2022

Magnaporthe Oryzae Telomeric Retrotransposon (moter) Relics Further Highlight Telomere Dynamics In A Rapidly Evolving Fungal Pathogen

Jane Ellen Dostart
Eastern Kentucky University

MAGNAPORTHE ORYZAE TELOMERIC RETROTRANSPOSONS (MOTER) RELICS FURTHER HIGHLIGHT TELOMERE DYNAMICS IN A RAPIDLY EVOLVING FUNGAL PATHOGEN

BY

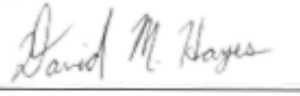
JANE E DOSTART

THESIS APPROVED:

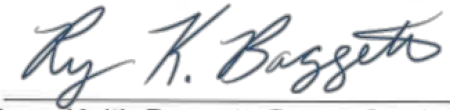
Dr. Patrick J. Calie, Chair, Advisory Committee

Dr. Mostafa Rahnama, Co-chair, Advisory Committee

Dr. Stephen Richter, Member, Advisory Committee



Dr. David Hayes, Member, Advisory Committee



Dr. Ryan Keith Baggett, Dean, Graduate School

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a Master of Science degree at Eastern Kentucky University, I agree that the Library shall make it available to borrowers under rules of the Library. Brief quotations from this document are allowable without special permission, provided that accurate acknowledgements of the source are made. Permission for extensive quotation from or reproduction of this document may be granted by my major professor. In [his/her] absence, by the Head of Interlibrary Services when, in the opinion of either, the proposed use of the material is for scholarly purposes. Any copying or use of the material in this document for financial gain shall not be allowed without my written permission.

Signature:

Date: 10/28/2022

MAGNAPORTHE ORYZAE TELOMERIC RETROTRANSPOSON (MOTER) RELICS FURTHER HIGHLIGHT TELOMERE DYNAMICS IN A RAPIDLY EVOLVING FUNGAL PATHOGEN

BY

JANE E DOSTART

Submitted to the Faculty of the Graduate School of
Eastern Kentucky University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

2022

ACKNOWLEDGEMENTS

I would like to firstly thank my two advisors, Dr. Patrick Calie and Dr. Mostafa Rahnama for their continued support and guidance throughout the work that went into this thesis. They agreed to many a weekend or evening call/visit to discuss science, writing, and my general well-being for which I am profoundly grateful. I would also like to thank Dr. Mark Farman for allowing me to participate in this project and for his wonderful collaboration between the University of Kentucky and Eastern Kentucky University. I thank my committee members, Dr. Stephen Richter and Dr. David Hayes for their review and comments as well as Dr. Melanie Link-Perez for her initial input and insightful comments for my thesis proposal. I would also like to thank Dr. Lindsay Cormier for pulling me into this wonderful opportunity. For that, I am eternally grateful. I want to thank Dr. Marcia Pierce for all the pep talks and humor that helped me through some difficult times as well as the opportunity to learn and grow on the other side of the classroom in BIO 111 and behind the scenes in the always entertaining microkitchen. I also thank all of those in the Biological Organization for Graduate Students (BOGS) including Reni Gull, Jeni Knight, Aaron Devine, Emily Jones, Ted Brancheau, and Hope Dull as well as all the students in the cohorts before and after mine. I want to thank my family for being incredibly supportive and interested in my research. And lastly, I would like to thank my partner Nate Hooven for all his love and support.

ABSTRACT

The telomeres of *M. oryzae* can experience unique rearrangements within a single generation. These rearrangements are frequently attributed to the presence of two retrotransposons (MoTeR 1 and MoTeR 2) that are endemic to the telomeres. These rearrangements can leave footprints within the interior of the genome that document previous telomere breakage events. The purpose of this study was to examine MoTeR relics within 10 genomes of strains of *M. oryzae* to document telomere instability among strain lineages, as well as to uncover MoTeR relic distribution, shared evolutionary history, and associated genome rearrangements. A series of local BLASTn and grep searches in the Unix Command Line were used to find and describe MoTeR relics. We found that MoTeR relics were found, on average, within 250 kb of the telomere and were often flanked by duplicate sequences that also mapped closely to telomeres. Our initial hypothesis for this study that MoTeR relics would largely represent unique strain or lineage-specific rearrangements was neither unequivocally supported nor entirely refuted. Intergenomic comparisons of MoTeR relics revealed several regions of shared synteny among distantly related strains that points towards their existence within ancestral strains, as well as seemingly strain- and lineage-specific relics that may be the result of more recent rearrangements. MoTeR relics serve as markers for investigating telomere dynamics in *M. oryzae* and further study might elucidate whether telomere instability within M. oryzae might play an adaptive potential in being able to quickly evolve and spread into new host plants as well as overcome resistance in others.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF SUPPLEMENTARY TABLES

LIST OF SUPPLEMENTARY FIGURES

INTRODUCTION

Agricultural importance - blast disease

Magnaporthe oryzae is a global fungal pathogen commonly referred to as the blast fungus, frequently used as a model organism for studying plant pathogen and host interactions. Genetic strains of *M. oryzae* form pathotypes specific to a wide variety of cereal grasses including, but not limited to, perennial ryegrass (*Lolium perenne*), rice (*Oryza sativa*), wheat (*Triticum aestivum*), finger millet (*Eleusine corocana*), oat (*Avena sativa*) and barley (*Hordeum vulgare*). Each fungal strain is typically restricted to one grass species or Genus in the Family *Poaceae*. Over 50 different species of grasses are hosts to strains of this fungus (Ou 1985). One host plant, rice, receives significant research funding as it serves as a staple food item to ~50% of the global population (Skamnioti and Gurr 2009). In 2009 it was reported that rice blast disease destroyed 10 – 30% of the world's annual yield, 10% of which would feed approximately 60 million people (Skamnioti and Gurr).

The earliest record of rice infection by *M. oryzae* was documented in China (1637), and subsequently documented in Japan (1704), Italy (1828), the United States (1906), and India (1913), and has been reported in more arid regions such as Iraq (Ou 1985). *M. oryzae* displays a wide range of tissue affinity with the ability to infect every above ground organ of the plant (e.g., leaves, panicles, stems, and nodes) during any stage of plant development (Ou 1985; Talbot and Wilson 2009). An easily recognizable symptom of infection is in the form of lesions on grass leaves (blades) that present yellow, and more interiorly

brown, margins and either a brown or gray section in the interior during the later stages of infection (Figure 1b, c). Infected seeds have helped facilitate the spread of this fungus to 85 countries and six continents (Kato 2001).

M. oryzae poses a threat to global wheat production as it spread rapidly through the Brazilian wheat crop and into adjacent South American countries in 1985 gaining the name “wheat blast” (Inoue et al. 2017). A wheat-infecting isolate was discovered by investigators at the University of Kentucky in the United States in 2011. The isolate did not come from South America but is thought to have instead originated from a *Lolium* pathotype (Farman et al. 2017), a pathotype being a variant of *M. oryzae* that infects a specific group of hosts. *Lolium* pathotypes infect species like perennial ryegrass while other pathotypes, such as *Oryza* and *Avena,* infect rice and oat species respectively (Inoue et al. 2017). The isolate in Kentucky was less virulent than the Brazilian strain (Farman et al. 2017). More recently an outbreak of wheat blast originating from South America had impacts in Bangladesh resulting in a range of 10-100% loss of yield in wheat crop (Inoue et al. 2017).

The genus Magnaporthe

M. oryzae is a filamentous ascomycete, the body of the organism consisting of filamentous hyphae and the sexual spores contained in a sac termed an “ascus”. It phylogenetically groups within the most species abundant phylum in the fungal kingdom. *M. oryzae* belongs to the Genus *Magnaporthe* nested in the Family *Magnaporthaceae*, Order *Magnaporthales*, Class *Sordariomycetes*, Subphylum *Pezizomycota,* and Phylum *Ascomycota* (Figure

1). Numerous species of ascomycetes are human pathogens, such as *Aspergillus fumigatus* which can grow in the lungs of immunocompromised individuals (Latgé 1999), *Candida albicans*, which can cause infections in the urinary tract (Sudbery 2011), and species within the genus *Trichophyton* which cause several skin infections like ringworm and athlete's foot (Gnat *et. al.* 2020). Like *M. oryzae*, several other plant pathogens belong to the Ascomycota including several rusts, powdery mildews, chestnut blight, and ergots (Berbee 2001).

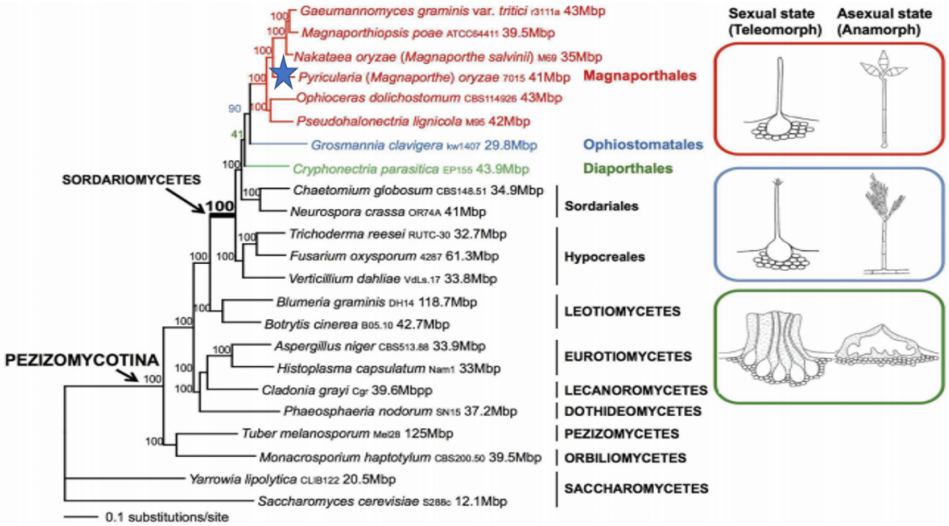


Figure 1 Maximum likelihood tree built from transcriptome data of 21 species from 7 Classes within the Subphylum Pezizomycotina against the outgroup Saccharomycetes. The species Magnaporthe oryzae within the Order Magnaporthales is labeled with a star. Each species is followed by its strain number and genome size. The teleomorph and anamorph structures for the Orders Magnaporthales, Ophiostomatales, and Diaporthales within the Class Sordariomycetes are pictured on the right. Reproduced from Luo et al. 2015.

The Fungi

Fungi possess life history strategies that allow for frequent transitions to parasitic lifestyles. Fungi are natural chemists, as their largely immobile state has

added evolutionary selective pressure on their ability to acquire nutrients, defend themselves, and colonize new territories. Although, as demonstrated earlier, several fungi are harmful to humans, many other species use biochemical strategies that are exceedingly important to the global economy and human health. The classic example of this can be seen in the antibiotic penicillin. This chemical is produced by the common bread mold, *Penicillium rubens* (Fleming 1941). Humans have also utilized *yeasts* (a noun that broadly encompasses numerous species of single-celled fungi) for hundreds of years to make breads, wines, and beers; some species are beautiful demonstrations of phylogeography and have formed clades that correspond to the specific regions of the world such as yeasts used to make sake in Asia and others whose ancestry traces the history of beer making in Europe and subsequent colonialization later of the New World (Gallone et al. 2016).

The Kingdom Fungi shares a common ancestor with animals and is estimated to have diverged from the animal lineage about 900 to 1500 million years ago (Figure 2; Gan et al. 2021). Two commonly recognized phyla within this kingdom are the Basidiomycetes and the Ascomycetes. The Ascomycota are the most species rich phylum of the fungal kingdom, yet members of the fungal kingdom that are more readily recognized are the mushrooms (Basidiomycota). The phyla Ascomycota and Basidiomycota comprise the Subkingdom Dikarya as they both form dikaryotic hyphae during sexual reproduction (Hibbett et al. 2007). Ascomycetes possess a sac-like reproductive structure called the ascus which houses ascospores – the asci are grouped together in a larger structure called

the perithecium. Basidiomycetes are characterized by their reproductive structure known as the basidium, which is comprised of a club-like structure topped with four sexually reproduced basidiospores. During sexual reproduction for members of dikarya the normally haploid genome of one mating type will fuse with the haploid genome of another mating type during a process called karyogamy. The fused nuclei will undergo several meiotic divisions as a diploid cell wherein the chromosomes from each mating type may undergo recombination. After several stages of replication and division of the diploid nucleus, the cells will divide into haploid daughter cells and form ascospores which will be released and propagated as a new fungus.

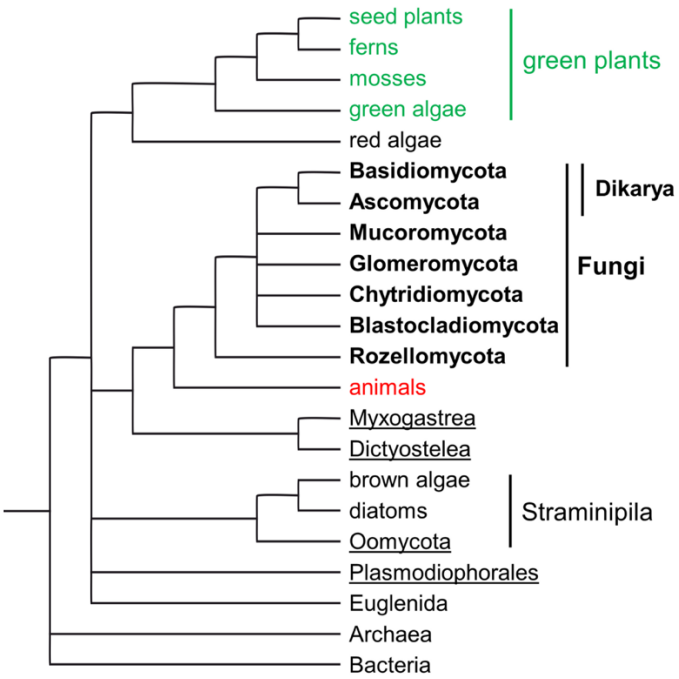


Figure 2 Simplified diagram of fungi (names in bold) by Meike Piepenbring 2021. Notice that fungi are more closely related to animals, sharing a more recent common ancestor than with plants.

The Order *Magnaporthales* can reproduce through mitotic divisions in their anamorphs through specialized structures known as conidiophores that extend from the vegetative, septate hyphae – these form haploid conidiospores (Brock, Smith, and Madigan 1984) (Figure 3a). The conidiospores of the species *M. oryzae* are composed of three cells and have a tear-drop shape (Talbot and Wilson 2009) (Figure 3a).

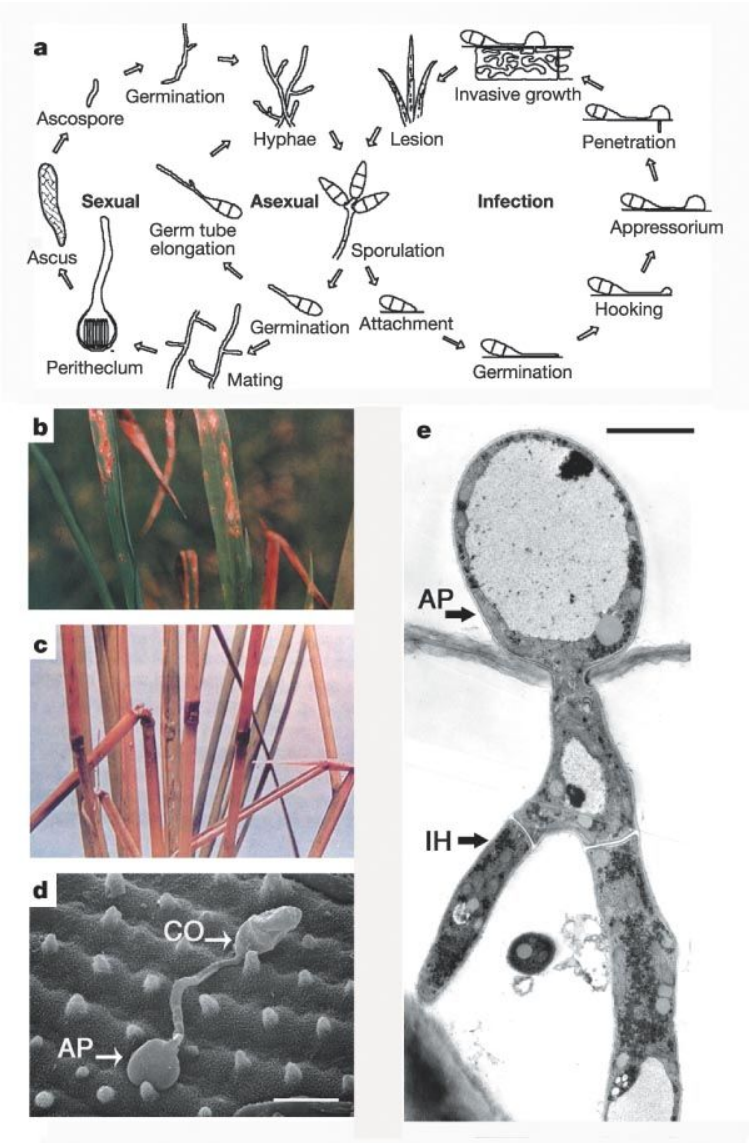


Figure 3 Life cycle of Magnaporthe oryzae. **a**, Two mating types of different strains sexually reproduce and develop ascospores within a perithecium. Conidiospores produced asexually

attach to host plant tissue and germinate to produce a structure called the appressorium which utilizes turgor pressure (~8MPa) to puncture plant tissue and grow invasive hyphae that produce lesions from which new spores will be released. **b**, Rice (Oryzae sativa) leaf lesions. **c**, Rice stem node infection. **d**, Image from a scanning electron microscope of a conidium (CO) and developing appressorium (AP) on the surface of a rice leaf. Scale bar, 10 μm. **e**, Image from a transmission electron microscope illustrating an appressorium on the surface of a rice leaf and invasive hyphae (IH) growing within the leaf. Scale bar, 5 μm. Reproduced from Dean et al. 2005.

One key feature that has facilitated fungal evolutionary success is their possession of a haploid genome. This might not seem intuitive as having less genetic material to manipulate through recombination or mutation might seem a disadvantage, however, the nature of only having one copy of each chromosome means that mutations to genes in their genome will be expressed as they will not be masked by another gene copy. This serves as a form of bet-hedging as a single fungus might produce millions of meiotic spores all with different genomes, e.g., SNPs, and those that persist should exhibit desirable traits that could increase the fitness of the next generation, while those with deleterious traits will be purged from the population (Orr and Otto 1994).

Magnaporthe as a plant pathogen

In the case of *M. oryzae*, its spores could land on a vegetative structure (e.g., leaf) on one of the >50 species of grasses it infects. When its spores land and adhere to a suitable host plant, the spore then germinates and grows to form an infective structure called an appressorium that utilizes turgor pressure to penetrate the vegetative tissue of its host (Talbot and Wilson 2009) (Figure 3d, e). During this initial assault of the host tissue, *M. oryzae* hyphae release a cocktail of effector proteins that aid in host infection (Talbot and Wilson 2009).

M. oryzae is a facultative parasite: it does not require its host plant to carry out certain life stages, but it does depend on its host plants for survival. This interaction has applied selective pressure for a suite of defense mechanisms generated by the host plant that help to detect and defend against invasion by *M. oryzae*. Much of the pathogenic success in *M. oryzae* is due to the rapid evolution of its effector proteins. Several mechanisms exist in the genome that facilitate the mutation and evolution of effector proteins. The mutation or loss of function of fungal *AVR* genes allows for increased infectivity and loss of recognition by the host plant's resistance (r) protein receptors (Sánchez-Vallet *et al.* 2018). Avr proteins often map to structurally unstable regions of the fungal genome, i.e., proximal to telomeres, adjacent to transposable elements (TEs), or within regions of repeat sequence (Sánchez-Vallet *et al.* 2018). These locations in the genome make *AVR* genes more susceptible to high mutation rates and sequence instability (Sánchez-Vallet *et al.* 2018). Most resistant plant cultivars do not maintain resistance as *M. oryzae* strains can quickly mutate or lose their *AVR* genes that would otherwise bind to the plant's R protein receptors and signal fungal infection (Dangl, Horvath, and Staskawicz 2013). Modifications to, or loss of, *AVR* genes has been inferred as a mechanism for "host jumping" as seen with the appearance of wheat blast where strains specific to certain host plants may acquire the ability to infect a different host plant (Inoue *et al*. 2017). Likewise, rotation of crops lacking R genes specific to *M*. *oryzae* pathotypes and populations could allow for devastating outbreaks as is thought to be the case for

new wheat variants that were planted in South America in 1985 that lacked the Rwt3 resistance gene (Inoue et al. 2017).

The Magnaporthe genome

M. oryzae strains have dynamic and variable genomes in size and architecture (e.g., chromosomal structure). The haploid genome contains ~41 Mbp organized into 7 core chromosomes; however, several strains vary in the precise size of their core chromosomes, and some have additional supernumerary chromosomes, or minichromosomes (Dean *et al.* 2005; Luo *et al.* 2015; Peng *et al*. 2019). The linear chromosomes are flanked in their terminal regions by telomeres that consist of short, repeat sequences ($^{5'}$CCCTAA$^{3'}$) bound by a shelterin protein complex to prevent chromosome degradation and potential loss of DNA between cycles of replication (De Lange, Lundblad and Blackburn 2006).

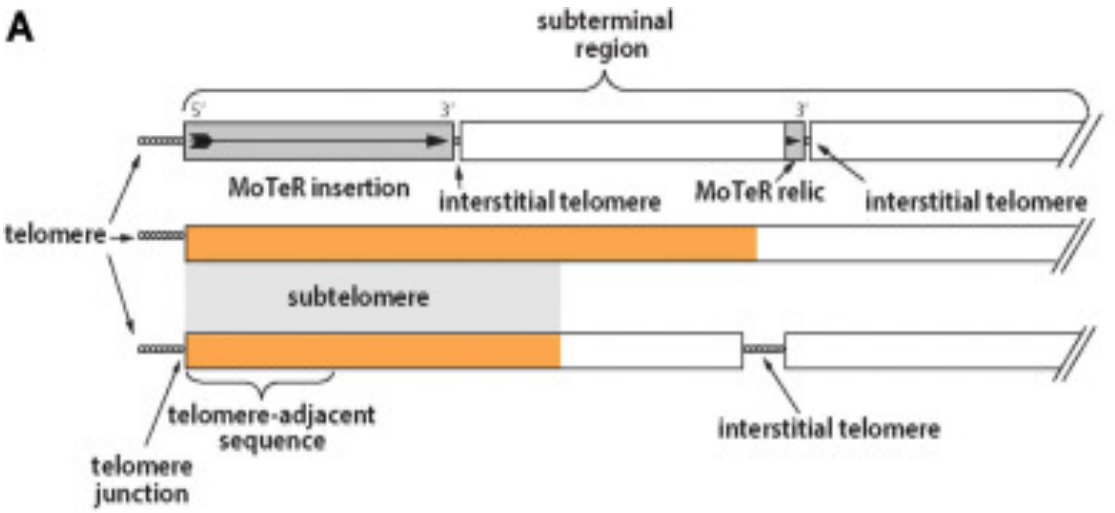


Figure 4 General layout of chromosome end architecture and terminology in M. oryzae. The telomere is illustrated as a string of circles representing its composition of telomeric ($^{5'}$CCCTAA$^{3'}$) repeats. The subterminal region is a general term used to describe domains near chromosome

ends. Within the subterminal region there can be subtelomeres which are duplicate sequences shared at other chromosome ends. Synteny between the two subtelomeres is represented by the gray box connecting the two areas in orange. Interstitial telomeres are telomere repeat sequences that are not continuous and surrounded by non-telomere repeat sequences. In the first row a full MoTeR element is inserted within the telomere with its 5' end facing towards the end of the chromosome, and its 3' end towards the chromosome interior. Further into the subterminal region is a truncated MoTeR element whose 3' end has been preserved. This truncated MoTeR is outside of the telomere of this chromosome, so it is considered a MoTeR relic. Lastly, a telomere junction indicates the boundary between the telomere and non-telomere sequence, whereas the telomere-adjacent sequence is found immediately next to the telomere junction (Adapted from Rahnama et al. 2021).

All background information in the introduction has led to the more refined arena of the thesis here-in that takes place in the telomeres and subtelomeres of *M. oryzae*. A study of the telomeres and subtelomeres within rice-infecting (*Oryza)* and perennial ryegrass-infecting (*Lolium*) strains revealed that telomere sequence-containing restriction fragments within perennial ryegrass-infecting strains were highly variable between progeny and parent strains compared to parents and progeny in rice-infecting strains (Starnes et al. 2012). Nucleotide sequence analysis revealed two mobile genetic elements (MGEs) residing within the telomeres of the perennial-ryegrass infecting strains that were inferred to be the causal agents of telomere instability within perennial ryegrass (Starnes *et al.* 2012). These MGEs were named *Magnaporthe oryzae* telomere retrotransposon elements (MoTeRs) 1 and 2 (Starnes *et al.* 2012).

Magnaporthe retrotransposons

Magnaporthe oryzae Telomere Retrotransposons (MoTeR) 1 & 2 are MGEs first discovered within telomeres in the *M. oryzae* genome of a strain infecting perennial ryegrass (prg) (Figure 5; Starnes *et al.* 2012). Telomeric sequences of prg infecting strains were investigated after southern blot analyses of telomere restriction fragments of strains infecting rice and prg revealed a high

degree of variability in telomere restriction fragment length (RFL) between parent and progeny in prg (Starnes *et al.* 2012). Telomeres and their internal sequences were cloned and sequenced to reveal two MGEs inserted within telomeres within prg strains (Starnes *et al.* 2012). Rice strains had little to no variation within RFL of parent and progeny and subsequent sequencing of rice strain parents and progeny did not display MGEs within the telomere (Starnes *et al.* 2012). The MGEs were determined to function as agents of genomic instability causing frequent genome rearrangements in prg strains as these frequent rearrangements were not observed in rice infecting strains (Starnes *et al.* 2012).

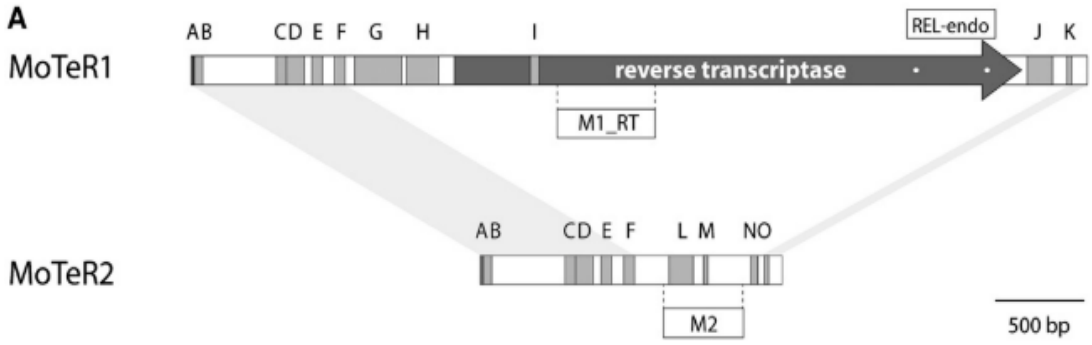


Figure 5 Schematic diagram of the M. oryzae MoTeR elements. MoTeR1 and MoTer2 are drawn to scale. Repeated non-coding sequences are indicated by the medium-grey boxes. The terminal shared sequences between MoTeR1 and MoTeR2 are indicated by light-grey bridging the two elements. The coding region for reverse transcriptase is embedded within the dark arrow also coding for a restriction-like endonuclease domain (REL-ENDO). The molecular probes for MoTeR1 and MoTeR2 are M1_RT and M2 respectively. Reproduced from Starnes et al. 2012.

Sequencing of the MoTeRs revealed that MoTeR1 is ~5 kb in length while MoTeR2 is ~1.7 kb in length (Starnes *et al.* 2012) (Figure 5). They both share an identical 860 bp sequence at their 5' ends and an identical 77 bp sequence at their 3' ends. Both MoTeRs are oriented with their 5' end proximal to the chromosome terminus as either a tandem array or as a solitary element (Starnes

et al. 2012). MoTeR1 contains an open reading-frame (ORF) that putatively encodes a protein with a reverse transcriptase domain that is 1,070 amino acids in length (Starnes *et al.* 2012). The reverse transcriptase domain contained statistically significant identity as determined by a BLASTx search (i.e., an acceptable e-value) to those found in retrotransposons in *Trypanosoma brucei gambiense* (Aksoy *et al.* 1990) and *Crithidia fasciculata* (Gabriel *et al.* 1990; Starnes *et al*. 2012). MoTeR2 contains an ORF that putatively encodes a protein of unknown function that is 280 amino acids in length (Starnes *et al.* 2012). MoTeR1 is a non-LTR retrotransposon while MoTeR2 is non-autonomous as it seemingly lacks the genes for transposition (Starnes e*t al.* 2012). The retrotransposons in *Trypanosoma brucei gambiense* (Aksoy *et al.* 1990) and *Crithidia fasciculata* (Gabriel *et al.* 1990) both contain restriction enzyme-like endonuclease domains (REL-ENDO) that insert only within splice leader sequence genes. The predicted reverse transcriptase in MoTeR1 has a putative REL-ENDO that is proposed to target telomere repeats ($^{5'}$CCCTAA$^{3'}$) for DNA strand cleavage (Starnes *et al*. 2012). The exposed strand of DNA is predicted to serve as a primer for reverse transcription as the 3' terminus of MoTeR1 contains telomere-like sequence that could complementarily anneal (Starnes *et al.* 2012) (Figure 6). MoTeRs insert within telomere repeats.

Figure 6 The theorized mechanism for MoTeR replication within M. oryzae. **A**. Potential cleavage site within the upper strand of the telomere repeat containing ($^{5'}$TTAGGG$^{3'}$) DNA. **B**. Top strand cleavage by the restriction enzyme-like endonuclease domain of reverse transcriptase and annealing of the MoTeR transcript 3' terminal sequence to the free upper strand of the DNA. **C**. Synthesis of complementary DNA (cDNA). The annealing of the MoTeR transcript 3' end to the upper strand of DNA will result in terminal duplications. The MoTeR transcript's 5' region could also anneal to a nick within the bottom strand of DNA resulting in terminal deletions. **D**. The 3' region of the cDNA anneals to the 3' region of the nicked bottom DNA strand followed by the synthesis of the second strand of MoTeR DNA. **E**, the nicks in the DNA strands are ligated resulting in complete MoTeR sequence insertion. Reproduced from Starnes et al. 2012.

The other interrogated fungal strains containing full-length MoTeR sequences were those infecting wheat and millet (Starnes e*t al.* 2012). Consistent with the proposed telomere insertional sites the full MoTeR sequences were present within telomere sequences (Starnes *et al*. 2012; Rahnama *et al.* 2020). Subsequent interrogation of *M. oryzae* genomes containing MoTeR sequences demonstrated that telomeres containing MoTeR sequences were not inherently unstable as first proposed by Starnes e*t al.* (2012) but variation in sequence length separating MoTeRs in an array allowed for

instability to occur (Rahnama e*t al.* 2020). Telomeres that contained MoTeR elements separated by interstitial telomere sequence repeats of 3 or more were found to be less stable than sequences that contained up to 2 repeats (Rahnama e*t al.* 2020). Further evidence of telomere instability associated with MoTeR elements was the presence of truncated MoTeR sequences located within the interior of the genome that were identified by their maintained 3' end followed by one or more telomere repeats ($^{5'}$…**CGCGAATTAAAA**$\underline{\text{CCCTAA}}_{(n)}{}^{3'}$); these internalized truncated MoTeRs were named MoTeR relics (Rahnama e*t al.* 2020). These relics were evidence of rearrangements in the telomere and served as indicators of potentially historical or more recent invasions of the telomere into the surrounding genome. Interestingly, several MoTeR relics were associated with sequence duplications likely formed through DNA-repair mechanisms following telomere breakage near MoTeRs (Rahnama e*t al.* 2020).

A BLASTn analysis of the fully assembled LpKY97 fungal genome (the strain infecting perennial ryegrass; Rahnama et al. 2020) revealed that of the 18 MoTeR relics ten have 5' flanking sequence duplications (56%), one is a duplication of a relic and its 5' flanking sequence (5%), and two are whole locus duplications (11%) in which the relic and both of its flanking sequences are duplicated elsewhere in the genome. The remaining five relics do not have any type of duplication (28%). The frequency of MoTeR relic 5' flanking sequence duplications can be explained by frequent chromosomal repair events that may occur at the 5' boundary of the MoTeR in the telomere where MoTeRs are present. Interstitial telomere repeats of 3 or more were associated with causing

telomere instability. They proposed that a double stranded break in a MoTeR array could lead to degradation of MoTeR DNA creating a truncated MoTeR (Rahnama *et al.* 2020). The naked end of the double strand could then be repaired in one way by the addition of internal DNA that becomes duplicated in the process. The MoTeR relics with no duplications associated with them were the next most common (28%) and were explained by possible translocation events or potential loss of duplicate sequences. Whole locus (11%) and relic + 5' flank duplications (5%) were the least abundant. The exact origins of these duplications and their length or sequence composition remains uncharacterized. This information could further explain how MoTeRs come to be truncated and moved out of the telomeres and which regions of the genome might be more prone to MoTeR repair. It is speculated that these duplications could serve an adaptive potential if they were to duplicate genes where-in changes to the sequence could allow for differences in protein function or merely allow them to escape recognition by the host plants that the fungus infects (Rahnama *et al.* 2020).

MoTeR relics are the footprints of genome rearrangements tied to their origin in the telomere. The purpose of this study was to explore the evolutionary history of MoTeR relics and how they arose in different strains of *M. oryzae* by examining the internalized MoTeR sequences of 10 fully assembled fungal genomes isolated from strains infecting the host grasses *Eleusine* (goosegrass), *Triticum* (wheat), *Lolium* (ryegrass), *Oryza* (rice), *Setaria* (foxtail), and *Stenotaphrum* (St. Augustine's grass) pathotypes.

METHODS

Genome Sequence Data

NCBI accession numbers and sources for each of the 10 fully assembled genome examined in this study are catalogued in Supplementary Table 1.

Local BLAST analyses

Local BLASTn (-e-value 1e-1, -task BLASTn-short, %identity >80%) searches of full MoTeR1 and MoTeR2 sequences (~5 kb and ~1.7 kb respectively) as queries against each genome were used to reveal the location, length, and orientation of MoTeR relics. In the initial BLASTn results aberrant hits of < 20 nt were filtered using the awk command and excluded from the analysis (Appendix 1). MoTeR sequences were considered relics if they were found outside of the terminal telomeric arrays (chromosomal ends) and contained a 3' terminal sequence of the MoTeR attached to the telomeric repeat/s ($^{5'}$CGCGAATTAAAA**CCCTAA**$^{3'}$; $^{5'}$**TTAGGG**TTTTAATTCGCG$^{3'}$). BLASTn hits that did not contain a 3' end were excluded from the analysis.

BLASTn searches (-e-value 1e-20) of each genome against each of the other 9 genomes were conducted to determine if MoTeR relics were shared/lost among strains. MoTeR relics were considered shared if the sequences flanking MoTeR relics were >90% identical for at least 500 bp on either side. If the sequence was broken up by a transposon insertion it was still considered to maintain synteny if the flanking sequence around the insertion had the same identity (>90%; Appendix 1).

UNIX Command Line – Grepping sequences

In addition to local BLASTn searches, the UNIX Command Line Interface was used to conduct grep searches of the 3' end sequences against the genome ($^{5'}$CGCGAATTAAAA**CCCTAA**$^{3'}$; $^{5'}$**TTAGGG**TTTTAATTCGCG$^{3'}$). The grep command can be used to search for exact matches of a query sequence given to it within a .fasta file. The results of grep searches were blasted (-task BLASTn-short -e-value 1e-1) against full MoTeR 1 and 2 sequences (~5 kb and 1.7 kb respectively) to support MoTeR relic identity (> 20 nt). New MoTeR relics found using grep were then blasted against the genome (e-value 1e-20 % identity = 100) to find their location and orientation within the genome (Appendix 1).

MoTeR Relic Flanking Duplicate Sequence Analysis

BLASTn (-e-value 1e-20) interrogations of each genome against itself were used to generate .gff files to search for duplicate sequences. Duplications adjacent to MoTeR relics were manually interrogated using the Integrative Genomics Viewer (IGV; Thorvaldsdottir et. al., 2013). Duplications were considered in the analysis if they were within 20 nt of the 3' or 5' boundaries of the MoTeR relic. Sequences that indicated a relic + 5' flanking sequence duplication were considered if the length of the duplication was > 500 nt. Flanking duplicate sequences in IGV were extended if adjacent hits mapped to the same region of the matching chromosome. In these cases, duplicate sequences were extended even if they were separated by short regions of unique sequence and/or transposable elements.

Integrative Genome Viewer

Chromosome ends of each strain's genome were manually examined using the Integrative Genomics Viewer (IGV; Thorvaldsdottir et. al., 2013). Chromosome termini of each strain were examined for the presence/absence (Y/N) of terminal MoTeR sequences and fully assembled telomere sequences. Pos_start represents the left-side of the chromosome while pos_end represents the right-side of the chromosome. The purpose of this information was to calculate the distance of individual MoTeR relics to the nearest terminal MoTeR or telomere sequence. If a MoTeR was present in the telomere, then the position of the last MoTeR leading into the subtelomere was reported. If there was no MoTeR present within the telomere, then the position recorded was the end of the telomere at the edge of the subtelomere. If the telomere was missing from the assembly, the position was defined as the first or last nucleotide in the chromosome. Chromosome ends with a single, or partial, telomere repeat ($^{5'}$**CCCTAA**$^{3'}$;$^{3'}$**TTAGGG**$^{5'}$) were not considered fully assembled telomeres.

Graphic visualization – Circos plots

MoTeR relics and adjacent duplicate flanking sequences were visualized using a custom R code (Appendix 2) applying the package Circos (Krzywinski et. al., 2009). In the generated figures MoTeR relic orientations are illustrated as triangles where the tip of the triangle corresponds to the relic's 3' end. Relics with 3' flanking sequence duplications are blue, those with relic + 5' sequence duplications are grey, those with no flanking sequence duplications are white, and relics with a combination of 3' flanking sequence duplications and relic + 5' end duplications are grey with a blue border. The color of links corresponds to

the chromosome in which the duplicate sequence is found flanking the MoTeR relic. Some relics have been shifted in the plot to aid in their visualization.

RESULTS

MoTeR Relic Mapping and Distributions

MoTeR relics are truncated MoTeR 1 and 2 elements that are found in the interior of the genome or exclusive of the telomeres. MoTeR relics investigated in this study were those that contained a conserved 3' sequence ($^{5'}$CGCGAATTAAAA$^{3'}$; $^{5'}$TTTTAACGCG$^{3'}$) flanked by one or more telomere repeats ($^{5'}$**CCCTAA**$^{3'}$; $^{5'}$**TTAGGG**$^{3'}$). The minimal length for relics found was 26 bp except for a relic in chromosome 3 of CD156 (15 bp) whose identity was supported by local BLASTn searches of MoTeR relics found in chromosome 3 in strains LpKY, FH, and B71 (Table S2). No full MoTeR 1 (5,034 bp) or MoTeR 2 (1,723 bp) sequences were found in the interior of any of the chromosomes or strains interrogated. The longest relic (4,277 bp) was found in chromosome 1 of strain CD156 which resided approximately 750 kb away from telomere 1 (Table S2). In general, MoTeR relic 3' sequences were flanked by one to two telomere repeats with LpKY having the shortest containing a relic with no telomere repeat in Chromosome 2, and Guy11 having perhaps the longest in Chr2, although its telomere repeats contain several point mutations (Table S2).

In the 10 assembled genomes a range from one to 17 MoTeR relics were detected in the seven core chromosomes (Table 1; Figure 7). The highest number of relics (17) was found in strain LpKY with the highest concentration on chromosome 3 (5 relics; Figure 7A). Strain U233 only contained one MoTeR relic

(347 bp) which is approximately 55 kb away from telomere 2 (end of chromosome 1) flanked by a single telomere repeat ($^{5'}$CCCTAA$^{3'}$, Figure 7F, Table S2). The strains with the highest occurrence of MoTeR relics following LpKY (17) were FH, CD156, and US71 containing 14, 15, and 13 relics respectively (Table 1; Figure 7B, C, J). The strains with the lowest concentrations of MoTeR relics following U233 were 70-15, Guy11, and Bm88324 containing 3, 4, and 7 relics respectively (Table 1; Figure 7G, E, I). Arcadia2 and B71 shared a median number of 11 MoTeR relics (Figure 7H, D).

LpKY, FH, CD156, and B71 appeared to share two populations of relics within chromosomes 3 and 6 that mapped approximately 2.6 Mb and 2.4 Mb to the nearest telomere or terminal MoTeR. Guy11 contained a MoTeR relic in chromosome 2 that mapped approximately 3.1 Mb from the nearest telomere or terminal MoTeR (TableS1). Typically, excluding the centralized relic populations in chromosomes 3 and 6 in strains LpKY, FH, CD156, and B71, MoTeR relics mapped within 252 kb of a terminal MoTeR or terminal telomeres. Including the distances for relics in the chromosomes of the examined strains gives a median distance for MoTeR relics of approximately 203 kb from the nearest terminal MoTeR or telomere illustrating that MoTeR relics appear to be more likely to be found near chromosome ends.

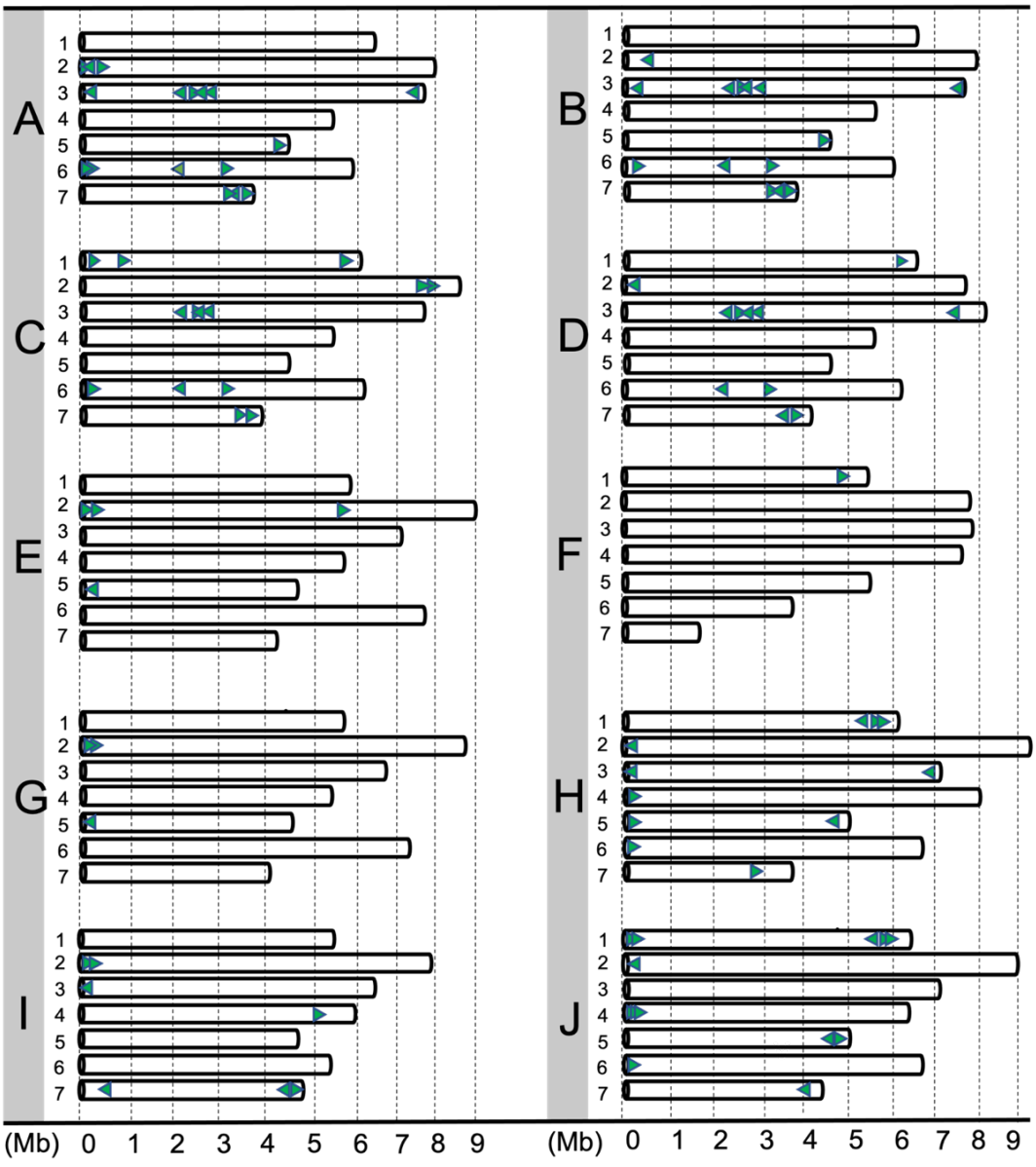


Figure 7 Map locations of 3' MoTeR relics in the 10 examined Magnaporthe oryzae strain genomes. **A** LpKY (Lolium; adapted from Rahnama et al. 2020), **B** FH (Lolium), **C** CD156 (Eleusine), **D** B71 (Triticum), **E** Guy11 (Oryza), **F** U233 (Stenotaphrum), **G** 70-15 (Oryza), **H** Arcadia (Setaria), **I** Bm88324 (Brachiaria), and **J** US71 (Setaria). MoTeR relics are represente as green triangles, the base representing the 5' terminus. The seven core chromosomes are plotted based on length and are not aligned to each other.

Table 1 Examined Magnaporthe oryzae strains. Total number of MoTeR relics identified per strain, telomeres containing full and partial MoTeR 1 and 2 sequences, and the ratio of fully assembled telomeres to unassembled telomeres. This table summarizes results from chromosomal end composition in Supplementary Table 2.

Genome	Pathotype/Host	Number of 3' MoTeR Relics	Number of Telomeres Containing MoTeR Sequence	Assembled/Unassembled Telomeres
LpKY97	*Lolium*	17+8*	13*	18/0
FH	*Lolium*	14	14*	15/1
U233	*Stenotaphrum*	1	4	10/4
US71	*Setaria*	13	0	7 /7
Arcadia2	*Setaria*	11	2	10/4
CD156	*Eleusine*	15	9	10/4
Guy11	*Oryza*	4	0	9/5
B71	*Triticum*	11	11	9/5
70-15	*Oryza*	3	0	11/3
Bm88324	*Brachiaria*	7	0	7/7

*Includes minichromosomes (LpKY: 2 minichromosomes, FH: 1 minichromosome), second number after + represent the number of relics on the mini chromosomes

Chromosomal End Composition

The chromosomal ends (telomeres) of each strain were examined to see if there was a correlation between the concentration of relics in each strain with the presence or absence of terminal MoTeRs. Telomeres of strains LpKY and FH were the most populated with MoTeR 1 and 2 sequences with 13/18 telomeres in strain LpKY97 containing MoTeR sequences (includes minichromosomes 1 & 2) and 14/16 telomeres in strain FH (includes minichromosome 1; Table 1). These strains contained the highest number of MoTeR relics (Table 1). Likewise, consistent with the findings of Starnes *et al.* (2012) telomeres in strains 70-15 and Guy11 did not contain MoTeR 1 or 2 sequences (Table 1) and these strains had some of the lowest occurrences of relics (Table 1). However, U233, the strain with the least amount of MoTeR relics (1) had MoTeR sequences present in four of its telomeres while US71, a strain containing 13 MoTeR relics, was found to not contain MoTeR sequences in any of its telomeres. Of the strains lacking MoTeRs within their telomeres (US71, Guy11, 70-15, and Bm88324), several of their telomeres were underrepresented and in the final nucleotide sequence data set were likely lost during sequencing or genome assembly (Table 1). Due to their sequence composition telomeres are recalcitrant to the DNA sequencing enzymology in short read (Illumina) sequencing used to generate genomic data as their sequences are extremely repetitive. Long read (MinIon) sequencing is able to ameliorate this by producing longer read lengths that can encompass kilobases of nucleotides facilitating more complete genome assemblies (Kim et al. 2021).

MoTeR Relic Comparisons

As summarized in figure 8 all relics in each chromosome are aligned to compare relic distribution and retention based on their flanking sequences. Although relic populations are variable among the ten examined strains the results showed patterns of relic retention that are seen in strains that share a most recent common ancestor (Figure S1). Those within the same lineage, such as strains 70-15 and Guy11 (the *Oryza* lineages) share synteny in all but one of their relics and US71 and Arcadia (the *Setaria* lineages) share eight relics. Arcadia possesses three unique relics and US71 contains five, all of which are found within approximately 540 kb or less from the nearest telomere. Four strains, LpKY, FH (the *Lolium* lineages), B71 (*Triticum*), and CD156 (*Eleusine*) share several relic populations in chromosomes 3, 6, 7, and 2 (Figure 8). Despite MoTeR relic synteny among the more closely related lineages, several strains contain a differential retention of relics across several lineages (Figure S1).

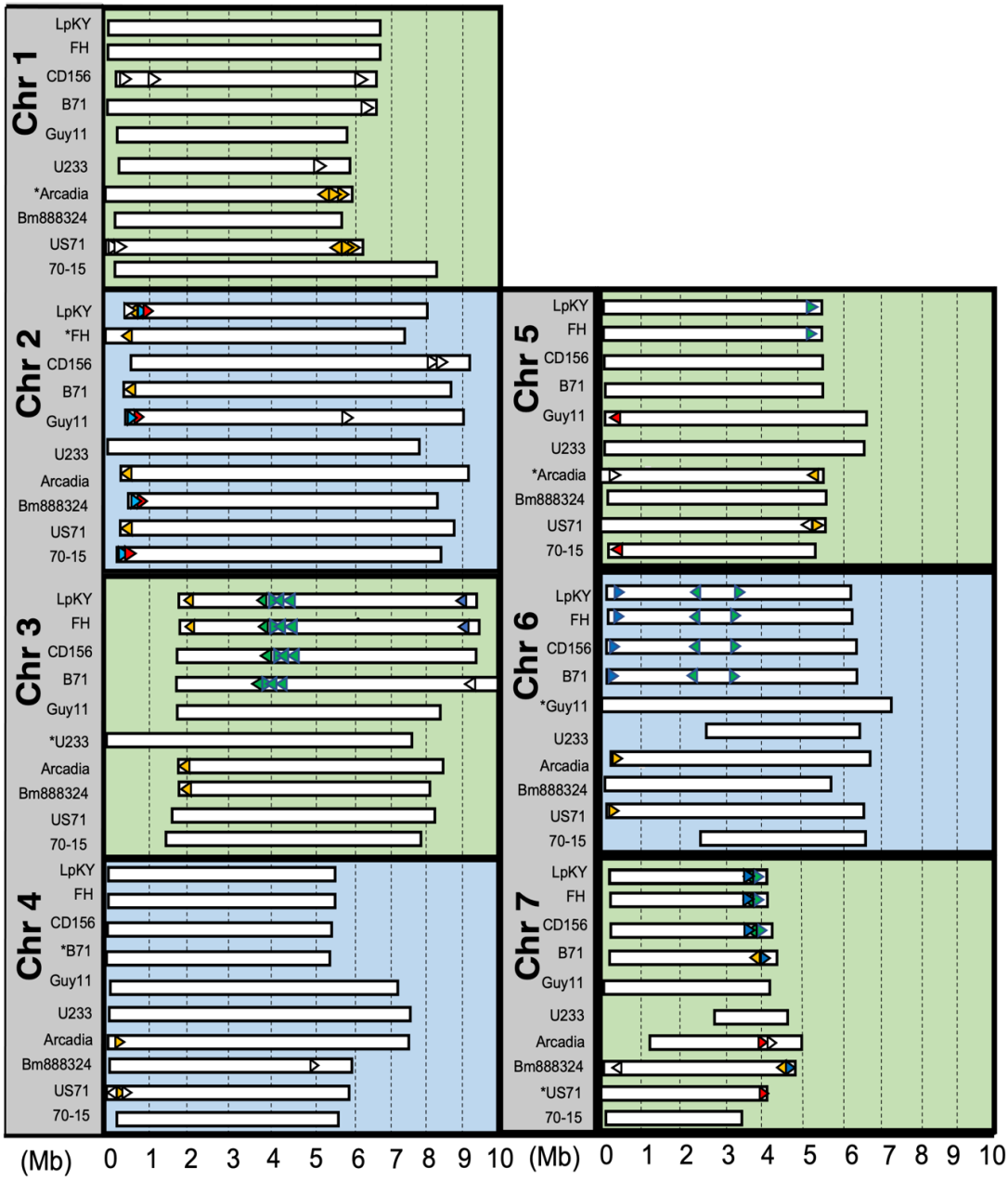


Figure 8 Mapping of MoTeR relics to the seven core chromosomes of the ten assembled Magnaporthe oryzae strain genomes LpKY, FH, CD156, B71, Guy11, U233, Arcadia, US71, and 70-15 and Bm88324. Triangles denote the orientation of relics where the apex of the triangle corresponds to the 3' end of the relic. For all seven chromosomes unique (strain-specific) relics are coded as white. All chromosomes are aligned to the chromosome of the strain with the left-most unique sequence denoted by an * next to the strain name. Colors are used to indicate which relics are shared among different strains. Similar color represent similar sequences are in the flanking side of the strains.

Chromosome 2 has the most differential retention of relics. All strains, excluding CD156 and U233, shared one or more relics with strain LpKY at the left end of Chromosome 2 (Figures 8 & 9). None of the strains contained all 4 of the relics found in strain LpKY that were captured within a 25 kb search window (Figure 9) and part of chromosome 2 in this region is truncated in Bm88324 (Figure 9). The two relics shared in strains LpKY, Guy11, 70-15, and Bm88324 have been lost in Arcadia and US71 towards the end of chromosome 2 (telomere 3; Figure 9). In chromosome 7, there is a pattern of synteny among strains Bm88324, B71, CD156, FH, and LpKY in which all five share a relic (in blue) in Chromosome 7 (Figures 8 and 10). Strains LpKY, FH, and CD156 do not share the relic found in B71 and Bm88324 (in yellow) which lies ca. 750 bp away (Figures 8 and 10).

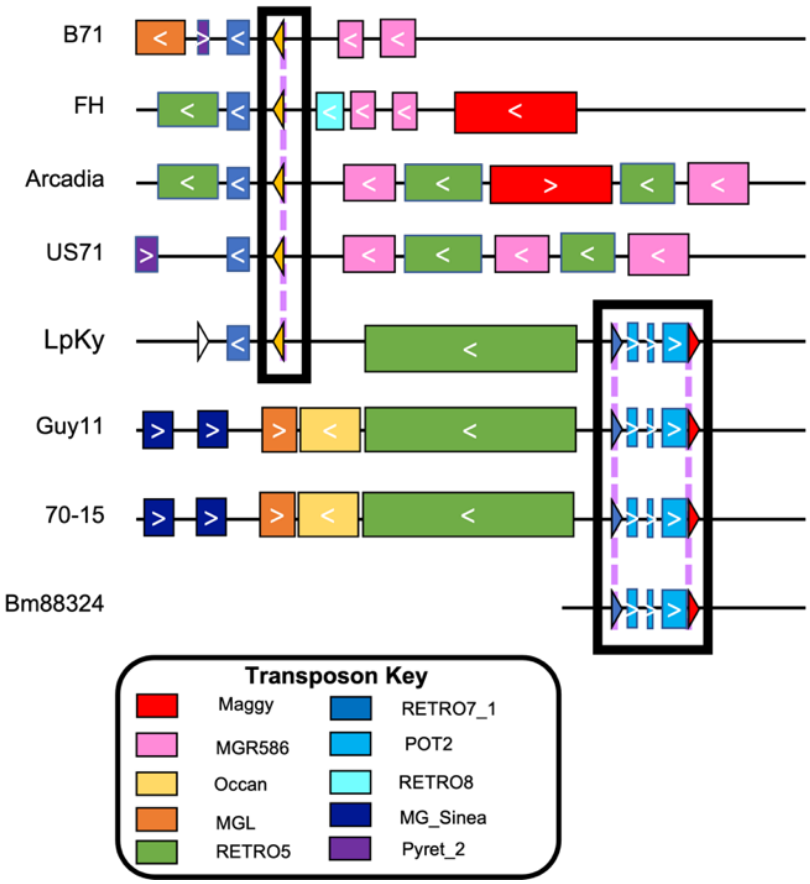


Figure 9 25 kb window of chromosomal rearrangements in a relic landscape of M. oryzae strains that share partial synteny with relics in chromosome 2 of strain LpKY but only show retention of one of two relic populations. Relic populations are surrounded by a black box and connected by dotted lines. Transposable elements are illustrated as boxes with arrows denoting their 5' to 3' orientation on chromosome 2. The beginning of the horizontal black lines does not indicate the beginning of the chromosome with the exception of strain Bm88324.

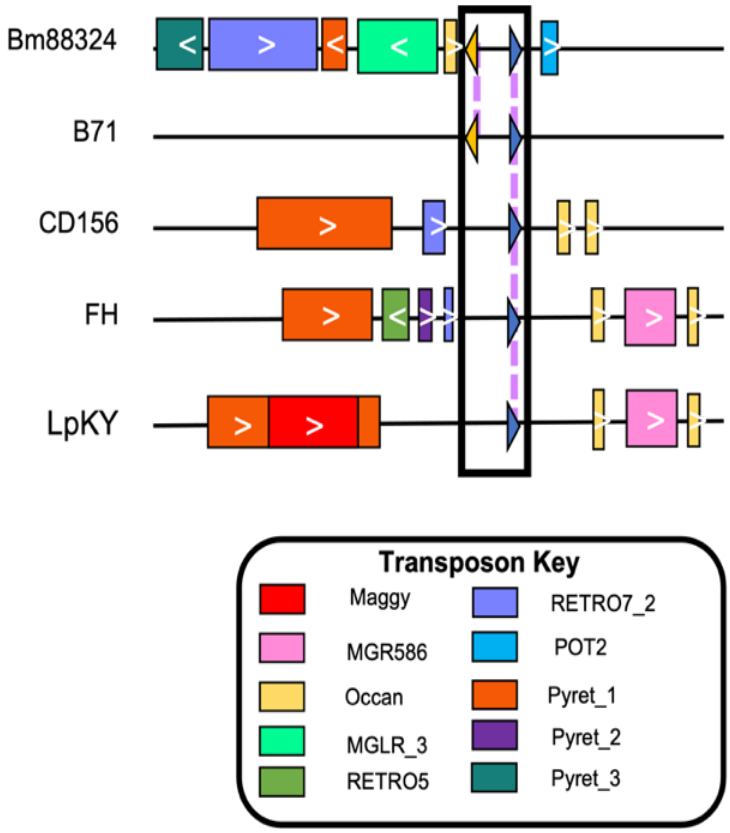


Figure 10 30 kb window of chromosomal rearrangements in a relic landscape of M. oryzae strains that share partial relic synteny in chromosome 7 of strains Bm88324, B71, CD156, FH, and LpKY.Bm88324 and B71 share two relics that are approximately 750 bp from each other, but only one of the two relics (in blue) is maintained in strains CD156, FH, and LpKY. The 5' to 3' direction of transposable elements is noted by a white arrow. Notice there are different transposable elements here that were not pictured in Figure 8, distinguished by their difference in color in the transposon key. MoTeR relics are represented as triangles and are surrounded by a black box connected by dotted lines.

MoTeR Relic Associated Duplications

Flanking sequences of MoTeR relics were examined to see if they contained duplicate sequences which could be indicative of how MoTeR relics moved from the telomere to the interior chromosomal regions. Consistent with a previous study (Rahnama *et al.* 2020), four types of duplications were detected:

1) 3' duplication: duplications that are flanking the 3' region of the relic,

2) 5’ duplication: duplications that are flanking the 5’ end of the relic,

3) relic + 5’ duplication: duplications that spanned both the relic and the sequence flanking its 5’ end, and

4) whole locus duplications: duplications in which the relic and the sequences on either side of it were duplicated (Figure 11).

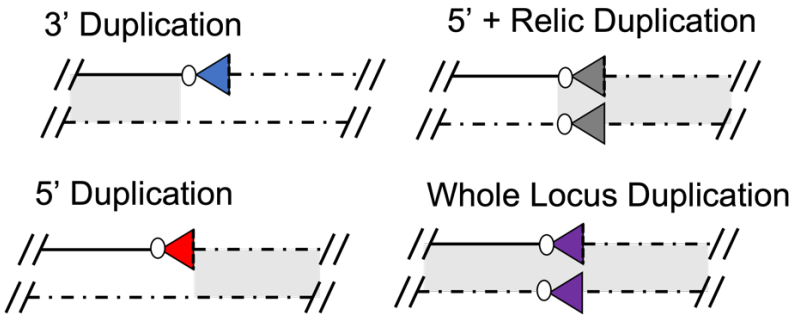


Figure 11 Duplications associated with MoTeR relic sequences. MoTeR relics are represented as triangles pointing in the 5’ -> 3’ direction. The 3’ duplication represents a duplication beginning at the 3’ boundary of the MoTeR relic and extending out. The 5’ + relic duplication indicates the duplication of the MoTeR relic and 5’ flanking sequence. The 5’ duplication begins and extends from the 5’ boundary of the relic. The whole locus duplication indicates a duplication that extends beyond either side of the MoTeR relic.

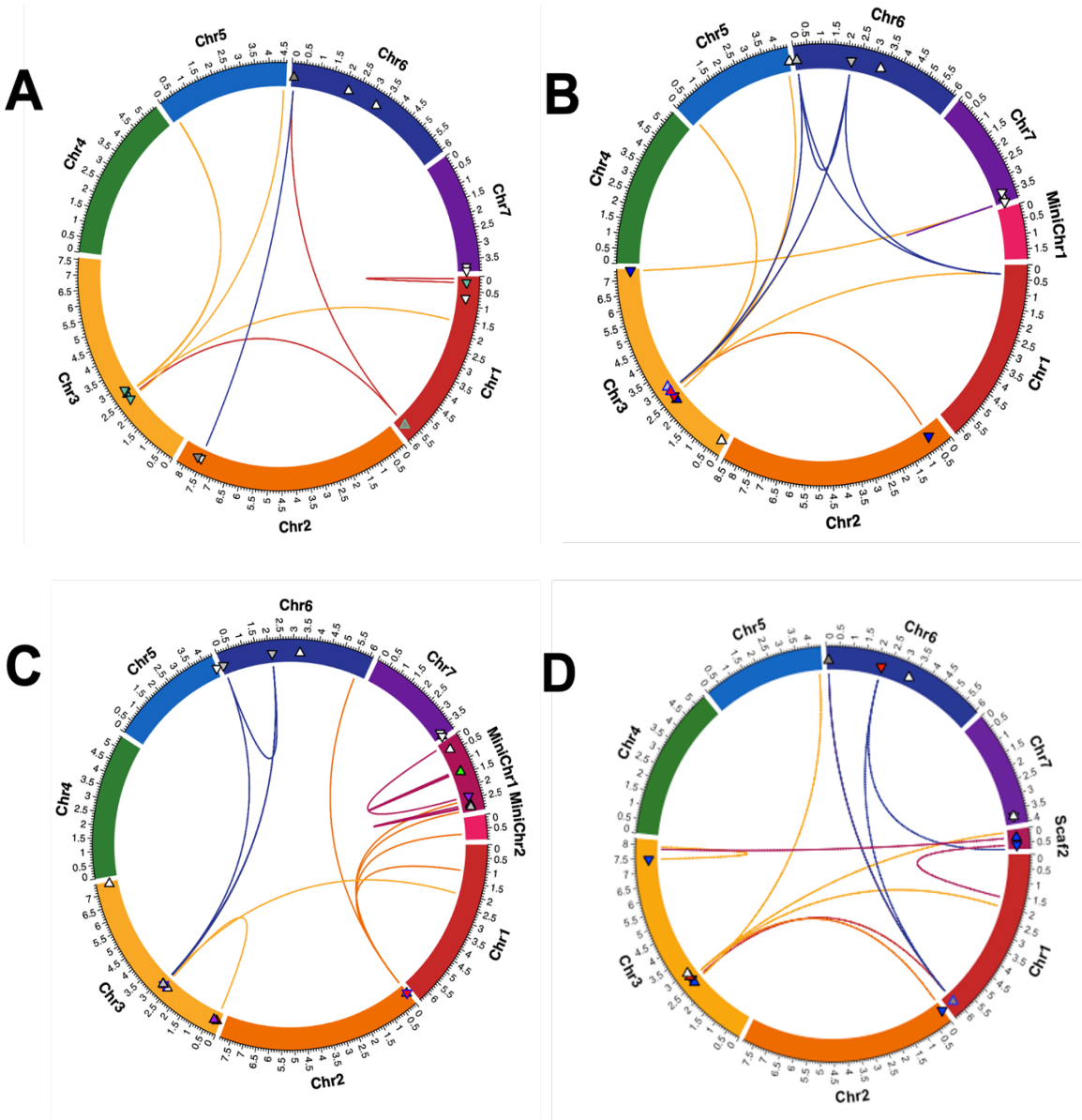
The results showed MoTeR relics were sometimes associated with sequences that had a copy elsewhere in the genome suggesting potential rearrangements could be attributed to the MoTeRs (Table 2, figure 9). The results of association of MoTeR relics with duplications (or no duplications) in 10 strain genomes are summarized in table 2. In strains 70-15, Bm88324, and U233 none of the MoTeR relics were associated with a duplicate sequence (Figure 9G, I, and J; Table 2). All duplicate sequences mapped to terminal regions of a chromosome or to another MoTeR relic except for those in strains CD156, LpKY, B71, and Arcadia that had duplications mapping from 1.5 to 2 Mb into Chromosome 1 (Figure 10A, C, D, and E). The Arcadia strain also contained a

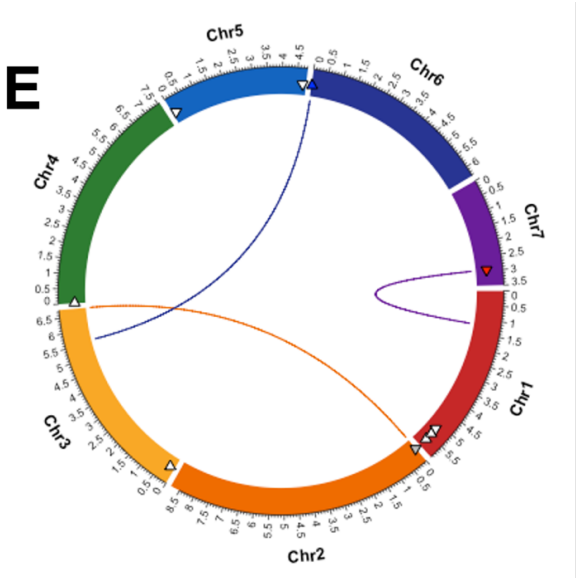
duplicate sequence that mapped to approximately 5.5 Mb in Chr 3 (Figure 10E). The largest duplication recorded was a whole locus duplication in strain US71 that was 23,985 bp long (Table S2). Overall, most duplications were <1,000 bp long (Table S2).

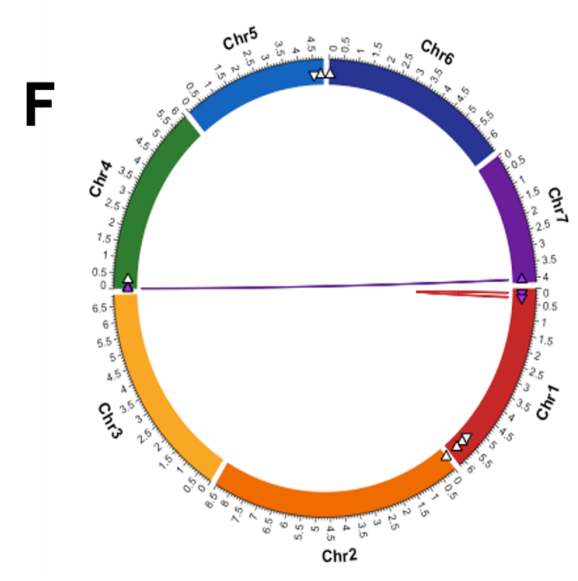
Duplicate sequences are the likely indicators of MoTeR relic formation. 5' duplications likely resulted from a break in the telomere that led to the 5' degradation of the MoTeR. The break was likely then healed through non-homologous end-joining (NHEJ) that duplicated sequences in the process.

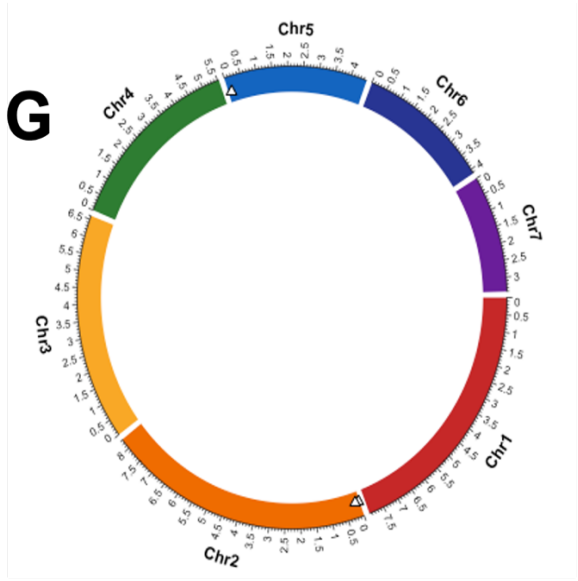
Table 2 Types of duplications in each strain of M. oryzae. Several relics have multiple duplications associated with them, so the total number of duplications recorded may exceed the number of relics per strain.

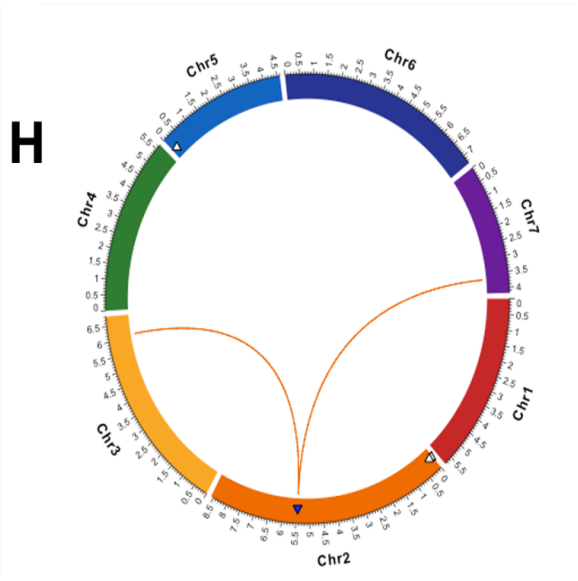
Strain	Number of Relics	No Duplications	3' Duplications	5' duplications	Relic+5' Duplications	Whole Locus Duplications
CD156	15	6	8	0	4	0
FH	14	6	10	5	7	0
LpKY	25	9	3	4	6	2
B71	14	4	7	3	2	0
Arcadia	10	7	1	1	1	0
US71	13	9	0	0	0	4
70-15	3	3	0	0	0	0
Guy11	4	3	2	0	0	0
Bm88324	7	0	0	0	0	0
U233	1	0	0	0	0	0


A
Chr5
Chr6
Chr7
Chr1
Chr2
Chr3
Chr4
B
Chr5
Chr6
Chr7
MiniChr1
Chr1
Chr2
Chr3
Chr4
C
Chr6
Chr5
Chr7
MiniChr1 MiniChr2
Chr1
Chr2
Chr3
Chr4
D
Chr5
Chr6
Chr7
Scaf2
Chr1
Chr2
Chr3
Chr4


E
Chr5
Chr6
Chr7
Chr1
Chr2
Chr3
Chr4


F
Chr5
Chr6
Chr7
Chr1
Chr2
Chr3
Chr4


G
Chr5
Chr6
Chr7
Chr1
Chr2
Chr3
Chr4


H
Chr5
Chr6
Chr7
Chr1
Chr2
Chr3
Chr4

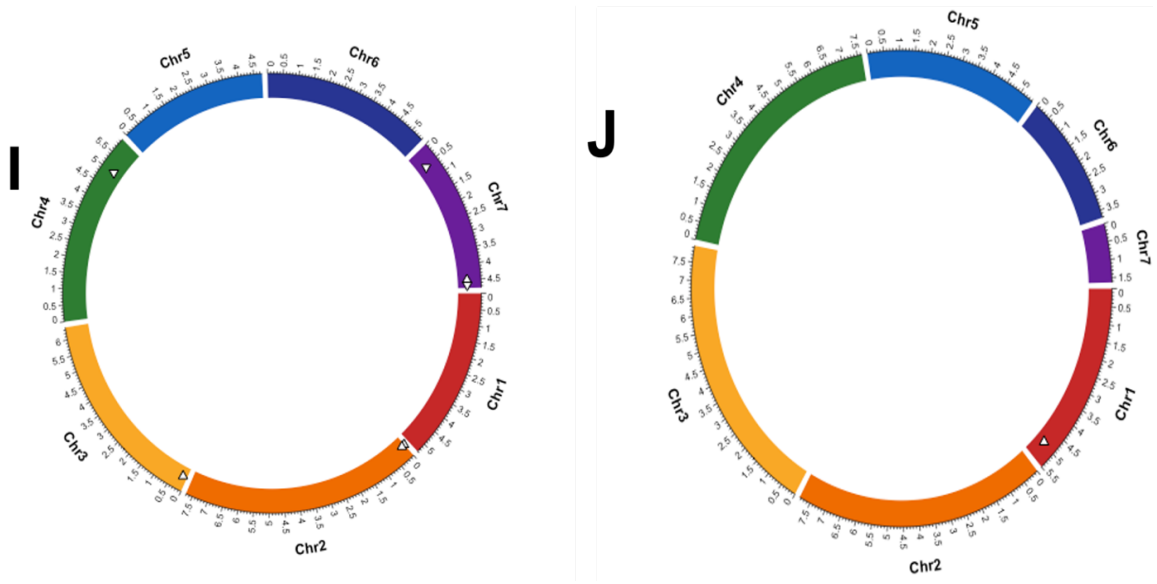


Figure 12 Duplications associated with MoTeR relic sequences. MoTeR relics are represented as triangles pointing in the 5' -> 3' direction. The color of the relic illustrates the type(s) of duplication(s) associated with it shown in the key. Links between chromosomes connect duplicate sequences associated with a MoTeR relic(s). Link colors are coded to that chromosome in which their corresponding MoTeR relic is found. **A.** CD156. **B.** FH. **C.** LpKY. **D.** B71. **E.** Arcadia. **F.** US71. **G.** 70-15. **H.** Guy-11. **I.** Bm88324. **J.** U233.

DISCUSSION

Initial Hypothesis

The purpose of this study was to explore the evolutionary history of MoTeR relics and how they arose in different strains of *M. oryzae* by examining the internalized MoTeR sequences of 10 fully assembled fungal genomes isolated from strains infecting the host grasses *Eleusine* (goosegrass), *Triticum* (wheat), *Lolium* (ryegrass), *Oryza* (rice), *Setaria* (foxtail), and *Stenotaphrum* (St. Augustine's grass) pathotypes. Our results showed that MoTeR relics were found, on average, within 250 kb of the telomere, illustrating a strong association between the two. Intergenomic comparisons of ten strains revealed differential

retention of relics across strain lineages with retention sometimes among distantly related lineages, but not more closely related ones as was the case for the relics found at the beginning of chromosome 2. MoTeR relics were also found flanked by several duplicate sequences that are likely markers for the rearrangements that took place to form MoTeR relics and push them outside of the telomere. These results further highlight the inherent instability associated with *M. oryzae* telomeres in the form of ancient rearrangements that took place to form MoTeR relics.

Our initial prediction for this study was that MoTeR relics would reveal a history of frequent chromosomal rearrangements influenced by the telomere based on evidence of telomere instability associated with MoTeRs (Starnes et al. 2012; Rahnama et al. 2020; Rahnama et al. 2021). If MoTeRs influence structural instability in the telomere, it would be reasonable to posit that the formation of MoTeR relics would not be a rare event and would likely cause the formation of unique, strain-specific profiles that would deviate from strain lineage phylogenies. Results of the frequency and distribution of MoTeR relics among strains were variable. The number of relics differed widely among strains from one relic in strain U233 to 25 in strain LpKY. Cross examination of strains revealed sections of conserved populations of relics, notably among the *Triticum*, *Eleusine*, and *Lolium* strains. These populations were found in chromosomes 3, 6, and 7, and it appears that these regions may have been inherited during a recent series of hybridization events from an *Eleusine* ancestor (Rahnama et al. 2022). Relics were also shared between other strains that were found within the

same lineage, such as the two *Oryza* strains (Guy11 and 70-15) and the two *Setaria* strains (Arcadia and US71).

I anticipated some conservation of relic sequences among more closely related strains as well as frequent deviations. What was notable was the presence of relics that were shared across several distantly related strains, e.g. LpKY (*Lolium*) and Guy11 (*Oryza*), but not consistently among all. For example, in strain LpKY the beginning of chromosome 2 contained 3 relics that were shared to varying degrees among all but two strains (U233 and CD156). These relics were found within 25 kb of each other and strain LpKY was unique in that it contained all three relics while all the other strains sharing relics in this region only shared one or two of the three. This region was populated by transposable elements whose presence could have caused a deletion of the missing relic(s) (Figure 9). Likewise, the pair of relics shared among strains B71 and Bm88324 in chromosome 7 that were separated by merely 700 bp may have been disrupted by transposon activity. MoTeR relic deletions are likely quite common as highlighted by Rahnama *et al.* in intergenomic comparisons of the MoTeR relics of strain CD156 against strains Arcadia, US71, and U233 (2021). In almost all cases the strains compared to CD156 indicated that the corresponding regions were sites of relic deletions, sometimes associated with other rearrangements. This comparison of the four (out of ten) strains examined in this study is notable as the results are most likely explained by the existence of the MoTeR relics within an ancestral population of *M. oryzae* before the divergence of the examined strain lineages. As regions found within proximity of the telomere are

not likely to experience selective pressure, their presence or absence in those regions is likely unimportant (Muszewska et al. 2019). To determine if other relics are representative of ancestral, lineage, or strain-specific sequences will require further investigation in examining potential deletions, rearrangements, and translocations that may have occurred. Thus, our hypothesis has not yet been falsified.

The comparison of strain CD156 to strains Arcadia, US71, and U233 provides strong evidence that many, if not all, of the relics present in this study have been differentially retained over evolutionary time from their ancestral population. I would predict that some MoTeR relics will still be specific to certain strains as many map as close as ca. 2,000 bp from the telomere and are potentially the result of a much more recent rearrangement. However, evidence from this study already shows that proximity to telomere ends is not equivalent to a recent rearrangement as seen on chromosome 2 in strain Bm88324 that are approximately 3-4 kb from the chromosome end (Figure 9). It appears likely that this proximity to the telomere may have been exacerbated by a recent truncation event that occurred after the divergence of this strain from its ancestor.

MoTeR relic distribution

In some cases, relics were found more distal to the telomeres than initially anticipated. These sets of relics, as mentioned earlier, were likely inherited from an *Eleusine* common ancestor as several of these strains in the *Triticum*, *Eleusine*, and Lolium lineages have genomes that indicate recent, and rapid, admixture (Rahnama et al. 2022). The majority of 3’ MoTeR relics were found

near terminal regions of the chromosome, on average 230 kb distal. Duplicate sequences associated with MoTeR relics were found to have their corresponding copy near chromosome ends. Prior analyses revealed that telomere adjacent sequences, i.e., sequences immediately next to the telomere end, were often duplicates of other terminal sequences (Rahnama et al. 2021). Telomeres often exchange information with each other through recombination events and utilize nearby telomeres for sequence repair. This should not be surprising as telomeres are often physically clustered together in the nucleus of the cell, facilitating sequence exchange through homologous recombination with neighboring telomeres (Linardopoulou et al 2005; Rahnama et al. 2021). Recombination and rearrangements among telomeres are functionally less disruptive than if a telomere were to use loci distal to the telomeres that could encode proteins necessary for cellular function (Linardopoulou et al. 2005).

MoTeR relics were often found in regions populated by other transposable elements that are highly concentrated near chromosome ends (Muszewska *et al*. 2019). It is likely that MoTeR relics are retained in these regions because they do not experience intense purifying selection as do other regions of the genome (Muszewska et al. 2019). It is postulated that the sequences associated with *M. oryzae* telomeres, including MoTeR relics, could be treated as potential borders for telomeric rearrangements that could occur without having detrimental impacts (e.g., loss of essential genes) on the genome and the organism (Rahnama et al. 2021). Extended repeat-rich regions subtending telomeres could serve as a buffer for the more strictly maintained regions of the genome. The genes that

experience intense purifying selection are not often found near the subtelomere and these sequences would not be ideal targets for telomere repair if the genome were to experience damage or replicative stress (Linardopoulou et al. 2005).

Origin of MoTeR relics?

This research effort did not entail an analysis of the mechanisms of formation of MoTeR relics, but an explanation of their possible origin is worth discussion. Surveyed chromosome ends from each genome suggest that the abundance of MoTeR elements present in the telomere does not necessarily correspond to MoTeR relic abundance (Table 1). However, of the ten examined genomes only LpKY had a fully represented telomere assembly. It is possible that telomeres may have not been fully sequenced or did not have enough statistical support for assembly, so unequivocal conclusions on MoTeR and telomere landscapes in each strain cannot be made with the current genome assemblies.

Telomere damage and repair is the most probable explanation for the formation of MoTeR relics. This is supported by their frequent association with duplicated sequences which are often formed during double-stranded break repair (Rahnama et al. 2020; Rahnama et al. 2021). While MoTeR 1 is theorized to be transcriptionally active it would require a sequence of telomere repeats to transpose to a region outside of the telomere (Starnes et al. 2012). Relics were searched for signs of transposition (i.e., target site duplications (TSD) but none were found (Rahnama *et al.* 2021). It is not improbable that these TSDs may have once been present and have since been deleted.

Telomere damage and repair is the most likely explanation for the formation of MoTeR relics, as posed by Rahnama et al. (2020; 2021), because their profiles match that of a break that might happen at interstitial telomere repeats between MoTeR elements which have previously been shown to be associated with telomere instability and breakage. This breakage could lead to repair mechanisms like non-homologous end-joining (NHEJ) that could result in sequence duplications in the process and shuffle the now truncated MoTeR out of the telomere and closer to the interior of the genome (Rahnama et al 2020). The frequent presence of duplicate sequences closely associated with MoTeR relics, the truncated ends of the relics themselves, and the lack of evidence for transposition support this. Several relics were also found to not be associated with a duplicate sequence. In these cases, it is possible that a duplicate sequence may have been lost, that perhaps the MoTeR had transposed and evidence for it had been lost (i.e., TSDs), or that maybe the relic was a passenger of another rearrangement that took place (Rahanama et al. 2020).

The presence of all these rearrangements associated with the telomere begs the question of why this might happen frequently in *M. oryzae*. The purpose of the telomere is to protect chromosome ends from degradation but there is clear evidence that *M. oryzae* has experienced frequent bouts of telomere failure throughout its evolutionary history (Rahnama et al. 2021). Telomere failure and repair within *M. oryzae* can also be observed within a single generation (Starnes et al. 2012; Rahnama et al. 2020). MoTeRs do play a role in causing instability in the telomere, but foxtail strains of *M. oryzae* have no or very few MoTeRs within

their telomeres and they still show high rates of recombination (Farman et al. 2014). In fact, while MoTeRs can cause instability within the telomeres of M. oryzae, they are not the only source of instability, and their mere presence does not always lead to telomere breakage (Rahnama et al. 2020). The telomere itself could become deprotected or vulnerable in several ways, whether that be an abnormality in the expression of the reverse transcriptase enzyme or faulty telomere capping protein complexes (Chan and Blackburn 2004).

It is thought that higher rates of instability within the telomere of *M. oryzae* could provide an adaptive advantage (Rahnama et al 2020; Rahnama et al 2021). Several human pathogens house genes key for adaptation to their host within the subtelomere which allows for stochastic silencing and activation as well as higher rates of mutation within genes housed in this region (Berriman *et al.* 2005; Freitas-Junior *et al.* 2000). It is also probable that a crucial resistance gene in some *Oryza* strains of *M. oryzae,* known as Avr-Pita, was lost due to telomere failure and subsequent truncation or loss of the gene that allowed for strains lacking this protein to infect rice hosts with the complimentary resistance proteins. As a result, the rice grasses would not easily recognize fungal strains without the fungus' Avr-pita protein to signal their infection (Orbach et al. 2000). Additionally, mechanisms for repair following double stranded breaks in the genome often duplicate other sequences to fill in the gaps of the broken sequence through non-homologous end-joining (NHEJ; Linardopoulou et al. 2005; Rahnama et al. 2020). It has been posed that if a break in the telomere were to lead to the duplication of a gene, or a stretch of genes, the duplicate

copy of that gene would then find itself in a more dynamic region of the genome where it might experience adaptive mutations. It would also have a “back-up” copy in the chromosome if the twin that migrated to the subtelomere were to gain some non-advantageous mutations (Rahnama et al. 2020; Rahnama et al 2021).

Limitations

There were several limitations to this study that should be noted as this study does not present a complete analysis of all MoTeR relics and their relationships with the chromosomes and strains in which they reside. Some of these limitations open avenues for future studies and proper utilization of the UNIX Command Line and appropriate bioinformatics software to detect some of the more problematic/questionable nucleotide sequences that were potentially overlooked in this thesis.

1) It is likely that not all 3’ MoTeR relics were found in this study as both grep and BLASTn failed to identify some MoTeR relics that were present in strains. BLASTn is the more advanced approach for finding MoTeR sequences in the genome as it can handle mismatches in the sequences when it comes to differences in sequence length as well as a certain number of nucleotide mismatches. The threshold for how many mismatches BLAST will allow depends on the length of the sequence as well as the defined e-value. For short BLASTn I used an e-value of 1e-1 because I was attempting to capture a shorter sequence (~20 nt minimum) and I wanted to allow for some mismatches. An e-value higher than that

for sequences this short would have returned far fewer results for many of the shorter relics present in the genome.

Several new relics were discovered using subsequent BLASTn pairwise intergenomic comparisons. For example, a local BLASTn alignment of one strain's genome against another might reveal the positions of two matching MoTeR relics. One such example can be seen in the chromosome 3 of strain CD156 where a 15 nt MoTeR relic was found matching relics found in the same location in strains LpKY and FH. This relic was easily missed during BLASTn and grep searches because matches less than 20 nucleotides were filtered out. However, this 15 nt relic is supported by the presence of longer relics in the same location in strains FH and LpKY.

2) In some strains there were initial patterns of mismatch in relic populations that had a clear relationship with one another. For example, a relic found in chromosome 3 in strains LpKY, CD156, and B71 was shared between strains LpKY and CD156 and between strains CD156 and B71 but was not detected using BLASTn between strains LpKY and B71. One explanation could be that some point mutations were between strains LpKY and CD156 and CD156 and B71, but the point mutations in strains LpKY and B71 were sufficiently different enough using short BLASTn. That is, strains LpKY and B71 had an excess of nucleotide differences in their MoTeR sequences to reach a statistically significant (and detectable) match using BLASTn but had sufficient identity to CD156 in that their true

identities were discernable by BLASTn. It should be noted that the BLAST algorithm utilizes both the length of the two sequences being aligned and their overall sequence identity to arrive at a positive match based on the specific e value chosen by the investigator. Future methods might approach using “fuzzy grep” that would be able to substitute mismatches that might not be detected by BLASTn. Depending on the length of sequence this strategy could become laborious. yt

3) Minichromosomes were not investigated in this study because there were issues with BLASTn and grep detecting matches. This could be due to the minichromosome assemblies which are problematic as they are often comprised of sequences found in the core chromosomes and are difficult to ascertain without first isolating them on an agarose gel. We know that the minichromosome of strain LpKY is full of MoTeR relics from previous studies (Rahnama et al. 2020). Additional analysis is needed to elucidate the presence of MoTeR relics in the strains that have sequenced minichromosomes. These could certainly reveal dynamic inter-specific rearrangements between the minichromosomes and the core chromosomes as previous studies have shown they have been known to trade sequences (Peng et al. 2019). Further analyses targeting minichromosomes would also require better genome assemblies as was the case for B71.

Future directions

Our initial hypothesis for this study that MoTeR relics would largely represent unique strain or lineage-specific rearrangements was neither unequivocally supported nor entirely refuted. In many cases strains shared MoTeR relics in some chromosomes across distantly related lineages that signified their formation within ancestral populations before the divergence of the strains examined. One study also revealed that even seemingly unique relic profiles were sites of differentially retained ancestral relics (Rahnama et al 2021). It is tempting to think that all relics belonged to ancestral lineages, however, this will require more in-depth comparisons of strains to see if sites might have once contained a corresponding MoTeR relic or if the relic itself might be the site for a unique rearrangement. This is an analysis that I intend to explore but for the sake of time excluded from the current thesis.

Future directions for this study could also pursue the identity of duplicate sequences for the presence of protein coding regions that may serve as an adaptive potential. Likewise, reasoning for the formation of 5' duplications were posed in a 2020 study (Rahnama et al.) but the dynamics behind how a 3' flanking duplication might have occurred are not entirely clear. Only 3' end MoTeR relics were investigated in this study and the methods for searching for these shorter sequences that have likely undergone mutations are not unequivocal. Finally, future investigations of MoTeR relic dynamics in minichromosomes could reveal an interesting narrative of core and minichromosome "cross-talk" (Peng et al. 2019) that could potentially be related to MoTeR dynamics if minichromosome and core chromosome telomeres

interact, or even if MoTeRs themselves could play a role in the formation of minichromosomes (Rahnama et al. 2020). MoTeR relics and their relationship with telomeres continue to present doorways for investigating telomere dynamics in *M. oryzae* in a system where telomere rearrangements can be observed in a single generation (Starnes et al. 2012; Rahnama et al. 2020) and their kinetics could play a role in the rapid adaptation that is so integral to the success of this fungal pathogen.

LITERATURE CITED

Aksoy, S., S. Williams, S. Chang, and F.F. Richards, 1990 SLACS retrotransposon from *Trypanosoma brucei gambiense* is similar to mammalian LINEs. Nucleic Acids Res. 18: 785–792.

Berbee, M.L., 2001 The phylogeny of plant and animal pathogens in the Ascomycota. Physiological and Molecular Plant Pathology. 59(4):165-187.

Berriman, M., E. Ghedin, C. Hertz-Fowler, G. Blandin, H. Renauld, D.C. Bartholomeu, N.J. Lennard, E. Caler, N.E. Hamlin, B. Haas, U. Böhme, L. Hannick, M.A. Aslett, J. Shallom, L. Marcello, L. Hou, B. Wickstead, U.C.M. Alsmark, C. Arrowsmith, R.J. Atkin, A.J. Barron, F. Bringaud, K. Brooks, M. Carrington, I. Cherevach, T.J. Chillingworth, C. Churcher, L.N. Clark, C.H. Corton, A. Cronin, R.M. Davies, J. Doggett, A. Djikeng, T. Feldblyum, M.C. Field, A. Fraser, I. Goodhead, Z. Hance, D. Harper, B.R. Harris, H. Hauser, J. Hostetler, A. Ivens, K. Jagels, D. Johnson, J. Johnson, K. Jones, A.X. Kerhornou, H. Koo, N. Larke, S. Landfear, C. Larkin, V. Leech, A. Line, A. Lord, A. MacLeod, P.J. Mooney, S. Moule, D.M.A. Martin, G.W. Morgan, K. Mungall, H. Norbertczak, D. Ormond, G. Pai, C.S. Peacock, J. Peterson, M.A. Quail, E. Rabbinowitsch, M.A. Rajandream, C. Reitter, S.L. Salzberg, M. Sanders, S. Schobel, S. Sharp, M. Simmonds, A.J. Simpson, L. Tallon, C.M.R. Turner, A. Tait, A.R. Tivey, S. Van Aken, D. Walker, D. Wanless, S. Wang, B. White, O. White, S. Whitehead, J. Woodward, J. Wortman, M.D. Adams, T.M. Embley, K. Gull, E. Ullu, J.D. Barry, A.H. Fairlamb, F. Opperdoes, B.G. Barrell, J.E.

Donelson, N. Hall, C.M. Fraser, S.E. Melville, and N.M. El-Sayed, 2005 The genome of the African trypanosome *Trypanosoma brucei.* Science 309(5733): 416-22.

Brock, T. D., D. W. Smith, and M. T. Madigan, 1984 *Biology of Microorganisms.* Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Chan, S.R.W.L and E.H. Blackburn, 2004 Telomeres and telomerase. Royal Society 359(1441).

Dangl, J.L., D.M. Horvath, and B.J. Staskawicz, 2013 Pivoting the plant immune system from dissection to deployment. Science 341(6147): 746-751.

Dean, R. A., N. J. Talbot, D. J. Ebbole, M. L. Farman, T. K. Mitchell, M.J. Orbach, M. Thon, R. Kulkarni, J.R. Xu, H. Pan, N.D. Read, Y.I. Lee, I. Carbone, D. Brown, Y.O. Yeon, N. Donofrio, S.J. Jun, D.M. Soanes, S. Djonovic, E. Kolomlots, C. Rehmeyer, W. Li, M. Harding, S. Kim, M.H. Lebrun, H. Bohnert, S. Coughlan, J. Butler, S. Calvo, L. Ma, R. Nicol, S. Purcell, C. Nusbaum, J.E. Galagan, and B.W. Birren, 2005 The genome sequence of the rice blast fungus *Magnaporthe grisea.* Nature 434: 980-986.

De Lange, T., V. Lundblad, and E.H. Blackburn, 2006 A history of telomere biology. In: Telomeres. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, p. 1-19.

Farman, M., O.S. Novikova, J.H. Starnes, and D.W. Thornbury, 2014 “Subtelomere organization, evolution, and dynamics in the rice blast fungus *Magnaporthe oryzae*,” in Subtelomeres, eds E.J. Louis and M.M.

Becker (Berlin: Springer-Verlag), 71-99. Doi: 10.1007/978-3-642-41566-1_4

Farman, M.L., G. Peterson, L. Chen, J. Starnes, B. Valent, P. Bachi, L. Murdock, D. Hershman, K. Pedley, J.M. Fernandes, and J. Bavaresco, 2017 The *Lolium* Pathotype of *Magnaporthe oryzae* recovered from a single blasted wheat plant in the United States. Plant Disease 101: 684-692.

Fleming, A., 1941 Penicillin. Br. Med. J. 2(4210):386.

Freitas-Junior, L. H., E. Bottius, L. A. Pirrit, K. W. Deitsch, C. Scheidig, F. Guinet, U. Nehrbass, T.E. Wellems, and A. Scherf*et*, 2000 Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. Nature 407: 1018-1022.

Gabriel, A., T.J. Yen, D.C. Schwartz, C.L. Smith, J.D. Boeke, B. Sollner-Webb, and D.W. Cleveland, 1990 A rapidly rearranging retrotransposon within the miniexon gene locus of *Crithidia fasciculata*. Mol. Cell. Biol. 10: 615–624.

Gallone, B., J. Steensels, T. Prahl, L. Soriaga, V. Saels, B. Herrera-Malaver, A. Merlevede, M. Roncoroni, K. Voordeckers, L. Miraglia, C. Teiling, B. Steffy, M. Taylor, A. Schwartz, T. Richardson, C. White, G. Baele, S. Maere, and K.J. Verstrepen, 2016 Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts. Cell 166(6):1397-1410.e16

Gan, T., T. Luo, K. Pang, C. Zhou, G. Zhou, B. Wan, G. Li, Q. Yi, A.D. Czaja, and S. Xiao, 2021 Cryptic terrestrial fungus-like fossils of the early Ediacaran Period. Nature Comm 12:641.

Gnat, S., D. Lagowski, and A. Nowakiewicz, 2020 Major challenges and perspectives in the diagnostics and treatment of dermatophyte infections. J. Applied Microbiol. 129(2):212-232.

Gladieux, P., B. Condon, S. Ravel, D. Soanes, J.L.N. Maciel, A. Nhani Jr., R. Terauchi, M. Lebrun, D. Tharreau, T. Mitchell, K.F. Pedley, B. Valent, N.J. Talbot, M. Farman, and E. Fournier, 2018 Gene flow between divergent cereal- and grass-specific lineages of the rice blast fungus *Magnaporthe oryzae*. American Soc. Microbiol. 9(1) e01219-17

Hibbett, D.S., M. Binder, J.F. Bischoff, M. Blackwell, P.F. Cannon, O.E. Eriksson, S. Huhndorf, T. James, P.M. Kirk, R. Lücking, H.T. Lumbsch, F. Lutzoni, P.B. Matheny, D.J. McLaughlin, M.J. Powell, S. Redhead, C.L. Schoch, J.W. Spatafora, J.A. Stalpers, R. Vilgalys, M.C. Aime, A. Aptroot, R. Bauer, D. Begerow, G.L. Benny, L.A. Castlebury, P.W. Crous, Y.C. Dai, W. Gams, D.M. Geiser, G.W. Griffith, C. Gueidan, D.L. Hawksworth, G. Hestmark, K. Hosaka, R.A. Humber, K.D. Hyde, J.E. Ironside, U. Kõljalg, C.P. Kurtzman, K.H. Larsson, R. Lichtwardt, J. Longcore, J. Miadlikowska, A. Miller, J.M. Moncalvo, S. Mozley-Standridge, F. Oberwinkler, E. Parmasto, V. Reeb, J.D. Rodgers, C. Roux, L. Ryvarden, J.P. Sampaio, A. Schüßler, J. Sugiyama, R.G. Thorn, L. Tibell, W.A. Untereiner, C. Walker, Z. Weng, A. Weir, M. Weiss, M.M. White, K. Winka, Y.J. Yao, and N. Zhang, 2007 A higher-level phylogenetic classification of the Fungi. Mycological Res. 111(5): 509-547.

Inoue, Y., T.T.P. Vy, K. Yoshida, H. Asano, C. Mitsuoka, S. Asuke, V.L. Anh, C.J.R. Cumagun, I. Chuma, R. Terauchi, K. Kato, T. Mitchell, B. Valent, M. Farman, and Y. Tosa, 2017 Evolution of the wheat blast fungus through functional losses in a host specificity determinant. Science 357: 80-83.

Kato, H. 2001 Rice blast disease. Pesticide Outlook 12: 23-25.

Kim, E., J. Kim, C. Kim, and J. Lee, 2021 Long-read sequencing and *de novo* genome assemblies reveal complex chromosome end structures caused by telomere dysfunction at the single nucleotide level. Nucleic Acids Res 49(6): 3338-3353.

Krzywinski, M., J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S.J. Jones, and M.A. Marra, 2009 Circos: an information aesthetic for comparative genomics. Genome Res. 19:1639-1645.

Latgé, J., 1999 *Aspergillus fumigatus* and aspergillosis. Clinical Micro. Rev. 12(2): 310-350.

Levis, R.W., R. Ganesan, K. Houtchens, L.A. Tolar, and F. Sheen, 1993 Transposons in place of telomeric repeats at a *Drosophila* telomere. Cell 75(6): 1083-1093.

Linardopoulou, E.V., E.M. Williams, Y. Fan, C. Friedman, J.M. Young, and B.J. Trask, 2005 Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. Nature 437: 94-100.

Luo, J., H. Qiu, G. Cai, N.E. Wagner, D. Bhattacharya, and N. Zhang, 2015 Phylogenomic analysis uncovers the evolutionary history of nutrition and infection mode in rice blast fungus and other Magnaporthales. Sci. Reports 5(9448):1-6.

Muszewska, A., K. Steczkiewicz, M. Stepniewska-Dziubinska, and K. Ginalski, 2019 Transposable elements contribute to fungal genes and impact fungal lifestyle. Scientific Reports. 9: 4307.

Okazaki, S., H. Ishikawa, and H. Fujiwara, 1995 Structural analysis of TRAS1, a novel family of telomeric repeat-associated retrotransposons in the silkworm, *Bombyx mori*. Mol. Cell Biol. 15(8):4545-4552.

Orbach, M.J., L. Farrall, J.A. Sweigard, F.G. Chumley, and B. Valent, 2000 A telomeric avirulence gene determines efficacy for the rice blast resistance gene Pi-ta. Plant Cell 12(11):2019-32.

Orr, H.A., and S.P. Otto, 1994 Does diploidy increase the rate of adaptation? Genetics 136(4): 1475-1480.

Ou, S.H., 1985 *Rice Diseases.* Commonwealth Agricultural Bureaux, Slough, UK.

Peng, Z., E. Oliveira-Garcia, G. Lin, Y. Hu, M. Dalby, P. Migeon, H. Tang, M. Farman, D. Cook, F. F. White, B. Valent, and S. Liu, 2019 Effector gene reshuffling involves dispensable mini-chromosomes in the wheat blast fungus. PLoS Genetics 15(9): 1-23.

Rahnama, M., O. Novikova, J. Starnes, L. Chen, S. Zhang, and M. Farman, 2020 Transposon-mediated destabilization: a driver of genome evolution in the blast fungus. Nucleic Acids Res. 48(13):7197-7217.

Rahnama, M., B. Wang, J. Dostart, O. Novikova, D. Yackzan, A. Yackzan, H. Bruss, M. Baker, H. Jacob, X. Zhang, A. Stewart, M. Heist, J. Hoover, P. Calie, L. Chen, J. Liu, M.L. Farman, 2021 Telomere roles in fungal genome evolution and adaptation. Front. Genet. e676751

Rahnama, M., B. Condon, J.P. Ascari, J.R. Dupuis, E. Del Ponte, K.F. Pedley, S. Martinez, B. Valent, and M.L. Farman, 2022 Recombination of standing

variation in a multi-hybrid swarm drove adaptive radiation in a fungal pathogen and gave rise to two pandemic plant diseases. Nature Comm. *In Review*

Sánchez-Vallet, A., S. Fouché, I. Fudal, F.E. Hartmann, J.L. Soyer, A. Tellier, and D. Croll, 2018 The Genome Biology of Effector Gene Evolution in Filamentous Plant Pathogens. Ann. Rev. Phytopath. 56:21-40.

Skamnioti, P., and S.J. Gurr, 2009 Against the grain: safeguarding rice from rice blast disease. Trends Biotech. 27: 141-150.

Starnes, J. H., D.W. Thornbury, O.S. Novikova, C.J. Rehmeyer, and M.L. Farman, 2012 Telomere-targeted retrotransposons in the rice blast fungus *Magnaporthe oryzae*: agents of telomere instability. Genetics. 191: 389-406.

Sudbery, P.E., 2011 Growth of *Candida albicans* hyphae. Nature Reviews Microbiology. 9: 737-748.

Takahashi, H., S. Okazaki, and H. Fujiwara, 1997 A new family of site-specific retrotransposons, SART1, is inserted into telomeric repeats of the silkworm, *Bombyx mori*. Nucleic Acids Res. 25(8): 1578-1584.

Talbot, NJ., and R.A. Wilson, 2009 Under pressure: investigating the biology of plant infection by *Magnaporthe oryza.* Nature Rev. Microbiol. 7:185-195.

Thorvaldsdottir, H., J.T. Robinson, and J.P. Mesirov, 2013 Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics. 14(2):178-192.

SUPPLEMENTARY MATERIALS

Supplementary Table 1 Sources of genome assemblies including the host plant they were collected from, their lineage, NCBI Accession Number, and the reference in which their assembly is published.

Strain	Host	Lineage	Reference	NCBI Accession #
LpKY97	*Lolium perenne* (perennial ryegrass)	*Lolium*	Farman et al. 2017	SAMN08009564
FH	*L. perenne*	*Lolium*	Pieck et al. 2017	SAMN08009551
CD156	*Eleusine indica* (goose grass)	*Eleusine*	Chiapello et al. 2015	SAMEA4708261
B71	*Triticum aestivum* (wheat)	*Triticum*	Inoue et al. 2017	SAMN04942725
Arcadia	*Setaria viridis* (green foxtail)	*Setaria*	Farman et al. 2014	SAMN14167122
US71	*Setaria spp.*	*Setaria*	Chiapello et al. 2015	SAMEA3373385
Bm88324	*Brachiaria mutica* (Buffalo grass)	*Brachiaria*	Borromeo et al. 1993	SAMN08009544
U233	*Stenotaphrum secundatum* (St. Augustine grass)	*Stenotaphrum*	Yasuhara-Bell et al. 2018	SAMN19488846
Guy11	*Oryza sativa* (rice)	*Oryza*	Islam et al. 2016	SAMN06050151
70-15	O. sativa	*Oryza*	Dean et al. 2005	SAMN02953596

References for genome assemblies

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

Chiapello, H., L. Mallet, C. Guérin, G. Aguileta, J. Amselem, T. Kroj, E. Ortega-Abboud, M. Lebrun, B. Henrissat, A. Gendrault, F. Rodolphe, D. Tharreau, and E. Fournier, 2015 Deciphering genome content and evolutionary relationships of isolates from the fungus *Magnaporthe oryzae* attacking different host plants. Genome Biol. And Evol. 7(10): 2896-2912.

Dean, R. A., N. J. Talbot, D. J. Ebbole, M. L. Farman, T. K. Mitchell, M.J. Orbach, M. Thon, R. Kulkarni, J.R. Xu, H. Pan, N.D. Read, Y.I. Lee, I. Carbone, D. Brown, Y.O. Yeon, N. Donofrio, S.J. Jun, D.M. Soanes, S. Djonovic, E. Kolomlots, C. Rehmeyer, W. Li, M. Harding, S. Kim, M.H. Lebrun, H. Bohnert, S. Coughlan, J. Butler, S. Calvo, L. Ma, R. Nicol, S. Purcell, C. Nusbaum, J.E. Galagan, and B.W. Birren, 2005 The genome sequence of the rice blast fungus *Magnaporthe grisea.* Nature 434: 980-986.

Farman, M., O.S. Novikova, J.H. Starnes, and D.W. Thornbury, 2014 "Subtelomere organization, evolution, and dynamics in the rice blast fungus *Magnaporthe oryzae*," in Subtelomeres, eds E.J. Louis and M.M. Becker (Berlin: Springer-Verlag), 71-99. Doi: 10.1007/978-3-642-41566-1_4

Farman, M.L., G. Peterson, L. Chen, J. Starnes, B. Valent, P. Bachi, L. Murdock, D. Hershman, K. Pedley, J.M. Fernandes, and J. Bavaresco, 2017 The *Lolium* Pathotype of *Magnaporthe oryzae* recovered from a single blasted wheat plant in the United States. Plant Disease 101: 684-692.

Inoue, Y., T.T.P. Vy, K. Yoshida, H. Asano, C. Mitsuoka, S. Asuke, V.L. Anh, C.J.R. Cumagun, I. Chuma, R. Terauchi, K. Kato, T. Mitchell, B. Valent, M. Farman, and Y. Tosa, 2017 Evolution of the wheat blast fungus through functional losses in a host specificity determinant. Science 357: 80-83.

Islam, M.T., D. Croll, P. Gladieux, D.M. Soanes, A. Persoons, P. Bhattacharjee, M.S. Hossain, D.R. Gupta, M.M. Rahman, M.G. Mahboob, N. Cook, M.U. Salam, M.Z. Surovy, V.B. Sancho, J.L.N. Maciel, A. NhaniJúnior, V.L. Castroagudín, J.T. de Assis Reges, P.C. Ceresini, S. Ravel, R. Kellner, E. Fournier, D. Tharreau, M. Lebrun, B.A. McDonald, T. Stitt, D. Swan, N.J. Talbot, D.G.O. Saunders, J. Win, and S. Kamoun, 2016. Emergence of wheat blast in Bangladesh was caused by a South American lineage of *Magnaporthe oryzae.* BMC Biol. 14:84

Pieck, M.L., A. Ruck, M.L. Farman, G.L. Peterson, J.P. Stack, B. Valent, and K.F. Pedley, 2016 Genomics-based marker discovery and diagnostic assay development for wheat blast. Plant Dis. 101: 103-109.

Yasuhara-Bell, J., K. F. Pedley, M. Farman, B. Valent and J. P. Stack, 2018 Specific Detection of the Wheat Blast Pathogen (*Magnaporthe oryzae Triticum*) by Loop-Mediated Isothermal Amplification. Plant Dis. 102**:** 2550-2559.

Supplementary Table 2 MoTeR relic comparisons among 10 fully assembled Magnaporthe oryzae strain genomes (LpKY, FH, CD156, B71, Bm88324, U233, US71, Guy11, 70-15, Arcadia). MoTeR relics are truncated MoTeR 1 and 2 elements that are found in the interior of the genome, or more specifically, outside of telomeres. MoTeR relics investigated were only those that contained a conserved 3' sequence ($^{5'}$CGCGAATTAAAA$^{3'}$;$^{3'}$TTTTAACGCG$^{5'}$) flanked by one or more telomere repeats ($^{5'}$**CCCTAA**$^{3'}$;$^{3'}$**TTAGGG**$^{5'}$). 35 of 80 3' MoTeR relics identified contain single nucleotide polymorphisms in the MoTeR sequence and/or telomere repeat(s).

Chr	Strain	Shared with	3' end position	Distance from Nearest Telomere or Terminal MoTeR	Relic length	3' end sequence
1	CD156	-	215205	215205	40	CGCGAATTAAAA**CCCTAACCCTTA**
1	CD156	-	753727	753727	4,277	GCGCGAATTAAGA**CCCAT**
1	CD156	-	5749080	259256	40	CGCGAATTAAAA**CCCTA**TA
1	B71	-	6127382	309277	40	CGCGAATTAAAA**CCCTA**
1	U233	-	4947590	552589	347	CGCGAATTAAAA**CCCTAA**
1	Arcadia	US71	5692246	295109	26	CGCGAATTAAAA**CCCT**
1	Arcadia	US71	5425062	562293	97	CGCGAATTAAAA**CCCTAACCCT**

Chr	Strain	Shared with	3' end position	Distance from Nearest Telomere or Terminal MoTeR	Relic length	3' end sequence
1	Arcadia	US71	5374231	613124	140	**TTATGGTGAGGG**TTTTAATTCGCA
1	US71	-	158743	158743	38	CGCGAATTAAAA**CCTTAACCCTAA**
1	US71	-	313516	313516	38	CGCGAATTAAAA**CCTTAACCCTAA**
1	US71	Arcadia	5599609	623091	440	**TTATGGTGAGGG**TTTTAATTCGCA
1	US71	Arcadia	5650676	572024	97	CGCGAATTAAAA**CCCTAACCCT**
1	US71	Arcadia	5918660	304040	26	CGCGAATTAAAA**CCCTGA**
2	LpKY97	-	25940	12467	98	ACGCGAATTAAAA**CCCTAA**
2	LpKY97	FH, B71 & Arcadia	28007	14534	117	*AAAATTAAGCGC
2	LpKY97	Guy11 & Bm88324, 70-15	42376	28903	40	GCGCGAATTAAAA**CCCTAACCCTAAC**
2	LpKY97	Guy11 & Bm88324, 70-15	43519	30046	102	CGCGAAACAAAA**CCCAAA**
2	FH	LpKY97, B71 & Arcadia	645364	633874	109	**GGTTAGGG**TTTTAATTCGCG

Chr	Strain	Shared with	3' end position	Distance from Nearest Telomere or Terminal MoTeR	Relic length	3' end sequence
2	CD156	-	7550702	714299	31	CGCGAATTAAAA**CCCTAAC**
2	CD156	-	7553167	711834	40	CGCGAATTAAAA**CCCTA**
2	B71	LpKY97, FH & Arcadia	39554	33103	118	**GGTTAGGG**TTTTAATTCGCG
2	Guy11	LpKY97 & Bm88324, 70-15	44503	44402	40	CGCGAATTAAAA**CCCTAACCGTAA**
2	Guy11	LpKY97, Bm88324, 70-15	45642	45541	102	CGCGAAATAAAA**CCCAAACTCTCTCCCTA**
2	Guy11	-	5458604	3132066	40	CGCGAATTAAAA**CCCTAA**
2	Arcadia	US71, LpKY97, FH & B71	11850	11850	118	**GGTTAGGG**TTTTAATTCGCG
2	Bm88324	LpKY97 & Guy11, 70-15	3176	3176	39	CGCGAATTAAAA**CCCTAACCCTAA**
2	Bm88324	LpKY97 & Guy11, 70-15	4311	4311	98	CGCGAAATAAAG**CCCAAAA**
2	US71	Arcadia, LpKY97, FH & B71	3026	3026	118	**TTAACGGTTAGGG**TTTTAATTCGCG
2	70-15	LpKY97 & Bm88324, Guy11	35652	35652	40	CGCGAATTAAAA**CCCTAACCGTAA**

Chr	Strain	Shared with	3' end position	Distance from Nearest Telomere or Terminal MoTeR	Relic length	3' end sequence
2	70-15	LpKY97 & Bm88324, Guy11	36813	36813	63	CGCGAAATAAAA**CCCAAA**
3	LpKY97	FH, Arcadia & Bm88324	152722	146667	40	**GGGTTAGGG**TTTTAATTCGCGT
3	LpKY97	FH, CD156 & B71	2548296	2542241	41	**AGGGTTAGGG**TTTTAATTCGCGT
3	LpKY97	FH, CD156 & B71	2742283	2736228	40	ACGCGAATTAAAA**CCCTAACCCTA**
3	LpKY97	FH, CD156 & B71	2751680	2745625	41	**GTTAGG**TTTTAATTCGCG
3	LpKY97	FH, CD156 & B71	2787030	2780975	31	**GGTTGGGG**TTTTAATTCGCG
3	LPKY97	FH & *CD156	7355697	59179	114	**TTAGGG**TTTTAATTCGCGC
3	FH	LpKY97, Arcadia & Bm88324	169997	161496	40	**GGGTTAGGG**TTTTAATTCGCG
3	FH	LpKY97, CD156 & B71	2553536	2545035	41	**GGGTTAGGG**TTTTAATTCGCG
3	FH	LpKY97, CD156 & B71	2749138	2740637	40	CGCGAATTAAAA**CCCTAACCCTA**

Chr	Strain	Shared with	3' end position	Distance from Nearest Telomere or Terminal MoTeR	Relic length	3' end sequence
3	FH	LpKY97, CD156 & B71	2758545	2750044	41	**GTTAGGG**TTTTAATTCGCG
3	FH	LpKY97, CD156 & B71	2793917	2785416	40	**GGTTGGGG**TTTTAATTCGCG
3	FH	LpKY97, CD156 & B71	7315089	57264	119	**TTAGGG**TTTTAATTCGCG
3	CD156	LpKY97, FH & B71	2424482	2424482	40	**AGGGTTAGGG**TTTTAATTCGCG
3	CD156	LpKY97, FH & B71	2621949	2621949	40	CGCGAATTAAAA**CCCTAACCCTA**
3	CD156	LpKY97, FH & B71	2625549	2625549	41	**GTTAGGG**TTTTAATTCGCG
3	CD156	LpKY97, FH & B71	2660716	2660716	31	**GGTTGGGG**TTTTAATTCGCG
3	CD156	LpKY97 & FH	7124927	426559	*15	**TTAGGG**TTTAATTCGCT
3	B71	LpKY97, FH & CD156	2523720	2523720	41	**GTCAGGGTTAGGG**TTTTAATTCGCG
3	B71	LpKY97, FH & CD156	2719208	2719208	40	CGCGAATTAAAA**CCCTAACCCTA**

Chr	Strain	Shared with	3' end position	Distance from Nearest Telomere or Terminal MoTeR	Relic length	3' end sequence
3	B71	LpKY97, FH & CD156	2722828	2722828	41	**GTTAGGG**TTTTAATTCGCG
3	B71	LpKY97, FH & CD156	2758440	2758440	31	**GGTTGGGG**TTTTAATTCGCG
3	B71	-	7468533	738187	40	**TAGGGTTAGGG**TTTTAATTCGCG
3	Arcadia	LpKY97, FH & Bm88324	165684	165560	40	**GGGTTAGGG**TTTTAATTCGCG
3	Bm88324	LpKY97, FH & Arcadia	148531	148531	40	**GGGTTAGGG**TTTTAATTCGCG
4	Arcadia	US71	15722	15312	239	CGCGAATTAAAA**CCCTAACCCTAA**
4	Bm88324	-	5007462	983162	131	CGCGAATTAAAA**CCCTAA**
4	US71	-	7328	7328	40	**TTAGGG**TTTTAATTCGCG
4	US71	Arcadia	51572	51572	241	CGCGAATTAAAA**CCCTAAC**
4	US71	-	51745	51745	165	CGCGAATTAAAA**CCCTAACCCTAA**

Chr	Strain	Shared with	3' end position	Distance from Nearest Telomere or Terminal MoTeR	Relic length	3' end sequence
5	LpKY97	FH	4395439	1701	147	CGCGAATTAAAA**CCCTAACCC**
5	FH	LpKY97	4423681	2139	149	CGCGAATTAAAA**CCCTAACCC**
5	Guy11	70-15	185204	185079	140	**TTATGGTGAGGG**TTTTAATTCGCG
5	Arcadia	-	122817	122697	105	CGCGAGTTAGAA**CCCTAACCCTAA**
5	Arcadia	US71	4700151	120119	166	**AGGGTTAGGG**TTTTAATTCGCG
5	US71	-	4630653	245799	370	**TTAGGGTTAGGG**TTTTAATTCGCG
5	US71	Arcadia	4756255	120197	1539	CGCGAATTAAAA**CCCTAACCCT**
5	70-15	Guy11	102769	102769	141	**TTATGGTGAGGG**TTTTAATTCGCG
6	LpKY97	FH, CD156 & B71	83149	78782	31	CGCGAATTAAAA**CCCTAA**
6	LpKY97	FH, CD156 & B71	2132281	2127914	40	**GTTGGGG**TTTTAATTCGCG

Chr	Strain	Shared with	3' end position	Distance from Nearest Telomere or Terminal MoTeR	Relic length	3' end sequence
6	LpKY97	FH, CD156 & B71	3271965	2851743	131	CGCGAATTAAAA**CCTAACCATCCCA**
6	FH	LpKY97, CD156 & B71	83149	70615	31	CGCGAATTAAAA**CCCTAA**
6	FH	LpKY97, CD156 & B71	2144774	2132240	39	**TTGGGG**TTTTAATTCGCG
6	FH	LpKY97, CD156 & B71	3282558	2820790	131	CGCGAATTAAAA**CCCTAA**
6	CD156	LpKY97, FH & B71	95554	77505	31	CGCGAATTAAAA**CCCTAA**
6	CD156	LpKY97, FH & B71	2153119	2135070	40	**TTGGGG**TTTTAATTCGCG
6	CD156	LpKY97, FH & B71	3282214	2779397	132	CGCGAATTAAAA**CCCTAACC**
6	B71	LpKY97, FH & CD156	78239	71228	31	CGCGAATTAAAA**CCCTAA**
6	B71	LpKY97, FH & CD156	2138379	2138379	40	**GTTGGGG**TTTTAATTCGCG
6	B71	LpKY97, FH & CD156	3275257	2808416	135	CGCGAATTAAAA**CCCTAACC**

Chr	Strain	Shared with	3' end position	Distance from Nearest Telomere or Terminal MoTeR	Relic length	3' end sequence
6	Arcadia	US71	22696	22696	1146	CGCGAATTAAAA**CCCTAACCCTA**
6	US71	Arcadia	42833	42833	1146	CGCGAATTAAAA**CCCTAACCCTA**
7	LpKY97	FH, CD156, Bm88324 & B71	3688512	161862	38	CGCGAATTAAAA**CCTTAACCCTAA**
7	LpKY97	FH & CD156	3825400	24974	63	**TTAGGTGTGGG**TTTTAAATCGCG
7	LpKY97	FH & CD156	3835928	14446	110	CGCGAATTGAAA**CCCTAACCCTAA**
7	FH	LpKY97, CD156, Bm88324 & B71	3694340	154907	37	CGCGAATTAAAA**CCTTAACCCTAA**
7	FH	LpKY97 & CD156	3836980	12267	63	**TTAGGGTGTGGG**TTTTAAATCGCG
7	FH	LpKY97 & CD156	3841696	7551	116	CGCGAATTGAAA**CCCTAACCCTAA**
7	CD156	LpKY97, FH, Bm88324 & B71	3761013	143103	38	CGCGAATTAAAA**CCTTAACCCTAA**
7	CD156	LpKY97 & FH	3893775	10341	63	**TTAGGGTGTGGG**TTTTAAATCGCG

Chr	Strain	Shared with	3' end position	Distance from Nearest Telomere or Terminal MoTeR	Relic length	3' end sequence
7	CD156	LpKY97 & FH	3896604	7512	105	CGCGAATTGAAA**CCCTAACCCTAA**
7	B71	Bm88324	3755038	280948	59	**GGAAGTTAAGGTTAGG**TTTTAAATTCGCG
7	B71	LpKY97, FH, Bm88324 & CD156	3755790	280196	59	CGCGAATTAAAA**CCTTAACCCTAA**
7	Arcadia	-	2960437	541093	132	CGCGAATTAAAA**CCCTAA**
7	Arcadia	US71	2820959	680571	40	CGCGAATTAAAA**CCCTAA**
7	Bm88324	-	560250	560250	94	**AGGGTTAGGG**TTTTAATTCGCG
7	Bm88324	B71	4570424	191128	112	**GTTAAGGTTAGG**TTTTAAATTCGCG
7	Bm88324	B71, LpKY97, FH, & CD156	4571156	190396	124	CGCGAATTAAAA**CCTTAAACCCAAA**
7	US71	Arcadia	4051648	107847	40	CGCGAATTAAAA**CCCTAA**

Supplementary Table 3 Telomere assembly and terminal MoTeR presence/absence and position. Chromosome ends of each strain were examined for the presence/absence (Y/N) of terminal MoTeR sequences and fully assembled telomere sequences. Pos_start represents the left-side of the chromosome while pos_end represents the right-side of the chromosome. The purpose of this information was to calculate the distance of individual MoTeR relics to the nearest terminal MoTeR or telomere sequence. If a MoTeR was present in the telomere, then the position of the last MoTeR leading into the subtelomere was reported. If there was no MoTeR present within the telomere, then the position recorded was the end of the telomere at the edge of the subtelomere. If the telomere itself was missing from the assembly the position was defined as the first or last nucleotide in the chromosome. Chromosome ends with a single, or partial, telomere repeat ($^{5'}$**CCCTAA**$^{3'}$;$^{3'}$**TTAGGG**$^{5'}$) were not considered fully assembled telomeres.

Strain	Chromosome	Terminal MoTeR pos_start?	pos_start	Fully assembled telomere pos_start?	Terminal MoTeR pos_end?	pos_end	Fully assembled telomere pos_end?
CD156	Chr1	N	0	N	Y	6008336	Y
CD156	Chr2	N	0	N	N	8265001	N
CD156	Chr3	N	0	N	Y	7551486	Y
CD156	Chr4	Y	9778	Y	Y	5494330	Y
CD156	Chr5	N	0	Y	Y	4604659	Y
CD156	Chr6	Y	18049	Y	Y	6061611	Y
CD156	Chr7	Y	8843	Y	Y	3904116	Y
FH	Chr1	N	0	N	Y	6243661	Y

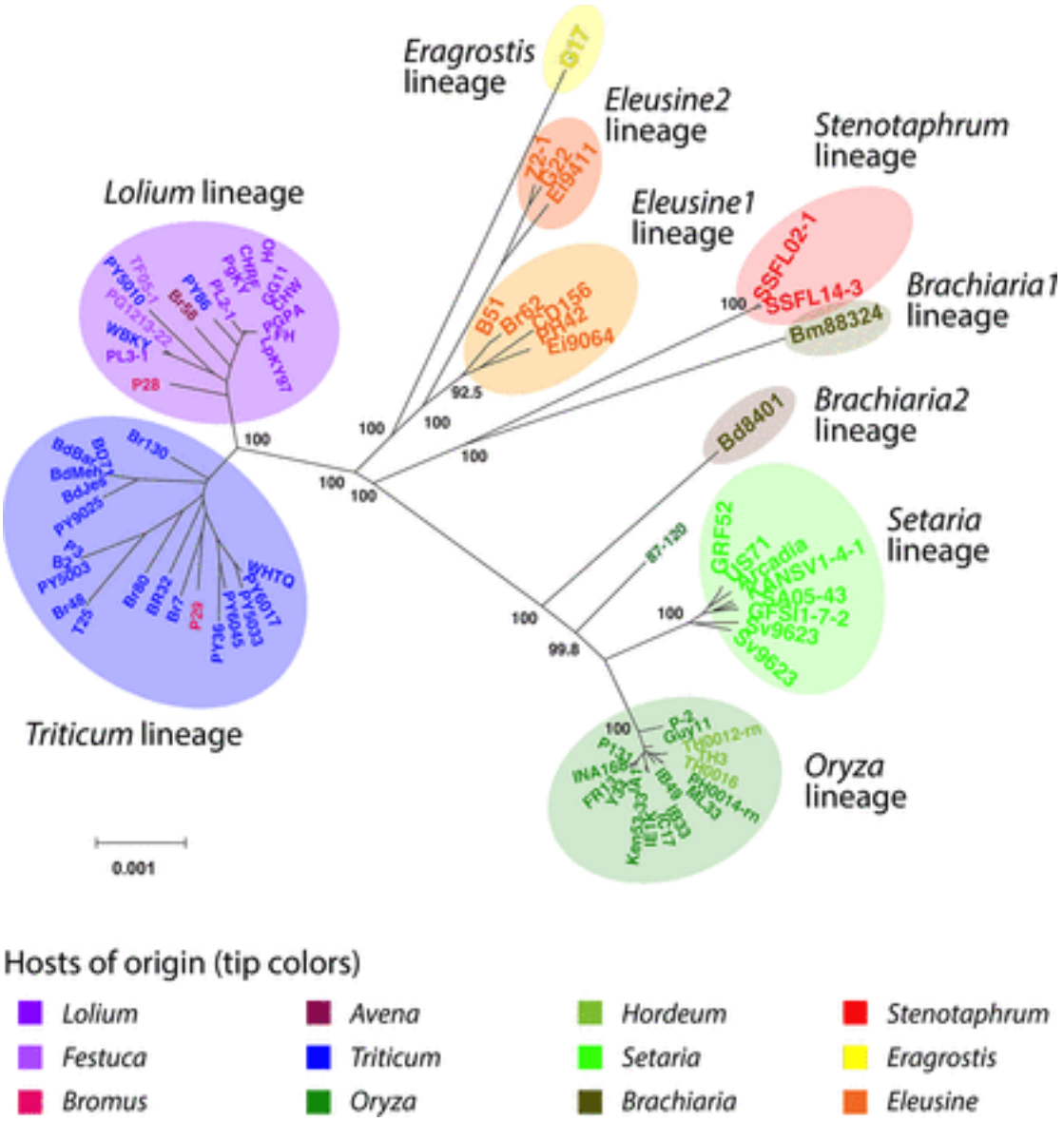
Strain	Chromosome	Terminal MoTeR pos_start?	pos_start	Fully assembled telomere pos_start?	Terminal MoTeR pos_end?	pos_end	Fully assembled telomere pos_end?
FH	Chr2	Y	11490	Y	N	8522316	Y
FH	Chr3	Y	8501	Y	Y	7372353	Y
FH	Chr4	Y	11502	Y	Y	5445612	Y
FH	Chr5	Y	21432	Y	Y	4425820	Y
FH	Chr6	Y	12534	Y	Y	6103348	Y
FH	Chr7	Y	5142	Y	Y	3849247	Y
FH	MinChr1_V3	Y	3461	Y	Y	1828878	Y
Arcadia	Chr1	N	111	Y	N	5987355	Y
Arcadia	Chr2	N	0	N	N	8805158	Y
Arcadia	Chr3	N	124	Y	Y	6770325	Y
Arcadia	Chr4	Y	410	Y	N	7569745	Y
Arcadia	Chr5	N	120	Y	N	4820270	Y
Arcadia	Chr6	N	0	N	N	6337561	N
Arcadia	Chr7	N	0	N	N	3501530	Y
Guy11	Chr1	N	0	N	N	5788060	N
Guy11	Chr2	N	101	Y	N	8590670	Y
Guy11	Chr3	N	192	Y	N	6810598	Y

Strain	Chromosome	Terminal MoTeR pos_start?	pos_start	Fully assembled telomere pos_start?	Terminal MoTeR pos_end?	pos_end	Fully assembled telomere pos_end?
Guy11	Chr4	N	0	N	N	5685480	Y
Guy11	Chr5	N	125	Y	N	4603878	N
Guy11	Chr6	N	0	N	N	7294679	Y
Guy11	Chr7	N	110	Y	N	4182135	Y
LpKY	Chr1	N	176	Y	Y	6326480	Y
LpKY	Chr2	Y	13473	Y	N	7953596	Y
LpKY	Chr3	Y	6055	Y	Y	7414876	Y
LpKY	Chr4	Y	11475	Y	N	5421679	Y
LpKY	Chr5	Y	5260	Y	Y	4397140	Y
LpKY	Chr6	Y	4367	Y	Y	6123708	Y
LpKY	Chr7	N	202	Y	N	3850374	Y
LpKY	MiniChr1	Y	13431	Y	Y	2980065	Y
LpKY	MiniChr2	Y	6577	Y	Y	860985	Y
U233	Chr1	N	5	N	N	5500179	N
U233	Chr2	N	132	Y	Y	7852337	Y
U233	Chr3	N	5	N	Y	7954589	Y
U233	Chr4	N	170	Y	N	7641527	Y

Strain	Chromosome	Terminal MoTeR pos_start?	pos_start	Fully assembled telomere pos_start?	Terminal MoTeR pos_end?	pos_end	Fully assembled telomere pos_end?
U233	Chr5	N	157	Y	Y	5438331	Y
U233	Chr6	N	166	Y	N	3716274	Y
U233	Chr7	Y	323	Y	N	1631780	N
B71	Chr1	N	0	N	Y	6436659	Y
B71	Chr2	Y	6451	N	Y	7898963	N
B71	Chr3	N	0	N	N	8206720	N
B71	Chr4	Y	15215	Y	Y	5392353	N
B71	Chr5	Y	5056	Y	Y	4434190	Y
B71	Chr6	Y	7011	N	Y	6083673	N
B71	Chr7	Y	12722	N	Y	4035986	N
70-15	Chr1	N	0	N	N	7978600	N
70-15	Chr2	N	0	Y	N	8319960	Y
70-15	Chr3	N	0	Y	N	6606598	N
70-15	Chr4	N	0	Y	N	5546960	Y
70-15	Chr5	N	0	Y	N	4490059	Y
70-15	Chr6	N	0	Y	N	4133993	Y
70-15	Chr7	N	0	Y	N	3415785	Y

Strain	Chromosome	Terminal MoTeR pos_start?	pos_start	Fully assembled telomere pos_start?	Terminal MoTeR pos_end?	pos_end	Fully assembled telomere pos_end?
Bm88324	Chr1	N	0	N	N	5302774	Y
Bm88324	Chr2	N	0	N	N	7716396	Y
Bm88324	Chr3	N	0	Y	N	6380525	N
Bm88324	Chr4	N	0	N	N	5990624	Y
Bm88324	Chr5	N	0	N	N	4768407	Y
Bm88324	Chr6	N	0	Y	N	5423066	N
Bm88324	Chr7	N	0	Y	N	4761552	N
US71	Chr1	N	0	N	N	6222700	Y
US71	Chr2	N	0	Y	N	8539580	N
US71	Chr3	N	0	Y	N	6862639	N
US71	Chr4	N	0	N	N	6008318	Y
US71	Chr5	N	0	N	N	4876452	N
US71	Chr6	N	0	Y	N	6422269	Y

Strain	Chromosome	Terminal MoTeR pos_start?	pos_start	Fully assembled telomere pos_start?	Terminal MoTeR pos_end?	pos_end	Fully assembled telomere pos_end?
US71	Chr7	N	0	Y	N	4159495	N



Supplementary Figure 1 Neighbor-joining distance tree based off number of SNPs per kb in repeat-masked genome assemblies. Colored circles are used to highlight Magnaporthe oryzae lineages. Labeled nodes have >80% confidence. Adapted from Gladieux et al. 2018.

Supplementary Table 4 MoTeR relic positions and their associated duplications. The positions for the duplicate sequences adjacent to the relics are listed first. The chromosome and position that each duplicate sequence maps to is noted with an *. 3' duplications are duplicate sequences that are found flanking the 3' sequence of MoTeR relics. Relic + 5' duplications are duplicate sequences that contain both the relic itself and its flanking 5' sequence.

Chr	Strain	Relic 3' position	Duplication type	Start (5')	End (3')	Length	Chr*	Start (5')*	End (3')*
Chr1	CD156	215205	3'	215231	216153	922	Chr1	82487	83397
Chr1	CD156	753727	----------	----------	----------	----------	---------	----------	----------
Chr1	CD156	5749080	3'	5749041	5749080	242	Chr3	2488147	2488390
Chr1	CD156	5749080	Relic + 5'	5748567	5749085	518	Chr6	95111	95570
Chr2	CD156	7550702	Relic + 5'	7550142	7550719	577	Chr6	94948	95571
Chr2	CD156	7550702	Relic + 5'	7550142	7550720	578	Chr3	2660698	2661270
Chr2	CD156	7551004	3'	7551022	7552228	1206	Chr6	6024169	6025375

Chr	Strain	Relic 3' position	Duplication type	Start (5')	End (3')	Length	Chr*	Start (5')*	End (3')*
Chr2	CD156	7553167	----------	----------	----------	----------	---------	----------	----------
Chr3	CD156	2424482	3'	2423615	2424461	849	Chr5	4504267	4505293
Chr3	CD156	2621949	3'	2621959	2622189	230	Chr1	1600219	1600452
Chr3	CD156	2625538	3'	2622713	2625525	2812	Chr5	191310	204348
Chr3	CD156	2660705	3'	2658586	2660695	2230	Chr5	213434	215664
Chr6	CD156	95565	Relic + 5'	94948	95571	623	Chr2	7550142	7550719
Chr6	CD156	2153119	----------	----------	----------	----------	---------	----------	----------
Chr6	CD156	3282214	----------	----------	----------	----------	---------	----------	----------
Chr7	CD156	3761013	----------	----------	----------	----------	---------	----------	----------
Chr7	CD156	3896608	----------	----------	----------	----------	---------	----------	----------

Chr	Strain	Relic 3' position	Duplication type	Start (5')	End (3')	Length	Chr*	Start (5')*	End (3')*
Chr2	FH	645364	**3'**	645105	645357	252	**Chr3**	2737140	2737390
Chr3	FH	169997	----------	----------	----------	---------	----------	----------	----------
Chr3	FH	2553539	**3'**	2552748	2553514	766	**Chr5**	4314329	4315093
Chr3	FH	2749138	**5'**	2748956	2749089	**133**	**Chr6**	83016	83152
Chr3	FH	2749138	**5'**	2748981	2749092	**111**	**Chr1**	345699	345812
Chr3	FH	2749138	**5'**	2748981	2749090	**109**	**Chr6**	2144767	2144876
Chr3	FH	2758548	**3'**	2758162	2758535	**373**	**Chr5**	213344	213716

Chr	Strain	Relic 3' position	Duplication type	Start (5')	End (3')	Length	Chr*	Start (5')*	End (3')*
Chr3	FH	2758548	**3'**	2758162	2758519	**357**	**MinChr1**	1069191	1069546
Chr3	FH	2758548	**5'**	2758604	2758649	**45**	Chr2	645093	645138
Chr3	FH	2758548	**5'**	2758605	2758783	**178**	**Chr6**	2144589	2144767
Chr3	FH	2793917	**Relic+5'**	2793911	2794087	**176**	**Chr6**	82971	83152
Chr3	FH	2793917	**3'**	2793783	2793907	**124**	**Chr3**	224782	224907
Chr3	FH	7315089	**3'**	7315218	7315452	234	**Chr7**	3851320	3851555
Chr3	FH	7315089	**3'**	7315218	7315462	**244**	**Chr7**	4812	5058

Chr	Strain	Relic 3' position	Duplication type	Start (5')	End (3')	Length	Chr*	Start (5')*	End (3')*
Chr3	FH	7315089	**3'**	7315233	7315452	219	**Chr5**	4425912	7315452
Chr3	FH	7315089	**3'**	7315233	7315441	208	**Chr4**	5445680	5445886
Chr3	FH	7315089	**3'**	7315233	7315275	42	**Chr6**	6103302	6103344
Chr5	FH	4423681	----------	----------	----------	---------	----------	----------	----------
Chr6	FH	83146	**Relic+5'**	82971	83149	**178**	**Chr3**	2793914	2794087
Chr6	FH	83146	**Relic+5'**	83043	83149	**106**	Chr6	2144771	2144875
Chr6	FH	83146	**Relic+5'**	83043	83148	**105**	**Chr1**	345706	345811

Chr	Strain	Relic 3' position	Duplication type	Start (5')	End (3')	Length	Chr*	Start (5')*	End (3')*
Chr6	FH	2144774	**Relic+5'**	2144772	2144772	**105**	**Chr1**	345706	345813
Chr6	FH	2144774	**Relic+5'**	2144767	2144876	**109**	**Chr3**	2793910	2794019
Chr6	FH	2144774	**Relic+5'**	2144768	2144875	**107**	**Chr6**	83043	83152
Chr6	FH	3282558	----------	----------	----------	---------	----------	----------	----------
Chr7	FH	3694340	----------	----------	----------	---------	----------	----------	----------
Chr7	FH	3836980	----------	----------	----------	---------	----------	----------	----------
Chr7	FH	3841696	----------	----------	----------	---------	----------	----------	----------
Chr2	LpKY	25940	3'	25959	(72754??)	126 or (1299	Chr6	5728681	5737809???

Chr	Strain	Relic 3' position	Duplication type	Start (5')	End (3')	Length	Chr*	Start (5')*	End (3')*
Chr2	LpKY	25940	5'	23390	25821	2431	MiniChr1	2933499	2935944
Chr2	LpKY	25940	5'	23390	25821	2431	Chr1	1077981	1080427
Chr2	LpKY	25940	5'	24796	25821	1025	MiniChr1	2531374	2532411
Chr2	LpKY	27585	Whole locus	27585	29145	1569	MiniChr2	680023	681615
Chr2	LpKY	42376	----------	----------	----------	---------	----------	----------	----------
Chr3	LpKY	152722	5'	152774	152934	160	Chr3	2672245	2672405
Chr3	LpKY	2548296	----------	----------	----------	---------	----------	----------	----------
Chr3	LpKY	2742283	5'	2742138	2742247	109	Chr6	2132274	2132383
Chr3	LpKY	2742283	3'	2742311	2742536	225	Chr1	2065426	2065654
Chr3	LpKY	2751680	5'	2751737	2751915	178	Chr6	2132096	2132274

Chr	Strain	Relic 3' position	Duplication type	Start (5')	End (3')	Length	Chr*	Start (5')*	End (3')*
Chr3	LpKY	2787030	Relic+5'	2787027	2787200	**652	Chr6	82985	82333
Chr3	LpKY	7355697	----------	----------	----------	---------	----------	----------	----------
Chr5	LpKY	4395439	----------	----------	----------	---------	----------	----------	----------
Chr6	LpKY	83149	Relic+5'	83058	83164	106	Chr6	2132278	2132382
Chr6	LpKY	83149	Relic+5'	82985	83164	179	Chr3	2787027	2787200
Chr6	LpKY	2132281	Relic+5'	2132274	2132383	109	Chr3	2787023	2787132
Chr6	LpKY	2132281	Relic+5'	2132278	2132382	104	Chr6	83058	83164
Chr6	LpKY	2132281	3'	2132096	2132274	178	Chr3	2751737	2751915
Chr6	LpKY	3271965	----------	----------	----------	---------	----------	----------	----------
Chr7	LpKY	3688512	----------	----------	----------	---------	----------	----------	----------

Chr	Strain	Relic 3' position	Duplication type	Start (5')	End (3')	Length	Chr*	Start (5')*	End (3')*
Chr7	LpKY	3825400	----------	----------	----------	---------	----------	----------	----------
Chr7	LpKY	3835928	----------	----------	----------	---------	----------	----------	----------
MiniChr1	LpKY	398360	----------	----------	----------	---------	----------	----------	----------
MiniChr1	LpKY	1359742	BIR	1357639	1360022	2383	MiniChr1	1360023	1362409
MiniChr1	LpKY	1362129	BIR	1360023	1362409	2386	MiniChr1	1362410	1364797
MiniChr1	LpKY	1364517	BIR	1362410	1364797	2387	MiniChr1	1264798	1367192
MiniChr1	LpKY	1366906	BIR	1264798	1367192	2395	----------	----------	----------
MiniChr1	LpKY	2383001	Whole locus	2380800	2386022	5222	MiniChr1	138552	143705
MiniChr1	LpKY	2700125	Relic + 5'	2700110	2700214	104	MiniChr1	2806582	2806687
MiniChr1	LpKY	2783081	Relic+5'	2783082	2783425	343	MiniChr1	2806592	2806931

Chr	Strain	Relic 3' position	Duplication type	Start (5')	End (3')	Length	Chr*	Start (5')*	End (3')*
MiniChr1	LpKY	2806591	Relic+5'	2806582	2806687	105	MiniChr1	2700110	2700214
MiniChr1	LpKY	2806591	Relic+5'	2806592	2806931	339	MiniChr1	2783082	2783425
Chr1	B71	6127382	**Relic + 5'**	6127375	6127903	**336**	**Chr6**	77778	78244
Chr1	B71	6127382	**3'**	6127906	6128151	**245**	**Chr3**	2587832	2588077
Chr2	B71	39554	**3'**	39295	39547	**252**	**Chr3**	2707333	2717585
Chr3	B71	2523720	**3'**	2522904	2523688	**784**	**Chr5**	4325026	4325809
Chr3	B71	2719426	**3'**	2719436	2719668	**232**	**Chr1**	2062719	2062952
Chr3	B71	2723035	**5'**	2723103	2723643	**540**	**Scaf2**	90691	91231

Chr	Strain	Relic 3' position	Duplication type	Start (5')	End (3')	Length	Chr*	Start (5')*	End (3')*
Chr3	B71	2758440	---	---	---	---	---	---	---
Chr3	B71	7468533	3'	7468299	7468938	**639**	**Chr3**	7943524	7944163
Chr6	B71	78239	**Relic + 5'**	77902	78244	**342**	**Chr1**	6127567	6127903
Chr6	B71	2138958	**5'**	2139056	2139212	**156**	**Scaf2**	739014	739167
Chr6	B71	2138958	**5'**	2139056	2139212	**156**	**Chr1**	6126568	6126724
Chr6	B71	3275257	---	---	---	---	---	---	---
Chr7	B71	3754755	---	---	---	---	---	---	---

Chr	Strain	Relic 3' position	Duplication type	Start (5')	End (3')	Length	Chr*	Start (5')*	End (3')*
Chr7	B71	3755518	---	---	---	---	---	---	---
Scaf2	B71	309325	**3'**	309099	309319	**220**	**Chr3**	7859255	7859473
Scaf2	B71	588650	**3'**	588657	592197	**3540**	**Chr1**	1690306	1693831
Scaf2	B71	588650	**3'**	588657	592197	**3540**	**Chr3**	1960673	1964198
Chr1	Arcadia	5692246	---	---	---	---	---	---	---
Chr1	Arcadia	5425062	---	---	---	---	---	---	---
Chr1	Arcadia	5374231	---	---	---	---	---	---	---

Chr	Strain	Relic 3' position	Duplication type	Start (5')	End (3')	Length	Chr*	Start (5')*	End (3')*
Chr2	Arcadia	11850	---------	----------	----------	---------	--------	----------	----------
Chr3	Arcadia	165684	5'	165879	165729	150	Chr1	60594	60444
Chr4	Arcadia	15722	---------	----------	----------	---------	--------	----------	----------
Chr5	Arcadia	122817	---------	----------	----------	---------	--------	----------	----------
Chr5	Arcadia	4700151	---------	----------	----------	---------	--------	----------	----------
Chr6	Arcadia	22696	3'	22706	22936	230	Chr3	5587685	5587917
Chr7	Arcadia	2960437	5'	2959707	2960308	601	Chr1	1208662	1209263

Chr	Strain	Relic 3' position	Duplication type	Start (5')	End (3')	Length	Chr*	Start (5')*	End (3')*
Chr1	US71	158743	Whole locus	149010	172995	23985	Chr1	303813	327767
Chr1	US71	313516	Whole locus	303813	327767	23954	Chr1	149010	172995
Chr1	US71	5599609	---	---	---	---	---	---	---
Chr1	US71	5650676	---	---	---	---	---	---	---
Chr1	US71	5918660	---	---	---	---	---	---	---
Chr2	US71	3026	---	---	---	---	---	---	---
Chr4	US71	7328	Whole locus	1	16824	16823	Chr7	4042072	4059008
Chr4	US71	51572	---	---	---	---	---	---	---
Chr4	US71	51745	---	---	---	---	---	---	---
Chr5	US71	4630653	---	---	---	---	---	---	---

Chr	Strain	Relic 3' position	Duplication type	Start (5')	End (3')	Length	Chr*	Start (5')*	End (3')*
Chr2	70-15	35652	----------	----------	----------	---------	----------	----------	----------
Chr2	70-15	36813	----------	----------	----------	---------	----------	----------	----------
Chr5	70-15	102769	----------	----------	----------	---------	----------	----------	----------
Chr2	Bm88324	3176	----------	----------	----------	----------	---------	----------	----------
Chr2	Bm88324	4311	----------	----------	----------	----------	---------	----------	----------
Chr3	Bm88324	148531	----------	----------	----------	----------	---------	----------	----------
Chr4	Bm88324	5007462	----------	----------	----------	----------	---------	----------	----------
Chr7	Bm88324	560250	----------	----------	----------	----------	---------	----------	----------
Chr7	Bm88324	4570424	----------	----------	----------	----------	---------	----------	----------
Chr7	Bm88324	4571156	----------	----------	----------	----------	---------	----------	----------

Chr	Strain	Relic 3' position	Duplication type	Start (5')	End (3')	Length	Chr*	Start (5')*	End (3')*
Chr1	U233	4947590	----------	----------	----------	----------	---------	----------	----------

APPENDICES

Appendix 1: *Unix Command Line*

List of commands:

>grep	>sort	>head	>cd	>tail
>egrep	>sed	>ls	>mkdir	>less
>awk	>cat	>pwd	>man	>nano

Installing local BLAST on the command line:

https://www.ncbi.nlm.nih.gov/books/NBK569861/

- Download .fasta files for query and subject sequences in your working directory
- Make local databases for each .fasta subject file
    - For example: Making a database for the Guy11 genome
        - > makeblastdb -in Guy11.fasta -out Guy11_db -dbtype nucl
            - "-in _____" represents the sequence file you will be using to make a database
            - "-out _____" is the name you are giving your database
            - "-dbtype _____" is the type of data within your database, in this case it's made up of nucleotide

- Once you've made your database you can start using BLASTn to search for matches to your query sequences within the database
    - For example: When looking for MoTeR relics one method I used was blasting full MoTeR1 and MoTeR2 sequences against my database for each genome
        - The query sequences needed to be in .fasta format
            - The name of each sequence should follow a ">" and comprise the first line of the text file. The nucleotides for that file will then follow. You can have multiple query sequences within a file as long as they are prefaced by a ">".

```
>MoTER1
CCCGAACCCGAACCCAA
ACCCAAACCCAAACCCAAACCCAAACCCAAACCCAAACCCAAACCCGGAG
GGTTCCCAAGTCGCCTAAACCCGAAGGGTTTAGGATATTATTTCGTTTAT
TAGAATTGGATAATTATTTACCCCTGTTGGACAGGGGGGTTGCAGGGGTT
AAATTAAGGTTTTTTATTATTTATGCGCCGTTTATTTGTTTACCCCCCCA
AATATTATAAAAGCGCGTTCCATCCTCTTAGGAAAAGCGAAGCTTTTCCT
TGTAAAAGTCGCTAGACTTTTACTATAAAAGTCGCTAGACTTTTATACCA
ATCTTTTAACAAAAAGCGTAGCTTTTTGTTGCCAATCTATTAAAAAAAGC
GGAGCTTTTTTTAACTTTTTCTTTTTTTTTTTTTTTTCTTTTTTTTTTTT
TTTTTTCTTTTTTTTTTTTTTTTTTTTTTTATATATATTATTATTATTATT
ATTAGCGGTGGGGCTATTTATGCGCTTTAATTTGTGCGGGGCTATTTATG
CGCTTTAATTTGTGCGGGGCTATTAATGCGCTTTAACTTTACAAATTTTA
TTTATGCGCTTTAATTGCTGCGGGCCTGTTAATGCGCTTTAATTTACAAA
TTTCATTAATGCGCTTTAACTTTTATATTTACTAATGCGTTATTTATATA
ATTGCTATTATTATCGTTGCTATTATTATTATTGCTATTATTATCGTTAT
TATTATTGCAATTTTATTATATAAACCCTCGTTTGTCCCTCGATTTATCC
CGTTTCTTTTCCATCCCATCGCGCGTTTTCGTAAGCTTTGGTTTTCGTAG
GATTTGCTTTCGTAGGCTTTGCTTTCGTAGGCTTTCGTCAGCTTTTACCT
GCTTTTATTTTTTCTTTTTCTTTTTATTCCCCCCCCTTTTTTTTACCTGG
```

- To search for query matches in your database (in this example I used short blastn to search for MoTeR relics)
  - >blastn -query MoTeRs.fasta -db Guy11_db -task "blastn-short" -out Guy_11_MoTeR_shortblast.txt -e-value 1e-1 -outfmt 6
    - "-query _____" the sequences you are searching for
    - "-db _____" the database you are searching for matching sequences within
    - "-task 'blastn-short'" is specific to this particular blastn search looking for MoTeR relics as some of our matches may be particularly short ****
    - "-out _____.txt" The file name for your blast results. I typically save them as a .txt file
    - "-e-value _____" Manipulating the e-value may narrow your output or expand it. However, narrowing and/or expanding to either cause you to miss important matches or receive too many aberrant matches that you don't care about. For short blastn I used an e-value of 1e-1 as some of the matches could be short (< 40 nt). A few mismatches within a short sequence can be picked up by a lower designated e-value, but are easily lost with higher e-values.
    - -"outfmt _____" The format in which your output file will be organized. All output files consist of 12 columns. For output 6 they are organized as:
      - https://www.metagenomics.wiki/tools/blast/blastn-output-format-6

- **qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore**
    1. **qseqid** – query or source (e.g., gene) sequence id
    2. **sseqid** – subject or target (e.g., reference genome) sequence id
    3. **pident** – percentage of identical matches
    4. **length** – alignment length
    5. **mismatch** – number of mismatches
    6. **gapopen** – number of gap openings
    7. **qstart** – start of alignment in query
    8. **qend** – end of alignment in query
    9. **sstart** – start of alignment in subject
    10. **send** – end of alignment in subject
    11. **evalue** – expect value
    12. **bitscore** – bitscore

- Example output:

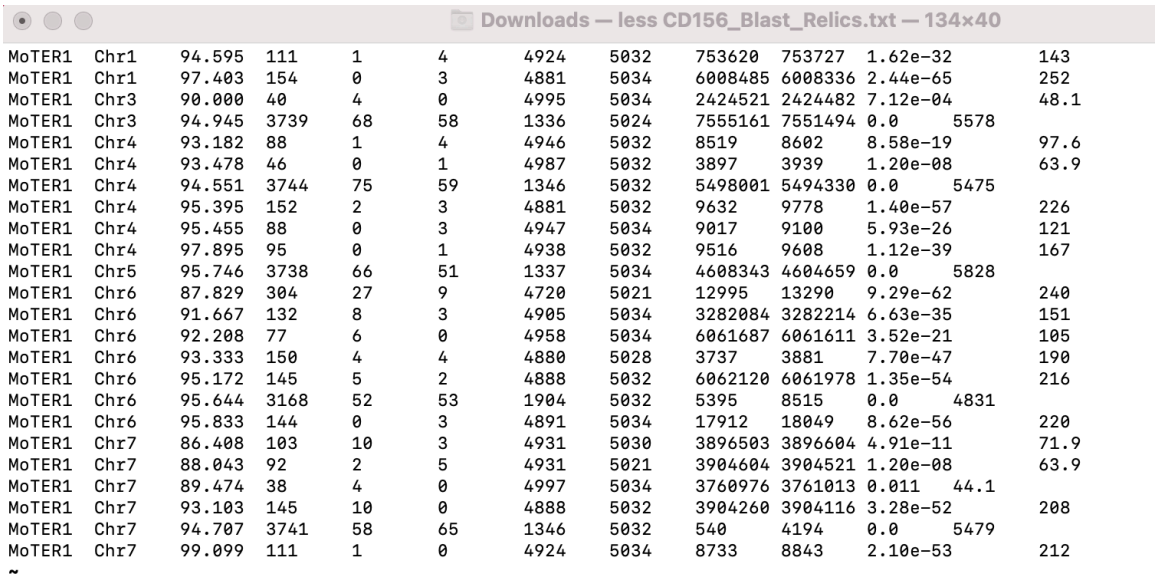
CD156_MoTeR_blast.txt

MoTER1	Chr4	83.888	571	32	21	622	1174	5498612	5498084	7.73e-84 313
MoTER1	Chr4	90.110	273	14	7	4505	4777	8608	8867	6.91e-75 283
MoTER1	Chr4	94.660	206	5	6	1050	1255	4771	4970	6.65e-72 274
MoTER1	Chr4	93.617	188	8	4	4549	4735	9142	9326	1.51e-63 246
MoTER1	Chr4	87.917	240	16	4	909	1137	100	337	2.26e-59 232
MoTER1	Chr4	95.395	152	2	3	4881	5032	9632	9778	1.40e-57 226
MoTER1	Chr4	95.522	134	5	1	908	1041	205	337	8.30e-53 210
MoTER1	Chr4	86.957	230	11	6	909	1126	5498259	5498037	1.25e-48 196
MoTER1	Chr4	95.935	123	3	2	908	1030	5498157	5498037	3.04e-46 188
MoTER1	Chr4	97.895	95	0	1	4938	5032	9516	9608	1.12e-39 167
MoTER1	Chr4	93.443	122	4	4	909	1030	4774	4891	1.04e-33 147
MoTER1	Chr4	84.549	233	17	7	957	1174	3	231	4.09e-33 145
MoTER1	Chr4	84.047	257	16	9	907	1150	4579	4823	4.09e-33 145
MoTER1	Chr4	92.800	125	5	4	1002	1126	4771	4891	4.09e-33 145

- To be able to efficiently look through this output you need to use commands that will help you filter and sort through all of the matches
    - My first criterion was finding matches that contained the 3' of MoTeR1 or MoTeR2
        - For that I could use sort to comb through the **qstart** and **qend** columns by sorting matches in that column numerically
            - > sort -k 7n CD156_MoTeR_blast.txt
                - **-k** used to designate a specific column (In this case column 7)

        - **n** used to tell the sort command it is sorting things numerically
    - OR I could use ask to search for matches that met a certain criterion
        - >awk '$7 >5000 || $8 >5000' CD156_MoTeR_blast.txt
            - **$** designates a specific column – in this case we are looking in columns 7 and 8
            - **||** is part of an OR statement. We want an output of matches in columns 7 and 8 that are greater than 5000 as the position of the 3' end in MoTeR1 is at nucleotide 5034 in the query sequence. (Position will differ for MoTeR2) I provided a buffer, so we didn't miss any potential matches.
        - Awk can be used even further to organize and filter through blast outputs and can be used in conjunction with sort (and other commands) to result in a final output. There are multiple ways to approach a dataset with these commands.

- >awk '$7 >5000 || $8 >5000' CD156_MoTeR_blast.txt | sort -k2 > CD156_Blast_Relics.txt




```
MoTER1  Chr1    94.595  111     1       4       4924    5032    753620  753727  1.62e-32        143
MoTER1  Chr1    97.403  154     0       3       4881    5034    6008485 6008336 2.44e-65        252
MoTER1  Chr3    90.000  40      4       0       4995    5034    2424521 2424482 7.12e-04        48.1
MoTER1  Chr3    94.945  3739    68      58      1336    5024    7555161 7551494 0.0     5578
MoTER1  Chr4    93.182  88      1       4       4946    5032    8519    8602    8.58e-19        97.6
MoTER1  Chr4    93.478  46      0       1       4987    5032    3897    3939    1.20e-08        63.9
MoTER1  Chr4    94.551  3744    75      59      1346    5032    5498001 5494330 0.0     5475
MoTER1  Chr4    95.395  152     2       3       4881    5032    9632    9778    1.40e-57        226
MoTER1  Chr4    95.455  88      0       3       4947    5034    9017    9100    5.93e-26        121
MoTER1  Chr4    97.895  95      0       1       4938    5032    9516    9608    1.12e-39        167
MoTER1  Chr5    95.746  3738    66      51      1337    5034    4608343 4604659 0.0     5828
MoTER1  Chr6    87.829  304     27      9       4720    5021    12995   13290   9.29e-62        240
MoTER1  Chr6    91.667  132     8       3       4905    5034    3282084 3282214 6.63e-35        151
MoTER1  Chr6    92.208  77      6       0       4958    5034    6061687 6061611 3.52e-21        105
MoTER1  Chr6    93.333  150     4       4       4880    5028    3737    3881    7.70e-47        190
MoTER1  Chr6    95.172  145     5       2       4888    5032    6062120 6061978 1.35e-54        216
MoTER1  Chr6    95.644  3168    52      53      1904    5032    5395    8515    0.0     4831
MoTER1  Chr6    95.833  144     0       3       4891    5034    17912   18049   8.62e-56        220
MoTER1  Chr7    86.408  103     10      3       4931    5030    3896503 3896604 4.91e-11        71.9
MoTER1  Chr7    88.043  92      2       5       4931    5021    3904604 3904521 1.20e-08        63.9
MoTER1  Chr7    89.474  38      4       0       4997    5034    3760976 3761013 0.011   44.1
MoTER1  Chr7    93.103  145     10      0       4888    5032    3904260 3904116 3.28e-52        208
MoTER1  Chr7    94.707  3741    58      65      1346    5032    540     4194    0.0     5479
MoTER1  Chr7    99.099  111     1       0       4924    5034    8733    8843    2.10e-53        212
~
```

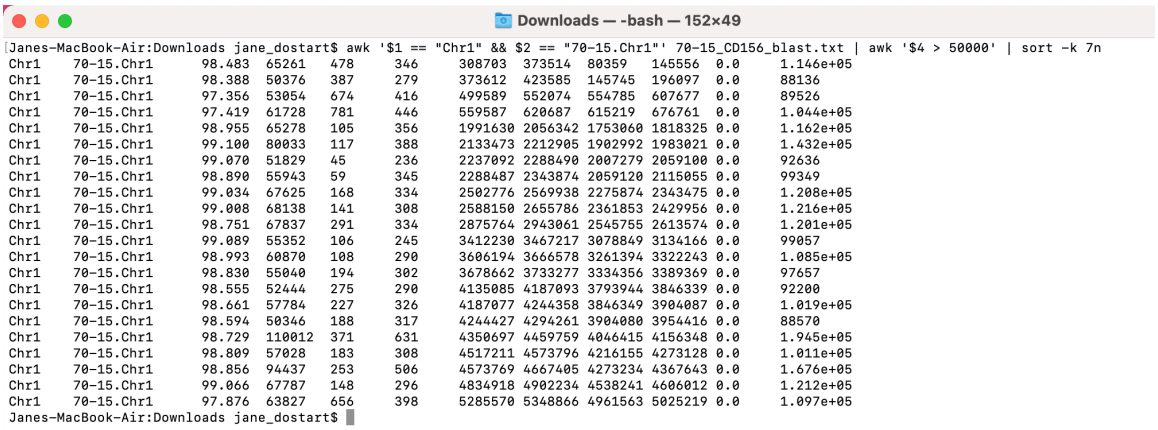
Making comparisons between strains to search for matching and unique relics was done in a similar manner. Genomes were blasted against each other using an e-value of 1e-20 and using regular BLASTn rather than short blast to look for much longer matches.

- blastn -query CD156.fasta -db 70-15_db -evalue 1e-20 outfmt 6 -out 70-15_CD156_blast.txt

After using nucleotide blast the output can be approached from a variety of ways such as using awk to look at specific positions or isolating certain chromosomes; probably the most efficient approach I've come to is known as "genome-walking".

For this I created a simple pipeline that selects the chromosomes I would like to compare using awk (in this case Chr1 and 70-15.Chr1). Then I set a criterion that the length of the match should be at least X (in this case 50000). This is purely subjective and can and should be adjusted, but it allows you to quickly look through what the largest alignments of the genomes are and whether or not the sequences and surrounding regions you are interested in align at a glance and in what orientation. Finally, I would sort one column (in this case 7) of positions initially to start "walking" from one direction along a chromosome. Following this I would always inspect the sequences in IGV, especially in the case that an alignment indicated that both genomes should have the same relic(s) but my BLASTn and grep analyses did not initially reveal this.

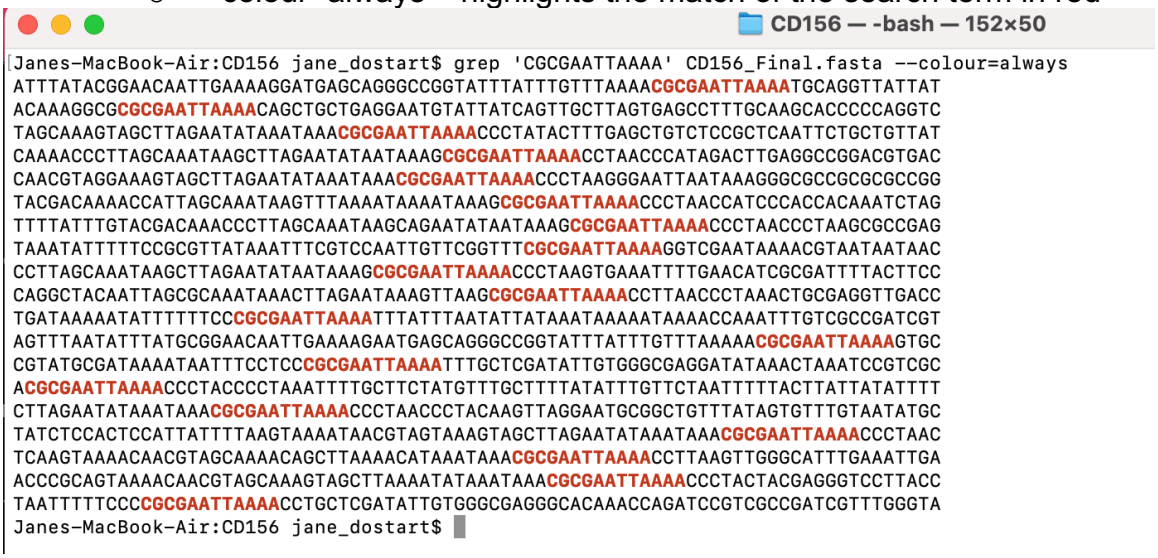
- awk '$1 == "Chr1" && $2 == "70-15.Chr1"' 70-15_CD156_blast.txt | awk '$4 > 50000' | sort -k 7n



Using grep to search for MoTeR relics

Grep is a command that can be used to look for exact matches in a file. In the cases that I used grep I would search for MoTeR 3' end sequences in .fasta files for genomes. The simplest grep search is shown below. Grep returns the search term and the line it was found on.

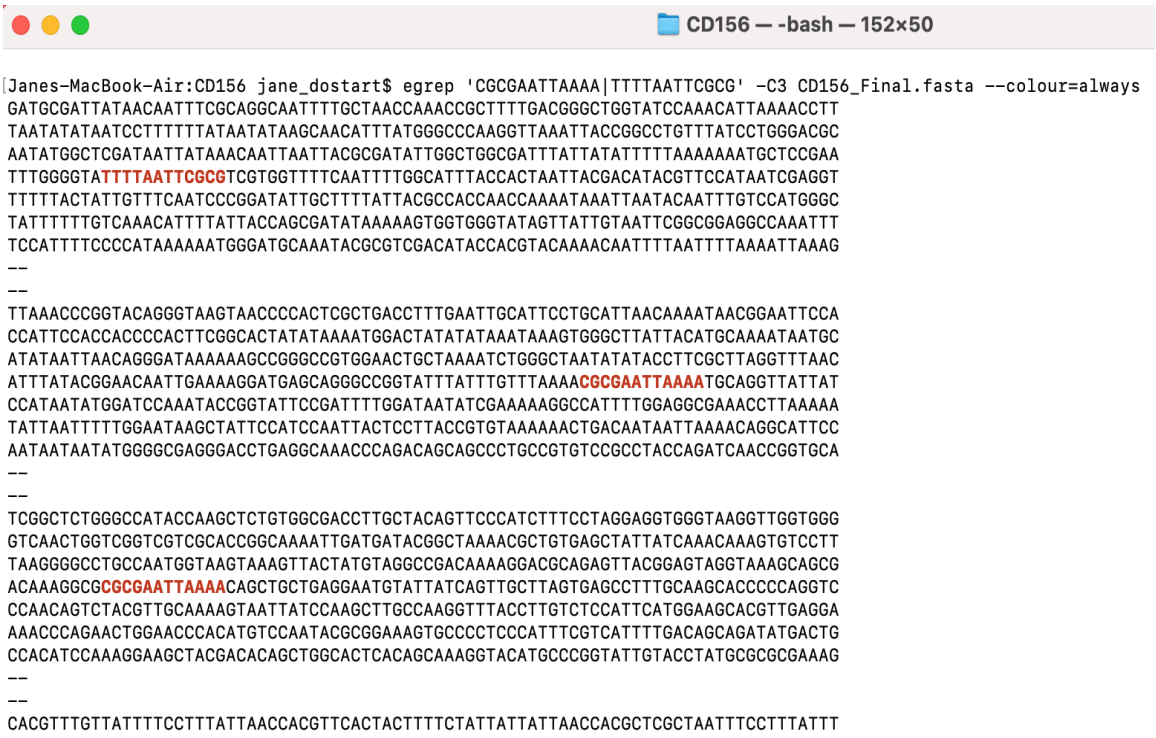
- grep 'CGCGAATTAAAA' CD156_Final.fasta –colour=always
  - '_____' – Your search term/sequence
  - _____.fasta - The file you are searching in
  - – colour=always – highlights the match of the search term in red



Because DNA is comprised of two strands which are complimentary and run antiparallel to each other, there are two sequences that make up the 3' end of MoTeRs. Using **egrep** allows you to look for multiple search terms within a file separated by a | sign.

- egrep 'CGCGAATTAAAA|TTTTAATTCGCG' CD156_Final.fasta --colour=always

Finally, with grep I needed to extend the sequences surrounding my search terms to first confirm that they were part of a MoTeR sequence and were not due to random chance. I also needed to extend the surrounding sequence to facilitate mapping them within the genome. To include lines surrounding the search term you can include the options: -A, -B, or -C with the number of lines you would like to include in the output. -A stands for after the search term, -B for before the search term, and -C includes both before and after. In the case below you can see how extending -C by 3 lines gives you a more detailed output.

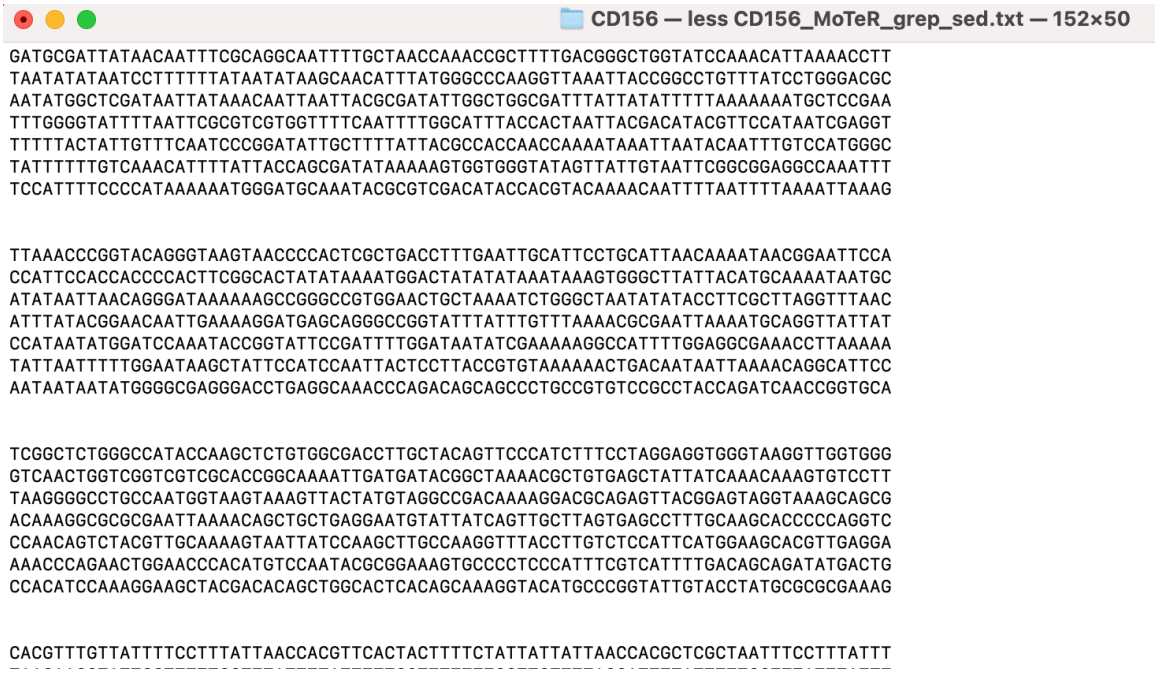


To both confirm the identity of the sequence as well as its position in the genome you need to convert your grep output to .fasta format and trim the sequences. Save your grep output as a text file (without the –colour=always option – this would change the text file format).

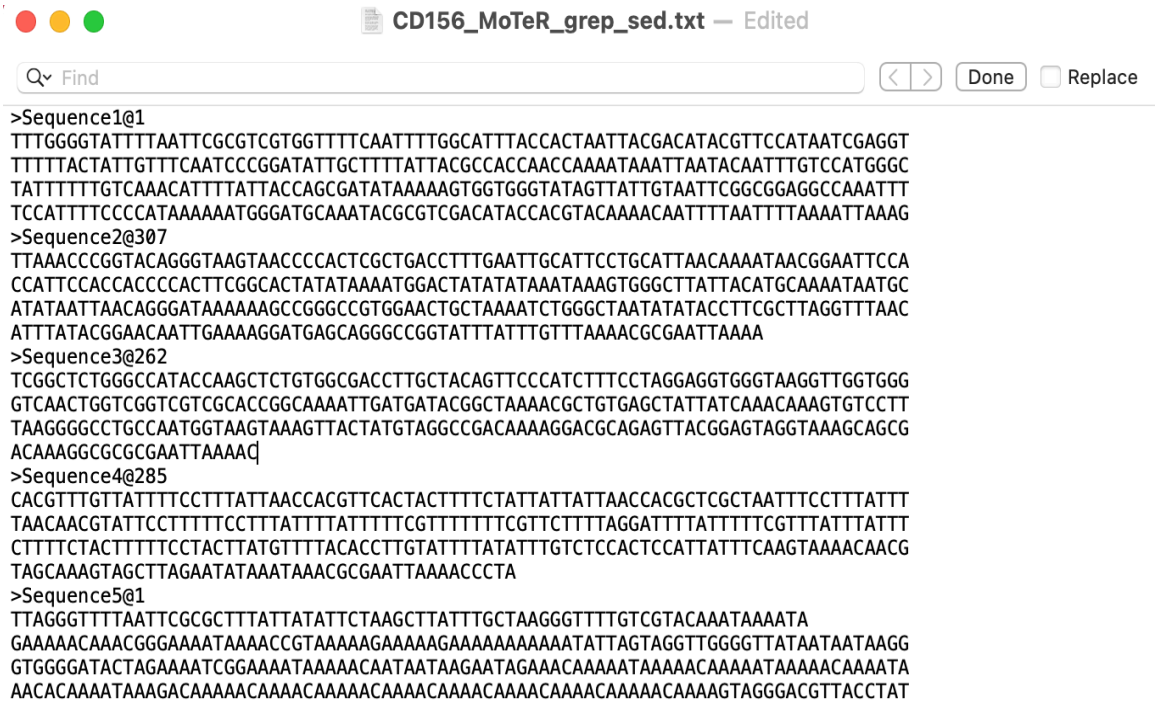
- egrep 'CGCGAATTAAAA|TTTTAATTCGCG' -C3 CD156_Final.fasta > CD156_MoTeR_grep.txt

Next you can remove the dashes in your text file by using the sed command. This command is used to substitute or replace a subject with something else in a file. In this case we will be replacing the '- -' with nothing, so we are essentially erasing it and replacing it with a space.

- sed 's/--//' CD156_MoTeR_grep.txt > CD156_MoTeR_grep_sed.txt
  - 's/__replace this/___with this/'
  - The s is an option in sed used for substituting



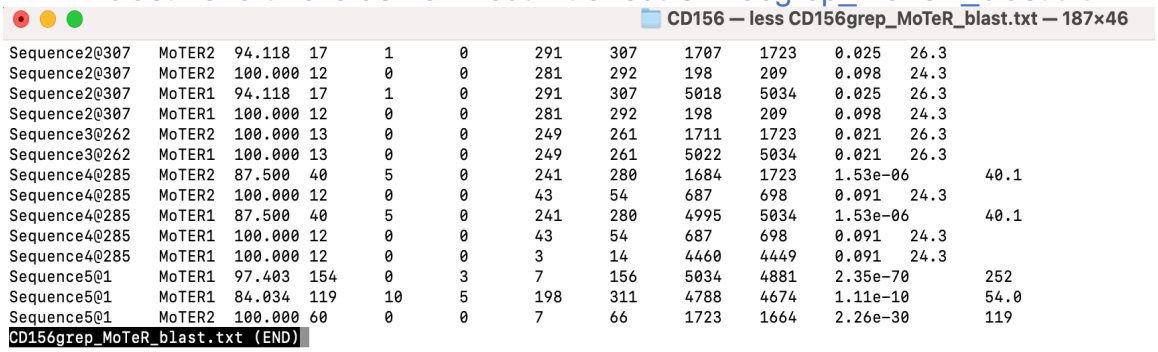
For the following steps I manually changed the file to a fasta format. For this you will need to exit the command line and work within the text file. There is likely a way to code this, but it took longer for me to try and figure out how than just doing this step manually myself. For BLAST it is easier to make comparisons between the grepped sequences and both MoTeRs 1 and 2 as well as the genome as blastn will only give you an output with the length of the match as well as the positions. To facilitate the process,-it is better to trim the sequences to include only the boundary of the 3’ sequence as well as what would be the continued MoTeR sequence in the 5’ direction. Include the position of the first or last nucleotide in the 3’ end sequence in the sequence name.



Once this has been formatted you can blast it against MoTeRs 1 and 2 as well as the genome it belongs to.

BLASTing grep sequences against MoTeR sequences (in this example I only used 5 sequences from the grep output)

- blastn -query CD156_MoTeR_grep_sed.txt -db MoTeR_database -task 'blastn-short' -evalue 1e-1 -outfmt 6 -out CD156grep_MoTeR_blast.txt

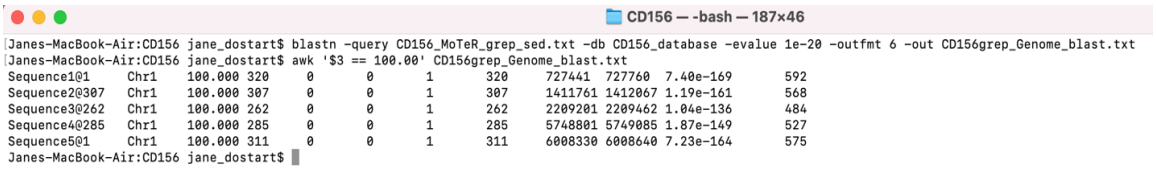


From this output you can use the position of the 3' end in the sequence to match it to the 3' end sequence of MoTeR 1 (at position 5034) and of MoTeR 2 (at position 1723). You can also filter out any hits that are particularly short. I didn't include anything less than 20 nt. Following this you can then blast the grepped sequences against the genome to find their position.

- blastn -query CD156_MoTeR_grep_sed.txt -db CD156_database -evalue 1e-20 -outfmt 6 -out CD156grep_Genome_blast.txt

To organize the output the easiest thing you can do is to awk for a 100% match in column 3 as we grepped these sequences from the genome, so they should each have an exact match. This will allow you to find the position in the genome as well as its orientation.

- awk '$3 == 100.00' CD156grep_Genome_blast.txt



After this step you should compare your grep matches that contained MoTeR relics, as confirmed by the grep-MoTeR blast, with your already recorded MoTeR relics that were captured by blasting the MoTeRs against the genome. I would then follow this by examining the sequences in IGV. In my case my .gff files for IGV did not show the MoTeR relics found using grep, so I had to designate their boundaries myself to investigate if they were associated with any rearrangements/duplications in the genome.

Appendix 2: *Graphics in R Studio*

```
###Plotting relics and their associated duplications on chromosomes in Circlize
using R Studio Version 4.0.2
## By: Jane E. Dostart

library(readxl)
library(dplyr)
library(tidyr)
library(circlize)
library(readr)


#Each color in this list corresponds to a specific chromosome. This list will
change in length
# As you account for differences in numbers of chromosomes (ie.
minichromosomes)

color2 = c("#C62828", "#EF6C00", "#F9A825", "#2E7D32", "#1565C0",
"#283593", "#6A1B9A", "#AD1457", "#E91E63")

#Read in the data for lengths of chromosomes, positions of relics, and positions
of duplicate sequences.
LpKY <- read_xlsx("LpKY_MoTeR_Duplications.xlsx", col_names = TRUE)
```

```
#This sets the first track with the corresponding length of each chromosome
df <-
  LpKY %>%
  select(Chromosome, Chr_start, Chr_end) %>%
  drop_na()
head(df)

#This reads in your duplicate sequences. Source_links and Target_links
correspond to each other and
#should be in order in each file as to the corresponding regions a duplicate
sequence is found in in
#each chromosome

source_links <-
  LpKY %>%
  select(chr = Chr_relic, start = relic_dup_start, end = relic_dup_end) %>%
  as.data.frame(table(unlist(source_links))) %>%
  drop_na()

target_links <-
  LpKY %>%
  select(chr = Chr_dup, start = o.chr_dup_start, end = o.chr_dup_end) %>%
```

```
  as.data.frame(table(unlist(target_links))) %>%
  drop_na()

#Reading in relic positions
Relics <-
  LpKY %>%
  select(Chromosome_relic, `end_position(3')`) %>%
  drop_na()

circos.clear()

#This helps to position the plot where you want it to be in the window. If part of
your figure
#is being cut off, manipulating this code (specifically your ylim) will help.
circos.par("track.height" = 0.8, gap.degree = 1.5, cell.padding = c(0,0,0,0),
          canvas.ylim = c(-1.1, 1.1), canvas.xlim = c(-1.1, 1.1))
#This actually sets up your track to plot your data onto. The lengths of your
chromosomes are now
#used to lay down the canvas or base of your plot. You won't see a figure until
the next bit of code.
circos.initialize(factors = df$Chromosome,
                  xlim = matrix (c(rep(0,9), df$Chr_end), ncol = 2))
```

```
#Visualizing the chromosomes - This creates rectangles that represent the
chromosome. You can manipulate
#interior and border color, size, and labels here.
circos.track(ylim=c(0,1),panel.fun=function(x,y) {
  chr=CELL_META$sector.index
  xlim=CELL_META$xlim
  ylim=CELL_META$ylim
  circos.text(mean(xlim),2.5,chr,cex=0.75,col="black",
            facing="bending.inside",niceFacing=TRUE, font=2)
},bg.col=color2,bg.border=F,track.height=0.11)


#Adding axis labels - I found this to be useful as it seemed the default axis labels
were in
#kilobases instead of megabases
brk <- c(0,0.5,1,1.5,2,2.5,3,3.5,4,4.5,5,5.5,6,6.5,7,7.5,8,8.5)*10^6
circos.track(track.index = get.current.track.index(), panel.fun = function(x, y) {
  circos.axis(h="top",major.at=brk,labels=round(brk/10^6,1),labels.cex=0.5,
            col="black",labels.col="black",lwd=0.7,labels.facing="clockwise")
},bg.border=F)

#Adding target and source links - genome duplications - For my figures the
duplicate sequences correspond
```

```
#specifically to the relic they were found adjacent to.
#original - circos.genomicLink(source_links, target_links, col = c("#D53E4F",
"#D53E4F", "#D53E4F", "#FEE08B", "#FEE08B", "#FEE08B", "#FEE08B",
"#99D594"))
circos.genomicLink(source_links, target_links, col =
c("#EF6C00","#EF6C00","#EF6C00","#EF6C00","#EF6C00","#F9A825","#F9A82
5","#F9A825","#F9A825","#F9A825","#283593","#283593","#283593","#283593",
"#283593”,"#AD1457","#AD1457","#AD1457","#AD1457","#AD1457",
"#AD1457", "#AD1457", "#AD1457"))

#color2 = c("#C62828", "#EF6C00", "#F9A825", "#2E7D32", "#1565C0",
"#283593", "#6A1B9A", "#AD1457", "#E91E63")

#Plotting Relics with No duplications. You could technically separate the types of
relics into different
#columns in your dataset, but where I used triangles to represent the direction of
the relics here
# "pch = 24 OR pch =25" I had to manipulate things a bit more directly, so I
entered in the position for each.
#Some of the triangles don't look great, so there may be a better way in R, or you
might use Adobe Illustrator.
#You will also find that many of the relics will overlap and be difficult to
distinguish, so manipulating
```

```
# the x coordinates allows you to visualize them. The true positions are found in
the excel file.

circos.trackPoints(factors = c("Chr2", "Chr3", "Chr3", "Chr6", "Chr7", "MiniChr1"),
cex = 0.8,
                x = c(42376, 2548296, 7355697, 3271965, 3825400, 398360),
                y = c(0.4, 0.4, 0.4, 0.4, 0.4, 0.4),
                pch = 24, bg = "white")

circos.trackPoints(factors = c("Chr5", "Chr7", "Chr7"), cex = 0.8,
                x = c(4395439, 3688512, 3835928),
                y = c(0.4, 0.4, 0.4),
                pch = 25, bg = "white")

#Plotting Relics with 5' Duplications
circos.trackPoints(factors = c("Chr3", "Chr3"), cex = 0.8,
                x = c(152722, 2751680),
                y = c(0.4, 0.4),
                pch = 24, bg = "red")

#Plotting Relic + 5' Duplications
circos.trackPoints(factors = c("Chr3", "MiniChr1", "MiniChr1", "MiniChr1"), cex =
0.8,
```

```
                x = c(2787030, 2700125, 2783081, 2806591),
                y = c(0.4, 0.4, 0.4, 0.4),
                pch = 24, bg = "grey")

circos.trackPoints(factors = c("Chr6", "Chr6"), cex = 0.8,
                x = c(83149, 2132281),
                y = c(0.4, 0.4),
                pch = 25, bg = "grey")
#Plotting relics with Whole locus duplications
circos.trackPoints(factors = "Chr3", cex = 0.8,
                x = 2787030,
                y = 0.4,
                pch = 24, bg = "purple")
circos.trackPoints(factors = "MiniChr1", cex = 0.8,
                x = 2382001,
                y = 0.4,
                pch = 25, bg = "purple")
#Plotting relics with 3' and 5' Duplications
circos.trackPoints(factors = c("Chr2", "Chr3"), cex = 0.8,
                x = c(25940, 2742283),
                y = c(0.4, 0.4),
                pch = 25, bg = "red", col = "blue")
#plotting areas with patterns of Breakage-induced-replication
```

```
circos.trackPoints(factors = c("MiniChr1", "MiniChr1", "MiniChr1", "MiniChr1"),
cex = 0.8,
                x = c(1359742, 1362129,1364517, 1366906),
                y = c(0.4, 0.4, 0.4, 0.4),
                pch = 24, bg = "green")
```