Organization of chromosome ends in the rice blast fungus, *Magnaporthe oryzae*

Cathryn Rehmeyer[1], Weixi Li[2], Motoaki Kusaba[1], Yun-Sik Kim[1], Doug Brown[3], Chuck Staben[2], Ralph Dean[3] and Mark Farman[1,*]

[1]Department of Plant Pathology and [2]Department of Biology, University of Kentucky, Lexington, KY 40546 USA and [3]Center for Integrated Fungal Research, North Carolina State University, Raleigh, NC 27695 USA

ABSTRACT

Eukaryotic pathogens of humans often evade the immune system by switching the expression of surface proteins encoded by subtelomeric gene families. To determine if plant pathogenic fungi use a similar mechanism to avoid host defenses, we sequenced the 14 chromosome ends of the rice blast pathogen, *Magnaporthe oryzae*. One telomere is directly joined to ribosomal RNA-encoding genes, at the end of the ∼2 Mb rDNA array. Two are attached to chromosome-unique sequences, and the remainder adjoin a distinct subtelomere region, consisting of a telomere-linked RecQ-helicase (*TLH*) gene flanked by several blocks of tandem repeats. Unlike other microbes, *M.oryzae* exhibits very little gene amplification in the subtelomere regions—out of 261 predicted genes found within 100 kb of the telomeres, only four were present at more than one chromosome end. Therefore, it seems unlikely that *M.oryzae* uses switching mechanisms to evade host defenses. Instead, the *M.oryzae* telomeres have undergone frequent terminal truncation, and there is evidence of extensive ectopic recombination among transposons in these regions. We propose that the *M.oryzae* chromosome termini play more subtle roles in host adaptation by promoting the loss of terminally-positioned genes that tend to trigger host defenses.

INTRODUCTION

Telomeres are the DNA sequences that form the ends of eukaryotic chromosomes and are a solution to the inherent problem associated with the replication of DNA termini. In most eukaryotes, the telomeres consist of tandem arrays of simple sequence repeats whose exact composition varies between organisms but usually consist of a GT-rich sequence on the strand that runs $5'$ to $3'$ toward the chromosome end (1). The most common motif in filamentous fungi and vertebrates is $(TTAGGG)_n$ (2), budding yeast *Saccharomyces cerevisiae* has the sequence $(TG_{1-3})_n$ (3), and many plants have $(TTTAGGG)_n$ (4). New repeats are added to the existing telomeres by the enzyme telomerase (5), thereby preventing the loss of terminal sequences as a consequence of DNA replication. The telomere repeat is bound by a number of proteins which protect the chromosome ends from degradation and prevent the ends from being recognized as double-strand breaks (6). These proteins are also involved in regulating telomere length (7–9), and have roles in the silencing of neighboring genes—a phenomenon known as the telomere position effect (TPE) (10–12). Silencing activity is associated with the sequestration of telomeres and their associated proteins in discrete nuclear foci (13–18). Such physical proximity between non-homologous chromosome ends within the nucleus is also thought to promote ectopic recombination between shared sequences in the subtelomeric regions, leading to the amplification and diversification of telomere-linked genes (16,19,20).

The chromosome regions found immediately internal to the telomere repeats are known as subtelomeres. These regions have a general organization that is widely conserved among organisms ranging from protists to humans (21). In many cases, two distinct subtelomere domains are discernable. The region that is immediately adjacent to the telomere repeats is known as the distal domain and usually consists of one or more core sequences that are present at many, and sometimes all, chromosome ends. In *S.cerevisiae* and humans, the distal domains are bordered by distinct boundary elements, beyond which lie the proximal domains. These, too, contain sequences that are duplicated at different chromosome termini, although, the sequences that make up the proximal domains tend not to be as widely spread as the ones in the distal domains. Eventually, the proximal domains transition

---

into sequences that are unique to each individual chromosome end (chromosome-unique sequences). The sizes of proximal domain duplications can be quite large, extending up to ∼30 kb in *S.cerevisiae* (22,23) and up to more than 200 kb in humans (24). In some cases, the patterns of sequence duplication and dispersal among different subtelomeres can be quite complex, causing the proximal domains to appear as patchworks of sequence identity (19,25,26). The most significant proximal domain feature is the presence of large blocks of genes, whose duplication and dispersal among different chromosome ends results in the generation of subtelomeric gene families whose constituent members often exhibit unusually high levels of sequence diversity (16,27–29).

The potential for rapid evolution of genes that reside in telomere-proximal locations has been particularly well exploited by a number of fungi and protists, several of which have proximal domains that contain families of genes involved in niche adaptation. Such genes have been termed 'contingency genes' (27,30) because they tend to be non-essential, yet confer phenotypes that provide the capability to respond to specific ecological challenges. In prokaryotes, contingency genes often exhibit elevated mutation rates, which results in a reservoir of phenotypic diversity that may be used to meet such challenges (30). It should be pointed out, however, that although telomere-linked genes in eukaryotic microbes often show extremely high levels of allelic diversity, there is, yet, no evidence to suggest that they have higher mutations rates.

The proximal subtelomeres of *S.cerevisiae* contain families of genes involved in carbon utilization and other traits that are important for fermentation. These include the *MAL* locus, a block of three genes involved in the uptake and utilization of maltose (31,32), *MEL* genes which encode α-galactosidases (33,34), *FLO* genes which control flocculation (35) and are involved in environmental stress responses (36), and the *PAU* genes, which are induced during anaerobic growth (37,38). The adaptive potential of proximal domains is illustrated by the fact that the genes in these regions are differentially amplified in *S.cerevisiae* strains from different ecological niches. For example, *MEL* genes are absent from most strains but are highly amplified in the proximal domains of strains from olive oil processing wastes and animal feces (34,39).

Subtelomere roles in adaptation are particularly evident in microbial human pathogens. In the protists *Plasmodium falciparum* (malaria), *Trypanosoma brucei* (African sleeping sickness) and the ascomycete fungus *Pneumocystis carinii* (pneumonia), the proximal subtelomere domains contain large gene families, which code for extracellular proteins that cover the cell surface, or, in the case of *P.falciparum*, coat the red blood cells in which the pathogen hides (40–42). These surface proteins are immunogenic and so would normally bring about eradication of the pathogen, when the host develops an appropriate antibody. However, these organisms undergo antigenic variation, which involves controlling surface protein expression so that only one form is produced at a time, together with a periodic switching of the expressed copy. This strategy enables certain members of the pathogen population to avoid detection and sustain an infection. Two basic mechanisms underlying antigenic variation have been described. *P.carinii* uses recombination to swap silent genes

into a unique subtelomeric expression site (43), whereas *P.falciparum* regulates *var* gene expression by switching transcription between different subtelomeric genes (44). *T.brucei* uses a combination of both strategies (27,45–47). The fact that a number of diverse pathogens utilize subtelomere regions for the amplification and switching of genes with critical roles in pathogenesis suggests that there is a distinct advantage to this arrangement. One obvious benefit is that the genes near telomeres tend to accumulate unusually high levels of allelic variability (16). As such, the subtelomeres are ideal locations for generating the raw materials for antigenic switching, namely rich reservoirs of variant gene copies. In addition, the tendency for subtelomeres to undergo ectopic exchanges with other chromosome ends, can lead to recombination between different gene copies, thereby accelerating the generation of novel gene variants, and, in the cases of *T.brucei* and *P.carinii*, facilitating the recombination of silent genes into the subtelomeric expression sites (27,43,45–47). Finally, for *T.brucei*, there is evidence that the TPE silencing machinery plays a role in the regulation of subtelomeric expression sites (48,49), further implicating telomeres in the switching process itself. Interestingly, the TPE is also responsible for controlling expression of subtelomeric adhesin genes, which are important virulence factors in the pathogenic fungus *Candida glabrata* (50).

Human pathogens may not be the only parasites that exploit the dynamic telomere environment for generation of variation in contingency genes with crucial roles in interactions with their hosts. Studies of host specificity in the fungus, *Magnaporthe oryzae*, a devastating pathogen of rice and other graminaceous hosts, suggest that the telomeres of plant pathogenic fungi might also contain contingency genes. *M.oryzae* is a filamentous ascomycete, formerly known as *Magnaporthe grisea*, until recent studies (51) revealed that strains causing diseases of rice, millet and other crop plants are taxonomically distinct from the type isolates, which are pathogens of crabgrass (52). *M.oryzae* populations are notorious for their pathogenic variability, which enables them to overcome resistant rice cultivars within a very short period of time after deployment (53). In *M.oryzae*, the ability to infect different hosts is controlled by avirulence genes ('*Avr*' genes), whose protein products directly, or indirectly, trigger host defense responses in plants that possess corresponding resistance genes. Significantly, 7 of the 17 *M.oryzae Avr* genes that have been identified so far map very near to telomeres (54–56), and in some cases, this positioning is associated with pathogenic variability. For example, strains carrying the *Avr* genes *Avr-Pita* and *Avr-Tsuy* are normally avirulent to rice cultivars Yashiro-mochi and Tsuyake, respectively (54), yet virulent mutants that are able to avoid recognition by rice plants containing the cognate resistance genes appear at a high frequency (54). Cloning of the *Avr-Pita* gene revealed that it is located at the very tip of the chromosome, with its 3′-untranslated region (3′-UTR) ending just 48 bp from the telomere repeat. Moreover, analysis of spontaneous mutants that had gained virulence to Yashiro-mochi revealed that many had experienced truncation of the chromosome end (57).

Telomeres are among the most variable regions of the *M.oryzae* genome. Telomeric restriction fragments tend to

be more polymorphic than internal fragments (58,59), and novel telomeres arise frequently during vegetative culture (58) and in meiosis (59). Thus, it appears that the dynamic environment of the *M.oryzae* chromosome ends is also used to generate variation in genes with critical roles in the interaction with the host plant. Considering that the subtelomeres of other pathogenic microbes often contain whole families of 'niche adaptation' genes (16,27,31,43,60,61), we were particularly interested in determining if the same is true for *M.oryzae*, as this could potentially lead to the discovery of previously unidentified *Avr* genes, as well as other genes with important functions in pathogen-host interactions. Unfortunately, chromosome ends were poorly represented in the *M.oryzae* genome sequence, with only one of the 14 telomeres in this fungus being present in the assembly that was available when this project was started (Version 2). Nevertheless, screening of the raw sequence data present in the National Center for Biotechnology Information (NCBI) Trace Archive (www.ncbi.nlm.nih.gov/Traces/trace.cgi) led to the identification of several hundred telomere-containing reads that had been excluded from the assembly (62). Extraction and characterization of these reads and their mate-pairs identified a set of fosmid clones that contain all 14 of the *M.oryzae* telomeres. Here, we describe the organization of *M.oryzae* chromosome ends as determined by analysis of the fosmid sequences.

MATERIALS AND METHODS

Identification of fosmids contain the 14 *M.oryzae* telomeres

Detailed methods describing how the telomeric fosmids were identified and verified will be published elsewhere. Briefly, telomeric reads were identified among the *M.oryzae* raw sequence traces derived from plasmid and fosmid inserts (62). The telomeric sequences and the mate-pair reads derived from the opposite ends of the inserts were then assembled into telomeric contigs and subtelomeric contigs, respectively, creating a set of telomeric scaffolds, with each scaffold corresponding to a separate chromosome end. The scaffolds were then linked to the genome assembly using BLASTn searches. The BLAST results were examined to ensure that fosmids containing the same telomere had mate-pair reads that matched the same genomic contig. Clones that did not meet this criterion were discarded due to possible chimerism. Three fosmids per chromosome end were further characterized to confirm that they were not rearranged, and that they contain terminal fragments matching specific telomeres in genomic DNA of 70-15.

Sequencing of fosmid clones

Each clone was cultured overnight in 500 ml of Luria–Bertani (LB) plus 25 µg/ml chloramphenicol. Fosmid DNA was extracted using a midiprep kit (Qiagen, Valencia, CA). Each DNA sample was sheared to a median size of ∼2 kb by passing 20 times through a standard shearing assembly (part number JHSH204004) on a HydroShear machine (Genomic Solutions, Ann Arbor, MI), using speed setting 6. The sheared fragments were end-repaired using the End-It

kit (Epicentre Technologies, Madison, WI) and resolved on a 0.8% agarose gel (100 V, 1 h). Fragments of ∼1.8–2.2 kb were then excised from the gel and purified by QIAquick gel extraction (Qiagen). The DNA was concentrated by ethanol precipitation and re-dissolved overnight in 6.5 µl T0.1E [10 mM Tris-HCl (pH 8.0) and 0.1 mM EDTA]. Approximately 1 µg of sheared DNA was ligated to the CloneSmart HCamp vector (Lucigen Corp., Middleton, WI), according to the manufacturer's protocol. After 2 h ligation, 1 µl of the ligation mix was used to transform DH5α-T1R chemically competent cells (Invitrogen, Carlsbad, CA), which were then plated on LB agar amended with 100 µg/ml ampicillin. Colonies were picked into 384-well plates containing LB and 10% glycerol, grown for 12 h at 37°C, and then sent as frozen stocks to Agencourt Biosciences (Beverly, MA), or deposited with the University of Kentucky Advanced Genetic Technologies Center (UK-AGTC) for sequencing with ABI BigDye V3 (Applied Biosystems, Foster City, CA) or CEQ DTCS (Beckman Coulter, Fullerton, CA) chemistries. Sequences were assembled using phred (63,64), phrap (www.phrap.org) and Consed (65) and the gaps remaining after shotgun assembly were filled by primer walking.

The fosmid sequences are deposited in the GenBank database under the accession nos DQ493941 through DQ493955.

BLAST—TruMatch

The fosmid sequences were linked to the genome assembly by using the complete (unfiltered) sequences to query the genome sequence using local BLAST. Due to the repetitive nature of the subtelomere regions, several queries yielded multiple matches. This, combined with the presence of numerous subtelomeric gaps in the assembly, often resulted in the correct alignment appearing at a very low position in the BLAST report, where it was easily overlooked. Therefore, we developed a perl script (TruMatch), which examines BLAST outputs to identify correct matches based on unique alignments (62). All of the initial BLAST results were post-processed with TruMatch, to reveal bona fide overlaps to the genome assembly.

Gene prediction and analysis

Fosmid assemblies were screened using CrossMatch (www.phrap.org) to mask known transposable elements (TEs) and other repeats with $\geq$10 copies in the *M.oryzae* genome (66). The masked sequences were then used for gene prediction with Fgenesh (67) trained on *M.oryzae* sequences (Softberry, www.softberry.com). The resulting protein sequences were searched against GenBank and the *M.oryzae* whole genome assembly using the blastx and blastp algorithms (68). In the absence of BLAST similarity, predicted proteins shorter than 100 amino acids were excluded from further analysis. Retained protein sequences were searched against Pfam [(69), www.sanger.ac.uk/Software/Pfam/], using both global and fragment alignments, with an E-value less than 10^{-4}. MEME (70) was used to identify motifs common to proteins encoded by subtelomeric genes, using minimum and maximum motif width values of 6 and 300 amino acids, respectively. Identification of gene and protein families was performed by clustering nucleotide and

amino acid sequences with BLASTClust (BLAST package, ftp://ftp.ncbi.nih.gov/blast/), using threshold values of 60% length and 30% similarity over both sequences.

The predicted amino acid sequences in the subtelomeric gene set were analyzed with SignalP [(71), www.cbs.dtu.dk/services/SignalP/], and secreted proteins defined when both the eukaryotic Neural Network and Hidden Markov Model scores exceeded a minimum threshold score of 0.5. Cellular localization prediction for secreted proteins was performed with ProtComp version 6.0 for animal and fungi (www.softberry.com). Tandem Repeats Finder (72) was utilized to identify tandem repeats. To identify new, dispersed repeats, stand-alone BLAST (68) was performed using the fosmids as both the reference sequences and the queries.

RESULTS

Terminology

Telomeres: used to describe the sequences that form the chromosome tips. In the present case, these consist of tandem repeats of the hexanucleotide TTAGGG (59).

Subtelomeres: this term is reserved for distinct domains that occur immediately adjacent to the telomeres. These domains consist of sequences that are duplicated at a number of chromosome ends.

Terminal sequences. We use this term to describe sequences that are near to the chromosome ends but are not part of the subtelomere domains. Unless otherwise noted, the operational definition of 'terminal' for the present study is being contained within a telomeric fosmid clone (i.e. within ~40 kb of the telomere).

M.oryzae strain 70-15

M.oryzae strain 70-15 is one of the number of progeny that were isolated during a series of crosses that were performed to develop rice-infecting strains with increased fertility (73). Therefore, when interpreting the telomere/subtelomere

organization of 70-15, especially with regard to pathogenic adaptation, it is important to consider that this strain is not a natural, field isolate. Nevertheless, 70-15 is closely related to the rice-infecting, field isolate, Guy11, which was the recurrent parent in the last three crosses used to generate the former strain (73). Moreover, analysis of 70-15's telomeres by Southern hybridization indicated that at least 10 of them were inherited from Guy11 (data not shown). Therefore, analysis of telomere organization in 70-15 is expected to provide a reasonable insight into telomere organization in natural, rice-infecting isolates of *M.oryzae*.

M.oryzae telomeres

The first criterion used to identify fosmids containing telomeres was the presence of TTAGGG repeats at one end of the insert. The number of copies of this sequence varied among the selected clones, ranging from 6 to 30. Analysis of the complete set of telomere-containing, shotgun reads obtained in the *Magnaporthe* genome sequencing project revealed that the average number of repeats across all chromosomes is 26, suggesting an average telomere length of ~150 bp. Some fosmids also had short TTAGGG motifs in the subtelomere region. However, these internal arrays never had more than two copies of this sequence in tandem (see below). There also don't appear to be any large telomeric arrays at other genomic locations, because Southern hybridization analysis of genomic DNA samples digested with various restriction enzymes never produced more than 14 hybridizing fragments (data not shown).

Telomere-adjacent sequences

Analysis of the 'internal' sequences that are found adjacent to the telomere repeats revealed three basic organizations:

(i) *rDNA sequences*. Fosmid 61J04, representing telomere 3, contains four tandem iterations of the ~8.4 kb ribosomal DNA repeat (Figure 1A), with the terminal rDNA unit being capped by telomere repeats in the
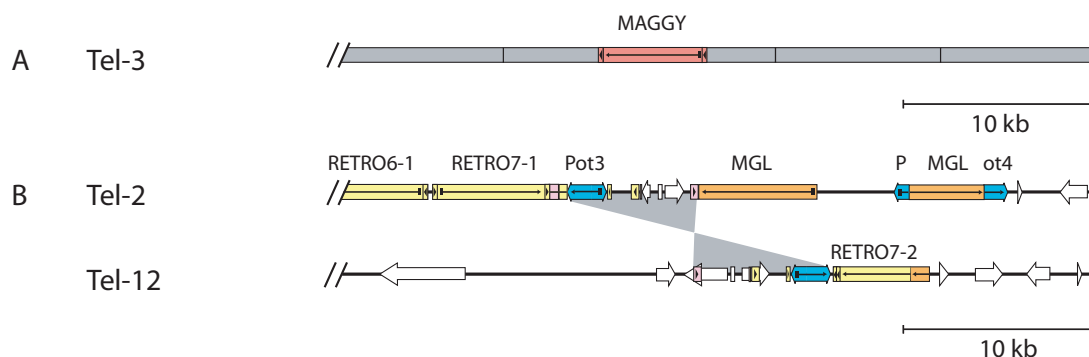

Figure 1. *M.oryzae* telomeres joined to the rDNA array and chromosome-unique sequences. The telomere repeats are represented by small black boxes at the right-hand end of each map. (**A**) The ribosomal RNA gene array on chromosome II extends all the way to the telomere. The individual repeat units are shown as gray boxes. A MAGGY retroelement insertion is represented as a red box, flanked by smaller boxes representing the LTRs. The orientation of the element is shown with an arrow. (**B**) The TTAGGG repeats of telomeres 2 and 12 are attached directly to chromosome-unique sequences. White arrows show predicted genes, and the black lines represent intergenic sequences. Transposon insertions are indicated as colored boxes and are distinguished as follows: double-headed arrows = inverted repeat transposons; colored boxes = non-LTR retrotranposons; colored boxes with small boxes at ends = LTR retrotranposons. Arrows inside the boxes represent orientation, and elements shown without an arrowhead and/or tail are truncated versions. The identity of each transposon is noted in the figure. A region of sequence alignment (in opposing orientation) between the two chromosome ends is shown as gray shading connecting the two maps.

intergenic spacer region (nt position 7810). Interestingly, the third rDNA unit from the telomere contains an insertion of the MAGGY LTR-retrotransposon (74). Long-range restriction mapping revealed that the telomeric rDNA sequences are contiguous with the major rDNA array, which extends ~2 Mb toward the centromere on chromosome 2 (data not shown).

(ii) *Chromosome-unique sequences.* Fosmids 25N20 and 72H05 contain telomeres 2 and 12, respectively. The sequences that occur immediately adjacent to the telomeres in these fosmids both exist as a single copy in the 70-15 genome and, therefore, we refer to them as chromosome-unique sequences (Figure 1B). Clones 25N20 and 72H05 both contain several transposon insertions but, otherwise, they lack sequence features that would suggest proximity to a chromosome landmark with such functional importance as the telomere. There are no tandem repeat arrays, or other repetitive motifs, in these clones (Figure 1B), and there is no obvious change in base composition as one approaches the telomere repeats (data not shown). The only feature of note is a 6.5 kb stretch of shared sequence that occurs in the opposite orientation at the two ends. In each case, the duplicated segment is bordered on one side by a Pot3 element (75,76) and by a solo-LTR of the Pyret retroelement (77) on the other (Figure 1B).

(iii) *Telomere-linked helicase genes.* Eleven of the fosmid clones share a significant amount of sequence similarity in the region adjacent to the telomere repeats. Figure 2 shows an alignment of physical maps for the
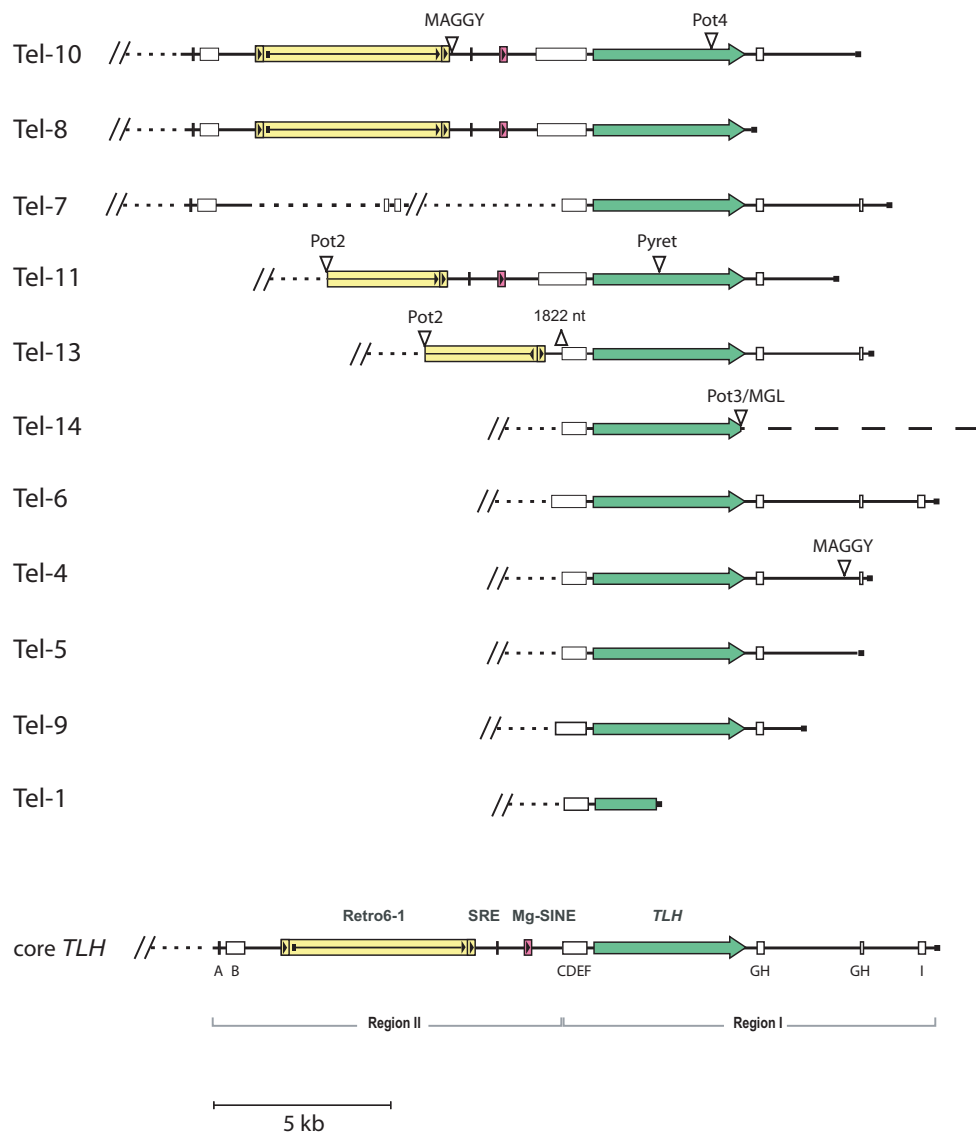

Figure 2. Telomeres adjoining a subtelomere region containing a telomere-linked helicase gene. The solid black lines represent a core *TLH* sequence that is shared by several chromosome ends. The chromosome ends are drawn with the *TLH* genes (green arrows) aligned to show the extent of subtelomere truncation at each telomere. HARs are shown as white boxes and are labeled in the map of the core *TLH* region shown at the bottom of the panel. Conserved transposon insertions are shown as colored boxes, and the positions of insertions that are specific to certain ends are noted with downward-pointing arrowheads above the corresponding maps. Chromosome-unique sequences are represented with a dotted line on the left-hand end of each map. Each of the *TLH* ends has a different sequence in this region. Note that the scales of the maps drawn in panels A, B and C are specific to each panel.

corresponding chromosome ends. A consensus map representing the 'core' sequence is shown at the bottom. All eleven ends contain at least a portion of the segment labeled region I, which contains a copy of the telomere-linked RecQ-helicase (*TLH*) gene previously described by Gao and coworkers (78). Using Southern hybridization analysis, we determined that 70-15 contains only 11 *TLH* copies (data not shown) and, therefore, the sequence data confirm that all of the *TLH* genes in 70-15 are near to telomeres. Based on the presence of the *TLH*-containing core sequence at several chromosome ends, we consider it to be a distinct subtelomere sequence.

With the exceptions of telomeres 1 and 14, all of the chromosome ends that contain *TLH* genes (henceforth termed 'TLH-ends') have almost identical sequences between the *TLH* gene and the telomere repeat. This sequence exhibits various degrees of truncation, however, so the distance between the 3′ end of the *TLH* gene and the telomere varies—ranging from ∼5.5 kb for telomere 6 to ∼−2.2 kb at telomere 1 (Figure 2). The core *TLH* to telomere sequence is AT-rich (66%) and contains three blocks of tandem repeats. These repeats are not found elsewhere in the genome and, therefore, we refer to them as 'helicase-associated repeats' (HARs). Subtelomere 1 lacks the core sequence due to truncation of the chromosome within the *TLH* open reading frame (ORF). The other TLH-end that lacks the sequence is telomere 14, which represents a novel chromosome end that

arose in the *M.oryzae* subculture used in the genome sequencing project (C. Rehmeyer and M. Farman, unpublished data). The sequence adjacent to telomere 14 (shown as a dashed line in Figure 2) was captured from an internal genomic location, presumably following loss of the original telomere.

All of the TLH-ends have a similar sequence organization upstream of the *TLH* genes, up to and including the large block of tandem repeats labeled HAR-CDEF. These repeats form a distinct boundary (the 'near' boundary) beyond which seven of the subtelomeres transition into chromosome-unique sequences. Alignments of the junctions between HAR-C and the chromosome-unique sequences are shown in Figure 3A. Subtelomeres 8, 10 and 11 have almost identical sequences beyond the HAR-C repeat (see below), whereas the other TLH-ends diverge in sequence upstream of the first or second cytosine in the HAR-C repeat unit (CCTAACCAGTAATA).

Four subtelomeres (8, 10, 11 and 13) have extensive sequence similarity beyond HAR-CDEF. The alignment between the four ends is interrupted by a 1822 bp deletion upstream of the HAR-CDEF block in subtelomere 13 (Figure 2). Beyond this deletion, however, the four ends share almost identical sequences for ∼5 kb, with the only disruption in colinearity being due to a MAGGY insertion in subtelomere 10. Subtelomeres 11 and 13 diverge first, transitioning into chromosome-unique sequences beyond a shared Pot2 insertion located ∼7.3 kb upstream of the *TLH* genes.
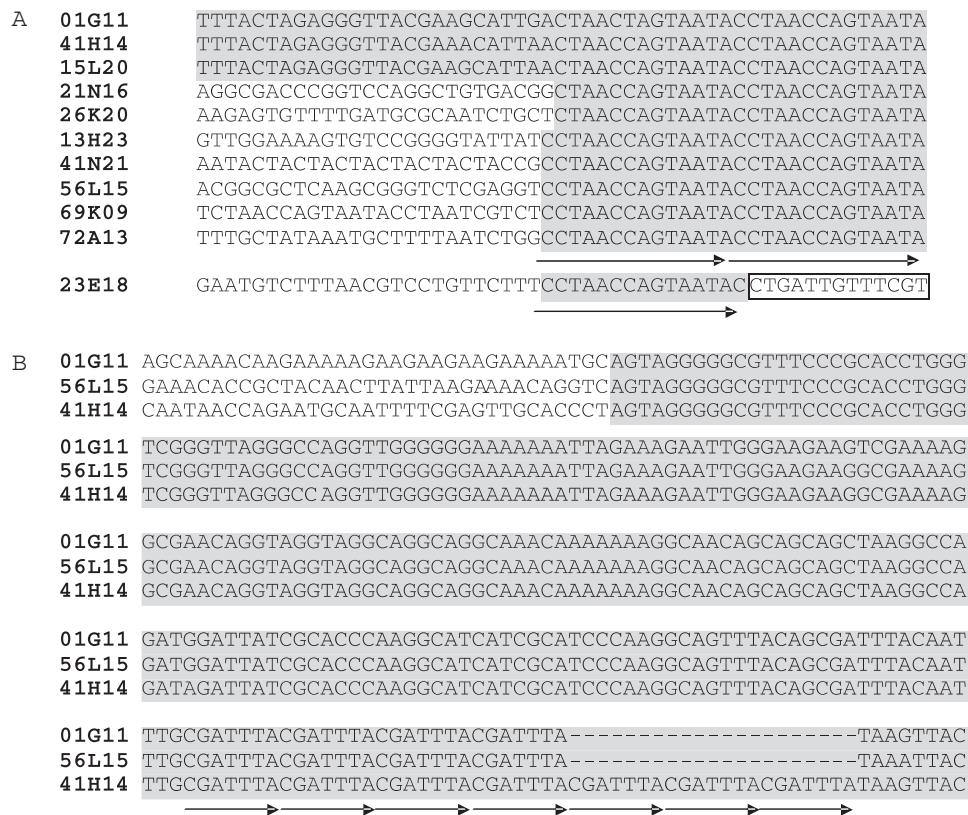
```
A  01G11   TTTACTAGAGGGTTACGAAGCATTGACTAACTAGTAATACCTAACCAGTAATA
   41H14   TTTACTAGAGGGTTACGAAACATTAACTAACCAGTAATACCTAACCAGTAATA
   15L20   TTTACTAGAGGGTTACGAAGCATTAACTAACCAGTAATACCTAACCAGTAATA
   21N16   AGGCGACCCGGTCCAGGCTGTGACGGCTAACCAGTAATACCTAACCAGTAATA
   26K20   AAGAGTGTTTTGATGCGCAATCTGCTCTAACCAGTAATACCTAACCAGTAATA
   13H23   GTTGGAAAAGTGTCCGGGGTATTATCCTAACCAGTAATACCTAACCAGTAATA
   41N21   AATACTACTACTACTACTACTACCGCCTAACCAGTAATACCTAACCAGTAATA
   56L15   ACGGCGCTCAAGCGGGTCTCGAGGTCCTAACCAGTAATACCTAACCAGTAATA
   69K09   TCTAACCAGTAATACCTAATCGTCTCCTAACCAGTAATACCTAACCAGTAATA
   72A13   TTTGCTATAAATGCTTTTAATCTGGCCTAACCAGTAATACCTAACCAGTAATA
                                       ───────────────→──────────────→

   23E18   GAATGTCTTTAACGTCCTGTTCTTTCCTAACCAGTAATAC CTGATTGTTTCGT
                                       ───────────────→


B  01G11   AGCAAAACAAGAAAAAGAAGAAGAAGAAAAAATGCAGTAGGGGGCGTTTCCCGCACCTGGG
   56L15   GAAACACCGCTACAACTTATTAAGAAAACAGGTCAGTAGGGGGCGTTTCCCGCACCTGGG
   41H14   CAATAACCAGAATGCAATTTTCGAGTTGCACCCTAGTAGGGGGCGTTTCCCGCACCTGGG

   01G11   TCGGGTTAGGGCCAGGTTGGGGGGGAAAAAAATTAGAAAGAATTGGGAAGAAGTCGAAAAG
   56L15   TCGGGTTAGGGCCAGGTTGGGGGGGAAAAAAATTAGAAAGAATTGGGAAGAAGGCGAAAAG
   41H14   TCGGGTTAGGGCC AGGTTGGGGGGGAAAAAAATTAGAAAGAATTGGGAAGAAGGCGAAAAG

   01G11   GCGAACAGGTAGGTAGGCAGGCAGGCAAACAAAAAAAGGCAACAGCAGCAGCTAAGGCCA
   56L15   GCGAACAGGTAGGTAGGCAGGCAGGCAAACAAAAAAAGGCAACAGCAGCAGCTAAGGCCA
   41H14   GCGAACAGGTAGGTAGGCAGGCAGGCAAACAAAAAAAGGCAACAGCAGCAGCTAAGGCCA

   01G11   GATGGATTATCGCACCCAAGGCATCATCGCATCCCAAGGCAGTTTACAGCGATTTACAAT
   56L15   GATGGATTATCGCACCCAAGGCATCATCGCATCCCAAGGCAGTTTACAGCGATTTACAAT
   41H14   GATAGATTATCGCACCCAAGGCATCATCGCATCCCAAGGCAGTTTACAGCGATTTACAAT

   01G11   TTGCGATTTACGATTTACGATTTACGATTTA--------------------TAAGTTAC
   56L15   TTGCGATTTACGATTTACGATTTACGATTTA--------------------TAAATTAC
   41H14   TTGCGATTTACGATTTACGATTTACGATTTACGATTTACGATTTACGATTTATAAGTTAC
           ───────→────────→────────→────────→────────→────────→
```

Figure 3. Junction sequences at the boundaries of *TLH* regions. *TLH*-associated sequences are shaded gray and chromosome-specific sequences are unshaded. (**A**) Shows the sequences at region I boundaries. HAR-C repeats units are indicated with arrows. The start of a HAR-D repeat is boxed. (**B**) Shows sequences spanning the centromere-proximal region II boundaries. Arrows represent tandemly-organized HAR-A motifs.

Table 1. Nucleotide sequences of individual repeat units making up the HAR arrays

HAR	Sequence
A	cgattta
B	tgttgcacccggggggcaagaaacctc
C[a]	**cctaacc**agtaata
D[b]	[ctg<u>attgtttcgttgaacgaagcaat</u>ggattttgtgtggacgcca- aggct]gcagggatgatggcagggggtacgaaacattggattta
E	attagaaggttacggggcattgactttttttaatcgtatcgcttc- aaaatttgaaccagcgatagatttaggcgccatgcacaatcaa- gtaaaaggacgattgatttggagcagaaatgcgggagttgtg- cgtggtttaaaattatattattaatattaatggggtaacgg- attggggtggacgagtcagggttacgagcaggccaacgcaa- cagacggc
F[b]	cgaggatgaa[ggac**gggttagggt**a]tcgcaggccagggcaacga
G[a]	agcgcccccggaacccgtataatggcgcggagcgcccatgtcc- ggcggcgtggaggccggaa**ggttagggtta**aaaataattat- ggtttaaaaaaacgcgg
H	ggttataaggaggcgggagggtgag
I[a]	gtcggagggtca**gggttag**aggtggttg**gggtta**- ttggcagtagggt

[a]Sequence in bold type indicates units of the telomere repeat motif (TTAGGG) embedded within a HAR sequence (the reverse complement of this sequence occurs in HAR-C).
[b]Brackets represent core sequences, which resist truncation. An imperfect palindrome within HAR-D is underlined.

Subtelomeres 8 and 10 have almost identical sequences for another 3.5 kb beyond this point, finally diverging at a 'far' boundary located at least 17 kb from the nearest telomere repeat (the one at telomere 8). Subtelomere 7 is unusual because, although it has chromosome-unique sequences beyond HAR-CDEF (and was, therefore, mentioned above as part of the first group of seven subtelomeres), it possesses a small stretch of similarity to subtelomeres 8 and 10. The similarity extends to the far boundary, before transitioning again into chromosome-unique sequences (Figure 2). Alignment of the far boundary sequences in subtelomeres 7, 8 and 10 reveals a discrete transition between the shared sequence and chromosome-unique DNA (Figure 3B). This also suggests that the sequences between the near and far boundaries are a distinct part of the TLH region, which we refer to as region II.

Nine different HARs were identified within regions I and II, and their consensus sequences are listed in Table 1. Most HARs occur as tandemly-organized, unit-length copies, with the exceptions of HARs D, F and G, which can also be found in truncated forms (Figure 4). Among the nine different repeating units, A, B, C, F, G, H and I occur as tandem, head-to-tail arrays. HARs D and E do not occur in tandem but are interspersed, in whole or in part, with the tandemly-arranged HARs. Four HARs (C, D, E and F) are found in close association with one another in a compound array that ends just 181 bp away from the predicted *TLH* start codon. HARs G and H also occur in compound arrays, with the HAR-G repeat unit that is adjacent to HAR-H always being a truncated copy. The lower panel of Figure 4 shows the organization of HAR sequences at the original TLH-end that came from strain O-137 (78), and from a third strain, ML33. This illustrates the conservation of HAR block organization among different *M.oryzae* strains.

Some of the individual HARs have sequence motifs that suggest they might interact with the telomere, or with telomere-binding proteins. These motifs are highlighted with bold type in Table 1. For example, the first 7 nt of each HAR-C repeat unit form the sequence CCTAACC, which is a reverse complement of the telomere repeat sequence. HARs F and G contain TTAGGG motifs and, although some copies of HAR-F are truncated, all retain a core repeat region, which contains 1.5 iterations of the repeat motif, GGGTTAGGG. HAR-D also contains a significant feature in the form of a 23 bp imperfect palindrome (underlined in Table 1). This occurs within a 50 bp subsequence (bracketed in Table 1) that is conserved in all repeat units, even ones that have experienced truncation.

Terminal gene predictions

Initial gene predictions identified 128 putative genes among the subtelomeric fosmid assemblies, which were subsequently filtered to remove TEs and short proteins (see Materials and Methods) to produce a terminal/subtelomeric gene set comprising 113 putative genes. These genes code for proteins with lengths ranging from 44 to 2381 amino acids, with an average length of 434 amino acids. The percentage of subtelomeric sequence that is predicted to code for proteins is ∼38%, a figure that is very similar to the protein-coding proportion of the whole genome (66).

Thirty of the genes identified in the present study are not represented in the genome assembly. Sixteen of these correspond to *TLH* sequences, with some of them matching full-length genes and others coinciding with smaller ORFs that are found in mutated *TLH* copies. The 14 remaining genes were previously unidentified, either because they were in gaps that existed between the telomeres and the assembly, or because they were located in gaps between genomic contigs. Among the terminal gene set, 34 are single-copy genes, and 46 encode predicted proteins with significant alignments to at least one domain in the Pfam database (Supplementary Table S1). Many of these genes are of undetermined function and only two appear to encode obvious 'housekeeping' genes. Both of these genes occur at least 35 kb centromere-proximal to telomere 10, and are predicted to code for GTP cyclohydrolase and adenylsuccinate synthetase, which are involved in the biosynthesis of riboflavin and purines, respectively.

Secondary metabolism gene clusters in terminal regions

Three of the chromosome termini contain clusters of genes that are predicted to be involved in the biosynthesis or detoxification of secondary metabolites. For example, Pfam similarities indicate that the terminal sequences associated with telomere 14 contain four genes encoding CYP450s; and these are interspersed with genes encoding a putative FAD-oxidoreductase, a major facilitator superfamily (MFS) protein, as well as several proteins of unknown function. A BLASTx search revealed the presence of two additional CYP450 pseudogenes. A second putative secondary metabolism gene cluster near telomere 1 encodes two proteins having domains typical of transporters: one being an MFS protein and the other having canonical ABC transporter domains (Supplementary Table S1). Inspection of the genome sequence in the region flanking the fosmid insert indicated
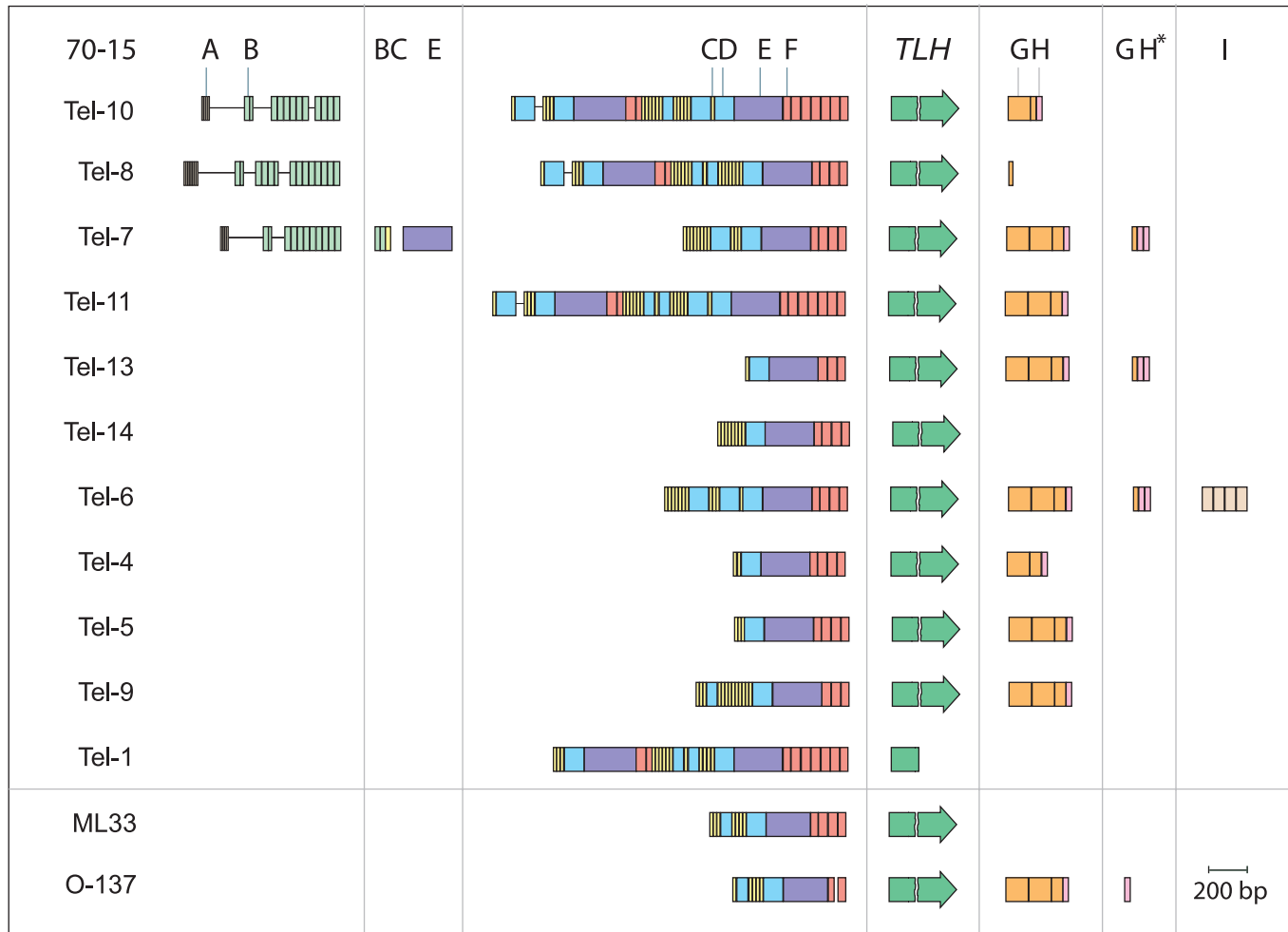
Figure 4. Variation in HAR organization at different chromosome ends. The HAR sequences listed in Table 1 are depicted in this figure as different-colored rectangles, with each individual rectangle representing a single iteration of a particular repeat sequence, and like-colors representing copies of the same repeat. *Additional, highly-mutated HAR-H units are not shown. The identity of each repeat is noted at the top of the figure. The position of each repeat block within its respective subtelomere region is shown in Figure 2C, where it is represented as a white box). Rectangles are drawn to scale, with narrower ones representing incomplete repeat units. The *TLH* genes are shown as green block arrows, with a gap included to indicate that it is not drawn to scale. The lower panel shows HAR organization in two additional *M.oryzae* strains.

that the cluster also codes for an oxidoreductase, as well as a methyltransferase that resembles the *Aspergillus fumigatus* LaeA protein, a global regulator of secondary metabolite gene clusters (79). Finally, analysis of genomic sequences that are closely linked to the insert in fosmid 72H05 (Tel-12) revealed a number of genes in a cluster that resembles the gibberellin biosynthetic cluster of *Gibberella fujikuroi* (80).

Terminal genes encoding secreted proteins

Avr genes usually encode secreted proteins that directly, or indirectly, trigger defense responses in plants and are, therefore, analogous to the variant surface protein genes in the various microbial human pathogens. None of the known *Avr* genes in *M.oryzae* are near to telomeres in 70-15. *Avr-Pita*, which is immediately adjacent to the telomere in isolate O-137, has two copies in 70-15, neither of which is present in the telomeric fosmids. The *Pwl2* and *Pwl3* genes, which confer avirulence to weeping lovegrass, also occupy

internal locations. The *Pwl1* gene is not present in 70-15 (66), although an AT-rich region associated with this gene is present in telomere-associated genomic contig 2.2068.

Three of the four cloned *M.oryzae Avr* genes encode proteins that are predicted to be secreted and which are, therefore, expected to come into direct contact with plant cells during infection. This prompted us to identify telomere-linked genes that code for secreted proteins. Among the terminal gene set, Signal 3.0P identified 23 that encode proteins with a probable signal peptide, of which 13 are predicted by ProtComp to be exported to the cell exterior (Supplementary Table S1). Two of the putative secreted proteins contain cellulose-binding motifs and glycosyl hydrolase domains and, therefore, likely have cell-wall-degrading capabilities. The remaining secreted proteins are of unknown function.

Gene duplications near chromosome termini

Proximal subtelomere domains of other organisms are characterized by the presence of genes that have become

amplified through spread to multiple chromosome ends, resulting in subtelomeric gene families (21). To determine if the *M.oryzae* subtelomeres/chromosome termini have played roles in the expansion of gene families, we used BlastClust to detect genes (other than *TLH*) that are present at more than one chromosome end. Surprisingly, we identified only three genes that meet this criterion. One occurs upstream of the HAR-CDEF array in region II of subtelomeres 7, 8 and 10, with a partial copy being present at an internal locus. The two other duplicated genes are present on the 6.5 kb segment that is duplicated in the non-*TLH* subtelomeres, 2 and 12. None of the gene duplications shows much sequence divergence subsequent to these duplications (maximum of 2 nt substitutions in any gene) and, although none of the mutations cause premature termination, Fgenesh predicts that some of the changes result in alternative splicing. However, we interpret these variable intron predictions with caution because the mutations do not occur in critical boundary sequences.

The paucity of duplicated genes in the fosmid sequence dataset suggests that the *M.oryzae* subtelomeres are not widely involved in the amplification of contingency genes. However, we also considered the possibility that the *M.oryzae* proximal domains are longer than the 10–40 kb of chromosome-unique sequence contained within these clones, and that more terminal sequence needed examination. To address this issue, we linked the fosmid sequences to the genome assembly using BLAST, and then referred to the genome sequence to search for gene duplications within 100 kb of the chromosome ends. Remarkably, among 148 new genes that were added to the terminal gene dataset, only one additional terminal duplication was discovered. In this case, a single gene located 15 kb from telomere 7 has a copy ∼54 kb away from telomere 8. We also sought to determine how many of the terminal genes have copies at internal genomic locations. BLAST searches of the whole genome sequence using the terminal gene set as a query revealed only nine genes having full-length duplications at internal loci and six showing evidence of partial duplication.

Finally, to ensure that we had not overlooked extensive segmental duplication by restricting the above analyses to predicted genes, the masked fosmid sequences were used to query the genome sequence using BLAST. Only one additional duplication greater than 1 kb in length was identified in this manner. Thus, with the exception of transposons and *TLH*-associated sequences, the *M.oryzae* chromosome termini show very little evidence of sequence duplication.

Transposon sequences at *M.oryzae* chromosome termini

The terminal regions of the *M.oryzae* chromosomes are particularly rich in transposon sequences. The colored boxes in Figure 5 show TE sequences that were identified in each of the fosmid clones. All 14 of the terminal regions contain transposon insertions within 40 kb of the chromosome tip. As mentioned above, there was even a MAGGY insertion in the ribosomal DNA array that terminates at telomere 3. This end is somewhat exceptional, however, because the others all contain sequences from several different transposons, with the result that at least one full-length member from each of the 11 transposon families in the *M.oryzae* genome

was found within 40 kb of at least one chromosome end (Table 2). A consequence of transposon abundance in the terminal regions is that repeated sequences comprise 24% of the terminal DNA in the fosmid clones characterized here. In contrast, only 9.7% of the sequence in the whole genome assembly consists of repeated DNA. With the exceptions of Mg-SINE and Occan, all transposon families that are present in the genome of 70-15 were over-represented in the terminal regions of the chromosomes (Figure 6).

Many of the terminal transposon insertions shown in Figure 5 appear to have undergone subsequent rearrangements. Several elements have other transposons inserted inside of them, with Pot2 and Pot4 having been particularly prone to disruption. In many cases, the secondary insertions are associated with the disappearance of one side of the disrupted element. A total of 12 elements show evidence of such rearrangement, and there are eight different chromosome ends that have been affected in this manner (Figure 5). Two good examples of this are the RETRO6-1 elements near telomeres 11 and 13, whose 5′ ends have been lost in association with a secondary Pot2 insertion (Figures 2 and 5). Similar insertion-rerrangement events appear to have been the predominant causes of truncations that have left 31 incomplete transposon copies scattered throughout the terminal regions. Additional transposon remnants are in the forms of 23 solo-LTRs, which were likely left behind when the full-length elements were excised from the genome via recombination.

DISCUSSION

We report here the sequencing of all 14 chromosome ends in the fungus, *M.oryzae*, an important pathogen of rice and other cereal grasses, and a model organism for the study of fungus–plant interactions, infection-related development and signaling during plant pathogenesis. This study is novel in its completeness, and makes *M.oryzae* the first organism to have all of its chromosome ends sequenced from the chromosome-unique regions out to the very tips. In this study, we present the results of an exhaustive analysis of the terminal sequences, and provide a comprehensive picture of the organization of these important chromosome regions.

Telomere structure in *M.oryzae*

In most vertebrates and plants, the telomere repeat array extends several kilobases from the chromosome tip (21,81). In contrast, the few fungal telomeres that have been characterized to date are much shorter, averaging ∼350 bp in *S.cerevisiae* (1,3) and ∼130 bp in *M.oryzae*'s closer relative, *Aspergillus nidulans* (82). Based on the raw sequence data, we estimate the average length of the *M.oryzae* telomere repeats to be ∼150 bp. Although, their lengths varied widely among fosmid clones (6 to 30 repeats), it is possible that this was due to slippage during replication in *Escherichia coli*.

Several organisms, including many plants, humans and *S.cerevisiae*, have large arrays of telomere repeats at internal chromosome locations (24,83). There was no evidence for the existence of extended, interstitial telomere arrays in the genome of the reference isolate of *M.oryzae* studied here because only 14 fragments were detected in Southern hybridization
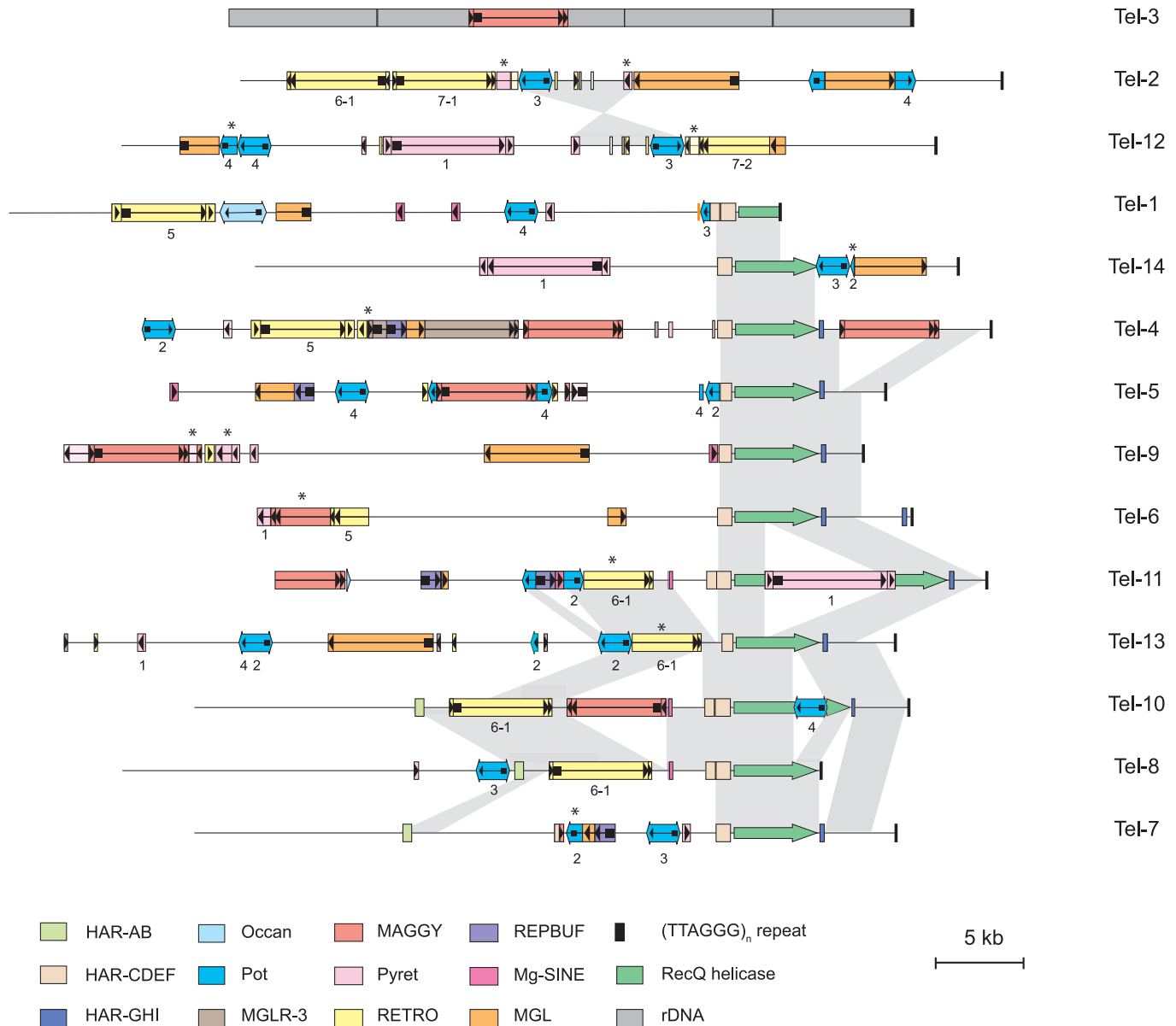
Figure 5. Transposon insertions in the terminal regions of *M.oryzae* chromosomes. Shown are the terminal segments of all 14 *M.oryzae* chromosomes, as represented by the full-length insert in each of the telomeric fosmid clones. Where possible, the sequences are aligned at the starting codon of the *TLH* ORF (green arrow). Transposable elements are depicted in different ways to illustrate the type of element. Inverted repeat transposons are shown as blue boxes with double-headed arrows. Retrotransposons are illustrated as rectangles, with internal arrows that show orientation. LTR-retrotransposons are drawn as rectangles with small boxes at each end, representing the LTRs. Small boxes that are not attached to rectangles represent solo-LTRs. Colored boxes lacking arrowheads and/or arrowtails correspond to truncated elements. The key at the bottom of the figure shows the color used for each transposon family. For transposon families that contain more than one member element, the identity of the element is noted on the figure (e.g. RETRO5, RETRO6-1, RETRO7-1). Partial elements are not labeled because it was not possible to distinguish between element subfamilies. Asterisks indicate elements showing evidence of having undergone ectopic recombination. Gray areas connecting different chromosome ends represent regions of sequence identity. The scale is shown at the bottom of the figure.

studies and no telomere-like arrays were detected in the genome sequence, apart from one sequence that we have shown to be a mis-assembly (C. Rehmeyer and M. Farman, unpublished data).

Chromosome ends that lack helicase genes

The *M.oryzae* rDNA array was previously mapped near telomere 3 on chromosome II (84,85). Here we show that, in 70-15, the array extends all the way to the telomere, and is interrupted near its terminus by an insertion of the MAGGY retroelement. Although transposons have been found in the rDNA of other organisms (86–88), this is the first reported example of an LTR-retrotransposon insertion. The MAGGY element appears to be the only transposon present in the ~2 Mb rDNA array, indicating that the majority of the repeat units are resistant to transposon insertion and that, perhaps, the subtelomeric portion has a higher-level organization that differs from the rest of the array. Alternatively, it is possible that the internal regions of the array have

been homogenized via unequal crossing over between rDNA repeats. Ribosomal RNA-encoding genes have been found at telomeres in a number of organisms, including the fungus *Neurospora crassa* (89), Cryptomonad and Chlorarachnio-phyte algae (90,91), the parasites *Giardia lamblia* (92) and *Encephalitozoon cuniculi* (93), *Caenorhabditis elegans* (94) and *Arabidopsis thaliana* (95). The human rDNA clusters occur on the p-arms of the acrocentric chromosomes but, in this case, the arrays are separated from the telomere by an incompletely characterized sequence (86). Bioinformatic mining of unassembled sequences, using TERMINUS, revealed that rDNA sequences are associated with telomeres in several other fungi, including *Fusarium graminearum* and *Coprinus cinereus*, although it is not yet known if the telomere-linked sequences are contiguous with the main rDNA array (C. Rehmeyer and M. Farman, unpublished data).

Telomeres 2 and 12 are unusual because they lack the *TLH* genes and/or HARs that are found at all of the other (non-rDNA) chromosome ends. However, analysis of a homologous version of telomere 2 from another *M.oryzae* strain (ML33) revealed that it contains a TLH region. Therefore, it would appear that telomere 2 in 70-15 (and possibly telomere 12 also) is a truncated version of a chromosome end that was once a TLH-end. Despite the absence of TLH regions, telomeres 2 and 12 have been faithfully propagated through multiple generations (M. Farman, unpublished data), and in the case of telomere 2, through a genetic cross (59). Thus, it is clear that the *TLH* genes, their associated repeats and surrounding sequences do not need to be present at a chromosome end for proper telomere function and/or maintenance.

Subtelomere domains in *M.oryzae*

The majority of *M.oryzae* chromosome ends contain at least a part of the core sequence we refer to as TLH region I. Based on its presence at multiple chromosome ends, region I is analogous to the subtelomere regions that have been described for other organisms (21). In addition, this region contains a number of short tandem repeats, which are a dominant feature of subtelomeres in almost all organisms, including the protists *T.brucei* (96) and *P.falciparum* (97–99), the fungi *P.carinii* (25) and *S.cerevisiae* (http://www.nottingham.ac.uk/biology/contact/academics/louis/endsdata.phtml), plants (100) and humans (101). The sole exception so far is *C.elegans*, which lacks a defined subtelomere sequence (94). The specific organization of TLH region I is strikingly similar to the distal portion of the *S.cerevisiae* subtelomeres, which consist of two sequences, known as the X and Y′ elements (102). X-elements consist of various short repeats (STRs) fused into a single array. They are present at all of the 32
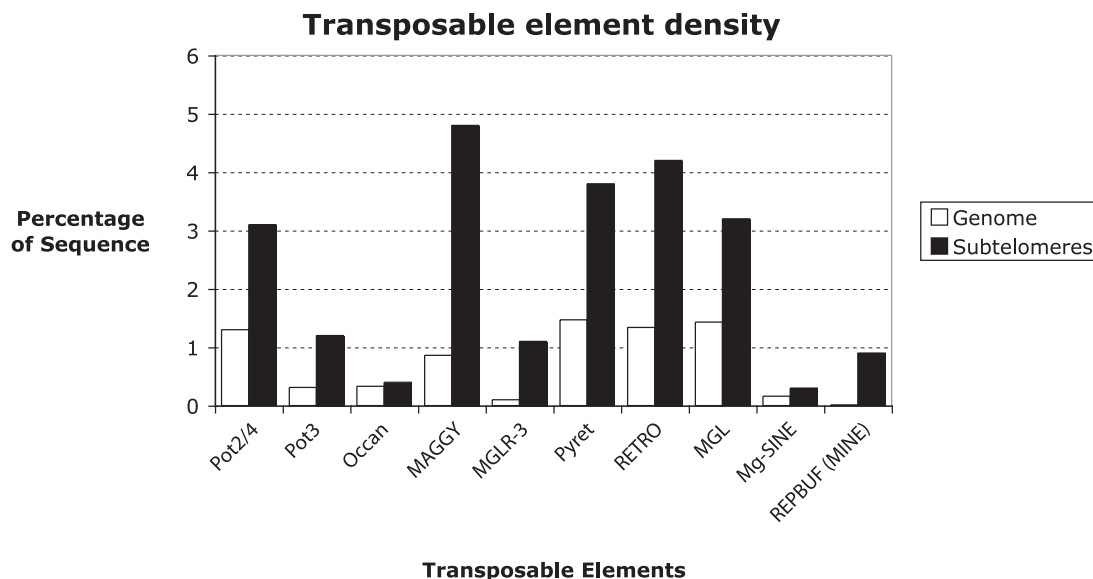
Table 2. Distribution of transposable elements in the terminal regions of *M.oryzae* chromosomes

Repeat name	Full-length	Truncated	Solo LTRs
LTR retrotransposons			
MAGGY	6	2	3
MGLR-3	(1)[a]	0	3
Pyret	3	4	8
Retro5	1	1	9
Retro6	3	2	0
Retro7	1	1	0
Non-LTR retrotransposons			
MGL	3	10	—
Mg-SINE-A	5	3	—
REPBUF	5	0	—
Inverted Repeat DNA transposons			
Pot2	3 (1)	3	—
Pot3	5	1	—
Pot4	4 (1)	3	—
Occan	1	1	—
TOTAL	42	31	21

[a]Parentheses indicate elements that are complete but are interrupted by another transposon inserted inside them.

Figure 6. Comparison of transposon density at subtelomeres versus the rest of the genome. The subtelomere regions are defined as the sequences contained in the 14 fosmids. In calculating the percentage of sequence occupied by each element, both full-length and partial copies were considered.

S.cerevisiae chromosome ends (103), where they either abut the telomere repeats directly, or are separated from the telomere by one or more copies of the Y′ element. Y′ elements consist of a 5–6 kb sequence that codes for a helicase gene. Y′ is not found at all chromosome ends but, when it is present, it can exist in tandem arrays, with up to four copies being present (104). The general organization of the HAR-CDEF arrays resembles that of the X-elements—both consist of different sequence motifs fused into a single block, and both possess repeat units that contain telomere-like sequences (105). HAR-CDEF and X also occupy equivalent positions within the subtelomere, occurring upstream of the helicase genes. In *S.cerevisiae*, X forms a boundary between the distal and proximal subtelomere domains, whereas HAR-CDEF either marks the transition into TLH region II, or into chromosome-unique sequences. Based on the above parallels, it is clear that TLH region I is analogous to the distal subtelomere domains present in *S.cerevisiae*.

The organization of TLH region I near the telomere repeats suggests that most of the *M.oryzae* chromosome termini have undergone extensive truncation. Indeed, as mentioned above, telomere 2 is clearly a truncated TLH-end that has lost at least 8 kb of core *TLH* sequences. Furthermore, analysis of telomeric sequences captured in the *Magnaporthe* genome sequencing project identified five *de novo* telomere truncations that had arisen in the fungal culture for DNA isolation (C. Rehmeyer and M. Farman, unpublished data). In one case, the deletion encompassed ∼35 kb of terminal sequence, again extending well into the chromosome-unique sequences. Therefore, it is clear that the frequent deletion of terminal sequences is not specific to the telomere that contains *Avr-Pita* (54,57) but is a general phenomenon in *M.oryzae*.

Most of the *M.oryzae* subtelomeres lack proximal domains

In other organisms, proximal domains are also recognized as sequences duplicated at different chromosome ends. However, they are distinct from their distal neighbors because they are separated from the latter sequences by specific boundary elements, and their constituent sequences tend to be much less widely dispersed than those found in the distal domains. Most importantly, the proximal domains usually contain genes (21). Most of the *M.oryzae* subtelomeres transition into chromosome-unique sequences upstream of HAR-CDEF and, therefore, based on the lack of sequence duplication, it is clear that these chromosome ends lack proximal domains. On the other hand, five TLH-ends have stretches of sequence similarity that extend beyond the HAR-CDEF block into region II. Initially, the less extensive duplication of region II and its position relative to the HAR-CDEF boundary, suggested that it might correspond to a proximal domain. However, if one disregards the RETRO6-1 insertion, region II contains only one predicted gene (of unknown function), and only three of the subtelomeres showing extended region II similarity have a copy. As such, the duplication of this region has not given rise to the extensive gene amplification that is characteristic of the proximal domains seen in other organisms. Instead, the presence of additional tandem repeat blocks within region II,

as well as its discrete boundary, suggests that it may simply be a part of an extended distal domain.

Aside from region II, there is very little evidence for the existence of proximal domains in 70-15. Not counting the ribosomal RNA genes at telomere 3, or the abundant transposon sequences, only two additional segmental duplications were detected within 100 kb of the chromosome termini, encompassing only three genes. Based on this information, it would appear that *M.oryzae* does not use subtelomere duplication to amplify genes with roles in pathogenesis. However, as mentioned above, the hybrid nature of 70-15 means that the subtelomere structures presented here do not give a perfectly accurate reflection of subtelomere organization in natural, rice-infecting strains of *M.oryzae*. Previous studies of telomere-linked sequences in the natural field isolate, Guy11—a recurrent parent in the crosses that gave rise to 70-15, indicated that a sequence linked to telomere 2 was duplicated near telomere 8 (59). Recently, we have shown that ∼6 kb of 'chromosome-unique' sequence immediately upstream of the *TLH* region at telomere 8 are identical to the telomere-adjacent sequences in Guy11/70-15 telomere 2. This shows that Guy11 has at least one subtelomeric duplication that is not represented in 70-15, and suggests that the 'true' proximal domains, when they are present, begin upstream of the far boundary sequence. In addition to telomere 8, three more telomeres in Guy11 were either not inherited by 70-15, or were rearranged during meiosis. One of these is the rDNA telomere, which leaves only two additional telomeres in Guy11 that potentially could exhibit proximal domain duplications. As such, it would appear that Guy11 also lacks the extensive subtelomeric gene amplification that, until now, have been found in all eukaryotic microbes whose telomeres have been studied in detail.

Telomere-linked helicase genes

One of the most striking features of the *M.oryzae* distal domains regions is the presence of the telomere-linked RecQ-helicase gene (*TLH1*) described by Gao *et al*. (78). Only four of the 11 *TLH* gene copies in *M.oryzae* appear to encode full-length products, with the remaining copies having been interrupted by chromosome truncation and transposon insertions, or mutated by nucleotide substitutions/deletions, resulting in stop codons and frameshifts, respectively (C. Rehmeyer and M. Farman, unpublished data). Telomere-linked helicase genes have been reported for a number of other ascomycete and basidiomycete fungi, including *Schizosaccharomyces pombe* (106), *Ustilago maydis* (107) and *Metarhizium anisopliae* (108). Recently, we have identified telomere-linked helicase genes in 10 additional fungi by searching fungal genome sequences. This amounts to a total of 15 *TLH*-positive fungi among 29 genomes analyzed. Among these is the *S.cerevisiae* Y-help protein, which, under careful scrutiny, appears to be a highly-diverged RecQ-helicase (C. Rehmeyer and M. Farman, unpublished data). Interestingly, we have not found *TLH* genes in organisms outside the kingdom fungi.

The cellular function of *TLH* genes is, yet unknown. However, the most closely related, non-telomeric human RecQ-helicases, WRN and BLM, are involved in telomere function and maintenance, and their mutation results in

Werner's and Bloom's syndromes, respectively (109–112). Considering the relationship to BLM and WRN and, especially their physical association with telomeres, it seems reasonable to speculate that the *TLH* genes are also involved in telomere function and/or maintenance. This begs the question, 'Why do these genes tend to occur in physical proximity to the chromosome regions that are the presumed targets of their products?' The answer to this question may lie in the recent observation that the *TLH* genes in *S.pombe* are expressed only during telomere crisis, induced by experimental deletion of the telomerase gene (106). Thus, it may be that the proximity of the *TLH* genes to chromosome ends provides a mechanism for their regulation based on telomere integrity—possibly one that takes advantage of the 'TPE' (12) to inhibit *TLH* gene expression until the telomere is compromised. The telomere-like sequences in the HAR-CDEF array are, perhaps, significant in this regard, as they potentially could function in the propagation of a silencing effect at *M.oryzae* chromosome ends, as has been reported for the X-elements in *S.cerevisiae* (113).

Telomere-linked genes in *M.oryzae*

Based on precedents set by the subtelomeres of other eukaryotic microbial pathogens, we had expected the corresponding regions in *M.oryzae* to contain large families of genes with roles in pathogenesis. However, within the 14 fosmids, only two genes were identified that are predicted to have functions with obvious relevance to the pathogenic process. These are the glycolsyl hydrolases near telomeres 9 and 12, both of which are predicted to have cell wall-degrading activity. None of the *Avr* genes that are known to be present in isolate 70-15 are found in subtelomeric locations. Neither are there genes encoding homologs of known avirulence proteins in these regions. Nevertheless, despite the absence of recognizable *Avr* genes, 13 subtelomeric genes are predicted to encode secreted proteins, which, by virtue of their extracellular locations, have the potential to interact with host plant cells to trigger defense or, conversely, to facilitate pathogenesis. Two of these genes are the glycosyl hydrolases mentioned above, while the remaining 11 have unknown functions. Given the reasonable expectation that any 'foreign' protein encountered by a plant cell could elicit a defense response, together with the precedents for telomere-proximity of *Avr* genes in *M.oryzae* (57) and of analogous surface antigen genes in other microbial pathogens (25,114,115), the telomere-linked 'secretome' of *M.oryzae* deserves closer investigation.

The chromosome ends terminating in telomeres 1, 12 and 14 contain gene clusters that are predicted to be involved in secondary metabolism. Fungi are unique among the eukaryotes, in that secondary metabolites tend to be synthesized by clusters consisting of closely-linked genes, which code for all the enzymes necessary to synthesize the end-products from primary metabolism-derived precursors. In addition, the clusters usually code for regulators of their own expression and for efflux pumps that deliver the end-product(s) out of the cell (116–120). A wide variety of gene clusters have been identified in a number of plant pathogenic fungi and many code for the synthesis of toxins that suppress host defenses, promote nutrient leakage or alter plant metabolism

(121,122). The gene cluster near *M.oryzae* telomere 12 is similar to the gibberellin biosynthetic cluster of *G.fujikuroi* (80,123) and contains the key gene that codes for the pathway-specific enzyme copalyl diphosphate synthase (123). This particular cluster might code for the production of a metabolite similar to gibberellin, a phytohormone with profound effects on rice development (53). The other telomere-linked gene clusters are quite small and, therefore, it is unlikely that they synthesize complex molecules. Instead, it is possible, for example, that the constituent genes function to detoxify secondary metabolites that are produced as part of the host plant's defense, or they may inactivate antibiotics released by other microbes co-colonizing the lesions caused by *M.oryzae* infections.

As the name suggests, the chemical products of secondary metabolism pathways are non-essential, although they often provide adaptive advantages under certain conditions (124–126). Therefore, the genes contained within these clusters could be considered contingency genes and, as such, the general association between contingency genes and telomeres (27) would appear still to hold true for *M.oryzae*. In fact, many fungi contain secondary metabolism clusters near to telomeres (C. Rehmeyer, W. Li and M. Farman, unpublished data), suggesting that there is an advantage to this arrangement. One possibility is that the subtelomeres provide an environment that favors the generation of cluster diversity through accelerated evolution of the constituent genes.

It was not possible to predict functions for the majority of the genes identified in the subtelomere regions, due either to a lack of database matches or because the only similarities were to other predicted or hypothetical proteins. It is quite possible that a number of these genes are involved in pathogenesis but remain cryptic due to the limited information that is available for fungal pathogenicity genes. In this regard, it is worth noting that two of the four cloned *M.oryzae Avr* genes lack matches to the GenBank and Pfam databases, and a large proportion of non-pathogenic insertion mutants of *M.oryzae* were found to have insertions in genes of unknown function (M. Farman and R. Dean, unpublished data).

M.oryzae chromosome ends are littered with transposons and transposon relics

The terminal regions of the *M.oryzae* chromosome exhibit a superabundance of TE insertions. Nevertheless, with the exception of a single MAGGY insertion near telomere 4, transposons are notably absent from the core sequence between the telomere-linked helicase and the telomere repeat. Given that the TLH region does not appear to be essential for telomere function and/or maintenance, this suggests that the *TLH* to telomere interval is inaccessible to mobile elements. A possible reason for this may be that it adopts a heterochromatic structure, as has been shown for a number of other organisms (49,127–129).

Some organisms possess families of transposons that are exclusively subtelomeric in their distribution. Examples are the Ty5 elements in *S.cerevisiae* (130), the non-LTR retrotransposons, GilM and GilT, in *G.lambia* (131) and a LINE-like retroelement, Zepp, in Chlorella (132). The telomeres of *Drosophila* completely lack the simple tandem repeats present at the chromosome ends of most organisms.

Instead, they possess telomere-specific retrotransposons, whose mobilization maintains a buffer against terminal degradation (133,134). In the *M.oryzae* strain sequenced here, all of the TEs that occupy the subtelomeres are also found scattered throughout the genome, so that none are exclusively associated with chromosome termini.

The *M.oryzae* chromosome ends contain large numbers of incomplete transposon copies, many of which are interrupted by other elements inserted within them (super-insertions). In several cases, one terminus of the invading element is flanked by sequences from the recipient transposon but the other terminus, instead of transitioning into sequences from the other end of the recipient element, is bordered by unrelated sequences. Such arrangements are most economically explained by the occurrence of ectopic recombination between the invading element and another copy nearby, resulting in deletion, or inversion, of the intervening DNA sequences. Strain 70-15 exhibits 12 such rearrangements at eight chromosome ends. Therefore, if the patterns of transposon insertion differ among *M.oryzae* strains, and ectopic, inter-element recombination occurs in a stochastic fashion, we would expect this to result in significant strain-to-strain variation in terminal gene content. Evidence in support of this prediction comes from observations that transposons containing super-insertions (and super-super-insertions, etc.) often mark boundaries of synteny between 70-15 chromosome termini and homologous chromosome ends in another *M.oryzae* strain (M. Farman, unpublished data).

Do the *M.oryzae* subtelomeres have roles in pathogenic adaptation?

M.oryzae Avr genes that reside in subtelomeric locations tend to be genetically unstable, which leads to the frequent appearance of mutants that are able to infect a formerly resistant host cultivar. This observation, together with the finding that telomere-associated sequences are among the most highly variable regions of the *M.oryzae* genome, originally led us to suspect that a telomere-based switching mechanism might underlie the pathogenic variability of this fungus. However, analysis of the subtelomere regions failed to identify the massively amplified families of surface protein genes that are characteristic of the above-mentioned human pathogens that use this method of host adaptation. In fact, none of the terminal genes that are predicted to code for secreted proteins exhibited subtelomeric amplification. As such, it seems highly unlikely that *M.oryzae* switches the expression of subtelomeric surface proteins to evade recognition by the host plant's surveillance system. One caveat to this conclusion is that several of the terminally-located genes in 70-15 have copies at internal locations. This raises the possibility that the evolution of these genes could be accelerated through recombination with interstitial copies, a phenomenon that sometimes occurs during antigenic switching in *T.brucei* (47).

Despite the lack of evidence for surface protein switching, the fact remains that several *M.oryzae Avr* genes map to the terminal regions of chromosomes (54–56). Based on our analysis of the organization of these regions, we propose that this arrangement may still provide an adaptive advantage. First,

it appears that all of the *M.oryzae* chromosome termini are susceptible to truncation and, therefore, we would predict that any gene residing near to a chromosome tip should exhibit a high level of genetic instability, similar to that of *Avr-Pita*. Second, the more centromere-proximal regions of the chromosome ends contain numerous transposon sequences, many of which show evidence of having undergone rearrangements that resulted in the deletion of intervening sequences. Clearly, any genes that reside in these sub-terminal regions would be susceptible to deletion through ectopic recombination between flanking transposon copies, or prone to inactivation by transposon insertion. Based on these findings, we predict that different *M.oryzae* strains will very enormously in their sub-terminal gene content. Moreover, we hypothesize that the terminal regions of the *M.oryzae* chromosome ends function in niche adaptation by promoting the deletion/disruption of genes whose non-uniform distribution within the pathogen population can be adaptive. *Avr* genes are excellent examples in this regard because an *M.oryzae* population consisting of strains with different *Avr* gene profiles will be able to colonize a more diverse collection of host genotypes than will one that is genetically homogeneous. Of course, one might expect *Avr* genes that reside in transposon-rich, internal regions of the genome to be prone to deletion also. However, ectopic recombination tends to be especially frequent near chromosome ends (16,19,20), and terminal rearrangements are less likely to have catastrophic effects on genome architecture. Accordingly, we predict that characterization of telomere-linked genes that vary in their presence/absence among different *M.oryzae* strains will lead to the identification of novel *Avr* genes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

REFERENCES

1. Zakian,V.A. (1996) Structure, function, and replication of *Saccharomyces cerevisiae* telomeres. *Annu. Rev. Genet.*, **30**, 141–172.
2. Meyne,J., Ratliff,R.L. and Moyzis,R.K. (1989) Conservation of the human telomere sequence (TTAGGG)n among vertebrates. *Proc. Natl Acad. Sci. USA*, **86**, 7049–7053.
3. Walmsley,R.W., Chan,C.S., Tye,B.K. and Petes,T.D. (1984) Unusual DNA sequences associated with the ends of yeast chromosomes. *Nature*, **310**, 157–160.
4. Richards,E.J. and Ausubel,F.M. (1988) Isolation of a higher eukaryotic telomere from *Arabidopsis thaliana*. *Cell*, **53**, 127–136.
5. Nakamura,T.M., Morin,G.B., Chapman,K.B., Weinrich,S.L., Andrews,W.H., Lingner,J., Harley,C.B. and Cech,T.R. (1997) Telomerase catalytic subunit homologs from fission yeast and human. *Science*, **277**, 955–959.
6. Shore,D. (2001) Telomeric chromatin: replicating and wrapping up chromosome ends. *Curr. Opin. Genet. Dev.*, **11**, 189–198.
7. Cooper,J.P., Nimmo,E.R., Allshire,R.C. and Cech,T.R. (1997) Regulation of telomere length and function by a Myb-domain protein in fission yeast. *Nature*, **385**, 744–747.
8. Krauskopf,A. and Blackburn,E.H. (1996) Control of telomere growth by interactions of RAP1 with the most distal telomeric repeats. *Nature*, **383**, 354–357.
9. van Steensel,B. and de Lange,T. (1997) Control of telomere length by the human telomeric protein TRF1. *Nature*, **385**, 740–743.
10. Aparicio,O.M., Billington,B.L. and Gottschling,D.E. (1991) Modifiers of position effect are shared between telomeric and silent mating-type loci in *S.cerevisiae*. *Cell*, **66**, 1279–1287.
11. Baur,J.A., Zou,Y., Shay,J.W. and Wright,W.E. (2001) Telomere position effect in human cells. *Science*, **292**, 2075–2077.
12. Gottschling,D.E., Aparicio,O.M., Billington,B.L. and Zakian,V.A. (1990) Position effect at *S. cerevisiae* telomeres: reversible repression of Pol II transcription. *Cell*, **64**, 751–762.
13. Andrulis,E.D., Neiman,A.M., Zappulla,D.C. and Sternglanz,R. (1998) Perinuclear localization of chromatin facilitates transcriptional silencing. *Nature*, **394**, 592–595.
14. Bass,H.W., Marshall,W.F., Sedat,J.W., Agard,D.A. and Cande,W.Z. (1997) Telomeres cluster *de novo* before the initiation of synapsis: a three-dimensional spatial analysis of telomere positions before and during meiotic prophase. *J. Cell. Biol.*, **137**, 5–18.
15. Dernburg,A.F., Sedat,J.W., Cande,W.Z. and Bass,H.W. (1995) Cytology of telomeres. In Blackburn,E.H. and Greider,C.W. (eds), *Telomeres*. Cold Spring Harbor Press, Cold Spring Harbor, NY, pp. 295–338.
16. Freitas-Junior,L., Bottius,E., Pirrit,L.A., Deitsch,K.W., Scheidig,C., Guinet,F., Nehrbass,U., Wellems,T.E. and Scherf,A. (2000) Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum.. Nature*, **407**, 1018–1022.
17. Gotta,M., Laroche,T., Formenton,A., Maillet,L., Scherthan,H. and Gasser,S.M. (1996) The clustering of telomeres and colocalization with Rap1, Sir3, and Sir4 proteins in wild-type *Saccharomyces cerevisiae*. *J. Cell. Biol.*, **134**, 1349–1363.
18. Laroche,T., Martin,S.G., Gotta,M., Gorham,H.C., Pryde,F.E., Louis,E.J. and Gasser,S.M. (1998) Mutation of yeast *Ku* genes disrupts the subnuclear organization of telomeres. *Curr. Biol.*, **8**, 653–656.
19. Mefford,H.C. and Trask,B.J. (2002) The complex structure and dynamic evolution of human subtelomeres. *Nature Rev. Genet.*, **3**, 91–102.
20. Scherf,A., Figueiredo,L.M. and Freitas-Junior,L.H. (2001) Plasmodium telomeres: a pathogen's perspective. *Curr. Opin. Microbiol.*, **4**, 409–414.
21. Pryde,F.E., Gorham,H.C. and Louis,E.J. (1997) Chromosome ends: all the same under their caps? *Curr. Opin. Genet. Dev.*, **7**, 822–828.
22. Britten,R.J. (1998) Precise sequence complementarity between yeast chromosome ends and two classes of just-subtelomeric sequences. *Proc. Natl Acad. Sci. USA*, **95**, 5906–5912.
23. Flint,J., Bates,G.P., Clark,K., Dorman,A., Willingham,D., Roe,B.A., Micklem,G., Higgs,D.R. and Louis,E.J. (1997) Sequence comparison of human and yeast telomeres identifies structurally distinct subtelomeric domains. *Hum. Mol. Genet.*, **6**, 1305–1314.
24. Riethman,H.C., Ambrosini,A., Castaneda,C., Finklestein,J., Hu,X.-L., Mudunuri,U., Paul,S. and Wei,J. (2004) Mapping and initial analysis of human subtelomeric sequence assemblies. *Genome Res.*, **14**, 18–28.
25. Keely,S.P., Renauld,H., Wakefield,A.E., Cushion,M.T., Smulian,A.G., Fosker,N., Fraser,A., Harris,D., Murphy,L., Price,C. *et al.* (2005) Gene arrays at *Pneumocystis carinii* telomeres. *Genetics*, **105**, 1589–1600.
26. Trask,B.J., Friedman,C., Martin-Gallardo,A., Rowen,L., Akinbami,C., Blankenship,J., Collins,C., Giorgi,D., Iadonato,S., Johnson,F. *et al.* (1998) Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Hum. Mol. Genet.*, **7**, 13–26.
27. Barry,J.D., Ginger,M.L., Burton,P. and McCulloch,R. (2003) Why are parasite contingency genes often associated with telomeres? *Int. J. Parasitol.*, **33**, 29–45.
28. Hernandez-Rivas,R., Hinterberg,K. and Scherf,A. (1996) Compartmentalization of genes coding for immunodominant antigens to fragile chromosome ends leads to dispersed subtelomeric gene families and rapid gene evolution in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.*, **78**, 137–148.
29. Naumov,G.I., Naumova,E.S., Sancho,E.D. and Korhola,M.P. (1996) Polymeric *SUC* genes in natural populations of *Saccharomyces cerevisiae*. *FEMS Microbiol. Lett.*, **135**, 31–35.
30. Moxon,E.R., Rainey,P.B., Nowak,M.A. and Lenski,R.E. (1994) Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol.*, **4**, 24–33.
31. Charron,M.J., Read,E., Haut,S.R. and Michels,C.A. (1989) Molecular evolution of the telomere-associated *MAL* loci of *Saccharomyces*. *Genetics*, **122**, 307–316.
32. Naumov,G.I., Naumova,E.S. and Michels,C.A. (1994) Genetic variation of the repeated *MAL* loci in natural populations of *Saccharomyces cerevisiae* and *Saccharomyces paradoxus*. *Genetics*, **136**, 803–812.
33. Naumov,G., Turakainen,H., Naumova,E., Aho,S. and Korhola,M. (1990) A new family of polymorphic genes in *Saccharomyces cerevisiae*: alpha-galactosidase genes *MEL1-MEL7*. *Mol. Genet. Genomics*, **224**, 119–128.
34. Naumov,G.I., Naumova,E.S. and Louis,E.J. (1995) Genetic mapping of the alpha-galactosidase *MEL* gene family on right and left telomeres of *Saccharomyces cerevisiae*. *Yeast*, **11**, 481–483.
35. Teunissen,A.W. and Steensma,H.Y. (1995) Review: the dominant flocculation genes of *Saccharomyces cerevisiae* constitute a new subtelomeric gene family. *Yeast*, **11**, 1001–1013.
36. Gancelo,J.M. (2001) Control of pseudohyphae formation in *Saccharomyces cerevisiae*. *FEMS Microbiol. Rev.*, **25**, 107–123.
37. Rachidi,N., Martinez,M.J., Barre,P. and Blondin,B. (2000) *Saccharomyces cerevisiae PAU* genes are induced by anaerobiosis. *Mol. Microbiol.*, **35**, 1421–1430.
38. Viswanathan,M., Muthukumar,G., Cong,Y.S. and Lenard,J. (1994) Seripauperins of *Saccharomyces cerevisiae*: a new multigene family encoding serine-poor relatives of serine-rich proteins. *Gene*, **148**, 149–153.
39. Naumov,G.I., Naumova,E.S., Sancho,E.D. and Korhola,M.P. (1996) Polymeric *SUC* genes in natural populations of *Saccharomyces cerevisiae*. *FEMS Microbiol. Lett.*, **135**, 31–35.
40. Leech,J.H., Barnwell,J.W., Miller,L.H. and Howard,R.J. (1984) Identification of a strain-specific malarial antigen exposed on the surface of *Plasmodium falciparum*-infected erythrocytes. *J. Exp. Med.*, **159**, 1567–1575.
41. Nakamura,Y. (1998) The major surface antigen of *Pneumocystis carinii*. *FEMS Immunol. Med. Lett.*, **22**, 67–74.
42. Vanhamme,L., Lecordier,L. and Pays,E. (2001) Control and function of the bloodstream variant surface glycoprotein expression sites in *Trypanosoma brucei. Int. J. Parasitol.*, **31**, 523–531.
43. Wada,M. and Nakamura,Y. (1996) Unique telomeric expression site of major-surface-glycoprotein genes of *Pneumocystis carinii*. *DNA Res.*, **3**, 55–64.
44. Scherf,A., Hernandez-Rivas,R., Buffet,P., Bottius,E., Benatar,C., Pouvelle,B., Gysin,J. and Lanzer,M. (1998) Antigenic variation in malaria: *in situ* switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in *Plasmodium falciparum*. *EMBO J.*, **17**, 5418–5426.
45. Barry,J.D., Graham,S.V., Fotheringham,M., Graham,V.S., Kobryn,K. and Wymer,B. (1998) VSG gene control and infectivity strategy of metacyclic stage *Trypanosoma brucei*. *Mol. Biochem. Parasitol.*, **91**, 93–105.

46. Donelson,J.E., Hill,K.L. and El-Sayed,N.M. (1998) Multiple mechanisms of immune evasion by African trypanosomes. *Mol. Biochem. Parasitol.*, **91**, 51–66.

47. Robinson,N.P., Burman,N., Melville,S.E. and Barry,J.D. (1999) Predominance of duplicative *VSG* gene conversion in antigenic variation in African trypanosomes. *Mol. Cell. Biol.*, **19**, 5839–5846.

48. Duraisingh,M.T., Voss,T.S., Marty,A.J., Duffy,M.F., Good,R.T., Thompson,J.K., Freitas-Junior,L.H., Scherf,A., Crabb,B.S. and Cowman,A.F. (2005) Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in *Plasmodium falciparum*. *Cell*, **121**, 13–24.

49. Freitas-Junior,L.H., Hernandez-Rivas,R., Ralph,S.A., Montiel-Condado,D., Ruvalcab-Salazar,O.K., Rojas-Meza,A.P., Mancio-Silva,L., Leal-Silvestre,R.J., Gontijo,A.M., Shorte,S. *et al.* (2005) Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites. *Cell*, **121**, 25–36.

50. De Las Penas,A., Pan,S.J., Castano,I., Alder,J., Cregg,R. and Cormack,B.P. (2003) Virulence-related surface glycoproteins in the yeast pathogen *Candida glabrata* are encoded in subtelomeric clusters and subject to RAP1- and SIR-dependent transcriptional silencing. *Genes Dev.*, **17**, 2245–2258.

51. Couch,B.C. and 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.

52. Hebert,T.T. (1971) The perfect state of *Pyricularia grisea*. *Phytopathology*, **61**, 83–87.

53. Bonman,J.M. (1992) Durable resistance to rice blast disease—environmental influences. *Euphytica*, **63**, 115–123.

54. Valent,B. and Chumley,F.G. (1994) A virulence genes and genetic instability. In Zeigler,R.S., Leong,S.A. and Teng,P.S. (eds), *Rice Blast Disease*. CAB International, Wallingford, UK, pp. 111–153.

55. Luo,X.-L., Yin,L.-F., Koyanagi,S., Farman,M.L., Kusaba,M. and Yaegashi,H. (2005) Genetic mapping and chromosomal assignment of *Magnaporthe oryzae* avirulence genes *AvrPik*, *AvrPiz*, and *AvrPiz-t* controlling cultivar specificity on rice. *Phytopathology*, **95**, 640–647.

56. Dioh,W., Tharreau,D., Notteghem,J.L., Orbach,M.J. and Lebrun,M.-H. (2000) Mapping of avirulence genes in the rice blast fungus, *Magnaporthe grisea*, with RFLP and RAPD markers. *Mol. Plant-Microbe Int.*, **13**, 217–227.

57. Orbach,M.J., Farrall,L., Sweigard,J.A., Chumley,F.G. and Valent,B. (2000) A telomeric avirulence gene determines efficacy for the rice blast resistance gene *Pi-ta*. *Plant Cell*, **12**, 2019–2032.

58. Farman,M.L. and Kim,Y.-S. (2005) Telomere hypervariability in *Magnaporthe oryzae*. *Mol. Plant Pathol.*, **6**, 287–298.

59. Farman,M.L. and Leong,S.A. (1995) Genetic and physical mapping of telomeres in the rice blast fungus, *Magnaporthe grisea. Genetics*, **140**, 479–492.

60. Carlson,M., Celenza,J.L. and Eng,F.J. (1985) Evolution of the dispersed *SUC* gene family of *Saccharomyces* by rearrangements of chromosome telomeres. *Mol. Cell. Biol.*, **5**, 2894–2902.

61. Schaffzin,J.K., Sunkin,S.M. and Stringer,J.R. (1999) A new family of *Pneumocystis carinii* genes related to those encoding the major surface glycoprotein. *Curr. Genet.*, **35**, 134–143.

62. Li,W., Rehmeyer,C.J., Staben,C. and Farman,M.L. (2005) TruMatch--a BLAST post-processor that identifies and reports bona fide sequence matches to a genome assembly. *Bioinformatics*, **21**, 2097–2098.

63. Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.

64. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.

65. Gordon,D., Abajian,C. and Green,P. (1998) Consed: a graphical tool for sequence finishing. *Genome Res.*, **8**, 195–202.

66. Dean,R.A., Talbot,N.J., Ebbole,D.J., Farman,M.L., Mitchell,T.K., Orbach,M.J., Thon,M., Kulkarni,R., Xu,J.R., Pan,H. *et al.* (2005) The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature*, **434**, 980–986.

67. Salamov,A.A. and Solovyev,V.V. (2000) *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res.*, **10**, 516–522.

68. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **213**, 403–410.

69. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, 138–141.

70. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, California, pp. 28–36.

71. Bendtsen,J.D., Nielsen,H., von Heijne,G. and Brunak,S. (2004) Improved detection of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.

72. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.

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

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

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

76. Kang,S., Lebrun,M.H., Farrall,L. and 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.

77. Nakayashiki,H., Matsuo,H., Chuma,I., Ikeda,K., Betsuyaku,S., Kusaba,M., Tosa,Y. and 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.

78. Gao,W., Khang,C.H., Park,S.-Y., Lee,Y.-H. and Kang,S.K. (2002) Evolution and organization of a highly dynamic, subtelomeric helicase gene family in the rice blast fungus *Magnaporthe grisea*. *Genetics*, **162**, 103–112.

79. Bok,J.W. and Keller,N.P. (2004) LaeA, a regulator of secondary metabolism in *Aspergillus* spp. *Euk. Cell*, **3**, 527–535.

80. Tudzynski,B. and Holter,K. (1998) Gibberellin biosynthetic pathway in *Gibberella fujikuroi*: evidence for a gene cluster. *Fungal Genet. Biol.*, **25**, 157–170.

81. McKnight,T.D., Riha,K. and Shippen,D.E. (2002) Telomeres, telomerase, and stability of the plant genome. *Plant Mol. Biol.*, **48**, 331–337.

82. Bhattacharyya,A. and Blackburn,E.H. (1997) *Aspergillus nidulans* maintains short telomeres throughout development. *Nucleic Acids Res.*, **25**, 1426–1431.

83. Regad,F., Lebas,M. and Lescure,B. (1994) Interstitial telomeric repeats within the *Arabidopsis thaliana* genome. *J. Mol. Biol.*, **239**, 163–169.

84. Nitta,N., Farman,M.L. and 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.

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

86. Gonzalez,I.L. and Sylvester,J.E. (1997) Beyond ribosomal DNA: on towards the telomere. *Chromosoma*, **105**, 431–437.

87. Penton,E.H. and Crease,T.J. (2004) Evolution of the transposable element *Pokey* in the ribosomal DNA of species in the subgenus *Daphnia* (Crustacea: Cladocera). *Mol. Biol. Evol.*, **21**, 1727–1739.

88. Perez-Gonzalez,C.E. and Eickbush,T.H. (2001) Dynamics of R1 and R2 elements in the rDNA locus of *Drosophila simulans*. *Genetics*, **158**, 1557–1567.

89. Butler,D.K. (1992) Ribosomal DNA is a site of chromosome breakage in aneuploid strains of *Neurospora. Genetics*, **131**, 581–592.

90. Gilson,P.R. and McFadden,G.I. (1996) The miniaturized nuclear genome of a eukaryotic endosymbiont contains genes that overlap, genes that are cotranscribed and smallest known spliceosomal introns. *Proc. Natl Acad. Sci. USA*, **93**, 7737–7742.

91. Zauner,S., Fraunholz,M., Wastl,J., Penny,S., Beaton,M., Cavalier-Smith,T., Maier,U.-G. and Douglas,S. (2000) Chloroplast protein and centrosomal genes, a tRNA intron, and odd telomeres in

an unusually compact eukaryotic genome, the cryptomonad nucleomorph. *Proc. Natl Acad. Sci. USA*, **97**, 200–205.

92. Adam,R.D., Nash,T.E. and Wellems,T.E. (1991) Telomeric location of *Giardia* rDNA genes. *Mol. Cell. Biol.*, **11**, 3326–3330.

93. Peyret,P., Katinka,M.D., Duprat,S., Duffleux,S., Barbe,V., Barbazanges,M., Weissenbach,J., Saurin,W. and Vivarès,C.P. (2001) Sequence and analysis of chromosome I of the amitochondriate intracellular parasite *Encephalitozoon cuniculi* (Microspora). *Genome Res.*, **11**, 198–207.

94. Wicky,C., Villeneuve,A.M., Lauper,N., Codourey,L., Tobler,H. and Muller,F. (1996) Telomeric repeats (TTAGGC)ₙ are sufficient for chromosome capping function in *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA*, **93**, 8983–8988.

95. Copenhaver,G.P. and Pikaard,C.S. (1996) RFLP and physical mapping with an rDNA-specific endonuclease reveals that nucleolus organizer regions of *Arabidopsis thaliana* adjoin the telomeres on chromosomes 2 and 4. *Plant J.*, **9**, 259–272.

96. Pays,E. and Steinert,M. (1988) Control of antigen gene expression in African trypanosomes. *Annu. Rev. Genet.*, **22**, 107–126.

97. Corcoran,L.M., Thompson,J.K., Walliker,D. and Kemp,D.J. (1988) Homologous recombination within subtelomeric repeat sequences generates chromosome size polymorphisms in *P. falciparum*. *Cell*, **53**, 807–813.

98. Dore,E., Pace,T., Ponzi,M., Picci,L. and Frontali,C. (1990) Organization of subtelomeric repeats in *Plasmodium berghei*. *Mol. Cell. Biol.*, **10**, 2423–2427.

99. Pace,T., Ponzi,M., Dore,E., Janse,C., Mons,B. and Frontali,C. (1990) Long insertions within telomeres contribute to chromosome size polymorphism in *Plasmodium berghei*. *Mol. Cell. Biol.*, **10**, 6759–6764.

100. Broun,P., Ganal,M.W. and Tanksley,S. (1992) Telomeric arrays display high levels of heritable polymorphism among closely related plant varieties. *Proc. Natl Acad. Sci. USA*, **89**, 1354–1357.

101. Royle,N.J. (1995) The proterminal regions and telomeres of human chromosomes. *Adv. Genet.*, **32**, 273–315.

102. Chan,C.S. and Tye,B.K. (1983) Organization of DNA sequences and replication origins at yeast telomeres. *Cell*, **33**, 563–573.

103. Liti,G., Peruffo,A., James,S.A., Roberts,I.N. and Louis,E.J. (2005) Inferences of evolutionary relationships from a population survey of LTR-retrotransposons and telomeric-associated sequences in the *Saccharomyces sensu stricto* complex. *Yeast*, **22**, 177–192.

104. Louis,E.J. and Haber,J.E. (1992) The structure and evolution of subtelomeric Y′ repeats in *Saccharomyces cerevisiae*. *Genetics*, **131**, 559–574.

105. Louis,E.J., Naumova,E.S., Lee,A., Naumov,G. and Haber,J.E. (1994) The chromosome end in yeast: its mosaic nature and influence on recombinational dynamics. *Genetics*, **136**, 789–802.

106. Mandell,J.G., Goodrich,K.J., Bahler,J. and Cech,T.R. (2004) Expression of a RecQ helicase homolog affects progression through crisis in fission yeast lacking telomerase. *J. Biol. Chem.*, **280**, 5249–5257.

107. Sanchez-Alonso,P. and Guzman,P. (1998) Organization of chromosome ends in *Ustilago maydis*. RecQ-like helicase motifs at telomeric regions.. *Genetics*, **148**, 1043–1054.

108. Inglis,P.W., Rigden,D.J., Mello,L.V., Louis,E.J. and Valadares_Inglis,M.C. (2005) Monomorphic subtelomeric DNA in the filamentous fungus, *Metarhizium anisopliae*, contains a RecQ helicase-like gene. *Mol. Genet. Genomics*, **274**, 79–90.

109. Bachrati,C.Z. and Hickson,I.D. (2003) RecQ helicases: suppressors of tumorigenesis and premature aging. *Biochem J.*, **374**, 577–606.

110. Ellis,N.A., Groden,J., Ye,T.Z., Straughen,J., Lennon,D.J., Ciocci,S., Proytcheva,M. and German,J. (1995) The Bloom's syndrome gene product is homologous to RecQ helicases. *Cell*, **83**, 655–666.

111. van Brabant,A.J., Stan,R. and Ellis,N.A. (2000) DNA helicases, genomic instability, and human genetic disease. *Annu. Rev. Genomics Hum. Genet.*, **1**, 409–459.

112. Yu,C.E., Oshima,J., Fu,Y.H., Wijsman,E.M., Hisama,F., Alisch,R., Matthews,S., Nakura,J., Miki,T., Ouais,S. *et al.* (1996) Positional cloning of the Werner's syndrome gene. *Science*, **272**, 258–262.

113. Pryde,F.E. and Louis,E.J. (1999) Limitations of silencing at native yeast telomeres. *EMBO J.*, **18**, 2538–2550.

114. Berriman,M., Ghedin,E., Hertz-Fowler,C., Blandin,G., Renauld,H., Bartholomeu,D.C., Lennard,N.J., Caler,E., Hamlin,N.E., Haas,B. *et al.* (2005) The genome of the African trypanosome *Trypanosoma brucei*. *Science*, **309**, 416–422.

115. Hernandez-Rivas,R., Mattei,D., Sterkers,Y., Peterson,D.S., Wellems,T.E. and Scherf,A. (1997) Expressed var genes are found in *Plasmodium falciparum* subtelomeric regions. *Mol. Cell. Biol.*, **17**, 604–611.

116. Brown,D.W., Yu,J.-H., Kelkar,H., Fernandes,M., Nesbitt,T.C., Keller,N.P., Adams,T.H. and Leonard,T.J. (1996) Twenty-five regulated transcripts define the sterigmatocystin gene cluster in *Aspergillus nidulans*. *Proc. Natl Acad. Sci. USA*, **93**, 1418–1422.

117. Gardiner,D.M., Ciozijnsen,A.J., Wilson,L.M., Pedras,M.S.C. and Howlett,B.J. (2004) The sirodesmin biosynthetic gene cluster of the plant pathogenic fungus *Leptosphaeria maculans*. *Mol. Microbiol.*, **53**, 1307–1318.

118. Pedley,K.F. and Walton,J.D. (2001) Regulation of cyclic peptide biosynthesis in a plant pathogenic fungus by a novel transcription factor. *Proc. Natl Acad. Sci. USA*, **98**, 14174–14179.

119. Proctor,R.H., Brown,D.W., Plattner,R.D. and Desjardins,A.E. (2003) Co-expression of 15 contiguous genes delineates a fumonisin biosynthetic gene cluster in *Gibberella moniliformi*s. *Fungal Genet. Biol.*, **38**, 237–249.

120. Yu,J.H. and Keller,N.P. (2005) Regulation of secondary metabolism in filamentous fungi. *Annu. Rev. Phytopathol.*, **43**, 437–458.

121. Ransom,R.F. and Walton,J. (1997) Histone hyperacetylation in maize in response to treatment with HC-toxin or infection by the filamentous fungus *Cochliobolus carbonum*. *Plant Phys.*, **115**, 1021–1027.

122. Wolpert,T.J., Dunkle,L.D. and Ciuffetti,L.M. (2002) Host-selective toxins and avirulence determinants: what's in a name? *Annu. Rev. Phytopathol.*, **40**, 251–285.

123. Tudzynski,B. and Holter,K. (1998) Gibberellin biosynthetic pathway in *Gibberella fujikuroi*: evidence for a gene cluster. *Fungal Genet. Biol.*, **25**, 157–170.

124. Panaccione,D.G., Scott-Craig,J.S., Pocard,J.A. and Walton,J.D. (1992) A cyclic peptide synthetase gene required for pathogenicity of the fungus *Cochliobolus carbonum* on maize. *Proc. Natl Acad. Sci. USA*, **89**, 6590–6594.

125. Tanaka,A., Shiotani,H., Yamamoto,M. and Tsuge,T. (1999) Insertional mutagenesis and cloning of the genes required for biosynthesis of the host-specific AK-toxin of *Alternaria alternata*. *Mol. Plant-Microbe Int.*, **12**, 691–702.

126. Yang,G., Rose,M.S., Turgeon,B.G. and Yoder,O.C. (1996) A polyketide synthase is required for fungal virulence and production of the polyketide T-toxin. *Plant Cell*, **8**, 2139–2150.

127. Fajkus,J., Kovarik,A., Kravolics,R. and Bezdek,M. (1995) Organization of telomeric and subtelomeric chromatin in the higher plant *Nicotiana tabacum*. *Mol. Gen. Genet.*, **247**, 633–638.

128. Hecht,A., Strahl-Bolsinger,S. and Grunstein,M. (1996) Spreading of transcriptional repressor SIR3 from telomeric heterochromatin. *Nature*, **383**, 92–96.

129. Wallrath,L.L. and Elgin,S.C. (1995) Position effect variegation in *Drosophila* is associated with altered chromatin structure. *Genes Dev.*, **9**, 1263–1277.

130. Zou,S., Ke,N., Kim,J.M. and Voytas,D.F. (1996) The *Saccharomyces* retrotransposon Ty5 integrated preferentially into regions of silent chromatin at the telomeres and mating loci. *Genes Dev.*, **10**, 634–645.

131. Arkhipova,I.R. and Morrison,H.G. (2001) Three retrotransposon families in the genome of *Giardia lamblia*: two telomeric, one dead. *Proc. Natl Acad. Sci. USA*, **98**, 14497–14502.

132. Noutoshi,Y., Arai,R., Fujie,M. and Yamada,T. (1998) Structure of the Chlorella Zepp retrotransposon: nested Zepp clusters in the genome. *Mol. Gen. Genet.*, **259**, 256–263.

133. Biessmann,H., Mason,J.M., Ferry,K., d'Hulst,M., Valgeirsdottir,K., Traverse,K.L. and Pardue,M.L. (1990) Addition of telomere-associated HeT DNA sequences 'heals' broken chromosome ends in *Drosophila*. *Cell*, **61**, 663–673.

134. Levis,R.W., Ganesan,R., Houtchens,K., Tolar,L.A. and Sheen,F.-M. (1993) Transposons in place of telomeric repeats at a *Drosophila* telomere. *Cell*, **75**, 1083–1093.