2-86, 170
USA Atlanta	24	8	5-136; 3-159
USA San Francisco	25	3	3-136
Subpopulation Sum	**294**	**42**	**12**
Total Population	**294**	**68**	**20**

Table 2.2 Microsatellite loci have 9 to 46 alleles and heterozygosities H_O between 0.38 and 0.74. Repeat indicates the nucleotide sequence of the tandemly repeated motif and the alleles column gives the number of alleles found for each locus.

Locus	**Primers**	**Repeat**	**Alleles**	**H_O**
CEF3	F TTTCCTCTTCCTTTCATATAGAA	TTC/ TTTC	27	0.74
	R GGATTCACTAGCAGCAGACA			
ERK1	F CGACCACGTCATCAATACAAATCG	$(GCWSMM)_x$	46	0.60
	R CGTTGAATGAAACTTGACGAGGGG	$(CWW)_y$		
KRE6	F CAAGCTTATAGTGGCTACTA	AAT	9	0.63
	R CCAACACTGATACATCTCG			
LOCUS 4	F GTAATGATTACGGCAATGAC	GAA	13	0.51
	R AGAACGACGTGTACTATTGG			
ZNF1	F CCATTACAGCTGAACCAGCGAGGG	CAA	11	0.38
	R CGCTAGGTAACCTACAGATTGTGGC			

Table 2.3. Populations in Hardy-Weinberg equilibrium by microsatellite loci. The reported p values are evaluated at $\alpha = 0.05$.

Population	CEF3	ERK1	KRE6	Locus 4	ZNF1	Overall
AUSTRIA Linz	0.00	0.00	0.04	0.02	0.00	**0**
AUSTRIA Vienna	0.01	0.01	0.00	0.00	0.00	**0**
BARBADOS	0.00	0.00	0.04	0.00	0.60	**1**
BRAZIL	0.09	0.00	0.00	0.03	0.00	**1**
CHINA	0.00	0.00	0.00	0.21	0.00	**1**
ENGLAND	0.00	0.00	1.00	0.26	0.06	**3**
GERMANY	0.00	0.00	0.54	0.08	0.01	**2**
INDIA	0.00	0.00	0.91	0.54	0.00	**2**
KUWAIT	0.00	0.00	1.00	0.74	0.01	**2**
NORWAY	0.00	0.00	0.03	0.48	0.00	**1**
SOUTH AFRICA	0.00	0.00	0.16	0.16	0.01	**2**
USA Atlanta	0.00	0.00	0.00	0.14	0.05	**2**
USA San Francisco	0.00	0.01	0.03	0.02	0.00	**0**
Pops. In Equilibrium	**1**	**0**	**5**	**8**	**3**	**17**

Table 2.4. About 80% of locus pairs are in pairwise linkage equilibrium. The number of populations in linkage equilibrium for each locus pair and number of locus pairs in linkage equilibrium for each population are given in the margins . Exact tests were performed by Genepop 3.1d and the p-values for each test are indicated: ** for p £ 0.005, * for 0.005 < p £ 0.05, + for 0.05 < p £ 0.1, and – for p > 0.1. Since 10 tests were performed for each population and data was used in more than test, $\alpha = 0.005$ (indicated by ** in the table) was used to determine significance. All locus pairs are in equilibrium in some populations.

Table 2.4. Linkage Equilibrium

Population	CEF3 ERK1	CEF3 KRE6	CEF3 LOC4	CEF3 ZNF1	ERK1 KRE6	ERK1 LOC4	ERK1 ZNF1	KRE6 LOC4	KRE6 ZNF1	LOC4 ZNF1	Overall by Pop
AUS Linz	*	-	-	-	-	-	-	**	-	-	**9**
AUS Vienna	*	-	-	**	-	-	**	*	-	-	**8**
BARBADOS	**	-	*	*	-	-	**	**	-	-	**7**
BRAZIL	-	-	-	*	-	-	**	+	-	**	**8**
CHINA	*	**	*	*	**	*	*	**	**	*	**6**
ENGLAND	+	*	*	*	+	**	+	*	-	*	**9**
GERMANY	+	**	**	*	-	+	-	**	*	*	**7**
INDIA	-	-	+	*	**	*	**	**	**	**	**5**
KUWAIT	-	**	**	*	-	*	*	**	+	-	**7**
NORWAY	**	-	**	-	-	**	-	**	-	+	**6**
S AFRICA	-	-	-	-	*	*	-	+	-	-	**10**
USA Atlanta	**	+	*	*	-	*	*	-	-	-	**9**
USA San Fran	+	*	*	+	-	-	-	-	-	+	**10**
Overall by Loci	**10**	**10**	**10**	**12**	**11**	**11**	**9**	**6**	**11**	**11**	**101**

Table 2.5 Genetic variation within and between populations.

Locus	**F_{IS}**	**F_{ST}**	**Rho_{ST}**
CEF3	0.126	0.079	0.180
ERK1	0.240	0.056	0.153
KRE6	0.00	0.027	0.033
LOCUS 4	0.100	0.024	0.011
ZNF1	0.169	0.064	0.211
Overall	**0.132**	**0.052**	**0.154**

Table 2.6. Gene flow (Nm) is 5.6 migrants per generation on average. For all pairwise population comparisons, Nm estimates are given above the diagonal and Rho_{ST} values, an estimator of RST (Goodman, 1997), are given below the diagonal. Estimates were calculated using RSTCALC (Goodman, 1997). Negative values indicate gene flow is so high an estimate cannot be calculated. Rho_{ST} indicates the percent of genetic variance due to differences between each population. Nm indicates the degree of gene flow between the populations, with Nm=1 sufficient to prevent fixation (Wright, 1969). Statistically significant Rho_{ST} values at $P<0.01$ are in bold and indicate these populations are somewhat differentiated; the corresponding Nm values are also significant.

Table 2.6. Gene flow (Nm).

	AUS Linz	**AUS** Vienna	**ENGLAND**	**GERMANY**	**NORWAY**	**CHINA**	**INDIA**	**KUWAIT**	**S AFRICA**	**BARBADOS**	**BRAZIL**	**USA** Atlanta	**USA** San Fran
AUS Linz		-376.2	6.18	-32.22	2.95	-25.28	7.97	-142.7	39.36	1.20	1.96	1.25	1.18
AUS Vienna	0.00		5.76	-11.84	2.59	12.94	2.77	6.79	16.47	0.94	1.86	0.92	0.86
ENGLAND	0.04	0.04		16.11	147.18	2.63	6.54	17.11	2.07	4.98	10.37	0.94	0.86
GERMANY	-0.01	-0.02	0.02		9.75	33.37	3.73	13.11	169.69	2.38	3.40	1.62	1.60
NORWAY	**0.08**	0.09	0.00	0.03		1.56	2.23	3.69	1.37	4.79	-507.1	1.11	1.04
CHINA	-0.01	0.02	**0.09**	0.01	**0.14**		3.91	14.57	-22.25	0.89	1.21	1.29	1.23
INDIA	0.03	0.08	0.04	0.06	**0.10**	0.06		-45.66	2.22	1.09	2.37	0.87	0.80
KUWAIT	0.00	0.04	0.01	0.02	**0.06**	0.02	-0.01		4.83	1.81	2.46	1.09	1.03
S AFRICA	0.01	0.01	**0.11**	0.00	**0.15**	-0.01	**0.10**	0.05		0.73	1.09	1.12	1.10
BARBADOS	**0.17**	**0.21**	0.05	**0.10**	0.05	**0.22**	**0.19**	**0.12**	**0.26**		3.67	0.90	0.83
BRAZIL	**0.11**	0.12	0.02	0.07	0.00	**0.17**	0.10	**0.09**	**0.19**	**0.06**		1.24	1.08
USA Atlanta	**0.17**	**0.21**	**0.21**	**0.13**	**0.18**	**0.16**	**0.22**	**0.19**	**0.18**	**0.22**	**0.17**		-21.67
USA San Fran	**0.17**	**0.23**	**0.22**	**0.14**	**0.19**	**0.17**	**0.24**	**0.20**	**0.19**	**0.23**	**0.19**	-0.01	

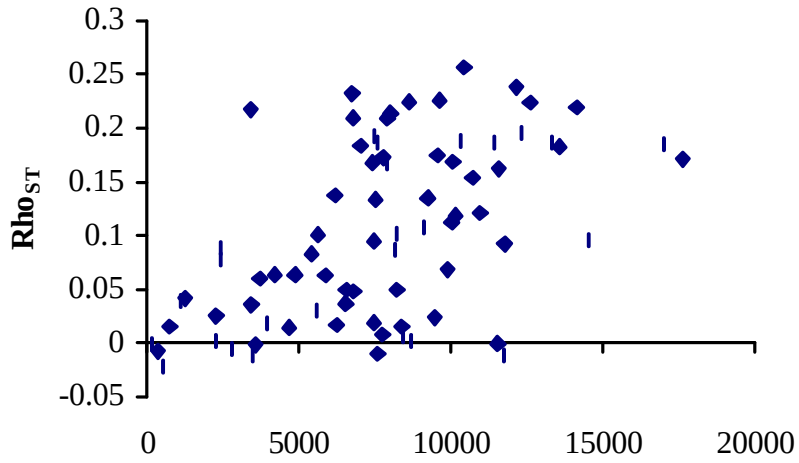

Figure 2.1. Isolation by distance accounts for 29% of genetic variation between populations, with Pearson's correlation coefficient r = 0.54. The relationship between geographic distance in kilometers and Rho_{ST} for population pairs is shown before clone correction. The pairwise Rho_{ST} values are shown in Table 2.2.

CHAPTER 3

WRIGHT'S F-STATISTICS IN CLONAL POPULATIONS[1]

[1] Fundyga, R.E and J. Arnold. 2001. To be submitted to Heredity

ABSTRACT

Clone-correction is a data analysis procedure used in clonally reproducing eukaryotes in which each genotype surveyed is counted once in the analysis regardless of how many times it actually appears. Our goal is to understand the effects of clone correction on perceived population structure. Clone correction was developed to avoid bias in population genetic analyses introduced by counting related individuals or the same genetic individual more than once. The utility of this procedure has not been formally assessed. The ratios of F_{IS} and F_{ST} before and after clone-correction were analytically determined to examine if clone-correction itself alters the perception of population structure. For one locus with two alleles, F_{IS} is corrected to $\pm$ 1/3 and F_{ST} to near 0, indicating non random mating in a hypothetical population which is assumed to be purely sexual with high gene flow between demes. As the number of alleles increased to 5, F_{IS} decreased to 0.17 after clone correction. A model was developed to simulate a population with 10 demes and specified F_{IS} and F_{ST} values to examine the effects of clone-correction at one locus. The F_{ST} values of 0.05 and 0.1 examined in the model did not affect the F_{IS} ratios. For the 2 alleles case, clone correction increased the magnitude of F_{IS} by 5 to 50 fold as the pre-correction F_{IS} increased from 0 to 1. In the 5 allele case, clone correction decreased F_{IS} up to 5 fold over the same range of F_{IS} values. The effect of clone-correction was smaller than in the two-allele case. Finally, the effects of clone-correction are illustrated in a microsatellite population survey of *Candida albicans*, with clone correction increasing F_{IS} by 0.009 to 0.54 at individual loci and 0.033 overall, and decreasing F_{ST} by 0.004 overall.

INTRODUCTION

The mating system and population structure of an organism provide basic information on how genes move between individuals and between populations (Wright, 1969). Such information is needed to manage and conserve wildlife, understand evolutionary processes, harness natural resources, and control pathogens. Understanding how genes move is critical to understanding how traits such as virulence, pathogenicity, and susceptibility are generated, maintained, and spread (Taylor et al., 1999a; Taylor et al., 1999b).

The mating system and population structure of organisms can be described through genotypic surveys that are analyzed within the framework of theoretical population genetics. The theories were developed for diploid systems with only random sexual reproduction and to avoid bias, each individual must be counted only once. It is difficult to apply the theories when clonal reproduction occurs in addition to sexual reproduction, as with many plants, fungi, and protozoa. Methods developed to account for clonality include clonal richness (Weider, et al. 1996) which is the number of unique clones in a population, clonal diversity which is the effective number of clones in a population (Parker, 1979), Simpson's diversity index D (Simpson, 1949; Pielou, 1969), and the Index of Association (Maynard Smith et al., 1993).

A method used in fungal population analysis is clone-correction, in which each genotype is counted once in the analysis no matter how many times it actually is sampled (Milgroom et al., 1992; Chen et al. 1994). It is performed to avoid repeated sampling of the same genetic individual while using the classic framework provided by population

genetics to describe population structure and the mating system. It makes rare alleles more common in the population. Whether this procedure is appropriate, necessary, beneficial or even detrimental has not been assessed formally. The goal of this analysis was to explore mathematically the effects of clone-correction itself on the perception of mating system and population structure. This was done at one locus for a diploid organism, such as a fungus, and was assessed through Wright's F statistics. These are commonly used parameters that describe how heterozygosity is distributed within and between demes, and in the total population.

How does clone-correction affect the perception of population structure using Wright's F statistics? To answer this question, five analyses were performed and they all assumed random, sexual mating and a single diploid locus with co-dominant, neutral alleles. Clone-correction was performed in demes, not on the total population. First an analytical model was derived to see under what conditions the ratios F_{IS}/F_{IS}' and F_{ST}/F_{ST}' would be less than, equal to, or greater than one. Primes (') are used to indicate values after clone-correction. When the ratios equal one, clone-correction has no effect; otherwise it causes a deviation from the uncorrected value. How clone-correction alters observed and expected heterozygosity, F_{IS}, and F_{ST} for 2,5, and 10 alleles was explored next. Third, how incomplete sampling of an allele in some demes affects the outcome of clone-correction was investigated, as might occur with rare alleles. This was done by performing simulations of four population structures, defined by the number of alleles at one locus and F_{ST}, to determine how allele frequency within these structures affect perception of population structure. Finally, the effects of clone-correction on microsatellite data were illustrated.

Population Structure: A Review

Sewall Wright (1969) thought genetic drift to be a major factor shaping populations, and that understanding and measuring genotypic frequencies would allow us to understand the population as a whole. With random mating and no evolutionary forces, such as drift, selection, or migration acting, allele frequencies will not change between generations and genotypic proportions will remain constant. Wright tracked population structure by comparing the observed and expected levels of heterozygosity in populations, with heterozygosity being the proportion of individuals that have two different alleles at a locus.

Populations have three main levels of structure: the individual, the deme, and the metapopulation. Many individuals together form a deme (or subpopulation), and several demes together form the metapopulation. The highest level of structure is the metapopulation and within this metapopulation demes can go extinct and be recolonized. This study does not consider extinction and recolonization processes, so the highest level of structure will be called the total population. Wright (1969) measured differences between the observed and expected heterozygosities within demes to assess nonrandom mating within demes. He also compared expected heterozygosity differences between demes and the total population to measure differentiation between demes. For example, observing fewer heterozygotes than expected within a population may indicate inbreeding, while more heterozygotes than expected may indicate heterosis.

With the advent of molecular genotyping protocols, Wright's F statistics have become common measures of population structure (Avise 1994; Cavalli-Sforza et al.,

1994). The F-statistic F_{IS} is defined as $(H_S–H_O)/H_S$, where H_O is the observed heterozygosity and H_S is the expected heterozygosity within a subpopulation. The observed heterozygosity H_O is estimated directly from the genotypic data, and H_S is calculated as 2pq, where p and q are the observed allele frequencies assuming two alleles. The F-statistic F_{IS} can range from –1 to +1; negative values occur when more heterozygotes were observed than expected, as with selection, migration or clonality. Positive values occur when fewer heterozygotes were observed than expected, as occurs with inbreeding, selfing, migration, or clonality.

The other F-statistic is defined as $F_{ST} = (H_T-H_S)/H_T$, where H_T is the expected heterozygosity in the total population and is calculated from the allele frequencies in the total population. The F-statistic F_{ST} is positive and ranges from 0 to infinity. When all subpopulations have the same genotypic frequencies, $F_{ST} = 0$ implying sufficient migration between subpopulations to create essentially one population. As subpopulations become more genetically isolated, F_{ST} increases but one migrant per generation prevents demes from differentiating (Wright, 1969).

A SIMPLE CASE OF ONE LOCUS WITH TWO ALLELES

First, the effects of clone-correction were examined assuming one locus with two alleles (A and a), pure sexual reproduction, and no factors of evolution acting. Since this is a one locus, diploid model, there are three genotypes possible: AA, Aa and aa. To examine F_{IS}/F_{IS}', all subpopulations are assumed to have the same genotypic composition after clone-correction, and F_{IS}' was directly calculated for the cases of 2 and 3 genotypes.

The conditions under which F_{IS}/F_{IS}' would be less than, equal to or greater than 1 were found by setting the equation $\frac{1-\frac{H_O}{2pq}}{F_{IS}'}$ to 1 and solving for the observed heterozygosity, H_O (Table 3.1).

To examine F_{ST}/F_{ST}', the genotypic composition of subpopulations were assumed to differ. Some proportion m_1 have the post-correction genotypes of AA/Aa, m_2 have AA/aa, m_3 have Aa/aa, and m_4 have all three genotypes. The sum of frequencies of all demic types m_1 to m_4 present in the total population is 1. We assumed demes were not fixed for an allele as markers with little variation are unlikely to be used. Expected heterozygosities in each of the four demic types were directly calculated and weighted by the corresponding proportion to find Hs'. The conditions under which F_{ST}/F_{ST}' would be less than, equal to or greater than 1 were found by setting $\frac{1-\frac{2pq}{H_S}}{F_{ST}'}$ to 1 and examining under which conditions allowable values of m produced the required result (Table 3.2).

For example, in the three-genotype case there are three genotypes of equal frequency after clone-correction (Table 3.1). The observed heterozygosity is 1/3 because one of the three genotypes is a heterozygote. Each allele is found twice in the homozygote and once in the heterozygote, for a total of three copies each out of the six alleles at the locus. The allele frequencies are p = q = 3/6 = 0.5. The expected heterozygosity is 2pq = 1/2. Since $F_{IS} = 1 - (H_O/H_S)$, $F_{IS}' = 1 - (0.33/0.5) = 1/3$. Clone-correction alone forced the perceived F_{IS} to 1/3, away from the true value of zero, and led to the interpretation of inbreeding even though this case assumed only sexual

reproduction. The ratio $F_{IS}/F_{IS}' < 1$ when $-1 < F_{IS} < 1/3$, and by setting the ratio to less than 1 with 1/3 substituted for F_{IS}', it can be found that this occurs when $H_O > 4pq/3$. The other cases were solved similarly.

If all demes have the same genotypic composition after clone-correction, they all have the same H_S and H_T, F_{ST}' is 0, and the ratio F_{ST}/F_{ST}' becomes $F_{ST}/0$ = undefined. The F-statistic F_{ST}' is not zero when demes have different compositions. For example, in the case where the total population consists of a proportion m_1 of demes that are AA/Aa and m_2 of demes that are AA/aa, $H_S' = 2(3/4)(1/4)m_1 + 2(1/2)(1/2)m_2$. All individuals from the clone-corrected demes are included in the total population, and allele frequencies for H_T are calculated from these individuals. The number of each genotype in the total population depends on the proportion m of demes with a given genotypic composition. The allele frequencies in the total population depend on the demic proportions, so in this example $H_T' = 2pq = 2(3/4\ m_1 + 1/2\ m_2)(1/4\ m_1 + 1/2\ m_2)$. These values of H_S' and H_T' were used to calculate F_{ST}', the ratio set to 1, and conditions under which the ratio could occur were found (Table 3.2).

Results:

The number of genotypes in all demes affects F_{IS}' (Table 3.1). When all demes have three genotypes and only random sexual reproduction, F_{IS} is forced to 1/3 after clone-correction, giving the appearance of fewer heterozygotes than expected, as would occur with inbreeding. Clone-correction doesn't change the sign of F_{IS}. The known case when $F_{IS} = 0$ produces a ratio of 0 when the observed and expected heterozygosities are

the same. Clone-correction increases F_{IS}' relative to F_{IS} when $H_O > 4pq/3$, giving the appearance of more inbreeding if F_{IS} was initially positive or more heterosis if F_{IS} was initially negative. Clone-correction decreases F_{IS}' when $H_O < 4pq/3$, implying a smaller deviation from random mating than actually exists. Clone-correction has no effect on F_{IS} when $H_O = 4pq/3$.

When all demes contain only the two homozygotes, the ratio F_{IS}/F_{IS}' will equal F_{IS} (Table 3.1). When all demes contain either AA/Aa or Aa/aa, as with a rare allele, clone-correction forces F_{IS}' = -1/3. If F_{IS} is actually positive, the sign and magnitude of deviation will change. This will change how the mating system is perceived and indicate inbreeding or selfing probably are not occurring when the uncorrected value indicates otherwise. If $-1 < F_{IS} < -1/3$ or $1/3 < F_{IS} < 1$, then F_{IS}' moves away from random mating towards more inbreeding or heterosis. If $-1/3 < F_{IS} < 1/3$, clone-correction makes the mating system appear more like random mating. These results indicate that clone-correction itself can alter the perception of the mating system by changing the type of reproduction inferred or by making the mating system appear more extreme than with the true value.

Clone-correction tends to reduce F_{ST}', forcing it to 0 or to some value less than that actually occurring (Table 3.2). The F-statistic F_{ST}' increases only in the case where the population consists of some combination of demes with genotypes AA/Aa, AA/aa, and AA/Aa/aa. These results indicate that in most cases clone-correction increases the perceived genetic similarity, and by inference increases estimates of gene flow, between demes.

HOW DOES THE NUMBER OF ALLELES AFFECT CLONE-CORRECTION?

How the number of alleles at one locus affected the corrected values of H_O, H_S, F_{IS}, and F_{ST} was examined. A numerical analysis was performed with the statistics calculated directly, assuming a randomly mating sexual population with all demes of the same genotypic composition and with all alleles of equal frequency. As the number of alleles increased from 2 to 5, the observed heterozygosity H_O' increased from 0.33 to 0.67, and allele frequency decreased from 0.5 to 0.2, as expected (Figure 3.1). The F-statistic F_{IS}' decreased from 0.33 to 0.17, and F_{ST}' remained 0 as all demes had the same composition.

Clone-correction affected the perception of the mating system and population structure, with the appearance of random sexual mating and high migration rates being approached as the number of alleles increased. This is consistent with clone-correction making rare alleles seem more frequent than they really are. The ratio F_{IS}/F_{IS}' decreases as observed heterozygosity increases (Figure3.2), with a steeper slope, meaning a smaller area where the ratio is approximately one, for larger numbers of alleles. A shallow slope between 0 and 1 still indicates significant deviations from one, however. At an intermediate number of alleles, as would be expected with isozymes, clone-correction has less effect as the ratio has a shallow slope and is close to 1 for a wider range of heterozygosities than for 2 and 10 alleles.

HOW DO MISSING ALLELES AFFECT CLONE-CORRECTION?

To account specifically for rare alleles and sampling error, the effect of clone-correction when some genotypes are missing was examined in a population for 2,3,4,5,10 and 15 alleles for H_O' and for 2 to 5 alleles for H_S'. This was done analytically by calculating genotype and allele frequencies for all combinations that sequentially led to 0, 1,…, all copies of one allele being absent from a population. Allele frequencies were calculated by directly counting the number of copies of an allele present. In the case with all genotypes present, all alleles are of equal frequency. All demes were assumed to have the same genotypic composition. The effect of clone-correction on both H_O and H_S decreased as the number of alleles increased, as expected (Figure 3.3). Heterozygosity is greater as the number of alleles increases, so the effect of missing genotypes is reduced because each missing heterozygote accounts for a smaller proportion of the overall heterozygosity.

SIMULATION

To determine how clone-correction would affect perceived population structure when each deme has different allele frequencies, four models were simulated in MATLAB release 12 (The MathWorks, Inc., Massachusetts) and 1000 simulated total populations were generated for each. A deme contained sexual, diploid individuals, and a population consisted of 10 demes. Calculations were based on frequencies so the number of individuals sampled per deme did not need to be specified. The model was modified

to allow 2 or 5 alleles at one locus and an F_{ST} of 0.05 or 0.1. It allowed alleles to be missing in some populations as with rare alleles and corresponds to RFLP and isozyme genotypes in an organism with moderate migration.

Allele frequencies for each deme in a simulated population were drawn from a Dirichlet distribution to incorporate population structure through F_{ST} (Wright 1969). The Dirichlet is a multivariate distribution shown to be a reasonable approximation to the distribution of allele frequencies (Slatkin and Muirhead, 2000). The density function for the allele frequencies is:

$$p(x) = \frac{G(Sa_j)}{PG(a_j)} Px_j^{a_j - 1} (1 - x_1 - x_2 - ... - x_d)^{a_{d+1}}$$

for $0 \le x_j \le 1$ and where $G(a) = (a-1)!$ (Gentle, 1998).

MacArthur's broken stick method was used to generate random frequencies over a [0,1] uniform distribution, obtaining one frequency per allele. Population structure was incorporated as $q = 4Nm = \frac{1 - F_{ST}}{F_{ST}}$, for small F_{ST} (Wright, 1969) and multiplying each of the 2 or 5 random frequencies by theta. These new variables, the alphas, were used to generate one gamma distribution per allele.

To generate a random Dirichlet vector of *I* allele frequencies, random numbers were drawn from *I* gamma distributions, one number per deme such that allele *i*'s gamma distribution produced the allele *i* frequencies in all 10 demes. When gamma random numbers were generated for all alleles in all demes, within deme numbers were summed, and each realization of a gamma random variable divided by this sum for the *j*th

particular deme to produce an allele frequency. The sum of these allele frequencies within each deme was 1, and these frequencies were used for the remainder of the calculations.

To obtain genotypic counts for analysis, F_{IS} was specified as 0, 0.1, 0.5, 0.9 and 1 within each of the four models. The F-statistic F_{IS} describes the relative proportions of heterozygotes and was incorporated as $2pq(1-F_{IS})$ for the heterozygotes and as $p^2(1-F_{IS})+p(F_{IS})$ for the homozygotes (Hedrick, 1985; Wright, 1969). Using a program written in MATLAB release 12, F_{IS} and F_{ST} were calculated before and after clone-correction. The F-statistic F_{IS} was calculated as $\frac{H_S - H_O}{H_S}$ and F_{ST} as $\frac{H_T - H_S}{H_T}$. Allele and genotype frequencies summed to 1 in all cases both before and after clone-correction. The pre-correction F_{IS} was also used as a check and in all cases summed to 2.5 (the sum of 0, 0.1, 0.5, 0.9 and 1) as expected. While F_{IS} can range from –1 to +1, only positive F_{IS} values were used to ensure positive genotype frequencies with small allele frequencies.

Results:

F_{ST} was calculated because it was needed in the analysis, but also it was examined to confirm how well the randomly generated population structures matched what was being simulated. For the F_{ST} set at 0.05 in theta, the median F_{ST} was 0.04 and for the F_{ST} set at 0.1, the median F_{ST} was 0.08 in the 2 allele case and 0.09 in the 5 allele case (Table

3.3). The F_{IS} ratio does not vary with the F_{ST} values examined here but does vary with the number of alleles.

For 2 alleles, clone-correction changes the sign of F_{IS} and increases its magnitude from 5 to 50 fold as the assigned F_{IS} moves from 0 to 1. For 5 alleles, the sign of F_{IS} remains the same but the value decreases by at most about 5 fold as the assigned F_{IS} value moves from 0 to 1. For $F_{IS} = 0$ or 1, the ratio equals the assigned F_{IS} value. The F_{ST} ratio was undefined, implying a post-correction F_{ST} of 0, when all alleles were present in all populations.

If alleles were missing in some populations, F_{ST} was a positive number and this is the number reported (Table 3.4). The ratio approximately doubled as the F_{ST} used in theta doubled. In the 2 allele models, F_{ST}' was about 2 to 5 times greater than the precorrection value but for the 5 allele case F_{ST}' was about 5 to 7 times smaller. The effects of clone-correction were greater in the 2 allele models than the 5 allele models.

DATA

The effect of clone-correction on microsatellite data in *Candida albican*s was examined (Table 3.5). Five microsatellite loci were used to genotype *C. albicans* isolates from 13 clinical populations in diverse geographic locations and analyzed using the programs GENEPOP (Chapter 2). Clone-correction was calculated in demes with each genotype being counted once in the analysis and the data reanalyzed. In this multiple allele, multiple locus example, clone-correction increased F_{IS} by 0.033 over all loci, but increased values at individual loci by 0.08 to a maximum of 0.54, which is slightly more

than a quarter of the range of F_{IS}. Clone correction had a much smaller effect on FST, altering the pre-corrected values by ± 0.1.

DISCUSSION

Clone-correction was developed to assist in analysis and interpretation of clonal fungal populations using the power of classical population genetic theories. The utility of this approach has not been formally assessed. Some clonal eukaryotic populations are significant to human health and it is critical to understand their genetic structure in order to implement better control measures (Taylor et al., 1999a; Taylor et al., 1999b). The goal of this paper was to determine how clone-correction affects the perception of population structure through Wright's F statistics. The effects of clone-correction were mathematically examined assuming one locus with two alleles in a randomly mating organism and found the conditions under which F_{IS}/F_{IS}' and F_{ST}/F_{ST}' were greater than, less than or equal to one. The effect of increasing the number of alleles was determined at one locus and of having alleles missing at one locus, both assuming all demes have the same genetic composition. To account for higher-level population structure, populations of 10 demes each were randomly generated based on 2 allele levels, 2 F_{ST} levels and 5 F_{IS} levels and the ratios F_{IS}/F_{IS}' and F_{ST}/F_{ST}' were analytically determined. Finally, the effects of clone-correction were examined on multilocus data obtained from an organism with clonal reproduction and the capacity for sexual reproduction.

When there is one locus with two alleles, clone-correction can dramatically alter the perception of population structure by forcing F_{IS} to +/- 1/3 and reducing F_{ST} to near 0.

This case would be similar to an RFLP genotype at one locus and indicates that even in a randomly mating organism, clone-correction disguises the genetic signatures of sexual recombination and produces a structure that would be expected in a clonal population. The effect decreases by about half for F_{IS}/F_{IS}' as the number of alleles increases from 2 to 5, as expected, because heterozygotes represent a larger proportion of the population as the number of alleles per locus increases. The ratio is closer to one over a greater range of observed heterozygosities for an intermediate number of alleles, as with isozymes. Similarly, the effect of missing alleles on the clone-corrected observed and expected heterozygosities decreases as the number of alleles increases because each heterozygote represents a smaller fraction of the total heterozygosity. The trend continued in the population structure model, with clone-correction's effect decreasing as the number of alleles increased and with a larger effect for F_{IS} than F_{ST}. Finally, as the number of alleles increased in the analysis of an actual fungal population, clone-correction had little effect on F_{IS} when averaged over all loci but increased the value for locus KRE6 by 0.54, which is approximately 1/4 the range of possible F_{IS} values.

Clone-correction makes populations look more clonal with higher gene flow when considering one locus with two alleles. Its effects become less pronounced as the number of alleles increases. When multiple loci were considered, the effects averaged out across loci although they were still more pronounced in F_{IS} than F_{ST}. It takes considerable effort to remove duplicated genotypes within demes and reanalyze the data with the new genotypic composition. This effort resulted in no practical difference when averaged over the multiple loci considered here, but can still have a great effect on F_{IS} at individual loci even in the context of multilocus genotypes.

It is appealing to have a numerical index such as F statistics that provide a way to compare the mating system and genetic population structure between organisms. Wright's F statistics may still be applicable to partially clonal organisms with positive F_{IS} and it may be best to represent clonality as the proportion of repeated genotypes within a deme, similar to F statistics, with the difference between expected and observed proportions normalized by the expected proportion. This way the occurrence of clonality could be balanced by the probability of obtaining that many identical genotypes due to allele frequencies. This clonality could be used to help interpret the F statistics: if clonality is high, selfing or heterosis may not be indicated.

ACKNOWLEDGEMENTS

Daniel Promislow suggested clone correction might produce statistical bias and that it could be examined mathematically. For their comments, we thank John Avise, Sarah Covert, Paul Peteet, Daniel Promislow, and Timothy Lott. REF was supported by a University of Georgia Fellowship and a Training Grant in Molecular and Cellular Mycology (T32-AI-07373) from the National Institutes of Health. JA was supported by grant MCB-9630910 from the National Science Foundation.

REFERENCES

Avise JC. 1994. Molecular Markers, Natural History, and Evolution. Chapman and Hall, New York.

Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. The History and Geography of Human Genes, Abridged. Princeton University Press, Princeton.

Chen RS, Boeger JM, McDonald BA. 1994. Genetic stability in a population of a plant-pathogenic fungus over time. Molecular Ecology 3: 209-218.

Gentle JE. 1998. Random Number Generation and Monte Carlo Methods. Springer, New York.

Hardy, GH. 1908. Mendelian proportions in a mixed population. Science 28: 49-50.

Hedrick, PW. 1985. Genetics of Populations. Jones and Bartlett Publishers, Inc., Boston.

Maynard Smith J, Sitj NH, O'Rourke M, Spratt BG. 1993. How clonal are bacteria? Proceedings of the National Academy of Sciences USA 90: 4384-4388.

Milgroom MG, Lipari SE, Powell WA. 1992. DNA fingerprinting and analysis of population structure in the chestnut blight fungus, *Cryphonectria parasitica*. Genetics 131: 297-306.

Parker ED. 1979. Ecological implications of clonal diversity in parthenogenetic morphospecies. American Zoologist 19: 753-762.

Pielou. 1969. An Introduction to Mathematical Ecology. Wiley, New York.

Simpson EH. 1949. Measurement of diversity. Nature 163: 688.

Slatkin M, Muirhead CA. 2000. A method for estimating the intensity of overdominant selection from the distribution of allele frequencies. Genetics 156: 2119-2126.

Taylor JW, Geiser DM, Burt A. Koufopanou V. 1999a. The evolutionary biology and population genetics underlying fungal strain typing. Clinical Microbiology Reviews 12: 126-146.

Taylor JW, Jacobson DJ, Fisher MC. 1999b. The evolution of asexual fungi: Reproduction, speciation and classification. Annual Review of Phytopathology 37: 197-246.

Weider, LJ, Hoback A, Crease TJ, Stibor H. 1996. Molecular characterization of clonal population structure and biogeography of arctic apomictic *Daphnia* from Greenland and Iceland. Molecular Ecology 5: 107-118.

Weinberg, W. 1908. Über den Nachweis der Vererbung beim Menschen. Jahreshelfts Ver. vaterl. Naturf. Württemberg 64: 3699-382.

Wright, S. 1969. Evolution and the Genetics of Populations Vol. 2: The Theory of Gene Frequencies. The University of Chicago Press, Chicago.

Table 3.1. The ratio F_{IS}/F_{IS}' for a one locus, two allele model. The conditions for which the ratio F_{IS}/F_{IS}' is less than, equal to or greater than one for a one locus, two-allele model assuming sexual reproduction and no evolutionary forces. All demes are assumed to have the same genotypic composition.

Genotypes	F_{IS}'	**Case**	F_{IS}/F_{IS}'	**Condition**
3	1/3	$-1 < F_{IS} < 1/3$	<1	$H_O > 4pq/3$
AA,Aa,aa		$F_{IS} = 1/3$	$= 1$	$H_O = 4pq/3$
		$1/3 < F_{IS} < 1$	>1	$H_O < 4pq/3$
		$F_{IS} = 0$	$= 0$	$H_O = 2pq$
2 AA, aa	1		F_{IS}	always
2	-1/3	$1/3 < F_{IS} < 1$	< -1	$H_O < 4pq/3$
AA,Aa		$F_{IS} = 1/3$	$= -1$	$H_O = 4pq/3$
Or		$0 < F_{IS} < 1/3$	$-1 < F_{IS}/F_{IS}' < 0$	$H_O > 4pq/3$
Aa,aa		$F_{IS} = 0$	$= 0$	$H_O = 2pq$
		$-1/3 < F_{IS} < 0$	$0 < F_{IS}/F_{IS}' < 1$	$H_O < 8pq/3$
		$F_{IS} = -1/3$	$= 1$	$H_O = 8pq/3$
		$-1 < F_{IS} < -1/3$	>1	$H_O > 8pq/3$

Table 3.2 The conditions for which the ratio F_{ST}/F_{ST}' is less than, equal to or greater than one for a one locus, two allele model assuming sexual reproduction and no evolutionary forces. The genotypic composition of demes after clone-correction is described in the first column. The proportion of demes in the total population that are AA, Aa is m_1 ; that are AA, aa is m_2; that are Aa, aa is m_3; and that are AA,Aa,aa is m_4. F_{ST}' is defined based on the proportion of each demic type present in the total population, m_1 to m_4, and the sum of all m is 1. F_{ST}' is either 0 or some expression 1 + Z where Z is a fraction based on the proportion of deme types. The column F_{ST}/F_{ST}' refers to the numerator, the denominator, or the expression Z itself.

Table 3.2. The ratio F_{ST}/F_{ST}' for a one locus, two allele model assuming sexual reproduction.

Case	Genotypes	F_{ST}'	F_{ST}/F_{ST}'	Result
I. All demes	i) AA/Aa (*m1*)	0	undefined	
have 2 genotypes	ii) AA/aa (m_2)	0	undefined	
	iii) AA/Aa, AA/aa	$1+\frac{m_2+3}{m_2^2-2m_2-3}$	< 1 when denominator >0	never occurs
			= 1 when numerator = 0	never occurs
			>1 when denominator < 0	always occurs
	iv) AA/Aa, AA/aa, Aa/aa (m_3)	$1+\frac{m_2+3}{m_2^2+4m_3^2-2m_2-4m_3-3}$	<1 when denominator >0	never occurs
			= 1 when numerator = 0	never occurs
			>1 when denominator < 0	always occurs

Case	Genotypes	F_{ST}'	F_{ST}/F_{ST}'	Result
II. All demes have 3 genotypes	AA/Aa/aa (m_4)	0	undefined	
III. Demes have either 2 or 3 genotypes	i) AA/Aa, AA/Aa/aa	$1+\frac{m_4+3}{m_4^2-2m_4-3}$	<1 when denominator > 0	never occurs
			= 1 when numerator = 0	never occurs
			> 1 when denominator < 0	always occurs
	ii) AA/aa, AA/Aa/aa	0	undefined	

Case	Genotypes	F_{ST}'	F_{ST}/F_{ST}'	Result
	iii) AA/Aa, AA/aa, AA/Aa/aa	$1+\frac{m_1}{4-m_1^2}$	<1 when m1/(4-m1^2)>0	always occurs
			= 1 when m1 = 0	never occurs
			>1 when m1 < 0	never occurs
	iv) AA/Aa, Aa/aa, AA/Aa/aa	$1-\frac{m_4+3}{(m_4+2m_1+1)(m_4+2m_3+1)}$	<1 when Z < 0	never occurs
			= 1 when Z = 0	never occurs
			> 1 when Z > 0	always occurs

Case	Genotypes	F_{ST}'	F_{ST}/F_{ST}'	Result
	v) AA/Aa, AA/aa, Aa/aa, AA/Aa/aa	$1-\frac{4-m_1-m_3}{4-m_1^2-m_3^2}$	< 1 when Z < 0	never occurs
			= 1 when numerator = 0	never occurs
			>1 when Z >0 always happens	

Table 3.3. F_{ST} values before clone-correction obtained from the randomly generated population

Model	2 alleles, F_{ST} = 0.05	5 alleles, F_{ST} = 0.05
Median	0.04	0.04
Average	0.04	0.05
Variance	0.0005	0.0002
Standard Deviation	0.02	0.01
Lowest Value	0.0008	0.014
Highest Value	0.14	0.11
Model	2 alleles, F_{ST} = 0.1	5 alleles, F_{ST} = 0.1
Median	0.08	0.09
Average	0.09	0.09
Variance	0.002	0.0007
Standard Deviation	0.04	0.02
Lowest Value	0.0009	0.026
Highest Value	0.30	0.21

Table 3.4. The effect of allele number, F_{IS} and F_{ST} on heterozygosities before clone-correction and F statistic ratios. Average results from 1000 runs of each population structure showing the effect of allele number, F_{IS} and F_{ST} on heterozygosities before clone-correction and F statistic ratios. The F_{ST}/F_{ST}' ratios for the 5 allele models with F_{IS} = 1 include error attributable to the model of one thousandth or less of one percent of the allele frequencies. Upon inspection, the F_{ST} calculations are 0/0 for practical purposes and are undefined. When all alleles are present in all demes, the F_{ST} ratio is undefined. The table reports the average ratio when some demes were randomly missing some alleles.

Table 3.4. The effect of allele number, F_{IS} and F_{ST} on heterozygosities before clone-correction and F statistic ratios.

	2 alleles, F_{ST} = 0.05				5 alleles, F_{ST} = 0.05			
	H_O	H_S	F_{IS}/F_{IS}'	F_{ST}/F_{ST}'	H_O	H_S	F_{IS}/F_{IS}'	F_{ST}/F_{ST}'
F_{IS} = 0	0.32	0.32	0	0.22	0.63	0.63	0	4.90
F_{IS} = 0.1	0.29	0.32	-0.02	0.22	0.57	0.63	0.60	4.90
F_{IS} = 0.5	0.16	0.32	-0.09	0.22	0.32	0.63	3.00	4.90
F_{IS} =0.9	0.03	0.32	-0.16	0.22	0.06	0.63	5.40	4.90
F_{IS} = 1	0	0.32	1	0.22	0	0.63	1	-0.2*
	2 alleles, F_{ST} = 0.1				5 alleles, F_{ST} = 0.1			
	H_O	H_S	F_{IS}/F_{IS}'	F_{ST}/F_{ST}'	H_O	H_S	F_{IS}/F_{IS}'	F_{ST}/F_{ST}'
F_{IS} = 0	0.30	0.30	0	0.49	0.60	0.60	0	7.67
F_{IS} = 0.1	0.27	0.30	-0.02	0.49	0.54	0.60	0.60	7.67
F_{IS} = 0.5	0.15	0.30	-0.09	0.49	0.30	0.60	2.99	7.67
F_{IS} =0.9	0.03	0.30	-0.16	0.49	0.06	0.60	5.38	7.67
F_{IS} = 1	0	0.30	1	0.49	0	0.60	1	4.64*

Table 3.5. In a study of *Candida albicans* populations using microsatellite loci, clone-correction increases F_{IS} slightly but has no affect on F_{ST}.

Locus	**F_{IS}**	**F_{ST}**
CEF3	.126 / .149	.079/.069
ERK1	.240 / .270	.056 / .050
KRE6	0.00 / 0.54	.027 / .024
LOCUS 4	.100 / .150	.024 / .031
ZNF1	.169 / .178	.064 / .057
Overall	**.132 / .165**	**.052 / .048**

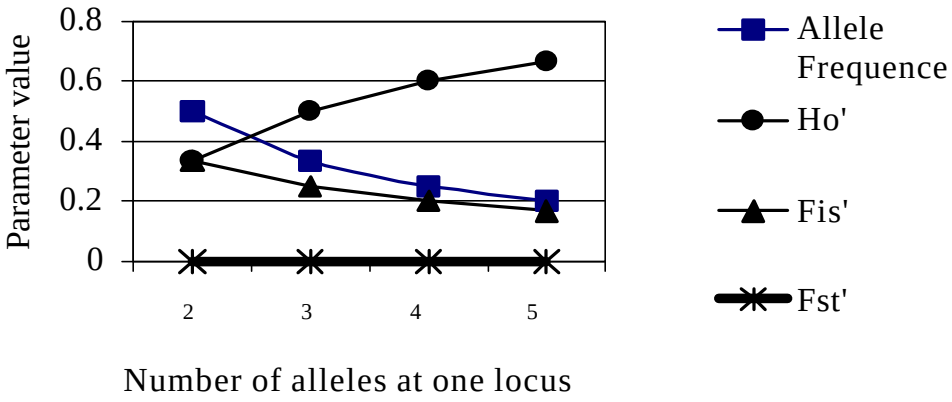

Figure 3.1 The effect of increasing number of alleles on clone-corrected values, assuming all demes contain all genotypes.

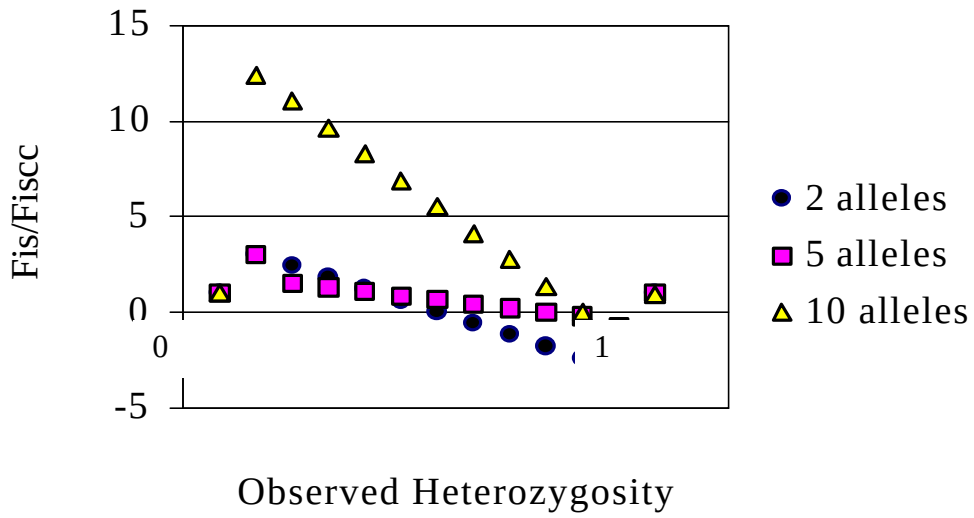

Figure 3.2 The effect of increasing number of alleles at one locus on the ratio of F_{IS} before and after clone-correction for 2, 5, and 10 alleles. All demes contain all genotypes, have the same genetic composition, and F_{ST}' is 0.

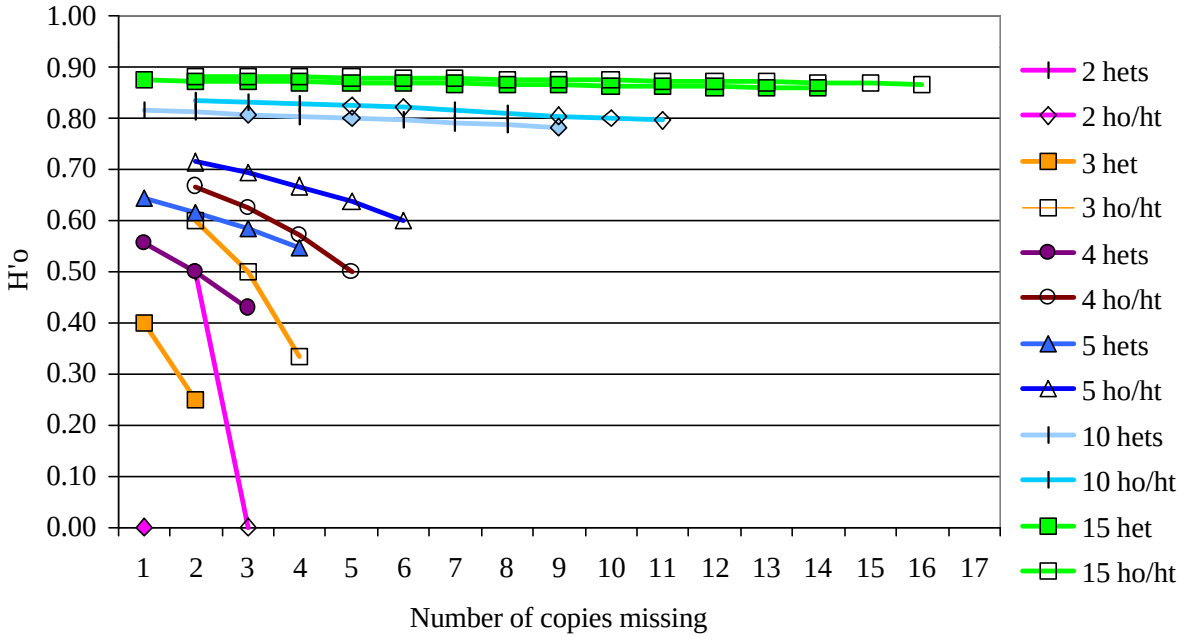

Figure 3.3 The effect of missing copies of one allele at a locus on the observed heterozygosity after clone-correction for 2,3,4,5,10 and 15 alleles. Solid markers indicate behavior of H_O' when copies are missing only in heterozygotes and open markers indicate behavior of H_O' when copies are missing in both homozygotes and heterozygotes. For example, "3 het" indicates 3 different heterozygotes are missing while "3 ho/het" indicates 2 copies are missing in a homozygote and 1 in a heterozygote.

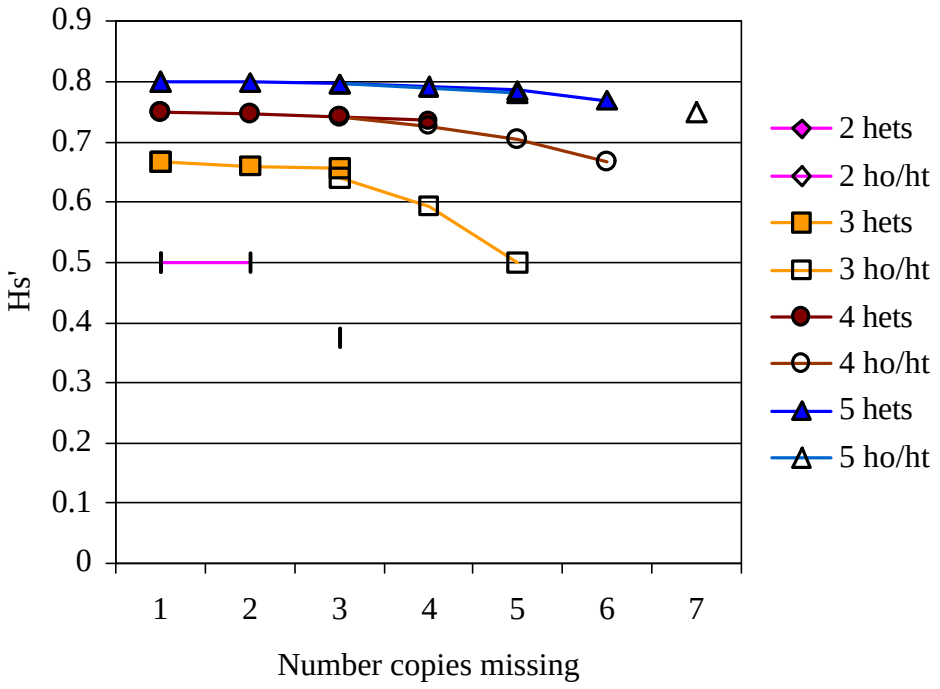

Figure 3.4. The effect of missing copies of one allele at a locus on the expected heterozygosity after clone-correction for 2,3,4, and 5 alleles. Solid markers indicate behavior of H_O' when copies are missing only in heterozygotes and open markers indicate behavior of H_O' when copies are missing in both homozygotes and heterozygotes. For example, "3 het" indicates 3 different heterozygotes are missing while "3 ho/het" indicates 2 copies are missing in a homozygote and 1 in a heterozygote

CHAPTER 4

THE EFFECTS OF CLONAL, SEXUAL, AND PARASEXUAL REPRODUCTION ON ALLELE FREQUENCIES[1]

[1] Fundyga, R.E and J. Arnold. 2001. To be submitted to Heredity

ABSTRACT

A model was developed to examine how clonal, sexual, and parasexual reproduction in fungi affect genotype frequencies across generations. It assumed one locus with two alleles. The model was an extension of the Hardy-Weinberg Law and incorporated disequilibria derived for clonal and parasexual reproduction. The parasexual disequilibria decay at a rate proportional to 1/3 the frequency of parasexual reproduction. This is faster than the decay rate of 1/2 under random mating. Even when clonal and/or parasexual reproduction accounts for 99% of all reproduction, disequilibria decay to nearly 0 in 10 generations.

INTRODUCTION

The manner in which genomes are transmitted during reproduction influences how allele frequencies are structured in a population. The basic assumption of classic population genetic models is that of random mating with biparental, meiotic recombination of genomes. This leads to random combinations of all alleles in the absence of evolutionary forces. Plants, fungi, some protozoa, and some animals do not reproduce according to this idealized model and also use selfing, clonality, and non-meiotic forms of biparental recombination to transmit their genes. Understanding how genes are transmitted is important for developing conservation plans, understanding how mating systems evolve, and surveying and controlling pathogen populations (Allard,

1975; Taylor et al., 1999). Examining how the relative proportions of clonal and sexual reproduction affect allele frequencies in medically important species is an important public health issue as the type of reproduction determines whether pathogenic genes or pathogenic genomes need to be monitored (Taylor et al., 1999).

How do these various forms of reproduction affect the observed allele frequencies? To answer these questions, the Hardy-Weinberg equations were modified to account for types of reproduction found in fungi and numerically analyzed to determine how the proportion of each reproduction mechanism and the allele frequencies interact in a hypothetical diploid fungus. The Hardy-Weinberg equations specify allele and genotype frequencies over time according to the probabilities of a given allele combination occurring.

This model assumed three types of reproduction such as occurs in fungi but can be modified easily to allow a more general examination of clonality and recombination. In sexual reproduction, it is assumed that there is fusion of two haploid gametes from different parents and that the gametes are produced by meiosis. In clonal reproduction, it is assumed that offspring are genetic duplicates of a single parent. Different definitions of clonality and selfing are used for fungi and plants. In fungi, selfing is considered a form of clonal reproduction because the offspring are genetically identical to the parent. In plants, selfing can lead to the production of offspring genetically different from the parent and is not considered a form of clonal reproduction. This model defines clonality as the production of genetically identical offspring by a single parent.

The last form of reproduction considered is parasexuality in which diploid, mitotic parent cells fuse to create a tetraploid, followed by random loss of chromosome to

restore diploidy (Poulter et al., 1981; Gu and Ko, 1998). It is assumed that recombination between homologous chromosomes and meiosis do not occur. Parasexual reproduction can lead to the formation of new genotypes because two parents contribute to the offspring, but the parents are not guaranteed equal contribution to the offspring as occurs in meiosis (Pontecorvo, 1956). In parasexual reproduction, random chromosome loss implies that one parent may contribute both alleles at a locus by chance.

THE MODEL

Let there be two alleles, *A* and *a*, at one locus in a diploid organism with frequencies p and q. This organism can reproduce by one of three mechanisms: clonally at a frequency *c*, sexually by haploid gametes at a frequency *g*, and parasexually by vegetative, diploid cells at a frequency *v*. Allele frequencies are the same for each reproductive class and there is no selection, migration, or drift. Clonal reproduction is defined as transmission of an exact copy of one parental genome to an offspring. Sexual reproduction is the random union of haploid, meiotically produced gametes and produces new genotypic associations; it will also be referred to as random mating. Parasexual reproduction is the random union of diploid, mitotically produced cells followed by random chromosome loss to restore diploidy. It is assumed that generations are discrete and that individuals can reproduce by one of the three mechanisms with probabilities *c*, *g*, and *v*. The proportions of each reproductive mechanism are constant and sum to one.

The proportion of each genotype in the next generation was found using a mating table (Table 4.1) that shows the frequency of each offspring genotype produced from a

given mating. The frequency of each parental genotype is *x* for AA, *y* for Aa, and *z* for aa. The frequency of a given offspring genotype from a particular mating is the product of the probability of the reproductive mechanism, the parental genotype(s), and the probability of producing a genotype given the parental genotype. The offspring were summed down each column (across all matings) to determine that genotype's frequency in the next generation. In clonal reproduction, only the parental genotype is reproduced and the frequency of each offspring type generated is the product of the genotype frequency and the probability of clonal reproduction. In sexual reproduction, it is assumed that there is random segregation of alleles and that one allele per parent is transmitted to the offspring; offspring frequencies were calculated using standard Mendelian probabilities. Offspring frequencies from parasexual reproduction were calculated assuming any two copies of a locus could be transmitted. The difference between parasexual and sexual reproduction can be illustrated by examining the progeny of the two homozygotes, AA and aa. In sexual reproduction, only heterozygous offspring are produced, but parasexual reproduction generates homozygotes with a frequency of 1/6 each because the probability of getting one allele A is 1/2 and the probability of again choosing A from the remaining three alleles is 1/3.

Disequilibria occur when the observed genotypes are different from those expected under random mating (Weir, 1996). For example, $x = p^2 + D$ and $y = 2pq - 2D$ (Hedrick, 1985), where *D* is the disequilibrium observed at that genotype. These definitions along with the results from the mating table were used to derive the recursions with clonal and parasexual reproduction included. The recursion equations for the three genotype frequencies are:

$$x' = p^2 + (c + \frac{1}{3}v)(xz - \frac{1}{4}y^2) \quad (1)$$

$$y' = 2pq - 2(c + \frac{1}{3}v)(xz - \frac{1}{4}y^2) \quad (2)$$

$$z' = q^2 + (c + \frac{1}{3}v)(xz - \frac{1}{4}y^2) \quad (3)$$

The solutions to these recursions are:

$$x_n = p^2 + \frac{1}{4}(c + \frac{1}{3}v)^n(4xz - y^2) \quad (4)$$

$$y_n = 2pq - \frac{1}{2}(c + \frac{1}{3}v)^n(4xz - y^2) \quad (5)$$

$$z_n = q^2 + \frac{1}{4}(c + \frac{1}{3}v)^n(4xz - y^2) \quad (6)$$

where n is the number of generations. The recursions sum to 1 and allele frequencies remain the same across generations. The solution for the allelic disequilibrium is:

$$D_n = (c + \frac{1}{3}v)^n D \quad (7)$$

RESULTS AND APPLICATIONS

The disequilibria in equations (1) through (6) consist of parts from sexual reproduction, and from clonal and parasexual reproduction. When genotype frequencies are in Hardy-Weinberg proportions, $xz - \frac{1}{4}y^2$ will be 0 and the recursions will yield the same frequencies across generations. Nonrandom mating will lead to $xz - \frac{1}{4}y^2$ being nonzero. Disequilibria occur under non-random mating and clonal and parasexual

reproduction. The rate of decay will be proportional to $(c+\frac{1}{3}v)$, and disequilibria will approach 0 as the exponential of $(c+\frac{1}{3}v)$ becomes large. The contribution of parasexual reproduction to disequilibria is a factor of 1/3, indicating the effects of biparental mixing. To investigate the rate of decay, $(c+\frac{1}{3}v)^n$ was plotted for values of $0 \le (c+\frac{1}{3}v) \le 1$, and for increasing number of generations *n* (Figure 4.1). The time for disequilibria to decay geometrically to 0 is at a rate $(c+\frac{1}{3}v)$. The rate of decay under parasexual reproduction, $\frac{1}{3}v$, is faster than the decay rate of ½ under random mating (Hedrick, 1985). For up to 99% of all reproduction occurring clonally and/or parasexually, disequilibria decay to nearly 0 within 10 generations. In this model, disequilibria are maintained only when meiotic, biparental mating never occurs.

The behavior was examined of 12 diploid loci in a published study of the population structure of *Candida albicans*, a member of the human normal flora that is associated with opportunistic infections (Gräser et al., 1996). This study found evidence of both clonal and sexual reproduction in clinical isolates using co-dominant PCR-RFLP markers. This and other studies using different markers have consistently found 1/4 to 1/2 of loci in agreement with Hardy-Weinberg proportions (see Chapter 1). The frequency of the most common homozygote genotype was considered the genotype AA and plotted using equations (4) through (6) assuming 50% and 99% parasexual and/or clonal reproduction (Figures 4.2 and 4.3). Assuming 50% parasexual reproduction, genotype frequencies were stable to 2 decimal places in 5 generations for all loci with a

range of 2 to 5 generations; assuming 99% parasexual reproduction, genotypes were stable in 10 generations for all loci with a range of 1 to 10 generations (Figures 4.2 and 4.3).

The observation that 1/4 to 1/2 of loci in *C. albicans*' populations are in agreement with Hardy-Weinberg expectations (Gräser et al., 1996; Lott and Effat, 2001; Chapter Two) may be partly due to variation in time for parasexual and sexual reproduction to decay disequilibria at the loci (Figure 4.4). Assuming either 50% or 99% parasexual or clonal reproduction, there is a range in time for the genotype frequencies to remain stable, indicating no disequilibria. However, even 99% clonal or parasexual reproduction is insufficient to retain disequilibria permanently. Since *C. albicans* populations are likely to be quite ancient (Hoyer et al., 2001; Lott and Effat, 2001), this implies other mechanisms are responsible for maintaining the disequilibria, such as selection, hitch-hiking, assortative mating (Wright, 1969) or mitotic recombination (Aguilera, et al., 2000; Chavez, et al., 2000; Ossman, et al., 2000).

Empirical studies have found random mating in fungi previously thought to be asexual (Gräser et al., 1996; Geiser et al., 1997), but the importance of sexual, clonal, and asexual reproductive mechanisms in structuring allele frequencies remains unknown. This model has demonstrated that in this simple case, parasexual and/or clonal reproduction in 99% of the population can drive genotypic disequilibria to nearly 0 in 10 generations. Low levels of sexual reproduction can provide substantial genotypic mixing even in the face of 99% clonal reproduction, as has been hypothesized elsewhere (Taylor et al., 1999).

REFERENCES

Aguilera A, Chávez S, Malagón. 2000. Mitotic recombination in yeast: elements controlling its incidence. Yeast 16: 731-754.

Allard RW. 1975. The mating system and microevolution. Genetics 79: 115-126.

Chavez S, Beilharz T, Rondon AG, Erdjument-Bromage H, Tempst P, Svejstrup JQ, Lithgow T, Aguilera A. 2000. A protein complex containing Tho2 Hpr1, Mft1 and a novel protein, Thp2, connects transcription elongation with mitotic recombination in *Saccharomyces cerevisiae*. EMBO Journal 19: 5824-5834.

Geiser DM, Arnold ML, Timberlake WE. 1994. Sexual origins of British *Aspergillus nidulans* isolates. Proceedings of the National Academy of Sciences USA. 91:2349-2352.

Gräser Y, Volovsek M, Arrington J, Schonian G, Presber W, Mitchell TG, Vilgalys R. 1996. Molecular markers reveal that population structure of the human pathogen *Candida albicans* exhibits both clonality and recombination. Proceedings of the National Academy of Sciences USA. 93:12473-12477.

Gu Y-H, Ko W-H. 1998. Occurrence of a parasexual cycle following the transfer of isolated nuclei into protoplasts of *Phytophthora parasitica*. Current Genetics 134: 120-123.

Hedrick, PW. 1985. Genetics of Populations. Jones and Bartlett Publishers, Inc., Boston.

Hoyer LL, Fundyga R, Hecht JE, Kapteyn JC, Klis FM, Arnold J. 2001. Characterization of agglutinin-like sequence genes from non-*albicans Candida* and phylogenetic analysis of the ALS family. Genetics 157:1555-1567.

Lott TJ, Effat MM. 2001. Evidence for a more recently evolved clade within a *Candida albicans* North American population. Microbiology 147:1687-1692

Ossman F, Adriance M, McCready S. 2000. The genetic control of spontaneous and UV-induced mitotic intrachromosomal recombination in the fission yeast *Schizosaccharomyces pombe*. Current Genetics 38: 113-125.

Pontecorvo G. 1956. The parasexual cycle in fungi. Annual Review of Microbiology 10: 393-400.

Poulter R, Jeffery K, Hubbard MJ, Shepherd MG, Sullivan PA. 1981. Parasexual genetic analysis of *Candida albicans* by spheroplast fusion. Journal of Bacteriology 146: 833-840.

Taylor JW, Jacobson DJ, Fisher MC. 1999. The evolution of asexual fungi: Reproduction, speciation and classification. Annual Review of Phytopathology 37: 197-246.

Weir B. 1996. Genetic Data Analysis II. Sinauer Associates, Incorporated, Massachusetts.

Wright S. 1969. Evolution and the Genetics of Natural Populations, vol. 2. The University of Chicago Press, Chicago.

Table 4.1 Mating Table

Reproduction	Parents	Offspring		
		AA	Aa	aa
Sexual	AA x AA	1	0	0
	AA x Aa	1/2	1/2	0
	AA x aa	0	1	0
	Aa x Aa	1/4	1/2	1/4
	Aa x aa	0	1/2	1/2
	aa x aa	0	0	1
Clonal	AA	1	0	0
	Aa	0	1	0
	Aa	0	0	1
Parasexual	AA x AA	1	0	0
	AA x Aa	1/2	1/2	0
	AA x aa	1/6	4/6	1/6
	Aa x Aa	1/6	4/6	1/6
	Aa x aa	0	1/2	1/2
	Aa x aa	0	0	1

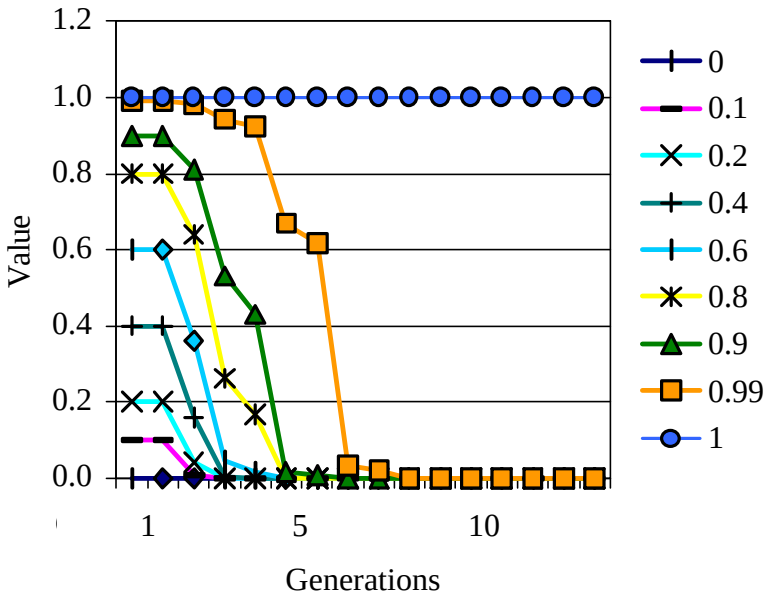

Figure 4.1 Number of generations *n* required for $(c+\frac{1}{3}v)^n$ to decay.

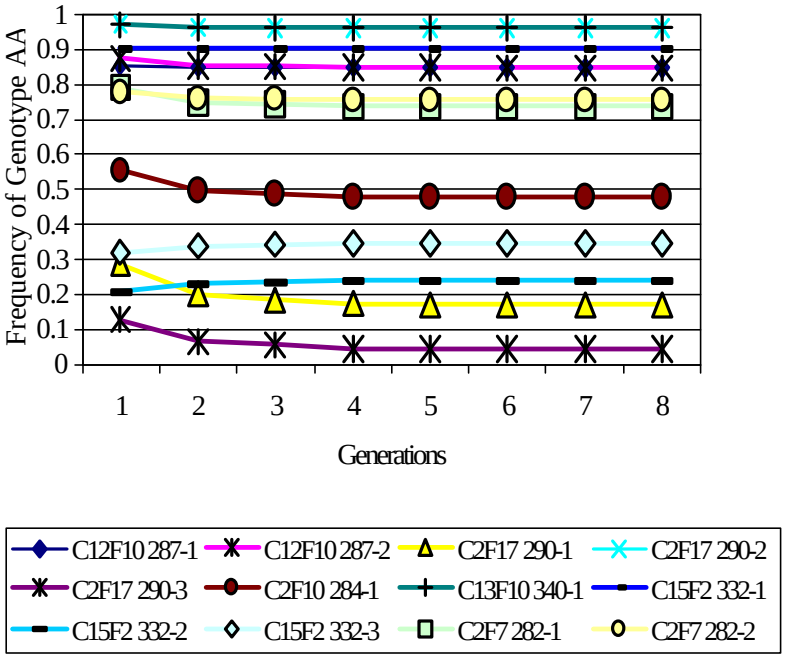

Figure 4.2 The number of generations required to reach stable genotype frequencies for 12 PCR-RFLP loci (Graser et al., 1996), assuming 50% of reproduction occurs by random, sexual mating.

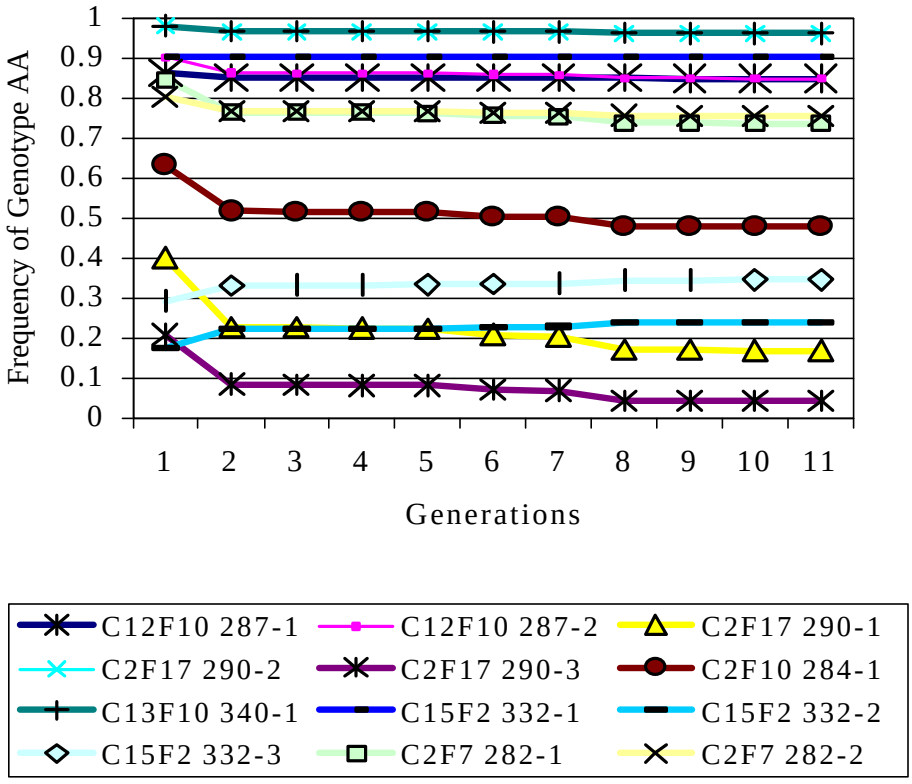

Figure 4.3 The number of generations required to reach stable genotype frequencies for 12 PCR-RFLP loci (Graser et al., 1996), assuming 1% of reproduction occurs by random, sexual mating.

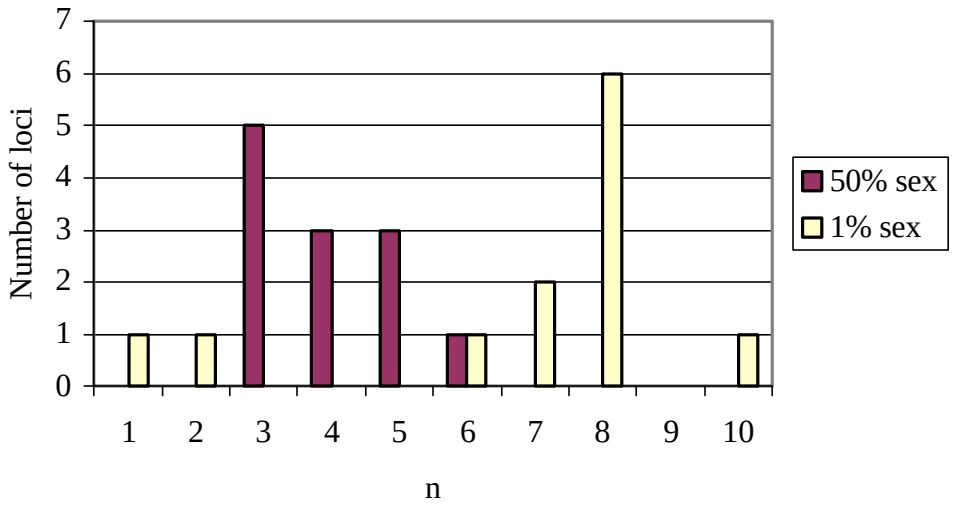

Figure 4.4 Number of generations n until the AA genotype is at stable frequencies to 2 decimal places.

CHAPTER 5

CONCLUSIONS

In the past decade, molecular epidemiologists have drawn from the theories of classical population genetics to understand the genetic structure of pathogen populations. This dissertation has furthered understanding of fungal population genetics through both empirical and theoretical studies, with an emphasis on diploid pathogens. The second chapter described the population structure of *Candida albicans*, accomplished through a survey of clinical isolates from various geographical locations. Five microsatellite loci were used to genotype 294 isolates from 13 cities. As in previous studies, no clear geographic patterns were found but there was a suggestion of division by patient population (Pfaller et al., 1998; Xu et al., 1999). The two United States populations were the most different from all of the other populations and were unique in being bloodstream isolates. The two U.S. populations were similar to each other, but statistically different from the Norway population, which consisted of half bloodstream isolates and was the most different population in Europe.

These results are consistent with those of Pfaller et al. (1998) who found hospital bloodstream isolates to be differentiated from mucosal isolates in the community. Xu et al. (1999) found isolates from HIV-positive patients in the U.S. and Brazil to be similar and that patient isolates grouped separately from those from healthy populations. Other studies have found no relation between body site, patient illness, and isolate genotype when comparing whether isolate genotypes are the same or different across body sites

and illnesses. The second chapter and the two other population studies described above imply that *C. albicans* isolates may genetically cluster by patient group and that broader population genetic surveys including both hospital, community, and healthy patient isolates are needed to track the genetic dynamics of *C. albicans* in at-risk patient groups.

One potential difficulty in elucidating the mating system and population structure of fungal populations is clonality. Population genetics methods commonly rely on sampling genetically unrelated individuals. To avoid biasing statistics by counting individuals with the same genotype more than once, some fungal researchers have used a process called clone correction in which a genotype is counted once in the analysis no matter how many times it is actually sampled (Milgroom et al., 1992; Chen et al., 1994). The third chapter is a formal mathematical analysis of clone correction to determine how it affects the perception of population structure. This was examined for one locus with two alleles by numerically calculating specific behaviors of F statistics, by simulating 1000 different 10-deme populations at two F_{ST} levels, and by performing clone correction on the data from the second chapter. The process of clone correction itself alters the perception of population structure. The statistic F_{IS} moves away from zero, giving the appearance of nonrandom mating, and F_{ST} moves toward zero implying higher levels of gene flow. The effects are reduced as the number of alleles at one locus increases from two to five, but increasing the number of alleles per locus alone is not a guarantee that clone correction will be without effect. The two F_{ST} values examined in the simulation did not themselves affect the results of clone correction on F_{IS}. When performed on the microsatellite data from the second chapter, clone correction increased F_{IS} by 0.09 to 0.54

for each of the five loci individually. The results of this chapter show that clone correction can alter the perception of population structure even in an organism that reproduces entirely by random mating, and that results of clone correction need to be interpreted with caution.

The fourth chapter was a mathematical model examining the effects of different proportions of sexual, clonal, and parasexual reproduction on genotype frequencies. As fungi can produce clonally, parasexually, and sexually, formal mathematical predictions of their genotype frequencies are needed to help interpret population genetic data. The model in the fourth chapter is an extension of the Hardy-Weinberg Law, but the disequilibria had to be derived specifically assuming clonal and parasexual reproduction. In this one locus, two allele model, the disequilibria decay in 10 generations even with 99% of the reproduction occurring by clonal and/or parasexual mechanisms. The rate of decay in this model is proportional to $(c + \frac{1}{3}v)$ indicating that decay due to parasexual reproduction can be faster or slower than that under mixed mating $\frac{1+S}{2}$ (Weir et al., 1972).

This dissertation examined the population genetics of diploid fungi from both empirical and theoretical perspectives. Future work to further clarify the role of different reproductive mechanisms in genetically structuring fungal populations is needed, particularly with regard to fungi isolated and cultured in a haploid state. Such efforts combining epidemiology with population genetics will assist our efforts to monitor, control, and prevent severe fungal infections in susceptible patients.

REFERENCES

Aguilera A, Chávez S, Malagón. 2000. Mitotic recombination in yeast: elements controlling its incidence. Yeast 16: 731-754.

Alexopoulos CJ, Mims CW, Blackwell M. 1996. Introductory Mycology. John Wiley and Sons, Inc., New York.

Allard RW. 1975. The mating system and microevolution. Genetics 79: 115-126.

Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reeb CA, Saunders NC. 1987. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annual Review of Ecology and Systematics 18: 489-522.

Avise JC. 1994. Molecular Markers, Natural History and Evolution. Chapman and Hall, New York.

Bertout S, Renaud F, Barton R, Symoens F, Burnod J, Piens MA, Lebeau B,Viviani MA, Chapuis F, Bastide JM, Grillot R, Mallie M. 2001. Genetic polymorphism of Aspergillus fumigatus in clinical samples from patients with invasive Aspergillosis: Investigation using multiple typing methods. Journal of Clinical Microbiology 39: 1731-7

Burt A, Dechairo BM, KoenigGL, Carter DA, White TJ, Taylor JW. 1997. Molecular markers reveal differentiation among isolates of *Coccidiodes immitis* from California, Arizona and Texas. Molecular Ecology 6: 781-786.

Carter DA, Burt A, Taylor JW, Koenig GL, White TJ. 1996. Clinical isolates of *Histoplasma capsulatum* from Indianapolis, Indiana, have a recombining population structure. Journal of Clinical Microbiology 34: 2577-2584.

Chavez S, Beilharz T, Rondon AG, Erdjument-Bromage H, Tempst P, Svejstrup JQ, Lithgow T, Aguilera A. 2000. A protein complex containing Tho2 Hpr1, Mft1 and a novel protein, Thp2, connects transcription elongation with mitotic recombination in *Saccharomyces cerevisiae*. EMBO Journal 19: 5824-5834.

Chen RS, Boeger JM, McDonald BA. 1994. Genetic stability in a population of a plant-pathogenic fungus over time. Molecular Ecology 3: 209-218

Clutterbuck AJ. 1992. Sexual and parasexual genetics of *Aspergillus* species. Biotechnology 23: 3-18.

Cornwell EE, Belzberg H. 1995. The pattern of fungal infections in critically ill surgical patients. American Surgeon 61: 847-851.

Dean DA, Burchard KW. 1998. Surgical perspective on invasive *Candida* infections. World Journal of Surgery 22: 127-134.

Denning DW, Clemons KV, Hanson LH, Stevens DA. 1990. Restriction endonuclease analysis of total cellular DNA of Aspergillus fumigatus isolates of geographically and epidemiologically diverse origin. Journal of Infectious Diseases 162: 1151-1158.

Dobzhansky T. 1937. Genetics and the Origins of Species. Columbia University Press, New York.

Durand N, Reymond P, Fevre M. 1993. Randomly amplified polymorphic DNAs assess recombination following an induced parasexual cycle in *Penicillium roqueforti*. Current Genetics 24: 417-420.

Edmond MB, Wallace SE, McClish DK, Pfaller MA, Jones RN, Wenzel RP. 1999. Nosocomial bloodstream infections in United States hospitals: a three-year analysis. Clinical Infectious Diseases 29: 239-244.

Fisher RA. 1930. The Genetical Theory of Natural Selection. Oxford University Press, Oxford.

Fisher F, Cook NB. 1998. Fundamentals of Diagnostic Mycology. W.B. Saunders Company, Philadelphia.

Fisher MC, Koenig GL, White TJ, Taylor JW. 2000. Pathogenic clones versus environmentally driven population increase: analysis of an epidemic of the human fungal pathogen *Coccidiodes immitis*. Journal of Clinical Microbiology 38: 807-813.

Fisher MC, Koenig GL, White TJ, San-Blas G, Negroni R, Alvarex IG, Wanke B, Taylor JW. 2001. Biogeographic range expansions into South America by *Coccidiodes immitis* mirrors New World patterns of human migration. Proceedings of the National Academy of Sciences USA 98:4558-4562.

Franzot AP, Hamdan JS, Currie BP, Casadevall A. 1997. Molecular epidemiology of *Cryptococcus neoformans* in Brazil and the United States: Evidence for both local genetic differences and a global clonal population structure. Journal of Clinical Microbiology 35: 2243-2251.

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

Goldman DL, Khine H, Abadi J, Lindenberg DJ, Pirofski L, Niang R, Casadevall A. 2001. Serologic evidence for *Cryptococcus neoformans* infection in early childhood. Pediatrics 107.

Gräser Y, Volovsek M, Arrington J, Schonian G, Presber W, Mitchell TG, Vilgalys R. 1996. Molecular markers reveal that population structure of the human pathogen *Candida albicans* exhibits both clonality and recombination. Proceedings of the National Academy of Sciences USA. 93:12473-12477.

.

Gu Y-H, Ko W-H. 1998. Occurrence of a parasexual cycle following the transfer of isolated nuclei into protoplasts of *Phytophthora parasitica*. Current Genetics 134: 120-123.

Hewitt GM. 2000. The genetic legacy of the Quaternary ice ages. Nature 405: 907-913.

Karthaus M, Doellmann T, Klimasch T, Elser C, Rosenthal C, Ganser A, Heil G. 2000. Intensive intravenous amphotericin B for prophylaxis of systemic fungal infections. Results of a prospective controlled pilot study in acute leukemia patients. Chemotherapy 46: 293-302.

Kwon-Chung J. Morphogenesis of *Filobasidiella neoformans*, the sexual state of *Cryptococcus neoformans*. Mycologia 68: 821-833.

Lachance MA, Bowles JM, Mueller C, Starmer WT. 2000. On the biogeography of yeasts in the *Wickerhamiella* clade and description of *Wickerhamiella lipophila sp. nov.*, the teleomorph of *Candida lipophila*. Canadian Journal of Microbiology 46: 1145-1148.

Marques SA, Robles AM, Tortorano AM, Tuculet MA, Negroni R, Mendes RP. 2000. Mycoses associated with AIDS in the Third World. Medical Mycology 38 S1: 269-279.

Milgroom MG. 1996. Recombination and the multilocus structure of fungal populations. Annual Review of Phytopathology 34: 457-477.

Oerke EC, Dehne HW, Schönbeck, Weber A. 1999. Crop production and crop protection: Estimated losses in major food and cash crops. Elsevier Science, New York.

Ossman F, Adriance M, McCready S. 2000. The genetic control of spontaneous and UV-induced mitotic intrachromosomal recombination in the fission yeast *Schizosaccharomyces pombe*. Current Genetics 38: 113-125.

Pfaller MA, Lockhart SR, Pujol C, Swails-Wenger JA, Messer SA, Edmond MB, Jones RN, Wenzel RP, Soll DR (1998) Hospital specificity, region specificity, and fluconazole resistance of *Candida albicans* bloodstream isolates. Journal of Clinical Microbiology 36, 1518-29.

Poulter R, Jeffery K, Hubbard MJ, Shepherd MG, Sullivan PA. 1981. Parasexual genetic analysis of *Candida albicans* by spheroplast fusion. Journal of Bacteriology 146: 833-840.

Puig S, Querol A, Barrio E, Pérez-Ortín J. 2000. Mitotic recombination and genetic changes in *Saccharomyces cerevisiae* during wine fermentation. Applied and Environmental Microbiology 66: 2057-2061.

Rattray AJ, McGill CB, Shafer BK, Strathern JN. 2001. Fidelity of mitotic double-strand-break repair in *Saccharomyces cerevisiae*: a role for SAE2/COM1. Genetics 158: 109-122.

Rees JR, Pinner RW, Hajjeh RA, Brandt ME, Reingold AL. 1997. The epidemiological features of invasive mycotic infections in the San Francisco Bay area, 1992-1993: Results of population-based laboratory active surveillance. Clinical Infectious Diseases 27: 1138-1147.

Rodrigues E, Meeüs TD Mallie M, Renaud F, Symoens F, Mondon P, Piens MA, Lebeau B, Viviani MA Grillot R, Nolard N, Chapuis F, Tortorano AM, Bastide JM. 1996. Multicentric epidemiological study of *Aspergillus fumigatus* isolates by multilocus enzyme electrophoresis. Journal of Clinical Microbiology 34: 2559-2568.

Rosa CA, Lachance MA. 1998. The yeast genus *Starmerella gen. nov*. and *Starmerella bombicola sp. nov.*, the teleomorph of *Candida bombicola* (Spencer, Gorin, and Tullock) Meyer and Yarrow. International Journal of Systematic Bacteriology 48: 1413-1417.

Schonian G, Forche A, Tietz HJ, Muller M, Graser Y, Vilgalys R, Mitchell TG, Presber W. 2000. Genetic structure of geographically different populations of *Candida albicans*. Mycoses 43 Suppl 2: 51-56.

Sigler L. 1996. *Ajellomyces crescens sp. nov.*, taxonomy of *Emmonsia spp.*, and relatedness with *Blastomyces dermatitidis* (teleomorph *Ajellomyces dermatitidis*). Journal of Medical and Veterinary Mycology 34: 303-314.

Slatkin M. 1985. Gene flow in natural populations. Annual Review of Ecology and Systematics 16: 393-430.

Slatkin M. 1987. Gene flow and the geographic structure of natural populations. Science 236: 787-792.

Taylor JW, Geiser DM, Burt A. Koufopanou V. 1999a. The evolutionary biology and population genetics underlying fungal strain typing. Clinical Microbiology Reviews 12: 126-146.

Taylor JW, Jacobson DJ, Fisher MC. 1999b. The evolution of asexual fungi: Reproduction, speciation and classification. Annual Review of Phytopathology 37: 197-246.

Tibayrenc M, Kjellberg F, Arnaud J, Oury B, Breniere SF, et al., 1991. Are eukaryotic microorganisms clonal or sexual? A populations genetics vantage. Proceedings of the National Academy of Sciences USA 88: 5129-5133.

Weir B. 1996. Genetic Data Analysis II. Sinauer Associates, Incorporated, Massachusetts.

Wenzel RP, Edmond MB. 2001. The impact of hospital-acquired bloodstream infections. Emerging Infectious Diseases 7: 174-177.

Wright S. 1969. Evolution and the Genetics of Natural Populations, vol 2. The University of Chicago Press, Chicago.

Xu JP, Vilgalys R, Mitchell TG. 1999. Multiple gene genealogies reveal recent dispersion and hybridization in the human pathogenic fungus *Cryptococcus neoformans*. Molecular Ecology 9: 1471-1481.