Evidence for balancing selection operating at the *het*-c heterokaryon incompatibility locus in a group of filamentous fungi

JENNIFER WU, SVEN J. SAUPE*, AND N. LOUISE GLASS†

Biotechnology Laboratory and Botany Department, University of British Columbia, Vancouver, BC V6T 1Z3, Canada

Communicated by David D. Perkins, Stanford University, Stanford, CA, August 5, 1998 (received for review May 15, 1998)

ABSTRACT In filamentous fungi, *het* loci (for *het*erokaryon incompatibility) are believed to regulate self/nonself-recognition during vegetative growth. As filamentous fungi grow, hyphal fusion occurs within an individual colony to form a network. Hyphal fusion can occur also between different individuals to form a heterokaryon, in which genetically distinct nuclei occupy a common cytoplasm. However, heterokaryotic cells are viable only if the individuals involved have identical alleles at all *het* loci. One *het* locus, *het-c*, has been characterized at the molecular level in *Neurospora crassa* and encodes a glycine-rich protein. In an effort to understand the role of this locus in filamentous fungi, we chose to study its evolution by analyzing *het-c* sequence variability in species within *Neurospora* and related genera. We determined that the *het-c* locus was polymorphic in a field population of *N. crassa* with close to equal frequency of each of the three allelic types. Different species and even genera within the Sordariaceae shared *het-c* polymorphisms, indicating that these polymorphisms originated in an ancestral species. Finally, an analysis of the *het-c* specificity region shows a high occurrence of nonsynonymous substitution. The persistence of allelic lineages, the nearly equal allelic distribution within populations, and the high frequency of nonsynonymous substitutions in the *het-c* specificity region suggest that balancing selection has operated to maintain allelic diversity at *het-c*. *Het-c* shares this particular evolutionary characteristic of departing from neutrality with other self/nonself-recognition systems such as major histocompatibility complex loci in mammals and the *S* (self-incompatibility) locus in angiosperms.

Self/nonself-recognition is essential for sexual reproduction, defense against pathogen invasion, and maintenance of individuality for an organism. In vertebrates, self/nonself-recognition relies on the major histocompatibility complex (MHC), which is an array of genetic loci that generate proteins important in pathogen recognition and activation of defense mechanisms. The human MHC consists of over 100 highly polymorphic genes that fall into two classes, MHC I and II. Allelic polymorphisms at the MHC II *DRB1* locus, represented by 58 alleles in humans, have existed for at least 30 million years and are shared by humans, apes, and other primates (1). Self/nonself-recognition during sexual reproduction in many plant species is mediated by the gametophytic or sporophytic self-incompatibility locus, *S*, which elicits recognition and rejection of self-pollen (2). Similar to alleles at the *DRB1* locus, different alleles at the *S* locus show long-term persistence, such that an allele from one species is often more closely related to an allele in a different species, rather than to another *S* allele in the same species (3, 4). Thus, alleles have been passed from ancestral to descendent species and are shared among contemporary species, a phenomenon referred to as "trans-species

polymorphism" (5). Allele polymorphisms at loci in the MHC and *S* locus are thought to be maintained by balancing selection, either by overdominance (heterozygotes have a higher fitness than homozygotes) or by frequency dependent selection (rare alleles are at a selective advantage, but become disadvantageous when common) (4, 6, 7).

In filamentous fungi, the requirement for and biological significance of self/nonself-recognition during vegetative growth is unclear. During vegetative growth, filamentous fungi possess the remarkable feature of being able to undergo hyphal fusion between different individuals to form vegetative heterokaryons that contain genetically distinct nuclei within a common cytoplasm. The viability of such heterokaryons is determined by the genetic constitution at *het* (*het*erokaryon) or *vic* (*v*egetative *inc*ompatibility) loci (8, 9, 10). The phenomenon of heterokaryon (vegetative, heterogenic, or somatic) incompatibility has been described in numerous filamentous ascomycetes and basidiomycetes (9, 10, 11). Nuclear components of a heterokaryon that contain different specificities at any of several *het* loci can give rise to unstable nuclear ratios or can result in growth inhibition of heterokaryotic cells, which are frequently destroyed by a lytic process (8, 9, 12, 13).

Heterokaryon incompatibility in filamentous fungi may confer selective advantages by preventing hyphal fusions that could spread mycoviruses, debilitated organelles, and deleterious plasmids throughout a fungal population (14). It also has been proposed that *het* genes may play a role in limiting outbreeding in certain fungal species (11). However, it is unclear whether heterokaryon incompatibility loci are bona fide self/nonself-recognition systems in filamentous fungi or whether their existence is a simple consequence of variation in genes with critical cellular functions (15). If a primary function of *het* loci is to mediate self/nonself-recognition during vegetative growth, selection may have favored polymorphisms at *het* loci. Alleles at *het* loci would then exhibit evolutionary features similar to those found in other loci that regulate self/nonself-recognition, such as trans-species polymorphism. Alternatively, *het* loci could be polymorphic in populations as a result of genetic variation and encode products for which heteroallelism becomes lethal or detrimental. In this case, the number and constitution of *het* loci might vary in different fungal species, including the types and frequency of polymorphism at individual *het* loci.

Genes involved in heterokaryon incompatibility have been cloned and characterized from two haploid filamentous asco- mycetes, *Podospora anserina* (for review, see ref. 15) and *Neurospora crassa* (8, 16). In *N. crassa*, the *het-c* locus is one of 11 genetically identified *het* loci (17). Individuals that are nearly isogenic, but carry different alleles at *het-c*, do not form viable heterokaryons (18). The *het-c* locus encodes a glycine- rich polypeptide that contains a coiled-coil domain (16). Three distinct and mutually incompatible alleles termed *het-c*[OR], *het-c*[PA], and *het-c*[GR] have been characterized (19, 20), and a region of 34–48 amino acids that is polymorphic among HET-C[OR], HET-C[PA], and HET-C[GR] (Fig. 1) determines *het-c* allelic specificity. Mutants in *het-c* are unaffected in vegetative or sexual phenotype but have lost the capacity to mediate incompatibility via differences at *het-c* (16). The identification of polymorphisms within a domain that confers *het-c* specificity has allowed us to examine the evolutionary pattern of alleles at this locus in species within the Sordariaceae.

In this study, we find that *het-c* shares evolutionary features with loci in the MHC and the *S* systems, in which allelic polymorphisms are believed to be maintained by balancing selection. We find that the specificity region of *het-c* displays trans-species polymorphism, that *het-c* allele frequencies are approximately equal within a *N. crassa* population sample, and that high frequency of nonsynonymous substitutions are present within the *het-c* specificity region. These findings support our assertion that *het-c* functions as a self/nonself-

recognition locus and that this function is critical during the vegetative growth phase of filamentous fungi.

MATERIALS AND METHODS

Strains and Cultural Conditions. For *het-c* phylogenetic analysis, we used 40 strains representing 11 species and three different genera within the Sordariaceae. All strains were obtained from the Fungal Genetics Stock Center (FGSC; Department of Microbiology, University of Kansas Medical Center, Kansas City, KS). *N. crassa* strains were: FGSC 1130 (Panama), FGSC 4711 (Haiti), FGSC 4832 (Ivory Coast), FGSC 2489 (Louisiana), FGSC 1455 (unknown), FGSC 4709 (Haiti), FGSC 430 (Ivory Coast), FGSC 847 (Louisiana), FGSC 1693 (Louisiana), FGSC 1824 (Pakistan), FGSC 1945 (Florida), and FGSC 967 (Liberia). *N. intermedia* strains were: FGSC 2316 (Florida), FGSC 1940 (Florida), FGSC 3290 (Oahu), FGSC 3721 (Kauai), and FGSC 1799 (Malaya). *N. discreta* strains were: FGSC 3228 (Texas), FGSC 5923 (Flor- ida), FGSC 6788 (New Guinea), and FGSC 6793 (Brazil). *N. sitophila* strains were: FGSC 963 (France), FGSC 5940 (Ta- hiti), and FGSC 5941 (Tahiti). *Gelasinospora* sp. strains were: FGSC 8241, FGSC 8239, FGSC 8242, and FGSC 8243 (all from the Yucatan Peninsula.). *Sordaria brevicollis* strains were: FGSC 7140 (New York) and FGSC 1904 (New York). *S. heterothallis* strain was: FGSC 2739 (Texas). *S. sclerogenia* strains were: FGSC 2741 (Ceylon) and FGSC 2740 (Ceylon). All of the above strains are heterothallic in mating behavior.

```
PA-type
Nc1130    EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGGH--SPVFPHVGTATKLKL-----ENRQFRRVRPGEGYDSG-AKYAWPLVTGTFGG PA
Nc430     EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGGH--SPVFPHVGTATKLKL-----ENRQFRRVRPGEGYDSG-AKYAWPLVTGTFGG PA
Nc4711    EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGGH--SPVFPHVGTATKLKL-----ENRQFRRVRPGEGYDSG-AKYAWPLVTGTFGG
Nc4832    EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGGH--SPVFPHVGTATKLKL-----ENRQFLRTRPGE-HDSG-AKYVWPLVTGTFGG
Ns5941    EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGGH--SPVFPHVGTETILKL-----ENGQFSRPKPGERHDSR-AKYVWPLVTGTFGG
Ni2316    EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGGH--SPVFPHVGTATKLKL-----ENKQFLPTRPGE-HDSG-AKYVWPLVTGTFGG
Ni3290    EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGGH--SPVFPHVGTATKLKV-----ENRQFLPTRPGE-HDSR-AKYVWPLVTGTFGG
Nt6583    EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGGH--SPVFPHVGTETKLKL-----ENGQFRRVRPGEGYDSR-AKYAWPLVTGTFGG
Nt1270    EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGGH--SPVFPHVGTATKLKL-----ENRQFRRVRPGEGYDSG-AKYAWPLVTGTFGG
Ndi3228   EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGGH--SPVFPHVGTATKLKL-----ENRQFRRVRPGEGYDSG-AKYAWPLVTGTFGG
Nc4486    EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGGH--SPVFPHVGTATKLKL-----ENRQFRRVRPGEGYDSG-AKYAWPLVTGTFGG
Ndi5923   EALRCLGQALHTLEDFPAHSNYCELVLIDQKEDRS-KSRIFPHVGTRTKLKR-----ENGMFVPAPYEERD------YVWPLVTGTFGG
Np7221    EALRCLGQALHTLEDFPAHSNYCELALMDIHKKETGE-ESQIFPHVGAHTRLKL-----KNGKFRRVQGGERDSSPPDKYVWPLVTGTFGG
Gsp8243   EALRCLGQALHTLEDFPAHSNYCELVLIDIYKNETGR-ESQIFPHVGAHTRLRY-----ENGKYSRVDGGGRDYSSPGKYVWPLVTGTFGG
Ss2740    EALRCLGQALHTFEDFPAHSNYCELALMDIHKKETGE-ESQIFPHVGAHTRLTL-----KNGKFTRVQGGERDDSSPDKYVWPLVTGTFGG
GR-type
Ndi6788   EALRCLGQALHTLEDFPAHSNYCELVLIDIHEKETCSCESRIFPHVGTVTRITL-----NNG---------------KLVWPLVTGTFGG
Nc1945    EALRCLGQALHTLEDFPAHSNYCELVLIDIHEKETRS-ESRIFPHVGTATRITL-----NNG---------------KLVWPLVTGTFGG GR
Nc1455    EALRCLGQALHTLEDFPAHSNYCELVLIDIHEKETRS-ESRIFPHVGTATQITL-----NNG---------------KLVWPLVTGTFGG GR
Nc967     EALRCLGQALHTLEDFPAHSNYCELVLIDIHEKETRS-ESRIFPHVGTATQITL-----NNG---------------KLVWPLVTGTFGG GR
Nc4499    EALRCLGQALHTLEDFPAHSNYCELALIDIHEKETRS-ESRIFPHVGTATQITL-----NNG---------------KLVWPLVTGTFGG
Ndo1692   EALRCLGQALHTLEDFPAHSNYCELVLIDIYENENPRRESQIFPHVGAHTRKTL-----NNG---------------KYVWPLVTGTFGG
Ss2741    EALRCLGQALHTLEDFPAHSNYCELALIDIHEKETRS-ESRIFPHVGTATRITL-----NNG---------------KLVWPLVTGTFGG
Sb7140    EALRCLGQALHTLEDFPAHSNYCELALIYIHEKETRS-ESRIFPHVGTATRITL-----NNG---------------KLVWPLVTGTFGG
OR-type
Ni3721    EALRCLGQALHTLEDFPAHSNYCELVLIDIHEKETGS-ESRIFPHVGTATRITLRNDTRNNG---------------KSVWPLVTGTFGG
Nc1824    EALRCLGQALHTLEDFPAHSNYCELVLIDIHEKETGS-ESRIFPHVGTATRVTLRNDTRNNG---------------KSVWPLVTGTFGG OR
Ndi6793   EALRCLGQALHTLEDFPAHSNYCELVLIDMGERRGQH--SPVFPHVGTNTRVTLRNDTRNNG---------------KSVWPLVTGTFGG
Sb1903    EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGQH--SPVFPHVGTNTRVTLRNDTRNNG---------------KSVWPLVTGTFGG
Sh2739    EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGGH--SPVFPHVGTDTRITLRNDTRNNE---------------KSVWPLVTGTFGG
Gsp8241   EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGQH--SPVFPHVGTNTRVTLRNDTRNNG---------------KSVWPLVTGTFGG
Gsp8239   EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGGH--SPVFPHVGTNTRVTLRNDTRNNG---------------KSVWPLVTGTFGG
Gsp8242   EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGGH--SPVFPHVGTDTRITLTNDTRNNG---------------KSVWPLVTGTFGG
Nc4481    EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGGH--SPVFPHVGTDTRITLRNDTRNNG---------------KSVWPLVTGTFGG
Nc2489    EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGGH--SPVFPHVGTDTRVTLRNDTRNNG---------------KSVWPLVTGTFGG OR
Nc4709    EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGGH--SPVFPHVGTDTRVTLRNDTRNNG---------------KSVWPLVTGTFGG
Nc847     EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGGH--SPVFPHVGTDTRVTLRNDTRNNG---------------KSVWPLVTGTFGG OR
Nc1693    EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGGH--SPVFPHVGTDTRVTLRNDTRNNG---------------KSVWPRITGTFGG OR
Ns963     EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGQH--SPVFPHVGTDTRVTLRNDTRNNG---------------KSVWPLVTGTFGG
Ns5940    EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGQH--SPVFPHVGTDTRVTLRNDTRNNG---------------KSVWPLVTGTFGG
Ni1940    EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGGH--SPVFPHVGTDTRVTLENDTRNNG---------------KLVWPLVTGTFGG
Ni1799    EALRCLGQALHTLEDFPAHSNYCELVLIDMEERRGGH--SPVFPHVGTDTRVTLENDTRNNG---------------KLVWPLVTGTFGG OR
          *********************** * *    * ***** *         *              *********
```

FIG. 1.   Inferred amino acid sequences of PCR amplified *het-c* allelic specificity motif from 40 isolates. The alignment was obtained by CLUSTALW 6.1 and modified manually. Nc, *Neurospora crassa*; Ni, *N. intermedia*; Ndi, *N. discreta*; Ns, *N. sitophila*; Nt, *N. tetrasperma*; Np, *N. pannonica*; Ndo, *N. dodgei*; Ss, *Sordaria sclerogenia*; Sh, *S. heterothallis*; Sb, *S. brevicollis*; Gsp, *Gelasinospora* sp. Numbers following indicate the Fungal Genetics Stock Center (FGSC) number or P number (4481, 4486, and 4499 strains). PA (Panama), GR (Groveland), and OR (Oakridge) refer to the three molecularly characterized alleles that confer different *het-c* specificities (20). The *het-c* specificity of alleles designated in the figure was determined by partial diploid analysis or heterokaryon tests (refs. 19 and 20 and unpublished results). Asterisk and dashes indicate identical and deletion residue sites, respectively. Underlined section is *het-c* specificity variable region. Bold region is polymorphic block.

The following pseudohomothallic *N. tetrasperma* strains were used: FGSC 6583 (Tahiti) and FGSC 1270 (unknown). Homothallic species used were: *N. pannonica*, FGSC 7221 (Hungary), and *N. dodgei*, FGSC 1692 (Puerto Rico).

For population analysis, an additional 36 *N. crassa* strains isolated by D. J. Jacobson (Dept. of Biological Sciences, Stanford University) from a 5-hectare sugar cane field in Louisiana (ref. 21; P4448-P4501) were obtained from A. J. F. Griffiths, Botany Department, University of British Columbia. All strains were maintained on vegetative growth medium (22).

PCR and DNA Sequence Analysis. Genomic DNA from all isolates was prepared as described by Oakley *et al.* (23). Genomic DNA was digested with restriction enzymes and subjected to Southern blotting as described (16). The region of *het-c* allelic specificity was amplified by PCR from genomic DNA with two primers designed from the sequence of *het-c*^OR allele (20): 5′-(598)GGAGACATGGCGATATCG(615)-3′ and 5′-(1441)GTGAGGCACAACCCACTC(1424)-3′. Single amplification products corresponding to *het-c* were obtained for all isolates. The PCR products were cloned into the pCRII vector by using the TA cloning kit (Invitrogen) and subjected to DNA sequence analysis by using the ABI automatic sequencing procedure at the Nucleic Acid and Protein Sequencing Unit (University of British Columbia). For restriction fragment length polymorphism (RFLP) analysis, the PCR products from the *het-c* specificity domain were subjected to *Hae*III digestion and probed with *het-c*^OR.

Phylogenetic and Statistical Analyses. DNA and deduced amino acid sequences of the specificity region from the 40 isolates described above were aligned by CLUSTALW 6.1 (Institute for Biomedical Computing, Washington University, St. Louis, Mo.) and adjusted by eye. Phylogenetic analyses were performed by the neighbor-joining and maximum-likelihood methods by using PHYLIP (24). Genetic distances were estimated based on Kimura's two-parameter method (24). Insertions and deletions were excluded from analysis in pairwise comparisons. The number of nonsynonymous substitutions per nonsynonymous site (d_N) and synonymous substitutions per synonymous site (d_S) were calculated by using MEGA software (25), with pairwise deletion of gap sites. A standard z test was used to determine whether (*i*) mean d_S = mean d_N and (*ii*) mean d_N (of the *het-c* specificity region) = mean d_N (of the *het-c* nonspecificity region). The expected values of d_N:d_S and the associated standard errors were calculated by using the delta method (26, 27).

RESULTS

The Three *het-c* Allelic-Types Are About Equally Frequent Both Within and Between *Neurospora crassa* Populations. A variable domain characterized by particular insertion/deletion motifs (20) determines allelic specificity at *het-c*. Of 15 *N. crassa* isolates in which the DNA sequence of the *het-c* specificity region was determined, five alleles were *het-c*^PA-type, four were *het-c*^GR-type, and six were *het-c*^OR-type (Fig. 1). There was no obvious correlation between geographical location and *het-c* allelic type. If the different *het-c* allelic types are maintained by selection, the expectation is that a near-equal frequency of each type of allele would be observed within a fungal population, assuming equivalent reactions among alleles and no intrinsic fitness differences. The near-equal frequency of the three *het-c* allelic types among geographically dispersed *N. crassa* isolates suggested the possibility that each allelic type also might be present at equal frequencies within a population of *N. crassa*.

Although a large number of wild-type isolates of *N. crassa* are available, most had been isolated from scattered regions in the sub-tropics. However, a population sample of 39 individuals had been recovered from a 5-hectare sugarcane field in Louisiana (21). RFLP analysis of the *het-c* specificity region PCR product from 36 isolates (three isolates did not grow) showed that 14 of the *N. crassa* isolates had a *het-c*^OR pattern, 14 had a *het-c*^PA pattern, and eight had a *het-c*^GR pattern (data not shown). The DNA sequence of the *het-c* specificity region was determined for a representative of each polymorphic class, which confirmed the allele designations made by RFLP analysis (Fig. 1; Nc4481, Nc4486, and Nc4499). When digested genomic DNA from the 36 isolates was probed with a 35-kbp cosmid containing *het-c*, twenty-four different RFLP patterns could be identified, demonstrating that the population was not clonal or, if clonal, was not of recent origin.

The HET-C Sequences from Different Species and Genera in the Sordariaceae Were Either HET-C^OR, HET-C^PA, or HET-C^GR-Type. If *het-c* polymorphisms are present because of genetic variation in *N. crassa* and if they are not under selection, the expectation is that *het-c* polymorphisms would be lost by genetic drift in species distantly related to *N. crassa.* To determine whether polymorphisms at *het-c* are restricted to *N. crassa*, the *het-c* specificity regions from an additional 25 isolates representing 11 different species and three different genera in the Sordariaceae were analyzed. DNA hybridization analysis was consistent with the presence of only a single *het-c* allele within the genome of each isolate (data not shown). Analysis was restricted mainly to heterothallic species. Unlike *Neurospora*, most other heterothallic species in the Sordariaceae are not represented in collections by large numbers of individual strains that could be analyzed.

The DNA sequences of the *het-c* specificity region were divergent among the 40 isolates, with the two most distantly related members of the 40 isolate data set (*N. pannonica* 7221 and *S. heterothallis* 2739) showing only 66% identity. Although the DNA sequences of the *het-c* specificity region were divergent, the deduced amino acid sequences from the 40 isolates grouped into one of the three HET-C specificity types based on the pattern of insertion/deletion first observed in the *N. crassa het-c* alleles (Fig. 1).

A comparison of the HET-C region from the 40 isolates showed conserved and variable domains. The regions flanking the HET-C specificity domain showed 100% amino acid identity among the 40 isolates (Fig. 1). A highly polymorphic amino acid block [MEERRGG(Q)H and IH(Q)EKET-(D)R(G/C)S] is highlighted in Fig. 1. The 40 isolates have either of these polymorphic blocks, although the HET-C^GR-type isolates all contained a similar motif [IHEK(N)ET-(N)R(C, P)S]. The amino acid sequence in this polymorphic block apparently does not affect *het-c* specificity in *N. crassa* because Nc1824 and Nc2489 have identical *het-c* specificities based on partial diploid analysis (19, 20). Following this polymorphic region is a five-amino acid block (FPHVG) that is conserved in all 40 isolates.

For all species in which more than one isolate was characterized (with the exception of the pseudohomothallic species, *N. tetrasperma*), at least two of the three *het-c* allelic variants were found. In the heterothallic species *S. brevicollis*, the two interfertile isolates from this species (both isolated from zebra dung in the New York Zoological Park) contained *het-c* alleles with different HET-C types (HET-C^GR and HET-C^OR) (Fig. 1). The two *S. sclerogenia* interfertile isolates from Ceylon also have different HET-C allelic types, one being HET-C^PA (Ss2740) and the other HET-C^GR (Ss2741). Similarly, the HET-C specificity domain from interfertile *Gelasinospora sp.* isolates from the Yucatan Peninsula fell into either HET-C^PA or HET-C^OR types. All three of the *het-c* allelic types were identified in all of the heterothallic *Neurospora* species. From the amino acid sequence alignment shown in Fig. 1, it is evident that HET-C polymorphism (as characterized by insertion/deletion motif and the associated highly polymorphic block within *het-c*) is shared between species and genera within the Sordariaceae.

The DNA Sequence of the *het-c* Specificity Region Shows Trans-Species Polymorphism. Evidence for shared polymorphism at the DNA level between species within the Sordariaceae would indicate that allelic specificities at *het-c* were generated in an ancestral species for genera within the Sordariaceae and have been maintained through multiple speciation events. Alternatively, each of the species within the Sordariaceae may have independently evolved similar insertion/deletion polymorphisms within the *het-c* specificity region that are only evident at the amino acid level. To distinguish between these two alternatives, the relationship of the DNA sequence of the *het-c* region from the 40 alleles was examined by phylogenetic analysis using a neighbor-joining algorithm based on Kimura's two-parameter distance with pairwise elimination of insertion/deletion sites (24). Nearly identical trees were produced by DNA maximum likelihood and DNA parsimony analyses (data not shown). In addition, a neighbor-joining analysis that excluded the entire deletion/insertion motif [from the beginning of the Oakridge (OR)-type insertion to the end of the Panama (PA)-type insertion] also gave a similar tree, indicating that tree topology was not an artifact of alignment.

From Fig. 2, it is apparent that the DNA sequences of the 40 alleles fell into groups by *het-c* allelic type rather than according to species. Thus, the DNA sequences of the specificity region of *het-c* displayed trans-species polymorphism. Although only three allelic specificities have been identified in *N. crassa* (*het-c*OR, *het-c*PA, and *het-c*GR-type), phylogenetic analysis of the *het-c* DNA sequences from the 40 isolates showed additional lineages. In all of the trees, two *het-c*OR-type strains, Ni3721 and Nc1824, were included in the Groveland (GR)-type lineage. The clustering of these two OR-type strains with the GR-type strains was due to the intragenic combinatorial association of the MEERRGGH/IHEKETRS polymorphic block (bold in Fig. 1) with the *het-c*OR insertion/deletion type. Similarly, strains Ndi5923, Gsp8243, Np7221, and Ss2740 were separated from the PA-lineage because of the linkage of the IHEKETRS-type polymorphic block with the *het-c*PA insertion. Because gap sites were eliminated in pairwise comparisons in genetic distance estimation, the polymorphic block in *het-c* affected placement of these strains in phylogenetic analysis. Combinatorial associations of polymorphic blocks also have been observed in MHC class I DNA polymorphisms (28).

Although strains from the different species/genera are included in each *het-c* lineage (OR, PA, or GR-type), the tips of lineages often show the same or closely related species. These data indicate that a number of closely related alleles with similar *het-c* polymorphisms are present within a single species, which one would expect if these alleles have been retained in populations and if gene flow among species is rare. However, representatives of a single species within a single allelic type do not form discrete groupings, suggesting that selection also may be operating to maintain diversity even within *het-c* allelic types.

In both the MHC and the *S* locus, trans-species polymorphisms have been retained during multiple speciation events. The allele polymorphisms at the *DRB1* locus have been calculated to have been retained for >30 million years (1), whereas the *S* locus polymorphisms have been retained for at least 36 million years (3). To determine how long *het-c* polymorphisms have been retained in species in the Sordariaceae, we calculated the divergence time between *Neurospora* and *Sordaria*. The molecular clock for the 18S rRNA sequences has been calibrated by using fossil evidence and ages of fungal hosts and symbionts to estimate the rate of substitution (29). The DNA sequence of the 18S rRNA gene from *Sordaria fimicola* (X69851) and *N. crassa* (29) show a 1.46% difference, indicating a 0.73% substitution per lineage. Assuming a reasonably high substitution rate of 2% per 100 million years (29),
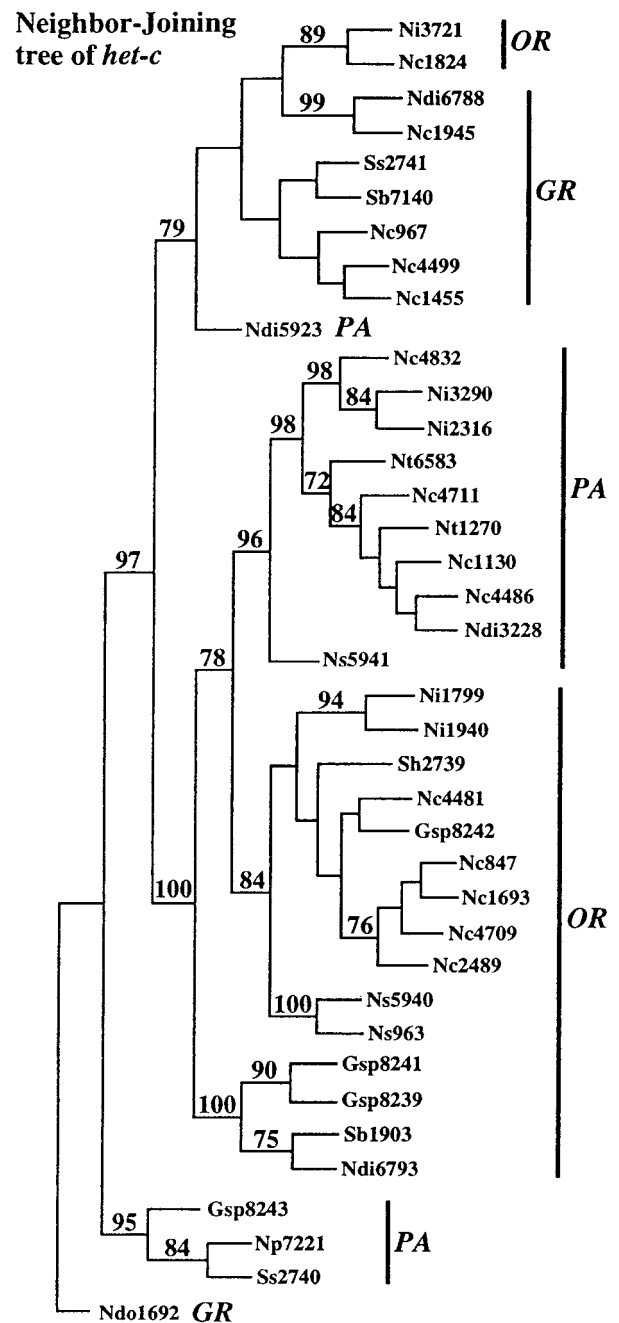

FIG. 2. Neighbor-joining tree using 390-bp nucleotide sequences spanning the *het-c* specificity motif from 39 isolates (*N. crassa* FGSC 1130 (Panama) and 430 (Ivory Coast) have identical nucleotide sequences in this region). The *het-c* sequence from the homothallic species, *N. dodgei* (Ndo1692), was used as an outgroup. Distance was calculated by Kimura's two parameter method in the phylogenetic analysis package PHYLIP 3.5 (24). Numbers are bootstrap percentages (out of 1,000 resamplings) in support of each node. Bootstrap percentages below 70% were not included. Each insertion and deletion was treated as a single event in the analysis. Species abbreviations are indicated as in Fig. 1. The designation OR, *het-c*OR allelic-type; PA, *het-c*PA allelic-type; and GR, *het-c*GR allelic-type is based on alignment in Fig. 1.

Sordaria diverged from *Neurospora* at least 36 million years ago. The persistence of *het-c* polymorphisms in different genera during such a long time period indicate that the allelic variants have been maintained by selection in populations across multiple speciation events.

HET-C Shows Constrained and Unconstrained Regions. Pairwise sequence comparisons of different *S*- or MHC-

encoded products show that the level of amino acid divergence is highly constrained in certain regions of the proteins, whereas other regions appear not only to be unconstrained, but under selection for diversity. This aspect is evidenced by an excess of nonsynonymous over synonymous substitutions in the hyper-variable regions conferring specificity, consistent with a diversifying selection thought to operate at these loci (4, 30, 31). A frequent occurrence of nonsynonymous substitutions also was observed among alleles of the *het-s* (32) and among alleles at *het-c* (33) loci in *P. anserina*. Among the 40 HET-C sequences in this study, the amino acids in the region that flanked the specificity region were 100% identical, whereas those in the specificity region were variable (Fig. 1). A comparison of synonymous to nonsynonymous substitution ratios in the *het-c* specificity domain among the 40 isolates also showed variability. The frequency of nonsynonymous substitutions significantly increased both within and around the insertion/deletion motif of the *het-c* specificity region (Fig. 3).

The numbers of nucleotide substitutions per synonymous (d_S) site and per nonsynonymous (d_N) site (26) were estimated among all 40 *het-c* alleles (Table 1). For the region flanking the specificity domain (non-specificity region), the mean d_S significantly exceeded the mean d_N ($d_N/d_S = 0.11$ for all alleles; z test $P < 0.05$). However, the value of d_N increased significantly within the 57 codons that make up the *het-c* specificity region (underlined in Fig. 1; $d_N/d_S = 0.54$ for all alleles; z test $P < 0.05$). The most variable alleles were within the *het-c*^{PA}-type insertion, where d_N/d_S ratio was 1.3. The analysis of d_N/d_S of *het-c* indicated that the polymorphic alleles are derived from an evolutionarily highly conserved gene and suggest that most amino acid differences between the polypeptides are not tolerated. However, a high occurrence of nonsynonymous substitutions and a significant increase in nonsynonymous substitutions per site near and within the insertion/deletion

Table 1. Comparison of d_S and d_N substitutions per 100 sites in *het-c*

Alleles/ species	n	Specificity region		Non-specificity region	
		d_S	d_N	d_S	d_N
OR-type*	17	25.5 (5.6)	10.3 (1.7)	19.2 (3.5)	2.1 (0.6)
GR-type*	8	16.2 (5.1)	6.0 (1.5)	14.0 (3.3)	1.4 (0.6)
PA-type[†]	15	16.3 (4.5)	21.5 (2.9)	13.8 (3.0)	1.2 (0.4)
All[†]	40	38.0 (7.1)	20.4 (2.8)	24.5 (3.5)	2.6 (0.6)
N. crassa[‡]	15	23.9 (7.8)	17.0 (3.3)	20.5 (4.3)	2.1 (0.7)
Sordaria[†]	5	37.0 (8.9)	20.3 (3.8)	27.3 (5.5)	1.5 (0.7)

Specificity and non-specificity region refer to the codons of filled and clear bar regions in Fig. 3, respectively. *n*, the number of alleles surveyed. Standard errors are in parentheses. GR, OR, and PA-types refer to *het-c*^{GR}, *het-c*^{OR}, and *het-c*^{PA} allelic specificity types, respectively. Test of hypothesis: d_N/d_S ratio has no significant difference in the two regions (26, 27). *$P < 0.05$, [†]$P < 0.001$, and [‡]$P < 0.01$.

motif resulted in amino acid diversity of *het-c* encoded products.

DISCUSSION

The *S* locus and the MHC systems are examples in which balancing selection has caused the retention of ancestrally polymorphic alleles, resulting in several common evolutionary features. First, each has a large number of alleles present within a single species. Second, these alleles differ at a number of nucleotide sites, often displaying large divergences in DNA sequence. Third, trans-species polymorphisms are apparent. Fourth, the ratio of nonsynonymous to synonymous substitutions within the specificity domain exceeds that outside of the specificity domain. The ratio of nonsynonymous to synonymous mutations (d_N/d_S) within the hypervariable domain (peptide-binding region) of the *DRB1* alleles can be as high as 10 for within species (salmon species; ref. 34), whereas a 40-amino acid variable domain of *S* alleles in the Solanaceae has a d_N/d_S ratio that is closer to 1 (31). In this study, we examined the evolutionary features of the *het-c* specificity region in different filamentous fungal species in the Sordariaceae. We show that the *het-c* region conferring specificity displays large divergences in DNA sequence, that DNA polymorphisms present in the *het-c* specificity region exhibit trans-species polymorphisms, and that this region has an increased frequency of nonsynonymous substitutions. The *het-c* locus thus shares with loci in the MHC and the *S* locus particular evolutionary features that are unexpected under the neutral model of evolution but that are predicted under balancing selection. Although the *het-c* locus shares these features with loci in the MHC and *S* locus, only three *het-c* allelic specificities have been identified in *N. crassa*, which are characterized by particular insertion/deletion motifs (19, 20). In *N. crassa*, vegetative incompatibility can be mediated by any of 11 different loci (17). If even a portion of these loci are relevant to self/nonself-recognition in *N. crassa*, the number of incompatible individuals in a segregating population would be extremely large. In a population analysis of *N. crassa*, heterokaryon compatible individuals were not identified, indicating that the population was highly heteromorphic for *het* loci (35). It is possible that the generation of additional alleles at *het-c* is constrained by requirements for protein function and that multiple alleles at *het-c* are not necessary because of multiplicity of *het* loci found within a species.

Retention of ancestral polymorphisms in neutral genes can be expected if species divergence times have been short and the population size is large. As judged by comparison of 18S rRNA sequences, *Sordaria* and *Neurospora* diverged over 36 million years ago. Theoretical analyses indicate that alleles at a locus under balancing selection will have long residence times and
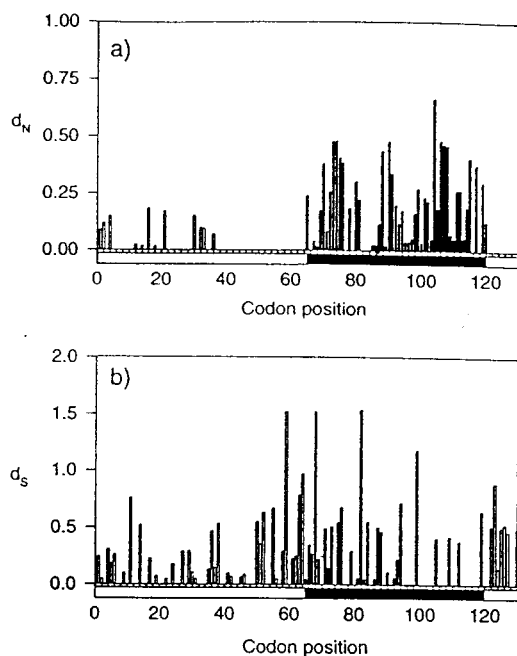

FIG. 3. Histograms of d_N and d_S substitutions per site mapped onto the codons for both the polymorphic and conserved regions in *het-c* from the 40 isolates. Sites 40–129 correspond to the sequences aligned in Fig. 1. Filled bar sites (codons 65–119) are polymorphic and correspond to the underlined region in Fig. 1. (*a*) The average number of nonsynonymous substitutions per nonsynonymous site (d_N) grouped by codons, and (*b*) the average number of synonymous substitutions per synonymous site (d_S) grouped by codons. Calculations were done by MEGA according to the Cantor–Jukes correction (25).

expected coalescence times far exceeding those of neutral alleles (1). Based on the data in this study, *het-c* polymorphisms must have originated in a species ancestral to *Sordaria* and *Neurospora* over 36 million years ago and have been retained during subsequent multiple speciation events. We believe that the wide occurrence of *het-c* trans-species polymorphisms within the Sordariaceae and the approximately equal frequency of each allelic type within a *N. crassa* population sample make it extremely unlikely that *het-c* diversity has been maintained under neutrality.

An alternative explanation for the occurrence of trans-species polymorphisms in *het-c* is that horizontal transfer occurs among species in the Sordariaceae. DNA sequences from a number of other genes and RFLP analysis support the hypothesis that strains assigned to *N. sitophila*, *N. tetrasperma*, and *N. discreta* form respective monophyletic groups within the genus *Neurospora* (36). It also is possible that the maintenance of *het-c* polymorphism results from "hitchhiking" after selection on a tightly linked locus. This hypothesis is not supported, however, because it predicts that a high ratio of d_N/d_S would be present throughout *het-c* and not just within the specificity region.

In *P. anserina*, an excess of nonsynonymous substitutions was observed between two *het-s* alleles (involved in allelic incompatibility; ref. 32) and between four *het-c* alleles (involved in non-allelic incompatibility; ref. 33). In this study, we observed a significant increase in the frequency of nonsynonymous substitutions within the *het-c* specificity region of the 40 alleles as compared with regions flanking the specificity domain. The fact that an excess of nonsynonymous substitutions has been observed in alleles from three different *het* loci suggests that selection to maintain diversity is a common evolutionary feature of these loci. The *het-c* diversity and presence of lineages within the OR, PA, and GR groups in our phylogenetic analysis also suggest that some of the *het-c* variants observed in other species and genera within the Sordariaceae may confer *het-c* specificities other than those identified in *N. crassa*. It will be of interest to determine whether *het-c* functions as a vegetative incompatibility locus in other species and whether balancing selection operates to maintain diversity of allelic-types at other *het* loci in filamentous fungi.

Allelic polymorphisms can be maintained through speciation events either by over-dominance in a diploid organism (heterozygote superiority) or by frequency dependent selection (6, 7); controversy surrounds the selection pressure for maintenance of MHC diversity (37, 38). If *het-c* allelic differences are expressed only in the haploid vegetative phase, then heterozygote superiority cannot be the driving force for maintenance of *het-c* polymorphisms, and the maintenance of allelic diversity would be attributed to frequency-dependent selection. However, until *het-c* is shown to be inactive in the sexual diplophase or dikaryophase, overdominance cannot be ruled out. Our analysis of the *het-c* specificity domain in the genera of the Sordariaceae supports a role for this locus as a self/nonself-recognition system limiting heterokaryon formation between unlike individuals during vegetative growth. These data shed light not only on the biological role of vegetative incompatibility in filamentous fungi but also reinforce the view that balancing selection is an evolutionary mechanism common to different self/nonself-recognition systems.

1. Ayala, F. J., Escalante, A., O'hUigin, C. & Klein, J. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 6787–6794.
2. Charlesworth, D. (1995) *BioEssays* **17,** 31–38.
3. Ioerger, T. R., Clark, A. G. & Kao, T. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 9732–9735.
4. Hinata, K., Watanabe, M., Yamakawa, S., Satta, Y. & Isogai A. (1995) *Genetics* **140,** 1099–1104.
5. Klein, J. (1987) *Hum. Immunol.* **19,** 155–162.
6. Takahata, N., Satta, Y. & Klein, J. (1992) *Genetics* **130,** 925–938.
7. Takahata, N. & Nei, M. (1990) *Genetics* **124,** 967–978.
8. Glass, N. L. & Kuldau, G. A. (1992) *Annu. Rev. Phytopathol.* **30,** 201–224.
9. Leslie, J. F. (1993) *Annu. Rev. Phytopathol.* **31,** 127–150.
10. Worrall, J. J. (1997) *Mycologia* **89,** 24–36.
11. Esser, K. & Blaich, R. (1994) in *The Mycota I: Growth, Differentiation and Sexuality*, eds. Wessels, J. G. H. & Meinhardt, F. (Springer, Heidelberg, Germany), pp. 211–232.
12. Garnjobst, L. & Wilson, J. F. (1956) *Proc. Natl. Acad. Sci. USA* **42,** 613–618.
13. Begueret, J. & Bernet, J. (1973) *Nat. New Biol.* **243,** 94–96.
14. Caten, C. E. (1972) *J. Gen. Microbiol.* **72,** 221–229.
15. Begueret, J., Turcq, B. & Clave, C. (1994) *Trends Genet.* **10,** 441–446.
16. Saupe, S. J., Kuldau, G. A., Smith, M. L. & Glass, N. L. (1996) *Genetics* **143,** 1589–1600.
17. Perkins, D. D., Radford, A., Newmeyer, D. & Bjorkman, M. (1982) *Microbiol. Rev.* **46,** 426–570.
18. Garnjobst, L. (1953) *Am. J. Bot.* **40,** 607–614.
19. Howlett, B. J., Leslie, J. F. & Perkins, D. D. (1993) *Fungal Genet. Newsletter* **40,** 40–42.
20. Saupe S. J. & Glass, N. L. (1997) *Genetics* **146,** 1299–1309.
21. Marcinko-Kuehn, M., Yang, X., Debets, F., Jacobson, D. J. & Griffiths, A. J. F. (1994) *Curr. Gen.* **26,** 336–343.
22. Vogel, H. J. (1964) *Am. Nat.* **98,** 435–446.
23. Oakley, C. E., Weil, P. L., Dretz, P. L. & Oakley, B. R. (1987) *Gene* **53,** 293–298.
24. Felsenstein, J. (1993) PHYLIP (Department of Genetics, University of Washington, WA), Version 3.5.
25. Kumar, S., Tamura, M. & Nei, M. (1993) *Molecular Evolutionary Genetics Analysis* (Pennsylvania State Univ. Press, University Park, PA).
26. Kotz, S. & Johnson, N. L. (1989) in *The Encyclopedia of Statistical Sciences*, (Wiley, New York), Vol. 8., p. 646.
27. Haydon, D., Lea, S., Fry, L., Knowles, N., Samuel, A. R., Stuart, D. & Woolhouse, M. E. J. (1998) *J. Mol. Evol.* **46,** 465–475.
28. Yuhki, N. & O'Brien, S. J. (1994) *J. Mol. Evol.* **39,** 22–33.
29. Berbee, M. L. & Taylor, J. W. (1993) *Can. J. Bot.* **71,** 1114–1127.
30. Hughes, A. L. & Nei, M. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 958–962.
31. Clark, A. G. & Kao, T. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 9823–9827.
32. Turcq, B., Deleu, C., Denayrolles, M. & Begueret, J. (1991) *Mol. Gen. Genet.* **228,** 265–269.
33. Saupe, S., Turcq, B. & Begueret, J. (1995) *Curr. Genet.* **27,** 466–471.
34. Miller, K. M. & Withler, R. E. (1996) *Immunogenetics* **43,** 337–351.
35. Mylyk, O. M. (1976) *Genetics* **83,** 275–284.
36. Skupski, M. P., Jackson, D. A. & Natvig, D. O. (1997) *Fungal Gen. Biol.* **21,** 153–162.
37. Slade, R. W. & McCallum, H. I. (1992) *Genetics* **132,** 861–862.
38. Hughes, A. L. & Nei, M. (1992) *Nature (London)* **355,** 402–403.