Recombination, balancing selection and adaptive evolution in the aflatoxin gene cluster of Aspergillus parasiticus

IGNAZIO CARBONE,* JUDY L. JAKOBEK,*JORGE H. RAMIREZ-PRADO* and BRUCE W. HORN†
*Center for Integrated Fungal Research, Department of Plant Pathology, North Carolina State University, Raleigh, NC 27695, USA, †National Peanut Research Laboratory, USDA, ARS, Dawson, GA 39842, USA

Abstract

Aflatoxins are toxic and carcinogenic polyketides produced by several Aspergillus species that are known to contaminate agricultural commodities, posing a serious threat to animal and human health. Aflatoxin (AF) biosynthesis is almost fully characterized and involves the coordinated expression of approximately 25 genes clustered in a 70-kb DNA region. Aspergillus parasiticus is an economically important and common agent of AF contamination. Naturally occurring nonaflatoxigenic strains of A. parasiticus are rarely found and generally produce O-methylsterigmatocystin (OMST), the immediate precursor of AF. To elucidate the evolutionary forces acting to retain AF and OMST pathway extrolites (chemotypes), we sequenced 21 intergenic regions spanning the entire cluster in 24 A. parasiticus isolates chosen to represent the genetic diversity within a single Georgia field population. Linkage disequilibrium analyses revealed five distinct recombination blocks in the A. parasiticus cluster. Phylogenetic network analyses showed a history of recombination between chemotype-specific haplotypes, as well as evidence of contemporary recombination. We performed coalescent simulations of variation in recombination blocks and found an approximately twofold deeper coalescence for cluster genealogies compared to noncluster genealogies, our internal standard of neutral evolution. Significantly deeper cluster genealogies are indicative of balancing selection in the AF cluster of A. parasiticus and are further corroborated by the existence of trans-species polymorphisms and common haplotypes in the cluster for several closely related species. Estimates of Ka/Ks for representative cluster genes provide evidence of selection for OMST and AF chemotypes, and indicate a possible role of chemotypes in ecological adaptation and speciation.

Keywords: coalescent, compatibility, linkage disequilibrium, recombination blocks

Received 1 March 2007; revision accepted 11 June 2007

Introduction

One of the most interesting concepts arising from analyses of complete eukaryotic genomes is that genes which have similar and/or coordinated expression, such as those involved in the synthesis of secondary metabolites, are not randomly distributed throughout the genome, as was once thought, but are in fact clustered (reviewed in Hurst et al. 2004; Yu & Keller 2005). This realization has led to the logical inquiry of the evolutionary mechanisms which fashioned biosynthetic clusters in plants, yeast and other eukaryotic microorganisms (Qi et al. 2004; Wong & Wolfe 2005). Although proposed explanations for this phenomenon abound, in general there is no clear understanding of the evolutionary processes driving formation and maintenance of metabolic pathway clusters. Several authors have asserted that if we are to understand gene regulation and evolution, we must first understand the evolutionary processes responsible for gene clustering (Trowsdale 2002; Hurst et al. 2004; Qi et al. 2004). This is a challenging task, as it requires that we separate cluster-specific processes such as recombination and selection from homogeneous processes such as selective sweeps, population size changes and demographic effects. The latter processes result in genome-wide changes in the mutation frequency spectrum that may confound our inferences on cluster-specific processes. In this study, we dissect the
contribution of each of these forces in the evolution of the aflatoxin (AF) gene cluster in the filamentous fungus, *Aspergillus parasiticus*, a common agent of AF contamination of food commodities, particularly peanuts (Cotty & Bhatanagar 1994; Ehrlich et al. 2003).

Aflatoxins are toxic and carcinogenic polyketides produced by several *Aspergillus* species, which include *Aspergillus flavus* and *A. parasiticus*, the species most responsible for aflatoxin contamination of oil-rich crops (Horn 2005). The biosynthesis of AF involves approximately 25 genes that convert acetate and malonate to the intermediates sterigmatocystin (ST) and O-methylsterigmatocystin (OMST). OMST is an immediate precursor of approximately 25 genes that convert acetate and malonate to the intermediates sterigmatocystin (ST) and O-methylsterigmatocystin (OMST). OMST is an immediate precursor of the genetic diversity within a single field population, we were able to simultaneously reconstruct the B-family of aflatoxins (B₁ and B₂); the conversion of OMST to the G-family of aflatoxins (G₁ and G₂) is a separate biosynthetic pathway that has recently been characterized (Ehrlich et al. 2004). The accumulation of aflatoxins and their precursor extrolites (chemotypes) varies among different *Aspergillus* species (Carbone et al. 2007) and on a population scale. For example, populations of *A. parasiticus* include strains that produce B and G aflatoxins, strains that accumulate only OMST (Horn et al. 1996), and strains with differences in G₁ production relative to B₁ (G₁/B₁ ratio). Naturally occurring nonaflatoxicigenic strains of *A. parasiticus* are rarely found and generally produce OMST, whereas nonaflatoxicigenic *A. flavus* strains are relatively common and produce no detectable ST or OMST (Horn et al. 1996; Horn & Dorner 1999). To examine the basis for the retention of specific chemotypes in *A. parasiticus* we focused our evolutionary and population genetic analyses on molecular sequence variation in the AF cluster.

The ST-AF pathway represents one of the best-studied pathways of fungal secondary metabolism. The ST cluster was the first to be fully sequenced and many of the steps for ST biosynthesis were first elucidated in *Aspergillus nidulans* (Brown et al. 1996). Completion of nucleotide sequencing of the entire 70-kb gene cluster for *A. parasiticus* (Yu et al. 2004a) and more recently several other *Aspergillus* species (Galagan et al. 2005; Machida et al. 2005; Nierman et al. 2005) provided us with a unique opportunity to examine evolutionary processes influencing this very important fungal secondary metabolic pathway. Previous studies of AF pathway evolution have primarily focused on exon or intergenic polymorphisms for one or a small number of genes within the cluster (Geiser et al. 1998; Peterson et al. 2001; Klich et al. 2003; Cary et al. 2005; Ehrlich et al. 2005b).

By focusing our analysis on 21 intergenic regions spanning the entire AF cluster in 24 *A. parasiticus* isolates chosen to represent the genetic diversity within a single field population, we were able to simultaneously reconstruct the complex history of mutation and recombination in the cluster.

Population genetic methods are increasingly providing evidence of recombination in several *Aspergillus* species. The report of mating-type genes in *Aspergillus fumigatus* (Paolotti et al. 2005) and *Aspergillus oryzae* (Galagan et al. 2005), two species that are not known to be sexual, indicates the possibility of recombination and a cryptic sexual state in nature. Similarly, the observed discordance among gene genealogies inferred from multiple loci in *A. flavus* is consistent with a history of recombination and possibly cryptic speciation (Geiser et al. 1998). There is no known sexual stage in *A. parasiticus* and *A. flavus*, but genetic exchange via asexual recombination has been observed under laboratory conditions (Papa 1973, 1978). Vegetatively compatible fungal cells (hyphae) can fuse with one another (anastomosis) forming a heterokaryon, which may ultimately result in fusion of unlike nuclei and parasexual genetic exchange. Another possible means of asexual genetic variation in *Aspergillus* may be horizontal gene transfer. Indeed, it has been long proposed that metabolic gene clusters may be transferred horizontally between organisms (Walton 2000; Andersson 2005); however, our recent examination of patterns of gene duplication and modularity in the cluster of several *Aspergillus* species points to vertical transmission as a more plausible mechanism (Carbone et al. 2007). Here we show that molecular sequence variation in the AF cluster of a single Georgia field population of *A. parasiticus* is consistent with a history of recombination, balancing selection and positive Darwinian selection.

**Materials and methods**

**Sample**

Our initial assay of molecular variation in the AF cluster included 76 single-spore isolates of *Aspergillus parasiticus* sampled from a single peanut field in southwestern Georgia in 1992 (Horn & Greene 1995) and 68 isolates from vegetative compatibility group (VCG) 1 collected in 1996 from 25 fields distributed along a transect in North Carolina (NC), South Carolina (SC), Georgia (GA), Alabama (AL), and Texas (TX) (Horn & Dorner 1998). Horn & Greene (1995) grouped the 76 isolates into nine VCGs (numbers 1–9) comprising two or more isolates and eight VCGs (numbers 10–17) containing single isolates.

**DNA isolation, PCR amplification and sequencing**

Fungal mycelia were grown on potato dextrose medium (Difco) supplemented with 75 µg/mL ampicillin for 3–5 days in darkness at 30 °C. Mycelia were then collected on cheesecloth, rinsed 3x with water and lyophilized until completely dry. The dried samples were held at ~80 °C. For DNA extraction, a portion of the lyophilized mycelia was ground to a fine powder with a mortar and pestle cooled with liquid nitrogen. Genomic DNA was isolated from...
powdered mycelia (30–50 mg) with the Plant DNeasy kit (QIAGEN). The final DNA suspension was diluted 1:10 with de-ionized water and held at 4 °C for polymerase chain reaction (PCR) amplification. PCR oligonucleotide primers (18–23-mers) were designed using primer3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi) to amplify approximately 1000 bp of the intergenic regions separating 21 genes in the AF gene cluster (see Table 1). Intergenic regions were targeted because (i) they are known to harbour more variation than coding regions, which are reported to be 90–99% similar between A. parasiticus and Aspergillus flavus (Ehrlich et al. 2005b), and

<table>
<thead>
<tr>
<th>Target region</th>
<th>Primer name</th>
<th>Primer sequence (5′–3′)</th>
<th>Intergenic/exon*</th>
</tr>
</thead>
<tbody>
<tr>
<td>aflT/aflC</td>
<td>aflC-F</td>
<td>GGTTCCTCAATGCTCACTGAT</td>
<td>397/0</td>
</tr>
<tr>
<td></td>
<td>aflC-R</td>
<td>ACCCTCTGCAATGATACGG</td>
<td></td>
</tr>
<tr>
<td>aflC/aflD</td>
<td>aflD-F</td>
<td>TTAATTCTCAGTTCACTGTAG</td>
<td>341/0</td>
</tr>
<tr>
<td></td>
<td>aflD-R</td>
<td>TATTTCTGCAATGATACGG</td>
<td></td>
</tr>
<tr>
<td>aflD/aflA</td>
<td>aflA-F</td>
<td>TTGTGCTGCAATGATACGG</td>
<td>297/0</td>
</tr>
<tr>
<td></td>
<td>aflA-R</td>
<td>TTTTCTGCAATGATACGG</td>
<td></td>
</tr>
<tr>
<td>aflA/aflB</td>
<td>aflB-F</td>
<td>CGGCTGCAATGCTGGATAG</td>
<td>334/219 (aflA)</td>
</tr>
<tr>
<td></td>
<td>aflB-R</td>
<td>TTGTGCTGCAATGATACGG</td>
<td></td>
</tr>
<tr>
<td>aflB/aflR</td>
<td>aflR-F</td>
<td>AGGATGCTGCAATGATACGG</td>
<td>355/0</td>
</tr>
<tr>
<td></td>
<td>aflR-R</td>
<td>TTTTCTGCAATGATACGG</td>
<td></td>
</tr>
<tr>
<td>aflR/aflS</td>
<td>aflS-F</td>
<td>TGCAGCTGCAATGCTGGAT</td>
<td>337/192+ (aflR)</td>
</tr>
<tr>
<td>aflS/aflH</td>
<td>aflH-F</td>
<td>GAAGAGGGCTGCAATGCTGGAT</td>
<td>303/93 (aflS)</td>
</tr>
<tr>
<td></td>
<td>aflH-R</td>
<td>GGTTCCTCAATGCTCACTGAT</td>
<td></td>
</tr>
<tr>
<td>aflH/aflJ</td>
<td>aflJ-F</td>
<td>GTCTCTGCAATGCTCACTGAT</td>
<td>459/63 (aflJ)</td>
</tr>
<tr>
<td></td>
<td>aflJ-R</td>
<td>TGTCCTGCAATGCTCACTGAT</td>
<td></td>
</tr>
<tr>
<td>aflJ/aflE</td>
<td>aflE-F</td>
<td>GGCAGCTGCAATGCTCACTGAT</td>
<td>0/456 (aflE)</td>
</tr>
<tr>
<td></td>
<td>aflE-R</td>
<td>CGGCTGCAATGCTCACTGAT</td>
<td></td>
</tr>
<tr>
<td>aflE/aflM</td>
<td>aflM-F</td>
<td>GTAACACCGCTGCAATGCTGGAT</td>
<td>283/59 (aflM)</td>
</tr>
<tr>
<td></td>
<td>aflM-R</td>
<td>TTGTGCTGCAATGATACGG</td>
<td></td>
</tr>
<tr>
<td>aflM/aflN</td>
<td>aflN-F</td>
<td>GGTTCCTCAATGCTCACTGAT</td>
<td>429/5 (aflN)</td>
</tr>
<tr>
<td></td>
<td>aflN-R</td>
<td>GGTTCCTCAATGCTCACTGAT</td>
<td></td>
</tr>
<tr>
<td>aflN/aflG</td>
<td>aflG-F</td>
<td>CGGCTGCAATGCTCACTGAT</td>
<td>291/186 (aflG)</td>
</tr>
<tr>
<td></td>
<td>aflG-R</td>
<td>CGGCTGCAATGCTCACTGAT</td>
<td></td>
</tr>
<tr>
<td>aflG/aflL</td>
<td>aflL-F</td>
<td>TTCCTGCAATGCTCACTGAT</td>
<td>443/215 (aflL)</td>
</tr>
<tr>
<td></td>
<td>aflL-R</td>
<td>CGGCTGCAATGCTCACTGAT</td>
<td></td>
</tr>
<tr>
<td>aflL/aflO</td>
<td>aflO-F</td>
<td>GTCCCTGCAATGCTCACTGAT</td>
<td>80/278 (aflO)</td>
</tr>
<tr>
<td></td>
<td>aflO-R</td>
<td>GGCCATGCAATGCTCACTGAT</td>
<td></td>
</tr>
<tr>
<td>aflO/aflP</td>
<td>aflP-F</td>
<td>TTGTGCTGCAATGCTCACTGAT</td>
<td>329/20 (aflP)</td>
</tr>
<tr>
<td></td>
<td>aflP-R</td>
<td>TTGTGCTGCAATGCTCACTGAT</td>
<td></td>
</tr>
<tr>
<td>aflP/aflQ</td>
<td>aflQ-F</td>
<td>CGGCTGCAATGCTCACTGAT</td>
<td>282/0</td>
</tr>
<tr>
<td></td>
<td>aflQ-R</td>
<td>CGGCTGCAATGCTCACTGAT</td>
<td></td>
</tr>
<tr>
<td>aflQ/aflK</td>
<td>aflK-F</td>
<td>GCCCTGCAATGCTCACTGAT</td>
<td>411/0</td>
</tr>
<tr>
<td></td>
<td>aflK-R</td>
<td>GCCCTGCAATGCTCACTGAT</td>
<td></td>
</tr>
<tr>
<td>aflK/aflV</td>
<td>aflV-F</td>
<td>TTGTGCTGCAATGCTCACTGAT</td>
<td>200/131 (aflV)</td>
</tr>
<tr>
<td></td>
<td>aflV-R</td>
<td>TTGTGCTGCAATGCTCACTGAT</td>
<td></td>
</tr>
<tr>
<td>aflV/aflW</td>
<td>aflW-F</td>
<td>TTGTGCTGCAATGCTCACTGAT</td>
<td>558/6 (aflW)</td>
</tr>
<tr>
<td></td>
<td>aflW-R</td>
<td>TTGTGCTGCAATGCTCACTGAT</td>
<td></td>
</tr>
</tbody>
</table>

*The ratio of the intergenic region to exon (including introns); †474 bp of the total length overlaps with an aflR antisense transcript (Woloshuk et al. 1994); ‡506 bp of the total length overlaps with a hypothetical gene, hypE (GenBank Accession no. CO133708). Of these 506 bp, only 414 bp are coding and 92 bp correspond to an intron; §21 bp and 155 bp of the total length overlap with aflL and hypB (http://www.aspergillusflavus.org/genomics/; feature name 77.m04242), respectively.

Table 1 Primers amplifying intergenic regions in the aflatoxin cluster of Aspergillus parasiticus
(ii) they are presumably neutrally evolving regions that can potentially serve as a molecular clock in coalescent analyses (described below). To compare the mutation frequency in cluster and noncluster regions, we amplified the following noncluster regions: the 5’ nontranscribed portion of tryptophan synthase (trpC; located on chromosome 4), portions of genes encoding acetamidase (amdS; chromosome 6) and beta-tubulin (benA; chromosome 6), and the internal transcribed spacer 1 (ITS-1; chromosome 7) region. The primer sequences for amdS, benA, trpC and ITS-1 were as published previously (White et al. 1990; Geiser et al. 1998). Primers were synthesized by Sigma-Genosys. Given the similarity between A. parasiticus, A. flavus and other species in section Flavi, the ITS-1 region was sequenced to further confirm the identity of our A. parasiticus isolates.

PCR amplification was performed in 20-µL reactions, which included 1 µL of the above DNA dilution, 1 U Red Taq DNA polymerase (Sigma-Aldrich), 2 µL Red Taq buffer, 10 mm deoxynucleotide triphosphates, and 0.5 µM of each primer. Reactions were run in a Mastercycler ep gradient S (Brinkmann Instruments) with the profile of 5 min at 94°C followed by 40 cycles of 30 s at 94 °C, 30 s at 63 °C [except for loci amdS (55 °C), trpC (60 °C), benA (58 °C), and ITS-1 (62 °C)] and 1 min at 72 °C. PCR-amplified DNA was purified using QIAquick multivell PCR purification kit (QIAGEN). DNA sequencing was performed by the North Carolina State University Genome Research Laboratory. Sequences were aligned and edited using SEQUENCER version 4.5 (Gene Codes). Complementary strands were only sequenced for strains with ambiguous bases at variable positions that could not be resolved by viewing chromatograms. Multiple DNA sequence alignments were exported in NEXUS format and imported into SNAP Workbench version 1.0 (Price & Carbone 2005) for further analysis.

Molecular variation

We initially sequenced five intergenic regions spanning the entire cluster: aflT/aflC, aflE/aflM, aflM/aflN, aflK/aflV, and aflV/aflW. All of the variation observed was between VCGs; there was no variation within VCGs and no variation in the expanded VCG 1 sample. The polymorphisms in aflM/aflN and aflK/aflV intergenic regions showed strong genetic subdivision between the O-methylsterigmatocystin-accumulating strains (will refer to as OMS strain) and one strain that accumulated predominantly G1 and had the highest G1/B1 ratio in the sample (will refer to as G1-dominant chemotype). Based on these preliminary results, we selected a smaller subsample of 24 isolates of A. parasiticus for in-depth population analyses (see Table 2). The subsample consisted of eight isolates from VCG 1 [three isolates from the single GA field and one from each of the five states (NC, SC, GA, AL, and TX) along the transect] and 16 isolates representing VCGs 2–17. It is important to note that this subsample includes sampling of population diversity on two scales. First, the subsample is representative of VCG diversity in the population and therefore can provide inferences on population processes. Second, it includes the sampled chemotype diversity in the population and thus can provide insights into the mechanisms underlying chemotype differences. We sequenced portions of 21 intergenic regions in the cluster for our subsample of 24 strains. We also sequenced portions of trpC, amdS, benA, and ITS-1. The sequencing of genes outside the AF cluster was important to determine whether our observation of strong genetic subdivision between OMS and G1-dominant chemotypes was corroborated by variation in noncluster regions.

Nucleotide diversity and neutrality tests

For each of the 21 intergenic regions, we calculated nucleotide diversity, which is the average number of nucleotide differences per site between two sequences, π (Nei 1987), and evaluated four simple neutral models of evolution: Tajima’s D (Tajima 1989), Fu and Li’s D* and F* (Fu & Li 1993) and Fu’s Fs (Tajima 1989; Fu & Li 1993; Fu 1997). Significance thresholds for neutrality tests were Bonferroni-corrected by dividing by the total number of tests performed (n = 84). Neutrality tests were used to provide a priori inference on population processes (e.g. population subdivision or balancing selection) and to detect deviations in population-size constancy that may result in departures from neutrality. These tests were performed on each intergenic region separately as they assume no recombination and no migration. Significant test results for Tajima’s D, Fu and Li’s D* and F* and nonsignificant values of Fu’s Fs across the entire cluster are indicative of balancing selection (Fu 1997). A significant test result for Fu’s Fs suggests population growth or genetic hitchhiking. Nonsignificant test results can also be informative by examining whether the values of several tests are uniformly positive or negative. For example, D and Fs tend to be negative when there is an excess of rare variants, an indication of genetic hitchhiking/selective sweep or population growth, and tend to be positive when there is an excess of intermediate-frequency alleles, as expected under a model of population subdivision or balancing selection.

Recombination and compatibility analysis

Multiple DNA sequence alignments generated separately for each intergenic region were concatenated into one alignment file using SNAP Combine (Aylor et al. 2006). Base substitutions and insertions or deletions (indels) were recoded as binary characters using SNAP Map (Aylor et al. © 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd
Table 2 Strain designations, haplotypes, VCGs and mean aflatoxin/OMST concentrations in Aspergillus parasiticus*  

<table>
<thead>
<tr>
<th>IC</th>
<th>NPL</th>
<th>NRRL</th>
<th>Haplotype</th>
<th>VCG</th>
<th>B₁</th>
<th>B₂</th>
<th>G₁</th>
<th>G₂</th>
<th>Total</th>
<th>G₁/B₁</th>
<th>OMST</th>
</tr>
</thead>
<tbody>
<tr>
<td>IC2</td>
<td>P9</td>
<td>29539</td>
<td>H7</td>
<td>1</td>
<td>95(9)</td>
<td>20(4)</td>
<td>72(6)</td>
<td>20(0)</td>
<td>171(15)</td>
<td>0.80(0.01)</td>
<td>0(0)</td>
</tr>
<tr>
<td>IC11</td>
<td>P51</td>
<td>29548</td>
<td>H7</td>
<td>1</td>
<td>92(3)</td>
<td>20(5)</td>
<td>80(8)</td>
<td>20(0)</td>
<td>176(8)</td>
<td>0.90(0.09)</td>
<td>0(0)</td>
</tr>
<tr>
<td>IC15</td>
<td>P69</td>
<td>29552</td>
<td>H7</td>
<td>1</td>
<td>91(15)</td>
<td>20(3)</td>
<td>74(12)</td>
<td>2(1)</td>
<td>170(27)</td>
<td>0.80(0.03)</td>
<td>0(0)</td>
</tr>
<tr>
<td>IC25</td>
<td>P36</td>
<td>29562</td>
<td>H1</td>
<td>2</td>
<td>112(24)</td>
<td>40(8)</td>
<td>158(37)</td>
<td>3(2)</td>
<td>278(59)</td>
<td>1.40(0.18)</td>
<td>0(0)</td>
</tr>
<tr>
<td>IC38</td>
<td>P52</td>
<td>29575</td>
<td>H7</td>
<td>3</td>
<td>91(7)</td>
<td>20(5)</td>
<td>96(7)</td>
<td>10(0)</td>
<td>191(14)</td>
<td>1.00(0.06)</td>
<td>0(0)</td>
</tr>
<tr>
<td>IC43</td>
<td>P13</td>
<td>29580</td>
<td>H10</td>
<td>4</td>
<td>377(70)</td>
<td>11(3.4)</td>
<td>36(11)</td>
<td>1(1)</td>
<td>426(85)</td>
<td>0.10(0.01)</td>
<td>0(0)</td>
</tr>
<tr>
<td>IC56</td>
<td>P48</td>
<td>29593</td>
<td>H7</td>
<td>5</td>
<td>177(12)</td>
<td>60(3)</td>
<td>203(13)</td>
<td>4(1)</td>
<td>390(25)</td>
<td>1.10(0.03)</td>
<td>0(0)</td>
</tr>
<tr>
<td>IC59</td>
<td>P77</td>
<td>29596</td>
<td>H3</td>
<td>6</td>
<td>124(11)</td>
<td>3(1.1)</td>
<td>229(23)</td>
<td>4(1)</td>
<td>361(32)</td>
<td>1.90(0.16)</td>
<td>0(0)</td>
</tr>
<tr>
<td>IC61</td>
<td>P63</td>
<td>29598</td>
<td>H9</td>
<td>7</td>
<td>142(24)</td>
<td>5(1.5)</td>
<td>191(41)</td>
<td>4(2)</td>
<td>343(78)</td>
<td>1.30(0.05)</td>
<td>0(0)</td>
</tr>
<tr>
<td>IC65</td>
<td>P20</td>
<td>29602</td>
<td>H4</td>
<td>8</td>
<td>0(0)</td>
<td>0(0.0)</td>
<td>0(0)</td>
<td>0(0)</td>
<td>0(0)</td>
<td>—</td>
<td>243(32)</td>
</tr>
<tr>
<td>IC67</td>
<td>P23</td>
<td>29604</td>
<td>H7</td>
<td>9</td>
<td>97(7)</td>
<td>15(1.0)</td>
<td>143(8)</td>
<td>16(1)</td>
<td>271(14)</td>
<td>1.50(0.03)</td>
<td>0(0)</td>
</tr>
<tr>
<td>IC69</td>
<td>P8</td>
<td>29606</td>
<td>H7</td>
<td>10</td>
<td>59(44)</td>
<td>2(1.5)</td>
<td>85(62)</td>
<td>2(1)</td>
<td>148(109)</td>
<td>1.50(0.11)</td>
<td>0(0)</td>
</tr>
<tr>
<td>IC70</td>
<td>P18</td>
<td>29607</td>
<td>H5</td>
<td>11</td>
<td>237(32)</td>
<td>14(1.7)</td>
<td>475(43)</td>
<td>22(2)</td>
<td>747(78)</td>
<td>2.00(0.10)</td>
<td>0(0)</td>
</tr>
<tr>
<td>IC71</td>
<td>P26</td>
<td>29608</td>
<td>H4</td>
<td>12</td>
<td>0(0)</td>
<td>0(0.0)</td>
<td>0(0)</td>
<td>0(0)</td>
<td>0(0)</td>
<td>—</td>
<td>83(15)</td>
</tr>
<tr>
<td>IC72</td>
<td>P42</td>
<td>29609</td>
<td>H3</td>
<td>13</td>
<td>167(15)</td>
<td>3(1.2)</td>
<td>253(34)</td>
<td>5(2)</td>
<td>429(51)</td>
<td>1.50(0.08)</td>
<td>0(0)</td>
</tr>
<tr>
<td>IC73</td>
<td>P44</td>
<td>29610</td>
<td>H2</td>
<td>14</td>
<td>74(4)</td>
<td>2(0.2)</td>
<td>378(22)</td>
<td>13(0)</td>
<td>467(26)</td>
<td>5.10(0.09)</td>
<td>0(0)</td>
</tr>
<tr>
<td>IC74</td>
<td>P50</td>
<td>29611</td>
<td>H6</td>
<td>15</td>
<td>98(5)</td>
<td>3(1.1)</td>
<td>189(9)</td>
<td>40(0)</td>
<td>293(8)</td>
<td>1.90(0.18)</td>
<td>0(0)</td>
</tr>
<tr>
<td>IC75</td>
<td>P59</td>
<td>29612</td>
<td>H7</td>
<td>16</td>
<td>142(2)</td>
<td>6(0.4)</td>
<td>203(6)</td>
<td>40(0)</td>
<td>355(8)</td>
<td>1.40(0.02)</td>
<td>0(0)</td>
</tr>
<tr>
<td>IC76</td>
<td>P61</td>
<td>29613</td>
<td>H4</td>
<td>17</td>
<td>0(0)</td>
<td>0(0.0)</td>
<td>0(0)</td>
<td>0(0)</td>
<td>0(0)</td>
<td>—</td>
<td>19(14)</td>
</tr>
<tr>
<td>IC77</td>
<td>NC1-3</td>
<td>35604</td>
<td>H7</td>
<td>1</td>
<td>59(4)</td>
<td>10(1.1)</td>
<td>40(4)</td>
<td>10(0)</td>
<td>101(8)</td>
<td>0.70(0.03)</td>
<td>0(0)</td>
</tr>
<tr>
<td>IC78</td>
<td>SC1-1</td>
<td>35605</td>
<td>H7</td>
<td>1</td>
<td>62(13)</td>
<td>20(5)</td>
<td>43(12)</td>
<td>10(0)</td>
<td>107(25)</td>
<td>0.70(0.08)</td>
<td>0(0)</td>
</tr>
<tr>
<td>IC98</td>
<td>GA6-17</td>
<td>35606</td>
<td>H7</td>
<td>1</td>
<td>46(1)</td>
<td>10(1.1)</td>
<td>28(2)</td>
<td>0(0)</td>
<td>75(2)</td>
<td>0.60(0.04)</td>
<td>0(0)</td>
</tr>
<tr>
<td>IC115</td>
<td>AL3-10</td>
<td>35607</td>
<td>H7</td>
<td>1</td>
<td>80(8)</td>
<td>3(0.2)</td>
<td>26(1)</td>
<td>10(0)</td>
<td>109(7)</td>
<td>0.30(0.04)</td>
<td>0(0)</td>
</tr>
<tr>
<td>IC123</td>
<td>TX11-1</td>
<td>35608</td>
<td>H7</td>
<td>1</td>
<td>48(16)</td>
<td>10(0.6)</td>
<td>36(16)</td>
<td>10(0)</td>
<td>87(33)</td>
<td>0.70(0.10)</td>
<td>0(0)</td>
</tr>
</tbody>
</table>

*Concentration units are µg/mL. Mean concentrations and ratios (n = 3) are shown; standard deviations are indicated in parentheses.

2006); segregating sites that violated an infinite-sites mutation model were excluded. Linkage disequilibrium (LD) across the entire cluster was quantified using r², the coefficient of determination between the allelic states at pairs of sites, and the strength of LD was assessed using a two-sided Fisher’s exact test implemented in Tassel version 1.1.0 (http://www.maizegenetics.net). An LD plot showing the physical location of segregating positions in the cluster was generated using the Matrix program in SNAP Workbench. The plot was examined to identify blocks of informative segregating sites that were highly correlated (0.8 < r² < 1) and significant (P < 0.01). We quantified the magnitude of recombination in the cluster by estimating the population recombination parameter using pairwise or composite likelihood methods (Hudson 2001; McVean et al. 2002; Fearnhead 2003) was computationally prohibitive because of the long cluster sequences (Fearnhead & Donnelly 2001).

Many recombination events in the cluster can go undetected if we restrict our analysis solely to patterns of LD. To enhance our detection of historical recombination events, we estimated the minimum (lower) bound on the number of recombination events using the programs RecMin (Myers & Griffiths 2003) and HapBound (Song et al. 2005). We also calculated a maximum (upper) bound on the number of recombination events using the program Shrub (Song et al. 2005). These analyses were performed on haplotypes that were inferred by collapsing sequence alignments using SNAP Map (Aylor et al. 2006). Our collapsing parameters were set to exclude segregating sites and indels that violated an infinite-sites mutation model. We then mapped the physical positions of all putative recombination events in the cluster. Some of these events
will coincide with the boundaries of blocks identified using LD but others will not. Because compatibility methods do not provide a statistical method for evaluating the position of recombination breaks, the LD analyses were used to determine the exact block boundaries. In subsequent analyses, we used the BEAGLE program (Lyngsø et al. 2005) to compute the minimum number of recombinations \( R_{\text{min}} \) and to simultaneously reconstruct the history of all mutation and recombination events in the cluster.

Phylogenetic analyses

We first inferred an unrooted genealogy using unweighted parsimony and PAUP* 4.0 (Swofford 1998). This approach assumes no recombination and is expected to introduce homoplasy to accommodate phylogenetic incongruence in a recombinating data set. If LD and compatibility analyses revealed evidence of recombination, we further reconstructed a minimal ancestral recombination graph (ARG) using a novel branch and bound algorithm implemented in BEAGLE (Lyngsø et al. 2005). The branch and bound-based method in BEAGLE is more efficient in computing an exact lower bound on \( R_{\text{min}} \) compared to previous methods (Song & Hein 2003) and the reconstructed ARG in BEAGLE is displayed as a hierarchical-directed graph, which can be viewed using open-source visualization and graph layout programs such as GT MINER version 1.19 (Brown DE, Powell AJ, Carbone I, Dean RA, unpublished), a component of SNAP Workbench. The ARG represents our most parsimonious reconstruction of cluster haplotypes assuming recombination and can be rooted using the most recent common ancestor (MRCA) of the entire sample, inferred from coalescent simulations (described below). The rooted ARG provides inferences on the relative order of recombination events, the contributions of mutation and recombination in the evolution of haplotypes, and the ancestral relationships among noncontiguous recombination blocks in the cluster. Phylogenetic congruency between AF cluster recombination blocks and noncluster regions was assessed using site-compatibility matrices (Carbone et al. 2004). In the absence of recombination, multiple sequence alignments for noncluster regions were concatenated into one alignment to increase the number of variable sites for phylogenetic inference.

Balancing selection vs. spatial structure or population growth

Demographic effects can mimic the signature of balancing selection. In fact, simulation studies have shown that balancing selection and strong spatial structure have similar influences on diversity at neutral loci (Navarro & Barton 2002). This suggests that relying entirely on neutrality tests for distinguishing between balancing selection and population subdivision is tenuous. Indeed, an excess of intermediate-frequency haplotypes in the AF cluster may indicate population subdivision or balancing selection acting on specific haplotypes or a combination of these processes. To investigate spatial differentiation, we tested the null hypothesis of no association between \( A.\ parasiticus \) cluster haplotypes and soil sampling location using Hudson’s test statistics (Hudson et al. 1992) implemented in SNAP Workbench. In the absence of population subdivision, we examined genealogies for evidence of balancing selection. One effect of balancing selection is to increase the depth of genealogies, resulting in a genealogy with two or more haplotypes or clades separated by long branches (May et al. 1999). In contrast, the signature of population growth is a star-like genealogy and the retention of an excess of rare alleles. Since both may have occurred in the ancestral history of the sample, we first tested for evidence of constant vs. exponential growth rates and then estimated the time to the most recent common ancestor (TMRCA) of cluster vs. noncluster genealogies under the appropriate demographic model, as described below.

We reconstructed the underlying genealogy of each inferred recombination block in the cluster using a coalescent model with panmixia as implemented in GENETREE version 9.0 (http://www.stats.ox.ac.uk/~griff/software.html). The population mutation parameter, \( \theta \), for coalescent analyses was estimated from the number of segregating sites (Watterson 1975) using GENETREE. Coalescent simulations were used to calculate the relative probabilities of all rooted genealogies, as well as the unrooted genealogy, for each recombination block. The rooted genealogy with the highest probability was selected as our best inference of ancestral relationships in a specified block. Because population size changes can alter the mutation frequency spectrum and mimic selective sweeps, we tested the null hypothesis that population size is not growing exponentially. To do this, we first found the MLE of the mutation parameter, \( \theta \), conditional on the observed rooted genealogy and assuming constant population size. Second, we estimated surface likelihoods for \( \theta \) for a range of fixed exponential growth parameters \( \beta = 0.1, 0.5, 1.0, 2.0 \) and \( 3.0 \) to determine the MLE of \( \theta \) for each fixed \( \beta \) value. Third, we selected \( \theta \) with the highest likelihood for our range of fixed \( \beta \). If the growth rate was close to 0.1, the lowest value in our range, we used the corresponding \( \theta \) and \( \beta \) values to estimate the mean TMRCA under a coalescent model with population growth. If the growth rate was greater than 0.1, we fixed \( \theta \) and generated a likelihood surface in the vicinity of the growth value to estimate the joint MLE of \( \theta \) and \( \beta \). This was necessary because the two parameters are dependent on each other. Finally, we estimated the mean TMRCA and the ages of mutations on the rooted genealogy, inferred for each recombination block in the
cluster and for the neutral noncluster genealogy assuming either a constant-size or growth model. We performed likelihood ratio tests using the \texttt{r} open source statistical package (http://www.r-project.org/) to assess the goodness-of-fit of our observed data under the two demographic models. A TMRCA in the cluster that was approximately two times greater than the mean TMRCA for noncluster genes was indicative of extended residence times for neutral mutations and of balancing selection in the cluster.

Trans-specific evolution and balancing selection

Inferences of balancing selection in the cluster were further corroborated by examining intergenic regions of the AF gene cluster for several species in section \textit{Flavi} for evidence of trans-species polymorphisms. Our analysis included all isolates in the \textit{A. parasiticus} sample (n = 24) plus \textit{A. flavus} AF13 (L strain, AY510451), AF36 (L strain, AY510455; non aflatoxin-producing strain; Ehrlich \& Cotty 2004), AF70 (S strain, AY510453, produces B and G aflatoxins; Ehrlich \textit{et al.} 2005a), NRRL 3357 (L strain; http://www.aspergillusflavus.org/), and NRRL 1957 (ex type; L strain; non aflatoxinogenic); \textit{Aspergillus oryzae} ATCC 42149 (Machida \textit{et al.} 2005) and NRRL 447; and \textit{A. sojae} (IMI 191300). We also included one geographically isolated \textit{A. parasiticus} strain from peanuts in Uganda (SU-1, AY371490). The existence of shared ancestral polymorphisms and common haplotypes among these species is evidence of trans-species evolution in support of balancing selection acting on the ancestor of extant species. If shared haplotypes were observed, we used the coalescent to provide estimates on the relative and absolute timing of speciation and chemotype divergence. Coalescent time was converted into real time assuming a TMRCA for chemotype divergence. Coalescent time was converted on the relative and absolute timing of speciation and observed, we used the coalescent to provide estimates on the ancestor of extant species. If shared haplotypes were haplotypes among these species is evidence of trans-specific evolution. Our analysis included all isolates in the \textit{A. parasiticus} sample (n = 24) plus \textit{A. flavus} AF13 (L strain, AY510451), AF36 (L strain, AY510455; non aflatoxin-producing strain; Ehrlich \& Cotty 2004), AF70 (S strain, AY510453, produces B and G aflatoxins; Ehrlich \textit{et al.} 2005a), NRRL 3357 (L strain; http://www.aspergillusflavus.org/), and NRRL 1957 (ex type; L strain; non aflatoxinogenic); \textit{Aspergillus oryzae} ATCC 42149 (Machida \textit{et al.} 2005) and NRRL 447; and \textit{A. sojae} (IMI 191300). We also included one geographically isolated \textit{A. parasiticus} strain from peanuts in Uganda (SU-1, AY371490). The existence of shared ancestral polymorphisms and common haplotypes among these species is evidence of trans-specific evolution in support of balancing selection acting on the ancestor of extant species. If shared haplotypes were observed, we used the coalescent to provide estimates on the relative and absolute timing of speciation and chemotype divergence. Coalescent time was converted into real time assuming a TMRCA for \textit{A. parasiticus}, \textit{A. flavus}, and \textit{A. oryzae} between 8 and 17 million years ago (Ehrlich \textit{et al.} 2005b).

Chemotype-specific adaptation

We calculated mean Ka/Ks for three distinct chemotype classes in \textit{A. parasiticus} (OMST, G$_i$/B$_i$ ratios > 5.0, G$_i$/B$_i$ ratios of 0.3–2.0) using \texttt{dnasp} (Rozas \textit{et al.} 2003). Selective processes acting on chemotypes can be either frequency-dependent, due to overdominance, or the result of a recent adaptation. Overdominant selection, a process where heterozygotes are more fit than homozygotes, has been implicated in the evolution of the major histocompatibility complex (Hughes \& Nei 1989), but is unlikely to apply here as \textit{A. parasiticus} is haploid. We tested whether there was a difference in mean Ka/Ks values between the three distinct chemotype classes in \textit{A. parasiticus} using a linear model to account for variation between genes. This model can be written as Ka/Ks = mean of all Ka/Ks values + gene effect + chemotype effect + error. We tested the null hypothesis that there is no difference in mean Ka/Ks between chemotypes by computing and testing chemotype contrasts. For example, a contrast of the form c(–1, 2, –1), where the chemotype order is G$_i$/B$_i$ = 0.3–2.0, and OMST, would compare the mean Ka/Ks values of G$_i$ dominant and OMST with the mean Ka/Ks of G$_i$/B$_i$ ratios of 0.3–2.0. Contrasts were computed using the fit.contrast function implemented by Gregory R. Warnes in the gmodels package in \texttt{r} (Venables \& Ripley 2002). The function returns a matrix containing the estimated regression coefficients, standard errors, t values and two-sided P values. A significant test result may indicate a difference in selective constraints or adaptive evolution between chemotype classes.

Results

Nucleotide diversity and neutrality tests

We sequenced 8807 bp in 21 intergenic regions of the AF gene cluster from all 24 strains in our \textit{Aspergillus parasiticus} subsample. Additional sequencing was from noncluster regions \texttt{trpC} (402 bp), \texttt{amds} (331 bp), \texttt{benA} (324 bp), and \texttt{ITS-1} (173 bp). Multiple DNA sequence alignments for each region have been deposited in GenBank under Accession nos DQ390585–DQ391184. Collapsing a concatenated multiple sequence alignment of our 21 cluster regions yielded 10 distinct haplotypes, abbreviated as H1 to H10 (Table 2; Fig. 1). The chemotype and VCG designation of all haplotypes are also shown in Table 2. Haplotypes H4 and H5 are OMST-specific haplotypes and contained the three OMST-accumulating strains (IC65, IC71, IC76), each belonging to a different VCG. Haplotypes H2 and H3 are G$_i$-dominant haplotypes and included the strain (IC73) with the highest G$_i$/B$_i$ ratio (5.1). Haplotypes H12 and H10 are G$_i$-dominant haplotypes and included the strain (IC43) with the lowest G$_i$/B$_i$ ratio (0.1) in the sample. All other haplotypes had G$_i$/B$_i$ ratios in the range of 0.3–2.0 (Table 2). There were four distinct haplotypes in \texttt{trpC} and four in \texttt{amds}. There was no variation in \texttt{benA} and \texttt{ITS-1}.

Nucleotide diversity estimates per site ranged from 0.00045 to 0.01622 in cluster regions, 0.00114 in \texttt{trpC}, and 0.01214 in \texttt{amds}. No statistical test of neutrality calculated for cluster and noncluster regions was significant after Bonferroni correction for multiple tests. Although we were unable to detect any significant departure from neutrality, D, D* and F* were predominantly negative (54 out of 63 tests) and F$_G$ was split between positive (11 out of 21 tests) and negative values across all cluster regions, indicating the existence of both rare and intermediate-frequency haplotypes. Based on these results, we cannot completely rule out population subdivision, balancing selection or population growth.
Recombination and compatibility

LD analysis was based on 161 base substitutions and 10 recoded indel polymorphisms spanning the 21 sequenced cluster regions. Polymorphisms were not evenly distributed across the cluster and approximately 90% of the variation was found downstream of aflB (Fig. 1). We identified five distinct recombination blocks; with the exception of one or two sites within each block, there was evidence of tight linkage, with $0.8 < r^2 < 1$ and $P < 0.01$ for most parsimony-informative sites. Block 1 is in the intergenic region separating the aflA/aflB and aflR/aflS gene modules (Carbone et al. 2007). The location and size of the other blocks are shown in Fig. 1. The overall population mean recombination rate, $\gamma$, was estimated to be $2Nc = 9$, which corresponds to a recombination rate of 0.0011 per generation per base pair. This estimate is inflated in regions upstream of aflB and overestimated in other regions, as recombination events are not evenly distributed across the entire cluster (Fig. 1). 

Phylogenetic analyses

Phylogenetic inference was based on 78 parsimony-informative and 93 parsimony-uninformative sites. An unweighted parsimonious analysis resulted in four equally parsimonious trees, each with a consistency index of 0.8680. Phylogenetic analyses using Fisher’s exact test whereas the lower triangular matrix represents the $P$ values calculated using Fisher’s exact test whereas the lower triangular matrix represents $r^2$, the coefficient of determination between the allelic states at pairs of sites. Coloured shading indicates statistical significance in linkage and strength of associations. The numbered boxes outline five blocks of segregating sites that are highly correlated ($0.8 < r^2 < 1$) and significantly linked ($P < 0.01$). The physical location of blocks with respect to gene modules in the cluster is shown above and below the plot. The distribution of the 171 polymorphic sites (SNPs) among the 10 distinct haplotypes is shown below the LD plot. Variable sites are characterized as transitions (t) or transversions (v), and as informative (i) or uninformative (–). Polymorphisms that match the consensus are indicated with a dot in the SNP map. Uninformative sites interrupt tightly linked sites in each block and are shown to identify highly conserved regions in the cluster. For example, all sites in the aflM/aflN region, which correspond to the hypE gene, are uninformative and H2 is a strongly differentiated allele in hypE. See Fig. 3 for the signature of balancing selection in hypE. Based on the ARG in Fig. 2B, blocks 1 and 4, which are noncontiguous in the cluster, share the same evolutionary history (shaded in grey). Similarly, blocks 3 and 5 share a common ancestor and are noncontiguous (shaded light green). The physical positions of the recombination breaks in the cluster, inferred using the beagle (B) and recmin (R) programs, are shown with vertical lines below the matrix. Blocks with asterisks include a portion of an exon or the block spans entire exons. Specifically, block 2 includes 63 bp of aflH and 456 bp of aflJ; block 3 includes 59 bp of aflE; and block 5 spans the aflK and aflV genes.

The ARG (Fig. 2B) was rooted by specifying H2 as the outgroup in beagle. The ARG shows the relative order of mutation and recombination events in the cluster and therefore captures the history of recombination blocks. For example, the recombination events that define blocks 1 and 4 at real positions 24 824 (site 13), 31 700 (site 29) and 48 206 (site 113) share a common history and can be concatenated such that block 1 is the prefix (P) of block 4, and block 4 is the suffix (S) of block 1 as shown in the ARG (shaded in grey in Figs 1 and 2B). This indicates that although these blocks are noncontiguous, they were at some point in their evolutionary history contiguous in an ancestral haplotype. Similarly, blocks 3 and 5 at real positions 39 602 (site 69), 41 327 (site 73) and 53 727 (site 127) share a common history and can be concatenated into a single contiguous block (shaded in green in Figs 1 and 2B).

The ARG (Fig. 2B) reveals extensive recombination in the history of the sample that can be traced back in time to either H2 or H4. Haplotypes H2 and H4 showed the deepest coalescence and excluding H2 and H4 eliminated our ability to detect recombination in this population sample. This translates to distinct chemotype classes G1, dominant (H2) and OMST (H4) as having undergone recombination in their evolutionary histories. The OMST lineage (H4) appears to have preferentially recombined with G1–dominant alleles (H2) in blocks 3 and 5; recombination events in other regions of the cluster are possible but will remain undetected if the parental sequences are very similar or identical. For example, haplotype H2 shows recent recombination with H7 and the recombinant subsequently gives rise to H1, an intermediate toxin-producing lineage (G1/B1 = 1.4, see Table 2). Such recombination events are more easily detected because one of the parent alleles (H2) is highly differentiated from the other (H7). Although all haplotypes share a history of recombination, mutation alone distinguishes haplotypes H7, H8 and H10 (Fig. 2B). Based on its position at the bottom of the ARG, haplotype

![Fig. 1](image-url) LD plot for 171 variable sites in the aflatoxin gene cluster of Aspergillus parasiticus (includes informative and uninformative sites). The numbers along the side and bottom of the plot are variable sites scaled to physical distance in units of base pairs. Gene modules are color coded in the diagram of the aflatoxin pathway (Carbone et al. 2007). The upper triangular matrix represents the $P$ values calculated using Fisher’s exact test whereas the lower triangular matrix represents $r^2$, the coefficient of determination between the allelic states at pairs of sites. Coloured shading indicates statistical significance in linkage and strength of associations. The numbered boxes outline five blocks of segregating sites that are highly correlated ($0.8 < r^2 < 1$) and significantly linked ($P < 0.01$). The physical location of blocks with respect to gene modules in the cluster is shown above and below the plot. The distribution of the 171 polymorphic sites (SNPs) among the 10 distinct haplotypes is shown below the LD plot. Variable sites are characterized as transitions (t) or transversions (v), and as informative (i) or uninformative (–). Polymorphisms that match the consensus are indicated with a dot in the SNP map. Uninformative sites interrupt tightly linked sites in each block and are shown to identify highly conserved regions in the cluster. For example, all sites in the aflM/aflN region, which correspond to the hypE gene, are uninformative and H2 is a strongly differentiated allele in hypE. See Fig. 3 for the signature of balancing selection in hypE. Based on the ARG in Fig. 2B, blocks 1 and 4, which are noncontiguous in the cluster, share the same evolutionary history (shaded in grey). Similarly, blocks 3 and 5 share a common ancestor and are noncontiguous (shaded light green). The physical positions of the recombination breaks in the cluster, inferred using the beagle (B) and recmin (R) programs, are shown with vertical lines below the matrix. Blocks with asterisks include a portion of an exon or the block spans entire exons. Specifically, block 2 includes 63 bp of aflH and 456 bp of aflJ; block 3 includes 59 bp of aflE; and block 5 spans the aflK and aflV genes.
H7 is the most recently evolved. Haplotype 7 is the most frequently represented in the sample and includes all strains from VCG 1 sampled along the transect.

**Balancing selection vs. spatial structure or population growth**

There was no evidence of spatial structure using Hudson’s tests that could account for the observed genetic differentiation between chemotype-specific strains across the entire cluster. To better understand the impact of population growth, if any, on cluster evolution, we performed coalescent simulations on rooted genealogies for cluster and noncluster regions, assuming either a constant-size or exponential-growth model. These simulations were performed on compatible partitions in the data. ARG analysis of the cluster region revealed three distinct partitions: blocks 1 and 4, block 2, and blocks 3 and 5. The noncluster regions trpC and amdS also are compatible and considered a single partition (data not

---

**Fig. 2** (A) Phylogenetic inference based on parsimony assuming no recombination. The numbers along the edges are for the number of mutations separating haplotypes (red ellipses). Coalescent nodes are shown in green. (B) Phylogenetic network inference. The ancestral recombination graph (ARG) was rooted using haplotype H2 as the most recent common ancestor of the sample sequences. In the ARG, the colour-coding scheme is a yellow dot (the highest point in the ARG) for the inferred ancestral (root) sequence and blue ellipses for recombination nodes. The recombinant consists of a prefix 'P' sequence concatenated with a suffix 'S' sequence. These designations appear on edges going into a recombination node. The numbers in the blue ellipses indicate the physical locations of recombination breakpoints in the cluster. For example, ‘31 700-’ means that a recombination breakpoint occurs between real positions 31 700 and 31 701 in the cluster. The recombination breaks that are shaded using the same colour (grey or light green) share a common evolutionary origin and coincide with the shaded blocks in Fig. 1.
shown). We also considered the hypE region in the cluster as a distinct partition because it showed evidence of trans-species evolution (described below). We used the MLE of θ for each partition to estimate the TMRCA under both demographic models. In all cases, we cannot reject the simpler constant-size demographic model (Table 3, P > 0.1, d.f. = 1). TMRCAs for each region were estimated using the conditional MLE of θ assuming a constant-size model in genetree. The ratio of the TMRCAs for each cluster partition to the trpC/amdS noncluster partition was 1.29, 1.11, 2.11, and 1.62 for the blocks 1 and 4, block 2, hypE, and blocks 3 and 5, respectively (Table 3). The estimated TMRCAs for blocks 3 and 5 and hypE are approximately twice that of the noncluster trpC/amdS, indicative of ancient polymorphisms in the cluster predating divergence in noncluster genes.

Trans-specific evolution and balancing selection

To further examine the possibility of trans-species polymorphisms in the AF gene cluster, we added sequences for closely related species within Aspergillus section Flavi and then recollapsed the alignments. Only the aflM/aflN intergenic region showed evidence of shared ancestral polymorphisms and haplotypes. This region corresponds to a putative hypothetical protein (hypE) identified from gene expression studies (G.A. Payne, unpublished). We inferred the rooted genealogy for hypE to examine more closely the ancestral relationships within section Flavi (Fig. 3). In the hypE genealogy, Aspergillus flavus (NRRL 1957) and Aspergillus oryzae (NRRL 447, ATCC 42149) share a common haplotype (H6) and a MRCA with two other A. flavus 1-strain haplotypes, H2 (AF36) and H3 (AF13). The A. parasiticus G1-dominant haplotype, H5 (IC73), shares a MRCA with H2, H3, and H6. In a separate and distinct clade, A. parasiticus (SU-1) and A. flavus (NRRL 3357) share a common haplotype (H1) and a MRCA with an A. parasiticus OMST-specific haplotype, H9 (IC65, IC71, IC76), A. sojae, H7 (IMI191300), and A. parasiticus, H8 (IC67), H10 (IC77, IC75, IC74, IC72, IC69, IC115, IC11, IC59, IC61, IC43, IC56, IC15, IC123, IC2, IC38, IC25, IC78, IC98), and H11 (IC70). Although our strongest evidence of trans-species evolution is in hypE, our observation of fixed polymorphisms and chemotype-specific differences across the entire cluster is consistent with a single-locus model of balancing selection acting to maintain chemotype differences. Our estimates of divergence times indicate that balancing selection in this species group is ancient (4.8 million years ago) and ongoing on a more recent timescale (1.2–2 million years ago). The A. flavus S-strain haplotype, H4 (AY510453), which produces B and G aflatoxins, is highly divergent and shares a distant common ancestor with the ingroup section Flavi species (8 million years ago).

Table 3 Population parameter estimates and statistics for constant size and exponential growth models

<table>
<thead>
<tr>
<th>Partition</th>
<th>π*</th>
<th>θ†</th>
<th>TMRCA (SD)</th>
<th>Likelihood (SE)</th>
<th>LR‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blocks 1 + 4</td>
<td>0.00297</td>
<td>4.00</td>
<td>3.77 (0.82)</td>
<td>3.30 × 10⁻¹² (1.30 × 10⁻¹⁴)</td>
<td>0.60</td>
</tr>
<tr>
<td>Block 2</td>
<td>0.00535</td>
<td>5.82</td>
<td>3.24 (0.59)</td>
<td>1.25 × 10⁻¹⁵ (1.90 × 10⁻¹⁷)</td>
<td>0.68</td>
</tr>
<tr>
<td>hypE¶</td>
<td>0.00125</td>
<td>6.38</td>
<td>6.17 (0.89)</td>
<td>7.05 × 10⁻¹¹ (2.22 × 10⁻¹²)</td>
<td>1.96</td>
</tr>
<tr>
<td>Blocks 3 + 5</td>
<td>0.00437</td>
<td>6.03</td>
<td>4.72 (0.78)</td>
<td>4.56 × 10⁻¹⁵ (8.39 × 10⁻¹⁷)</td>
<td>1.25</td>
</tr>
<tr>
<td>trpC¶ + amdS**</td>
<td>0.00610</td>
<td>2.85</td>
<td>2.92 (0.80)</td>
<td>8.67 × 10⁻⁰⁹ (4.60 × 10⁻¹¹)</td>
<td>0.19</td>
</tr>
</tbody>
</table>

*Average number of pairwise nucleotide differences per site for each partition; †MLE of θ assuming a constant-size model (upper row for each partition) or an exponential growth model (lower row) where β, the exponential growth rate per gene per generation, was fixed at 0.1; TMRCA, time to the most recent common ancestor; SD, standard deviation; SE, standard error; ¶LR is the likelihood ratio of the difference between constant and exponential growth demographic models. The LR test is not significant (P > 0.1) for all partitions; ¶hypE is a hypothetical protein-encoding region (414 bp) located in the aflM/aflN intergenic region; ¶trpC consists of 35 bp and 367 bp for coding and noncoding sequences, respectively; **amdS consists of 283 bp and 48 bp for coding and noncoding sequences, respectively.
representative portions of genes spanning the entire cluster: \textit{aflA}, \textit{aflR} (adjacent to block 1), \textit{aflJ} (within block 2), \textit{aflM} (adjacent to block 3), \textit{hypE} (region with evidence of ancient balanced polymorphisms), \textit{aflO} (between blocks 4 and 5), \textit{aflV} (within block 5), and one noncluster gene (\textit{amdS}). We designated three chemotype classes: OMST, \textit{G}_1/B_1 > 5 (\textit{G}_1 \text{ dominant}), and \textit{G}_1/B_1 ratios in the range of 0.3–2.0 (\textit{G}_1/B_1). With the exception of \textit{aflJ} and \textit{amdS} with mean $Ka/Ks$ values of 0.2 and 0.1, respectively, all the sequences in the \textit{G}_1/B_1 class are very similar or identical ($Ka/Ks = 0$) (Fig. 4). Although \textit{G}_1-dominant and OMST classes have mean $Ka/Ks > 0$, we detected no differences in mean $Ka/Ks$ between \textit{G}_1-dominant and \textit{G}_1/B_1 classes ($P > 0.1$). Overall, only mean $Ka/Ks$ values for OMST were different from \textit{G}_1/B_1 values ($P < 0.05$), suggesting a possible selective sweep in the evolution of OMST gene.

Fig. 3 The inferred rooted \textit{hypE} genealogy for closely related species in section \textit{Flavi} based on maximum-likelihood estimates of all possible rooted trees using \textsc{genetree} (likelihood of best tree = $1.1317 \times 10^{-22}$, SD = $1.1254 \times 10^{-19}$, $\theta = 20.0$). A total of 10 million coalescent simulations and five independent runs assuming a constant population size model were performed to ensure convergence. Each mutation along an edge in the tree is indicated with a dot and a number. The timescale for major bifurcations in the tree was generated by converting coalescent time into real time assuming a TMRCA for \textit{Aspergillus parasiticus}, \textit{A. flavus}, \textit{A. oryzae} between 8 and 17 million years ago (Ehrlich et al. 2005b). The upper range of real time estimates is shown in parentheses. The matrix below the tree shows the distribution of haplotypes among species.

\begin{table}
\centering
\begin{tabular}{cccccccccc}
H4 & H5 & H2 & H3 & H6 & H1 & H7 & H8 & H9 & H10 & H11 \\
\hline
1 & . & . & . & . & . & . & . & . & . & . \\
. & . & . & 1 & . & 1 & . & 18 & 1 & . & . \\
. & . & 1 & 1 & 1 & . & . & . & . & . & . \\
1 & . & . & . & . & 2 & . & . & . & . & . \\
\hline
\end{tabular}
\end{table}

\begin{itemize}
\item A. \textit{parasiticus} (G, dominant)
\item A. \textit{parasiticus} (OMST)
\item A. \textit{sojae}
\item A. \textit{flavus} L
\item A. \textit{flavus} S
\item A. \textit{oryzae}
\end{itemize}

% \textcopyright 2007 The Authors
\textit{Journal compilation} \textcopyright 2007 Blackwell Publishing Ltd
be greater between than within species, but we did not observe this. As shown in Fig. 3, the G<sub>1</sub>-dominant haplotype shares a recent common ancestor with Aspergillus flavus as opposed to a distant common ancestor with other A. parasiticus haplotypes, a phenomenon entirely consistent with trans-species evolution. A similar phenomenon has been observed in population studies of the AF gene cluster of A. flavus (G. Moore, unpublished data).

Recombination and phylogenetics

Our LD analysis revealed five distinct recombination blocks that when analysed simultaneously, showed a complex mutation–recombination history that could not be accurately captured by mutation alone. Comparison of the topology of one most parsimonious tree with the ARG in Fig. 2 shows how recombination introduces many mutational differences between haplotypes that otherwise share a very recent common ancestor. Highly divergent haplotypes sharing a distant common ancestor in the cluster may appear to share a recent common ancestor in some blocks due to recombination. For example, only four recombination events at real positions 24 824-, 31 700-, 48 206-, and 50 331- in the cluster separate haplotypes H1 and H7 in the ARG, whereas 16 polymorphisms separate these haplotypes in the parsimony tree. Indeed, if recombination is ignored, all polymorphisms within the block 1 are counted as independent mutational events. Overall, the joint effects of recombination and balancing selection may introduce significant phylogenetic errors in branch lengths and topology (Fig. 2). Similar signatures of recombination and balancing selection have been detected in other gene clusters. For example, in the trichothecene gene cluster of Fusarium, Ward et al. (2002) found that recombination and balancing selection did indeed render the trichothecene genes very misleading with respect to phylogeny. The present study suggests that recombination block analysis of the trichothecene gene cluster may reveal additional insights into the evolutionary processes underlying this important gene cluster.

Estimating the relative ages of recombination events in the AF cluster of A. parasiticus using coalescent simulation methods was challenging given the size and complexity of the data set. Instead, we examined the relative order of recombination events by inferring a rooted ARG and we estimated divergence times from a coalescent analysis of hypE, the cluster region with the longest history of neutral evolution. These analyses indicate that recombination in the cluster has been ongoing for approximately 4.8 million years (Fig. 3). This time estimate corresponds to the divergence of the G<sub>1</sub>-dominant haplotype from OMST and the other A. parasiticus haplotypes in Fig. 3. Although mutation alone separates H7, H8, and H10 in the ARG (Fig. 2B), it is important to note that with the exception of

clusters. Further evidence in support of a selective sweep are significant differences in mean Ka/Ks values between OMST and G<sub>1</sub>-dominant haplotypes (P < 0.001) and a reduced amount of neutral variation in OMST gene clusters (Ramirez-Prado, unpublished data).

Discussion

Our population genetic analyses of molecular variation in the AF gene cluster of Aspergillus parasiticus indicate a complex interplay of recombination, balancing selection and positive selection acting to maintain OMST and G<sub>1</sub>-dominant haplotypes. These analyses revealed extensive LD organized in distinct recombination blocks and necessitated that we look for the signature balancing selection in each distinct block, as well as in noncluster regions. Our coalescent analyses of the TMRCA in cluster vs. noncluster regions showed a twofold difference in depth that is indicative of ancient balancing selection. Balancing selection was also corroborated by the existence of trans-species haplotypes among geographically isolated strains of A. parasiticus and among closely related species. Trans-species analyses in particular provided strong evidence for balancing selection in the cluster. Although we do not have estimates of the expected variance in the TMRCA for neutral loci throughout the genome of A. parasiticus, a twofold greater depth on average for cluster genes is not simply an artifact of ‘noise’ associated with variation in neutral coalescent depth. Under a neutral model, we would expect mean estimates of the TMRCA to
the recombination events in blocks 1 and 4, haplotype H7 is identical to H1. Indeed, recombination may be occurring on a contemporary timescale. Among A. parasiticus strains in the present study, H7 was the most abundant haplotype in the Georgia peanut field (Horn & Greene 1995) and was widely distributed in peanut fields along a transect through the southern United States (Horn & Dorner 1998). The recent expansion of H7 and the strong clonal component to A. parasiticus population biology may be the result of local increases in effective population size with the advent of agriculture. Larger effective population sizes tend to increase mean population mutation and recombination rates (Hartl & Clark 1997), further driving the evolution of novel VCGs. The exact mechanism of recombination is not known, but may include sexual and/or parasexual reproduction. Preliminary data indicate the presence of both mating-type idiomorphs in the A. parasiticus population examined in this study (Ramirez-Prado, unpublished data). This suggests that sexual reproduction could be an underlying mechanism for the observed recombination in the cluster.

Balancing selection in the evolution of the AF gene cluster
As there was no evidence of spatial structure or population growth within the sampled field, we hypothesize that the strong genetic differentiation between OMST and G1-dominant haplotypes is the result of balancing selection. Based on the assumption of constant population size, estimates of the TMRCA in cluster regions appear to be approximately twice that of noncluster regions amdS and trpC. The twofold deeper coalescence for hypE provides evidence for balancing selection in the A. parasiticus cluster, whereas a similar TMRCA for the blocks 1, 2 and 4 compared to noncluster genes suggests that genes in these blocks may have been acquired recently, possibly via recruitment from other genomic locations or recent reorganization of genes from an ancestral cluster. One intriguing idea is that gene modules (e.g. aflG/aflL) that are contiguous in one species (e.g. A. flavus) and noncontiguous in others (e.g. Aspergillus nidulans) are the result of cluster rearrangements in an ancestral ST species (Carbone et al. 2007). There is some evidence in support of this from the compatibility relationships among noncontiguous blocks in this study. For example, if we splice block 4 and unite with block 1, then aflB is adjacent to aflE, as observed in the A. nidulans ST cluster. Similarly, if we swap block 5 with block 3, then aflE/aflM is adjacent to aflW as in A. nidulans.

In addition to cluster reorganization from an ST ancestor, the magnitude of recombination and selection on cluster genes may also influence cluster stability. This is supported by a recent study reporting partial deletions in the AF cluster of nonaflatoxicogenic strains of A. flavus from nature (Chang et al. 2005). The location of deletion breakpoints in strains with partial deletion patterns is similar to the boundaries of recombination blocks in A. parasiticus. The deletion patterns suggest that regions in the cluster under strong balancing selection, such as hypE and blocks 3 and 5, are retained. Interestingly, there were no genes corresponding to blocks 1 and 2 in any of the strains with partial deletions. One possible explanation is that the missing genes were progressively lost from a parental strain with a full gene cluster as reported in Aspergillus oryzae (Kusumoto et al. 2000). Alternatively, the partial-deletion strains may represent the ancestral state, and full clusters arose from the acquisition of additional genes via recruitment of gene modules from other genomic locations or through genetic exchange and recombination. This cannot be ascertained without performing similar population studies in A. flavus.

Balancing selection on specific chemotypes also results in trans-species polymorphisms, a phenomenon whereby the number of shared polymorphisms is greater between species than within a single species (Takahata 1990; Ward et al. 2002). This process can explain the conservation of chemotype classes and polymorphisms among different Aspergillus species. Two striking examples in the AF cluster are the sharing of a common haplotype in hypE between A. parasiticus and A. flavus (H1) and a shared MRCA between a G1-dominant A. parasiticus haplotype (H5) and A. flavus/A. oryzae haplotypes (H2, H3, H6). The presence of trans-specific haplotypes indicates that the signatures of balancing selection in A. parasiticus (two divergent chemotype lineages) may also be present in other species in Aspergillus section Flavi. Indeed, two distinct evolutionary lineages have been reported in A. flavus (Geiser et al. 1998) and Aspergillus nomius (Ehrlich et al. 2003). In A. flavus, Geiser et al. (1998) reported two highly divergent lineages defined as groups I and II. A disproportionately large number of fixed polymorphisms between groups I and II were from the omi12 region which resides within aflP in the present study. Geiser and coworkers showed that the omi12 region had a high nucleotide diversity compared to housekeeping genes, consistent with our observations of longer residence times of chemotype-specific alleles in cluster genealogies relative to the combined trpC and amdS gene genealogy. Although a large number of fixed polymorphisms may be interpreted as reproductive isolation (Geiser et al. 1998), our examination of variation across the entire AF cluster in A. parasiticus populations provides evidence of recombination between highly divergent chemotype lineages. The possibility of distinct chemotype lineages underlying divergence and speciation in Aspergillus needs to be explored further. Indeed, phylogenies that are based on one or more loci harbouring trans-species polymorphisms may show ancestral relationships that confound species delimitations.
Purifying selection, gene modules and adaptive evolution

The functional constraints on aflR, the AF pathway-specific transcriptional regulator, are obvious. However, the reason for the strong functional constraints on other genes is unclear, but such constraints may be important in regulating OMST and the C1/B1 ratio. In Aspergillus parasiticus, there appears to be an evolutionary balance in gene modules, such that one gene is functionally constrained by purifying selection and another gene is enhanced by positive Darwinian selection. This is consistent with a mechanism of gene duplication followed by neofunctionalization in which mutations in one gene enhance the activity of the other gene in the module (Rodriguez-Trelles et al. 2003). For example, aflS is known to interact with aflR to modulate the transcription of certain AF pathway genes (Yu et al. 2004b). Although subfunctionalization appears to be a more plausible mechanism for aflA/aflB (Hitchman et al. 2001), the fixation of advantageous mutations in aflA of C1-dominant strains is the signature of adaptive processes. While both neo- and sub-functionalization may be determining the evolutionary fate of gene modules, the mechanism of selection in A. parasiticus that maintains a high fraction of polymorphisms in LD for two low frequency chemotype variants is not known. Are different ecological niches selecting for specific chemotypes, or is recombination simply a means of spreading beneficial alleles (in terms of aflatoxicity) or removing deleterious alleles from the population? Selection against deleterious alleles can result in extensive LD such that a favoured chemotype is swept to fixation. Selective sweeps in the cluster may be important in maintaining multiple functional polymorphisms across several intergenic regions and exons. Indeed, the recombination observed between OMST (H4) and C1-dominant strains (H2) suggests that selection should be highly effective in maintaining these chemotypes.

The overall strong purifying selection acting on the AF gene cluster, coupled with the considerable expenditure of energy needed to synthesize the major pathway extrolites, suggests that these compounds are advantageous to the fungus in some manner. Several theories on the role of aflatoxins in nature have been proposed. The toxicity of these extrolites toward arthropods (Al-Adil et al. 1972; Moore et al. 1978; Jarvis et al. 1984) indicates that aflatoxins may have originally evolved to parasitize insects or to protect the fungus against insect feeding. Alternatively, the link between AF/ST production and reproductive structures, such as conidia and sclerotia, suggests a possible role of these toxins in fungal developmental processes (Wilkinson et al. 2004). Another theory proposes that these toxins may be important in alleviating oxidative stress (Kim et al. 2005). The toxicity of aflatoxins to maturing seeds and grain has not been examined and deserves further investigation. It is clear that we do not fully understand the role of these toxins in nature. Further population sampling and elucidation of evolutionary processes in the AF gene cluster of different species in section Flavi may shed light on the role of aflatoxins in ecological adaptation and speciation.

Conclusion

Population-level molecular sequence variation in the AF cluster of Aspergillus parasiticus allowed us to reconstruct the history of mutation and recombination in the cluster. Our analyses revealed five distinct recombination blocks, as well as chemotype-specific selection across the entire A. parasiticus cluster. This study clearly shows that population studies are critical in testing hypotheses on the evolution of AF gene clusters and for identifying the strains that should be examined more closely to elucidate the specific mechanisms regulating and maintaining aflatoxicity. Whether the strength of positive selection on specific genes in the cluster is directly responsible for adjusting the C1/B1 ratio in nature is unknown and will require further sampling of this and other evolutionarily distinct chemotype classes.

Acknowledgements

We thank Deepak Bhatnagar for providing us with the aflatoxin gene cluster sequence for Aspergillus parasiticus prior to its publication, Julie E. Cox for technical help, Doug Brown for bioinformatics support, and David Aylor (Bioinformatics Research Center, NC State University) for help in developing the correlation tests and linear models in r. This work was supported in part by the University of North Carolina General Administration under an award for High Performance Computing (HPC) and Computational Sciences. Coalescent simulations were performed, in part, on HPC resources provided by the NC State Information Technology Division with support from the Office of the Provost and Office of Research and Graduate Studies. This work was funded by the North Carolina Cooperative State Research, Education, and Extension Service, grant nos 2004-35400-14429, 2005-34500-15893, 2006-35604-16666, and by the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service, grant no. 2005-35319-16126 to I. C.

References


Ignazio Carbone is an Assistant Professor interested in the ecology and evolution of fungal pathogens on both macro and microevolutionary timescales. His main focus is the evolutionary genetics of aflatoxin biosynthesis in \textit{Aspergillus}. He is also developing software tools to facilitate and enhance our understanding of the underlying evolutionary processes shaping diversity in pathogen populations. Judy Jakobek is a Research Analyst for the Carbone laboratory interested in fungal molecular population genetics.

Jorge Ramirez-Prado is a Postdoctoral Researcher exploring the population genetics of mating system evolution in \textit{Aspergillus}. Bruce Horn is a Microbiologist interested in the ecology and population biology of aflatoxigenic fungi.