Available online at www.sciencedirect.com

SCIENCE DIRECT®

Fungal Genetics and Biology 41 (2004) 657–666

FUNGAL GENETICS AND BIOLOGY

www.elsevier.com/locate/yfgbi

BAC end sequences and a physical map reveal transposable element content and clustering patterns in the genome of *Magnaporthe grisea*

Michael R. Thon,[a] Stanton L. Martin,[a] Stephen Goff,[b] Rod A. Wing,[c] and Ralph A. Dean[a,*]

[a] *Center for Integrated Fungal Research, Department of Plant Pathology, North Carolina State University, Raleigh, NC 27695-7251, USA*
[b] *Syngenta, 3054 Cornwallis Rd, Research Triangle Park, NC 27709-2257, USA*
[c] *Arizona Genomics Institute, Department of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA*

Received 21 May 2003; accepted 19 February 2004
Available online 18 March 2004

Abstract

Transposable elements (TEs) are viewed as major contributors to the evolution of fungal genomes. Genomic resources such as BAC libraries are an underutilized resource for studying genome-wide TE distribution. Using the BAC end sequences and physical map that are available for the rice blast fungus, *Magnaporthe grisea*, we describe a likelihood ratio test designed to identify clustering of TEs in the genome. A significant variation in the distribution of three TEs, MAGGY, MGL, and Pot2 was observed among the fingerprint contigs of the physical map. We utilized a draft sequence of *M. grisea* chromosome 7 to validate our results and found a similar pattern of clustering. By examining individual BAC end sequences, we found evidence for 11 unique integrations of MAGGY or MGL into Pot2 but no evidence for the reciprocal integration of Pot2 into another TE. This suggests that: (a) the presence of Pot2 in the genome predates that of the other TEs, (b) Pot2 was less transpositionally active than other TEs, or (c) that MAGGY and MGL have integration site preference for Pot2. High transition/transversion mutation ratios as well as bias in transition site context was observed in MAGGY and MGL elements, but not in Pot2 elements. These features are consistent with the effects of a Repeat-Induced Point (RIP) mutation-like process occurring in MAGGY and MGL elements. This study illustrates the general utility of a physical map and BAC end sequences for the study of genome-wide repetitive DNA content and organization.

Index Descriptors: Transposon; Transposable element; Rice blast; *Magnaporthe grisea*; *Pyricularia grisea*; BAC library; Physical map

1. Introduction

The filamentous fungus *Magnaporthe grisea* is the causal agent of rice blast disease, one of the most important pathological threats to rice supplies worldwide (Ou, 1987). It has been the focus of intense genetic and molecular biological studies that have increased our understanding of the molecular determinants of pathogenesis and biology for this and related fungi. Studies of genetic diversity among field isolates of the fungus have resulted in the identification of several classes of transposable elements (TEs) within the genome (Farman et al., 1996b; Kachroo et al., 1995; Skinner et al., 1993). These highly repeated sequences have made useful probes for identification of restriction fragment length polymorphisms. In addition, some TEs show a restricted distribution among strains that follows host range (Borromeo et al., 1993; Dobinson et al., 1993; Hamer et al., 1989). This work has led to an interest in the study of the role of TEs in the evolution of the genome of *M. grisea*.

Repetitive DNA elements have been shown to make up a considerable portion of eukaryotic genomes and

[*] Corresponding author. Fax: 1-919-513-0024.
E-mail address: ralph_dean@ncsu.edu (R.A. Dean).

1087-1845/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.fgb.2004.02.003

have elicited a great deal of interest because of their potential effects on genome structure and mutation (Kidwell and Lisch, 1997). Within the genome of *M. grisea*, a considerable portion of the repetitive sequences are transposable elements (TEs). TEs have been implicated as a major source of genetic mutations in the *M. grisea* genome. Their repetitive nature can serve as recombination sites and their ability to transpose can cause insertional mutations. Mutations caused by transposons are known to affect pathogenicity and host range of *M. grisea* (Kang et al., 2001).

TEs can be divided into two classes, depending on their mechanism of transposition (Kidwell and Lisch, 1997). Class I elements transpose via an RNA intermediate employing reverse transcriptase. At least three types of class I elements have been described in *M. grisea*. The most commonly reported belong to the long terminal repeat (LTR) retrotransposons group and contain one or two genes (Gag and Pol) and terminal repeats. Other groups of class I elements include the long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). SINEs typically are less than 500 bp and contain an RNA polymerase III promoter but lack open readings frames while LINEs are longer and encode a reverse transcriptase. Class II elements do not utilize an RNA intermediate, contain terminal inverted repeats and have at least one open reading frame that encodes a transposase.

While several studies have characterized variation in the distribution of TEs among strains of the fungus, little is known about the distribution of TEs within the genome. A growing body of evidence shows that TEs are not distributed randomly in the genomes of many species, and may be localized to specific chromosomal landmarks, such as centromeres or intergenic regions (Bartolomé et al., 2002; Daboussi and Capy, 2003; El Amrani et al., 2002). In the plant pathogenic fungus *Fusarium oxysporum*, class II TEs appear to be arranged in tightly packed clusters (Hua-Van et al., 2000). Evidence from hybridizations studies in *M. grisea* suggests that it has a similar arrangement of some TE families. In the process of cloning and characterizing TEs in *M. grisea*, several authors have noted that fragments of other TEs often occur on the same genomic clones, suggesting that TEs may be clustered in the genome (Kang, 2001; Nishimura et al., 1998; Nitta et al., 1997; Shull and Hamer, 1996). In addition, hybridization studies using genomic libraries have shown that TEs tend to be clustered on BAC and cosmid clones (Nishimura et al., 1998, 2000; Nitta et al., 1997; Zhu et al., 1997). These results suggest that TE integrations may be subject to site specificity or site preference during transposition.

Whole genome sequences represent the ultimate resource for studying the distribution of TEs within genomes. However, genomic resources such as BAC libraries and fingerprint contigs can be better exploited to understand TE distribution and genome organization. In this study, we describe a technique to identify clustered TE distribution using BAC end sequences and a physical map for *M. grisea*. Our results show that the three most common TEs, MAGGY, MGL, and Pot2, are not distributed randomly among the fingerprint contigs and appear to be clustered in distinct regions of the chromosomes. We validated our results by identifying a similar pattern of TE clustering in a draft sequence the *M. grisea* chromosome 7. We observed in the BAC end sequences integration patterns that suggest the presence of the TE Pot2 prior to the invasion of MAGGY and MGL. In addition, these data provided us with an opportunity to evaluate the evidence for the presence of Repeat-Induced Point mutation (RIP) in the *M. grisea* genome. Our results show evidence for the presence of RIP in MAGGY and MGL, but not in Pot2. This study demonstrates the utility of a BAC library and BAC end sequences for studying genome-wide TE organization.

2. Materials and methods

2.1. Magnaporthe grisea physical map and sequences

The BAC library and other genomic resources are derived from *M. grisea* strain 70–15, a domesticated strain that is the result of a breeding program designed to improve mating competence (Chao and Ellingboe, 1991; Lau et al., 1993). It is derived from a cross between the rice infecting isolate Guy 11 and a weeping lovegrass isolate followed by several backcrosses to Guy 11.

The *M. grisea* BAC library and physical map were reported previously (Zhu et al., 1997). The BAC library contains 9216 clones with an average insert size of 130 kb and represents 25× coverage of the *M. grisea* genome. The BAC clones were assembled into fingerprint contigs by digesting the clones with *Hin*dIII, estimating fragment sizes on agarose gels, and assembling the resulting fingerprints into contigs using the software package FPC (Zhu et al., 1997). By hybridizing genetic markers (Nitta et al., 1997) to the BAC library, specific clones and their corresponding fingerprint contigs were assigned to chromosomes. End sequences were derived from both ends of the clones in the BAC library. BAC DNA preparation and sequencing was performed as described previously (Mao et al., 2000). After base calling using the program phred, the low quality bases were trimmed from individual sequencing reads (Ewing et al., 1998). Sequencing reads that did not have at least 100 bases of phred quality value 20 or higher were removed from the analysis. After removing the low quality sequence reads and vector sequences, 15,209 sequences were retained for further analysis.

Genomic sequences derived from end sequencing the BAC library are limited to sites flanking *Hin*dIII sites in the genome. Since the genome size of *M. grisea* is estimated to be 40 Mb and *Hin*dIII sites should, if randomly distributed in the genome, be found on average every 4096 bp, then it is estimated that there are 9765.6 *Hin*dIII sites in the genome and 19,531 potential sites that can be obtained by BAC end sequencing. Since the cloned sites are sampled 'with replacement,' the probability that any specific site was sequenced at least once (its inclusion probability) is $1 - (1 - (1/N))^m = 0.541$ where N equals the total number of potential sites to be sequenced in the genome and m equals the number of samples taken. The expected number of genomic sites present in the BAC end sequences is $0.541 \times 15,209 = 8228$. Some sites have therefore been sequenced multiple times and the average number of sequences per site (depth of coverage) is 1.8.

The physical map, genetic map, and BAC end sequences have been integrated into a searchable database by Martin et al. (2002). The chromosome 7 sequence is a draft assembly obtained from the sequencing project that is currently underway in our laboratory. It is based on 5X shotgun sequences of BAC clones and contigs from the version 2 assembly of the whole genome shotgun sequence provided by the Whitehead Institute Center for Genome Research (http://www.broad.mit.edu/annotation/fungi/magnaporthe). The BAC clone sequences are available from GenBank and the draft assembly of chromosome 7 is available at http://www.fungalgenomics.ncsu.edu.

2.2. Transposable element identification

GenBank accessions for TEs known to occur in the genome of *M. grisea* were used as reference sequences (Table 1). Using the reference sequences as queries, we used the FASTA version 3.4 software package (Pearson, 2000), with the -A command line parameter to utilize the Smith–Waterman alignment algorithm, to identify BAC end sequences containing known TEs. BAC end sequences that aligned to a query sequence over a length of at least 75 bases and 80% sequence similarity were considered matches. The match criteria were determined empirically by manually examining the FASTA search results. To identify novel TEs, a library of 63 known TE sequence from filamentous fungi (Daboussi and Capy, 2003) was searched using TBLASTX (Altschul et al., 1997) using the default parameters.

2.3. Description of the test statistic

We developed a likelihood ratio technique in order to determine whether the frequency of TE integrations varied among the fingerprint contigs. The observed data was divided into fingerprint contigs comprised of BAC end sequences that may or may not contain TEs. Let t be the total number of fingerprint contigs and s_i be the total

Table 1
Occurrence of TEs in *M. grisea* BAC end sequences identified by sequence similarity searches

TE	Number of BAC end sequences	Reference sequence	
		GenBank Accession No.	References
Class I			
LTR retrotransposons			
MAGGY	1277	L35053	Farman et al. (1996b)
Grasshopper	0	M77661	Dobinson et al. (1993)
fosbury[a]	N/A	U15189, U15190	Shull and Hamer (1996)
Pyret	167	AB062507	Nakayashiki et al. (2001)
MGLR-3	37	AF314096 (bases 1121–7462)	Kang (2001)
Occan	11	AB074754	Kito et al., unpublished GenBank entry
LINE like elements			
MGL (MGR583)	437	AF018033	Meyn et al, unpublished GenBank entry
SINE like elements			
Mg-SINE	167	U35313	Kachroo et al. (1995)
MGSR-1	N/A[b]	S65049	Sone et al. (1993)
Class II			
Pot2	288	Z33638	Kachroo et al. (1994)
Pot3 (MGR586)	26	U60989	Farman et al. (1996a)
Unclassified			
MGR608	12	U36923 (bases 202–327)	Kang et al. (1995)
MGR619	5	U36923 (bases 328–406)	Kang et al. (1995)

[a] Only fosbury LTR sequences are available, and the sequences are identical to MAGGY LTRs.

[b] Based on sequence similarity to Pyret (Nakayashiki et al., 2001) and examination of BAC end sequences, we conclude that MGSR-1 is a subsequence of Pyret and removed it from this analysis. See text for details.

number of BAC end sequences collected from fingerprint contig i. Among these s_i sequences, the number that contain a TE can be treated as a random variable N_i. The observed value of the random variable N_i will be denoted by n_i. To represent the probability that a randomly selected BAC end sequence from fingerprint contig i contains a TE, we will use R_i. The expected value of N_i is $R_i s_i$. Because the frequency of TE integrations was low, the distribution of N_i can be approximated with a Poisson distribution

$$P(N_i = x) = \frac{\mathrm{e}^{-R_i s_i}(R_i s_i)^x}{x!}.$$

We considered the null hypothesis that postulates TE integrations have no tendency to cluster within the genome. According to this hypothesis, $R_1 = R_2 = \cdots = R_t$ and we used R to represent the probability that is shared among contigs. A maximum likelihood estimate of R for this null hypothesis is

$$\hat{R} = \frac{\sum_{i=1}^{t} n_i}{\sum_{i=1}^{t} s_i}$$

and the value of the log-likelihood at this maximum likelihood estimate is

$$\log L_0 = \sum_{i=1}^{t} \log\left(\frac{\mathrm{e}^{-\hat{R}s^i}\left(\hat{R}s_i\right)^{n_i}}{n_i!}\right).$$

The alternative hypothesis postulates that each contig can have its own value of R_i. For the alternative hypothesis, the maximum likelihood estimate of R_i is

$$\hat{R}_i = \frac{n_i}{s_i}.$$

The value of the maximum log-likelihood for the alternative hypothesis is

$$\log L_A = \sum_{i=1}^{t} \log\left(\frac{\mathrm{e}^{-n_i} n_i^{n_i}}{n_i!}\right).$$

As a test statistic, we adopted $L = \log(L_A/L_0) = \log L_A - \log L_0$.

The distribution of the test statistic under the null hypothesis of no tendency for clustering can be approximated via a parametric bootstrap approach in which 1000 data sets are simulated. For each simulated data set, the 'observed' number of BAC end sequences in contig i that contain a TE was sampled from a Poisson distribution with mean $\hat{R}s_i$. Following simulation of each of the 1000 data sets, the test statistic value corresponding to that data set was computed. By comparing the observed value of the test statistic to its simulated null distribution, the null hypothesis of no clustering can be evaluated. A Perl script to implement the test statistic calculations and parametric bootstrap approach is available upon request from the authors.

2.4. Multiple sequence alignments and phylogenetic analysis

Sequences were aligned with ClustalW (Thompson et al., 1994) and Jalview (http://www.ebi.ac.uk/jalview/) was used to manually inspect and edit the alignments. We computed percent similarity and transition/transversion (t/v) ratios using PAUP* version 4.0.0 included in the Wisconsin Package version 10 (Genetics Computer Group, Madison, WI) using the default parameters.

3. Results

3.1. TE content of BAC end sequences

We obtained reference sequences of 11 known TEs in the genome of *M. grisea* from GenBank (Table 1) and verified their identity by investigating the relevant literature and by performing BLASTN searches to the GenBank non-redundant nucleotide database, nt. Several of the GenBank entries contained genomic sequences flanking the repetitive elements but only the regions annotated in the GenBank entries as the repetitive element were used in this analysis. During this investigation, we found that the SINE-like element MGSR-1 was over 91% similar to the retrotransposon Pyret suggesting that either MGSR-1 is, in fact, a fragment of Pyret, or that the sequence of Pyret in GenBank contains an insertion of MGSR-1 (Nakayashiki et al., 2001). To test the first of these two hypotheses, we attempted to identify copies of MGSR-1 in the BAC end sequences that do not have flanking Pyret sequences. Using MGSR-1 as a query sequence in a FASTA search of the BAC end sequences, we identified 12 BAC end sequences with significant matches. However, after comparing them to the Pyret reference sequence, we concluded that all 12 BAC end sequences were derived from Pyret and not from independent insertions of MGSR-1. Thus, MGSR-1 appears to be a fragment of Pyret and not a SINE element. We also found that the 3′ end of the SINE element Mg-SINE and the 3′ end of the LINE element MGL are over 99% similar, suggesting that these two elements comprise a LINE–SINE pair. Such pairings have been described in several other organisms and has lead to the hypothesis that the reverse transcriptase encoded by the LINE element is also active in reverse transcribing its partner SINE (Okada et al., 1997).

We utilized the FASTA program to search 15,209 *M. grisea* BAC end sequences using the set of reference sequences as queries and identified 2427 (15.9%) BAC end sequences with significant similarity to the known TEs (Table 1). Over 52% of these (1277 sequences) matched the retrotransposon MAGGY, which was the most abundant. The LTR regions of MAGGY are over

99% similar to the LTRs of fosbury (Shull and Hamer, 1996). Because a full length sequence of fosbury has not been reported, we were not able to distinguish sequences belonging to fosbury. The second and third most abundant TEs were the LINE element MGL (437 BAC ends) and the class II element Pot2 (288 BAC ends), respectively. Together, these three elements made up over 82% of all the TE occurrences that we identified in our study. It should be noted that the occurrences of TEs identified among the BAC end sequences cannot be directly used as an indication of TE copy number in the genome. TE length, and number of *Hin*dIII sites within each element can greatly affect the observed number of TEs.

To identify novel TEs among the BAC end sequences, we created a database of 63 known TE sequences from filamentous fungi (including those from *M. grisea* listed in Table 1) as reported by Daboussi and Capy (2003) and searched the database using TBLASTX. In addition to the known *M. grisea* TEs already identified using the FASTA algorithm, this strategy allowed us to identify several novel TEs. Using an e value cutoff of 1e – 20, we found 6 BAC end sequences with similarity to *REAL*, a retrotransposon from the genome of *Alternaria alternata*. Closer examination of these BAC end sequences using BLASTN searches revealed that they were nearly identical to *RETRO6* and *RETRO7*, novel retrotransposons from the genome of *M. grisea* recently identified by Farman (unpublished results). Twenty-five BAC end sequences were found with similarity to *grasshopper (grh)*, a retrotransposon identified from *M. grisea* but not from rice-infecting strains. These too, when more closely examined, were found to be copies of *RETRO6* and *RETRO7*. A single BAC end sequence with similarity to *Hop* was also identified, consistent with the results of Chalvet et al. (2003) who report a single degenerate copy in the genome sequence of *M. grisea*.

3.2. TEs are not randomly distributed in the M. grisea genome

Our initial examination of the distribution of TEs among the fingerprint contigs (Fig. 1) revealed that there was considerable variation in the proportion of BAC end sequences containing TEs among the fingerprint contigs. This suggests that the frequency of TE integration is not the same at all points in the genome and that some regions, represented by the fingerprint contigs, have a higher frequency of TE integration than others. To test this hypothesis, we developed a likelihood ratio method that compares the frequency of TE integrations in each fingerprint contig. The total number of BAC end sequences and the number that contain each of the three most frequently occurring TEs were computed for each fingerprint contig. In each of the three cases, the test statistic (L) was very high, strongly favoring the alternative hypothesis that the frequency of integration differs among the fingerprint

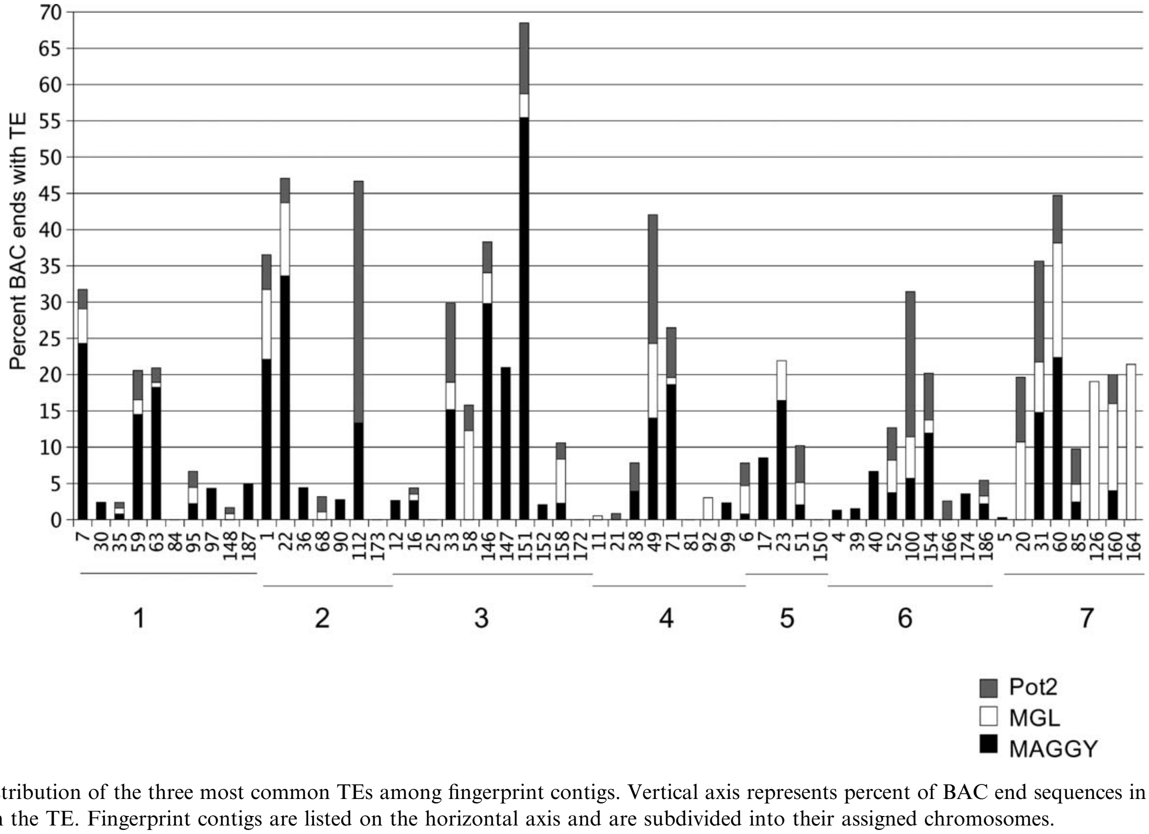

Fig. 1. Distribution of the three most common TEs among fingerprint contigs. Vertical axis represents percent of BAC end sequences in the contig that match the TE. Fingerprint contigs are listed on the horizontal axis and are subdivided into their assigned chromosomes.

Table 2
Test statistics calculated from observed data and 1000 bootstrap simulations

TE	Observed test statistic (L)	Summary of test statistics from 1000 bootstrap simulations		
		Mean	Min	Max
Fingerprint contigs				
MAGGY	383	119	97	150
MGL	160	103	83	129
Pot2	165	105	85	131
Chromosome 7 sequence				
MAGGY	50	21	10	30
MGL	53	20	11	31
Pot2	41	21	11	30

contigs (Table 2). The test statistic was validated by generating 1000 parametic bootstrap replicates that conform to the null hypothesis. For each dataset, the mean frequency of TE observations (R) was computed. Then, for each contig in the dataset, a random sample was drawn from a Poisson distribution where parameter x is equal to the expected number of contigs with TE in contig i (n^i). A summary of the test statistics computed from the bootstrap replicates is shown in Table 2. In each case, the observed value of L was higher than the mean and maximum values of L found in the bootstrap replicates, providing support that the observed values of L were not likely to have been obtained by chance.

We performed a second validation of this dataset by examining a draft assembly of the *M. grisea* chromosome 7 sequence, which was obtained from the chromosome 7 sequencing project presently underway in our laboratory. For this analysis, TEs in the chromosome sequence were identified with RepeatMasker (A. Smit and P. Green, unpublished results, http://ftp.genome.washington.edu/RM/RepeatMasker.html) using the reference sequences described here as the repeat database. The sequence was divided into 100 kb intervals and each interval was treated in the same fashion as the fingerprint contigs. Since the draft sequence still contains gaps, s^i was computed as 100 kb minus the number of gap characters. The resulting test statistics derived from the chromosome 7 sequence dataset also supports the alternative hypothesis that the frequency of TE integrations is not the same among all intervals (Table 2). As shown in Fig. 2, when the fingerprint contigs were aligned to the chromosome 7 draft sequence, there was a strong correlation between the occurrence of TE-rich fingerprint contigs and sequence intervals.

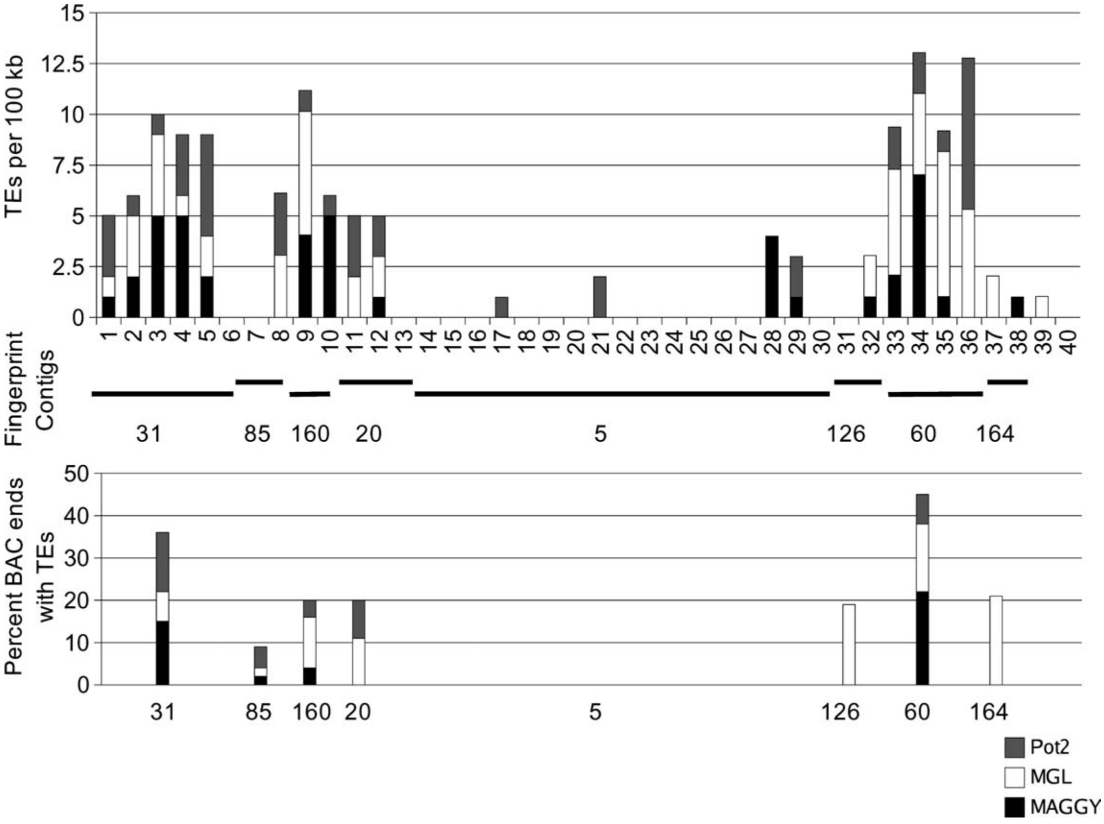

Fig. 2. Distribution of TEs on chromosome 7. Upper graph represents the chromosome 7 draft sequence divided into intervals of 100 kb. Horizontal bars below the graph indicate the locations of fingerprint contigs within the sequence. Lower graph represents the occurrence of TEs within BAC end sequences assigned to the fingerprint contigs.

3.3. TE integration site analysis

The three most common TEs tended to co-occur on the same fingerprint contig. A striking example of this shown in chromosome 7 (Fig. 2) where contigs 31 and 60 contain high levels of all three TEs. This close clustering suggested that TEs may also co-occur in individual BAC end sequences. Observations of two different TEs occurring on one BAC end sequence were rare and only 58 sequences that contain more than one type of TE were identified (Table 3). We selected the 3 most common pairings, MGL-Pot2, MGL-Pyret, and MAGGY-Pot2, for closer examination to identify instances where a TE integrated into a preexisting TE. Since the average number of BAC end sequences per genomic site is 1.8 (see Section 2), we expected that not all of the sequence listed in Table 3 represent unique integration events. We examined the BAC end sequences and the corresponding BAC fingerprints in order to identify those that potentially represent multiple sequences of the same site in the genome. Based on the evidence available to us, we concluded that the 25 sequences we examined represent 17 unique genomic sites. This corresponds to an average depth of coverage of 1.47 sequences per site, which is very close to the expected value of 1.8. Upon examination of the 11 sequences that contain both Pot2 and MGL, we could distinguish 7 unique sites in the genome, one having been sequenced 3 times. Interestingly, all 7 sites suggest the integration of MGL into a preexisting Pot2 element. Five unique genomic sites were evident among the 8 sequences that contain both MGL and Pyret. Four of these suggest that MGL integrated into preexisting Pyret elements while the fifth was not readily distinguishable. The 7 sequences that contain both MAGGY and Pot2 comprise 4 unique genomic sites, all of which suggest that MAGGY integrated into preexisting Pot2 elements.

3.4. Sequence diversity of MAGGY elements

Multiple sequence alignments of the most abundant TE, MAGGY, were performed to characterize sequence diversity among them. Since the BAC library was prepared from *Hin*dIII digested DNA, the three *Hin*dIII sites within the reference MAGGY sequence were identified. We searched the BAC end sequences using FASTA and identified sequences that correspond to each of the six regions delineated by the *Hin*dIII sites. The number of BAC end sequences identified in each region ranged from 82 to 448. This variation may be due to mutations at the *Hin*dIII sites among the copies of MAGGY in the genome and/or the presence of incomplete copies of MAGGY. The overall sequence similarity among the aligned sequences within each region ranged from 95 to 99%.

Inspection of the multiple sequence alignments revealed what appeared to be an unusually high rate of nucleotide transitions. The average transition/transversion (t/v) ratio over all six sequence alignments was 6.8, while the average t/v ratio that we calculated in actin, β-tubulin, and calmodulin genes among *M. grisea* isolates reported by Couch and Kohn (2002) was 2.2. We performed multiple sequence alignments of the MGL and Pot2 sequences using the same procedures used for MAGGY. The average t/v ratio for MGL and Pot2 were 4.2 and 2.3, respectively. This unusually high t/v ratio among the MAGGY elements is consistent with a phenomenon known as Repeat-Induced Point mutation (RIP), originally described in *Neurospora crassa* (Cambareri et al., 1989). Since its discovery in the *Neurospora* genome, RIP-like mutations have been described in *Podospora* (Graia et al., 2001), *Magnaporthe* (Ikeda et al., 2002), *Fusarium* (Daboussi et al., 2002), and several other fungi.

Bias in the sequence context of RIP-mutated sites has frequently been observed and the preferred sequence context reported for *M. grisea* is (A/T)pCp(A/T) (Nakayashiki et al., 1999a,b). We analyzed the multiple sequence alignments of the BAC end sequences to identify whether the preferred sequence context is recognizable. The multiple sequence alignments were scanned for sites that contain transition mutations (C–T or G–A) between any two sequences in the alignment, indicating that site had been mutated in at least one copy of the sequence. The nucleotides 5′ and 3′ of the mutations are reported in Table 4. No bias in the composition of the nucleotides flanking the transition mutations was

Table 3
Number of BAC end sequences that contain each TE pair

	Pyret	MGLR-3	occan	MGL	Mg-SINE	Pot2	Pot3	MGR608	MGR619
MAGGY	1	0	0	0	0	7	0	0	0
Pyret		1	0	8	5	1	1	1	1
MGLR-3			0	1	1	1	0	3	2
occan				0	0	0	0	0	0
MGL					*	11	1	5	0
Pot2							2	0	0
Pot3								0	0
MGR608									5

[*] Mg-SINE and MGL have sequence similarity. See text for details.

Table 4
Sequence context of transitions in multiple sequence alignments

	5′					3′				
	A	T	C	G	AT to CG ratio	A	T	C	G	AT to CG ratio
BAC end sequences										
MAGGY	72	110	94	138	0.784	81	115	132	100	0.845
Pot2	193	247	151	159	1.419	194	229	182	145	1.294
MGL	89	129	109	152	0.835	97	148	126	102	1.075
Chromosome 7										
MAGGY	35	55	15	29	2.045	42	64	15	13	3.786
Pot2	61	69	54	44	1.327	57	84	51	35	1.64
MGL	32	68	39	46	1.176	58	86	22	17	3.692

Values represent the number of times each nucleotide was observed flanking a transition mutation. G–A transitions are reported as the complementary C–T transition.

identified among the BAC end sequences, possibly because of sequencing errors inherent in single pass sequences. We created a second set of multiple sequence alignments comprised of full-length copies of MAGGY (10 copies), MGL (7 copies), and Pot2 (30 copies) identified in the draft chromosome 7 sequence. As is shown in Table 4, there was a bias toward A/T 5′ and 3′ of the mutated C nucleotide in the sequences derived from the chromosome 7 sequence. The strongest bias is evident in the 3′ nucleotide of transitions in the MAGGY and MGL sequences where A and T were 3.8 and 3.7 times more likely to be observed than C or G.

4. Discussion

As in other organisms, the TEs of *M. grisea* do not appear to be distributed randomly in the genome. By comparing the proportion of BAC end sequences assigned to fingerprint contigs, we found that the three most common TEs, MAGGY, MGL, and Pot2 are not randomly distributed among the contigs. Instead, each chromosome appears to have one or more regions that contain an unusually high density of TEs. By hybridizing TE probes to genomic libraries of strain 2539, Nitta et al. (1997) also found that several families of TEs often hybridized to the same cosmids. Likewise, Nishimura et al. (1998, 2000) found a much higher than expected number of clones from a BAC library that contain more than one family of TE and proposed that "transposon islands" appear to occur on *M. grisea* chromosomes. Indeed, it appears that this arrangement might be common among TE-containing fungal genomes since TEs in the *F. oxysporum* genome appear have a similar arrangement (Hua-Van et al., 2000). These observations are consistent with the experimental results of Raina et al. (2002) who found that Ds transposons in *Arabidopsis thaliana* tend to integrate near the donor site and with Singleton and Levin (2002) who found that the Tf1 element of *Schizosaccharomyces pombe* targeted specific sites in intergenic regions.

Several possible models may be used to describe the observed TE distribution. Since the *A. thaliana* retrotransposon Ds tends to integrate near the donor site, the TEs of *M. grisea* may also integrate near their donor site, leading to a cluster of TEs near the original integration. TEs may rarely integrate at more distant loci, leading to the formation of new TE clusters. If this scenario were true, then there would likely be higher sequence similarity within clusters than between clusters. Examination of pairwise sequence similarities and phylogenetic trees revealed no such pattern (data not shown), however, the high overall degree of sequence similarity among the sequences and the fact that we utilized single pass sequences that could contain sequencing errors may obscure such patterns in our data set. Several alternative scenarios are also presented by Singleton and Levin (2002), including subnuclear localization of chromosomes, chromatin composition within chromosomes, and variation in the timing of chromosome replication during meiosis. Selective pressures, too, have been implicated in driving TEs toward clustered distribution patterns (Bartolomé et al., 2002; Charlesworth et al., 1994). In this scenario, selective pressure from insertions may lead to the accumulation of TEs in regions of low gene density. Alternatively, it has been proposed that deleterious ectopic chromosomal rearrangements induced by the presence of repetitive DNA may lead to the accumulation of TEs in regions of low recombination. Further experimental evidence will likely be required to determine if direct site specificity or selective pressures play a role in determining the distribution of TEs in fungal genomes.

Close examination of individual BAC end sequences provides clues that help to understand the evolution of TEs in the genome. We found four independent occurrences of MAGGY integrations that occurred within a previously existing Pot2 element, while in MGL, seven integrations occurred within a previously existing Pot2 element. Kachroo et al. (1995) also reported the cloning of Mg-SINE as an insertion into a Pot2 element, suggesting that other TEs have integration site preference for Pot2. MAGGY elements have previously been found embedded in AT-rich sequences (Farman et al., 1996b), suggesting that AT content may play a role in integration site specificity. The AT content of the Pot2 reference sequence is 59%, slightly higher than the 50% AT content of the BAC end sequences overall, lending credence to this hypothesis. However, the relatively small number of MAGGY–genomic DNA junctions observed in the BAC end sequences in this study and by Farman et al. (1996b) prevents us from making definitive conclusions as to the role of AT content in integration site preference. Transposable element integration site preference has been described in TEs of various families from both plants and fungi (El Amrani et al., 2002; Singleton and Levin, 2002) suggesting that this may be a general characteristic of TEs. A second hypothesis is that Pot2 is less transpositionally active than MAGGY and MGL, leading to few occurrences of a Pot2 element integrating into another TE. While this may also be a valid explanation of the apparent TE integration patterns, there is little evidence to support or refute this theory. A third possibility is that the presence of Pot2 in the genome may predate the presence of MAGGY and MGL. If Pot2 is an ancient component of the *M. grisea* genome then it could serve as integration sites for more recent invaders, such as MAGGY and MGL. This view is supported by host range information for these elements (Farman et al., 1996b; Kachroo et al., 1995). Since Mg-SINE and MGL have sequence identity at their 3′ ends, and the hybridization probe used by Kachroo et al. (1995) overlaps this region of identity, the data from the genomic Southern blots can be used to infer the host range of both Mg-SINE and MGL. Thus, both MAGGY and MGL have a restricted host range within *M. grisea*, suggesting either selective loss of both of these elements in certain populations of *M. grisea*, or a more recent origin, through horizontal transfer, in certain populations. The host range information, and the integration patterns described here are consistent with the view that MAGGY and MGL were horizontally transferred into the *M. grisea* genome after the introgression of Pot2.

We identified an unusually high t/v ratio among the MAGGY sequences that suggests the presence of a RIP-like process similar to the process that has been extensively studied in *N. crassa* and reported in several other fungi, including *M. grisea* (Ikeda et al., 2002; Irelan and Selker, 1996). The t/v ratio among the MAGGY sequences was 3× higher than the reference genes, and the t/v ratio of the MGL sequences are nearly 2× higher, which would be expected if RIP were occurring. There was also a corresponding bias in preferred sequence context of transitions observed in the MAGGY and MGL sequences. The bias was strongest in the nucleotide 3′ of the transition mutation, which tended to be A or T, consistent with the report of (Nakayashiki et al., 1999a,b). No bias in t/v ratio or in transition sequence context was found for transitions in the Pot2 alignments, suggesting that RIP is either not active in Pot2 or was active in ancestral populations but is no longer functioning.

Acknowledgments

We are grateful to Drs. Thomas Mitchell, Mark Farman, Jeffrey L. Thorne, and Brian M. Wiegmann for helpful discussion and suggestions. This work was supported with funds from Syngenta (formerly, the Novartis Foundation), the National Science Foundation, the United States Department of Agriculture, Clemson University, and North Carolina State University.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Bartolomé, C., Maside, X., Charlesworth, B., 2002. On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. Mol. Biol. Evol. 19, 926–937.

Borromeo, E.S., Nelson, R.J., Bonman, J.M., Leung, H., 1993. Genetic Differentiation among isolates of *Pyricularia* infecting rice and weed hosts. Phytopathology 83, 393–399.

Cambareri, E.B., Jensen, B.C., Schabtach, E., Selker, E.U., 1989. Repeat-induced G–C to A–T mutations in *Neurospora*. Science 244, 1571–1575.

Chalvet, F., Grimaldi, C., Kaper, F., Langin, T., Daboussi, M.J., 2003. *Hop*, an active *Mutator* -like element in the genome of the fungus *Fusarium oxysporum*. Mol. Biol. Evol. 20, 1362–1375.

Chao, C.T., Ellingboe, A.H., 1991. Selection for mating competence in *Magnaporthe grisea* pathogenic to rice. Can. J. Bot. 69, 2130–2134.

Charlesworth, B., Sniegowski, P.D., Stephan, W., 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 271, 215–220.

Couch, B.C., Kohn, L.M., 2002. A multilocus gene genealogy concordant with host preference indicates segregation of a new species, *Magnaporthe oryzae*, from *M. grisea*. Mycologia 94, 683–693.

Daboussi, M.J., Capy, P., 2003. Transposable elements in filamentous fungi. Annu. Rev. Microbiol. 57, 275–299.

Daboussi, M.J., Daviere, J.M., Graziani, S., Langin, T., 2002. Evolution of the *FotI* transposon in the genus *Fusarium*: discontinuous distribution and epigenetic inactivation. Mol. Biol. Evol. 19, 510–520.

Dobinson, K.F., Harris, R.E., Hamer, J.E., 1993. Grasshopper, a long terminal repeat (LTR) retroelement in the phytopathogenic fungus *Magnaporthe grisea*. Mol. Plant–Microbe. Interact. 6, 114–126.

El Amrani, A., Marie, L., Ainouche, A., Nicolas, J., Couee, I., 2002. Genome-wide distribution and potential regulatory functions of AtATE, a novel family of miniature inverted-repeat transposable elements in *Arabidopsis thaliana*. Mol. Genet. Genomics 267, 459–471.

Ewing, B., Hillier, L., Wendl, M.C., Green, P., 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8, 175–178.

Farman, M.L., Taura, S., Leong, S.A., 1996a. The *Magnaporthe grisea* DNA fingerprinting probe MGR586 contains the 3′ end of an inverted repeat transposon. Mol. Gen. Genet. 251, 675–681.

Farman, M.L., Tosa, Y., Nitta, N., Leong, S.A., 1996b. MAGGY, a retrotransposon in the genome of the rice blast fungus *Magnaporthe grisea*. Mol. Gen. Genet. 251, 665–674.

Graia, F., Lespinet, O., Rimbault, B., Dequard-Chablat, M., Coppin, E., Picard, M., 2001. Genome quality control: RIP (repeat-induced point mutation) comes to Podospora. Mol. Microbiol. 40, 586–595.

Hamer, J.E., Farrall, L., Orbach, M.J., Valent, B., Chumley, F.G., 1989. Host species-specific conservation of a family of repeated DNA sequences in the genome of a fungal plant pathogen. Proc. Natl. Acad. Sci. USA 86, 9981–9985.

Hua-Van, A., Daviere, J.M., Kaper, F., Langin, T., Daboussi, M.J., 2000. Genome organization in *Fusarium oxysporum*: clusters of class II transposons. Curr. Genet. 37, 339–347.

Ikeda, K., Nakayashiki, H., Kataoka, T., Tamba, H., Hashimoto, Y., Tosa, Y., Mayama, S., 2002. Repeat-induced point mutation (RIP) in *Magnaporthe grisea*: implications for its sexual cycle in the natural field context. Mol. Microbiol. 45, 1355–1364.

Irelan, J.T., Selker, E.U., 1996. Gene silencing in filamentous fungi: RIP, MIP and quelling. J. Genet. 75, 313–324.

Kachroo, P., Leong, S.A., Chattoo, B.B., 1994. Pot2, an inverted repeat transposon from the rice blast fungus *Magnaporthe grisea*. Mol. Gen. Genet. 245, 339–348.

Kachroo, P., Leong, S.A., Chattoo, B.B., 1995. Mg-SINE—a short interspersed nuclear-element from the rice blast fungus, *Magnaporthe grisea*. Proc. Natl. Acad. Sci. USA 92, 11125–11129.

Kang, S., 2001. Organization and distribution pattern of MGLR-3, a novel retrotransposon in the rice blast fungus *Magnaporthe grisea*. Fungal Genet. Biol. 32, 11–19.

Kang, S.C., Sweigard, J.A., Valent, B., 1995. The PWL host specificity gene family in the blast fungus *Magnaporthe grisea*. Mol. Plant–Microbe Interact. 8, 939–948.

Kang, S., Lebrun, M.H., Farrall, L., Valent, B., 2001. Gain of virulence caused by insertion of a Pot3 transposon in a *Magnaporthe grisea* avirulence gene. Mol. Plant–Microbe. Interact. 14, 671–674.

Kidwell, M.G., Lisch, D., 1997. Transposable elements as sources of variation in animals and plants. Proc. Natl. Acad. Sci. USA 94, 7704–7711.

Lau, G.W., Chao, C.T., Ellingboe, A.H., 1993. Interaction of genes controlling avirulence/virulence of *Magnaporthe grisea* on rice cultivar Katy. Phytopathology 83, 375–382.

Mao, L., Wood, T., Yu, Y., Budiman, M.A., Tomkins, J., Woo, S.-s., Sasinowski, M., Presting, G., Frisch, D., Goff, S., Dean, R.A., Wing, R.A., 2000. Rice transposable elements: a survey of 73,000 sequence-tagged-connectors. Genome Res. 10, 982–990.

Martin, S.L., Blackmon, B.P., Rajagopalan, R., Houfek, T.D., Sceeles, R.G., Denn, S.O., Mitchell, T.K., Brown, D.E., Wing, R.A., Dean, R.A., 2002. MagnaportheDB: a federated solution for integrating physical and genetic map data with BAC end derived sequences for the rice blast fungus *Magnaporthe grisea*. Nucleic Acids Res. 30, 121–124.

Nakayashiki, H., Kiyotomi, K., Tosa, Y., Mayama, S., 1999a. Transposition of the retrotransposon MAGGY in heterologous species of filamentous fungi. Genetics 153, 693–703.

Nakayashiki, H., Nishimoto, N., Ikeda, K., Tosa, Y., Mayama, S., 1999b. Degenerate MAGGY elements in a subgroup of *Pyricularia grisea*: a possible example of successful capture of a genetic invader by a fungal genome. Mol. Gen. Genet. 261, 958–966.

Nakayashiki, H., Matsuo, H., Chuma, I., Ikeda, K., Betsuyaku, S., Kusaba, M., Tosa, Y., Mayama, S., 2001. Pyret, a Ty3/Gypsy retrotransposon in *Magnaporthe grisea* contains an extra domain between the nucleocapsid and protease domains. Nucleic Acids Res. 29, 4106–4113.

Nishimura, M., Nakamura, S., Hayashi, N., Asakawa, S., Shimizu, N., Kaku, H., Hasebe, A., Kawasaki, S., 1998. Construction of a BAC library of the rice blast fungus *Magnaporthe grisea* and finding specific genome regions in which its transposons tend to cluster. Biosci. Biotechnol. Biochem. 62, 1515–1521.

Nishimura, M., Nakamura, S., Hayashi, N., Masuya, M., Asakawa, S., Shimizu, N., Kaku, H., Hasebe, A., Kawasaki, S., 2000. Analysis of transposable-element clustering patterns in *Magnaporthe grisea* genome using BAC library. In: Tharreau, D. (Ed.), Advances in Rice Blast Research. Kluwer Academic Publishers, The Netherlands, pp. 316–322.

Nitta, N., Farman, M.L., Leong, S.A., 1997. Genome organization of *Magnaporthe grisea*: integration of genetic maps, clustering of transposable elements and identification of genome duplications and rearrangements. Theor. Appl. Genet. 95, 20–32.

Okada, N., Hamada, M., Ogiwara, I., Ohshima, K., 1997. SINEs and LINEs share common 3′ sequences: A review. Gene 205, 229–243.

Ou, S.H., 1987. Rice Diseases. Commonwealth Mycological Institute, Surrey.

Pearson, W.R., 2000. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol. Biol., 185–219.

Raina, S., Mahalingam, R., Chen, F.Q., Fedoroff, N., 2002. A collection of sequenced and mapped Ds transposon insertion sites in *Arabidopsis thaliana*. Plant Mol. Biol. 50, 93–110.

Shull, V., Hamer, J.E., 1996. Genetic differentiation in the rice blast fungus revealed by the distribution of the Fosbury retrotransposon. Fungal Genet. Biol. 20, 59–69.

Singleton, T.L., Levin, H.L., 2002. A long terminal repeat retrotransposon of fission yeast has strong preferences for specific sites of insertion. Eukaryot. Cell 1, 44–55.

Skinner, D.Z., Budde, A.D., Farman, M.L., Smith, J.R., Leung, H., Leong, S.A., 1993. Genome organization of *Magnaporthe grisea*—genetic map, electrophoretic karyotype, and occurrence of repeated DNAs. Theor. Appl. Genet. 87, 545–557.

Sone, T., Suto, M., Tomita, F., 1993. Host species specific repetitive DNA sequence in the genome of *Magnaporthe grisea*, the rice blast fungus. Biosci. Biotechnol. Biochem. 57, 1228–1230.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Zhu, H., Choi, S.D., Johnston, A.K., Wing, R.A., Dean, R.A., 1997. A large-insert (130 kbp) bacterial artificial chromosome library of the rice blast fungus *Magnaporthe grisea*: genome analysis, contig assembly, and gene cloning. Fungal Genet. Biol. 21, 337–347.