INTRODUCTION

Transposable elements (TEs) are an important component of many genomes. In eukaryotes, TEs often make up a large proportion of the genome, and are involved in a wide range of genetic and genomic rearrangements. TEs are traditionally classified into two broad classes, I and II. Class I elements, or retrotransposons, transpose via RNA intermediates and encode the enzyme reverse transcriptase (RT). These can be further subdivided into the long terminal repeat (LTR) retrotransposons and the non-LTR retrotransposons, which differ in both structure and replication mechanisms (Eickbush & Malik, 2002). In class II elements, or DNA transposons, the genomic DNA itself is the mobile intermediate. These elements typically encode a transposase, which is involved in the excision of the element from the host genome and its reinsertion at a new site. DNA transposons can be classified into a number of large superfamilies, such as Tcl/mariner (Doak et al., 1994) and hAT (Calvi et al., 1991), on the basis of differences in structure and in the sequence of the transposase.

In most LTR retrotransposons, and many different DNA transposons, the integration/transposition reaction is catalysed by an enzyme bearing a characteristic triad of aspartate and glutamate residues, known as the DDE domain (Fayet et al., 1990; Khan et al., 1991; Kulkosky et al., 1992). Recently, however, we reported that members of the DIRS1 group of LTR retrotransposons lack genes for DDE-type integrases, but instead encode tyrosine (or lambda) recombinases (Goodwin & Poulter, 2001a). These enzymes are likely to be involved in the insertion of DIRS1-like elements into their hosts’ genomes. Tyrosine recombinase genes had previously been found in a variety of prokaryotic elements, such as bacteriophage lambda, and various prokaryotic transposons and plasmids, but had seldom been found in eukaryotes, the only well-characterized eukaryotic examples being the FLP recombinases of yeast 2-micron circle plasmids. Since our initial report we have identified a second group of LTR retrotransposons, Ngaro1-like elements, that encode tyrosine recombinases (T. J. D Goodwin & R. T. M. Poulter, unpublished). In addition, several eukaryotic DNA transposons which bear tyrosine recombinase genes have very recently been described (Doak et al., 2003; Jacobs et al., 2003). These
elements were found in the ciliate *Euplotes crassus*, and like most other DNA transposons they are flanked by long terminal inverted repeats. Tec1 and Tec2 (Doak et al., 2003) each encode a Tc1/mariner-like DDE-type transposase as well as the tyrosine recombinase, and it is possible that the recombinase serves as a resolvase rather than as a transposase. The full complement of coding regions in Tec3 (Jacobs et al., 2003) is not yet known.

In this report we describe a new group of tyrosine-recombinase-encoding DNA transposons. The hosts of these elements are all pathogenic fungi, and include the ascomycetes *Coccidioides posadasii* and *Histoplasma capsulatum*, and the basidiomycete *Cryptoccus neoformans*. *C. posadasii* is a highly virulent species that can cause serious, sometimes fatal, systemic infections (Fisher et al., 2002; Kirkland & Fierer, 1996). *H. capsulatum* is a common cause of respiratory tract infections, and can be life-threatening (Woods, 2002). *C. neoformans* is an important opportunistic pathogen of humans that poses a significant threat to immunocompromised individuals (Kwon-Chung & Bennett, 1992). Isolates of *C. neoformans* are classified into three varieties known as *grubii* (serotype A), *neofor mans* (serotype D) and *gattii* (serotypes B and C), based on antigenic differences in the polysaccharide capsule that surrounds the fungal cells (Kwon-Chung et al., 1982). Serotypes A and D, and a hybrid AD serotype, are found worldwide. *C. neoformans* var. *gattii* strains are mainly restricted to tropical and subtropical regions (Hull & Heitman, 2002). Molecular phylogenetic work revealed that the *grubii* and *neofor mans* varieties are separated by ~18.5 million years of evolution, and these varieties diverged from *gattii* ~37 million years ago (Xu et al., 2000). Due to their importance as human pathogens, and in some cases as model organisms, each of the above species is currently the focus of a genome sequencing project (briefly outlined in Methods).

We refer to the new tyrosine-recombinase-encoding transposons that we have identified in these fungi as cryptons. These elements are very different in structure from any previously described TEs. Some appear to have been recently mobile and may still be active, which should facilitate their experimental characterization. Sequence analyses suggest that cryptons may be related to the tyrosine-recombinase-encoding DIRS1 and *Ngaro1* groups of retrotransposons.

**METHODS**

**Strains and culture conditions.** *C. neoformans* strains used were JEC21 (Heitman et al., 1999); Cn3511 (CBS1312; Ikeda et al., 1985); WM148 (Weiland Meyer, University of Sydney); IUM92-4755, IUM91-6422, IUM93-1545 and IUM88-3921 (M. A. Viviani, University of Milan). Strains were grown on YPD medium (1% yeast extract, 2% peptone, 2% glucose) at 37 °C.

**DNA manipulations.** *C. neoformans* genomic DNA was isolated essentially by the method of Philippson et al. (1991). PCR was performed on an Eppendorf Mastercycler Gradient instrument, using the Expand Long Template or High Fidelity PCR Systems (Roche). Oligonucleotide primers were from Prolog, Singapore. Sequencing was performed using an ABI377 DNA sequencer at the University of Otago. Southern blots were performed as described previously (Goodwin & Poulter, 2000). Autoradiographs were scanned using a Bio-Rad GS-800 Calibrated Densitometer and Quantity One software.

**Sequence analyses.** General sequence analyses were performed using the programs of the Wisconsin GCG package (Genetics Computer Group) and the Australian National Genomic Information Service node located at the University of Otago (http://angis.otago.ac.nz/). Sequence similarity searches were performed at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). Multiple sequence alignments were constructed using CLUSTAL_X (Thompson et al., 1997) and refined using SEAVIEW (Galtier et al., 1996). Phylogenetic trees were constructed using PAUP*4b10 (Swofford, 1998).

**Fungal genome sequencing projects.** *Cryptoccus neoformans.* The genome of serotype D strain B-3501A is being sequenced at the Stanford Genome Technology Center (SGTC; http://www-sequence.stanford.edu/group/C.neoformans/index.html). Results described in this paper refer to release cneoformans030328 (March 31, 2003). Serotype D strain JEC21 is being sequenced at The Institute for Genomic Research (TIGR; http://www.tigr.org/tdb/e2k1/cnf1/). Serotype A strain H99 is being sequenced at the Duke Center for Genome Technology (http://cneo.genetics.duke.edu/) and the Vancouver Genome Sequence Centre, BC Cancer Research Centre; http://www.bcgsc.bc.ca/). In addition, large numbers of cDNA sequences for strains B-3501 and H99 have been obtained by the University of Oklahoma’s Advanced Center for Genome Technology (http://www.genome.ou.edu/cneo.html).

*Coccidioides posadasii.* Genomic sequence data were from TIGR (http://www.tigr.org/tdb/tgi/cgi/GenInfo.html).

*Histoplasma capsulatum.* Genomic sequence data were from the Genome Sequencing Center at Washington University in St Louis (http://www.genome.wustl.edu/projects/Hcapsulatum/). Two distinct strains of *H. capsulatum*, G217B and G186AR, are being sequenced.

*Candida albicans.* Sequence data for *Candida albicans* were obtained from the Stanford Genome Technology Center website at http://www.sequence.stanford.edu/group/candida. Sequencing of *Candida albicans* was accomplished with the support of the NIDR and the Burroughs Wellcome Fund.

**Sequences.** The following sequences have been deposited in the GenBank/EMBL/DDJB databases: crypton Cnl, AJ488893; related empty site for crypton Cnl, AJ488894.


Sources of the additional tyrosine recombinase sequences used in the phylogenetic analyses are as follows: all *Apo loo* sequences, *Phanerochaete chrysosporium*, http://www.jgi.doe.gov/programs/whiteroot.htm; CbRec2m, Cae noharbattis briggsae, AC084491; DrDirs2, Danio rerio, AL645756; EgRecorn1, Euglena gracilis, L39772; Goliath, Lytechinus variegatus, AC131494; Kangaroo1, Volvox carteri, AY137241; Lvdirs1, L. variegatus, AC131505; M1, 04370.40.11

On: Sun, 16 Dec 2018 02:58:55
**RESULTS**

Tyrosine recombinase genes in *C. neoformans*

Using the putative tyrosine recombinase sequences of some *Ngaro1*-like retrotransposons as queries in BLAST searches we detected several distinct tyrosine recombinase genes in *C. neoformans*. These genes were initially found in the sequence database provided by the Stanford Genome Technology Center (http://www-sequence.stanford.edu/group/C.neoformans/index.html); similar sequences were subsequently also detected in the TIGR (http://www.tigr.org/tdb/e2k1/cna1/) and Duke University (http://cneo.genetics.duke.edu/) Cryptococcus sequence databases.

In the *C. neoformans* genes, the matches to tyrosine recombinases were frequently found in different reading frames, and often with stop codons in the intervening regions. This might be because these are degenerate pseudogenes, or it might result from the presence of some small introns. To test the latter possibility we compared the genes with a database containing a large number of *C. neoformans* cDNA sequences (http://www.genome.ou.edu/blast/B3501.html). For one of the genes several matching cDNAs were found. Comparison of the genomic and cDNA sequences revealed that there are indeed several introns in the gene. These appear to be typical spliceosomal introns, similar to those found in other fungi (Bon et al., 2003). Removal of these introns from the genomic sequence creates an uninterrupted ORF. By comparison with this sequence we were able to predict introns in the other genes as well. In several additional cases removal of these introns also results in intact ORFs. Some other sequences, however, do appear to be pseudogenes.

Four of the putative *C. neoformans* tyrosine recombinases (predicted from the reconstructed ORFs) are shown aligned with other tyrosine recombinases in Fig. 1(a). It can be seen that the *C. neoformans* sequences share significant sequence similarities with members of the tyrosine recombinase family. In particular, most of them contain the four very highly conserved residues characteristic of these enzymes: the RHRY tetrad (Nunes-Duby et al., 1998). They also contain sequence similarities in the regions flanking these conserved residues. The overall level of conservation clearly indicates that these *C. neoformans* sequences are members of the tyrosine recombinase family.

The various *C. neoformans* tyrosine recombinase genes often differ quite considerably from each other in sequence (Fig. 1a). For instance, figures of ~40% amino acid identity between various pairs, over the region encompassing the conserved RHRY tetrad, are typical. This suggests that these sequences diverged from a common ancestor a long time ago. The *C. neoformans* sequences do not appear to be particularly closely related to any previously identified tyrosine recombinases: they seldom display greater than 20% overall amino acid identity with other sequences. Close examination of the sequence alignment (Fig. 1a), however, reveals that they share several short motifs specifically with tyrosine recombinases from *Ngaro1* and DIRS1 retrotransposons, and some bacterial sequences previously found to be related to these retrotransposons. Most striking among these is a motif 50–60 amino acid residues downstream of the first conserved Arg residue that often contains the sequence Cys-Pro-Val. Unlike the recombinases from *Ngaro1* and DIRS1 retrotransposons, however, we did not find the *C. neoformans* sequences in close association with any recognizable RT or ribonuclease H (RNH) genes – characteristic features of retrotransposons.

The presence of introns in the *C. neoformans* sequences, and the absence of associated RT or RNH genes, suggests that, unlike the majority of previously identified eukaryotic tyrosine recombinase genes, they are not components of retrotransposons. Instead, the above results suggest that they may be cellular genes, or else be parts of some novel type of transposable element.

**Novel transposons in *C. neoformans***

Transposable elements often have a number of characteristic features which enable them to be distinguished from normal cellular genes: they usually exist in multiple copies per genome; these copies may reside at different loci; the elements may have distinct ends; empty sites related to the inserted sites may be found; and there may be different patterns of insertion in different strains.

We found that many of the *C. neoformans* tyrosine recombinases genes appear to be parts of larger elements which have many of the above features, and which are, therefore, likely to be transposable elements. We refer to these elements in general as cryptons. Elements from a particular species are given an additional two-letter species identifier, e.g. *Cn* for *C. neoformans*, and a number to indicate the particular family to which they belong (elements sharing greater than ~80% identity at the DNA level were considered to be a family).

For an element that we call crypton *Cn1* we found four distinct 5’ ends and two distinct 3’ ends in the various *C. neoformans* sequence databases (Fig. 2a). Distinct insertion sites were identified both within a single strain and in different strains. For each end, the elements at different loci are all highly similar in sequence up to a certain point, after which the sequences abruptly diverge. These points probably correspond to the precise ends of the
A sequence contig from strain JEC21 contains both the 5' and 3' ends of a single crypton Cn1 element (the other crypton Cn1 elements appear to be truncated or to lie at the end of a contig, perhaps reflecting difficulties in contig assembly posed by repeat elements). Using PCR primers designed to the regions predicted to flank this element we were able to obtain a related empty site from another strain. Comparison of the inserted and related empty site sequences (Fig. 2b) confirms that the termini of the element are as predicted. Interestingly, we found that the position in the related empty site that corresponds to the position of the crypton insertion contains a 4 bp sequence, 5'-TGTT-3', that is repeated at each end of the element. The presence of this short sequence at both termini might result from the element inserting into the host genome via a recombination reaction between a 4 bp donor sequence in the element and a similar sequence in the target site (see Discussion).

The termini of several other cryptons Cn could also be defined. A second example of a comparison between the sequences of inserted and related empty sites is shown for crypton Cn2, one copy of which is inserted in a retrotransposon LTR (Goodwin & Poulter, 2001b; Fig. 2c). Once again, the related empty site can be seen to contain a short sequence, 5'-TGTT-3', which is repeated at the ends of the element. Interestingly, the extreme termini of this element are very similar in sequence to those of crypton Cn1 (Fig. 2), despite the two elements not being highly similar in sequence overall. This conservation in sequence suggests that the terminal sequences have an important role in the transposition of these elements.

For several cryptons Cn the termini could not be defined with the available data, presumably either because these elements exist as just single copies in the sequencing project strains, or because the sequences of alternative insertion sites were excluded during the contig assembly processes. The similarity in structures between these elements, and...
those for which many of the characteristic features of mobile elements are apparent (see below), nevertheless suggests that they are also likely to be transposons.

Structures of cryptons

Several apparently full-length and potentially intact crypton Cn sequences are available in the various C. neoformans sequence databases. To confirm the available data we completely resequenced a copy of crypton Cn1 from strain JEC21. Our sequence (GenBank accession no. AY248893) was found to be identical to that in the JEC21 sequence database (http://www.tigr.org/tdb/e2k1/cna1/). The element is ~3-8 kb long and contains a single long gene interrupted by a number of introns (Fig. 3). The three introns closest to the 3' end were detected by comparison with cDNA sequences, as described above. Two further introns at the 5' end were predicted by a combination of

![Fig. 2. The termini of cryptons. (a) An alignment of four 5' termini and two 3' termini of crypton Cn1. (b) Comparison of the sequences flanking a crypton Cn1 insertion and the sequence of a related empty site obtained by PCR. (c) Comparison of the sequences flanking a crypton Cn2 insertion and the sequence of a related empty site. (d) Termini and flanking sequences of some Coccidioides crypton Cp1 elements. (e) Termini and flanking sequences of some Histoplasma crypton Hc1 elements. In all panels, residues corresponding to the cryptons are shown in boldface. Short direct repeats at the extreme termini are underlined. In some cases these are not perfect repeats (see Discussion). Flanking sequences are shown in standard type. Dashes represent internal regions of the cryptons which are not shown. Names of the sequences are shown on the right. Some of the names in panels (a), (d) and (e) refer to contigs from the relevant sequencing project.

![Fig. 3. Structures of cryptons. Thick lines represent 5' and 3' untranslated regions; narrow lines represent flanking sequences. Boxes represent coding regions; intervening V-shaped lines represent introns. Stippled areas correspond to the tyrosine recombinase domains. Hatched areas correspond to the putative DNA-binding domains. Vertical lines within the coding regions indicate premature stop codons. A split, boxed triangle represents a retrotransposon LTR into which crypton Cn2 has become inserted. Question marks indicate that the locations of the termini of some elements have not been determined. Note that some small 5' exons may have escaped detection, so the 5' ends of the coding regions may not be exactly as shown. The particular elements depicted are those identified in Methods.](http://mic.sgmjournals.org)
methods, including intron-prediction software optimized for *C. neoformans* (http://www.tigr.org/tdb/glimmerm/glimr_form.html) and comparisons with putative coding regions of the other cryptons *Cn*. [We should note, however, that some additional 5′ introns might have escaped detection, so the exact 5′ end of the coding region cannot be identified with certainty at present.] Removal of the identified introns unites six exons into a single long ORF with the potential to code for a 691-amino acid residue protein (starting from the first available Met codon). The three other full-length crypton *Cn* elements have structures similar to that of crypton *Cn1* (Fig. 3). The reconstructed ORFs of three of the elements appear to be intact. In the fourth the ORF is disrupted by a single stop codon in the fifth exon.

The tyrosine recombinase domains lie in the central parts of the crypton coding regions (Fig. 3). The coding regions upstream of these domains contain little in the way of conserved sequences (Fig. S1 in the supplementary data available with the online version of this paper at http://mic.sgmjournals.org.uk), and do not have any significant similarity to previously described proteins. The majority of the regions downstream of the recombinase domain are also poorly conserved in sequence among the elements (Fig. S1) and have no significant matches to previously identified sequences. The C-terminal-most region of each crypton gene, however, contains a conserved domain which is similar in sequence to several fungal transcription factors (Fig. 1b). The matching proteins include *Saccharomyces cerevisiae* Msn1p (multi-copy suppressor of SNF1; Estruch & Carlson, 1990), Hot1p (high osmolarity-induced transcription factor; Rep et al., 1999), and Gcr1p (transcriptional regulator of glycolytic genes; Holland et al., 1987) and the *Kluyveromyces lactis* homologue (KIGcr1p; Haw et al., 2001). In the case of Gcr1p, the region that matches the crypton proteins has been shown to bind DNA with high affinity (Huie et al., 1992; Huie & Baker, 1996). Msn1p is also known to be a DNA-binding protein (Estruch & Carlson, 1990). The observed sequence similarity (Fig. 1b) suggests that the C-terminal domains of the crypton proteins are likely to be involved in binding to DNA. None of the other genes commonly found in TEs, such as DDE-type transposases/integrase proteins are likely to be involved in binding to DNA.

Unlike the majority of known TEs, none of the cryptons was found to contain extensive repeats, either inverted or direct, at their termini, or at any other location. Indeed, the only repeat sequences consistently found with cryptons were the short sequences, mentioned above, which appear as direct repeats at each extremity.

**Southern analysis**

The number of crypton *Cn1* elements in each of a variety of *C. neoformans* strains was estimated by Southern blotting (Fig. 4). For this analysis we included strains of both serotypes A and D, as well as an apparent mixed-serotype (AD) strain. The probe was derived from the crypton *Cn1* element that we sequenced from serotype D strain JEC21. The location of the probe, and the restriction enzyme used to cut the genomic DNA, were chosen so as to detect only those fragments that contain the 3′ boundaries between the elements and the flanking genomic sequences. In this situation the numbers of bands detected provide estimates of the element’s abundance. We found that the number of bands, and their sizes and intensities, varied among the different strains. The two serotype D strains both gave relatively bright bands. In one strain, six distinct bands, of varying intensity, were apparent. The variable band intensities might result from the brighter bands representing more than one element, or the fainter bands corresponding to diverged or partial copies. The other serotype D strain gave just a single, relatively bright band, of a different size to any detected in the first strain. Two serotype A strains were analysed (Fig. 4, and not shown). Each gave just a single relatively faint band, of the same size in each strain. The relative faintness of the bands might be due to divergence of the elements in these strains from the serotype D-derived
element used as the probe. The mixed-serotype (AD) strain had an identical banding pattern to the serotype A strains, suggesting that at this locus it is more like serotype A than serotype D strains.

The apparent multi-copy nature of crypton Cnt1 in some strains, and the interstrain variation in banding pattern, are as expected for a transposable element. The differences in banding pattern between serotypes A and D, and among serotype D strains, suggest that the element has been mobile since these serotypes/strains diverged.

Cryptons in other species

Several crypton-like elements were detected in the genome sequences of the ascomycetes Coccidioides posadasii and Histoplasma capsulatum. The elements from C. posadasii all appear to be members of the one family, crypton Cp1, typically sharing 85–95% identity at the DNA level with each other. The elements are ~4·4 kb long, appear at various locations in the C. posadasii genome, have distinct termini, and at the extremities of these termini are short (sometimes imperfect) direct repeats (Fig. 2d). In several cases related empty sites could be detected, suggesting recent mobility. The elements are very similar in structure to their counterparts in C. neoformans (Fig. 3), and encode similar tyrosine recombinases and putative DNA-binding domains (Fig. 5; see also Fig. S1 in the supplementary data available with the online version of this paper at http://mic.sgmjournals.org.uk). An interesting difference, however, is that while introns are common in the C. neoformans elements, they appear to be absent from (or rare in) the elements of C. posadasii.

The elements from H. capsulatum are most similar in sequence and structure to those from their fellow ascomycete C. posadasii (Figs 3 and 5). They also have well-defined termini and appear at multiple sites in the genome (examples in Fig. 2e). The extreme termini are short direct repeats, and, as with the elements from C. neoformans, comparisons with related empty sites reveal short regions (typically 3–8 bp) of sequence similarity between the termini and the target sites (not shown). Cryptons appear to be fairly abundant in H. capsulatum, with one family, crypton Hc1, having 35–40 distinct copies in strain G186AR and about 10 in strain G217B. None of the insertion sites are shared between the two strains, suggesting that the elements have been very active since these strains diverged from a common ancestor.

The coding regions of all the detected cryptons from C. posadasii and H. capsulatum are corrupted by multiple in-frame stop codons (examples in Figs 3 and 5). Close examinations of the sequences (not shown) suggest that these mutations are not the result of the elements being long-dead remnants of once-active elements, which have, over time, accumulated many debilitating mutations. Rather, it appears that these are recently active elements.
which have been subject to a hyper-mutational process akin to the RIP (repeat-induced point mutation) system first described in the ascomycete Neurospora crassa (Cambareri et al., 1989). In Neurospora, RIP identifies any sequences longer than ~400 bp (Watters et al., 1999) and with greater than ~80% identity (Cambareri et al., 1991) that appear at more than one copy in the genome, and introduces a large number of G-C to A-T transitions in both copies of the repeated sequence. This process occurs during the sexual cycle and prefentially causes C to T transitions at CpA dinucleotides. RIP is thought to act as a defence against the proliferation of TEs. The features that suggest that a process similar to RIP has acted on the cryptons in C. posadasii and H. capsulatum will be described in more detail elsewhere (T. J. D. Goodwin & R. T. M. Poulter, unpublished), but, briefly, they are that the vast majority of the differences among the elements are transitions, with other mutations (transversions, insertions, deletions) being exceedingly rare; the elements exhibit extreme depletions in the frequency of CpG dinucleotides, and the differences in the coding capacities of the various C. posadasii and H. capsulatum cryptons are consistent with a high frequency of transitions at CpG dinucleotides. These elements thus exhibit many of the features of Neurospora sequences that have been subject to RIP, with the exception that in C. posadasii and H. capsulatum the preferential target for the process appears to be CpG dinucleotides, rather than CpA.

A sequence related to cryptons was also detected in the ascomycete Candida albicans. This sequence contains a long ORF, the product of which matches the predicted protein from near the start of the tyrosine recombinase domain to the end of the putative DNA-binding domain (Fig. 5, Fig. S1). This C. albicans sequence does not appear to be repetitive, however, and displays no evidence of recent mobility. In addition, the predicted protein lacks several of the highly conserved residues of the tyrosine recombinase that are present in the other elements. This C. albicans ORF may no longer be part of a mobile element, but instead may have been domesticated by the host to perform some other function.

The appearance of cryptons in both ascomycetes and basidiomycetes suggests that these elements are an ancient component of fungal genomes, probably in existence prior to the divergence of these two phyla, approximately 400 million years ago (Berbee & Taylor, 1993). Cryptons were not, however, detected in several other fungi whose genomes have been sequenced, such as the ascomycetes Saccharomyces cerevisiae, Schizosaccharomyces pombe and N. crassa, or the basidiomycete Phanerochaete chrysosporium, indicating that cryptons have a patchy distribution, and have probably been lost from a number of lineages which used to contain them.

Finally, we should note that we found several cryptons in GenBank sequences (including AAAA01023386 and AAAA01025685) that are annotated as being derived from the indica cultivar of rice (Oryza sativa). These sequences were produced as part of the effort to sequence the indica rice genome (Yu et al., 2002). We have several reasons, however, for believing that these elements might not be genuine rice sequences, but instead might represent fungal DNA contaminating the indica rice sequence data: (1) several distinct copies of the element appear in the database, but all are on short contigs and none are physically linked to any sequences that are clearly genuine rice sequences; (2) no similar elements were detected in the virtually complete sequence of the japonica cultivar of rice, nor in any other available plant sequence; and (3) we have identified a number of other TE sequences in the indica rice database which are much more similar to fungal elements than to any other known plant element. For instance, the sequence with accession no. AAAA01039796, annotated as being derived from indica rice, contains a partially transposase sequence which is 83% identical over 244 amino acid residues to a transposon from the plant-pathogenic fungus Cochliobolus carbonum (not shown), but which is only weakly similar to known plant sequences.

Evolutionary analyses

The relationships between previously identified tyrosine recombinases, and those identified here in cryptons, were studied by phylogenetic analysis. Trees were constructed based on an alignment of the regions encompassing the RHRY tetrads of a large number of bacterial, archaeal and phage recombinases, and representatives of the diversity of tyrosine recombinases found in eukaryotes. These latter sequences include such enzymes from the DIRS1 and Ngarol groups of LTR retrotransposons, yeast 2-micron circle plasmids, and the recently described Tec transposons of Euplotes crassus, as well as some tyrosine recombinase-like sequences from baculoviruses and mitochondria. A representative tree obtained by the neighbour-joining method is shown (Fig. 6). As has been noted earlier (Esposito & Scocca, 1997; Goodwin & Poulter, 2001a; Jacobs et al., 2003), phylogenetic analyses of tyrosine recombinases are hindered by the high level of sequence diversity within the group. This usually means that the nature of the more distant relationships within the group cannot be resolved with certainty using current techniques. Nevertheless, on trees constructed by a variety of methods, we found that the crypton recombinases consistently formed a well-supported monophyletic group within a larger clade that contains the recombinases of the DIRS1 and Ngarol LTR retrotransposons, the Cre recombinase of bacteriophage P1 and several other prokaryotic sequences. This grouping, while it does not receive high levels of bootstrap support, is in agreement with the sequence similarities between the crypton recombinases and those of the retrotransposons and related prokaryotic elements, that was noted earlier. The crypton recombinases do not appear to be closely related to the other eukaryotic recombinases, including those of the Tec DNA transposons from E. crassus. They also do not appear to be closely related to the recombinases of prokaryotic transposons which are thought to employ the enzyme for
a transposase (rather than a resolvase) function, such as Tn916 from Enterococcus faecalis (Storrs et al., 1991).

Within the crypton group the relationships among the elements are consistent with the host phylogeny: the elements from the ascomycetous fungi (C. posadasii and H. capsulatum) group together (100% bootstrap support), as do all the elements from C. neoformans (93% support). The crypton identified in indica rice sequences, which may be derived from a fungal contaminant, was also included on the tree. This element groups among the fungal sequences, and appears to be more closely related to the C. neoformans cryptons than to those from the ascomycetes.

**DISCUSSION**

Cryptons are a novel type of transposon, quite unlike any previously described elements. Various features of these elements, which we have outlined above, indicate that they are transposable elements. However, the presence of spliceosomal introns and the absence of RT and RNH genes indicate that, unlike the majority of tyrosine recombinase-encoding eukaryotic TEs, cryptons are not retrotransposons. Cryptons are also very different from the only previously described eukaryotic, tyrosine-recombinase-encoding DNA transposons – the Tec elements of E. crassus. The two best-characterized Tec elements, Tec1 and Tec2, are highly abundant (~12,000 copies per genome), are flanked by long terminal inverted repeats (TIRs), and in addition to the tyrosine recombinase, they encode a typical Tc1/mariner DDE-type transposase (Doak et al., 2003). Tec3 is less well characterized, but is also known to have long TIRs (Jacobs et al., 2003). In contrast, cryptons appear at lower copy numbers, they do not have TIRs, do not encode a DDE-type transposase, and they contain some protein-coding domains that are not found in Tec elements. Phylogenetic analyses

![Diagram](http://mic.sgmjournals.org.uk)
also do not suggest a close relationship between cryptons and Tec elements.

Several lines of evidence instead suggest that cryptons may be related to the DIRS1 and Ngaro1 groups of LTR retrotransposons. For instance, comparisons of tyrosine recombinase sequences reveal blocks of conserved amino acids shared by these elements, in addition to those common to the majority of these enzymes. Phylogenetic analyses group the crypton recombinases into a putative clade that includes the DIRS1 and Ngaro1 recombinases (as well as several bacterial enzymes and Cre from bacteriophage P1). Analyses of the termini and the insertion sites of cryptons also reveal some interesting similarities with DIRS1 and Ngaro1 retrotransposons. For instance, for many cryptons we found that the extreme termini contain a short sequence (4–6 bp) which is repeated at either end, and this same sequence appears in the related empty sites. A similar situation was found for DIRS1-like elements (Goodwin & Poulter, 2001a) and also occurs with at least some elements from the Ngaro1 group (T. J. D. Goodwin, unpublished data). For the DIRS1 elements, we interpreted these findings as suggesting that the extrachromosomal transposition intermediates of these elements integrate into the host genome by recombination between a short sequence at the circular junction of the element’s termini and an identical sequence in the target site. For the cryptons we propose (in part by analogy with the transposition of prokaryotic elements) that an element might transpose in a similar fashion, as follows. (1) A copy of the element is excised from the host genome by recombination, presumably involving the encoded tyrosine recombinase. The recombination reaction occurs at the short direct repeats flanking the element; the recognition of these sequences by the recombinase might be assisted by additional sequences in the immediately subterminal regions. The result is a circular, extrachromosomal, double-stranded DNA molecule. The circular junction of the crypton’s termini would contain a single copy of the repeat which flanks the integrated copies (we refer to the single copy of this sequence in the extrachromosomal intermediate as the ‘donor’ sequence). (2) The extrachromosomal DNA (and associated enzymes) moves to a new chromosomal site that contains a sequence identical (or very similar) to the donor sequence (referred to as the ‘target’ sequence). (3) The crypton is inserted into the new site by recombination between the donor and the target sequences. If the donor and target sequences are identical then the element ends up flanked by perfect direct repeats. Differences between the donor and target sequences would result in the element being flanked by imperfect repeats.

Whatever the exact mechanism, the data suggest that transposition/integration may be similar in the cryptons and the tyrosine-recombinase-encoding retrotransposons. The similarities in sequence, and the putative mechanistic similarities, between cryptons on the one hand, and DIRS1 and Ngaro1 retrotransposons on the other, might suggest that these elements are evolutionarily related. For instance, the retrotransposons might have arisen from the combination of a crypton-like element and the RT/RNH gene of a pre-existing retrotransposon. Alternatively, cryptons could be descended in part from a DIRS1- or Ngaro1-like element.

Several of the cryptons identified in C. neoformans appear to be intact, as judged by their intact ORFs and the presence of all the expected highly conserved residues. Some also appear to have transposed recently, as suggested by the existence of similar elements with distinct insertion sites, and the identification of empty sites highly similar in sequence to the regions flanking element insertions. The existence of such elements, and their presence in a relatively tractable host such as C. neoformans, should facilitate their experimental characterization. We anticipate that such work could lead to valuable advances in our understanding of recombination and transposition mechanisms, and in the evolutionary processes leading to the great diversity amongst transposable elements. Novel mobile elements such as these might also find applications as insertional mutagens in Cryptococcus and other pathogenic fungi, and/or be useful for strain typing and evolutionary studies.

ACKNOWLEDGEMENTS

We thank members of the following genome sequencing projects for provision of sequence data: the C. neoformans Genome Project, Stanford Genome Technology Center, funded by the NIAID/NHL under cooperative agreement AI47087; and The Institute for Genomic Research, funded by the NIAID/NHL under cooperative agreement U01 AI48594, for C. neoformans serotype D genomic sequence data; the C. neoformans H99 sequencing project, Duke Center for Genome Technology, and the Genome Sequence Centre, BC Cancer Research Centre, for C. neoformans serotype A genomic sequence data; the Cryptococcus neoformans cDNA Sequencing Project, NIH-NIAID grant number AI147079 and Bruce A. Roe, Doris Kupfer, Heather Bell, Sun So, Yuong Tang, Jennifer Lewis, Sola Yu, Kent Buchanan, Dave Dyer and Juneann Murphy at the University of Oklahoma, for C. neoformans cDNA sequence data; The Institute for Genomic Research for C. posadasii genomic sequence data; and the Genome Sequencing Center at Washington University in St Louis for H. capsulatum genomic sequence data. Sequence data for Candida albicans were obtained from the Stanford Genome Technology Center website at http://www-sequence.stanford.edu/group/candida. Sequencing of Candida albicans was accomplished with the support of the NIDR and the Burroughs Wellcome Fund. Phanerochaete chrysosporium sequence data were provided freely by the Joint Genome Institute for use in this publication/correspondence only. T.G. is supported by a Research Grant to Microbiology Research from the New Zealand Lottery Grants Board.

REFERENCES


