INTER-VARIETY HYBRIDIZATION IN *CRYPTOCOCCUS NEOFORMANS*

GENETIC ANALYSES OF INTER-VARIETY HYBRIDIZATION IN THE HUMAN PATHOGENIC FUNGUS *CRYPTOCOCCUS NEOFORMANS*

By
SHENG SUN, M.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

McMaster University

DOCTOR OF PHILOSOPHY (2009) MCMASTER UNIVERSITY
(DEPARTMENT OF BIOLOGY) HAMILTON, ONTARIO

TITLE: GENETIC ANALYSES OF INTER-VARIETY HYBRIDIZATION IN THE HUMAN PATHOGENIC FUNGUS *CRYPTOCOCCUS NEOFORMANS*

AUTHOR: SHENG SUN, M. Sc. (MCMASTER UNIVERSITY)

SUPERVISOR: DR. JIANPING XU

NUMBER OF PAGES: ix, 175

ABSTRACT

Cryptococcus neoformans is among the most common human pathogenic fungi and the leading cause of fungal meningitis. It consists of two divergent haploid serotypes A and D, as well as their hybrid serotypes AD. Hybridization between the two serotypes A and D is a widespread phenomenon in natural populations of this species. Studies have shown that serotype AD strains possess unique properties in several medically important traits: they are more virulent in animal models, more tolerant to antifungal drugs, more tolerant to UV radiation and other stresses than one or both parental serotypes. Despite the potential medical and evolutionary importance, relatively little is known about the phenotypic and genotypic consequences of the hybridization in *C. neoformans*. In my thesis, I present a series of studies that show: (1) recombination occurs at a very low frequency during hybridization between serotypes A and D; (2) serotypes A and D in *C. neoformans* have diverged from each other not only at the DNA sequence level, but also at the chromosomal structure level, both of which contributed to suppressed recombination and limited genetic introgression between the two serotypes; (3) though there is widespread hybridization, the hybrids seemed largely distinct and the genomes of the parental serotypes A and D populations are relatively unaffected by the hybrids in natural populations; and (4) the fixation of these serotype specific chromosomal rearrangements in the two serotypes, together with evidences of selective sweeps observed at both inversion junctions and genes located close to chromosomal inversions suggest that these rearranged regions (or genes surrounding these regions) were likely involved in the diversification between the two serotypes. My studies provide a genetic framework for future quantitative analyses of medically important traits of *C. neoformans*.

ACKNOLEDGEMENTS

First and foremost I want to thank my supervisor Dr. Jianping (JP) Xu, to whom I am deeply in dept for his mentoring during my M.Sc. and Ph.D. study. It has been an honor to be his student. And I deeply appreciate all his contributions of time, ideas, insights, suggestions and funding to make my Ph.D. pursuit a fulfilling and overall joyful experience. I cannot say enough about how thankful I am for his patience, support and encouragement that have carried me on, especially through the difficult times during my study. Without him, this dissertation would not have been possible.

I thank the rest of my thesis committee members: Dr. Rama Singh, Dr. Jon Stone and Dr. Ben Evans. Their valuable feedback helped me to improve the dissertation in many ways. My sincere gratitude also goes to my former committee members, Dr. Brian Golding and Dr. Turlough Finan, from whom I have learned a great deal.

I am fortunate to have the opportunity to work with a group of energetic people in Dr. Xu's lab. I have enjoyed every moment that we have shared together, and I have learned many things about science, as well as life from all of you. I want especially to thank Timothy James, Irina Skosireva, Mochan (Sammi) Li, Hong Guo, Zhun Yan, Jianrong Wu, Ping Zhang, Mori Shahid, Susan Han, Rovena Tey and Xiaogang Li.

Many people on the faculty and staff of the Biology Department have helped, assisted and encouraged me in various ways during my course of study. I want to take this opportunity to say thank you to all of them.

Finally, it is impossible to have my research career without the love, support, and encouragement from my parents, my family, and my friends. I want to thank my wife, Li Xu, and two sons, Dennis and Austin, for their love and understanding. This dissertation is dedicated to them.

Sheng Sun
McMaster University
Summer, 2009

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

GENERAL INTRODUCTION

Hybridization and Speciation

Hybridization is conventionally defined as successful mating between individuals from two populations, or groups of populations, which are distinguishable on the basis of one or more heritable characters (HARRISON, 1993; ARNOLD, 1997). Hybridization has long been observed in animals and plants. However, the role of hybridization in evolution has been debated for more than a century and for a long time there are two opposite viewpoints among biologists.

The first regards hybridization as evolutionary noise with only transient effects in evolution, based on the facts that plant and animal hybrids are often not viable or sterile, due to the genetic incompatibility generated during population divergence. The widely accepted theoretical basis of the hybrid incompatibilities is the well-known Dobzhansky-Muller model, which proposes that hybrid incompatibilities evolve as a consequence of interactions between or among genes that have diverged in each of the hybridizing populations (COYNE and ORR 2004). Specifically, because hybridization brings together genetic materials that have experienced different adaptation and natural selection, the divergent genetic materials may not be compatible with each other when they co-exist within the same cell or individual. As a result, many hybrid progeny are either sterile or inviable (i.e. zero fitness and an evolutionary dead end), or have severely reduced fitness and are eventually out-competed by the parental populations (i.e. transient existence). Under this scenario, hybridization events will reinforce the reproductive isolation between the two parental populations. Many examples consistent with this view of hybridization have been reported in animals (e.g. *Drosophila*; BRIDEAU *et al.* 2006) and in plants (e.g. Minulus; SWEIGART *et al.* 2006).

The contrasting position considers hybridization as a potent evolutionary force that creates opportunities for speciation and adaptive evolution (SCHEMSKE 2000; BARTON 2001). The argument is that because hybridization brings together divergent and separately adapted genomes into the same cell, many novel gene interactions could occur (e.g. complementary action of additive alleles, overdominance, epistasis and unmasking of rare recessive alleles), thus generating extreme (transgressive segregation) and/or novel phenotypes. These extreme or new phenotypes could be of significant adaptive potential and result in introgression between divergent populations. In heterogeneous environments and with niche separation between the hybrids and parental populations, the hybrids could be permanently established (RIESEBERG et al. 1999a; RIESEBERG et al. 2003). Under certain circumstances, the hybrids can even replace one or both of the original parental populations. Examples consistent with this argument have been observed in many plant species, such as Irises (Johnston et al. 2003), as well as in some animal species, such as Darwin's Finches (GRANT and GRANT 2002).

Studies have also shown that hybridization has occurred in many microbial populations, such as the opportunistic human pathogenic yeast *Cryptococcus neoformans* (XU et al. 2000, 2002; XU and MITCHELL 2003) and many plant pathogenic fungi (see BRASIER 2000). Three unique characteristics of microorganisms may allow hybridization to have even more impact in their evolution than those in plants and animals. First, most microorganisms can reproduce either sexually or asexually, which means there is no traditional "evolutionary dead end" in microorganisms and even the "sterile" hybrid progeny can have continuous evolutionary effects. Second, most microorganisms have short life span and large population sizes. Such features generate a greater probability that during microbial hybridization, hybrid progeny with fitness higher than parental populations will emerge. Third, microorganisms are usually small and thus more sensitive to environmental changes. The small niches and high sensitivity and responsiveness mean that there are more heterogeneous environments for natural selection to work and more novel niches in which hybrid progeny can out-compete the existing populations.

Hybridization and Recombination

The importance of hybridization in evolution is closely related to the rate of recombination that occurs during hybridization events. This is because given the linkage disequilibrium established between the parental species, the higher recombination rates will generate hybrids that are more genetically heterogeneous and will increase chances to produce progeny with novel or extreme phenotypes that have adaptive potential to new environments. Here, recombination refers to the process of genetic mixing, and that may include either the acquisition of new genes (e.g. allopolyploidy hybridization) or, more commonly, the generation of new combinations of existing alleles at multiple homologous loci (MICHOD and LEVIN, 1988).

Within species, most recombination occurs in meiosis during sexual reproduction. The crossing-over and recombination during meiosis have several important functions: 1) assuring proper chromosome pairing and segregation during meiosis (HUNTER et al., 1996), 2) reshuffling alleles from different individuals to produce progeny with genotypes that differ from either parent. Such reshuffling may generate progeny with genotypes better able to adjust to temporal or spatial environmental changes, 3) facilitating the incorporation of beneficial mutations and the removal of deleterious mutations, and 4) preventing populations from diverging into subpopulations (HARTL and CLARK, 1997).

Many studies have shown that recombination frequencies during hybridization are much lower than they would normally be in crosses between closely related individuals. The low rate of recombination might be due to chromosomal length and structural polymorphisms between the divergent populations and such polymorphisms have also been argued to

be the reason why many hybrid progeny are aneuploid (HUNTER et al., 1996). Decreased frequency of recombination during hybridization has also been considered as a driving force behind speciation (COYNE and ORR, 2004). There are three mutually non-exclusive possible explanations for the reduced level of recombination observed in hybridization events. The first is that the lower sequence similarity between strains involved in hybridization may prevent effective homologous DNA pairing, through the action of mismatch repair system, and thus reduce the occurrence of recombination (SELVA et al. 1995; CHAMBERS et al. 1996; DATTA et al. 1996; HUNTER et al. 1996; DATTA et al. 1997; CHEN et al. 1999; HESS et al. 2002; reviewed in EVANS and ALANI 2000). Secondly, recombination may be repressed by the chromosomal rearrangements between the diverging parental populations (RIESEBERG 2001; See below). And thirdly, the disassociation of epistatic interactions between different alleles established within each population results in hybrids that are inviable or that have decreased fitness. As a result, non-recombinant progeny are more likely to be recovered from hybridization events, thus reducing the observed effective recombination frequency.

Chromosomal Rearrangements

In a broad sense, chromosomal rearrangements refer to any changes in nucleotide sequences other than point mutations and small insertion/deletions. The broad-sense definition thus includes fusion or fission of chromosomes, duplication or deletion of chromosomal segments, inversion of a chromosomal segment, change of position of a chromosomal segment between homologous chromosomes (i.e. transposition) or non-homologous chromosomes (i.e. translocation). Chromosomal rearrangements are ubiquitous. They have been found in bacteria (e.g. see review by Hughes 2000) as well as in diverse groups of eukaryotic organisms such as Baker's yeast (*Saccharomyces cerevisiae*), fruit flies (*Drosophila melanogaster*), mosquitoes (*Anopheles gambiae*), and humans (FISCHER et al. 2006; COLUZZI et al. 2002; RANZ et al. 2007; for recent reviews, see RIESEBERG 2001, HOFFMANN et al. 2004, and COGHLAN et al. 2005).

Similar to hybridization, there has been a long debate about the role of chromosomal rearrangements in evolution and speciation. The widely held view is that the accumulation of chromosomal rearrangements between diverging populations is incidental to speciation (FUTUYMA and MAYER 1980; SITES and MORITZ 1987; COYNE and ORR 1998). This argument is mainly based on the fitness effects of chromosomal rearrangements, and the main reasons include: (1) empirical evidence showing that many chromosomal rearrangements have little direct fitness effects (SITES and MORITZ 1987; COYNE et al. 1993); (2) theoretical difficulties associated with fixing chromosomal rearrangements that are strongly underdominant (WALSH 1982; LANDE 1985).

On the other hand, even if chromosomal rearrangements do not have significant direct phenotypic effects, their other features, e.g. recombination repression, can play important roles in speciation process. The reasoning is that because chromosomal rearrangement can effectively reduce the recombination frequency in their surrounding chromosomal regions, any lineage-specific adaptations accumulated within those regions will be protected from recombination and gene flow between the diverging populations, thus facilitating the differentiation process. Indeed, significantly lower levels of recombination frequency and gene introgression have been reported in regions surrounding chromosomal rearrangements compared to collinear chromosomal regions in several different species, such as *Drosophila* (MACHADO et al. 2007) and sunflowers (*Helianthus*) (RIESEBERG et al. 1995; 1999b).

The repression of recombination by chromosomal rearrangement during hybridization may be achieved through two processes. First, chromosomal rearrangements can reduce the efficiency in pairing up between homologous chromosomes. As a result, crossing over and recombination occur at low levels during hybridization, especially around the rearrangement regions. Second, recombination occurring within chromosomal rearrangements usually result in chromosomes that have abnormal structures (e.g. large duplication/deletion, non-centromeric or bi-centromeric). Progeny possessing these types of chromosomes usually have very low fitness or are inviable, leading to the selective recovery of non-recombinant chromosomes in the hybrid progeny and an effective reduction in recombination frequency.

Cryptococcus neoformans

Cryptococcus neoformans is an encapsulated basidiomycetous yeast that can infect the central nervous system to cause meningoencephalitis mostly in immunocompromised hosts. As the result of the pandemic of HIV infections, *C. neoformans* has become a major, worldwide pathogen in humans and has recently received considerable research attention. The phenotypic characteristics proven to be associated with *C. neoformans* virulence include: 1) the presence of a polysaccharide capsule, 2) the production of a cellular pigment melanin, and 3) the ability to grow at 37°C (CASADEVALL and PERFECT, 1998).

Most of the *C. neoformans* strains are haploid. Based on the difference in their capsular polysaccharides, *C. neoformans* strains can be classified into three serotypes: serotype A (i.e. variety *grubii*), serotype D (i.e. variety *neoformans*) and serotype AD. All three serotypes are distributed worldwide, and their natural habitats include trees, bird droppings and soils. Molecular phylogenetic analyses have shown that var. *grubii* and var. *neoformans* diverged from each other about 18.5 - 40 million years ago, and *C. neoformans* has diverged from its closest related species, *Cryptococcus gattii*, about 37-80 million years ago (XU et al. 2000;

SHARPTON et al. 2008). The genomic sequence of one serotype A strain, H99, has been sequenced and fully assembled, while the genomic sequence of one serotype D strain, JEC21, has already been published (LOFTUS et al., 2005). Both genomes are about 19 Mb in size, and are comprised of 14 chromosomes. The overall nucleotide similarity between the two genomes is around 85-90% (KAVANAUGH et al. 2006).

Strains of *C. neoformans* normally grow as budding yeasts. Under certain environmental conditions, many strains can also undergo yeast-filament dimorphic transitions (ALSPAUGH et al. 2000; IDNURM et al., 2005; LIN and HEITMAN, 2006). *C. neoformans* has a defined sexual cycle with a teleomorph state called *Filobasidiella neoformans*. Strains of *C. neoformans* belong to one of two mating types (MAT) that are determined by one locus with two alternative alleles, MATa and MATalpha. Under suitable environmental conditions (i.e. nitrogen-limiting and low moisture), mating can occur between strains of the same serotype with opposite mating types (a with alpha). Typically, mating starts with the fusion of haploid cells of different mating types, followed by filamentous growth of the dikaryotic cells. At the tip of the dikaryotic cells, basidia may be formed. Nuclear fusion and meiosis then occur in the basidia and four chains of haploid spores are produced on the surface of basidia (KWON-CHUNG, 1975, 1976).

For both serotypes A and D, the majority of natural isolates have been identified to be MATalpha, with MATa strains comprising less than 5% of the natural population (Kwon-Chung and Bennett 1978). Currently, the reason underlying the significant bias in mating types is not fully understood, although some studies have shown that MATalpha strains are more virulent or more competitive in colonizing animal models than MATa strains (KWON-CHUNG et al. 1992; NEILSON et al. 2003, 2005a, 2005b). Interestingly, mating between cells with same mating type, MATalpha, has recently been observed in laboratory (LIN et al. 2005). Subsequent studies have also shown evidence that same sex mating between MATalpha cells have occurred / are occurring in natural *C. neoformans* populations (FRASER et al. 2005; LIN et al. 2007).

There have been extensive genotyping studies of *C. neoformans* natural isolates. Using molecular markers, most *C. neoformans* isolates have been grouped into one of four major molecular genotypes. Specifically, VN-I and VN-II contain mostly serotype A, VN-III contains mostly serotype AD and VN-IV contains mostly serotype D (MEYER et al. 1999; ELLIS et al. 2000; BOEKHOUT et al. 2001; LITVINTSEVA et al. 2003, 2005, 2006).

However, despite the fact that *C. neoformans* can undergo sexual reproduction, most population genetics studies suggested natural *C. neoformans* populations are overall clonal (CURRIE et al. 1994; BRANDT et al. 1996; BOEKHOUT et al. 2001; MEYER et al. 2003; LITVINTSEVA et

al. 2005), with some evidence of recombination (BURT et al. 2000; XU et al. 2000; XU and MITCHELL 2003). However, sexual reproduction may be more common in certain geographic areas than others. For example, a distinct group of serotype A stains (VN-B) was discovered in Botswana, and studies have shown that these isolates are highly variable in their genotypes and show evidence of sexual recombination (LITVINTSEVA et al. 2003, 2005, 2006).

Natural Hybridization in *Cryptococcus neoformans*

In *C. neoformans*, strains of another serotype, serotype AD, have been found in natural environments and in patients. Molecular analyses have shown that most environmental and clinical serotype AD strains are diploid or aneuploid and contain alleles typical of both serotypes A and D. Gene genealogical analyses also indicated that serotype AD strains are recent hybrids between strains of serotypes A and D and that multiple hybridization events have occurred between strains of these two serotypes (XU et al. 2002; XU and MITCHELL 2003). Consistent with these results, studies have shown that serotypes A and D strains of *C. neoformans* can grow and successfully mate on medium containing pigeon guano, a natural habitat for strains of these two serotypes (STAIB, 1981; NIELSEN et al. 2007).

The fact that hybridization between serotypes A and D *C. neoformans* produces progeny that are mostly diploid or aneuploid, as well as the abundant heterozygosity observed in serotype AD strains suggest that the meiotic recombination during hybridization is likely abnormal, possibly due to sequence divergence, chromosomal rearrangements or epistatic interactions. Indeed, significant karyotypic variations have been found in *C. neoformans* that could have influenced chromosomal pairing and segregation during meiosis (KWON-CHUNG et al. 1992; WICKES et al. 1994; MARRA et al. 2004; FRASER et al. 2005).

Studying hybridization events in *C. neoformans* is important to better understand their population structure and dynamics, pathogenicity, speciation and evolution. For example, in some geographical areas, serotype AD strains are more prevalent in clinical isolates than haploid serotype D strains (BRANDT et al. 1996). It has also been shown that serotype AD hybrid strains are overall more tolerant to UV radiation, as well as the common anti-fungal drug Fluconazole (XU et al. 2001). However, despite the medical importance of the AD hybrids and the discovery of recent natural hybridization events between serotypes A and D strains, relatively little is known about the phenotypic and genotypic consequences of hybridization in *C. neoformans*.

Thesis Objectives

At present, although more evidence of past hybridization is being discovered in plants and animals (ARNOLD, 1997), and to a lesser extent in microorganisms (XU et al. 2000; XU et al. 2002; SCHARDL and CRAVEN, 2003; XU and MITCHELL, 2003), the genotypic and phenotypic consequences of hybridization remain largely unknown, and the role and mechanisms of natural hybridization in long-term evolution and adaptation remain poorly understood. Using the human pathogenic yeast *Cryptococcus neoformans* as a model organism, my thesis examines the following three sets of questions:

First, what are the patterns of meiotic segregation and recombination in an inter-variety cross of *C. neoformans*? Are there differences in segregation patterns and recombination rates among chromosomes, or among regions within chromosomes? [Chapter 2]

Second, how many unambiguous chromosomal rearrangements are there between serotypes A and D in *C. neoformans*? Are these regions correlated to the chromosomal regions showing significantly low recombination frequencies during inter-variety hybridization? And if they are, how much do they contribute to the repression of recombination during the inter-variety hybridization of *C. neoformans*? In addition, what are the distributions of different chromosomal types at these rearrangement regions in natural *C. neoformans* populations? [Chapter 3]

Third, what are the effects of the chromosomal rearrangements on DNA polymorphism and population structures in natural populations of serotypes A, D and AD *C. neoformans*? [Chapter 4]

Answers to these questions will significantly enhance our understanding of genetic differentiation, population structures and dynamics, and the effect of hybridization in *C. neoformans* populations, as well as in other fungal species.

CHAPTER 2

GENETIC ANALYSES OF A HYBRID CROSS BETWEEN SEROTYPES A AND D STRAINS OF HUMAN PATHOGENIC FUNGUS *CRYPTOCOCCUS NEOFORMANS*

PREFACE

Population and epidemiological surveys have identified widespread natural hybridization between strains of serotypes A and D *Cryptococcus neoformans*. At present, relatively little is known about the phenotypic and genotypic consequences of the hybridization. Because precise parental strains of natural hybrids of *C. neoformans* cannot be determined with certainty, the genotypic and phenotypic consequences of hybridization cannot be unambiguously inferred by examining only natural strains. Therefore, to determine the consequences of hybridization between strains of serotypes A and D, I examined a laboratory hybridization between two known strains and directly analyzed a large number of meiotic progeny for their genotypes

Chapter 2 has been published in the journal Genetics:
Sun, S., and J. Xu, 2007 Genetic analyses of A hybrid cross between serotypes A and D strains of the human pathogenic fungus *Cryptococcus neoformans*. Genetics 177: 1475-1486.

Sheng Sun was the primary researcher of this study. He carried out all the experiments and wrote the draft of the manuscript.

ABSTRACT

Cryptococcus neoformans has two varieties, var. *grubii* and var. *neoformans* that correspond to serotypes A and D, respectively. Molecular phylogenetic analyses suggest that these two varieties have diverged from each other for about 18 million years. The discovery of pathogenic serotype AD hybrid strains in nature indicates that inter-variety mating in *C. neoformans* occurs in natural environments. However, little is known about the genetic consequences of hybridization in *C. neoformans*. Here, we analyzed a hybrid population of 163 progeny from a cross between strains of serotypes A (CDC15) and D (JEC20) using 114 co-dominant nuclear PCR-RFLP markers and one direct PCR marker. These markers were distributed on all 14 chromosomes of the sequenced strain JEC21 that was isogenic to one of the parents (JEC20) in our cross. Our analyses identified that of the 163 progeny, five were heterozygous at all 115 loci, one was completely homozygous and identical to one of the parents (CDC15), and the remaining 157 each contained at least one heterozygous locus. Because all 163 progeny inherited mitochondria from the MATa parent JEC20, none of the progeny had a genotype identical to either of the two parents or a composite of the two parents. All 115 nuclear loci showed three different genotypes in the progeny population, consistent with Mendelian segregation during meiosis. While the linkage analysis showed independent re-assortment among loci on different linkage groups, there were significant differences in recombination frequencies among chromosomes and among regions within certain chromosomes. Overall, the linkage map length from this hybrid cross was much shorter and the recombination frequency much lower than those constructed using serotype D strains, consistent with suppressed recombination in the inter-variety cross between strains of serotypes A and D. We discuss the implications of our results in our understanding of the speciation and evolution of the *C. neoformans* species complex.

INTRODUCTION

Cryptococcus neoformans is an encapsulated basidiomycetous yeast that can infect the central nervous system to cause meningoencephalitis in immunocompromised hosts. Most of the *C. neoformans* strains are haploid and belong to two different serotypes, A and D, corresponding to variety *grubii* and variety *neoformans* respectively. Molecular phylogenetic analyses have shown that var. *grubii* and var. *neoformans* have diverged from each other for about 18.5 million years (XU et al. 2000). Because both varieties of *C. neoformans* are significant opportunistic pathogens of humans and other animals, in recent years, there have been significant research activities aimed at understanding the genotypic and phenotypic differences between the varieties. However, much remains unknown. The objective of this study is to analyze the patterns of molecular marker segregation in a hybrid cross between strains of var. *grubii* and var. *neoformans* in an effort to help improve our understanding of the genetic consequences of hybridization in this species.

Strains of *C. neoformans* normally grow as budding yeasts. Under certain environment conditions, many strains can also undergo filamentous dimorphic transitions (reviewed in ALSPAUGH et al. 2000). *C. neoformans* has a defined sexual cycle with a teleomorph state called *Filobasidiella neoformans*. Strains of *C. neoformans* belong to one of two mating types (MAT) that are determined by one locus with two alternative alleles, MATa and MATalpha. Under suitable environmental conditions (i.e. nitrogen-limiting and low moisture conditions), mating can occur between strains of opposite mating types. Typically, mating starts with the fusion of haploid cells of different mating types, and is followed by filamentous growth of the dikaryotic cells. At the tip of the dikaryotic cells, basidia may be formed, nuclear fusion and meiosis can occur and haploid spores are produced on the basidia (KWON-CHUNG, 1975; KWON-CHUNG, 1976).

It has been shown that strains of serotypes A and D in *C. neoformans* can grow and successfully mate on medium containing pigeon guano, a natural habitat for strains of these two serotypes (STAIB, 1981; NIELSEN et al. 2007). This result suggests that mating and sexual reproduction could occur in natural environments between serotypes A and D strains. Consistent with this hypothesis, serotype AD strains have been found in natural environments and in patients (e.g. BRANDT et al. 1996). These serotype AD strains are also virulent in the murine model of Cryptococcosis (LENGELER et al. 2001; CHATURVEDI et al. 2002; BARCHIESI et al. 2005). Indeed, gene genealogical analyses demonstrated that serotype AD strains are recent hybrids between strains of serotypes A and D and that multiple hybridization events have occurred between strains of these two serotypes (XU et al. 2002; XU and MITCHELL 2003).

Molecular analyses have shown that most environmental and clinical strains of serotype AD are diploid or aneuploid and contain alleles typical of both serotypes A and D (XU et al. 2000; BOEKHOUT et al. 2001; LENGELER et al. 2001; CHATURVEDI et al. 2002; XU et al. 2002; XU and MITCHELL 2003). The abundant heterozygosity in serotype AD strains suggests that meiosis in these hybrid zygotes might be impaired, due possibly to the large genomic differences among serotypes. Indeed, significant karyotypic variations have been found in *C. neoformans* (KWON-CHUNG et al. 1992; WICKES et al. 1994; MARRA et al. 2004; FRASER et al. 2005). However, despite the medical importance of the AD hybrids and the discovery of recent natural hybridization events between serotypes A and D strains, relatively little is known about the genotypic consequences of hybridization in *C. neoformans*.

Here, we analyzed a hybrid progeny population generated from an inter-variety cross between strains CDC15 (serotype A, MATalpha) and JEC20 (serotype D, MATa). These two strains differ in several phenotypic traits. For example, CDC15 is resistant to the anti-fungal drug Fluconazole (MIC = 64 μg/ml) while JEC20 is not (MIC = 4 μg/ml). We obtained genotype data for 115 broadly distributed codominant molecular markers for each of 163 progeny from this cross. The obtained hybrid linkage map was then compared to genetic linkage maps constructed from between serotype D strains reported previously (FORCHE et al. 2000; MARRA et al. 2004). The genotype and linkage map information were used to understand the patterns of marker segregation and recombination within and among chromosomes in this inter-variety cross in *C. neoformans*

MATERIALS AND METHODS

Mapping population

We used a mapping population of 163 progeny, generated by a cross between *C. neoformans* serotype A strain CDC15 (MATalpha) and serotype D strain JEC20 (MATa). To construct the mapping population, parental strains were first grown on YEPD medium (1% yeast extract, 2% dextrose, 2% Bacto-peptone, 1.5% agar) at 25°C for three days. About 10^8 cells from each parental strain were then thoroughly mixed on V8-juice agar medium (XU et al. 2000). After four weeks of incubation at 25°C, hyphal mats were visible at the edge of the mating mixture. Hyphae and basidiospores were scraped off the agar surface from the edge of the mating mixture (i.e. without any parental yeast cells), washed in sterile distilled water, diluted and spread-plated on YEPD medium. Plates were then incubated at 37°C for three days. Well-separated single colonies without any obvious hyphal filaments from the original mating were streaked onto a new YEPD plate to obtain pure cultures. Only one colony from the second plate was picked from each original colony so as to maximize the genotypic diversity of the progeny population. DNA was

extracted from each progeny as well as the parental strains according to an established procedure (XU et al. 2000).

Codominant Molecular Markers and Genotyping

PCR primers were designed based on the published sequence of *C. neoformans* strain JEC21 (LOFTUS et al. 2005), which is isogenic to one of the parental strains in our cross, JEC20, except at the mating type locus (JEC21 is MATalpha while JEC20 is MATa). For each chromosome of the annotated JEC21 genome, starting from one terminus, genes were selected at the frequency of about one gene every 150 kb. Primers were designed for each selected gene using an online program (http://seq.yeastgenome.org/cgi-bin/web-primer). PCR primer pairs that successfully amplified the expected sized-DNA fragments in both parental strains were selected. The PCR products from the two parental strains, as well as from an equal mixture of the genome DNA of the two parental strains (the positive control for heterozygosity), were then digested separately with each of 12 restriction enzymes. For each PCR product, the enzymatic digestion that produced band patterns easily distinguishable between the two parents was chosen to further genotype the entire mapping progeny population. For primer pairs that did not work with either parental strains or failed to produce co-dominant enzymatic digestion patterns, we designed new primer pairs from genes that were located close to the initially selected genes to try to find suitable PCR-RFLP markers. For some regions where suitable PCR-RFLP markers could not be found after several tries, no molecular marker was included in the analysis. Overall, we tried to have at least two markers from each chromosome located within 200 kb from both ends of the chromosome. The only exception was chromosome 5, for which the closest marker was about 300 kb away from one end of this chromosome. A total of 114 PCR-RFLP markers were developed for this study (Table 2.1).

Genotypes at the mating type locus (MAT) for these progeny were determined by direct PCR using the MATa and MATalpha - specific PCR primers for the ste20 gene, as described previously (LENGELER et al. 2001; YAN et al. 2002).

The 163 progeny were genotyped for each of the 115 markers. For each marker, we used "1" to represent the allele (i.e. the enzymatic digest pattern) from parent CDC15, "2" to represent the allele from JEC20, and "3" to represent the heterozygote that contains alleles from both parental strains (i.e. a composite enzymatic digestion pattern that includes DNA fragments from both parental strains). PCR amplification, enzyme digestion, gel electrophoresis and data scoring followed those in XU et al. (1999) and LAN and XU (2006).

Data analysis and linkage map construction

Of the 163 progeny, 162 have at least one heterozygous locus (see Results below), suggesting that these progeny were either diploid or

aneuploid. Therefore, in our marker segregation and linkage analysis, these progeny were treated as diploid. For each locus, the observed homozygosity for each of the two parental alleles as well as heterozygosity in the progeny population were calculated as simple ratios of the number of progeny in each genotype over the total number of analyzed progeny. The potential bias of the two parental alleles in the progeny population was examined for each locus using the Chi-square goodness of fit test (χ^2 test). The null hypothesis in these tests was that the two parental alleles in the progeny population should be in equal frequency.

Linkage and mapping analyses were performed using the MAPMAKER software, version 3.0 (LANDER and GREEN 1987; LANDER et al. 1987). To generate the linkage map, different values of LOD thresholds and maximum mapping distances (MD) between adjacent markers were tested to determine the optimal parameters for the MAPMAKER program, using the chromosomal organization and genome sequence of JEC21 as a guide. For maximum MD between adjacent markers, we first tried 25 cM, which was similar to those that have been used previously in linkage mapping of serotype D *C. neoformans* (FORCHE et al. 2000; MARRA et al. 2004). With this cut-off, the MAPMAKER program generated a linkage map containing 14 linkage groups (LGs) that were in overall agreement with the 14 chromosomes of the published genome sequence of JEC21 (LOFTUS et al. 2005).

Marker segregation

Of the 115 marker loci, 102 had allele frequencies not significantly different from 1:1 in the progeny population ($\chi^2 < 3.84$, $P > 0.05$) while the remaining 13 exhibited significantly skewed segregation ratios ($\chi^2 > 3.84$, $P < 0.05$) (Figure 2.3). Of these 13 loci that showed skewed allele frequencies, 10 were located on chromosome 2, with alleles from parent JEC20 significantly more prevalent in the progeny population than those from parent CDC15. Indeed, all analyzed markers located on chromosome 2 showed skewed segregations toward JEC20 (Figure 2.3). The remaining three loci were skewed toward CDC15, with one located on chromosome 11 and two on chromosome 14 (Figure 2.3).

Compared to the two genetic linkage maps constructed using serotype D strains by FORCHE et al. (2000) and MARRA et al. (2004), the percentage of markers showing distorted segregation in our study was higher than that found in the MARRA et al. study (1.7%), but lower than that found in the FORCHE et al. study (18.2%). In comparison to other fungal linkage mapping studies, the percentage of skewed markers found in our study was lower than most others. For example, in mapping the *Agaricus bisporus* genome, KERRIGAN et al. (1993) found that 32.8% loci had skewed ratios. In constructing the genetic linkage map of the maize pathogen *Cochliobolus heterostrophus*, TZENG et al. (1992) found segregation distortions in 15.9% of the polymorphic loci. ZHONG et al. (2002) found 19.6% of the polymorphic loci showed skewed segregation during linkage map construction for the ascomycete plant pathogen *Cochliobolus sativus*.

Table 2.1. Summary information of individual linkage groups in the inter-variety cross between strains of serotypes A and D in *C. neoformans.*

Linkage Group	Corresponding Chromosome in JEC21	Number of Markers	Length (cM)	Physical Distance Covered by Markers (kb)	Ratio between Physical Distance and Genetic Distance (kb/cM)
1	1	13	38.5	2135	55.45
2	2	9	12.5	1401	112.08
3	3	11	11.8	1604 [a]	195.61[b]
4	4	11	32.3	1612	49.91
5	5	9	13.9	1144	82.30
6	6	7	12.6	1307	103.73
7	7	12	11.5	1266	110.09
8	8 and 12	10	22.2	n.c. [c]	n.c. [c]
9	9	7	9.1	1146	125.93
10	10	7	7.4	986	133.24
11	11	8	13.6	948	69.71
12	13	6	8.3	722	86.99
13	14	2	2.2	177	80.45
14	14	2	0.9	102	113.33
	Total	114	196.8		

[a]: Marker CNC07180 was not included, because it was identified as a translocation (See results).
[b]: Genetic distance involving marker CNC07180 was not included (see reason above). Therefore, a genetic distance 8.2, not 11.8, was used to calculate the ratio.
[c]: Not calculated, because markers in LG8 were located on two different chromosomes in JEC21

Increasing MD to 30 cM or decreasing MD to 10 cM did not significantly influence the mapping results. However, when an MD of less than 10 cM was used, MAPMAKER generated many small linkage groups with each including only two to four markers.

To select for an optimum LOD threshold for linkage map construction, we first used the LOD score of 5, a value used in previous mapping studies of *C. neoformans* (FORCHE et al. 2000; MARRA et al. 2004). With this LOD value, we obtained a linkage map containing 10 LGs, in which eight LGs corresponded to eight chromosomes of JEC21 and the remaining two LGs corresponding to markers from multiple chromosomes, with one LG containing markers from two chromosomes and the other containing markers from five chromosomes of JEC21. We then increased the LOD threshold in an effort to resolve the two LGs with markers from multiple chromosomes. Overall, increasing the LOD score did not influence the LGs with markers from the eight chromosomes of JEC21. However, when the LOD score was at 20, the initial LG with markers from five chromosomes separated into three LGs, with two of them corresponding to two chromosomes of JEC21 and the other containing markers from three different chromosomes. When the LOD score was further increased to 25, the two remaining LGs containing markers from multiple chromosomes (when LOD of 20 was used) were further resolved into smaller LGs that corresponded to chromosomes of JEC21. However, one LG still contained markers from two chromosomes of JEC21 (chromosome 8 and chromosome 12). Further increasing the LOD threshold to 30 divided this LG into two LGs with each containing markers mostly from chromosome 8 and chromosome 12, respectively. With the incremental LOD score values, several markers became unassigned to any LG. Because with LOD of 25 and MD of 10 cM, all the markers (except one) were assigned to LGs and the linkage map produced by MAPMAKER was overall in good agreement with the 14 chromosomes of JEC21, these two parameters were chosen to construct the linkage map.

In order to analyze the genotype distribution among the 163 progeny and to examine their relationships among each other and with the parents, we performed a simple cluster analysis using the UPGMA algorithm through the PAUP* computer software (SWOFFORD 2002).

Confirmation of chromosomal rearrangements

To confirm the identified putative inversions and translocations (see Results below), we compared the locations and orders of the markers involved in the rearrangements in the published JEC21 genome and compared them to the locations and orders of their orthologs in the H99 genome using BLAST. Because the H99 genome sequence is still not completely assembled into chromosomes (there are 210 super contigs at present in the database), an exhaustive test of potential chromosomal rearrangements identified is not possible at present. Instead, we designed

specific primers to target regions that showed clear rearrangements between JEC21 and H99 genomes to confirm the putative chromosome rearrangement between JEC20 and CDC15 using PCR (See below).

Comparison of recombination frequency between inter-variety and intra-variety crosses

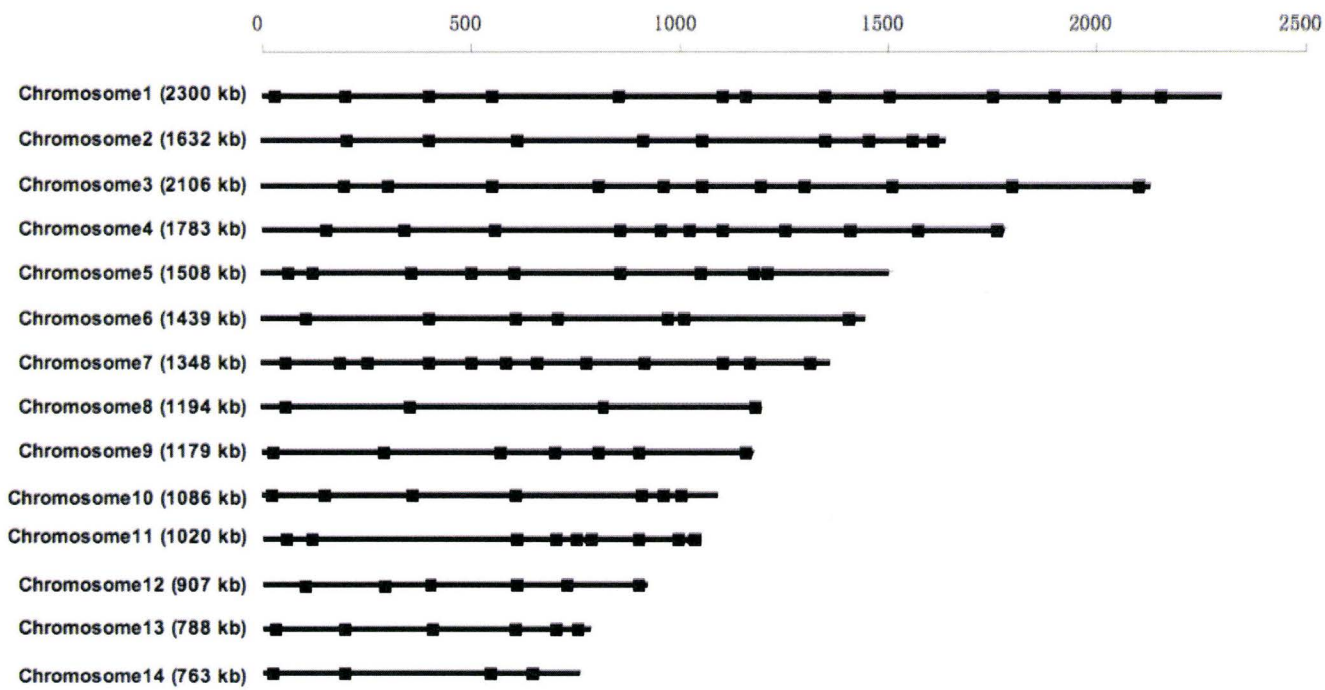
Among the 115 markers analyzed in our study, 15 were used previously in a mapping study of serotype D strains of *C. neoformans* (MARRA et al. 2004). This sharing of markers gave us an opportunity to make a direct comparison of recombination frequency between the inter-variety and the intra-variety crosses of *C. neoformans*. To make the comparison, we first obtained the genetic distance data for each marker pair in the two studies. Genetic distances from the same study were put into one group. Then the two groups of genetic distance data were compared using the Mann-Whitney U test (Wilcoxon rank sum test) to determine whether recombination frequency was significantly lower in the inter-variety cross than that in the intra-variety cross.

RESULTS AND DISCUSSION

PCR-RFLP marker development

In total, we successfully obtained 114 co-dominant PCR-RFLP markers that easily distinguish the two parental strains as well as the heterozygote. In addition, one marker, the ste20 gene located within the mating type region, was assayed based on direct PCR using mating type-specific primers. Alleles at this locus were detected based on the presence/absence of a PCR product using the MATa and MATalpha-specific PCR primers. The ste20 primers were originally designed based on the unique ste20 sequences at the MATa and MATalpha loci respectively by LENGLER et al. (2001). Overall, these 115 markers cover all 14 chromosomes of the genome of JEC21 (Figure 2.1). The physical distances between two adjacent markers ranged from 16kb to 488kb, with an average of 169kb (Supplemental Table S2.1).

Figure 2.1. Distribution of markers along the 14 chromosomes used in this study. The lengths of the chromosomes as well as the positions of the markers analyzed on each chromosome are shown.


0
500
1000
1500
2000
2500
Chromosome1 (2300 kb)
Chromosome2 (1632 kb)
Chromosome3 (2106 kb)
Chromosome4 (1783 kb)
Chromosome5 (1508 kb)
Chromosome6 (1439 kb)
Chromosome7 (1348 kb)
Chromosome8 (1194 kb)
Chromosome9 (1179 kb)
Chromosome10 (1086 kb)
Chromosome11 (1020 kb)
Chromosome12 (907 kb)
Chromosome13 (788 kb)
Chromosome14 (763 kb)

Progeny genotypes

Our progeny population of 163 was derived from single colony isolation and not from single basidiospores dissected using a micromanipulator. Previous studies have shown that basidiospores from crosses between strains of serotypes A and D have very low viability (less than 10%, LENGELER et al. 2001). Therefore, in order to have sufficient number of progeny for linkage map construction and analyses, we used the alternative method of first growing single spores into colonies and then through streaking and purification to obtain pure cultures representing single spores. The size of our progeny population is similar to or larger than those used in linkage analyses in many other studies. For example, in the two studies in which genetic linkage maps of *C. neoformans* var. *neoformans* were generated, 100 and 94 progeny were analyzed, respectively (FORCHE et al. 2000; MARRA et al. 2004). The genetic linkage map for the Oomycete plant pathogen *Phytophthora infestans* was constructed using 73 progeny (VAN DER LEE et al. 1997). For the button mushroom *Agaricus bisporus,* 52 progeny were used to construct its linkage map (KERRIGAN et al. 1993).

All 163 progeny inherited their mitochondrial genomes from the JEC20 parent using PCR-RFLP markers developed previously (Xu 2002; data not shown). This result is consistent with uniparental mitochondrial inheritance in *C. neoformans* from the MATa parent (XU et al. 2000; YAN and XU 2003). UPGMA cluster analysis based on genotypic data of the 115 nuclear marker loci revealed that among the 163 progeny analyzed here, there were 119 different genotypes (Figure 2.2). 103 genotypes were each represented by only one progeny while two genotypes were each represented by 10 progeny (genotypes 1 and 2). Among the remaining 14 genotypes, one was represented by six progeny (genotype 3), one by five progeny (genotype 4), one by four progeny (genotype 5), three by three progeny each, and eight by two progeny each (Figure 2.2). Among the 163 progeny, one (progeny R010) had a nuclear genotype at the 115 loci identical to that of the parent strain CDC15 (but with a different mitochondrial genome – that of JEC20, see above), and five progeny were heterozygous (having alleles from both parents) at all 115 nuclear loci (genotype 4) screened (Figure 2.2). The prevalence of recombinant genotypes in our progeny population suggests widespread recombination during meiosis in this inter-variety cross.

Figure 2.2. The UPGMA phenogram showing the overall genetic similarity among the 163 progeny as well as their relationships to the two parental strains CDC15 and JEC20 and their composite genotype P1P2. Genotypes represented by more than 3 progeny were marked to the right.

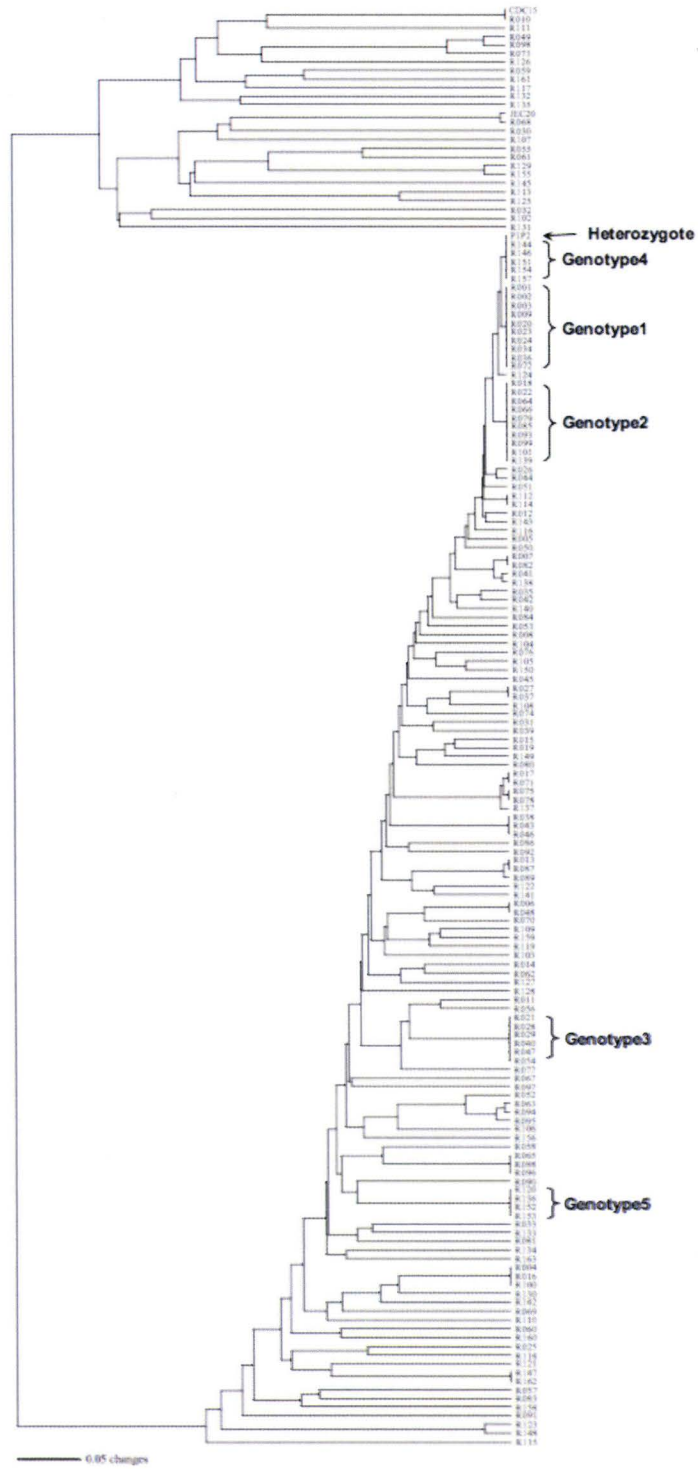
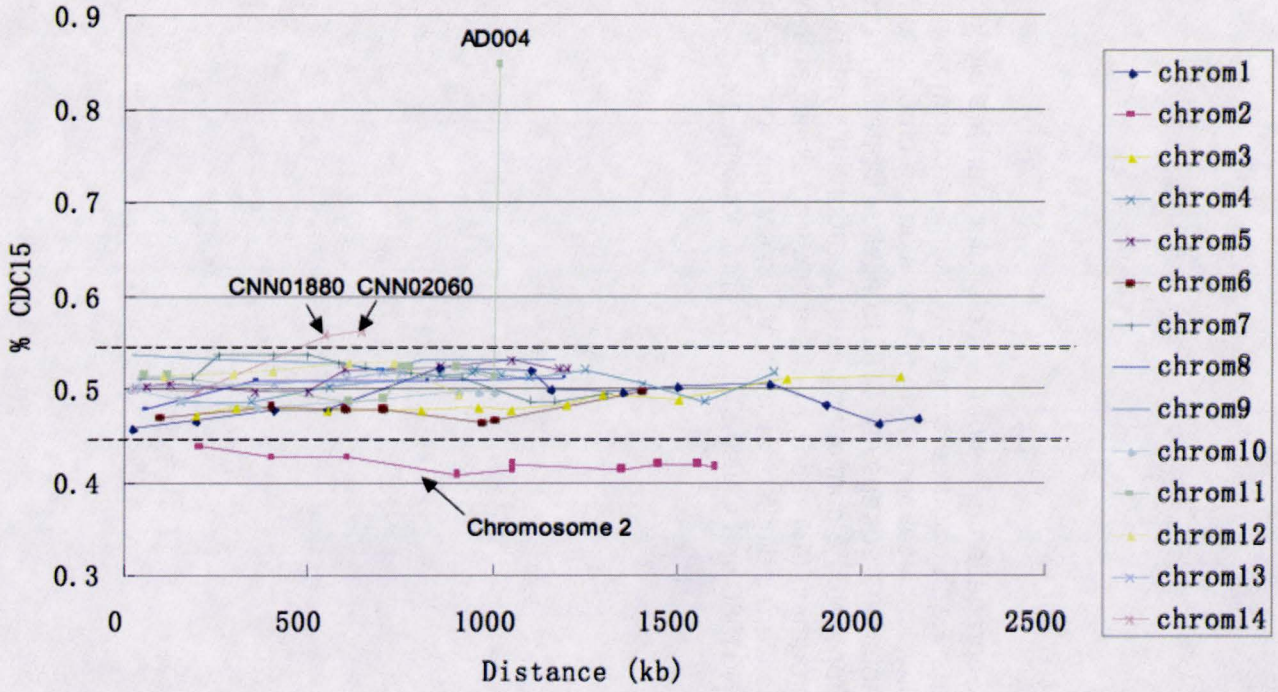

Heterozygote
Genotype4
Genotype1
Genotype2
Genotype3
Genotype5
0.05 changes

Figure 2.3. Segregation ratios of the molecular markers along chromosomes. Each coloured curve represents a chromosome. The x-axis shows the physical locations of the markers on the chromosome. The y-axis represents the frequency of CDC15 alleles. Dotted lines represent the boundaries for the 95% confidence interval around the null hypothesis of random segregation (1:1 ratio). Markers showing significantly skewed segregations were indicated: all markers on chromosome 2 and two marker, CNN01880 and CNN02060, on chromosome 12.


0.9
0.8
0.7
0.6
0.5
0.4
0.3
% CDC15
AD004
CNN01880 CNN02060
Chromosome 2
0
500
1000
1500
2000
2500
Distance (kb)
chrom1
chrom2
chrom3
chrom4
chrom5
chrom6
chrom7
chrom8
chrom9
chrom10
chrom11
chrom12
chrom13
chrom14

There may be two possibilities contributing to the observed segregation distortions. First, there might be different fitness properties between alleles from the two parents that could have directly influenced differential spore germination, viability or growth rates among progeny carrying different alleles. Second, there might be tight linkage between the distorted marker loci and other loci with alleles differing in fitness. In our study, we found all the markers located on LG2 (also chromosome 2) showed skewed allele frequencies in favour of the parent JEC20 allele, with the percentage of JEC20 alleles at each locus ranging between 56.1% and 59.2%. At present, we do not know the genetic basis for this observed chromosome-wide segregation distortion. A similar phenomenon has been reported by JURGENSON et al. (2002) in their linkage mapping analysis of the maize fungal pathogen *Gibberella zeae*.

Among the 115 nuclear loci, the heterozygosity (the percentage of heterozygotes in the progeny population of 163) at individual loci ranged between 25.16% and 88.34%. The percentage of homozygotes with allele from parent CDC15 ranged between 3.68% and 72.26% among the loci. The percentage of homozygotes with the JEC20 allele ranged between 2.58% and 26.38%. When marker loci with skewed segregation ratios were excluded, the heterozygosity at these loci ranged between 49.69% and 88.34%, the percentage of homozygotes with the CDC15 allele ranged between 6.75% and 25.15% and the percentage of homozygotes with the JEC20 allele ranged between 3.07% and 26.38% (Supplemental Table S2.1).

Interestingly, when heterozygosity was mapped onto the linkage map and compared to their chromosomal organizations, in 13 of the 14 chromosomes, the lowest heterozygosity was found closest to the proposed centromere regions in the published genome sequence of JEC21 (LOFTUS et al. 2005, Supplemental Figure 2.1). Lower levels of heterozygosity around these markers suggest higher levels of recombination frequency in the regions around them. The mechanism for such a low level heterozygosity surrounding the centromeres is not clear.

Linkage groups

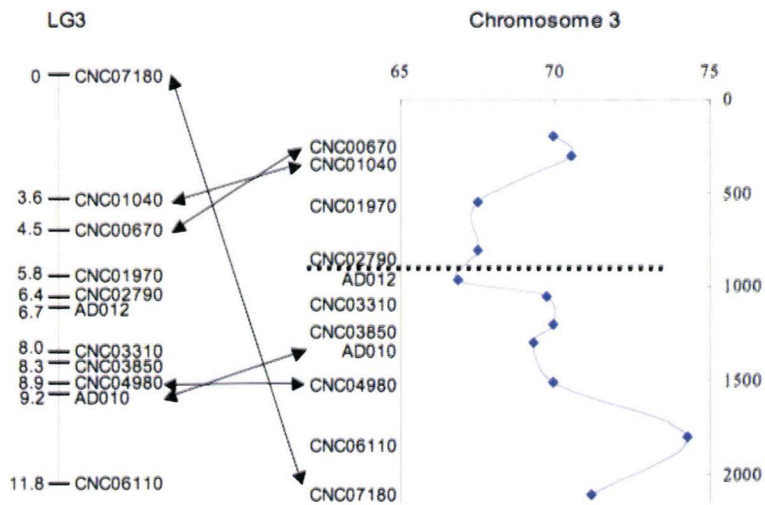
Linkage analysis was performed using MAPMAKER with a LOD threshold of 25 and maximum distances between adjacent markers of 10 cM (see Materials and Methods). Among the 115 markers, 114 were assigned to LGs. The only exception was marker AD004, which showed highly skewed segregation toward the allele from parent CDC15 (χ^2 = 150.5, $P < 0.0005$). Marker AD004 was then excluded from further linkage analyses. These remaining 114 markers were assigned to 14 LGs, with 12 LGs having at least six markers each and the other two LGs having two markers each (Table 2.1). Eleven of the 14 LGs corresponded to 11 chromosomes of JEC21. Overall, the orders of marker loci in these 11 LGs were congruent with their physical locations in each chromosome, except

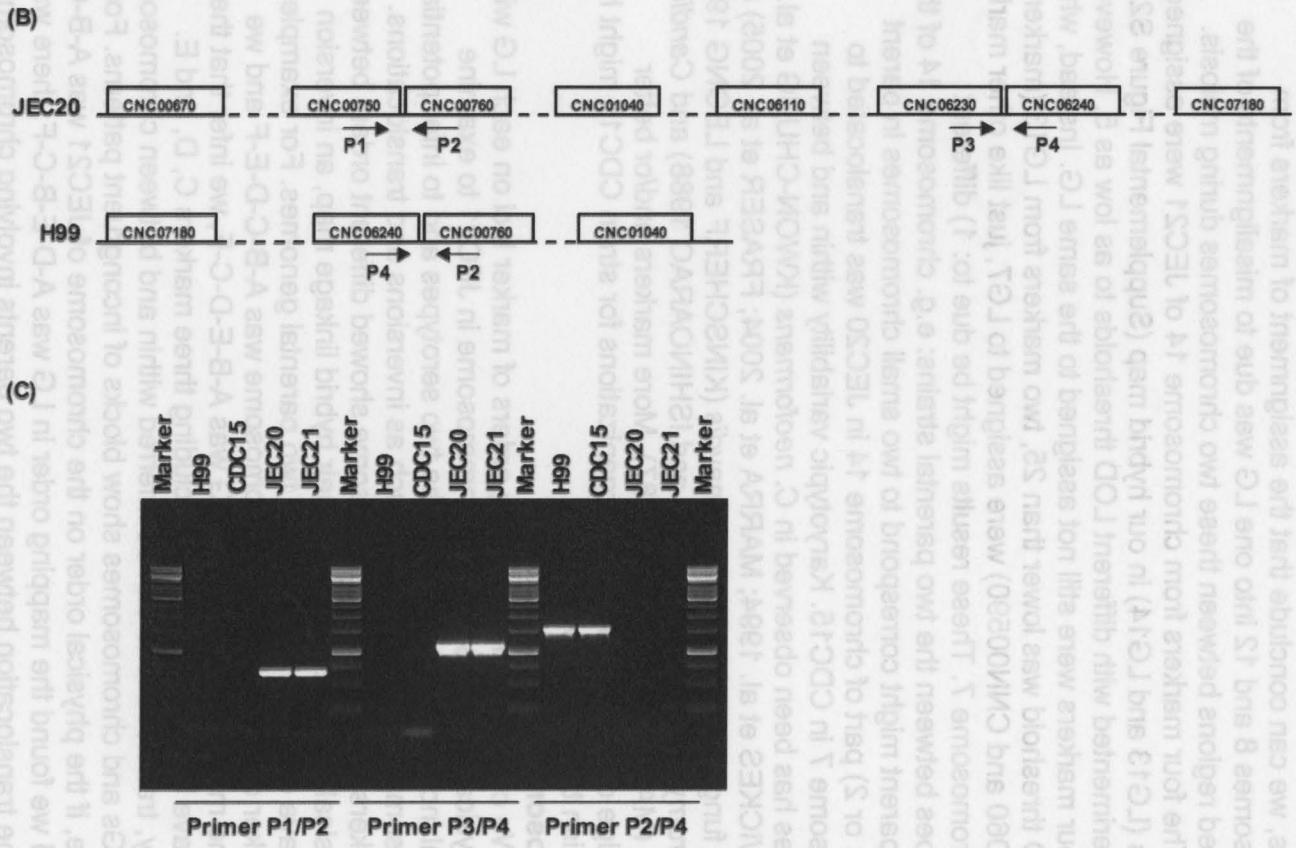
a few putative inversions and translocations with each involving one to three marker loci (Figure 2.4 and Supplemental Figure S2.1, see below). Our result is consistent with the physical mapping results by SCHEIN et al. (2002) in which they compared physical maps of *C. neoformans* serotype A and D strains based on restriction site mapping and Southern hybridizations. They found a high degree of conservation of synteny between strains JEC21 (serotype D) and H99 (serotype A) as well as some chromosomal rearrangements including both inversions and translocations.

Our LG8 contained markers from both chromosomes 8 and 12 of the JEC21 genome. Increasing LOD threshold to 30 separated LG8 into two LGs: one LG (LG8-A) contained three markers from chromosome 8 only and the other LG (LG8-B) contained all the markers from chromosome12 except one (marker CNH03700) from chromosome 8 (Supplemental Figure S2.1). It has been reported by FRASER et al. (2005) that chromosome 8 and chromosome 12 in JEC21 originated through the breakage of a dicentric chromosome resulting from telomere-telomere fusion of two chromosomes from strain B3501A. During this process, a chromosomal translocation occurred and as a result, chromosome 8 and chromosome 12 of JEC21 shared a large segmental duplication spanning ~62 kb and containing 22 predicted genes. In our study, marker CNH03700 was located in this duplicated region. When we increased the LOD threshold to 30, marker CNH03700 could be unambiguously assigned to a LG in which all the other markers were from chromosome 12 (Supplemental Figure S2.1). In our linkage map, CNH3700 was separated by 1.6 cM from another marker, CNL06830, which was identified as a putative translocation (see below). These two markers were well separated from other markers by 5.8 cM and 5.4 cM, respectively, from their flanking regions (Supplemental Figure S2.1).

Figure 2.4. A representative linkage group and the confirmation of a translocation breakpoint in this linkage group constructed from an inter-variety cross in *C. neoformans*.
(A) Linkage group 3 (LG3) of the hybrid genetic linkage map of *C. neoformans*.
Marker names were indicated on the right side of the linkage group. Numbers on the left side were genetic distances in centi-Morgan. Figure on the right side of the linkage group was the distributions of heterozygosity at each marker locus located on LG3, with the marker orders following their physical locations in the chromosome 3 of JEC21. X-axis showed the level of heterozygosity and y-axis showed the physical distance from one end of chromosome. Arrows indicated markers with positions in the linkage groups different from their physical locations in chromosomes 3 of JEC21 (putative inversions and translocations, see Results). Dotted lines indicated the approximate locations of candidate centromeric region in chromosome 3 of JEC21 as identified by LOFTUS et al. (2005).
(B) The locations of primers used to confirm the putative translocation of marker CNC07180 in LG3. Arrows indicated the positions and directions of primers in chromosome 3 of strain JEC21 and contig 32.1 of strain H99.
(C) Gel electrophoresis of PCR products amplified from different strains using different primer pairs to detect translocation. The strain names were indicated at the top of each lane. The primer pairs used for PCR were indicated at the bottom of the picture. Primer sequences are listed in Table 2.2.

(A)


LG3
Chromosome 3
0 — CNC07180
3.6 — CNC01040
4.5 — CNC00670
5.8 — CNC01970
6.4 — CNC02790
6.7 — AD012
8.0 — CNC03310
8.3 — CNC03850
8.9 — CNC04980
9.2 — AD010
11.8 — CNC06110
CNC00670
CNC01040
CNC01970
CNC02790
AD012
CNC03310
CNC03850
AD010
CNC04980
CNC06110
CNC07180
65
70
75
0
500
1000
1500
2000


(B)
JEC20
CNC00670
CNC00750
CNC00760
CNC01040
CNC06110
CNC06230
CNC06240
CNC07180
P1
P2
P3
P4
H99
CNC07180
CNC06240
CNC00760
CNC01040
P4
P2
(C)
Marker
H99
CDC15
JEC20
JEC21
Marker
H99
CDC15
JEC20
JEC21
Marker
H99
CDC15
JEC20
JEC21
Marker
Primer P1/P2
Primer P3/P4
Primer P2/P4

When these two markers were excluded in the linkage analysis, the remaining markers from chromosome 8 and chromosome 12 were assigned to two distinct LGs with each containing only markers from the same chromosome at LOD threshold 25 (data not shown). Based on these analyses, we can conclude that the assignment of markers from chromosomes 8 and 12 into one LG was due to misalignments of the duplicated regions between these two chromosomes during meiosis.

The four markers from chromosome 14 of JEC21 were assigned to two LGs (LG13 and LG14) in our hybrid map (Supplemental Figure S2.1). We experimented with different LOD thresholds to as low as 5. However, these four markers were still not assigned to the same LG. Instead, when the LOD threshold was lower than 25, two markers from LG13 (markers CNN00060 and CNN00590) were assigned to LG7, just like other markers from chromosome 7. These results might be due to: 1) different karyotypes between the two parental strains: e.g. chromosome 14 of the JEC20 parent might correspond to two small chromosomes in parent CDC15; or 2) part of chromosome 14 in JEC20 was translocated to chromosome 7 in CDC15. Karyotypic variability within and between serotypes has been observed in *C. neoformans* (KWON-CHUNG et al. 1992; WICKES et al. 1994; MARRA et al. 2004; FRASER et al. 2005) and in other fungi, including *Ustilago maydis* (KINSCHERF and LEONG 1988), *Saccharomyces cerevisiae* (ONO and ISHINOARAO 1988) and *Candida* species (MAGEE and MAGEE 1987). More markers and/or better knowledge of the chromosomal organizations for strain CDC15 might help distinguish these possibilities.

Chromosomal rearrangements

We compared the mapping orders of marker loci on each LG with their physical locations on each chromosome in JEC21 to examine potential incongruence between the two serotypes and to infer potential chromosomal re-arrangements such as inversions and translocations. If the markers on the same chromosome showed different orders between the physical map of JEC21 and their hybrid linkage map, an inversion might have occurred between the two parental genomes. For example, if the marker order on a JEC21 chromosome was A-B-C-D-E-F and we found the mapping order in this LG was A-B-E-D-C-F, we infer that there might have been one inversion including three markers C, D, and E. Similarly, translocations can be inferred within and between chromosomes when LGs and chromosomes show blocks of incongruent patterns. For example, if the physical order on the chromosome of JEC21 was A-B-C-D-E-F and we found the mapping order in LG was A-D-E-B-C-F, there was likely one translocation between the two parents involving chromosomal regions that included markers D-E with that of B-C.

Overall, the orders of markers on each LG matched well to their physical locations on the chromosomes in JEC21. However, there were a

few exceptions, likely involving putative inversions and translocations between the two parents. Specifically, eight LGs (LG1, LG2, LG4, LG5, LG6, LG10, LG11 and LG12) had the orders of marker loci identical to the physical locations of these markers in JEC21. LG13 and LG14 have only two markers each and their comparisons are thus meaningless. Among the other four LGs, LG7 had one putative inversion involving two markers (AD019 - CNG02290). LG3 had two putative inversions with each involving two markers (CNC00670 – CNC01040 and AD010 – CNC04980). In addition, LG3 had a putative rearrangement that involved one marker (CNC07180). LG9 had one putative inversion that involved three markers (CNI02950 – CNI03300 – CNI04370). LG8 had markers from both chromosomes 8 and 12. Comparing the marker order in LG8 to the linear aligned order of these markers on the two chromosomes, we found one putative inversion that involved two markers (CNL04620 – CNL04980) as well as one putative translocation that involved one marker (CNL06830) (Figure 2.4A; Supplemental Figure S2.1).

Confirmation of chromosomal rearrangements

To confirm the aforementioned putative translocations and inversions, we checked the locations of the markers involved in those putative translocations and inversions in genome databases of JEC21 and H99 to see if their orders in the two genomes are congruent. Because the H99 genome sequence assembly is not finished, we could only compare markers that were located in the same contig in H99 genome. Out of the aforementioned 7 putative inversions and translocation, two inversions (chromosome 3, CNC00670-CNC01040; chromosome 9, CNI02950-CNI03300-CNI04370) had markers that were located on different contigs in H99; three inversions (chromosome 3, AD010-CNC04980; chromosome 7, AD019-CNG02290; chromosome 8, CNL04620-CNL04980) had the same marker orders in both the JEC21 and H99 genomes. Although we could not find definitive evidence for marker order incongruence of the putative inversions between the JEC21 and H99 genomes, the orders of marker loci in the CDC15 genome might still be different from those of JEC20 and H99. The putative rearrangements inferred here is supported by the observation that even though the genetic distances between marker pairs involved in those putative inversions were relatively low (highest 0.9 cM and average 0.5 cM; Supplemental Figure S2.1), the ratio between physical distance and genetic distance were much higher than the overall average (lowest 119 kb/cM, highest 687 kb/cM, average 367 kb/cM; see below). These results are consistent with the hypothesis that inverted regions typically experience low levels of recombination (see below).

Blast analysis of the sequence of CNL06830 against the genome sequences of JEC21 and H99 showed that this putative translocated marker had two copies in the JEC21 genome that were located on both chromosome 8 (about 170 kb away from CNH03700) and chromosome 12,

respectively, but only one copy in contig 91.1 of the H99 genome. The duplicated region includes CNL06830 and spanned about 10kb. It has been shown in *S. cerevisiae* that ectopic recombination between artificial repeats of 5.5kb in size could compete efficiently with normal allelic crossovers during meiosis (JINKS-ROBERTSON et al. 1997). Therefore, the putative translocation of CNL06830 identified in our study could be the result of misalignment and ectopic recombination between the duplicated regions during meiosis.

Our comparison of the H99 and JEC21 genomes indicated that marker CNC07180 in H99 (chromosome 3, a putative translocation) might be located in a different chromosome region but is close to marker CNC01040 in the linkage map (Figure 2.4A and Figure 2.4B). We therefore attempted to identify the possible breakpoints of the translocation between the two genomes. To test whether such breakpoints exist among strains JEC20, H99 and CDC15, we designed PCR primers around these breakpoints based on the JEC21 genome sequence. Indeed, the translocation was confirmed using PCR (Table 2.2; Figure 2.4B and Figure 2.4C). Specifically, primer pairs P1/P2 and P3/P4 were separated from each other on chromosome 3 of JEC21 by more than 1.4 Mb. Both primer pairs could amplify PCR products with expected sizes from JEC20 and JEC21 but not from H99 and CDC15. Using primer pair P2/P4, which comprises one primer from each of the two primer pairs, we successfully amplified PCR products with expected size from both H99 and CDC15 but not from JEC20 and JEC21, consistent with a chromosomal rearrangement between JEC20 (also JEC21) and CDC15 (also H99) (Figure 2.4C).

Analysis of recombination frequency

The total length of this linkage map was 196.8 cM (Table 2.2). The largest LG was LG1, with 38.5 cM in length. The smallest LG was LG14, which was only 0.9 cM in length (Table 2.1). To calculate the ratio between genetic distance and physical distance as identified based on JEC21 genome sequence, we included only markers that showed clear associations between chromosomes in JEC21 and linkage groups (Supplemental Figure S2.1). In total, 103 of the 114 markers fit this criterion. The remaining 11 markers were excluded with 10 from LG 8 because they were from two different chromosomes and one (marker CNC07180) from LG3 because it was identified as a translocation. The 103 markers covered a total physical distance of 14550kb, about 76% of the whole genome. The total map length covering these 103 markers was 171 cM, corresponding to 1 cM for every 85kb.

The ratio of linkage map distance in this hybrid cross over physical distance derived from the JEC21 genome was highly variable among chromosomes and among regions within certain chromosomes. When the map length of each individual LG was compared to the physical distance

covered by its markers on chromosome of JEC21, the ratio between physical distance and genetic distance ranged between 49.91 kb/cM (LG 4) and 195.61 kb/cM (LG 3) among the chromosomes (Table 2.1). When each individual marker pair was considered, the ratio between physical distance and genetic distance ranged between 13 kb/cM (LG6, AD018 – CNF03420) and 686 kb/cM (LG6, CNF01350 – CNF02070). Twelve pairs of loci showed no recombination in this cross (i.e. their pairwise genetic distance was 0). In contrast, their physical distances ranged between 20kb and 149kb in the JEC21 genome.

Comparison between the inter-variety hybrid cross and a previously published intra-variety cross

Compared to the linkage maps constructed using *C. neoformans* serotype D strains, the total length of our hybrid linkage map (197 cM) was much shorter and the ratio between physical distance and genetic distance was much higher (85 kb/cM). For example, the linkage map constructed by FORCHE et al. (2000) had a total length of 1356.3 cM, corresponding to 13.6 kb/cM. The genetic map constructed by MARRA et al. (2004) had a total length of ~1500 cM, corresponding to ~13.2 kb/cM across the genome. Fifteen PCR-RFLP markers used in MARRA et al. (2004) study were also included in our study and we compared the genetic distances between these markers in these two studies (Table 2.3). Only marker pairs located on the same chromosomes were compared. Of the 9 marker pairs that could be compared between the two studies, one showed a smaller genetic distance in the serotype D linkage map than that in the hybrid map, a second pair showed similar genetic distances between these two maps, and the remaining seven pairs showed much greater genetic distances in the serotype D linkage map (Table 2.3). The differences in the seven pairs ranged from 3 folds to more than 30 folds. The combined analyses of all nine pairs identified that the genetic distance in the serotype D cross was significantly higher than that in the hybrid cross (Mann-Whitney U test, $p < 0.05$). Because the genetic distance is a function of recombination frequency between markers, this result suggested that meiotic recombination occurred at a much higher frequency in the intra-serotype D cross than in the inter-variety cross.

The shorter genome map length and much higher physical distance to genetic distance ratios for the inter-variety cross observed in our study indicated the recombination occurred at a much lower frequency in inter-variety cross than in intra-variety crosses of *C. neoformans*. The low level of recombination is expected during hybridization between divergent populations. Indeed, they have been considered as a driving force for speciation. Besides those aforementioned putative inversions and translocations that might have repressed recombination in those regions, several other mechanisms could also contribute to the observed low level of recombination frequency in the inter-variety cross of *C. neoformans*.

First, the DNA sequence divergence between serotype A and serotype D *C. neoformans* might have decreased the efficiency of homologous recombination.

Recombination suppression could be achieved through the activities of mismatch repair proteins, such as Pms1 and Msh2 (HUNTER et al. 1996). These genes are eukaryotic homologs of the bacterial MutS protein that have shown to play a significant role in recombination and speciation. In *Saccharomyces cerevisiae, pms1* and *msh6* mutants showed increased level of mismatch binding and meiotic recombination during meiosis (CHAMBERS et al. 1996; HESS et al. 2002). It has also been suggested that the reproductive isolation between species in *Saccharomyces sensu stricto* is mainly due to the mismatch repair system (HUNTER et al. 1996; GREIG et al. 2002; LITI et al. 2006). Genetic studies of homeologous recombination in yeast suggested that if sequences were too divergent (>10%), recombination in these region were severely repressed, presumably due to the inability to form sufficiently stable base-paired intermediates (SELVA et al. 1995; DATTA et al. 1996; DATTA et al. 1997; CHEN et al. 1999; reviewed in EVANS and ALANI 2000). Data collected in *C. neoformans* was consistent with this explanation. For example, sequence analyses of *C. neoformans* revealed that sequence divergences between serotype A and serotype D were more than 10 fold higher than those within serotype A and serotype D, up to over 10% in certain regions (XU et al. 2000). Second, theoretic studies have suggested that there might exist "recombination modifiers" that control the recombination frequency. During the divergence between the two varieties, recombination modifiers from different varieties might be less or not at all compatible with each other. As a result, recombination frequencies decrease. A third possibility is the existence of strong epistasis among alleles within each of the two serotypes and these alleles can no longer segregate randomly. These three mechanisms are not mutually exclusive and it is very likely that all three could have contributed to the observed low recombination frequency in the inter-variety cross of *C. neoformans*.

Although the overall recombination frequency observed in our study was very low, several regions showed comparable levels of recombination to the intra-variety crosses of *C. neoformans*. For example, two regions, AD018 - CNF03420 in chromosome 6 and CNK02410 – CNK02590 in chromosome 11, had physical/genetic distance ratios of 13.09 and 14.17 kb/cM, respectively, similar to those found by MARRA et al. (2004) and FORCHE et al. (2000). It has been shown that in *C. neoformans,* the two regions flanking the mating type locus were recombination hotspots (HSUEH et al. 2006). In our study, one of these two regions had physical/genetic distance ratio of 26.6 kb/cM, about 3 times lower than that of the genome average.

Table 2.2. Primers used for the confirmation of a putative rearrangement inferred from linkage mapping analysis of the inter-variety cross.

Primer	Primer sequence (5' – 3')
P1	AACCCTCGGTCCCCCAATTA
P2	TTTTATTTCCGGGCCTTTCGG
P3	AAGCAAGGAGCAAGAGGCGA
P4	GGCAATATTATGCAGAAGAG

Table 2.3. Comparison of genetic distances for the same marker pairs between the inter-variety hybridization analyzed here and the intra-variety cross reported in MARRA et al. (2004).

Marker Pairs [a]	Chromosome	Genetic Distance (intra-variety) [b]	Genetic Distance (inter-variety) [c]
AD012 (Xba11) AD010 (Eco4)	3	46.8	2.5
AD029 (Stu4) AD030 (Hind8)	4	2.2	2.2
AD014 (Eco29) AD028 (Stu3)	5	30 [d]	13.3
AD021 (Hind21) AD020 (Hind17)	7	59.4	3.8
AD020 (Hind17) AD019 (Eco30)	7	1.9	2.9
AD019 (Eco30) AD026 (Xba21)	7	70 [d]	4.2
AD005 (Hind7) AD006 (Xho5)	9	53.7	1.6
AD001 (Eco21) AD002 (Pst28)	11	49.5	5.8
AD002 (Pst28) AD003 (Xho18)	11	21.3	6.5

[a]: Marker names in parentheses were used in the serotype D mapping study (MARRA et al., 2004)..
[b]: Genetic distance calculated in serotype D mapping study (MARRA et al., 2004).
[c]: Genetic distance calculated in this study.
[d]: This pair of markers were located on different LGs in the serotype D mapping study (MARRA et al. 2004). Here the minimum possible genetic distance between this pair was derived by adding the shorter distances of each marker to the end of its LG, and then add 30 cM, the threshold swept radius used in the Marra et al. study.

CONCLUSION

In this study, we constructed and analyzed a hybrid linkage map containing 115 co-dominant markers using 163 progeny collected from an inter-variety cross between strain CDC15 (var. *grubii*, serotype A, MATalpha) and strain JEC20 (var. *neoformans*, serotype D, MATa) in the important human fungal pathogen *C. neoformans*. Overall, the marker positions were highly similar and syntenic to those inferred from the serotype D cross and to the physical positions in the genome of strain JEC21. While our analysis identified Mendelian segregation and independent re-assortment among markers during meiosis, the genetic distances between markers in the hybrid cross were much smaller than those in the serotype D cross, indicating suppressed recombination in the inter-varietal cross. The low recombination frequency coupled to the presence of heterozygotes at all loci in the progeny population suggests abnormal/incomplete nuclear disjuctions during meiosis. The constructed linkage map should help the genetic analyses of divergent phenotypic traits between these two varieties.

ACKNOWLEDGEMENTS

This research is supported by the Natural Science and Engineering Research Council (NSERC) of Canada and by the Premier's Research Excellence Award. Sheng Sun has been supported by the Ontario Graduate Scholarship and the postgraduate scholarship from NSERC.

Supplemental Table S2.1. Polymorphic PCR-RFLP markers developed in this study. Shown here are marker identifications, their chromosome locations in JEC21 genome, their starting nucleotide positions within each chromosome, the primer sequences, the enzymes that produced polymorphic restriction digest patterns distinguishing the two parental strains, and the frequencies of the three genotypes in the progeny population.

ID	Chromosome	Location (bp)	Forward Primer (5' - 3')	Reverse Primer (5' - 3')	Enzyme	Homozygous CDC15 (%)	Homozygous JEC20 (%)	Heterozygote (%)
CNA00050	1	18428	AGCAATTCCAAACCGACCCC	TTACGCCCACACCAAGGCAT	HaeIII	12.27	20.86	66.87
CNA00670	1	193301	TTGGGCCAGGAGTGAGTGAT	AAGGCAGAAAGAGCCGCGTT	HaeIII	12.88	19.63	67.48
CNA01490	1	402248	AGGCAGGCGCGTCAGCTTTTTG	AAACAGGCGACCATTGCGGGAG	HaeIII	15.34	19.63	65.03
CNA02040	1	547518	GGTGCGAGGGCTTCAATTGT	TGCCGTGATCGCCTGTTTTG	HaeIII	16.56	20.86	62.58
CNA03240	1	850012	AAGTGCGGCGCGATGAATGA	TATGGGAAAGGCGGGGATGT	AccI	25.15	20.25	54.60
CNA04100	1	1101662	CCACCAGCGCAAAAACGAC	TGATGCGCCGCCTTTCTCTT	HaeIII	23.93	19.63	56.44
CNA04280	1	1152675	CCGCAAAATCATCTCCGCTG	AAAGGGGCGAAAACGCTCGT	HinfI	20.25	20.25	59.51
CNA05090	1	1350514	CGGCGTTGTAACCCATGACC	ATGCTCCCCACCCACTTTCA	EcoRI	19.63	20.25	60.12
CNA05600	1	1500782	TGGGTTTGGCTTCAGGCAG	CCGCTCTTCACATGCAGCAA	HaeIII	17.79	17.18	65.03
CNA06430	1	1746918	CCAACCGAAGCCCAAGACAA	TTGAAGGATGATCCGGCCGA	HaeIII	14.72	13.50	71.78
CNA06990	1	1900000	AGCGAGCAAAAGGCGGACAA	ACGGCCCTTTCGCAAGTTGA	PvuII	10.43	13.50	76.07
CNA07470	1	2045366	TCCGGCATGCATGATCCGAA	AATCGCCTGCAACTGCGCAA	HinfI	8.02	15.43	76.54
CNA07790	1	2152763	TCCAATGGACGAGGACGATG	TGACCGGTGTGGGTTGCAAT	HinfI	6.75	12.88	80.37
CNB00700	2	202159	AGAAAGCGGGCTATGCGCAA	ACCCATTTCGCCCTTCTGCT	EcoRI	6.75	19.02	74.23
CNB01310	2	399106	CATCGGCGAATTGAGCACCA	AAGAGGGCGGTGGAGAACAA	HaeIII	6.75	21.47	71.78
CNB02080	2	607122	ACCCAATGTTGCAAGGCCCA	TCGTCACCGCCCAAGATCAAC	HaeIII	7.36	22.09	70.55
CNB02980	2	906260	GGGCCATGCAATGATGACCAG	CCCACGATCGCTTGCAATCACAA	EcoRI	6.75	25.15	68.10

(Supplemental Table S2.1 continued)

ID	Chromosome	Location (bp)	Forward Primer (5' - 3')	Reverse Primer (5' - 3')	Enzyme	Homozygous CDC15 (%)	Homozygous JEC20 (%)	Heterozygote (%)
CNB03520	2	1052632	TTTCGGGAGTTGGGAGCACA	TTGGCTGCGGTACTGGCATTCTT	HaeIII	6.75	23.93	69.33
CNB04740	2	1351027	ATCGCGGCAAGTGGCAAGAAG	GCAACCGGAACGGGAATCAAGC	HinfI	3.68	20.86	75.46
CNB05090	2	1450569	TTACCTCCGACGACCGCAAAC	TTGCCAACGCAGTGTCCCAAC	BsmI	3.70	19.75	76.54
CNB05530	2	1555440	AACGCCTTTGTCAGCGCCTTGG	AATGCATGGGCGGACAGGATTG	PvuII	3.68	19.63	76.69
CNB05710	2	1602683	TTGGACAACGCAAGACCCCAG	TGTTGCAAGCAACGATGCC	HindIII	3.68	20.25	76.07
CNC00670	3	192408	TGTGCGGCTTTGGGATTGGT	TTGCAGGTATGGCCGAATGG	HaeIII	12.27	17.79	69.94
CNC01040	3	299012	TTGGACAGCGCCGTTTGCAT	TTGGCGGCAAAGCGTCAAGA	HaeIII	12.88	16.56	70.55
CNC01970	3	546980	TGCATTGGCGAGATCAGGCA	TTCGATCCATCCCTGGAAGCCT	HaeIII	14.11	18.40	67.48
CNC02790	3	800550	GATGAAATGGCGAGGACGCA	TTCCGCGTTGCAACACAACC	EcoRI	14.11	18.40	67.48
AD012	3	959300	CATGGTAGAGGGAGTCGAAAAG	CTTAGTTACAAGGAGATG	HinfI	14.72	18.40	66.87
CNC03310	3	1048312	AGTAATCGTTGGTGGGGGGT	TTTGACGCTGCCTTCGAGGA	HindIII	12.96	17.28	69.75
CNC03850	3	1197448	CGGCACGCATGATTTGGA	TTGGGTTGGCAAGGTGGA	EcoRI	13.50	16.56	69.94
AD010	3	1296700	GACGAGTGGAAAGATGAAAA	GATACCTGCTGCAACCCAAG	TaqI	14.72	15.95	69.33
CNC04980	3	1502391	AAGCAGACTGCACATGGGCA	GGGGCAAAGGGTTCACCAAA	AluI	14.11	15.95	69.94
CNC06110	3	1796156	AAGGGCGTTGATCCGGCAAT	TTCCCCAGGTCGTTTGGGAT	HaeIII	14.11	11.66	74.23
CNC07180	3	2102063	TGGAGGCGTTGGGCGAAATAGAG	TTCAGCCGTCGCCTTTACCACAA	HaeIII	15.95	12.88	71.17
CND00510	4	151248	CGGTGCCGCTTTATTTGTGGC	TCTAGCGCCAAAGCGTGCAAG	HinfI	23.93	26.38	49.69
CND01190	4	343121	AGGAAGGTCGCTATTGACGCG	TTCTTCGCTGGCGCCTTCTTC	XhoI	23.93	26.38	49.69
CND02060	4	554559	TCAGGTTCAAACCGCCAGCA	AGTTTTGCCCCGCTTCGCTTGC	HaeIII	21.47	20.86	57.67
CND03160	4	853595	GCAAAGCGGCCCGAAACATT	TTCCATCATGCCCTGCGTCA	HaeIII	20.86	17.79	61.35
CND03480	4	944503	TCGAACACCGCCATCCATCA	ATTGCCAACGCTGGGAAGGA	PvuII	19.63	15.95	64.42

(Supplemental Table S2.1 continued)

ID	Chromosome	Location (bp)	Forward Primer (5' - 3')	Reverse Primer (5' - 3')	Enzyme	Homozygous CDC15 (%)	Homozygous JEC20 (%)	Heterozygote (%)
AD030	4	1022500	CTTGATGCTCTTTATGGGGAA	TGTGCCAAGGTTATGGAGATG	HinfI	19.63	16.56	63.80
CND03960	4	1096841	TCCAGTTCCGCTGGCGTTTTG	TCCGGGCAACTTTACCGCAA	HaeIII	17.79	15.34	66.87
CND04540	4	1250923	ACAAATCTGGTTCGGGGCCT	TCAGCGCGTTTTACGCCAAG	HinfI	17.18	12.88	69.94
CND05140	4	1407510	TCCCATCCGGCATGGAATGA	TTGACTTCAACGCCGCTCGACTT	HaeIII	14.72	13.50	71.78
Ste20-Aalpha (MAT)	4	1572597	CCAAAAGCTGATGCTGTGGA	AGGACATCTATAGCAGAT	N/A[a]	10.43	12.88	76.69
Ste20-Da (MAT)	4	1572597	GATCTGTCTCAGCAGCCAC	AATATCAGCTGCGCAGGTGA	N/A[a]			
CND06370	4	1762706	TCCCCATTCCCTCTCGTCAA	CGGCTTCCCAAAGCACATC	HaeIII	12.88	9.20	77.91
CNE00250	5	56469	TGGCGTCTCTTTGAACGCGATC	ATGGCGGAATGTCCGGCTTT	HaeIII	8.59	7.98	83.44
AD014	5	119500	GAACGAGAGTTGAGAGAACAACC	CTTGAGCGAGTCAGAGAGTGAG	HinfI	9.20	7.98	82.82
CNE01270	5	351269	CCCCACATTTGCCTTCAGCAG	AATTGGCGATACGCCGAGCA	HindIII	9.82	10.43	79.75
CNE01830	5	499184	AAAAGGCGACCCACCATCCA	TTCGTAAAGGCGGCGCTCAA	HaeIII	9.82	10.43	79.75
CNE02210	5	599253	TGCCGTTTCTGGAAGGGCTA	ACCACATTGCTGGCGGTGAT	HinfI	14.11	9.82	76.07
CNE03010	5	855983	TTGCTGGCAACCACCAGCTT	TTTCCGCACCGGTTGATGCT	HaeIII	14.72	9.82	75.46
CNE03700	5	1050908	AGGCCCGATGTCAGCGAAAT	TTCGGTTGCTTGAGCATGGG	PvuII	15.34	9.20	75.46
AD028	5	1179100	GCCCTAGACCAATCGAATAACC	TGATGACATTAGGACCTCAAACC	HaeIII	11.66	7.36	80.98
CNE04300	5	1200307	TCCCTCAAGACGAAACGGTC	TCCTCAAACCCACCTGGCAA	AluI	11.11	6.79	82.10
CNF00290	6	97505	TCATGCCCCTTCGCCTTCAT	TTCTCCTTCTCCCCATCCCA	HaeIII	8.59	14.72	76.69
CNF01350	6	397642	TCAGCAACGGTTCCGGCAAT	AACCTGCACAATGGCGGCGG	HinfI	9.82	13.50	76.69
CNF02070	6	603543	TTGCATGATCGTCGTCCGCAG	GCCACCCAAACGCGTCATTTCTG	HaeIII	9.82	14.11	76.07
CNF02400	6	702997	TCAAAGCTTCCGCCCGTGTT	CAGCCGCCTCAAATCACGAA	HindIII + PvuII[b]	10.43	14.72	74.85

(Supplemental Table S2.1 continued)

ID	Chromosome	Location (bp)	Forward Primer (5' - 3')	Reverse Primer (5' - 3')	Enzyme	Homozygous CDC15 (%)	Homozygous JEC20 (%)	Heterozygote (%)
AD018	6	970900	CCTCCTCTCCGGGTATTC	AGAGCAATGACCCTGTCCAC	HaeIII	9.82	17.18	73.01
CNF03420	6	1008849	AAATCCAGTTGCGCGGCACA	TTGGCAGCTTTTCGTCCCCA	HaeIII	8.64	15.43	75.93
CNF04830	6	1405304	TTGATCGCCGCCGGTTTCTT	TTGGCCGTCCAAATCATGCCC	HinfI	11.04	11.66	77.30
CNG00170	7	44241	TTTCTTCCGCCGCTTCTCAC	ACAGCGCGTTGAGTTTCGGT	HaeIII	10.43	7.98	81.60
AD026	7	182750	CTTGGTGGTAAGGAGATGG	TAACCTTGGCATTTGATGTGT C	AluI	10.43	7.98	81.60
CNG00900	7	251372	TGGTATGACTGCCATCGCCA	AAACGCTCCGTTCCGTTCCT	HaeIII	14.72	7.36	77.91
CNG01370	7	400685	TTCACCGGTTTCCTGGTCGA	TGCCGGACGATCATGCCAA	HaeIII	14.72	7.36	77.91
CNG01750	7	496693	TCATATCGCCGCCAGCGAAA	TCACGGGGTTGAGCATTGGA	EcoRI	14.72	7.36	77.91
AD019	7	584200	AGACAGGGCTGGACAAGAAG	GAGATGTATGCTGGCGTAT	HinfI	14.11	8.59	77.30
CNG02290	7	651617	TGGAGCCGGAATGGCTATCA	TTCGGGATCTTGCCGTTCGTG	PstI	14.11	9.20	76.69
AD020	7	771800	ACTGACATGGAGTAGCC	GATCGAAGTGGAGGTTTTGAA G	TaqI	14.72	11.04	74.23
CNG03250	7	917464	TTGCCAATAACGTGGCACGG	AAAGGGAGGCGGCTGATGAT A	EcorI + HindIII[b]	14.72	12.27	73.01
CNG03900	7	1102910	TCAGCCAGCCCCACCAATAA	TTCCCATTGCCGCCCAAACT	HaeIII	10.43	12.88	76.69
AD021	7	1167200	TGGGTATGCTTGGATGGAG	ACTCGGCCTTGTAGGATTTAC C	HaeIII	10.43	12.88	76.69
CNG04610	7	1309639	TGTTTCCCACAGGCCAAGGA CT	CGTGCGCGAATGCATCGATAT	HinfI	11.04	12.27	76.69
CNH03700	8	50036	AAGCGCCGATGGCAGGAAAA	AACGGGCAAACCCTCGCAAT A	BamHI	3.68	7.98	88.34
CNH02750	8	350270	TTGGATCGCTTGCTCGCGAA	AGGCCCGAGCAAAGGAATGA	EcoRI	7.98	6.13	85.89
AD024	8	815700	TCATAGTCTCGGCGTATGTCT C	ATGGGTTGGCTCTGTTTGTC	HaeIII	11.04	9.20	79.75
CNH00030	8	1187106	TGTCGATGTGCTTCTCGGCA	CTCCCTCCCATCCCAAAACAc	PstI	11.04	8.59	80.37

(Supplemental Table S2.1 continued)

ID	Chromosome	Location (bp)	Forward Primer (5' - 3')	Reverse Primer (5' - 3')	Enzyme	Homozygous CDC15 (%)	Homozygous JEC20 (%)	Heterozygote (%)
CNI00070	9	20878	CCGCCTGCACACCTTTCTt	TGTCTTCGGTTTGGATGGG	XhoI	10.43	3.07	86.50
AD006	9	291300	TTTGTCCAAGAACGCAAGTC	CTTTCCAACCGTCAAATCTACC	HaeIII	9.82	3.68	86.50
AD005	9	564000	ATGAGCCTGGAGTAGTGGTGTC	CAAGGCCAGAACCTTCCTAAAC	HaeIII + HinfI[b]	8.59	3.07	88.34
CNI02550	9	693186	TTTGGCAGGAAGGCGGTTGA	ACATCGATCGCGGTGCATCT	HaeIII	7.98	3.68	88.34
CNI02950	9	801862	TGCTTTCATCGCAGCGTCCA	GCAAGCCGCCGCACAAAA	HaeIII	9.20	3.07	87.73
CNI03300	9	898152	TGCCGCTTCCTTCAACAACG	CGACATCCAGTTGGCGGAAA	XhoI	9.20	3.07	87.73
CNI04370	9	1167156	AGCGGCTACAGCAAAAGCGA	AACATGTCCGCCTCACCCAA	HinfI	9.20	3.07	87.73
CNJ00070	10	14987	ATGGCGGAAGAGGCGTATGA	CCTGTCCAGTGCGCATTTCG	HinfI	8.59	8.59	82.82
CNJ00540	10	145833	TTCCGGGGTCAACATGCGTA	TGGAATTGGCTTCGAGCCGA	HaeIII	7.41	9.26	83.33
CNJ01260	10	355274	TGGCCTTGGCCTTGGCATTA	TGGCGCACCATAGGCTTTGA	XhoI	6.75	10.43	82.82
CNJ02080	10	605055	TGGGTCGCATGGAATGAGCA	ACCAGCAGCAGAGAACATTT	HinfI	9.20	11.66	79.14
CNJ02920	10	905762	TGGGGGAGAAAGGACATTGG	CAAATGCCGAGCTCCCTTC	HinfI	11.04	11.66	77.30
CNJ03090	10	957253	TCTGGGGCTATCCCAAGCAA	TTTCGCTCCCGGAAGCATCA	HaeIII	11.04	11.66	77.30
CNJ03190	10	1001358	ATTTGGCGTTTGGCCGCTCTC	ATGTTCGCCGGTAAAGCGGTCT	XhoI	11.04	11.66	77.30
CNK00170	11	55090	AACATGGCATCTCCCCCCAA	TCGTGCTGACCATGCGGTTT	HinfI	14.72	11.66	73.62
AD001	11	118930	CCAGTACCTACAACGCTTTCCT	GGTAAGTCCCAAATCATCCATC	HaeIII	14.72	11.66	73.62
AD002	11	607300	GACGTTCAGCGGTTTCTGTC	AGCTGCCTATCAACATCTCTCC	HinfI	11.66	14.11	74.23
CNK02410	11	701076	TTTCGAAGCTGCGAGGGATG	TCTTCTTCACGACGGCGGAA	PvuII	9.20	11.04	79.75
CNK02590	11	752082	TCCTCGGTGCAGGTCACAAA	TGGCGTTGATAGCATCCGCA	HaeIII	11.04	6.13	82.82
AD003	11	771750	CTACACCAAACACCATGTCACC	CTGAGACCGAAGGAAAGAAGAG	HaeIII + HinfI[b]	11.04	6.13	82.82

(Supplemental Table S2.1 continued)

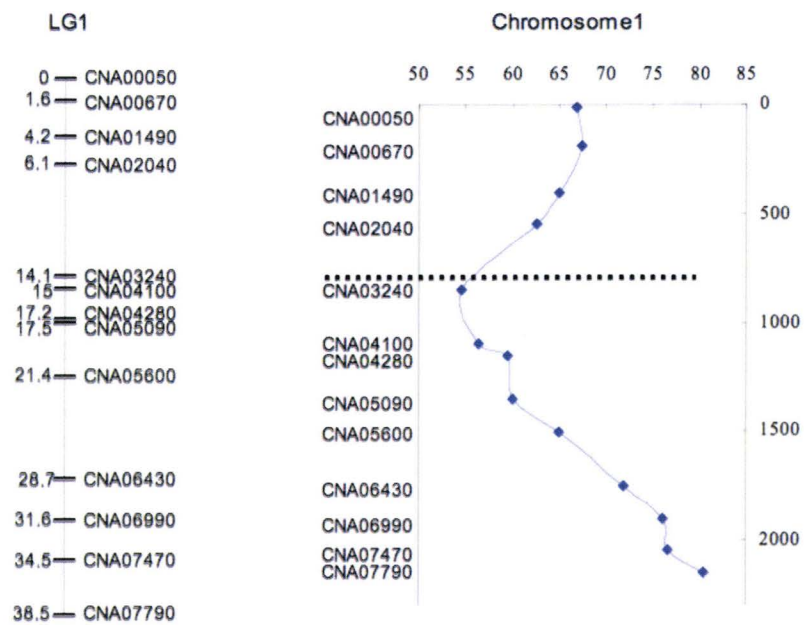
ID	Chromosome	Location (bp)	Forward Primer (5' - 3')	Reverse Primer (5' - 3')	Enzyme	Homozygous CDC15 (%)	Homozygous JEC20 (%)	Heterozygote (%)
CNK03060	11	901632	TAAGCGGTTCGGCGAATGCTC	TGCTGCTGGCGGCGATAATA	SalI	11.04	6.13	82.82
CNK03410	11	1003207	TTTCGCCGCACCCCCTTTTT	CCTCGCCGCCCAATAATTCA	HinfI	9.82	7.36	82.82
AD004	11	1018800	CAAGGGGTCAAGTGATAACCAG	TGCCATGAAGTCTTGTAGCATC	TaqI	72.26	2.58	25.16
CNL04030	12	101857	GAGCGAACAAGCATCATACC	ACAAGGCCAAGGCCAAAGCA	HaeIII	11.66	8.59	79.75
CNL04620	12	290771	TTCGTGGCGACAGGTTTTGGG	TTCAGCGATGGGTTGAGGCA	SalI	12.27	8.59	79.14
CNL04980	12	398737	AGCGTGGATCAAACGGGCAT	CAAGCCTTTTGGTCCCGAGA	HinfI	12.27	7.98	79.75
CNL05760	12	604685	TGCTACTATCAAAGTTCCCAA	ATGCGTCGGGTGCTTGTTTGGC	HaeIII	10.43	4.91	84.66
AD007	12	726600	AAGCGGATAGTGCAAATAGAGG	TCGGTATCACTACCAGAGACAAAC	TaqI	10.43	4.91	84.66
CNL06830	12	902847	TAGTGGCAAGCTTCGCATCGG	TCGGTTCCAAAGACCTCGCA	HinfI	6.75	7.98	85.28
AD009	13	29750	CAAACGAGAAGATACGGCAAAG	CGTGAAGATACTGGAGCAAACC	HaeIII	12.88	12.27	74.85
CNM00630	13	195550	TGCCAATTGCAAGGGTGGCT	TGCGTTGAACAACGCGACCT	PstI	12.88	12.27	74.85
CNM01380	13	406059	AAAGCTGGCCCGCAAAGAAG	TTGGTCGGTTTGATGGCAGG	SalI	12.27	11.04	76.69
CNM01960	13	601704	TCCACGCATCCGGCAAAT	CCCAACCCCAAATGCTCCAT	HaeIII	9.20	6.75	84.05
CNM02290	13	699388	TTCGCTGGTCATCAGCACCA	TTGAGAGAAAGGCGCAGGGT	HaeIII	9.20	5.52	85.28
CNM02490	13	752378	TGGTGTTTCGGCCAGCTTCA	TTGGCGCCTGTTGATGTCGA	PvuII	9.20	5.52	85.28
CNN00060	14	19684	CCCAACCTCATCCCACCTC	ACAGAACCCATTGAGCCCGA	XhoI	9.20	9.20	81.60
CNN00590	14	196771	ACGAGCCGGTGGAACGATAAAC	CTGGAAATTGGTGGCAACGC	HindIII	7.98	8.59	83.44
CNN01880	14	544310	TGAAGAAGGCCGCTTTGGGT	TCGGATTTGTCGGGCAGCAT	HaeIII	23.93	12.27	63.80
CNN02060	14	645541	TTGGAACAGGCCACTCGGAA	ACCGCCAAGGATTCTTGCGA	HaeIII	24.54	12.27	63.19

[a]: Mating type of each progeny was determined using the mating type-specific primers of the ste20 gene through direct PCR and gel electrophoresis.
[b]: PCR product was digested using the combination of two restriction enzymes.

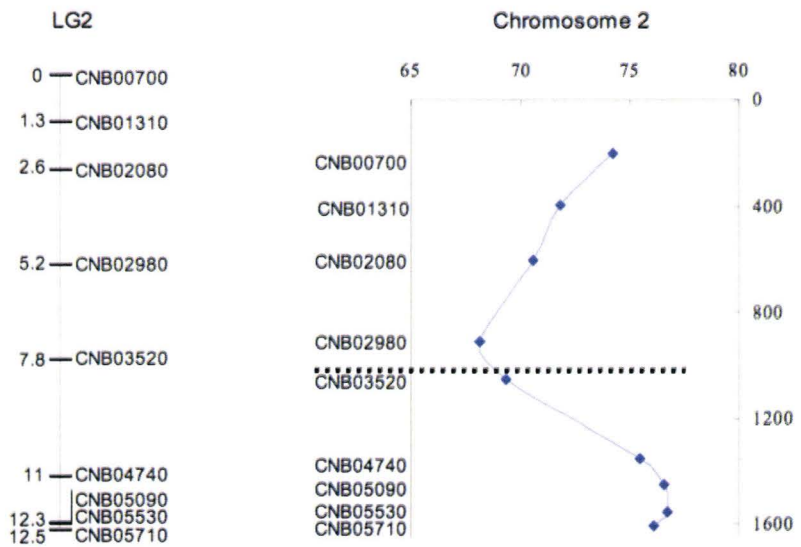
Supplement Figure Legend

Supplement Figure S2.1. The hybrid genetic linkage map of *C. neoformans* generated using a mapping population of 163 progeny and 115 molecular markers. Marker names were indicated on the right side of each linkage group. Numbers on the left side of each linkage group were genetic distances in centi-Morgan. Figures on the right side of each linkage group were the distributions of heterozygosity at each marker locus located on that linkage group, with the marker orders following their physical locations in the chromosome of JEC21. X-axis shows the level of heterozygosity and y-axis shows the physical distance from one end of chromosome. Arrows indicate markers with positions in the linkage groups different from their physical locations in chromosomes of JEC21 (inversions and translocations, see Results). Dotted lines indicated the approximate locations of candidate centromeric regions in each chromosome of JEC21 as identified by LOFTUS et al. 2005. The dash line in the middle of LG8 indicated the position from where LG8 was divided into two LGs, LG8-A and LG8-B,and were two resolved linkage groups when the LOD threshold was increased to 30.

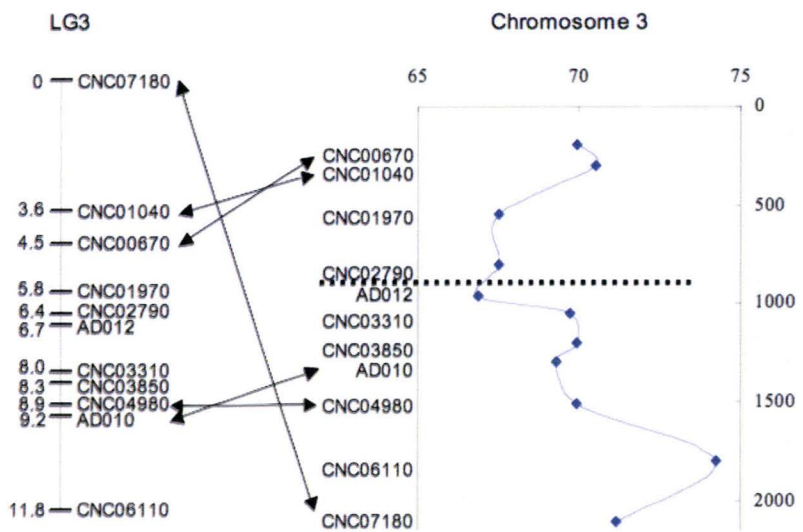
Supplemental Figure S2.1



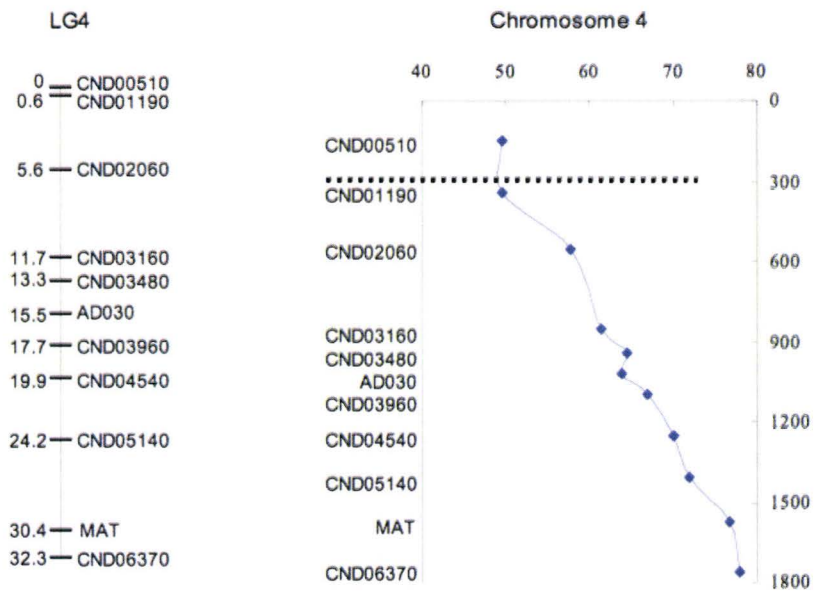
Supplemental Figure S2.1 (continued)



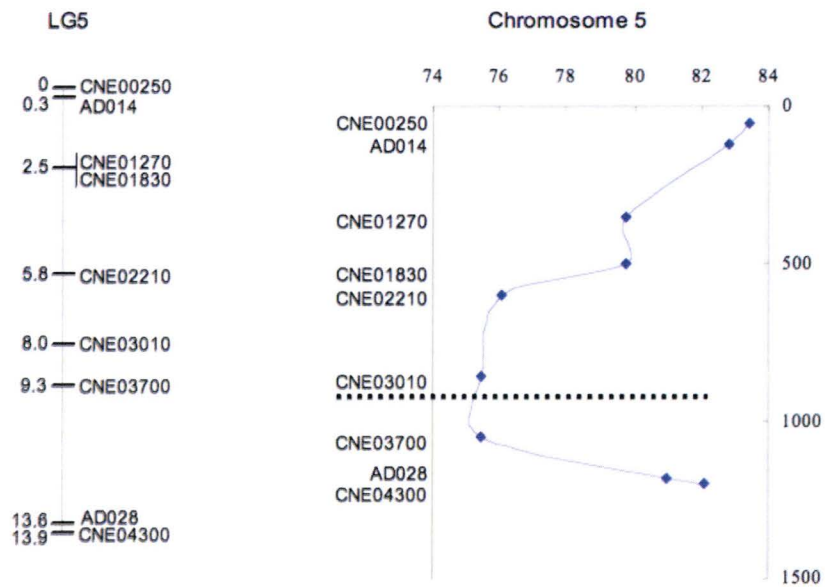
Supplemental Figure S2.1 (continued)



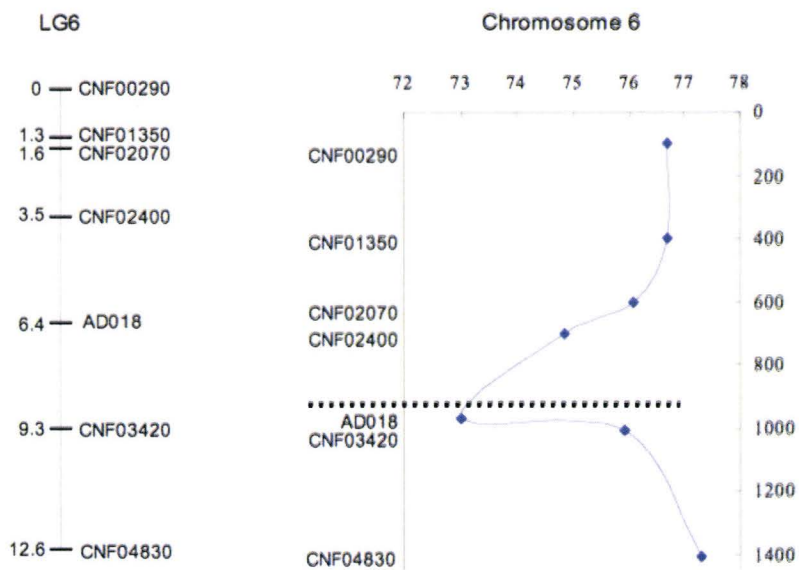
Supplemental Figure S2.1 (continued)



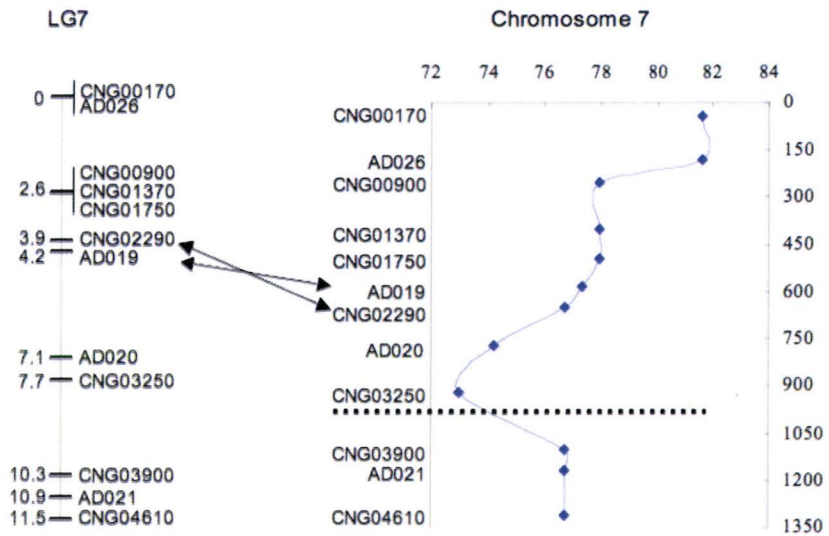
Supplemental Figure S2.1 (continued)



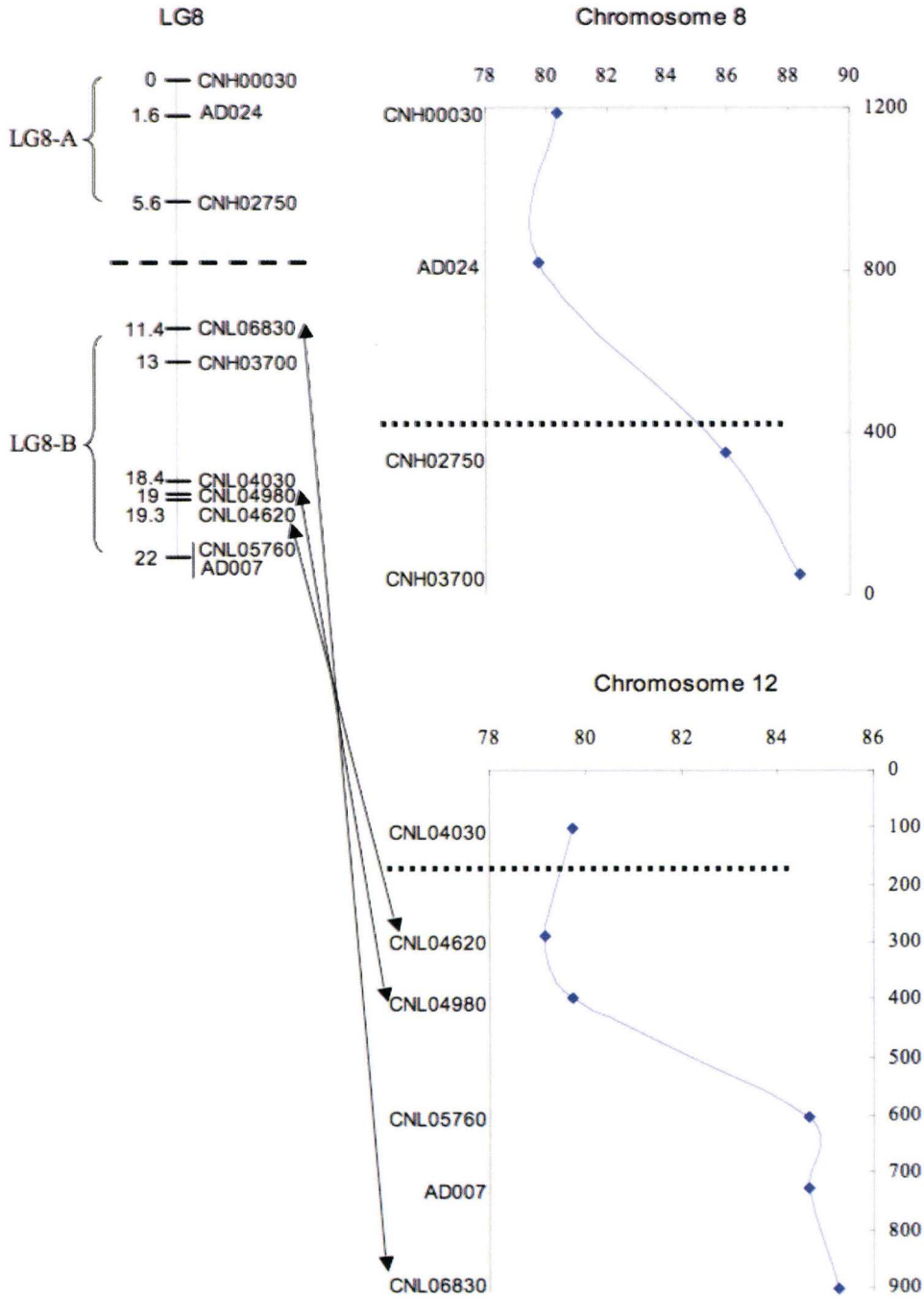
Supplemental Figure S2.1 (continued)



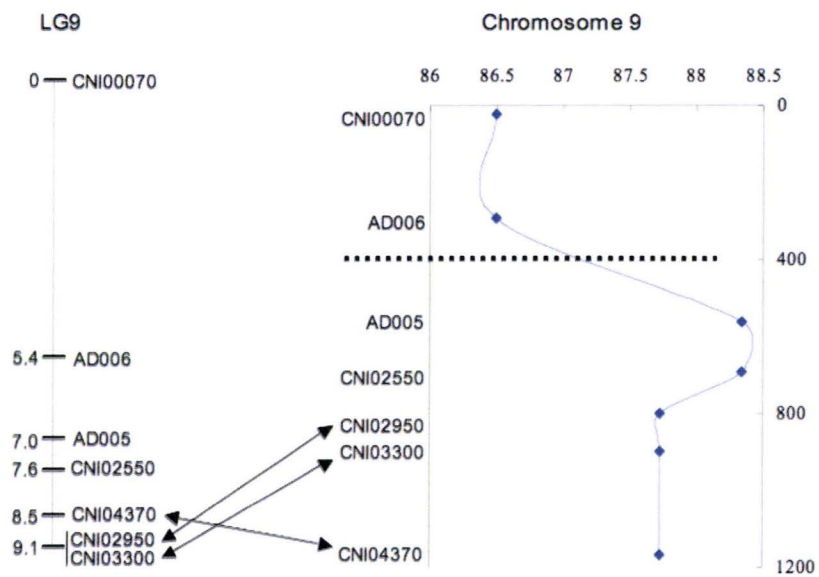
Supplemental Figure S2.1 (continued)



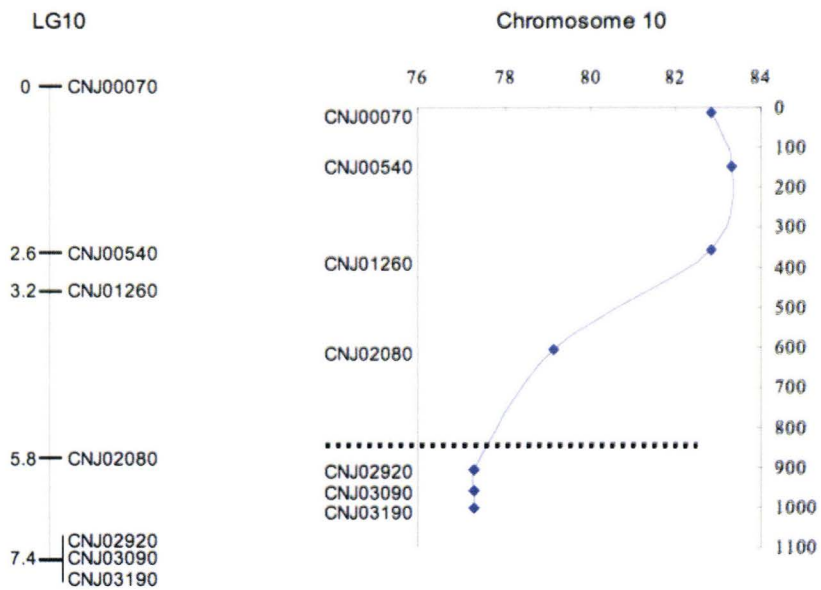
Supplemental Figure S2.1 (continued)



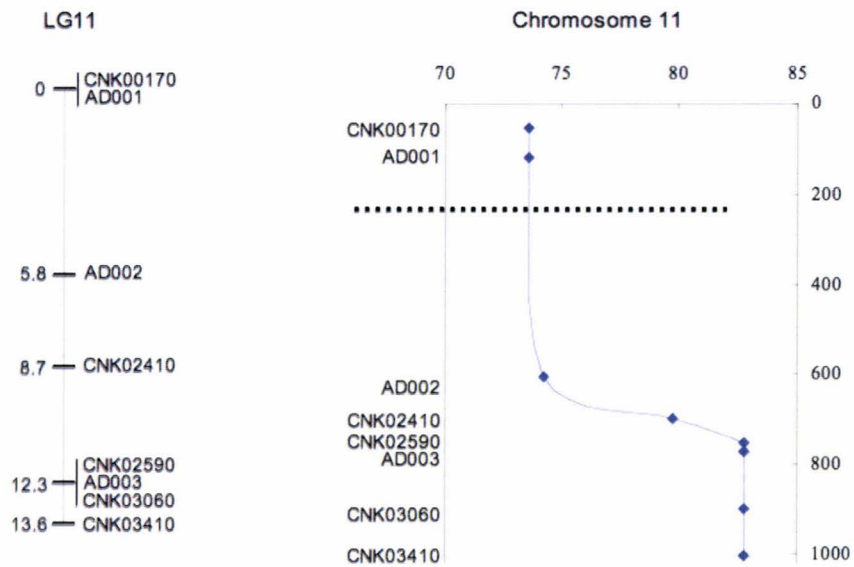
Supplemental Figure S2.1 (continued)



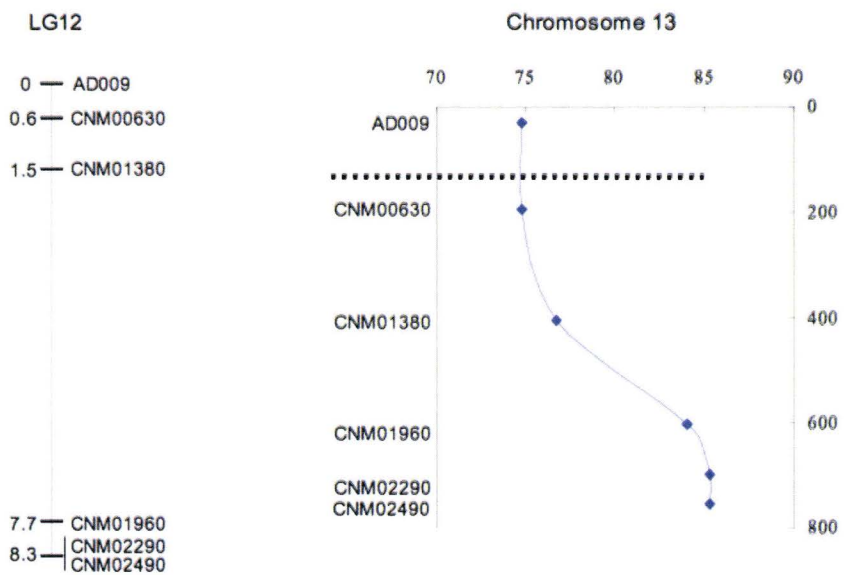
Supplemental Figure S2.1 (continued)



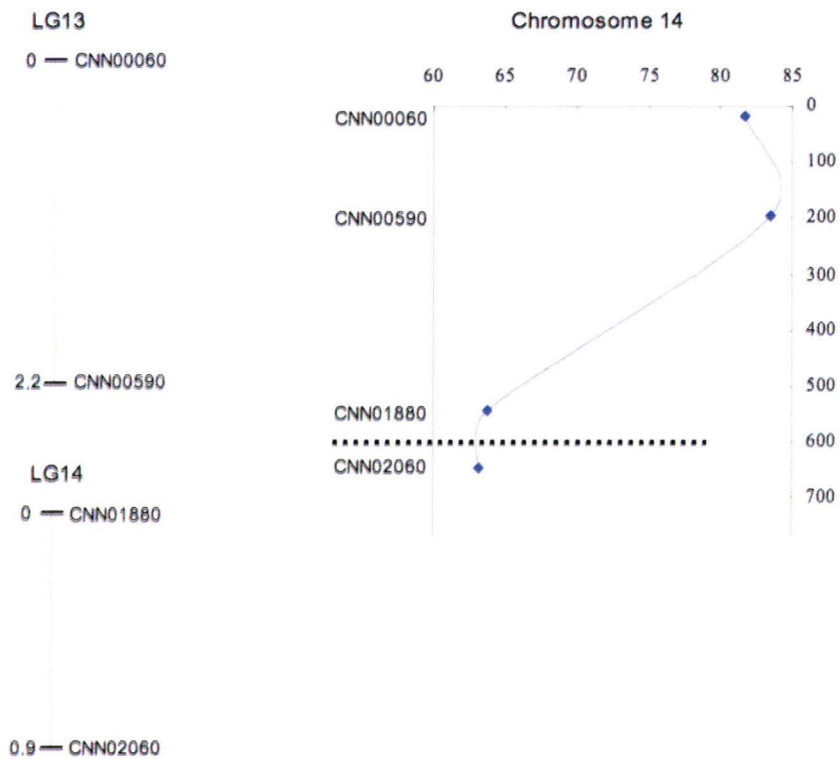
Supplemental Figure S2.1 (continued)



Supplemental Figure S2.1 (continued)



Supplemental Figure S2.1 (continued)



CHAPTER 3

CHROMOSOMAL REARRANGEMENGS BETWEEN SEROTYPES A AND D STRAINS IN *CRYPTOCOCCUS NEOFORMANS*

PREFACE

In the previous chapter (Chapter 2), I found that recombination frequencies were highly reduced in the inter-variety hybridization compared to the intra-variety cross. Also, the recombination frequency was not evenly distributed among chromosomes and among regions within certain chromosomes. As mentioned in the General Introduction, the repression of recombination frequency during inter-variety hybridization could be due to several reasons, including low sequence similarity between the two varieties, the existence of chromosomal rearrangements (e.g. inversions), as well as the existence of epistasis. However, little is known about the relative contributions of these mechanisms to the repression of recombination during hybridization.

In this chapter, I examined the number of chromosomal rearrangements between serotypes A and D genomes, and estimated the contributions of these chromosomal rearrangements to the repression of recombination during inter-variety hybridization. In addition, I surveyed the distributions of the identified chromosomal rearrangements in natural *C. neoformans* populations. Answers to these questions can help us to better understand the genome structure and evolution of *C. neoformans*.

Chapter 3 has been published in the journal PLoS ONE:
Sun, S., and J. Xu, 2009 Chromosomal rearrangements between serotype A and D strains in *Cryptococcus neoformans*. PLoS ONE 45: e5524.

Sheng Sun was the primary researcher of this study. He carried out all the experiments and wrote the draft of the manuscript.

ABSTRACT

Cryptococcus neoformans is a major human pathogenic fungus that can cause meningoencephalitis in immunocompromised hosts. It contains two divergent varieties, var. *grubii* (serotype A) and var. *neoformans* (serotype D), as well as hybrids (serotype AD) between these two varieties. In this study, we investigated the extent of chromosomal rearrangements between the two varieties, estimated the effects of chromosomal rearrangements on recombination frequencies, and surveyed the potential polymorphisms of the rearrangements among natural strains of the three serotypes. Through the analyses of two sequenced genomes from strains H99 (representing var. *grubii*) and JEC21 (representing var. *neoformans*), we revealed a total of 32 unambiguous chromosome rearrangements, including five translocations, nine simple inversions, and 18 complex rearrangements. Our analyses identified that overall, rearranged regions had recombination frequencies about half of those around syntenic regions. Using a direct PCR screening strategy, we examined the potential polymorphisms of 11 rearrangements among 64 natural *C. neoformans* strains from five countries. We found no polymorphism within var. *neoformans* and very limited polymorphism within var. *grubii*. However, strains of serotype AD showed significant polymorphism, consistent with their hybrid origins coupled with differential loss of heterozygosity. We discuss the implications of these results on the genome structure, ecology, and evolution of *C. neoformans*.

INTRODUCTION

When populations become isolated, genetic differences may accumulate as a result of differential fixation of spontaneous mutations by genetic drift and/or natural selection. Mutations may be classified into two broad types. The first type is small-scale point mutations, including nucleotide substitutions and short insertions or deletions. The second type is large-scale changes that include large duplications, deletions, and chromosomal rearrangements such as inversions and translocations. When large-scale changes occur, the genomic size and gene content of the diverging genomes may remain similar, but the physical locations and/or orientations of certain chromosomal segments can differ among the diverging lineages.

Large-scale genetic changes can arise spontaneously and have been observed in natural populations of many organisms. For example, chromosomal rearrangements are commonly found in bacteria (e.g. see review by HUGHES 2000) as well as in diverse groups of eukaryotic organisms such as Baker's yeast (*Saccharomyces cerevisiae*), fruit flies (*Drosophila melanogaster*), common mosquitoes (*Anopheles gambiae*), and humans (FISCHER et al. 2006; COLUZZI et al. 2002; RANZ et al. 2007; for recent reviews, see RIESEBERG 2001, HOFFMANN et al. 2004, and COGHLAN et al. 2005). Some of these rearrangements significantly impacted the phenotypes of these organisms. For example, in species of *Drosophila*, inversions have been linked to the variation in a diversity of traits such as body size, tolerance and resistance to extreme temperatures, wing size, female fecundity, and male mating success (HOFFMANN et al. 2004). In humans, chromosomal rearrangements have been linked to a large number of diseases, including cancer. As a result, chromosomal rearrangements can serve as direct targets for natural selection and be accumulated in different lineages due to such selection (or by genetic drift in small populations), contributing to the divergence between lineages. It has been known for a long time that chromosomal rearrangements can also play a critical role in speciation. Specifically, homologous but rearranged chromosomes may not be able to undergo proper pairing and/or disjunction during meiosis, thus can suppress recombination and accelerate divergence and speciation by reducing hybrid fitness (RIESEBERG, 2001).

Speciation commonly refers to the complete fixation of alternative alleles in loci involved in reproductive isolation between lineages. These loci, also called isolation loci, can act on a variety of phenotypic and physiological traits to ensure pre-zygotic and/or post-zygotic isolation. However, unless the diverging lineages are allopatric (i.e. they are separated by geographic barriers), the fixation of different alleles at the isolation loci among lineages could be interrupted or eroded by gene flow between lineages through inter-lineage hybridization. During inter-lineage

hybridizations, meiosis and recombination could cause introgression of fixed alleles among the lineages, leading to homogenization and preventing lineage divergence. The locations of these isolation loci could influence the rate of fixation. For example, if the isolation loci are located within or tightly linked to rearranged chromosomal regions, their fixation in different lineages can be significantly facilitated. This is because chromosomal rearrangements can suppress recombination during inter-lineage hybridization by interrupting proper pairing between homologous chromosomes during meiosis or by the generation of non-viable progeny because of irregularities in chromosome number (e.g. aneuploidy) and/or chromosome structure (e.g., chromosomes with no or two centromeres, RIESEBERG, 2001).

Recent studies have shown that both small-scale point mutations and large-scale genome changes can play important roles in shaping the evolutionary history and population structure of closely related lineages/species (RIESEBERG et al. 1995; RIESEBERG et al. 1999; DEPAULIS et al. 1999; for recent reviews, see RIESEBERG 2001; MITCHELL-OLDS et al. 2007). Therefore, identifying chromosome rearrangements between diverging lineages and studying the distributions of these rearrangements in natural populations can provide significant information about the evolutionary histories of different lineages and help us understand the population structure and dynamics of the organisms.

Cryptococcus neoformans is an opportunistic human fungal pathogen that can cause meningitis, mainly in immuno-compromised patients (e.g. AIDS and transplant surgery patients). In AIDS patients, it has been estimated that *C. neoformans* accounts for up to 15% of total fatality (CASADEVALL and PERFECT 1998). Because of its medical importance and ease of genetic manipulation, *C. neoformans* has been one of the most extensively studied human fungal pathogens in the past couple of decades. *C. neoformans* contains strains belonging to three serotypes, A, D, and AD. Strains of serotypes A and D are generally haploid and correspond to two varieties, var. *grubii* and var. *neoformans*, respectively; while strains of serotype AD are recent natural hybrids between strains of serotypes A and D and are mostly diploid or aneuploid (LENGELER et al. 2001; XU et al. 2002; 2003). Gene genealogies have suggested that these two varieties have diverged from each other for at least 18.5 million years (XU et al. 2000), and DNA sequence divergence between serotype A and D strains is between 10-15% (KAVANAUGH et al. 2006). Despite their significant divergence, mating between strains of the two varieties can occur. However, the viability of meiotic progeny from the hybrid cross is typically low (LENGELER et al. 2001). Furthermore, recombination frequencies across many chromosomal regions were significantly lower in an inter-variety cross than in intra-variety crosses (FORCHE et al. 2001; MARRA et al. 2004; SUN and XU, 2007). These

results suggest that certain types of genetic differences between the two varieties compromised meiosis and recombination during inter-variety hybridization and that partial reproductive isolation has been established between the two varieties. At present, the type, number, and distributions of specific genes involved in the partial reproductive isolation between the two varieties remain unknown.

While most population and epidemiological studies of *C. neoformans* have focused on point mutations in DNA sequences, several studies using pulse-field gel electrophoresis (PFGE) have shown that extensive karyotype variation (i.e., chromosome number and size) exists among natural *C. neoformans* strains (FRIES et al. 1996; KLEPSER and PFALLER, 1998; BOEKHOUT and BELKUM, 1997; SUKROONGREUNG et al. 2001). FRASER et al. (2005) showed that during the process of constructing the isogenic *C. neoformans* JEC20/JEC21 laboratory strains, telomere-telomere fusion and chromosomal breakage likely had occurred, resulting in a large translocation and segmental duplication in the JEC20/JEC21 genome compared to their parental strains. These studies suggest that chromosomal rearrangements might occur frequently in *C. neoformans*. However, while PFGE can only detect large insertions/deletions and translocations that involve more than one chromosome, it cannot detect intra-chromosomal translocations or inversions.

Since the genomes of two representative strains of *C. neoformans* are available, one for a serotype A strain H99 and the other for a serotype D strain JEC21, the most efficient way to study potential inter-variety chromosomal rearrangements would be to directly compare genome sequences and chromosome organizations between the two strains. Indeed, a recent comparison of the genomes of strains H99 and JEC21 identified that these two genomes were overall highly syntenic (KAVANAUGH et al. 2006). Interestingly, two large regions of high sequence identity (~95%; likely due to introgression) and three inversions were identified between the two strains (KAVANAUGH et al. 2006). However, their criterion of sequence identity (>94%) for identifying inversions was higher than the whole-genome average. As a result, the number of inversions identified in their study is likely an underestimate. Indeed, because inversions and other types of chromosome rearrangements would likely accelerate sequence divergence between lineages relative to adjacent genomic regions, the regions around rearrangements might have lower nucleotide identity than other regions, further causing underestimates of potential rearrangements. Furthermore, the distributions of the rearrangements identified between the two sequenced genomes have not been analyzed among natural strains of *C. neoformans* to determine whether such rearrangements are strain-specific or serotype-specific. The number and distribution of rearrangements

could provide valuable information for understanding the evolution of genomic architecture in *C. neoformans* and for understanding the genetic basis of partial reproductive isolation between the two varieties.

In this study, we used a set of flexible criteria to identify the number, location and distribution of chromosome rearrangements between the two varieties of *C. neoformans*. We then examined the potential polymorphisms of the non-centromeric chromosomal rearrangement regions, including both simple inversions and complex rearrangements (see below), in a collection of natural *C. neoformans* strains. We were specifically interested in the following questions. First, what types of chromosomal rearrangements are there between the two sequenced serotypes A and D genomes? And, how many unambiguous rearrangements can we detect? Second, do regions with chromosome rearrangements show lower levels of recombination frequency than those without rearrangements? And third, what is the pattern of distribution for the chromosome rearrangements among natural *C. neoformans* strains? Will we see rearrangement polymorphisms among strains within the same serotype?

RESULTS AND DISCUSSION

Overall genome structure comparisons between H99 and JEC21

The two genomes that we compared, H99 (18874 kb) and JEC21 (19052 kb), were less than 1% different in size. Each of the two genomes has 14 chromosomes, and the blastn results showed that there was an overall one-to-one correspondence between chromosomes from the two genomes (Table 3.1). The exceptions were JEC21 chromosomes 3 and 11, which have been shown to be involved in large-scale translocations (TRs, see below). Ten of the 12 homologous chromosome-pairs had size differences that were less than 6% of the respective JEC21 chromosome. The other two chromosome-pairs showed relatively large size variations (19.6% and 22.4% for chromosome pairs involving JEC21 chromosomes 8 and 12, respectively; Table 3.1) due to the existence of translocations (see below).

Chromosomal rearrangements between H99 and JEC21 genomes

We found that the genome structures of JEC21 and H99 were mostly syntenic. A total of 32 chromosomal regions showed unambiguous rearrangements between these two genomes (Table 3.2). Lowering the sequence identity from 85% to 75% and reducing the match lengths to less than 200bp did not increase the number of unambiguous chromosome rearrangement regions (data not shown). Because the JEC21 genome has already been published while the H99 genome is not yet, we used the JEC21 chromosomes as the reference and refer to the H99 chromosomes as the rearranged types. Below we describe each of three types of chromosomal rearrangements found between the two genomes.

Large translocations (TR)

Karyotypic variation has been reported previously among *C. neoformans* environmental and clinical isolates (BOEKHOUT and BELKUM, 1997; KLEPSER and PFALLER, 1998; SUKROONGREUNG et al. 2001). Some of this variation was found among isolates from the same patient at different episodes of infection and from samples collected before and after passage in mice (FRIES et al. 1996). While the mechanisms for these observed variations were unknown, such studies suggested that translocations and large-scale deletions and duplications might occur frequently in *C. neoformans*.

Table 3.1. The One-to-One correspondence between chromosomes from H99 and JEC21 based on reciprocal blast searches.

JEC21		H99		
Chromosome	Size (bp)	Chromosome	Size (bp)	% Difference[a]
1	2300533	1	2291499	-0.39
2	1632307	2	1621675	-0.65
4	1783081	5	1814975	+1.79
5	1507550	6	1422463	-5.64
6	1438950	7	1399503	-2.74
7	1347793	8	1398693	+3.78
8[b]	1194300	14[b]	926563	-22.42[b]
9	1178688	9	1186808	+0.69
10	1085720	10	1059964	-2.37
12[b]	906719	4[b]	1084805	+19.64[b]
13	787999	12	774062	-1.77
14	762694	13	756744	-0.78
3[c]	2105742	3[c]	1575141	n.a.[d]
11[c]	1019846	11[c]	1561194	n.a.[d]
Total	19051922		18874089	-0.93

[a]: Percentages were calculated by dividing the size differences between the two chromosomes using the sizes of respective JEC21 chromosomes. A positive number indicates that the H99 chromosome is larger than the corresponding JEC21 chromosome. A negative number indicates that the H99 chromosome is smaller than the corresponding JEC21 chromosome.

[b]: Chromosomes in which there are large translocations regions.

[c]: Chromosomes for which homologous chromosomes are not established due to the existence of large scale of translocations.

[d]: Not calculated due to the existence of the large scale of translocations

Table 3.2. Specific chromosomal rearrangements between H99 and JEC21 genomes

Region [a]	JEC21 (kb)				H99 (kb)			
	Start	End	Size[b]	Adjacent to transposable element [c]	Start	End	Size[b]	Adjacent to transposable element [c]
SI(1)A	0	27	27	No	0	49	49	No
SI(1)B	852	932	80	Yes	879	960	81	Yes
SI(1)C	1874	1881	7	No	1851	1858	7	No
SI(1)D	2289	2301	12	No	2267	2291	24	Yes
SI(3)	1218	1228	10	No	771	781	10	No
SI(4)	1621	1633	12	No	1613	1623	10	Yes
SI(5)	1395	1406	11	No	1371	1381	10	No
SI(8)	846	860	14	Yes	571	588	17	No
SI(9)	721	1115	394	Yes	716	1106	390	No
CR(1)[c]	937	998	61	Yes	965	1007	42	Yes
CR(2)[c]	855	905	50	Yes	835	893	58	Yes
CR(3)[c]	745	911	166	Yes	109	235	126	Yes
CR(4)A[c]	217	279	62	Yes	233	256	23	Yes
CR(4)B	768	782	14	No	762	778	16	No
CR(4)C	1525	1621	96	Yes	1534	1613	79	Yes
CR(5)[c]	775	856	81	Yes	780	823	43	Yes
CR(6)A	75	117	42	No	69	109	40	No
CR(6)B[c]	863	939	76	Yes	828	874	46	Yes
CR(7)[c]	882	912	30	Yes	893	948	55	Yes
CR(8)[c]	706	762	56	Yes	464	485	21	Yes
CR(9)[c]	324	389	65	Yes	346	386	40	Yes
CR(10)[c]	802	879	77	Yes	829	858	29	Yes
CR(11)[c]	143	172	29	Yes	871	922	51	Yes
CR(12)[c]	129	177	48	Yes	331	376	45	Yes
CR(13)[c]	122	183	61	Yes	139	171	32	Yes
CR(14)A	45	63	18	Yes	3	26	23	Yes
CR(14)B[c]	567	645	78	Yes	579	633	54	Yes
TR(3)A	0	212	212	No	1357	1575	218	No
TR(3)B	212	1080	868	No	550	1106	556	No
TR(3)C	1815	2105	290	No	0	550	550	No
TR(8)	0	245	245	Yes	0	202	202	Yes
TR(11)	0	592	592	No	0	642	642	No

[a]: SI: simple inversion; CR: complex rearrangement; TR: translocation. Numbers in parentheses indicate the specific chromosomes in the JEC21 genome on which the rearrangements were located, and their corresponding chromosomes can be found in Table 3.1.

[b]: Sizes were calculated as the physical distances between the two syntenic chromosomal regions flanking the chromosomal rearrangements.
[c]: CR regions corresponding to the proposed centromeric regions in LOFTUS et al. (2005).

Using the blast search strategy of chromosome against chromosome, we identified that five large chromosomal regions in the JEC21 genome were likely translocated in the H99 or JEC21 genomes. One translocation is intra-chromosomal, which involves a region [TR(3)A] that has been inversely translocated from the beginning to the end of the chromosome 3 (Figure 3.1, chromosome 3). The other four regions, located on JEC21 chromosomes 3, 8 and 11 ranging in sizes between 212kb and 868kb (Table 3.2 and Figure 3.1), did not have homologous sequences in their corresponding chromosomes, indicating that they might have been involved in inter-chromosomal translocations.

To confirm that these five TR regions were indeed translocated (i.e. they were not large insertion/deletion in one of the genomes), we used nuclear sequences of these regions from JEC21 as queries and blasted them against the whole H99 genome. The locations of the homologous regions of these "putative" chromosomal translocations in the H99 genome were determined. Our results showed that each TR region had one and only one corresponding homologous chromosomal region in the H99 genome (Figure 3.1, chromosomes 3, 8, 11 and 12), confirming that these TR regions were indeed translocations.

One of the four inter-chromosomal translocations occurred between chromosomes 8 and 12 [TR(8); Pink bar in Figure 3.1], which has been previously described by FRASER et al. (2005) and has been shown to be the result of chromosomal fusion and re-breakage events that occurred during the process of producing the isogenic JEC20/JEC21 strains.

For two of the other three TR regions, blast hits for the middle and one end of JEC21 chromosome 3 (Figure 3.1, represented by grey and green bars on chromosome 3, respectively) were located at one end of chromosome 11 in the H99 genome. For the fourth translocation, one end of chromosome 11 in the JEC21 genome (Figure 3.1, chromosome 11 yellow bar) had significant blast hits located at one end of chromosome 3 in the H99 genome. Interestingly, the orientations of some of these chromosomal segments on the two chromosomes also differed between the two genomes.

Figure 3.1. The one-to-one comparisons between chromosomes of H99 and JEC21. For each comparison, the chromosome from JEC21 is shown on top and that of H99 at the bottom. Colored lines between the paired chromosomes represent the correspondence between regions from the two chromosomes, and different colors indicate different lengths of the blast hits (red – more than 10000 bp; blue – 5000-10000 bp; light blue – 1000-5000 bp; yellow – 500-1000 bp; pink – less than 500 bp). Chromosomal rearrangements (SI, CR and TR, see MATERIALS and METHODS) are indicated in brackets. The CRs with two stars (**) are regions corresponding to proposed centromeric regions in the JEC21 genome (LOFTUS et al. 2005). The mating-type locus (MAT) is indicated on chromosome 4. Chromosomes 3 and 11 in the two genomes were compared, despite the existence of large-scale translocations in these chromosomes. The colored bars in chromosomes 3 and 11 correspond to those colored block arrows used in Figure 3.2 (see below).

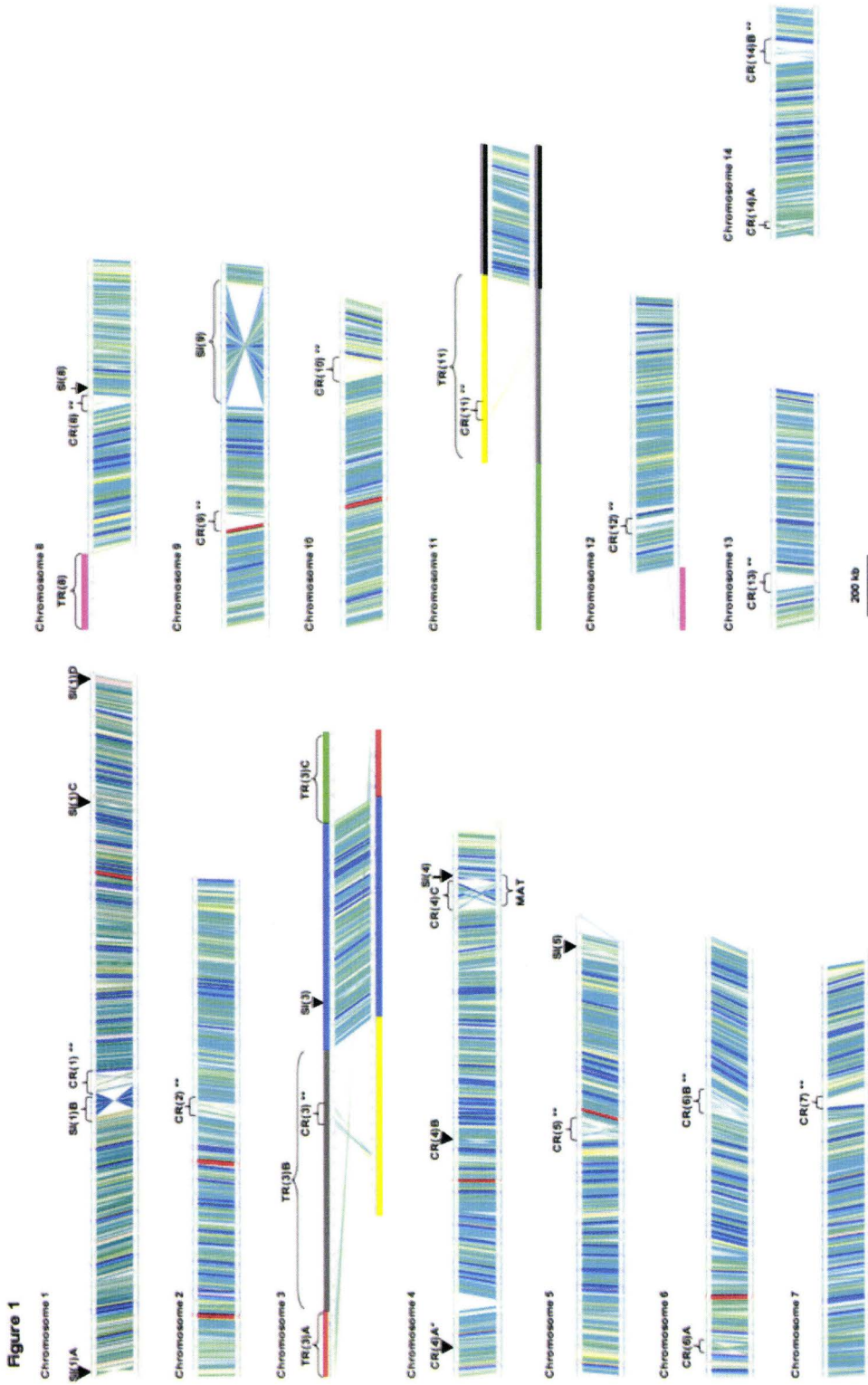


Figure 1
Chromosome 1
Chromosome 2
Chromosome 3
Chromosome 4
Chromosome 5
Chromosome 6
Chromosome 7
Chromosome 8
Chromosome 9
Chromosome 10
Chromosome 11
Chromosome 12
Chromosome 13
Chromosome 14
TR(6)
CR(6) **
SI(6)
CR(9) **
SI(9)
CR(10) **
TR(11)
CR(11) **
CR(12) **
CR(13) **
CR(14)A
CR(14)B **
200 kb
SI(1)A
SI(1)B
CR(1) **
SI(1)C
SI(1)D
CR(2) **
TR(3)A
TR(3)B
CR(3) **
SI(3)
TR(3)C
CR(4)A*
CR(4)B
CR(4)C
SI(4)
MAT
CR(5) **
SI(5)
CR(6)A
CR(6)B **
CR(7) **

To further infer the evolutionary history of chromosomes 3 and 11, we blasted these two chromosomes in strains JEC21 and H99 against the genome sequence of strain R265 (http://www.broad.mit.edu/annotation/genome/cryptococcus_neoformans_b/Home.html), which belongs to *Cryptococcus gattii*, a closely related species of *C. neoformans*. Though the R265 genome is not completely assembled and annotated, its preliminary assembled status organized in more than 700 contigs allowed the identification of four junctions among the segments of the translocated regions. These regions are shown in Figure 3.2 with different colors indicating the chromosomal blocks across the three strains. The results suggest that the R265 chromosomal structures at these junctions likely represent the ancestral chromosomal structures of H99 (serotype A) and JEC21 (serotype D). Based on this inference, we propose a most parsimonious ancestral organization for chromosomes 3 and 11 and deduce the likely events that could have been responsible for generating the chromosomal structural polymorphisms between strains JEC21 and H99 (Figure 3.2).

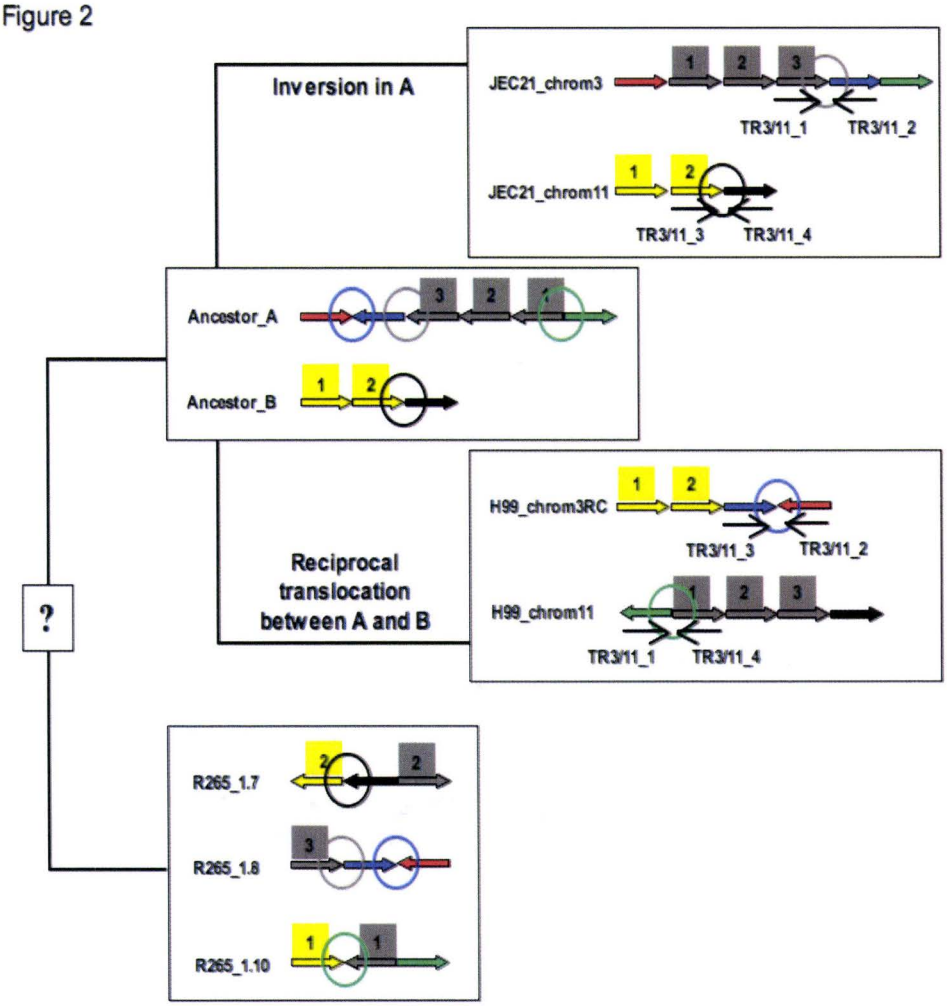
Using our direct PCR strategy, we analyzed the chromosomal states at these two junctions (the black and gray circles in Figure 3.2) among 64 natural *C. neoformans* strains. Our results indicated all the strains had chromosomal structures identical to those of JEC21 at these two junctions. These results suggest JEC21 likely represents the ancestral states of the chromosomal structures at these two regions in *C. neoformans* and that the H99 chromosomal structures were likely generated by unique recent translocation events (Table 3.3).

Complex rearrangements (CR)

Our genome comparison revealed 18 chromosomal regions showing complex rearrangements between H99 and JEC21 (CRs; Table 3.2 and Figure 3.1), with each including both inversions and small translocations. The sizes of the complex rearranged regions between the two genomes varied between 13kb and 166kb (Table 3.2). Not surprisingly, fourteen of the 18 CRs were located in the proposed centromere regions of the chromosomes in the JEC21 genome (LOFTUS et al. 2005), consistent with previous studies showing that centromeric regions are involved in extensive chromosomal rearrangements. The other four CRs were located on JEC21 chromosomes 4, 6 and 14.

Figure 3.2. A hypothetical evolutionary history of chromosomes 3 and 11 in H99 and JEC21
The colored block arrows indicate the homologous chromosomal regions, and their relative orientations to each other, among H99, JEC21 and R265. The colors correspond to the colored bars in chromosomes 3 and 11 of Figure 3.1. The numbers above the block arrows represent the segments within that block. The colored circles indicate the junctions of chromosomal segments in R265 that have been found in H99 (serotype A) or JEC21 (serotype D). Arrows labeled with TR3/11_1, 2, 3 and 4 indicated the locations and orientations of the primers used for PCR confirmation of the chromosomal types of the translocation in natural isolates. The table at the bottom lists the confirmed PCR results (positive and negative) from different primer combinations in H99 and JEC21.

Figure 2



Primer	Position (JEC21, bp)	Sequence (5' - 3')
TR3/11_1	Chrom3: 1080016	GCCTCAGATGCCTCATGGAT
TR3/11_2	Chrom3: 1081117	GCTCGGAAACACGGACCAAA
TR3/11_3	Chrom11: 591433	ACAGCTTATAGCCCTCGCACC
TR3/11_4	Chrom11: 592075	TGTCGCCACTGGTTGGGAAT

Confirmation of translocations between the genomes of JEC21 and H99

Primer pair	(1) TR3/11_1&2	(2) TR3/11_3&4	(3) TR3/11_1&4	(4) TR3/11_2&3
Serotype A (H99)	-	-	+	+
Serotype D (JEC21)	+	+	-	-

Table 3.3. Summary of strains and their chromosome types at chromosomal rearrangement regions

Strain	Origin	ST [a]	SI(1)A	SI(1)B	SI(1)C	SI(1)D	SI(3)	SI(4)	SI(5)	SI(8)	SI(9)	CR(4)B	CR(6)A	CR(14)A	TR3/11
ATCC34869	ATCC	A	A	A	A	A	A	A	A	A	A	A	A	A	D
Y195-90	BRAZIL	A	A	A	A	A[3] [b]	A	A	A	A	A	A	A	A	D
Y288-90	CANADA	A	A[3] [b]	A[4] [b]	A	n.a.[c]	A	A	A	A	A	A	A	A[11] [b]	D
Y289-90	CANADA	A	A	A	A	A	A	A	A	A	A	A	A	A	D
INN3	INDIA	A	A	A	A	A[3] [b]	A	A	A	A	A	A	A	A	D
Y408-91	USA	A	A	A	A	A[2] [b]	A	A	A	A	A	A	A	A	D
CDC92_001	USA	A	A	A	A	A	A	A	A	A	A	A	A	A	D
CDC92_002	USA	A	A[3] [b]	A[4] [b]	A	n.a.[c]	A	A[4, 5] [b]	A	A	A	A	A	A[11] [b]	D
CDC92_003	USA	A	A	A	A	A	A	A	A	A	A	A	A	A	D
CDC92_004	USA	A	A	A	A	A	A	A	A	A	A	A	A	A	D
CDC92_007	USA	A	A[3] [b]	A[4] [b]	A	n.a.[c]	A	A	A	A	A	A	A	A[11] [b]	D
CDC92_014	USA	A	A[3] [b]	A[4] [b]	A	n.a.[c]	A	A	A	A	A	A	A	A[11] [b]	D
CDC92_015	USA	A	A	A	A	A[3] [b]	A	A	A	A	A	A	A	A	D
CDC92_016	USA	A	A	A	A	A	A	A	A	A	A	A	A	A	D
CDC92_204	USA	A	A	A	A	A	A	A	A	A	A	A	A	A	D
CDC92_205	USA	A	A	A	A	A	A	A	A	A	A	A	A	A	D
CDC92_236	USA	A	A[3, 5] [b]	A	A	A	A	A	A	A	A	A	A	A	D
Y367-91	USA	A	A	A	A	A	A	A	A	A	A	A	A	A	D
Y370-91	USA	A	A	A	A	A	A	A	A	A	A	A	A	A	D
Y393-91	USA	A	A	A	A	A	A	A	A	A	A	A	A	A	D
Y490-91	USA	A	A	A	A	A	A	A	A	A	A	A	A	A	D
Y504-91	USA	A	A[3, 4] [b]	A	A	A	A	A	A	A	A	A	A	A	D

Table 3.3 (continued)

Strain	Origin	ST [a]	SI(1)A	SI(1)B	SI(1)C	SI(1)D	SI(3)	SI(4)	SI(5)	SI(8)	SI(9)	CR(4)B	CR(6)A	CR(14)A	TR3/11
B4962	ZAIRE	A	A	A	A	A[3] [b]	A	A	A	A	A	A	A	A	D
B4963	ZAIRE	A	A	A	A	A[2] [b]	A	A	A	A	A	A	A	A	D
B4964	ZAIRE	A	A	A	A	A[3] [b]	A	A	A	A	A	A	A	A	D
B4968	ZAIRE	A	A	A	A	A[3] [b]	A	A	A	A	A	A	A	A	D
ATCC24067	ATCC	D	D	D	D	D	D	D	D	D	D	D	D[2, 3, 4] [b]	D	D
ATCC34875	ATCC	D	D	D	D	D	D[2] [b]	n.a.[Δ]	D	D	D	D	D	D	D
Y286-90	CANADA	D	D	D	D	D	D	D	D	D	D	D	D[2, 3, 4] [b]	D	D
Y290-90	CANADA	D	D	D	D	D	D	D[1,2] [b]	D	D	D	D	D[2, 3, 4] [b]	D	D
CAP67-2	USA	D	D	D	D	D	D[2] [b]	D[1,2] [b]	D	D	D	D	D	D	D
CDC92_027	USA	D	D	D[1,3] [b]	D	D	D[2] [b]	D	D	D	D	D	D	D	D
CDC92_032	USA	D	D	D	D	D	D	D	D	D	D	D	D	D	D
CDC92_076	USA	D	D	D	D	D	D	D	D	D	D	D	D[2, 3, 4] [b]	D	D
CDC92_077	USA	D	D	D	D	D	D	D[1,2] [b]	D	D	D	D	D[2, 3, 4] [b]	D	D
CDC92_119	USA	D	D	D	D	D	D	D	D[2, 3] [b]	D	D	D	D[2, 3, 4] [b]	D	D
CDC92_134	USA	D	D	D[1,3] [b]	D	D	D[2] [b]	D	D	D	D	D	D	D	D
CDC92_138	USA	D	D	D[1,3] [b]	D	D	D[2] [b]	D	D	D	D	D	D	D	D
CDC92_170	USA	D	D	D	D	D	D	D	D	D	D	D	D[2, 3, 4] [b]	D	D
CDC92_178	USA	D	D	D	D	D[1] [b]	D[1] [b]	D	D	D	D	D	D	D	D
CDC92_198	USA	D	D	D	D	D	D	D[1,2] [b]	D	D	D	D	D[2, 3, 4] [b]	D	D
CDC92_337	USA	D	D	D	D	D	D	D	D	D	D	D	D[2, 3, 4] [b]	D	D
Y494-91	USA	D	D	D	D	D	D	D	D	D	D	D	D	D	D
ATCC48184	ATCC	AD	AD	AD	AD	AD	AD	D[2, 3] [b]	D	AD	AD	AD	A[7, 8]D [b]	A[10, 11] [b]	D

Table 3.3 (continued)

Strain	Origin	ST [a]	SI(1)A	SI(1)B	SI(1)C	SI(1)D	SI(3)	SI(4)	SI(5)	SI(8)	SI(9)	CR(4)B	CR(6)A	CR(14)A	TR3/11
ATCC32719	ATCC	AD	D	D	D	D	AD	A	AD	D	A	AD	AD[1,2, 4] b	D	D
CDC92-005	USA	AD	AD	A[4, 6]D b	AD	AD	AD	A	AD[1,2] b	AD	AD	AD	AD	AD	D
CDC92-026	USA	AD	A	A	A[4] b	A	AD	D[2, 3] b	AD	AD	AD	AD	A[7, 8]D b	A[9]D[2, 5] b	D
CDC92-046	USA	AD	A	A	A[4] b	A	AD	D[2, 3] b	AD	AD	AD	AD	A[7, 8]D b	AD	D
CDC92-047	USA	AD	A	A	A[4] b	A[2, 3, 5] b	AD	D[2, 3] b	AD	AD	AD	AD	A[7, 8]D b	AD	D
CDC92-062	USA	AD	A	A	A[4] b	A	AD	D[2, 3] b	AD	AD	AD	AD	A[7, 8]D b	AD	D
CDC92-066	USA	AD	A	A	A[4] b	A[2, 3, 5] b	AD	D[2, 3] b	AD	AD	AD	AD	A[7, 8]D b	AD	D
CDC92-074	USA	AD	A	A	A[4] b	A	AD	D[2, 3] b	D	AD	A	D	A[7, 8]D b	AD	D
CDC92-174	USA	AD	AD	AD	A[4]D b	AD	AD	D[2, 3] b	D	AD	AD	AD	A[7, 8]D b	AD	D
CDC92-181	USA	AD	AD	AD	AD	AD	AD	A	AD[1,2] b	AD	A	AD	AD	A[9, 11]D b	D
CDC92-190	USA	AD	AD	AD	A	A	AD	AD[2, 3] b	AD[1,2] b	AD	AD	AD	AD[2, 3, 4] b	AD	D
CDC92-228	USA	AD	A	A	A[4] b	A	AD	D[2, 3] b	AD	D	AD	AD	A[7, 8]D b	AD	D
CDC92-280	USA	AD	A	A	A[4] b	A	AD	D[2, 3] b	AD	D	A	AD	A[7, 8]D b	AD	D
CDC92-283	USA	AD	A	A	A[4] b	A	AD	D[2, 3] b	AD	AD	AD	AD	A[7, 8]D b	AD	D
CDC92-304	USA	AD	AD	AD	A[4]D b	AD	AD	D[2, 3] b	D	AD	AD	AD	A[7, 8]D b	A[9, 11]D b	D
CDC92-328	USA	AD	A	A	A[4] b	A	AD	D[2, 3] b	AD	AD	AD	AD	A[7, 8]D b	AD	D
CDC92-354	USA	AD	A	A	A[4] b	A	AD	D[2, 3] b	AD	AD	AD	AD	A[7, 8]D b	AD	D
CDC92-355	USA	AD	A	A	A[4] b	A	AD	D[2, 3] b	AD	AD	AD	AD	A[7, 8]D b	AD	D
CDC92-383	USA	AD	AD	AD	AD	A[2, 3]D[1] b	AD	n.a. [c]	AD[1,2] b	AD	AD	D	AD	AD	D
Y520-91	USA	AD	A	A	A[4] b	A	AD	D[2, 3] b	D	AD	AD	AD	A[7, 8]D b	AD	D

[a]: ST: Serotype. Serotypes identified by traditional method.
[b]: Superscript number(s) within brackets refers to the numbers of primer pairs that worked for that strain at that rearrangement region (see Supplemental Figures). Character (i.e. A or D) without a superscript number indicates all the primer pairs expected to work with that strain worked.
[c]: No primer pair worked.

CR(4)C was located within the mating type (MAT) locus (Figure 3.1) and our observation is consistent with results from previous studies showing that the MAT locus in *C. neoformans* contained extensive rearrangements (LENGELER et al. 2002, LOFTUS et al. 2005). CR(4)B and CR(6)A each contained one inversion and one local translocation, and CR(14)A contained two inversions and two local translocations (Figure 3.1 and Supplemental Figures S3.A to S3L.).

Simple inversions (SI)

We identified nine chromosomal segments flanked by syntenic sequences but were in reverse orientations between the H99 and JEC21 genomes. These nine simple inversions (SIs) were all paracentric and they were located on six different chromosomes (Table 3.2; Figure 3.1). There were significant size differences among the nine SI regions. The smallest SI region, SI(1)D on chromosome 1, was only 3 kb in size and contained three genes, while the largest SI region, SI(9) on chromosome 9, was about 394 kb in size and contained 151 genes.

In the study by KAVANAUGH et al. (2006), three inversions were identified between these two genomes. Two of those three inversions corresponded to two SI regions identified in our study: SI(1)A and SI(1)B. We also found one SI region in the location of the third inversion reported by KAVANAUGH et al. (2006), but the size of our SI was smaller (12 kb compared to 70 kb) than they reported. The reason that we identified more SI regions was probably because the criteria used to filter blast hits in our study were more relaxed than those applied in the KAVANUGH et al. study (2006). In their study, only regions longer than 1000 bp with nucleotide identity higher than 94% were considered and inversions with lower sequence similarities and/or shorter lengths would have been missed.

The nine SIs identified between the H99 and JEC21 genomes in our study were fewer than those reported in other genome comparison studies involving other fungal species. For example, FISCHER et al. (2006) found that the numbers of small inversions ranged from 59 to 773 between pairs of hemiascomycetous yeast species. They identified that the lowest number of inversions was between the most closely related species pair, *Kluyveromyces lactis* and *Ashbya gossypii*, which had 59 small inversions. One reason for the small number of inversions observed here might be because we did not include the ambiguous putative small inversions located within the four large translocations and within the CR regions [e.g. CR(4)B, Figure 3.1]. Another reason might be related to the length of time the compared genomes have diverged from each other. The genome pairs analyzed in the study by FISCHER et al. (2006) were from different, reproductively isolated species that have probably diverged from each other for a lot longer than between varieties *grubii* and *neoformans* of *C. neoformans*.

Interestingly, though the number of inversions identified between the JEC21 and H99 genomes seemed fewer than those found between the hemiascomycetous species, the sizes of the inversions were bigger between JEC21 and H99 than those between the yeast species. Specifically, FISCHER et al. (2006) found that the average numbers of genes per inversion were usually less than three in their comparisons. In contrast, the average number of genes within the 11 SI regions was ~27 between JEC21 and H99, with one inversion, SI(9), containing 151 genes (detailed data not shown). The contrasting patterns in the number and size distributions found between this study and those in FISCHER et al. (2006) suggest that inversions arising at the beginning of speciation may be relatively larger than those typically observed between the genomes that have diverged for long periods of time. Over time, other events such as hybridization and chromosomal introgression could have broken up the originally inverted regions, by double recombination, gene conversions, and/or additional inversions. Breaking large inversions into smaller ones might be favored by natural selection if these chromosome segments contained or were tightly linked to loci involved in reproductive isolation, as it has been suggested that multiple shorter inversions might have a stronger effect in suppressing recombination than one large inversion (NAVARRO et al. 1997; NAVARRO and RUIZ, 1997).

Association between chromosomal rearrangements and transposable elements

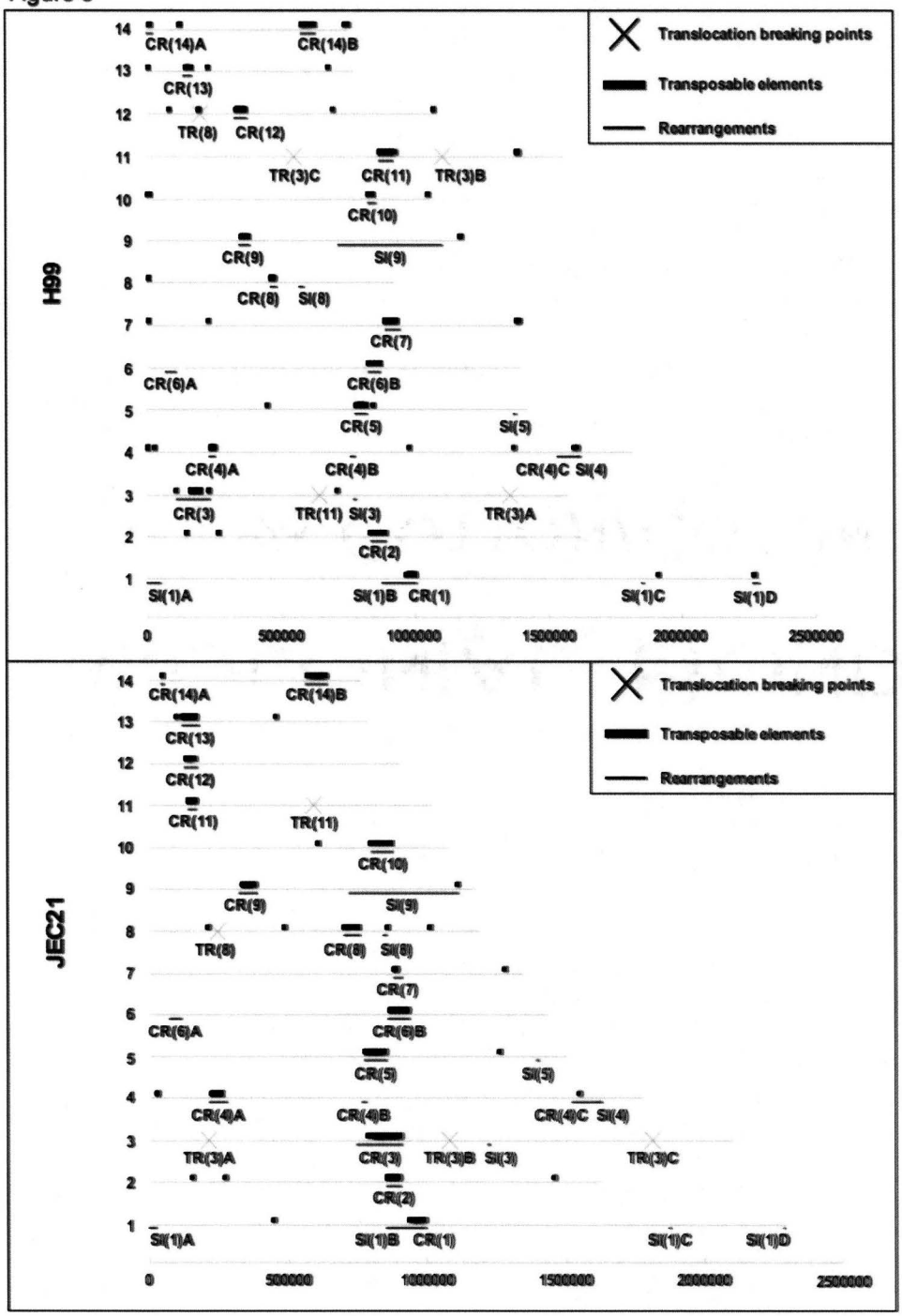
Since previous studies have suggested that transposable elements could facilitate the occurrence of chromosomal rearrangements (FRASER et al. 2005), we screened the genomes of H99 and JEC21 for the existence of homologous sequences of 30 transposable elements that have been identified in *C. neoformans* so far, and some of these transposable elements had been described previously by Goodwin and Poulter (2001; see Materials and Methods).

We found that for the boundaries of TR regions, only that of TR(8) was located in close proximity of transposable element in both H99 and JEC21 (Table 3.2 and Figure 3.3). This is consistent with the proposal in a previous study suggesting that the presence of transposable elements close to the TR(8) region might be responsible for the occurrence of that translocation event (through non-homologous recombination; FRASER et al. 2005).

Of the 18 CR regions, all except CR(4)B and CR(6)A overlapped with the chromosomal regions that had high densities of transposable elements in both H99 and JEC21 (Table 3.2 and Figure 3.3).

Figure 3.3. Distribution of chromosomal rearrangements reported in this study and transposable elements in the H99 and JEC21 genomes
X axis indicates the length of the chromosomes (bp). On the Y-axis, each line represents a chromosome. The black blocks above the line indicate the presence of transposable elements. The bold line segments under the line indicate the locations of chromosomal rearrangements. The crosses on the line indicate the breaking points of the translocation regions.

Figure 3


H99
JEC21
Translocation breaking points
Transposable elements
Rearrangements
CR(14)A
CR(14)B
CR(13)
TR(8)
CR(12)
TR(3)C
CR(11)
TR(3)B
CR(10)
CR(9)
SI(9)
CR(8)
SI(8)
CR(7)
CR(6)A
CR(6)B
CR(5)
SI(5)
CR(4)A
CR(4)B
CR(4)C
SI(4)
CR(3)
TR(11)
SI(3)
TR(3)A
CR(2)
SI(1)A
SI(1)B
CR(1)
SI(1)C
SI(1)D
TR(3)B
TR(3)C
0
500000
1000000
1500000
2000000
2500000

While the majority of the CR regions were located close to transposable elements, only five of the nine SI regions were found to have transposable elements nearby in the H99 and/or JEC21 genome. Specifically, SI(1)B had a transposable element located in its close proximity in both the H99 and JEC21 genomes; SI(1)D and SI(4) had transposable elements located close to one of their respective boundaries in the H99 genome; and SI(8) and SI(9) had one transposable element each located close to one of their respective boundaries in the JEC21 genome (Table 3.2 and Figure 3.3).

These results suggest that simple inversions in these genomes were probably less affected by transposable elements than the CR regions. The dearth of transposable elements around SI regions with low complexity of rearrangements in comparison to the high frequency of transposable elements around CR regions with extensive rearrangement are consistent with the roles of transposable elements in facilitating chromosomal rearrangements. The potential role of transposable elements in chromosomal rearrangements is also supported by the observation that in most cases, the transposable elements identified in the close proximity of a rearrangement region belong to same type of transposons (detailed results not shown). However, it should be pointed out that factors other than transposable elements (e.g. the presence of repetitive sequences and tRNAs) could also facilitate chromosomal rearrangements.

Low recombination frequencies surrounding the regions with CR and SI

Using the genetic linkage map constructed recently by SUN and XU (2007), we calculated the recombination frequencies between each adjacent pair of markers (Table 3.4). In that study, the authors used 115 PCR-RFLP markers to construct a hybrid genetic linkage map between serotypes A and D *C. neoformans*. For each marker, we retrieved its physical location in the JEC21 genome (LOFTUS et al., 2004) and genetic location in the linkage map (SUN and XU, 2007). Then for each pair of adjacent markers, the recombination frequency between them was calculated as cM/kb, i.e. the ratio of genetic distance over physical distance between the two markers. The higher the value, the more recombination events occurred per unit of physical distance, and thus the higher the recombination frequency. We excluded marker pairs for which we were not able to calculate robust cM/kb ratios. The excluded marker pairs were: i) located within the TR regions identified in the current study; or ii) located in the regions that showed discrepancies between genetic linkage map and physical map in SUN and XU study (2007) (Table 3.4).

We then compared these scaled recombination frequencies between those located around rearranged chromosomal regions (i.e. SI and CR regions) and those located in syntenic regions. We found that

overall recombination frequencies surrounding SI and CR regions were significantly lower than the recombination frequencies in the syntenic chromosomal regions (Mann-Whitney U test, $P<0.05$). We did not find significant difference in recombination frequencies between marker pairs surrounding SI regions and CR regions (Mann-Whitney U test, $P>0.05$).

It has been suggested that the repression of recombination is achieved mainly through two processes. The first is through the function of mismatch repair systems, such as those shown in the budding yeast *S. cerevisiae* (HUNTER et al. 1996; GREIG et al. 2002; LITI et al. 2006). These "proof-reading" mechanisms interfere with the pairing between homologous chromosomes that have low sequence similarities during meiosis and thus suppress crossing over and recombination in these regions. The second mechanism for reduced recombination is through chromosome rearrangements, such as inversions and translocations. These rearrangements make the homologous chromosomes hard to pair with each other during meiosis. Even if crossing over occurs in these regions, recombination could result in progeny with abnormal chromosome structures (e.g. non-centromeric or di-centromeric chromosomes) and such progeny tend to be non-viable or if viable, have low fitness, thus reducing the observable recombinant genotypes and lower "effective" recombination frequency. It should be pointed out that these two processes are not mutually exclusive. Chromosomal rearrangements could facilitate sequence divergence around the rearranged regions, thus indirectly contributing to recombination repression through the mismatch repair system (RIESEBERG, 2001; NOOR et al. 2001; NAVARO and BARTON, 2003a; KIRKPATRICK and BARTON, 2006). Consistent with this hypothesis, rearranged regions have been found associated with speciation events and significantly diverged genomic sequences among many species, including humans, dogs and mice (NAVARO and BARTON, 2003b; ARMENGOL et al. 2005; LINDBLAD -TOH et al. 2005).

Table 3.4. Comparison of recombination frequencies between syntenic and rearranged regions

Marker Pair [a]			Chromosomal Rearrangement	Physical Distance (kb)	Genetic Distance (cM)	Recombination Frequency (cM/kb*100)
CNA00050	—	CNA00670	SI(1)A	175	1.6	0.91494709
CNA00670	—	CNA01490		209	2.6	1.244337665
CNA01490	—	CNA02040		145	1.9	1.30790941
CNA02040	—	CNA03240		302	8	2.644680556
CNA03240	—	CNA04100	SI(1)B//CR(1) **	252	0.9	0.357638868
CNA04100	—	CNA04280		51	2.2	4.312626193
CNA04280	—	CNA05090		198	0.3	0.151638837
CNA05090	—	CNA05600		150	4.2	2.795006255
CNA05600	—	CNA06430		246	7.3	2.965833999
CNA06430	—	CNA06990	SI(1)C	153	2.9	1.89441572
CNA06990	—	CNA07470		145	2.9	1.994964435
CNA07470	—	CNA07790		107	4	3.724481482
CNB00700	—	CNB01310		197	1.3	0.660077737
CNB01310	—	CNB02080		208	1.3	0.624951927
CNB02080	—	CNB02980	CR(2)**	299	2.6	0.869162612
CNB02980	—	CNB03520		146	2.6	1.776302081
CNB03520	—	CNB04740		298	3.2	1.073825503
CNB04740	—	CNB05090		100	1.3	1.305974835
CNB05090	—	CNB05530		105	0	0
CNB05530	—	CNB05710		47	0.3	0.635007991
CND00510	—	CND01190	CR(4)A**	192	0.6	0.312706843
CND01190	—	CND02060		211	5	2.364765001
CND02060	—	CND03160	SI(4)A	299	6.1	2.039888174
CND03160	—	CND03480		91	1.6	1.76001144
CND03480	—	AD030		78	2.2	2.820512821
AD030	—	CND03960		74	2.2	2.959336033
CND03960	—	CND04540		154	2.2	1.427811165
CND04540	—	CND05140		157	4.3	2.746086029
CND05140	—	MAT	CR(4)B	165	6.2	3.755595534
MAT	—	CND06370	SI(4)B//CR(4)B	190	1.9	0.999424016
CNE00250	—	AD014		63	0.3	0.4759525
AD014	—	CNE01270		232	2.2	0.949223039
CNE01270	—	CNE01830		148	0	0
CNE01830	—	CNE02210		100	3.3	3.29772457

Table 3.4 (continued)

Marker Pair [a]			Chromosomal Rearrangement	Physical Distance (kb)	Genetic Distance (cM)	Recombination Frequency (cM/kb*100)
CNE02210	—	CNE03010	CR(5)**	257	2.2	0.856933075
CNE03010	—	CNE03700		195	1.3	0.666921465
CNE03700	—	AD028		128	4.3	3.354343485
AD028	—	CNE04300		21	0.2	0.94308483
CNF00290	—	CNF01350	CR(6)A	300	1.3	0.433134813
CNF01350	—	CNF02070		206	0.3	0.145701089
CNF02070	—	CNF02400		99	1.9	1.910430953
CNF02400	—	AD018	CR(6)B**	268	2.9	1.082481346
AD018	—	CNF03420		38	2.9	7.641835094
CNF03420	—	CNF04830		396	3.3	0.832376941
CNG00170	—	AD026		139	0	0
AD026	—	CNG00900		69	2.6	3.78889998
CNG00900	—	CNG01370		149	0	0
CNG01370	—	CNG01750		96	0	0
CNG01750	—	AD019		88	N.C. [b]	
AD019	—	CNG02290		67	0.3	0.447761194
CNG02290	—	AD020		120	N.C. [b]	
AD020	—	CNG03250	CR(7)**	146	0.6	0.411906854
CNG03250	—	CNG03900		185	2.6	1.402029168
CNG03900	—	AD021		64	0.6	0.933263857
AD021	—	CNG04610		142	0.6	0.421234427
CNH03700	—	CNH02750		300	N.C. [c]	
CNH02750	—	AD024	CR(8)**	465	4	0.859420321
AD024	—	CNH00030	SI(8)	371	1.6	0.43079541
CNI00070	—	AD006		270	5.4	1.996878952
AD006	—	AD005	CR(9)**	273	1.6	0.586725339
AD005	—	CNI02550		129	0.6	0.464446612
CNI02550	—	CNI02950	SI(9)	109	N.C. [b]	
CNI02950	—	CNI03300	SI(9)	96	0	0
CNI03300	—	CNI04370	SI(9)	269	0.6	0.223048327
CNJ00070	—	CNJ00540		131	2.6	1.987076361
CNJ00540	—	CNJ01260		209	0.6	0.286476176
CNJ01260	—	CNJ02080		250	2.6	1.040911839
CNJ02080	—	CNJ02920	CR(10)**	301	1.6	0.532080284
CNJ02920	—	CNJ03090		51	0	0
CNJ03090	—	CNJ03190		44	0	0

Table 3.4 (continued)

Marker Pair [a]			Chromosomal Rearrangement	Physical Distance (kb)	Genetic Distance (cM)	Recombination Frequency (cM/kb*100)
CNL04620	—	CNL04980		108	0.3	0.277777778
CNL04980	—	CNL05760		206	N.C. [b]	
CNL05760	—	AD007		122	0	0
AD007	—	CNL06830		176	N.C. [b]	
AD009	—	CNM00630	CR(13)**	166	0.6	0.361881785
CNM00630	—	CNM01380		211	0.9	0.42753618
CNM01380	—	CNM01960		196	6.2	3.169005086
CNM01960	—	CNM02290		98	0.6	0.614222318
CNM02290	—	CNM02490		53	0	0
CNN00060	—	CNN00590		177	2.2	1.242330725
CNN00590	—	CNN01880	CR(14)A	348	N.C. [c]	
CNN01880	—	CNN02060	CR(14)B**	101	0.9	0.889051333

[a]: Marker pairs on JEC21 chromosomes 3 and 11 were excluded due to the existence of large scale of translocations on these chromosome

[b]: Not calculated due to the inconsistence between the marker orders in the linage groups and their physical locations on the chromosome in SUN and XU (2007).

[c]: Not calculated because the two markers are on separate linkage groups in SUN and XU (2007)

Chromosome rearrangements, especially inversions, are known to repress recombination, restrict gene flow, and play important roles in establishing new lineages/species (RIESEBERG 2001). For example, in three hybrid zones of wild sunflowers (*Helianthus*), the average frequencies of introgression across chromosomes with rearrangements were about 50% lower than those across syntenic chromosomes (RIESEBERG et al. 1999). Similar results have been found in a study of introgression frequency among backcrossed progeny between two sunflower species, *Helianthus annuus* and *H. petiolaris*. However, in this study, the authors found that the percentages of specific genomes that had been introgressed were about 17-fold higher in collinear regions than in rearranged regions (RIESEBERG et al. 1995).

In the SUN and XU (2007) study, they found that recombination occurred at a significantly lower level (overall ~7 fold lower) in an inter-variety hybrid cross than in the intra-variety cross in *C. neoformans* (MARRA et al. 2004). The reduction in recombination frequency during serotypes A and D hybridization could be the result of combined effects of the two aforementioned processes. In this study, we found that when only syntenic chromosomal regions were considered, the average recombination frequency (unit: cM/kb*100) was 1.45, about 5 folds lower than that reported in the intra-variety cross by MARRA et al. (2004). When SI and CR regions were compared to syntenic chromosomal regions, we found that both SI and CR regions showed significantly reduced recombination frequencies in the hybrid cross in *C. neoformans*. The average recombination frequencies in the SI and CR regions were 0.98 and 0.88 respectively, which were about 32% and 39% lower than the average recombination frequency in the syntenic chromosomal regions (1.45), respectively. The differences of recombination frequencies between chromosomal rearrangements and syntenic chromosomal regions observed here were comparable to that reported in RIESEBERG et al. (1999), but significantly lower than that reported in RIESEBERG et al. (1995).

It should be pointed out that in our study, most of the markers used for calculation of recombination frequencies for rearrangements were not located exactly at the boundaries of the rearranged regions. In other words, the recombination frequencies calculated for the rearrangement regions included some syntenic regions, which could have led to overestimates of recombination frequencies for the rearrangement regions. Also, because we used different serotype A strains for the genome comparison (H99) and the genetic linkage map construction (CDC15), it is possible that strain CDC15 may have unique chromosomal rearrangements that differ from both H99 and JEC21, which could also lead to underestimates of recombination frequencies for the syntenic chromosomal regions. For example, in the inter-variety cross between CDC15 and JEC20, ten pairs

of markers showed no recombination but appeared to be located in syntenic chromosomal regions between H99 and JEC21 (Table 3.4, marker pairs with zero genetic distances). It is possible that chromosomal rearrangements might be present within or around these regions in the CDC15 genome that could have suppressed recombination in these regions.

In summary, our results suggested that chromosomal rearrangements likely contributed over 30% reduction of recombination frequency in the hybrid cross between strains of serotypes A and D. In contrast, the majority of the reduction when comparing inter-variety and intra-variety crosses was likely due to other mechanisms such as the mismatch repair systems that contributed to about 5 folds reduction (SUN and Xu 2007; Marra et al. 2004).

Distribution of chromosomal rearrangements in natural serotype A, D and AD strains

We successfully designed PCR primers from which we were able to unambiguously differentiate the two chromosomal types of H99 (serotype A) and JEC21 (serotype D) for a total of 13 rearrangement regions, including nine SIs, three CRs and one TR (Figure 3.2; Supplemental Figure S3.A to S3.L). The 15 CR regions that were not screened were located in JEC21 either within the MAT locus [CR(4)C] or in the proposed centromeric regions (LOFTUS et al. 2005). All of these 15 CR regions contained extensive rearrangements and showed very low levels of sequence similarities between H99 and JEC21 (Figure 3.1). We were unable to find robust primers to unambiguously screen rearrangement polymorphisms among natural strains for these regions. All 13 rearranged regions that we were able to design proper primers for were screened for potential rearrangement polymorphisms in a collection of 64 natural strains of serotypes A, D and AD (Table 3.4).

For one translocation juncture that we screened, TR3/11, all the natural isolates had the same chromosomal type as JEC21, suggesting this translocation was an unique event that happened only in the evolution of the H99 genome (Table 3.4; Figure 3.2).

For the nine SI and three CR regions, we found that all strains belonging to serotype A had the same chromosome types and that they were different from all the serotype D strains that all had the alternate chromosome types (Table 3.4). Our results thus indicated that at these rearrangement regions, the two chromosomal types were likely each fixed within each of the two varieties.

Among the 21 serotype AD strains, we were able to obtain PCR products for all of the nine SI and three CR regions using at least one of the primer pairs that worked for respective serotype A or D strains. Our results indicated that the 21 serotype AD strains showed different levels of heterozygosity at different SI and CR regions (Table 3.4). Two rearranged

regions [SI(3) and CR(6)A] showed 100% heterozygosity (i.e. all the serotype AD strains possessed both types of the chromosome). In addition, we found that no strain was homozygous for the serotype D chromosomal type in region SI(9), and that no strain was homozygous for the serotype A chromosomal type in three regions [SI(5), SI(8) and CR(4)B]. In seven of the 12 SI and CR regions, the majority of the serotype AD strains were heterozygous (more than 16 out of 21 serotype AD strains, i.e. 76%). The five exceptions were all simple inversions, with four of them being located on chromosome 1 [SI(1)A-D, all biased toward the serotype A chromosomal type (Chi-square test, $P<0.05$)] and the other one on chromosome 4 [SI(4), biased toward serotype D chromosomal type (Chi-square test, $P<0.05$)].

For several rearrangement regions, our PCR primers did not work for some of the natural strains, suggesting potential sequence divergence from the JEC21 and H99 genomes, the two genomes used as references for primer designs (Table 3.4). In addition, we also noted PCR fragment size polymorphisms among the natural strains for some of the rearranged junctures, suggesting the existence of small insertion(s) and/or deletion(s) for the amplified regions (Table 3.4, detailed data not shown).

The result that each serotype AD strain was heterozygous for chromosomal types in at least four rearrangement regions was consistent with previous studies that serotype AD strains originated from inter-variety hybridization between serotypes A and D strains (XU et al. 2002). The high levels of heterozygosity observed for most of the rearrangement regions analyzed here for serotype AD strains were also consistent with the low levels of recombination observed around these regions (see Discussions above). However, we did not find any serotype AD strain heterozygous at all the 12 SI and CR regions. For these non-heterozygous rearrangement regions, the strain might contain either two copies of the same chromosome type or could be haploid for that specific chromosome(s) or chromosome region(s). The loss of a chromosome or chromosome segment could result from abnormal chromosomal segregation during the inter-variety hybridization or random loss during subsequent mitotic reproduction. Our data suggest that the losses of chromosomes or chromosomal segments are likely not random. Specifically, for the nine SI regions that had homozygous strains, five had chromosomal types biased toward serotype A (i.e., serotype AD strains are either heterozygous or homozygous for the serotype A chromosomal type, with the exception of strains CBS132) and the other four had chromosomal types biased toward serotype D. Another interesting observation is the very strong linkage disequilibrium among the four SI regions located on chromosome 1. All the chromosomal types for the four SI regions were significantly biased toward the serotype A type (Table 3.3). A propensity of serotype A chromosome 1 in serotype AD strains was

recently reported by Hu et al. using a comparative genome hybridization procedure (HU et al. 2008). The reason for the clear biases in the chromosomal types in serotype AD strains is not known. However, the fact that the four SI regions on chromosome 1 showed strong linkage disequilibrium among serotype AD strains suggested the existence of strong epistatic interactions among loci along chromosome 1.

Copy number changes for chromosomal segments or whole chromosomes, occurring either spontaneously or as a result of hybridization, have been reported previously in many different fungi, such as *Saccharomyces cerevisiae* (HUGHES et al. 2000; KOLODNER et al. 2002; TORRES et al. 2007), *Candida albicans* (ZOLAN 1995; CHEN et al. 2004; RUSTCHENKO 2007) and *C. neoformans* (HU et al. 2008). These large-scale genetic changes may allow the organisms to adapt rapidly to new/changing environments. Indeed, many of these changes have been found to have phenotypic consequences. For example, in *C. albicans*, phenotypic variation in colony morphology, virulence, as well as drug resistance have been found to be associated with certain ploidy changes (ZOLAN 1995; COSTE et al. 2007). Similarly, in *C. neoformans*, serotype AD strains have been shown to be more tolerant to the anti-fungal drug fluconazole (Xu et al. 2001), UV radiation (LITVINTSEVA et al. 2007), and high temperature (LIN et al. 2007; LIN et al. 2008), possibly due to the diploid/aneuploid nature of their genomes.

Unlike plants and animals, serotype AD hybrids are not "evolutionary dead end," despite producing meiotic progeny with low viability. This is because serotype AD strains can reproduce asexually and undergo somatic recombination to generate genetic variants. Such variants have the potential to influence the structure of natural *C. neoformans* populations. As was shown in previous studies, despite their recent origins (Xu et al., 2000), serotype AD strains can be prevalent in certain clinical samples (Xu et al., 2002). It is not clear whether or not the genes located in these chromosomal rearrangements are contributing directly to these medically important traits. However, because these chromosomal rearrangements are repressing recombination during hybridization between serotypes A and D, they contribute to the production of diploid/aneuploid serotype AD strains, which may directly or indirectly influence the medically important traits.

We found that all 12 SI and CR regions had fixed chromosome types in the two varieties for strains across broad geographic areas. Some of the isolates of serotypes A and D were isolated from the same geographic area (e.g. San Francisco, detailed data not shown). These results suggest that gene flow between the two varieties was historically limited and hybridizations between them were relatively recent events, consistent with earlier observations (Xu et al. 2002; 2003) and at least partially explains the significantly reduced viability of progeny produced by

serotypes A and D hybridization. While there are some differences in the geographic ranges of serotypes A and D strains, they do overlap in their current geographic distributions, with both found in many regions of the world. The observed fixation of inversions between the two varieties suggests that they most likely reflect their ancient divergence, possibly due to geographic and/or ecological niche separations. Their current geographic distributions were likely due to recent dispersals by humans and other animals, resulting their hybridization and the generation of serotypes AD hybrids (Xu et al. 2000; 2002).

MATERIALS AND METHODS

C. neoformans genomes used for comparison

For the genome comparison, we used the whole-genome shotgun sequence information of two strains, H99 and JEC21. Strains H99 and JEC21 belonged to *C. neoformans* var. *grubii* (serotype A) and *C. neoformans* var. *neoformans* (serotype D), respectively. The genome sequence of strain H99 was downloaded from Broad Institute website at (http://www.broad.mit.edu/annotation/genome/cryptococcus_neoformans/MultiHome.html) and the nuclear genome is organized in 14 supercontigs. The genome sequence of strain JEC21 was published in 2004 (LOFTUS et al., 2004). The JEC21 genome also contains 14 chromosomes and is 19 Mbp in size.

Types of chromosome rearrangements

In this study, we were mainly interested in large chromosomal rearrangements that could be identified unambiguously. We specifically looked at three types of chromosomal rearrangements.

The first type was translocation (TR). Translocations are identified when there are regions with no significant blast hits between corresponding homologous chromosomes but significant blast hits found between non-homologous chromosomes of the two genomes. The second type of chromosomal rearrangement was simple inversion (SI). As the name implies, a simple inversion is defined as one stretch of a chromosome, flanked by syntenic chromosomal regions but in reverse orientations in the H99 and JEC21 genomes. While it is relatively straightforward to infer these two types of chromosome rearrangements (TR and SI), the third type is complex and may involve multiple inversions and/or small-scale translocations. We call the third type of rearrangements complex rearrangements (CR) (See RESULTS AND DISCUSSION).

Identification of chromosomal rearrangements between JEC21 and H99

To identify the rearrangements, we first did nucleotide blast searches (blastn) using individual chromosomes of JEC21 as separate queries and the complete H99 genome as subject database. This search

identified the most likely corresponding chromosomes between the two genomes. After the one-to-one correspondence of chromosomes was established between the two genomes, we performed all-against-all blast searches (blastn), each time using one chromosome from JEC21 as a query and its corresponding chromosome from H99 as the subject database. Following JIN et al (2005), our blast search results were selected using the following two criteria: (i) that the e-value of the blast hits be lower than e^{-10}; and (ii) that the length of the blast hits be longer than 200bp and the sequence identity be higher than 85%. These two criteria were used to ensure stringency, identify robust blast hits, and reduce background noise. These selected blast results were then used as input files and imported into the GenomeComp program (YANG et al., 2003) to obtain graphic representations of the blast hits, following the program's instructions. In these graphic representations, colored lines connect homologous regions, and the lines connecting syntenic regions of two chromosomes are parallel. Potential chromosomal rearrangements were then identified through visual inspections. Each simple inversion (SI) would be marked by systematic crossings of the color lines connecting homologous sequences between H99 and JEC21, with all lines crossing one point of the graph. In contrast, a chromosomal region with a complex rearrangement (CR) would contain clusters of intersecting color lines in diverse orientations, with many different crossing points among the lines. A translocation (TR) was identified as a stretch of chromosomal region that did not have blast hits between the corresponding chromosomes. Reciprocal blast searches using H99 chromosomes as queries and JEC21 as the subject database were used to confirm the search results.

Identification of transposable elements

We first retrieved sequences of 30 transposable elements from the *C. neoformans* database at TIGR website (http://www.tigr.org/tdb/e2k1/cna1/; Table 3.5). These 30 elements included all the transposable elements identified in *C. neoformans* so far and some have been described before (Goodwin and Poulter, 2001). Each of the 30 elements has a unique DNA sequences. These elements were clustered to nine different Tcn groups (e.g. Tcn1, Tcn2, etc) based on their overall sequence similarities (results not shown). We then used these 30 sequences as queries and compared them against the genome sequences of H99 and JEC21. The blast results were evaluated by two criteria: (i) that the e-value of the blast hits must be lower than e^{-10}; and (ii) that the length of the blast hits must be longer than 30% of the query length. The locations of the transposable element sequences were then mapped onto H99 and JEC21 chromosomes (Figure 3.3).

Table 3.5. List of transposable elements screened in this study [a]

Locus [b]	Description/Category [c]
CNA01670	Transposable elements -Tcn760, putative
CNA03610	Transposable elements -Tcn1, putative
CNA03620	Transposable elements -Tcn3, putative
CNA03630	Transposable elements -Tcn2, putative
CNA03640	Transposable elements -Tcn4, putative
CNA03660	Transposable elements -Tcn6, putative
CNA03670	Transposable elements -Tcn3, putative
CNA03680	Transposable elements -Tcn4, putative
CNA03710	Transposable elements -Tcn7, putative
CNE02930	Transposable elements -Tcn5, putative
CNE02940	Transposable elements -Tcn1, putative
CNE02950	Transposable elements -Tcn4, putative
CNE02960	Transposable elements -Tcn6, putative
CNE02970	Transposable elements -Tcn2, putative
CNE02980	Transposable elements -Tcn6, putative
CNE02990	Transposable elements -Tcn3, putative
CNF03080	Transposable elements -Tcn2, putative
CNF03100	Transposable elements -Tcn4, putative
CNK00490	Transposable elements -Tcn6, putative
CNK00500	Transposable elements -Tcn5, putative
CNK00510	Transposable elements -Tcn3, putative
CNK00520	Transposable elements -Tcn3, putative
CNK00530	Transposable elements -LTR11, putative
CNM00500	Transposable elements -Tcn6, putative
CNM00510	Transposable elements -Tcn3, putative
CNM00520	Transposable elements -Tcn2, putative
CNM00530	Transposable elements -Tcn3, putative
CNM00550	Transposable elements -Tcn3, putative
CNM00560	Transposable elements -Tcn5, putative
CNM00570	Transposable elements -Tcn1, putative

[a]: The list was obtained by "Gene name search" using key words "transposable elements" at website http://www.tigr.org/tdb/e2k1/cna1/.
[b]: Names of loci are the same as those in the annotated JEC21 genome (Genbank AE017341-AE017353; AE017356).
[c]: Descriptions provided by TIGR website.

Correlation between chromosomal rearrangements and recombination frequencies

To study whether recombination frequencies were affected by chromosomal rearrangements, we compared recombination frequencies around the rearranged regions (TR, SI and CR) to those in syntenic regions between the JEC21 and H99 genomes. We took advantage of the hybrid genetic linkage map constructed using serotype A and D hybrid progeny by SUN and XU (2007) and used data presented in that study to calculate the ratio between genetic distance (cM) and physical distance (kb) for each marker pair. The simple ratio of cM/kb is used as an indicator of the amount of recombination occurring over one unit of physical distance (i.e. 1 kb). For each chromosomal rearrangement, we used the genetic distance (in cM) and the physical distance (in kb) between the two markers that were located just outside but most close to the two ends of the rearranged region to calculate the recombination frequency spanning the rearranged region (i.e. cM/kb ratio).

Survey of rearrangement polymorphisms among natural *C. neoformans* strains

To examine whether the identified chromosomal rearrangements were strain-specific or serotype specific, we surveyed the distributions of all non-centromeric chromosomal rearrangements (LOFTUS et al. 2005), including nine SIs, three CRs and one TR, among natural strains. To assay the 13 chromosomal rearrangements, we designed 4 to 12 PCR primers to flank the breaking points for each rearranged region. The alternative chromosomal arrangements were then determined by direct PCR using different primer combinations, with genomic DNA samples from strains H99 and JEC21 serving as positive/negative controls (depending on the primer combination). The details of the primer locations, their sequences, and the expected PCR products for different primer combinations in the 13 rearrangements are provided in the Figure 3.2 and Supplemental Figures S3.A to S3.L. A total of 64 other natural strains of *C. neoformans* were examined. These strains were collected from five countries and consisted of 26 serotype A strains, 17 serotype D strains, and 21 serotype AD strains (Table 3.1). The PCR conditions, gel electrophoresis, staining and scoring followed those in our previous studies (SUN and XU, 2007).

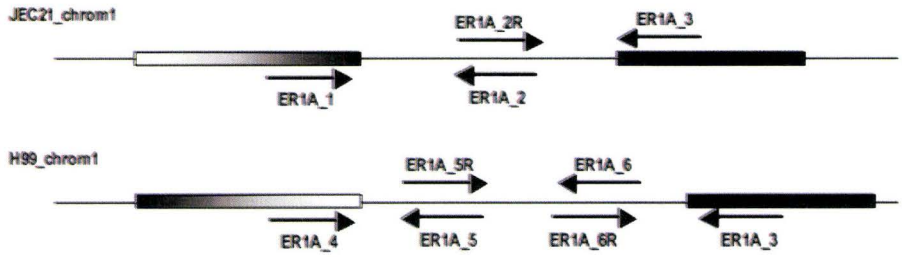
ACKNOWLEDGEMENT

We thank Dr. Jason Stajich for providing the assembled genome sequence of strain H99 and Dr. Tim James for critical comments on the manuscript. Financial support for this work was provided by NSERC and PREA.

Supplemental Figures S3.A to S3.L. Illustrations of the direct PCR strategy used to confirm different chromosomal types at each of 12 rearranged chromosome regions (9 SI regions and 3 CR regions).
For each figure, the drawing on the top shows the locations of primers used for chromosomal type determination in the two corresponding chromosomes of H99 and JEC21. The bars with uniform black or white colors indicate syntenic chromosomal regions flanking the rearranged regions. The bars with uniform blue or green colors (e.g. in Figures S3.K and S3.L) indicate transpositions with the same sequence orientations. The bars with gradient colors indicated inverted chromosomal regions that have sequences in opposite directions in the two genomes. Arrows indicate the locations and directions of the primers in the two genomes. The table in the middle of each figure lists the primer sequences. The table at the bottom presents the expected and confirmed PCR results (positive and negative) of different primer combinations for H99 (serotype A) and JEC21 (serotype D).

Supplemental Figure S3.A

Supplemental Figure A. SI(1)A



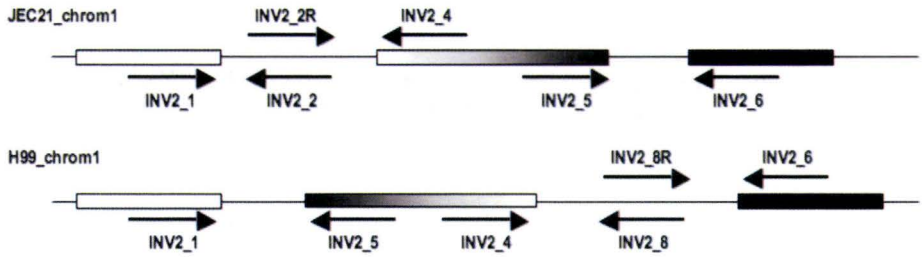
Primer	Position (H99 | JEC21, bp)	Sequence (5' - 3')
ER1A_1	25038 | 24041	TCGTCAAGGCTCAGGGTGAA
ER1A_2	N.A. | 25474	TTCAGCGACGGAGTGCTCAA
ER1A_3	48943 | 27033	CTCGTTCCTCGCAGTAAGCA
ER1A_4	42863 | 5677	CCCCTCCATCGCCGTAAAAA
ER1A_5	44681 | N.A.	TCGCAACCTCATCTGCTTCG
ER1A_6	46646 | N.A.	ATACTCCGCCCGTCGTTTCT
ER1A_2R	N.A. | 25454	TTGAGCACTCCGTCGCTGAA
ER1A_5R	44661 | N.A.	CGAAGCAGATGAGGTTGCGA
ER1A_6R	46626 | N.A.	AGAAACGACGGGCGGAGTAT

Expected and confirmed PCR results with different primer combinations for rearrangement [SI(1)A]

Primer pair	(1) 1/2	(2) 2R/3	(3) 4/5	(4) 5R/6	(5) 6R/3
Serotype A (H99)	-	-	+	+	+
Serotype D (JEC21)	+	+	-	-	-

Supplemental Figure S3.B

Supplemental Figure B. SI(1)B



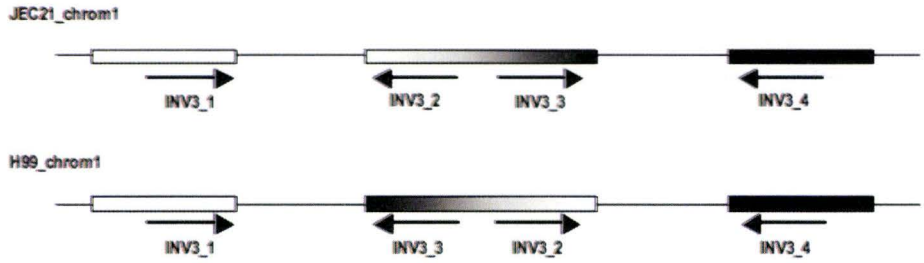
Primer	Position (H99 | JEC21, bp)	Sequence (5' - 3')
SS_NG_INV_2_1	878965 | 852680	CCAGTGAAAAACGCATCGGG
SS_NG_INV_2_2	N.A. | 854562	ACGAGATCATGCGGGAAAGG
SS_NG_INV_2_4	956411 | 856075	TCGCTTTGGCTTTGGAGTTGG
SS_NG_INV_2_5	880257 | 931867	TTTATCAAATCGCGGAACTCGG
SS_NG_INV_2_6	960670 | 932704	CTGGGATCGGGATTTGGACA
SS_NG_INV_2_8	957866 | N.A.	AATGTGCATTGGAGCTGGGG
SS_NG_INV_2_2R	N.A. | 854543	CCTTTCCCGCATGATCTCGT
SS_NG_INV_2_8R	957847 | N.A.	CCCCAGCTCCAATGCACATT

Expected and confirmed PCR results with different primer combinations for rearrangement [SI(1)B]

Primer pair	(1) 1/2	(2) 2R/4	(3) 5/6	(4) 1/5	(5) 4/8	(6) 8R/6
Serotype A (H99)	-	-	-	+	+	+
Serotype D (JEC21)	+	+	+	-	-	-

Supplemental Figure S3.C

Supplemental Figure C. SI(1)C



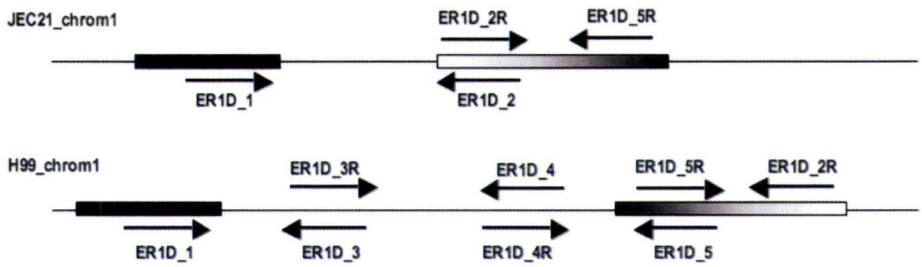
Primer	Position (H99 | JEC21, bp)	Sequence (5' - 3')
SS_NG_INV_3_1	1850883 | 1874018	AACGTCCCCAGTCGCGAAAA
SS_NG_INV_3_2	1857294 | 1875247	TTTGCCTTACGGCCCCATCG
SS_NG_INV_3_3	1852228 | 1880297	CAGCCAATATTCTCCCGGCA
SS_NG_INV_3_4	1858832 | 1881115	AAATAGCGGACTCAGGCCAT

Expected and confirmed PCR results with different primer combinations for rearrangement [SI(1)C]

Primer pair	(1) 1/2	(2) 3/4	(3) 1/3	(4) 2/4
Serotype A (H99)	-	-	+	+
Serotype D (JEC21)	+	+	-	-

Supplemental Figure S3.D

Supplemental Figure D. SI(1)D



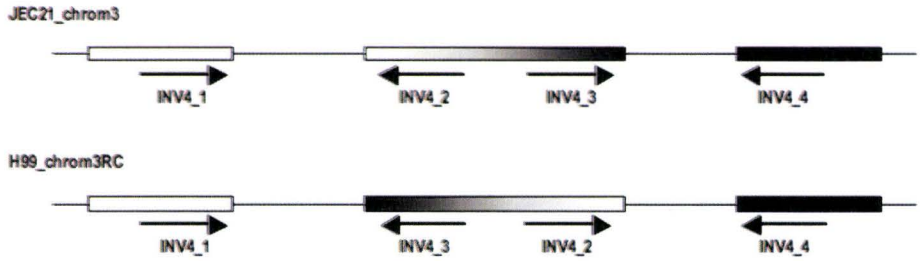
Primer	Position (H99 | JEC21, bp)	Sequence (5' - 3')
ER1D_1	2267418 | 2288647	TCCGACTGCGAGTTGGGATCT
ER1D_2	2272955 | 2290636	GAACTGGGACAAATTTGGCCG
ER1D_3	2268716 | N.A.	TGCGTTTGACATGTGGCAGG
ER1D_4	2270457 | N.A.	TCGAGCAAAGTCGTAAGCGG
ER1D_5	2272100 | 2291956	CAACCGCCGGTTATTAGGAA
ER1D_2R	2272976 | 2290615	CGGCCAAATTTGTCCCAGTTC
ER1D_3R	2268696 | N.A.	CCTGCCACATGTCAAACGCA
ER1D_4R	2270437 | N.A.	CCGCTTACGACTTTGCTCGA
ER1D_5R	2272080 | 2291976	TTCCTAATAACCGGCGGTTG
ER1D_1	2267418 | 2288647	TCCGACTGCGAGTTGGGATCT

Expected and confirmed PCR results with different primer combinations for rearrangement [SI(1)D]

Primer pair	(1) 1/2	(2) 2R/5R	(3) 1/3	(4) 3R/4	(5) 4R/5
Serotype A (H99)	-	+	+	+	+
Serotype D (JEC21)	+	+	-	-	-

Supplemental Figure S3.E

Supplemental Figure E. SI(3)



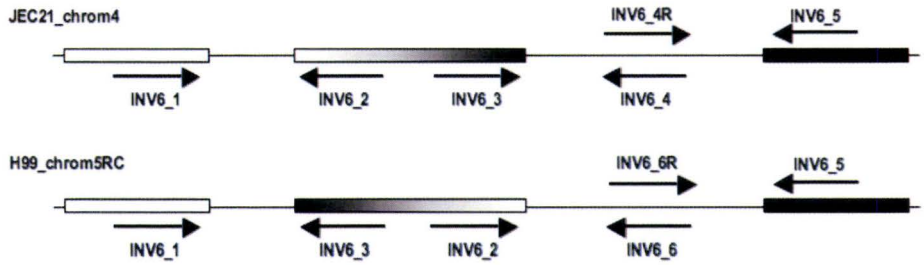
Primer	Position (H99 | JEC21, bp)	Sequence (5' - 3')
SS_NG_INV_4_1	771088 | 1217933	CGTCTCTTGAACTCTGCCCA
SS_NG_INV_4_2	779692 | 1219135	ACAAGGGCCCAGAAAGCCCAGAG
SS_NG_INV_4_3	772626 | 1226260	GTCGACCGAGCCCAAATGCCT
SS_NG_INV_4_4	780783 | 1227308	CATTCATGCAAGGGGCAAGC

Expected and confirmed PCR results with different primer combinations for rearrangement [SI(3)]

Primer pair	(1) 1/2	(2) 3/4	(3) 1/3	(4) 2/4
Serotype A (H99)	-	-	+	+
Serotype D (JEC21)	+	+	-	-

Supplemental Figure S3.F

Supplemental Figure F. SI(4)



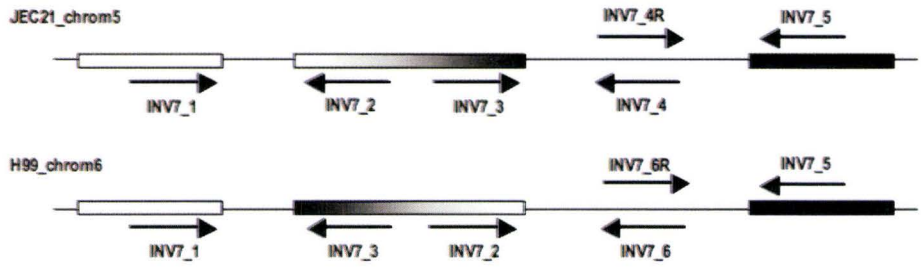
Primer	Position (H99 | JEC21, bp)	Sequence (5' - 3')
SS_NG_INV_6_1	1613690 | 1621729	GGGGCTTTCGGCTGTTTGAA
SS_NG_INV_6_2	1620411 | 1623578	TGGAAGCAGAAGACCTTGCGA
SS_NG_INV_6_3	1614396 | 1629575	CGCAAAGCCCCCGTTATGTA
SS_NG_INV_6_4	N.A. | 1631112	CCACCATTTCACCGTCGAAG
SS_NG_INV_6_5	1623027 | 1633024	ATGATGGATGCCGGCTTCCT
SS_NG_INV_6_6	1621557 | N.A.	TTCAACAGCTTTACTCCGCCG
SS_NG_INV_6_4R	N.A. | 1631093	CTTCGACGGTGAAATGGTGG
SS_NG_INV_6_6R	1621538 | N.A.	TCGGCGGAGTAAAGCTGTTG

Expected and confirmed PCR results with different primer combinations for rearrangement [SI(4)]

Primer pair	(1) 1/2	(2) 3/4	(3) 4R/5	(4) 1/3	(5) 2/6	(6) 6R/5
Serotype A (H99)	-	-	-	+	+	+
Serotype D (JEC21)	+	+	+	-	-	-

Supplemental Figure S3.G

Supplemental Figure G. SI(5)



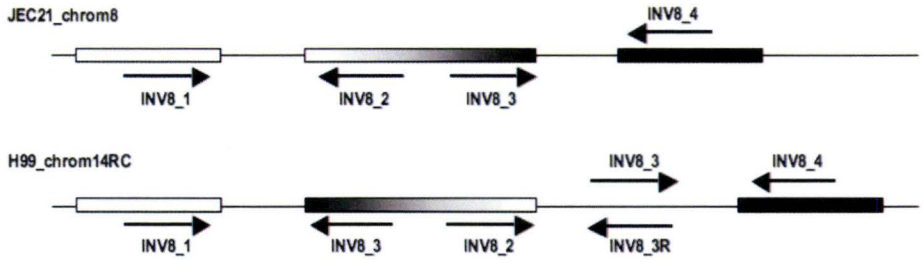
Primer	Position (H99 | JEC21, bp)	Sequence (5' - 3')
SS_NG_INV_7_1	1371283 | 1395141	ACATTGCCTTTCCCGAGCTG
SS_NG_INV_7_2	1378115 | 1397085	TAAAGCCGCGAAAGCGGAG
SS_NG_INV_7_3	1372833 | 1402399	TGGAGCTGGACCGCCTAAATG
SS_NG_INV_7_4	N.A. | 1404097	AATCCGCCCCGTTGATGATG
SS_NG_INV_7_5	1380828 | 1405492	TGCGGTTATTGTCCTGTGCC
SS_NG_INV_7_6	1379548 | N.A.	AGCCCGGTGGTTAGGAAGAA
SS_NG_INV_7_4R	N.A. | 1404078	CATCATCAACGGGGCGGATT
SS_NG_INV_7_6R	1379529 | N.A.	TTCTTCCTAACCACCGGGCT

Expected and confirmed PCR results with different primer combinations for rearrangement [SI(5)]

Primer pair	(1) 1/2	(2) 3/4	(3) 4R/5	(4) 1/3	(5) 2/6	(6) 6R/5
Serotype A (H99)	-	-	-	+	+	+
Serotype D (JEC21)	+	+	+	-	-	-

Supplemental Figure S3.H

Supplemental Figure H. SI(8)



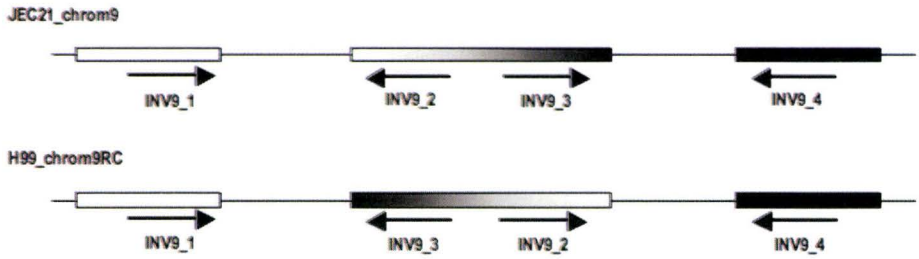
Primer	Position (H99 | JEC21, bp)	Sequence (5' - 3')
SS_NG_INV_8_1	571554 | 846443	GACGGTTGGCGTGGTTTTTTG
SS_NG_INV_8_2	580933 | 847453	TCCTGCGCGTAACGCTTCTTT
SS_NG_INV_8_3	573410/583532 | 854601	AAAAGGGAACGGAGCGAAGCA
SS_NG_INV_8_4	583966 | 855034	CCAAAATCTGCTGTCGCGTG
SS_NG_INV_8_3R	573390/593552 | 854621	TGCTTCGCTCCGTTCCCTTTT

Expected and confirmed PCR results with different primer combinations for rearrangement [SI(8)]

Primer pair	(1) 1/2	(2) 3/4	(3) 1/3	(4) 2/3R
Serotype A (H99)	-	+	+	+
Serotype D (JEC21)	+	+	-	-

Supplemental Figure S3.I

Supplemental Figure I. SI(9)



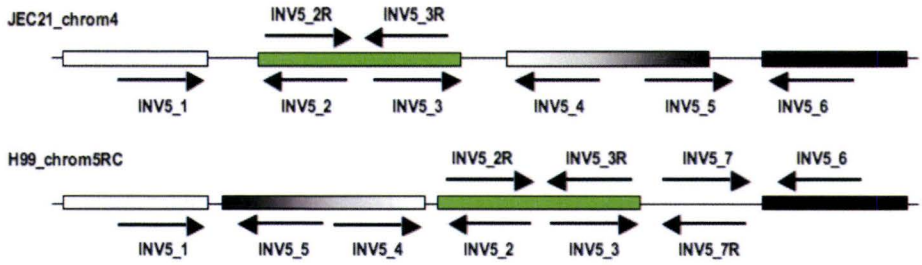
Primer	Position (H99 | JEC21, bp)	Sequence (5' - 3')
SS_NG_INV_9_1	716595 | 721678	TGCCGCAGATGGTCATTGAA
SS_NG_INV_9_2	1103074 | 723202	GCGATGGATAAGAAGGCGCA
SS_NG_INV_9_3	718283 | 1113573	GGAAAGTGCGACGGAGAAGTG
SS_NG_INV_9_4	1105896 | 1114934	TGGCAAGAAAGGTGGCTACCT

Expected and confirmed PCR results with different primer combinations for rearrangement [SI(9)]

Primer pair	(1) 1/2	(2) 3/4	(3) 1/3	(4) 2/4
Serotype A (H99)	-	-	+	+
Serotype D (JEC21)	+	+	-	-

Supplemental Figure S3.J

Supplemental Figure J. CR(4)B



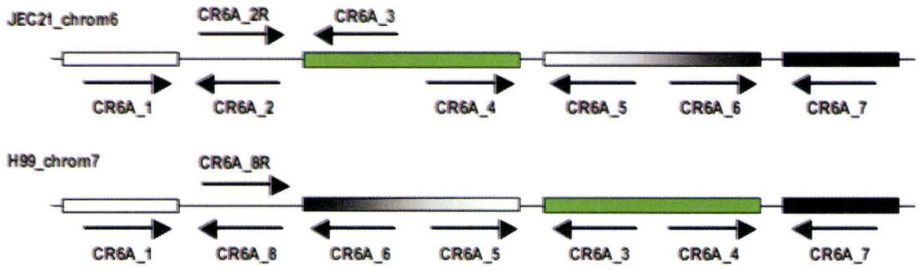
Primer	Position (H99 | JEC21, bp)	Sequence (5' - 3')
SS_NG_INV_5_1	762218 | 768379	GGACCCCTTTTTATGGTTCAA
SS_NG_INV_5_2	772758 | 769869	ATCGTGGTGGCGTGCTTTTCAG
SS_NG_INV_5_3	774197 | 771346	ATCGTTTGGGAGCCAGAGGA
SS_NG_INV_5_4	771245 | 773091	TGAAAGGTGCGAAGATGCGG
SS_NG_INV_5_5	763553 | 780799	TTGTCTGGCGCGAAACTGATG
SS_NG_INV_5_6	778071 | 781995	AGATGCATGACGACCTATTGAG
SS_NG_INV_5_7	776735 | N.A.	TAGTTTTGCAACAGGCAGGGG
SS_NG_INV_5_2R	772737 | 769848	CTGAAAAGCACGCCACCACGAT
SS_NG_INV_5_3R	774216 | 771365	TCCTCTGGCTCCCAAACGAT
SS_NG_INV_5_7R	776755 | N.A.	CCCCTGCCTGTTGCAAAACTA

Expected and confirmed PCR results with different primer combinations for rearrangement [CR(4)B]

Primer pair	(1) 1/2	(2) 2R/3R	(3) 3/4	(4) 5/6	(5) 1/5	(6) 4/2	(7) 3/7R	(8) 7/6
Serotype A (H99)	-	+	-	-	+	+	+	+
Serotype D (JEC21)	+	+	+	+	-	-	-	-

Supplemental Figure S3.K

Supplemental Figure K. CR(6)A



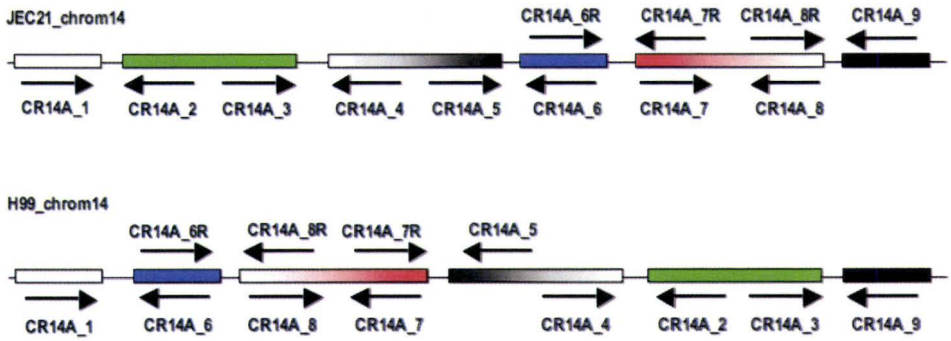
Primer	Position (H99 | JEC21, bp)	Sequence (5' - 3')
CR6A_1	69744 | 75585	AGCACGTCTTACAATGCCCCA
CR6A_2	N.A. | 77563	TCCATTCGCACAGGGCAATC
CR6A_3	95851 | 79715	AACAGGATCCGTTGGGGACGT
CR6A_4	107102 | 93746	AGGGGGATAAGCTGGAGCTTG
CR6A_5	93382 | 95327	GACGGACAATGCTTTCCGGTCC
CR6A_6	73243 | 115526	ATCGAGCGTCGGCGAAAAGT
CR6A_7	108430 | 116843	TCAAGGAGGGCATCCCATGA
CR6A_8	71566 | N.A.	ACTGGTGCCAGCGGTCTTCTT
CR6A_2R	N.A. | 77543	GATTGCCCTGTGCGAATGGA
CR6A_8R	71545 | N.A.	AAGAAGACCGCTGGCACCAGT

Expected and confirmed PCR results with different primer combinations for rearrangement [CR(6)A]

Primer pair	(1)1/2	(2)2R/3	(3)4/5	(4)6/7	(5)1/8	(6)8R/6	(7)3/5	(8)4/7
Serotype A (H99)	-	-	-	-	+	+	+	+
Serotype D (JEC21)	+	+	+	+	-	-	-	-

Supplemental Figure S3.L

Supplemental Figure L. CR(14)A



Primer	Position (H99 | JEC21, bp)	Sequence (5' - 3')
CR14A_1	1518 | 2757	TGAGATCACTTTGCCTCGGGA
CR14A_2	26136 | 5230	TGCTTGGTCGATTCAGTGCG
CR14A_3	68529 | 45043	AAACGCGAAGGAGGACCTGA
CR14A_4	23073 | 45690	TCCGATCACGGCATCATGT
CR14A_5	16203 | 52765	GCACACTTCCAAACCGGAAAG
CR14A_6	11259 | 53539	TGCTCGGAGGCAAGGTAGTGA
CR14A_7	14530 | 55136	ACGGGTTGACGATACTTCGC
CR14A_8	12573 | 57092	GCATTCGCCTTGGACAGTGAA
CR14A_9	70098 | 62842	GGAATCAGATTCAGCGGCGA
CR14A_6R	11239 | 53519	TCACTACCTTGCCTCCGAGCA
CR14A_7R	14511 | 55155	GCGAAGTATCGTCAACCCGT
CR14A_8R	12594 | 57071	TTCACTGTCCAAGGCGAATGC

Expected and confirmed PCR results with different primer combinations for rearrangement [CR(14)A]

Primer pair	(1) 1/2	(2) 3/4	(3) 5/6	(4) 6R/7R	(5) 7/8	(6) 8R/9	(7) 1/6	(8) 6R/8R	(9) 5/7R	(10) 2/4	(11) 3/9
Serotype A (H99)	-	-	-	-	+	-	+	+	+	+	+
Serotype D (JEC21)	+	+	+	+	+	+	-	-	-	-	-

CHAPTER 4

CHROMOSOMAL INVERSIONS AND MOLECULAR EVOLUTION OF CHROMOSOME 1 IN THE HUMAN PATHOGENIC FUNGUS *CRYPTOCOCCUS NEOFORMANS*

PREFACE

In the previous two chapters, we identified many chromosomal rearrangements and showed that the majority are serotype specific and have significant impact on recombination frequency during inter-variety hybridization. However, it is not clear how these rearrangements might be established within each serotype and what effects they might have on the nucleotide polymorphism and structures in natural *C. neoformans* populations. In this chapter, I addressed these questions by sequencing a large number of loci on Chromosome 1 and performing population genetics analyses.

ABSTRACT

Chromosomal rearrangements can facilitate the formation and persistence of divergent lineages (or species) by repressing gene flow and accelerating the accumulation of lineage-specific genetic adaptations. In the human pathogenic fungus *Cryptococcus neoformans*, 9 chromosomal inversions have been identified between the two varieties, var. *grubii* (serotype A) and var. *neoformans* (serotype D), and these inversions have been shown to suppress recombination during inter-variety hybridization. However, the effects of these rearrangements on the patterns of molecular evolution and structures of natural *C. neoformans* populations remain unknown. Here, we examine the effects of four inversions on nucleotide polymorphisms along chromosome 1 in natural serotypes A, D and their hybrid AD populations. A total of 18 loci along chromosome 1 were sequenced, including the junctions of four inversions, as well as coding genes located within and outside these inverted regions. Overall, the loci further away from inverted regions showed significantly lower levels of nucleotide polymorphisms than did those close to inversions. Although no evidence of recombination was detected between serotypes A and D, a signature of recombination was observed within each serotype, with more robust evidence for loci located further away from inversions. The distributions of polymorphisms are consistent with the sequential establishment of inversions, followed by recombination and selective sweeps. The higher haploid diversities in serotype AD than those in serotypes A and D indicate that serotype AD strains might be experiencing faster mutation rates than their haploid serotype A and D counterparts, suggesting that these hybrid strains will continue to impact the structure and dynamics of natural *C. neoformans* populations.

INTRODUCTION

Speciation refers to the process that establishes genetic differences between diverging lineages that can completely prevent genetic introgression and gene flow. This process can be constantly interrupted by inter-lineage hybridization, which acts to eliminate the established divergence. Studies have shown that chromosomal rearrangements may facilitate the speciation process, by repressing recombination and lowering hybrid fitness during the hybridization events. For example, higher rates of gene flow have been reported between collinear chromosomes than between rearranged chromosomes in sunflower hybrid zones (RIESEBERG et al. 1999), between races of *Rhagoletis* fruit flies (FEDER et al. 2003, 2005), and between *Drosophila* species (MACHADO et al, 2002; MACHADO and HEY 2003; HEY and NIELSEN 2004; MACHADO et al. 2007). Chromosomal rearrangements could arise spontaneously and be fixed between lineages through mutation and random genetic drift. However, the establishment of chromosomal rearrangements between lineages can be significantly facilitated if these rearrangements contain or are linked to alternative alleles that confer differences in adaptation, mate discrimination and/or hybrid dysfunction in different lineages (BUTLIN 2005).

Cryptococcus neoformans is a basidiomycetous fungus that can cause meningitis, affecting a significant proportion of immunocompromised people. Based on the difference in their capsular polysaccharides, *C. neoformans* can be classified into three serotypes, serotype A (i.e. var. *grubii*), serotype D (i.e. var. *neoformans*), and serotype AD. *C. neoformans* has a bipolar mating system, with one large mating type locus (MAT) containing two alternative functional alleles, MATa and MATalpha. Within serotypes A and D, mating can occur between cells with opposite mating types, in the laboratory as well as in natural environments (XUE et al. 2007). In addition, mating between cells of the same mating type (MATalpha) can also occur (LIN et al. 2005, 2007; FRASER et al. 2005). In contrast to the haploid serotypes A and D strains, serotype AD strains are diploid/aneuploid and possess alleles typical for both serotypes A and D strains. Population genetic and phylogenetic studies identified that strains of serotype AD are recent hybrids between strains of serotypes A and D and that multiple hybridization events have occurred in natural populations of this species (XU et al. 2000, 2002; LENGELER et al. 2001; XU and Mitchell 2003). Based on the analyses of alleles at the MAT locus, most of the serotype AD strains can be classified as alphaADa (i.e. hybrids between serotype A MATalpha strain and serotype D MATa strain), aADalpha or alphaADalpha (i.e. same sex hybridization). However, some of the serotype AD strains contained only one mating type, MATa or MATalpha, likely having resulted from random loss of the other mating type allele after hybridization.

The genome-wide DNA sequence similarity between the two sequenced serotypes A and D strains H99 and JEC21 is around 85-90% (KAVANAUGH et al. 2006). Assuming a mutation rate of $2X10^{-9}$ per base per year, the time since their divergence is estimated to be about 18 to 40 million years (XU et al. 2000; SHARPTON et al. 2008). Our recent study has shown that, in an inter-variety cross, recombination occurred at a significantly lower level compared to intra-variety matings (~7 fold reduction), suggesting that genetic introgression between the two serotypes is likely to be highly repressed (SUN and XU, 2007). In addition, we (SUN and XU, 2009) further identified a total of 32 rearranged chromosomal regions between the two completely sequenced serotypes A and D genomes. Compared to collinear chromosomal regions, rearranged regions displayed significantly lower recombination frequencies in the inter-variety cross. In addition, our surveys of 66 natural strains of serotypes A, D and AD demonstrated that these chromosomal rearrangements have likely been fixed between the two serotypes (SUN and XU, 2009).

The identified chromosomal rearrangements between serotypes A and D strains have likely contributed to the diploid or anueploid nature of their hybrid serotype AD strains. The diploidy or aneuploidy of serotype AD strains may also impact their patterns of molecular evolution. Specifically, the existence of an extra copy of the sequences could have relaxed the selection pressure acting on the genes in serotype AD strains relative to those of haploid serotypes A or D strains. If true, a greater rate of substitutions might be observed among serotype AD strains than serotypes A and D strains.

The identification of a large number of chromosomal rearrangements fixed between the two serotypes also raises questions about their origins, spread, and fixation. Specifically, it is not clear how these chromosomal rearrangements were established in the two serotypes and how they affected the nucleotide diversity in natural *C. neoformans* populations. The fixation of these chromosomal rearrangements in the two serotypes suggests that they might have been established through selective sweeps. If this were true, the nucleotide polymorphism surrounding the rearranged regions would have been highly reduced. As a result, *C. neoformans* populations will appear to be clonal in regions surrounding inversions, while in regions further away from inversions there might still contain signatures of recombination and gene flow. Consistent with this line of reasoning, most studies of natural *C. neoformans* populations indicated widespread clonality (CURRIE et al. 1994; BRANDT et al. 1996; BOEKHOUT et al. 2001; MEYER et al. 2003; LITVINTSEVA et al. 2005). However, several studies have also shown evidence of recombination in both serotypes A and D *C. neoformans* populations (XU et al. 2000, 2002; XU and MITCHELL 2003; BURT et al. 2000;

LITVINTSEVA et al. 2003, 2005). However, in these studies, molecular markers were predominantly located on different chromosomes, and DNA polymorphisms were not analyzed for loci from the same chromosome nor considered the effect of chromosomal rearrangements on the patterns of nucleotide variations.

The main objective of the current study was to study sequence polymorphisms at loci located at various distances to four rearranged chromosomal regions (all simple inversions) on the largest chromosome, chromosome 1, in a collection of natural serotypes A, D and AD strains. Specifically, we were interested in answering the following questions. First, are there differences in the patterns of sequence polymorphisms among loci distributed at different distances to the rearranged chromosomal regions? Second, do all four loci show similar patterns of sequence divergence? Third, is there any evidence for recombination for loci located on chromosome 1 within each serotype? And last, is there any evidence for gene flow between the serotypes A and D? Answers to these questions will help us better understand the population structure and dynamics of *C. neoformans*, as well as the evolutionary histories of serotypes A, D and AD in *C. neoformans*.

MATERIALS AND METHODS

C. neoformans strains

A total of 26 natural *C. neoformans* strains were used in this study, including eight serotype A strains, eight serotype D strains and ten serotype AD strains (Table 4.1). These strains were selected from a larger collection to represent their broad geographic distributions as well as genetic diversities based on previous studies (BRANDT et al. 1993; XU et al. 2000, 2002; 2003; HIREMATH et al. 2008; SUN and XU 2009). The serotypes of these strains were determined based on the slide agglutination test using monoclonal antibodies. The mating type of each strain was determined by the direct PCR method using mating type specific primers described previously by LENGELER et al. (2001) and YAN et al. (2002). Specifically, four pairs of primers with each specific for one of the four serotype and mating types combinations (i.e. Aa, Aalpha, Da, Dalpha) were designed for Ste20, a gene located within the mating type locus. Based on the presence and absence of PCR products with these primer pairs, the serotype and mating type of each strain were determined (LENGELER et al. 2001; YAN et al. 2002).

Chromosome 1 loci sequenced

C. neoformans chromosome 1 has been shown to contain four serotype specific simple inversions (SUN and XU, 2009). Here we sequenced a total of 18 DNA fragments distributed throughout the chromosome. The specifics about each of the 18 DNA fragments are provided in Table 4.2 and Figure 4.1. Twelve of the 18 DNA fragments

were located within protein-coding genes and six were flanking the inversion junctions of the four simple inversions identified recently by SUN and XU (2009). PCR products were cleaned and prepared for sequencing following procedures provided at the website of Cornell University Life Sciences Core Laboratories Center (http://cores.lifesciences.cornell.edu/brcinfo/), where the sequencing was performed.

Table 4.1. Strains used in this study

Strain	Origin	Serotype[a]	Mating Type[b]
B4962	ZAIRE	A	Aalpha
CDC92_1	SF, USA	A	Aalpha
CDC92_16	GA, USA	A	Aalpha
CDC92_205	AL, USA	A	Aalpha
CDC92_236	TX, USA	A	Aalpha
INN3	INDIA	A	Aalpha-Dalpha
Y195-90	BRAZIL	A	Aalpha
Y288-90	CANADA	A	Dalpha
B3176	ATCC34875	D	Da
CAP67-2	USA	D	Dalpha
CDC92_32	GA, USA	D	Dalpha
CDC92_76	GA, USA	D	Dalpha
CDC92_77	SF, USA	D	Dalpha
CDC92_119	SF, USA	D	Dalpha
CDC92_337	TX, USA	D	Dalpha
Y286-90	CANADA	D	Dalpha
ATCC48184	ATCC	AD	Aa-Dalpha
CBS132	ATCC32719	AD	Aalpha-Da
CDC92-5	SF, USA	AD	Aalpha-Da
CDC92-74	SF, USA	AD	Dalpha
CDC92-181	SF, USA	AD	Aalpha-Da
CDC92-190	SF, USA	AD	Aalpha-Dalpha
CDC92-228	SF, USA	AD	Aa-Dalpha
CDC92-304	SF, USA	AD	Aa-Dalpha
CDC92-383	TX, USA	AD	Da
Y520-91	GA, USA	AD	Aa-Dalpha

[a]: Serotypes were previously determined by traditional agglutination tests.
[b]: Mating types were determined by direct PCR method (See Materials and Method).

Table 4.2. Chromosome 1 loci sequenced in this study

Locus	Location relative to inversion [a]	Description [b]	Location on chromosome 1 (kb) [a]		Distance to closest inversion breaking point (kb) [a]		Primer (5' - 3') [c]
			Serotype A	Serotype D	Serotype A	Serotype D	
SI(1)A_right	Junction	n.a.	49	27	n.a.	n.a.	Forward: CTCGTTCCTCGCAGTAAGCA
							Reverse [A]: AGAAACGACGGGCGGAGTAT
							Reverse [D]: TTGAGCACTCCGTCGCTGAA
SI(1)B_left	Junction	n.a.	879	852.7	n.a.	n.a.	Forward: CCAGTGAAAAACGCATCGGG
							Reverse [A]: TTTATCAAATCGCGGAACTCGG
							Reverse [D]: ACGAGATCATGCGGGAAAGG
SI(1)B_right	Junction	n.a.	960	932.7	n.a.	n.a.	Forward: CTGGGATCGGGATTTGGACA
							Reverse [A]: CCCCAGCTCCAATGCACATT
							Reverse [D]: TTTATCAAATCGCGGAACTCGG
SI(1)C_left	Junction	n.a.	1851	1874	n.a.	n.a.	Forward: AACGTCCCCAGTCGCGAAAA
							Reverse [A]: CAGCCAATATTCTCCCGGCA
							Reverse [D]: TTTGCCTTACGGCCCCATCG
SI(1)C_right	Junction	n.a.	1858	1881.1	n.a.	n.a.	Forward: AAATAGCGGACTCAGGCCAT
							Reverse [A]: TTTGCCTTACGGCCCCATCG
							Reverse [D]: CAGCCAATATTCTCCCGGCA
SI(1)D_left	Junction	n.a.	2267	2288.6	n.a.	n.a.	Forward: TCCGACTGCGAGTTGGGATCT
							Reverse [A]: TGCGTTTGACATGTGGCAGG
							Reverse [D]: GAACTGGGACAAATTTGGCCG

(Table 4.2 continued)

Locus	Location relative to inversion [a]	Description [b]	Location on chromosome 1 (kb) [a]		Distance to closest inversion breaking point (kb) [a]		Primer (5' - 3') [c]
			Serotype A	Serotype D	Serotype A	Serotype D	
CNA01890	Outside	Conserved expressed protein	543.8	509	335.2	347.1	Forward: ATCACGCAACCCGTTGCCAT
							Reverse: TGCCTTGGCGTTGGCTCTTT
CNA03050	Outside	Ubiquitin-protein ligase, putative	819.8	795.9	59.2	60.2	Forward: CCCCCATTGTCGCCAAACAT
							Reverse: AAGCAATTCCCCACGGCGAT
CNA03240	Outside]	Hypothetical protein	876.2	850	2.8	6.1	Forward: AAGTGCGGCGCGATGAATGA
							Reverse: TATGGGAAAGGCGGGGATGT
CNA03280	Inside SI(1)B	Transcriptional activator gcn5, putative	955.1	857.4	4.9	1.3	Forward: CTGCCGGGAAAACGACCATT
							Reverse: TTGGGTGTGCTTTTCGGACG
CNA04280	Outside	t-complex protein 1, beta subunit (tcp-1-beta), putative	1162.1	1152.7	155.1	154.6	Forward: CCGCAAAATCATCTCCGCTG
							Reverse: AAAGGGGCGAAAACGCTCGT
CNA05600	Outside	Catalase A, putative	1484.8	1500.8	366.2	374.5	Forward: TGGGTTTGGCTTCAGGCAG
							Reverse: CCGCTCTTCACATGCAGCAA
CNA06430	Outside	Cytosolic small ribosomal subunit protein, putative	1727.4	1746.9	123.6	128.3	Forward: CCAACCGAAGCCCAAGACAA
							Reverse: TTGAAGGATGATCCGGCCGA
CNA06890	Inside SI(1)C	Cell wall organization and biogenesis-related protein, putative	1852.7	1879.8	1.7	0.5	Forward: CGTGCCCTTCCCCAAACAAT
							Reverse: TCATTGTGCGAGATGCCTGC

(Table 4.2 continued)

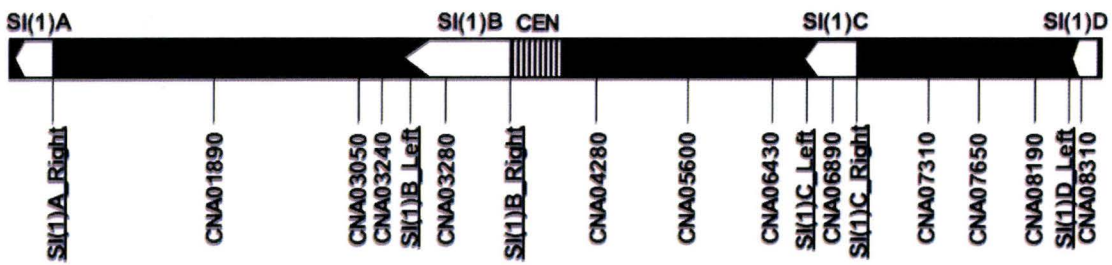
Locus	Location relative to inversion [a]	Description [b]	Location on chromosome 1 (kb) [a]		Distance to closest inversion breaking point (kb) [a]		Primer (5' - 3') [c]
			Serotype A	Serotype D	Serotype A	Serotype D	
CNA07310	Outside	Conserved hypothetical protein	1973.7	2000.9	115.7	120.6	Forward: ACACCCCAAATTCCCCAACC
							Reverse: ACATCCCAAACGAACCCAC
CNA07650	Outside	NADH dehydrogenase, putative	2073.9	2099.5	193.1	191.1	Forward: AGAAGAGCATGGCAGCGGAA
							Reverse: ATGTCGTCGTCGTTTCGCCC
CNA08190	Outside	Conserved hypothetical protein	2230.6	2251.8	36.4	38.8	Forward: TCGCTGGCGTTCACACTCAT
							Reverse: TCACAAAAGGTGGCGGGAGA
CNA08310	Outside	Ligase, putative	2262.3	2286.7	4.7	7.4	Forward: GCGTTTTCACACGACCCGTT
							Reverse: CGGGTGAGTTCCATTTGGTTG

[a]: Locations were based on the published genome sequence of strain JEC21 (LOFTUS et al. 2005), the fully assembled genome sequence H99 (http://www.broad.mit.edu/annotation/genome/cryptococcus_neoformans/MultiHome.html), and Sun and Xu (2009).

[b]: The description was retrieved from GenBank (http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?TAXID=5207&CHR=1&MAPS=cntg-r,rna,genes[1.00%3A2300533.00]&CMD=TXT#1).

[c]: Each locus was sequenced using the forward primer. For each simple inversion junctions, alleles from serotypes A and D backgrounds were amplified using primer combinations "Forward/Reverse [A]" and "Forward/Reverse [D]", respectively.

Figure 4.1. Distributions of the chromosome 1 loci analyzed in this study SI(1)A to D are the four simple inversion described in SUN and XU (2009). CEN is the centromeric region proposed by LOFTUS et al. (2005). The loci underlined are inversion junctions and the loci that are not underlined are coding genes.


Sl(1)A
Sl(1)B
CEN
Sl(1)C
Sl(1)D
Sl(1)A_Right
CNA01890
CNA03050
CNA03240
Sl(1)B_Left
CNA03280
Sl(1)B_Right
CNA04280
CNA05600
CNA06430
Sl(1)C_Left
CNA06890
Sl(1)C_Right
CNA07310
CNA07650
CNA08190
Sl(1)D_Left
CNA08310

Data analyses

Sequences were aligned with the program ClustalX version 2 (LARKIN et al. 2007) and further checked and edited with the programs Finch TV version 1.4.0 (Geospiza, Inc.) and MacClade version 4.06 (Sinauer Associates, Inc.). For each protein-coding gene, the cDNA sequence from the published genome of *C. neoformans* strain JEC21 was retrieved from GenBank and used as a guide for alignment. The analyses of nucleotide polymorphisms were performed using the program DnaSP version 4.90 (ROZAS et al. 2003) with indels excluded. For all the polymorphism estimates, the significance of departure from neutral model was determined with 5000 coalescent simulations without recombination, as implemented in the program DnaSP. Two specific statistics were also performed using commands implemented in DnaSP. In the first, neutrality of the mutations at each locus was estimated using Tajima's D (TAJIMA 1989), and the existence of natural selection at each coding gene was estimated by McDonald-Kreitman test (McDONALD and KREITMAN 1991). In the second, the average intra-locus linkage disequilibrium between serotypes A and D was estimated using Z_{nS} (KELLY 1997).

For population genetic analyses, two complementary tests were conducted to examine evidence for clonality and recombination in the populations. In the first test, we examined evidence for genealogical incongruence using the partition homogeneity test (PHT) or incongruence length difference (ILD) test (FARRIS et al.1995a, 1995b) implemented in PAUP*4.0b10 (SWOFFORD, 2003). The null hypothesis of PHT is that genealogies among different genes are congruent. The significance of the test was determined by comparing the observed statistic to the null distribution that was generated by random partitioning the original dataset. PHT was intended to detect the presence of strongly supported character conflicts among individual data sets (FARRIS et al. 1995a, 1995b). It has been shown that relying on this test alone for inferences of recombination could produce false positive results (i.e. congruence) when the consistency indices for individual gene trees are not high (BARKER and LUTZONI 2002). To avoid these potential problems, we also conducted another set of common population genetic tests, the index of association (I_A, MAYNARD SMITH et al., 1993) and RbarD (AGAPOW and BURT, 2001) tests using only the polymorphic sites. I_A test examines the associations among alleles from different loci, and RbarD is a modified measure of I_A that was adjusted for the number of loci. Both I_A and RbarD were estimated using the program Multilocus version 1.3b (AGAPOW and BURT, 2001) following the program instructions. The null hypothesis of I_A and RbarD test is that the population is recombining, which is opposite to the PHT. The null hypothesis is rejected if the observed allelic association is significantly higher than the majority (95%) of the null distribution that was generated by randomly recombining the alleles. We applied I_A and

RbarD tests to the whole dataset that includes all loci, as well as to all possible pair-wise combinations of loci. We analyzed serotypes A and D together, as well as separately.

RESULTS

Mating types of the strains used in this study

Using the direct PCR amplification strategy, we determined the mating type allele(s) for each of the strains analyzed in this study. Of the eight serotype A strains, six were identified as serotype A MATalpha (i.e. Aalpha; Table 4.1). The other two strains were different. Strain INN3 had both serotypes A and D MATalpha alleles amplified at the MAT locus (i.e. alphaADalpha; Table 4.1). However, at all the coding genes and inversion junctions analyzed in this study, one unambiguous allele was obtained for each of the genes and their sequences were more similar to those from other serotype A strains than to serotype D strains (see results below). The other exception was strain Y288-90. It was amplified by the serotype D-specific MATalpha primers (i.e. Dalpha; Table 4.1). However, similar to strain INN3, all sequenced nuclear loci indicated strain Y288-90 was more similar to other serotype A strains. It is possible that both strains INN3 and Y288-90 were of hybrid origin between strains of serotypes A and D through either same-sex mating or opposite-sex mating but have since undergone haploidization and maintained chromosome 1 from only the serotype A parent. As a result, these two strains will remain in the serotype A group in subsequent analyses.

Of the eight serotype D strains, all but one were identified as MATalpha (i.e. Dalpha). The only exception was strain B3176, which was identified as MATa (i.e. Da; Table 4.1). Sequences at all 18 marker loci also placed them in the serotype D group.

Of the ten serotype AD strains, we identified four aADalpha strains (i.e. each of them has two alleles at the MAT locus, with one allele from serotype A MATa background and the other allele from serotype D MATalpha background), three alphaADa strains, and one strain each for alphaADalpha, Dalpha and Da (Table 4.1).

DNA polymorphism in chromosome 1

The summary statistics for polymorphisms for the 18 sequenced fragments on chromosome 1 are presented in Tables 4.3 and 4.4. For the 12 protein-coding genes, the weighted average values of nucleotide diversity per site (estimated using the number of segregating mutations), θ, and the average number of pairwise nucleotide differences, π, were 0.006 ± 0.0043 (Mean ± SD) and 0.0058 ± 0.0056 for serotype A strains; 0.0052 ± 0.0046 and 0.0059 ± 0.0052 for serotype D strains, and 0.0233 ± 0.0057 and 0.0383 ± 0.0086 when serotypes A and D strains were analyzed together (Table 4.3). The average pairwise difference per base pair between serotypes A and D alleles ranged between 0.0365 and 0.0898

(Mean: 0.0676; SD: 0.0152; Table 4.3). For the six simple inversion junctions, the mean θ and π were 0.0036 and 0.0042 for serotype A strains, and 0.0060 and 0.0050 for serotype D strains (Table 4.4).

Distributions of θ in protein coding genes and in inversion junctions are presented in Figures 4.2 and 4.3. Among the six inversion junctions that we analyzed, θ was lower or similar to that of the closest protein-coding gene that we analyzed for each of the junctions within both the serotypes A and D populations. There is however one noticeable variation between the two junctions surrounding SI(1)B. Specifically, the SI(1)B_left junction had a polymorphism about 10 times higher than those of the SI(1)B_right junction and its closest coding gene CNA03240 in both the serotypes A and D populations (Figure 4.3).

The observed polymorphism was not distributed evenly along chromosome 1. As shown in Figure 4.2, among the coding genes, the polymorphism level (θ) was significantly lower for loci closer to chromosomal inversions [especially SI(1)B and SI(1)C] than those located far apart from these regions in both the serotypes A and D populations. Indeed, there is a significant negative correlation between the level of nucleotide polymorphism and the physical distance between loci and their closest inversion junction for the serotype A population ($P < 0.001$), and marginally insignificant for the serotype D population ($P = 0.052$) (Figure 4.4). In addition, for both the serotypes A and D samples, the gene with the lowest polymorphism was CNA03280, located within an inverted region, SI(1)B (Table 4.3).

Table 4.3. Polymorphism statistics for the analyzed chromosome 1 protein-coding genes

Locus	Serotype(s) [a]	n [b]	Length (in bp) [c]	S (singleton/informative) [d]	H [e]	Hd [f]	Theta [g]	Pi [h]	Tajima's D [i]	Z_{ns} [j]	MK Test [k]	Diff [l]
CNA01890	A	10	915	9 (8/1)	3	0.511	0.00349	0.00216	-1.68719**			
	D	9	915	7 (2/5)	6	0.833	0.00281	0.00298	0.25402			
	A and D	19	915	72 (2/70)	9 **	0.836 **	0.02322	0.04119	2.84249	0.8003**	n.s.	0.07226
CNA03050	A	10	759	12 (5/7)	9	0.978	0.00559	0.00518	-0.35215			
	D	8	759	12 (4/8)	6	0.929	0.0061	0.00639	0.21018			
	A and D	18	759	70 (1/69)	15	0.98	0.02681	0.04628	2.65785	0.6719**	n.s.	0.07889
CNA03240	A	10	726	6 (3/3)	5	0.8	0.00295	0.00285	-0.14916			
	D	9	726	3 (0/3)	6	0.833	0.00153	0.00193	0.98666			
	A and D	19	726	55 (0/55)	11	0.912	0.02204	0.04069	3.10061	0.8256**	n.s.	0.07146
CNA03280	A	10	756	3 (3/0)	2	0.2	0.00141	0.0008	-1.56222			
	D	9	756	3 (0/3)	3	0.722	0.00146	0.00184	0.98666			
	A and D	19	756	56 (0/56)	5**	0.731**	0.02128	0.04021	3.26657	0.9055**	**	0.07161
CNA04280	A	10	627	7 (5/2)	5	0.756	0.00395	0.00292	-1.11638			
	D	9	627	4 (2/2)	4	0.75	0.00235	0.00249	0.23146			
	A and D	19	627	47 (0/47)	9	0.883	0.02145	0.04016	3.18861	0.8339**	n.s.	0.07032
CNA05600	A	10	669	34 (3/31)	10	1	0.01796	0.02153	0.84036			
	D	9	669	29 (6/23)	9	1	0.01595	0.0186	0.72854			
	A and D	19	669	82 (2/80)	19	1	0.03507	0.05209	1.44219	0.3365	**	0.07638
CNA06430	A	10	618	7 (4/3)	7	0.867	0.00404	0.00306	-1.0403			
	D	9	618	3 (1/2)	4	0.694	0.0018	0.00191	0.21755			
	A and D	19	618	50 (1/49)	11	0.901	0.02338	0.04277	2.87036	0.8068**	n.s.	0.07495

(Table 4.3 continued)

Locus	Serotype(s) [a]	n [b]	Length (in bp) [c]	S (singleton/informative) [d]	H [e]	Hd [f]	Theta [g]	Pi [h]	Tajima's D [i]	Z_{nS} [j]	MK Test [k]	Diff [l]
CNA06890	A	10	492	8 (6/2)	5	0.756	0.00575	0.00432	-1.09355			
	D	9	492	5 (0/5)	5	0.806	0.00374	0.00465	1.00871			
	A and D	19	492	23 (0/23)	10	0.895	0.01338	0.02184	2.06459	0.5172**	n.s.	0.03654
CNA07310	A	10	681	11 (2/9)	8	0.956	0.00571	0.0067	0.74277			
	D	7	681	3 (2/1)	4	0.81	0.0018	0.00168	-0.30187			
	A and D	17	681	37 (0/37)	12	0.956	0.01607	0.02593	2.17594	0.5869**	n.s.	0.0442
CNA07650	A	9	618	17 (4/13)	9	1	0.01012	0.01179	0.75283			
	D	8	618	14 (2/12)	5	0.857	0.00874	0.01009	0.73474			
	A and D	17	618	46 (0/46)	14	0.971	0.02202	0.03427	1.97597	0.4025	n.s.	0.05293
CNA08190	A	9	723	11 (6/5)	8	0.972	0.0056	0.00572	-0.32636			
	D	9	723	11 (3/8)	9	1	0.0056	0.00657	0.78373			
	A and D	18	723	60 (1/59)	17	0.993	0.02413	0.04285	2.65574	0.6912**	**	0.0718
CNA08310	A	10	501	7 (6/1)	5	0.667	0.00494	0.00334	-1.38265			
	D	9	501	15 (3/12)	9	1	0.01102	0.01254	0.61478			
	A and D	19	501	53 (2/51)	14	0.912	0.03027	0.05386	2.28473	0.6496**	n.s.	0.08978

n.a., not available; n.s., not significant; **, significant at P<0.05

[a]: Data was separated into different serotypes, as well as grouped together, for the analyses.

[b]: Number of strains sequenced.

[c]: Length (base pairs) of the sequences used for the analyses.

[d]: Number of polymorphic sites. Within the parentheses are the numbers of singletons and phylogenetic informative polymorphic sites.

[e]: Number of haplotype (gene).

[f]: Estimate of Haplotype (gene) Diversity (NEI 1987). Significance was determined using 5000 coalescent simulations with no recombination that was implemented in program DnaSP 4.90 (ROZAS et al. 2003).

[g]: Estimate of 4Nu/bp using the number of polymorphic sites (WATTERSON 1975).
[h]: Estimate of 4Nu using the average number of nucleotide differences per site (TAJIMA 1983).
[i]: Tajima's statistic (1989). Significance was determined using 5000 coalescent simulations with no recombination that was implemented in program DnaSP 4.90 (ROZAS et al. 2003).
[j]: Average intralocus linkage disequilibrium (KELLY 1997). Significance was determined using 5000 coalescent simulations with no recombination that was implemented in program DnaSP 4.90 (ROZAS et al. 2003).
[k]: McDonald and Kreitman test of neutrality hypothesis (McDONALD and KREITMAN 1991). Significance was determined using two-tailed Fisher's exact test that was implemented in program DnaSP 4.90 (ROZAS et al. 2003).
[l]: Average pairwise difference per base pair between alleles from serotypes A and D strains.

Table 4.4. Polymorphism statistics for the analyzed chromosome 1 inversion junctions

Locus	Allele Group [a]	n [b]	Length (in bp) [c]	S (singleton/informative) [d]	H [e]	Hd [f]	Theta [g]	Pi [h]	Tajima's D [i]	Z_{nS} [j]
SI(1)A_Right	TypeA_all	17	982	23 (0/23)	5**	0.743	0.00748	0.01004	1.29211	0.7312**
	SeroA_Aalpha	6	982	22 (0/22)	2**	0.533**	0.0106	0.01312	1.36568	1**
	SeroAD_Aa	4	982	1 (0/1)	2	0.667	0.0006	0.00073	1.63299	n.a.
	SeroAD_Aalpha	2	982	0	1	0	0	0	n.a.	n.a.
	TypeD_all	16	982	37 (3/34)	7	0.792**	0.01209	0.0153	0.88	0.4054
	SeroD_Dalpha	7	982	23 (10/13)	3**	0.524**	0.01018	0.00994	-0.20469	0.6196
	SeroAD_Da	4	982	34 (34/0)	2**	0.5**	0.01992	0.01872	-0.86319**	1
	SeroAD_Dalpha	3	982	26 (26/0)	2	0.667	0.01862	0.01897	n.a.	1
SI(1)B_Left	TypeA_all	9	972	25 (3/22)	5	0.722**	0.01614	0.0182	0.52375	0.5145
	SeroA_Aalpha	1	977	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
	SeroAD_Aa	4	977	0	1	0	0	0	n.a.	n.a.
	SeroAD_Aalpha	3	977	18 (18/0)	3	1	0.02102	0.02145	n.a.	0.9167
	TypeD_all	16	977	41 (8/33)	9	0.925	0.0144	0.01636	0.2896	0.1969
	SeroD_Dalpha	7	977	34 (30/4)	5**	0.857	0.01599	0.01224	-1.40378	0.3916
	SeroAD_Da	4	977	23 (23/0)	3	0.833	0.01442	0.01343	-0.85708	0.8524
	SeroAD_Dalpha	3	977	16 (16/0)	2	0.667	0.01233	0.01249	n.a.	1
SI(1)B_Right	TypeA_all	13	861	4 (0/4)	3	0.41	0.00159	0.00178	0.38711	0.803
	SeroA_Aalpha	5	861	0	1	0	0	0	n.a.	n.a.
	SeroAD_Aa	2	861	0	1	0	0	0	n.a.	n.a.
	SeroAD_Aalpha	2	861	0	1	0	0	0	n.a.	n.a.
	TypeD_all	15	861	10 (0/10)	7	0.857	0.00391	0.00357	-0.34089	0.3669
	SeroD_Dalpha	6	861	2 (2/0)	2	0.333	0.00111	0.00085	-1.13197	1
	SeroAD_Da	4	861	7 (7/0)	3	0.833	0.00485	0.00447	-0.81734	0.746
	SeroAD_Dalpha	3	861	8 (8/0)	3	1	0.00678	0.00682	n.a.	0.8125

(Table 4.4 continued)

Locus	Allele Group [a]	n [b]	Length (in bp) [c]	S (singleton/informative) [d]	H [e]	Hd [f]	Theta [g]	Pi [h]	Tajima's D [i]	Z_{nS} [j]
SI(1)C_Left	TypeA_all	12	1172	30 (25/5)	10	0.97	0.01217	0.00807	-1.64608**	0.3617
	SeroA_Aalpha	5	1172	10 (6/4)	5	1	0.00515	0.00517	0	0.3017
	SeroAD_Aa	0	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
	SeroAD_Aalpha	2	1171	3 (3/0)	2	1	0.00322	0.00322	n.a.	1
	TypeD_all	15	1172	4 (0/4)	5	0.829	0.00134	0.00148	0.31845	0.1241
	SeroD_Dalpha	7	1172	3 (3/0)	4	0.714**	0.00132	0.00092	-1.35841	0.0278
	SeroAD_Da	4	1172	2 (1/1)	3	0.833	0.00118	0.00126	0.59158	0.3333
	SeroAD_Dalpha	2	1172	0	1	0	0	0	n.a.	n.a.
SI(1)C_Right	TypeA_all	19	915	159 (17/142)	9**	0.871**	0.06074	0.1017	1.69371	0.7704**
	SeroA_Aalpha	6	915	3 (1/2)	4	0.867	0.00148	0.00166	0.60031	0.2667
	SeroAD_Aa	4	915	0	1	0	0	0	n.a.	n.a.
	SeroAD_Aalpha	3	915	4 (4/0)	3	1	0.00301	0.00302	n.a.	0.625
	TypeD_all	13	915	4 (2/2)	4	0.679	0.00183	0.00197	0.25198	0.224
	SeroD_Dalpha	6	915	4 (4/0)	2	0.333**	0.00235	0.0018	-1.29503	1
	SeroAD_Da	3	915	0	1	0	0	0	n.a.	n.a.
	SeroAD_Dalpha	2	915	1 (1/0)	2	1	0.00134	0	n.a.	n.a.
SI(1)D_Left	TypeA_all	16	964	12 (3/9)	10	0.933	0.00412	0.00468	0.49324	0.2596
	SeroA_Aalpha	6	964	2 (1/1)	3	0.733	0.001	0.00099	-0.05002	0.1
	SeroAD_Aa	4	964	1 (1/0)	2	0.5	0.00062	0.00057	-0.61237	n.a.
	SeroAD_Aalpha	3	964	3 (3/0)	3	1	0.00227	0.00228	n.a.	0.5
	TypeD_all	15	964	16 (5/11)	10	0.943**	0.00545	0.00555	-0.18374	0.1296
	SeroD_Dalpha	7	964	11 (8/3)	5	0.905	0.00497	0.00412	-0.93421	0.2673
	SeroAD_Da	4	964	5 (5/0)	2	0.5	0.00299	0.00275	-0.79684	1
	SeroAD_Dalpha	2	964	1 (1/0)	2	1	0.0011	0.0011	n.a.	n.a.

n.a., not available; n.p., not present; **, significant at P<0.05

[a]: For the analyses, alleles were separated into different groups based on the serotype and mating type backgrounds from which the alleles were amplified (Table 4.1). Specifically,
TypeA_all: all alleles from serotype A strains,
SeroA_Aalpha: alleles from serotype A MATalpha strains,
SeroAD_Aa: alleles from serotype AD strains that were identified as MATa,
SeroAD_Aalpha: alleles from serotype AD strains that were identified as MATalpha,
TypeD_all: all alleles from serotype D strains,
SeroD_Dalpha: alleles from serotype D MATalpha strains,
SeroAD_Da: alleles from serotype AD strains that were identified as MATa, and
SeroAD_Dalpha: alleles from serotype AD strains that were identified as MATalpha.

[b]: Number of strains sequenced.

[c]: Length (base pairs) of the sequences used for the analyses.

[d]: Number of polymorphic sites. Within the parentheses are the numbers of singletons and phylogenetic informative polymorphic sites.

[e]: Number of haplotype (gene).

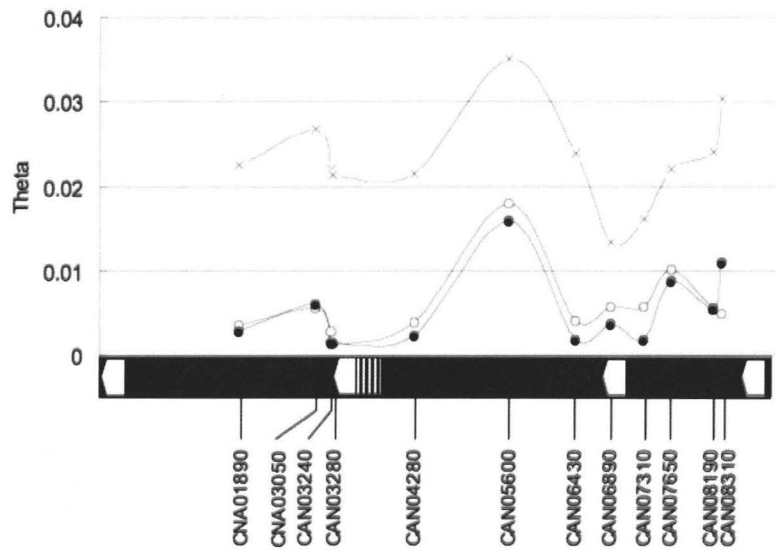
[g]: Estimate of 4Nu/bp using the number of polymorphic sites (WATTERSON 1975).

[h]: Estimate of 4Nu using the average number of nucleotide differences per site (TAJIMA 1983).

[i]: Tajima's statistic (1989). Significance was determined using 5000 coalescent simulations with no recombination that was implemented in program DnaSP 4.90 (ROZAS et al. 2003).

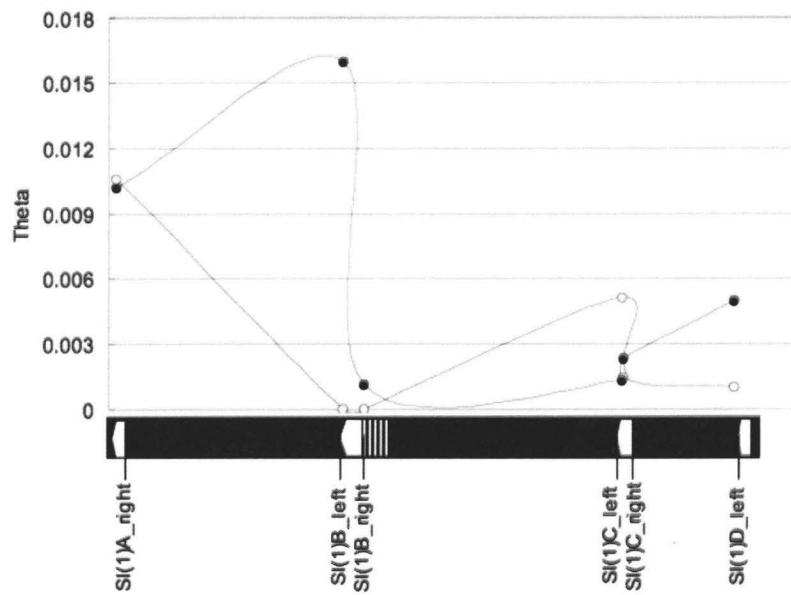
[j]: Average intralocus linkage disequilibrium (Kelly 1997). Significance was determined using 5000 coalescent simulations with no recombination that was implemented in program DnaSP 4.90 (ROZAS et al. 2003).

Figure 4.2. Distribution of Theta (θ) of the protein coding genes
The white dots are those of serotype A strains; the black dots are those of serotype D strains; and the crosses are those when serotypes A and D strains were analyzed together.



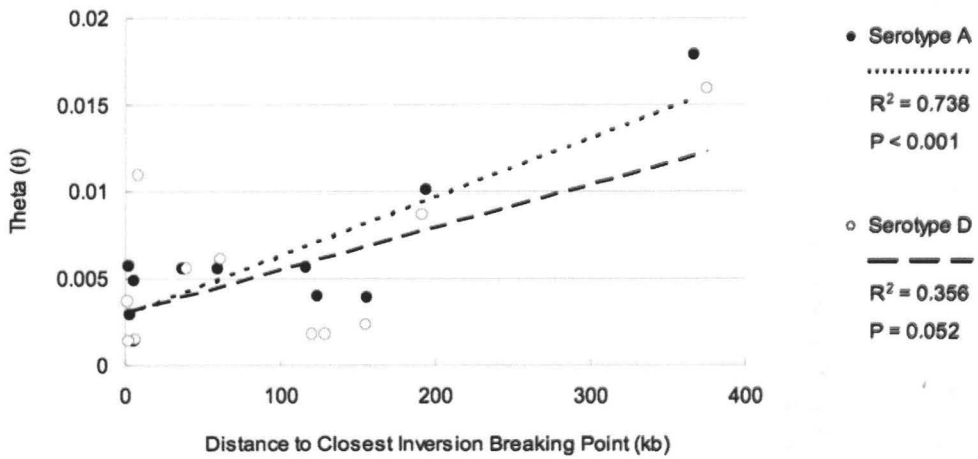
Coding genes Theta

Figure 4.3. Distribution of Theta (θ) of the simple inversion junctions
The white dots are those of serotype A strains; the black dots are those of serotype D strains; and the crosses are those when serotypes A and D strains were analyzed together


0.018
0.015
0.012
0.009
0.006
0.003
0
Theta
Sl(1)A_right
Sl(1)B_left
Sl(1)B_right
Sl(1)C_left
Sl(1)C_right
Sl(1)D_left

Sl Theta

Figure 4.4. Regression of nucleotide polymorphisms on the physical distances between loci and their respective closest located inversion breaking point
Black dots and dotted line are serotype A and open dots and dashed line are serotype D. Loci CNA01890 was excluded due to its statistically significant Tajima's D value (see results).


0.02
0.015
0.01
0.005
0
Theta (θ)
0
100
200
300
400
Distance to Closest Inversion Breaking Point (kb)
Serotype A
R² = 0.738
P < 0.001
Serotype D
R² = 0.356
P = 0.052

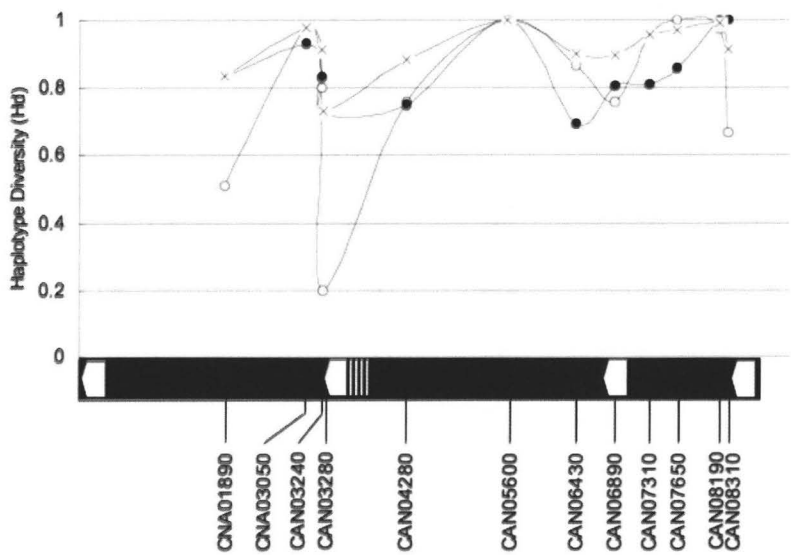
Haplotype diversity at coding genes and inversion junctions

Haplotype diversity (Hd), also called heterozygosity, is a measure of nucleotide polymorphism. Different from θ, Hd takes into account of the frequencies of each allele in the population and thus is less likely to be biased by the existence of excessive number of rare alleles, which is expected for populations undergoing frequent selective sweeps or rapid expansions. We calculated Hd for all the protein coding genes and the inversion junctions (Tables 4.3 and 4.4). The distributions of haplotype diversity values are shown in Figure 4.5 and 4.6 for the protein coding genes and inversion junctions, respectively.

The analyses showed that none of the protein coding genes had Hd values significantly lower than the neutral expectations calculated based on 5000 coalescent simulations ($P>0.05$). One gene, CNA03280, located within simple inversion SI(1)B, had a haplotype diversity three times higher in the serotype D samples than in the serotype A samples (Table 4.2 and Figure 4.5).

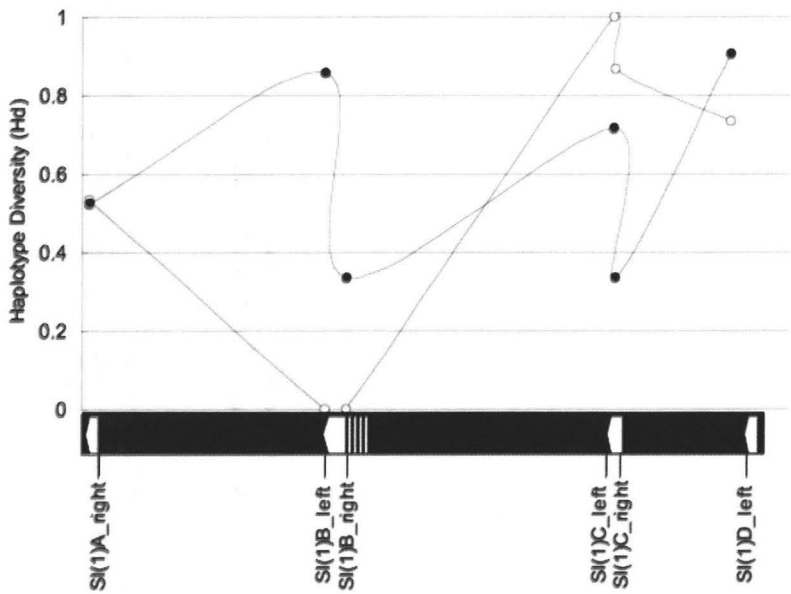
For the inversion junctions, when alleles from only haploid strains were analyzed, we found significantly lower than expected Hd values (based on 5000 coalescent simulations) in both the serotypes A and D chromosomal types at locus SI(1)A_right. Similarly, for SI(1)B, we found significantly lower than expected Hd at both junctions for the serotype A sample. In contrast, the serotype D population had Hd values not significantly different from neutral expectations at both loci (Table 4.4 and Figure 4.6). For the two junctions around inversion SI(1)C, we did not observe Hd levels significantly different from neutral expectations for the serotype A populations. However, we did find marginally lower level of Hd ($P=0.06$) among the serotype D strains. When serotype AD strains were included in the analyses, the two junctions showed slightly different results: the SI(1)C_left junction had a significantly lower level of Hd than expected for the serotype A chromosomal type. In contrast, both the serotypes A and D chromosomal types had Hd values similar to those expected under the neutral model for the SI(1)D_left junction.

Figure 4.5. Distribution of Haplotype Diversity (Hd) of the protein coding genes
The white dots are those of serotype A strains and the black dots are those of serotype D strains.



Coding genes Hd

Figure 4.6. Distribution of Haplotype Diversity (Hd) of the simple inversion junctions
The white dots are those of serotype A strains and the black dots are those of serotype D strains.


Haplotype Diversity (Hd)
1
0.8
0.6
0.4
0.2
0
Sl(1)A_right
Sl(1)B_left
Sl(1)B_right
Sl(1)C_left
Sl(1)C_right
Sl(1)D_left

Sl Hd

Evidence of purifying selection

We found that all the protein-coding genes around the simple inversions had very low Ka/Ks ratios, with most of them lower than 0.5. In addition, the ratios were lower in serotype A populations than in the serotype D populations (Figure 4.7). These results are consistent with the hypotheses that genes close to simple inversions are under strong purifying selection pressure, with the serotype A population experiencing stronger pressure than the serotype D population around these regions. The low level of polymorphisms around the inversion regions is also consistent with the hypothesis of purifying selection around these regions.

Although the Ka/Ks ratio was much less than 1 for the majority of the analyzed protein-coding genes (Figure 4.7), this ratio was close to or higher than 1 for two genes, CNA04280 and CNA05600. Specifically, the gene CNA04280 codes for the beta subunit of the t-complex protein 1 (tcp-1-beta) and it had Ka/Ks ratios of 1.76 and 0.84 respectively for serotypes A and D populations. The second of the two genes CNA05600 codes for catalase A and had Ka/Ks ratios of 1.42 and 1.00 for serotypes A and D populations respectively. The high Ka/Ks ratios for these two genes suggest they are under strong positive selection, a result supported also by the McDonald and Kreitman test (Table 4.3).

Linkage disequilibrium between and within serotypes A and D

To investigate the relationships among alleles for genes located on Chromosome 1, we first examined the linkage disequilibrium that included all the strains and allelic data together. Three statistics, Z_{nS} test, PHT and I_A/RbarD tests, were used to detect LD at three different levels: all genes combined (PHT and I_A/RbarD tests), single gene separately (Z_{nS} test and I_A/RbarD tests; among polymorphic nucleotide sites within each gene), and all possible pair-wise gene combinations (PHT and I_A/RbarD tests) (see Materials and Methods). When all the genes were analyzed together, the entire sample was found to be in significant LD by both PHT and I_A/RbarD tests (detailed data not shown). When single protein coding genes were considered separately, Z_{nS} test suggested 10 of the 12 genes were in significant LD (except CNA05600 and CNA07650; Table 4.3); I_A/RbarD tests suggested each of the genes were in significant LD ($P < 0.01$; detailed data not shown). When all possible pair-wise gene combinations were analyzed, PHT failed to reject the null hypothesis of clonal population in the majority of gene combinations (Figure 4.8) while I_A/RbarD tests suggested all the gene combinations are in LD ($P < 0.01$; detailed data not shown). Overall, our results are consistent with clonal divergence between the two serotypes and that mixing of the alleles from the two serotypes had not occurred at an appreciable frequency to obscure their divergence.

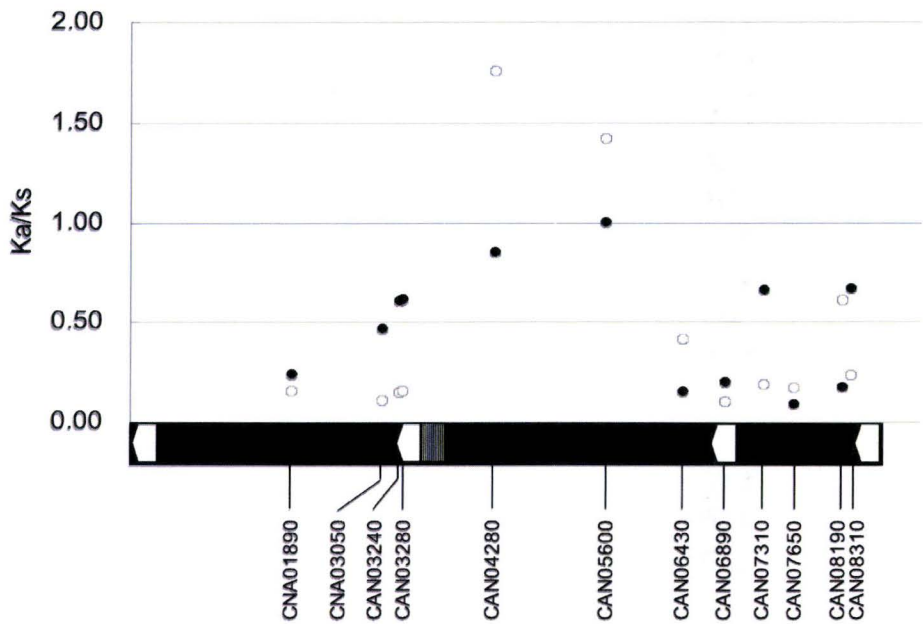
When the LD within serotype A or D were analyzed separately, the I_A/RbarD tests showed that within serotype A, most protein-coding genes

were in significant LD. The only exception was locus CNA06430 (Figure 4.9). However, within serotype D, I_A/RbarD tests failed to reject the null hypothesis of linkage equilibrium (LE) in the following four of the 12 protein-coding genes: CNA03240, CNA03280, CNA06430 and CNA07310 (Figure 4.9).

When all the pair-wise combinations of protein coding genes were analyzed, the PHT test showed more gene combinations in LD (clonal) ($P > 0.05$) than the I_A/RbarD tests ($P < 0.05$) for both the serotypes A and D (Figures 4.8 and 4.9) samples. However, it should be pointed out (as mentioned in Materials and Methods), the PHT is intended to detect "hard incongruence" between data sets that are strongly supported (BARKER and LUTZONI 2002). Because the polymorphic levels are overall low within serotypes, branches separating different allelic sequences are statistically not as robust as those separating the two serotypes.

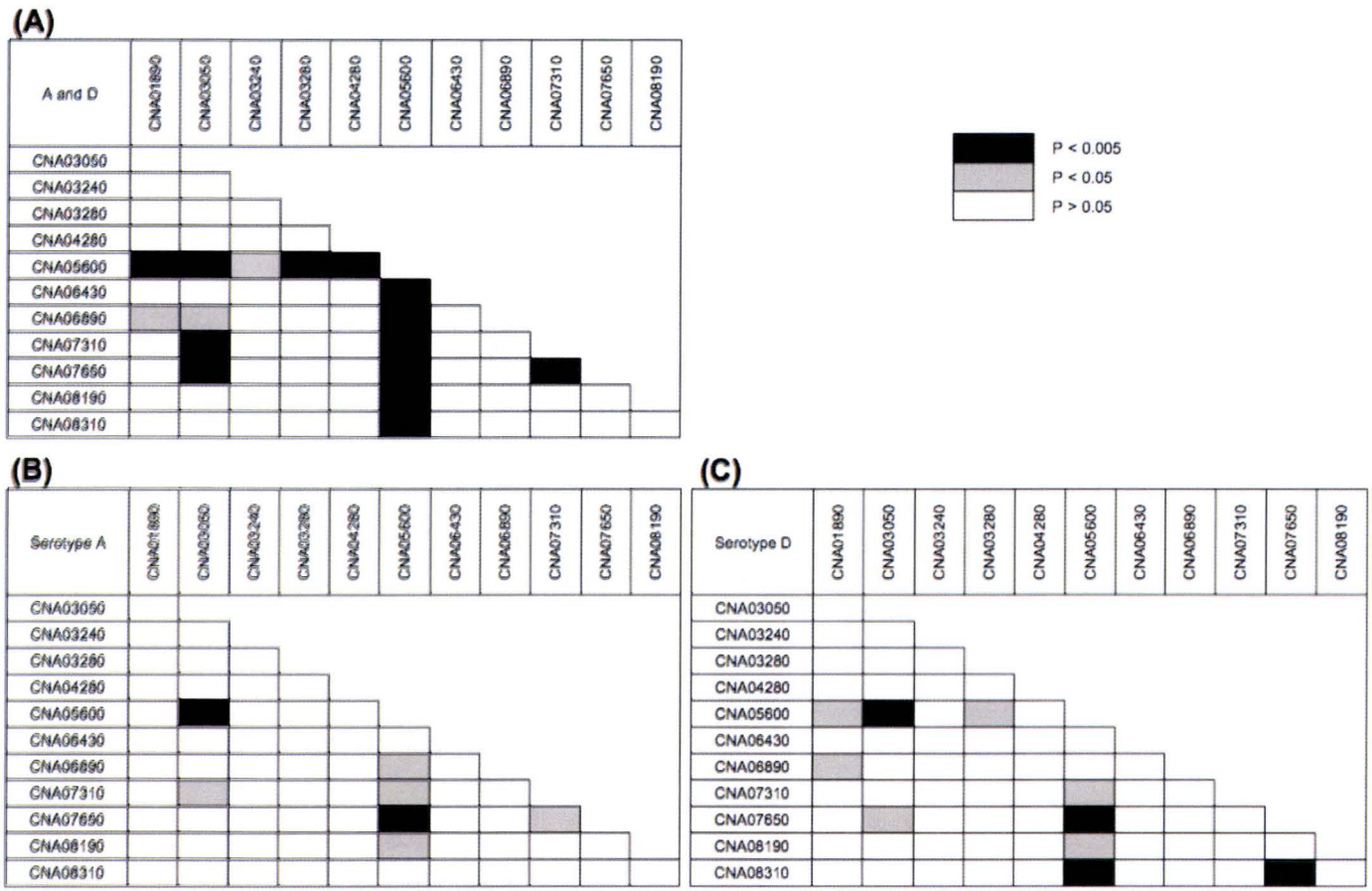
Using the I_A/RbarD tests, we found that more pairwise gene combinations were in LD in serotype D ($n = 26$) than in serotype A ($n = 20$) (Figure 4.9, $P < 0.05$). Although many of the pairwise comparisons showed inconsistent results between the two serotypes (27, or 41%, of the total 66 gene combinations showed different results between serotypes A and D; Figure 4.9), several genes showed relatively consistent results in the I_A/RbarD tests between populations of the two serotypes. For example, gene CNA07650 showed no LD with any of the other 11 genes in both serotypes A and D; CNA07310 showed no LD with other genes in serotype D, but showed LD with only one gene (CNA05600) in serotype A; and CNA05600 showed LD with one gene (CNA07310) in serotype A and two genes (CNA04280 and CNA08310) in serotype D (Figure 4.9). Interestingly, all the three genes (CNA05600, CNA07310 and CNA07650) were located far away from the inversion regions (>100kb, Table 4.2).

Figure 4.7. Distribution of Ka/Ks ratios of the protein coding genes
The white dots are those of serotype A strains and the black dots are those of serotype D strains


2.00
1.50
1.00
0.50
0.00
Ka/Ks
CNA01890
CNA03050
CAN03240
CAN03280
CAN04280
CAN05600
CAN06430
CAN06890
CAN07310
CAN07650
CAN08190
CAN08310

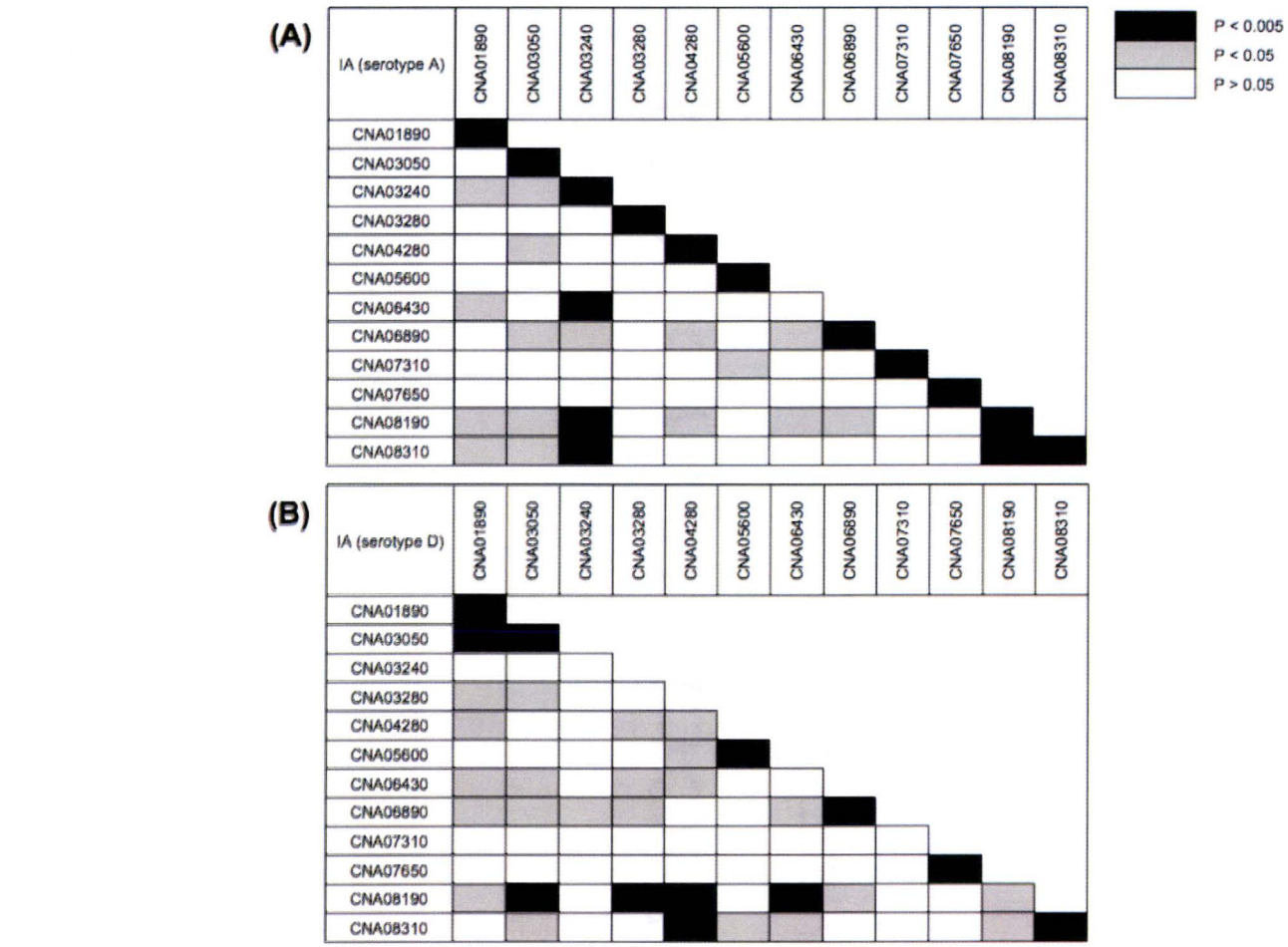
Ka/Ks ratio

Figure 4.8. Summary of PHT for all the pair-wise gene combinations
Figures 4.8A to 4.8C summarized PHT results when 1) serotypes A and D were analyzed together; 2) serotype A was analyzed alone; and 3) serotype D was analyzed alone, respectively. Each square represents PHT result of one gene combination. Black indicated $P < 0.005$; gray indicated $P < 0.05$; and white indicated $P > 0.05$.



PH-test_Pairwise_Results

Figure 4.9. Summary of IA test for all the pair-wise gene combinations Figure 4.9A and 4.9B summarized IA test results for serotypes A and D, respectively. The far right square of each row represents the intra-locus result for that gene. Each of the rest of the squares represents IA test result of one gene combination. Black indicated $P < 0.005$; gray indicated $P < 0.05$; and white indicated $P > 0.05$.



Multilocus_Pairwise_Results

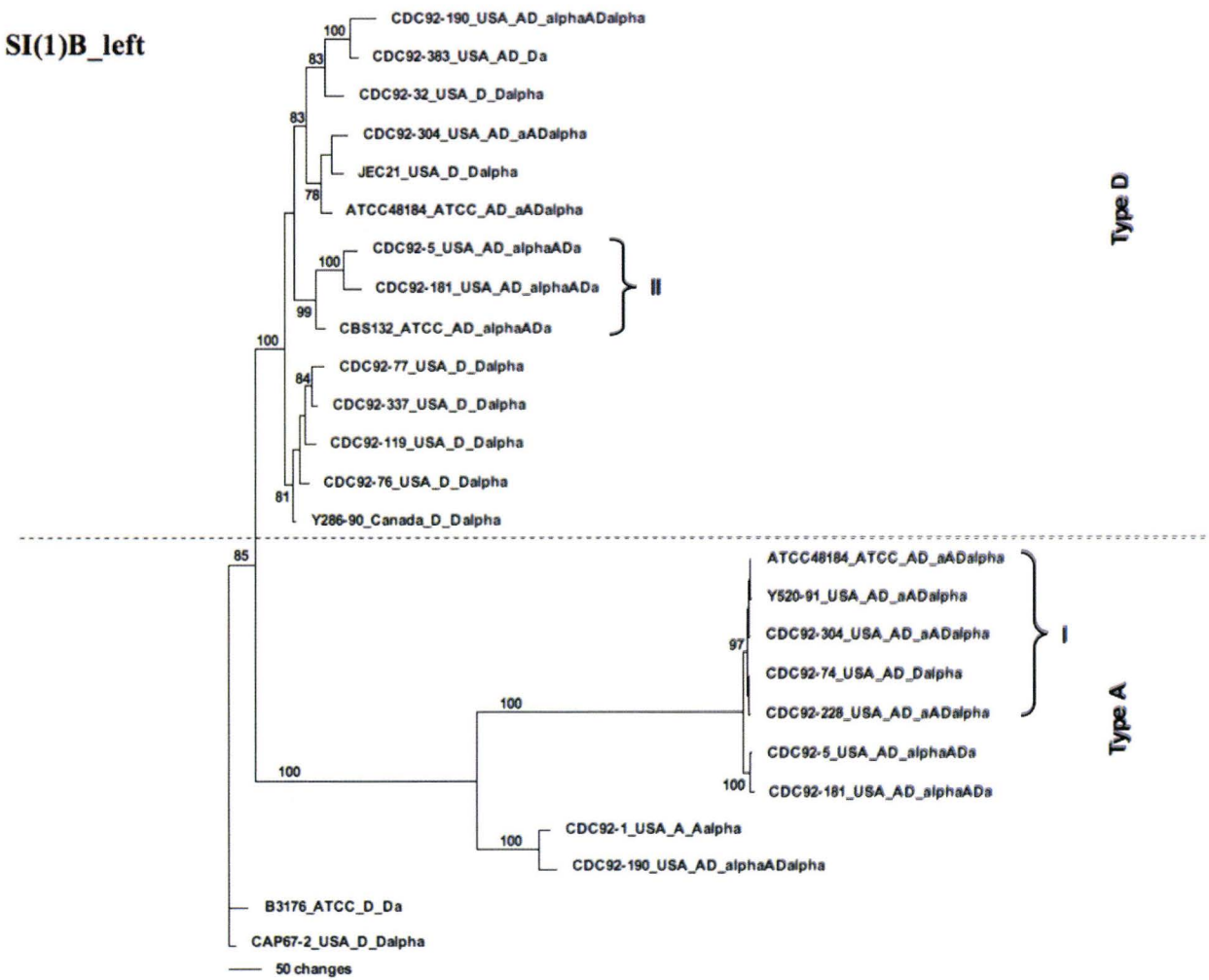
Differentiation between MATa and MATalpha alleles in serotype AD strains at the inversion junctions

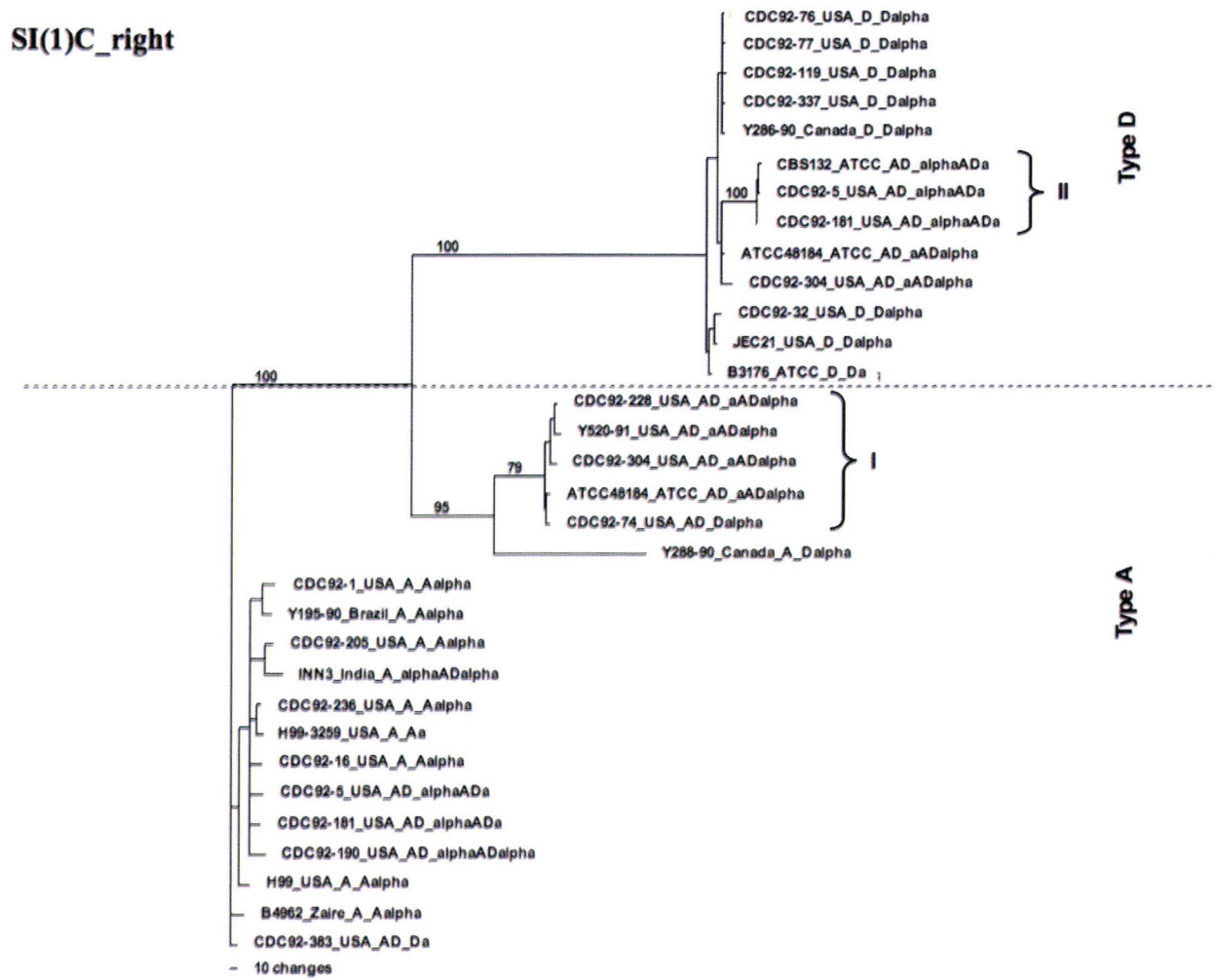
For five of the six inversion junctions that we analyzed, subsets of alleles from serotype AD strains tended to cluster together. Figure 4.10 showed two such clusters, I and II, that corresponded to the serotypes A and D allelic groups respectively. As can be seen from Table 4.1 and Figure 4.10, these two clusters seemed to be correlated to the presence of the Aa and Da serotype-mating type strain background respectively. The distinctiveness of these two clusters was supported by high bootstrap values. Specifically, cluster I contained the unique group of serotype A alleles from the aADalpha strains while cluster II contained the unique group of serotype D alleles mainly from the alphaADa strains. These clustering patterns indicate that alleles from MATa-containing strains contain unique polymorphisms not found in either Aalpha or Dalpha haploid strains.

High levels of polymorphism among serotype-specific alleles from serotype AD strains

We compared the amount of sequence variation and haplotype diversity among the serotype-specific alleles from serotype AD strains with those from haploid serotype A or D strains. A significantly greater diversity was observed for alleles from the serotype AD strains than those from serotypes A or D strains. This pattern holds true even when adjusted to the same mating type background. For example, for the SI(1)A_right junction, the haplotype diversity of alleles from serotype D MATalpha haploid strains (i.e. group "SeroD_Dalpha") was significantly lower than the serotype D alleles from serotype AD strains (i.e. group "SeroAD_Dalpha"). Among the five junctions that we were able to analyze (i.e. each with $n \geq 3$; Table 4.3), four had lower levels of Hd for alleles from haploid strains than those from comparable groups of alleles in the serotype AD strains. The difference was statistically significant (Wilcoxon Signed-Rank Test, $P<0.05$).

Figure 4.10. Maximum Parsimony trees of loci SI(1)B_left and SI(1)C_right
Figures 4.10A and 4.10B showed the maximum parsimony trees of inversion junctions SI(1)B_left and SI(1)C_right, respectively. Branches with >75% bootstrap vales (based on 1000 replications) were labeled. Dashed line separated alleles from serotypes A and D backgrounds. The format of the sequence tag is as
"(Strain Name)_(Origin)_(Serotype)_(Mating Type)"


SI(1)B_left
CDC92-190_USA_AD_alphaADalpha
CDC92-383_USA_AD_Da
CDC92-32_USA_D_Dalpha
CDC92-304_USA_AD_aADalpha
JEC21_USA_D_Dalpha
ATCC48184_ATCC_AD_aADalpha
CDC92-5_USA_AD_alphaADa
CDC92-181_USA_AD_alphaADa
CBS132_ATCC_AD_alphaADa
II
Type D
CDC92-77_USA_D_Dalpha
CDC92-337_USA_D_Dalpha
CDC92-119_USA_D_Dalpha
CDC92-76_USA_D_Dalpha
Y286-90_Canada_D_Dalpha
ATCC48184_ATCC_AD_aADalpha
Y520-91_USA_AD_aADalpha
CDC92-304_USA_AD_aADalpha
CDC92-74_USA_AD_Dalpha
CDC92-228_USA_AD_aADalpha
I
Type A
CDC92-5_USA_AD_alphaADa
CDC92-181_USA_AD_alphaADa
CDC92-1_USA_A_Aalpha
CDC92-190_USA_AD_alphaADalpha
B3176_ATCC_D_Da
CAP67-2_USA_D_Dalpha
50 changes


SI(1)C_right
CDC92-76_USA_D_Dalpha
CDC92-77_USA_D_Dalpha
CDC92-119_USA_D_Dalpha
CDC92-337_USA_D_Dalpha
Y286-90_Canada_D_Dalpha
CBS132_ATCC_AD_alphaADa
CDC92-5_USA_AD_alphaADa
CDC92-181_USA_AD_alphaADa
ATCC48184_ATCC_AD_aADalpha
CDC92-304_USA_AD_aADalpha
CDC92-32_USA_D_Dalpha
JEC21_USA_D_Dalpha
B3176_ATCC_D_Da
CDC92-228_USA_AD_aADalpha
Y520-91_USA_AD_aADalpha
CDC92-304_USA_AD_aADalpha
ATCC48184_ATCC_AD_aADalpha
CDC92-74_USA_AD_Dalpha
Y288-90_Canada_A_Dalpha
CDC92-1_USA_A_Aalpha
Y195-90_Brazil_A_Aalpha
CDC92-205_USA_A_Aalpha
INN3_India_A_alphaADalpha
CDC92-236_USA_A_Aalpha
H99-3259_USA_A_Aa
CDC92-16_USA_A_Aalpha
CDC92-5_USA_AD_alphaADa
CDC92-181_USA_AD_alphaADa
CDC92-190_USA_AD_alphaADalpha
H99_USA_A_Aalpha
B4962_Zaire_A_Aalpha
CDC92-383_USA_AD_Da
10 changes
100
100
100
79
95
II
I
Type D
Type A

DISCUSSION

Our results showed that sequence variation (θ) was not distributed uniformly along chromosome 1 in either serotype A or serotype D populations (Figures 4.2, 4.3 and 4.4). Within each serotype, the loci located closer to inverted regions showed lower levels of polymorphism (Figure 4.4). In addition, we observed low frequencies of LD in gene combinations involving loci located far away from inversion regions (Figure 4.9; see Results). The low sequence polymorphism and high frequency of LD observed in genes surrounding the inverted regions are consistent with a scenario in which each of the four simple inversion arose in one serotype, and the two chromosomal types were later fixed for each of the two serotypes. Furthermore, their fixation likely occurred through a series of selective sweeps, with each re-enforcing the accumulated genetic differentiation and limiting genetic introgression between the two serotypes. Consistent with this hypothesis, negative Tajima's D values were observed at ten of the 12 coding genes (except CNA05600 and CNA07650), as well as at inversion junctions in both serotype A and/or serotype D populations (Table 4.4). In addition, the significantly lower than expected haplotype diversity (Hd) in serotypes A and D populations for five of the six inversion junctions [except that of SI(1)D_Left] further supported the hypothesis that selective sweeps have occurred at these gene regions. The signatures of positive selection for gene CNA03280 located within the inverted region SI(1)B is consistent with differential selection pressures between the two serotypes, which could have facilitated the establishment of the alternative chromosomal types in the two serotypes (Table 4.3). Gene CNA03280 is described as a transcriptional activator (gcn5) in GenBank. Previous studies have shown that transcriptional regulators are involved in sexual reproduction in *C. neoformans* (Hull et al. 2005). Whether CNA03280 is related to reproductive isolation and/or differential adaptation between the two serotypes remain to be examined.

For the inversion junctions, we found that the nucleotide polymorphisms were significantly lower among alleles from haploid strains than comparable groups of alleles from serotype AD strains (Wilcoxon Signed-Rank Test, n=5, P=0.05). These results suggest that alleles in serotype AD strains are likely experiencing accelerated rates of mutation accumulation, resulting from either higher nucleotide substitution rate or lower purifying selection pressure. The hypothesis of faster rate of mutation accumulation is also consistent with the observation that the majority of the polymorphism for alleles from serotype AD strains were contributed by singletons (Table 4.4).

The high mutation rate for alleles in serotype AD strains might be explained by two none mutually exclusive processes. The first one is "relaxed selection constraint." Because serotype AD strains were diploid or aneuploid, two copies of the same gene should be present, at least initially

after hybridization. The redundancy in gene copy number would allow these strains to be more tolerant to mutational pressure than haploid serotypes A and D strains, resulting in a faster rate of mutation accumulation in one or both of the alleles. This is similar to those that have been reported for duplicated genes, which have shown faster rates of evolution, possibly due to relaxed selection pressure (e.g. LYNCH and CONERY 2001). However, this hypothesis cannot explain the higher-level polymorphisms observed at the inversion junctions of serotype AD strains because these regions were no-coding sequences and mutations should accumulate at equal rates among haploid and diploid strain backgrounds.

The second possibility is related to the role of recombination for loci around inversions. Specifically, if the loci around inversions were involved in ensuring proper intra-variety mating and sexual reproduction, high sequence identity and low nucleotide polymorphism would be expected for these regions within each serotype. Because most serotype AD strains are diploid or aneuploid, they are unlikely to mate and reproduce sexually and are thus likely experiencing less pressure of ensuring proper sexual reproduction. As a result, mutations could have accumulated at these loci at higher rates in serotype AD strains. Regardless of the mechanism, the higher sequence polymorphism in alleles from serotype AD strains suggest the importance and potential of these hybrid strains in shaping the structure and dynamics of natural *C. neoformans* populations.

We observed unique signature sequences for each of the two mating types within each serotype (Figure 4.10) at the inverted regions. These sequences were contained within the serotype AD strains and mostly associated with MATa mating type backgrounds. Such differences could have resulted from two processes. In the first, the serotype AD strains might have resulted from ancient hybridization events followed by clonal dispersion. The differentiation we observed between alleles from different mating type backgrounds might be reflecting the divergence between alleles from different mating types at the time when hybridization occurred. The second explanation is that the serotype AD strains were derived from recent hybridizations. In this case, the high sequence diversity and unique signature sequences observed in serotype AD strains reflected a sub-population of *C. neoformans* that has not been analyzed by typical population genetic studies, which focus on predominantly haploid serotypes A and D strains (Xu et al. 2000, 2002). The current data favor the second possibility. For example, unlike the serotype AD strains that contained both mating types within each strain, the commonly analyzed serotypes A and D population samples were highly biased toward haploid MATalpha strains (BRANDT et al. 1996; BOVERS et al. 2008; YAN et al. 2002; Table 4.1). In addition, gene genealogical analyses have indicated that alleles within serotype AD strains often have identical counterparts in the current serotypes A and D strains and populations,

consistent with recent hybridization (XU et al. 2002). Indeed, multiple recent hybridization events have occurred for the generation of the current serotype AD strains (XU et al. 2002, 2003).

Previous studies have shown that the population structure of natural *C. neofromans* is mostly clonal (CURRIE et al. 1994; BRANDT et al. 1996; BOEKHOUT et al. 2001; MEYER et al. 2003; LITVINTSEVA et al. 2005), with limited signatures of sexual reproduction in natural populations of serotypes A and D (BURT et al. 2000; XU et al. 2000; LITVINTSEVA et al. 2005). Consistent with those studies, our study found an overall linkage disequilibrium between serotypes A and D. However, when serotypes A and D were analyzed separately, although PHT showed that each population still had congruent gene genealogies, a significant number of gene combinations did not show evidence of LD in the I_A/RbarD tests, more so in the serotype A population (46 of 66 gene combinations, $P > 0.05$, Figure 4.9) than in the serotype D population (40 of 66 gene combinations, $P > 0.05$, Figure 4.9). Interestingly, three of the six genes that were located more than 100 kb away from inversions showed no evidence of LD with other genes in I_A/RbarD tests. A possible explanation is that regions flanking (but not very close to) chromosomal rearrangements had enhanced recombination, which in turn can facilitate the random association among their alleles in the population. For example, HSUEH et al. (2006) found that recombination "hot spots" flanking the mating type locus of *C. neoformans* that contained extensive chromosomal rearrangements. There is likely a threshold distance between the marker gene and rearranged regions for enhanced recombination as our earlier study found suppressed recombination for loci located close to chromosomal rearrangements (SUN and XU 2009).

In conclusion, we found loci located away from inversions showed higher level nucleotide polymorphism than those close to inversions on Chromosome 1 in both the serotypes A and D populations of *C. neoformans*. Interestingly, we found evidence of recombination among markers located on the same chromosome within each serotype, and lower LD was observed between pairs of genes located further away from chromosomal inversions. Our results suggest that chromosomal inversions can have significant effects on population structure and dynamics of *C. neoformans* and should be taken into consideration in studies of *C. neoformans* natural populations. In addition, the elevated mutation rate for alleles from serotype AD strains suggest these hybrids may have significant adaptive advantage over parental serotypes A and D strains of this important human fungal pathogen.

CHAPTER 5

GENERAL CONCLUSION AND PERSPECTIVES

In this thesis, I presented a series of studies that showed: (1) serotypes A and D *Cryptococcus neoformans* are genetically well diverged from each other (Chapters 3 and 4); (2) though there is widespread hybridization, the hybrids seemed largely distinct and the genomes of the parental serotypes A and D populations are relatively unaffected by the hybrids in natural populations (Chapters 2, 3 and 4); and (3) the repression of genetic introgression between the two serotypes could be partially explained by the existence of serotype-specific chromosomal rearrangements (Chapter 3 and 4). These results are consistent with previous studies that showed significant divergence between serotypes A and D (for about18~40 million years) and the predominantly clonal population structure of *C. neoformans*. However, the widely distributed serotype-specific chromosomal rearrangements (e.g. chromosomal inversions; Chapter 3), their repression on recombination during inter-variety hybridization (Chapter 3) and the evidence of selective sweep at the inversion junctions on chromosome 1 (Chapter 4) suggest that the genes around rearrangements could have phenotypic effects, contributing to the differentiation between the two serotypes through enhancing reproduction isolation and/or niche-specific adaptations. For example, the gene CNA03280, which encodes a transcription activator, was found to be located within a chromosomal inversion (Chapter 3) and was shown to be under significant positive selection between the two serotypes (Chapter 4). It is possible that changes in this gene could have dramatic effects on the cellular processes and thus result in significant differentiation between the two serotypes. More molecular genetics studies of these regions and genes surrounding these regions will help us to better understand the diverging process between the two serotypes.

In my thesis, I have focused on the genotypic consequences of inter-variety hybridization between serotypes A and D of *C. neoformans*. However, this is only one side of the story. Natural selection ultimately works on phenotype. Transgressive segregation has been observed in inter-variety hybridization of *C. neoformans* (Shahid et al. 2008). And, as mentioned in General Introduction, unlike those in plants and animals, there is no evolutionary "dead end" in microbial hybrids, even if they are sterile, as they can propagate effectively through asexual reproduction. As such, the hybrids can continuously impact the structure of microbial populations. In addition, several medically important traits such as virulence and tolerance to high temperature and UV irradiation in *C. neoformans* have shown to be or are likely quantitative traits (Lin et al. 2006; Shahid et al. 2008). The two parental strains of the hybrid population used in the genetic linkage map construction in Chapter 2 also differed from each other significantly in their resistance to the common anti-fungal drug, Fluconazole. The patterns of fluconazole susceptibility among the progeny in this cross are consistent with multiple genes

controlling fluconazole resistance. Future analysis of phenotypic traits together with the genetic data, e.g. through quantitative trait locus (QTL) analysis, will help us identify the chromosomal regions that are responsible for the phenotypic differentiation between the two serotypes, and provide us with valuable knowledge on how to fight against this leading human pathogenic fungus.

In Chapter 4, we analyzed the effects of chromosomal inversions on the nucleotide polymorphism and population structure of chromosome 1 in natural *C. neoformans*. Our analyses focused on the largest chromosome, chromosome 1, that contained four of the nine chromosomal inversions in the genome. Whether inversions on other chromosomes or other types of chromosome rearrangements will show a similar pattern of molecular evolution remains to be determined. Such analyses will provide us with a more robust picture of the impact of chromosomal inversions on the population genetics of *C. neoformans*. In addition, we found evidence of accelerated mutation accumulation rate at the inversion junctions in serotype AD hybrids (Chapter 4). It would be interesting to study whether or not protein coding genes are experiencing higher evolution rates in serotype AD hybrids.

Finally, our study suggested there might be a large portion of natural *C. neofromans* serotypes A and D strains that have not been isolated and analyzed (Chapter 4). The extensive sampling in Botswana uncovered genetic variations not found in other geographic areas (Litvintseva et al. 2006). Whether other geographic regions contain unique genetic variations remain to be examined, through more extensive sampling and sequencing. More efforts should be made in identifying the natural niche of *C. neoformans*, finding the natural environments where they complete the sexual reproduction, and collecting *C. neoformans* isolates from environments as well as patients for population genetic studies. Also, the population genetics analyses should take into consideration the chromosomal rearrangements identified in Chapter 3, as we have shown that these rearrangements can have significant impacts on the population structure of natural *C. neoformans*.

REFERENCES

Agapow, P.M., and A. Burt, 2001 Indices of multilocus linkage disequilibrium. Mol. Ecol. Notes 1: 101-102.

Alspaugh, J. A., R. C. Davidson and J. Heitman, 2000 Morphogenesis of *Cryptococcus neoformans*, pp. 217-238 in *Dimorphism in Human Pathogenic and Apathogenic Yeasts*, edited by J. F. ERNST and A. SCHMIDT, Basel: Karger Contrib. Microbiol.

Armengol, L., T. Marquès-Bonet, J. Cheung, R. Khaja, J. R. González *et al.*, 2005 Murine segmental duplications are hot spots for chromosome and gene evolution. Genomics 86(6): 692-700.

Arnold, M. L., 1997 Natural Hybridization and Evolution. Oxford University Press, Oxford, England.

Barchiesi, F., M. Cogliati, M. C. Esposto, E. Spreghini, A. M. Schimizzi *et al.*, 2005 Comparative analysis of pathogenicity of *Cryptococcus neoformans* serotypes A, D and AD in murine cryptococcosis. J. Infect. 51: 10-16.

Barker, F.K., and F. M. Lutzoni, 2002 The utility of the incongruence length difference test. Syst Biol. 51: 625-637.

Barton, N. H., 2001 The role of hybridization in evolution. Mol Ecol. 10(3): 551-68.

Boekhout, T., and A. van Belkum, 1997 Variability of karyotypes and RAPD types in genetically related strains of *Cryptococcus neoformans*. Curr. Genet. 32: 203-208.

Boekhout, T., B. Theelen, M. R. Diaz, J. W. Fell, W. C. Hop *et al.*, 2001 Hybrid genotypes in the pathogenic yeast *Cryptococcus neoformans*. Microbiology 147: 891-907.

Bovers, M., F. Hagen, E. E. Kuramae and T. Boekhout, 2008 Six monophyletic lineages identified within *Cryptococcus neoformans* and *Cryptococcus gattii* by multi-locus sequence typing. Fungal Genet. Biol. 45: 400-421.

Brandt, M. E., L. C. Hutwagner, L. A. Klug, W. S. Baughman, D. Rimland *et al.*, 1996 Molecular subtype distribution of *Cryptococcus neoformans* in four areas of the United States. *Cryptococcal* Disease Active Surveillance Group. J. clin. Microbiol. 34: 912-917.

Brasier, C., 2000 Plant pathology: The rise of the hybrid fungi. Nature 405: 134-135.

Brideau, N. J., H. A. Flores, J. Wang, S. Maheshwari, X. Wang and D. A. Barbash, 2006 Two Dobzhansky-Muller Genes Interact to Cause Hybrid Lethality in *Drosophila*. Science 314: 1292-1295.

Burt, A., V. Koufopanou and J. W. Taylor, 2000 Population genetics of human-pathogenic fungi, pp. 229-244 in Molecular Epidemiology of Infectious Diseases, edited by R. A. Thompson. Edward Arnold, London.

Butlin, R. K., 2005 Recombination and speciation. Mol. Ecol. 14: 2621-2635.

Casadevall, A., and J. R. Perfect, 1998 *Cryptococcus neoformans*. ASM Press.

Chambers, S. R., N. Hunter, E. J. Louis and R. H. Borts, 1996 The mismatch repair system reduces meiotic homeologous recombination and stimulates recombination-dependent chromosome loss. Mol. Cell. Biol. 16: 6110-6120.

Chaturvedi, V., J. Fan, B. Stein, M. J. Behr, W. A. Samsonoff *et al.*, 2002 Molecular genetic analyses of mating pheromones reveal intervariety mating or hybridization in *Cryptococcus neoformans*. Infect. Immun. 70: 5225-5235.

Chen, W., and S. Jinks-Robertson, 1999 The role of the mismatch repair machinery in regulating mitotic and meiotic recombination between diverged sequences in yeast. Genetics 151: 1299-1313.

Chen, X., B. B. Magee, D. Dawson, P. T. Magee and C. A. Kumamoto, 2004 Chromosome 1 trisomy compromises the virulence of *Candida albicans*. Mol. Microbiol. 51: 551-565.

Coghlan, A., E. E. Eichler, S. G. Oliver, A. H. Paterson and L. Stein, 2005 Chromosome evolution in eukaryotes: a multi-kingdom perspective. Trends in Genetics 21: 673-682.

Coluzzi, M., A. Sabatini, A. della Torre, M. A. Di Deco and V. Petrarca, 2002 A polytene chromosome analysis of the *Anopheles gambiae* species complex. Science 298: 1415-1418.

Coste, A., A. Selmecki, A. Forche, D. Diogo, M. E. Bougnoux *et al.*, 2007 Genotypic evolution of azole resistance mechanisms in sequential *Candida albicans* isolates. Eukaryot. Cell. 6: 1889-1904.

Coyne, J. A., W. Meyers, A. P. Crittenden and P. Sniegowski, 1993 The fertility effects of pericentric inversions in *Drosophila melanogaster*. Genetics 134: 487-496.

Coyne, J. A., and H. A. Orr, 1998 The evolutionary genetics of speciation. Philos. Trans. R. Soc. London Ser. B 353: 287-305.

Coyne, J. A., and H. A. Orr, 2004 Speciation. Sinauer Associates.

Currie, B. P., L. F. Freundlich and A. Casadevall, 1994 Restriction fragment length polymorphism analysis of *Cryptococcus neoformans* isolates from environmental (pigeon excreta) and clinical sources in New York City. J. Clin. Microbiol. 32: 1188-1192.

Datta, A., A. Adjiri, L. New, C. F. Crouse and S. Jinks-Robertson, 1996 Crossovers between diverged sequences are regulated by mismatch repair proteins in yeast. Mol. Cell. Biol. 16: 1085-1093.

Datta, A., M. Hendrix, M. Lipsitch and S. Jinks-Robertson, 1997 Dual roles for DNA sequence identity and the mismatch repair system in the regulation of mitotic crossing-over in yeast. Proc. Natl. Acad. Sci. USA 94: 9757-9762.

Depaulis, F., L. Brazier and M. Veuille, 1999 Selective sweep at the *Drosophila melanogaster* suppressor of hairless locus and its association with the In(2L)t inverstion polymorphism. Genetics 152: 1017-1024.

Ding, K., K. Zhou, F. He and Y. Shen, 2003 LDA – a java-based linkage disequilibrium analyzer. Bioinformatics 19: 2147-2148.

Ellis, D., D. Marriott, R. A. Hajjeh, D. Warnock, W. Meyer and R. Barton, 2000 Epidemiology: surveillance of fungal infections. Med. Mycol. 38: 173-182.

Evans, E., and E. Alani, 2000 Roles for mismatch repair factors in regulating genetic recombination. Mol. Cell. Biol. 20: 7839-7844.

Farris, J. S., M. Kallersjo, A. G. Kluge and C. Bult, 1995a Testing significance of incongruence. Cladistics 10: 315-319.

Farris, J. S., M. Kallersjo, A. G. Kluge and C. Bult, 1995b Construction a significance test for incongruence. Syst. Biol. 44: 570-572.

Feder, J. L., J. B. Roethelle, K. Filchak, J. Niedbalski and J. Romero-Severson, 2003 Evidence for inversion polymorphism related to sympatric host race formation in the apple maggot fly, *Rhagoletis pomonella*. Genetics 163: 939-953.

Feder, J. L., X. Xie, J. Rull, S. Vellez, A. Forbes *et al.*, 2005 Mayr, Dobzhansky, and Bush and the complexities of sympatric speciation in Rhagoletis. Proc. Natl. Acad. Sci. USA 102: 6573-6580.

Fischer, G., E. P. Rocha, F. Brunet, M. Vergassola and B. Dujon, 2006 Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. PLoS Genetics 2: e32.

Forche, A., J. Xu, R. J. Vilgalys and T. G. Mitchell 2000 Development and characterization of a genetic linkage map of *Cryptococcus neoformans* var. *neoformans* using amplified fragment length polymorphisms and other markers. Fungal Genet. Biol. 31: 189-203.

Fraser, J. A., J. C. Huang, R. Pukkila-Worley, J. A. Alspaugh, T. G. Mitchell *et al.*, 2005 Chromosomal translocation and segmental duplication in *Cryptococcus neoformans*. Eukaryot. Cell 4: 401-406.

Fraser, J. A., S. S. Giles, E. C. Wenink, S. G. Geunes-Boyer, J. R. Wright *et al.*, 2005 Same-sex mating and the origin of the Vancouver island *Cryptococcus gattii* outbreak. Nature 437: 1360-1364.

Fries, B. C., F. Chen, B. P. Currie and A. Casadevall, 1996 Karyotype instability in *Cryptococcus neoformans* infection. Journal of Clinical Microbiology 34: 1531-1534.

Futuyma, D. J., and G. C. Mayer, 1980 Non-allpatric speciation in animals. Syst. Zool. 19: 254-271.

Goodwin, T. J., and R. T. Poulter, 2001 The diversity of retrotransposons in the yeast *Cryptococcus neoformans*. Yeast 18:865-880.

Grant, P. R., and B. R. Grant, 2002 Unpredictable Evolution in a 30-Year Study of Darwin's Finches. Science 296: 707-711.

Greig, D., R. H. Borts, E. J. Louis and M. Travisano, 2002 Epistasis and hybrid sterility in *Saccharomyces*. Proc. R. Soc. Lond. Ser B 269: 1167-1171.

Harrison, R. G., 1993 Hybrid Zones and the Evolutionary Process. Oxford University Press, Oxford, England.

Hartl, D. L., and A. G. Clark, 1997 Principles of Population Genetics. Sinauer Associates.

Hess, M. T., R. D. Gupta and R. D. Kolodner, 2002 Dominant *Saccharomyces cerevisiae* msh6 mutations cause increased mispair binding and decreased dissociation from mispairs by Msh2-Msh6 in the presence of ATP. J. Biol. Chem. 277: 25545-25553.

Hey, J., and R. Nielsen, 2004 Multilocus methods for estimating population sizes, migration rates and divergence time, with application to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167: 747-760.

Hoffmann, A. A., C. M. Sgrò and A. R. Weeks, 2004 Chromosomal inversion polymorphisms and adaptation. Trends in Ecology and Evolution 19: 482-488.

Hsueh, Y. P., A. Idnurm and J. Heitman, 2006 Recombination hotspots flank the *Cryptococcus* mating-type locus: implications for the evolution of a fungal sex chromosome. PLoS Genet. 2: e184.

Hu, G., I. Liu, A. Sham, J. E. Stajich, F. S. Dietrich and J. W. Kronstad, 2008 Comparative hybridization reveals extensive genome variation in the AIDS-associated pathogen *Cryptococcus neoformans*. Genome Biology 9: R41.

Hughes, D., 2000 Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes. Genome Biology 1: 1-8.

Hughes, T. R., C. J. Roberts, H. Dai, A. R. Jones, M. R. Meyer *et al.*, 2000 Widespread aneuploidy revealed by DNA microarray expression profiling. Nat. Genet. 25: 333-337.

Hull, C. M., M. J. Boily and J. Heitman, 2005 Sex-specific homeodomain proteins *Sxi1alpha* and *Sxi2a* coordinately regulate sexual development in *Cryptococcus neoformans*. Eukaryot Cell 4: 526-535.

Hunter, N., S. R. Chambers, E. J. Louis and R. H. Borts, 1996 The mismatch repair system contributes to meiotic sterility in an interspecific yeast hybrid. EMBO J. 15: 1726-1733.

Idnurm, A., Y. S. Bahn, K. Nielsen, X. Lin, J. A. Fraser and J. Heitman, 2005 Deciphering the model pathogenic fungus *Cryptococcus neoformans*. Nat. Rev. Microbiol. 3: 753-64.

Jin, Q., Z. Yuan, J. Xu, Y. Wang, Y. Shen *et al.*, 2002 Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through

comparison with genomes of *Escherichia coli* K12 and O157. Nucleic Acids Research 30: 4432-4441.

Jinks-Robertson, S., S. Shariq and T. Murphy, 1997 Meiotic crossing over between nonhomologous chromosomes affects chromosome segregation in yeast. Genetics 146: 69-78.

Johnston, J. A., M. L. Arnold and L. A. Donovan, 2003 High hybrid fitness at seed and seedling life history stages in Louisiana irises. Journal of Ecology 91: 438-446.

Jurgenson, J. E., R. L. Bowden, K. A. Zeller, J. F. Leslie, N. J. Alexander *et al.*, 2002 A genetic map of *Gibberella zeae* (*Fusarium graminearum*). Genetics 160: 1451-1460.

Kavanaugh, L. A., J. A. Fraser and F. S. Dietrich, 2006 Recent evolution of the human pathogen *Cryptococcus neoformans* by intervarietal transfer of a 14-gene fragment. Mol. Biol. Evol. 23: 1879-1890.

Kelly, J., 1997 A test of neutrality based on interlocus associations. Genetics 146: 1197-1206.

Kerrigan, R. W., J. C. Royer, L. M. Baller, Y. Kohli, P. A. Horgen *et al.*, 1993 Meiotic behaviour and linkage relationships in the secondarily homothallic fungus *Agaricus bisporus*. Genetics 133: 225-236.

Kinscherf, T. G., and S. A. Leong, 1988 Molecular analysis of the karyotype of *Ustilago maydis*. Chromosoma 96: 427-433.

Kirkpatrick, M., and N. Barton, 2006 Chromosome inversions, local adaptation and speciation. Genetics 173: 419-434.

Klepser, M. E., and M. A. Pfaller, 1998 Variation in Electrophoretic karyotype and antifungal susceptibility of clinical isolates of *Cryptococcus neoformans* at a University-Affiliated teaching hospital from 1987 to 1994. Journal of Clinical Microbiology 36: 3653-3656.

Kolodner, R. D., C. D. Putnam and K. Myung, 2002 Maintenance of genome stability in *Saccharomyces cerevisiae*. Science 297: 552-557.

Kwon-Chung, K. J., 1975 A new genus, *Filobasidiella*, the perfect state of *Cryptococcus neoformans*. Mycologia 67: 1197-1200.

Kwon-Chung, K. J., 1976 Morphogenesis of *Filobasidiella neoformans*, the sexual state of *Cryptococcus neoformans*. Mycologia 68: 821-823.

Kwon-Chung, K. J., and J. E. Bennett, 1978 Distribution of alpha and a mating types of *Cryptococcus neoformans* among natural and clinical isolates. Am. J. Epidemiol. 108: 337-340.

Kwon-Chung, K. J., J. C. Edman and B. L. Wickes, 1992 Genetic association of mating types and virulence in *Cryptococcus neoformans*. Infect. Immun. 60: 602-605.

Lan, L., and J. XU, 2006 Multiple gene genealogical analyses suggest divergence and recent clonal dispersal in the opportunistic human pathogen *Candida guilliermondii.* Microbiology. 152: 1539-1549.

Lande, R., 1985 The fixation of chromosomal rearrangements in a subdivided population with local extinction and colonization. Heredity 54: 323-332.

Lander, E. S., and P. Green, 1987 Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci. UAS 84: 2363-2367.

Lander, E. S., P. Green, A. Abrahamson, A. Barlow and M. J. Daly, 1987 MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1: 174-181.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan *et al.*, 2007 Clustal W and Clustal X version 2.0. Bioinformatics, 23: 2947-2948.

Lengeler, K. B., G. M. Cox and J. Heitman, 2001 Serotype AD strains of *Cryptococcus neoformans* are diploid or aneuploid and are heterozygous at the mating-type locus. Infect. Immun. 69: 115-122.

Lengeler, K. B., D. S. Fox, J. A. Fraser, A. Allen, K. Forrester et al., 2002 Mating-type locus of *Cryptococcus neoformans*: a step in the evolution of sex chromosomes. Eukaryotic Cell. 1: 704-718.

Lewontin, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics, 49: 49-67.

Lin, X., C. M. Hull and J. Heitman, 2005 Sexual reproduction between partners of the same mating type in *Cryptococcus neoformans*. Nature 434: 1017-1021.

Lin, X., and J. Heitman, 2006 The biology of the *Cryptococcus neoformans* species complex. Annu. Rev. Microbiol. 60: 69-105.

Lin, X., J. C. Huang, T. G. Mitchell and J. Heitman, 2006 Virulence Attributes and hyphal growth of *Cryptococcus neoformans* are quantitative traits and the MATalpha allele enhances filamentation. PLoS Genetics 2: e187.

Lin, X., A. P. Litvintseva, K. Nielsen, S. Patel, A. Floyd *et al.*, 2007 αADα hybrids of *Cryptococcus neoformans*: evidence of same-sex mating in nature and hybrid fitness. PLoS Genetics 3: e186.

Lin, X., K. Nielsen, S. Patel and J. Heitman, 2008 Impact of mating type, serotype, and ploidy on the virulence of *Cryptococcus neoformans*. Infection and Immunity 76: 2923-2938.

Lindblad-Toh, K, C. M. Wade, T. S. Mikkelsen, E. K. Karlsson, D. B. Jaffe *et al.*, 2005 Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438: 803-819.

Liti, G., D. B. Barton and E. J. Louis, 2006 Sequence diversity reproductive isolation, and species concepts in *Saccharomyces*. Genetics 174: 839-850.

Litvintseva, A. P., R. E. Marra, K. Nielsen, J. Heitman, R. Vilgalys *et al.*, 2003 Evidence of sexual recombination among *Cryptococcus*

neoformans serotype A isolates in sub-Saharan Africa. Eukaryot Cell 2: 1162-1168.

Litvintseva, A. P., L. Kestenbaum, R. Vilgalys and T. G. Mitchell, 2005 Comparative analysis of environmental and clinical populations of *Cryptococcus neoformans*. J. Clin. Microbiol. 43: 556-564.

Litvintseva, A. P., R. Thakur, R. Vilgalys and T. G. Mitchell, 2006 Multilocus sequence typing reveals three genetic subpopulations of *Cryptococcus neoformans* var. *grubii* (serotype A), including a unique population in Botswana. Genetics 172: 2223-2238.

Litvintseva, A. P., X. Lin, I. Templeton, J. Heitman and T. G. Mitchell, 2007 Many Globally Isolated AD Hybrid Strains of *Cryptococcus neoformans* Originated in Africa. PLoS Pathog. 3(8): e114.

Loftus, B. J., E. Fung, P. Roncaglia, D. Rowley, P. Amedeo et al., 2005 The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. Science 307: 1321-1324.

Lynch, M., and J. S. Conery, 2001 The Evolutionary Fate and Consequences of Duplicate Genes. Science 290: 1151-1155.

Machado, C. A., R. M. Kliman, J. A. Markert and J. Hey, 2002 Inferring the history of speciation using multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives. Mol. Biol. Evol. 19: 472-488.

Machado, C. A., and J. Hey, 2003 The causes of phylogenetic conflict in a classic *Drosophila* species group. Proc. R. Soc. Lond. Ser. B 270: 1193-1202.

Machado, C. A., T. S. Haselkorn and M. A. Noor, 2007 Evaluation of the genomic extent of effects of fixed inversion differences on intraspecific variation and interspecific gene flow in *Drosophila pseudoobscura* and *D. persimilis*. Genetics 175: 1289-1306.

Magee, B. B., and P. T. Magee, 1987 Electrophoretic karyotypes and chromosome numbers in *Candida* species. J. Gen. Microbiol. 133: 425-430.

Marra, R. E., J. C. Huang, E. Fung, K. Nielsen, J. Heitman *et al.*, 2004 A genetic linkage map of *Cryptococcus neoformans* variety *neoformans* serotype D (*Filobasidiella neoformans*). Genetics 167: 619-631.

Maynard-Smith, J., N. H. Smith, M. O'Rourke and B. G. Spratt, 1993 How clonal are bacteria? Proc Natl Acad Sci USA 90: 4384-4388.

McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351: 652-654.

Meyer, W., K. Marszewska, M. Amirmostofian, R. P. Igreja, C. Hardtke, K. Methling *et al.*, 1999 Molecular typing of global isolates of *Cryptococcus neoformans var. neoformans* by polymerase chain reaction fingerprinting and randomly amplified polymorphic DNA – a

pilot study to standardize techniques on which to base a detailed epidemiological survey. Electrophoresis 20: 1790-1799.

Meyer, W., A. Castaneda, S. Jackson, M. Huynh and E. Castaneda, 2003 Molecular typing of IberoAmerican *Cryptococcus neoformans* isolates. Emerg. Infect. Dis. 9: 189-195.

Michod, R. E., and B. R. Levin, 1988 The Evolution of Sex. Sinauer Associates.

Mitchell-Olds, T., J. H. Willis and D. B. Goldstein, 2007 Which evolutionary processes influence natural genetic variation for phenotypic traits? Nature Review Genetics 8: 845-856.

Navarro, A., and A. Ruiz, 1997 On the fertility effects of pericentric inversion. Genetics 147: 931-933.

Navarro, A., E. Betrán, A. Barbadilla and A. Ruiz, 1997 Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. Genetics 146: 695-709.

Navarro, A., and N. H. Barton, 2003a Accumulating postzygotic isolation genes in parapatry: A new twist on chromosomal speciation. Evolution 57(3): 447-459.

Navarro, A., and N. H. Barton, 2003b Chromosomal speciation and molecular divergence - Accelerated evolution in rearranged chromosomes. Science 300: 321-324.

Nei, M., 1987 Molecular Evolutionary Genetics. Columbia University Press, New York.

Nielsen, K, G. M. Cox, P. Wang, D. L. Toffaletti, J. R. Perfect *et al.*, 2003 Sexual cycle of *Cryptococcus neoformans* var. *grubii* and virulence of congenic a and a isolates. Infect. Immun. 71: 4831–4841.

Nielsen, K., G. M. Cox, A. P. Litvintseva, E. Mylonakis, S. D. Malliaris, *et al.*, 2005a *Cryptococcus neoformans* a strains preferentially disseminate to the central nervous system during coinfection. Infect. Immun. 73: 4922–4933.

Nielsen, K., R. E. Marra, F. Hagen, T. Boekhout, T. G. Mitchell, G. M. Cox and J. Heitman, 2005b Interaction between genetic background and the mating-type locus in *Cryptococcus neoformans* virulence potential. Genetics 171: 975–983.

Nielsen, K., A. L. De Obaldia and J. Heitman, 2007 *Cryptococcus neoformans* mates on pigeon guano: implications for the realized ecological niche and globalization. Eukaryot. Cell 6: 949-959.

Noor, M. A., K. L. Grams, L. A. Bertucci and J. Reiland, 2001 Chromosomal inversions and the reproductive isolation of species. PNAS. 98(21): 12084-12088.

Ono, B., and Y. Ishinoarao, 1988 Inheritance of chromosome length polymorphisms in *Saccharomyces cerevisiae*. Curr. Genet. 14: 413-418.

Ranz, J. M., D. Maurin, Y. S. Chan, M. von Grotthuss, L. W. Hillier *et al.*, 2007 Principles of genome evolution in the *Drosophila melanogaster* species group. PLoS Biology 5: e152.

Rieseberg, L. H., C. R. Linder and G. J. Seiler, 1995 Chromosomal and genic barriers to introgression in *Helianthus*. Genetics 141: 1163-1171.

Rieseberg, L. H., M. A. Archer and R. K. Wayne, 1999a Transgressive segregation, adaptation and speciation. Heridity 83: 363-372.

Rieseberg, L. H., J. Whitton and K. Gardner, 1999b Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. Genetics 152: 713-727.

Rieseberg, L. H., 2001 Chromosomal rearrangements and speciation. Trends in Ecology and Evolution 16: 351-358.

Rieseberg, L. H., O. Raymond, D. M. Rosenthal, Z. Lai, K. Livingstone, T. Nakazato, *et al.*, 2003 Major ecological transitions in wild sunflowers facilitated by hybridization. Science 301: 1211-1216.

Rozas, J., J. C. Sánchez-DelBarrio, X. Messeguer and R. Rozas, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496-2497.

Rustchenko, E., 2007 Chromosome instability in *Candida albicans*. FEMS Yeast Res. 7: 2-11.

Schardl, C. L., and K. D. Craven, 2003 Interspecific hybridization in plant-associated fungi and oomycetes: a review. Mol. Ecol. 12: 2861-2873.

Schein, J. E., K. L. Tangen, R. Chiu, H. Shin, K. B. Lengeler *et al.*, 2002 Physical maps for genome analysis of serotype A and D strains of the fungal pathogen *Cryptococcus neoformans*. Genome Res. 12: 1445-1453.

Schemske, D. W., 2000 Understanding the origin of species. Evolution 54: 1069-1073.

Selva, E. M., L. New, G. F. Crouse and R. S. Lahue, 1995 Mismatch correction acts as a barrier to homeologous recombination in *Saccharomyces cerevisiae*. Genetics 139: 1175-1188.

Shahid, M., S. Han, H. Yoell and J. Xu, 2008 Fitness distribution and transgressive segregation across 40 environments in a hybrid progeny population of the human-pathogenic yeast *Cryptococcus neoformans*. Genome 51: 272-281.

Sharpton, T. J., D. E. Neafsey, J. E. Galagan and J. W. Taylor, 2008 Mechanisms of intron gain and loss in *Cryptococcus*. Genome Biology 9: R24.

Sites, J. W., and C. Moritz, 1987 Chromosomal evolution and speciation revisited. Syst. Zool. 36: 153-174.

Staib, F., 1981 The perfect state of *Cryptococcus neoformans*, *Filobasidiella neoformans*, on pigeon manure filtrate agar. Zentbl. Bakteriol. A 248: 575-578.

Sukroongreung, S., S. Lim, S. Tantimavanich, B. Eampokalap, D. Carter *et al.*, 2001 Phenotypic switching and genetic diversity of *Cryptococcus neoformans*. Journal of Clinical Microbiology, 39: 2060-2064.

Sun, S., and J. Xu, 2007 Genetic analyses of A hybrid cross between serotypes A and D strains of the human pathogenic fungus *Cryptococcus neoformans*. Genetics 177: 1475-1486.

Sun, S., and J. Xu, 2009 Chromosomal rearrangements between serotype A and D strains in *Cryptococcus neoformans*. PLoS ONE 45: e5524.

Sweigart, A. L., L. Fishman and J. H. Willis, 2006 A simple genetic incompatibility causes hybrid male sterility in *Mimulus*. Genetics 172: 2465–2479.

Swofford, D. L., 2002 PAUP*: *Phylogenetic Analysis Using Parsimony. Sunderland*, MA: Sinauer Associates.

Swofford, D. L., 2003 PAUP*, Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Tajima, F., 1983 Evolutionary relationships of DNA sequences in finite populations. Genetics 105: 437-460.

Tajima, F., 1989 Statistical method for testing the neutral mutationhypothesis by DNA polymorphism. Genetics 123: 585-595.

Torres, E. M., T. Sokolsky, C. M. Tucker, L. Y. Chan, M. Boselli *et al.*, 2007 Effects of aneuploidy on cellular physiology and cell division in haploid yeast. Science 317: 916-924.

Tzeng, T. H., L. K. Lyngholm, C. F. Ford and C. R. Bronson, 1992 A restriction fragment length polymorphism map and electrophoretic karyotype of the fungal maize pathogen *Cochliobolus heterostrophus*. Genetics 130: 81-96.

Van Der Lee, T., I. Dewitte, A. Drenth, C. Alfonso and F. Govers, 1997 AFLP linkage map of the Oomycete *Phytophthora infestans*. Fungal Genet. Biol. 21: 278-291.

Walsh, J. B., 1982 Rate of sccumulation of reproductive isolation by chromosomal rearrangements. Am. Nat. 120: 510-532.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7: 256-276.

Wickes, B. L., T. D. E. Moore and K. J. Kwon-Chung, 1994 Comparison of the electrophoretic karyotypes and chromosomal location of ten genes in the two varieties of *Cryptococcus neoformans*. Microbiology 140: 543-550.

Xu, J., T. G. Mitchell and R. Vilgalys, 1999 PCR-RFLP analyses reveal both extensive clonality and local genetic differentiation in *Candida albicans*. Mol. Ecol. 8:59-74

Xu, J., R. J. Vilgalys and T. G. Mitchell, 2000 Multiple gene genealogies reveal recent dispersion and hybridization in the human pathogenic fungus *Cryptococcus neoformans*. Mol. Ecol. 9: 1471-1481.

Xu, J., C. Onyewu, H. J. Yoell, R. Y. Ali, R. J. Vilgalys, *et al.*, 2001 Dynamic and heterogeneous mutations to fluconazole resistance in *Cryptococcus neoformans*. Antimicrobial Agents and Chemotherapy. 45: 420-427.

Xu, J., 2002 Mitochondrial DNA polymorphisms in the human pathogenic fungus *Cryptococcus neoformans*. Curr Genet. 41:43-47

Xu, J., G. Luo, R. J. Vilgalys, M. E. Brandt and T. G. Mitchell, 2002 Multiple origins of hybrid strains of *Cryptococcus neoformans* with serotype AD. Microbiology 148: 203-212.

Xu, J., and T. G. Mitchell, 2003 Comparative gene genealogical analyses of strains of serotype AD identify recombination in populations of serotypes A and D in the human pathogenic yeast *Cryptococcus neoformans*. Microbiology 149: 2147-2154.

Xue, C., Y. Tada, X. Dong and J. Heitman, 2007 The human fungal pathogen *Cryptococcus* can complete its sexual cycle during a pathogenic association with plants. Cell Host & Microbe 1: 263-273.

Yan, Z., X. Li and J. Xu, 2002 Geographic distribution of mating type alleles of *Cryptococcus neoformans* in four areas of the United States. J. Clin. Microbiol. 40: 965-72.

Yan, Z., and J. Xu, 2003 Mitochondria are transmitted from MATa parent in sexual crosses of the human pathogenic fungus *Cryptococcus neoformans*. Genetics 163: 1315-1325.

Yang, J., J. Wang, Z. J. Yao, Q. Jin, Y. Shen *et al.*, 2003 GenomeComp: a visualization tool for microbial genome comparison. J. Microbiol. Meth. 54: 423-426.

Zhong, S., B. J. Steffenson, J. P. Martinez and L. M. Ciufetti, 2002 A molecular genetic map and electrophoretic karyotype of the plant pathogenic fungus *Cochliobolus sativus*. Mol. Plant-Microbe Interact. 15: 481-492.

Zolan, M. E., 1995 Chromosome-length polymorphism in fungi. Microbiol. Rev. 59: 686-698.